БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ

Институт по Информационни и Комуникационни Технологии

асист. маг. Дорина Петрова Кабакчиева

на Тема:

„Изследване на Data Mining модели за класификация"

Дисертация

За присъждане на образователна и научна степен

"Доктор"

Професионално направление 4.6. "Информатика и компютърни науки"

(по научна специалност 01.01.12. "Информатика")

Научен ръководител

акад. проф. д.н. Иван Попчев

София

Декември 2012г.

Съдържание

Списък на използваните термини

Brunch	При „Дърво на решенията” – клон на дървото - резултат от тестова процедура в даден възел на дървото
Business Intelligence	Бизнес интелигентност - обобщаващо понятие, включващо архитектури, инструменти, бази данни, приложения и методологии, използвани за преобразуването на данните в информация и знание, с цел подпомагане взимането на управленски решения.
Clustering	Клъстерен анализ - разделянето на данните на групи (клъстери) чрез прилагане на принципа за максимално сходство между елементите на един клъстер и минимално сходство между елементите на различни клъстери.
Confusion Matrix/ Contingency Table	Матрица на класификация – резултат от оценката на получения модел чрез приложения алгоритъм - показва броя на правилно/неправилно класифицираните обекти от тестовата извадка във всеки клас
Covering Rules	Покриващи правила – правила, които са валидни за част от обектите в дадена съвкупност от данни
Data Mining	Извличане на знания от големи обеми данни
Data Mining for Prediction	Извличане на знания от големи обеми данни чрез решаване на задачи за предсказване, при които целта е да се предсказват стойностите на една от променливите на база на известните стойности на други променливи
Data Mining for Classification	Извличане на знания от големи обеми данни чрез решаване на задачи за предсказване на цифрови променливи с дискретни стойности и на променливи от номинален тип, чрез обучаване на модели, описващи съществуващи класове от обекти, с цел използването им за определяне класа на обекти с неизвестен клас.
Data Mining for Regression	Извличане на знания от големи обеми данни чрез решаване на задачи за предсказване на цифрови променливи с непрекъснати стойности.
Directed Data Mining	Извличане на знания от големи обеми данни чрез обучение, т.е. чрез извличане на функционална взаимовръзка между предсказваната променлива и входните променливи
Educational Data Mining	Прилагане на методи и средства за извличане на знания от големи обеми данни и подпомагане решаването на проблеми на образователни институции (университети, колежи, училища, държавни агенции и др.)
Exploratory Data Analysis (EDA)	Опознаване и изследване на данните
Forward Scattering Radar	Радари гледащи напред
Heterogeneity	При „Дърво на решенията” – хетерогенност - нехомогенност

	или нееднородност на производните възли по отношение на предсказваната променлива
Homogeneity	При „Дърво на решенията” – хомогенност или еднородност на производните възли по отношение на предсказваната променлива
Leaf Node	При „Дърво на решенията” – листо на дървото – краен възел, който представлява търсения клас
Node	При „Дърво на решенията” – възел – тестова процедура за даден атрибут
Overfitting	Пренагаждане, плътно прилягане на модела към обучаващите данни
Purity	При „Дърво на решенията” – чистота или хомогенност на производните възли по отношение на предсказваната променлива
Root Node	При „Дърво на решенията” – корен на дървото – начален възел на дървото
Supervised Learning	Откриване на зависимости чрез обучение (Класификация и регресия)
Test Data	Тестова извадка - извадка от общата съвкупност от данни, използвана за оценка
Training Data	Обучаваща извадка - извадка от общата съвкупност от данни, използвана за обучение
Undirected Data Mining	Извличане на знания от големи обеми данни без обучение
Unsupervised Learning	Откриване на зависимости без обучение

Увод

Извличането на знания от големи обеми данни (Data Mining) е сравнително ново научно направление - възникнало през последните 20-25 години. В действителност, изследванията в тази област не са напълно нови и непознати за научната общност, тъй като се базират на постиженията на редица други познати и добре развити научни дисциплини като статистика, машинно обучение, изкуствен интелект, бази данни и др. Това, което отличава Извличането на знания от големи обеми данни (Data Mining) от останалите дисциплини, е интердисциплинарният характер, който се заключава в интегрирането на знанията за бази данни, аналитични методи и средства, и бизнес познанията.

Настоящият дисертационен труд е посветен на изследване на приложимостта и ефективността на различни Data Mining модели за класификация. Под *Data Mining модели за класификация* в дисертацията се разбира *обучени класификатори, получени чрез прилагане на различни методи за класификация при извличане на знания от големи обеми данни*. Терминът *Data Mining* често се използва в текста на дисертацията вместо неговия български еквивалент - *извличане на знания от големи обеми данни*, с цел постигане на по-кратък и ясен изказ.

Изследванията са осъществени върху данни, произхождащи от две различни предметни области - данни за студенти и данни за открити морски цели, с което се цели да се потвърди още един път твърдението, че Data Mining методите и средствата могат да бъдат прилагани върху различни видове данни от различни научни области.

През последните години българските университети претърпяват различни промени и се изправят пред сериозни проблеми. Намалява броят на потенциалните кандидати за висшите учебни заведения и се засилва конкуренцията между образователните институции, което води до необходимостта от актуална информация, която да подпомага ръководствата при взимането на важни и своевременни управленски решения. Съвременните университети разполагат с големи обеми данни, които в повечето случаи не са ефективно анализирани и използвани. Това може да бъде постигнато чрез използване на Data Mining методи и средства.

„Educational Data Mining" е едно от най-новите направления в областта на извличането на знания от големи обеми данни, посветено на решаването на проблеми, свързани с привличането на подходящи студенти (targeted marketing), задържане на студенти и предотвратяване на тяхното отпадане от обучение (retention of students), повишаване качеството на образователния процес, управление на кандидатстудентските кампании, подобряване на организационната ефективност, управление на взаимоотношенията с бивши възпитаници (alumni management). Осъщественият анализ на

публикуваните научни изследвания в областта на Educational Data Mining показва, че предсказването успеха на студентите е много важен и актуален проблем за университетите. Успехът на студентите се предсказва най-често чрез решаване на задача за предсказване чрез Data Mining методи и средства, като в повечето случаи предсказваната величина е номинална променлива, т.е. решава се задача за класификация. Най-често използваните методи за генериране на модели за предсказване включват „Дърво на решенията", „Невронни мрежи", Байесови методи - наивен Байесов класификатор, Байесови мрежи, и логистична регресия. Във всички случаи обаче, се прилагат няколко различни метода или няколко различни алгоритми, за да се достигне до създаването на подходящ модел за предсказване.

Проблемът, свързан с класификацията на открити радарни цели, занимава радарната научна общност от много време. Задачите за откриване и класификация на сигнали могат да се решават с методите за разпознаване на образи, защото радарните сигнали могат да се разглеждат като образи от данни. При много методи за разпознаване на образи се използва статистическата теория на решенията. Съществува голямо разнообразие от методи за разпознаване на изображения, разработени и в други научни области, например Data Mining, които биха могли да бъдат приложени за класификация на радарни цели, което е и работната хипотеза в настоящата дисертация.

С въпросите за класификация на наземни и морски цели при условията на forward scattering ефект научната общност се занимава през последните няколко години. Един от водещите изследователски колективи в областта е екипът на проф. Черняков от Университета в Бирмингам, Великобритания. За момента не са известни изследвания, свързани с използването на Data Mining методи за класификация на радарни морски цели. Изследванията върху такъв тип данни в настоящия дисертационен труд имат за цел да проверят възможностите за ефективно прилагане на Data Mining алгоритми за решаване на задачата за класификация и по отношение на този нестандартен тип данни – forward scattering радиолокационни сигнали, и по-конкретно – за класификация на открити морски forward scattering радиолокационни цели. Особеностите тук са свързани с избор на информативни параметри на сигналите и необходимостта от предварителна обработка на записите на сигналите с цел оценка на параметрите им, за да се получат от тях данни, подходящи за аналитична обработка с Data Mining методи и средства. Друг проблем е невъзможността на този етап да бъдат получени голям обем от записи, което до голяма степен влияе върху качеството на създаваните класификационни модели.

На базата на извършените проучвания, описани по-горе, и благодарение на осигурения достъп до двете съвкупности от данни – данни за студенти, предоставени от ръководството на УНСС, и записи за морски цели, предоставени от изследователите от Бирмингамския университет, в рамките на действащи научноизследователски проекти, са дефинирани целта и основните задачи на дисертацията.

Целта на настоящия дисертационен труд е да се генерират и изследват Data Mining модели за класификация (обучени класификатори, получени чрез прилагане на различни методи за класификация при извличане на знания от големи обеми данни), за предсказване успеваемостта на студентите в университета на базата на техни персонални характеристики преди и след приемане в университета, и за класификация на морски обекти.

Дисертацията е структурирана и организирана в три глави, както следва:

В ***Първа глава*** е разгледана актуалността на темата на изследванията, направен е обзор на Data Mining областта, като е акцентирано върху решаването на задачата за класификация. Разгледани са четири Data Mining метода за генериране на модели за класификация, описани са различни мерки за оценка и сравнение на моделите. Направен е обзор на състоянието на научните изследвания относно използването на Data Mining методи и средства в избраните конкретни приложни области – обучението на студенти и класификацията на морски цели. Дефинирани са целта и задачите на изследването.

Във ***Втора глава*** е представена методиката за провеждане на изследването, включваща избор на подход за реализация на Data Mining проекта, избор на подходящи софтуерни средства за осъществяване на поставените задачи, и описание на конкретните дейности, които ще бъдат осъществени в рамките на изследването. Описани са и извършените дейности, свързани с предварителна подготовка и изследване на двете съвкупности от данни – за студенти и за морски цели, които се използват при генерирането на Data Mining моделите за класификация.

В ***Трета глава*** са представени резултатите от изследването, осъществено чрез прилагане на избраните Data Mining алгоритми за класификация върху двете подготвени съвкупности от данни, за студенти и за морски цели, с помощта на Data Mining софтуер WEKA. Оценени и сравнени са генерираните Data Mining модели за класификация чрез избраните мерки за оценка.

Дисертацията започва с ***Увод***, в който накратко е описано съдържанието на материала, и завършва със ***Заключение***, в което са представени основните изводи от осъщественото изследване, като са изброени научните и научно-приложните приноси на автора.

В началото на Дисертацията е дадено ***Съдържание*** и ***Списък на използваните термини***, а в края – ***Списък на фигурите, Списък на таблиците*** и ***Библиография***.

Получените резултати са публикувани в 1 статия в международно електронно списание (International Journal of Computer Science and Management Research, 2012), 3 статии, докладвани на реномирани международни конференции, 2 от които в чужбина (EDM 2011, IRS 2011) и 1 в България (S3T 2010), 1 статия - приета за публикуване в специализирано национално списание (International Journal of Cybernetics and Information

Technologies), и 1 статия - представена и публикувана на международен семинар в България.

Глава 1: Обзор и актуално състояние на областта Извличане на знания от големи обеми данни (Data Mining) и методите за класификация

Първа глава на дисертационния труд е организирана в шест части. В първата част (т.1.1) е представена актуалността на темата на изследванията. Във втората част (т.1.2) е направен обзор на областта Извличане на знания от големи обеми данни (Data Mining), като е акцентирано върху задачата за класификация, разгледани са четири Data Mining метода за генериране на модели за класификация, и са описани мерки за оценка и сравнение на моделите. Обзор на състоянието на научните изследвания относно използването на Data Mining методи и средства в избраните конкретни приложни области – сферата на образованието и FS сигналната обработка, е направен в третата и четвъртата част на главата (т.1.3 и т.1.4). В петата част (т.1.5) са дефинирани целта и научните задачи на изследването. Главата завършва с изводи (т.1.6), на които се базират дейностите, описани в следващите две глави на дисертацията.

1.1. Актуалност на темата на изследвания

Живеем в силно наситена с информация среда. Всяка организация, за осъществяване на своите ежедневни дейности, обикновено използва различни информационни източници и често генерира големи количества данни.

Успешно работещите организации в съвременното „информационно общество" или „общество базирано на знания" са осъзнали, че човешкият капитал и данните, с които разполагат, са техните най-ценни ресурси. Ефективното използване на наличните данни и информация е важен фактор за нормалното функциониране и просперирането на организациите.

За да се оползотворяват ефективно огромните обеми от налични данни и за да бъдат превръщани в полезна информация и знания, които да подпомагат успешното развитие на организациите, трябва да се прилагат нови мощни информационни технологии за обработка и анализ. Голяма част от тези технологии през последните години се обединяват от понятието „Business Intelligence (BI)" - обобщаващо понятие, включващо архитектури, инструменти, бази данни, приложения и методологии, използвани за преобразуването на данните в информация и знание с цел подпомагане взимането на управленски решения. Едни от новите инструменти за анализиране и ефективно използване на големи обеми от данни с различни формати, са методите и средствата за извличане на знания, известни като Data Mining.

Настоящият дисертационен труд е посветен на актуален научен проблем - изследване на приложимостта и ефективността на различни методи за извличане на знания от големи обеми данни (Data Mining), за решаване на задачата за класификация. Изследванията са осъществени върху данни, произхождащи от две различни предметни области и с различни обеми. Едната съвкупност от данни е от сферата на образованието и представлява извадка от наличните данни за студентите, приети да учат в университет през три последователни години (2007-2009), в обем около 10000 записа. Втората съвкупност от данни е от сферата на защита на морските граници от несанкциониран достъп на плувци и малки плавателни съдове, с помощта на радарни бариери инсталирани на морски буйове. Представлява извадка от данни, получени чрез обработка на Forward Scattering радарни сигнали, и включва набор от параметри, характеризиращи сигналите, получени чрез реални записи на няколко вида морски радиолокационни цели, в обем около 80 записа.

Анализ на университетски данни

През последните 20 години българските университети претърпяват различни промени и се изправят пред сериозни проблеми. От една страна, процесите на демократизация доведоха до значително нарастване броя на образователните институции в областта на висшето образование. В момента в България официално съществуват 43 университета и висши училища, и 10 колежа, около 70% от тях - държавни, а останалите 30% - с частно финансиране[1]. От друга страна, забавеното икономическо развитие предизвика негативни демографски промени. Много млади хора напуснаха страната, за да търсят реализация в чужбина, и все по-често техните деца получават своето образование извън страната. Освен това, с присъединяването на България към Европейския съюз, се увеличиха възможностите за получаване на висококачествено образование в европейските държави. Тези процеси доведоха до значително намаляване броя на потенциалните кандидати за висшите учебни заведения в България. През последните няколко години дори водещи университети не успяха да привлекат обичайния брой кандидат-студенти и изпитаха сериозни затруднения да попълнят обявените свободните места.

Изброените фактори оказват сериозно влияние върху пазара на висшето образование у нас, като особено силен е натискът върху държавните университети, които са свикнали да работят в много по-устойчива среда, в която липсва силна конкуренция. Българските университети са изправени пред нови проблеми, свързани с разработването и осъществяването на подходяща политика и стратегия. От една страна, те трябва да разработват и предлагат учебни програми, съобразени с нуждите на бизнеса и обществото, а от друга, да привличат най-подходящите студенти, които успешно ще завършат

[1] Report on the Status Analysis of Research in Bulgaria, Bulgarian Ministry of Education and Youth (2010), http://www.minedu.government.bg

обучението си и ще подпомогнат постигането на поставените от учебното заведение цели. Отчитайки промените, настъпващи в образователните процеси в рамките на разширена Европа, университетите трябва да насочат своите усилия към откриване и засилване на своята уникалност и към привличането на най-подходящите студенти.

При дискусиите относно модернизацията на електронната среда в един от водещите български университети се стига до изводите, че непрекъснато събираните и натрупвани данни не се използват ефективно. Изготвят се само ограничен брой стандартни справки, но те не осигуряват достатъчно информация за задълбочено анализиране работата на университета, за оптимизиране на дейностите и взимане на навременни и адекватни управленски решения. Ръководителите на университета ясно осъзнават необходимостта от използването на нови подходи и средства за анализиране на наличните данни. Така се стига до заключението да се проведат научни изследвания относно възможностите за използване на Data Mining методи за анализ на университетските данни.

„Educational Data Mining" (EDM) е едно от най-новите направления в научната област Извличане на знания от големи обеми данни (Data Mining), чието начало се поставя през 2004г. с организирането на първия научен семинар, първата конференция по Educational Data Mining се организира през юни 2008г. в Монреал, Канада, след което се провежда ежегодно, от октомври 2009г. започва издаването и на електронно списание – Journal of Educational Data Mining[2]. Научните изследвания са насочено към решаването на сложни проблеми, свързани с повишаване ефективността на управление, качеството на образование и подобряване процесите на взимане на управленски решения, на организациите в образователния сектор. Научната общност- International Educational Data Mining Society[3], е съвсем наскоро формирана (юли 2011) и се разраства с бързи темпове. Публикувани са множество научни статии, в които се дискутират различни проблеми в образователния сектор и се предлагат идеи за тяхното успешно решаване с помощта на data mining методи и средства. Класификационните data mining алгоритми се прилагат в по-голямата част от проучените изследвания в тази област.

Изводи: Educational Data Mining" (EDM) е едно от най-новите направления в научната област Извличане на знания от големи обеми данни (Data Mining). Съгласно публикациите, ***алгоритмите*** за откриване на знания в данни чрез обучение на класификатори много често се използват за решаването на различни проблеми при обучението на студентите.

[2] Journal of Educational Data Mining, http://www.educationaldatamining.org/JEDM/
[3] International Educational Data Mining Society, http://www.educationaldatamining.org/

Класификация на радарни сигнали

В новото Европейско пространство особено актуален става напоследък въпросът за охрана на границите му, както на въздушните и сухоземните, така и на морските. В момента се търсят нови иновативни технологии за защитата им и се инвестират много средства както от самите страни, така и от Европеската общност (7 Рамкова Програма на Европейската Комисия, Приоритетна област 10 "Sequrity").

По една такава нова технология за защита на морските граници от несанкциониран достъп на плувци и малки плавателни съдове, с помощта на радарни бариери инсталирани на морски буйове, в момента се работи в Бирмингамския университет - колективът на проф. М.Черняков. Охраната се състои в създаването на мрежа от евтини, автономни и автоматизирани радарни бариери, охраняващи морски зони или брегова линия. Технологичната иновативна същност на тези нови радарни бариери се състои в това, че те се изграждат на принципа на комуникационните канали, т.е. чрез пространствено разделение на предавателя и приемника на радара.

Тогава, откриването и класификацията на целите се извършва само при пресичането от целта на тесния електромагнитен лъч между предавателя и приемника на радара. Този нестандартен нов вид радари използват нов ефект, слабо изследван, и се наричат радари гледащи напред (forward scattering radar). На практика до сега не са се използвали подобен вид радарни системи. Естествено, при осъществяване на защита на морски граници е особено важно охраната да може да установи факта, че границата е нарушена, да определи мястото на нарушението, както и вида на нарушителя (плувец или плавателен съд). Следователно, от особена важност се явява задачата за класификация на морски обекти на база на радарни сигнали в системи, използващи ефекта на "forward scattering".

Проблемът, свързан с класификацията на откритите радарни цели, занимава радарната научна общност от много време. За решаването му се използват предимно различни статистически методи. Но, с въпросите за класификация на морски и наземни цели при условията на forward scattering ефект научната общност се занимава само през последните няколко години. Един от водещите изследователски колективи в областта е споменатият по-горе екип на проф. Черняков от Бирмингамския университет. През последните две години в тази научноизследвателска дейност се включи и български колектив с учени от СУ „Св. Кл.Охридски”, УНИБИТ, БАН, които съвместно с изследователите от Бирмингамския университет, в рамките на съвместен научен проект, финансиран от Националния фонд „Научни изследвания”, работят по автоматичното откриването, оценката на параметрите на сигналите и класификация на морски цели.

Трябва да се отбележи, че за момента не са известни изследвания, свързани с използването на data mining методи за класификацията на радарни морски цели. Изследванията върху такъв тип данни в настоящия дисертационен труд имат за цел да проверят възможностите за ефективно прилагане на тези методи и по отношение на този нестандартен тип данни – forward scattering радиолокационни сигнали, и по-конкретно – за класификация на открити морски forward scattering радиолокационни цели. Особеностите тук са свързани със сложните преобразувания, през които трябва да преминат сигналите, за да се получат данни, подходящи за аналитична обработка с Data Mining методи и средства, както и с невъзможността на този етап да бъдат получени голям обем от записи (примери), което до голяма степен влияе върху качеството на създаваните класификационни модели.

Изводи: До момента не са намерени резултати за използване на *алгоритми за откриване на знания в данни чрез обучение на класификатори* за класификация на радарни морски цели. Възможността за прилагането на такива методи върху този специфичен вид данни е една от работните хипотези, използвани в настоящата дисертация.

1.2. Обзор на методите за откриване на знания в данни (Data Mining) чрез обучение на класификатори

В тази част на дисертацията са представени резултатите от теоретичните проучвания. В т.1.2.1 е разгледано актуалното състояние на областта Извличане на знания от големи обеми данни (Data Mining), като са описани понятието Data Mining и основните видове Data Mining задачи. В т.1.2.2 е разгледана задачата за класификация - основната задача, решавана в дисертационния труд чрез Data Mining методи и средства, а в т.1.2.3 са описани основните Data Mining методи за класификация, използвани в дисертацията. Основните критерии за оценка и сравнение на моделите за класификация са представени в .1.2.4. В последната част, т.1.2.5, са дадени и някои съществуващи подходи за реализация на Data Mining проекти, сред които е и подходът, избран за осъществяване на изследванията в дисертацията.

1.2.1. Състояние на научната област Извличане на знания от големи обеми данни (Data Mining)

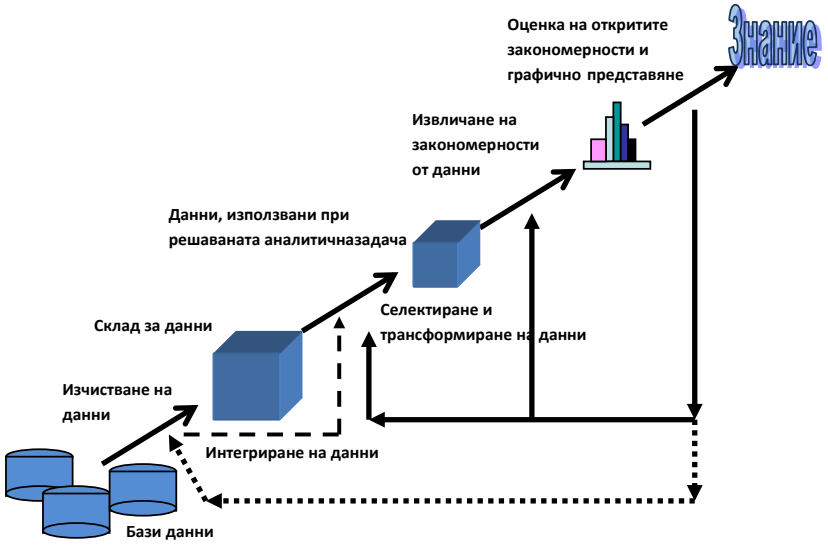
1.2.1.1. Понятието Data Mining

Извличането на знания от големи обеми данни (Data Mining) е сравнително ново научно направление - възникнало през последните 20-25 години. В действителност, изследванията в тази област не са напълно нови и непознати за научната общост, тъй като се базират на постиженията на редица други познати и добре развити научни дисциплини

като статистика, машинно обучение, изкуствен интелект, бази данни и др. Това, което отличава Data Mining от останалите дисциплини, е интердисциплинарният характер, който се заключава в интегрирането на знанията за бази данни, аналитични методи и средства, и бизнес познания.

Терминът „Data Mining" се появява в края на 80-те години на 20 век. Първите книги по тази тематика излизат през 90-години и представляват сборници от статии, посветени на откриването на знания в бази данни (Piatetsky-Shapiro and Frawley, 1991; Fayyad et al. 1996). Малко по-късно се появяват и книги, които са предимно бизнес ориентирани и засягат практични въпроси, свързани с приложението на data mining методите, а не толкова с теоретични аспекти (Adriaans and Zantige, 1996; Berry and Linoff, 1997; Cabena et al., 1998). През 2001г. е публикувана интердисциплинарна книга, в която data mining методите са представени от гледна точка на базите данни и е написана от международен авторски екип (Hand et al., 2001). В последствие се появяват множество книги, които се фокусират върху теоретични аспекти на Data Mining (Han&Kamber – 2000; Ye - 2003; Larose – 2005, 2006; Sumathi et al. – 2006; Tan et al. - 2006), както и върху практическата реализация на много от дейностите, свързани с извличане на закономерности – изследване и предварителна подготовка на данните, моделиране и оценяване (Pyle - 1999, 2003; Guidici – 2003; Witten&Frank - 2005). Много от авторите наблягат върху практическото приложение на Data Mining методи и средства в различни области – например за маркетинг, продажби, управление на взаимоотношенията с клиенти, в производството, комуникациите и т.н. (Berry&Linoff – 2004; Shmueli et al. – 2007; Vercellis - 2009).

Понятието Data Mining първоначално се използва за обозначаване на една от основните стъпки в процеса на "откриване на знания в бази данни", представен на Фиг.1.1.



Фиг.1.1: "Извличането на знания" (Data Mining) - важна стъпка в процеса на "откриване на знания в бази данни"

На практика, терминът Data Mining е станал по-популярен от термина „откриване на знания в бази данни" не само в индустрията и медиите, но и сред научноизследователските среди. Затова, понятието ***Data mining*** се използва в по-широк смисъл, като ***процес на откриване на полезни знания сред големи количества данни***.

Въпреки голямото разнообразие от дефиниции, давани от изследователи и експерти, представляващи академичната общност и бизнеса, всички те съдържат общи елементи. Можем да заключим, че:

> *Data Mining е* ***процес на подбор, изследване и моделиране на огромни количества данни, водещи до откриване на предварително неизвестни закономерности и взаимовръзки, с цел получаване на разбираеми и полезни резултати за притежателя на данните.***

Именно тази работна дефиниция е използвана за целите на изследванията в настоящата дисертация.

1.2.1.2. Основни видове задачи, решавани при извличане на знания от големи обеми данни (Data Mining)

Задачите, които се решават при извличането на знания от данни, най-общо могат да бъдат разделени на две категории: *задачи за описание и обобщение на данни* и *задачи за предсказване*. При описателните задачи се представят в обобщен вид основните свойства на наличните данни. При задачите за предсказване се правят заключения по отношение на съществуващите данни с цел предсказване по отношение на нови данни.

Това условно разделяне на data mining задачите на две категории е свързано и с разглеждането на два основни типа извличане на закономерности – *с обучение* (supervised/directed learning) и *без обучение* (unsupervised/undirected learning). Извличането на закономерности чрез обучение (directed data mining) води до обясняване или категоризиране на избрана променлива, наречена изходна или целева. Извличането на закономерности без обучение (undirected data mining) води до откриване на закономерности или подобия сред групи от записи, без да се използва предварително избрана (целева) променлива и без да са предварително известни класовете, в които се разпределят обектите в изследваната съвкупност от данни.

За изграждането на добри модели за предсказване от изключително важно значение е доброто познаване и разбиране на данните. Затова, при извличането на закономерности обикновено първо се решават задачи, свързани с описание и обобщение на данните, а след

това се прилагат алгоритми и методи за генериране на модели за предсказване. Понякога задачите за описание и обобщение на данни могат да бъдат основна цел при извличането на закономерности и да бъдат решавани самостоятелно.

Основна цел при ***задачите за описание и обобщение на данни*** е намирането на подходящо описание на основните характеристики на данните, което позволява на потребителя да получи представа за тяхната структура. Този процес е известен под наименованието *„Изследователски анализ на данните" (Exploratory Data Analysis - EDA).* Използват се различни графични методи и средства за визуализация – диаграми, графики, таблици и правила, които най-често позволяват интерактивно участие на изследователя.

Към задачите за описание и обобщение на данни спада и ***клъстерният анализ (Clustering)***, чиято основна цел е разделянето на данните на групи (клъстери). Клъстерите се формират чрез прилагане на принципа за максимално сходство между елементите на един клъстер и минимално сходство между елементите на различни клъстери. Всеки клъстер може да се разглежда като клас от обекти, от който могат да се извличат различни правила. При клъстерният анализ, за разлика от предсказването, предварително не е известно на какви групи ще бъдат разделени данните.

Втората категория Data Mining задачи са ***задачите за предсказване***, при които целта е да се намери модел, чрез който да се предсказват стойностите на една от променливите на база на известните стойности на други променливи. Задачите за предсказване могат да бъдат разделени на два вида – ***класификация и регресия***. ***Класификацията*** е процес на изграждане на модели, описващи съществуващи класове от обекти, с цел използването им за определяне класа на обекти с неизвестен клас. При класификацията, предсказваната променлива (наричана още целева променлива/клас) е от категориен тип и приема краен брой дискретни стойности. Моделите се изграждат чрез анализ на извадка от т.нар. обучаващи данни (training data) – обекти с известен клас на принадлежност. ***Регресията*** е задача подобна на класификацията, като основната разлика се състои в това, че при регресията предсказваната променлива е числова величина, която приема не дискретни, а непрекъснати стойности. Това означава, че целта на предсказването е да се намери числовата стойност на целевата променлива за нови обекти. Когато предсказването се осъществява чрез анализиране на данни, организирани във времеви серии, тази задача носи названието ***прогнозиране*** (forecasting).

Задачите, описани по-горе, са свързани предимно с изграждането на модели. Друг тип задачи са свързани с откриване на правила. ***Извличането на асоциативни правила*** е популярен и добре изследван метод за откриване на интересни взаимовръзки между променливите в големи бази данни, който може да бъде използван както при описание и обобщение на данни, така и при решаване на задачи за предсказване.

Данните могат да съдържат определени обекти, които не отговарят на общия модел на данните. Такива обекти са изключения, наричат се още крайни или екстремни (outliers). Повечето от Data Mining методите разглеждат екстремните обекти като шум, който обикновено се пренебрегва. Съществуват и случаи, при които именно изключенията, които не се подчиняват на общите закономерности, са от особен интерес, например при разкриването на измами. Задачата за намиране на подобни обекти се нарича ***анализ на крайностите/отклоненията/изключенията*** (outlier/deviation detection).

Това е само една от съществуващите класификации на най-често решаваните Data Mining задачи (Hand et al., 2001). Различните автори използват и други подходи при определяне на типовете Data Mining задачи, базирайки се на разнообразието от цели, които могат да бъдат постигани, както и типовете данни, от които се извличат закономерностите. За целите на настоящия дисертационен труд този начин на класификация е подходящ, тъй като достатъчно добре дефинира и разграничава задачата за предсказване, предмет на настоящите изследвания, от останалите Data Mining задачи.

Изводи: Избрана е работна дефиниция на понятието Data Mining. Идентифицирана е задачата за извличане на знания от данни чрез обучение на класификатори като подходяща за постигане на целта на дисертацията.

1.2.2. Същност на класификацията

От направения обзор на Data Mining изследванията в областта на Educational Data Mining (т.1.3) следва, че най-често използваният подход е търсенето на класификационни модели за предсказване на "качеството" на студентите, като особено популярни са следните методи за класификация - „Дърво на решенията”, „Невронни мрежи”, „Генератор на правила”, „К-най-близък съсед”, Байесови класификатори и др. В обзора на областта, свързана с класификацията на морски цели (т.1.4), също е подчертана особената важност на задачата за класификация. Затова:

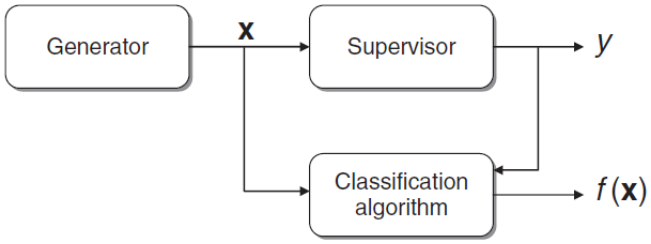
Основната Data Mining задача, решавана в настоящия дисертационен труд, е задачата за предсказване на променлива с дискретни стойности, т.е. задачата за извличане на знания от данни чрез обучение на класификатори.

Предсказването в този случай се разглежда като процес на изследване на реални данни, съдържащи вече класифицирани обекти, и създаване на модели, описващи закономерностите и взаимовръзките в изследваните данни, с цел да се определи (т.е. да се предскаже) класа, към който принадлежи даден некласифициран обект.

Основната цел на класификационния модел е да идентифицира повтарящи се взаимовръзки между входните променливи, които описват обектите/записите, принадлежащи на един и същи клас. Откритите взаимовръзки впоследствие се преобразуват в класификационни правила, които се използват за предсказване класа на обекти/записи, за които са известни стойностите на предсказващите/входните променливи. Класификационните правила могат да бъдат представяни под различна форма в зависимост от типа на използвания модел.

От математическа гледна точка, при решаването на задачата за класификация имаме m на брой примера, описани с двойката $(\boldsymbol{x}_i, y_i)$, $i \in M$, където $\boldsymbol{x}_i \in R^n$ е вектор от стойностите на n предсказващи променливи за i-тия пример, а с $y_i \in H = \{v_1, v_2, \ldots, v_H\}$ се обозначава съответния клас на целевата променлива. Всеки компонент x_{ij} на вектора $\boldsymbol{x}_i$ се разглежда като реализация на случайната променлива X_j, $j \in N$, представляваща атрибута a_j в съвкупността от данни D. При класификация с двоична предсказвана променлива $H=2$, т.е. имаме два класа, обозначавани например $H=\{0,1\}$ или $H=\{-1;1\}$, без да се загубва общия характер (without loss of generality).

Нека F е клас от функции $f(\boldsymbol{x})$: $R^n \rightarrow H$, наричани хипотези и представляващи хипотетичните взаимовръзки между y_i и $\boldsymbol{x}_i$. Класификационната задача се състои в дефинирането на подходящо хипотетично пространство F и на алгоритъм A_F, чрез който се открива функция $f^* \in F$, която да описва по оптимален начин взаимовръзката между предсказващите променливи и предсказваната променлива – класа. Вероятностното разпределение $\boldsymbol{P}_{\boldsymbol{x},y}(\boldsymbol{x},y)$ на примерите в съвкупността от данни D, дефинирано в пространството $R^n \times H$, обикновено е неизвестно и повечето класификационни модели са непараметрични, тъй като не правят предварителни предположения за формата на разпределението $\boldsymbol{P}_{\boldsymbol{x},y}(\boldsymbol{x},y)$. Процесът е представен на Фиг.1.2 и включва три компонента – генератор на наблюдения, учител (supervisor) на класа и класификационен алгоритъм.



Фиг.1.2: Процес на обучение при класификация (Vercellis, 2008)

Задачата на генератора е да извлича случайни вектори $\boldsymbol{x}$, съгласно неизвестно вероятностно разпределение $\boldsymbol{P}_{\boldsymbol{x}}(\boldsymbol{x})$. За всеки вектор $\boldsymbol{x}$ учителят (supervisor) връща класа, съгласно условна вероятност $\boldsymbol{P}_{y|\boldsymbol{x}}(y|\boldsymbol{x})$, която също е неизвестна. Класификационен алгоритъм A_F, наричан още класификатор, избира функция $f^* \in F$ в хипотетичното

пространство така, че да се минимизира подходящо избрана функция на загубите (loss function).

Класификационните модели са подходящи както за описание и интерпретация на данните, така и за предсказване, като и в двата случая имаме извличане на закономерности чрез обучение (supervised learning). По-простите модели обикновено генерират интуитивни класификационни правила, които могат да бъдат лесно интерпретирани. По-сложните модели създават по-трудни за интерпретиране правила, но обикновено постигат по-висока точност на предсказване.

Част от примерите в съвкупността от данни D се използва за *обучаване*, т.е. за извличане на функционална взаимовръзка между целевата променлива и входните променливи, изразена посредством хипотезата $f^* \in F$. Останалите данни се използват за оценка точността на генерирания модел, както и за избор на най-добър модел сред генерираните с различни класификационни методи.

Следователно, създаването на класификационен модел включва три основни фази:

- *Обучение*

  Във фазата на обучение класификационният алгоритъм се прилага върху примерите, принадлежащи на подсъвкупност T на съвкупността от данни D, наричана обучаваща извадка (training data), с цел извличане на класификационни модели и правила, чрез които да се определи целевия клас y за всяко наблюдение $\boldsymbol{x}$.

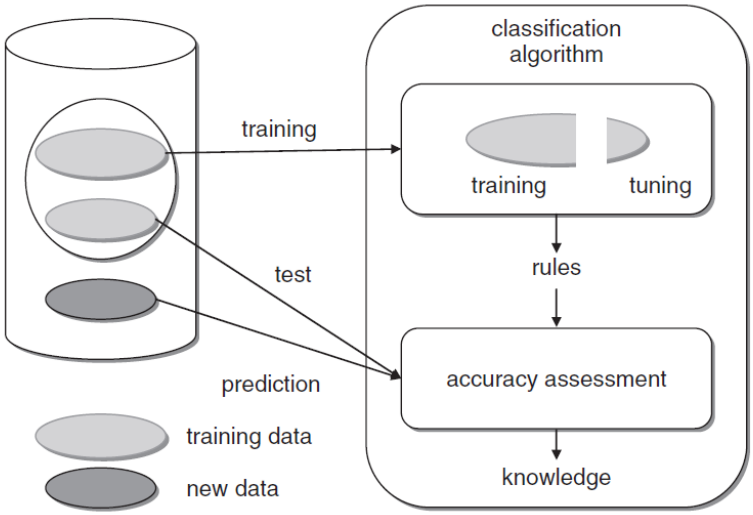
- *Тестване*

  Във фазата на тестване правилата, генерирани по време на обучението, се използват за класифициране на примерите от съвкупността D, които не са били включени в обучаващата извадка и за които вече е известен класът (тестова извадка – test data). За да се оцени точността на класификационния модел, действителният клас на всеки пример от тестовата извадка $V=D-T$ се сравнява с класа, предсказан от класификатора. Обучаващата и тестовата извадки трябва да са напълно независими, т.е. примери от обучаващата извадка не трябва да се съдържат в тестовата извадка и обратно, за да се постигне реалистична оценка.

- *Предсказване*

  Фазата на предсказване представлява реалното използване на класификационния модел за определяне класа на нови примери, които ще бъдат регистрирани в бъдещ момент. Предсказване се получава чрез

прилагане на моделите и правилата, генерирани по време на обучението, върху входните променливи, описващи новите примери.



Фиг.1.3: Фази в процеса на обучение при класификационен алгоритъм (Vercellis, 2008)

На Фиг.1.3 е представен процесът на обучение при класификационните алгоритми. Допълнително към фазите описани по-горе, от обучаващата извадка може да се отдели част, наречена извадка за настройка (tuning set), която се използва при някои класификационни алгоритми за намиране на оптималните стойности на някои параметри, участващи във функциите $f \epsilon F$.

Основните стъпки, които се предприемат при решаването на задачи за класификация, са:

- *Дефиниране целта на задачата* – в зависимост от целта на задачата могат да бъдат създадени различни модели за предсказване;
- *Определяне типа на модела и метода за класификация* – класификацията може да бъде реализирана чрез различни методи, всеки от които води до получаването на различен тип модел за предсказване (напр. дърво на решенията, невронна мрежа, класификационни правила и т.н.).
- *Избор на алгоритми* – Съществуват различни алгоритми за изграждане на модели от даден тип (напр. за изграждането на дърво на решенията могат да се използват алгоритмите CART, C5.0, CHAID).
- *Прилагане на избраните методи и алгоритми, и оценка на получените резултати от класификацията*

Поставените цели най-често се постигат чрез прилагането на няколко различни методи и алгоритми при решаването на една и съща задача. Обикновено не може да се

прецени кой от методите е най-подходящ, преди да се изпробват различни подходи и да се направи оценка на получените резултати от предсказването.

Изводи: Обяснена е същността на задачата за *извличане на знания от данни (Data Mining) чрез обучение на класификатори* и са описани основните стъпки, които се предприемат при решаването на тази задача.

1.2.3. Методи за извличане на знания от данни чрез обучение на класификатори, използвани в дисертацията

Съществува голямо разнообразие от методи за класификация. За целите на настоящия дисертационен труд са разгледани четири основни метода – „Дърво на решенията", „Генератор на правила", „К-най-близък съсед" и „Невронни мрежи".

Методите „Дърво на решенията", „Генератор на правила" и „К-най-близък съсед" използват класификационни процедури, базирани на сравнително прости и интуитивни алгоритми. При метода „К-най-близък съсед" класификацията се извършва чрез определяне на разстоянието между обектите в изследваните данни. При дървото на решенията се прилага итеративно разделяне на обектите, съдържащи се в изследваните данни, на все по-хомогенни групи по отношение на предсказваната променлива. Генераторите на правила работят чрез генериране на „покриващи" правила, валидни за по-голямата част от обектите в изследваните данни.

При метода „Невронни мрежи" атрибутното пространство се разделя на отделни области на базата на стойностите на предсказваната променлива. Тъй като на практика е много трудно да се постигне напълно точно разделяне на пространството на области, съответстващи на стойностите на предсказваната величина, се дефинира функция на загубите (loss function), която взема предвид грешно класифицираните обекти. Следва решаването на оптимизационна задача – оптимално разделяне на области се получава чрез минимизиране на общите загуби.

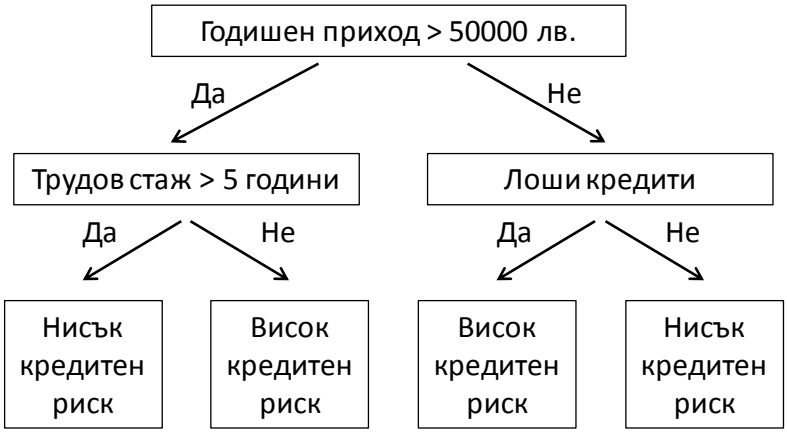
1.2.3.1. Метод „Дърво на решенията"

„Дърво на решенията" е най-известния и най-често използван метод при data mining приложенията. Популярността му се дължи на разбираемостта, лесното използване, бързината, устойчивостта по отношение на липсващи данни и изключения, и най-вече на лесната интерпретация на генерираните модели и класификационни правила. Чрез алгоритмите „Дърво на решенията" се създават модели от дървовиден тип, на базата на които се дефинират и разбираеми правила, описващи взаимовръзките между целевата

променлива и предсказващите входни променливи, чрез които се разделят обектите, принадлежащи към различни класове.

Създаването на класификационно дърво съответства на фазата на обучение на модела и се осъществява чрез повтаряща се процедура. Дървото на решенията израства чрез итеративно разделяне на обучаващите данни на все по-малки групи, съгласно критерии за разделяне (division criteria), като целта е получените подгрупи да бъдат с максимална „чистота" (purity) или "еднородност" (хомогенност - homogeneity) по отношение на предсказваната целева променлива.

Дървото на решенията представлява диаграма с дървовидна структура, при която всеки вътрешен възел (node) представлява тестова процедура за даден атрибут (входна променлива), всеки клон (branch) представлява резултат от теста, а крайните възли – листа на дървото (leaf nodes), представляват търсения клас. Началният възел се нарича корен на дървото (root node). Пример за дърво на решенията е показан на Фиг.1.4.



Фиг.1.4: Дърво на решенията – пример

Основните параметри на алгоритмите, чрез които се генерира дърво на решенията, са свързани с начините за избиране на най-добър разделящ тест (правило за разделяне), както и правила за спиране нарастването на дървото.

Правило за разделяне. За всеки възел на дървото се избира оптимално правило за разделяне на обектите и създаването на производни възли, включващо избор на атрибут за разделяне и стойности на този атрибут, по които да стане разделянето. Използват се различни критерии за избор на подходящо правило за разделяне, които се различават по броя на производните възли, броя на атрибутите и мерките за оценка на „чистотата" на създадените производни възли.

Правилото за разделяне се използва за намиране на най-подходящата обяснителна променлива сред всички налични и за избиране на най-ефективен критерий за разделяне сред всички възможни. Обикновено и двете решения се взимат чрез изчисляване на т.нар.

функция за оценка (evaluation function), за всеки атрибут и за всяко възможно разделяне, което води до получаването на мярка за нееднородност (хетерогенност) на обектите, принадлежащи на разделяния възел (parent node) и на производните възли (child nodes), по отношение стойностите на предсказваната променлива. Максималната стойност на функцията за оценка определя разделянето, което води до получаването на производни възли с по-голяма еднородност, отколкото в разделяния възел.

Нека p_h е пропорцията от обекти от клас v_h, $h \in H$, в даден възел q и нека Q е общия брой обекти в q. Тогава

$$\sum_{h=1}^{H} p_h = 1 \tag{1.1}$$

Мярката за нееднородност (хетерогенност) (heterogeneity index) $I(q)$ за даден възел обикновено е функция на относителните честоти p_h, $h \in H$, с които се срещат стойностите на предсказваната променлива (класа) в този възел, и трябва да отговаря на три критерия: да приема максимална стойност, когато обектите във възела са разпределени еднородно/хомогенно между всички класове; да приема минимална стойност, когато всички обекти във възела принадлежат на един и същи клас; да бъде симетрична функция по отношение на относителните честоти p_h, $h \in H$.

Сред най-популярните мерки за нееднородност (хетерогенност) за даден възел q, наричани още мерки за „нечистота" (impurity) или "нееднородност/нехомохенност" (inhomogeneity), които удовлетворяват тези свойства, са Misclassification Index, Entropy Index и Gini Index.

Мярката за грешно класифициране Misclassification Index се определя като

$$Miscl(q) = 1 - \max p_h \tag{1.2}$$

и измерва пропорцията от неправилно класифицирани обекти, когато за всички обекти във възел q се определя класът, към който принадлежат по-голямата част от тях (съгласно т.нар „демократичен принцип" - majority voting).

Мярката Entropy Index се определя като

$$Entropy(q) = -\sum_{h=1}^{H} p_h log_2 p_h \tag{1.3}$$

като е известно, че $0 log_2 0 = 0$.

Мярката Gini Index се определя като

$$Gini(q) = 1 - \sum_{h=1}^{H} p_h^2 \tag{1.4}$$

Критерии за спиране нарастването на дървото. Във всеки възел на дървото се прилагат различни критерии за спиране, за да се прецени дали нарастването на дървото трябва да продължи или възелът трябва да се счита за листо. В зависимост от използваните критерии за спиране на нарастването, при равни други условия, се получават класификационни дървета с различна топология.

Критерии за подрязване на дървото. Накрая е подходящо да се приложи някакъв критерий за подрязване на дървото, с цел да се избегне плътното прилягане на модела към обучаващите данни (overfitting). Подрязването на дървото може да се извърши по време на фазата на обучение (предварително подрязване), или след като е достигнат максималния размер на дървото (последващо подрязване).

При някои алгоритми правилата за спиране се базират на максимален брой листа на дървото или на максимален брой стъпки в процеса на разделяне. Други алгоритми използват вероятностни хипотези (probabilistic assumptions) за променливите и подходящи статистически тестове. При липсата на вероятностни хипотези нарастването спира, когато намаляването на «нечистотата» е незначително. Резултатите от класификацията, получени чрез дърво на решенията, могат да бъдат много чувствителни към избора на критерий за спиране нарастването на дървото.

Използват се два основни метода за подрязване на дървото на решенията – предварително подрязване, спиращо нарастването на дървото, когато се изпълни предварително зададен критерий, и завършващо подрязване, когато подрязването на дървото става след завършване на процеса на построяване на голямо и вероятно преспецифицирано дърво. Предварителното подрязване има предимства от изчислителната гледна точка в сравнение със завършващото, но то може да доведе до прекалено рано спиране на ръста на дървото, което води до получаване на неоптимални решения. По тази причина най-често използвания метод за избягване на преспецификацията е завършващото подрязване.

Особености при използване на метода „Дърво на решенията”

Преди да започне формирането на дървото, обикновено е добре да се извърши прецизно изследване на данните (preliminary exploratory analysis). Трябва да се убедим, че обучаващата извадка от данни е достатъчно голяма, тъй като при последващите разделяния ще има все по-малко обекти в производните възли, за които ще съществува голямо разнообразие от предсказани стойности. Освен това, разумно е да се извърши прецизно изследване на предсказваната променлива, особено ако се установи наличието на аномалии (наблюдения, които силно се различават от останалите данни), които може съществено да повлияят на резултатите от анализа. Особено внимание трябва да се обърне на формата на разпределението на предсказваната променлива. Например, ако разпределението е силно несиметрично, процедурата по изграждане на дървото може да

доведе до получаването на изолирани групи с много малко на брой обекти от «опашката» на разпределението. Освен това, когато зависимата величина е променлива от номинален тип, обикновено броят на стойностите, които тя може да приеме, не трябва да бъде много голям. Големият брой стойности трябва да се намали, за да се подобри стабилността на дървото и да се постигне по-добра предсказваща способност.

След предварителния анализ на данните трябва да се избере подходящ алгоритъм за генериране на дървото. Два са важните аспекти – критериите за разделяне и използваните методи за намаляване размерността на дървото.

Дърво на решението може да бъде изградено чрез използване на различни алгоритми, които се различават по критериите за разделяне и използваните методи за намаляване размерността на дървото. Най-често използваният алгоритъм от статистическата общност е CART (Classification And Regression Trees, Breiman et al., 1984). Други популярни алгоритми са CHAID (Chi-Squared Automatic Interaction Detection, Kass, 1980), C4.5 и неговата по-късна версия C5.0 (Quinlan, 1993). Алгоритмите C4.5 и C5.0 са широко използвани от компютърната общност. Първите версии на C4.5 и C5.0 са били предназначени за предсказване само на качествени променливи, но последните версии са подобни на CART.

Предимства и недостатъци на метода „Дърво на решенията”

Предимства:

- Разбираеми са и лесни за интерпретация;
- Много гъвкави модели, защото са приложими в различни случаи и могат да работят с различен тип променливи (например, с непрекъснати и дискретни), тъй като разделянето на обектното пространство (при най-опростения вид дървета) се извършва чрез двоични тестове (чрез прагови стойности за непрекъснатите променливи и тестове за принадлежност към дадена подгрупа при дискретните променливи);
- Бързо може да се определи класа на нови обекти.

Недостатъци:

- По-големи изчислителни ресурси.
- Зависимост на моделите от изследваните данни - Дори малка промяна в данните може да доведе до съществена промяна в структурата на дървото.
- Последователна природа – Може да доведе до недостатъчно добър избор на предсказващи променливи и до създаването на напълно различни модели за предсказване.
- Дискретизацията на променливите, които са с непрекъснати стойности, води до известна загуба на информация, съдържаща се в оригиналните данни.

- Понякога моделите могат да станат много сложни, особено при анализирането на огромни масиви от данни. Дори подрязването на дървото може да не бъде достатъчно ефективно, за да се получи лесен за разбиране и интерпретация модел.
- Не се взимат предвид взаимодействията между входните променливи, а те могат да бъдат решаващи за създаването на висококачествен модел за предсказване. Затова, при откриването на такива взаимодействия при изследването на данните, те трябва да бъдат отразени в данните чрез създаване и добавяне на нови променливи от изследователя.

Независимо от изброените недостатъци, дървото на решенията е бърз и удобен метод за предсказване, който работи добре при голям брой предсказващи променливи и за големи обеми от данни.

1.2.3.2. Метод „Генератор на покриващи правила”

Както бе описано по-горе, при дървовидните модели за осъществяване на класификацията се използва подходът „разделяй и владей”, т.е. общата съвкупност от данни последователно се разделя на все по-малки и хомогенни групи по отношение на предсказваната величина, като на всяка стъпка се избира атрибут и стойности, по които да се извърши разделянето. Чрез тази стратегия се генерира дърво на решенията, което при необходимост лесно може да се превърне в съвкупност от класификационни правила. Алтернативен подход е да се избере всеки един от класовете и да се потърси начин да се „покрият” всички обекти в него, като същевременно се изключат обектите от другите класове. Този подход се използва от генераторите на правила.

Извличането на правила се осъществява чрез генерирането на т.нар. „покриващи” правила (covering rules) – правила, които са валидни за определена част от обектите. Повечето алгоритми за извличане на правила започват работата си с откриването на „най-покриващото” правило, т.е. правилото, което е валидно за най-много обекти. След това тези обекти се премахват от изследваните данни и алгоритъмът се прилага отново с цел да се открие правилото, „покриващо” най-много обекти, и т.н. Генерираните правила на практика също разделят обектното пространство на области, подобно на дървовидните модели за предсказване. Отличават се по начина на разделяне на обектите и по графичното представяне.

Извлечените правила обикновено са от типа „Ако (условие/условия), то (резултат)” (If – Then). Условната част на правилото (частта IF) може да е много дълга и сложна, и да съдържа множество условия, които трябва да бъдат изпълнени, за да се получи резултатът от резултантната част на правилото (частта THEN).

Алгоритъмът 1R

Един от често използваните алгоритми за извличане на правила за класификация е алгоритъмът 1R (1-rule), чрез който се генерира дърво на решенията на едно ниво, представено като съвкупност от правила, всяко от които съдържа тест на един единствен атрибут (входна променлива).

Идеята на алгоритъма 1R е следната: генерират се правила, с които се тества една единствена входна променлива и се получават съответни разклонения. Всеки „клон" съответства на различна стойност на входната променлива. За всеки „клон" се определя класа, към който принадлежат най-голям брой обекти, попаднали в това разклонение. Определянето на грешката се извършва като за всяко разклонение се преброяват обектите, които в действителност принадлежат към различен клас от определения за този „клон".

За всяка входна променлива се генерира различна съвкупност от правила, по едно правило за всяка стойност на променливата. Накрая се определя стойността на грешката за всяка съвкупност от правила. Резултатът на класификационния алгоритъм е съвкупността от правила за тази единствена входна променлива, за която се получава най-малка грешка при класификацията.

Предимства и недостатъци на метода „Генератор на покриващи правила"

Предимства:

- Разбираеми са и лесни за интерпретация;
- Често дават добра точност на предсказване за сложни съвкупности от данни.

Недостатъци:

- Дискретизацията на променливите, които са непрекъснати стойности, води до частична загуба на информация, съдържаща се в първоначалните данни.
- Качеството на генерираните правила до голяма степен зависи от избраната стратегия за дискретизация.
- Понякога е невъзможно предсказването за нови обекти - Понякога генерираните правила не покриват цялото обектно пространство, т.е. за нови обекти, които не са участвали в обучителната извадка, става невъзможно предсказването, тъй като за тях не са открити валидни правила.
- Подобно на метода „Дърво на решенията", те не взимат предвид взаимодействията между входните променливи в изследваната съвкупност от данни. Ако наличието на такива взаимодействия между входните променливи е открито при изследването на данните, то те трябва да бъдат отразени в данните чрез създаване и добавяне на нови променливи от самия изследовател.

1.2.3.3. Метод „К-най-близък съсед”

Методът „К-най-близък съсед” е известен още като Memory Based Reasoning метод, тъй като заключенията по отношение на нови обекти, които трябва да бъдат класифицирани, се правят на базата на минал опит. При този метод се използва фактът, че всеки обект може да бъде представен като точка в многомерното пространство, чрез неговия вектор - съвкупността от стойности, заемани от описващите го променливи. Алгоритъмът приема, че близостта на обектите в многомерното пространство може да се използва за дефиниране на подобността между техните вектори.

Методът „К-най-близък съсед” (kNN – K-Nearest Neighbour) е data mining метод за предсказване. Предсказването се заключава в определяне стойностите на някоя от променливите (предсказвана променлива), описващи обектите от изследваната съвкупност от данни, като се използват известните стойности на останалите (предсказващи) променливи.

Прилагане на метода:

- Определя се броят К на най-близките обекти, които ще бъдат използвани при класификацията;
- Намират се най-близките К на брой съседни обекти до обекта с неизвестен клас на принадлежност, като се използва подходяща функция за измерване на разстоянието между два обекта;
- Определя се класът на новия обект на базата на заключенията, направени по отношение на намерените К най-близки съседи.

Разстоянието е начин за определяне на подобност при методите “най-близък съсед”. В случая трябва да се определя разстоянието между обекти в многомерното пространство, описвани чрез множество променливи от различен тип – цифрови и категорийни. Разстоянието между два обекта се определя по следния начин. Първо се намира разстоянието между двата обекта по отношение на всяка от променливите чрез използване на подходяща функция за разстояние. След това намерените разстояния се комбинират в една обобщена функция за разстояние, която представлява търсеното разстояние между двата обекта.

Четирите най-често използвани функции за разстояние при числови променливи са:

- Абсолютна стойност на разликата: $|A-B|$
- Квадрат на разликата: $(A-B)^2$
- Нормализирана абсолютна стойност:

$$|A-B|/(\text{максималната разлика между } A \text{ и } B) \qquad (1.5)$$

- Абсолютната стойност на разликата между стандартизираните стойности:

(A-mean)/(standard deviation)-(B-mean)/(standard deviation) *(1.6*

което е еквивалентно на

(A-B)/(standard deviation) *(1.7)*

По-трудно е намирането на подходяща функция за разстояние при променливи от категориен тип (напр. какво е разстоянието между стойностите „мъжки" и „женски" на променливата „Пол", или какво е разстоянието между стойностите „червен" и „син" на променливата „цвят"). Най-простата функция за разстояние при категорийни променливи е „идентично на", която приема стойност 1, когато стойностите на изследваната променлива са еднакви за двата обекта, а стойност 0, когато стойностите на изследваната променлива са различни.

След като бъдат намерени разстоянията между стойностите на отделните променливи, те се обединяват в обща функция за разстоянието между два обекта. Най-често се използват три основни начина за сумиране:

- Сумиране (разстояние Манхатън):

$$d_{sum}(A,B) = d_{Var1}(A,B) + d_{Var2}(A,B) + ... + d_{VarN}(A,B) \quad (1.8)$$

- Нормализирано сумиране:

$$d_{norm}(A,B) = d_{sum}(A,B) / \max(d_{sum}) \quad (1.9)$$

- Евклидово разстояние:

$$d_{Euclid}(A,B) = sqrt\left(d_{Var1}(A,B)^2 + d_{Var2}(A,B)^2 + ... + d_{Var3}(A,B)^2\right) \quad (1.10)$$

След като са намерени търсените К на брой най-близки обекти, определянето на търсения клас за нов обект най-често се основава на т.нар. "демократичен принцип". При класификация всеки от съседите гласува за своя клас. Пропорцията гласове за всеки клас съответства на вероятността класифицираният обект да принадлежи към този клас. Когато целта на задачата е да се определи точно един клас за новия обект, се избира класът, получил най-много гласове.

Предимства и недостатъци на метода „К-най-близък съсед"

Предимства:

- Лесно разбираеми за потребителите, особено при наличието на малък брой променливи в изследваните данни.
- Дават възможност за работа с различни типове данни - изображения, свободен текст, географско положение, с които обикновено много трудно се работи при другите методи за предсказване.

- Лесно адаптиране на създадените модели при добавянето на нови обекти към изследваната съвкупност от данни, което прави възможно получаването на „минал опит” за определяне на нови класове или за различно тълкуване и обяснение на съществуващи класове.

Недостатъци:

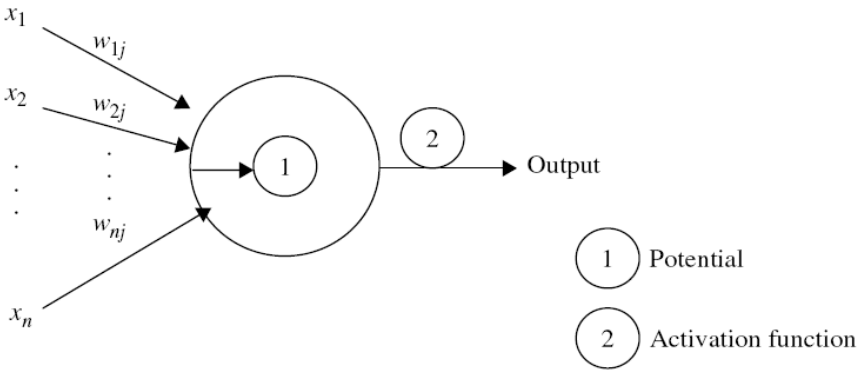
- Необходимост от големи изчислителни ресурси, особено при наличието на голям обем данни и голям брой предсказващи променливи.
- Необходимо е по-продължително време за прилагане на метода, тъй като се налага извършването на изчисления за всеки нов обект (за разлика от методите „Дърво на решенията” и „Невронни мрежи”, които се прилагат бързо при появата на нов обект, тъй като моделът за предсказване вече е генериран през фазата на обучение).
- Необходимо е наличие на много големи обеми от исторически данни, за да се гарантира откриването на К-най-близки съседни обекти.
- Трудности при избор на подходящо число К и откриването на подходящи функции за разстояние при по-сложни типове променливи, което често се осъществява чрез многократни проби и грешки, както и чрез интуитивно мислене.

1.2.3.4. Метод „Невронни мрежи”

Невронните мрежи представляват особен интерес, защото предлагат средства за ефективно моделиране на големи и сложни проблеми, при които може да участват голям брой предсказващи променливи да съществуват множество взаимодействия. Невронните мрежи могат да бъдат използвани както за описание на данните, така и за решаване на задачи за предсказване. Първоначално са възникнали в областта на машинното обучение с цел да имитират нервно физиологичната дейност на човешкия мозък чрез комбинация от елементарни изчислителни елементи (неврони), работещи в една силно свързана система.

Невронната мрежа е съставена от множество елементарни изчислителни елементи, наречени неврони, свързани помежду си чрез претеглени връзки и организирани в слоеве. Формално процесите се описват по следния начин. Невронът j с прагова стойност θ_j, получава входни сигнали $x=[x_1,x_2,...,x_n]$ от елементите от предишния слой, с които е свързан. Всеки сигнал е свързан с тегло $w_j=[w_{1j},w_{2j},...,w_{nj}]$, определящо неговата важност.

Невронът обработва едновременно входните сигнали, техните тегла и праговата стойност чрез т.нар. комбинативна функция (combination function). Комбинативната функция произвежда стойност, наречена потенциал (potential) или нетен входен сигнал (net input). Активираща функция (activation function) трансформира потенциала в изходен сигнал. На Фиг.1.5 схематично е представена работата на един неврон.



Фиг.1.5: Представяне действието на неврон в невронна мрежа

Комбинативната функция обикновено е линейна, затова потенциалът е претеглена сума на входните сигнали, умножени по теглата на съответните връзки. Тази сума се сравнява с праговата стойност. Потенциалът на неврон j се определя чрез следната линейна комбинация:

$$P_j = \sum_{i=1}^{n} (x_i w_{ij} - \theta_j) \tag{1.11}$$

За да се опрости изборът за потенциала, праговата стойност може да се абсорбира чрез добавяне на още един входен сигнал с постоянна стойност $x_0=1$, свързан с неврона j чрез тегло $w_{0j} = -\theta_j$:

$$P_j = \sum_{i=0}^{n} (x_i w_{ij}) \tag{1.12}$$

Изходният сигнал y_j на неврон j, се получава чрез прилагане на активиращата функция към потенциала P_j:

$$y_j = f(x, w_j) = f(P_j) = f\left(\sum_{i=0}^{n} x_i w_{ij} \right) \tag{1.13}$$

Величините x и w във функцията $f(x,w_j)$ са векторни величини.

Един от елементите, който трябва да бъде определен при дефиниране модела на невронната мрежа, е активиращата функция. Обикновено се използват три типа активираща функция: линейна (linear), стъпаловидна (stepwise) и сигмоидална (sigmoidal).

Линейната активираща функция се дефинира по следния начин:

$$f(P_j) = \alpha + \beta P_j \tag{1.14}$$

където P_j е реално число, а α и β са реални константи; в частния случай, когато α=0 и β=1, се получава идентичната функция (identity function), която обикновено се използва, когато моделът изисква изходният сигнал на неврона да е равен на неговото активиращо ниво (неговия потенциал).

Стъпаловидната активираща функция се дефинира като:

$$f(P_j) = \begin{cases} \alpha, P_j \geq \theta_j \\ \beta, P_j < \theta_j \end{cases} \quad (1.15)$$

В този случай активиращата функция може да заема само две стойности в зависимост от това дали потенциалът превишава или не праговата стойност θ_j. При $\alpha=1$, $\beta=0$ и $\theta_j=0$ получаваме т.нар. знакова активираща функция (sign activation function), която приема стойност 0, когато потенциалът е отрицателен, и стойност +1, когато потенциалът е положителен.

Сигмоидалната, или S-образна активираща функция, е може би най-често използвана, тъй като е нелинейна и лесно разбираема. При нея винаги се получава положителен изходен сигнал в интервала [0,1]. *Сигмоидалната активираща функция* се определя като:

$$f(P_j) = \frac{1}{1+e^{-\alpha P_j}} \quad (1.16)$$

където α е положителен параметър, който определя наклона на функцията.

Архитектура на невронната мрежа

Невроните на невронната мрежа са организирани в слоеве. Тези слоеве могат да бъдат три вида: входящ, изходящ или скрит. Входящият слой получава информация само от външната среда, като на всеки неврон обикновено съответства входна (предсказваща) променлива. Във входящия слой не се извършват никакви изчисления, той предава информация към следващия слой. Изходящият слой произвежда крайните резултати, които се подават от мрежата навън от системата. Всеки от неговите неврони съответства на стойности на предсказваната променлива. Между входящия и изходящия слоеве може да има един или повече междинни слоеве, наречени скрити, тъй като те не влизат в директен контакт с външната среда. Тези слоеве се използват изцяло за аналитични цели, тяхната основна функция е да открият съществуващите взаимовръзки между входните и изходната променливи.

Под “архитектура” на невронната мрежа се разбира нейната организация: брой слоеве, брой елементи (неврони) принадлежащи на всеки слой, както и начина, по който са свързани елементите. За класификация на възможните топологии на една невронна мрежа се използват четири признака:

- Степен на разграничаване (обособяване) на входящия и изходящия слоеве
- Брой слоеве
- Посока, в която се извършват изчисленията
- Тип връзки между елементите

Невронните мрежи с повече от един слоеве с претеглени неврони, съдържащи един или повече скрити слоеве, се наричат многослойни персептрони (multilayer perceptrons) и те са най-често използваните невронни мрежи.

Различните информационни потоци водят до получаването на различни типове невронни мрежи. При невронни мрежи с прав поток (feedforward networks) информацията се придвижва само в една посока, от входящия слой през скритите слоеве към изходящия слой, и няма обратни цикли. При невронни мрежи с обратна връзка (feedback networks) информацията се придвижва в двете посоки – от входящия към изходящия слой и обратно.

Ако всеки елемент от един слой е свързан с всички елементи от следващия слой, мрежата се нарича напълно взаимно свързана (totally interconnected); ако всеки елемент е свързан с всеки елемент от всеки слой, мрежата се нарича напълно свързана (totally connected).

При изграждането на невронни мрежи или потребителят, или самият софтуер трябва да изберат броя на възлите и скритите слоеве, активиращата функция, ограниченията за теглата. За тази цел не съществуват общовалидни правила, в повечето случаи се налага експериментиране с тези параметри.

Особености при използването на невронни мрежи

За да бъде ефективно и успешно прилагането на невронни мрежи, е необходимо предварително изследване на данните. Например, тези модели са особено чувствителни към лисващи стойности и наличието на шум в данните. Освен това, прилагането на невронните мрежи изисква сериозна предварителна подготовка на данните, включваща кодиране и трансформиране на променливите, както и намаляване на тяхната размерност (особено при големи съвкупности от данни с много входни променливи).

Тъй като невронните мрежи работят само с цифрови променливи, се налага кодиране на качествените променливи. Количествените променливи се представят с един неврон, а качествените променливи се представят в двоичен вид чрез използването на няколко неврона за всяка променлива (броят на невроните е равен на броя на стойностите,

които приема всяка променлива). На практика, не е необходимо броят на невроните, представящи дадена променлива, да е точно равен на нейните нива. Препоръчително е да се елиминира едно от нивата, и съответно един неврон, тъй като стойността на този неврон е еднозначно определена от останалите.

След кодирането на променливите предварителният анализ може да покаже необходимост от някакъв тип трансформиране, например стандартизиране на входните променливи (с цел претеглянето им по подходящ начин, за да се предотврати по-силното влияние на променливи с по-големи стойности). Ако мрежата е била обучавана с трансформирани входни или изходни променливи, когато се използва за предсказване, трябва изходните сигнали да се възстановят съгласно оригиналната ска́ла на измерване.

Една от най-важните форми на предварителна подготовка на данните е намаляване размерността на входните променливи. Най-простият подход е да се елиминират част от входните променливи на базата на експертна оценка. Други подходи включват използването на линейни или нелинейни комбинации на оригиналните променливи като входни сигнали за мрежата.

Друго важно решение, което трябва да се вземе при прилагането на невронни мрежи, е изборът на архитектура на мрежата, тъй като тя има съществено влияние върху тоността на предсказване. Много от алгоритмите, работещи с невронни мрежи, съдържат оптимизация на архитектурата като част от обучаващия процес.

При наличието на специфична архитектура на невронната мрежа, основна цел при генерирането на модела за предсказване е намирането на теглата на връзките. Избраният алгоритъм изчислява теглата чрез използване на обучаваща извадка от наличните данни, като се минимизира функцията на грешката между реалните и предсказаните стойности (тя може да е класическа функция за разстояние, напр. Евклидово разстояние, или класификационна грешка). Един от най-често използваните методи за обучение е методът на обратно разпространение (back propagation).

Алгоритъмът за обучение на модела чрез обратно разпространение работи по следния начин. Стойностите на входните променливи се предават чрез входящите неврони и първоначално избрани тегла на връзките през скритите слоеве до изходящия слой, където се изчисляват стойностите на изходящите неврони. Определя се грешката, получена в изходящите неврони, като разлика между изчислената стойност и действителната стойност (която се съдържа в обучаващата извадка). Следва обратно разпространение (back propagation), грешката от изходящите неврони се предава обратно към скритите слоеве, и се използва от алгоритъма за настройка на теглата на връзките. Основната цел е минимизиране на грешката. Този процес се повтаря повтаря за всеки ред от обучаващата извадка. Всяко преминаване през всички редове на обучаващата извадка се нарича „епоха” (epock). Обучаващата извадка се използва многократно, докато

грешката престане да намалява. В този момент се счита, че невронната мрежа вече е обучена да открива закономерностите в обучаващата извадка.

За да се предотврати плътното прилягане (пренагаждане) на невронната мрежа към данните от обучаващата извадка (overfitting), трябва да се прецени кога да се прекрати обучението. В някои случаи се използва допълнителна тестова извадка, чрез която периодично се проверява невронната мрежа по време на обучението. Докато намалява грешката при тестовата извадка, обучението продължава. Ако се повиши грешката при тестовата извадка, независимо че грешката при обучаващата извадка все още продължава да пада, обучението се прекратява, за да не се получи пренагаждане (overfitting) на невронната мрежа.

Съществуват различни варианти за спиране работата на обучаващия алгоритъм: да се спре след достигане на определен брой итерации; да се спре след изразходването на определен изчислителен ресурс (CPU usage); да се спре, когато функцията на грешката падне под предварително зададена стойност; да се спре, когато разликата между две последователни стойности на функцията на грешката е по-малка от предварително зададена стойност; да се спре, когато грешката на предсказване или класификация, измерена по отношение на подходяща валидираща съвкупност от данни, започва да нараства (ранно спиране), подобно на подрязване на дърво на решенията. По принцип не е възможно да се установи кой е най-добрият алгоритъм; работата на алгоритмите се променя при решаването на различни задачи.

Предимства и недостатъци на невронните мрежи

Предимства:

- Ефективен метод за създаване на модели с голяма точност на предсказване.
- Справят се успешно при наличието на взаимовръзки между входните променливи, за разлика от дървовидните модели, при които тези взаимодействия трябва да се открият от изследователя и да се добавят в наличните данни.
- Много бързо правят предсказвания с висока точност.

Недостатъци:

- Изискват големи изчислителни ресурси и време за генериране на модел, но лесно могат да бъдат реализирани на паралелни компютри, което позволява едновременно извършване на изчисленията във всички неврони на мрежата.
- Трудно се интерпретират и не предоставят категорична обосновка за предсказванията, които правят.
- Преспециализират се бързо, затова трябва да се вземат сериозни мерки за предотвратяване на този проблем.

- Чувствителени са към липсващи стойности в данните, тъй като това може силно да влоши работата на алгоритъма, което води до получаването на модел с влошена точност на предсказване.
- Предварителното прилагане на неподходящи методи за попълване на липсващите стойности също може да доведе до влошаване качеството на модела.
- Необходимост от активното участие на потребителя при използването на този метод.
- Необходима е предварителна подготовка на данните и избор на топология на мрежата.

Предимствата и недостатъците на четирите разгледани метода за класификация са обобщени и представени в Табл.1.1.

Таблица 1.1: Сравнение на избраните методи за класификация

	Алгоритми за класификация			
	Дърво на решенията	**Невронна мрежа**	**К-най-близък съсед**	**Класификационни правила**
Разбираемост, интерпретация и използване	+ Разбираеми и лесни за интерпретация от потребителите. Обученият модел се прилага бързо по отношение на нови обекти.	+ Създават се модели с голяма точност на предсказване. Обученият модел се прилага бързо по отношение на нови обекти. - Трудно се интерпретират. Не предоставят категорична обосновка за предсказванията.	+ Лесно разбираеми за потребителите, особено при малък брой променливи. - Необходимо е по-продължително време за прилагане на метода, тъй като се налага извършването на изчисления за всеки нов обект	+ Разбираеми и лесни за интерпретация от потребителите. Често дават добра точност на предсказване за сложни съвкупности от данни.
Гъвкавост	+ Приложими при големи обеми данни и при голям брой атрибути. Могат да работят с различен тип променливи – непрекъснати, дискретни.	- Работят само с променливи от числов тип, затова се налага предварителното преобразуване на номиналните променливи в цифрови.	+ Дават възможност за работа с различни типове данни - изображения, свободен текст, географско положение и др.	+ Често дават добра точност на предсказване за сложни съвкупности от данни.
Изчислителни ресурси	- Необходими големи изчислителни ресурси, особено при по-големи обеми данни.	- Изискват големи изчислителни ресурси и време за генериране на модел. + Могат да бъдат реализирани на паралелни компютри.	- Необходимост от големи изчислителни ресурси, особено при наличието на голям обем данни и голям брой променливи.	+ Не се изискват много големи изчислителни ресурси.

Сложност	- Могат да станат много сложни, особено при анализирането на огромни масиви от данни.	- Необходимост от активното участие на потребителя за създаването на подходящ модел.	- Трудности при избор на подходящо число K и откриването на подходящи функции за разстояние при по-сложни типове променливи.	+ По-опростени алгоритми, въпреки че сложността се увеличава при големи обеми данни и много променливи.
Зависимост от изследваните данни	- Дори малка промяна в данните може да доведе до съществена промяна в структурата на дървото. Не взимат предвид взаимодействията между входните променливи и те трябва да бъдат отразени в данните чрез създаване и добавяне на нови променливи от самия изследовател.	- Преспециализират се бързо, затова трябва да се вземат мерки за предотвратяване на този проблем. Чувствителени са към липсващи стойности в данните. + Справят се успешно при наличието на взаимовръзки между входните променливи.	+ Лесно адаптиране на създадените модели при добавянето на нови обекти към изследваните данни. - Необходимо е наличие на много големи обеми от исторически данни, за да се гарантира откриването на K-най-близки съседни обекти.	- Невъзможно е предсказването на нови обекти, които не са участвали в обучаващата извадка и за които няма открити покриващи правила. Не взимат предвид взаимодействията между входните променливи и те трябва да бъдат отразени в данните чрез създаване и добавяне на нови променливи от самия изследовател.

Изводи: Разгледани са предимствата и недостатъците на четири метода за класификация: „Дърво на решенията", „Генератор на правила", „K-най-близък съсед" и „Невронни мрежи", които ще се използват в дисертацията.

1.2.4. Оценка и сравнение на Data Mining модели за класификация (обучени класификатори)

При решаването на задачата за класификация е необходимо да се разработят различни Data Mining модели (обучени класификатори) и накрая да се избере този от тях, при който се получава най-висока точност на предсказване (Vercellis, Guidici, etc.). Различни модели за предсказване могат да бъдат получени чрез използване на различни методи за класификация, напр. „Генератор на правила", „Дърво на решенията", „Невронни мрежи" и т.н., както и чрез промяна на параметрите на алгоритмите, с които се генерират моделите.

Критерии за сравнение на методите за класификация

Методите за класификация обикновено са оценявани и сравнявани на базата на следните критерии:

- Точност на предсказване (Predictive accuracy) – способността на модела правилно да предсказва класа за нови обекти (неизползвани при обучението на модела).

- Скорост (Speed) - необходимите изчислителни ресурси при генерирането и прилагането на модела.
- Устойчивост (Robustness) – устойчивостта на модела при наличие на шум, липсващи данни, различни типове променливи и др.
- Скалируемост (Scalability) - възможността на генерирания модел да работи ефективно при наличието на големи обеми данни.
- Разбираемост (Interpretability) - степента на лесно разбиране и използване на моделите от потребителите.

Методи за формиране на оценките на генерираните модели за класификация

Оценката на генерираните модели за класификация може да се извърши чрез използването на няколко различни метода:

- *Разделяне на общата съвкупност от данни на обучаваща и тестова извадка (Holdout method)*

  При този метод на оценка, обектите/записите от общата съвкупност от данни се разделят на две части – едната част се използва за обучението на модела и се нарича *обучаваща извадка*, а другата – за тестване на модела и се нарича *тестова извадка*. Обикновено, обектите в обучаващата извадка се избират на случаен принцип, останалите обекти попадат в тестовата извадка. При някои алгоритми се постига еднакво пропорционално разпределението на обектите от различните класове в обучаващата и тестовата извадки. Големината на обучаващата извадка може да варира в зависимост от големината на общата съвкупност от данни. Най-често обучаващата извадка включва от 1/2 до 2/3 от общия брой записи в изследваната съвкупност от данни.

  Точността на оценявания класификационен алгоритъм в този случай зависи от избора на тестова извадка, затова може да се получи надценяване или подценяване спрямо действителната точност на предсказване, при промяна на тестовата извадка. За да се получи по-устойчива оценка, и съответно по-надеждно да се оцени точността на класификационния модел, се използват някои от стратегиите, описани по-долу.

- *Многократно случайно избиране на обектите в обучаващата и тестовата извадки* (*Repeated random sampling*)
  При този метод процедурата по разделянето на общата съвкупност от данни на две части – обучаваща и тестова извадки, се извършва многократно. При всяко повторение двете извадки се избират на случаен принцип и се определя точността на класификация на получения модел за съответната итерация.

Накрая, точността на модела се определя като средно-ариметично на получените оценки за отделните итерации.

Броят на повторенията може да бъде определен чрез предварително прилагане на различни статистически техники като се взима предвид големината на изследваната обща съвкупност от данни.

Този метод в много случаи се предпочита пред Holdout метода, тъй като при него може да бъде получена по-надеждна оценка за точността на класификационните модели. Но, при него не се контролира броят на попаденията на даден обект от общата съвкупност от данни в обучаващата и тестовата извадки. Това може да окаже неблагоприятен ефект върху генерираните класификационни правила и оценката на точността, особено при наличието на доминиращ обект, за който една или повече променливи приемат необичайни стойности (изключения).

- *Крос-валидиране* (*Cross-Validation*)

  Методът на крос-валидирането е разновидност на Repeated Random Sampling метода, като при него се гарантира еднакъв брой участия на всеки обект от изследваната обща съвкупност от данни в обучаващата извадка и едно единствено участие – в тестовата извадка.

  В практиката най-често се използва т.нар. tenfold cross-validation, което означава, че общата съвкупност от данни се разделя на 10 части и алгоритъмът се изпълнява 10 пъти, като всеки път 9 от 10-те части се използват като обучаваща извадка, а останалата 1 част се използва за тестова извадка.

Мерки за оценка на генерираните модели за класификация

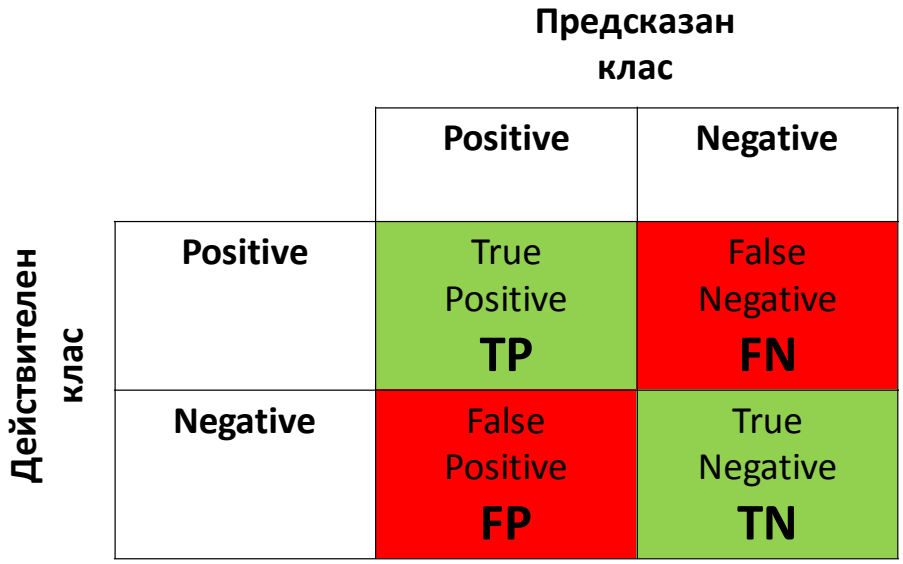
Най-често използваните мерки за оценка на генерираните модели за класификация включват Матрица на класификация (Confusion Matrix), ROC крива, Lift диаграма и др.

Матрица на класификация (Confusion Matrix)

Предсказването на бинарна променлива (променлива, приемаща само *2* възможни стойности), е най-простия и най-често срещан случай при решаването на Data Mining задачи за класификация. Затова именно за този случай се разглежда матрицата на класификация по-долу (матрица с размерност *2x2*). Когато предсказваната величина не е бинарна променлива (т.е. може да приема *N* различни стойности), матрицата на класификация се получава по аналогичен начин и е с размерност *NxN*.

При наличието на бинарна предсказвана променлива класификационният модел разпределя всеки обект в един от два възможни класа, например Positive и Negative,

съответно вярно (True) или невярно (False). Това води до четири възможни варианта за класификация на всеки обект: True Positive (TP) – действителен клас Positive вярно предсказан като клас Positive, True Negative (TN) – действителен клас Negative вярно предсказан като клас Negative, False Positive (FP) - действителен клас Negative невярно предсказан като клас Positive, и False Negative (FN) - действителен клас Positive невярно предсказан като клас Negative. Тези възможности могат да бъдат представени във вид на матрица (Фиг.1.6), наричана *Confusion Matrix* или *Contingency Table*.



Фиг.1.6: Матрица на класификация (Confusion Matrix)

В елементите от главния диагонал на матрицата се намират правилно предсказаните обекти (TP и TN), а в останалите елементи на матрицата – грешно предсказаните обекти (FP и FN).

На базата на матрицата на класификация се определят множество различни мерки, които се използват за оценка на генерираните модели:

- *Точност* (*Classification Accuracy*) и *грешка* (*Classification Error*) на класификация

  При решаването на Data Mining задача за класификация е естествено да се измерва работата на класификатора чрез стойността на грешката или чрез точността на предсказване. Тези две мерки са взаимно свързани.

  Оценката се извършва чрез сравняване на предсказания клас за всеки обект от тестовата извадка с неговата действителна стойност. Ако тези две стойности съвпадат – се отчита правилно предсказване, ако не съвпадат – се отчита грешка.

Точността на класификация се определя като отношение между броя обекти, чийто клас е правилно предсказан от класификатора, и общия брой обекти в тестовата извадка. Често се представя в %.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{1.17}$$

Класификационната грешка се определя като съотношение между броя обекти, чийто клас е грешно предсказан от класификатора, и общия брой обекти в тестовата извадка. Често се представя в %.

$$Error\ Rate = \frac{FP+FN}{TP+FP+TN+FN} \tag{1.18}$$

- Точност на предсказване на всеки отделен клас

TP Rate показва каква част са обектите, правилно класифицирани в клас Positive, от общия брой обекти, които действително принадлежат към клас Positive, т.е. показва каква част от обектите в клас Positive са разпознати от класификационния модел. Тази мярка е еквивалентна на мярката Recall (представена по-долу):

$$TP\ Rate = \frac{TP}{TP+FN} \tag{1.19}$$

FP Rate показва каква част са обектите, които са неправилно класифицирани в клас Positive, от общия брой обекти, които в действителност принадлежат към клас Negative:

$$FP\ Rate = \frac{FP}{FP+TN} \tag{1.20}$$

Аналогично се определят *TN Rate* и *FN Rate*:

$$TN\ Rate = \frac{TN}{TN+FP} \tag{1.21}$$

$$FN\ Rate = \frac{FN}{FN+TP} \tag{1.22}$$

- *Precision, Recall* и *F-Measure*

Мярката *Precision* показва каква част представляват обектите, които действително принадлежат към клас Positive, от общия брой обекти, които са класифицирани от модела в клас Positive:

$$Precision = \frac{TP}{TP+FP} \tag{1.23}$$

Мярката *Recall* е еквивалентна на мярката *TP Rate*:

$$Recall = \frac{TP}{TP+FN} = TP\ Rate \tag{1.24}$$

Мярката *F-Measure* е комбинирана мярка, обединяваща *Precision* и *Recall*:

$$F - Measure = \frac{2}{1/Precision+1/Recall} \tag{1.25}$$

Kappa Statistic

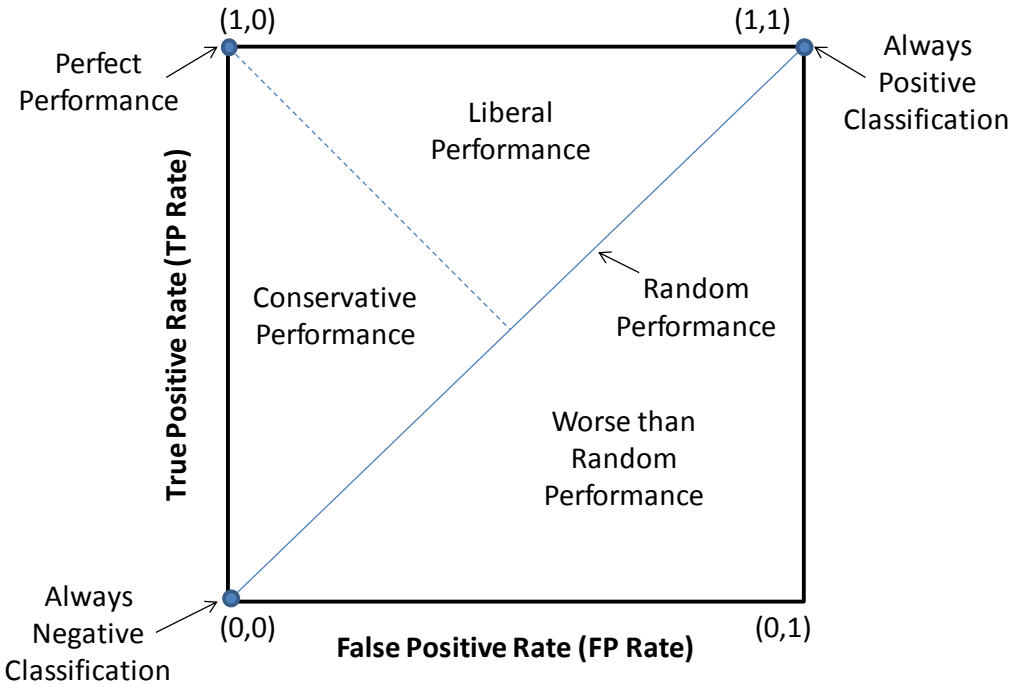
Друга мярка, която често се използва за оценка на класификационните модели, е *Kappa Statistic,* и измерва степента на съгласуваност между предсказвания и действителния клас. Стойността на Kappa Statistic е в интервала [0;1]. Когато стойността на Kappa Statistic е 1.0, това означава, че има пълно съгласуване между предсказвания и действителния клас.

ROC крива

Receiver Operating Characteristic (ROC) кривите са визуални средства, произхождащи от теорията на комуникациите, и използвани за определяне на оптималните параметри на филтрите, отделящи полезните сигнали от смущенията. В областта на Data Mining ROC кривите се използват за визуална оценка на точността на класификаторите, както и за сравняване на различни класификационни модели.

ROC кривата е двумерна диаграма, която се получава чрез изобразяване на True Positive Rate (TP Rate) по вертикалната ос и False Positive Rate (FP Rate) по хоризонталната ос (Фиг.1.7). С диагонала, свързващ долния ляв ъгъл и горния десен ъгъл на диаграмата, се изобразява работата на случаен класификатор, т.е. на класификационен модел, при който броят на правилно предсказаните обекти от клас Positive е равен на броя на неправилно предсказаните обекти от клас Positive (TP Rate=FP Rate). Точката (0,1) на ROC кривата (Фиг.1.7) представлява идеалният класификатор, тъй като се отнася за модел, който не прави грешка при предсказването (FP Rate=0 и TP Rate=1). Точка (0,0) съответства на класификатор, който предсказва клас Negative за всички обекти (т.е. винаги предсказва клас Negative), а точка (1,1) съответства на класификатор, който предсказва клас Positive за всички обекти (т.е. винаги предсказва клас Positive). Класификатор, който попада в областта вдясно от случайния класификатор, работи по-слабо от него. Областта над случайния класификатор се разделя на две части – в лявата част попадат класификаторите, които работят консервативно, а в дясната част – класификаторите които работят по-либерално. При консервативните класификатори стойностите на FP Rate са ниски, т.е. те правят минимален брой грешки при предсказването на клас Positive. При либералните класификатори стойностите на TP Rate са по високи, но са по високи и стойностите на FP

Rate, т.е. тези класификатори предсказват правилно по-голям брой обекти от клас Positive, но правят и повече грешки при предсказването на клас Positive.



Фиг.1.7: ROC Крива

Повечето класификатори съдържат параметри, които могат да бъдат променяни. Целта на настойката на тези параметри е да се подобри работата на модела, т.е. да се намери оптималното съотношение между TP Rate и FP Rate.

Чрез ROC кривата визуално се изобразява информацията от няколко последователно получени матрици на класификация за един и същи класификационен модел. За да се получи траекторията на ROC кривата за даден класификационен модел, трябва да се използват двойки стойности (FP Rate, TP Rate), получени за различни стойности на параметрите на модела.

Често като мярка за оценка работата на класификационните модели се използва площта над ROC кривата - *ROC Area*, приемаща стойности между 0 и 1. Колкото по-голяма е площта над ROC кривата, толкова по-добър е полученият класификационен модел. Ако ROC Area<0.5, то оценяваният модел работи по-лошо от случайния класификатор (с ROC Area=0.5).

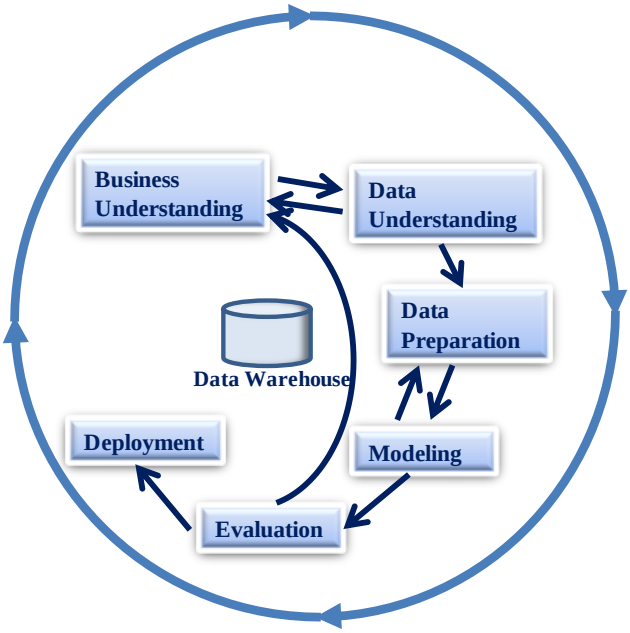
Изводи: Разгледани и предложени са множество различни метрики, които се използват в дисертацията за оценка на обучените класификатори, както следва: *Точност на класификация; Грешка при класификацията; Матрица на класификация*; *Параметрите*

TP Rate, FP Rate, Precission, Recall и F-measure; *Kappa Statistic*; *ROC Area*; *Model Correct/Incorrect Ratio*; *Model/Naïve Correct/Incorrect Ratio.*

1.2.5. Съществуващи подходи за реализация на Data Mining проекти

Извличането на знания от големи обеми данни е сложен процес, чиято успешна реализация изисква прилагането на системен подход, който да дефинира последователността от необходимите етапи и дейности, които трябва да бъдат осъществени от изследователите. До тези изводи са стигнали много от фирмите, разработващи и предлагащи на пазара софтуерни решения за Data Mining, както и различни консултантски компании, затова са създадени различни подходи за реализация на Data Mining проекти, които стъпка по стъпка да направляват потребителите (особено тези, които нямат голям опит в тази област) към постигането на крайни добри резултати. Някои от най-известните подходи са:

- ***Подходът 5A's – Assess, Access, Analyze, Act and Automate*** – предложен от фирмата SPSS (SPSS Clementine), включва 5 основни етапа: разбиране на данните и предметната област (Assess); използването на технологии за ефективно събиране и намиране на данни (Access); задълбоченото анализиране на данните чрез прилагане на различни Data mining методи и средства и взимане на информирани решения (Analyze); предприемане на съответни действия въз основа на взетите решения (Act); внедряване и къстамизация в съответствие с нуждите на различните потребители (Automate).
- **Подходът SEMMA: Sample, Explore, Modify, Model, Assess** – предложен от SAS Institute, друга водеща фирма, разработваща софтуерни решения за Data Mining, също включва пет етапа като фокусирането е основно върху техническите дейности, които се извършват при реализацията на Data Mining проекти. Включва: селектиране и извличане на данните (Sample); изследване на данните (Explore); предварителна подготовка и модифициране на данните (Modify); генериране на модели чрез прилагане на различни Data Mining методи върху данните (Model); оценка на получените модели (Assess).
- ***Подходът CRISP-DM – Cross-Industry Standard Process for Data Mining*** – създаден през 1999-2000г. от консорциум от разработчици, консултанти и потребители на Data Mining софтуер – включва шест етапа и представлява цикличен процес с множество обратни връзки (Фиг.1.8).



Фиг.1.8: Подходът CRISP-DM – Cross-Industry Standard Process for Data Mining

Съгласно този модел, един цикъл на изследвания в областта на Data Mining може да се представи като процес, състоящ се от шест основни етапа. Последователността на етапите не е строга: често се налага връщане на предишен етап с цел подобряване на резултатите от следващ етап.

- ***Разбирането на проблемната област*** (Business Understanding) е началният етап, който се фокусира върху разбирането на целите на изследванията и формулиране на изисквания от гледната точка на потребителя. След завършването на етапа, получените знания трябва да бъдат използвани за дефиниране на задачите за извличане на закономерности и за съставяне на предварителния план за постигане на целите.
- ***Разбирането на данните*** (Data Understanding) започва с първоначално събиране на данни и продължава с дейностите, целящи задълбочаване на знанията на изследователя за естеството на данните. На този етап е необходимо да бъдат идентифицирани проблеми, свързани с качеството на данни, да бъде получено първоначално мнение за характера на данните, да бъдат намерени интересните подмножества на данните, за да могат да бъдат формирани първоначални хипотези за скритата в данните информация.
- ***Подготовката на данните*** (Data Preprocessing) включва всички дейности по създаване от първоначални "сурови" данни на крайното множество от данни (т.е. данни, които ще бъдат използвани от моделиращите средства). Етапът на подготовката на данни често се налага да бъде изпълняван многократно и в различни моменти от изпълнението на data mining анализа.

Подготовката на данни включва изчистване на данните (data cleaning), интегриране на данните (data integration), трансформиране на данните (data transformation) и намаляване обема на данните (data reduction). Чрез процедурите за *изчистване на данните* се попълват липсващи стойности, коригират се грешки, заглажда се „шума", премахват се несъответствия. При *интегрирането на данните* се обединяват данни от различни източници с цел да се създаде единна кохерентна база данни, върху която да се прилагат техниките и методите за извличане на закономерности. Чрез процедурите за *трансформиране*, данните се преобразуват във вид удобен за прилагане на избраните методи и средства за извличане на знания. Това се налага поради спецификата на различните методи и особеностите на различните типове данни. Техниките за *намаляване обема на данните* включват агрегиране на кубове данни, компресиране на данни, дискретизация и др., чрез които се постига намалява обема на данните, без да се загубва важно съдържание.

- ***Етапът на моделиране*** (Modelling) се състои в избора и прилагането на различни методи за моделиране, целящи извличане на закономерности от данните. Параметрите на моделите се калибрират до оптимални стойности. Тъй като някои модели поставят специфични изисквания по отношение формата на данните, на този етап често се налага връщане към етапа за подготовка на данните.
- ***Оценка на моделите*** (Evaluation) се прави с цел по-задълбочено разбиране на създадените модели от гледната точка не само на изследователя, но и на потребителя. Важно е да бъдат внимателно прегледани всички стъпки, изпълнени при създаването на конкретния модел, за да се получи увереност, че те постигат поставените цели.
- ***Прилагане на моделите*** (Deployment). Готовите модели могат да се използват по два основни начина. Анализаторът може да препоръча предприемането на конкретни действия на базата на заключения от изградения модел и получените резултати, или моделът може да се приложи към други масиви от данни.

Изводи: Разгледани са различни съществуващи подходи за реализация на Data Mining проекти и е избран подходът CRISP-DM – Cross-Industry Standard Process for Data Mining, който ще се използва в настоящата дисертация.

1.3. Обзор на Извличане на знания от данни в сферата на образованието (Educational Data Mining)

Съвременните университети са организации, в които се събират и съхраняват огромни обеми от данни, свързани със студентите, с организацията и управлението на образователния процес, както и с цялостното администриране на дейността.

Анализирането на тези уникални типове данни чрез прилагането на аналитични методи и средства за извличане на знания (data mining) могат съществено да допринесат за подпомагане на цялостното управление чрез ускоряване процесите на взимане на информирани и правилни управленски решения, нарастване на аналитичните възможности, подобряване качеството на образователния процес чрез предоставяне на възможности за анализиране работата на преподаватели и студенти.

Данните, с които разполагат организациите в образователния сектор, произлизат от два основни типа обучение – традиционно и дистанционно (Romero&Ventura, 2007). В институциите, които предлагат традиционно обучение, данни се събират при приемането на нови студенти, при организирането и осъществяването на обучението, за целите на управлението и др., и обикновено са организирани в бази данни или склад за данни (data warehouse). Основните източници на данни при дистанционното обучение са логовите файлове, съдържащи информация за начина на използване на уеб-базираните образователни системи от обучаемите. Прилагането на data mining методи и средства се различава за двата типа обучение, тъй като се използват различни източници на данни и информационни системи, различни са и поставяните цели и решаваните проблеми. В настоящия дисертационен труд изследванията са насочени към прилагане на data mining методи и средства за извличане на знания от данни, получени при традиционно обучение, което е основна и най-често срещана форма на обучение в българските университети и висши училища.

Приложението на data mining методи и средства може да бъде насочено към различни потребители в образователните институции (Romero&Ventura, 2007) – студенти, преподаватели, мениджъри, администратори. Студентите могат да бъдат подпомагани като им бъдат препоръчвани различни информационни ресурси, дейности, задачи или дори различни пътища за овладяване на знанията. Преподавателите биха могли да получават по-обективна обратна връзка и осъществяването на задълбочен анализ на предлагания образователен процес, което би им помогнало да подобрят съдържанието на своите курсове, да подберат по-ефективни методи за поднасяне на учебния материал, да прилагат диференциран подход към студенти с различни характеристики на учене, различни възможности и особености. Предимствата за администраторите са свързани с получаването на необходимата информация за взимането на управленски решения в подходящия момент и основаваща се на извършените анализи на наличните данни. Мениджърите могат да бъдат подпомагани при осъществяването на техните стратегически задачи, чрез предоставянето на задълбочени анализи, разкриващи съществуващи тенденции и възможности за подобряване ефективността и качеството на управление.

До момента са публикувани две научни статии, в които е направен обзор на научните изследвания в областта на Educational Data Mining. Първата статия излиза през 2007г. и представя анализ за периода 1995-2005г. (Romero&Ventura, 2007), а втората се появява през 2009г., дава исторически преглед за периода след 2005г. и очертава

актуалните проблеми и тенденции в развитието на това ново научно направление (Baker&Yacef, 2009). Проблемите на институциите, работещи в сферата на висшето образование, които най-често представляват интерес за изследователите и нерядко се превръщат в основен фактор за иницииране на data mining проекти, са свързани с привличането на подходящи студенти (targeted marketing), задържане на студенти и предотвратяване на тяхното отпадане от обучение (retention of students), повишаване качеството на образователния процес, управление на кандидат-студентските кампании, подобряване на организационната ефективност, управление на взаимоотношенията с бивши възпитаници (alumni management).

Редица автори предлагат анализи на основните проблеми в образователния сектор, чието решаване може да бъде успешно подпомагано чрез прилагане на Data Mining методи и средства. Още през 2002г. се очертават четири възможни области на приложение с голям потенциал (бивши възпитаници, институционална ефективност, маркетинг и управление на кандидат-студентски кампании), и се обяснява как използването на подобна технология води до спестяване на ресурси и повишаване ефективността на академичната дейност (Luan, 2002, 2004). През 2008г. се предлагат разработени насоки за избор и прилагане на различните data mining методи с цел подпомагане взимането на решения във висшите образователни институции (Delavari et al., 2008). През 2009г. се дискутира непрекъснато нарастващият брой приложения на Data Mining в образованието – кандидат-студентски кампании, академично представяне, web-базирано обучение, запазване на студенти и др. (Nandeshwar&Chandhari, 2009). През 2010г. е публикувана обзорна статия относно използването на data mining методи в сферата на висшето образование в Индия – подобряване функционирането на университетите (управление представянето на студентите, намаляване процента на отпадащи студенти, предоставяне на допълнителна помощ на нуждаещи се студенти, избор на курсове, по-добро разпределение на лекторите), по-ефективно управление на субсидии (дарения от бивши възпитаници, средства за реализиране на изследователски програми, грантове и др.), управление жизнения цикъл на студентите, намаляване на разходите на институцията (Ranjan&Randjan, 2010). Анализ относно потенциални нови приложения на data mining методите в образователния сектор е направен и през 2011г. – организиране на учебните програми, предсказване записването на студентите в дадена програма, предсказване представянето на студентите по време на обучението в университета, откриване на измами при електронен формат на изпитване, откриване на грешки в събраните данни (Kumar&Chadha, 2011).

Използването на data mining методи за предсказване на слабо представящи се студенти, които впоследствие отпадат от обучението, за откриване на факторите, които в най-голяма степен влияят върху отпадането и предприемането на съответни мерки с цел постигане на минимален брой отпаднали студенти, са проблеми, които се разглеждат в голям брой научни статии (Luan 2004, Gao 2005, Herzog 2006, Superby et al. 2006,

Shyamala&Rajagopalan 2007, Wang 2007, Yu et al. 2007, Vandamme et al. 2007, Cortez&Silva 2008, Yu et al. 2010, Kovačić 2010). Много от публикациите са посветени на подобряване представянето на студентите, откриване на факторите за успех, разработване на система за препоръки относно избора на курсове и подобряване на институционалната ефективност (Luan 2002, Minaeli-Bidgoli et al. 2003, Kotsiantis&Pintelas 2004, Kotsiantis et al. 2004, DeLong et al. 2007, Noel-Levitz 2008, Delavari et al. 2008, Kumar&Uma 2009, Vialardi et al. 2009, Dekker et al. 2009, Wook et al. 2009, Ranjan&Ranjan 2010, Kovačić 2010, Ramaswami&Bhaskaran 2010, Kumar&Chadha 2011, Vialardi et al. 2011). Създаването на модели за предсказване, свързани с организирането и провеждането на кандидат-студентски кампании и маркетинг, ориентиран към най-желаните студенти – такива, които се представят най-добре в процеса на обучение в университета, също е често срещата цел при научните изследвания в областта на Educational Data Mining (Ma et al 2000, Antons&Maltz 2006, Noel-Levitz 2008, Nandeshwar&Chaudhari 2009).

Предприетата научноизследователска дейност, представена в настоящия дисертационен труд, е посветена на откриване на информация в съществуващите данни, която да подпомогне ръководството на университета по-добре да опознае своите студенти и да организира по-ефективно своята маркетингова политика. Тези проблеми са били обект на изследване на множество автори, работещи в областта на educational data mining, през последните години. В много публикувани научни статии се дискутира разработването на модели за предсказване представянето на студентите на различни нива, както и сравняване работата на различни Data Mining алгоритми.

Още през 2000г. е публикувано проучване, свързано с използването на Data Mining за откриване на най-слабите студенти с цел включването им в курсове за допълнително обучение (Ma et al., 2000). Решаваната Data Mining задача е „извличане на асоциативни правила", определящи слабите студенти в зависимост от техния пол, регион и успех през предишните години в училище. През 2002г. проблемът с отпадането на студенти от обучението и необходимостта от тяхното ранно идентифициране с цел предприемане на необходимите мерки за подпомагане, както и извършването на промени в маркетинговата политика, е обект на изследване и на Luan (2002). Целта на изследването е да се класифицират студентите в *две групи – продължаващи и отпадащи*, като са използвани няколко различни data mining методи – *клъстеризация, невронни мрежи и дърво на решенията (C5.0 и C&RT алгоритми)* (Luan, 2002). През 2004г. отново Luan представя изследвания, свързани с определянето на типове студенти в зависимост от крайните цели на обучението (усвояване на основни знания и умения, придобиване на специфична квалификация, продължаване на обучението и т.н.) чрез използване на data mining методи за клъстеризация (TwoStep и K-Means алгоритми) (Luan, 2004), както и с ранно откриване на студенти в риск от отпадане чрез използване на data mining методи за класификация (невронни мрежи и дърво на решенията).

През 2003г. е публикувано изследване (Minaeli-Bidgoli et al., 2003), насочено към разработване на модели за предсказване оценките на студентите в Държавния университет на Мичиган чрез решаване на data mining задачата за класификация при три различни варианта на предсказваната променлива/класа (двоична величина - pass/fail; номинална величина с три стойности - low, middle, high; и номинална величина с 9 различни стойности – от 1, съответстваща на най-ниска оценка, до 9, съответстваща на най-висока оценка). Използвани са 227 записа от електронна система за обучение, съдържаща различни характеристики като напр. „брой коригирани отговори", „брой опити за изпълнение на домашна работа" и др. Приложени са различни Data Mining алгоритми, но *най-добри резултати са получени с комбинирани класификатори (classifier ensemble), включващи напр. Дърво на решенията и Невронна мрежа, а получените точности на предсказване са съответно 94% (binary), 72% (3-classes) и 62% (9-classes).*

Kotsiantis et al. (2004) предлагат логическа архитектура за разработването на софтуер за автоматично откриване на студенти, обучавани дистанционно и застрашени от отпадане (Kotsiantis, S., Pintelas, P., 2004). Предвиждат се възможности за използване на различни data mining методи в зависимост от използваните данни и целите на предсказване. Входните параметри, описващи студентите, включват демографски характеристики (напр. пол, възраст, семейно положение) и оценки от обучението (напр. оценка от определено задание), а предсказваната величина е двоична променлива с *две стойности pass/fail*. Използвани са различни методи за класификация (дърво на решенията, невронна мрежа, байесов класификатор, логистична регресия и др.), но най-добри резултати от предсказването са получени чрез метода на Наивна Байесова класификация - 74% точност. Освен това изследователите са установили, че *оценките от училище влияят в по-голяма степен върху класификацията, отколкото демографските фактори* (Kotsiantis, Pierrakeas & Pintelas, 2004).

С анализиране на данни от електронна система за обучение се занимават и Pardos и др. (Pardos et al., 2006), които решават регресионна data mining задача с цел да предсказват оценките на тест по математика за ученици от 8 клас в зависимост от техни индивидуални умения. Най-добри резултати са получени с *алгоритъма Байесови мрежи (Bayesian Networks) – грешка на предсказване 15%.*

Superby, Vandamme и др. (2006) прилагат няколко различни метода за класификация на студентите в три групи в зависимост от нивото на риска от отпадане – дърво на решението, невронни мрежи, линеен дискриминантен анализ (Superby et al., 2006, Vandamme et al., 2007). Резултатите показват, че само 20% от използваните входни променливи оказват съществено влияние върху академичния успех. *Моделите за предсказване, генерирани чрез различните data mining методи, не се отличават с особено висока точност (40.63%-57.35%).* Наблюдават се и различия в поведението на алгоритмите относно предсказването на трите различни групи студенти.

Yu и др. (Yu et al., 2007) използват data mining методи за откриване на факторите, чрез които може да се влияе с цел предпазване на студентите от отпадане през първата година от обучението им в Държавния университет на щата Аризона. Изследването показва, че „cumulated earned hours" (общо натрупани часове) е най-важният фактор, оказващ влияние върху запазването на студентите, а полът и етническият произход не са от съществено значение. Същите автори изследват проблема със задържането на студенти и през 2012 (Yu et al., 2010), използвайки няколко data mining метода - *класификационно дърво на решенията, невронни мрежи и* MARS (multivariate adaptive regression splines) подход.

Cortez и Silva (2008) насочват усилията си към разработването на модел за предсказване оценките по два предмета на ученици в горен курс на обучение, чрез прилагане и сравняване на четири различни data mining алгоритми – дърво на решението, невронни мрежи, Random Forest, Support Vector Machine (Cortez and Silva, 2008). Задачата за предсказване е решена за *три варианта на предсказваната променлива - номинална променлива с две/пет стойности (класификация), и цифрова променлива (регресия).* Получени са сравнително високи точности на класификация, като изследванията са проведени за различни входни параметри – понякога са включвани демографски данни, предишни оценки на учениците и т.нат., в зависимост от наличността на тези данни.

Delavari и др. (Delavari et al., 2008) представят разработването на модели за предсказване успеваемостта на студентите по отношение на индивидуален студент или индивидуален лектор (модел за предсказване) чрез използване на data mining методите „дърво на решенията" и невронни мрежи, създаване на модел за характеризиране на студентите които се записват в даден курс (описателен модел) чрез използване на метода асоциативен анализ, създаване на модел за характеризиране на различните типове преподаватели (описателен модел) водещи даден курс (Delavari et al., 2008) чрез прилагане на метод за клъстерен анализ.

Wook и др. (Wook et al., 2009) сравняват два метода, *невронни мрежи и комбиниран подход включващ разделяне на клъстери и дърво на решенията, за предсказване представянето на студентите, въз основа на резултатите от изпитите, чрез класифицирането им в две групи – успешни и неуспешни* (Wook et al., 2009).

Целта на изследването на Dekker и др. (Dekker et al., 2009) е да се създаде модел за предсказване на успешните и неуспешните студенти. Проучването е осъществено чрез data mining софтуера с отворен код WEKA, като се сравнява работата на алгоритми за класификация, работещи на различни принципи – дърво на решенията, байесов метод, генериране на правила, логистичен метод. Точността на избрания модел е подобрена чрез прилагане на т.нар. cost-sensitive learning (предварително задаване на тегла).

Kovačić (Kovačić, 2010) също разглежда проблема със запазването/отпадането на студентите като създава модели за предсказване на студенти в риск, на базата на техни социално-демографски характеристики (възраст, пол, етническа принадлежност, образование, трудова заетост, инвалидност и др.) и параметри на образователната среда (курсове в учебните програми, блокове курсове и др.). Избраният data mining метод е „дърво на решенията", като са приложени различни Decision Tree алгоритми (използващи различни критерии за нарастване на дървото). Целта на изследването е да се открият най-важните фактори, които оказват влияние върху отпадането на студентите и чрез които може да се повлияе върху тяхното задържане, както и да се изградят профили на типични успешни и неуспешни студенти.

Ramaswami и др. (Ramaswami et al., 2010) се фокусират върху генерирането на предсказващи data mining модели за идентифициране на студенти, които учат по-бавно (slow learners), както и върху изследване влиянието на основните фактори върху успеха им в университета, като използват *метода „дърво на решенията" - популярния CHAID алгоритъм*.

<u>Изводи:</u>

- Осъщественият анализ на публикуваните научни изследвания в областта на Educational Data Mining показва, че предсказването успеха на студентите е много важен и актуален проблем за университетите. Резултатите от провежданите изследвания се използват за различни цели:
    - *Откриват се изоставащи студенти*, които евентуално ще отпаднат от обучение поради слабото си представяне, и на тези студенти се предоставя допълнителна помощ, за да бъдат задържани в университета.
    - *Откриват се добрите студенти* – това са студентите, които се справят най-успешно с обучението и които са най-желани за университета, и именно към такъв тип студенти се насочват бъдещите маркетингови кампании.
    - *Откриват се факторите, които в най-голяма степен влияят върху успеха* или слабото представяне на студентите, и това се превръща в ценна информация, на базата на която се взимат управленски решения за извършването на промени и за подобряване ефективността и качеството на предлаганото обучение.
- *Успехът на студентите се предсказва най-често чрез решаване на задача за предсказване, като*:
    - в повечето случаи *предсказваната величина е номинална променлива*, т.е. решава се задача за класификация
    - в по-редки случаи предсказваната променлива е цифрова, т.е. решава се задача за регресия.
- Най-често използваните методи за обучение на класификатори включват:
    - „Дърво на решенията"

- „Невронни мрежи”
- Байесови методи - Наивен Байесов класификатор, Байесови мрежи
- Логистична регресия.

- Във всички случаи обаче, се прилагат няколко различни метода или няколко различни алгоритми, за да се достигне до създаването на подходящ обучен класификатор, а не се използва един единствен алгоритъм.
- Впечатление прави и внимателното изучаване и подготовка на данните, както и изборът и форматирането на предсказваната величина.
- Често задачата за извличане на знания от данни чрез обучение на класификатори се решава за различни варианти на предсказваната променлива – дискретна величина, приемаща различен брой стойности (напр. 2,3,5,9).
- Представените резултати от изследванията обикновено показват, че при по-малък брой възможни стойности на предсказваната променлива често се получава и по-висока точност на предсказване.

1.4. Обзор на класификацията на радарни цели

Проблемът, свързан с класификацията на откритите радарни цели, занимава радарната научна общност от много време (от Втората световна война). За решаването му се използват предимно различни статистически методи. В теорията на статистическата радиолокация първо са разработени въпросите на откриването, тъй като с тази фундаментална задача започва радиолокацията. Развитието на многоалтернартивните методи за откриване на сигнали при различни техни енергетични и информативни параметри довежда до появата на съвместната задача откриване с класификация на радарните цели. Изборът на решение в случая се основава на характера на промяна на сигнала, отразен от целта. Само в случай, че различните цели предизвикват добре различими промени в отразените сигнали, е възможна тяхната класификация.

В радарите, един от разпространените подходи за съвместно откриване и класификация на сигналите от целите е класификацията с използване на филтрова или корелационна обработка на сигналите. Например, във филтровата (корелационната) класификация се използва информацията за енергетичния спектър на ехо сигнала от конкретна цел. Изгражда се многоканална система за класификация, като всеки от каналите се състои от филтър с последващо решаващо прагово устройство. Филтрите са настроени на параметрите на спектрите на сигналите. В случай, че през тази система преминава сигнал от конкретна известна цел, на изхода на канала, настроен на тази цел, в решаващото му устройство, сигналът ще прескочи прага и ще се извърши класификация на целта.

При радар с непрекъснато излъчване входният сигнал към класификационната многоканална система може да е необработено изходно напрежение на приемника. В този случай многоканалната класификационна система ще се състои от филтри, настроени на различни доплерови честоти, или от филтри, съгласувани с различни спектри на сигнали от целите.

Задачите за откриване и класификация на сигнали могат да се решават с методите за разпознаване на образи, защото радарните сигнали могат да се разглеждат като образи от данни. Тъй като при много методи за разпознаване на образи се използва статистическата теория на решенията, при априорна определеност и неопределеност, затова разпознаването или класификацията на образите се свежда всъщност до метода за многоалтернативно изпитване на хипотези. Съществува голямо разнообразие от методи за разпознаване на изображения, разработени в други научни области, например Data Mining, които биха могли да бъдат приложени за класификация на радарни цели, което е и работната хипотеза в настоящата дисертация.

С въпросите за класификация на наземни и морски цели при условията на forward scattering ефект научната общност се занимава само през последните няколко години. Един от водещите изследователски колективи в областта е екипът на проф. Черняков от Университета в Бирмингам, Великобритания.

Raja A. (2005), представител на школата на проф. Черняков от Бирмингамския Университет, в PhD дисертацията си разглежда възможността да се използва методът Shadow Inverse Synthetic Aperture Forward Scattering Radar (Инверсна Синтетична Апертура на Сянката в Радари Гледащи Напред) за автоматична класификация на автомобили (наземни цели). Разглеждат се две класификационни задачи: класификация на автомобилите по техния специфичен тип (марка) - разпознаване, и класификация по категория - категоризация по големина и форма.

За извличане на класификационни характеристики се използват спектрите на сигналите от целите и централната честота на сигнала. Оценява се и точността на измерване на скоростта на целите, тъй като тя се използва при класификацията и силно влияе върху грешката на класификация.

Предлага се филтрова концепция за класификация на сигналите. С помощта на БФТ (Бърза Фурие Трансформация - Fast Fourie Transform, FFT) се получават спектрите на входните сигнали от целите. Извършва се предварителна обработка на спектрите на сигналите, като се нормират по носеща честота и по мощност; и се извършва "орязване" на спектъра, т.е. използва се най информативната му част в ниските доплерови честоти и максимални мощности около -20dB. Извършва се трениране на еталоните за отделните видове цели, с помощта на еталонираните сигнали. Многоканална класификация на

входния сигнал на целта се извършва като се сравни входният сигнал с всички еталони, след което се получава решение на този канал, в който сигналът е най-близък до еталона.

За да се редуцира параметричното пространство на спектъра, се използва методът Principle Component Analysis (PCA). Класификацията на целите се извършва чрез алгоритъма K-най-близък съсед (kNN), и се изследва влиянието на параметъра K върху качеството на класификация. Преценено е, че този метод дава най-добри резултати при K=4, но точността на класификация на автомобилите в трите класа - малки, средни и големи, не е висока (съответно 35,8%, 44% и 69.4%).

За подобряване на точността на класификация се използват многосензорни радарни системи с известен и неизвестен ъгъл на преминаване на автомобилите през базовата линия, между предавателя и приемника на FSR (Forward Scattering Radar). Получените стойности за обобщената точност на класификация (средна точност на класификация за трите вида автомобили - малки, средни и големи) са: за единична FSR система - при известен ъгъл - 70,4%, а при неизвестен ъгъл - 53%; за многопозиционна FSR система - при известен ъгъл - 78,4%, при неизвестен ъгъл - 70,9%. Получените резултати показват, че Forward Scattering радари могат да бъдат успешно използвани за класификация на движещи се наземни цели.

По тази дисертация са публикувани няколко статии, представящи получените резултати, както следва.

Cherniakov et al. (2005) прави обобщение на предложения подход Shadow Inverse Synthetic Aperture Forward Scattering Radar за класификация на движещи се автомобили, на база няколко предишни публикации. Изследвана е работата на FSR класификатора, разгледани са основните етапи на работа му и подготовката на данните. Използвани са подходът PCA (Principal Component Analyses) и K-най-близък съсед (kNN) класификатор. Представени са редица оригинални резултати за обобщената точност на предсказване чрез FSR системи. Когато в класификатора се използва информация за ъгъла на пресичане на базовата линия от целта, класификацията се извършва с по-голяма точност (при ъгъл на пресичане 90° - 79.7%, при 45° - 83.4%; при 22.5° - 48.2%). Когато в класификатора не се използва информация за ъгъла на пресичане на базовата линия от целта, точността на класификация намалява (при ъгъл 90° - 64.2%, при 45° - 60.9%, при 22.5° - 34.0%). Обобщената класификационна точност за единична FSR система при известен ъгъл е 70.4%, а при неизвестен ъгъл е 53%, а за многопозиционна FSR система - при известен ъгъл е 78.4%, а при неизвестен ъгъл е 70.9%.

Rashid et al. (2008) предлагат блокова схема на АТС (автоматичен класификатор на целите), включваща следните компоненти: Power Spectrum Calculator, предварителна обработка с нормиране на скоростта на автомобилите, PCA и класификатор K-най-близък съсед (кNN). Показана е възможността този класификатор да работи на различни радио

честоти (64 MHz, 151MHz и 434 MHz), като резултатите показват, че класовете се разделят най-добре при честота 151MHz.

В изследване на Ibrahim N. et al. (2009) се използва същият АТС (автоматичен класификатор на целите) за класифициране на четири вида автомобили с Forward Scattering Radar (FSR). Отново са приложени два подхода - за разпознаване на марката на автомобила и за класифициране (категоризиране) на автомобилите по големина и форма. За класификация се използват няколко вида невронни мрежи (Multi-Layer Perceptron) - Levenberg-Marquadt (LM) back propagation; Quasi-Newton (BFG) back propagation; Scaled Conjugate Gradient (SCG) back propagation. Предложените невронни класификатори са сравнени по точност с класификатора К-най-близък съсед (kNN). Изследваните невронни мрежи се различават по броя на невроните във входния слой (два или три), скритият слой съдържа 10 неврона, а изходният слой съдържа 3 неврона, съответстващи на трите класа автомобили - малъки, средни и големи. За изследването се използват 45 записа на сигнали от автомобили. Получената точност на класификация е 100% за класове "малки" и "средни", и 95% - за клас "големи". Получените резултати са много по-добри от тези за класификация чрез използването на PCA и К-най-близък съсед (кNN) класификатор.

За момента не са известни изследвания, свързани с използването на Data Mining методи за класификация на радарни морски цели. Изследванията върху такъв тип данни в настоящия дисертационен труд имат за цел да проверят възможностите за ефективно прилагане на Data Mining алгоритми за класификация и по отношение на този нестандартен тип данни – forward scattering радиолокационни сигнали, и по-конкретно – за класификация на открити морски forward scattering радиолокационни цели. Особеностите тук са свързани с избор на информативни параметри на сигналите и необходимостта от предварителна обработка на записите на сигналите с цел оценка на параметрите им, за да се получат от тях данни, подходящи за аналитична обработка с Data Mining методи и средства. Друг проблем е невъзможността на този етап да бъдат получени голям обем от записи, което до голяма степен влияе върху качеството на създаваните класификационни модели.

<u>Изводи:</u>

- При обзора на състоянието на използваните методи за класификация на FSR цели бяха открити публикувани резултати само на представители на изследователския екип от Бирмингамския университет. В тези публикации се използват основно алгоритмите "К-най-близък съсед" (kNN) и "Невронни мрежи" (Multi-Layer Perceptron (MLP), след предварителна обработка на спектрите на сигналите от целите.
- Спектрите на сигналите от целите се обработват предварително с цел да се получат обучаващи сигнали по следния начин: първоначално спектърът на сигнала се нормира по носеща честота, по мощност и по скорост, а след това се "орязва", за да се използва най-информативната му част в ниските доплерови честоти.

- Използва се и методът Principle Component Analysis (PCA) за намаляване на размерността на параметричното пространство на спектрите на получените сигнали.

1.5. Цел и задачи на дисертационния труд

Настоящият дисертационен труд е посветен на изследване на приложимостта и ефективността на различни методи за извличане на знания от големи обеми данни чрез обучение на класификатори. Изследванията са осъществени върху данни, произхождащи от две различни предметни области - данни за студенти и данни за открити морски цели.

Основните проблеми, свързани с извличане на полезна информация за обучаемите в университетите в България и разгледани в конкретния пример на УНСС, които представляват интерес за ръководството, могат да бъдат формулират както следва:

- Прогнозиране за успешно/слабо представящи се студенти в университета, в зависимост от техни персонални особености преди и след приемане във висшето учебно заведение.
- Определяне на факторите, от които най-силно зависи успеваемостта на студентите в университета.
- Подобряване на събирането и организирането на данните за студентите, с цел по-добро управление на качеството на обучение.

Решаването на тези проблеми може да бъде подпомогнато чрез използване на подходящи методи и средства за извличане на знания от наличните данни за студентите, и по-конкретно чрез решаване на задача за класификация.

Важен проблем, който възниква при осъществяване на защита на морски граници, е определянето на вида на нарушителя (класа на плавателен съд). За решаването на този проблем също се предполага, че могат да бъдат използвани Data Mining методи за класификация.

Целта на настоящия дисертационен труд е да се генерират и изследват Data Mining модели за класификация (обучени класификатори, получени чрез прилагане на различни методи за класификация при извличане на знания от големи обеми данни), за предсказване успеваемостта на студентите в университета на базата на техни персонални характеристики преди и след приемане в университета, и за класификация на морски обекти.

За постигане на посочената цел се поставят ***следните основни задачи***:

- Разработване на методика за решаване на задача за извличане на знания от данни (Data Mining) чрез обучение на класификатори.

- Избор на подходящи софтуерни средства за реализация на задачата за извличане на знания от данни (Data Mining) чрез обучение на класификатори.
- Генериране на Data Mining модели за класификация (обучени класификатори) чрез избрани алгоритми.
- Оценяване на генерираните Data Mining модели за класификация (обучени класификатори) чрез избрани мерки за оценка.
- Сравняване на изследваните Data Mining модели за класификация (обучени класификатори).

1.6. Изводи по Първа Глава

В настоящата глава е постигнато следното:

- Направен е обзор на областта Извличане на знания от големи обеми данни (Data Mining), като е акцентирано върху задачата за класификация - основна задача, решавана в дисертационния труд чрез Data Mining методи и средства.
- Представени са четири метода - „Дърво на решенията", „Генератор на правила", „К-най-близък съсед" и „Невронни мрежи", които се използват за генериране на Data Mining модели за класификация (обучени класификатори) в дисертацията.
- Предложени са различни мерки за оценка и сравнение на генерираните модели за класификация (обучените класификатори).
- Разгледани са различни подходи за реализация на Data Mining проекти.
- Направен е обзор на състоянието на Извличането на знания от големи обеми данни в областта на образованието (Educational Data Mining - EDM).
    - Анализът на публикуваните научни изследвания показва, че предсказването успеха на студентите е много важен и актуален проблем за университетите. Резултатите се използват за различни цели: *Откриват се изоставащи студенти*, които евентуално ще отпаднат от обучение поради слабото си представяне, и на тези студенти се предоставя допълнителна помощ; *Откриват се добрите студенти* – това са студентите, които се справят най-успешно с обучението и които са най-желани за университета, и именно към такъв тип студенти се насочват бъдещите маркетингови кампании; *Откриват се факторите, които в най-голяма степен влияят върху успеха* или слабото представяне на студентите, и това се превръща в ценна информация, на базата на която се взимат управленски

решения за извършването на промени и за подобряване ефективността и качеството на предлаганото обучение.

- *Успехът на студентите се предсказва най-често чрез решаване на задача извличане на знания от данни чрез решаване на задача за предсказване, като* в повечето случаи *предсказваната величина е номинална променлива*, т.е. решава се задача за класификация.
- Най-често използваните методи за обучение на класификатори включват „Дърво на решенията”, „Невронни мрежи”, Байесови методи - Наивен Байесов класификатор и Байесови мрежи, и Логистична регресия. Във всички случаи обаче, се прилагат няколко различни метода или няколко различни алгоритми, за да се достигне до създаването на подходящ модел за предсказване (обучен класификатор), а не се използва един единствен алгоритъм.
- Често класификационната задача се решава за различни варианти на предсказваната променлива, като резултатите обикновено показват, че при по-малък брой възможни стойности на предсказваната променлива се получава и по-висока точност на предсказване.

- Направен е обзор относно използваните методи за класификация на морски цели.
  - При обзора на състоянието на използваните методи за класификация на FSR цели бяха открити публикувани резултати само на представители на изследователския екип от Бирмингамския университет. В тези публикации се използват основно алгоритмите "K-най-близък съсед" (kNN) и "Невронни мрежи" (Multi-Layer Perceptron - MLP), след предварителна обработка на спектрите на сигналите от целите.
  - Спектрите на сигналите от целите се обработват предварително с цел да се получат обучаващи сигнали по следния начин: първоначално спектърът на сигнала се нормира по носеща честота, по мощност и по скорост, а след това се "орязва", за да се използва най-информативната му част в ниските доплерови честоти.
- Формулирани са целта и задачите на дисертацията.

Глава 2: Методика за провеждане на изследването. Подготовка на данните.

Втора глава от дисертационния труд включва две части. В първата част (т.2.1) е представена методиката за провеждане на изследването, включваща избор на подход за реализация на Data Mining проекта, избор на подходящи софтуерни средства за осъществяване на поставените задачи и описание на конкретните дейности, които ще бъдат осъществени в рамките на изследването. Във втората част (т.2.2) са описани извършените дейности, свързани с предварителна подготовка и изследване на наличните данни – в т.2.2.1 на данните за студенти, а в т.2.2.2 на данните за морски цели, които се използват при генерирането на Data Mining моделите за класификация (обучени класификатори).

2.1. Методика за провеждане на изследването

2.1.1. Избор на подход за реализация на задачата за извличане на знания от данни (Data Mining) чрез обучение на класификатори

Решаваната научна задача в този дисертационен труд е задачата за извличане на знания от данни (Data Mining) чрез обучение на класификатори. Затова, за целите на проведеното научно изследване в настоящия дисертационен труд е избран подходът CRISP-DM (Cross-Industry Standard Process for Data Mining), по-подробно описан в т.1.2.5, който се приема за стандартен подход за реализация на Data Mining проекти от специалистите в тази научна област, неутрален е по отношение на областтите на приложение и е най-често използвания подход за провеждане на Data Mining анализи през последните години.

Извличането на знания от двете съвкупности от данни (университетски данни и FS радиолокационни данни за движещи се морски обекти), избрани за провеждане на научните изследвания, е извършено чрез последователно преминаване през петте основни етапа на CRISP-DM процеса - Business Understanding, Data Understanding, Data Preprocessing, Modeling and Evaluation, не са изпълнени само дейностите предвидени в последния етап – Deployment, тъй като те са свързани с конкретното практическо прилагане на получените аналитични резултати в производствена среда (такава не е използвана за целите на настоящия дисертационен труд).

2.1.2. Избор на софтуерни средства за провеждане на изследването

За осъществяването на научните изследвания, представени в настоящия дисертационен труд, ще се използват различни софтуерни решения на различни етапи, като избор и подготовка на данните, изследване и описание на данните, моделиране и оценка на получените модели, при реализацията на основни дейности в CRISP-DM подхода.

Изборът на софтуерни средства е продиктуван от следните съображения.

Microsoft Excel е свободно наличен, добре познат и широко използван софтуер, предоставящ разнообразни възможности за работа с големи обеми данни, организирани в табличен вид. Използвани са основно функционалностите, позволяващи сортиране, филтриране и графично представяне на данните.

QlikView е Business Intelligence (BI) софтуер, подходящ за лесно представяне и опознаване на данните, предоставящ възможности за изследване на данните по различни параметри на различни нива на детайлност, с добри графични възможности за визуализация. QlikView е най-бързо развиващото се BI приложение в световен мащаб, позволяващо бързо и лесно да се консолидират данни от различни източници, да се изследват връзките между данните, да се визуализира информацията под формата на разбираеми графики и диаграми. Свободен достъп до персонална версия (версия за лична употреба и разработване на собствени приложения) на този софтуер е осигурен от фирмата-разработчик QlikTech (http://www.qlikview.com/).

Съществува голямо разнообразие от софтуерни решения за реализацията на Data Mining анализи. По-голямата част от тях са разработени от водещи световни фирми, производители на софтуер, и се разпространяват с комерсиална цел. Най-известните и най-разпространето софтуерни решения за Data Mining се предоставят от фирмите-лидери в ИКТ областта и включват (www.kdnuggets.com/software/suites.html):

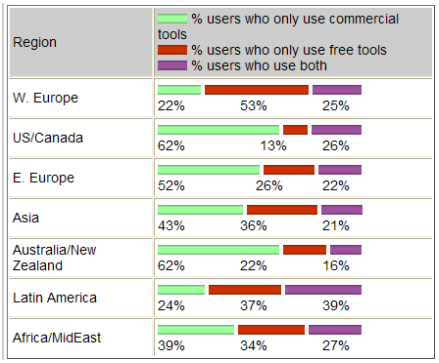
- ***Microsoft*** BI platform, DBMiner 2.0 (Enterprise), XLMiner (Data Mining Add-In For Excel), KnowledgeMiner (yX) for Excel
- ***IBM*** Intelligent Miner Data Mining Suite, IBM Cognos
- ***Oracle*** Data Mining (ODM)
- ***SAP*** Business Objects/NetWeaver
- ***SPSS*** Clementine
- ***SAS*** Enterprise Miner
- ***Statistica*** Data Miner
- ***Pentaho*** open-source BI suite (data mining based on Weka)

Достъпът до повечето от тези средства за чисто научноизследователски цели обикновено е ограничен. Затова, много по-често се използват софтуерните решения с отворен код, които се предоставят свободно. Сред по-известните са:

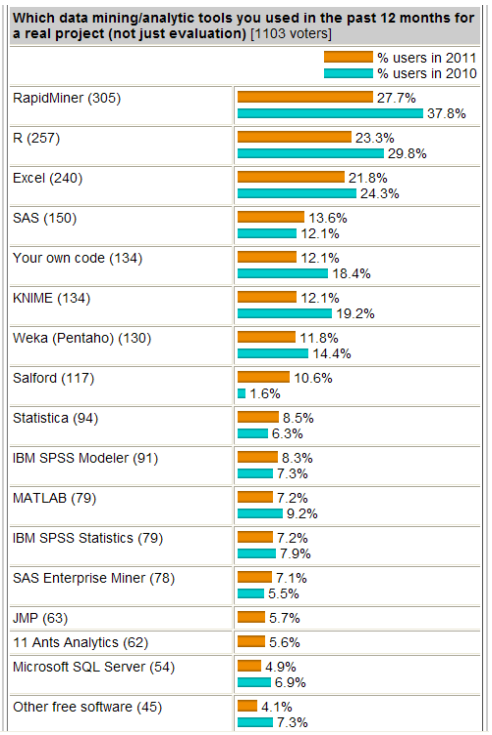
- Weka

- RapidMiner
- R

Съгласно последното (13-то) годишно онлайн проучване на KDNUGGETS.com (основен интернет ресурс на членовете на Data Mining общността, предоставящ информация за Data Mining и аналитични софтуерни средства, налични свободни работни позиции, консултантски услуги, курсове за обучение, новини, фирми и др.), относно най-използваните софтуерни решения за Data Mining от потребителите през 2011г., става ясно, че предпочитани са решенията с отворен код (RapidMiner, R, WEKA) и комерсиалните предложения на по-големите и известни фирми (SAS, IBM, Microsoft) (Фиг.2.1):



Фиг.2.1а



Фиг.2.1б

Фиг.2.1а,б: Резултати от годишно онлайн проучване на KDNUGGETS.com за 2011г. (Източник: http://www.kdnuggets.com/polls/2011/tools-analytics-data-mining.html)

В Табл.2.1 е представен сравнителен анализ между най-често използваните софтуерни решения за Data Mining с отворен код, които са достъпни за изследователски цели.

Таблица 2.1: Сравнителен анализ на софтуерни решения за Data Mining с отворен код

	WEKA	**RapidMiner**	**R**
Потребителски интерфейс	„+" Добра графична визуализация „-" Недостатъчно добър интерфейс	„+" Много добър потребителски интерфейс „+" Много добра графична визуализация	„+" Добра графична визуализация „-" Недостатъчна интерактивност на потребителския интерфейс
Алгоритми	„+" Голямо разнообразие от Data Mining алгоритми „+" Добри възможности за сравняване работата на алгоритмите „+" Добър за Text Mining	„+" Голямо разнообразие от Data Mining алгоритми, включва много от WEKA алгоритмите „-" По-трудно е сравнението на алгоритмите	„+" Богата статистическа библиотека
Работа със софтуера	„+" Лесно се усвоява „+" Специализиран за Data Mining анализи	„-" Не се усвоява много лесно „+" Предлага автоматично фиксиране на установени проблеми	„-" По-трудно се усвоява „-" По-слабо ориентиран към Data Mining анализи
Програмен език	Java „+"	Java „+"	Array language „-" Много различен от популярните програмни езици C++, Java, PHP
Интегриране на данни с различен формат	„-" Работи с arff формат Excel файловете се преобразуват във csv формат; Затруднена е работата с бази данни, които не са на Java.	„+" Работи добре с Excel файлове и бази данни	

Въз основа на извършения сравнителен анализ, за целите на дисертацията е избран Data Mining софтуер WEKA.

Изводи:

За целите на научните изследвания, осъществени в настоящия дисертационен труд, ще се използват:

- ***Microsoft Excel*** (Microsoft Office 2007) - за избор и интегриране на данни от различни източници, за изчистване и предварителна подготовка на данните, за подготовка на данните във формат, подходящ за използване в Data Mining софтуер.
- ***Microsoft Excel*** (Microsoft Office 2007), ***QlikView*** (v9.0) и ***WEKA*** (v3.6) - за изследване, опознаване и описание на данните.
- ***WEKA*** (v3.6) - за моделиране и оценка на получените резултати.

2.1.3. Описание на методиката на изследването

Научните изследвания, представени в настоящия дисертационен труд, са осъществени в следната последователност:

- ***Проучване на литературни източници*** в ***областта Data Mining*** (книги, научни публикации, описания на софтуерни средства), ***с цел избор и формулировка на решавните задачи, методи, класификационни модели, програмни продукти.***

  В резултат на извършените проучвания се стига до избора на научна задача, която да бъде реализирана в настоящия дисертационен труд – Data Mining задачата за класификация, както и на практически подход за реализация на изследванията – CRISP-DM моделът за реализация на Data Mining проекти. Избират се и конкретни приложни области, като до голяма степен изборът тук е свързан с достъпа до подходящи данни, които да бъдат използвани за реализация на изследванията. Една от избраните приложни области в случая е сферата на висшето образование, където изследователят работи през последните 10 години, сравнително добре познава спецификата и има осигурен достъп до данни и експертни знания от УНСС. Другата приложна област е свързана с обработката на сигнали, и по-конкретно с класификацията на открити радиолокационни цели от морски обекти. Дисертантът е член на международен екип работещ по национален проект, финансиран от ФНИ в тази област.

- ***Подбор на данни от избраните приложни области и изучаване на тяхната същност, произход, начин на събиране, организиране и съхранение*.**

  Основните научни методи, използвани на този етап от изследването, са: проучване на наличен опит представен в намерени научни публикации, участие в научни конференции, дискусии с експерти в избраните приложни области, проучване на налична документация за описание на данните.

  Като резултат от тези дейности се стига до получаването на двете първоначални съвкупности от данни, които се използват за реализация на научните изследвания. Избират се променливите, които ще бъдат предсказвани чрез избраните класификационни алгоритми, и се дефинират техните стойности – и за двете приложни области предсказваните променливи са дискретни величини от категориен тип.

- ***Изучаване и предварителна подготовка на наличните съвкупности от данни*.**

  Изучаването и предварителната подготовка на данните е изключително важен, труден и продължителен етап от реализацията на всеки Data Mining проект. От

неговата реализация до голяма степен зависи и качеството на получените крайни резултати от извършваните анализи.

На този етап са използвани различни изследователски методи:

- ***Визуално изследване на данните*** чрез софтуерните средства Excel, QlikView и WEKA.

  За всяка от наличните променливи в данните се изследва типа величина и нейните основни характеристики. При цифровите променливи се изследват техните обобщени статистически характеристики (обхват, средно аритметично, стандартно отклонение и др.), а при променливите от категориен тип – брой заемани стойности, медиана, мода и др. Изучава се и разпределението на променливите (честотно разпределение и хистограми). Всички променливи се изследват за наличието на грешки (необичайно стойности, имайки предвид същността на променливите), изключения (стойности, които силно се различават от общата съвкупност от дани), липсващи стойности. При необходимост, се предприемат мерки за отстраняването на установените проблеми в данните.

- ***Изследване индивидуалното влияние на всяка от входните променливи върху изходната/предсказваната променлива (класа).***

  Този изследователски метод е използван само за втория вариант на предсказваната променлива (двоична вели чина) при университетската съвкупност от данни.

  За целта се използва алгоритъм „Дърво на решението” (WEKA J48), който се прилага последователно само върху две от променливите, съдържащи се в данните – изследваната входна променлива и изходната/предсказвана променлива. Във всеки от тези случаи на прилагане на агоритъма, се получава опростено дърво на решенията, чрез което данните се разделят на два класа „Слаби” и „Силни” (двете възможни стойности на изходната променлива) само чрез стойностите на изследваната входна променлива (категориите, които заема, ако е дискретна променлива от категориен тип, или прагови стойности, ако е непрекъсната числова променлива).

  В резултат на извършените дейности се получават крайните съвкупности от данни, върху които се прилагат избраните методи за Data Mining анализи.

- ***Разработване на модели за класификация чрез прилагане на избрани Data Mining алгоритми върху данните.***

  На този етап върху данните са приложени четири класификационни алгоритми – генератор на правила 1R, дърво на решението, невронна мрежа и К-най-близък съсед. Основната причина за избора на тези алгоритми е тяхната различна природа, т.е. различния начин на разделяне на съвкупността от данни на отделни класове. Освен това, всеки от тези алгоритми има свои особености на прилагане, както и различни предимства и недостатъци.

  Въведени са критерии и метрики за оценка на качеството на генерираните класификационни правила или модели (виж по-долу).

  Целта на изследователя е да се провери кой от алгоритмите работи най-добре по избраните критерите при различните данни и в различни условия, как параметрите на тези алгоритми влияят върху точността на предсказване при осъществяваната класификация.

- ***Оценка на разработените модели за класификация (обучени класификатори)***

  Получените модели за класификация на изследваните съвкупности от данни се оценяват на базата на резултатите, получени от работата на всеки от избраните алгоритми, по критерии и метрики за оценка на качеството на генерираните класификационни правила или модели, (по-подробно описани в т.1.2.4), включващи:

  - *Точност на класификация* - % правилно класифицирани обекти;
  - *Грешка при класификацията* - % грешно класифицирани обекти;
  - *Матрица на класификация* (Confusion Matrix), в която са представени резултатите от извършената класификация чрез приложения класификационен алгоритъм, за всеки от класовете.
  - *Параметрите TP Rate, FP Rate, Precission, Recall и F-measure,* чиито стойности се изчисляват на базата на Матрицата на класификация.
  - *Kappa Statistic* - параметър, чийто стойности показват степента на съгласуваност между предсказания и реалния клас на изследваните обекти;
  - *ROC Area* – графичен метод за оценка на точността на класификаторите, както и за сравняване на различни класификационни модели.
  - *Model Correct/Incorrect Ratio* - съотношение между броя на правилно и неправилно класифицираните обекти от избрания алгоритъм – параметър, който показва доколко добре се справя генерираният модел с класификацията.
  - *Model/Naïve Correct/Incorrect Ratio* – параметър, чрез който се оценява работата на избрания алгоритъм и който показва в каква степен се подобрява класификацията чрез използване на генерирания модел, в сравнение с наивната

класификация (т.е. класификация, която би била извършена без използване на модел, създаден от наличните данни, а само въз основа на тяхното вероятностно разпределение в съответните класове).

- ***Формулиране на изводи и препоръки към крайния потребител на резултатите от Data Mining анализите.***

Осъществените научни изследвания завършват с обощено представяне на изводите относно извършената оценка на създадените класификационни модели. Формулирани са и препоръки към потенциалните крайни потребители на моделите, които се отнасят както за прилагането на моделите в реални условия (какви са ограниченията, как да бъдат развивани и подобрявани и т.н.), така и за самите данни (как да се подобри събирането на данните, за да се разширят възможностите за тяхното по-пълноценно използване и анализиране).

Описанието на проведените изследвания и получените научни резултати е осъществено поотделно за двете избрани приложения – анализ на университетски данни и анализ на данни от радиолокационни сигнали, като последователно са представени основните дейности, извършени на всеки етап на работа (базирано на CRISP-DM подхода).

Изводи: На база CRISP-DM подхода са избрани необходимите стъпки в методиката за провеждането на изследванията в дисертацията, както следва:

- ***Проучване на литературни източници*** в ***областта Извличане на знания от големи обеми данни (Data Mining)***
- ***Изучаване и предварителна подготовка на наличните съвкупности от данни***.
  - ***Визуално изследване на данните*** чрез софтуерните средства Excel, QlikView и WEKA.
  - ***Изследване индивидуалното влияние на всяка от входните променливи върху изходната/предсказваната променлива (класа)***.
- ***Разработване на модели за класификация чрез прилагане на избрани Data Mining алгоритми върху данните***.
- ***Оценка на разработените модели за класификация (обучени класификатори);***
- ***Формулиране на изводи и препоръки към крайния потребител на резултатите от Data Mining анализите.***

2.2. Подготовка на данните

Научните изследвания, представени в настоящия дисертационен труд, са осъществени за две различни съвкупности от данни – данни за студенти и данни за сигнали от радарни морски цели. Научните резултати за двата вида данни са получени чрез използване на една и съща методика, описана в т.2.1.3.

2.2.1. Анализ на данните за студенти

2.2.1.1. Запознаване с университетските бази данни за студентите

Научноизследователската дейност по настоящата дисертация започва с проучване на предметната област, което се осъществява чрез събиране и преглед на научни публикации, свързани с приложението на Data Mining методи и средства за анализ на данни, налични в различни организации от образователния сектор (университети, колежи, училища, държавни институции). По-нататък проучването е фокусирано върху изучаване и анализиране на съществуващи проблеми в университетите, за чието решаване са прилагани разнообразни Data Mining методи и техники. Резултатите от това проучване са описани по-подробно в Глава 1, т.1.3, където е представен анализ на актуалното състояние на изследванията в избраната област на научните изследвания от настоящия дисертационен труд.

Друга важна стъпка на този етап в дисертационния труд е провеждането на формални разговори с представители на ръководството на избрания университет – УНСС, относно необходимостта от по-задълбочено изследване и анализ на данните за студентите, с които разполага университетът. Идентифицирани са актуални и важни за университета проблеми (свързани с по-ефективното е ефикасно управление на университета и по-доброто познаване на студентите), чието по-успешно решаване би могло да бъде подпомогнато чрез анализ на наличните данни с data mining методи и средства. Установена е необходимост от дефиниране на насоки за подобряване събирането на данни с цел повишаване на аналитичните възможностите и ефективното подпомагане взимане на решения. Допълнителна информация е събрана и чрез неформални срещи и разговори с преподаватели, студенти и представители на администрацията (ИКТ експерти и ръководители).

На базата на извършените на този етап дейности са дефинирани целите, които трябва да бъдат постигнати от гледна точка на потребителите – в случая ръководителите на университета. Тези цели впоследствие са трансформирани в научни задачи, които се решават чрез прилагане на избрани data mining методи и средства. Целите и задачите на научното изследване са представени в Глава 1, т.1.5.

2.2.1.2. Избор на данни за изследването

Съгласно посочената по-горе методология на изследването, дисертационният труд включва избор и опознаване на данните, които ще бъдат използвани при последващите анализи. Работата на този етап започна със запознаване с правила, процедури и документи за кандидатстване и приемане на студентите в УНСС, както и с правила и процедури за събиране и съхранение на информация, свързана с академичното обучение на приетите студенти, за да се установи с какви данни в електронен формат разполага университетът. Проведени са и дискусии с представителите на администрацията, които са отговорни за събирането, обработката, съхранението и поддръжката на университетските данни.

В резултат от извършените дейности се стига до следните изводи – университетските данни се съхраняват в две отделни релационни бази данни – в едната се съдържат данните, свързани с провеждането на кандидат-студентските кампании, а в другата – данните за обучението на приетите студенти.

Базата данни „Кандидат-студентски кампании” (КСК) съдържа следните данни:

- Лични данни за кандидат-студента (напр. имена, адреси, информация за полученото до момента образование - профил, училище, успех);
- Данни за организацията и провеждането на кандидат-студентските изпити (напр. избран предмет за кандидатстване в университета, оценки от положените изпити, бал за кандидадстване и др.).

Базата данни „Студент” съдържа данни за всички приети студенти, които се обучават в УНСС, и включва:

- Лични данни (част от личните данни, които се прехвърлят от базата данни КСК, направление, специалност, поток, група, факултетен номер и т.н.);
- Административни данни (информация за положени изпити, оценки получени на различни изпитни сесии, заверки и т.н.).

Резултатът от извършените на този етап дейности е изборът на съвкупност от данни, които да бъдат използвани за целите на провежданите научни изследвания. Тъй като избраните атрибути на данните се съдържат в двете отделни бази данни, крайната съвкупност от данни е получена чрез екстракция на необходимите параметри от двете бази и интегрирането им в един общ файл в Excel формат.

2.2.1.3. Предварителна подготовка на данните

Изследването и предварителната подготовка на данните е един от най-важните етапи в реализацията както във всеки data mining проект, така и в настоящата дисертация. Доброто познаване на данните и правилното извършване на дейностите по предварителната подготовка на данните (премахването на грешки и несъответствия, попълване на липсващи стойности, преобразуване на променливи, създаване на нови променливи, редуциране на променливи и др.) в голяма степен влияят върху качеството на получените резултати от анализите.

Предоставената от университета съвкупност от данни съдържа 10330 записа за студентите в УНСС, приети по време на кандидат-студентските кампании (КСК) през 2007, 2008 и 2009 години. Информацията за всеки студент в оригиналните данни съдържа следните 20 променливи (атрибути):

- Вид Направление/Поднаправление/Специалност
- Наименование на Направление/Поднаправление/Специалност
- Семестър, в който учи студента в момента на извличане на данните (началото на 2011г.)
- Година на раждане
- Месторождение
- Селище на местоживеене
- Държава
- Вид образование, получено до момента на кандидатстване
- Селище, в което се намира учебното заведение, където е получено образованието
- Наименование на училището
- Профил на училището
- Година на приемане в УНСС
- Пол
- Среден успех от средното образование
- Среден успех от I курс без 2-ки
- Среден успех от I курс с 2-ки
- Брой невзети изпити
- Бал от КСК
- Изпит на приемане при КСК
- Оценка от приемния изпит

Извършени са различни преобразувания на оригиналните данни с цел, от една страна, да се повишат възможностите за извършване на по-задълбочени анализи, които да подпомагат реалните потребители при взимане на решения (в случая, ръководството на

университета), и от друга страна, да се използват по-пълноценно аналитичните възможности на използваните BI и data mining методи и средства.

А. Премахване на променливи или стойности на променливи

Част от променливите, съдържащи се в оригиналните данни, не представляват интерес за целите на изследването в настоящата дисертация. Например, променливите „Държава" и „Вид образование, получено до момента на кандидатстване" приемат по една единствена стойност за всички записи (съответно, „България" и „Средно"), тъй като данните са за българските студенти в УНСС и всички те са завършили средно образование до момента на кандидатстване. Променливите „Месторождение" и „Селище на местоживеене" също не се взимат под внимание при анализите, тъй като „Селище, в което се намира учебното заведение, където е получено образованието" се счита най-важна от характеристиките, свързани с местоположението на кандидат-студентите. Затова тези 4 променливи са изключени от крайната съвкупност от данни, използвана при изследванията.

В крайната съвкупност от данни се извключва и променливата „Среден успех от I курс без 2-ки", тъй като стойностите на тази величина не отразяват действителния среден успех на студентите в края на I курс. Запазва се само променливата „Среден успех от I курс с 2-ки".

При изследването на данните се установява, че променливата „Изпит на приемане при КСК" заема 9 различни стойности. В действителност, студентите се приемат в УНСС с пет различни изпита – български език, математика, география, история и икономика. За някои от направленията се държат и допълнителни изпити, затова в данните се появяват и изпити по английски език, френски език, журналистика – писмен и устен. За да се извършат коректно анализите, записите, които съдържат стойностите, съответстващи на допълнителните изпити, се премахват от крайната съвкупност от данни (така крайната съвкупност от данни намалява от 10330 записа на 10067 записа).

Б. Преобразуване на променливи от тип „свободен текст"

Една от основните стъпки при предварителната подготовка на предоставените данни е обработката на променливите от типа „свободен текст" (т.е. стойностите, които приемат тези променливи, не са предварително структурирани и представляват свободен текст, въвеждан от служителите при приемане на документите на кандидат-студентите), които се считат за особено важни за осъществяване на предвижданите анализи. Такъв тип данни трудно могат да бъдат анализирани така, че да се получат смислени резултати от анализите. Поради това е извършена предварителна обработка на стойностите на тези параметри – свободният текст е заменен с определен брой стойности, извлечени по смисъл

от наличния текст. Такъв вид преобразуване е извършено за променливите „Профил на училището” и „Селище, в което се намира учебното заведение, където е получено образованието”.

Променливата „Профил на училището”, величина от типа „свободен текст” в оригиналните данни, е трансформирана в номинална променлива, приемаща краен брой стойности, съответстващи на различните профили на средните училища в България. При преобразуването е взето предвид профилирането на училищата, в които получават средното си образование учениците в България (непрофилирани СОУ и профилирани училища, в които се извършва прием след завършване на 7 и 8 клас), като информация е взета основно от интернет страницата на Министерството на образованието, младежта и науката (www.minedu.government.bg).

Променливата „Селище на образование”, съдържаща наименования на различни селища в България, е преобразувана във величина с краен брой стойности. Първоначално са запазени стойностите, съответстващи на 12-те най-големи града в България (София, Пловдив, Варна, Бургас, Русе, Велико Търново, Сливен, Стара Загора, Шумен, Добрич, Благоевград, Плевен), а останалите са заменени с нови 7 стойности, съответстващи на 7 региона в България - София-област, северозападен (СЗ), северен централен (СЦ), североизточен (СИ), югозападен (ЮЗ), южен централен (ЮЦ) и югоизточен (ЮИ), на базата на принадлежността на селищата към съответните региони на страната (използвана е основно информацията, публикувана на интернет страницата http://bg.guide-bulgaria.com/). В последствие, след като при анализирането на данните се установява, че големият брой стойности, които заема тази променлива, не водят до съществени изводи, техният брой се намалява до 7 – София и шестте региона в страната (ЮЗ, ЮЦ, ЮИ, СЗ, СЦ, СИ).

Преобразувания се извършват и на променливата „Наименование на Направление/Поднаправление/Специалност”. Оригиналните данни са взети в началото на 2011г., което означава, че студентите, приети през периода 2007-2009г., в този момент са в различни семестри на обучение (1-10 семестър, в зависимост от направление, прекъсване на обучението и др.). Специфичното в случая е, че студентите в УНСС първоначално (1-4 семестър) се приемат в съответното направление или поднаправление, а в последствие се разпределят по специалности. Поради това, в данните се съдържат наименования както на направления и поднаправления, така и на конкретни специалности, което е причина за наличието на голям брой различни стойности на променливата „Наименование на Направление/Поднаправление/Специалност”, което силно затруднява анализите. За да се получат по-смислени анализи, в крайната съвкупност от данни стойностите на тази променлива са редуцирани. Запазени са стойностите на 5 от 6-те професионални направления - „Социология, антропология и наука за културата”, „Политически науки”, „Обществени комуникации и информационни науки”, „Право” и „Администрация и управление”, а шестото направление „Икономика” (2/3 от общата съвкупност от данни се

отнася за студенти в това направление) е представено чрез своите 5 поднаправления - „Икономика и бизнес”, „Приложна информатика, комуникации и иконометрия”, „Икономика с чуждоезиково обучение”, „Икономика с преподаване на английски език”, „Финанси, счетоводство и контрол”.

Този вид преобразувания на данните са извършвани на няколко етапа, затова в някои от публикуваните резултати може да са използвани варианти на данните, които се различават от крайната съвкупност от данни, използвани при получените резултати в настоящич дисертационен труд. Тъй като промените не са толкова големи (отнасят се до малка част от записите в цялата съвкупност от данни), счита се че различията не са толкова съществени и не оказват силно влияние върху получените резултати.

В. Създаване на нови променливи

При изследването и подготовката на данните често се стига до необходимостта от създаването на нови променливи, които се считат важни за извършването на предвидените анализи.

В случая, за целите на анализа в настоящата дисертация е създадена новата променлива „Възраст на кандидат-студентите”, която е формирана като разлика между величините „Година на приемане в УНСС” и „Година на раждане”.

Г. Преобразуване на количествени в качествени променливи

При решаването на дефинираните data mining задачи и при прилагането на избраните data mining методи и средства се извършват и някои допълнителни преобразувания на променливите.

Например, част от променливите заемат числови стойности, но на практика не носят числов смисъл (т.е. не е логично да участват в различни математически изчисления), затова те се трансформират от числови в номинални променливи (напр. „Година на приемане в УНСС”, „Семестър”, „Възраст”), тъй като са много по-информативни при извършването на анализите.

Една от целите в настоящата дисертация е да се генерират и изследват Data Mining модели за класификация, за предсказване успеваемостта на студентите в университета на базата на техни персонални характеристики преди и след приемане в университета. Постигането на тази цел се осъществява чрез решаването на data mining задача за класификация, а при класификацията изходната (целевата) променлива, т.е. променливата, която ще бъде предсказване, е дискретна величина. Затова, оригиналната променлива „Среден успех от I курс с 2-ки”, която е непрекъсната числова величина, е преобразувана в променлива с краен брой дискретни стойности. Преобразуването е извършено по два начина – в единия случай дискретната величина заема 5 различни стойности, а във втория

случай – 2 различни стойности, като е изследвано поведението на приложените data mining алгоритми при тези различни условия.

Преобразуването на непрекъсната числова променлива „Среден успех от I курс с 2-ки” в дискретна променлива с 5 стойности е извършено въз основа на възприетата в българското образование шестобална система за оценяване, както е показано в Табл.2.2:

Таблица 2.2: Преобразуване на променливата „Среден успех от I курс с 2-ки” в дискретна променлива с 5 стойности

Оригинални стойности на променливата „Среден успех от I курс с 2-ки”	**Нови стойности на променливата „Среден успех от I курс с 2-ки”**
[0.00;2.49]	Слаб (Bad)
[2.50;3.49]	Среден (Average)
[3.50;4.49]	Добър (Good)
[4.50;5.49]	Много добър (Very Good)
[5.50;6.00]	Отличен (Excellent)

Преобразуването на непрекъсната числова променлива „Среден успех от I курс с 2-ки” в дискретна променлива с 2 стойности се извършва с цел да се провери поведението на избраните data mining алгоритми при наличието на двоична изходна променлива. В този случай изходната променлива представлява величина с две отчетливи стойности, т.е. по този параметър студентите се разделят на две групи (два класа) – „силни” и „слаби”. Праговата стойност за формиране на тези два класа е 4.50, както е показано в Табл.2.3:

Таблица 2.3: Преобразуване на променливата „Среден успех от I курс с 2-ки” в дискретна променлива с 2 стойности

Оригинални стойности на променливата „Среден успех от I курс с 2-ки”	**Нови стойности на променливата „Среден успех от I курс с 2-ки”**
[0.00;4.49)	Слаби (Weak)
[4.50;6.00]	Силни (Strong)

Д. Други преобразувания на променливите

При изследването на данните, проведено в дисертацията, са открити някои грешки (наличие на стойности, които са извън допустимия интервал за съответната променлива), които са коригирани след консултиране с представителите на университетската администрация, отговорни за данните.

При прилагането на data mining алгоритмите са извършвани и други преобразувания на основната съвкупност от данни, използвани при проведените изследвания, които са описани при представянето на резултатите (Глава 3).

Друго преобразуване на данните, което е извършено поради спецификата на данните и използваните софтуерни средства, е замяната на оригиналните текстови стойности на променливите, на български език, с техните еквиваленти - на английски език, за да е

възможно разчитането на получените резултати, а и за да бъдат използвани резултатите при разработените научни публикации.

2.2.1.4. Формиране на съвкупност от данни за изследванията

Като резултат от описаните по-горе преобразувания на оригиналните данни, се достига до крайната съвкупност от данни, използвана при проведените изследвания, представени в настоящия дисертационен труд.

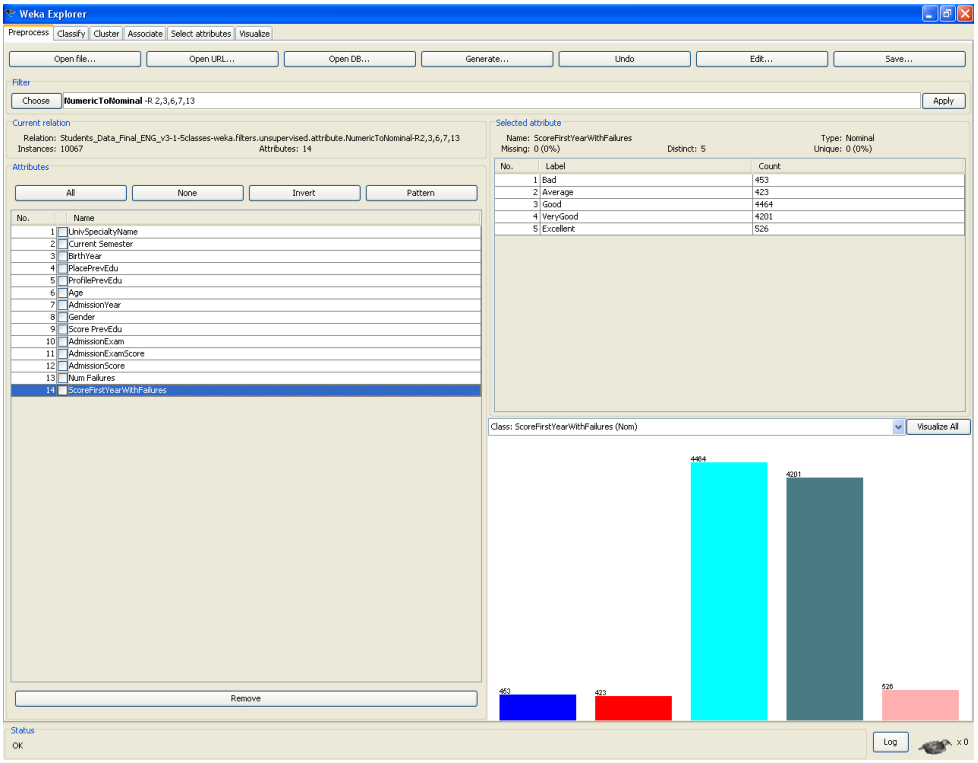
Съвкупността от данни, използвана при анализите, съдържа 10067 записа (т.е. данни за 10067 студента, приети в УНСС през периода 2007-2009г.) и 14 атрибута (променливи, които характеризират всеки от приетите студенти), и е представена в обобщен вид в Табл.2.4 и на Фиг.2.3.

Таблица 2.4: Обобщено описание на данните, използвани при анализите на данните за студенти

No.	Име на Атрибут	Тип променлива	Разпределение по стойности	Липсващи стойности
1	Наименование на направление/поднаправление	Номинална	10 различни стойности – съответстващи на 5 професионални направления и 5 поднаправления в УНСС	0 (0%)
2	Семестър	Номинална	10 различни стойности – 1,2,3,4,5,6,7,8,9,10	0 (0%)
3	Година на раждане	Номинална	29 различни стойности в интервала 1961-1991	0 (0%)
4	Местоположение (средно образование)	Номинална	7 различни стойности	9 (0%)
5	Профил (средно образование)	Номинална	9 различни стойности	123 (1%)
6	Възраст	Номинална	29 различни стойности в интервала 17-48	0 (0%)
7	Година на приемане в УНСС	Номинална	3 различни стойности – 2007, 2008, 2009	0 (0%)
8	Пол	Номинална	2 различни стойности – м/ж	0 (0%)
9	Среден успех (средно образование)	Числова	Min=3.4, Max=6, Mean=5.559, StdDev=0.391	219 (2%)
10	Изпит на приемане при КСК	Номинална	5 различни стойности	677 (7%)
11	Оценка от изпита на приемане	Числова	Min=0, Max=6, Mean=5.431, StdDev=0.55	677 (7%)
12	Общ бал при приемане	Числова	Min=0, Max=35.91, Mean=28.282, StdDev=3.41	682 (7%)
13	Брой невзети изпити	Номинална	13 различни стойности – 0-12	1 (0%)
14	Среден успех от I курс в УНСС	Номинална	1 Вариант: 5 различни стойности – слаб, среден, добър, много добър, отличен 2 Вариант: 2 различни стойности – слаби и силни	0 (0%)

От Табл.2.4 се вижда, че в данните се съдържат 14 променливи, 11 от тях са качествени променливи (nominal variables), приемащи ограничен брой дискретни/номинални стойности), останалите 3 са числови непрекъснати променливи (numeric variables), чието разпределение е описано с техни основни характеристики (минимална, максимална и средна стойности, и стандартно отклонение). Много малко са липсващите стойности (за 3 от атрибутите - около 7%, за други 2 атрибута - съответно 2% и 1%), което означава, че това няма да се отрази съществено върху резултатите от изследванията и затова не е необходимо да се предприемат действия по отношение попълването на тези липсващи стойности.

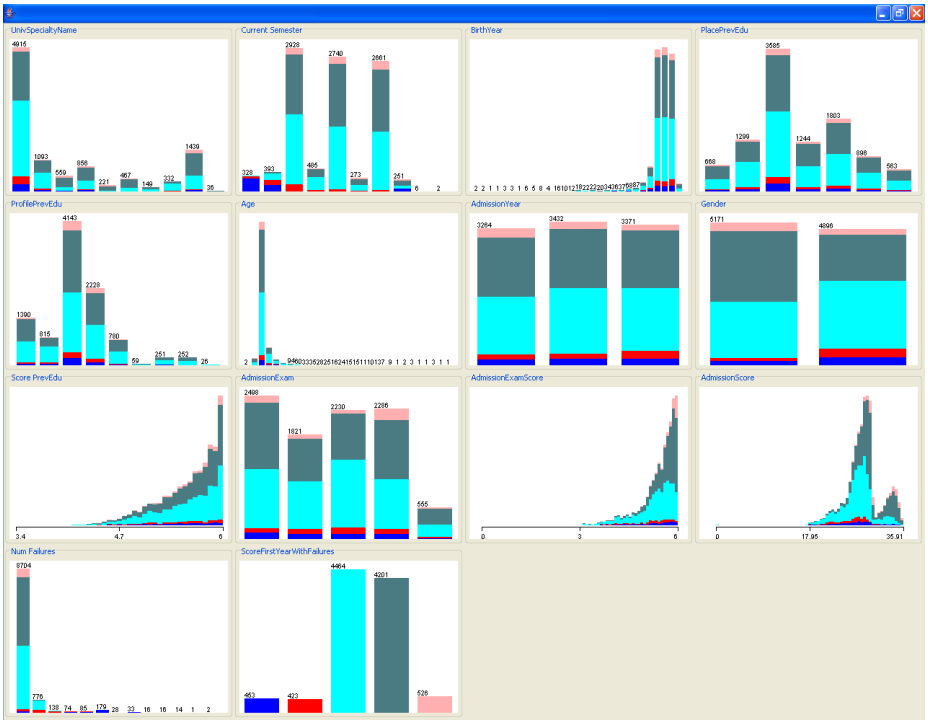
На Фиг.2.2 е представено обобщеното описание на данните в WEKA – съвкупност от 10067 записа и 14 атрибута, изходната величина (клас) „Среден успех от I курс" (Вариант 1 - номинална променлива с пет стойности - слаб, среден, добър, много добър, отличен) и нейното разпределение. Атрибути с номера 2,3,6,7 и 13, са преобразувани от числов в номинален тип променливи (приложен е WEKA Preprocessing Unsupervised Attribute Filter – NumericToNominal).



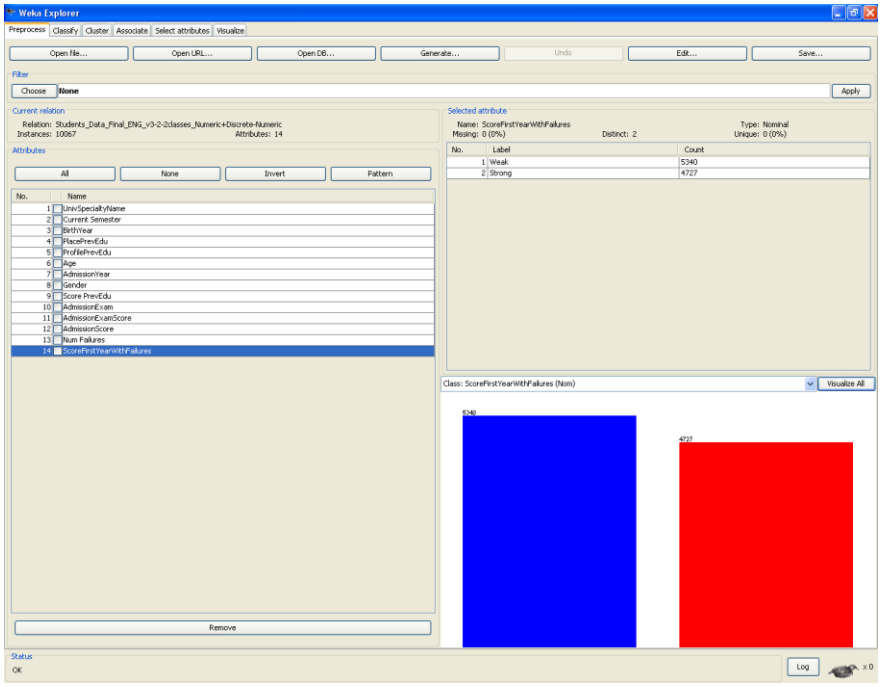
Фиг.2.2: Обобщеното описание на данните в WEKA за предсказвана променлива с 5 стойности

На Фиг.2.3 е представено разпределението на 14-те атрибута, които се съдържат в данните, по отношение на петте стойности на изходната променлива (Вариант 1) -

„Среден успех от I курс” (слаб, среден, добър, много добър, отличен – изобразени с различни цветове).

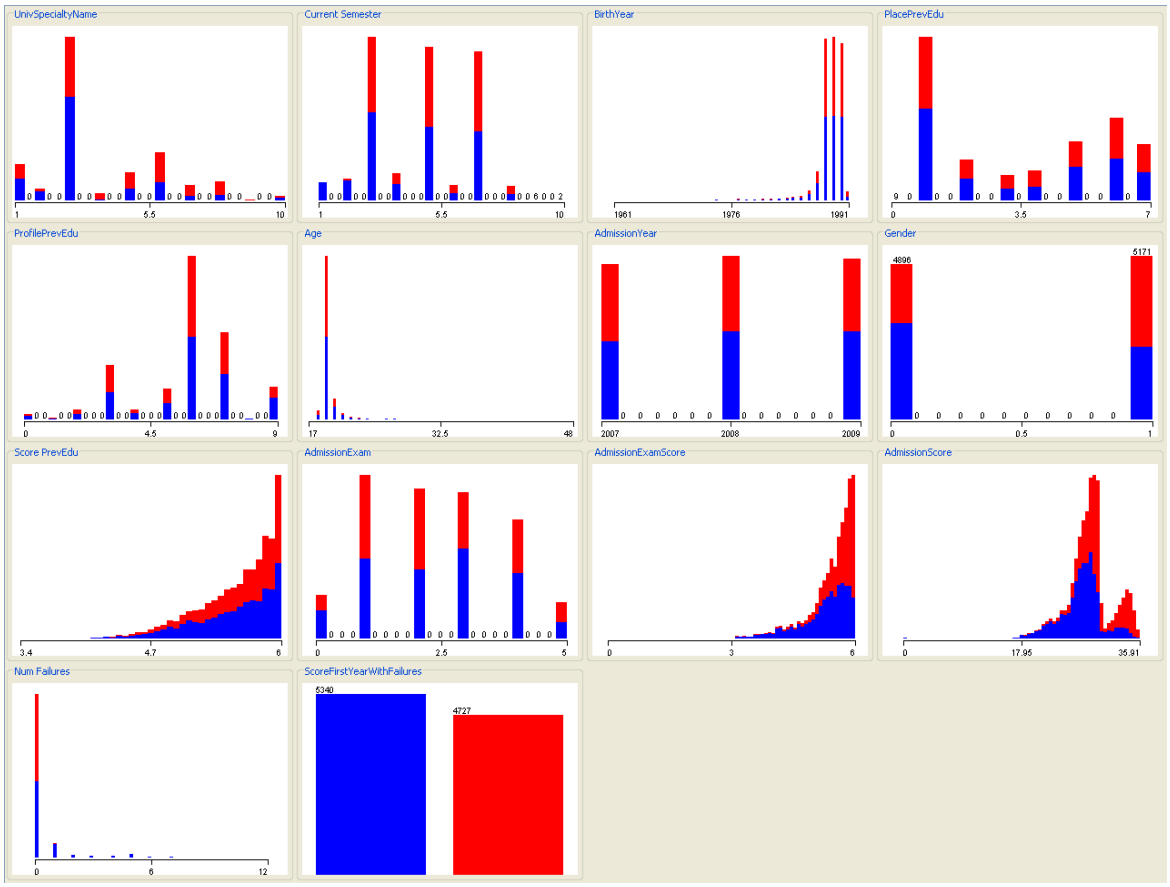


Фиг.2.3: Разпределение на атрибутите в данните относно 5-те стойности на предсказваната променлива в WEKA



Фиг.2.4: Обобщеното описание на данните в WEKA за предсказвана променлива с 2 стойности

На Фиг.2.4 и Фиг.2.5 са показани съответно обобщеното описание на данните в WEKA за предсказвана променлива с 2 стойности и разпределението на атрибутите в данните относно 2-те стойности на предсказваната променлива.



Фиг.2.5: Разпределение на атрибутите в данните относно 2-те стойности на предсказваната променлива в WEKA

Изводи:

- Получена е крайна съвкупност от данни за студентите, която ще се използва за генериране на Data Mining модели за класификация (обучение на класификатори) чрез софтуер WEKA.

- Съвкупността от данни съдържа 10067 записа за студенти, описани чрез 14 атрибута (параметъра), 11 качествени променливи (nominal variables)и 3 числови непрекъснати променливи (numeric variables).

- Много малко са липсващите стойности (за 3 от атрибутите - около 7%, за други 2 атрибута - съответно 2% и 1%), което означава, че това няма да се отрази съществено върху резултатите от изследванията и затова не е необходимо да се предприемат действия по отношение попълването на тези липсващи стойности.

- Изходната/предсказваната променлива „Среден успех от I курс" е преобразувана в номинална променлива по два начина (Вариант 1 - номинална променлива с пет стойности - слаб, среден, добър, много добър, отличен; Вариант 2 - номинална променлива с 2 стойности - слаби и силни).

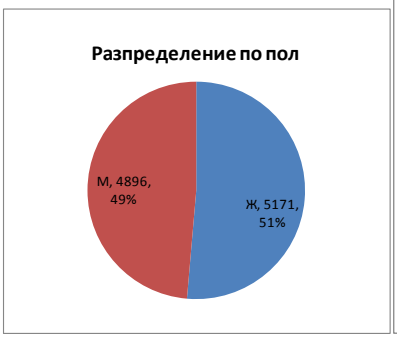
2.2.1.5. Изводи от анализа на данните

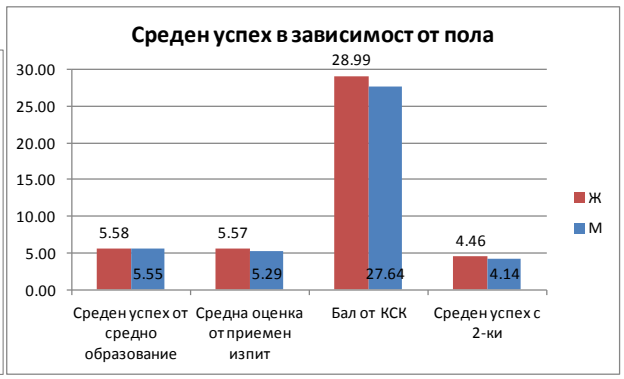
Задача на изследването е да се получи по-ясна представа за разпределението на кандидат-студентите, участващи в КСК през периода 2007-2009 и приети в УНСС, по различни параметри.

Първоначално анализът е осъществен чрез последователно изследване на разпределението на кандидат-студентите по избрани техни характеристики, които представляват интерес за ръководството на университета, както следва:

- Разпределение по Пол и Възраст (А)
- Разпределение в зависимост от Изпита, с който са били приети в университета (Б)
- Разпределение в зависимост от Профила на училището, в което кандидат-студентите са получили средно образование (В)
- Разпределение в зависимост от Региона, където се намира училището, в което кандидат-студентите са получили средно образование (Г)
- Разпределение по Професионални направления и поднаправленията на направление „Икономика", в които са приети кандидат-студентите (Д)
- Разпределение в зависимост от Успех на студентите в I курс от обучението в университета (Е)

А. Пол и Възраст





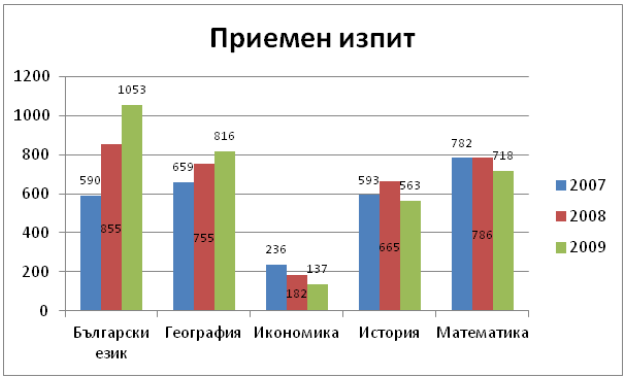
Фиг.2.6а Фиг.2.6б

Фиг.2.6а,б: Разпределение по „Пол”

От Фиг.2.6а и Фиг.2.6б става ясно, че наличните данни се отнасят до приблизително равен брой мъже (49%) и жени (51%), средният успех от средното образование е почти еднакъв за мъжете (5.55) и жените (5.58), средната оценка от приемен изпит е по-висока за жените (5.57) отколкото за мъжете (5.29), което естествено е валидно и за средния бал от КСК (28.99 за жените и 27.64 за мъжете). Средният успех в края на първата година от обучението в УНСС също е по-висок за жените (4.46) отколкото за мъжете (4.14).

Студентите, приети в УНСС през периода 2007-2009г., са на възраст между 17 и 48 години (при приемане). Преобладават студентите на възраст 18-20 години (92% от общия брой студенти 10067), като основната маса е на възраст 19 години - 78%, около 10% са студентите на възраст 20 години и малко над 4% са студенти на възраст 18 години. Студентите на възраст между 21 и 30 години представляват около 7.5% от общия брой, а студентите над 30 години – малко под 1% от общия брой.

Б. Изпит, с който са били приети студентите

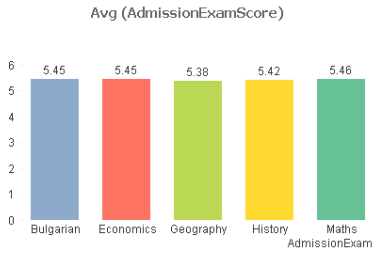


Фиг.2.7: Разпределение по приемен изпит **Фиг.2.8**: Разпределение по приемен изпит и

по години

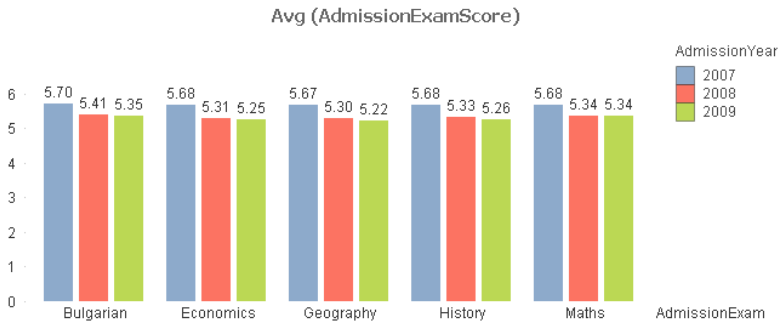
От приетите през 2007-2009 години студенти в УНСС, най-много са успелите с изпити по български език (27%), математика (24%) и география (24%), по-малко – с история (19%) и най-малко – с икономика (6%) (виж Фиг.2.7). Това разпределение се отнася до данните за 9390 студенти (при общо 10067 записа, за 677 записа стойността на тази променлива липсва).

Както е показано на Фиг.2.8, през трите години 2007-2009 значително е нараствал броят на студентите приети в УНСС с изпит по български език, увеличавали са се и студентите приети с изпит по география. Броят на студентите, приемани с изпит по математика, се е запазвал сравнително постоянен през периода. Съществено е намалявал броят на студентите, приети с изпит по икономика. Липсва ясно изразена тенденция по отношение изпита по история – най-голям е броят на приетите студенти с история в средата на периода (през 2008).



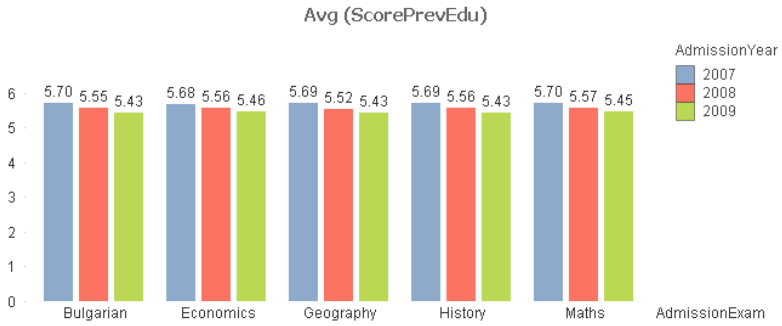
Фиг.2.9: Среден успех на кандидат-студентите, постигнат на приемен изпит

Няма съществени разлики в средния успех, постигнат от кандидат-студентите на петте изпита (Фиг.2.9) за разглеждания период 2007-2009. Но, забелязва се трайна тенденция за намаляване на средния успех по всички предмети през трите години (Фиг.2.10) – 2007, 2008 и 2009.



Фиг.2.10: Среден успех на кандидат-студентите, постигнат на приемен изпит, за периода 2007-2009

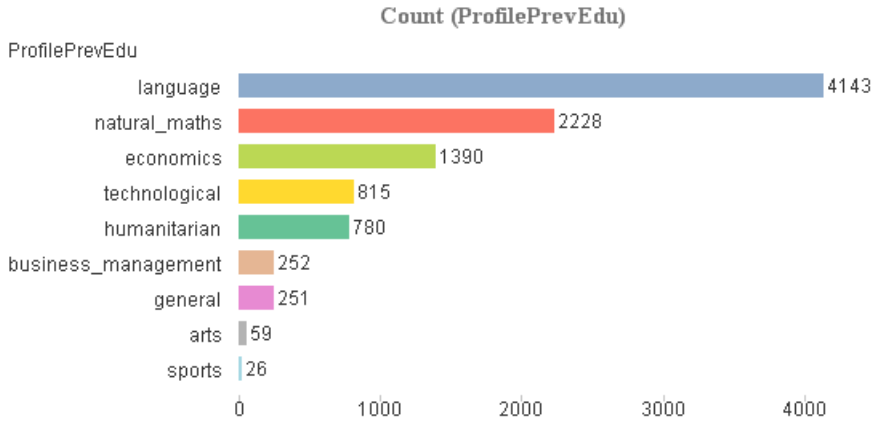
Подобна тенденция се наблюдава и по отношение на средния успех от средното образование на кандидат-студентите (Фиг.2.11), който намалява през трите последователни години.



Фиг.2.11: Среден успех от средното образование на кандидат-студентите, приети с различен приемен изпит, за периода 2007-2009

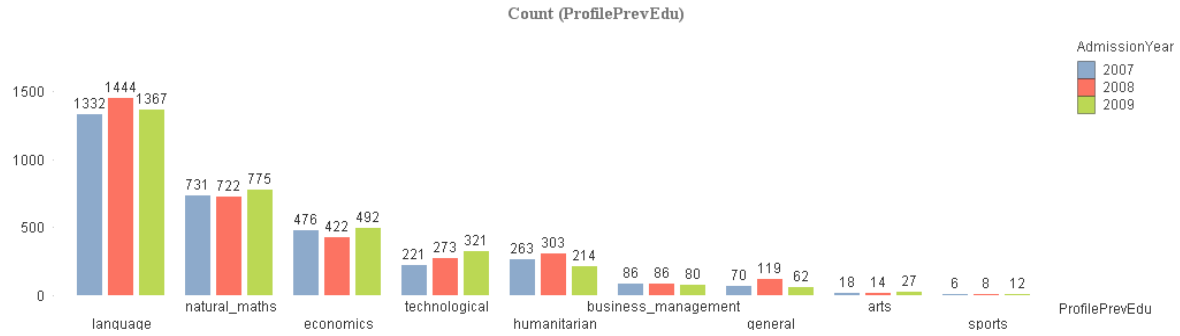
В. Профил на училището, в което кандидат-студентите са получили средно образование

За периода 2007-2009 най-много кандидат-студенти са завършили средни училища със следните профили: чуждоезиков (41.7%), природо-математически (22.4%), икономически (14%), технологичен (8.2%), хуманитарен (7.8%), бизнес и мениджмънт (2.5%), изкуства (0.6%) и спорт (0.2%), непрофилиран (2.5%) (Фиг.2.12).



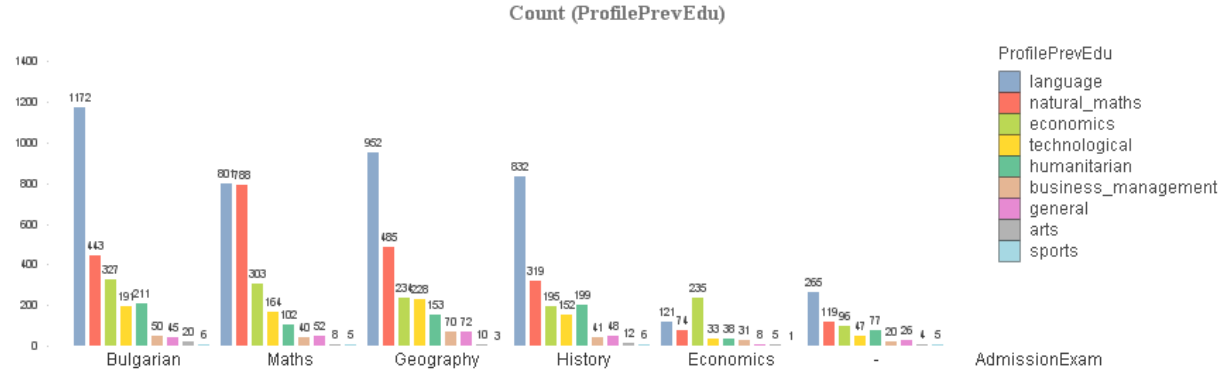
Фиг.2.12: Разпределение на кандидат-студентите по „Профил от средно образование”

Ситуацията е подобна през трите години на изследвания период 2007-2009 (Фиг.2.13), забелязва се по-трайна тенденция за нарастване броя на кандидат-студентите с технологичен профил и известно намаляване броя на кандидат-студентите с хуманитарен профил (по-съществено през 2009).

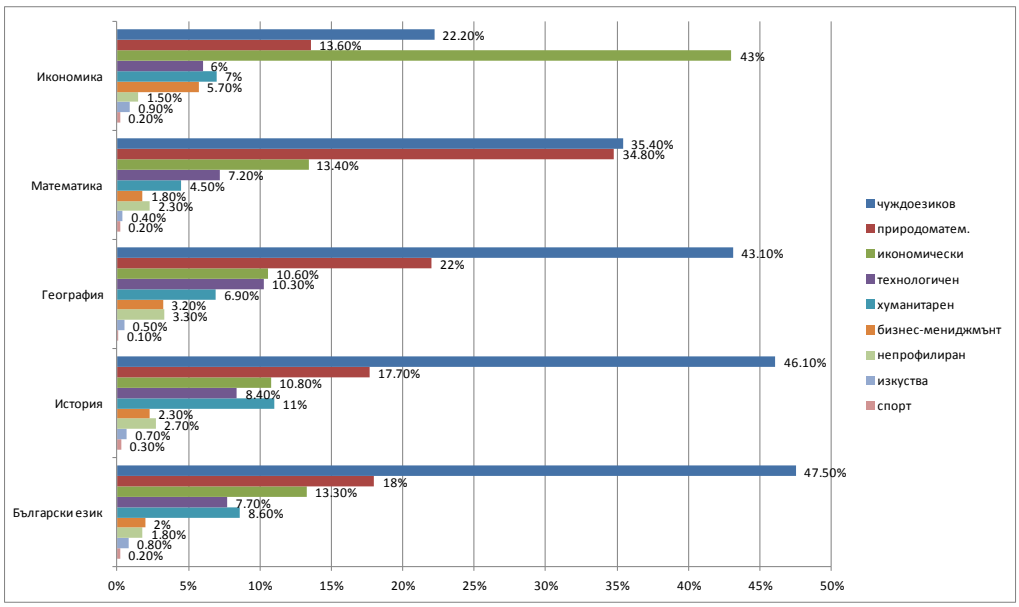


Фиг.2.13: Разпределение на кандидат-студентите по „Профил от средно образование", за периода 2007-2009

Както се вижда на Фиг.2.14, най-голям брой кандидат-студенти с чуждоезиков профил от средното образование са приети с изпит по български език (1172), следват изпитите по география (952) и история (832). С изпит по математика са приети най-вече кандидат-студенти с природоматематически (788) и чуждоезиков (801) профил от средното образование (приблизително еднакъв брой). С география и история също са приети най-много кандидат-студенти с чуждоезиков (съответно 952 и 832) и природоматематически (съответно 485 и 319) профил. Очаквано, най-предпочетен е изпитът по икономика от кандидат-студентите с икономически профил.



Фиг.2.14: Разпределение на кандидат-студентите по „Профил от средно образование" и „Приемен изпит"



Фиг.2.15: Разпределение на кандидат-студентите по „Приемен изпит" и „Профил от средно образование"

На Фиг.2.15 е представено разпределението на студентите в зависимост от приемния изпит и профила от средното образование. Подобни са установените закономерности при студентите, приети с изпити по български език, география и история. Различни са закономерностите при приетите с изпити по математика и икономика.

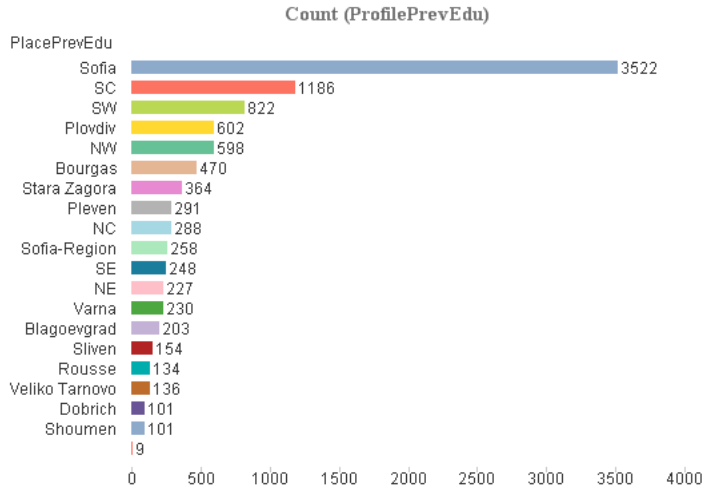
Почти половината от студентите приети с изпити по български език (47.5%), история (46.1%) и география (43.1%), са с чуждоезиков профил от средното образование. Около 1/5 от студентите (18-22%, приети с тези изпити, са с природоматематически профил - български език (18%), история (17.7%) и география (22%). Около 11-13% от студентите, приети с тези изпити, са с икономически профил - български език (13.3%), история (10.8%) и география (10.6%).

Приблизително по 1/3 от студентите, приети с изпит по математика, са с чуждоезиков (35.4%) и природоматематически (34.8%) профили. Останалат 1/3 от студентите, приети с изпит по математика, са с икономически (13.4%), технологичен (7.2%) и хуманитарен профил (4.5%).

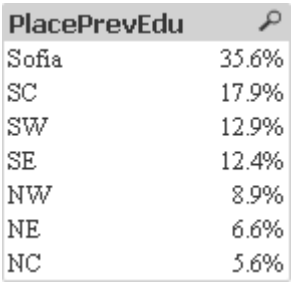
Около половината от студентите, приети с изпит по икономика, са с икономически (43%) и бизнес-мениджмънт (5.7) профили. Около ¼ са приетите студенти с чуждоезиков профил (22.2%). Останалите приети и икономика са с хуманитарен (7%), технологичен (6%) и бизнес-мениджмънт (5.7) профили.

Студентите с общообразователен профил и профили изкуства и спорт са сравнително равномерно разпределени между 5-те изпита, с които са приети студентите в УНСС през периода 2007-2009, т.е. не се наблюдават съществени изразени тенденции.

Г. Регион, където се намира училището, в което кандидат-студентите са получили средно образование



Фиг.2.16а



PlacePrevEdu	
Sofia	35.6%
SC	17.9%
SW	12.9%
SE	12.4%
NW	8.9%
NE	6.6%
NC	5.6%

Фиг.2.16б

Фиг.2.16а,б: Разпределение на кандидат-студентите по „Регион - средно образование”

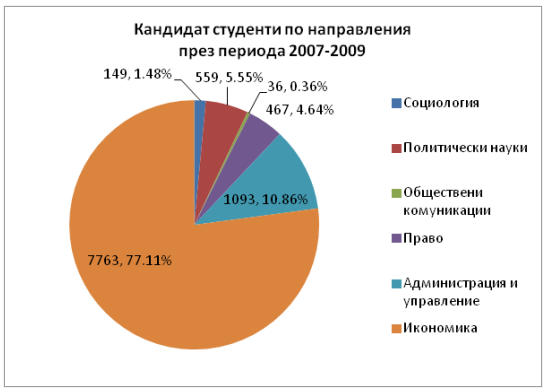
По отношение на региона, където кандидат-студентите са завършили средното си образование (Фиг.2.16а и Фиг.2.16б), най-много са завършилите в София-град - 35.6%, южен централен регион (ЮЦ) – 11.9%, югозападен регион (ЮЗ) – 8.2%, Пловдив и северозападен регион (СЗ) – по 6%, Бургас - 4.7%, Стара Загра – 3.7%, Плевен и северен централен регион (СЦ) – 2.9%, София-област – 2.6%, югоизточен регион (ЮИ) –2.5% и североизточен регион (СИ) – 2.3%. Следват останалите големи градове в България – Варна (2.3%), Благоевград (2%), Сливен (1.5%), Русе и Велико Търново (1.4%), Добрич и Шумен (1%). Ситуацията е подобна за трите години на изследвания период 2007-2009.

Не се забелязва някаква съществена закономерност при избора на кандидат-студентски изпит от кандидат-студенти, завършили средното си образование в различни градове и региони на страната.

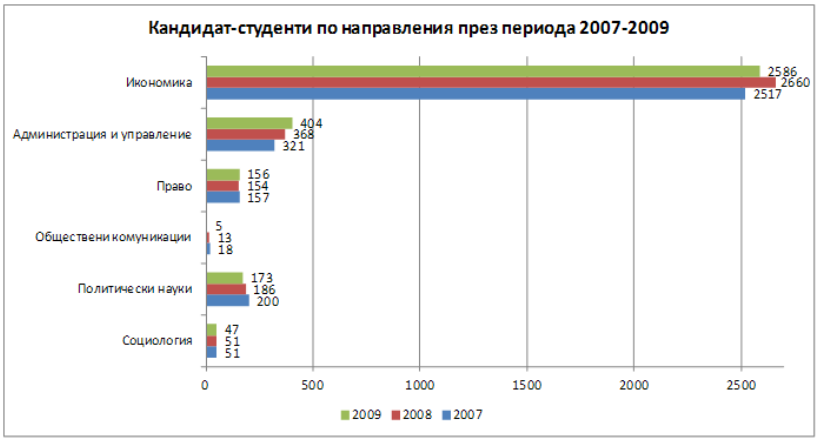
Д. Професионални направления, в които са приети студентите

На Фиг.2.17 и Фиг.2.18 е представено разпределението на кандидат-студентите по направления през изследвания период 2007-2009г. Най-голям е процентът на студентите, приети в направление „Икономика” – 77.11%, следват направленията „Администрация и

управление – 10.86%, ”„Политически науки” – 5.55%, „Право” – 4.64%, „Социология” – 1.48% и „Обществени комуникации” – 0.36%.



Фиг.2.17: Разпределение на кандидат-студентите по направления

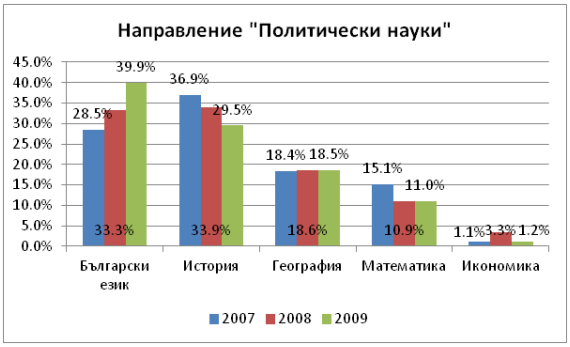


Фиг.2.18: Разпределение на кандидат-студентите по направления за периода 2007-2009

През трите години от периода 2007-2009 не се забелязват особени различия относно броя на приетите студенти в направления „Икономика”, „Право” и „Социология”. Съществува обаче тенденция за увеличаване броя на студентите в направление „Администрация и управление” и намаляване броя на студентите в направления „Политически науки” и „Обществени комуникации”.

В направление „Социология” преобладават студентите от женски пол (57.7%). Приети са предимно с изпит по български език (41,4%) и история (33.6%), по-малко - с география (16.4%), и най-малко - с математика (6.4%) и икономика (2.1%). Тази тенденция се запазва и през трите години на изследвания период (Фиг.2.19).

Фиг.2.19: Разпределение по „Приемен изпит" в направление „Социология"



Фиг.2.20: Разпределение по „Приемен изпит" в направление „Политически науки"

Студентите в направление „Политически науки" (Фиг.2.20) също са преобладаващо от женски пол (54.4%). Най-много през изследвания период 2007-2009 са приетите с изпити по български език (33.8%) и история (33.5)%, по-малко – с география (18.5%) и математика (12.3%), и най-малко с икономика (1.9%). Забелязва се тенденция за нарастване броя на приетите с изпит по български език, намаляване броя на приетите с изпити по история и математика, и запазване броя на приетите с изпит по география.

Фиг.2.21: Разпределение по „Приемен изпит" в направление „Обществени комуникации и информационни науки"

Фиг.2.22: Разпределение по „Приемен изпит" в направление „Право"

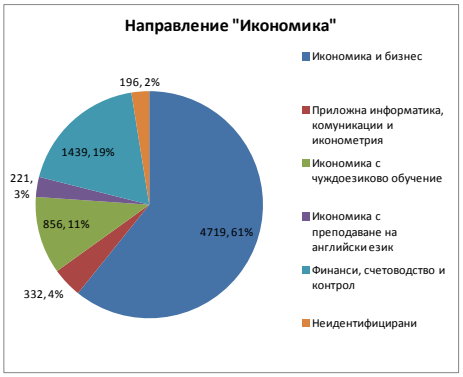
В направление „Обществени комуникации и информационни науки" през изследвания период 2007-2009 (Фиг.2.21) са приети равен брой студенти от двата пола – 50/50%, като през 2007 са приети по равен брой, през 2008 са приети повече мъже (61.5%), а през 2009 – повече жени (80%). С изпит по български език са приети 60.7% от студентите, следват приетите с изпити по математика – 14.3%, география - 10.7%, история и икономика – по 7.1%. Както е показано на Фиг.2.21, няма ясно изразена тенденция през трите години на изследвания период. И през трите години най-много са студентите приети с български език, но техният брой значително пада от 2007 (64.3% - 18 студента) до 2009 (39.9% - 5 студента). В някои години е голям броя на студентите приети с изпит по математика, история и икономика, но през други години няма нито един студент приет с тези изпити.

В направление „Право” през периода 2007-2009 (Фиг.2.22) са приети почти равен брой студенти от двата пола (51.2% жени и 48.8% мъже), като няма съществени различия през трите години. Най-голям е броят на студентите приети с изпит по история – 61.2% и български език – 37.9%, незначителен е броят на приетите с математика – 0.7% и география – 0.2%, нито един няма приет с изпит по икономика. Тази тенденция се запазва и през трите години – 2007 (български език – 49.6%, история – 47.2%), 2008 (български език – 34.2%, история – 65.8%), 2009 (български език – 31.5%, история – 68.5%).

Фиг.2.23: Разпределение по „Приемен изпит” в направление „Администрация и управление”

В направление „Администрация и управление” (Фиг.2.23) са приети почти равен брой стиденти от двата пола (51.9% жени и 48.1% мъже). Най-много са приетите с изпит по български език – 31,1%, история – 25.9% и география – 23.8%, по-малко с математика – 16.9% и най-малко с икономика – 2.2%. Тази тенденция се запазва през трите години на изследвания период.

Студентите от направление „Икономика” (7763 човека) представляват 77% от общата съвкупност от данни, използвани в изследването. Направление „Икономика” включва 5 поднаправления, разпределението на студентите е представено на Фиг.2.24. Най-много студенти (4719 човека, 61%) са в поднаправление „Икономика и бизнес”, студентите в поднаправление „Финанси, счетоводство и контрол” са 19% (1439 студенти), в поднаправление „Икономика с чуждоезиково обучение” попадат 11% от студентите, в поднаправления „Приложна информатика, комуникации и иконометрия” и „Икономика с преподаване на английски език” са съответно 4% и 3%, останалите 2% са студенти в специалности, които в момента не са сред обявените специалности за кандидатстване в УНСС (обозначени на Фиг.2.24 като „неидентифицирани”).



Фиг.2.24: Разпределение по поднаправления на студентите в направление „Икономика”

Фиг.2.25: Разпределение по „Приемен изпит” на студентите от направление „Икономика”

В направление „Икономика” има приблизително еднакъв брой студенти от двата пола (51% жени и 49% мъже), като и през трите години на изследвания период 2007-2009г. има незначителен превес на жените. Най-много студенти са приети с изпити по математика (28%), география (25.7%) и български език (24.4%), следват изпитите по история (14.8%) и икономика (7.1%). Както е представено на Фиг.2.25, вижда се ясно изразена тенденция за нарастване броя на студентите приети с изпит по български език и география, и леко намаляване броя на студентите приети с математика (въпреки че този брой се запазва висок и през трите години). Неособено голям е броят на студентите приети с изпити по история и икономика, и се забелязва тенденция за намаляване на този брой през трите години на изследвания период 2007-2009 (от 16.6% до 12% за изпита по история, и от 10.2% до 5.2% за изпита по икономика.

Фиг.2.26: Разпределение по „Приемен изпит” на студентите от поднаправление „Икономика и бизнес”

Фиг.2.27: Разпределение по „Приемен изпит” на студентите от поднаправление „Приложна информатика, комуникации и иконометрия”

В поднаправление „Икономика и бизнес” (Фиг.2.26) лек превес имат студентите от женски пол (51.1%), като това е характерно за трите години на изследвания период 2007-2009г. Най-голям е броят на приетите с изпит по география (26.6%), математика (25.3%) и български език (25.3%), по-малък е броят на студентите приети с изпит по история

(14.9%) и най-малък – с изпит по икономика (7.6%). За периода 2007-2009г. се наблюдават ясно изразени тенденции на увеличаване броя на студентите приети с изпит по български език и география, и намаляване броя на студентите приети с изпит по математика, история и икономика.

Наличните данни за поднаправление „Приложна информатика, комуникации и икометрия" съдържат информация за студенти, приети през 2007 и 2008г. (Фиг.2.27). В това поднаправление преобладават студентите от женски пол (52.7%). Най-много са студентите приети с изпит по математика (33.7%), следват изпитите по география (25.4%) и български език (20.1%), по-малко – с история (14.5%) и най-малко – с икономика (6.3%). През 2007г. приетите с изпит по математика са 42.5% от всички студенти в това поднаправление, а студентите приети с география и през двете години 2007 и 2008 са около 25% (¼) от всички приети в поднаправлението. Нараства броят на приетите с български език (15%-23.9%) и история (11%-17%).

Фиг.2.28: Разпределение по „Приемен изпит" на студентите от поднаправление „Икономика с чуждоезиково обучение"

Фиг.2.29: Разпределение по „Приемен изпит" на студентите от поднаправление „Икономика с преподаване на английски език"

В поднаправление „Икономика с чуждоезиково обучение" (Фиг.2.28) са приети равен брой студенти от двата пола. Най-много са приетите с изпит по география (27.5%), български език (27.4%) и математика (26%), следват приетите с история (17%) и почти незначителен брой – с икономика (2%). Както е представено на Фиг.24, през трите години на изследвания период 2007-2009г. броят на приетите с география намалява съвсем малко, по-съществено пада броят на приетите с математика (с 5-6%) и история (от 22% до 11.9%), за сметка на значително нарастване на броя на приетите с изпит по български език (от 18.2% през 2007 до 36.4% през 2009). Съвсем очаквано, 2/3 (70.7%) от приетите студенти в това поднаправление са завършили средното си образование в училища с чуждоезиков профил, останалите са предимно с природоматематически профил (18%).

И в поднаправление „Икономика с английски език" (Фиг.2.29) са приети равен брой момичета и момчета. Преобладават приетите с изпит по математика (37%), следват

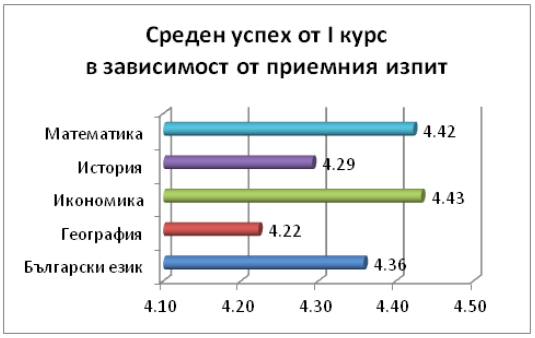
приетите с изпит по български език (25.7%), география (16.7%) и история (15.2%), най-малко са приетите с икономика (5.2). През трите години на изследвания период 2007-2009 (Фиг.2.29) се забелязва тенденция за намаляване броят на студентите приети с математика и нарастване броят на студентите приети с български език, броят на приетите с изпити по история и география варира в рамките на 10-20%. И в това поднаправление болшинството от приетите студенти са получили средното си образование в училища с чуждоезиков (64.1%) и природоматематически (26.8%) профил.

Фиг.2.30: Разпределение по „Приемен изпит" на студентите от поднаправление „Финанси, счетоводство и контрол"

В поднаправление „Финанси, счетоводство и контрол" са приети почти равен брой момчета и момичета (Фиг.2.30). Най-много са приетите с изпит по математика (36.3%), следват изпитите по география (23%) и български език (19.2), история (12.8%) и икономика (8.7%). За изследвания период 2007-2009 се наблюдават тенденции за нарастване броят на приетите с изпит по български език (10%-25%) и география (21-25%), и намаляване броят на приетите с изпит по математика (42%-32%), история (15%-9.7%) и икономика (10.2%-7.6%). Приетите студенти в това поднаправление са предимно с чуждоезиков (33.8%), природоматематически (30.1%) и икономически (21.8%) профили от средното образование, като съществено впечатление прави големият брой студенти с икономически профил.

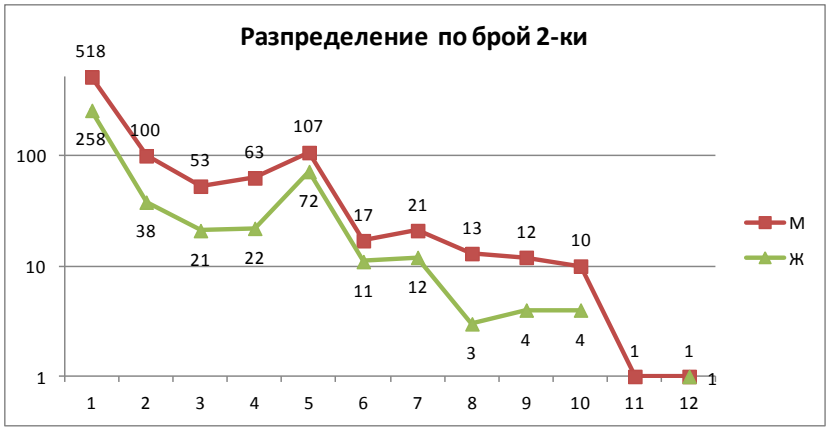
Е. Успех на студентите в I курс от обучението в университета

Средният успех на студентите, постигнат през първата учебна година в УНСС (Фиг.2.31), е най-висок за приетите с изпити по икономика (4.43) и математика (4.42), следват български език (4.36), история (4.29) и география (4.22). Но, тези разлики не са особено съществени (максималната разлика е 0.20).



Фиг.2.31: Среден успех на студентите в университета в зависимост от „Приемен изпит”

От общият брой студенти (10067), 86.5% (8704) не са получили слаби оценки (2-ки) през първата година на обучението си в университета, а останалите 13.5% (1362) са получили от 1 до 12 слаби оценки (разпределението по брой получени слаби оценки в зависимост от пола на студента е показано на Фиг.2.32. Сред студентите без слаби оценки преобладават жените (54.3%), а сред студентите с двойки значително преобладават мъжете.



Фиг.2.32: Разпределение по „Брой слаби оценки” и „Пол”

Изводи

При първоначалното изучаване на данните, използвани в настоящата дисертация , бяха направени ***следните изводи за студентите приети в УНСС***:

А. Относно Пола и Възраста

- Приемат се приблизително еднакъв брой студенти от двата пола.
- Лек превес на студентите от женски пол, приети в професионални направления „Социология” и „Политически науки”.

- Преобладават студентите на възраст 18-20 години, но основната част от студентите започват обучението си в университета на 19 години (78%).

Б. Относно Изпита, с който са приети

- Най-голям брой приети с изпит по български език, математика и география.
- Нараства броят на приеманите с изпит по български и география, и се запазва броят на приеманите по математика
- С изпити по история и икономика се приемат по-малко, като този брой намалява за икономика и е променлив за история.
- Средният успех, постигнат на петте изпита, е подобен, но се забелязва трайна тенденция за намаляване по всички предмети през изследвания период.
- Намалява и средният успех от средното образование на приетите с петте изпита за изследвания период.

В. Относно Профила на училището, в което са получили средно си образование

- Най-много за завършилите средни училища с чуждоезиков и природо-математически профил, по-малко са тези с икономически, хуманитарен и технологичен профил.
- Трайна тенденция за нарастване на броя с технологичен профил и известно намаляване на броя с хуманитарен профил.
- От чуждоезиков профил са приемани предимно с изпити по български език, география и история, но не е малък и броят на приетите с изпит по математика.
- От природоматематически профил са приемани предимно с изпит по математика.
- За икономически профил най-предпочитан е изпитът по икономика.
- От технологичен и хуманитарен профил са разпределени пропорционално между петте изпита.
- С хуманитарен профил - по-голяма успеваемост на изпитите по история и български език.
- С технологичен профил - по-добро представяне на изпитите по география и математика.

Г. Относно Региона на училището, в което е получено средно образование

- Над 1/3 от студентите в университета са от София (36%).
- Около 44% от студентите са от южните региони на страната, като преобладават представителите на южен централен регион (около 18%).
- Най-малко са приетите от северните региони (около 21%), като най-малък е броят на студентите от северен централен регион (5.6%).

- Няма ясно изразени тенденции по отношение на изпитите, с които са приети, профила от средното образование и успеха на студентите от различните региони на страната.

Д. Относно Професионалните направления, в които са приети студентите

- Най-голям е броят на студентите в направление „Икономика" (77%), а най-малък – в направление „Обществени комуникации" (0.4%).
- За изследвания период няма особени разлики относно броя на приетите в направления „Икономика", „Право" и „Социология", но се увеличава броят на студентите в направление „Администрация и управление" и намалява броят на студентите в направления „Политически науки" и „Обществени комуникации".
- В направление „Икономика" най-много приети с изпити по математика, география и български език, като нараства броят на приеманите с български език и география, и леко намалява броят на приеманите с математика.
- В направление „Икономика" не е голям броят на приетите с история и икономика, и този брой намалява през изследвания период.
- Наблюдават се различия в поднаправленията на направление „Икономика".
    - В „Икономика и бизнес" и „Икономика с чуждоезиково обучение" най-голям е броят на приеманите с география.
    - В „Икономика с английски език", „Финанси, счетоводство и контрол" и „Приложна информатика, комуникации и иконометрия" са приемани най-много с математика.
    - За всички поднаправления нараства броят на студентите приемани с български език и намаляване броят на студентите приемани с математика.
- В направление „Администрация и управление" най-много са приеманите с български език, история и география.
- В направления „Социология" и „Политически науки" са приемани предимно с български език и история.
- В направление „Политически науки" нараства броят на приеманите с изпит по български език и намалява броят на приеманите с изпити по история и математика.
- В направление „Обществени комуникации и информационни науки" са приемани предимно с български език.
- В направление „Право" са приемани предимно с история, като намалява броят на приеманите с български език, а нараства броят на приеманите с история.

Е. Относно Успеха на студентите в I курс от обучението в университета

- Средният успех на студентите, постигнат през първата учебна година в УНСС, е най-висок за приетите с икономика и математика, следват приетите с български език, история и география. Но, тези разлики не са особено съществени (максималната разлика е 0.20).
- Успехът на студентите от женски пол е по-висок от успеха на студентите от мъжки пол, което се отнася както за успеха от средното образование и оценките от приемните изпити, така и за средния успех, постигнат през първата година на обучение в университета.
- Около 13.5% от студентите са получили слаби оценки (броят на 2-ките е между 1 и 12) през първата година на обучение в университета.
- Сред студентите без слаби оценки преобладават жените, а сред студентите с двойки значително преобладават мъжете.

2.2.2. Анализ на данните за класификация на морски цели

Изходните данни са събрани от изследователския екип на Бирмингамския университет в периода Февруари-Март 2010г. Параметрите, които характеризират всеки запис на радарна морска цел, включват както цифрови, така и номинални променливи, и описват различни аспекти - разстояние между радарите в използваната тестова радарна система, параметри на антената, метеорологични характеристики като скорост и посока на вятъра, записи на Доплеровия сигнал получен при преминаването на лодката през базовата линия между приемника и предавателя на радара.

Суровите данни, или записите на Доплеровия сигнал за засечените морски цели, събирани по време на реалните изпитания с разработената за целта FS радарна система, са обработени допълнително така, че от тях да се получи съвкупност от данни, подходяща за Data Mining анализи. Смисълът на тази предварителна обработка на сигнали от движещи се морски цели се състои в това, че по наличните записи на FS доплерови сигнали, получени при пресичането на базовата линия между радарите, се извършва предварително задачата на автоматично откриване на целта и последваща оценка на информативни параметри на сигнала или целта, необходими за последващото й класифициране с методите на Data Mining. Тези параметри на сигналите са: дължина на целта - косвено измервана чрез броя на семплите на открития Доплеров сигнал, и енергия на отразения Доплеров сигнал - косвено определяща общата големина на морския обект, измервана като произведение на средната мощност на отразения сигнал и неговата продължителност (по време). Времетраенето на отразения от лодките сигнал се изчислява като произведение на времевия интервал между импулсите, облъчващи целта, и техния брой.

Параметрите на сигнала са измерени на изхода на оригинална структура на MTI CA CFAR K/M-L процесор, работещ във времевата област, разработена от екипа на

българските изследователи. Това е алгоритъм за автоматично откриване и измерване на параметрите на сигнала на FS Доплерови цели, използващ:

- MTI (Moving Target Indicator) - декорелира смущението от морето чрез еднократно презпериодно изваждане на семплите на сигналната поредица;
- CA-CFAR (Cell Average Constant False Alarm Ratio) процесор - осредняващ откривател на единични семпли за поддържане на постоянна честота на лъжлива тревога;
- K/M-L тест - измервател на големината на откритията пакет от семпли с помощтта на K/M-L тест, на базата на който се извършва измерване на началото и края на Доплеровия сигнал (пакета).

За изчисляването на броя на откритите семпли на изхода на K/M-L детектора е използван стандартен математически оператор в MATLAB - преброяват се импулсите, които превишават адаптивната прагова стойност на осредняващ откривател на единични семпли, подържащ постоянна величина на лъжливата тревога.

Стандартна статистическа процедура за намиране на средно аритметично в MATLAB е използвана за приблизителното определяне на средната енергия на целта. Затова се използва само амплитудата след премахване на шума, т.е. разликата между всички амплитуди на сигналния пакет и усреднения праг на откриване на CA CFAR.

За да се изследва робастността на MTI CA CFAR K/M-L детектора, се използват оценки и на други параметри във времевата област. Това са оценки, които се получават на изхода на K/M-L детектора и включват коефициент на корелация и отношение сигнал/шум. Параметърът SNR се изчислява чрез стандартна функция в MATLAB като отношение между две стандартни отклонения на откритията пакет импулси след CA CFAR филтрирането и шума от тестовия прозорец. Тези параметри се оценяват като се използват едни и същи записи на сигнали от морски цели, но различни варианти на обработка на сигналите. Например, наличие или отсъствие на MTI (Moving Target Indicator).

Така извлечените данни са организирани в плосък файл – в случая Excel файл, тъй като този формат е подходящ за работа с избрания Data Mining софтуер WEKA, а и данните са ограничени. В случай, че радарната система работи в реални условия, количеството събирани данни ще бъде доста голямо, което ще доведе до необходимост от организирането им в база от данни.

Съвкупността от данни, с която са проведени Data Mining анализите, съдържа 80 записа на радарни морски цели, описани чрез 16 параметъра (включващи и предсказваната величина), представени в Табл.2.5.

Таблица 2.5: Обобщено описание на данните, използвани при класифицирането на радарни морски цели

Variable Name	VarType	Values	Missing
Trial Date	Nom	17/02/10 (43), 18/02/10 (10), 21/03/10 (14), 23/03/10 (13)	0 (0%)
Distance Between Radars	Num	Min=300m, Max=500m, Mean=330.6m, StdDev=57.22 (300m, 316m, 500m)	0 (0%)
Antenna	Nom	A1/2/V/A2/1/V; A1/3/H/A2/1/H	0 (0%)
Weather	Nom	Sunny (56), Gloomy (11), Raining (13)	0 (0%)
Wind Speed	Num	1 - 5.1 m/s	2 (3%)
Wind Direction	Nom	SE (42), S (22), NW (1), SW (10), W (3)	2 (3%)
Boat Direction	Nom	South (11), North (12)	57 (71%)
S/N Ratio Before PC	Num	0 – 93.04, Mean=37.246, StdDev=26.207	12 (15%)
S/N Ratio After PC	Num	0 – 65.59, Mean=17.937, StdDev=22.605	14 (18%)
Number of Pulses Before PC	Num	0 – 2557, Mean=1148.892, StdDev=720.049	15 (19%)
Number of Pulses After PC	Num	0 – 4361, Mean=863.424, StdDev=1113.801	14 (18%)
Correlation Before PC	Num	0.62 – 1, Mean=0.954, StdDev=0.099	17 (21%)
Correlation After PC	Num	0.008 – 0.982, Mean=0.676, StdDev=0.322	18 (23%)
Energy Before PC	Num	0 – 2.939, Mean=0.599, StdDev=0.712	14 (18%)
Energy After PC	Num	0 – 0.499, Mean=0.024, StdDev=0.067	13 (16%)
Target	Nom	BigBoat (11), MISL_Boat (62), AverageBoat (6)	1 (1%)

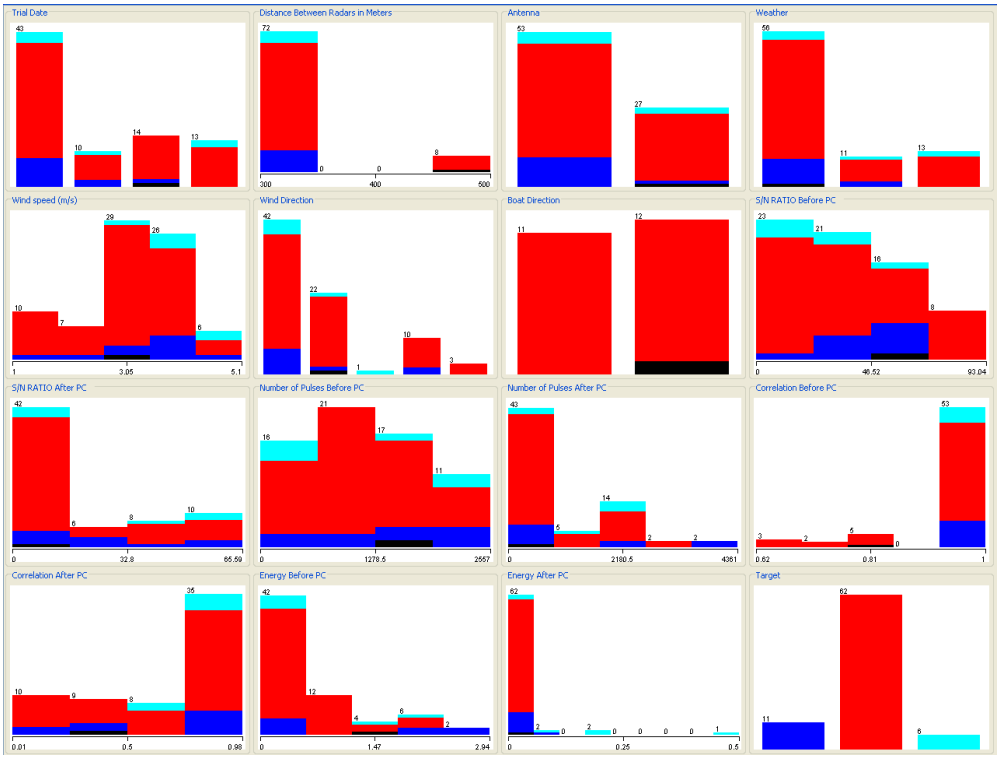
Както се вижда от Табл.2.5, за голяма част от параметрите има много липсващи стойности. Това се дължи както на липсваща информация в самите сурови данни от изпитанията, както и на затруднения при измерването на някои от параметрите.

Предсказваната целева променлива е *вида на откритата радарна цел*, който трябва да бъде определен чрез решаването на задачата за класификация. Оригиналните данни от изпитанията съдържат записи за 14 различни морски цели. Тъй като наличният обем от данни е много ограничен (80 записа), е счетено за целесъобразно тези 14 морски цели да бъдат обединени в три класа и така е получена предсказвана променлива от категориен тип с три различни стойности - *MISL Boat*, *Big Boat* и *Average Boat*.

В класа *MISL Boat* попадат записите на малка гумена моторна лодка, която изследователският екип от Бирмингамския университет използва за реалните изпитания.

В класа *Big Boat* попадат записите за по-големи лодки, които са били регистрирани по време на изпитанията и които в оригиналните данни фигурират като *Big boat*, *Big fishing boat*, *Big white boat*, *Huge boat*.

В класа *Average Boat* попадат записите за средно големи лодки, фигуриращи в оригиналните данни като *Life boat*, *Police boat*, *Speed boat*, *Medium yacht*, *Fast red boat*. Разпределянето на записите в избраните три класа е извършено чрез консултации с членовете на екипа, участващи в реалните изпитания.



Фиг.2.33: Разпределение на атрибутите в данните относно стойностите на предсказваната променлива в WEKA

На Фиг.2.33 е представено разпределението на 16-те атрибута (променливи), които се съдържат в данните, по отношение на трите стойности на изходната предсказвана променлива (Target – вида на откритата морска цел) – *MISL Boat* (с червен цвят), *Big Boat* (тъмно син цвят), *Average Boat* (светло син цвят).

Броят на записите в трите класа е различен. Най-много данни са налични за клас *MISL Boat* (62 записа), а доста по-ограничен е броят са записите за клас *Big Boat* (11) и клас *Average Boat* (6). За целевата променлива има само една липсваща стойност (Target Variable - 1% Missing Values - виж Табл.2.4), затова общият брой на записите в изследваната съвкупност от данни е 79.

Изводи:

- Получена е крайна съвкупност от данни за морски цели, която ще се използва за генериране на Data Mining модели за класификация (обучение на класификатори) чрез софтуер WEKA.
- Съвкупността от данни съдържа 80 записа за радарни морски цели, описани чрез 16 параметъра (включващи и предсказваната променлива).
- За голяма част от параметрите има много липсващи стойности, което се дължи както на липсваща информация в самите сурови данни от изпитанията, така и на затруднения при измерването на някои от параметрите.

- Предсказваната целева променлива е вида на откритата радарна цел - променлива от номинален тип с три различни стойности - *MISL Boat*, *Big Boat* и *Average Boat*. Броят на записите в трите класа е различен. Най-много данни са налични за клас *MISL Boat* (62 записа), а доста по-ограничен е броят са записите за клас *Big Boat* (11) и клас *Average Boat* (6).

2.3. Изводи по Втора Глава

В настоящата глава е постигнато следното:

- Избран е CRISP-DM подход за решаването на избраната Data Mining задача.
- Разгледани са различни софтуерни средства, които са подходящи за осъществяване на формулираните задачи. Избрани са *WEKA, QlikView* и *Excel*, които се използват за предварителна подготовка на данните и генериране на Data Mining модели за класификация (обучение на класификатори).
- Разработена е методика за решаване на задачата за извличане на знания от данни (Data Mining) чрез обучение на класификатори. Избрани са мерки за оценка и сравнение на генерираните Data Mining модели за класификация (обучените класификатори).
- Съгласно по-горе посочените подход и методика, е осъществена задачата за предварителна подготовка и изследване на данните за студентите, предоставени от УНСС. Оценено е разпределението на студентите, приети в УНСС през периода 2007-2009г., в зависимост от: пол и възраст; изпита, с който са били приети в университета, и получените оценки на този изпит; профила и региона на училището, в което кандидат-студентите са получили средно образование; професионални направления и поднаправления, в които са приети кандидат-студентите; успеха на студентите в I курс от обучението в университета.
- Получена е крайна съвкупност от данни за студентите, която ще се използва за генериране на Data Mining модели за класификация (обучение на класификатори) чрез софтуер WEKA.
    - Съвкупността от данни съдържа 10067 записа за студенти, описани чрез 14 атрибута (параметъра).
    - Изходната/предсказваната променлива „Среден успех от I курс” е преобразувана в номинална променлива по два начина (Вариант 1 - номинална променлива с пет стойности - слаб, среден, добър, много добър, отличен; Вариант 2 - номинална променлива с 2 стойности - слаби и силни).
- Получена е крайна съвкупност от данни за морски цели, която ще се използва за генериране на Data Mining модели за класификация (обучение на класификатори) чрез софтуер WEKA.

- Данните са в подходящ формат за прилагане на Data Mining методи и средства, получен в резултат от специфична обработка на оригиналните записи на сигнали от морски цели.
- Съвкупността от данни съдържа 80 записа за радарни морски цели, описани чрез 16 параметъра.
- Предсказваната целева променлива е вида на откритата радарна цел - променлива от номинален тип с три различни стойности - *MISL Boat*, *Big Boat* и *Average Boat*. Броят на записите в трите класа е различен. Най-много данни са налични за клас *MISL Boat* (62 записа), а доста по-ограничен е броят са записите за клас *Big Boat* (11) и клас *Average Boat* (6).

Глава 3: Резултати от изследването на Data Mining моделите за класификация (обучените класификатори)

В ***Трета глава*** са представени резултатите от изследването, осъществено чрез прилагане на избраните Data Mining алгоритми за класификация върху двете подготвени съвкупности от данни, за студенти и за морски цели, с помощта на Data Mining софтуер WEKA. Оценени и сравнени са генерираните Data Mining модели за класификация (обучените класификатори) чрез избраните мерки за оценка. Резултатите относно данните за студентите са описани в т.3.1, а резултатите относно данните за морски цели са представени в т.3.2.

3.1. Генериране и оценка на Data Mining класификационни модели (обучени класификатори) за предсказване успеваемостта на студентите

3.1.1. Класификационни алгоритми, реализирани в Data Mining софтуер WEKA

За решаването на Data mining задачата за класификация са използвани различни алгоритми – генератор на правила (Rule Learner), дърво на решението (Decision Tree), невронна мрежа (Neural Network) и К-най-близък съсед (K-Nearest Neighbour).

Тези алгоритми са избрани, тъй като всеки от тях работи на различен принцип, т.е. по различен начин генерира модела за предсказване на целевата променлива. Затова, прилагането на тези алгоритми върху едни и същи данни може да доведе до получаването на различни резултати – в някои случаи определени алгоритми дават по-добри резултати, а в други случаи – други алгоритми се справят по-добре, по отношение предсказването на класа (това действително се вижда от получените резултати в представеното изследване). При решаването на data mining задачи винаги се препоръчва да не се разчита на прилагането на един единствен метод, дори когато той дава добри резултати, а да се сравни работата на няколко различни метода и да се избере оптималното решение за последващо прилагане върху нови данни. Друга причина за избора на тези алгоритми е тяхното често използване при решаването на различни Data Mining задачи в сферата на образованието (виж т.1.3).

Работата на избраните data mining алгоритми за класификация се изследва за два варианта на целевата променлива (с 5 и с 2 стойности), тъй като много често точността на алгоритмите зависи от броя на стойностите, които заема целевата променлива, което е пряко свързано и с представителността на всеки от предсказваните класове в общата съвкупност от данни.

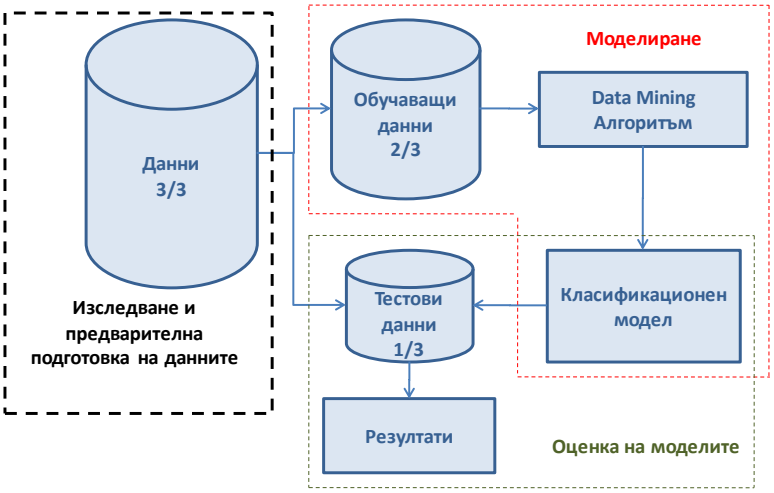
Data mining анализите в дисертацията се реализират със софтуерния пакет WEKA, предлагащ богато разнообразие от Data Mining алгоритми. Кратко описание на използваните алгоритми е представено в Табл.3.1.

Таблица 3.1: Описание на използваните Data Mining алгоритми в WEKA

Име на алгоритъма	Описание	Параметри
Генератор на правила – 1R WEKA ***OneR***	Алгоритъмът извършва предсказването на класа като използва само един атрибут – този с най-малка грешка при предсказване, като генерира дърво на решението на едно ниво. Числовите променливи се дискретизират.	Стойности по подразбиране (WEKA default parameters)
Дърво на решенията – C4.5 WEKA ***J48***	Класификаторът J48 се базира на алгоритъм C4.5 – изгражда дърво на решенията от налична съвкупност от данни като използва критерий за оценка чистотата на производните възли - Information Gain (всеки възел може да се разклонява на повече от 2 подвъзела).	1) Стойности по подразбиране (M=2) (WEKA default parameters) 2) Минималният брой записи във всяко листо M=2, 5,10,15,20,25,30,40, 50,70,100,200,500
Невронна мрежа WEKA ***MultilayerPerceptron***	Невронната мрежа представлява силно свързана система, съставена от множество елементарни изчислителни елементи (неврони) свързани чрез претеглени връзки и организирани в слоеве. Многослойният персептрон е невронна мрежа с един или повече скрити слоеве, с двупосочно извършване на изчисленията (feedback network).	За различни стойности на H - брой на невроните в скрития слой: 1) H=1 2) Стойност по подразбиране (WEKA default parameters) H=a=(attributes+classes)/2)
K-най-близък съсед (k-NN) WEKA ***IBk***	Метод за класифициране на обекти въз основа класа на принадлежност на най-близко разположените обекти в обектното пространство. Чрез алгоритъма k-NN класът на нов обект се определя чрез мнозинството от гласовете на намерените K на брой най-близки съседни обекти (всеки обект гласува за своя клас).	1) Стойности по подразбиране (WEKA default parameters) 2) За различни стойности на параметъра K: K=1,10,20,30,40,50,60,70, 80,90,100,150,200,250,300, 500,750,1000,1200

Избраните data mining алгоритми за класификация са приложени върху наличните данни при различни условия (test options): “percentage split” и „stratified 10-fold cross validation”.

При “percentage split test option” (holdout method - разделяне на общата съвкупност от данни на обучаваща и тестова извадка), 2/3 (66%) от данните се използват за обучаваща извадка, а останалата 1/3 (34%) – за оценка на генерирания модел. Тестовата схема “percentage split” (66%/34%) е представена на Фиг.3.1.



Фиг.3.1: Тестова схема “Percentage Split”

При “stratified 10-fold cross validation” общата съвкупност от данни се разделя на 10 части и алгоритъмът се прилага 10 пъти, като всеки път 9 от 10-те части се използват като обучаващи данни (training data), а останалата 1 част се използва за оценка на модела. Това е стандартен подход, който се прилага за определяне грешката на предсказване на класификационните алгоритми при наличието на една постоянна съвкупност от данни. Стратификацията на данните означава, че съвкупността от данни ще бъде разделена произволно на 10 части, но във всяка част всеки клас ще е представен приблизително в същите пропорции, както в цялата съвкупност от данни. Крайната грешка на алгоритъма се определя като средно аритметично от грешките, получени при 10-те изпълнения на алгоритъма.

При реализацията на класификационната задача се генерира модел, на базата на наличните стойности както за входните, така и за изходната променлива (класа), с който впоследствие се предсказва класът на принадлежност на нов обект.

За целите на изследването в дисертацията, за целева променлива (променлива, която ще бъде предсказвана) е избрана величината „Среден успех от I курс”. В оригиналната версия на наличните университетски данни тази променлива е от числов тип, затова е извършено нейното предварително преобразуване във величина от номинален тип

(дискретна качествена променлива). Това преобразуване е осъществено в два варианта (описано по-подробно в т.2.2.1.3.Г):

- ***Вариант 1***: целевата променлива „Среден успех от I курс" заема 5 различни стойности (слаб, среден, добър, много добър, отличен) (виж Табл.2.1)
- ***Вариант 2***: целевата променлива „Среден успех от I курс" заема 2 различни стойности (слаби, силни) (виж Табл.2.2)

Останалите променливи, които се съдържат в данните, се разглеждат като входни или предсказващи променливи (описание на променливите е дадено в т.2.2.1.4).

Реализацията на класификационната задача в двата случая, за двата варианта на предсказваната променлива, позволява да се изследва устойчивостта на предсказващия модел към промяна на данните, и/или да се избере по-ефективно класификационно правило.

3.1.2. Резултати за предсказвана променлива с пет стойности

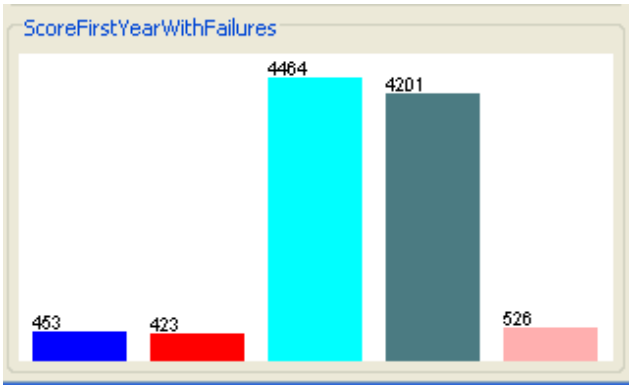
Предсказваната променлива е от *категориен тип* и заема *5 различни стойности* – „Слаб" (Bad), "Среден" (Average), „Добър" (Good), „Много добър" (Very Good) и „Отличен" (Excellent), създадени на базата на правилата в използваната шестобална система за оценяване в България (подробно обяснение е дадено в т.2.2.1.3.Г).

Използвани са четири Data Mining метода за класификация – генератор на правила (WEKA алгоритъм 1R), дърво на решенията (WEKA алгоритъм J48), невронна мрежа (WEKA алгоритъм MultilayerPerceptron) и К-най-близък съсед (WEKA алгоритъм IBk).

Генерираните модели за класификация (обучени класификатори) са оценени и сравнени чрез получените стойности на избрани параметри за оценка – *точност/грешка на предсказване, матрица на класификация и базираните на нея параметри, Kappa Statistic, ROC Area, Model Correct/Incorrect Ratio* (по-подробно описани в т.1.2.4 и т.2.1.3). Оценено е и подобрението по отношение точността на предсказване, което се получава чрез генерираните класификационни модели с избраните алгоритми, спрямо наивната класификация (Naïve Classification) - класификация на случаен принцип без използване на модел за предсказване, изготвен на базата на обучаваща извадка, чрез параметъра *Model/Naïve Correct/Incorrect Ratio*. Това се прави с цел да се получи реална оценка за качеството на получените модели за класификация (обучени класификатори), т.е. за да се види реалното подобряване на класификацията чрез генерираните модели.

Матрица на наивна класификация

На Фиг.3.2 е представена хистограмата на изходната/предсказваната променлива.



Фиг.3.2: Хистограмата на изходната (предсказвана) променлива в WEKA

Разпределението на изходната/предсказваната променлива – Среден успех на студентите през първата година от обучението им в университета (ScoreFirstYearWithFailures), е описано в Табл.3.2, за общата съвкупност от данни и за тестовата извадка (при тестова опция %-съотношение, тестовата извадка е 34%, т.е. 1/3 от общата съвкупност 10067, или 3423 записа).

Таблица 3.2: Разпределението на променливата „ScoreFirstYearWithFailures" за общата съвкупност от данни и за тестовата извадка

	Full Dataset		**Test Dataset**	
	Number	**%**	**Number**	**%**
Class Bad	453	4.5%	154	4.5%
Class Average	423	4.2%	145	4.2%
Class Good	4464	44.3%	1517	44.3%
Class Very Good	4201	41.7%	1428	41.7%
Class Excellent	526	5.2%	179	5.2%
Total	**10067**		**3423**	

Матрицата при наивна класификация е представена в Табл.3.3. Методът за изчисляване матрицата при наивна класификация е подробно обяснен при получаването на матрицата при наивна класификация за случая на двоична изходна/предсказвана променлива (т.3.2).

Таблица 3.3: Матрица на случайна класификация при предсказвана променлива с 5 стойности

	Classified as Bad	**Classified as Average**	**Classified as Good**	**Classified as Very Good**	**Classified as Excellent**	**Total**
Is Bad	**7**	7	68	64	8	154
Is Average	7	**6**	64	60	8	145
Is Good	68	64	**672**	633	80	1517
Is Very Good	64	60	633	**597**	74	1428
Is Excellent	8	8	80	74	**9**	179

Total	154	145	1517	1428	179	3423

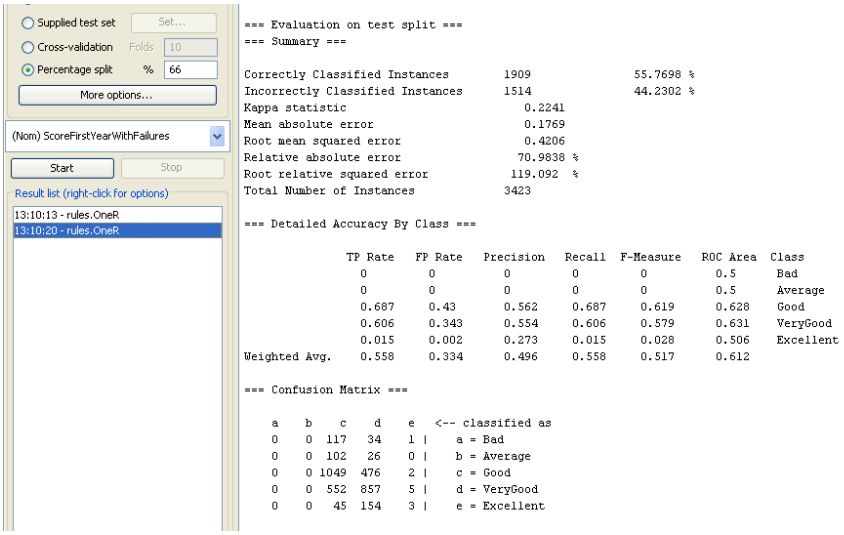
Naïve Correct/Incorrect Ratio**=(7+6+672+597+9)/(3423-1291)=1291/2132=**0.61

Получената стойност 0.61 на отношението Правилно/Неправилно класифицирани записи за наивната класификация показва, че броят на правилно класифицирани обекти е по-малък от броя на неправилно класифицирани обекти (39/61%), т.е. на всеки 0.61 правилно класифицирани обекти се получава една грешка.

Резултатите от класификацията за четирите избрани алгоритъма, представени по-долу, са получени за тестова опция „% разделяне", т.е. при разделяне на общата съвкупност от данни на две части – 2/3 (66%) се използват като обучаваща извадка, т.е. за генериране на модела за предсказване, а 1/3 (34%) се използват като тестова извадка, т.е. за тестване на модела.

Алгоритъм 1R

Резултатите от работата на 1R алгоритъма са представени на Фиг.3.3.



Фиг.3.3: WEKA резултати от работата на *1R* алгоритъм

Входната променлива, която се използва от *1R* класификатора, е *AdmissionScore*, т.е. *Общ бал при приемане на студентите в университета* е атрибутът, който дава минимална грешка при класификацията. Това е същата входна променлива, която е използвана от 1R алгоритъма и при случая с двоична изходна променлива.

Точността на класификация на алгоритъма *1R* е само 55.77%, т.е. само около 56% от общата съвкупност от записи са правилно класифицирани. Освен това, забелязват се проблеми при предсказването на класове Слаб (Bad), Среден (Average) и Отличен

(Excellent), стойностите на параметъра TP Rate (True Positive Rate) за тези три класа са съответно 0, 0 и 0.015, т.е. на практика няма правилно предсказани записи за класове Слаб и Среден, и има само 3 правилно предсказани записи от клас Отличен. Това са трите класа, които са по-слабо представени в общата съвкупност от данни.

Параметърът Kappa Statistic, който показва степента на съгласуваност между предсказването и реалния клас, също има много ниска стойност – 0.2241 (много по-ниска стойност от 1, съответстваща на идеалната съгласуваност). Площта под ROC кривата е ROC Area=0.612 – стойност, която е по-ниска от 0.7, което означава че този модел не е добър за класификация.

Отношението между правилно и неправилно класифицираните обекти за генерирания модел на класификация чрез този алгоритъм е:

Model Correct/Incorrect Ratio=1909/1514=**1.26**

Получената стойност на съотношението Правилно/Неправилно класифицирани записи 1.26 показва, че броят на правилно класифицираните записи е 1.26 пъти по-голям от броя на неправилно класифицираните записи, т.е. на всеки 1.26 правилно класифицирани записи се получава една грешка.

За да се оцени работата на генерирания модел за класификация чрез алгоритъма *1R*, се сравняват получените отношения Правилно/Неправилно класифицирани записи за модела и при наивна класификация (стойностите за наивна класификация за получени по-горе):

$$\frac{Model\ \frac{Correct}{Incorrect}\ Ratio}{Naive\ \frac{Correct}{Incorrect}\ Ratio} = \frac{1.26}{0.61} = 2.07$$

Получената стойност 2.07 показва, че генерираният модел извършва класификацията 2.07 пъти по-добре от наивната класификация (т.е. при липса на модел за предсказване).

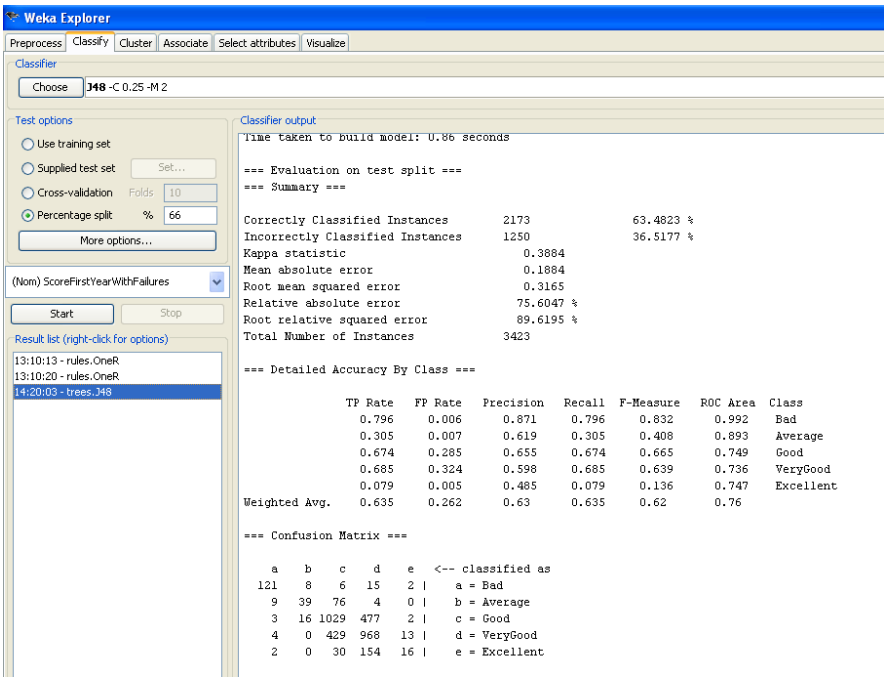
<u>Изводи за модела на класификация, получен чрез алгоритъма 1R:</u>

- Точността на класификация = 55.77%
- Моделът не предсказва успешно класове Слаб (Bad), Среден (Average) и Отличен (Excellent).
- Kappa Statistic=0.2241
- ROC Area=0.612
- Model Correct/Incorrect Ratio=1.26

- $\frac{Model\frac{Correct}{Incorrect}Ratio}{Naive\frac{Correct}{Incorrect}Ratio} = 2.07$

Алгоритъм „Дърво на решенията" (WEKA J48)

Резултатите от работата на J48 алгоритъма са представени на Фиг.3.4.



Фиг.3.4: WEKA резултати от работата на *J48* алгоритъм

Точността на класификация на алгоритъма J48 е 63.48%, т.е. 63.48% от общата съвкупност от записи са правилно класифицирани. Алгоритъмът работи добре по отношение на клас Слаб (Bad), Добър (Good) и Много Добър (Very Good), но не дава добра точност на предсказване по отношение на клас Среден (Average) и Отличен (Excellent). Параметърът Kappa Statistic, който показва степента на съгласуваност между предсказването и реалния клас, също има ниска стойност – 0.3884 (доста под 1, съответстваща на идеалната съгласуваност). Площта под ROC кривата е ROC Area=0.76 – стойност, която е по-висока от 0.7, което означава че този модел е по-добър от модела, генериран чрез *1R* алгоритъма.

Отношението между правилно и неправилно класифицираните обекти за генерирания модел на класификация чрез алгоритъма J48 е:

Model Correct/Incorrect Ratio=2173/1250=**1.74**

Получената стойност на съотношението Правилно/Неправилно класифицирани записи 1.74 показва, че броят на правилно класифицираните записи е 1.74 пъти по-голям

от броя на неправилно класифицираните записи, т.е. на всеки 1.74 правилно класифицирани записи се получава една грешка.

За да се оцени работата на генерирания модел за класификация чрез *дърво на решението*, се сравняват получените отношения Правилно/Неправилно класифицирани записи за модела и при наивна класификация:

$$\frac{Model\,\frac{Correct}{Incorrect}\,Ratio}{Naive\,\frac{Correct}{Incorrect}\,Ratio} = \frac{1.74}{0.61} = 2.85$$

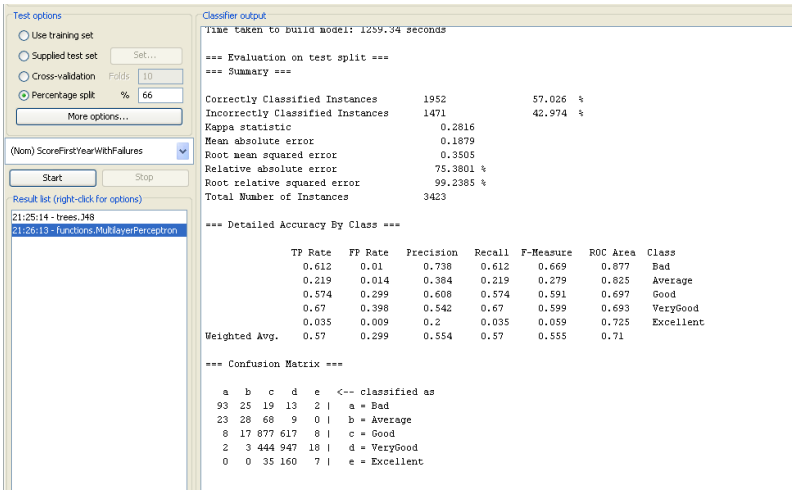
Получената стойност 2.85 показва, че генерираният модел извършва класификацията 2.85 пъти по-добре от наивната класификация (т.е. при липса на модел за предсказване).

Изводи за модела на класификация, получен чрез алгоритъма „Дърво на решението”:

- Точността на класификация = 63.48%
- Моделът предсказва с добра точност класове Слаб (Bad), Добър (Good) и Много Добър (Very Good), но не дава добра точност на предсказване по отношение на клас Среден (Average) и Отличен (Excellent).
- Kappa Statistic = 0.3884
- ROC Area = 0.76
- Model Correct/Incorrect Ratio=1.74
- $\frac{Model\,\frac{Correct}{Incorrect}\,Ratio}{Naive\,\frac{Correct}{Incorrect}\,Ratio} = 2.85$

Алгоритъм „Невронна мрежа” (WEKA Multilayer Perceptron)

Резултатите от работата на невронната мрежа са представени на Фиг.3.5.



Фиг.3.5: WEKA резултати от работата на *MultiLayerPerceptron* алгоритъм

Точността на класификация на получената невронна мрежа е 57.026%, т.е. само 57% от общата съвкупност от записи са правилно класифицирани. Алгоритъмът работи сравнително добре по отношение на клас Слаб (Bad), Добър (Good) и Много Добър (Very Good), не дава добра точност на предсказване по отношение на клас Среден (Average) и не се справя успешно с предсказването на клас Отличен (Excellent). Параметърът Kappa Statistic, който показва степента на съгласуваност между предсказването и реалния клас, също има ниска стойност – 0.2816 (доста под 1, съответстваща на идеалната съгласуваност). Площта под ROC кривата е ROC Area=0.71 – стойност, която е съвсем близка до 0.7, което означава че този модел също е по-добър от модела, генериран чрез 1R алгоритъма.

Отношението между правилно и неправилно класифицираните обекти за генерирания модел на класификация чрез невронната мрежа е:

Model Correct/Incorrect Ratio=1952/1471=**1.33**

Получената стойност на съотношението Правилно/Неправилно класифицирани записи 1.33 показва, че броят на правилно класифицираните записи е 1.33 пъти по-голям от броя на неправилно класифицираните записи, т.е. на всеки 1.33 правилно класифицирани записи се получава една грешка.

За да се оцени работата на генерирания модел за класификация чрез невронната мрежа, се сравняват получените отношения Правилно/Неправилно класифицирани записи за модела и при наивна класификация:

$$\frac{Model\ \frac{Correct}{Incorrect}\ Ratio}{Naive\ \frac{Correct}{Incorrect}\ Ratio} = \frac{1.33}{0.61} = 2.18$$

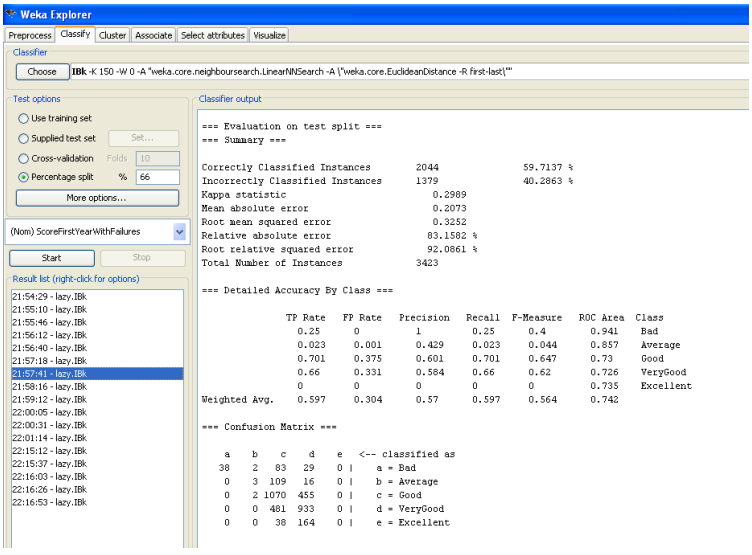
Получената стойност 2.18 показва, че генерираният модел извършва класификацията 2.18 пъти по-добре от наивната класификация (т.е. при липса на модел за предсказване).

Изводи за модела на класификация, получен чрез Невронна мрежа:

- Точността на класификация = 57.026%
- Моделът предсказва с по-висока точност класовете Слаб (Bad), Добър (Good) и Много Добър (Very Good), с по-ниска точност клас Среден (Average) и не се справя успешно с предсказването на клас Отличен (Excellent).
- Kappa Statistic = 0.2816
- ROC Area = 0.71
- Model Correct/Incorrect Ratio = 1.33
- $$\frac{Model\ \frac{Correct}{Incorrect}\ Ratio}{Naive\ \frac{Correct}{Incorrect}\ Ratio} = 2.18$$

Алгоритъм К-най-близък съсед (k-NN)

Алгоритъмът *К-най-близък съсед* е приложен върху изследваната съвкупност от данни за различни стойности на параметъра K=1,5,10,15,20,25,30,35,40,45,50,60,70,80,90,100,150. Най-голяма точност на класификация е получена при K=30. Резултатите от работата на kNN алгоритъма при K=30 са представени на Фиг.3.6.



Фиг.3.6: WEKA резултати от работата на *К-най-близък съсед* алгоритъм

Точността на класификация на алгоритъма е 59.71%, т.е. почти 60% от общата съвкупност от записи са правилно класифицирани. Алгоритъмът работи най-добре по отношение на класовете Добър (Good) и Много Добър (Very Good), не дава добра точност на предсказване по отношение на клас Слаб (Bad), а изобщо не се справя с предсказването на класове Среден (Average) и Отличен (Excellent). Параметърът Kappa Statistic, който показва степента на съгласуваност между предсказването и реалния клас, също има ниска стойност – 0.2989 (доста под 1, съответстваща на идеалната съгласуваност). Площта под ROC кривата е ROC Area=0.742 – стойност, която е по-висока от 0.7, което означава че този модел също е по-добър от модела, генериран чрез 1R алгоритъма.

Отношението между правилно и неправилно класифицираните обекти за генерирания модел на класификация чрез алгоритъма k-NN е:

Model Correct/Incorrect Ratio=2044/1379=**1.48**

Получената стойност на съотношението Правилно/Неправилно класифицирани записи 1.48 показва, че броят на правилно класифицираните записи е 1.48 пъти по-голям от броя на неправилно класифицираните записи, т.е. на всеки 1.48 правилно класифицирани записи се получава една грешка.

За да се оцени работата на генерирания модел за класификация чрез алгоритъма k-NN, се сравняват получените отношения Правилно/Неправилно класифицирани записи за модела и при наивна класификация:

$$\frac{Model\ \frac{Correct}{Incorrect}\ Ratio}{Naive\ \frac{Correct}{Incorrect}\ Ratio} = \frac{1.48}{0.61} = 2.43$$

Получената стойност 2.43 показва, че генерираният модел извършва класификацията 2.43 пъти по-добре от наивната класификация (т.е. при липса на модел за предсказване).

Изводи за модела на класификация, получен чрез алгоритъма К-най-близък съсед:

- Точността на класификация =59.71% (най-висока при К=30).
- Моделът предсказва с висока точност класовете Добър (Good) и Много Добър (Very Good), с по-ниска точност клас Слаб (Bad), а изобщо не се справя с предсказването на класове Среден (Average) и Отличен (Excellent).
- Kappa Statistic = 0.2989
- ROC Area = 0.742
- Model Correct/Incorrect Ratio=1.48
- $\frac{Model\ \frac{Correct}{Incorrect}\ Ratio}{Naive\ \frac{Correct}{Incorrect}\ Ratio} = 2.43$

Сравнение на получените модели за класификация

В Табл.3.4 са представени резултатите, получени при прилагането на избраните четири различни класификационни алгоритми върху изследваната съвкупност от данни (изходната променлива е номинална от тип категория - с 5 различни стойности) – генератор на правила 1R (WEKA OneR), дърво на решенията (WEKA J48), невронна мрежа (WEKA MultilayerPerceptron) и К-най-близък съсед (kNN – WEKA IBk).

Таблица 3.4: Резултати, получени в WEKA с Data Mining алгоритми *OneR, J48, MultiLayer Perceptron* и *IBk* при предсказвана променлива с 5 стойности

	1R Classifier	**Decision Tree**	**Neural Network**	**kNN Algorithm**
Correctly classified instances	55.77%	63.48%	57.026%	59.71%
Kappa Statistic	0.2241	0.3884	0.2816	0.2989
ROC Area	0.612	0.76	0.71	0.742
Model Correct/Incorrect Ratio	1.26	1.74	1.33	1.48
Model/Naïve Correct/Incorrect Ratio	2.07	2.85	2.18	2.43

С най-голяма точност на предсказване работи алгоритъмът „Дърво на решенията” (≈64%), следват алгротимите kNN (≈60%), невронна мрежа (≈57%) и 1R (≈56%). Генерираното дърво на решенията е моделът, за който са най-високи стойностите на параметрите Kappa Statistic (0.3884) и ROC Area (0.76), следователно при този модел се постига най-голямо съответствие между предсказания и действителния клас на обектите в изследваната съвкупност от данни, и е сравнително добър модел за предсказване (ROC Area>0.7). Стойностите на параметъра ROC Area са над 0.7 и за моделите, получени чрез алгоритмите kNN (0.742) и невронна мрежа (0.71), следователно и тези два алгоритъма биха могли да се използват за успешно предсказване. Алгоритъмът 1R е с много ниски стойности както за точността на предсказване (56%), така и за параметрите Roc Area (0.612), и за Kappa Statistic (0.2241), следователно моделът, генериран чрез този алгоритъм не е добър за предсказване класа на обектите в изследваната съвкупност от данни. Въпреки че получените стойности за отношението Правилно/Неправилно предсказани обекти не е особено висока и за четирите алгоритъма, се получава поне двойно подобрение при класифицирането на обектите в сравнение с наивната класификация (стойностите на отношението Model/Naive Correct/Incorrect Ratio са >2 за всички използвани алгоритми). Най-високи стойности за тези параметри отново се получават за

модела, генериран чрез дърво на решенията (Model Correct/Incorrect Ratio=1.74, Model/Naïve Correct/Incorrect Ratio=2.85), при който се получава 2.85 пъти подобрение в сравнение с наивната класификация.

Независимо от получените приемливите стойности на параметрите, чрез които се оценява общото качество на получените модели, трябва да се отбележи, че алгоритмите не работят еднакво добре по отношение предсказването на всичките пет класа. И четирите алгоритъма се справят добре при предсказването на класове „Добър" и „Много Добър". Дървото на решенията и невронната мрежа работят добре и по отношение на клас „Слаб". Всички алгоритми не предсказват точно клас „Среден", а най-ниска е точността на предсказване за клас „Отличен". Една от вероятните причини за по-големите неточности при предсказването на класовете „Отличен", „Среден" и „Слаб", вероятно е неравностойното представяне на тези класове в общата съвкупност от данни. Но, независимо от малкия брой обекти, принадлежащи към клас „Слаб" в общата съвкупност от данни (4.5%), все пак дървото на решенията и невронната мрежа показват висока точност на предсказване за този клас, доближаваща се до точностите на предсказване на клас „Добър" и „Много Добър". За предсказването на клас „Среден" алгоритмите „Дърво на решенията" и „Невронна мрежа" също постигат по-висока точност, отколкото за клас „Отличен", въпреки че клас „Среден" (4.2%) е по-малко представен от клас „Отличен (5.2%) в общата съвкупност от данни. Следователно, неравномерното представяне на петте класа в изследваната съвкупност от данни не е единствената причина за разликите в точността на предсказване. Тъй като в конкретния случай точността на предсказване на студентите, които принадлежат към клас „Отличен", е важна за крайните потребители на резултатите от изследването, това означава, че и четирите модела не се справят достатъчно успешно.

От сравнението на моделите за класификация, генерирани с избраните Data Mining алгоритми за предсказвана номинална променлива с 5 стойности, могат да се направят следните изводи.

<u>Изводи</u>:

- Най-добър е моделът, получен чрез алгоритъма *Дърво на решенията*, тъй като дава най-висока точност на предсказване ≈64%, както и най-високи стойности на параметрите Kappa Statistic=0.3884 и ROC Area=0.76;
- Моделите, получени чрез алгоритмите *kNN* и *Невронна мрежа,* работят по-слабо от полученото *Дърво на решенията.* Но, тъй като и за тези модели стойностите на параметъра ROC Area са >0.7 (за kNN 0.742, за Невронната мрежа 0.71), това означава, че и тези два модела могат да се използват успешно за предсказване.
- И четирите модела осигуряват поне *2 пъти по-точно предсказване* спрямо наивната класификация (класификация на случаен принцип) (стойностите на

отношението Model/Naive Correct/Incorrect Ratio са >2 за всички използвани алгоритми).

- Най-добрият модел, генериран чрез алгоритъма *Дърво на решенията*, подобрява предсказването *2.85 пъти* спрямо наивната класификация.
- Моделите не предсказват с еднаква точност петте класа:
  - *Четирите модела* предсказват с по-висока точност класовете „Добър" и „Много Добър".
  - *Дървото на решенията* и *Невронната мрежа* предсказват с по-висока точност и клас „Слаб".
  - *Четирите модела* не предсказват с висока точност клас „Среден".
  - Най-ниска е точността на предсказване за клас „Отличен".
- Проведените изследвания показват, че точността на предсказване при всички методи за класификация не е достатъчно висока, за да бъде използван един от тях на практика;

3.1.3. Резултати за предсказвана променлива с две стойности

Проучените до момента научни изследвания, както и проведените в настоящата дисертация, показват, че точността на предсказване при голяма част от съществуващите методи за класификация се подобрява, когато предсказваната променлива заема по-малък брой стойности. Тъй като получените резултати от класификацията при предсказвана променлива с пет различни стойности са недостатъчно добри, е предприето трансформиране на предсказваната променлива в двоична величина и провеждане на същите Data Mining анализи при тези нови условия. В този случай, изходната предсказвана променлива е двоична величина от тип категория, приемаща две стойности „Слаби" (Weak) и „Силни" (Strong), съответстващи на студенти с общ успех постигнат в I курс от обучението си в университета, съответно <4.50 и ≥4.50 (виж т.2.2.1.3.Г, Фиг.2.4 и Фиг.2.5).

На етап „Изследване на данните", допълнително в този случай е извършено изследване на *индивидуалното влияние на всяка от входните променливи върху изходната/предсказваната променлива*. За целта е използван алгоритъм „Дърво на решението" (WEKA J48), който се прилага последователно само върху две от променливите, съдържащи се в данните – изследваната входна променлива и изходната/предсказвана променлива. Във всеки от тези случаи на прилагане на алгоритъма, се получава опростено дърво на решенията, чрез което данните се разделят на два класа „Слаби" и „Силни" (двете възможни стойности на изходната променлива) само чрез стойностите на изследваната входна променлива (категориите, които заема, ако е дискретна променлива от категориен тип, или прагови стойности, ако е непрекъсната числова променлива).

На етап „Моделиране” са използвани същите четири метода за класификация – генератор на правила (WEKA алгоритъм 1R), дърво на решенията (WEKA алгоритъм J48), невронна мрежа (WEKA алгоритъм MultilayerPerceptron) и К-най-близък съсед (WEKA алгоритъм IBk), използвани в случая на предсказвана променлива с 5 различни стойности. По аналогичен начин са оценени и сравнени генерираните модели за класификация чрез получените стойности на избраните параметри за оценка – *точност/грешка на предсказване*, *матрица на класификация* и *базираните на нея параметри*, *Kappa Statistic*, *ROC Area* и *Model Correct/Incorrect Ratio*. Оценено е и подобрението по отношение точността на предсказване, което се получава чрез генерираните класификационни модели с избраните алгоритми, спрямо наивната (случайна) класификация, чрез параметъра *Model/Naïve Correct/Incorrect Ratio*.

В този случай (при двоична предсказвана променлива) *допълнително е изследвана работата на алгоритмите при различни стойности на избрани техни параметри*, чрез които може да се подобряват генерираните модели за класификация.

3.1.3.1. Изследване влиянието на всяка входна променлива върху предсказването

При тези изследвания изследователят не цели висока точност на класификация, а изследва на влиянието на всеки от входните параметри върху задачата на класификация.

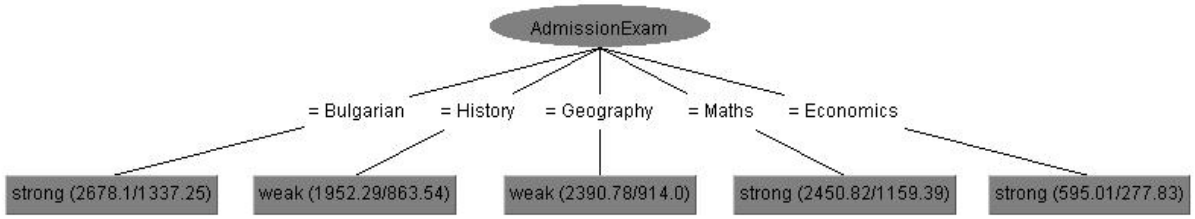
Осъществява се по следния начин:

- От съвкупността от данни се премахват всички променливи, с изключение на изследваната входна променлива и предсказваната променлива (класа).
- Прилага се избран алгоритъм за класификация „Дърво на решенията” – в случая това е алгоритъма WEKA J48 (алгоритъм C4.5), със стойности по подразбиране на параметрите на алгоритъма (default parameters), и тестване чрез крос-валидиране (10-fold Cross Validation Test Option).
- Оценява се качеството на класификация по числовите характеристики получени в листата на едно-нивово дърво на решенията, за всяка входна пороменлива.
- В изследваната съвкупност от данни се съдържат общо 13 входни променливи и една изходна променлива – предсказваният клас.
  Изследва се влиянието на 9 от наличните 13 входните променливи (*А-И*). Влиянието на останалите 4 входни променливи - *Година на раждане*, *Възраст*, *Година на приемане* и *Семестър*, не е изследвано. Счита се, че параметрите *Година на приемане* и *Семестър* не носят полезна информация относно успеха на студентите. Параметърът *Възраст* е производна променлива на параметъра *Година на раждане*, освен това, тъй като разпределението на записите в категориите на тези променливи (общо 29 на брой – цели числа в интервала 17-48) е много

неравномерно (по-голямата част от студентите са на възраст 18-20 години), също се приема, че влиянието им върху класификацията не представлява интерес.

А. Изследване влиянието на променливата „Приемен изпит" върху класификацията

На Фиг.3.7 е представено дървото на решенията, получено при класификацията с единствената входна променлива в случая за „Приемен изпит".



Ф

иг.3.7: Получено *Дърво на решенията* в WEKA за променливата „*Приемен изпит*"

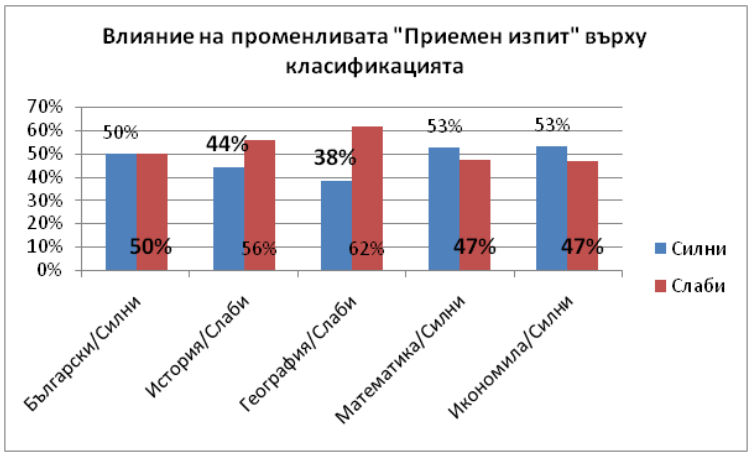
В Табл.3.5 е представена информация относно броя на приетите студенти със съответния приемен изпит, в изследваната съвкупност от данни (реалните данни) и в листата на класификационното дърво. Наблюдава се разлика между общия брой студенти в реалните данни и получените данни на класификатора. Класификационният алгоритъм разпределя всички 10067 студенти в петте листа на дървото, съответстващи на петте приемни изпита. Разликата от 677 студента се дължи на наличието на липсващи стойности за променливата „Приемен изпит" в реалните данни – те са точно 677 (7%), което напълно съвпада с първоначалното изследване на променливите (виж Табл.2.3). Студентите, за които липсва стойност за променливата „Приемен изпит" (677), са разпределени от класификационния алгоритъм пропорционално (по 7%) в петте листа на дървото, както се вижда от Табл.3.5.

Таблица 3.5: Брой студенти, приети със съответния приемен изпит, в реалните данни и в полученото класификационно дърво

Приемен изпит	Реални данни	Класификатор	Разлика	Разлика, %
Български език	2498	2678	180	7%
История	1821	1952	131	7%
География	2230	2391	161	7%
Математика	2286	2451	165	7%
Икономика	555	595	40	7%
Общо	9390	10067	677	7%

Трябва да се отбележи, че студентите, приети с изпити по Български език, Математика и Икономика, са разпределени от класификационния алгоритъм в листа от

клас „Силни”, а студентите приети с изпит по История и География – в листа от клас „Слаби” (Фиг.3.8). Разбира се, точността на класификация на алгоритъма в този случай е много ниска – едва 55% са правилно класифицираните студенти от алгоритъма, което е очаквано, ***затова създаденият модел не може да се използва успешно за класификация*** (а и това не е целта на това изследване – изследва се влиянието на всяка от променливите върху класификацията).



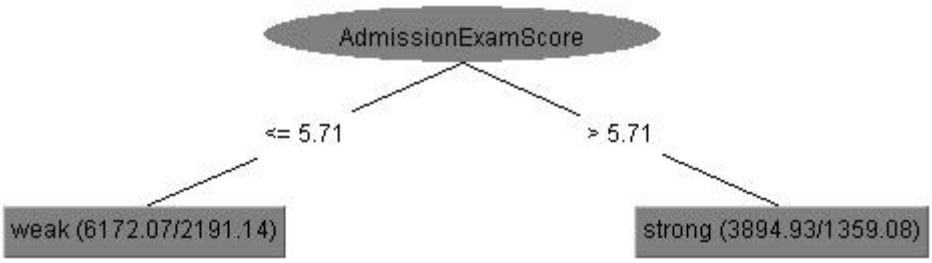
Фиг.3.8: Влияние на променливата „*Приемен изпит*” върху класификацията

Забелязва се разлика и по отношение точността на класификация в петте листа на дървото (Фиг.3.8) – най-голям е % на грешно класифицираните студенти, приети с изпит по Български език (50%), следват приетите с изпит по Математика и Икономика (47%), История (44%) и География (38%). По-голям % грешно класифицирани студенти има за клас „Силни” (2774 грешно предсказани от общо 5724, т.е. 49%), а по-малко – за клас „Слаби” (1778 грешно предсказани от общо 4343, т.е. 41%).

Изводи: Следователно, в този случай, за променливата „Приемен изпит”, алгоритъмът ***предсказва с по-голяма точност клас „Слаби”, отколкото клас „Силни”***.

Б. Изследване влиянието на променливата „Оценка от приемен изпит” върху класификацията

На Фиг.3.9 е представено дървото на решенията, получено при класификацията с единствената входна променлива „Оценка от приемен изпит” (AdmissionExamScore).



Фиг.3.9: Получено *Дърво на решенията* в WEKA за променливата „*Оценка от приемен изпит*”

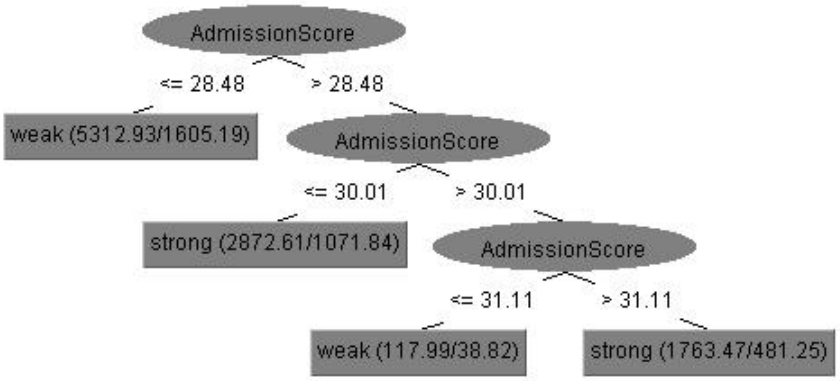
Праговата стойност за разделяне на записите в двата класа е 5.71. Студентите с оценка от приемен изпит ≤5.71 са класифицирани като „Слаби”, а студентите с оценка от приемен изпит >5.71 са класифицирани като „Силни”. Точността на класификация е 66% (34% са грешно класифицираните записи), т.е. правилно предсказаните записи са почти два пъти повече от грешно предсказаните. Това не е неочаквано, тъй като се предполага, че кандидат-студентите, които са се представили най-добре на приемните изпити, ще се представят добре и по време на обучението си в университета, особено през първата година (именно успехът от първата година на следването е величината, която се използва за формирането на изходната предсказвана променлива).

В реалните данни съотношението Слаби/Силни студенти е 53%/47%. Съгласно полученото дърво на решенията на Фиг.3.9, студентите в клас „Слаби” са 61%, а в клас „Силни” – 39%.

Изводи: Това означава, че класификационният алгоритъм, на базата на ***оценката от приемния изпит***, определя клас „Слаби” за по-голяма част от студентите, отколкото реално се съдържат в данните, т.е. ***алгоритъмът приоритетно определя клас „Слаби”***.

В. Изследване влиянието на променливата „Общ бал при приемане” върху класификацията

На Фиг.3.10 е представено дървото на решенията, получено при класификацията с единствената входна променлива „Общ бал при приемане” (AdmissionScore). Това е по-сложно дърво – с 3 нива, т.е. последователно са използвани три прагови стойности на входната променлива за разделяне на общата съвкупност от данни на два класа (28.48, 30.01 и 31.11).



Фиг.3.10: Получено *Дърво на решенията* в WEKA за променливата „*Общ бал при приемане*”

В този случай се получават две листа на дървото с клас Слаби” и две листа с клас „Силни”. Силните студенти са с общ бал над 31.11 (1764 студента са класифицирани в това листо на дървото като броят на грешно предсказаните записи е 27%), т.е. това са студентите с най-висок бал, което е очаквана закономерност. В другото листо с клас „Силни” са попаднали студенти с 28.48<Бал≤30.01 (2873 студента са класифицирани в това листо на дървото като броят на грешно предсказаните записи е 37%). В първото листо с клас „Слаби” са попаднали студенти с Бал≤28.48 (5313 студента общо, от които 30% са грешно предсказани), което също е очаквана закономерност, защото това са студентите, приети с най-нисък бал при кандидатстването. В другото листо с клас „Слаби” са попаднали само 118 студенти, от които 33% са грешно класифицирани, и това са студенти с 30.01<Бал≤31.11. Броят на тези студенти е много малък спрямо общия брой студенти, класифицирани от избрания алгоритъм в клас „Слаби” (118 представлява само 2% от общия брой 5431=5313+118). Грешките на предсказване както за клас „Силни”, така и за клас „Слаби” са в интервала 27-37%, т.е. в този случай класификационният алгоритъм дава сравнително висока точност на предсказване и за двата класа.

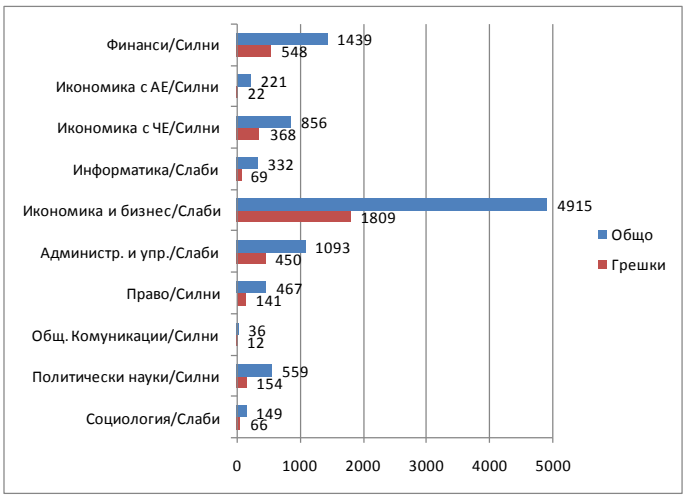
Изводи: Общата точност на алгоритъма е 69% (69% правилно класифицирани записи), следователно ***входният параметър „Бал на приемане”*** е величина, която сравнително ***добре разделя общата съвкупност от данни на двата класа „Слаби” и „Силни”***.

Г. Изследване влиянието на променливата „Направление/Поднаправление” върху класификацията

Дървото на решенията, получено от изследване на взаимодействието между входната променлива “Направление/Поднаправление” (UnivSpecialtyName) и изходната променлива (класа), съдържа 10 листа, съответстващи на 10-те

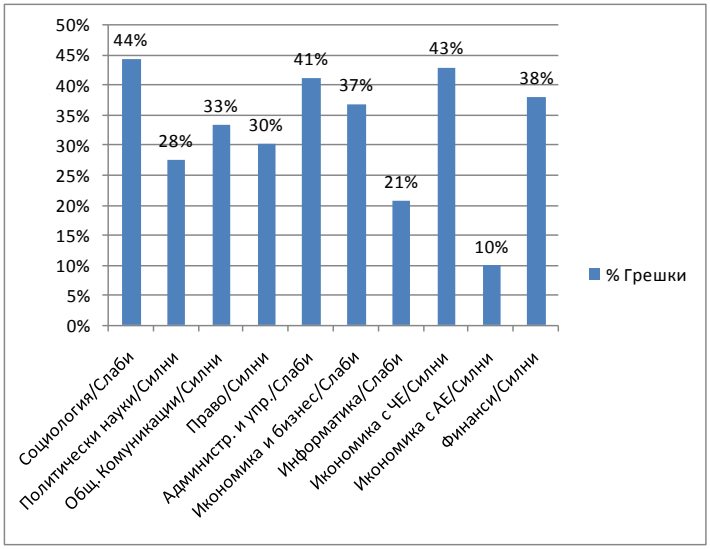
направления/поднаправления, в които се обучават студентите в университета. Студентите от направления/поднаправления Политически науки, Обществени комуникации, Право, Икономика с чуждоезиково обучение, Икономика с английски език и Финанси са разпределени от класификационния алгоритъм в листа с клас „Силни", а студентите от направления/поднаправления Администрация и управление, Икономика и бизнес, Социология и Приложна информатика - в листа с клас „Слаби". Класът е правилно предсказан за 64% от записите в съвкупността от данни, неправилна е класификацията за 36% от записите.

На Фиг.3.11 е представено разпределението на записите в десетте листа на дървото. За всяко листо е показан общият брой на записите, достигнали до съответното листо, както и броят на грешно класифицираните записи. Класът на всяко листо е показан заедно със стойностите на променливата Направление/Поднаправление.



Фиг.3.11: Разпределението на записите в десетте листа на полученото *Дърво на решенията* в WEKA

На Фиг.3.12 е показана и големината на грешката, т.е. процента на грешно класифицираните записи, във всяко от десетте листа на дървото.



Фиг.3.12: Грешка при класификация на записите в десетте листа на полученото *Дърво на решенията* в WEKA

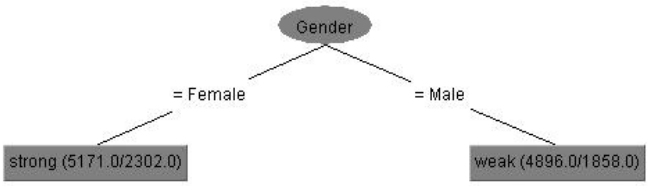
На Фиг.3.12 се вижда, че грешката е най-голяма за направление Социология (44%). Това означава, че броят на „Слабите" студенти от това направление (56%) е близък до броя на „Силните" студенти, който представлява 44-те % грешно класифицирани записи в клас „Слаби". Аналогично, грешката за поднаправление Икономика с чуждоезиково обучение е 43%, което означава, че „Силните" студенти тук са 57%, а „Слабите" – 43%, т.е. независимо че това листо на дървото е с клас „Силни", броят на „силните" студенти не е много по-голям от броя на „слабите" студенти.

Подобни изводи могат да бъдат направени и за останалите направления/поднаправления с по-голяма % стойност на грешно класифицираните записи.

Изводи: ***Променливата „Направление/Поднаправление" не е подходящ параметър за класификация в двата класа „Слаби" и „Силни".***

Д. Изследване влиянието на променливата „Пол" върху класификацията

На Фиг.3.13 е представено дървото на решенията, получено при класификацията с единствена входна променлива „Пол" (Gender). Студентите от женски пол са класифицирани в листо на дървото с клас „Силни", а студентите от мъжки пол – в листо на дървото с клас „Слаби".



Фиг.3.13: Получено *Дърво на решенията* в WEKA за променливата „*Пол*”

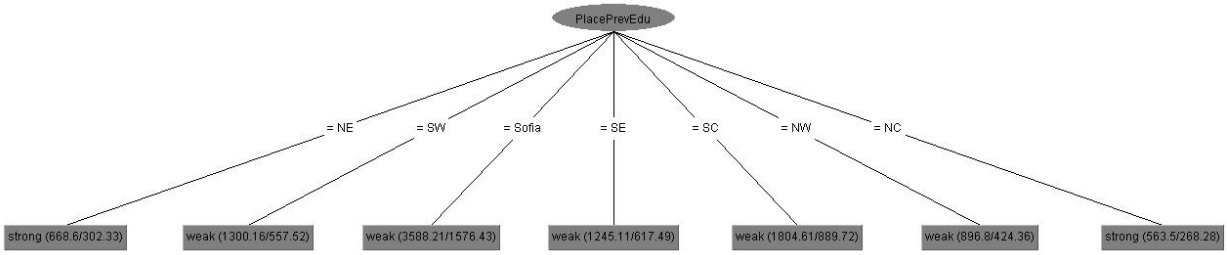
Точността на класификация на избрания алгоритъм в този случай е 59%, т.е. 59% са правилно класифицираните записи. Количеството на грешно класифицираните обекти е по-голямо за клас „Силни” (45%, 2301/5171), отколкото за клас „Слаби” (38%, 1858/4896).

Изводи: Това означава, че в действителност не всички студенти от женски пол са „силни” (45% са всъщност „слаби”), както и че не всички студенти от мъжки пол са „слаби” (38% са всъщност „силни”).

Изводи: ***Променливата „Пол” не е подходящ параметър за класификация в двата класа „Слаби” и „Силни”.***

Е. Изследване влиянието на променливата „Местоположение (средно образование)” върху класификацията

На Фиг.3.14 е представено дървото на решенията, получено при класификацията с единствена входна променлива „Регион, в който студентите са завършили средно образование” (PlacePrevEdu).



Фиг.3.14: Получено *Дърво на решенията* в WEKA за променливата „*Регион - средно образование*”

Студентите от само два региона – североизточен (NE) и северен централен (NC), са класифицирани в листа с клас „Силни”. Но, за тези две листа на дървото броят на грешно класифицираните записи е съответно 45% и 48%, което означава, че броят на слабите студенти е почти колкото броят на силните студенти в тези два региона. Студентите от останалите 5 региона – София, северозападен, югозападен, южен централен и югоизточен, са класифицирани в листа с клас „Слаби”. Количеството на грешно класифицирани записи

в тези листа е съответно 44%, 47%, 43%, 49% и 50%, което означава, че и за тези региони броят на „слабите" и „силните" студенти е приблизително равен.

Изводи: ***Точността на класификация в този случай е само 54%,*** следователно тази входна променлива ***не разделя качествено*** общата съвкупност от данни ***на двата класа – слаби и силни***.

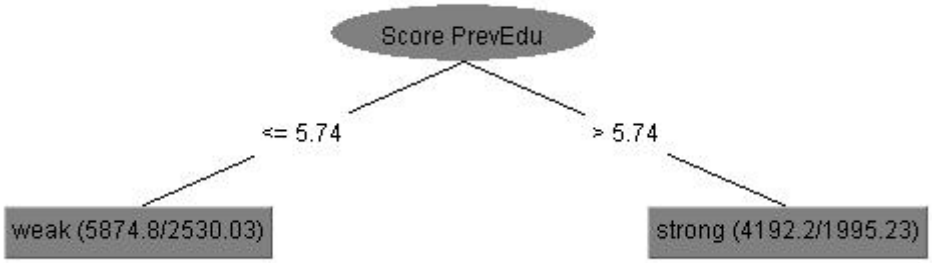
Ж. Изследване влиянието на променливата „Профил от средното образование" (ProfilePrevEdu) върху класификацията

Дървото на решенията, получено при класификацията с единствена входна променлива „Профил от средното образование" (PlacePrevEdu), съдържа само едно единствено листо – всички студенти от съвкупността от данни са класифицирани като „Слаби". Но, точността на класификация на алгоритъма е 53% (47% са грешно класифицираните записи в листото).

Изводи: Следователно входната променлива ***„Профил от средното образование" не позволява добро разделяне на студентите на „Слаби" и „Силни".***

3. Изследване влиянието на променливата „Успех от средното образование" (ScorePrevEdu) върху класификацията

На Фиг.3.15 е представено дървото на решенията, получено при класификацията с единствена входна променлива „*Успех от средното образование*" (ScorePrevEdu). Полученото дърво е с едно ниво и съдържа две листа.



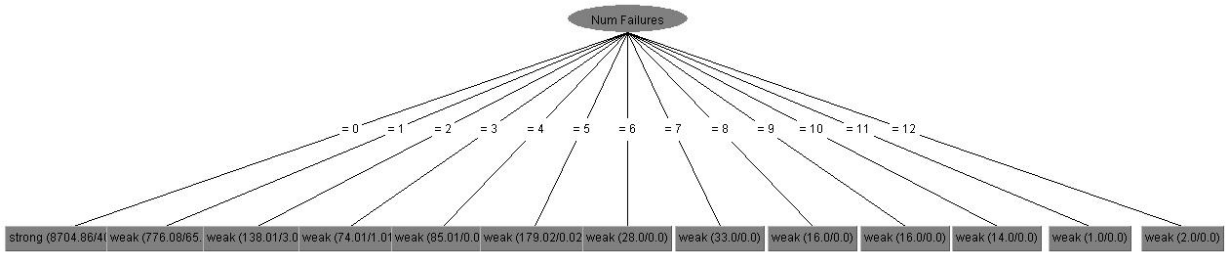
Фиг.3.15: Получено Дърво на решенията в WEKA за променливата „Успех от средното образование"

В листото с клас „Слаби” са класифицирани студентите с успех от средното образование ≤5.74, а в листото с клас „Силни” са разпределени студентите с успех от средното образование >5.74. Това е очаквана закономерност, тъй като се очаква кандидат-студентите с по-висок успех от средното образование да се справят по-добре в университета от тези с по-нисък успех. Точността на класификация за клас „Слаби” е 57% (43% са грешно класифицираните обекти в това листо), а за клас „Силни” – 52% (48% са грешно класифицираните обекти в това листо).

Изводи: Общата точност на класификационния алгоритъм в този случай е 55%, което означава, че входната ***променлива „Успех от средното образование” не е подходящ параметър за разделяне*** на съвкупността от данни ***на двата класа „Слаби и „Силни”.***

И. Изследване влиянието на променливата „Брой 2-ки” върху класификацията

На Фиг.3.16 и Фиг.3.17 е представено дървото на решенията, получено при класификацията с единствена входна променлива „*Брой 2-ки*” (NumFailures), съответно в графичен и текстови формат. Полученото дърво е с едно ниво и съдържа 13 листа, съответстващи на броя слаби оценки (2-ки), които имат студентите, включени в изследваната съвкупност от данни, в края на първата година от обучението си в университета – параметър, съдържащ цели стойности в интервала [0-12].



Фиг.3.16: Получено *Дърво на решенията* в WEKA за променливата „*Брой 2-ки*” в графичен формат

```
J48 pruned tree
------------------

Num Failures = 0: strong (8704.86/4047.0)
Num Failures = 1: weak (776.08/65.08)
Num Failures = 2: weak (138.01/3.01)
Num Failures = 3: weak (74.01/1.01)
Num Failures = 4: weak (85.01/0.01)
Num Failures = 5: weak (179.02/0.02)
Num Failures = 6: weak (28.0/0.0)
Num Failures = 7: weak (33.0/0.0)
Num Failures = 8: weak (16.0/0.0)
Num Failures = 9: weak (16.0/0.0)
Num Failures = 10: weak (14.0/0.0)
Num Failures = 11: weak (1.0/0.0)
Num Failures = 12: weak (2.0/0.0)

Number of Leaves  :     13

Size of the tree :      14
```

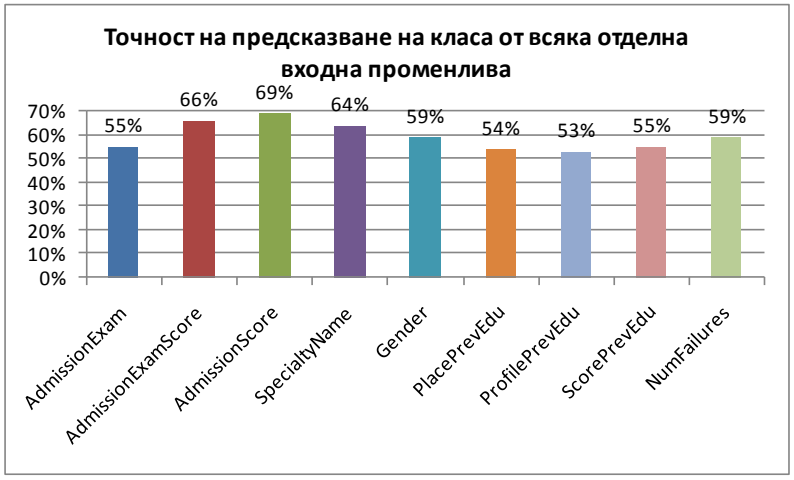
Фиг.3.17: Получено *Дърво на решенията* в WEKA за променливата „*Брой 2-ки*” в текстови формат

Единствено студентите без 2-ки (със стойност 0 на входната променлива „Брой 2-ки”) са класифицирани от алгоритъма в клас „Силни”, останалите са разпределени в листа с клас „Слаби”, което е очаквано и разбираемо, тъй като изходната променлива – класа, е формирана на базата на общия успех на студентите в края на първата година от обучението им в университета. Но, общата грешката на класификация на алгоритъма е 41% (правилно предсказаните обекти са 59%), като тя е много по-голяма за клас „Силни” (4047 грешно предсказани от общо 8705, т.е. 47%), отколкото за клас „Слаби” (69 грешно предсказани от общо 1362, т.е. 5%; тази грешка се получава основно от листото с клас „Слаби”, за което стойността на параметъра „Брой 2-ки” е 1).

Изводи: Следователно, в този случай класификационният алгоритъм работи добре само по отношение на клас „Силни”, но дава много голяма грешка на предсказване за клас „Слаби”, т.е. ***входната променлива „Брой 2-ки” не може да се използва за качествено предсказване на класа.***

Сравнение и изводи

Както беше споменато по-горе, в изследваната съвкупност от данни се съдържат общо 13 входни променливи и една изходна променлива – класа. На Фиг.3.18 са представени обобщените резултати от изследването на индивидуалното влияние на 9 от входните променливи върху класификацията, и по-точно резултатите по отношение точността на класификация на избрания класификационен алгоритъм (Дърво на решенията).



Фиг.3.18: Резултати от от изследването на индивидуалното влияние на входните променливи върху класификацията

Изводи:

- ***Най-висока точност на предсказване*** се получава в случая, когато като единствена входна променлива е използвана величината „***Бал на приемане”*** (AdmissionScore).
- ***Точността на предсказване чрез параметъра „Бал на приемане” е 69%***, което означава, че броят на правилно класифицираните записи е 2.23 пъти (69%/31%) по-голям от броя на неправилно класифицираните записи. Това е една доста висока точност на класификация като се има предвид, че само една единствена входна променлива (от общо 13) се използва за предсказване.
- Другите входни параметри, при които се получава сравнително висока точност на предсказване са *Оценка от приемен изпит* (AdmissionExamScore, 66%), *Направление/поднаправление* в което е приет кандидат-студента (SpecialtyName, 64%), *Пол* (Gender, 59%) и *Брой слаби оценки – 2-ки* (NumFailures, 59%).
- Получените резултати потвърждават предварителното предположение, че ***индивидуалното въздействие на всяка от входните променливи***, свързани с *Профила*, *Региона* и *Успеха от средното образование* на кандидат-студентите, и *Изпита*, с който са приети в университета, ***върху класификацията е по-слабо*** (получената точност на предсказване е 53-55%). Следователно, ***тези параметри не могат да бъдат самостоятелно използвани за качествено разделяне на студентите на „Слаби и „Силни”.***

3.1.3.2. Генериране на модели за класификация с избраните Data Mining алгоритми в WEKA

Изследванията са осъществени върху цялата съвкупност от данни (с всички 14 променливи – 13 входни и 1 изходна). *Предсказвана променлива* е от *категориен тип* и заема *2 стойности* – „Слаби" (Weak) и „Силни" (Strong) (виж т.2.2.1.3.Г, Фиг.2.4 и Фиг.2.5).

Използвани са четири Data Mining метода за класификация – генератор на правила (WEKA алгоритъм 1R), дърво на решенията (WEKA алгоритъм J48), невронна мрежа (WEKA алгоритъм MultilayerPerceptron) и К-най-близък съсед (WEKA алгоритъм IBk), както и в случая за предсказвана променлива с 5 стойности. Избраните алгоритми са приложени при различни условия – за различни стойности на някои от параметрите за настройка на самите класификационни алгоритми, с цел да се изследват допълнителни възможности за повишаване на точността на предсказване.

Генерираните модели за класификация са оценени и сравнени чрез получените стойности на избрани параметри за оценка – *точност/грешка на предсказване, матрица на класификация и базираните на нея параметри, Kappa Statistic, ROC Area, Model Correct/Incorrect Ratio* (по-подробно описани в т.1.2.4 и т.2.1.3), както това бе направено в случая за предсказвана променлива с 5 стойности (виж. т.3.1.3). Оценено е и подобрението по отношение точността на предсказване, което се получава чрез генерираните класификационни модели с избраните алгоритми, спрямо наивната класификация (Naïve Classification) - класификация на случаен принцип без използване на модел за предсказване, изготвен на базата на обучаваща извадка, чрез параметъра *Model/Naïve Correct/Incorrect Ratio*. Това се прави с цел да се получи реална оценка за качеството на получените модели за класификация, т.е. за да се види реалното подобряване на класификацията чрез генерираните модели.

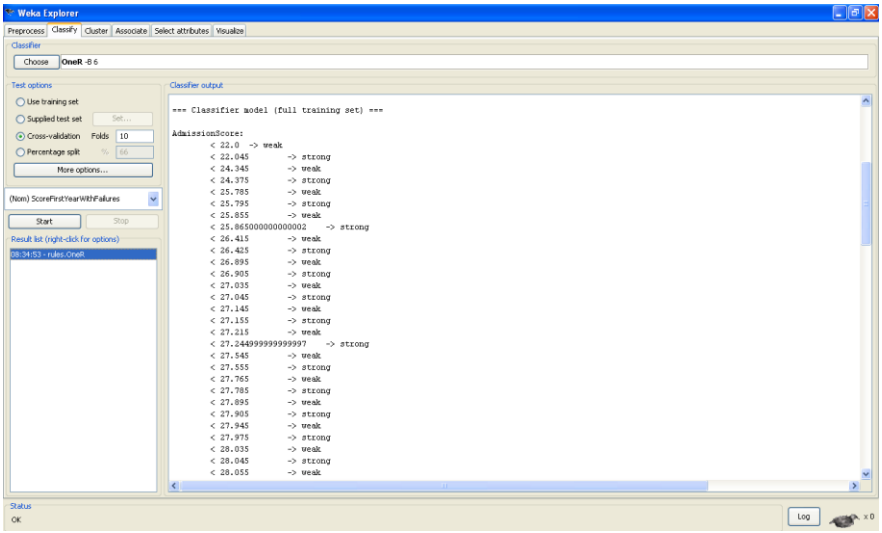
Алгоритъм за класификация 1R

Алгоритъмът *1R* (WEKA OneR Classifier) избира и използва една единствена входна променлива за извършването на класификацията – това е променливата с минимална грешка на предсказване (minimum-error attribute). Числовите променливи се дискретизират преди извършването на класификацията.

Алгоритъмът *1R* е приложен за цялата съвкупност от данни с всички налични 14 променливи (13 входни променливи и 1 изходна променлива – класа), като 5 от входните променливи с числови стойности са трансформирани в номинални (Семестър, *Година на раждане*, *Възраст*, *Година на приемане* и *Брой невзети изпити*). Резултатите от изпълнението на алгоритъма в избраното софтуерно Data Mining решение – WEKA,

включват генерирания модел за класификация и стойностите на параметрите за оценка на получения модел.

Моделът за класификация, получен чрез *1R* алгоритъма, представлява списък от стойности на избраната променлива, които се използват като прагови стойности за разделяне на обектите в общата съвкупност от данни на двата класа „Слаби" и Силни" На Фиг.3.19 е представена част от генерирания модел за класификация.



Фиг.3.19: Получен модел за класификация чрез алгоритъма *OneR* в WEKA

Резултатите за оценка на генерирания модел са получени за две тестови опции – крос-валидиране и %-разделяне, и са представени в Табл.3.6.

Таблица 3.6: Резултати за оценка на получения модел за класификация чрез алгоритъма *OneR* в WEKA

***1R* Algorithm**	**Test Options**					
	10-fold Cross Validation			**% Split (66/34)**		
	Weak	**Strong**	**Weighted Average**	**Weak**	**Strong**	**Weighted Average**
Correctly Classified Instances	66.4349%			67.4554%		
Incorrectly Classified Instances	33.5651%			32.5446%		
Kappa Statistic	0.322			0.3439		
TP Rate	0.731	0.589	0.664	0.7333	0.61	0.675
FP Rate	0.411	0.269	0.344	0.39	0.267	0.332
Precision	0.668	0.66	0.664	0.677	0.671	0.674
Recall	0.731	0.589	0.664	0.733	0.61	0.675
F-Measure	0.698	0.622	0.662	0.704	0.639	0.673
ROC Area	0.66	0.66	0.66	0.671	0.671	0.671

Входната променлива, която се използва от *1R* класификатора, е *AdmissionScore*, т.е. *"Общ бал"* при приемане на студентите в университета е атрибутът, който дава минимална грешка при класификацията.

От стойностите в Табл.3.6 се вижда, че при втората тестова опция - % разделяне на съвкупността от данни на две части (66% за обучение и генериране на модела, и 34% за тестване), *1R* класификаторът работи с около 1% по-голяма точност (67.4554% правилно класифицирани студенти), отколкото при първата тестова опция – крос-валидирането (67.4554% правилно класифицирани студенти). Тази малка разлика се отразява почти несъществено и върху останалите резултати на класификатора.

Параметърът Kappa Statistic показва степента на съгласуваност между предсказването и реалния клас. В случая неговата стойност за двете тестови опции е 0.322-0.3439 (доста под 1.0 - пълно съгласуване), което означава, че точността на предсказване на класификатора е ниска.

Получените по-горе резултати (в т.3.1.4.1) от класификацията на изследваната съвкупност от данни чрез използване на единствена входна променлива *Бал от приемането* (*AdmissionScore*) с WEKA алгоритъма *J48* показват, че точността на класификация с модела, генериран чрез алгоритъма „Дърво на решенията", е 69%. Алгоритъмът *1R (rule classifier)* също използва за извършване на класификацията една единствена входна променлива (променливата с минимална грешка на класификация), и това отново е променливата Бал от приемането (AdmissionScore), но дава по-ниска точност на предсказване – 66% (взимаме получения резултат за крос-валидиране, тъй като алгоритъмът J48 също е приложен за тази тестова опция). Следователно, алгоритъмът генериращ „Дърво на решенията" дава по-висока точност (с 3%) от генератора на правила 1R, което се дължи от една страна на разликата в същността на самите алгоритми, но и поради дискретизацията на числовите променливи, която се извършва преди осъществяването на класификацията от 1R алгоритъма и също води до загубване на част от информацията, която се съдържа в съвкупността от данни относно взаимовръзката между входните променливи и изходната/предсказвана променлива – класа.

Матрица на класификацията (Confusion Matrix)

Друг много важен резултат от работата на класификационните алгоритми, който се използва за оценка на точността на предсказване на алгоритмите, е матрицата с резултатите от класификацията – Confusion Matrix.

Матрицата с получените резултатите от класификацията (Confusion Matrix) с *1R* алгоритъма, при % разделяне на съвкупността от данни на обучаваща (66%) и тестова (34%) извадки, е представена в Табл.3.7.

Таблица 3.7: Матрица на класификация за получения модел чрез алгоритъма *OneR* в WEKA

1R Clasifier %-Split Test Option	Class Weak	Class Strong	Total

Is Weak	1324	483	1807
Is Strong	631	985	1616
Total	1955	1468	3423

Точността на предсказване на алгоритъма, представени в Табл.3.7, се изчислява чрез броя на правилно и неправилно класифицираните записи (стойностите са представени в Табл.3.8). От матрицата могат да бъдат изчислени и точностите на класификация за двата класа. В случая, точността на предсказване за клас „Слаби" (1324/1807=0.73, т.е. 73%) е по-висока от точността на предсказване на клас „Силни" (985/1616=0.61, т.е. 61%).

Таблица 3.8: Брой правилно/неправилно класифицирани записи за получения модел чрез алгоритъма *OneR* в WEKA

Модел	**Брой**	**% от тестовата извадка**
Правилно предсказани	1324+985=2309	68% (2309x3423=0.68)
Неправилно предсказани	631+483=1114	32% (1114x3423=0.32)

Отношението между правилно и неправилно класифицираните обекти за генерирания модел на класификация се изчислява от Табл.3.8:

Model Correct/Incorrect Ratio=2309/1114=**2.07**

Получената стойност на съотношението Правилно/Неправилно класифицирани записи 2.07 показва, че броят на правилно класифицираните записи е два пъти по-голям от броя на неправилно класифицираните записи, т.е. на всеки два правилно класифицирани записа се получава една грешка.

Матрица на наивна класификация

Тестовата извадка представлява 34% от общата съвкупност от данни и съдържа 3423 записа, чието разпределени в двата класа „Слаби" (1807, 53%) и „Силни" (1616, 47%) съответства напълно на разпределението на записите в общата съвкупност от данни (10067), показано на Фиг.2.4 („Слаби" – 5340, 53% и „Силни" – 4727, 47%).

За тестовата съвкупност от данни се изготвя подобна матрица, която да показва какви биха били резултатите при наивна класификация (Naïve Classification), т.е. при класификация на случаен принцип без използване на модел за предсказване, изготвен на базата на обучаваща извадка. Това се прави с цел да се получи реална оценка за качеството на получения модел за класификация, т.е. за да се види реалното подобряване на класификацията чрез генерирания модел. Матрицата на резултатите при наивна класификация е представена в Табл.3.9.

Таблица 3.9: Матрица на случайна класификация за предсказвана променлива с 2 стойности

Naïve Classification	Class Weak	Class Strong	Total
Is Weak	962	853	1815
Is Strong	853	756	1609
Total	1815	1609	3424

Матрицата при наивна класификация се изчислява по следния начин: В тестовата извадка има общо 3423 записа, разпределени съответно в клас „Слаби" (Weak) – 53% (1807 записа) и в клас „Силни" (Strong) – 47% (1606 записа). Стойностите в клетките на матрицата се изчисляват като:

- ***Случай 1***: Class Weak Is Weak
  (0.53x0.53)x3423=0.2809x3423=**962**
- ***Случай 2***: Class Weak Is Strong
  (0.53x0.47)x3423=0.2491x3423=**853**
- ***Случай 3***: Class Strong Is Strong
  (0.47x0.47)x3423=0.2209x3423=**756**
- ***Случай 4***: Class Strong Is Weak
  (0.47x0.53)x3423=0.2491x3423=**853**

Разликата между общия брой записи в матрицата на класификационния модел (3423 записа) и матрицата на наивна класификация (3424) е 1 запис и се дължи на закръгленията, направени при изчисляването на матрицата на наивна класификация.

Отношението Правилно/Неправилно класифицирани записи за наивната класификация се получава аналогично на получения модел за класификация:

Naïve Correct/Incorrect Ratio=(962+756)/(853+853)=1718/1706=**1.01**

Получената стойност 1.01 на отношението Правилно/Неправилно класифицирани записи за наивната класификация показва, че броят на правилно класифицирани обекти е равен на броя на неправилно класифицирани обекти (50/50%), т.е. на всеки правилно класифициран обект се получава една грешка.

За да се оцени работата на генерирания модел за класификация се сравняват получените отношения Правилно/Неправилно класифицирани записи за модела и при наивна класификация:

$$\frac{Model\ \frac{Correct}{Incorrect}\ Ratio}{Naive\ \frac{Correct}{Incorrect}\ Ratio} = \frac{2.07}{1.01} = 2.05$$

Получената стойност 2.05 показва, че генерираният модел извършва класификацията два пъти по-добре от наивната класификация.

Изводи за модела на класификация, получен чрез алгоритъма 1R:

- *Алгоритъмът 1R* извършва класификацията чрез входната променлива *AdmissionScore*, т.е. "*Общ бал*" при приемане на студентите в университета е атрибутът, който дава минимална грешка при класификацията.
- 1R класификаторът работи с малко по-голяма точност (около 1%) при % разделяне на съвкупността от данни на две части (66% за обучение и генериране на модела, и 34% за тестване), отколкото при крос-валидиране (десетократно изпълнение на алгоритъма като всеки път 9/10 от данните се използват за обучение и генериране на модела, а 1/10 - за тестване).
- Точността на класификация= 66%.
- *Дървото на решението*, получено чрез единствената входна променлива "*Бал от приемането*" (*AdmissionScore*) (виж т.3.1.4.1), извършва класификацията с точност 69%, т.е. с 3% по-висока точност от 1R алгоритъма.
- Kappa Statistic=0.3439
- Model Correct/Incorrect Ratio=2.07
- $\frac{Model\frac{Correct}{Incorrect}Ratio}{Naive\frac{Correct}{Incorrect}Ratio} = 2.05$

Алгоритъм за класификация „Дърво на решенията" - WEKA J48

Алгоритъмът *J48* е приложен за цялата съвкупност от данни с всички налични 14 променливи (13 входни променливи и 1 изходна променлива – класа), като 5 от входните променливи с числови стойности са трансформирани в номинални (Семестър, Година на раждане, Възраст, Година на приемане и Брой невзети изпити). Резултатите са получени за две тестови опции – крос-валидиране и %-разделяне, представени са в Табл.3.10.

От стойностите в Табл.3.10 се вижда, че и при двете тестови опциия – крос-валидиране и % разделяне на съвкупността от данни на две части (66% за обучение и генериране на модела, и 34% за тестване), алгоритъмът J48 работи с приблизително еднаква точност. Разликата в точността на класификация е само 0.33% и е в полза на втората тестова опция- % разделяне. Тази малка разлика се отразява почти несъществено и върху останалите резултати на класификатора.

Таблица 3.10: Резултати за оценка на полученото *Дърво на решенията* чрез алгоритъма *J48* в WEKA

***J48* Algorithm**	**Test Options**	
	10-fold Cross Validation	**% Split (66/34)**

	Weak	Strong	Weighted Average	Weak	Strong	Weighted Average
Correctly Classified Instances	72.4148%			72.7432		
Incorrectly Classified Instances	27.5852%			27.2568		
Kappa Statistic	0.4468			0.4524		
TP Rate	0.732	0.715	0.724	0.753	0.699	0.727
FP Rate	0.285	0.268	0.277	0.301	0.247	0.276
Precision	0.744	0.703	0.724	0.736	0.717	0.727
Recall	0.732	0.715	0.724	0.753	0.699	0.727
F-Measure	0.738	0.709	0.724	0.745	0.708	0.727
ROC Area	0.791	0.791	0.791	0.784	0.784	0.784

Параметърът Kappa Statistic показва степента на съгласуваност между предсказването и реалния клас. В случая неговата стойност за двете тестови опции е 0.45 (максималната възможна стойност е 1.0, показваща пълно съгласуване). Това означава, че точността на предсказване на класификатора не е много висока.

Проведени са експерименти при използване на различен брой входни променливи (последователно са премахвани част от променливите, които от експертна гледна точка се считат недостатъчно съществени по отношение на предсказването – например година на раждане на кандидат-студента, година на приемане в университета, семестър в който учи студента в момента на получаване на изследваната съвкупност от данни), както и за различни тестови опции (крос-валидиране и процентно разделяне). При всички тези експерименти е установено, че точността на предсказване е по-ниска отколкото получената точност при използване на всички входни променливи, затова следващите изследвания са проведени с пълната налична съвкупност от данни.

Матрицата с получените резултатите от класификацията (Confusion Matrix) с *J48* алгоритъма, при % разделяне на съвкупността от данни на обучаваща (66%) и тестова (34%) извадки, е представена в Табл.3.11.

Таблица 3.11: Матрица на класификация за полученото *Дърво на решенията* чрез алгоритъма *J48* в WEKA

J48 Classifier %-Split Test Option	Class Weak	Class Strong	Total
Is Weak	1361	446	1807
Is Strong	487	1129	1616
Total	1848	1575	3423

Точността на предсказване на алгоритъма се изчислява чрез броя на правилно и неправилно класифицираните записи (общо за алгоритъма и поотделно за всеки клас). В случая, общата точност на алгоритъма е 73% ((1361+1129)/3423), като точността на

предсказване за клас „Слаби” (1361/1807=0.75, т.е. 75%) е по-висока от точността на предсказване за клас „Силни” (1129/1616=0.70, т.е. 70%).

Отношението между правилно и неправилно класифицираните обекти за генерирания модел на класификация чрез алгоритъма J48 е:

Model Correct/Incorrect Ratio=(1361+1129)/(487+446)=2490/933=**2.67**

Получената стойност на съотношението Правилно/Неправилно класифицирани записи 2.67 показва, че броят на правилно класифицираните записи е 2.67 пъти по-голям от броя на неправилно класифицираните записи, т.е. на всеки 2.67 правилно класифицирани записа се получава една грешка.

За да се оцени работата на генерирания модел за класификация чрез алгоритъма J48, се сравняват получените отношения Правилно/Неправилно класифицирани записи за модела и при наивна класификация (стойностите за наивна класификация за получени в т.2.1):

$$\frac{Model\,\frac{Correct}{Incorrect}\,Ratio}{Naive\,\frac{Correct}{Incorrect}\,Ratio} = \frac{2.67}{1.01} = 2.64$$

Получената стойност 2.64 показва, че генерираният модел извършва класификацията 2.64 пъти по-добре от наивната класификация.

Полученото дърво на решенията

Полученото дърво на решенията е с размер 558, 467 листа и е изградено на 14 нива, т.е. то е голямо и сложно. На първо ниво в дървото е използвана входната променлива *Брой 2-ки* (*NumFailures*), като за всички стойности на този параметър (1-12), различни от 0, записите са класифицирани в листа с клас „Слаби”. Записите, за които стойността на параметъра *NumFailures*=0, се разпределят в двата класа на следващите нива на дървото. На второ ниво е избрана входната променлива *Бал на приемане* (*AdmissionScore*) с прагова стойност 28.79, като за *AdmissionScore*≤28.79 следва разделяне на трето ниво по параметъра *Пол* (*Gender*), а за *AdmissionScore*>28.79 следва разделяне на трето ниво отново по параметъра *Бал на приемане* (*AdmissionScor*e) с друга прагова стойност 33.88. На четвърто ниво в дървото са използвани параметрите *AdmissionYear*, *AdmissionScore* и U*nivSpecialtyName*. На пето ниво – *UnivSpecialtyName*, *CurrectSemester*, *AdmissionExam*, *AdmissionScore*, и т.н. Полученото дърво показва, че най-добро разделяне на записите от изследваната съвкупност от данни на двата класа - „Слаби” и „Силни”, се получава чрез входните променливи *NumFailures*, *AdmissionScore*, *Gender* и *UnivSpecialtyName* (променливите, които се появяват на по-високите нива на дървото) – това са същите

параметри, за които бе установено, че оказват по-съществено индивидуално въздействие върху класификацията (т.3.1.4.1, Фиг.3.18).

Представените до тук резултати са получените при стойности по подразбиране (default parameters) за параметрите на избрания алгоритъм. Един от параметрите, който може да бъде променян, е минималният брой записи M, които трябва да съдържа всяко от листата на дървото. Стойността по подразбиране е M=2. Самата природа на алгоритъма „Дърво на решенията" е такава, че в корена на дървото (който се намира най-отгоре, на първото ниво на дървото) се поставя цялата изследвана съвкупност от записи, а на всяка следваща стъпка от работата на алгоритъма записите се разделят на все по-малки и по-хомогенни групи чрез избор на подходяща входна променлива и подходящи нейни стойности. По този начин, в листата, които са по-отдалечени от корена на дървото, попадат все по-малко на брой записи от общата съвкупност и тези листа стават все по-малко представителни по отношение на първоначалната обща съвкупност от записи. Това означава, че много отдалечените от корена листа на дървото характеризират наличието на шум, а не на реални взаимовръзки в общата съвкупност от данни. Именно затова изборът на стойности на параметъра M на алгоритъма е важен по отношение на получените резултати от класификацията.

В Табл.3.12 са представени резултатите от класификацията чрез избрания алгоритъм „Дърво на решението" (WEKA J48), при различни стойности на параметъра M на алгоритъма.

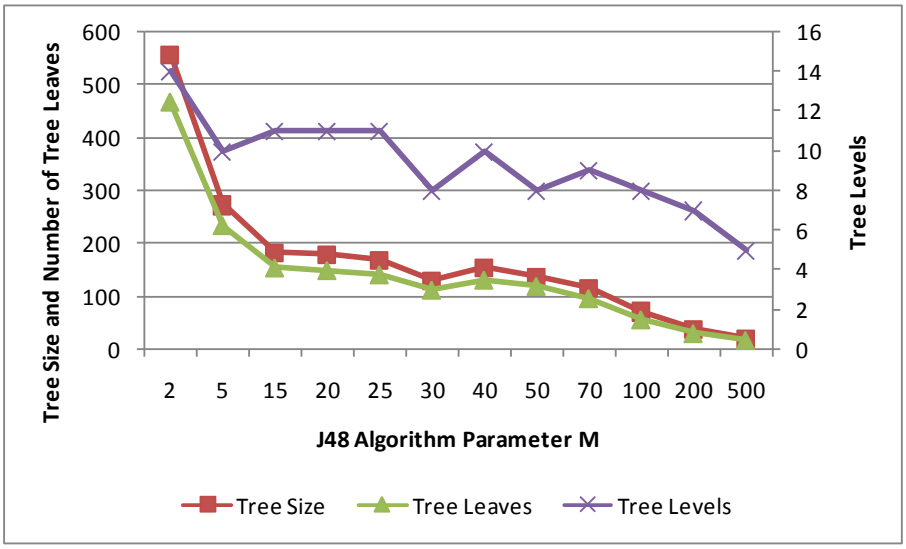
Таблица 3.12: Получени резултати от класификацията чрез алгоритъма *J48* в WEKA при различни стойности на параметъра M (минимален брой записи в листо на дървото)

J48 Classifier	Min No of objects in a leaf	TP Rate	FP Rate	% of Correctly classified instances	ROC Area	Tree Size	Number of Leaves	Tree Levels
Case 1	M=2	0.727	0.276	72.74% Max	0.784	558	467	14
Case 2	M=5	0.725	0.279	72.45%	0.789	274	234	10
Case 3	M=15	0.725	0.278	72.48%	0.790	183	154	11
Case 4	M=20	0.719	0.284	71.93%	0.790	179	149	11
Case 5	M=25	0.725	0.278	72.48%	0.792	168	141	11
Case 6	M=30	0.721	0.282	72.13%	0.790	131	112	8
Case 7	M=40	0.717	0.283	71.75%	0.790	155	131	10
Case 8	M=50	0.719	0.285	71.87%	0.791	138	119	8
Case 9	M=70	0.720	0.282	71.98%	0.790	115	96	9
Case 10	M=100	0.717	0.283	71.69%	0.786	72	56	8
Case 11	M=200	0.702	0.297	70.23%	0.760	39	30	7
Case 12	M=500	0.702	0.298	70.17% Min	0.762	21	17	5

От данните в Табл.3.12 се вижда, че с увеличаване на минималния брой записи в листата на дървото (M) се променя точността на предсказване на класа. За промяна на M в

интервала [2;500], промяната на точността на предсказване е в рамките на 2.57%. Много малка е промяната в стойностите на параметъра ROC Area – в рамките на 0.032, което означава, че класификаторът работи с приблизително еднаква точност, в сравнение с наивната класификация, при всички изследвани стойности на M.

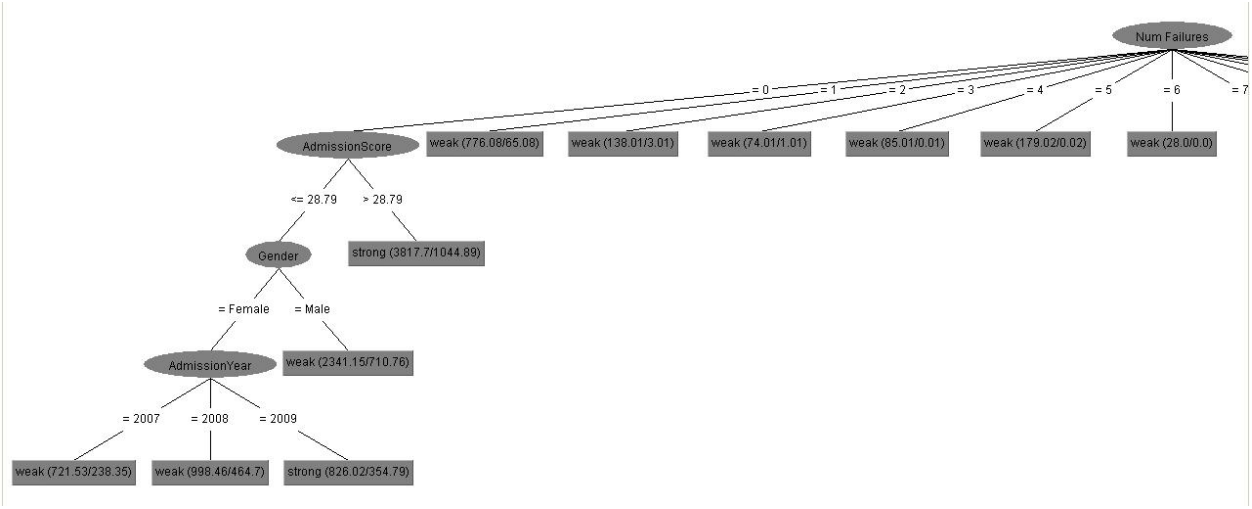
На Фиг.3.20 е представена зависимостта на големината на полученото дърво на решенията от параметъра M (минимален брой записи в едно листо на дървото) на избрания алгоритъм.



Фиг.3.20: Зависимост на големината на полученото *Дърво на решенията* от параметъра M (минимален брой записи в едно листо на дървото)

От данните в Табл.3.12 и на Фиг.3.20 се вижда, че с увеличаване на минималния брой записи в листата, намалява общият размер на дървото и броят на листата (от 12 случая има 1 изключение). Като цяло, има тенденция и за намаляване броя на нивата на дървото, но тя не е толкова ясно изразена (от 12 случая има 3 изключения).

При по-малки стойности на параметъра M на алгоритъма се получава дърво на решенията с голяма размерност. Такова дърво не може да бъде представено така, че лесно да бъде обхванато с поглед и анализирано. При стойност M=500 се получава дърво на решенията организирано на 5 нива, с общо 21 възела, 17 от които са листа на дървото. Именно това дърво (липсва малка част) е представено на Фиг.3.21.



Фиг.3.21: Полученото *Дърво на решенията* чрез алгоритъма *J48* в WEKA (при M=500)

Точността на предсказване при M=500 намалява до 70.17%, ROC Area=0.762, но това дърво е много по-лесно за интерпретиране. От него могат да бъдат извлечени следните правила за класифициране на студентите в двата класа „Слаби" и „Силни":

IF NumFailures=1,2,3,4,5,6,7,8,9,10,11,12, THEN Class=Weak

IF NumFailures=0 AND AdmissionScore>28.79, THEN Class=Strong

IF NumFailures=0 AND AdmissionScore≤28.79 AND Gender=Male, THEN Class=Weak

IF NumFailures=0 AND AdmissionScore≤28.79 AND Gender=Female AND AdmissionYear=2009, THEN Class=Strong

IF NumFailures=0 AND AdmissionScore≤28.79 AND Gender=Female AND AdmissionYear=2007,2008, THEN Class=Weak

Съгласно получените класификационни правила:

- В клас „Силни" се определят студентите, които:
  - Нямат слаби оценки (2-ки) на изпитите през първата учебна година в университета и са приети с Бал >28.79;
  - Нямат слаби оценки (2-ки) на изпитите през първата учебна година в университета, приети са с Бал ≤28.79, от женски пол са и са приети през 2009г.
- В клас „Слаби" се определят студенти, които:
  - Имат слаби оценки (2-ки), получени на изпитите в края на първата учебна година в университета;
  - Нямат слаби оценки (2-ки) на изпитите през първата учебна година в университета, приети са с Бал ≤28.78 и са от мъжки пол;
  - Нямат слаби оценки (2-ки) на изпитите през първата учебна година в университета, приети са с Бал ≤28.78, от женски пол са и са приети през 2007 или през 2008г.

По аналогичен начин може да бъде интерпретирано всяко дърво, което се получава като резултат от прилагането на избран алгоритъм „Дърво на решението" върху изследваните данни. Дървото на решения с по-голяма размерност и по-голям брой листа (и съответно по-малък брой записи в листата) характеризира по-детайлно съществуващите взаимовръзки между променливите в изследваната съвкупност от данни, но при него е по-голям риска от описване и на наличния шум в данните. Обратно, дърво на решения с по-малка размерност и по-малък брой листа (съответно по-голям брой записи в листата) описва по-общо съществуващите взаимовръзки между променливите в изследваната съвкупност от данни, но е по-лесно за интерпретация и е с по-малък риск относно характеризирането на наличния в данните шум.

Изводи за модела на класификация, получен чрез алгоритъма "Дърво на решенията":

- Най-добри параметри на модела за класификация чрез алгоритъма "Дърво на решенията" се получават при % разделяне на изследваната съвкупност от данни на 2 части (2/3 за генериране на модела и 1/3 за тестване.
- Постигната най-висока Точност на класификация =72.7%
- Най-висока стойност на параметъра Kappa Statistic=0.4524
- Най-висока стойност на параметъра ROC Area=0.784
- Model Correct/Incorrect Ratio=2.67
- $\frac{Model\frac{Correct}{Incorrect}Ratio}{Naive\frac{Correct}{Incorrect}Ratio} = 2.64$
- Входните променливи, които най-добре разделят студентите на два класа „Слаби" и „Силни", са *Брой 2-ки (NumFailures), Бал на приемане* (*AdmissionScore), Пол* (*Gender)* и *Направление*/*Поднаправление* (*UnivSpecialtyName).*
- Промяната на параметъра M (минимален брой записи в едно листо на дървото) на алгоритъма „Дърво на решенията" води до промяна и на моделът за класификация, като увеличаването на M води до намаляване на общия размер на дървото, броя на листата и броя на нивата в дървото.
- Точността на предсказване намалява с намаляване размерността на дървото.
- Дърветата с по-малка размерност се интерпретират по-лесно – от тях се извличат разбираеми за потребителя класификационни правила.

Алгоритъм за класификация „Невронна мрежа" - WEKA MultilayerPerceptron

Една от основните особености при прилагането на класификационен алгоритъм от типа „Невронна мрежа" е необходимостта от преобразуване на променливите, тъй като невронните мрежи работят с числови величини и са особено чувствителни при наличието на колинеарност между входните променливи.

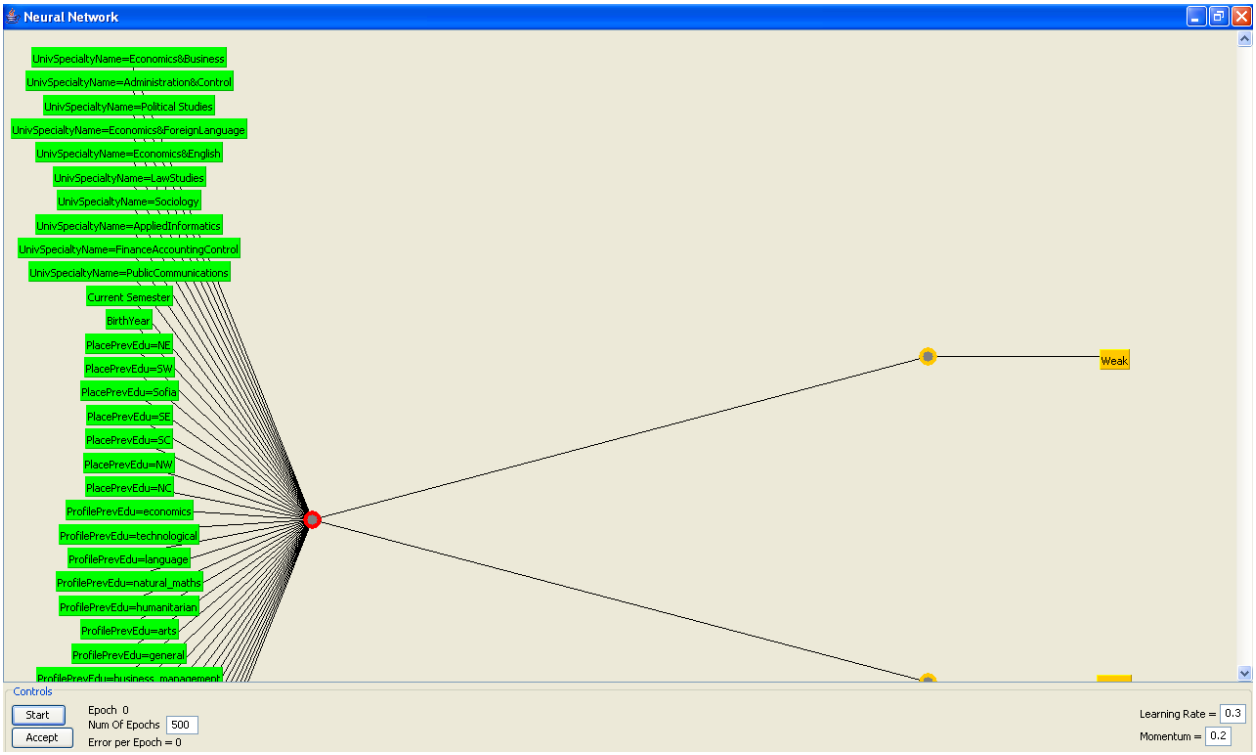
Поради споменатите особености на невронните мрежи, са извършени следните преобразувания на изследваната съвкупност от данни:

- Използвана е първоначалната съвкупност от данни, в която са запазени оргиналните цифрови стойности на входните променливи CurrentSemester, BirthYear, Age, AdmissionYear и NumFailures.
- Невронната мрежа е създадена за два случая. В първия случай (обозначен „12 променливи" в таблицата с резултатите) са премахнати променливите BirthYear (Година на раждане) и AdmissionYear (Година на приемане), тъй като стойностите на тези две променливи са използвани за дефинирането на новата променлива Age (Възраст), а този алгоритъм е чувствителен към наличието на силни взаимовръзки между входните променливи. Във втория случай (14 променливи) е използвана пълната съвкупност от данни с всички 14 променливи.
- Изходната променлива – *Среден успех през първата година* (ScoreFirtYearWithFailures), е трансформирана в двоична променлива, приемаща 2 дискретни стойности: Strong, съответстваща на клас „Силни", и Weak, съответстваща на клас „Слаби". За превръщането на оригиналните стойности на променливата е използвана прагова стойност 4.50 (т.е. студентите с успех <4.50 се считат „Слаби", т.е. за тях изходната променлива приема стойност Weak, а студентите с успех ≥4.50 се считат „Силни" и за тях изходната променлива приема стойност Strong).

В крайния си вид, използваната съвкупност от данни, върху която се прилага избраният алгоритъм – невронна мрежа, съдържа избраните общо 12/14 променливи (11/13 входни променливи и една изходна променлива – класа). Половината от променливите (6/8) са цифрови величини, а останалите 6 са променливи с дискретни стойности - едната от тях е изходната променлива – двоична величина, заемаща стойности Weak и Strong. Тъй като избраният алгоритъм, WEKA MultiLayer Perceptron, работи с цифрови величини, в самия алгоритъм е предвидено предварително преобразуване на дискретните променливи в цифрови - прилага се т.нар. NominalToBinaryFilter, т.е. променливите с номинални стойности (дискретни, неподредени стойности) се превръщат в двоични променливи. Това преобразуване се предлага като опция/параметър на самия алгоритъм и е включено в параметрите по подразбиране (default parameters). Стойностите на цифровите променливи се нормализират, така че да попадат в интервала [-1;+1], което е също опция/параметър на самия алгоритъм и е включено в параметрите по подразбиране.

Избраният алгоритъм „Невронна мрежа" (WEKA MultilayerPerceptron) е приложен със стойностите по подразбиране (default parameters), като е променена само стойността на параметъра HiddenLayers (H), чрез който се задава броят на невроните в скрития слой на невронната мрежа, и в случая е зададена стойност H=1. Алгоритъмът е реализиран при тестова опция „% разделяне" като отново изследваната съвкупност от данни е разделена

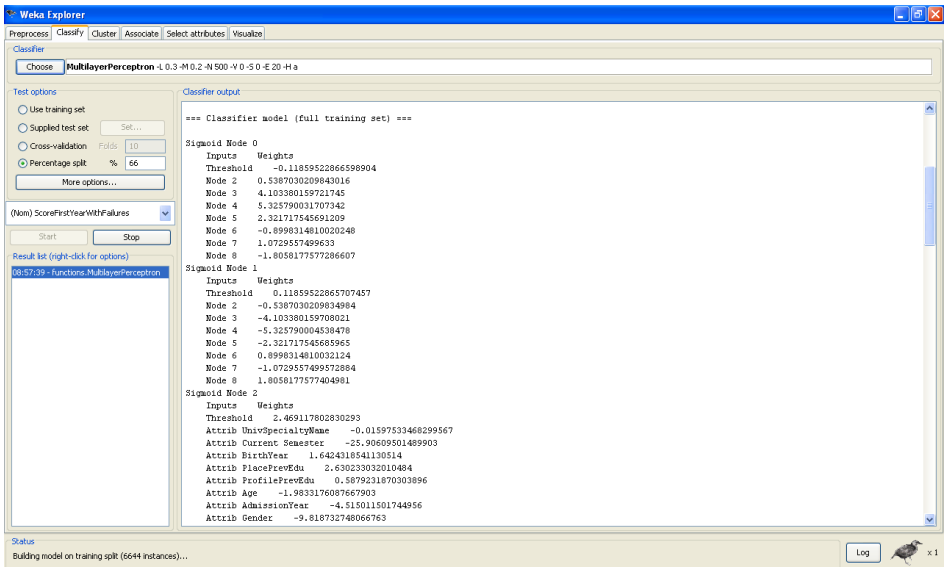
на две части – обучаваща извадка (66%, т.е.2/3 от общата съвкупност) и тестова извадка (34%, т.е. 1/3 от общата съвкупност).



Фиг.3.22: Получената *Невронна мрежа* чрез алгоритъма *Multilayer Perceptron* в WEKA (при H=1 - брой неврони в скрития слой)

На Фиг.3.22 е представена топологията на създадената от алгоритъма невронна мрежа. Тя съдържа входен слой с толкова на брой неврони, колкото са входните променливи, съдържащи се в изследваната съвкупност от данни. В случая, входните променливи с дискретни стойности са трансформирани в двоични променливи, приемащи стойности 0 и 1, затова на всяка стойност на дискретна променлива съответства неврон във входящия слой на мрежата (получава се невронна мрежа с много голям брой неврони във входния слой, затова е трудно да се визуализира по подходящ начин, затова на фигурата се вижда само част от получената невронна мрежа). Изходният слой на невронната мрежа съдържа два неврона, съответстващи на двата класа – „Слаби” и „Силни”. Невронната мрежа съдържа и един скрит слой, който в случая съдържа само един единствен неврон.

Моделът, генериран чрез алгоритъма „Невронна мрежа”, представлява съвкупност от изчислените тегла на невроните в създадената невронна мрежа (част от модела е представен на Фиг.3.23).



Фиг.3.23: Получената *Невронна мрежа* чрез алгоритъма *Multilayer Perceptron* в WEKA, в текстови формат (при H=1 - брой неврони в скрития слой)

Получените резултати от класификацията са представени в Табл.3.13.

Таблица 3.13: Резултати за оценка на получената *Невронна мрежа* чрез алгоритъма *Multilayer Perceptron* в WEKA (при H=1 - брой неврони в скрития слой)

MultiLayer Perceptron	**Test Option - % Split (66/34)**					
	12 Variables			**14 Variables**		
	Weak	**Strong**	**Weighted Average**	**Weak**	**Strong**	**Weighted Average**
Correctly Classified Instances	71.4286%			71.78%		
Incorrectly Classified Instances	28.5714%			28.22%		
Kappa Statistic	0.4257			0.434		
TP Rate	0.745	0.68	0.714	0.73	0.704	0.718
FP Rate	0.32	0.255	0.289	0.296	0.27	0.284
Precision	0.722	0.704	0.714	0.734	0.7	0.718
Recall	0.745	0.68	0.714	0.73	0.704	0.718
F-Measure	0.734	0.692	0.714	0.732	0.702	0.718
ROC Area	0.8	0.8	0.8	0.8	0.8	0.8

Получените резултати показват, че по-голяма точност на класификация се получава с невронната мрежа, когато са използвани всичките 14 променливи, т.е. цялата налична съвкупност от данни. Разликата в точността на класификация е само 0.35%, отразява се почти несъществено и върху останалите резултати на класификатора.

Матрицата с получените резултатите от класификацията (Confusion Matrix) с *MultiLayerPerceptron* алгоритъма, при % разделяне на съвкупността от данни на обучаваща (66%) и тестова (34%), при използвани всички 14 променливи, е представена в Табл.3.14.

Таблица 3.14: Матрица на класификация за получената *Невронна мрежа* чрез алгоритъма *Multilayer Perceptron* в WEKA (при H=1 - брой неврони в скрития слой)

MultiLayerPerceptron **%-Split Test Option**	Class Weak	Class Strong	Total
Is Weak	1319	488	1807
Is Strong	478	1138	1616
Total	1797	1626	3423

Точността на предсказване на алгоритъма се изчислява чрез броя на правилно и неправилно класифицираните записи (общо за алгоритъма и поотделно за всеки клас). В случая, общата точност на алгоритъма е 72% ((1319+1138)/3423), като точността на предсказване за клас „Слаби" (1319/1807=0.73, т.е. 73%) е по-висока от точността на предсказване за клас „Силни" (1138/1616=0.70, т.е. 70%).

Отношението между правилно и неправилно класифицираните обекти за генерирания модел на класификация чрез невронната мрежа е:

Model Correct/Incorrect Ratio=(1319+1138)/(478+488)=2457/966=**2.54**

Получената стойност на съотношението Правилно/Неправилно класифицирани записи 2.54 показва, че броят на правилно класифицираните записи е 2.54 пъти по-голям от броя на неправилно класифицираните записи, т.е. на всеки 2.54 правилно класифицирани записа се получава една грешка.

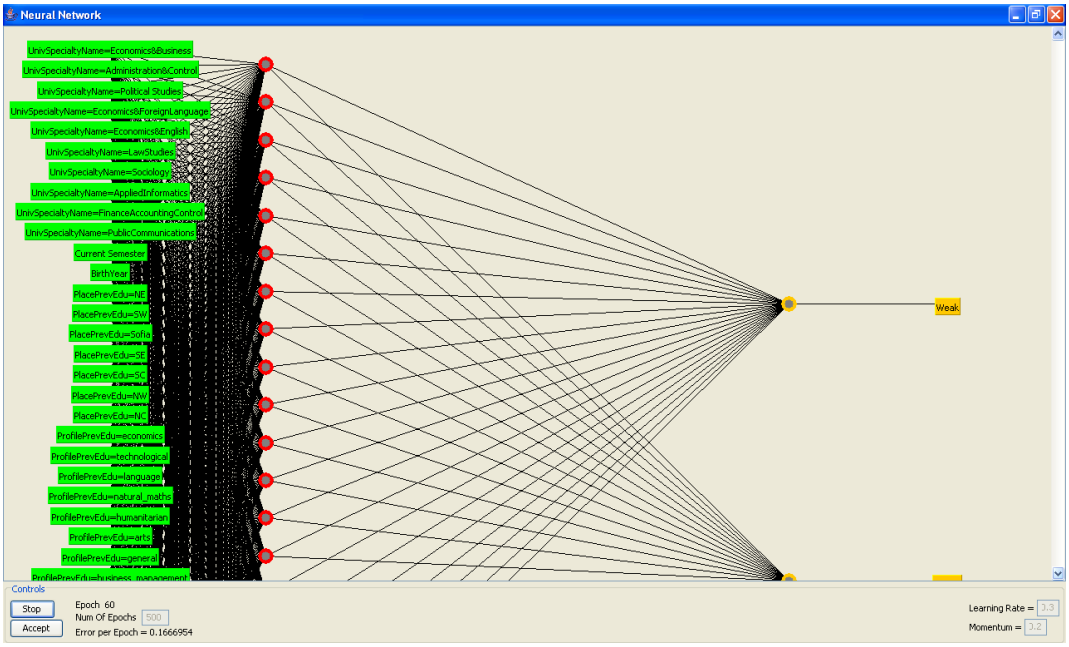
За да се оцени работата на генерирания модел за класификация чрез невронната мрежа, се сравняват получените отношения Правилно/Неправилно класифицирани записи за модела и при наивна класификация (стойностите за наивна класификация за получени в т.2.1):

$$\frac{Model\,\frac{Correct}{Incorrect}\,Ratio}{Naive\,\frac{Correct}{Incorrect}\,Ratio} = \frac{2.54}{1.01} = 2.52$$

Получената стойност 2.52 показва, че генерираният модел извършва класификацията 2.52 пъти по-добре от наивната класификация.

Параметърът Kappa Statistic показва степента на съгласуваност между предсказването и реалния клас. В случая получената стойност е 0.434 (максималната възможна стойност е 1.0, показваща пълно съгласуване). Това означава, че точността на предсказване на класификатора не е много висока. Стойността на площта под ROC кривата е ROC Area=0.8 (т.е. над 0.7), което означава, че генерираният модел за предсказване е добър.

Един от начините за подобряване точността на предсказване на невронните мрежи е да се експериментира с нейната топология. Затова, избраният алгоритъм, невронна мрежа MultilayerPerceptron, е приложен и при стойност по подразбиране за параметъра (H)HiddenLayers=a (a=(attributes+classes)/2). Автоматично избраната топология на невронната мрежа е доста по-сложна и е показана на Фиг.3.24 (частично, тъй като е много голяма и не може да бъде подходящо визуализирана). Разликата с предишния случай е в броя на невроните в скрития слой на невронната мрежа, който в случая е доста по-голям и се определя съгласно посочената по-горе формула H=a=(attributes+classes)/2.



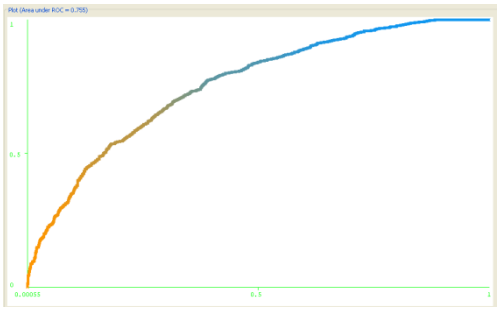
Фиг.3.24: Получената Невронна мрежа чрез алгоритъма Multilayer Perceptron в WEKA (при H(HiddenLayers)=a, a=(attributes+classes)/2)

Получените резултати от класификацията в този случай са представени в Табл.3.15.

Таблица 3.15: Резултати за оценка на получената *Невронна мрежа* чрез алгоритъма *Multilayer Perceptron* в WEKA (при H(HiddenLayers)=a, a=(attributes+classes)/2)

MultiLayer Perceptron	**Test Option - % Split (66/34)**		
	Weak	**Strong**	**Weighted Average**
Correctly Classified Instances	68.6532%		
Incorrectly Classified Instances	31.3468%		
Kappa Statistic	0.3732		
TP Rate	0.675	0.7	0.687
FP Rate	0.3	0.325	0.312
Precision	0.715	0.0.658	0.688
Recall	0.675	0.7	0.687
F-Measure	0.694	0.678	0.687
ROC Area	0.755	0.755	0.755

Резултатите показват, че при новата топология на невронната мрежа (повече неврони в скрития слой) точността на класификация намалява (68.65% са записите, за които правилно е предсказан съответният клас). По-малка е и стойността на параметъра Kappa Statistic – 0.3732, т.е. степента на съгласуваност между предсказването и реалния клас е доста под максималната възможна стойност е 1.0, показваща пълно съгласуване.



Фиг.3.25: ROC крива на получената *Невронна мрежа* чрез алгоритъма *Multilayer Perceptron* в WEKA (при H=a)

Получената ROC крива е представена на Фиг.3.25. Независимо, че точността на генерирания модел за класификация е по-ниска, площта под ROC кривата е с достатъчно голяма стойност (ROC Area=0.755), което означава, че и този модел е добър.

Върху точността на предсказване на невронната мрежа оказва влияние типа на променливите в изследваната съвкупност от данни. Затова е създаден нов файл с данни, в който категорийните променливи са трансформирани в цифрови, както е показано на Фиг.3.26.

UnivSpecialtyName Discrete-to-Numeric Transformation	
Administration&Control	1
AppliedInformatics	2
Economics&Business	3
Economics&English	4
Economics&ForeignLanguage	5
FinanceAccountingControl	6
LawStudies	7
PoliticalStudies	8
PublicCommunications	9
Sociology	10

PlacePrevEdu Discrete-to-Numeric Transformation	
Sofia	1
NW	2
NC	3
NE	4
SW	5
SC	6
SE	7
blanks	0

ProfilePrevEdu Discrete-to-Numeric Transformation	
arts	1
business_management	2
economics	3
general	4
humanitarian	5
language	6
natural_maths	7
sports	8
technological	9

AdmissionExam Discrete-to-Numeric Transformation	
Bulgarian	1
Maths	2
Geography	3
History	4
Economics	5
Blanks	0

Gender Discrete-to-Numeric Transformation	
Female	1
Male	0

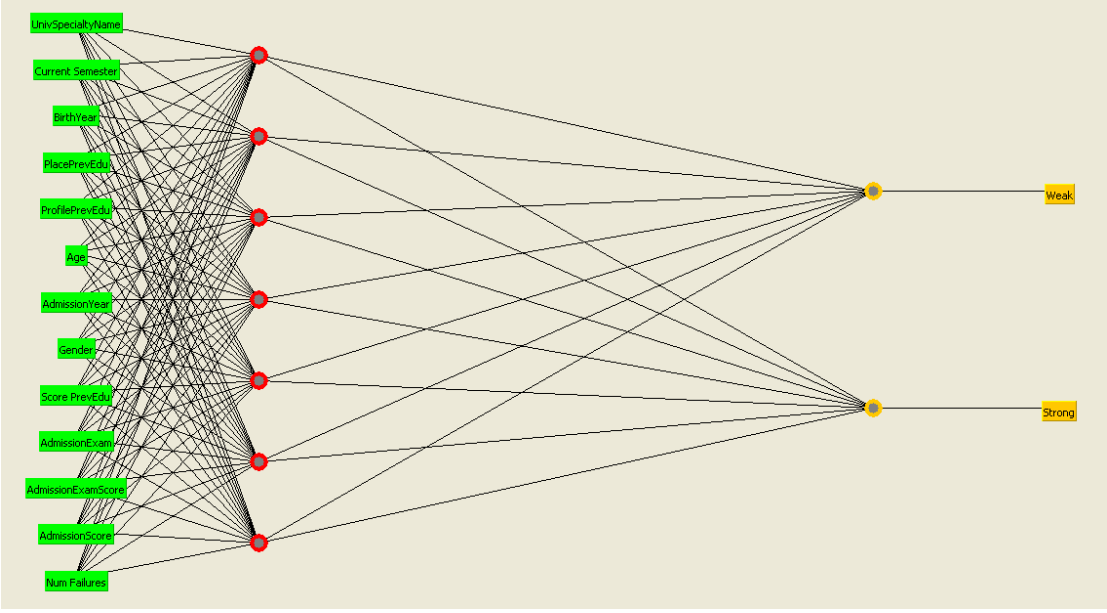
Фиг.3.26: Трансформиране на категорийните променливи в цифрови

Резултатите от класификацията са представени в Табл.3.16, при две различни стойности на параметъра H (HiddenLayers) – H=1 и H=a=(attribute+class)/2.

Таблица 3.16: Резултати за оценка на получената *Невронна мрежа* чрез алгоритъма *Multilayer Perceptron* в WEKA (при цифрови променливи, за H=1 и H=a)

MultiLayer Perceptron	**Test Option - % Split (66/34)**					
	H=1			**H=a**		
	Weak	**Strong**	**Weighted Average**	**Weak**	**Strong**	**Weighted Average**
Correctly Classified Instances	71.3701%			73.5904%		
Incorrectly Classified Instances	28.6299%			26.4096%		
Kappa Statistic	0.4326			0.4730		
TP Rate	0.625	0.813	0.714	0.704	0.772	0.736
FP Rate	0.188	0.375	0.276	0.228	0.296	0.26
Precision	0.789	0.66	0.728	0.775	0.7	0.74
Recall	0.625	0.813	0.714	0.704	0.772	0.736
F-Measure	0.698	0.728	0.712	0.738	0.734	0.736
ROC Area	0.784	0.784	0.784	0.823	0.823	0.823

Точността на класификация е по-висока за стойност H=a (73.59%), отколкото за стойност H=1 (71.31%). Топологията на мрежата при H=a е представена на Фиг.3.27 и съдържа входящ слой с 13 неврона, съответстващи на 13-те входни променливи, 1 скрит слой със 7 неврона (a=(13+1)/2=14/2=7) и изходящ слой с 2 неврона, съответстващи на двете стойности на изходната променлива – класа (Слаби и Силни).



Фиг.3.27: Получената *Невронна мрежа* чрез алгоритъма *Multilayer Perceptron* в WEKA (при цифрови променливи, за H=a)

В случая, когато категорийните променливи са предварително трансформирани в цифрови, се получава по-висока точност на класификация, отколкото ако трансформирането на категорийните променливи в двоични цифрови величини се извършва автоматично от невронната мрежа.

В този случай параметърът Kappa Statistic приема стойност 0.473, а ROC Area=0.823, което показва, че генерираният модел за класификация е доста добър.

Матрицата с получените резултатите от класификацията (Confusion Matrix) е представена в Табл.3.17.

Таблица 3.17: Матрица на класификация за получената *Невронна мрежа* чрез алгоритъма *Multilayer Perceptron* в WEKA (при цифрови променливи, за H=a)

MultiLayerPerceptron **%-Split Test Option**	Class Weak	Class Strong	Total
Is Weak	1272	535	1807
Is Strong	369	1247	1616
Total	1641	1782	3423

Точността на предсказване на алгоритъма е 74% ((1272+1247)/3423), като точността на предсказване за клас „Слаби" (1272/1807=0.704, т.е. 70%) е по-ниска от точността на предсказване за клас „Силни" (1247/1616=0.772, т.е. 77%). Това е единствения модел за предсказване, генериран до момента, при който точността на предсказване на клас „Силни" е по-висока от точността на предсказване на клас „Слаби". Това може да е и едно от важните основания именно този модел да бъде избран за последващо прилагане върху нови данни (deployment) с цел предсказване на класа.

Отношението между правилно и неправилно класифицираните обекти за генерирания модел на класификация чрез тази невронна мрежа е:

Model Correct/Incorrect Ratio=(1272+1247)/(369+535)=2519/904=**2.79**

Получената стойност на съотношението Правилно/Неправилно класифицирани записи 2.79 показва, че броят на правилно класифицираните записи е 2.79 пъти по-голям от броя на неправилно класифицираните записи, т.е. на всеки 2.79 правилно класифицирани записа се получава една грешка.

За да се оцени работата на генерирания модел за класификация чрез невронната мрежа, се сравняват получените отношения Правилно/Неправилно класифицирани записи за модела и при наивна класификация (стойностите за наивна класификация за получени в т.2.1):

$$\frac{Model\,\frac{Correct}{Incorrect}\,Ratio}{Naive\,\frac{Correct}{Incorrect}\,Ratio} = \frac{2.79}{1.01} = \mathbf{2.76}$$

Получената стойност 2.76 показва, че генерираният модел извършва класификацията 2.76 пъти по-добре от наивната класификация (т.е. при липса на модел за предсказване).

Изводи за модела на класификация, получен чрез алгоритъма "Невронна мрежа":

- При прилагането на алгоритъма „Невронна мрежа", по-добри модели за класификация се получават, когато се извърши подходящо предварително преобразуване на променливите от категориен тип в цифрови величини. Този извод напълно съответства на очакванията, тъй като самата природа на този алгоритъм изисква работа с цифрови величини, затова в този вид алгоритми е предвидена процедура за предварително преобразуване на качествените променливи.
- Премахването на две от променливите, които са взаимно свързани, от наличната съвкупност от данни, не води до получаването на модел за класификация с по-добри параметри. Обикновено очакванията в това отношение се различават от получените резултати, тъй като премахването на взаимосвързани променливи обикновено води до повишане на точността на предсказване.
- Върху точността на предсказване на моделите, генерирани чрез алгоритъм „Невронна мрежа", оказва влияние топологията на мрежата. В конкретните изследвания, невронната мрежа с по-голям брой неврони в скрития слой работи с по-голяма точност от невронната мрежа с един неврон в скрития слой.
- Най-добрият модел за класификация чрез алгоритъма „Невронна мрежа" се получава за пълната съвкупност от данни (с всички 14 променливи), при предварително подходящо преобразуване на всички дискретни променливи в цифрови величини, и при наличието на 7 неврона в скрития слой (стойност на параметъра H=(attribute+class)/2=7).
- Постигната най-висока Точност на класификация =73.59%
- Най-висока стойност на параметъра Kappa Statistic=0.473
- Най-висока стойност на параметъра ROC Area=0.823
- Model Correct/Incorrect Ratio=2.79
- $\frac{Model\,\frac{Correct}{Incorrect}\,Ratio}{Naive\,\frac{Correct}{Incorrect}\,Ratio} = 2.76$
- Невронната мрежа е единственият модел за предсказване, генериран в рамките на изследванията, при който точността на предсказване на клас „Силни" е по-висока от точността на предсказване на клас „Слаби".

Алгоритъм за класификация „К-най-близък съсед" – WEKA IBk

Алгоритъмът К-най-близък съсед (WEKA IBk) е приложен за различни стойности на параметъра К – предварително зададеният брой обекти, на базата на които се определя класът на нов обект. Това е един от основните параметри на алгоритъма, чрез които може да се променя неговата точност на класификация. Изследването е проведено за цялата изследвана съвкупност от данни (с дискретните категорийни стойности на петте променливи CurrentSemester, BirthYear, Age, AdmissionYear и NumFailures) при тестова опция - % разделяне (66% обучаващи данни и 34% тестови данни). Получените резултатите са представени в Табл.3.18.

Таблица 3.18: Резултати за оценка на получения модел за класификация чрез алгоритъма *IBk* в WEKA за различни стойности на параметъра *K* (брой най-близки съседи)

kNN Algorithm	**Test Option - % Split (66/34)**									
	K=1	**K=10**	**K=20**	**K=30**	**K=40**	**K=50**	**K=60**	**K=70**	**K=80**	**K=90**
Correctly Classified Instances, %	62.29	68.30	69.62	70.44	70.35	**70.47**	70.44	70.17	70.09	69.73
Incorrectly Classified Instances, %	37.71	31.70	30.38	29.56	29.65	**29.53**	29.57	29.83	29.91	30.27
Kappa Statistic	0.2448	0.3598	0.3896	0.4069	0.4057	**0.4085**	**0.4086**	0.4034	0.402	0.3952
ROC Area	0.623	0.752	0.771	0.779	0.781	**0.784**	**0.785**	**0.785**	**0.786**	0.782

kNN Algorithm	**Test Option - % Split (66/34)**									
	K=100	**K=150**	**K=200**	**K=250**	**K=300**	**K=350**	**K=500**	**K=750**	**K=1000**	**K=1200**
Correctly Classified Instances, %	69.59	69.21	69.59	69.73	69.91	69.27	69.73	68.77	68.92	67.86
Incorrectly Classified Instances, %	30.41	30.79	30.41	30.27	30.09	30.73	30.27	31.23	31.08	32.14
Kappa Statistic	0.3924	0.3857	0.3936	0.3964	0.4002	0.3878	0.3987	0.381	0.3855	0.3663
ROC Area	0.782	0.777	0.777	0.778	0.778	0.778	0.769	0.764	0.77	0.765

Получените резултати показват, че алгоритъмът работи с най-голяма точност на класификация при К=50. В този случай висока стойност имат и параметрите Kappa Statistic=0.4085 (неговата най-висока стойност е 0.4086 за К=60) и ROC Area=0.784 (неговите най-високи стойности са 0.786 за К=80 и 0.785 за К=60,70). Именно за случая К=50 са и следващите представени резултати. В Табл.3.19 са дадени резултатите от прилагането на алгоритъма за К=50

Матрицата с получените резултатите от класификацията (Confusion Matrix) е представена в Табл.3.20.

Таблица 3.19: Резултати за оценка на получения модел за класификация чрез алгоритъма *IBk* в WEKA (за *K=50)*

kNN Algorithm (IBk) ***k=50***	**Test Option - % Split (66/34)**		
	Weak	**Strong**	**Weighted Average**
Correctly Classified Instances	70.4645%		
Incorrectly Classified Instances	29.5355%		
Kappa Statistic	0.4085		
TP Rate	0.706	0.704	0.705
FP Rate	0.296	0.294	0.295
Precision	0.727	0.681	0.705
Recall	0.706	0.704	0.705
F-Measure	0.716	0.692	0.705
ROC Area	0.784	0.784	0.784

Таблица 3.20: Матрица на класификация за получения модел за класификация чрез алгоритъма IBk в WEKA (за K=50)

MultiLayerPerceptron **%-Split Test Option**	Class Weak	Class Strong	Total
Is Weak	1275	532	1807
Is Strong	479	1137	1616
Total	1754	1669	3423

Точността на предсказване на алгоритъма е 71% ((1275+1137)/3423), като точността на предсказване за клас „Слаби” (1275/1807=0.71, т.е. 71%) е приблизително равна на точността на предсказване за клас „Силни” (1137/1616=0.70, т.е. 70%).

Отношението между правилно и неправилно класифицираните обекти за генерирания модел на класификация чрез този алгоритъм е:

Model Correct/Incorrect Ratio=(1275+1137)/(479+532)=2412/1011=**2.39**

Получената стойност на съотношението Правилно/Неправилно класифицирани записи 2.39 показва, че броят на правилно класифицираните записи е 2.39 пъти по-голям от броя на неправилно класифицираните записи, т.е. на всеки 2.39 правилно класифицирани записи се получава една грешка.

За да се оцени работата на генерирания модел за класификация чрез невронната мрежа, се сравняват получените отношения Правилно/Неправилно класифицирани записи за модела и при наивна класификация (стойностите за наивна класификация за получени в т.2.1):

$$\frac{Model\,\frac{Correct}{Incorrect}\,Ratio}{Naive\,\frac{Correct}{Incorrect}\,Ratio} = \frac{2.39}{1.01} = 2.37$$

Получената стойност 2.37 показва, че генерираният модел извършва класификацията 2.37 пъти по-добре от наивната класификация (т.е. при липса на модел за предсказване).

Изводи за модела на класификация, получен чрез алгоритъма "К-Най-близък съсед":

- Параметърът К – предварително зададеният брой обекти, на базата на които се определя класът на нов обект, е един от основните параметри на алгоритъма, чрез които може да се променя неговата точност на класификация. В проведеното изследване, модел за класификация с най-висока точност на предсказване е получен при К=50.
- Точност на предсказване = 70.47%
- Kappa Statistic = 0.4085
- ROC Area = 0.784
- Model Correct/Incorrect Ratio=2.39
- $\frac{Model\,\frac{Correct}{Incorrect}\,Ratio}{Naive\,\frac{Correct}{Incorrect}\,Ratio} = 2.37$

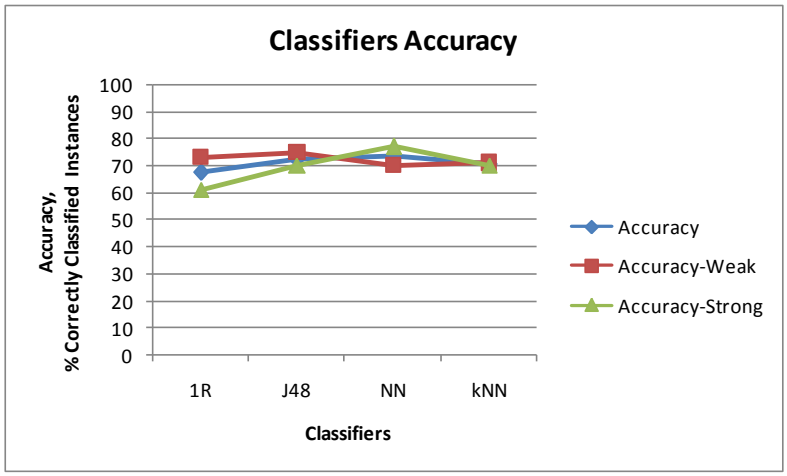
Сравнение на резултатите, получени за различните алгоритми на класификация

Таблица 3.21: Резултати, получени в WEKA с Data Mining алгоритми *OneR, J48, MultiLayer Perceptron* и *IBk* при предсказвана променлива с 2 стойности

	1R Classifier	**Decision Tree**	**Neural Network**	**kNN Algorithm**
Correctly classified instances	67.46%	72.74%	73.59%	70.47%
Correctly classified instances for Class Weak	73%	75%	70%	71%
Correctly Classified Instances for Class Strong	61%	70%	77%	70%
Kappa Statistic	0.3439	0.4524	0.4730	0.4085
ROC Area	0.671	0.784	0.823	0.784
Model Correct/Incorrect Ratio	2.07	2.67	2.79	2.39
Model/Naïve Correct/Incorrect Ratio	2.05	2.64	2.76	2.37

В Табл.3.21 са представени резултатите, получени при прилагането на избраните четири различни класификационни алгоритми върху изследваната съвкупност от данни – генератор на правила 1R (WEKA OneR), дърво на решенията (WEKA J48), невронна мрежа (WEKA MultilayerPerceptron) и К-най-близък съсед (kNN – WEKA IBk).

На Фиг.3.28 е представено сравнение на избраните класификационни алгоритми – генератор на правила (1R), дърво на решението (J48), невронна мрежа (NN) и К-най-близък съсед (kNN), по отношение на точността на предсказване.



Фиг.3.28: Сравнение на моделите за класификация, получени чрез алгоритмите *OneR, J48, Multilayer Perceptron* и *IBk* в WEKA, по отношение на *точността на предсказване*

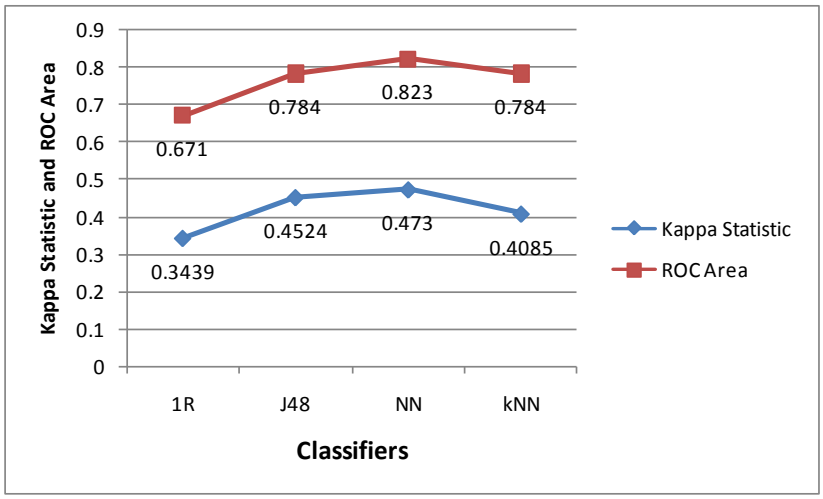
Сред избраните класификационни алгоритми, най-висока точност на предсказване е получена при модела, генериран чрез алгоритъма „*Невронна мрежа*” – 73.59%. Класификационният модел, създаден с невронна мрежа, е и единственият алгоритъм, при който точността на предсказване на клас „Силни” е по-висока от точността на предсказване на клас „Слаби”. Постигнатата точност за предсказване на клас „Силни” при невронната мрежа е 77%, което е и най-голямата постигната точност за предсказване на някой от двата класа. Затова моделът, получен с невронна мрежа, е най-подходящ за предсказване на „силните” студенти на базата на техни характеристики от средното образование и кандидат-студентската кампания. Основни недостатъци на метода са голямата сложност и трудната интерпретация на получения класификационен модел, необходимостта от по-продължително време за генериране на модела, силната чувствителност на невронните мрежи към липсващи стойности в данните, необходимостта от по-сложна предварителна подготовка на данните (трансформиране на дискретните променливи в цифрови, нормализация и др.)

Висока е точността на предсказване и на модела, получен чрез алгоритъма „*Дърво на решенията*” – 72.74%. При този модел точността на предсказване на клас „Слаби” е по-висока от точността на предсказване на клас „Силни”. Основно предимство на този

алгоритъм е неговата по-лесна интерпретация и възможността за извличане на разбираеми класификационни правила, голямата бързина на работа на алгоритъма, ниската чувствителност по отношение на липсващи стойности и добрите възможности за работа както с цифрови, така и с категорийни променливи. Следователно, полученото дърво на решенията е подходящ модел за предсказване, когато трябва да се изведат ясни и разбираеми правила, обясняващи извършваната класификация, както и при наличието на голям брой категорийни променливи и липсващи стойности в данните.

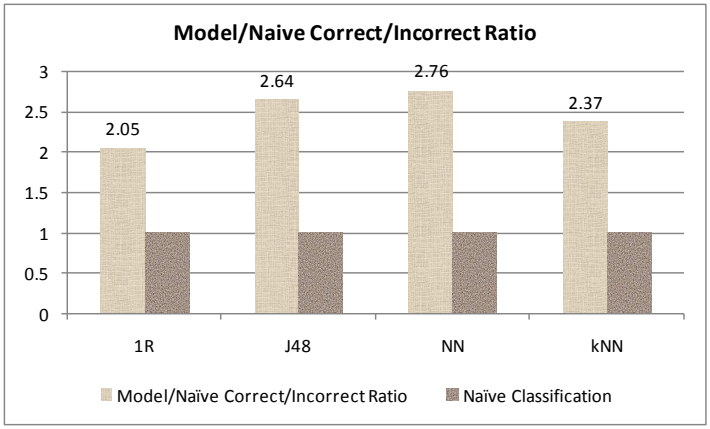
Моделът за предсказване, създаден чрез алгоритъма *К-най-близък съсед*, извършва класификацията с точност 70.5% и показва приблизително еднакви точности на предсказване на двата класа „Слаби" и „Силни".

Най-ниска точност на предсказване е получена за модела, създаден чрез генератора на правила *1R*. И при този модел, както при дървото на решенията, точността на предсказване на клас „Слаби" е по-висока от точността на предсказване на клас „Силни".



Фиг.3.29: Сравнение на моделите за класификация, получени чрез алгоритмите *OneR, J48, Multilayer Perceptron* и *IBk* в WEKA, по отношение на параметрите Kappa Statistic и *ROC Area*

На Фиг.3.29 са представени резултатите, получени за четирите класификационни алгоритми, относно параметрите за оценка на моделите за класификация – *Kappa Statistic* и *ROC Area*. Параметърът *Kappa Statistic* показва степента на съгласуваност между предсказването и реалния клас, и най-висока стойност е получена за модела от невронната мрежа - 0.473. Площта под *ROC* кривата е друг важен параметър за оценка на генерираните модели за класификация. Стойностите на този параметър за три от избраните алгоритми – *Невронна мрежа*, *Дърво на решенията* и *К-най-близък съсед*, са над 0.7, което означава, че и трите получени модела за класификация са добри. Най-висока стойност на параметъра *ROC Area* отново е получена за модела от невронната мрежа – 0.823.



Фиг.3.30: Сравнение на моделите за класификация, получени чрез алгоритмите OneR, J48, Multilayer Perceptron и IBk в WEKA, по отношение на коефициентите на подобрение на предсказването спрямо случайната класификация

На Фиг.3.30 са представени изчислените коефициенти, които показват колко се подобрява предсказването при използване на генерираните модели за класификация, в сравнение с наивната класификация (т.е. при класификацията без използване на предварително създаден модел на базата на наличните данни). Най-висок коефициент е получен за модела от невронната мрежа – 2.76, следван от дървото на решенията – 2.64. Трябва да се отбележи, че и при четирите използвани алгоритъма получените коефициенти са >2, т.е. и с четирите модела предсказването ще се извършва поне 2 пъти по-точно в сравнение с наивната класификация.

Изводи:

- Сред избраните класификационни алгоритми, *най-висока точност на предсказване* е получена при модела, генериран чрез алгоритъма „*Невронна мрежа*" *– 73.59%*.
- *Невронната мрежа* е единствения класификационен модел, при който *точността на предсказване на клас „Силни" (77%) е по-висока от точността на предсказване на клас „Слаби"*, и това е най-високата постигната точност за предсказване на някой от двата класа.
- Полученото *Дърво на решението* също е *модел с висока точност на класификация – 72.74%*, като този модел е *по-разбираем* и *лесен за интерпретация* за потребителите.
- Моделът, създаден чрез алгоритъма *К-най-близък съсед*, извършва класификацията с *точност 70.5%* и показва *приблизително еднакви точности на предсказване на двата класа „Слаби" и „Силни"*.
- *Най-ниска точност на предсказване* е получена за модела, създаден чрез *генератора на правила 1R.*

- *Най-висока стойност* на параметъра *Kappa Statistic=0.473* е получена за модела „*Невронна мрежа*”.
- *Най-висока стойност* на параметъра *ROC Area=0.823* е получена за модела „*Невронна мрежа*”.
- Стойностите на параметъра *ROC Area* за три от избраните алгоритми – *Невронна мрежа, Дърво на решенията* и *К-най-близък съсед*, са *>0.7*, което означава, че и трите получени модела за класификация са добри.
- *Най-висок коефициент за подобряване на предсказването спрямо наивна класификация* е получен за *Невронната мрежа – 2.76*, следва *Дърво на решенията – 2.64.*
- За *четирите използвани алгоритъма* получените *коефициенти са* >2, т.е. и с четирите модела предсказването ще се извършва поне 2 пъти по-точно в сравнение с наивната класификация.

3.1.4. Сравнение на получените Data Mining модели за класификация за двата варианта на предсказваната променлива

Въз основа на резултатите, получени при генерирането на Data Mining модели за класификация, върху съвкупността от данни за студентите, за двата варианта на предсказваната променлива - с пет и с две стойности, могат да бъдат направени следните изводи.

Изводи:

- Генерираните модели за класификация чрез четирите избрани алгоритми дават *по-голяма точност на предсказване*, когато *изходната променлива съдържа по-малък брой различни стойности*.
- За *двоична предсказвана променлива* (с две стойности - „Слаби” и „Силни”):
  - *Всички приложени алгоритми, с изключение на 1R класификатора*, работят с обща *точност на предсказване >70%* (weighted average accuracy), като се справят сравнително добре и с двата класа.
  - И за *трите алгоритъма (Невронна мрежа, Дърво на решението и К-Най-близък съсед*) се получават високи стойности на параметрите *Kappa Statistic (>0.40) и ROC Area (>0.78)*, което означава, че генерираните модели са подходящи за извършване на класификацията с висока точност.
  - *Невронната мрежа* е генерираният *модел за класификация с най-добри параметри*.
  - *Невронната мрежа* се справя най-добре при предсказването на *клас „Силни”*, което в конкретния бизнес контекст е особено важно.

- Чрез прилагането на създадената *невронна мрежа* в реални условия, върху нови данни, *точността на предсказване може да се подобри 2.76 пъти в сравнение с наивната класификация*.

- За *предсказвана променлива с пет различни стоности* – „Слаб", „Среден", „Добър", „Много Добър" и „Отличен":
  - Получена е *по-ниска точност на предсказване* и при *четирите използвани алгоритъма (<70%)*
  - Наблюдават се *съществени разлики в точността на предсказване на различните класове*
  - *Най-ниски точности на предсказване* се получават по отношение на *клас „Отличен"*.
  - *Най-добър модел* се получава чрез алгоритъма „*Дърво на решенията*", който дава *2.85 пъти подобрение в сравнение с наивната класификация*.

3.2. Генериране и оценка на Data Mining класификационни модели (обучени класификатори) за предсказване вида на открити морски цели

Data Mining задачата за класификация и в случая на радарни данни се осъществява на базата на CRISP-DM (Cross-Industry Standard Process for Data Mining) модела – стандартен подход за реализация на Data Mining проекти, и включва:

- Избор и извличане на данни за Data Mining анализи – описано в т.2.2.2 на дисертацията;
- Изследване и предварителна подготовка на данните - описано в т.2.2.2 на дисертацията;
- Избор на Data Mining методи и генериране на модели за предсказване вида на откритите радарни морски цели чрез софтуерен пакет WEKA;
- Сравнение и оценка на получените модели за предсказване.

Data Mining анализите са осъществени върху *съвкупността от данни*, съдържаща *80 записа* на радарни морски цели, описани чрез *16 параметъра* (описани в т.2.2.2 и Табл.2.4). *Предсказваната целева променлива* е *вида на откритата радарна цел - променлива от категориен тип* с *три различни стойности - MISL Boat, Big Boat* и *Average Boat*. Броят на записите в трите класа е различен. Най-много данни са налични за клас MISL Boat (62 записа), а доста по-ограничен е броят са записите за клас Big Boat (11) и клас Average Boat (6). За целевата променлива има само една липсваща стойност (Target Variable - 1% Missing Values - виж Табл.2.4), затова общият брой на записите в изследваната съвкупност от данни е 79.

За решаването на Data mining задачата за класификация са използвани различни алгоритми: *Генератори на правила* (Rule Learners) – WEKA *OneR* и *JRip*, *Дърво на решението* (Decision Tree) – WEKA *J48* (базиран на алгоритъма C4.5) и *RandomForest*, *Невронна мрежа* (Neural Network) – WEKA *Multilayer Perceptron*, *К-най-близък съсед* (K-Nearest Neighbour) – WEKA *IBk*, два Байесови класификатора (*NaiveBayes* и *BayesNet*) и Логистична регресия - *SimpleLogistic*. Тези алгоритми са избрани, тъй като всеки от тях работи на различен принцип, т.е. по различен начин генерира модела за предсказване на целевата променлива. Затова, прилагането на тези алгоритми върху едни и същи данни може да доведе до получаването на различни резултати – в някои случаи определени алгоритми дават по-добри резултати, а в други случаи – други алгоритми се справят по-добре, по отношение предсказването на класа (това действително се вижда от получените резултати в представеното изследване).

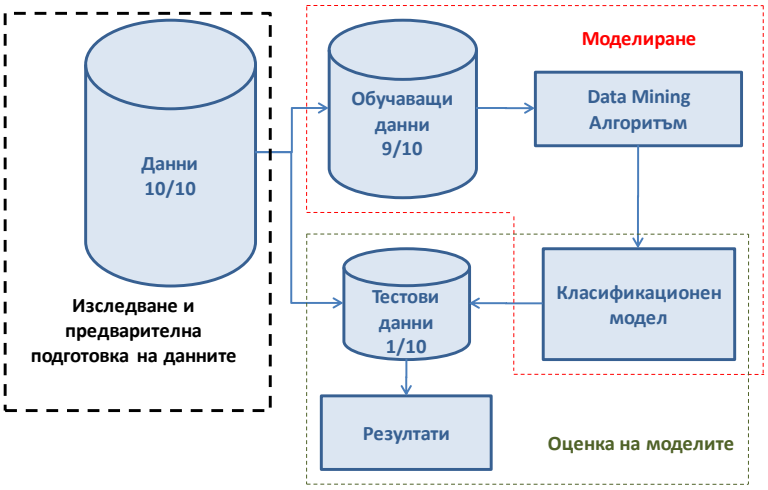
Кратко описание на използваните Data Mining алгоритми в софтуерния пакет WEKA е представено в Табл.3.22.

Таблица 3.22: Описание на използваните Data Mining алгоритми в WEKA

Име на алгоритъма	Описание	Параметри
Генератор на правила – 1R WEKA OneR	Алгоритъмът извършва предсказването на класа като използва само един атрибут – този с най-малка грешка при предсказване, като генерира дърво на решението на едно ниво.	Стойности по подразбиране (WEKA default parameters)
Генератор на правила – WEKA JRip	Класификаторът JRip прилага алгоритъма RIPPER (Repeated Incremental Pruning to Produce Error Reduction).	Стойности по подразбиране (WEKA default parameters)
Дърво на решенията – C4.5 WEKA J48	Класификаторът J48 се базира на алгоритъм C4.5 – изгражда дърво на решенията от налична съвкупност от данни като използва критерий за оценка чистотата на производните възли - Information Gain (взеки възел може да се разклонява на повече от 2 подвъзела).	Стойности по подразбиране (WEKA default parameters)
Дърво на решенията WEKA RandomForest	RandomForest (ensemble classifier) - класификатор, който се състои от множество дървета и предсказва класа, който се явява мода (най-често срещана стойност) за индивидуалните дървета.	Стойности по подразбиране (WEKA default parameters)
Байесов класификатор - Naive Bayes WEKA Naive Bayes	Статистически класификатор, който предсказва принадлежността към даден клас на базата на вероятности (вероятност, че даден обект принадлежи към определен клас). Алгоритъмът Naive Bayes приема, че ефектът на всяка от променливите върху класификацията не зависи от стойностите на останалите променливи.	Стойности по подразбиране (WEKA default parameters)
Байесов класификатор - Bayesian networks	Статистически класификатор Байесовите мрежи са графични модели, чрез които могат да бъдат изследвани вероятностни разпределения при съвместно въздействие на две и повече променливи.	Стойности по подразбиране (WEKA default parameters)

К-най-близък съсед (k-NN) WEKA IBk	Метод за класифициране на обекти въз основа класа на принадлежност на най-близко разположените обекти в обектното пространство. Чрез алгоритъма k-NN класът на нов обект се определя чрез мнозинството от гласовете на намерените К на брой най-близки съседни обекти (всеки обект гласува за своя клас).	Алгоритъмът е приложен за различни стойности на параметъра К: k=5,6,7,8,9,10
Невронна мрежа WEKA MultilayerPerceptron	Невронната мрежа представлява силно свързана система, съставена от множество елементарни изчислителни елементи (неврони) свързани чрез претеглени връзки и организирани в слоеве. Многослойният персептрон е невронна мрежа с един или повече скрити слоеве, с двупосочно извършване на изчисленията (feedback network).	Стойности по подразбиране (WEKA default parameters)
Логистична регресия Logistic Regression WEKA SimpleLogistics	Статистически метод за предсказване на двоична променлива. Логистичната регресия се използва за предсказване на вероятността за появата на дадено събитие чрез подходящо преобразуване на данните в голостична крива.	Стойности по подразбиране (WEKA default parameters)

Избраните data mining алгоритми за класификация са приложени върху наличните данни като са използвани стойностите по подразбиране на техните параметри за настройка (WEKA default parameters), освен в случаите, когато това е допълнително указано. Моделите за предсказване вида на откритата радарна морска цел са получени като е използвана тестова схема „stratified 10-fold cross validation" (крос-валидиране), представена на Фиг.3.31.



Фиг.3.31: Тестова схема за класификация крос-валидиране („10-fold Cross Validation")

При "stratified 10-fold cross validation test option" общата съвкупност от данни се разделя на 10 части и алгоритъмът се прилага 10 пъти, като всеки път 9 от 10-те части се използват като обучаващи данни (training data), а останалата 1 част се използва за оценка

на модела. Това е стандартен подход, който се прилага за определяне грешката на предсказване на класификационните алгоритми при наличието на една постоянна съвкупност от данни. Стратификацията на данните означава, че съвкупността от данни ще бъде разделена произволно на 10 части, но във всяка част всеки клас ще е представен приблизително в същите пропорции, както в цялата съвкупност от данни. Крайната грешка на алгоритъма се определя като средно аритметично от грешките, получени при 10-те изпълнения на алгоритъма.

Всеки от моделите за класификация, получени чрез приложените различни Data Mining алгоритми, е оценен по следните критерии (по-подробно описани в т.1.2.4 на дисертацията):

- *Точност/Грешка на предсказване*
- *Матрица на класификация* и *базираните на нея параметри*
- *Kappa Statistic*
- *ROC Area*
- *Model Correct/Incorrect Ratio*
- *Коефициент на подобрение на предсказването* спрямо случайната класификация $\frac{Model\frac{Correct}{Incorrect}Ratio}{Naive\frac{Correct}{Incorrect}Ratio}$

За изчисляване на коефициента на подобрение на предсказването с получените модели, спрямо случайната (наивна) класификация, се използва т.нар. *Матрица на наивна класификация*, т.е. класификация на случаен принцип, без използване на модел за предсказване, изготвен на базата на обучаваща извадка. В случая *Матрицата на наивна класификация* има следния вид (Табл.3.23):

Таблица 3.23: Матрица на случайна класификация за данните от сферата на FS сигнална обработка

Naïve Classification	Class *Big Boat*	Class *MISL Boat*	Class *Average Boat*	Total
Is *Big Boat*	1	9	1	11
Is *MISL Boat*	9	48	5	62
Is *Average Boat*	1	5	0	6
Total	11	62	6	79

Матрицата на наивна класификация се изчислява по следния начин: Тъй като алгоритмите се прилагат за тестова опция “10-fold Cross Validation”, резултатите от тестването на модела се получават за пълната съвкупност от данни. В случая в данните се съдържат 79 записа – 11 в клас *Big Boat* (14%), 62 в клас *MISL Boat* (78%) и 6 в клас *Average Boat* (8%). Стойностите в клетките на матрицата се изчисляват като:

- ***Случай 1***: Class *Big Boat* Is *Big Boat*
  (0.14x0.14)x79=0.0196x79=1.55~**1** или 2
- ***Случай 2***: Class *Big Boat* Is *MISL Boat*
  (0.14x0.78)x79=0.1092x79=8.63**~9**
- ***Случай 3***: Class *Big Boat* Is *Average Boat*
  (0.14x0.08)x79=0.0112x79=0.89**~**0 или **1**
- ***Случай 4***: Class *MISL Boat* Is *Big Boat*
  (0.78x0.14)x79=0.1092x79=8.63**~9**
- ***Случай 5***: Class *MISL Boat* Is *MISL Boat*

  (0.78x0.78)x79=0.6084x79=48.06~**48**

- ***Случай 6***: Class *MISL Boat* Is *Average Boat*
  (0.78x0.08)x79=0.0624x79=4.93~**5**
- ***Случай 7***: Class *Average Boat* Is *Big Boat*
  (0.08x0.14)x79=0.0112x79=0.89**~1**
- ***Случай 8***: Class *Average Boat* Is *MISL Boat*
  (0.08x0.78)x79=0.0624x79=4.93~**5**
- ***Случай 9***: Class *Average Boat* Is *Average Boat*

  (0.08x0.08)x79=0.0064x79=0.51~**0 или 1**

Отношението *Правилно/Неправилно класифицирани записи за наивната класификация* сее изчислява като:

Naïve Correct/Incorrect Ratio=(1+48+0)/(9+1+9+5+1+5)=49/30=**1.63**

Получената стойност 1.63 на отношението Правилно/Неправилно класифицирани записи за наивната класификация показва, че броят на правилно класифицирани обекти е 1.63 пъти по-голям от броя на неправилно класифицирани обекти, т.е. на всеки 1.63 правилно класифицирани обекти се получава една грешка.

Получените различни модели за класификация са сравнение по следните критерии за оценка:

- *Точност/Грешка на предсказване*
- *Точност на предсказване по отношение на трите клас*а (трите различни стойности на предсказваната променлива) чрез параметъра *True Positive (TP) Rate* (пропорцията от обекти, за които е предсказан клас X, сред всички обекти, които наистина принадлежат към клас X).
- *Коефициент на подобрение на предсказването* спрямо случайната класификация

Алгоритъм „Дърво на решенията”

Резултатите от прилагането на алгоритъма WEKA *J48* върху данните са представени в Табл.3.24:

Таблица 3.24: Резултати за оценка на полученото *Дърво на решенията* чрез алгоритъм *J48* в WEKA

***J48* Algorithm**	**Classes**			**Weighted Average**
	Big Boat	**MISL Boat**	**Average Boat**	
Correctly Classified Instances	81.0127%			
Incorrectly Classified Instances	18.9873%			
Kappa Statistic	0.3328			
TP Rate	0.091	0.968	0.5	0.81
FP Rate	0.029	0.647	0.027	0.514
Precision	0.333	0.845	0.6	0.755
Recall	0.091	0.968	0.5	0.81
F-Measure	0.143	0.902	0.545	0.769
ROC Area	0.515	0.617	0.527	0.596

Точността на предсказване на модела, генериран чрез алгоритъма J48 е много висока (81%), но от получените резултати се вижда, че това се дължи най-вече на високата точност на предсказване на клас *MISL Boat* (*TP Rate*=0.968). Точността на предсказване на другите два класа е значително по-ниска (*TP Rate*=0.5 за клас *Average Boat* и *TP Rate*=0.091 за клас *Big Boat*).

Параметърът *Kappa Statistic*, показващ степента на съгласуваност между предсказването и реалния клас, има стойност 0.3328 (максимална възможна стойност - 1.0, показваща пълно съгласуване), което означава, че точността на предсказване не е много добра за всички класове. Параметърът *ROC Area* има стойност 0.596, което също показва, че от този класификатор не може да се очаква голяма точност на предсказване за всички класове.

Матрицата с получените резултатите от класификацията (Confusion Matrix) с *J48* алгоритъма е представена в Табл.3.25.

Таблица 3.25: Матрица на класификацията за полученото *Дърво на решенията* чрез алгоритъм *J48* в WEKA

J48 Classifier	Class *Big Boat*	Class *MISL Boat*	Class *Average Boat*	Total
Is *Big Boat*	1	9	1	11
Is *MISL Boat*	1	60	1	62
Is *Average Boat*	1	2	3	6
Total	3	71	5	79

Точността на предсказване на алгоритъма се изчислява чрез броя на правилно и неправилно класифицираните записи (стойностите са представени в Табл.3.26).

Таблица 3.26: Брой правилно/неправилно класифицирани записи за полученото *Дърво на решенията* чрез алгоритъм *J48* в WEKA

Модел	Брой	% от тестовата извадка
Правилно предсказани	1+60+3=64	68% (2309x3423=0.68)
Неправилно предсказани	9+1+1+1+1+2=15	32% (1114x3423=0.32)

Отношението между правилно и неправилно класифицираните обекти за генерирания модел на класификация се изчислява от Табл.3.26:

Model Correct/Incorrect Ratio=64/15=**4.27**

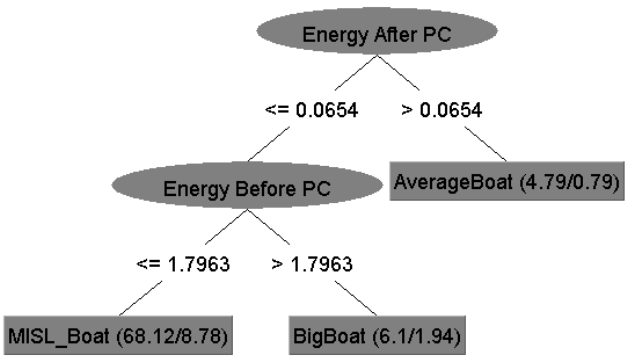
Получената стойност на съотношението Правилно/Неправилно класифицирани обекти 4.27 показва, че броят на правилно класифицираните обекти е над 4 пъти по-голям от броя на неправилно класифицираните обекти, т.е. на всеки 4.27 правилно класифицирани обекти се получава една грешка.

За да се оцени работата на генерирания модел за класификация се сравняват получените отношения Правилно/Неправилно класифицирани записи за модела и при наивна класификация:

$$\frac{Model\,\frac{Correct}{Incorrect}\,Ratio}{Naive\,\frac{Correct}{Incorrect}\,Ratio} = \frac{4.27}{1.63} = \mathbf{2.62}$$

Получената стойност 2.62 показва, че генерираният модел извършва класификацията 2.62 пъти по-добре от наивната класификация.

На Фиг.3.32 е представено полученото „Дърво на решенията" чрез алгоритъма J48:



Фиг.3.32: Полученото *Дърво на решенията* чрез алгоритъм *J48* в WEKA

Полученото дърво на решенията е с размер 5, 3 листа и е изградено на 3 нива. Променливите, които оказват най-съществено влияние при разделянето на обектите на трите класа са *Energy After PC* и *Energy Before PC, т.е. енергията на отразения сигнал най-силно влияе върху разделянето на морските цели в трите класа. Това е разбираемо и от физическа гледна точка, тъй като енергията на сигнала включва както средна мощност на полезния FSR сигнал, така и средното време на съществуване на сигнала от целта.*

Много подобни са и резултатите за модела на класификация, получен чрез прилагането на другия избран алгоритъм от типа „Дърво на решенията" – WEKA *RandomForest*. Получената точност на предсказване е 81%, но отново съществуват разлики по отношение предсказването на трите класа. Стойността на параметър Kappa Statistic=0.2653 е по-ниска отколкото за алгоритъма J48, но стойността на параметъра ROC Area=0.646 е малко по-висока. Това означава, че малко се влошава точността на предсказване на двата по-слабо представени класа (наистина, *TP Rate*=0.333 за клас *Average Boat* и *TP Rate*=0.091 за клас *Big Boat)* в общата съвкупност от данни, но леко се подобрява надеждността на предсказването на най-представения в данните клас *MISL Boat*.

Изводи за модела на класификация, получен чрез алгоритъма "Дърво на решенията":

- *Общата точност на предсказване* на модела е висока – 81%, но има съществени разлики по отношение предсказването на трите класа.
- Най-висока е *точността на предсказване* на клас *MISL Boat*, което най-вероятно се дължи на факта, че именно този клас има най-много представители в изследваната съвкупност от данни (78%).
- *Точността на предсказване* на другите два класа, *Big Boat* и *Average Boat*, е значително по-ниска.
- *Kappa Statistic*=0.3328 (при максимална възможна стойност 1.0), следователно степента на съгласуваност между предсказания и реалния клас не е много висока, т.е. някои класовете се предсказват с много ниска точност.
- *ROC Area*=0.596 (стойност <0.7), което също показва, че от този класификатор не може да се очаква голяма точност на предсказване за всички класове.
- *Model Correct/Incorrect Ratio*=4.27
- $\frac{Model\frac{Correct}{Incorrect}Ratio}{Naive\frac{Correct}{Incorrect}Ratio} = 2.62$, следователно се постига 2.62 пъти по-добро предсказване в сравнение с класификацията на случаен принцип (при наивна класификация).
- Променливите, които оказват най-съществено влияние при разделянето на обектите на трите класа са *Energy After PC* и *Energy Before PC, т.е. енергията на отразения сигнал най-силно влияе върху разделянето на морските цели в трите класа. Това е разбираемо и от физическа гледна точка, тъй като енергията на сигнала включва*

както средна мощност на полезния FSR сигнал, така и средното време на съществуване на сигнала от целта.

Алгоритъм за класификация JRip

Резултатите от прилагането на алгоритъма WEKA *1R* върху данните са представени в Табл.3.27:

Таблица 3.27: Резултати за оценка на получения модел за класификация чрез алгоритъм *JRip* в WEKA

Rule Learner JRip	**Classes**			**Weighted Average**
	Big Boat	**MISL Boat**	**Average Boat**	
Correctly Classified Instances	82.2785%			
Incorrectly Classified Instances	17.7215%			
Kappa Statistic	0.3333			
TP Rate	0.182	0.984	0.333	0.823
FP Rate	0.029	0.706	0	0.558
Precision	0.5	0.836	1	0.801
Recall	0.182	0.984	0.333	0.823
F-Measure	0.267	0.904	0.5	0.784
ROC Area	0.58	0.577	0.655	0.583

Моделът за предсказване, получен чрез алгоритъма *JRip*, е с обща точност на предсказване 82,3%, но отново се вижда, че това се дължи най-вече на много високата точност на предсказване на клас *MISL Boat* (*TP Rate*=0.984). Точността на предсказване на другите два класа е значително по-ниска (*TP Rate*=0.333 за клас *Average Boat* и *TP Rate*=0.182 за клас *Big Boat*).

Параметърът *Kappa Statistic*, показващ степента на съгласуваност между предсказания и реалния клас, има стойност 0.3333, което означава, че точността на предсказване не е много добра за всички класове. Параметърът *ROC Area* има стойност 0.583, което също показва, че от този класификатор не може да се очаква голяма точност на предсказване за всички класове.

Матрицата с получените резултатите от класификацията (Confusion Matrix) с *JRip* алгоритъма е представена в Табл.3.28.

Таблица 3.28: Матрица на класификацията за получения модел за класификация чрез алгоритъм *JRip* в WEKA

JRip Rule Learner	Class *Big Boat*	Class *MISL Boat*	Class *Average Boat*	Total
Is *Big Boat*	2	9	0	11
Is *MISL Boat*	1	61	0	62
Is *Average Boat*	1	3	2	6
Total	4	73	2	79

Model Correct/Incorrect Ratio=65/14=**4.64**

Получената стойност на съотношението Правилно/Неправилно класифицирани обекти 4.64 показва, че броят на правилно класифицираните обекти е над 4 пъти по-голям от броя на неправилно класифицираните обекти, т.е. на всеки 4.64 правилно класифицирани обекти се получава една грешка.

За да се оцени работата на генерирания модел за класификация се сравняват получените отношения Правилно/Неправилно класифицирани записи за модела и при наивна класификация:

$$\frac{Model\,\frac{Correct}{Incorrect}\,Ratio}{Naive\,\frac{Correct}{Incorrect}\,Ratio} = \frac{4.64}{1.63} = \mathbf{2.85}$$

Получената стойност 2.85 показва, че генерираният модел извършва класификацията 2.85 пъти по-добре от наивната класификация.

Изводи за модела на класификация, получен чрез алгоритъма JRip Rule Learner:

- *Общата точност на предсказване* на модела е висока – 82%, но има съществени разлики по отношение предсказването на трите класа.
- Най-висока е *точността на предсказване* на клас *MISL Boat*, което най-вероятно се дължи на факта, че именно този клас има най-много представители в изследваната съвкупност от данни (78%).
- *Точността на предсказване* на другите два класа, *Big Boat* и *Average Boat*, е значително по-ниска.
- *Kappa Statistic*=0.3333, следователно степента на съгласуваност между предсказания и реалния клас не е много висока, т.е. някои класовете се предсказват с много ниска точност.
- *ROC Area*=0.583 (стойност <0.7), което също показва, че от този класификатор не може да се очаква голяма точност на предсказване за всички класове.

- *Model Correct/Incorrect Ratio*=4.64
- $\frac{Model\frac{Correct}{Incorrect}Ratio}{Naive\frac{Correct}{Incorrect}Ratio} = 2.85$, следователно се постига 2.85 пъти по-добро предсказване в сравнение с класификацията на случаен принцип (при наивна класификация).

Алгоритъм за класификация SimpleLogistic

Резултатите от прилагането на алгоритъма WEKA *SimpleLogistic* върху данните са представени в Табл.3.29:

Таблица 3.29: Резултати за оценка на получения модел за класификация чрез алгоритъм *SimpleLogistic* в WEKA

SimpleLogistic Algorithm	**Classes**			**Weighted Average**
	Big Boat	**MISL Boat**	**Average Boat**	
Correctly Classified Instances	81.0127%			
Incorrectly Classified Instances	18.9873%			
Kappa Statistic	0.1828			
TP Rate	0.091	1	0.167	0.81
FP Rate	0	0.882	0	0.692
Precision	1	0.805	1	0.847
Recall	0.091	1	0.167	0.81
F-Measure	0.167	0.892	0.286	0.745
ROC Area	0.609	0.663	0.709	0.659

Моделът за предсказване, получен чрез алгоритъма *SimpleLogistic*, е с обща точност на предсказване 81%, но отново се вижда, че това се дължи най-вече на изключително високата точност на предсказване на клас *MISL Boat* (*TP Rate*=1, т.е. няма грешно предсказани обекти от клас *MISL Boat*). Точността на предсказване на другите два класа е значително по-ниска (*TP Rate*=0.167 за клас *Average Boat* и *TP Rate*=0.091 за клас *Big Boat*).

Параметърът *Kappa Statistic*, показващ степента на съгласуваност между предсказания и реалния клас, има много ниска стойност 0.1828 (<<1), което означава, че точността на предсказване не е добра за всички класове. Параметърът *ROC Area* има стойност 0.659, което показва, че от този класификатор не може да се очаква много голяма точност на предсказване за всички класове.

Матрицата с получените резултатите от класификацията (Confusion Matrix) с алгоритъма *SimpleLogistic* е представена в Табл.3.30.

Таблица 3.30: Матрица на класификацията за получения модел за класификация чрез алгоритъм *SimpleLogistic* в WEKA

JRip Rule Learner	Class *Big Boat*	Class *MISL Boat*	Class *Average Boat*	Total
Is *Big Boat*	1	10	0	11
Is *MISL Boat*	0	62	0	62
Is *Average Boat*	0	5	1	6
Total	1	77	1	79

Вижда се, че моделът предсказва с абсолютна точност клас *MISL Boat*, но предсказва клас *MISL Boat* и за повечето обекти от другите два класа.

Model Correct/Incorrect Ratio=64/15=**4.27**

Получената стойност на съотношението Правилно/Неправилно класифицирани обекти 4.27 показва, че броят на правилно класифицираните обекти е над 4 пъти по-голям от броя на неправилно класифицираните обекти, т.е. на всеки 4.27 правилно класифицирани обекти се получава една грешка.

За да се оцени работата на генерирания модел за класификация се сравняват получените отношения Правилно/Неправилно класифицирани записи за модела и при наивна класификация:

$$\frac{Model\,\frac{Correct}{Incorrect}\,Ratio}{Naive\,\frac{Correct}{Incorrect}\,Ratio} = \frac{4.27}{1.63} = \mathbf{2.62}$$

Получената стойност 2.62 показва, че генерираният модел извършва класификацията 2.62 пъти по-добре от наивната класификация.

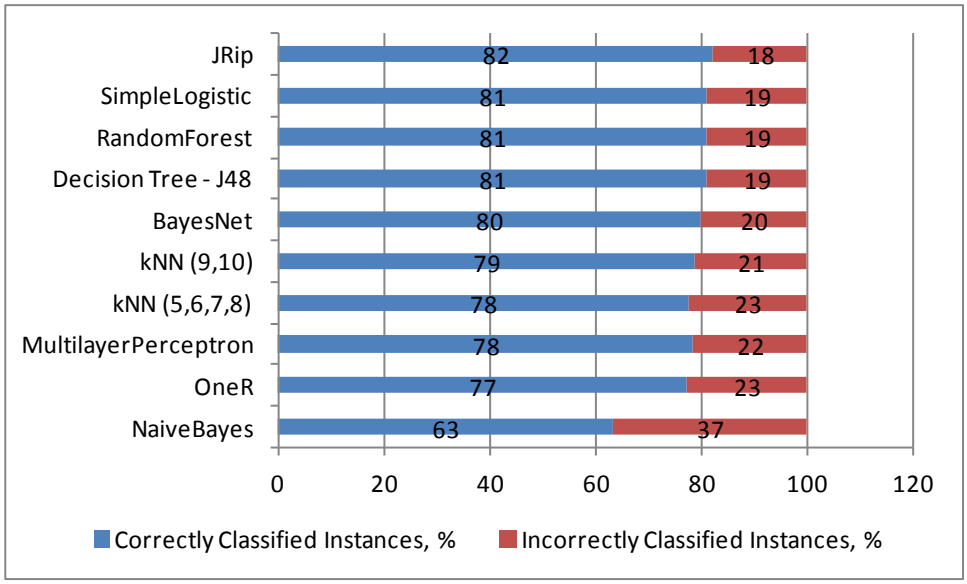
Изводи за модела на класификация, получен чрез алгоритъма JRip Rule Learner:

- *Общата точност на предсказване* на модела е висока – 82%, но има съществени разлики по отношение предсказването на трите класа.
- Алгоритъмът предсказва без грешка клас *MISL Boat*.
- Алгоритъмът предсказва най-често клас *MISL Boat* за обектите от другите два класа - *Big Boat* и *Average Boat*.
- *Kappa Statistic*=0.1828 (<<1), следователно степента на съгласуваност между предсказания и реалния клас е много ниска, т.е. потвърждава се неспособността на алгоритъма да предсказва точно някои от класовете.
- *ROC Area*=0.659 (стойност <0.7), което също показва, че от този класификатор не може да се очаква голяма точност на предсказване за всички класове.

- *Model Correct/Incorrect Ratio*=4.27
- $\frac{Model\frac{Correct}{Incorrect}Ratio}{Naive\frac{Correct}{Incorrect}Ratio} = 2.62$, следователно се постига 2.62 пъти по-добро предсказване в сравнение с класификацията на случаен принцип (при наивна класификация).

Сравнение на резултатите от класификацията, получени чрез прилагането на избраните Data Mining алгоритми

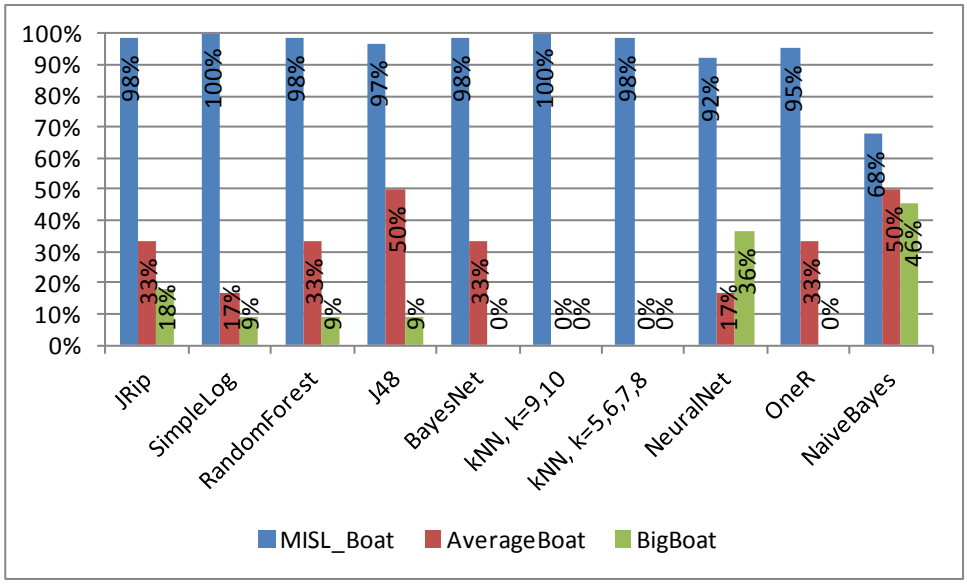
Получените резултати за Точността/Грешката, получена при предсказване на вида на откритите FS радарни морски цели (класа) чрез различните модели за класификация, генерирани с избраните Data Mining алгоритми, са представени на Фиг.3.33.



Фиг.3.33. Сравнение на моделите, получени с избраните класификационни алгоритми, по отношение на точността/грешката на предсказване

На Фиг.3.33 са показани пропорциите (% стойности) на правилно и неправилно класифицираните обекти (откритите морски цели) за всеки от приложените Data Mining алгоритми. Резултатите показват, че с най-голяма точност работят Генераторът на правила Jrip (82% правилно предсказани обекти), алгоритмите от типа Дърво на решенията – J48 и RandomForest (81% правилно предсказани обекти), и Логистичната регресия – SimpleLogistic (81% correctly classified instances). Останалите алгоритми също работят със сравнително висока точност (между 77% и 80%), с изключение на Наивния Байесов класификатор (NaiveBayes), който дава точност на предсказване 63%.

Получените резултати от сравнението на генерираните модели за класификация чрез различните Data Mining методи, относно способността им за предсказване на трите различни класа – *MISL Boat*, *Average Boat* и *Big Boat*, на базата на параметъра True Positive Rate (TP Rate), са представени на Фиг.3.34.



Фиг.3.34: Сравнение на моделите, получени с избраните класификационни алгоритми, по отношение на точността на класификация по класове

Резултатите показват, че всички класификационни алгоритми работят най-добре, т.е. предсказват с най-висока точност клас *MISL Boat,* което най-вероятно се дължи на факта, че в анализираните данни се съдържат най-много обекти, принадлежащи на този клас. Точността на предсказване е доста по-ниска за клас *Average Boat* и най-ниска за клас *Big Boat*. Тези резултати не са изненадващи, като се има предвид пропорционалното представяне на тези два класа в анализираната съвкупност от данни. Някои от генерираните модели, получени чрез алгоритмите *BayesNet, OneR* и *kNN*, са абсолютно неспособни да предсказват класовете, които са по-слабо представени в наличната съвкупност от данни.

Изводи:

- Всички избрани класификационни алгоритми - Генератори на правила OneR и JRip; Дърво на решението - J48 и RandomForest; Невронна мрежа - Multilayer Perceptron; К-най-близък съсед IBk; Байесови класификатори - NaiveBayes и BayesNet, и Логистична регресия - SimpleLogistic, предсказват с висока точност (68% - 100%)) клас MISL Boat.
- Точността на предсказване е доста по-ниска за клас Average Boat и най-ниска за клас Big Boat, които са по-слабо представени в общата съвкупност от данни

- Алгоритмите BayesNet, OneR и kNN, са абсолютно неспособни да предсказват класовете, които са по-слабо представени в наличната съвкупност от данни.
- Променливите, които оказват най-съществено влияние при разделянето на обектите на трите класа са *Energy After PC* и *Energy Before PC*, т.е. енергията на отразения сигнал най-силно влияе върху разделянето на морските цели в трите класа. Това е разбираемо и от физическа гледна точка, тъй като енергията на сигнала включва както средна мощност на полезния FSR сигнал, така и средното време на съществуване на сигнала от целта.
- Data Mining моделът за класификация, получен чрез алгоритъма Дърво на решенията (WEKA J48), е със сравнително висока точност на предсказване за два от класовете (97% за MISL Boat, 50% за Average Boat) и с ниска точност за третия клас (9% за Small Boat). Но, полученият модел е много разбираем, лесен за интерпретация от потребителите, и по-лесно може да се реализира за работа в реално време.
- Data Mining моделът за класификация, получен чрез Наивния Байесов класификатор (WEKA NaiveBayes), е с близки точности на предсказване за трите класа (68% за MISL Boat, 50% за Average Boat, 46% за Small Boat). Това е единственият модел, който дава сравнително приемлива точност на класификация и за трите вида морски цели.
- Получените резултати са съизмерими с резултатите, получени от изследователския колектив на Бирмингамския университет, отнасящи се за класификация на движещи се наземни цели (автомобили) чрез използване на Forward Scattering радарна система, представени в т.1.4.

3.3. Изводи по Трета Глава

В настоящата глава е постигнато следното:

- Обучени са класификатори за извличане на знания от данни (генерирани са Data Mining модели за класификация) чрез избрани класификационни алгоритми, за две съвкупности от данни, в Data Mining софтуер *WEKA*.
- Качеството на обучените класификатори е оценено по избраните мерки за оценка - *Точност/Грешка при класификацията; Матрица на класификация; Параметрите TP Rate, FP Rate, Precission, Recall и F-measure; Kappa Statistic; ROC Area; Model Correct/Incorrect Ratio; Model/Naïve Correct/Incorrect Ratio*.
    - Оценени и сравнени са генерираните Data Mining модели (обучените класификатори) за предсказване успеваемостта на студентите, чрез принадлежност към един от пет възможни класа.
    - Оценени и сравнени са генерираните Data Mining модели (обучените класификатори) за предсказване успеваемостта на студентите, чрез принадлежност към един от два възможни класа.

- Сравнени са получените резултати, относно предсказване успеваемостта на студентите, за двата изследвани варианта на предсказваната променлива - с пет и с две стойности.
- Оценени и сравнени са генерираните Data Mining модели (обучените класификатори) за предсказване класа на морски цели, представен като номинална променлива с три стойности, по параметрите на сигналите от тези цели.

- Обучените класификатори (генерираните Data Mining модели) за предсказване успеваемостта на студентите на базата на техни характеристики преди и след приемане в университета, са генерирани чрез избраните класификационни алгоритми (Генератор на правила *OneR*, Дърво на решението *J48*, Невронна мрежа *MultiLayer Perceptron* и К-най-близък съсед *IBк*) и са получени за два варианта на предсказваната променлива (с пет и с две стойности). За случая на предсказвана променлива с две стойности е предварително изследвано и влиянието на всяка входна променлива върху предсказването, с цел да се открият факторите, които в най-голяма степен влияят върху успеваемостта на студентите. Резултатите показват, че:
  - Генерираните модели за класификация (обучените класификатори), чрез четирите избрани алгоритми, дават по-голяма точност на предсказване, когато изходната променлива съдържа по-малък брой различни стойности.
  - За *предсказвана променлива с две стойности („Слаби" и „Силни")*, всички приложени алгоритми, с изключение на 1R класификатора, работят с обща точност на предсказване >70% (weighted average accuracy), като се справят сравнително добре и с двата класа. И за трите алгоритъма (*Невронна мрежа, Дърво на решението* и *К-Най-близък съсед*) се получават високи стойности на параметрите *Kappa Statistic (>0.40)* и *ROC Area (>0.78*), което означава, че генерираните модели са подходящи за извършване на класификацията с висока точност. *Невронната мрежа* е генерираният модел за класификация с най-добри параметри, той е единствения модел, който се справя по-добре с предсказването на клас „Силни", което в конкретния бизнес контекст е особено важно, и чрез използване на този модел точността на предсказване може да се подобри 2.76 пъти в сравнение с наивната класификация.
  - За *предсказвана променлива с пет стойности („Слаб", „Среден", „Добър", „Много Добър" и „Отличен")*, е получена по-ниска точност на предсказване и при четирите използвани алгоритъма (<70%). Наблюдават се съществени разлики в точността на предсказване на различните класове, най-ниска е точността за клас „Отличен". Най-добър модел се получава чрез алгоритъма „*Дърво на решенията*", който дава 2.85 пъти подобрение в сравнение с наивната класификация.
- Data Mining моделите за класификация (обучените класификатори), за предсказване класа на морски цели (представен като номинална променлива с три стойности, в

зависимост от параметрите на FSR сигналите на целите), са генерирани чрез избраните Data Mining алгоритми (*Генератори на правила - OneR* и *JRip, Дърво на решението - J48* и *RandomForest, Невронна мрежа – Multilayer Perceptron, К-най-близък съсед - IBk,* Байесови класификатори - *NaiveBayes* и *BayesNet,* и Логистична регресия - *SimpleLogistic*).

- Всички избрани класификационни алгоритми предсказват с добра точност (68% - 100%) клас MISL Boat. Точността на предсказване е доста по-ниска за клас Average Boat и най-ниска за клас Big Boat, които са по-слабо представени в общата съвкупност от данни.
- Data Mining моделът за класификация (обученият класификатор), получен чрез алгоритъма Дърво на решенията (WEKA J48), е със сравнително висока точност на предсказване за два от класовете (97% за MISL Boat, 50% за Average Boat) и с ниска точност за третия клас (9% за Small Boat). Но, полученият модел е много разбираем, лесен за интерпретация от потребителите, и по-лесно може да се реализира за работа в реално време. Променливите, които оказват най-съществено влияние при разделянето на обектите на трите класа са *Energy After PC* и *Energy Before PC* (Дърво на решенията), т.е. енергията на отразения сигнал най-силно влияе върху разделянето на морските цели в трите класа. Това е разбираемо и от физическа гледна точка, тъй като енергията на сигнала включва както средна мощност на полезния FSR сигнал, така и средното време на съществуване на сигнала от целта.
- Data Mining моделът за класификация (обученият класификатор), получен чрез Наивния Байесов класификатор (WEKA NaiveBayes), е с близки точности на предсказване за трите класа (68% за MISL Boat, 50% за Average Boat, 46% за Small Boat). Това е единственият модел, който дава сравнително приемлива точност на класификация и за трите вида морски цели.
- Получените резултати са съизмерими с резултатите, получени от изследователския колектив на Бирмингамския университет, отнасящи се за класификация на движещи се наземни цели (автомобили) чрез използване на Forward Scattering радарна система.

- В дисертацията са получени и изследвани Data Mining модели за класификация (обучени класификатори), на базата на една и съща методика, за две различни съвкупности от данни. Резултатите показват, че обучените класификатори могат да бъдат успешно използвани за предсказване класа на принадлежност както на студентите, така и на радарните морски цели, независимо че данните са много различни - произхождат от различни предметни области и са с различен обем. Следователно, в дисертацията отново се потвърждава факта, че Data Mining методите и средствата могат да бъдат ефективно прилагани върху различни видове данни от различни научни области.

Заключение

В настоящия дисертационен труд е решена и изследвана *задачата за откриване на знания в данни чрез обучение на класификатори*, за две съвкупности от данни в различни предметни области - *за предсказване успеваемостта на студентите в зависимост от техни характеристики преди и след приемане в университета*, и за *предсказване класа на морски обекти в зависимост от избрани параметри на отразените сигнали от тези обекти*.

Задача за предсказване успеваемостта на студентите

Осъщественият анализ на публикуваните научни изследвания в областта на Educational Data Mining показва, че предсказването успеха на студентите е много важен и актуален проблем за университетите. Резултатите от провежданите изследвания се използват за различни цели:

- *Откриват се изоставащи студенти*, които евентуално ще отпаднат от обучение поради слабото си представяне, и на тези студенти се предоставя допълнителна помощ, за да бъдат задържани в университета.
- *Откриват се добрите студенти* – това са студентите, които се справят най-успешно с обучението и които са най-желани за университета, и именно към такъв тип студенти се насочват бъдещите маркетингови кампании.
- *Откриват се факторите, които в най-голяма степен влияят върху успеха или слабото представяне на студентите*, и това се превръща в ценна информация, на базата на която се взимат управленски решения за извършването на промени и за подобряване ефективността и качеството на предлаганото обучение.

Успехът на студентите се предсказва най-често чрез откриване на знания в данни чрез обучение на класификатори (прилагане на *Data Mining методи и средства* за решаване на задача за предсказване), като в повечето случаи *предсказваната величина е номинална променлива*, т.е. решава се задача за *класификация*, а в по-редки случаи *предсказваната променлива е цифрова*, т.е. решава се задача за *регресия*.

Най-често използваните методи, в областта на Educational Data Mining, за обучение на класификатори включват *Дърво на решенията*, *Невронни мрежи*, *Байесови методи* - *Наивен Байесов класификатор*, *Байесови мрежи*, *Логистична регресия* и др. Във всички случаи обаче, се прилагат няколко различни метода или няколко различни алгоритми, за да се достигне до създаването на подходящ класификатор.

За да се получат класификатори с добри параметри, е необходимо да се извърши внимателно *изучаване и предварителна подготовка на данните*, както и да се избере и форматира по подходящ начин предсказваната променлива.

Изучаването на данните за студенти, с които разполагат университетите, е свързано с опознаване на източниците, от които са интегрирани тези данни, и с разбиране на значението и допустимите стойности на съдържащите се променливи в данните. В резултат на изучаването на данните могат да бъдат получени интересни изводи относно разпределението на студентите по важни техни характеристики, които впоследствие могат да бъдат използвани от ръководството на университите за подобряване на организацията на учебния процес и за повишаване качеството на предлаганото обучение.

Предварителната подготовка на данните е много важен етап от обучението на класификатори. От правилното извършване на дейностите по предварителната подготовка на данните (премахването на грешки и несъответствия, попълване на липсващи стойности, преобразуване на променливи, създаване на нови променливи, редуциране на променливи и др.) до голяма степен зависи качеството на получените класификатори. Предварителната подготовка на данните за студенти включва различни дейности като:

- *Редуциране на променливи*
    - *Премахване на променливи*, които съдържат една единствена стойност (например променливата „Държава”, когато се търсят съществуващи зависимости в данните само за обучаваните български студенти) или не представляват интерес за конкретните цели на изследванията (например „Месторождение” на студента).
    - *Избор на най-информативни променливи*, които да участват при обучението на класификаторите – това са променливите, които в най-голяма степен влияят върху разделянето на обектите в съвкупността от данни в избраните класове.
- *Преобразуване на променливи*
    - *Преобразуване на променливи от тип „свободен текст”*

      Много често събираните данни в университетите са неструктурирани, т.е. в много от съществуващите полета в базите данни се съдържа свободен текст. Работата с неструктурирани данни силно затруднява и ограничава възможностите за анализ, затова е препоръчително „свободният текст” да се формализира като се замени с определен брой дискретни стойности, извлечени по смисъл от наличния текст. Трябва да се има предвид, че наличието на голям брой дискретни стойности за дадена променлива също може да затрудни анализите и да не позволи получаването на съществени изводи, затова трябва да се направи добра преценка при преобразуването.

Например, променливата „Профил от средното образование” може да се получи от полето „Училище, в което е завършено средното образование”, където обикновено се въвежда име на училището, като „свободният текст” в оригиналните данни се трансформирана в номинална променлива, приемаща краен брой стойности, съответстващи на различните профили на средните училища в България. Това може да бъде постигнато като се вземе предвид профилирането на училищата, в които получават средното си образование учениците в България (непрофилирани СОУ и профилирани училища, в които се извършва прием след завършване на 7 и 8 клас), като такава информация може да бъде получена от Министерство на образованието, младежта и науката (www.minedu.government.bg).

Друга важна променлива, която може да представлява интерес за ръководството на университетите, е „Регион, от който произхождат студентите”. Тази променлива може да бъде получена отново чрез преобразуване на „свободния текст”, който обикновено се съдържа в поле „Населено място на училището, в което е завършено средното образование. В зависимост от целите на изследванията, трансформирането в номинална променлива с краен брой стойности може да се извърши по различни начини – да се запазят имената на по-големите селища, а по-малките да се заменят с регионите, в които се намират; да се използват само регионите, за да се правят по-обобщени анализи и т.н.

- *Преобразуване на цифрови в номинални променливи*

  Извършва се понякога за променливи, които приемат числови стойности, но на практика не носят числов смисъл (т.е. не е логично да участват в различни математически изчисления), затова се трансформират в номинални променливи.

  Много често се извършва такова преобразуване, за да се получи подходяща променлива за предсказване от номинален тип – при решаване на задача за класификация. Например, „Среден успех” на студентите е числова променлива, но преобразуването ѝ в номинална променлива с краен брой стойности може да доведе до извличане на много по-съществени закономерности от данните (при откриване на слаби студенти, на които да се окаже предварителна помощ, за да не отпаднат от обучението; при откриване на силни студенти, за да се насочат маркетинговите кампании към подобен тип студенти и т.н.).

- *Създаване на нови променливи*

Тази дейност е полезна, когато в наличните данни не се съдържат в явен вид важни характеристики на студентите. Например, в данните за студенти обикновено не се

съдържа поле „Възраст”, но този параметър може лесно да бъде изчислен от наличните полета „ЕГН” и „Година на приемане в университета”.

- *Изчистване на грешки*

  Свързано е с откриване и премахване/коригиране на стойности, които са извън допустимия интервал за съответната променлива. Препоръчително е при изчистването на грешките да се използва консултиране с представителите на университетската администрация, отговорни за събирането, съхранението и образотката на данните, тъй като те са най-добре познават данните и могат да дадат логични обяснения за наличието на стойности, които на пръв поглед изглеждат грешни, но в действителност могат да съдържат полезна информация за изследванията.

- *Други преобразувания*

  В зависимост от спецификата на използваните софтуерни средства, при осъществяването на Data Mining анализите може да се наложи преобразуването на данните от кирилица на латиница, за да са разпознаваеми стойностите на променливите.

Обучението на класификаторите може да се извърши чрез изпозване на *различни методи и алгоритми за класификация*. Получените класификатори могат да бъдат оценявани и сравнявани чрез различни метрики за оценка - Точност/Грешка при класификацията; Матрица на класификация; параметрите TP Rate, FP Rate, Precission, Recall и F-measure; Kappa Statistic; ROC Area; Model Correct/Incorrect Ratio; Model/Naïve Correct/Incorrect Ratio.

Получените резултати в настоящия дисертационен труд показват, както трябва да се очаква, че *класификатори с по-висока точност на предсказване* се получават, когато *изходната променлива съдържа по-малък брой различни стойности*.

При *предсказвана променлива с две стойности* - „Слаби” и „Силни”, всички приложени алгоритми (*Невронна мрежа, Дърво на решението, К-Най-близък съсед* и *1R*) работят с по-висока точност на предсказване (weighted average accuracy>70%), като се справят сравнително добре и с двата класа. И четирите получени класификатора подобряват повече от два пъти пресказването, в сравнение с класификацията на случаен принцип. *Невронната мрежа* е генерираният *класификатор с най-добри параметри* по отношение на предсказването и е единственият класификатор, който работи по-добре при предсказването на клас „Силни”. Останалите три класификатора предсказват с по-висока точност клас „Слаби”. Понякога може да се наложи да се направи компромис с точността

на предсказване, например когато е важно да бъдат получени разбираеми за потребителите класификатори и правила за предсказване. *Дърво на решение*, *К-Най-близък съсед* и *1R* са по-разбираеми и лесни за интерпретация класификатори, в сравнение с *Невронната мрежа*.

При *предсказвана променлива с пет различни стойности* – „Слаб", „Среден", „Добър", „Много Добър" и „Отличен", се получават класификатори с по-ниска точност на предсказване (<70%). И четирите модела осигуряват поне 2 пъти по-точно предсказване спрямо наивната класификация. Наблюдават се *съществени разлики в точността на предсказване на петте класа,* което най-вероятно се дължи на неравномерното представителство на тези класове в общата съвкупност от данни. И четирите класификатора предсказват с по-високи точности двата класа, които са най-силно представени в общата съвкупност от данни – класове „Добър" и „Много добър". Най-ниска е точността на предсказване на *клас „Отличен"*, въпреки че това не е класа, който е най-слабо представен в общата съвкупност от данни. *Дървото на решенията* и *Невронната мрежа* предсказват с по-висока точност и клас „Слаб", който е по-слабо представен от клас „Отличен" в общата съвкупност от данни. *Дърво на решението* е получения класификатор с най-висока средна точност на предсказване и подобрява предсказването *2.85 пъти* спрямо наивната класификация.

За да се редуцира броят на променливите, които се съдържат в общата съвкупност от данни и се използват при обучението на модела, е препоръчително да се изследва индивидуалното влияние на всяка от входните променливи върху предсказването. Резултатите от подобно изследване могат да се използват и за да се намерят факторите, които в най-голяма степен влияят върху класифицирането на студентите в някой от избраните класове, например „Слаби" или „Силни", което да се използва впоследствие от ръководството на университетите за взимането на важни управленски решения.

Изследванията могат да се задълбочат и чрез експериментиране с различни важни параметри на избраните алгоритми за класификация, например *минимален брой обекти в листо на дървото* при алгоритъма „Дърво на решенията", *брой неврони в скрития слой* при алгоритъма „Невронна мрежа", *брой на най-близките К обекти* при алгоритъма „К-най-близък съсед" и др. По този начин могат да бъдат получени класификатори с по-високи точности на предсказване.

Задача за предсказване класа на морски обекти

При обзора относно използваните методи за класификация на радарни цели, открити с радарна система от типа Forward Scattering Radar (FSR), бяха открити публикувани резултати само на представители на изследователския екип от Бирмингамския университет. В тези публикации се използват основно алгоритмите "K-най-близък съсед" (KNN) и "Невронна мрежа" (Multi-Layer Perceptron (MLP), след предварителна обработка

в честотната област, т.е. на спектрите на сигналите от целите. Спектрите на сигналите от целите се обработват предварително с цел да се получат обучаващи сигнали по следния начин: първоначално спектърът на сигнала се нормира по носеща честота, по мощност и по скорост, а след това се "орязва", за да се използва най-информативната му част в ниските доплерови честоти. Използва се и методът Principle Component Analysis (PCA) за намаляване на размерността на параметричното пространство на спектрите на получените сигнали.

За решаването на тази класификационна задача в дисертационния труд е използвана съвкупност от данни, получена чрез съвместните усилия на международен изследователски екип – изследователи от Бирмингамския университет са разработили Forward Scattering радарна система и са направили записи на морски обекти в реално извършени експерименти, след което изследователи от българския екип (СУ ”Св. Кл. Охридски” и ИИКТ-БАН) са обработили записите на Доплеровите сигнали чрез прилагане на CA CFAR алгоритъм с цел автоматично откриване на сигнала от целта и оценка на неговите времеви (енергетични) параметри.

Предварителната подготовка на данни за решаването на задачата за класификация на морски обекти, открити с Forward Scattering радарна система, силно се различава от описаната по-горе подготовка на данни за студенти. Най-важното в случая е да се намерят подходящи параметри на откритите сигнали от морските цели, чрез които по най-добрия начин да се характеризират самите морски обекти. За целите на изследването в дисертацията, оценката на параметрите на сигнала от целта се базира на хипотезата, че големината на целта е еквивалентна на енергията, отразена от нея, и може да се определи чрез обработка на откритите семпли в Доплеровия сигнал от целта. Оценените параметри на сигнала от целта включват брой семпли в откритията отразен сигнал, времетраене, средна мощност и енергия на отразения сигнал, отношение Сигнал/Смущение (S/N Ratio), коефициент на корелация на морското смущение, като параметрите са оценени преди и след подтискане на морското смущение (чрез Pulse Cancellation - алгоритъм за презпериодно изваждане на семплите на Доплеровия сигнал).

Формирането на предсказваната променлива – класа на откритите радарни морски цели, е също много важен проблем, на който трябва да се обърне сериозно внимание. В конкретния случай разпределянето на наличните записи в избраните три класа е извършено на базата на експертна оценка - чрез консултации с членовете на екипа, участващи в реалните изпитания. Добре би било да се използват по-обективни критерии при формирането на предсказваната променлива, например може да се използват методи и алгоритми за клъстерен анализ, които да доведат до формирането на добре изразени и по-равномерно разпределени обекти.

Обучението на класификаторите може да се извърши чрез изпозване на различни Data Mining методи и алгоритми за класификация. Получените класификатори могат да бъдат

оценявани и сравнявани чрез различни метрики за оценка - Точност/Грешка при класификацията; Матрица на класификация; параметрите TP Rate, FP Rate, Precission, Recall и F-measure; Kappa Statistic; ROC Area; Model Correct/Incorrect Ratio; Model/Naïve Correct/Incorrect Ratio.

Получените резултати в настоящия дисертационен труд показват, че всички избрани класификационни алгоритми – генератори на правила *OneR* и *JRIP*, дърво на решението - *J48* и *RandomForest*, Невронна мрежа - *Multilayer Perceptron*, К-най-близък съсед *IBk*; Байесови класификатори - *NaiveBayes* и *BayesNet*, и Логистична регресия - *SimpleLogistic*, предсказват с висока точност (68% - 100%) най-силно представения клас в общата съвкупност от данни - клас *MISL Boat* (гумена моторна лодка, използвана при реалните експерименти). Точността на предсказване е доста по-ниска за клас *Average Boat* и най-ниска за клас *Big Boat*, които са по-слабо представени в общата съвкупност от данни. Най-ниска е точността на предсказване за *клас Big Boat,* въпреки че най-слабо представения клас в общата съвкупност от данни е клас *Big Boat*. Някои алгоритми са абсолютно неспособни да предсказват класовете, които са по-слабо представени в наличната съвкупност от данни (BayesNet, OneR и kNN), което има своите логични обяснения като се вземе предвид самата същност на тези алгоритми, определяща и начина на разпределяне на обектите в наличните класове. Променливите, които оказват най-съществено влияние при разделянето на обектите на трите класа са *Energy After PC* и *Energy Before PC*, т.е. енергията на отразения сигнал най-силно влияе върху разделянето на морските цели в трите класа. Това е разбираемо и от физическа гледна точка, тъй като енергията на сигнала включва както средна мощност на полезния FSR сигнал, така и средното време на съществуване на сигнала от целта. Класификаторът, получен чрез алгоритъма *Дърво на решенията* е със сравнително висока точност на предсказване за два от класовете (97% за MISL Boat, 50% за Average Boat) и с ниска точност за третия клас (9% за Small Boat). Но, полученият модел е много разбираем, лесен за интерпретация от потребителите, и по-лесно може да се реализира за работа в реално време. Класификаторът, получен чрез *Наивния Байесов алгоритъм* (NaiveBayes), е с близки точности на предсказване за трите класа (68% за *MISL Boat*, 50% за *Average Boat*, 46% за *Small Boat*). Това е единственият модел, който дава сравнително приемлива точност на класификация и за трите вида морски цели.

Наличната съвкупност от данни за решаване на задачата за класификация на морски обекти в настоящата дисертация съдържа само 80 записа. Изпълването на Data Mining методи и средства при наличието на малки обеми от данни е необичайно, тъй като Data Mining обикновено се прилага за извличане на закономерности от големи обеми данни. Тъй като проведените изследвания са извършени в рамките на конкретен действащ научноизследователски проект, чиято крайна цел е да се достигне до получаването на реално работеща система за охраняване на морски граници, се предполага, че при внедряването на такава система за работа в реално време в нея ще се събират огромни

обеми от данни, което прави целесъобразно използването на Data Mining методи и средства за класификацията на откритите морски обекти.

Насоки на бъдещи изследвания

Научните изследвания, свързани с анализирането и извличането на знания от данните за студенти, ще се развиват в няколко посоки. Ще се експериментира с различни варианти на предсказваната променлива – класа, към който принадлежат студентите. Ще бъдат приложени различни методи за клъстърен анализ с цел дефиниране на класове, на които да бъдат разделяни студентите, различни от използваните до момента класове, определени чрез експертна оценка. Ще бъдат проучени и възможностите за предсказване успеха на студентите, без предварително преобразуване на числовите стойности на тази променлива, т.е. чрез решаване на задачата за регресия (предсказване на числова променлива). По-подробен анализ ще бъде направен относно влиянието на отделните предсказващи променливи, ще се експериментира с използването на по-малки съвкупности от данни, включващи само най-информативните променливи. Ще бъдат проведени и експерименти за извличане на знания от избрани съвкупности от данни, в които е по-равномерно разпределението на обектите в различните класове.

Научните изследвания в областта на класификацията на радарни цели ще бъдат развити в няколко направления. Ще бъде направен по-подробен анализ относно индивидуалното влияние на всяка от наличните входни променливи върху точността на предсказване на класа на морските обекти, както и избирането на най-подходящи променливи, които да бъдат използвани при решаване на задачата за класификация. Ще бъдат приложени и различни методи с цел преодоляване на проблемите, свързани с малкия размер на използваната съвкупност от данни и неравномерното разпределение на обектите в дефинираните класове. Например, класовете могат да бъдат сведени до два (един клас, който да съдържа обектите от класове *Average Boat* и *Big Boat*, и втори клас - *MISL Boat*), за да се получи по равномерно разпределение на данните по класове. Чрез методи за клъстърен анализ ще се експериментира разделянето на данните за морски обекти на класове, различни от избраните при настоящите изследвания класове на базата на експертна оценка.

Научно-приложни приноси

Най-важните *научно-приложни приноси*, с характер *обогатяване на съществуващи знания и приложение на научните постижения в практиката,* са представени в следните групи -

получаване и доказване на нови факти, подтвърдителни факти, приноси за внедряване, както следва:

А) Получаване на нови факти, зависимости, изводи и заключения

Решена и изследвана е задача за откриване на знания в данни чрез обучение на класификатори за две съвкупности от данни в различни предметни области:

- за предсказване успеваемостта на студентите, на базата на техни характеристики преди и след приемане в университета, за два варианта на предсказваната променлива (с пет и с две стойности);
- за предсказване на класа на морски обекти, представен като номинална променлива с три стойности, в зависимост от избрани времеви параметри на отразени сигнали от целите.

Б) Получаване на подтвърдителни факти

Получени са подтвърдителни факти за двете изследвани съвкупности от данни:

- Разпределянето на студентите в избраните класове (2 или 5) в най-голяма степен зависи от предсказващите променливи *Бал на приемане, Оценка от приемен изпит, Направление в което са приети студентите, Брой двойки*.
- Точността на предсказване на получените модели за класификация на морски цели е сравнима с резултатите, получени от изследователи в Бирмингамския университет, използвали подобни честотни алгоритми за класификация на движещи се наземни цели.
- Потвърдена е възможността за успешно прилагане на една и съща методика за откриване на знания в данни чрез обучение на класификатори върху две съвкупности от данни в различни предметни области.

В) Приноси за внедряване: методи, конструкции, реализация на по-ефективни средства за приложения

- Апробирана е методика на изследването, включваща предварителна подготовка на данните, обучение на класификатори за откриване на знания в данни, и оценка и сравнение на получените модели, която може да се използва и в други случаи, за извличане на закономерности от:
    - данни за учениците и студентите в различни училища или университети, за нуждите на подобряване на качеството на обучение в Университети и училища.

- данни за радарни и GPS сигнали от различни видове цели - наземни, въздушни и морски.

- Получени са множество резултати от откриване на знания в данни чрез обучение на класификатори за предсказване успеваемостта на студентите, описващи зависимости между успеха на студентите и техните индивидуални особености: пол, приемен изпит, профил и регион на училището, където е получено средно образование, професионално направление, в което се обучават, и др., в табличен и графичен формат, и под формата на изводи.

- Част от получените резултати се използват в проекти:
  - 2010-2013: Проект No.ДДВУ02/50/20.12.2010г. на СУ ”Св. Кл. Охридски” – НИС, финансиран от Фонд „Научни изследвания”, на тема „Разработка на програмна система за изследване и проектиране на радио мрежи, базирани на радиотехники за разпространение на сигнали „напред”, за лоциране на движещи се обекти на фона на море и електронни смущения с цел защита на морски зони и граници”.
  - 2009-2013: Проект No.ДТК 02/28/2009г. на ИИКТ-БАН, финансирани от Фонд „Научни изследвания”, на тема "Откриване, оценка на параметрите на слаби GPS сигнали и подобряване на ефективността на системата чрез подтискане на радиочестотни смущения и намаляване на навигационната грешка".
  - 2012-2013: Университетски проект № НИД НИ 2–1/2012г., финансиран от УНСС-НИД, на тема „Създаване на профил на студента в УНСС чрез средствата на Data Mining”.

Цитирания

Открито е цитиране на труд No.22 от Библиографията на дисертацията, в списание с импакт фактор около 1 (0,878 за 2011, виж www.bioxbio.com/.../IET-R):

- труд N22, Kabakchiev C., D. Kabakchieva, M. Cherniakov, M. Gasninova, V. Behar, I. Garvanov,”Maritime Target Detection, Estimation and Classification in Bistatic Ultra Wideband Forward Scattering Radar”, Intern. Radar Symp. IRS-2011,7-9 Sept., Leipzig, 2011, pp.79-84:

цитиран през 2012г. в списание "IET Radar, Sonar and Navigation"

- Ai, X. , Li, Y. , Wang, X. "Some results on characteristics of bistatic high-range resolution profiles for target classification"

Благодарности

Изследванията в дисертацията, свързани с въпросите за генериране на Data Mining модели за класификация за предсказване успеваемостта на студентите в университета на базата на техни персонални характеристики преди и след приемане в университета, са осъществени с подкрепата и в резултат от дейностите, проведени по Университетски проект № НИД НИ 2–1/2012г., финансиран от УНСС-НИД.

Изследванията в дисертацията, свързани с въпросите за генериране на Data Mining модели за класификация на морски цели по параметрите на FSR сигналите, получени от тях, са осъществени с подкрепата и в резултат от дейностите, проведени по Проект No.ДДВУ02/50/20.12.2010г. на СУ ”Св. Кл. Охридски” – НИС, и Проект ДТК 02/28/2009г. на ИИКТ-БАН, финансирани от Фонд „Научни изследвания”.

Дисертантът изказва благодарност на неговия научен ръководител акад.проф.д.н.Иван Попчев, както и на секция "Интелигентни системи" с ръководител доц.д-р Любка Дуковска.

Списък на Фигурите

Фиг.3.9: Получено *Дърво на решенията* в WEKA за променливата „*Оценка от приемен изпит*”

Фиг.3.10: Получено *Дърво на решенията* в WEKA за променливата „*Общ бал при приемане*”

Фиг.3.11: Разпределението на записите в десетте листа на полученото *Дърво на решенията* в WEKA

Фиг.3.12: Грешка при класификация на записите в десетте листа на полученото *Дърво на решенията* в WEKA

Фиг.3.13: Получено *Дърво на решенията* в WEKA за променливата „*Пол*”

Фиг.3.14: Получено *Дърво на решенията* в WEKA за променливата „*Регион - средно образование*”

Фиг.3.15: Получено Дърво на решенията в WEKA за променливата „Успех от средното образование”

Фиг.3.16: Получено *Дърво на решенията* в WEKA за променливата „*Брой 2-ки*” в графичен формат

Фиг.3.17: Получено *Дърво на решенията* в WEKA за променливата „*Брой 2-ки*” в текстови формат

Фиг.3.18: Резултати от от изследването на индивидуалното влияние на входните променливи върху класификацията

Фиг.3.19: Получен модел за класификация чрез алгоритъма *OneR* в WEKA

Фиг.3.20: Зависимост на големината на полученото *Дърво на решенията* от параметъра M (минимален брой записи в едно листо на дървото)

Фиг.3.21: Полученото *Дърво на решенията* чрез алгоритъма *J48* в WEKA (при M=500)

Фиг.3.22: Получената *Невронна мрежа* чрез алгоритъма *Multilayer Perceptron* в WEKA (при H=1 - брой неврони в скрития слой)

Фиг.3.23: Получената *Невронна мрежа* чрез алгоритъма *Multilayer Perceptron* в WEKA, в текстови формат (при H=1 - брой неврони в скрития слой)

Фиг.3.24: Получената Невронна мрежа чрез алгоритъма Multilayer Perceptron в WEKA (при H(HiddenLayers)=a, a=(attributes+classes)/2)

Фиг.3.25: ROC крива на получената *Невронна мрежа* чрез алгоритъма *Multilayer Perceptron* в WEKA (при H=a)

Фиг.3.26: Трансформиране на категорийните променливи в цифрови

Фиг.3.27: Получената *Невронна мрежа* чрез алгоритъма *Multilayer Perceptron* в WEKA (при цифрови променливи, за H=a)

Фиг.3.28: Сравнение на моделите за класификация, получени чрез алгоритмите *OneR, J48, Multilayer Perceptron* и *IBk* в WEKA, по отношение на *точността на предсказване*

Фиг.3.29: Сравнение на моделите за класификация, получени чрез алгоритмите *OneR, J48, Multilayer Perceptron* и *IBk* в WEKA, по отношение на параметрите Kappa Statistic и *ROC Area*

Фиг.3.30: Сравнение на моделите за класификация, получени чрез алгоритмите OneR, J48, Multilayer Perceptron и IBk в WEKA, по отношение на коефициентите на подобрение на предсказването спрямо случайната класификация

Фиг.3.31: Тестова схема за класификация крос-валидиране („10-fold Cross Validation")

Фиг.3.32: Полученото *Дърво на решенията* чрез алгоритъм *J48* в WEKA

Фиг.3.33. Сравнение на моделите, получени с избраните класификационни алгоритми, по отношение на точността/грешката на предсказване

Фиг.3.34: Сравнение на моделите, получени с избраните класификационни алгоритми, по отношение на точността на класификация по класове

Списък на Таблиците

Таблица 3.15: Резултати за оценка на получената *Невронна мрежа* чрез алгоритъма *Multilayer Perceptron* в WEKA (при H(HiddenLayers)=a, a=(attributes+classes)/2)

Таблица 3.16: Резултати за оценка на получената *Невронна мрежа* чрез алгоритъма *Multilayer Perceptron* в WEKA (при цифрови променливи, за H=1 и H=a)

Таблица 3.17: Матрица на класификация за получената *Невронна мрежа* чрез алгоритъма *Multilayer Perceptron* в WEKA (при цифрови променливи, за H=a)

Таблица 3.18: Резултати за оценка на получения модел за класификация чрез алгоритъма *IBk* в WEKA за различни стойности на параметъра *K* (брой най-близки съседи)

Таблица 3.19: Резултати за оценка на получения модел за класификация чрез алгоритъма *IBk* в WEKA (за *K=50)*

Таблица 3.20: Матрица на класификация за получения модел за класификация чрез алгоритъма IBk в WEKA (за K=50)

Таблица 3.21: Резултати, получени в WEKA с Data Mining алгоритми *OneR, J48, MultiLayer Perceptron* и *IBk* при предсказвана променлива с 2 стойности

Таблица 3.22: Описание на използваните Data Mining алгоритми в WEKA

Таблица 3.23: Матрица на случайна класификация за данните от сферата на FS сигнална обработка

Таблица 3.24: Резултати за оценка на полученото *Дърво на решенията* чрез алгоритъм *J48* в WEKA

Таблица 3.25: Матрица на класификацията за полученото *Дърво на решенията* чрез алгоритъм *J48* в WEKA

Таблица 3.26: Брой правилно/неправилно класифицирани записи за полученото *Дърво на решенията* чрез алгоритъм *J48* в WEKA

Таблица 3.27: Резултати за оценка на получения модел за класификация чрез алгоритъм *JRip* в WEKA

Таблица 3.28: Матрица на класификацията за получения модел за класификация чрез алгоритъм *JRip* в WEKA

Таблица 3.29: Резултати за оценка на получения модел за класификация чрез алгоритъм *SimpleLogistic* в WEKA

Таблица 3.30: Матрица на класификацията за получения модел за класификация чрез алгоритъм *SimpleLogistic* в WEKA

Библиография

1. Antons, C., Maltz, E. (2006). *Expanding the role of institutional research at small private universities: A case study in enrollment management using data mining*. New Directions for Institutional Research, 2006: 69–81. doi: 10.1002/ir.188.
2. Baker, R., Yacef, K. (2009). *The State of Educational Data mining in 2009: A Review and Future Visions*. Journal of Educational Data Mining, Vol.1, Issue 1, Oct. 2009, pp.3-17.
3. Berry, M., Linoff, G. (2004). *Data Mining Techniques: For Marketing, Sales and Customer Relationship Management*. Wiley Publishing Inc.
4. Berson, A., Smith, C., Thearling, K. (1999). *An Overview of Data Mining Techniques* (Excerpted from the book Building Data Mining Applications for CRM). Available at: http://www.thearling.com/text/dmtechniques/dmtechniques.htm
5. Berthold, M., Hand, D. (2007). *Intelligent Data Analysis*. Springer-Verlag Berlin Heidelberg.
6. Bulgarian Ministry of Education and Youth (2010). *Report on the Status Analysis of Research in Bulgaria*. Available at: http://www.minedu.government.bg
7. Cherniakov, M, Gashinova, M, Cheng, H, Antoniou, M, Sizov, V, Daniel, L.Y. (2007). *Ultra wideband forward scattering radar: Concept and prospective*. Proceedings of Int. Conf. Radar 2007, 1-5, (2007).
8. Cherniakov, M., Raja Abdullah, R.S.A., Jancovic, P., Salous, M. (2005). *Forward Scattering Micro Sensor for Vehicle Classification*. Proceedings of the IEEE International Radar Conference, Washington DC, US, pp. 184-189, 2005.
9. Cortez, P., Silva, A. (2008). *Using Data Mining to Predict Secondary School Student Performance*. EUROSIS, A. Brito and J. Teixeira (Eds.), 2008, pp.5-12.
10. Dekker, G., Pechenizkiy, M., Vleeshouwers, J. (2009). *Predicting Students Drop Out: A Case Study*. Conference Proceedings of the 2nd International Conference on Educational Data Mining (EDM'09), 1-3 July 2009, Cordoba, Spain, pp.41-50.
11. Delavari, N., Phon-Amnuaisuk, S., Beikzadeh, M. (2008). *Data Mining Application in Higher Learning Institutions*. Internationa Journal of Informatics in Education, published by the Institute of Mathematics and Informatics, Vilnius, 2008, Vol. 7, No. 1, pp.31–54.
12. DeLong, C., Radclie, P., Gorny, L. (2007). *Recruiting for retention: Using data mining and machine learning to leverage the admissions process for improved freshman retention*. Proceedings of the Nat. Symposium on Student Retention 2007.
13. Gao, H. (2005). *Who Are the Students Who L eft? Answers from the Answer Tree: A First-Year Retention Study from a Private 4-year College*. AIR 2005 Forum, 29 May – 1 June 2005, San Diego.
14. Guidici, P. (2003). *Applied Data Mining: Statistical Methods for Business and Industry*. John Wiley & Sons Ltd.
15. Han, J., Kamber, M. (2000). *Data Mining: Concepts and Techniques*. Simon Fraser University. Morgan Kaufmann Publishers.

16. Hand, D., Mannila, H., Smyth, P. (2001). *Principles of Data Mining*. The MIT Press.
17. Herzog, S. (2006). *Estimating Student Retention and Degree-Completion Time: Decision Trees and neural Networks Vis-à-vis Regression*. New Directions for Institutional Research, No.131, Fall 2006, Wiley Periodicals Inc., pp.17-33.
18. Ibrahim, N.K.,. Raja Abdullah, R.S.A, Saripan, M.I. (2009). *Artificial Neural Network Approach in Radar Target Classification*. Journal of Computer Science 5 (1): pp.23-32, 2009, ISSN 1549-3636.
19. Kabakchiev, C., Garvanov, I., Behar, V., Kabakchiev, A., Kabakchieva, D. (2012). *Forward Scatter Radar Detection and Estimation of Marine Targets*. Conference Proceedings of the International Radar Symposium (IRS) 2012, Warsaw, Poland, pp.79-84.
20. Kabakchiev, C., I Garvanov, M. Cherniakov, M. Gashinova, V. Behar, A. Kabakchiev, V. Kiovtorov (2011). *Bistatic UWB FSR CFAR for Maritime Target Detection and Estimation in Frequency Domain*. Proc. of the International Radar Symposium – IRS'11, Leipzig, Germany, pp.73-78, 2011.
21. Kabakchiev, C., I. Garvanov, V. Behar, M.Cherniakov, M. Gashinova, A. Kabakchiev (2011). *CFAR Detection and Parameter Estimation of Moving Marine Targets using Forward Scattering Radar*, Proc. of the International Radar Symposium – IRS'11, Leipzig, Germany, pp. 85-90, 2011.
22. Kabakchiev, H., Kabakchieva, D., Cherniakov, M., Gashinova, M., Behar, V., Garvanov, I. (2011). *Maritime Target Detection, Estimation and Classification in Bistatic Ultra Wideband Forward Scattering Radar.* Conference Proceedings of the International Radar Symposium (IRS 2011), 7-9 September 2011, Leipzig, Germany, pp.79-84.
23. Kabakchieva, D. (2012). *Predicting Student Performance by Using Data Mining Methods for Classification*. Submitted for Publication in the International Journal of Cybernetics and Information Technologies, Bulgarian Academy of Sciences.
24. Kabakchieva, D. (2012). *Student Performance Prediction by Using Data Mining Classification Algorithms.* International Journal of Computer Science and Management Research, Vol.1, Issue 4, November 2012, pp.686-690. Available at: www.ijcsmr.org
25. Kabakchieva, D., Stefanova, K., Kisimov, V. (2011). *Analyzing University Data for Determining Student Profiles and Predicting Performance.* Conference Proceedings of the 4th International Conference on Educational Data Mining (EDM 2011), 6-8 July 2011, Eindhoven, The Netherlands, pp.347-348.
26. Kabakchieva, D., Stefanova, K., Kisimov, V., Nikolov, R. (2010). *Research Phases of University Data Mining Project Development.* Conference Proceedings of the Second International Conference on Software, Services and Semantic Technologies (S3T), 11-12 Sept. 2010, Varna, Bulgaria, pp.231-239.
27. KDnuggets (2008). Availabele at: www.kdnuggets.com
28. Kotsiantis, S., Pierrakeas, C., Pintelas, P. (2004). *Prediction of Student's Performance in Distance Learning Using Machine Learning Techniques*. Applied Artificial Intelligence, Vol. 18, No. 5, 2004, pp. 411-426.

29. Kotsiantis, S., Pintelas, P. (2004). *A Decision Support Prototype Tool for Predicting Student Performance in an ODL Environment*. International Journal of Interactive Technology and Smart Education (ITSE), Vol. 1, Issue 4, Nov. 2004, pp 253-263.
30. Kovačić, Z. (2010). *Early Prediction of Student Success: Mining Students Enrolment Data*. Proceedings of Informing Science & IT Education Conference (InSITE) 2010, pp.647-665.
31. Kumar, N., Uma, G. (2009). *Improving Academic Performance of Students by Applying Data Mining techniques*. European Journal of Scientific Research, ISSN 1450-216X, Vol.34, No.4 (2009), pp.526-534.
32. Kumar, V., Chadha, A. (2011). *An Empirical Study of the Applications of Data Mining Techniques in Higher Education*. International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 2, No.3, March 2011, pp.80-84.
33. Larose, D. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley & Sons, Inc.
34. Larose, D. (2006). *Data Mining: Methods and Models*. John Wiley & Sons, Inc.
35. Luan, J. (2002). Data *Mining and Its Applications in Higher Education. New Directions for Institutional Research, Special Issue titled Knowledge Management: Building a Competitive Advantage in Higher Education*, Vol. 2002, Iss.113, pp.17–36.
36. Luan, J. (2004). *Data Mining Applications in Higher Education*. SPSS Executive Report, 2004.
37. Luan, L. (2002). *Data Mining and Knowledge management in Higher Education – Potential Applications*. Presentation at AIR Forum, Toronto, Canada, 2002. Available at: www.cabrillo.edu/services/pro/oir_reports/DM_KM2002AIR.pdf
38. Ma, Y., Liu, B., Wong, C. K., Yu, P. S., Lee, S. M. (2000). *Targeting the right students using data mining*. Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, Boston, pp 457-464.
39. Maimon, O., Rokach, L. (2005). *The Data Mining and Knowledge Discovery Handbook* (A Complete Guide for Research Scientists and Practitioners). Springer Science+Business Media, Inc.
40. MicroStrategy Inc. (2007). *Applications of Industrial-Strength Business Intelligence: Seven Leading Applications of Business Intelligence Software*. Available at: http://www.bisource.com
41. MicroStrategy Inc. (2008). *Summary Results from the BI Survey 7*. Available at: http://www.bisource.com
42. Minaeli-Bidgoli, B., Kashy, D., Kortemeyer, G., Punch, W. (2003). *Predicting Student Performance: An Application of Data Mining Methods with the Educational Web-Based System LON-CAPA*. 33rd ASEE/IEEE Frontiers in Education Conference, 5-8 Nov 2003, Boulder, CO
43. Nandeshwar, A., Chaudhari, S. (2009). *Enrollment Prediction Models Using Data Mining*. Available at: http://nandeshwar.info/wp-content/uploads/2008/11/DMWVU_Project.pdf

44. Noel-Levitz White Paper (2008). *Qualifying Enrollment Success: Maximizing Student Recruitment and Retention Through Predictive Modeling*. Noel-Levitz, Inc., 2008.
45. Pardos Z., Heffernan N., Anderson B., and Heffernan C. (2006). *Using Fine-Grained Skill Models to Fit Student Performance with Bayesian Networks*. In Proceedings of the Workshop in Educational Data Mining held at the 8th International Conference on Intelligent Tutoring Systems (ITS2006), June 26, 2006, Taiwan.
46. Pyle, D. (1999). *Data Preparation for Data Mining*. Morgan Kaufmann Publishers Inc.
47. Pyle, D. (2003). *Business Modeling and Data Mining*. Morgan Kaufmann Publishers Inc.
48. Raja Syamsul Azmir Bin Raja Abdulla (2005). *Forward Scattering Radar for Vehicle Classification*. PhD Thesis, July 2005, The University of Birmingham, UK.
49. Ramaswami, M., Bhaskaran, R. (2010). *A CHAID Based Performance Prediction Model in Educational Data Mining*. IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No.1, January 2010, pp.10-18.
50. Ranjan, J., Randjan, R. (2010). *Application of Data Mining Techniques in Higher Education in India*. Journal of Knowledge Management Practice, Vol. 11, Special Issue 1, January 2010. Available at: http://www.tlainc.com/articlsi10.htm
51. Rashid, N., Antoniou, M, Jancovic, P, Sizov, V, Abdullah, R, Cherniakov, M. (2008). *Automatic Target Classification in a Low Frequency FSR Network*. Proc. of EuRAD Conf. 2008, pp. 68-71.
52. Romero, C., Ventura, S. (2007). *Educational Data Mining: A Survey from 1995 to 2005*. Expert Systems with Applications 33, 2007, pp.135-146.
53. SAS Institute Inc. (2003). *Data Mining Using SAS® Enterprise Miner: A Case Study Approach*, Second Edition. SAS Institute Inc., Cary, NC, USA.
54. Shmueli, G., Patel, N., Bruce, P. (2007). *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*. John Wiley & Sons, Inc., Hoboken, New Jersey.
55. Shyamala, K., Rajagopalan, S. (2006). *Data Mining Model for a Better Higher Educational System*. Information Technology Journal 5 (3), 2006, pp.560-564.
56. Shyamala, K., Rajagopalan, S. (2007). *Mining Student Data to Characterize Drop out Feature Using Clustering and Decision Tree Techniques*. International Journal of Soft Computing 2 (1), 2007, pp.150-156.
57. SPSS Inc. (Chapman, P., et al.) (2000). *CRISP-DM 1.0: Step-by-step data mining guide*.
58. Sumathi, S., Sivanandam, S.N. (2006). *Introduction to Data Mining and its Applications*. Springer-Verlag Berlin Heidelberg.
59. Superby, J. Vandamme, J., Meskens, N. (2006). *Determination of Factors influencing the achievement of the first-year university students using data mining methods*. Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems (ITS 2006). Jhongli, Taiwan, pp.37-44.
60. Tan, P., Steinbach, M., Kumar, V. (2006). *Introduction to Data Mining*. Addison-Wesley.

61. Two Crows Corporation (1999). *Introduction to Data Mining and Knowledge Discovery*, Third Edition.
62. Vandamme, J., Meskens, N., Superby, J. (2007). *Predicting Academic Performance by Data Mining Methods*. Education Economics, 15(4), pp405-419.
63. Vercellis, C. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making*. John Wiley & Sons Ltd, Chichester, West Sussex, UK.
64. Vialardi, C., Bravo, J., Shafti, L., Ortigosa, A. (2009). *Recommendation in Higher Education Using Data Mining Techniques*. Conference Proceedings of the 2nd International Conference on Educational Data Mining (EDM'09), 1-3 July 2009, Cordoba, Spain, pp. 190-199.
65. Vialardi, C., Chue, J., Peche, J., Alvarado, G., Vinatea, B., Estrella, J., Ortigosa, A. (2011). *A data mining approach to guide students through the enrollment process based on academic performance*. User Model User-Adap Inter (2011) 21:217–248, Springer Science+Business Media B.V. 2011.
66. Wang, M. (2007). *Using Data Mining Techniques to Predict Student Development and Retention*. Presentation included in the SAS Conference Proceedings: Midwest SAS User Group 2007, 2007-10-28/2007-10-30, Des Moines, Iowa, USA.
67. Witten, I., Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, Elsevier Inc. 2005.
68. Wook, M., Yahaya, Y., Wahab, N., Isa, M., Awang, N., Seong, H. (2009). *Predicting NDUM Student's Academic Performance Using Data Mining Techniques*. Second International Conference on Computer and Electrical Engineering, 2009.
69. Ye, N. (2003). *The Handbook of Data Mining*. Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey.
70. Yu, C., DiGangi, S., Jannasch-Pennell, A., Kaprolet, C. (2010). *A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year*. Journal of Data Science 8(2010), pp.307-325.
71. Yu, C., DiGangi, S., Jannasch-Pennell, A., Lo, W., Kaprolet, C. (2007). *A Data-Mining Approach to Differentiate Predictors of Retention*. Educational Resources Information Center (ERIC), ERIC No.ED496657, February 2007. Available at: http://www.eric.ed.gov

Публикации по дисертацията:

1. Kabakchieva, D. (2012). *Student Performance Prediction by Using Data Mining Classification Algorithms.* International Journal of Computer Science and Management Research, Vol.1, Issue 4, November 2012, pp.686-690. Available at: www.ijcsmr.org
2. Kabakchiev, H., Kabakchieva, D., Cherniakov, M., Gashinova, M., Behar, V., Garvanov, I. (2011). *Maritime Target Detection, Estimation and Classification in Bistatic Ultra Wideband Forward Scattering Radar.* Conference Proceedings of the International Radar Symposium (IRS 2011), 7-9 September 2011, Leipzig, Germany, pp.79-84.

3. Kabakchieva, D., Stefanova, K., Kisimov, V. (2011). *Analyzing University Data for Determining Student Profiles and Predicting Performance.* Conference Proceedings of the 4th International Conference on Educational Data Mining (EDM 2011), 6-8 July 2011, Eindhoven, The Netherlands, pp.347-348.
4. Kabakchieva, D., Stefanova, K., Kisimov, V., Nikolov, R. (2010). *Research Phases of University Data Mining Project Development.* Conference Proceedings of the Second International Conference on Software, Services and Semantic Technologies (S3T), 11-12 Sept. 2010, Varna, Bulgaria, pp.231-239.
5. Kabakchieva, D. (2012). *Predicting Student Performance by Using Data Mining Methods for Classification*. Accepted for Publication in the International Journal of Cybernetics and Information Technologies, Bulgarian Academy of Sciences.
6. Kabakchieva, D., Popchev, I. (2006). *Business Intelligence Initiatives – Valuable Opportunity and Great Challenge*. Conference Proceedings of the International Workshop "Distributed Computer and Communication Networks", 30 Oct–2 Nov 2006, Sofia, Bulgaria, organized by the Russian Academy of Sciences and the Bulgarian Academy of Sciences, pp.254-263.