UNIVERZITET U NIŠU

ELEKTRONSKI FAKULTET

Gabrijela P. Dimić

RAZVOJ METODOLOGIJE ZA OTKRIVANJE ZNANJA U MOODLE SISTEMU ZA UPRAVLJANJE UČENJEM

DOKTORSKA DISERTACIJA

Niš, 2018.

UNIVERSITY OF NIŠ

FACULTY OF ELECTRONIC ENGINEERING

Gabrijela P. Dimić

DEVELOPMENT METHODOLOGY FOR DISCOVERING KNOWLEDGE IN THE MOODLE LEARNING MANAGEMENT SYSTEM

DOCTORAL DISSERTATION

Niš, 2018.

Podaci o doktorskoj disertaciji

Mentor:	Prof. dr Dejan D.Rančić, redovni profesor Univerzitet u Nišu, Elektronski fakultet
Naslov:	Razvoj metodologije za otkrivanje znanja u Moodle sistemu za upravljanje učenjem
Rezime:	Predviđanje uspeha studenata je jedna od tema istraživanja oblasti poznate pod nazivom otkrivanje znanja iz obrazovanih skupova podataka. Cij ove disertacije fokusiran je na utvrđivanje faktora uspešnosti studenata u mešanom okruženju učenja kombinovanom implementacijom metoda za otkrivanje znanja i dubinske analize podataka. Opisana je studija slučaja istraživanja sprovedenog na Visokoj školi elektrotehnike i računarstva u Beogradu. Dostupni podaci o pedagoškom procesu i akademskim rezultatima kursa Računarska grafika omogućili su komparativne analize realizovanih eksperimenata. Ishod istraživanja predstavlja utvrđivanje metodologije otkrivanja znanja iz mešanog okruženja učenja. Konceptualni model predviđanja zasnovan je na realizovanim aktivnostima studenata u okviru Moodle kursa i klasičnog načinu održavanja nastave
Naučna oblast:	Elektrotehničko i računarsko inženjerstvo (Računarstvo i informatika)
Naučna disciplina:	Rudarenja podataka u obrazovnom problemskom domenu
Ključne reči:	mešano okruženje učenja, sistem za upravljanje učenjem, otkrivanje znanja, mašinsko učenje, diskretizacija, izbor obeležja, distribucija, uzorkovanje, korelaciona analiza, modelovanje, zajednička informacija, pravila udruživanja, asocijativna analiza, klasifikacija, predviđanje
UDK:	((004.738.4:159.953)+004.41):004.032.6
CERIF klasifikacija:	T120: Sistemski inženjering, računarska tehnologija
Tip licence Kreativne zajednice:	**CC BY-NC-ND**

Data on Doctoral Dissertation

Mentor:	PhD Dejan D.Rančić, full professor University of Niš, Faculty of Electronic Engineering
Naslov:	Development methodology for discovering knowledge in the Moodle Learning Management System
Rezime:	Prediction of students' success is one of the topics of research of the area known as the discovery of knowledge from educational data sets. The aim of this dissertation is focused on determining the students' success factors in a blended learning environment by combined implementation of methods for knowledge discovery and in-depth analysis of data. A case study study conducted at the School of Electrical and Computer Engineering of Applied Studies in Belgrade has been described. Available data about the pedagogical process and the academic results of the Computer graphics course enabled comparative analyzes of the realized experiments. The outcome of the research is to establish a methodology for discovering knowledge from a blended learning environment. The conceptual model of prediction is based on realized student activities within the Moodle course and the classical way of teaching
Naučna oblast:	Electrical and Computer Engineering (Computer Science)
Naučna disciplina:	Educational Data Mining
Ključne reči:	blended learning enviroment, data mining, machine learning, discretization, features selection, data distribution, oversampling, correlation analysis, modeling, mutual information, association rules, clustering, classification, prediction
UDK:	((004.738.4:159.953)+004.41):004.032.6
CERIF klasifikacija:	T120: Systems engineering, computer technology
Tip licence Kreativne zajednice:	

ZAHVALNICA

Zahvaljujem se mentoru, profesoru dr Dejanu Rančiću, koji me je vodio kroz proces dugogodišnjeg naučnog rada i pružio šansu da sarađujem sa njim. Njegovo znanje, saveti i pomoć omogućili su mi da primenim nova saznanja i realizujem postupak izrade disertacije. Zahvaljujem se profesoru dr Ivanu Milentijeviću i profesoru dr Petru Spaleviću na savetima i pruženoj mogućnosti da dosta naučim od njih.

Koristim priliku da izrazim zahvalnost za svu ljubav i bezuslovnu podršku mojih najdražih koji su bili i ostali izvor moje snage i istrajnosti, sinu Lazaru kome posvećujem ovu disertaciju i suprugu Siniši.

Zahvaljujem se na podršci i pomoći kolegi dr Nemanji Maček, kao i ostalim kolegama na pruženoj pomoći i podršci.

Zahvaljujem se roditeljima koji su verovali u moj uspeh, posebno ocu koji nažalost nije doživeo da vidi završetak disertacije, a koji bi sigurno bio veoma ponosan.

Autor

SADRŽAJ

1. UVOD

Razvoj informaciono - komunikacionih tehnologija uticao je na promenu načina prenosa informacija i znanja. Oblast obrazovanja jedan je od domena na koji su ove promene imale veliki uticaj. Osnova koncepta novih obrazovnih okruženja bazirana je na sistemima za učenje zasnovanim na savremenim tehnologijama. E-učenje (engl. *e-learning*) podrazumeva primenu Web i Internet tehnologija za kreiranje različitih okruženja i materijala [1] što omogućava studentima da kontrolišu kontekst, okvir i dinamiku procesa učenja. U studiji grupe autora [2] razmatrana je činjenica kombinovanja e-učenja sa tradicionalnim načinom odvijanja nastave jer prezentacija didaktičkog materijala putem teksta u kombinaciji sa zvukom i grafikom ima pozitivan efekat i omogućava veću zainteresovanost studenata.

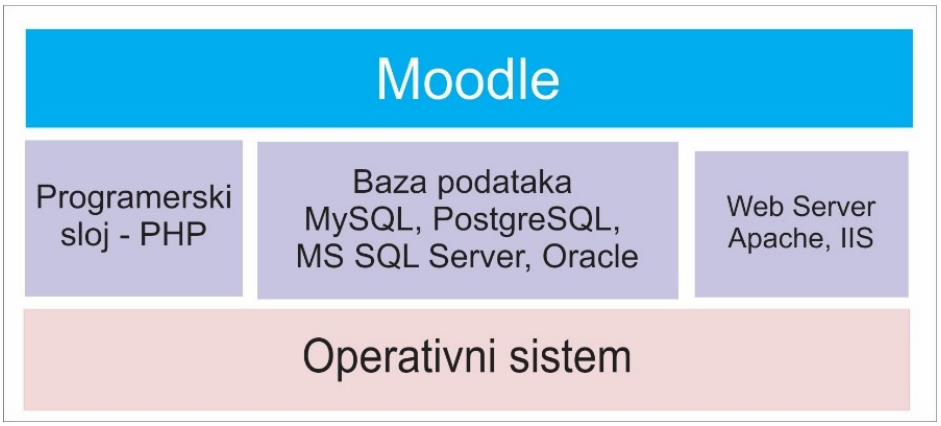
Kontekst kombinovanja takozvanog "*face-to-face*" okruženja i e-sistema za učenje uslovila je pojavu mešanog okruženja učenja (engl. *blended learning environment*). Mešano okruženje učenja integriše fizičke i virtuelne komponente i predstavlja kritičku strategiju za institucije visokog obrazovanja [3]. Efektivnost mešanog učenja prozilazi iz zapažanja da takvi kursevi omogućavaju studentima veću fleksibilnost pristupa.

P. Ramsden navodi [4] da mešano okruženje učenja pruža studentu veću mogućnost izbora što može da dovede do poboljšanja efikasnosti učenja. Oliver i Trigwell sugerišu [5] da mešano okruženje može ponuditi iskustva nedostupna u klasičnim obrazovnim sistemima i da priroda ovih različitih iskustava promoviše učenje. Pored istraživanja koja ukazuju na potencijal mešanog učenja, postoje i studije koje dokazuju da neki mešani kursevi za učenje ne ispunjavaju odgovarajuća očekivanja [6].

Upotreba sistema za učenje omogućila je kreiranje bogate riznice podataka [7]. Sistemi za upravljanje učenjem (engl. *Learning Management System*, *LMS*) podrazumevaju resurse za učenje, softverski aplikativni sloj koji omogućava kreiranje i administraciju kurseva. Jedan od najpopularnijih je sistem otvorenog koda (engl. *open-source*) Moodle (engl. *Modular Object Dynamic Learning Environment*) [8] prvenstveno zato što je besplatan, a osim toga i zbog fleksibilnog načina rada.

Moodle je softverska aplikacija koja se koristi za kreiranje, organizaciju, realizaciju, administriranje kurseva ili programa obuke za jednog ili više korisnika u učionici ili u virtuelnom okruženju.

Troslojna arhitektura Moodle sistema (slika 1.1) uključuje bazu podataka u kojoj se skladište sve sistemske informacije: lične informacije, rezultati, ocene, pristup i upotreba resursa i materijala za učenje, pokušaji testiranja, realizacija zadataka, komunikacija sa ostalim studentima na kursu. Moodle omogućava generisanje određenih izveštaja na osnovu unapred utvrđenih parametara.

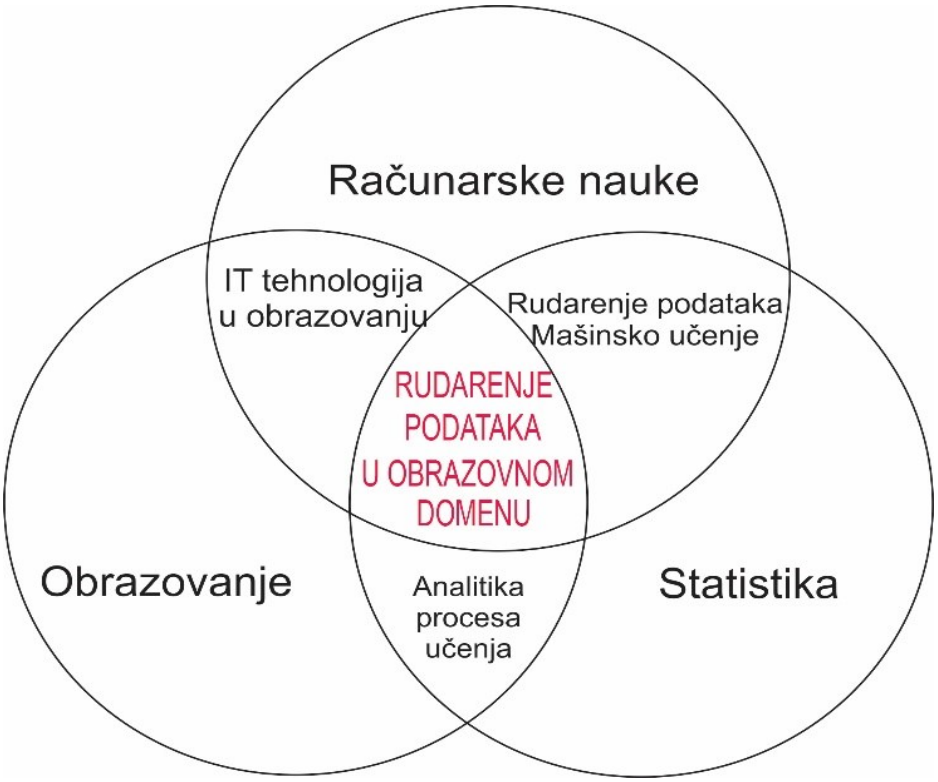


Slika 1.1: Troslojna arhitektura Moodle sistema

U slučaju primene tradicionalne analize, izdvajanje obrazaca ponašanja studenata na kursu, predstavljalo bi težak i dugotrajan posao. Alternativu tradicionalnoj analizi predstavlja proces otkrivanja sakrivenih informacija prisutnih u podacima. Poslednjih godina, istraživači u domenu obrazovanja primenjuju metode rudarenja podataka (engl. *data mining*, *DM*), otkrivanja znanja (engl. *knowledge discovery, KD*) i mašinskog učenja (engl. *Machine Learning*, *ML*) za dubinsku analizu skladištenih podataka, što omogućava otkrivanje informacija od značaja za proces učenja.

DM se definiše kao proces izdvajanja novih, interesantnih, razumljivih informacija ili uzoraka sadržanih u podacima, a sve u cilju donošenja ispravnih poslovnih odluka [9, 10]. Implementacija DM metoda u obrazovanju uslovila je pojavu nove istraživačke oblasti poznate kao rudarenje podataka u obrazovnom problemskom domenu (engl. *Educational Data Mining*, *EDM*).

Na slici 1.2, EDM označava presek više oblasti što omogućava integrisanu implementaciju metoda statistike, mašinskog učenja, data mininga i obrazovanja zasnovanog na primeni računara.



Slika 1.2: Prikaz oblasti usko povezanih sa domenom EDM [11]

Romero i Ventura [12] ukazuju da otkrivanje značajnih informacija iz obrazovnih podataka doprinosi boljem razumevanju procesa učenja i poboljšanju efikasnosti obrazovnih sistema. Baker i Yacef definišu [13] EDM proces kao implementaciju metode klasifikacije, grupisanja, asocijativne analize, izdvajanja značajnih obrazaca, kao i nove metodologije otkrivanja znanja iz psihometrijski modelovanih okruženja.

Primena rudarenja podataka u obrazovnim sistemima može se posmatrati kao neprekidni kružni proces (slika 1.3). Otkriveno znanje treba da na osnovu identifikovane ciklične petlje sistema omogući poboljšanje učenja u celini, ne samo ubacivanjem podataka u znanje nego i izdvajanjem informacija od značaja za donošenje konačne odluke.

Slika 1.3: Kružni proces rudarenja podataka u obrazovnim sistemima [12]

1.1. Metodološke osnove istaživanja

Problem koji je nametnut u ovom istraživanju sadržan je u segmentu oblasti rudarenja podataka koji se bavi razvojem metoda za analizu obrazovnih okruženja u cilju boljeg razumevanja studenata, procesa i okruženja učenja. Imajući u vidu kompleksnost predmeta istraživanja i opširnost materije, definisanost problema istraživanja može se predstaviti primenom određenih matematičkih i statističkih metoda, kao i EDM metodama klasifikacije i asocijativne analize, mašinskog učenja.

1.1.1. Predmet i cilj istraživanja

Predmet istraživanja ove disertacije odnosi se na proces rudarenja podataka mešanog okruženja učenja za studiju slučaja kursa Računarska grafika. Obrazovni sistem zasnovan na konceptu mešanog učenja podrazumevao je kreiranje online kursa na Moodle platformi. Korisničko okruženje kursa prilagođeno je implementiranom konceptu podrške i poboljšanju efikasnosti klasične nastave. Klasični oblik nastave obuhvatao je tri časa predavanja i dva časa laboratorijskih vežbi nedeljno. Na predavanjima su objašnjavani pojmovi i koncepti iz oblasti računarske grafike, a na vežbama su studenti samostalno realizovali praktične zadatke.

U toku semestra, studenti su polagali dva kolokvijuma preko kojih je vršena provera znanja iz oblasti tehnologija ulaza, izlaza, geometrijskih transformacija, osnovnih rasterskih algoritama. Završni test je obuhvatao praktične zadatke iz oblasti OpenGL biblioteke (engl.

Open Graphics Library) i održavao se u terminu ispita. Pored zvaničnih testova, bodovana je aktivnost studenata na laboratorijskim vežbama i učešće u diskusijama na predavanjima. U okviru Moodle kursa, studentima su bili dostupne lekcije, tutorijali, diskusije na forumima, privatne poruke za elektronske konsultacije, probni testovi za samotestiranje kao i zvanične provere znanja. Pristup kursu, izbor aktivnosti, upotreba resursa, trajanje sesije kao i ostale akcije u potpunosti su bile indivudualno određene od strane svakog studenata kursa. Podaci za istraživanje izdvojeni su iz više distribuiranih izvora: baze podataka Moodle sistema za upravljanje učenjem, informacionog sistema obrazovne ustanove, Google dokumenta za praćenje aktivnosti studenata u okviru klasične nastave.

Cilj istraživanja je bio da se razvije metodologija otkrivanja značajnih obrazaca korelacija iz prikupljenih podataka i kreiranje preciznog modela predviđanja za slučaj obrazovnog okruženja gde studenti mogu samostalno, po sopstvenom nahođenju, da biraju aktivnosti koje će da realizuju u toku procesa učenja.

Disertacija je imala zadatak da odgovori na sledeća pitanja:

- koji su postupci neophodni za pripremu i obradu skupa izdvojenog iz više obrazovnih distribuiranih skladišta podataka;
- na koji način je moguće ostvariti bolju tačnost i preciznost predviđanja performansi studenata zasnovanu na samostalno izabranim aktivnostima i upotrebljenom materijalu u mešanom okruženju učenja.
- koliki je dobitak u performansama ostvarenog uspeha studenata u slučaju primene predloženog koncepta otkrivanja značajnih obrazaca u procesu učenja studenata.

Svrha disertacije je da informiše stručnjake u oblasti informacionih i obrazovnih tehnologija o značaju i pravilnim načinima implementacije metoda rudarenja podataka i statističke analize na različite obrazovne sisteme i okruženja.

Naučni cilj disertacije je uslovljen naučnom deskripcijom istaživanja sa elementima statističke analize, mašinskog učenja, rudarenja podataka i analize sistema za upravljanje učenjem. Naučna opravdanost ovog istraživanja leži u sve većoj potrebi implementacije

informacionih tehnologija u oblasti obrazovanja sa ciljem realizacije novih okruženja za učenje.

1.1.2. Istraživačke hipoteze

Osnovna hipoteza disertacije odnosi se na mogućnost izdvajanja značajnih korelacija između relizovanih aktivnosti Moodle kursa mešanog okruženja učenja i ostvarenog uspeha studenta.

Upotrebljene su sledeće radne hipoteze:

- Kako se mogu koristiti metode rudarenja podataka i mašinskog učenja za kombinovanje karakteristika mešanog okruženja učenja i predviđanje uspešnosti studenata?
- Da li izbor karakteristika utiče na tačnost predviđanja?
- Koji faktori mogu biti integrisani u model predviđanja kako bi se obezbedila pouzdana mera procene performansi studenata?

1.1.3. Metode istraživanja i tok istraživačkog postupka

Način istraživanja u disertaciji uslovljen je predmetom istraživanja. Metod istraživanja baziran je na osnovu prethodnih teorijskih i novih naučnih saznanja iz oblasti rudarenja podataka, mašinskog učenja, statistike i praktičnih znanja o modelima elektronskog obrazovanja. Istraživanje je zasnovano na kombinaciji različitih istraživačkih metoda:

- osnovne metodičke,
- hipotetičko-deduktivne i
- komparativne statističke metode.

Prevashodno su korišćeni naučni izvori podataka. Primarni izvori koji su upotrebljeni su relevantni, eksperimentalno potvrđeni autorski radovi, u kojima su obrađene studije primene metoda mašinskog učenja i rudarenja podataka na koncept daljinskog učenja realizovanog na Moodle platformi. Većina analiziranih autorskih radova se bavila identifikovanjem faktora uspeha na osnovu akademskih, socijalnih i demografskih karakteristika studenata.

Istraživanje sprovedeno za potrebe ove disertacije karakteriše analiza koncepta mešanog okruženja učenja koje je realizovno kombinovanjem klasične nastave (predavanja i

laboratorijske vežbe) i Moodle kursa. Moodle kurs je kreiran u svrhu podrške tradicionalnom načinu nastave što je podrazumevalo da materijali za učenje (resursi, aktivnosti) budu dostupni studentima za upotrebu po sopstvenom načinu izbora.

U toku istraživanja izvršene su sledeće analize:

- kritička analiza relevantnih autorskih radova u kojima su opisane metode za kreiranje klasifikatorskih modela predviđanja u različitim obrazovnim okruženjima,
- analiza metoda i tehnika pogodnih za transformaciju i izbor optimalnog vektora ulaznih obeležja
- komparativna analiza performansi klasifikatora implementiranih na oskudni obučavajući obrazovni skup podataka.

1.2. Očekivani rezultati istraživanja i naučni doprinos

Osnovni cilj istraživanja je da se potvrdi da je predloženim metodama moguće značajno povećati performanse klasifikatora i formirati sistem visoke osetljivosti koji efikasno detektuje sve kategorije mešanog okruženja učenja, uključujući i kategorije koje su u obučavajućem skupu opisane malim brojem instanci.

U ovoj disertaciji očekivani su sledeći doprinosi:

- eksperimentalna potvrda da sprovođenje deskriptivne statističke analize orginalnog ulaznog skupa mešanog okruženja omogućava uvid u raspodelu domena vrednosti i način transformacije numeričkih obeležja.
- eksperimentalna potvrda da predloženi model predviđanja uključuje pristup personalizacije procesa učenja studenata zasnovan na definisanom klasnom više-dimenzionalnom obeležju.
- sintezom sistema zasnovanog na razvijenoj metodologiji moguće je izdvojiti obrasce od značaja Moodle kursa i identifikovati korelacije između određenih aktivnosti studenata i ostvarenog uspeha.

1.3. Pregled sadržaja po poglavljima

U uvodnom poglavlju dat je kratak uvod u područje i problem koje se istražuje. Navode se predmet, cilj, metodologija i očekivani rezultati istraživanja.

U **poglavlju 2** opisan je pojam e-učenja i sistema za upravljanje učenjem. Predstavljen je značaj rudarenja podataka u obrazovnim sistemima. Dati su osnovni koncepti razvoja istraživačke oblasti *Educational Data Mining* i opisane metode koje će se koristiti u samom radu.

U **poglavlju 3** je dat kritički osvrt i komparativna analiza potvrđenih autorskih radova u oblasti otkrivanja znanja, mašinskog učenja i primene najznačajnijih tehnika i metoda na obrazovne podatke.

U **poglavlju 4** opisan je postupak deskriptivne analize ulaznog skupa kreiranog integracijom podataka izdvojenih iz Moodle baze podataka, Google Excel dokumenata i informacionog sistema obrazovne ustanove. Izračunate su mere varijabilnosti i centralne tendencije. Analizirana je raspodela, distribuiranost i tip izdvojenih podataka.

U **poglavlju 5** prikazan je detaljan postupak faze predprocesiranja podataka. Analizirane su nadzirane (engl. *supervised*) i nenadzirane (engl. *unsupervised*) metode diskretizacije ulaznih i klasnog obeležja i metode izdvajanja optimalnog vektora obeležja (engl. *feature selection*). Prikazana je primena tehnika uzorkovanja (engl. *oversampling*) za obradu asimetrične raspodele podataka u obučavajućem skupu. Izvršena je analiza međusobnog uticaja obeležja primenom tehnike korelacione analize (engl.*correlation analysis*) i mere zajedničke informacije (engl. *mutual information*). Utvrđen je modul faze predprocesiranja za obučavajući skup podataka studije slučaja mešanog modela učenja.

U **poglavlju 6** razmatrane su klasifikacijske metode nadziranog učenja. Utvrđeni su osnovni principi izbora kandidata klasifikatora za studiju slučaja mešanog okruženja učenja. Opisani su postupci procene i evaluacije klasifikatora. Izvršena je komparativna analiza ostvarenih performansi u sprovedenim eksperimentima.

Predloženi model je zasnovan na integrisanju klasifikatora Naivnog Bajesa (engl. *Näive Bayes, NB)*, Skrivenog Naivnog Bajesa (engl. *Hidden Näive Bayes, HNB)*, J48 stabla odluke (engl. *J48 Decision Tree, J48DT*) i Slučajnih šuma (engl. *Random forest, RF)* u ansambl, primenom mehanizma većinskog glasanja (engl. *Majority Vote, MV*). Problem disbalansa podataka za klasno više-dimenzionalno obeležje rešen je primenom funkcije ponovnog uzorkovanja (engl. *Resample*). Ostvareno je značajno poboljšanje pouzdanosti i stabilnosti modela predviđanja, postignuta je tačnost od 98,94%, a greška klasifikacije smanjena na 1,06%.

U **poglavlju 7** je prikazana asocijativna analiza podataka izdvojenih iz Moodle pripremnih testova i testova za proveru znanja implementacijom Apriori algoritma. Cilj sprovedenog istraživanja je otkrivanje odnosa i veza između odgovora studenata na pitanja iz pripremnih i testova za proveru znanja, načina rešavanja i ostvarenih rezultata. Procena značaja pravila izvršena je primenom objektivnog i subjektivnog pristupa. Izdvojena značajna pravila nastavniku omogućavaju da bolje sagleda koncepte kreiranih testova i odluči na koji način treba da izvrši izmene i unapređenje sa ciljem poboljšanja rezultata testiranja.

U **poglavlju 8** izvršena je analiza rezultata sprovedenih eksperimenata, date su smernice za dalja istraživanja i rezime doprinosa. Potvrđena je osnovna hipoteza da realizacija faze pred-procesiranja oskudnog obrazovnog obučavajućeg skupa mešanog okruženja učenja podrazumeva primenu odgovarajućih metoda za izbor optimalnog vektora obeležja i tehniku uzrokovanja za problem iskrivljene distribucije i da se na taj način omogućava formiranje modela sa većom tačnošću previđanja. Izdvojen je ostvaren naučni doprinos rada u oblasti obrazovnog otkrivanja znanja i rudarenja podataka izdvojenih iz distribuiranih izvora. Razvijen je postupak za izgradnju preciznih klasifikatora za oskudni obučavajući skup obrazovnih podataka, utvrđene su najefikasnije metode za rangiranje i selekciju obeležja i ostvarena personalizacija Moodle kursa u mešanom okruženju učenja.

2. TEORIJSKE OSNOVE ISTRAŽIVANJA

E-učenje (engl. *e-learning*) sve više nalazi primenu u formalnom tipu obrazovanja (osnovno, srednje, visoko), kao i u neformalnom (učenje uz rad, prekvlalifikacija za nova zanimanja i slično). Može se definisati kao bilo koja upotreba komunikaciono-informacionih tehnologija za kreiranje različitih korisničkih okruženja i materijala za učenje [1]. Studija autora [2] ukazala je na činjenicu da tradicionalni način realizacije nastave u kombinaciji sa okruženjem e-učenja omogućava veću zainteresovanost studenata i ima pozitivan efekat na ostvarene rezultate.

Poslednjih godina razvijen je veći broj sistema za učenje na daljinu zasnovanih na savremenim komunikacionim tehnologijama i web servisima što predstavlja osnovu koncepta savremenih tehnologija učenja. LMS sistem za upravljanje učenjem omogućava distribuciju informacija i komunikacija između učesnika. Generalno, ovi sistemi predstavljaju softverske aplikacije koje se koriste za kreiranje, organizaciju, realizaciju, administriranje, kreiranje izveštaja i dokumentacije jednog ili više e-kurseva, programa obuke za jednog ili više korisnika u učionici ili u virtuelnom okruženju. Obezbeđuju bazu podataka u kojoj se skladište sve sistemske i lične informacije korisnika. Omogućavaju samoautentifikaciju, pristup sadržajima za učenje, jednostavno postavljanje dokumenata, vođenje diskusija, testiranje, anketiranje, snimanje ocena i sl. LMS sistemi akumuliraju veliku količinu podataka o registrovanim korisnicima i sve prijave na sistem. Arhiviraju se aktivnosti studenta kao što su pristup i upotreba materijala za učenje, pristupanje testovima, obavljanje raznih zadataka, pa čak i komunikacija sa ostalim studentima na kursu [7]. Na osnovu unapred definisanih ulaznih parametara, generišu više vrsta izveštaja. Obzirom na način prikaza i količinu podataka, postupak analiziranja generisanih izveštaja i izdvajanja korisnih informacija predstavlja izuzetno težak i dugotrajan posao. Zbog toga su neophodni alati koji bi LMS sistemima pomogli da lakše izvršavaju navedene zadatke. Iako neke platforme za elektronsko učenje nude specifične alate za kreiranje izveštaja, kada postoji veliki broj studenata i dalje je postupak izdvajanja korisnih informacija komplikovan.

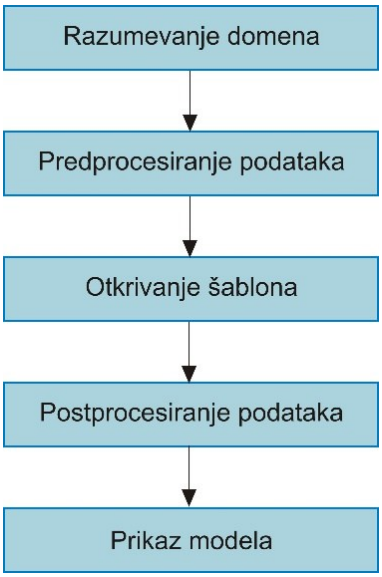
Da bi se kreirao efektivan sistem za realizaciju elektronskog učenja, neophodno je posmatrati studenta kao najvažniji deo paradigme. Elektronsko učenje podrazumeva aktivnu strategiju koja omogućava studentima da kontrolišu kontekst, okvir i dinamiku odvijanja procesa učenja. Sa sigurnošću se može tvrditi da se najbolje performanse u ovom procesu postižu u slučaju kada postoje prethodni podaci o predznanju, iskustvu i korišćenju kursa od strane korisnika. Kreiranje i razvoj kursa za sistem elektronskog učenja podrazumeva izbor i organizaciju sadržaja koji će biti prikazan u audio-vizuelnoj formi, formiranje strukture kursa, kao i identifikaciju odgovarajućih resursa namenjenih za različite tipove potencijalnih korisnika kursa. Generalno rečeno, dizajn kursa, bilo da se odnosi na strukturu, sadržaj, navigaciju ili korisnički interfejs, u većini slučajeva, podložan je stalnom unapređivanju i modifikaciji zasnovanoj na kontinualnom empirijskom razvojnom pristupu, praćenjem studentskih aktivnosti [14].

Moodle je tipičan primer LMS sistema otvorenog koda koja stiče sve veću popularnost, jer omogućava kreatorima kursa da efikasno organizuju on-line okruženje za učenje [15]. Modularan dizajn omogućava kreiranje kurseva koji mogu da podrže način učenja pod nazivom socijalna konstruktivna pedagogija [8]. Ovaj način učenja smatra da je proces učenja najefikasniji ukoliko je student u interakciji sa materijalom učenja i sa ostalim učesnicima kursa. Moodle ne zahteva eksplicitno upotrebu ovog stila prilikom kreiranja kurseva ali je ovaj način učenja najbolje podržan. Sadržaj u Moodle sistemu može da predstavlja materijal za učenje, proveru znanja ili aktivnost. Da bi se kreirao incijalni model, kreatori kursa treba da uspostave pravila koja povezuju stilove učenja i različite objekte sadržaja.

Obzirom na modularan dizajn Moodle sistema, kreator kursa ima na raspolaganju skup fleksibilnih resursa prilikom organizacije sadržaja i strukture kursa i to: statički materijal (tekstualne stranice, HTML stranice, vezu ka određenom dokumentu ili stranici na Web-u, ZIP arhive, IMS paketi), interaktivan materijal (zadaci, lekcije, izbor, testovi, upitnik), aktivnosti u kojima su studenti u međusobnoj interakciji (pričaonica, forum, rečnik, wiki, radionice, baze podataka). Moodle vodi detaljan zapis o svim aktivnostima i pristupima učesnika bilo da su studenti ili nastavnici. Kada se korisnik prijavi na sistem, registruje se svaki navigacijski klik koji izvrši korisnik i informacije se skladište u bazi podataka u odgovarajućoj tabeli. Kreator kursa ima mogućnost prikaza izveštaja u formi HTML strane o svim aktivnostima učesnika kao i mogućnost uključivanja određenih filtera. Generisani izveštaji se mogu koristiti da se utvrde aktivni učesnici, aktivnosti kao i vreme pristupa učesnika. Kada su u pitanju testovi, u

izveštajima su pored rezultata i utrošenog vremena po testu, na raspolaganju i detaljne analize odgovora na postavljena pitanja za svakog učesnika kao i analiza odgovora po pojedinačnom pitanju. Nastavnici mogu lako dobiti potpun izveštaj o svim aktivnostima za svakog pojedinačnog učesnika kao i potpun izveštaj o svim učesnicima za određene aktivnosti. Dostupni su izveštaji za svaku aktivnost studenta, detalji o svakom modulu (o poslednjem pristupu, broju čitanja sadržaja modula), kao i detaljan kompletan izveštaj o angažovanosti svakog studenta u toku semestra. Ovo može da bude od posebne koristi nastavniku kada želi da proveri da li su učesnici kursa uradili postavljen zadatak ili su proveli vreme učestvujući u nekim drugim aktivnostima. Tradicionalna analiza podataka u sistemima za e-učenje je hipoteza ili je vođena pretpostavkama. To znači da korisnik na osnovu polaznog pitanja istražuje podatke da bi potvrdio svoje opažanje i intuiciju [16]. Alternativa tradicionalnoj analizi podataka je dubinska analiza podataka kao induktivan pristup za automatsko otkrivanje skrivenih informacija prisutnih u podacima. Za razliku od tradicionalne analize, zasniva se na otkrivanju značajnih informacija iz podataka. To znači da se pretpostavka automatski izvodi iz podataka, odnosno da je vođena podacima, pre nego što je zasnovana na istraživanju ili je vođena čovekom. [17].

Otkrivanje znanja (engl. *knowledge discovery, KD*) u podacima predstavlja disciplinu primenljivu u raznim oblastima istraživanja, koja za cilj ima otkrivanje zanimljivih zakonitosti, obrazaca i koncepata u podacima. Otkrivanje znanja je interaktivan i iterativan proces koji se može predstaviti dijagramom prikazanim na slici 2.1.



Slika 2.1: Proces otkrivanja znanja

Obuhvata različite tehnike iz oblasti mašinskog učenja (engl. m*achine learning, ML*), rudarenja podataka (engl. *data mining, DM*) i statistike.

Prva faza podrazumeva razumevanje domena problema sa aspekta određivanja potreba, ciljeva i ograničenja, kao i utvrđivanja vrste i tipa dostupnih podataka.

Faza predprocesiranja obuhvata izbor odgovarajućih izvora i integraciju distribuiranih podataka, izdvajanje karakteristika iz nestruktuiranih podataka, otklanjanje nepravilnosti u podacima. Obzirom da obrazovni podaci mogu da budu u različitim formatima, postupak pripreme može da predstavlja dugotrajniji proces. Faza predprocesiranja je od izuzetne važnosti obzirom da ima značajan uticaj na krajnji rezultat sprovedenog procesa.

Treća faza otkrivanja obrazaca obuhvata implementaciju odgovarajućih DM metoda ili algoritama mašinskog učenja.

U fazi postprocesiranja izdvajaju se otkriveni značajni obrasci i prezentuju krajnjem korisniku.

U fazi prikaza modela, u okviru domena analiziranog problema, implementiraju se generisani rezultati, a proces izvršavanja ponovo započinje sa novim podacima.

U literaturi se naizmenično koriste pojmovi rudarenje podataka, mašinsko učenje i otkrivanje znanja. Često se pod procesom otkrivanja znanja podrazumeva rudarenje podataka, a mašinsko učenje se posmatra kao disciplina u okviru rudarenja podataka. Rudarenje podataka označava postupak identifikovanja validnih, novih, potencijalno korisnih i razumljivih obrazaca u podacima [18]. Može se posmatrati kao metodologija za automatizaciju procesa analize i istraživanja podataka sa ciljem otkrivanja novih informacija, obrazaca i odnosa. Analitičke metode koje se koriste su matematičke tehnike i algoritmi izvedeni iz statistike, mašinskog učenja i baza podataka [19].

Proces rudarenja podataka sastoji se od sledećih faza:

- sakupljanje podataka za analizu,
- priprema i predprocesiranje podataka,
- primene algoritama za rudarenje podataka,
- tumačenje i vrednovanje rezultata.

Sistematizacija u podeli na prediktivno i opisno modelovanje podudara se sa podelom mašinskog učenja na nadgledano i nenadgledano učenje. Prednost ovog pristupa ogleda se u sledećem:

- Prediktivni i opisni modeli međusobno su povezani.
- Metodama opisnog modelovanja se otkrivaju osnovni obrasci iz podataka i utiče na izbor adekvatne paradigme za proces prediktivnog modelovanja.
- Podela na prediktivno i opisno modelovanje daje smernice pri izboru tehnika za validaciju rezultata.
- Prediktivni i opisni modeli imaju različite postavke generisanja dobrog modela što utiče na funkciju pretrage rezultata.
- Primena prediktivnih modela u praksi omogućava prikupljanje podataka za nove deskriptivne modele.

Tehnike rudarenja podataka se mogu primeniti na širok spektar podataka uključujući baze podataka, prostorne podatake, multimedijalne podatake, Web orjentisane podatke i kompleksne objekte. Ove tehnike su, takođe, često korišćene kod komercijalnih sajtova za identifikaciju ključnih kupaca i povećanje efikasnosti online prodaje komercijalnih sajtova. Navedeni aspekti mogu biti prevedeni u oblast obrazovanja i sistema za upravljanje učenjem.

2.1. Tehnike rudarenja podataka u obrazovnim sistemima

Primena data mining tehnika u obrazovnim sistemima podrazumeva integraciju procesa za otkrivanje informacija sa razvojem i upotrebom metoda za izdvajanje značajnih obrazaca i korisnih znanja iz različitih tipova podataka. Rudarenje podataka u obrazovnom problemskom domenu (engl. *Educational Data Mining*, *EDM*) predstavlja novu istraživačku oblast koja se bavi pronalaskom prethodno nepoznatih značajnih obrazaca u cilju poboljšanja performansi procesa učenja. [20].

Zajednica *Educational Data Mining* definiše EDM na sledeći način: *"EDM je disciplina u nastajanju koja se bavi razvojem metoda za upoznavanje jedinstvene vrste podataka koje dolaze iz obrazovnih sredina i upotrebu tih metoda za bolje razumevanje studenata i okruženja za učenje "*[21].

Koristi tipične metode rudarenja podataka kao i nove metodologije za integraciju sa psihometrijski modelovanim okruženjem [13]. Sa stanovišta nastavnika, otkriveno znanje se može koristiti za uspostavljanje postupka odlučivanja kojim će se modifikovati proces nastave i nastavnih planova u smislu unapređenja efikasnosti procesa učenja i razumevanja ponašanja studenata. Sa stanovišta studenta, otkriveno znanje se može koristiti kao značajna smernica u poboljšanju i kreiranju individulanih okruženja učenja.

Autori [12] predstavljaju model za dubinsku analizu obrazovnih podataka koji pokazuje da je primena tehnika rudarenja u obrazovanju irevirzabilni ciklus postavljanja, testiranja hipoteze i procesiranja. Autori su zaključili da otkriveno znanje treba koristiti za olakšanje i poboljšanje kompletnog procesa učenja. Obrazovni podaci potrebni za dubinsku analizu u osnovi predstavljaju lične i akademske podatke o studentima. Moguće je prikupiti i dodatne podatke primenom metode anketiranja u okviru obrazovnog sistema [22]. Međutim, ukoliko se uzme u obzir specifičnost različitih modela strategije i načina odvijanja procesa učenja u obrazovnim sistemima za upravljanje učenjem, podaci prikupljeni putem anketa neće generisati potpune modele procene. Za primenu rudarenja podataka neophodne su informacije o realizovanim aktivnostima studenata zabeležene u dnevničkoj datoteci (engl. *log*) na Web serveru ili bazi podataka sistema. Data mining koristi napredne tehnike za otkrivanje obrazaca ponašanja iz postojećih podataka koje se mogu prilagoditi ispitivanju efikasnosti sistema za e-učenje [23].

U zavisnosti od svrhe i izvora podataka, može se primeniti veliki broj različitih metoda rudarenja podataka [19], a najčešće upotrebljivane su:

- Klasifikacija (engl. *classification*)– razdvajanje događaja ili korisnika u unapred definisane kategorije.
- Pravila udruživanja (engl. *association rules*) – otkrivanje obrazaca u kojima je jedan događaj povezan sa drugim događajem.
- Grupisanje, tj klasterovanje (engl. *clustering*)– grupisanje korisnika ili događaja bez prethodnog znanja kategorija.
- Predviđanje (engl. *prediction*) - otkrivanje uzoraka u podacima koji mogu dovesti do razumnog predviđanja o budućnosti.

Uspešnost procesa rudarenja podataka zasnovana je na detaljno sprovedenoj fazi predprocesiranja u cilju kreiranja što efikasnije pripremljenog skupa podataka za analizu. Ova faza podrazumeva pripremu obučavajućeg skupa u odgovarajuću formu, obradu šuma u podacima, odnosno, instanci sa pojavom nepravilnosti (engl. *noise*), nedostajućih vrednosti (engl. *missing values*), izuzetaka (engl. *outliers*), izbor relevantnog podskupa obeležja (engl. *feature selection*), transformaciju kontinulanih obeležja primenom odgovarajućih metoda diskretizacije (engl. *discretization*). U radu [24] autori predstavljaju slučaj otkrivanja znanja iz Moodle sistema za upravljanje učenjem primenom metoda rudarenja podataka. Opisana faza predrocesiranja obuhvata aktivnosti kreiranja sumarne tabele koja u jednom zapisu sadrži informacije o aktivnostima studenta na predmetnom Moodle kursu; diskretizaciju izdvojenih podataka; transformisanje diskretizovanih podataka u odgovarajući tip vrednosti za primenu algoritama.

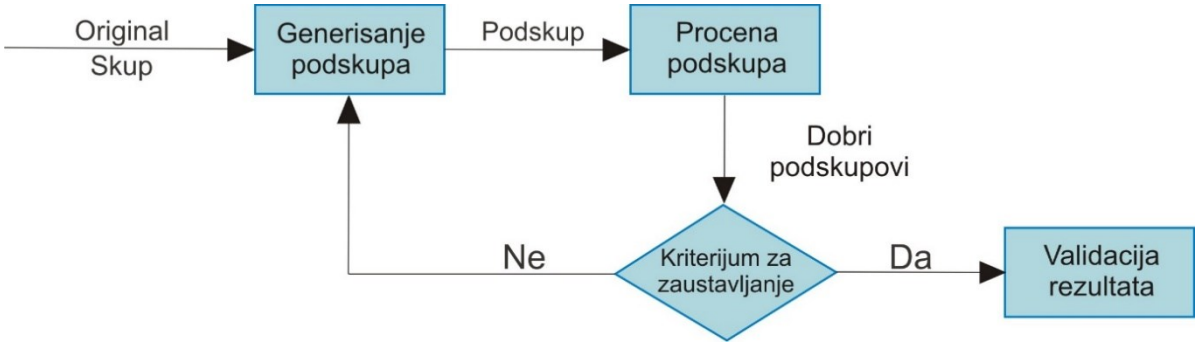
2.2. Izbor obeležja

Izbor obeležja (engl. *feature selection*) predstavlja jednu od značajnijih metoda faze predprocesiranja u procesu otkrivanja znanja [25]. Primenom odgovarajuće tehnike za analizirani skup podataka, uklanjaju se nebitna i suvišna obeležja. Na taj način se ubrzava generisanje klasifikacijskih algoritama, postiže veća tačnost predikcije, razumljivost rezultata i poboljšavaju performanse modela. Izbor obeležja zasniva se na rešavanju problema izdvajanja optimalnog vektora obeležja iz orginalnog izdvojenog skupa za analizu, potrebnih i dovoljnih da opišu ciljni koncept [26]. Pronalazak optimalnog podskupa je složen i težak postupak koji nosi sa sobom i veliki broj problema. Generalno, postupak izdvajanja podskupa značajih obeležja može se predstaviti pomoću četiri osnovna koraka [27]:

1. generisanje podskupa,
2. evaluacija izdvojenog podskupa,
3. kriterijum zaustavljanja i
4. validacija rezultata.

Na slici 2.2 prikazan je proces izbora obeležja ulaznog vektora. Generisanje podskupa bazira se na odgovarajućoj strategiji pretrage [28] koja proizvodi kandidate za procenu. Svaki podskup kandidata se procenjuje i poredi sa najboljim prethodnim odgovarajućim metodama evaluacije. Ako je novi podskup bolji, zamenjuje prethodno generisani. Proces se ponavlja sve

dok se ne zadovolji kriterijum zaustavljanja. Vrednovanjem izabranog najboljeg podskupa utvrđuje se značajnost obeležja, zanemarujući njihove međusobne moguće interakcije. Metode vrednovanja zasnovane su na statistici, teoriji informacija ili na funkcijama izlaza klasifikacije.



Slika 2.2: Proces izbora obeležja

Autori u radu [27] prikazuju prednosti i slabosti primenjenih metoda za izdvajanje optimalnog skupa obeležja. Primena odgovarajuće metode zasnovana je na postojanju različitih tipova podataka u analiziranom skupu, više-dimenzionalnosti klasnog obeležja, prisustvu nepravilnosti među podacima, a usmerena je ka poboljšanju prediktivne tačnosti. Metode izbora podskupa značajnih obeležja primenjuju se u raznim oblastima istraživanja kao što su: statističko prepoznavanje oblika [29], mašinsko učenje [30], [31], [32], kategorizacija teksta [33], [34], prepoznavanje slike [35], detekcija upada [36].

Prema opisu prirode mere za procenu vrednosti optimalnog vektora, algoritmi za izbor obeležja se mogu podelite u tri kategorije [37]: filter modeli [27]. [38], [39], [40], modeli omotača [41] i hibridni modeli [42].

Filter modeli se oslanjaju na opšte karakteristike podataka za procenu obeležja bez uključivanja algoritma učenja. Za skup podataka D, filter algoritam započinje pretragu incijalizacijom podskupa S_0 (prazan skup, pun skup ili nasumično izabran podskup) i pretražuje prostor vrednosti obeležja primenom određene strategije pretraživanja. Svaki generisan podskup S procenjuje se na osnovu nezavisne mere i poredi sa prethodnim najboljim. Ako se utvrdi da je bolji, smatra se trenutno najboljim podskupom. Pretraga se nastavlja sve dok se ne zadovolji prethodno definisan kriterijum za zaustavljanje. Izlaz algoritma je poslednji najbolji podskup i predstavlja konačan rezultat. Promenom strategije pretrage i mere za procenu moguće je realizovati različite algoritme u okviru filter modela.

Modeli omotača (engl. *wrapper)* modeli zahtevaju unapred definisan algoritam koristeći njegove performanse kao kriterijum procene i validacije. Ovi algoritmi izdvajaju onaj skup obeležja koji doprinosi poboljšanju performansi generisanog modela. Postupak izvršavanja *wrapper* algoritma je poprilično sličan proceduri izvršavanja filter algoritma osim što koristi unapred definisan algoritam rudarenja podataka umesto nezavisne mere za ocenjivanje podskupa obeležja. Primena različitih algoritama će proizvesti i različite podskupove. S obzirom da je izabrani podskup pogodan za unapred definisan algoritam, wrapper model ima tendenciju obezbeđivanja superiornijih performansi ali i mogućnost povećanja troškova u slučaju velikog broja obeležja obučavajučeg skupa.

Hibridni modeli nastoje da iskoriste prednosti pomenuta dva modela izbegavanjem unapred određenog kriterijuma zaustavljanja u različitim fazama pretrage i obezbeđivanjem rada sa visoko-dimenzionalnim podacima [43]. Tipični hibridni algoritam koristi nezavisnu meru i algoritam rudarenja za procenu podskupa. Nezavisnu meru koristi za određivanje najboljih podskupa za datu kardinalnost, a primenom algoritma izdvaja konačan najbolji podskup među predloženim podskupovima različite kardinalnosti. Pretraga započinje od datog podskupa S_0 (prazan skup u sekvencijalnim izboru unapred) i ponavlja se sve dok se ne pronađu najbolji podskupovi inkrementiranjem kardinalnosti. U svakoj iteraciji za najbolji podskup kardinalnosti c, pretražuju se svi mogući podskupovi kardinalnosti c+1 dodavanjem po jednog obeležja od preostalih. Svaki novi generisan podskup S kardinalnosti c+1 procenjuje se na osnovu nezavisne mere M i poredi sa prethodnim najboljim. Ako je S najbolji, postaje trenutno najbolji podskup S_{best} na nivou c+1. Na kraju svake iteracije, algoritam A se primenjuje na S_{best} na nivou c+1 i vrši se poređenje rezulata sa rezultatima dobijenim za najbolji podskup na nivou c. Ako je S_{best} bolji, algoritam nastavlja potragu na sledećem nivou, a ako nije, zaustavlja se pretraga i proglašava se tekući podskup za konačan najbolji. Kvalitet rezulata algoritma obezbeđuje prirodan kriterijum zaustavljanja u hibridnom modelu.

Na osnovu toga da li vrše procenu značaja svakog obeležja pojedinačno ili na nivou podskupa, filter modeli se mogu kategorizovati kao: algoritmi za rangiranje (engl. *feature weighting algorithms)* i algoritmi za izdvajanje podskupa obeležja (engl. *subset search algorithms)*. Algoritmi za rangiranje dodeljuju vrednosti svakom obeležju i rangiraju ih na osnovu njihovog značaja za ciljni koncept. Postoji veći broj različitih definicija značajnosti obeležja u navedenoj literaturi [25, 41]. Obeležje obučavajućeg skupa za analizu se smatra

kandidatom za izbor ukoliko je bolje rangirano, odnosno ako je vrednost njenegovog značaja veća od postavljenog donjeg praga granične vrednosti .

Algoritmi za izdvajanje podskupa obeležja vrše pretragu prostora koja je zasnovana na proceni kandidata [28]. Optimalan podskup je selektovan kada se pretraga završi. Neke od postojećih mera za procenu koje su se pokazale efikasnim u uklanjanju nebitnih i suvišnih obeležja uključuju meru doslednosti [44] i meru korelacije [38]. Mera doslednosti nastoji da pronađe minimalni broj obeležja koji dosledno razdvaja klasne oznake u kompletan skup. Nedoslednost se definiše za dve instance koje za iste vrednosti obeležja imaju različite klasne oznake.

2.2.1. Informaciona dobit

Informaciona dobit (engl. *information gain, IG*) je jedna od jednostavnijih metoda za rangiranje obeležja. Zasnovana je na meri entropije koja se koristi u oblasti teorije informacija.

Neka je za dati skup S, izvršena distribucija vrednosti klase C na osnovu vrednosti obeležja A. Entropija klase posmatrana u vezi sa podelom na osnovu vrednosti obeležja A manja je od entropije pre podele s obzirom na odnos između klase C i A. Entropija klase C nakon posmatranja obeležja A data je jednačinom (2.1):

$$H(C|A) = -\sum_{a\in A} p(a) \sum_{c\epsilon C} p(c|a)\, log_2(p(c|a)) \qquad (2.1)$$

gde je $p(c/a)$ uslovna verovatnoća klasne oznake (C=c) za datu vrednost obeležja A=a.

Imajući u vidu da se entropija može posmatrati kao kriterijum nečistoće, definiše se mera koja označava dodatne informacije o klasi C dobijene od obeležja A što predstavlja iznos za koji se entropija od klase C smanjuje. Ova mera je poznata kao informaciona dobit [45]. IG meri količinu informacije u bitovima o predviđanju klase, na osnovu informacija o prisutnosti obeležja za odgovarajuću distribuciju klase. Implementacijom ove metode svakom obeležju skupa S dodeljuje se rang izračunavanjem dobiti informacije prema jednačini (2.2).

$$IG(Class, Feature) = H(Class) - H(Class|Feature) = H(Feature) - H(Feature|Class) \qquad (2.2)$$

Iz jednačine (2.2) može se zaključiti da je IG simetrična mera. Dobit informacija za klasu C koja se dobije nakon posmatranja obeležja A jednaka je dobiti informacija za obeležje A posle posmatranja same klase.

2.2.2. Relief

Relief je metod rangiranja obeležja zasnovan na instancama [26] Zasniva se na principu slučajnog uzorka instance iz podataka, a zatim identifikovanjem najbliže susedne sa istom ili suprotnom klasom. Vrednosti obeležja najbližih suseda se porede sa instancom uzetom kao uzorak i koriste za ažuriranje relevantnosti. Ovaj postupak se ponavlja za korisnički specificiran broj instanci *m*. Značajno obeležje treba da pravi razliku u slučaju instanci sa različitom klasom, odnosno da ima jednake vrednosti za instance iste klase. Prvobitno, ovaj metod je bio definisan za problem klase sa dve vrednosti. ReliefF je proširen metod sa mogućnošću primene za slučaj skupa podataka sa multidimenzionalnom vrednošću klase i pojavom šuma u podacima. [46] ReliefF smanjuje uticaj šuma u podacima na osnovu prosečnog doprinosa od *k* najbližih suseda iste i suprotne klase svake uzorkovane instance umesto jednog najbližeg suseda. Za skup podataka sa problemom više-dimenzionalne klase, Relief pronalazi najbliže susede za svaku klasu koja se razlikuje od trenutno uzorkovane instance i odmerava doprinose od strane prethodne verovatnoće svake klase. Kononenko [46] napominjene da je u slučaju većeg broja uzorkovanih instanci veća pouzdanost procene ReliefF algoritma rangiranja, s tim da se uvećava i vreme izvršavanja. Preporuka autora je postavljanje broja instanci na 250 i parametra *k* na 10.

2.2.3. Simetrična procena nesigurnosti obeležja

Simetrična procena nesigurnosti obeležja (engl. *symmetrical uncertainty attribute evaluation*, *SummU*) je metod rangiranja baziran na merenju simetrične nepouzdanosti u odnosu na klasu. Procenjene vrednosti mogu biti u opsegu od nula do jedan, pri čemu jedan označava da je obeležje od značaja za klasu, a nula da nije [47]. Implementacijom ove metode za rangiranje, svakom obeležju dodeljuje se vrednost na osnovu jednačine (2.3):

$$SummU(Class, Feature) = 2\left[\frac{(H(Class) - H(Class \mid Feature))}{H(Class) + H(Feature)}\right] \quad (2.3)$$

H(Class) je mera entropije klase, a *H(Class|Feature)* uslovna entropija koja označava preostale nesigurnosti klase u odnosu na obeležje.

2.2.4. Izbor obeležja zasnovan na korelaciji

Izbor obeležja zasnovan na korelaciji (engl. *correlation-based feature selection*, *CFS*) [38] je metod za izdvajanje optimalnog skupa obeležja. Ovom metodom se razmatra prediktivna sposobnost i značaj svakog obeleležja i proverava postojanje redudantosti među njima. Izdvaja se podskup koji je u visokoj korelaciji sa klasom, a u slaboj međusobnoj korelaciji. Heurističkom procenom dodeljuje se visoka vrednost za izdvojeni podskup.

Neka je S skup podataka, a *k* broj obeležja skupa S. $Merit_S$ predstavlja heurističku vrednost izdvojenog podskupa za skup S i izračunava se jednačinom (2.4):

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}} \tag{2.4}$$

$\overline{r_{cf}}$ predstavlja prosečnu korelaciju obeležja i klase, a $\overline{r_{ff}}$ prosečnu međukorelaciju između obeležja. Izraz $k\overline{r_{cf}}$ može se posmatrati kao naznaka prediktivnosti, a izraz $\sqrt{k+k(k-1)\overline{r_{ff}}}$ označava koliko ima suvišnih odnosno redudantnih obeležja među njima koje se diskriminišu kako bi se sprečilo postojanje visoke korelacije sa jednom ili više drugih.

2.3. Diskretizacija

Efikasni modeli predviđanja zahtevaju detaljan pristup diskretizaciji podataka u fazi predprocesiranja. U oblastima mašinskog učenja i rudarenja podataka, postoji veliki broj algoritama koji su primarno orjentisani na rad sa diskretnim vrednostima. Međutim, u realnom svetu, podaci su mešovitog tipa, a u velikom broju slučajeva su kontinualnog tipa. Zbog toga je neophodno integrisati primenu algoritama sa metodama diskretizacije kako bi se izvršila transformacija kontinualnih podataka.

Osnovni cilj diskretizacije prvenstveno je usmeren na smanjenje broja mogućih vrednosti kontinualnih karakteristika čime se utiče na brzinu i efikasnost procesa [48]. Na taj način se redukuje obim podataka, a kontinualne vrednosti transformišu u odgovarajući skup diskretnih vrednosti relevantnijih za interpretaciju [49, 50]. Prednosti upotrebe diskretnih vrednosti se odnose na zahtevanje manje prostora u memoriji, razumljivost i jednostavnost upotrebom smislenih oznaka umesto tačnih vrednosti obeležja, regulisanje varijacija odstupanja u proceni manjih fragmentiranih podataka, smanjenje količine podataka

identifikovanjem i uklanjanjem redudantnih podataka, tačnost i brzinu algoritama [51]. Diskretizacija se može opisati kao postupak redukcije opsega kontinualnih obeležja deljenjem domena vrednosti na konačan skup disjunktnih intervala kojima se dodeljuju smislene oznake.

Neka je *A* obeležje skupa podataka *S*, a *n* broj instanci. Skup svih vrednosti obeležja *A* u skupu *S* predstavlja njegov aktivni domen *Dom(A)*, a $a = (a_1, a_2, .., a_n)$ vektor svih vrednosti obeležja *A* za *n* instanci.

Diskretizacija obeležja A koja je numeričkog tipa znači pronalazak *k* intervala aktivnog domena *Dom(A)* što podrazumeva određivanje *k-1* tački podele $t_0, t_1, t_2 \dots t_k$, $t_0 < t_1 < t_2 < \cdots < t_k$ takvih da skup $P = \{P_1, P_2, P_3 \dots P_k\}$ označava moguće intervale. P_i je definisano kao:

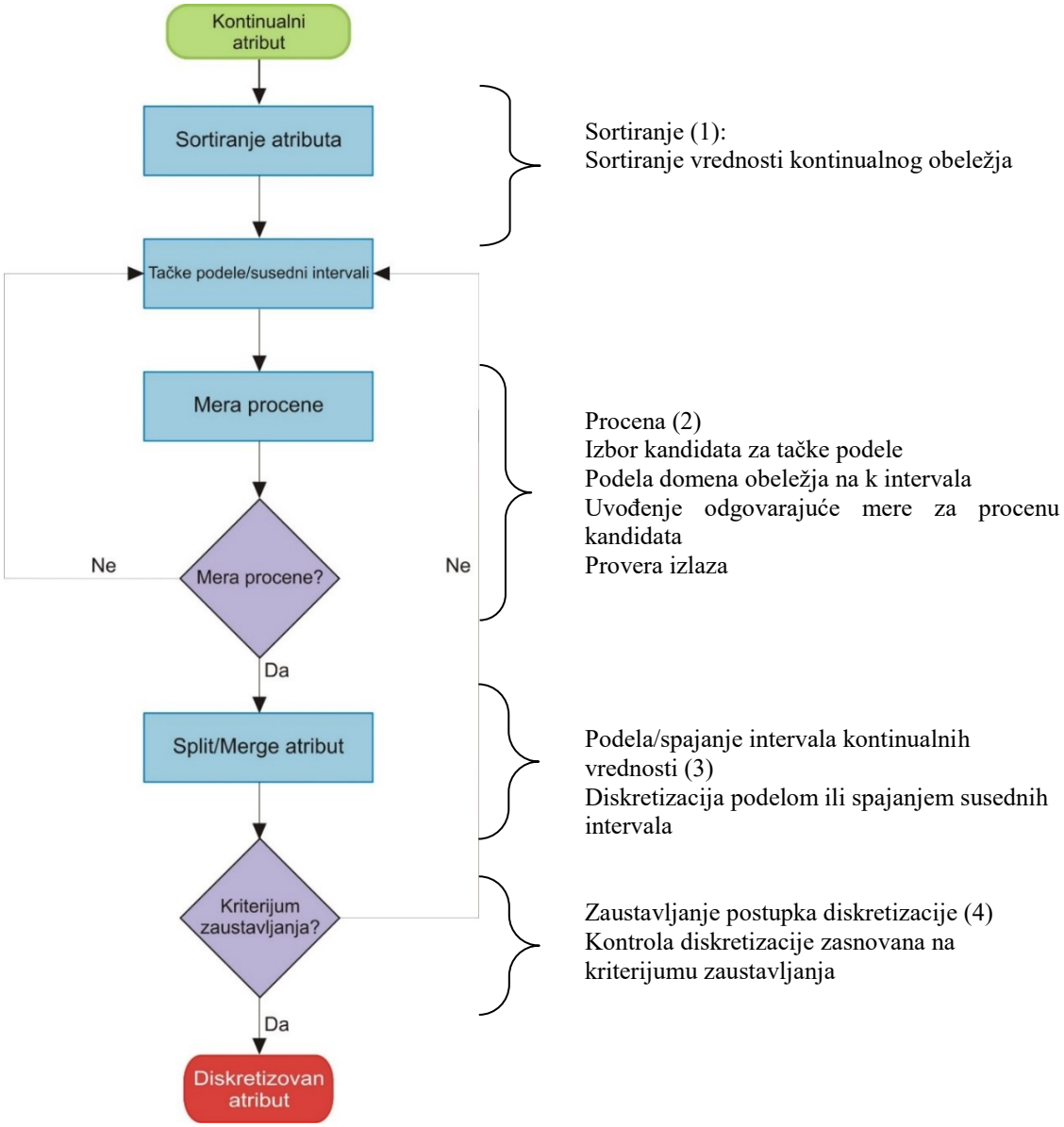
$P_i = \{a \in Dom(A): t_{i-1} \leq a \leq t_i\}$ za $i = \overline{0, k-1}$, $P_k = \{a \in Dom(A): t_{k-1} \leq a \leq t_k\}$, $t_0 = minDom(A)$, $t_k = maxDom(A)$.

Obeležje *A* se zamenjuje diskretizovanim obeležjem A^{disc} definisanim kao vektor $A^{disc} = \left(a_1^{disc}, a_2^{disc}, \dots, a_n^{disc}\right)$; $a_j^{disc} = i$ ako i samo ako $a_j \in P_i \; za \; j = \overline{1, n}$. Na osnovu navedenog, svaka vrednost obeležja *A* koja pripada P_i se zamenjuje sa i.

Tipičan proces diskretizacije može se opisati sa četri osnovna koraka:

1. sortiranje kontinualnih vrednosti obeležja koje se transformišu,
2. izbor, utvrđivanje mere procene, ispitivanje podobnosti kandidata za tačke podele odnosno broja *k* intervala podele domena,
3. podela ili spajanje intervala kontinualnih vrednosti prema odgovarajućem kriterijumu i
4. zaustavljanje postupka zasnovano na kriterijumu stopiranja. Pojam tačka odsecanja (engl. *cut-point*) označava realnu vrednost koja pripada domenu kontinualnog obeležja i kojom se opseg deli u dva intervala. Jedan interval obuhvata vrednosti manje ili jednake od tačke podele, a drugi vrednosti veće od tačke podele. Broj tačaka podele *k-1*, određen je brojem intervala podele *k* koji mogu biti korisnički definisani ili utvrđeni na osnovu postavljenog heurističkog pravila.

Na slici 2.3. dat je vizuelni prikaz tipičnog procesa diskretizacije.



Slika 2.3: Vizuelni prikaz postupka diskretizacije

Dobar algoritam diskretizacije treba suštinski da izbalansira gubitak informacija sa postupkom generisanja razumnog broja tački odsecanja za odgovarajući prostor pretraživanja. Kompromis mora biti pronađen između kvaliteta informacija, odnosno homogenih intervala s obzirom na karakteristiku predviđanja i statističkog kvaliteta koji predstavlja brojnost instanci u svakom intervalu za obezbeđivanje generalizacije. Obzirom da veliki broj mogućih vrednosti kontinualnih karakteristika usporava i dovodi do neefikasnosti primenjenog algoritama, osnovni cilj metoda diskretizacije je redukcija domena vrednosti podelom na diskretne intervale. U isto vreme, metod diskretizacije treba da maksimizuje nezavisnost između

različitih diskretnih vrednosti karakteristika i klasnih oznaka, a da minimizuje gubitak informacija u tom procesu. Izbor odgovarajuće metode diskretizacije podrazumeva postizanje kompromisa između ova dva cilja. Mnoge studije pokazuju uticaj diskretizacije na indukcione zadatke: pravila sa diskretnim vrednostima su razumljivija, a u slučajevima predviđanja i klasifikacije postiže se veća tačnost. Uticaj diskretizacije može se meriti sa aspekta tačnosti, vremena potrebnog za izvršavanje algoritma učenja i razumljivosti rezultata. Većina metoda diskretizacije primenjuje iterativnu pretragu prostora kandidata koristeći različite funkcije ocenjivanja za procenu rezultata. Ključno pitanje nije samo da li je neka metoda superiornija u odnosu na druge, nego pod kojim uslovima određena metoda može ostvariti značajno bolje performanse za dati problem.

Metode diskretizacije mogu se klasifikovati kao [51]:

- nadzirane (engl. *supervised*) i nenadzirane (engl. *unsupervised*),
- hijerarhijske (engl. *hierarchical*) i nehijerarhijske (engl. *non-hierarchical*),
- s vrha prema dnu (engl. *top-down*) i sa dna prema vrhu (engl. *bottom-up*),
- statičke (engl. *static*) i dinamičke (engl. *dynamic*),
- globalne (engl. *global*) i lokalne (engl. *local*),
- parametrizovane (engl. *parametric*) i neparametrizovane (engl. *non-parametric*),
- jednovarijantne (engl. *univariate*) i više-varijantne (engl. *multivariate*).

Nenadzirane metode podrazumevaju podelu domena kontinualnih vrednosti obeležja u podopsege, ne uzimajući u obzir informacije o klasi, tačnije na osnovu korisnički definisanih parametara. Lošiji rezultati ovih metoda mogu se javiti u slučaju neuniformne distribucije kontinualnih vrednosti i pojavu izuzetaka u skupu podataka. Međutim, ukoliko nije poznata informacija o klasi, njihova primena je jedina moguća. U nadziranim metodama diskretizacije, informacija o klasi se koristi za pronalazak odgovarajućih intervala određivanjem najoptimalnijih tački odsecanja Postoje različite metode koje koriste klasnu informaciju za pronalazak značajnih intervala. Nadzirana diskretizacija može se dalje kategorisati prema tome da li su intervali izabrani upotrebom metrika zasnovanih na greškama u obučavajućim podacima, meri dispariteta, odnosno, entropiji intervala ili nekim statističkim merama. U slučaju hijerarhijske diskretizacije izbor tački odsecanja se ostvaruje postepenim procesom, formiranjem implicitne hijerarhije u opsegu vrednosti. Primenjena procedura može biti podela

(engl. *split*) ili spajanje (engl. *merge*) [52]. Nehijerahijeske metode vrše skeniranje uređenih vrednosti samo jednom, sekvencijalno, formirajući intervale. Metode s vrha prema dnu kreću od jednog intervala koji se dalje deli u procesu diskretizacije. Metode s dna prema vrhu započinju postupak sa kompletnom listom vrednosti kontinualnog obeležja, posmatrajući ih kao tačke odsecanja. U toku postupka diskretizacije, ove metode vrše uklanjanje nekih tački odsecanja spajanjem intervala. Kao kriterijum zaustavljanja postupka koriste se različite prag vrednosti. Statičkim metodama diskretizacija se izvršava nad jednim obeležjem ne uzimajući u obzir ostale. Postupak se ponavlja za ostala obeležja onoliko puta koliko je potrebno. Dinamičkim metodama, u jednom trenutku se izvršava diskretizacija svih obeležja. Statički pristup podrazumeva primenu diskretizacije u okviru faze predprocesiranja, a dinamičkim pristupom diskretizacija kontinualnih vrednosti se vrši u toku kreiranja klasifikatora kao što je C4.5. Dinamičke metode su povezane sa onim metodama klasifikacije koje rade sa realnim vrednostima karakteristika. *Global /Local:* Lokalne metode vrše diskretizaciju u lokalizovanim regijama prostora instanci (podskup instanci). Globalne metode diskretizacije koriste ceo prostor instanci i rade sa svim obeležjima. Vrši se podela u particije koje su nezavisne i kreira se globalna mreža nad n-dimenzionalnim prostorom kontinualnih instanci [53]. Lokalne metode se često povezuju sa dinamičkim metodama diskretizacija kod kojih se koristi oblast prostora instanci. Parametarska metoda diskretizacije zahteva unos od strane korisnika, kao što je maksimalan broj intervala diskretizacije. Neparametarska diskretizacija koristi samo dostupan skup podataka i nije potreban ulaz od strane korisnika. Metoda jednovarijantne diskretizacije kvantifikuje vrednosti jednog kontinualnog obeležja, dok metoda više-varijantne diskretizuje simultano više obeležja.

2.3.1. Metoda jednakih intervala

Metoda jednakih intervala (engl. *equal-width binning*, *EWB*) se može svrstati u najednostavniju direktnu nenadziranu metodu diskretizacije. Postupak obuhvata sortiranje kontinualnih obeležja, a zatim podelu domena posmatranih obeležja na *k* intervala (engl. *bins*) iste veličine (δ) sa *k+1* tački podele.

Neka je *A* obeležje skupa, a aktivni domen vrednosti označen sa $Dom(A) = (a_1, a_2, \dots a_n)$, $a_{max} = \max\{a_1, a_2, \dots a_n\}$ i $a_{min} = \min\{a_1, a_2, \dots a_n\}$ granice domena vrednosti posmatranog obeležja. Veličina δ za *k* jednakih intervala određuje se prema jednačini (2.5):

$$\delta = \frac{(a_{max} - a_{min})}{k}, \quad (2.5)$$

gde a_{min} i a_{max} predstavljaju minimalnu i maksimalnu vrednost instanci, a tačke podele su a_{min}, $a_{min} + \delta$, ..., $a_{max} = a_{min} + k\delta$.

Broj intervala k za skup instanci n, a_{min} i a_{max}, respektivno, može biti unapred definisan od strane korisnika ili izračunat na osnovu formula (2.6), (2.7), (2.8):

$k = \log_2 n + 1$ [54], (2.6)

$$k = \frac{a_{max} - a_{min}}{h}, \quad (2.7)$$

h izračunava prema formuli $h = \frac{2 \times IQR}{\sqrt[3]{n}}$, IQR predstavlja unutrašnje kvartale skupa podataka [55]

$$k = \frac{a_{max} - a_{min}}{h}, \quad (2.8)$$

h se izračunava prema formuli $h = \frac{3.5 \times \sigma}{\sqrt[3]{n}}$, σ predstavlja standardnu devijaciju [56].

Broj intervala k je fiksno određen i nezavisan od specifičnih osobina obučavajućih podataka. Ova restrikcija može da dovede do nekih neželjenih sporednih efekata. U slučaju velikog skupa podataka, mali broj intervala podele može prouzrokovati grupisanje širokog spektra instanci što svakako neće imati povoljan uticaj na primenjeni algoritam učenja. S druge strane, ako je broj intervala podele suviše veliki, onda će intervali sadržati mali broj instanci pa se u tom slučaju ne može utvrditi značaj i uticaj sprovedene diskretizacije.

Pored navedenog, određivanje graničnih tačaka za intervale podele, bez uzimanja u obzir klasnog obeležja, može da dovede da neki od intervala sadrži kombinovane vrednosti obeležja koje su u zavisnosti sa različitim klasnim oznakama [52] što dalje uslovljava usložnjavanje postupka klasifikacije.

2.3.2. Metoda diskretizacije zasnovana na histogramu

Metoda diskretizacije zasnovane na histogramu spada u nenadzirane tehnike diskretizacije, obzirom da ne koristi informaciju o klasnom obeležju. Histogram predstavlja geometrijsku raspodelu tabele frekvencija koja olakšava statističku analizu podataka.

Neka obeležje *X* uzima *n* vrednosti: x_1 , x_2 ,..., x_n koje se u skupu od *N* instanci pojavljuje f_1 ,f_2 ,..., f_n puta. Veličine f_1 ,f_2 ,..., f_n zadovoljavaju jednačinu $f_1 + f_2 + \cdots + f_n = N$ i predstavljaju frekvencije. U slučaju velikog broja podataka *N*, domen vrednosti obeležja *X* se predstavlja sa k intervala, a frekvencije f_1 ,f_2 ,..., f_k sada označavaju broj vrednosti *X* koje pripadaju prvom, drugom, ..., *k*-tom intervalu. Intervali se ne preklapaju međusobno i imaju određene granične vrednosti. Definiše se ista širina (engl. *bin size, bar width, class size*) za sve intervale, a broj instanci posmatrane slučajne promenljive za svaki interval predstavlja frekvencije.

Histogram podrazumeva dostupnost svih podataka, odnosno isključuje rad sa podacima u kojima su zabeležene instance sa vrednostima koje nedostaju. Uzimajući u obzir ekstremne vrednosti i izuzetke pri kreiranju histograma definišu se tačke podele a_1 , ..., a_{k-1} i frekvencije instanci f_1 ,..., f_{k-1} , f_k tako da se može definisati *k* intervala u domenu vrednosti posmatrane slučajne promenljive $(-\infty, a_1]$, $(a_1, a_2]$, ..., $(a_{k-2}, a_{k-1}]$, $(a_{k-1}, +\infty)$. Vizuelnim prikazom u obliku pravougaonih grafikona između kojih nema praznina, histogram obezbeđuje informacije o distribuciji slučajno odabrane promenljive analiziranog skupa podataka. Algoritam za kreiranje histograma obuhvata sledeće korake:

- sortiranje vrednosti slučajne promenljive u rastućem redosledu,
- određivanje minimalne i maksimalne vrednosti,
- određivanje vrednosti *k* , broj intervala podele (engl. *numbers of bins*),
- izračunavanje širine intervala, *bin size*=$\frac{\max - min}{k}$,
- izračunavanje frekvencije instanci za svaki interval (engl. *frequency table*) i
- iscrtavanje pravougaonih grafikona tako da x - osa predstavlja intervale podele, a y-osa frekvenciju instanci za svaki interval.

Najčešće korišćeni histogrami su tipa jednakih širina (engl. *equal width*) gde je opseg posmatranih vrednosti podeljen u *k* intervala jednakih širina ili tipa jednakih frekvencija (engl. *equal frequency*), gde je opseg posmatranih vrednosti podeljen u *k* intervala koji sadrže isti broj instanci. Za oba pomenuta algoritma neophodno je definisati parametar *k*, odnosno broj intervala podele, što je ujedno i glavni problem.

U istraživačkoj analizi podataka, primena histograma podrazumeva rekurzivnu primenu na svaku particiju u cilju automatskog generisanja koncepta hijerahije sa više nivoa sve dok se ne dostigne unapred definisan broj nivoa. Za kontrolu rekurzivne procedure može se koristiti minimalna veličina intervala ili minimalni broj vrednosti po intervalu. Kreiranje histograma za različite vrednosti parametra *k* (broj intervala) omogućava izabor najpodesnijeg u zavisnosti od cilja njegove upotrebe.

2.3.3. Metoda diskretizacije zasnovana na entropiji

Nadzirana metoda diskretizacije zasnovana na entropiji koristi informaciju o entropiji klase kandidata prilikom određivanja tačke podele. Informacija entropije klase je mera čistoće koja meri količinu informacija potrebnih za utvrđivanje kojoj klasi instanca pripada. Posmatra se interval koji sadrži sve poznate vrednosti obeležja i rekurzivnom podelom se deli u manje podintervale sve dok ne bude zadovoljen postavljeni kriterijum zaustavljanja. Tačka podele će biti izabrana procenom mere dispariteta odnosno određivanjem entropije klasa za kandidate particija. U metodi diskretizacije zasnovanoj na entropiji, najbolja tačka se određuje na osnovu entropije potencijalnih kandidata tačke podele.

Neka je dat skup instanci S, atribut A i granica podele tačka T. Entropija podele T, označena sa $E(A,T;S)$ definisana je sledećim jednačinama (2.9) i (2.10):

$$E(A,T;S) = \frac{|S_1|}{|S|} Ent(S_1) - \frac{|S_2|}{|S|} Ent(S_2) \tag{2.9}$$

$$Ent(S_i) = -\sum_{i=1}^{k} p(C_j, \mathrm{S_i}) \log_2\big(p(C_j, S_i)\big) \tag{2.10}$$

Entropija podskupa S_1 i S_2 izračunava se prema formuli (2.10), gde $p(C_j\ , S_i)$ predstavlja procenat instanci u S_i koje imaju klasu C_j, k predstavlja broj klasa označenih sa $C_{1,}\ C_2\ \ldots C_k$.

U jednačini (2.9) skup instanci S podeljen je u dva intervala S_1 i S_2 koristeći tačku podele T za vrednost atributa A. Funkcija entropije *Ent* za dati skup izračunata je na osnovu raspodele uzoraka klase u skupu. Najbolji kandidat za tačku podele T među svim kandidatima za $E(A,T;S)$ je onaj koji ima minimalnu vrednost entropije [57]. Nakon izbora tačke podele,

vrednosti kontinualnih obeležja se dele u dva dela. Ovaj postupak rekurzivno se ponavlja sve dok ne bude dostignut postavljen kriterijum zaustavljanja.

U metodi diskretizacije zasnovanoj na entropiji, kriterijum zaustavljanja definisan je jednačinama (2.11), (2.12), (2.13):

$$Gain(A,T;S) < \frac{log_2(N-1)}{N} + \frac{\Delta(A,T;S)}{N}, \tag{2.11}$$

gde je N broj instanci skupa S

$$Gain(A,T;S) = Ent(S) - E(A,T;S) \tag{2.12}$$

$$\Delta(A,T;S) = \log_2(3^k - 2) - [k \cdot Ent(S) - k_1 \cdot Ent(S_1) - k_2 \cdot Ent(S_2)] \tag{2.13}$$

k_i je broj oznaka klase predstavljen u skupu S_i.

Uzimajući u obzir da se, na osnovu opisanog kriterijuma zaustavljanja, particije uz svaku granu rekurzivne diskretizacije procenjuju nezavisno, neki delovi prostora kontinualnih vrednosti biće particionisani veoma fino, dok će oni delovi sa relativno niskom entropijom biti grubo podeljeni. Navedeni kriterijum zaustavljanja poznat kao MDL (eng. *Minimal Description Length Principle*) uveden je i opisan u radu [58].

2.4. Klasifikacija

Klasifikacija (engl. *classification*) je jedan od najčešće proučavanih problema u oblasti rudarenja podataka i mašinskog učenja. Zadatak klasifikacije je predviđanje vrednosti klase na osnovu vrednosti ulaznih obeležja. Smatra se zadatkom nadgledanog učenja, što znači da su vrednosti funkcije zadate u skupu obučavanja. Skup instanci podataka podeljen je u obučavajući i test skup. U obučavajućem skupu svakoj instanci dodeljuje se klasna oznaka koja identifikuje klasu kojoj pripada.

Algoritmi nadgledanog mašinskog učenja koriste se da podstaknu klasifikator iz skupa pravilno klasifikovanih instanci odnosno skupa obučavanja. Skup testiranja koristi se za

merenje kvaliteta klasifikatora dobijenog nakon primenjenog procesa obučavanja. Generalno problem klasifikacije može se definisati na sledeći način:

Dat je skup karakteristika $R = \{A_1, ..., A_k; C\}$ pri čemu je C obeležje klase sa diskretnim vrednostima. Neka su vrednosti klase C definisane sa $Dom(C)=\{c_1, ..., c_k\}$, a S prostor obeležja koji obuhvata skup R. Problem klasifikacije predstavlja pronalazak mapiranja $f: Dom(A_1) x...x Dom(A_k) \text{-->} Dom(C)$ tako da za sve tačke podatka $p \in S$, važi jednakost (2.14).

$$f(p[A_1,...,A_k]) = p[C]. \tag{2.14}$$

2.4.1. Linerani klasifikatori

Linearna regresija (engl. *linear regression)* predstavlja najjednostavniji klasifikator kada su u pitanju numerički podaci. Podrazumeva pronalazak linearnog modela za ciljnu promenljivu Y a u zavisnosti od ulaznih, međusobno nezavisnih promenljivih $X_1,...,X_k$. Najednostavniji slučaj je kada Y linearno zavisi od samo jedne promenljive X. U slučaju više ulaznih promenljivih koristi se višestruka linearna regresija (engl. *multiple linear regression-MLR).* MLR je korisna deskriptivna tehnika obzirom da generalizuje koeficijent korelacije za više promenljivih i ukazuje na relativnu značajnost pojedinačnih ulaznih promenljivih.

Primenom višestruke linearne regresije, ciljna promenljiva Y se modeluje prema jednačini (2.15) kao linerna funkcija nezavisnih promenljivih X_1, ..., $X_{k.:}$

$$Y = \alpha_k X_k + \alpha_{k-1} X_{k-1} + ... + \alpha_1 X_1 + \alpha_0, \tag{2.15}$$

Bez obzira na numeričku vrednost rezultata, model se lako može tumačiti u slučaju kada je klasna promenljiva binarnog tipa (ocena studenta može da uzme jednu od dve vrednosti – pao ili položio). Jednačina (2.16) ukazuje da se linearna regresija može koristiti kao sredstvo za identifikaciju linearnih ili nelinearnih veza između promenljivih proverom vrednosti kvadrata koeficijenta višestruke korelacije

$$r^2 = \frac{Var(Y) - MSE}{Var(Y)} \tag{2.16}$$

Vrednost $Var(Y)$ predstavlja Y varijante, a MSE (engl. *mean squared error*) srednju vrednost kvadrata grešaka za skup podataka veličine n, odnosno $MSE = SSE/n$, pri čemu SSE

(engl. *error sum of squares)* predstavlja sumu kvadrata grešaka. Što je vrednost r^2 bliža 1, veća je linearna zavisnost. U domenu obrazovnih podatka, glavni problem u primeni linearne regresije tiče se kolinearnosti i podataka koji se ne uklapaju – izuzetaka (na primer, studenti koji polože ispit sa minimalno uloženog napora ili padnu ispit bez nekog vidljivog razloga).

Metoda vektora oslonaca (engl. *Support Vector Machines*, SVM) [25] predstavlja linearni klasifikator koji pronalazi optimalnu hiperravan u cilju razdvajanja primera koji pripadaju različitim klasama [59]. U slučaju linearno razdvojivih klasa, postoji beskonačno mnogo hiperravni koje klasifikuju instance skupa za obučavanje. Najefikasnije raspoređivanje novih instanci test skupa, izvršiće optimalna hiperravan sa maksimalnom marginom separacije. Vektori koji ograničavaju širinu margine predstavljaju granične primere vektora oslonaca. Linearna kombinacija vektora oslonca je rešenje modela; ostale tačke podataka se ignorišu. Broj odabranih vektora podrške od strane SVM algoritma je obično mali, te su ovi klasifikatori pogodniji za manje skupove sa većim brojem osobina podataka. Za većinu skupova podataka SVM klasifikator ne može da nađe optimalnu razdvajajuću hiperravan zbog postojanja pogrešno klasifikovanih instanci u podacima. Ovaj problem se rešava primenom mekih margina koje prihvataju pogrešnu klasifikaciju primeraka obučavajućeg skupa. U slučaju nerazdvojivih podataka, SVM klasifikator vrši preslikavanje podataka na više-dimenzionalni prostor obeležja (engl. *feature space)* i definiše razdvajajuće hiperravni u okviru njega. Nakon kreiranja hiperravni, koristi se posebna funkcija jezgra (engl. *kernel*) za mapiranje novih tačaka u više-dimenzionalnom prostor za klasifikaciju. S obzirom da se ovom funkcijom definiše više-dimenzionalni prostor obeležja u kome će instance obučavajućeg skupa biti klasifikovane, njen izbor je izuzetno značajan. U praksi, SVM se smatraju boljim klasifikatorima jer generišu odlične rezultate.

2.4.2. Stabla odlučivanja

Stablo odlučivanja (engl. *decision tree*) je specifičan tip klasifikatora. Predstavlja prediktivan model jednostavne hijerahijske strukture. Koreni čvor se nalazi na najvišem nivou. Svaki unutrašnji čvor označava testiranje vrednosti obeležja, grana predstavlja rezultat testa, a svaki čvor lista odgovarajuću klasnu oznaku [19]. Kriterijumom testiranja realizuje se grananje i formira binarno stablo ili stablo višeg reda. Izbor čvorova na najvišem nivou zasniva se na proveri kvaliteta obeležja izračunavanjem statističkih mera poput informacione dobiti, mere *gain ratio (GR)* i *gini indeks (GI)*. U slučaju binarnog stabla, roditeljski čvor može imati samo

dva čvora deteta. Unutrašnji čvor proverava vrednost obeležja i rezultat prosleđuje pripadajućoj grani. Formiraju se putanje stabla zasnovane na logičkom izrazu konjukcije $T_1 \wedge T_2 \wedge \ldots T_k$, gde je svaki test obeležja T_i u formi $A_i = 1$ ili $A_i = true$. Za kategorijski tip podataka, koristi se stablo višeg reda. Svaki roditeljski čvor može imati više od dvoje dece što omogućava kreiranje putanja kroz stablo. Kreiranje putanja kroz stablo podrazumeva testiranje primenom izraza logičke disjunkcije $A_1 = a_1 \vee \ldots \vee A_i = a_k$ i formiranje pripadajućih grana. Osnovna ideja algoritama stabla odluke usmerena je ka podeli prostora obeležja dok se ne dostigne kriterijum prekida u listu i ostvari pripadanje lista određenoj klasnoj oznaki. Kao posledica mogućih greški i nepotpunosti u podacima javlja se pojava nedoslednosti; obučavajući skup može da sadrži instance sa zajedničkim vrednostima ulaznih obeležja, ali različitih klasnih oznaka. U ovom slučaju bira se vrednost klasnog obeležja većinskog broja podataka datog čvora. Ovaj princip je poznat kao princip glasanja većine (engl. *majority vote principle*). Alternativni način je određivanje uslovne verovatnoće klase zasnovane na relativnim frekvencijama u čvoru. Uslovna verovatnoća klase predstavlja pouzdanost (engl. *confidence*) odgovarajućeg pravila odluke.

Kompleksnost prenaučenih modela predstavlja osnovni problem. Rešenje problema prenaučenosti moguće je primenom jednog od pristupa: rano zaustavljanje (engl. *early stopping*) i kresanje, tj. potkresavanje (engl. *pruning*). Pristupom ranog zaustavljanja prekida se proces kreiranja stabla pre nego što model postane previše složen. Problem se ogleda u donošenju odluke o odgovarajućoj optimalnoj tački zaustavljanja rasta stabla.

Potkresavanje stabla odluke predstavlja pristup kojim se ostvaruje optimizacija efikasnosti i postiže tačnost klasifikacije. Primena ove metode obično rezultira smanjenjem veličine stabla odnosno broja čvorova. Na taj način se izbegava nepotrebna kompleksnost modela kao i prenaučenost podataka, previše prilagođavanje skupa podataka prilikom klasifikacije novih podataka. Potkresavanje stabla se realizuje spajanjem jednog po jednog čvora listova ili zamenom podstabla sa listom.

Postoje dva načina za potkresavanje stabla: metod potkresavanja potpuno indukovanog stabla (engl. *post-pruning*) i metod potkresavanja tokom formiranja stabla (engl. *online pruning*). Osnovna razlika između navedenih realizacija ogleda se u vremenu implementacije i parametrima koji utiču na odluku o uklanjanju čvorova. U zavisnosti od implementiranog

metoda, stablo odluke mora biti u postupku indukcije kreiranja ili u potpunosti kreirano i spremno za potkresavanje.

Osnovni princip zasniva se na poređenju greški generisanih u slučaju pre i posle potkresavanja stabla, a zatim odlučivanja u skladu sa maksimalnim izbegavanjem greške. Metrika za opisivanje moguće greške, označena je procenom greške (E) i izračunava se sa formulom (2.17):

$$E = \frac{e+1}{N+m} \tag{2.17}$$

gde vrednost *e* označava broj pogrešno klasifikovanih instanci za dati čvor, *N* broj instance koje dostižu dati čvor i *m* sve instance obučavajućeg skupa.

Metod *post-pruning* se implementira na kompletno indukovanom stablu odluke za uklanjanje statistički beznačajnih čvorova. U smeru od dna prema vrhu, upoređuju se verovatnoće (relativna frekvencija) bratskih listova čvorova. Preovlađujuća dominacija određenog čvora lista rezultuje potkresavanjem tog čvora. Izračunava se procena greške svakog dete čvora i koristi za utvrđivanje ukupne greške roditelj čvora. Roditelj čvor se zatim obrezuje u skladu sa relativnim frekvencijama dece čvorova, a greška zamenjenog čvora poredi sa greškom starog roditeljskog čvora na koju utiču greške dece čvorova. Ovim poređenjem se utvrđuje da li je realizovano povoljno potkresavanje u datom čvoru.

Ukoliko se potkresavanje stabla obavlja u toku formiranja, odgovarajući algoritam dele skup na osnovu obeležja koja obezbeđuju najviše informacija o klasnoj oznaci. Navedeni faktor podele predstavlja odlučujuće obeležje za svaki test čvor. Svakom podelom koja formira dete list sa manje od minimalnog broja instanci obučavajućeg skupa podataka, roditelj čvor i njegova deca se grupišu u jedan čvor. Ovaj proces se nastavlja tokom kreiranja celog stabla.

Prednosti stabla odluke se odnose se na jednostavnost interpretacije, lakoću razumevanja, rada sa mešovitim podacima numeričkog i kategorijskog tipa, brzu klasifikaciju novih instanci, fleksibilnost u slučaju postojanja manje količine šuma i nedostajućih vrednosti obeležja.

Glavno ograničenje odnosi se na pretpostavku da vrednosti domena podataka mogu biti deterministički klasifikovane u tačno jednu klasu pa se sve nedoslednosti intrepretiraju kao greške. Ovo ograničenje se javlja u slučaju obrazovnog skupa podataka sa postojanjem izuzetaka. Stablo odluke je osetljivo na pojavu prenaučenosti, posebno u slučaju malog skupa.

Najčešće primenjivani algoritmi učenja stabla odluke koji koriste različite kriterijume podele u unutrašnjim čvorovima su ID3, C4.5 i CART. ID3 je jednostavan algoritam stabla odluke čiji je tvorac Ross Quinlan [60]. Osnovna ideja ID3 algoritma usmerena je na konstrukciju stabla odluke upotrebom pohlepne pretrage s vrha prema dnu. Sa ciljem izbora najznačajnijih, u svakom čvoru vrši se testiranje obeležja. Definisana je statistička mera dobiti informacija za merenje značaja obeležja. C4.5 algoritam [45] se može smatrati naslednikom ID3. Koristi meru odnosa dobiti (engl. *gain ratio, GR*) kao kriterijum podele skupa podataka. Algoritam primenjuje normalizaciju dobiti informacija kao vrednost za podelu.

CART [61] radi se kategorijskim i diskretnim podacima i nije osetljiv na postojanje vrednosti koje nedostaju. Koristi *gini index* kao meru za izbor obeležja od značaja. Za razliku od ID3 i C4.5 algoritama, CART realuzuje binarnu podelu. Mera *gini index* ne koristi probabilističku pretpostavku. CART koristi pristup potkresavanja kompleksnosti za otkanjanje nepotrebnih grana sa stabla odluke i poboljšanje tačnosti. J48 predstavlja verziju ID3 algoritma realizovanog u Weka softveru otvorenog koda [62]. Generiše se stablo višeg reda a mera odnosa dobiti se koristi kao kriterijum za utvrđivanje značaja obeležja. Najznačajnije obeležje postavlja se za koreni čvor i skup podataka se deli na osnovu vrednosti korenog elementa [63]. Izračunava se dobit informacija za sve pod-čvorove pojedinačno i proces se ponavlja sve dok se predviđanje ne završi.

J48 može da radi sa kontinualnim i diskretnim tipovima, obučava podatke sa nedostajućim vrednostima obeležja i omogućava preuređenje stabla nakon kreiranja. Potkresavanje zasnovano na greškama se vrši nakon faze rasta stabla. Postoji veliki broj parametara vezanih za kresanje stabla prilikom primene J48 algoritma koji značajno mogu uticati na kvalitet rezultata. J48 koristi dve metode: zamena podstabla i formiranje podstabla.

Problem prenaučenosti, na koji je stablo odlučivanja osetljivo u slučaju manjeg obrazovnog skupa, moguće je izbeći primenom pristupa učenja ansambla [64]. Pristup primene metoda učenja ansambla omogućava kombinovanje modela različitih struktura i značajno poboljšanje tačnosti klasifikacije.

Random forest [65] predstavlja klasifikator koji se sastoji od kolekcije klasifikatora stabla odlučivanja $f_m(X)$, *m=1,2..M,* koji zavise od nezavisno identično distribuiranih skupova obeležja, a svako stablo raspoređuje jediničan glas da klasifikuje ulazni vektor *X*. Predloženo je nekoliko načina za generisanje kolekcije slučajnih stabala [66,67]. Metod pakovanja (engl. *bagging*) [68] obučava svako stablo na nezavisnim, nasumično ravnomernim podskupovima uzoraka sa zamenom izdvojenim iz orginalnog obučavajućeg skupa. Sve moguće kombinacije obeležja realizuju se metodom pakovanja i vrši se klasifikacija uzimanjem najpopularnije izglasane klase od strane svih predviđajućih stabala u šumi. Dizajn stabla odluke zahteva odabir kriterijuma izbora obeležja i metoda potkresavanja.

Postoji mnogo pristupa izbora obeležja korišćenih za indukciju stabla odlučivanja. Većina pristupa zasniva se na dodeli mere značaja i kvaliteta direktno obeležju. Najčešće korišćene mere za utvrđivanje značaja obeležja su informaciona dobit, odnos dobiti i *gini index (GI)* [69]. Svaki put stablo naraste do maksimalne dubine na novom obučavajućem skupu generisanom koristeći kombinaciju obeležja. Ova u potpunosti formirana odrasla stabla nisu potkresana. To predstavlja osnovnu prednost *Random forest* klasifikatora u odnosu na ostale metode stabla odlučivanja predložene u [45]. Studije ukazuju da izbor metoda potkresavanja, a ne kriterijumi odabira obeležja, utiču na performance klasifikatora na kojima je zasnovano stablo [70]. Breiman predlaže da sa povećavanjem broja stabala, greška generalizacije konvergira čak i bez potkresavanja a prenaučenost ne predstavlja problem [71]. Za generisanje *Random forest* klasifikatora neophodna su dva korisnički definisana parametra: broj obeležja koji se koristi u svakom čvoru za formiranje stabla i broj stabala koji treba da raste. *Random forest* klasifikator predstavlja uspešan oblik kombinovanja pakovanja sa slučajnim izborom obeležja za svaki čvor svakog stabla u šumi [72]. Prednosti ovog klasifikatora odnose se na generisanje tačnijih modela, neosetljivost na prisustvo šuma i izuzetaka, internu procenu greške, snage, korelacije i značaja obeležja, jednostavnost i razumljivost.

2.4.3. Bajesove mreže

Bajesove mreže (engl. *Bayesian networks, BN*) predstavljaju dobro razvijenu tehniku u oblasti mašinskog učenja. Koristeći obučavajući skup podataka, Bajesovo učenje izračunava verovatnoće i vrši predviđanje zasnovano na verovatnoćama hipoteza. Statističke zavisnosti ovih klasifikatora predstavljaju se vizuelnom strukturom grafa [73]. Svaki čvor grafa odgovara jednoj karakteristici, dolazne ivice do čvora određene su zavisnim karakteristikama, a jačina te

zavisnosti definisana je uslovnim verovatnoćama. Kada se koriste Bajesove mreže za klasifikaciju, prvo treba naučiti mrežu o strukturi zavisnosti između karakteristika $A_1, ..,A_k$ i klase C. Nakon izbora strukture, parametri uče iz podataka i definišu uslovne distribucije klase za sve moguće tačke podataka i sve vrednosti klase.

Verovatnoća klasifikacije tačke podataka t u klasu klasne oznake c izračunava se prema Bajesovom pravilu koje je prikazano u jednačini (2.18):

$$P(C = c \mid t) = \frac{P(C = c)P(t \mid C = c)}{P(t)} \quad . \tag{2.18}$$

U praksi, problem je veliki broj verovatnoća koje treba proceniti. Na primer, ako sve karakteristike $A_1, ...,A_k$ imaju j različitih vrednosti i međusobno su zavisni, potrebno je definisati $O(j^k)$ verovatnoća. To znači da nam je potreban veliki obučavajući skup za procenu zajedničke verovatnoće tačnosti. Bajesove mreže su veoma atraktivne metod za obrazovni domen, ali je opšta mreža previše složena za skupove sa malim brojem podataka. Rešenje ovih problema je upotreba Näive Bayes klasifikatora koji generiše model ograničen snažnom nezavisnošću pretpostavke.

2.4.4. Näive Bayes

Näive-Bayes (NB) klasifikator [74] naziva se naivnim jer pretpostavlja uslovnu nezavisnost klase. To znači da je efekat vrednosti obeležja za datu klasnu oznaku nezavisan u odnosu na ostale. Kreira jednostavan model lak za interpretiranje. Pogodan je za male skupove podataka, kombinujući složenost sa fleksibilnim probabilističkim modelom. Radi sa kategorijskim i numeričkim podacima koje je neophodno transformisati primenom metode diskretizacije. Procenom gustine podataka umesto distribucije vrednosti moguće je obučiti kontinualan model. Ovaj pristup podrazumeva opšti oblik normalne distribucije podataka, što je u realnim obrazovnim okruženjima redak slučaj. Osim toga, diskretizacija pojednostavljuje način kreiranja modela tako da kod rezultujućeg modela retko dolazi do efekta preobučavanja. Struktura Näive Bayes mreže sastoji se od samo dva sloja, sloj osnovnog čvora u kome je klasna promenljiva i sloj čvorova listova u kojima su sve druge promenjive. Glavna ideja koja stoji iza Näive Bayes klasifikacije podataka je maksimiziranje vrednosti $P(X|C_i)P(C_i)$ gde je i indeks klase, koristeći Bajesovu teoremu posteriorne verovatnoće.

Neka je dat skup podataka predstavljen n-dimenzionalnim vektorom X= $(x_1, x_2, \dots, x_n)$ koji prikazuje n instanci sa k obeležja označenih sa $A = A_1, A_2, \dots, A_k$ i m klasnih oznaka $C = C_1, C_2, \dots, C_m$. Koristeći Bajesovu teoremu, Näive Bayes klasifikator izračunava verovatnoću svake klase za obeležje A i klasna oznaka se dodeljuje sa maksimalnom posteriornom verovatnoćom. Stoga se pokušava postići maksimalna vrednost verovatnoće $P(C_i|A)$ =$P(A|C_i)P(C_i)$ / P(A). Međutim kako je P(A) konstantno za sve klase, samo izraz $P(A|c_i)P(c_i)$ je potrebno maksimizovati. Ukoliko prethodne verovatnoće klase nisu poznate, pretpostavlja se da su klase jednako verovatne, tj. $P(C_1) = P(C_2) = \cdots = P(C_m)$ pa treba maksimizovati $P(A|C_i)$. U suprotnom maksimizujemo $P(A|C_i)P(C_i)$. Prethodna verovatnoća klase može se proceniti sa $P(C_i)=^{s_i}/_{S}$ gde je s_i broj instanci sa klasnom oznakom C_i i s ukupan broj instanci obučavajućeg skupa. NB pretpostavka klasne nezavisnosti podrazumeva uslovnu nezavisnost obeležja obzirom na klasnu oznaku, odnosno između obeležja nema odnosa zavisnosti. Ako je $A_k = a_k$ obeležje kategorijskog tipa tada $P(a_k|C_i)$ predstavlja odnos broja instanci klasne oznake C_i koje imaju vrednost a_k za obeležje A_k i ukupnog broja instanci sa klasnom oznakom C_i. U slučaju da je A_k kontinualno obeležje verovatnoća $P(a_k|C_i)$ se može izračunati primenom Gausove funkcije gustine. Pretpostavka o uslovnoj nezavisnosti obeležja podrazumeva da se verovatnoća klase $P(A|C_i)$ može izračunati na osnovu proizvoda dimenzionih verovatnoća i data je jednačinom (2.19):

$$P(A|C_i) = P(A = A_1, A_2, \dots, A_k|C_i) = \prod_{j=1}^{k} P(A_j|C_i), \quad za\ i=1,\dots,m \tag{2.19}$$

Jednostavanost, efikasanost, laka interpretacija, pogodnost za male skupove podataka osnovne su prednosti Näive Bayes modela. U obrazovnom domenu pretpostavka o uslovnoj nezavisnosti se često ignoriše i narušava. S obzirom da su promenljive međusobno povezane, Näive Bayes klasifikator može da toleriše snažne iznenađujuće zavisnosti između nezavisnih promenljivih. Smatra se da ovi klasifikatori mogu da nadmaše sofisticiranije kao što su stabla odluke i opšte Bajesovske klasifikatore, posebno u slučajevima skupova podataka sa manjim brojem zapisa. [75].

U slučaju primene skrivenih Naïve Bayes klasifikatora (engl. *Hidden Naïve Bayes*, HNB) [76], za svako obeležje kreira se skriveni roditeljski čvor u kome se kombinuju međusobni uticaji ostalih obeležja. Skriveni roditelj može se posmatrati kao agregiranje uticaja svih ostalih obeležja tako da se snažnijim uticajima dodeljuju veći težinski faktori. Eksplicitna

semantika HNB model čini razumljivijim i lakšim za upotrebu. Od Naïve Bayes mreže, nasleđuje jednostavnost strukture, pa je obučavanje HNB klasifikatora prilično olakšano i generalno se odnosi na procenu obeležja određivanjem pripadajućih težinskih vrednosti. Izračunavanje težinskih faktora je od presudnog značaja i realizuje se direktnim izračunavanjem procenjenih vrednosti iz podataka ili pristupom pretrage zasnovane na metodi unakrsne validacije. Algoritam HNB modela prikazan u nastavku zasnovan je na pristupu direktnog izračunavanja težina $W_{ij},\ i,j = 1..n, i \neq j$ koje pokazuju agregiranje uticaja obeležja. Uslovna međusobna zavisnost obeležja A_i, A_j označena je vrednošću mere zajedničke informacije I_p kao težina $P(A_i, A_j|c)$.

Algorithm *HNB*
Input: *D – set of training instances*
Output: *HNB model for D set*
for *each class value c of C*
Calculate $P(c)$
for *each pair of features* A_i, A_j
for *each value* a_i, a_j, c *of* A_i, A_j *and class C*
Calculate $P(a_i, a_j|c)$
for *each pair of features* A_i, A_j
Calculate $I_p(A_i, A_j|c)$
for *each feature* A_i
Calculate $W_i = \sum_{j=1, j\neq i}^{n} I_p(A_i, A_j|c)$
for *each feature* A_j *and* $j \neq i$
Calculate $W_{ij} = \frac{I_p(A_i, A_j|c)}{W_i}$

Prednost ovog modela je mogućnost obučavanja bez utvrđene strukture učenja. Uzimajući u obzir jednostavnost i razumljivost, HNB se može smatrati značajnim unapređenjem Naïve Bayes mreže i obećavajućim modelom koji ima veću primenu u praksi.

TAN Näive Bayes klasifikator [77] predstavlja proširenje Näive Bayes modela dajući i mogućnost dodatnih zavisnosti. Struktura TAN modela je ista kao i kod Näive Bayes mreže jedino što čvorovi listova mogu biti međusobno zavisni, pored zavisnosti prema klasnom obeležju. TAN model je često dobar kompromis između Näive Bayes i opšte Bajesove mreže. Struktura modela je dovoljno jednostavna da se izbegne prenaučenost ali bi trebalo uzeti u obzir i jaku zavisnost između obeležja.

AODE [78] (engl. *Aggregating One-Dependence Estimators*) je Bajesovski metod koji ostvaruje veoma preciznu klasifikaciju izračunavanjem proseka nad prostorom alternativnih Bajesovih modela koji imaju slabije, a samim tim i manje štetne nezavisne pretpostavke nego Näive Bayes. Rezultujući algoritam je računski efikasan uz istovremeno ostvarivanje preciznije klasifikacije u odnosu na Näive Bayes metod nad skupovima podataka sa karakteristikama koje nisu nezavisne.

2.4.5. Metod najbližih suseda

Metod najbližih suseda [78] (engl. *K-nearest neighboor*) spada u lenje metode učenja (engl. *lazy learner*). Metod učenja instancama predstavlja drugačiji pristup klasifikaciji pamteći primere i generalizuje ih tek unosom nove instance što je do tada odloženo. Ovi klasifikatori ne izgrađuju eksplicitni model. Model predstavljaju podaci koji se koriste u procesu klasifikacije. Osnovna ideja je klasifikacija novog objekta ispitivanjem vrednosti klase u K najsličnijim tačkama podataka. Izabrana klasa može biti najčešća klasa među susedima ili klasa distribucije u susedstvu. Zadatak obučavanja klasifikatora K-najbližih suseda odnosi se na izbor dva parametra, broj suseda K i metrika rastojanja d. Podešavanje parametra K i d je težak posao. Ako je postavljen mali broj suseda, klasifikacija se zasniva na samo nekoliko tačaka podataka. U tom slučaju dobijamo nestabilan klasifikator jer ovih nekoliko suseda može dosta da varira. S druge strane, ako je postavljen veliki broj suseda onda najčešće klase u susedstvu mogu mnogo odstupati od stvarne klase. Preporuka je da se za mali dimenzionalni skup podataka, broj suseda K postavlja na vrednosti između 5 i 10. Glavna mana je pravilan izbor funkcije rastojanja. Nedostatak eksplicitnog modela može biti prednost a i mana. Ako je model kompleksan, lakše je uskladiti približavanje na lokalnom nivou, a pri dodavanju novih podataka nije potrebno ažurirati klasifikator. Ako je skup analiziranih podataka veći, potrebno je primeniti indeksiranje za efikasno pronalaženje najbližih suseda. U svakom slučaju ova vrsta lenjih metoda učenja je sporija u klasifikaciji od pristupa zasnovanih na modelu. Takođe treba napomenuti da je generisanje eksplicitnog modela korisnije za procenu i analizu sistema.

2.4.6. Mehanizmi kombinovanja klasifikatora u ansambal

Velika pažnja je usmerena određivanju mehanizama kombinovanja klasifikatora u slučaju integrisanja u ansamblu. Pristup koji označava strategiju pakovanja [68] se odnosi na naglašavanje različitih delova skupa tokom obučavanja podataka. Druga strategija podrazumeva stvaranje raznolikosti ne na osnovu obučavajućih skupova, nego od samih

klasifikatora. Predloženo je nekoliko tehnika za merenje raznolikosti između klasifikatora [80]. Razmatrane su korelacije različitosti klasifikatora u ansamblu i tačnost koju ansambl može postići [81, 82]. Neka od istraživanja su usmerena na razvoj strategije pretrage za dinamičko otkrivanje skupa klasifikatora primenljvih za određenu problematiku zadatka [83, 84]. Glasanje (engl. *voting*) predstavlja mehanizam kombinovanja odluka više klasifikatora zasnovan na tehnici agregacije. Odluka je zasnovana na pluralizmu ili većinskom glasanju, svaki pojedinačni klasifikator doprinosi sa jednim glasom [85]. Predviđanje agregacijom predstavlja donošenje odluke na osnovu većine glasova, klasa sa najvećim brojem glasova predstavlja predviđanje. Konačna predikcija se zasniva na sumiranju svih glasova i izboru klase sa najvećom vrednošću agregacije.

Glasanje može biti realizovano konsenzusom, konsenzusom osim uzdržanih, prostom većinom glasova, kvalifikovanom većinom. Konsenzus predstavlja tehniku glasanja kojom se odluka o pripadnosti instance klasi *Cj* donosi ako i samo ako je to odluka svih klasifikatora integrisanih u ansamblu. Konsenzus osim uzdržanih predstavlja strategiju donošenja odluke o pripadnosti date instance odgovarajućoj klasi *Cj* ako i samo ako nijedan klasifikator ne klasifikuje navedenu instancu u neku drugu klasu različitu od klase *Cj*. Prosta većina glasova podrazumeva donošenje odluke o pripadnosti instance klasi *Cj* ako i samo ako je to odluka unapred definisane većine klasifikatora u ansamblu (više od polovine). Kvalifikovana većina ukazuje na tehniku glasanja kojom se donosi odluka o pripadnosti instance klasi *Cj* ako i samo ako je to odluka zadane većine klasifikatora koja je značajno veća od broja klasifikatora koji klasifikuju navedenu instancu u klase različite od klase *Cj*.

Tehnika glasanja koje koriste najčešće primenjivane slabe klasifikatore OneR, Decision Stump, Naïve Bayes opisane su u radu [86]. Prednosti slabih klasifikatora opisane su u radovima [87, 88]. Obzirom da izbegavaju obuku izuzetaka, za prednosti slabih klasifikatora navodi se manja verovatnoća pojave problema prenaučenosti ili moguće granice pojave nedoslednosti i šuma u podacima. Vreme obučavanja je često kraće u slučaju generisanja klasifikatora ansambla.

2.5. Pravila udruživanja

Pravila udruživanja (engl. *association rules*) zasnovana su na pronalaženju interesantnih odnosa između stavki u analiziranom obučavajućem skupu podataka [89].

Problem otkrivanja pravila udruživanja opisan je i u radu [90]. Neka je $I = \{i_1; i_2; i_3; \ldots; i_n\}$ skup stavki, a D skup svih transakcija u relaciji. Asocijativno pravilo je ako-onda izraz (IF-THEN) ili implikativna forma $X \Longrightarrow Y$, Y gde je $X \subset I$, $Y \subset I$ i $X \cap Y = \emptyset$, odnosno prethodnik X i sledbenik Y su skupovi koji nemaju zajedničkih stavki. Značenje pravila udruživanja se ogleda u tome da ako je sledbenik X zadovoljen, onda je vrlo verovatno da će i sledbenik Y biti zadovoljen. Procena generisanih pravila udruživanja zasniva se na postavljanju donjeg praga vrednosti parametara *minsup* i *minconf* i izračunavanju osnovnih mera podrške (*support, sup, s*) i poverenja (*confidence, conf, c*).

Podrška (*support, sup*) pravila udruživanja predstavlja procenat transakcija koje uključuju stavke X i Y iz D. Izračunava se po formuli datoj u jednačini (2.20). Može se posmatrati kao verovatnoća da transakcija sadrži stavke iz skupa X i stavke iz skupa Y odnosno ima osobinu simetričnosti što je prikazano jednačinom (2.21).

$$sup(X \Rightarrow Y) = \frac{|X,Y|}{n} \tag{2.20}$$

n - broj transakcija u skupu D,

$\mathrm{n} = |\mathrm{D}|$,

$|X,Y| = |\{\mathrm{X} \cup \mathrm{Y} \subseteq \mathrm{T}, \mathrm{T} \in \mathrm{D}\}|$ broj transakcija koje sadrže stavke iz skupa X i skupa Y.

$$sup(X \Rightarrow Y) = sup(Y \Rightarrow X) \tag{2.21}$$

Poverenje (engl. *confidence*) pravila $X \rightarrow Y$ je udeo transakcija koje sadrže stavke skupa Y među transakcijama koje sadrže stavke skupa X i izračunava se po formulama prikazanim jednačinama u (2.22) i (2.23).

$$conf(X \Rightarrow Y) = \frac{|X,Y|}{|X|} \tag{2.22}$$

$$P(X) = \frac{|X|}{n} \tag{2.23}$$

$|X|$ broj transakcija koje sadrže stavke iz skupa X.

Poverenje se može definisati kao uslovna verovatnoća odnosno da transakcija sadrži Y znajući da već sadrži i X. Ova mera nema osobinu simetričnosti što daje pravac pravila udruživanja.

Pored podrške i poverenja, značajna je i lift metrika [91]. Lift parametar predstavlja odnos verovatnoće prethodnika i sledbenika koji se dešavaju zajedno i verovatnoće prethodnika i sledbenika koji se dešavaju samostalno. Lift parametar može se predstaviti formulom prikazanom jednačinom (2.24).

$$lift\ (X \Rightarrow Y) = \frac{P(X,Y)}{P(X)P(Y)} = \frac{sup(X \cup Y)}{sup(X)sup(Y)} = \frac{conf(X \Rightarrow Y)}{sup(Y)} \tag{2.24}$$

Kao što se može videti iz jednačine (2.23), parametar lift se definiše za pravilo $X \Longrightarrow Y$. Označava meru značaja generisanog pravila udruživanja i odstupanje od statističke nezavisnosti između prethodnika i sledbenika. U pogledu verovatnoće, to znači da pojava X i pojava Y u istoj tranasakciji su nezavisni događaji, odnosno da prethodnik i sledbenik nisu u korelaciji. Vrednost lift mere manja ili jednaka 1 ukazuje na pravila bez značaja jer se sledbenik češče pojavljuje od prethodnika. Slučaj kada je vrednost lift mere veća od 1 ukazuje na pozitivnu korelaciju između prethodnika i sledbenika. Ova pravila se označavaju kao značajna za predviđanje posledica u skupovima podataka. Metrike poverenje i podrška se uobičajeno koriste u proceni pravila udruživanja. Međutim, čak i pravila sa jakom podrškom i poverenjem mogu biti bez značaja za analiziran problem. Zbog toga je potrebno uzeti u obzir jačinu povezanosti X i Y primenom drugih metrika. Postoji veliki broj mera predloženih u literaturi. Međutim, iako se ne može reći da je jedna mera bolja od druge za različite situacije, u slučaju visoke vrednosti podrške moguće je podudaranje rezultata dobijenih primenom različitih mera.

Generisanje pravila udruživanja može da se podeli u dva zadatka. Prvi zadatak je pronalaženje onih skupova stavki čija verovatnoća dešavanja dostiže predhodno definisani prag podrške i nazivaju se frekventni skupovi stavki. Drugi zadatak je generisanje pravila udruživanja iz predhodno dobijenih velikih skupova stavki, a koja su ograničena sa minimalnom podrškom.

Apriori algoritam [92] je standardni metod za pametno pretraživanje prostora pravila udruživanja iz skupa transakcija. Neophodno je po staviti minimalne vrednosti za parametre

podrške i poverenje. Pronalazi sva pravila koja zadovoljavaju postavljene uslove tako što iterativno smanjuje minimalnu podršku dok ne pronađe pravila koja imaju poverenje jednako ili veće od postavljenog minimalnog poverenja.

Pseudokod Apriori algoritma je:

```
procedure Apriori (T, minSupport)
{//T=baza podataka transakcija t, minSupport je minimalna podrška za L1= frekventne stavke;
    for (k= 2; Lk-1 !=∅; k++)
      { Ck= candidates generated from Lk-1
      //kartezijski proizvod Lk-1 x Lk-1 i uklanjanje svih stavki k-1 veličine koje nisu česte //;
        for each transaction t in database do
          {
           #increment the count of all candidates in Ck that are contained in t
            Lk = candidates in Ck with minSupport }
        //end for each
          }
  }//end for
return ∪k Lk;
}
```

U prvoj fazi, Apriori algoritam koristi velike skupove L_{k-1} koji se sastoje od k-1 stavki pronađene u k-1 koraku, za stvaranje skupa kandidata C_k. U drugoj fazi se utvrđuje značaj svakog pojedinog kandidata iz skupa kandidata C_k. Kandidati koji imaju značaj veći od minimalnog i čiji se svi podskupovi nalaze u L_{k-1}, postaju elementi skupa L_k . Interesantnost pravila zavisi i od samog problema koji treba da se reši. U mnogim slučajevima, korisniku je potrebno da pronađe pravila kod kojih postoji veza u slučaju realnog okruženja.

Prediktivni Apriori algoritam [93] generiše n pravila za koje može biti postignuta maksimalna tačnost predikcije. Neophodno je unapred definisati broj pravila (n), ali nije potrebno postavljati vrednosti za minimalnu podršku i poverenje kao u slučaju Apriori algoritma. Značajnost Prediktivnog Aprirori algoritma je u činjenici da je u pitanju dinamička tehnika određivanja pravila koja koriste gornje granice tačnosti svih pravila.

3. PREGLED ISTRAŽIVANJA U OBLASTI RUDARENJA PODATAKA U OBRAZOVNOM PROBLEMSKOM DOMENU

Od 1995. godine, oblast rudarenja podataka u obrazovnom problemskom domenu razvija se velikom brzinom kao interdisciplinarno istraživačko područje koje se bavi razvojem metoda rudarenja podataka iz različitih obrazovnih okruženja. Za potrebe sprovedenog istraživanja, koje je opisano u ovoj disertaciji, pronađen je veliki broj srodnih referenci. Izdvojene su i analizirane relevantne reference koje se odnose na osnovne zadatke primene prediktivnih i deskriptivnih metoda i u kojima su prikazani eksperimenti sa realnim obučavajućim skupom podataka.

Baker i dr. [94] izdvajaju četri ključna zadatka primene rudarenja podataka u obrazovnom okruženju: poboljšanje modela studenta, poboljšanje modela domena, analiza sistema i softvera za učenje, analiza načina ponašanja studenata/učenika tokom procesa učenja. Primena dubinske analize obrazovnih podataka koja se odnosi na poboljšanje modela studenta obuhvatala je postupak utvrđivanja obrazaca ponašanja u sistemima za učenje, proces otkrivanja faktora koji predviđaju neuspeh studenata, analiziranje lošijeg uticaja na proces učenja, utvrđivanje nezainteresovanosti studenta. Druga ključna primena je usmerena na otkrivanje i unapređenje domena modela znanja. Kombinovanjem psihometrijskog modelovanja okruženja sa algoritmima pretrage mašinskog učenja, veliki broj istraživača je razvio pristup za otkrivanje domena modela, direktno iz podataka. Neki od istraživača razvili su algoritme za pronalaženje delimično uređenih modela strukture znanja koji objašnjavaju međusobne interakcije u okviru modela. Treća ključna primena odnosi se na izučavanje pedagoške podrške u obrazovnom softveru ili u drugim sistemima za učenje i utvrđivanje koji je tip pedagoške podrške najefikasniji za različite grupe studenata u različitim situacijama. Četvrta ključna primena je usmerena na pronalaženje empirijskih dokaza koji utiču na poboljšanje i proširenje obrazovnih teorija, kako bi se što bolje razumeli ključni faktori ponašanja studenata, a sve u cilju poboljšanja efikasnosti procesa učenja.

Kategorizacija primene dubinske analize u obrazovnim okruženjima može se izvršiti prema ciljevima usmerenim ka različitim korisnicima i na osnovu najčešće implementiranih metoda. Prema ciljevima fokusiranim ka korisnicima, primene metoda rudarenja podataka moguće je svrstati u sledećih pet kategorija:

- Student: ciljevi primene odnose se na otkrivanje efikasnih sekvenci u procesu učenja, izdvajanje korisnih resursa, aktivnosti koje utiču na konačnu ocenu, utvrđivanje materijala za učenje koje treba poboljšati, predlaganje interesantnih procesa i okruženja učenja zasnovanih na pozitivnim iskustvima studenata, personalizaciju okruženja učenja;
- Nastavnik: ciljevi primene odnose se na obezbeđivanje povratnih informacija, klasifikaciju studenata na osnovu ponašanja i potreba, utvrđivanje efikasnih obrazaca učenja, detektovanje studenata sa tendencijom odustajanja, unapređenje procesa učenja i strukture sadržaja kursa, izdvajanje studenata koji zahtevaju podršku; predviđanje ostvarenih rezultata; pronalazak najčešćih grešaka;
- Kreator kursa: ciljevi primene odnose se na komparativne analize efikasnosti rezultata metoda i tehnika u cilju utvrđivanja najpodesnijih za svaki zadatak; razvoj specifičnih alata i okruženja za realizaciju dubinske analize u obrazovne svrhe.
- Administrator: ciljevi primene odnose se na efikasniju upotrebu raspoloživih resursa; poboljšanje obrazovnih programa; efikasniji pristup učenju na daljinu i ostalim okruženjima za e-učenje.

Kategorizacija primene metoda rudarenja podataka na obrazovna okruženja data je u [95]. Navedene su opšte kategorije koje su u velikoj meri priznate kao univerzalne i u ostalim oblastima primene i to: klasifikacija, grupisanje, predviđanje, pravila udruživanja, neuronske mreže, stabla odlučivanja i metoda najbližih suseda.

Postoje dve glavne taksonomije kategorizacije tehnika rudarenja podataka u obrazovnim sistemima: Romero-Ventura [12] i Baker-Yassef [13]. Sa slike 3.1 može se videti da su neke od kategorija povezane, dok neke nisu u vezi ni sa jednom od ostalih navedenih.

Slika 3.1: EDM taksonomija [96]

Romero-Ventura taksonomija obuhvata četri kategorije i više je fokusirana na primeni obrazovnog otkrivanja znanja na web podatke. Iako se kategorija "statistika i vizualizacija" ne mogu formalno smatrati metodama rudarenja podataka, autori ih uključuju u kategorizaciju jer označavaju polaznu tačku većine studija. Druga i treća kategorija predstavljaju uobičajne zadatke većine projekata u oblasti rudarenja podataka, dok se četvrta kategorija može smatrati proširenjem na tekstualno struktuirane podatke i bliska je sa pojmom rudarenja Web sadržaja (engl. *Web content mining*). Baker-Yassef's taksonomija pokriva Maimon i Rokach [97] DM metode (druga, treća i čevrta kategorija) i uvodi dve dodatne. Prva odgovara Romero-Ventura kategoriji "Statistika i vizualizacija", a peta se smatra najmanje zastupljenom u istraživanjima.

Sistematska analiza sprovedenih istraživanja u oblasti rudarenja obrazovnih podataka prikazana je u radu [12]. Autori predstavljaju najrelevantnije studije sprovedene u ovoj oblasti u periodu između 1995. do 2005. godine. Donose zaključak da je oblast obrazovnog rudarenja podataka obećavajuće područje za istraživanje. Pored toga navode i važne razlike u načinu primene na obrazovne podatke u odnosu na podatke iz drugih okruženja. Autori [98] predstavljaju istraživanje studija sprovedenih do 2010.godine. Opisuju primene u zavisnosti od grupe korisnika, vrste obrazovnih okruženja i podataka koji su na raspolaganju. Navodi i najčešće zadatke rešavane tehnikama za rukovanje podacima. U radu [99] istražuje različite pristupe i tehnike za rukovanje podacima koji se mogu primeniti na obrazovne podatke kako

bi se izgradilo novo okruženje za predviđanje. Ova studija takođe razmatra primene tehnologija obrade velikih količina podataka (engl. *big data*) u obrazovanju i predstavlja pregled literature o istraživanju obrazovnih podataka i procesa učenja. Dutt i saradnici [100] prikazuju analizu relevantne literature za period od 1983. do 2016.godine, primenljivosti i upotrebljivosti u kontekstu rudarenja obrazovnih podataka.

Analiziranjem dostupne literature iz oblasti rudarenja obrazovnih podataka sa ciljem otkrivanje znanja o uticaju upotrebe različitih materijala i resursa tokom procesa učenja, izdvojene su relevante reference sa najvećim brojem citiranosti. Izdvajanje informacija o materijalu koji je student koristio tokom procesa učenja, u radu [101] autori zasnivaju na primeni odgovarajućih metoda na podatke dnevnika zapisa i baza podataka sistema za učenje. Beck i Woolf [102] opisuju primenu metoda rudarenja podataka u obrazovnom problemskom domenu za razvoj modela studenta.

U radu [103] autori prikazuju izdvajanje upotrebljenih resursa tokom procesa učenja. Analiza upotrebe diskusionih foruma i interakcije studenta učešćem u diskusijama opisana je u [104]. Merceron i Yacef [105] prikazuju kako upotreba metoda rudarenja podataka može da pomogne za otkrivanje pedagoških informacija iz baze podataka obrazovnih Web sistema. Otkrivene informacije i znanje su od izuzetnog značaja za nastavnike kako bi bolje razumeli različite obrasce ponašanja studenata u procesu učenja i postupili u skladu sa tim. U radu [106] prikazana je optimizacija sadržaja portala za elektronsko učenje. Prikazane su osnovne faze predprocesiranja podataka, otkrivanja obrazaca ponašanja i analize otkrivenih obrazaca. Obučavajući skup podataka kreiran je od podataka prikupljenih iz dnevničkih datoteka skladištenih na serveru

Á.Horváth i dr. [107] prikazuju primenu metode rudarenja podataka korišćenu za komercijalne sajtove u cilju pronalaženje ključnih kupaca i pobošanja efikasnosti e-prodavnica. Obučavajući skup kreiran je na osnovu podataka o realizovanim aktivnostima korisnika koje su skladištene u dnevniku zapisa Moodle sistema za upravljanje učenjem i ostvarenih ocena. Identifikovanjem ponašanja studenata utvrđene su uspešne strategije učenja. Rezultati ove studije pokazuju značaj primene metode dubinske analize podataka za procenu kvaliteta i korisnosti e-learning sistema. U radu [108] autori analiziraju specifična ponašanja studenata u *on-line* kursu na Moodlu sistemu. Za izgradnju modela koji prati ponašanje studenata korišćene su veštačke neuronske mreže i metoda vektora oslonca. Obučavajući skup je kreiran na osnovu

podataka prikupljenih iz dnevničkih datoteka. Predložen model predviđanja ukazuje na uspešnost studenata i generiše informacije od značaja za proces učenja u toku realizacije kursa. Castro i dr. [109] predlažu primenu dubinske obrazovne analize u postupku procene efikasnosti učenja studenata. Preporuku odgovarajućeg nivoa *on-line* kursa autori zasnivaju na analizi ponašanja i generišu povratnu informaciju za nastavnike i studente. Predlažu različite pristupe za procenu materijala za učenje i detektuju karakteristično i drugačije ponašanje studenata koje se izdvaja od ostalih. Analiza ponašanja studenata i otkrivanje različitih web stranica koje posećuju tokom procesa učenja prikazana je u [110].

Romero i Ventura u radu [24] detaljno prikazuju faze primene dubinske analize podataka Moodle sistema za upravljanje učenjem. Opisan je kompletan proces "rudarenja" podataka, primena glavnih tehnika statistike, klasifikacija, grupisanje, pravila udruživanja, vizualizacije upotrebom besplatnih alata za rudarenje podataka. Koriste metod grupisanja za razlikovanje aktivnih i neaktivnih studenata u skladu sa stepenom izvršavanja predviđenih aktivnosti.

3.1. Primena metoda rudarenja na obrazovne podatke

Predviđanje performansi studenata u okruženju daljinskog učenja opisano u radu [86] realizovano je sprovođenjem eksperimentalne uporedne analize klasifikacijskih algoritama Naïve Bayes, stabla odlučivanja C4.5, veštačkih neuronskim mreža, metode vektora oslonca, metode najbližih suseda i logističke regresije. Obučavajući skup je kreiran na osnovu demografskih podataka, rezultata realizovanih zadataka, informacija o učešće na grupnim sastancima. Obučavajući skup sadržao je 354 zapisa o studentima. Podaci su bili numerički i kategorijalni. Model najboljih performansi formiran je od strane Naïve Bayes klasifikatora.

U radu [111] autori implementiraju veštačke neuronske mreže za previđanje konačne ocene studenata klasifikovanjem u pet klasa (B, C, D, E, F). Delgado i dr. u radu [112] koriste metode neuronskih mreža radijalnih bazičnih funkcija za kreiranje modela predviđanje uspešnosti studenata (položio/pao) analiziranjem podataka Moodle dnevnika. U radu [113] autori Wang i Mitrović opisuju sistem za predviđanje grešaka studenata zasnovan na metodama neuronskih mreža.

Bajesove mreže korišćene su u slučaju modelovanja znanja i previđanja performansi studenata u okviru tutorskog sistema za podučavanje [114]; za previđanje proseka ocena

diplomcima na osnovu podataka datih prilikom prijave na fakultet [115]; za predviđanje performansi grupe studenata u klasičnom kolaborativnom okruženju za učenje [116].

U radu [117] opisana je primena metoda rudarenja podataka za analizu i procenu kvaliteta Moodle kursa. Autori prikazuju tip izdvojenih podataka, procesiranje podataka, postupak kreiranja obučavajućeg skupa za primenu metoda rudarenja i primenljivost otkrivenog znanja. Autori [118] u radu prikazuju upotrebu metode stabla odlučivanja za predviđanje uspešnosti studenata i blagovremeno obezbeđivanje materijala u formi nastavnih lekcija u okruženju web sistema za realizaciju e-učenja.

Autori u radu [119] opisuju eksperiment primene tri popularne metode mašinskog učenja za predviđanje studenata koji odustaju od programa e-učenja. Implementirane su veštačke neuronske mreže sa propagiranjem signala unapred, metoda vektora oslonca i pojednostavljeni probabilistički rasplinuti ansambl ARTMAP. Metode su ispitane u smislu ukupne tačnosti, osetljivosti i preciznosti, a rezultati su bili znatno bolji od navedenih u relevantnoj literaturi. U radu [120] predložen je model predviđanja konačnih ocena studenata u kontekstu višeg obrazovanja na osnovu prikupljenih informacija iz baze podataka. Klasifikacija je realizovana metodom stabla odlučivanja, a zasnovana je na podacima o ocenama i aktivnostima studenata ostvarenim u prethodnom semestru. Model predviđa konačne ocene, studente koji imaju tendenciju da odustanu, ukazuje da timski i grupni rada studenta ima značajan uticaj na predviđanje ostvarenih performansi u procesu učenja.

Istraživanje opisano u radu Dekker i dr. [121] ima za cilj predviđanje studenata koji odustaju od studiranja nakon prvog semestra i identifikaciju faktora uspeha specifičnih za analizirani program. Obučavajući skup za analizu je kreiran od podataka o brucošima prikupljenim tokom perioda od 2000. do 2009. godine na Tehnološkom univerzitetu Eindhoven, Holandija. Izvršena je komparativna studija poređenja rezultata ostvarenih primenom algoritma stabla odlučivanja, Bajesovog klasifikatora, logističkog modela, algoritma Random Forest i OneR klasifikatora. Kao indikator prediktivnog značaja određenih obeležja posmatran je OneR klasifikator. Eksperimentalni rezultati su ukazali na algoritme stabla odlučivanja sa ostvarenom preciznošću između 75 i 80%. Zaključeno je da se poboljšanje tačnosti predviđanja može ostvariti bez potrebe za prikupljanjem dodatnih podataka o studentima.

Dimić i dr. u radu [122] prikazuju studiju uporedne analize metoda Bajesovih klasifikatora, stabla odlučivanja, najbližih suseda implementiranih nad obrazovnim skupom Moodle podataka. Procena formiranih modela izvršena je izračunavanjem statističkih mera TP (eng. *True Positive Rate*), FP (eng. *False Positive Rate*), F-mere (eng. *F-measure*), preciznosti (eng. *Precision*), i matrice grešaka (eng. *Confusion Matrix*). Na osnovu rezultata sprovedenog istraživanja, autori zaključuju da metode stabla odlučivanja generišu model najveće tačnosti predviđanja.

Kovačić u radu [123] istražuje demografske karakteristike kao što su: godine, pol, etnička pripadnost, obrazovanje, radni status, invaliditet i okruženje u kome studenti studiraju, a koje može uticati na odustajanje. Skup za analizu kreiran je od podatke podataka sakupljenih u periodu od 2006. do 2009. godine o 450 studenata programa Fakulteta "*Open Polytechnic*" Velington, Novi Zelend. Pristup se sastojao od unakrsne tabularne analize, odnosno tabele za nepredviđene situacije, izbora obeležja i predviđanja realizovanih sa četri različita algoritma stabla odlučivanja. Zaključeno je da su najvažniji faktori koji razdvajaju uspešne od neuspešnih studenata etnička pripadnost i program. Razmatranjem primenjene metodologije, ističe se da klasifikacija i regresijsko drvo (*CART*) generiše najuspešniji model sa ukupnim procentom tačnosti klasifikacije od 60,5%.

Autori [124] identifikuju potencijalne probleme kako bi se sprečilo odustajanje studenata od nastavka studija. Kreiran je sistem koji uključuje metode za dubinsku analizu podataka i to: Naïve Bayes, metoda vektora oslonca i stabla odlučivanja. Obučavajući skup kreiran je na osnovu podataka sakupljenih sa Univerziteta Thames Valley iz Engleske. Sistem je pratio i analizirao ponašanje studenata sa ciljem obezbeđivanja efikasnih strategija za poboljšanje procesa učenja. Autori zaključuju da Naïve Bayes klasifikator ostvaruje najveću tačnost predviđanja koja prevazilazi ostvarene rezulate postignute primenom vektora podrške i stabla odlučivanja.

Autor u radu [125] prikazuje visok potencijal primene rudarenja podataka u cilju poboljšanja upisne kampanje. Istraživanje je bilo fokusirano na razvoj modela predviđanja performansi studenata na osnovu ličnih i podataka o prethodnom obrazovanju. Implementirano je više klasifikatora i to: stabla odlučivanja, veštačke neuronske mreže, klasifikator zasnovan na pravilima za obučavanje (engl. *rule based learner classsificator*) i metoda najbližih suseda. Skup za analizu kreiran je od podataka o studentima upisanim na Univerzitet za nacionalnu i

svetsku ekonomiju u Sofiji, Bugarska, u tri uzastopne godine. Na osnovu rezultata je zaključeno da je najveća tačnost predviđanja postignuta metodom neuronskih mreža. Stabla odlučivanja i metoda najbližih suseda ostvarili su nešto lošiju preciznost. Obeležja koja se odnose na rezultat prijemnog i broj neuspešnih prethodnih prijemnih ispita su jedan od faktora koji najviše utiče na proces klasifikacije.

Hämäläinen i dr. u radu [126] sprovode eksperimente poređenja pet klasifikatorskih algoritama za studiju slučaja malog obučavajućeg skupa podataka Inteligentnog Tutorskog Sistema. Za prikaz ostvarenog uspeha na kursu definisano je klasno obeležje sa oznakama pao ili položio. Metoda vektora oslonca i logistička regresija implementirani su za numeričke podatke, a za nominalne podataka tri varijante Naïve Bayes klasifikatora. Autori ukazuju na Naïve Bayes klasifikator kao najpodesniji za slučaj malog obučavajućeg skupa mešovitog tipa podataka. Cocea i Weibelzahl [127] prikazuju testiranje performansi osam klasifikatora za predviđanje angažovanja studenata u virtuelnim kursevima. Dva obučavajuća skupa izdvojena su iz dnevničkih datoteka i sadržala su 341 i 450 zapisa. Sva obeležja, osim klasne promenljive, bila su numeričkog tipa. Rezultati primenjenih tehnike pokazuju dobar nivo predviđanja, s tim da IBK algoritam daje najpreciznije rezultate.

Romero i dr. [128] sprovede komparativnu studiju primene različitih metoda i tehnika za klasifikaciju studenata na osnovu njihovih podataka o aktivnostima u okviru kursa Moodle sistema za upravljanje učenjem i konačnih ostvarenih ocena. Za sprovedenu studiju utvrđen je najefikasniji klasifikatorski model obrazovnog okruženja generisan implementacijom metode stabla odlučivanja. Za kreirani model zaključeno je da mora biti tačan, precizan i razumljiv za nastavnike kako bi bio od koristi pri donošenju odluka. U radu [129] je prikazano poređenje metoda stabla odlučivanja i Bajesovih mreža za predviđanje prosečne ocene studenata. Rezultati su pokazali da je klasifikator zasnovan na stablu odlučivanja bio bolji od Bajesovih mreža u svim klasama. Tačnost je poboljšana upotrebom tehnika ponovnog uzimanja uzorka posebno u slučaju primene stabla odlučivanja za sve slučajeve klase. Pored toga smanjena je i pogrešna klasifikacija u manjinskim klasama nebalansiranih skupova podataka.

Diego Garcia-Saiz i Marta Zorrilla [130] sprovode eksperimente poređenja performansi i interpretacije izlaza različitih algoritama klasifikacije OneR, J48, Naïve Bayes, BayesNet TAN, NNge. Performanse klasifikatora su testirane na tri obučavajuća skupa i definisanim klasnim obeležjem Mark kao predviđajućom varijablom. Skupovi Mu0910 i Mu0710 sadržala

su 65 i 164 instanci respektivno, a Mu0910S 65 instanci. Ni jedan od implementiranih klasifikatora nije se izdvojio sa značajno ostvarenom tačnošću kreiranog modela. Poboljšanje preciznosti modela za specifične skupove obrazovnih podataka ostvareno je kombinovanjem J48 i Naïve Bayes klasifikatora primenom meta algoritma.

Zang i Lin [131] primenjuju kombinovanje slabih klasifikatora zasnovano na boosting pristupu za predviđanje konačne ocene studenata. Slabi klasifikator koristi samo jedno od 74 obeležja za predviđanje ishoda kursa. Kombinovanjem je postignuto samo 69% tačnosti ali je primenjeni pristup kombinovanja klasifikatora otkrio faktore koji su najviše uticali na uspeh kursa. Ribeiro i Cardoso [132] analiziraju specifično ponašanja studenata u okviru Moodle kursa. Prikazna je primena klasifikatora zasnovanih na veštačkim neuronskim mrežama i metodi vektora oslonca nad podacima izdvojenim iz dnevničkih datoteka. Generisan je model predviđanja uspešnosti studenata koji prikazuje korisne povratne informacije tokom odvijanja kursa.

Dimić i dr. [133] prikazuju postupak kombinovanja metoda za izbor optimalnog vektora obeležja i tehnika ponovnog uzorkovanja obučavajućeg obrazovnog skupa podataka. Rezultati opisane studije ukazuju na uticaj izbora obeležja i veličine obučavajućeg skupa na tačnost modela predviđanja. Thai-Nghe, N. i dr. [134] porede osnovne klasifikatore: stabla odlučivanja, Bajesove mreže i metodu vektora oslonca. Problem degradacije klasifikatora rešavaju uzimanjem u obzir disbalansa klase. Poboljšanje rezultata predviđanja ostvaruju upotrebom *over-sampling* tehnika i *cost-sensitive* učenja (engl. *cost-sensitive learning*, CSL).

Santana i dr. [135] sprovode komparativnu studiju efikasnosti obrazovnih tehnika otkrivanja znanja za rano predviđanje neuspeha studenata u slučaju kurseva sa tematikom koja se odnosila na Uvod u programiranje. Osnovni cilj istraživanja je bio usmeren ka smanjivanju stope neuspeha. Rezultati su ukazali da model zasnovan na metodi vektora oslonca ostvaruje značajnu tačnost ranog predviđanja.

Dimić i dr. u radu [136] identifikuju metodologiju za poboljšanje efikasnosti nenadzirane metode diskretizacije obrazovnog obučavajućeg skupa podataka. Autori prikazuju značaj faze predprocesiranja dovodeći u vezu poboljšanje nekih metoda klasifikacije sa postupkom diskretizacije i rebalansiranjem obučavajućeg skupa podataka.

Ge i dr. [137] predstavljaju istraživanje zasnovano na poređenju klasifikacijskih algoritama. Predložen je model za utvrđivanje osobina ličnosti studenata. Obučavajući skup sadržao je 79 zapisa o studentima koji još nisu diplomirali i testiran sa tri klasifikatora. Rezultati ukazuju da SVM klasifikator ostvaruje najbolje rezultate sa ostvarenom tačnošću od 72% i balansiranom F-merom od 77%. U radu [138] autori prikazuju sprovedenu komparativnu studiju poređenjem preciznosti šest klasifikatora: Naïve Bayes, stabla odlučivanja, veštačke neuronske mreže sa propagiranjem signala unapred, metoda vektora oslonca, metoda najbližih suseda i logistička regresija. Obučavajući skup za analizu sadržio je 350 instanci zapisa o demografskim podacima, rezultatima zadatka i učešću na grupnim sastancima za svakog studenta. Podaci su bili numeričkog i kategorijalnog tipa. Predložena su dva klasifikatora, Naïve Bayes i veštačke neuronske mreže, koji su bili u stanju da predvide slučajeve odustajanja sa 80% tačnosti.

Autori [139] implementiraju Naïve Bayes klasifikator, Bajesovski metod *Aggregating One-Dependence Estimators*, stablo odlučivanja i metod vektora oslonca na obučavajuće obrazovne skupove različite kardinalnosti sa ciljem identifikovanja efikasne metode za izbor optimalnog vektora obeležja u mešanom okruženju učenja. U radu [87] autori predlažu mehanizam kombinovanja klasifikatora u ansambal primenom lokalizovanog glasanja slabih klasifikatora. Razvijen model pokazuje veću preciznost i tačnost predviđanja.

3.3. Primena pravila udruživanja

U obrazovnim sistemima metode pravila udruživanja primenjuju se za identifikaciju nekompatibilnih karakteristika između različitih grupa učenika [140], za pronalaženje studenatskih grešaka koje se javljaju zajedno [141], za optimizaciju sadržaja e-učenja određivanjem šta najviše interesuje korisnika [142], za otkrivanje značajnih veza između obrazovnih resursa i učenja studenata

U cilju poboljšanja kurseva e-učenja i omogućavanje saradnje nastavnika u podeli otkrivenih informacija García i dr. [144] predlažu sistem za dubinsku analizu obrazovnih podataka zasnovan na metodama pravila udruživanja. Sistem je opisan na nivou detaljnih instrukcija za korišćenje. Kreiran je za upotrebu od strane stručnih lica, a nastavniku je ostavljena mogućnost analize rezultata i donošenja odluka o tome kako poboljšati kurs e-učenja. Merceron i Yacef [91] su pokazali primenu pravila udruživanja na podatke izdvojene za online kurs realizovan na Moodle sistemu. Utvrđene su tipične vrednosti mera cosine i lift

za izdvajanje značajnijih pravila iz obrazovnih podataka arhiviranih u sistemu za upravljanje učenjem. U radu [145] autori prikazuju novi pristup u postupku analize ponašanja studenata u procesu učenja. Opisana je primena pravila udruživanja na podatke koji se zapisuju u dnevničke datoteke korišćenog sistema za upravljanje učenjem.

Dimić i dr. u radu [146] su prikazali implementaciju pravila udruživanja u postupku procene kvaliteta interaktivnih zadataka kao jednog od materijala za učenje u okviru Moodle kursa Programabilna logička kola. U radu [147] prikazana je primena pravila udruživanja za poboljšanje efikasnosti Moodle testova i kursa. Predložen algoritam G3PARM zasnovan je na grammar-guided genetskom programiranju. Rezultati studije pokazali su efikasnost primene pravila udruživanja za otkrivanje značajnih pravila i obrazaca upotrebljivih u procesu unapređenja testova za proveru znanja. Dimić i dr. u radu [148] predlažu primenu asocijativne analize za unapređenje procesa Moodle e-testiranja u mešanom okruženju učenja. Procena značaja izvršena je primenom objektivnih i subjektivnih mera estimacije. Rezultati studije ukazali su na korisne i interesantne obrasce povratnih informacijama koje omogućavaju nastavniku da bolje sagleda koncepte kreiranih testova i odluči na koji način treba da izvrši izmene s ciljem unapređenja postupka e-testiranja.

Na osnovu kritičkog osvrta sprovedene analize navedenih referenci i literature, zaključeno je da rudarenje podataka u obrazovom problemskom domenu doprinosi poboljašnju efikasnosti procesa učenja nezavisno od okruženja i ciljeva primene.

4. DESKRIPTIVNA ANALIZA SKUPA PODATAKA

Podaci predstavljaju glavnu stavku procesa dubinske analize i otkrivanja znanja. Karakteristike podataka, različiti tip i opseg vrednosti značajno utiču na tačnost modela. U ovoj glavi, prikazan je postupak prikupljanja podataka iz različitih izvora. Definisani su osnovni koncepti, tipovi i karakteristike obrazovnih podataka. Obučavajući skup za mešano okruženje učenja kreiran je integrisanjem izdvojenih podataka iz distribuiranih izvora. Priprema obučavajućeg skupa podrazumevala je popunjavanje vrednosti koje nedostaju, ispravljanje nedoslednosti, ublažavanje problema pogrešnih vrednosti identifikovanjem izuzetaka.

U istraživanju opisanom u disertaciji, faza pripreme je proširena postupkom deskriptivne statističke analize obučavajućeg skupa. Utvrđena varijabilnost vrednosti podataka omogućila je postizanje veće efikasnosti u fazi predprocesiranja.

4.1. Originalni skup podataka

Prvi problem prikazanog istraživanja je bio prikupljanje podataka iz nekoliko izvora sa ciljem kreiranja preciznih modela dubinske analize za mešano okruženje učenja.

Originalan skup podataka zasnovan je na informacijama o aktivnostima studenata koji su izabrali predmet Računarska grafika na Visokoj školi elektrotehnike i računarstva strukovnih studija u Beogradu. Sistem za upravljanje učenjem, Moodle, korišćen je kao platforma za realizaciju daljinskog i mešanog okruženja učenja. Model mešanog okruženja učenja (engl. *blended learning enviroment, BLE*) predstavlja mešavinu fleksibilnosti interaktivnih elektronskih materijala i direktnog učešća nastavnika u nastavi. Ovaj program omogućio je podršku klasičnom načinu organizacije nastave obezbeđivanjem dodatnih resursa, aktivnosti i materijala za učenje. Klasični oblik nastave na predmetu Računarska grafika podrazumevao je tri časa predavanja i dva časa laboratorijskih vežbi nedeljno. Na predavanjima, objašnjavani su pojmovi i koncepti iz oblasti računarske grafike, a na vežbama održanim u računarskim laboratorijama studenti su samostalno realizovali praktične zadatke. Zadaci su bodovani u skladu sa procentom uspešno ostvarene realizacije, a osvojeni poeni su

upisivani u *Google Doc Excel* evidenciju. Za razliku od laboratorijskih vežbi, prisustvo na predavanjima nije bilo obavezno. Na kraju svakog predavanja, studenti su mogli da rade kratak test u papirnoj formi i na taj način osvoje određene poene koji su takođe upisivani u Google Doc evidenciju.

U okviru Moodle kursa dostupni su bili različiti statički, interaktivni resursi i aktivnosti kao što su: PDF dokumenti, lekcije sa pitanjima za proveru znanja, audio - video tutorijali, forumi, privatne poruke, domaći zadaci, testovi za samotestiranje kao i zvanične provere znanja. Realizacija elektronskih konsultacija ostvarena je putem javnog komunikacionog kanala - foruma i privatnih Moodle poruka. U okviru foruma, studenti su mogli da postave pitanje, uključe se u diskusiju sa ostalim studentima i nastavnikom. Drugi način elektronskih konsultacija je bio omogućen upotrebom privatnih Moodle poruka koje su studenti mogli da koriste među sobom kao i za konsultacije sa nastavnikom. Video tutorijali i PDF datoteke korišćene su za pripremu laboratorijskih vežbi, a materijal za kolokvijume i završni ispit bio je dostupan u formi Moodle lekcija sa pitanjima za proveru znanja. Studenti su tokom semestra imali pet domaćih zadataka koji su predstavljali testove i zadatake za proveru znanja sa vežbi i predavanja i bili su aktivni do početka ispitnog roka. Testovi su bili podešeni na dva pokušaja rešavanja, a uzimao se u obzir pokušaj sa bolje ostvarenim rezultatom. U toku semestra, studenti su polagali dva kolokvijuma preko kojih je vršena provera znanja iz oblasti tehnologija ulaza, izlaza, geometrijskih transformacija, osnovnih rasterskih algoritama. Završni test obuhvatao je praktične zadatke iz oblasti OpenGL-a (*Open Graphics Library*) okruženja i održavao se u terminu ispita. Provere znanja nisu bile eliminatornog karaktera, već su služile da studenti prikupe poene na osnovu predispitnih obaveza. Na raspolaganju za rešavanje su bili i probni pripremni testovi kreirani po uzoru na zvanične, a sa ciljem da simuliraju realna okruženja testiranja. Pripremni testovi su bili dostupni tokom održavanja kursa i studenti su mogli da im pristupe neograničen broj puta. Moodle resursi i aktivnosti za učenje bili su preporučeni studentima kao podrška klasičnoj nastavi, s tim da su studenti pristupali ponuđenom elektronskom materijalu prema sopstvenom nahođenju.

Nakon održanog ispita u ispitnom roku, konačna ocena je formirana sumiranjem osvojenih poena i u skladu sa bodovnom skalom. Student je položio ispit i osvojio ocenu šest ukoliko je minimalno sakupio 51 poen. Podaci o konačnim ocenama upisivani su formu elektronskog zapisnika putem web servisa informacionog sistema obrazovne ustanove.

4.1. Osnovni koncepti podataka

Osnovni koncept podataka se može predstaviti kao $n \times d$ matrica sa n redova i d kolona. Redovi predstavljaju entitete skupa, a kolone atribute od značaja. Svaki red u matrici podataka označava zapis vrednosti atributa za određeni entitet.

Skup podataka *P* predstavljen $n \times d$ matričnom formom dat je u nastavku jednačinom (4.1)

$$P = \begin{array}{c|c} & \begin{matrix} X_1 & X_2 & \dots & X_d \end{matrix} \\ \hline \begin{matrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{matrix} & \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{21} & \cdots & x_{2d} \\ \vdots & \ddots & & \vdots \\ x_{n1} & x_{n2} & & x_{nd} \end{pmatrix} \end{array} \tag{4.1}$$

x_i označava i-ti red, prikazan kao vektor $x_i = (x_{i1}, x_{i2}, \dots, x_{id})$

X_j označava j-tu kolonu datu kao vektor $X_j = (x_{1j}, x_{2j}, \dots, x_{nj})$

U zavisnosti od domena primene, za pojam red u literaturi se koriste pojmovi entitet, instanca, primer, zapis, transakcija, objekat, vektor-karakteristika i slično. Takođe, za pojam kolone može se upotrebiti pojam atribut, osobina, obeležje, karakteristika, varijabla, polje i slično. Broj instanci *n* označava veličinu, a broj obeležja *d* dimenzionalnost skupa podataka. Podaci mogu biti struktuirani i nestruktuirani.

Podaci skupa P su struktuirani, ako se svi elementi $d \in P$ mogu predstaviti kao niz obeležje - vrednost parova $d = \{O_1 = o_1,\ O_2 = o_2, \dots, O_k = o_k\}$ pri čemu $o_k \in D(O_i)$. U suprotnom, podaci su nestruktuirani. Nestrukturirani podaci poput teksta, zvuka, videa, slike nemaju nikakve atribute, a stavka podatak je samo jedan atomski entitet, odnosno niz znakova. Kompleksnost strukture podataka zavisi od njihove reprezentacije.

Podaci kursa Računarska grafika korišćeni u opisanom istraživanju su struktuirani. Prikazani su relacionom šemom, a strukturu čini skup parova obeležje - vrednost. Utvrđeni pojmovi iz koncepta podataka koji će se koristiti u disertaciji su: za red pojam *instanca*, za kolonu pojam *obeležje*.

4.1.1. Tipovi podataka

Tipovi podataka klasifikuju obeležja prema njihovim domenima vrednosti. Osnovna podela odnosi se na numeričke (kvantitativne) i kategorijske (kvalitativne) tipove podataka. Numerički podaci imaju značenje kao brojevi i može se meriti red, rastojanje i jednakost dve numeričke varijable. Nasuprot tome, kategorijski podaci nemaju značenje kao brojevi i može se meriti samo jednakost između oznaka kategorija.

Numerički podaci mogu biti klasifikovani kao neprekidni (kontinuirani) i prekidni (diskontinuirani). Neprekidni podaci mogu imati bilo koju numerički vrednost iz skupa realnih brojeva. U slučaju prekidnih podataka, domen vrednosti sadrži samo izolovane tačke. Prekidni podaci mogu biti celi brojevi i decimalni brojevi sa unapred definisanom tačnošću.

Kategorijski podaci se mogu klasifikovati kao nominalni ili redni. Nominalne vrednosti predstavljaju oznake kategorija i nemaju nikavo drugo značenje sem kategorizacije. Treba napomenuti da se kategorijski podaci mogu smatrati diskontinuiranim u smislu da kategorijske varijable mogu imati nebrojivo mnogo vrednosti.

Na osnovu studija slučaja opisanih u pregledu literature (Poglavlje 3), može se zaključiti da većina obrazovnih podataka spada u numeričke prekidne.

4.1.2. Karakteristike obrazovnih podataka

Osnovna karakteristika obrazovnih podataka odnosi se na manji broj instanci skupa što utiče na preciznost formiranih modela. Potrebno je utvrditi veličinu prostora koje karakterišu vrednosti obeležja (eng. *data space*) i gustinu podataka (eng. *data density*) u ćeliji prostora.

Neka je $M= \{ O_1, ..., O_k\}$ skup obeležja, a $D(O_1),.., D(O_k)$ domeni vrednosti. Obeležja $O_1, ..., O_k$ predstavljaju dimenzije prostora skupa $S = D(O_1) \times D(O_2) \times ... \times D(O_k)$. Prostor S karakterišu osobine obeležja $O_i, i=1,...,k,$. Zapremina prostora obeležja izračunava se kao $|S| = |D(O_1)| ... |D(O_k)|$.

Neka je $M= \{O_1, ..., O_k\}$ skup obeležja koji karakteriše prostor S i n broj redova zapisa. Prosečna gustina podataka izračunava se po formuli datoj u jednačini (4.2)

$$density(r) = \frac{n}{|S|} \tag{4.2}$$

Oskudan skup (engl. *sparse data*) je skup u kome je vrednost prosečne gustine podataka isuviše mala. Tačna granica između oskudnog i gustog skupa zavisi od konteksta, ali se generalno smatra da bi trebalo da bude najmanje 5, naročito u slučaju prediktivnog modelovanja. U slučaju oskudnog skupa podataka nemoguće je otkriti uobičajne obrasce u postupku deskriptivnog modelovanja. Oskudnost skupa ne utiče na postupak klasterovanja, ali zato je teže razdvojiti klastere ako je skup isuviše gust.

Gustina podataka obično varira u različitim delovima prostora obeležja. Mera stepena asimetrije u distribuciji podataka označava pojavu odstupanja od uniformne gustine podataka u prostoru obeležja. U slučaju pojave iskrivljenosti u podacima (engl. *skewness*) uobičajena je pojava tzv. "repa iskrivljenosti" na desnu ili levu stranu. Neka je sa S označen prostor karakterisan skupom obeležja R i $|R| > 1$. Prosečnu gustinu podataka skupa S predstavlja funkcija $density(r)$, a $density_T(r)$ označava prosečnu gustinu podataka podprostora T za koji važi $T \subseteq$S. Ako prostor obeležja S sadrži podprostor T, $T \subseteq$S, tako da za granične parametre definisane od strane korisnika $\theta > 0$ i $\gamma > 0$ važi jednačina (4.3) smatra se da su podaci iskrivljeni, odnosno da je distribucija podataka asimetrična.

$$\frac{|T|}{|S|} = \theta \, , \frac{(density(r) - density_T(r))^2}{density(S)} > \gamma \tag{4.3}$$

U slučaju obrazovnih podataka, pojava iskrivljenosti se javlja kao posledica zavisnosti između obeležja. Na primer, neka studenti rade test iz oblasti geometrijskih 2D transformacija u kome su pitanja međusobno zavisna. Studenti koji ostvare lošije rezultate na pitanju koji se odnosi na definiciju translacije, ostvarili su i loše rezultate na pitanju koje se odnosi na zadatak sa primenom metode translacije. Posledica navedenog je pojava oskudnog prostora u levom gornjem uglu grafika raspodele vrednosti podataka. S druge strane, studenti koji su tačno odgovorili na pitanje u vezi definicije rotacije, ostvarili su dobre rezultate i na zadatku koji sadrži primenu metode translacije. U ovom slučaju, dolazi do pojave gustog prostora u desnom gornjem uglu grafika raspodele vrednosti podataka.

Pojava gustog područja u prostoru obeležja otkrivaju zanimljive tendencije u podacima i ne predstavljaju problem prilikom generisanja modela. Međutim, razređena područja su problematičnija. U tim područjima parametri modela se ne mogu precizno proceniti.

Za obrazovne podatake uobičajna je pojava izuzetaka. Prema definiciji, izuzetak predstavlja "posmatranje koje toliko odstupa od drugih zapažanja kako bi izazvalo sumnju da je generisano drugačijim mehanizmima". Opšta definicija za pojavu izuzetka data je u nastavku.

Definicija 1: Neka je $P(D|M)$ verovatnoća podataka D za dati model M, $P(o|M)$ verovatnoća tačke $o \in D$ za dati model i γ korisnički definisana konstanta. Tačka o se smatra izuzetkom ukoliko važi nejednakost (4.4)

$$P(o|M) < \gamma\, P(D|M) \tag{4.4}$$

Pojava izuzetaka u podacima ukazuje na mogućnost kreiranja neodgovarajućih i nepreciznih modela.

Na osnovu pregleda literature koja je data u poglavlju 3, obrazovni skup podataka se može smatrati oskudnim i manjim u odnosu na druge oblasti primene dubinske analize. Utvrđena je pojava izuzetaka i asimetrične raspodele vrednosti u obrazovnim podacima. Uzimajući u obzir složenost organizacije i strukture visokoškolskih obrazovnih sistema moguće je izdvojiti skupove podataka dovoljne brojnosti i gustine za realizaciju odgovarajućih metoda modelovanja.

4.1.3. Deskriptivna analiza podataka

Obrazovni skup podataka uobičajno sadrži numeričke i kategorijske podatke različitih domena i opsega vrednosti. U slučaju prikupljanja podataka iz različitih izvora, evidentna je pojava nesimetrične raspodele vrednosti. Preciznost modela zasniva se na dobro pripremljenom obučavajućem skupu za primenu metoda otkrivanja znanja. Deskriptivna staistička analiza podrazumeva postupak izračunavanja mera centralne tendencije i varijabilnosti podataka. Mere centralne tendencije ukazuju na centralnu vrednost, a varijabilnosti na raspodelu vrednosti u odnosu na utvrđenu centralnu vrednost. U mere centralne tendencije spadaju aritmetička sredina, medina, mod. Varijabilnost podataka utvrđuje se na osnovu mera varijanse, standardne devijacije i raspona vrednosti.

Aritmetička sredina ($\bar{X}$) predstavlja jednu od najosnovnijih mera centralne tendencije. Predstavlja optimalni prikaz prosečne vrednosti distribucije posmatranog obeležja i izračunava se po formuli prikazanoj jednačinom (4.5).

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{4.5}$$

$\bar{X}$ - aritmetička sredina, $\sum_{i=1}^{N} x_i$ - suma svih vrednosti obeležja, *N* je broj instanci skupa.

Informacija o centralnom položaju sortiranih vrednosti obeležja označena je merom median (M_e). Median predstavlja vrednost koja se nalazi tačno u sredini domena. Najfrekventnija vrednost u domenu posmatranog obeležja naziva se mera mod (M_o). Da bi se odredila vrednost mod, potrebno je izvršiti rangiranje vrednosti. U nekim distribucijama postoje više od jedne modalne vrednosti. Na primer, kod bimodalne distribucije postoje dve vrednosti koje su najfrekventnije.

Varijabilnost podataka može se posmatrati izračunavanjem mere standardne devijacije, raspona rezultata i mera asimetrije. Širina ili raspon vrednosti (engl. *range*) predstavlja najjednostavniju meru. Izračunava se oduzimanjem minimalne od maksimalne vrednosti obeležja. Ova mera je izuzetno osetljiva na ekstremne vrednosti u domenu. Kada se pojavi ekstremno visoka ili niska vrednost, veštački se povećava varijaciona širina, a raspon ne prikazuje realnu procenu varijabilnosti koja se analizira.

Standardna devijacija (*S*) predstavlja prosečno odstupanje vrednosti od aritmetičke sredine. Izračunava se kao kvadratni koren iz varijanse koja predstavlja meru variranja vrednosti obeležja u odnosu na aritmetičku sredinu. Može se posmatrati i kao mera disperzije vrednosti koju bi trebalo navoditi uz aritmetičku sredinu. Formula za izračunavanje standardne devijacije prikazana je jednačinom (4.6):

$$S = \sqrt{\sigma^2} = \frac{\sqrt{\sum_{i=1}^{N}(x_i - \bar{X})^2}}{N} \tag{4.6}$$

S standardna devijacija, σ^2 je varijansa, $\bar{X}$ je aritmetička sredina, x_i je i-to obeležje i N je broj zapisa obučavajućeg skupa.

Mere asimetrije označavaju način rasporeda i distribuciju podataka u odnosu na srednju vrednost. Postoje više mera asimetrije: α_3 – treći momenat oko aritmetičke sredine - koeficijent asimetrije, Pirsonov koeficijent asimetrije (S_k) i kvartalni koeficijenat asimetrije (S_{kQ}). Koeficijent asimetrije (α_3) izračunava se po formuli prikazanoj u jednačini (4.7):

$$\alpha_3 = \frac{\sum_{i=1}^{N}(x_i - \bar{X})^3}{NS^3}, \quad -2 < \alpha_3 < 2 \tag{4.7}$$

Ako je $|\alpha_3| < 0{,}1$smatra se da nema asimetrije, a ako je $0{,}1 < |\alpha_3| < 0{,}25$, asimetrija je mala. Srednja asimetrija se javlja u slučajevima kada je $0{,}25 < |\alpha_3| < 0{,}5$, dok $|\alpha_3| > 0{,}5$ ukazuje na jaku asimetriju.

Pirsonov koeficijent asimetrije predstavlja odnos iskrivljenja i mera aritmetičke sredine, medijana i moda. Izračunava se po formulama prikazanim u jednačini (4.8).

$$S_k = \frac{\bar{X} - M_o}{S} \approx \frac{3(\bar{X} - M_e)}{S}, \quad -3 < S_k < 3 \tag{4.8}$$

Kvartalni koeficijent asimetrije prikazan je jednačinom (4.9):

$S_{kQ} = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}$, Q_1, Q_2, Q_3 označavaju prvi, drugi i treći kvartal (4.9)

Desnostranu (pozitivnu) asimetričnu distribuciju (engl. *right skewed distribution*) karakteriše pozitivna vrednost mera asimetrije i veća vrednost aritmetičke sredine u odnosu na median, $\bar{X} > M_e$.

Levostranu (negativnu) asimetričnu distribuciju (engl. *left skewed distribution*) karakteriše negativna vrednost mera asimetrije i manja vrednost aritmetičke sredine u odnosu na median odnosno važi $\bar{X} < M_e$.

U slučaju simetrične distribucije (engl. *normal distribution*) važi jednakost $\bar{X} = M_e = M_o$, a koeficijent asimetrije se nalazi u opsegu od 0 do 0,25.

Histogram predstavlja geometrijsku raspodelu tabele frekvencija koja olakšava deskriptivnu statističku analizu podataka. Vizuelnim prikazom u obliku pravougaonih grafikona između kojih nema praznina, histogram obezbeđuje informacije o distribuciji slučajno odabrane promenljive analiziranog skupa podataka.

Neka obeležje *X* uzima *n* vrednosti: x_1 , x_2 ,..., x_n koje se u skupu od *N* instanci pojavljuje f_1 ,f_2 ,..., f_n puta. Veličine f_1 ,f_2 ,..., f_n zadovoljavaju jednačinu $f_1 + f_2 + \cdots + f_n = N$ i predstavljaju frekvencije. U slučaju velikog broja podataka *N*, domen vrednosti obeležja *X* se predstavlja sa *k* intervala, a frekvencije f_1 ,f_2 ,..., f_k sada označavaju broj vrednosti *X* koje pripadaju prvom, drugom, ..., *k*-tom intervalu. Intervali se ne preklapaju međusobno i imaju određene granične vrednosti. Definiše se ista širina *bin size* (takođe se pominje i kao *bar width* ili *class size*) za sve intervale, a broj instanci posmatrane slučajne promenljive za svaki interval predstavlja frekvencije. Histogram podrazumeva dostupnost svih podataka, odnosno nepostojanje „vrednosti kojih nema" (engl. *missing values*) u analiziranom skupu. Uzimajući u obzir ekstremne vrednosti i izuzetke, pri kreiranju histograma definišu se tačke podele a_1 , ..., a_{k-1} i frekvencije instanci f_1 ,..., f_{k-1} , f_k tako da se može definisati *k* intervala u domenu vrednosti posmatrane slučajne promenljive $(-\infty, a_1]$, $(a_1, a_2]$, ..., $(a_{k-2}, a_{k-1}]$, $(a_{k-1}, +\infty)$.

Algoritam za kreiranje histograma obuhvata sledeće korake:

- sortiranje vrednosti slučajne promenljive u rastućem redosledu,
- određivanje minimalne i maksimalne vrednosti,
- određivanje vrednosti *k ,* broj intervala podele (engl. *numbers of bins*),
- izračunavanje širine intervala po formuli *bin width*=$\frac{\max - min}{k}$,
- izračunavanje frekvencije instanci za svaki interval (engl. *frequency table*) i
- iscrtavanje pravougaonih grafikona tako da x osa predstavlja intervale podele, a y osa frekvenciju instanci za svaki interval.

Određivanje broja intervala podele značajno utiče na oblik histograma. Ako *k* ima isuviše malu vrednost, histogram neće biti korektnog oblika. Ukoliko *k* ima previše veliku vrednost, histogram će prikazati i intervale koji ne sadrže ni jednu vrednost obeležja. U literaturi je zabeleženo više pravila na osnovu kojih se može izračunati vrednost parametra *k.*

Najčešće korišćena pravila – Štrugerovo pravilo (engl. *Strugers rule*) i Rajsovo pravilo (eng. *Rice rule*) data su jednačinama (4.10) i (4.11), respektivno.

$$k = 1 + 3{,}3 \log_{10} n \tag{4.10}$$

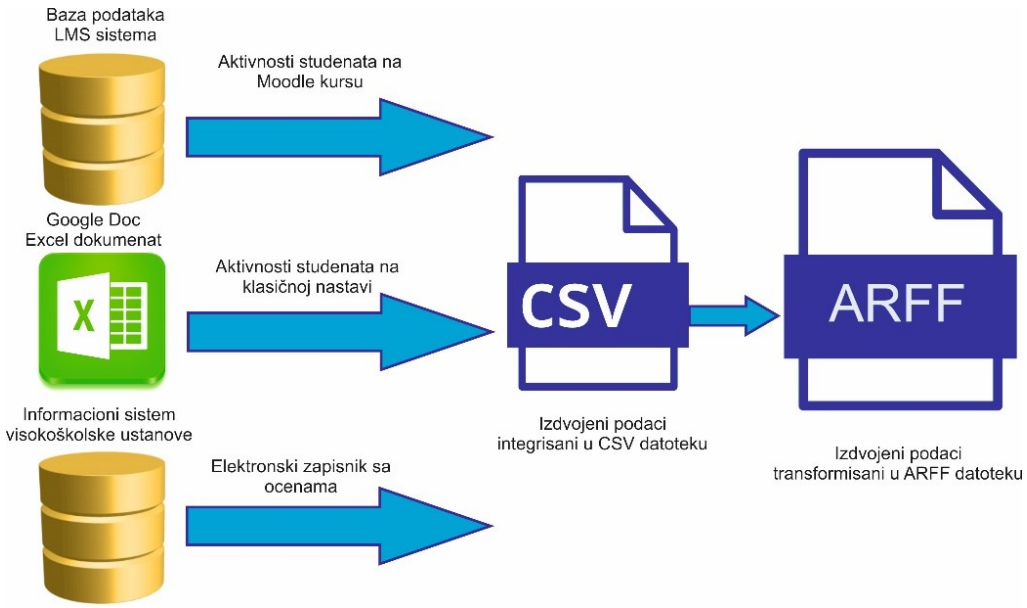
$$k = 2\sqrt[3]{n} \tag{4.11}$$

n - ukupan broj merenja u uzorku, k - broj intervala podele u histogramu.

Generalna preporuka statističara je eksperimentisanje sa nekoliko različitih vrednosti k i izbor broja intervala podele tako da formira najznačajniji histogram u smislu prikaza suštine posmatranog obeležja.

4.2. Izdvajanje podataka

Obučavajući skup za analizirani predmet Računarska grafika kreiran je integracijom podataka izdvojenih iz distribuiranih izvora u odgovarajuću organizacionu formu. Na slici 4.1 prikazana su tri izvora korišćena za potrebe istraživanja opisanog u ovoj disertaciji: baza podataka Moodle sistema, *Google Doc Excel* dokumenat, baza podataka informacionog sistema obrazovne visokoškolske ustanove.



Slika 4.1: Intetegracija podataka izdvojenih iz distribuiranih izvora

Izdvojeni podaci su integrisani u CSV datoteku (engl. *Comma-Separated Values*) pogodnu za čuvanje tabelarnih podataka. CSV datoteka transformisana je u tekstualan *ARFF*

format (engl. *Attribute-Relation File Format*) datoteke pogodan za postupak dubinske analize. ARFF datoteka je u *ASCII* tekstualnom formatu i sadrži listu instanci deljenih među skupom obeležja [57]. Organizovana je u dve odvojene sekcije: sekciju sa nazivom relacije, listom obeležja i tipovima podataka, i sekciju koja sadrži informacije o deklaraciji podataka i stvarne redove zapisa sa instancama.

Baza podataka Moodla sistema podržana je PostgreSQL platformom. Sadržala je 226 zavisnih tabela iz kojih su izdvojeni podaci o osvojenim bodovima na kolokvijumima, ispitu, domaćim zadacima. Evidencija aktivnosti klasične nastave vođena je upotrebom *Google Doc Excel* deljenog dokumenta u koji su upisivani podaci o ostvarenim bodovima na predavanjima i laboratorijskim vežbama. Nakon održanog ispita, formirana je konačna ocena i upisana u elektronski zapisnik. Informacije o načinu upotrebe Moodle resursa, samotestiranju, pristupu interaktivnom i statičkom materijalu, korišćenju kanala komunikacije nisu bili lako dostupni putem klasičnih izveštaja u okviru administracije kursa. U radovima [24, 117] autori opisuju redosled aktivnosti predprocesiranja podataka Moodle kursa:

- kreiranje obučavajućeg skupa podataka u formi sumarne tabele tako da redovi predstavljaju zapis informacija o svim realizovanim aktivnostima studenta analiziranog kursa,
- diskretizacija obučavajućeg skupa podataka, i
- transformisanje diskretizovanih podataka u odgovarajući format za primenu algoritama dubinske analize.

Prikupljanje zapisa o aktivnostima studenata na kursu Računarska grafika izvršeno je generisanjem SQL upita nad odgovarajućim tabelama Moodle baze podataka. Osvojeni bodovi na laboratorijskim vežbama i predavanjima preuzeti su iz Google Doc evidencije. Nakon održanog ispita, profesor je unosio ocene u elektronski zapisnik informacionog sistema obrazovne ustanove. Upis ocena podrazumevao je izbor jedne od ponuđenih vrednosti koje su označavale konačan ostvaren rezultat: 1- poništio, 2 - udaljen sa ispita, 3 - nije izašao, 5 - pao, 6 - šest, 7 - sedam, 8 - osam, 9 - devet, 10 - deset. Preuzimanje unetih podataka izvršeno je generisanjem izveštaja iz informacionog sistema nakon izvršenog upisa ocena. Organizacija podataka pogodna za rudarenje obrazovnih podataka ostvarena je integracijom prikupljenih podataka u jedinstvenu formu. Integrisani podaci iz distribuiranih izvora su bili u tabelarnoj

formi. Kolone (obeležja) su označavale realizovane aktivnosti studenata u mešanom okruženju učenja. Redovi (instance) su predstavljali kompletan zapis podataka za svakog studenta kursa. Naziv i opis obeležja prikazan je u tabeli 4.1.

Tabela 4.1. Naziv i opis obeležja ulaznog skupa podataka

Obeležje	Opis
FD	Učešće u diskusijama na forumima
MM	Upotreba Moodle poruka za elektronske konsultacije
LVT	Upotreba video tutorijala
PDF	Upotreba PDF materijala
LESS_AC	Akcije u lekcijama kojima je pristupano
DZ1, DZ2, DZ3, DZ4, DZ5	Bodovi osvojeni rešavanjem domaćih zadataka
P1, P2, P3	Prosečni bodovi svih pokušaja rešavanja pripremnih testova
T1, T2	Bodovi ostvareni na prvom i drugom kolokvijumu
ISPIT	Bodovi ostvareni na ispitnom testu
LAB	Bodovi osvojeni na laboratorijskim vežbama
BB	Bodovi osvojeni na predavanjima
OCENA	Konačna ocena

Podaci sumarne tabele bili su numeričkog i kategorijskog tipa. Numerička obeležja *LAB, BB, DZ1, DZ2, DZ3, DZ4, DZ5, P1, P2, P3, T1, T2, ISPIT* imala su različite realne vrednosti te su kategorisana kao neprekidna. Vrednosti obeležja *PDF, LVT, OCENA* su celi brojevi pa su označena kao prekidna. Domen mogućih vrednosti za *LESS_AC* sadržao je oznake odgovarajućih osobina što je upućivalo na opisno polinominalno obeležje. Za obeležja koja su označavala aktivnosti studenata, zabeležena je pojava vrednosti koje nedostaju. Korišćene su globalne vrednosti, odnosno vrednost nula i na taj način izvršeno identifikovanje da student nije realizovao određenu aktivnost.

Konačan ulazni skup kreiran je od instanci studenata koji su pali ili položili ispit sa odgovarajućom ocenom. Skup je sadržao 276 redova instanci, numeričkih obeležja i obeležja čije su vrednosti predstavljale oznake određenih osobina.

4.3. Priprema podataka

U istraživanju opisanom u disertaciji, faza pripreme podataka je obuhvatala sledeće postupke:

- popunjavanje vrednosti koje nedostaju,

- otklanjanje problema pogrešnih vrednosti, odnosno šuma u podacima,
- identifikovanje izuzetaka,
- deskriptivnu statističku analizu numeričkih obeležja i
- vizuelni prikaz vrednosti obeležja formiranjem histograma raspodele frekvencija.

U podacima izdvojenim iz baze podataka obrazovne ustanove nije utvrđeno postojanje pogrešnih vrednosti i izuzetaka. Pronađene greške prilikom unosa bodova u *Google Doc Excel* evidenciju su ispravljene ručno. U zapisima o aktivnostima studenata na Moodlu kursu, izdvojeni su redovi koji su označavali neaktivne studente. Za popunjavanje vrednosti koje nedostaju, korišćene su globalne vrednosti. Vrednosti koje nedostaju popunjene su nulom i na taj način označile nerealizovanu aktivnost.

Deskriptivna statistička analiza izvršena je za numerička obeležja obučavajućeg skupa. Izračunate su mere centralne tendencije i varijabilnosti: minimalna vrednost (*Min*), maksimalna vrednost (*Max*), raspon (*Range*), aritmetička sredina ($\bar{X}$), median (M_e), standardna devijacija (S), mod(M_o), koeficijent simetrije, odnosno asimetrije (α_3). Rezultati sprovedene analize su prikazani u Tabeli 4.2.

Tabela 4.2. Deskriptivna statistička analiza obučavajućeg skupa

Obeležje	*Min*	*Max*	*Range*	$\bar{X}$	M_e	M_o	S	α_3
LAB	8	14,9	6,9	12,49	12,8	11	1,487	-0,314
BB	0	4,495	4,95	3,32	3,7	4,4	1,228	-0,774
DZ1	0	2	2	1,36	1,6	2	0,699	-0,999
DZ2	0	2	2	1,33	1,8	2	0,808	-0,788
DZ3	0	2	2	1,32	1,7	2	0,796	-0,831
DZ4	0	2	2	1,14	1,35	0	0,733	-0,518
DZ5	0	2	2	1,2	1,57	2	0,819	-0,502
P1	0	20	20	6,33	3	0	7,218	0,611
P2	0	20	20	9,85	11,2	0	8,133	-0,165
P3	0	30	30	11,72	8,06	0	11,857	0, 257
T1	0	20	20	13,09	13,5	20	4,953	-0,447
T2	1	20	19	13,29	13,25	19	4,370	-0,45
ISPIT	0,63	30	29,37	20,74	22,56	30	6,724	-0,414
PDF	0	8	8	4,61	5,9	6	2,697	-0,338
LVT	0	12	12	6,56	7	12	3,712	-0,232
OCENA	5	10	5	7,56	8	6	1,733	0,051

Neposrednim posmatranjem izračunatih statističkih parametara teško je uočiti određenu zakonitost u podacima. Vizuelni prikaz zasnovan na odgovarajućoj formi prethodno pripremljenih podataka izuzetno olakšava analizu i uvid u distribuciju vrednosti posmatranog obeležja. Uzimajući u obzir broj instanci obučavajućeg skupa i veći broj različitih vrednosti koji karakteriše domene numeričkih obeležja, izvršena je vizuelna ilustracija kreiranjem histograma raspodele frekvencija. Ne menjajući suštinu podataka, za svako numeričko obeležje, formirana je tabela raspodele frekvencija po intervalima opsega vrednosti jednake širine. Korišćeno je Rajsovo i Štrugersovo pravilo za izračunavanje broja intervala. U većini slučajeva oba pravila su davala slične rezultate. Na osnovu utvrđenog broja intervala, određivana je širina tako da se što jasnije uoči karakter posmatranog obeležja i obuhvate sve vrednosti domena. Za svako numeričko obeležje formirane su tabele frekvencija i generisani histogrami raspodele. Postavljene su vrednosti centralnih mera: žuta oznaka - M_o, zelena oznaka - $\bar{X}$ i crvena oznaka za M_e.

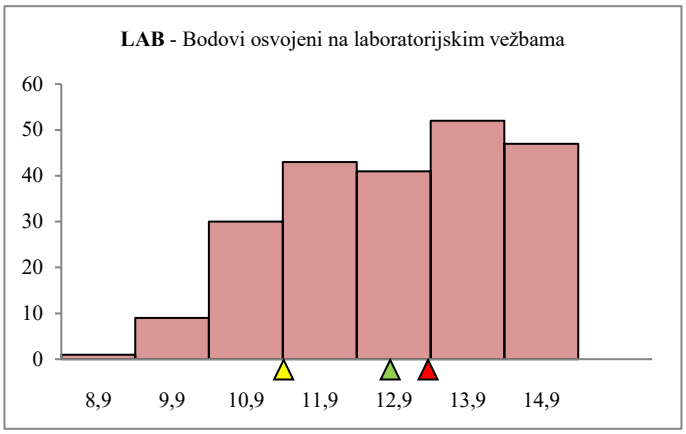
Na osnovu rezulata prikazanih u tabeli 5 za numerička obeležja *LAB, BB, DZ1, DZ2, DZ3, DZ4, DZ5, P2, T1, T2, ISPIT, PDF, LVT* zaključeno je da je vrednost mere median veća u odnosu na aritmetičku sredinu, a koeficijent asimetrije negativan. Na histogramu raspodele za svako obeležje prikazane su mere centralne tendencije i mera M_e posmatrana kao centar iskrivljenja.

Za obeležje *LAB* vrednosti mere median M_e = *12,8* i aritmetičke sredine $\bar{X}$= *12,49* se neznatno razlikuju. Distribucija vrednosti nije simetrična: rezultati se šire na levo dalje nego što to rade desno. Histogram pokazuje da je većina rezultata u blizini centra distribucije, sa manje rezultata u ekstremnim situacijama na kraju iskrivljenja. Interval opsega od 13 do 14 sadrži najveći broj frekvencija i prikazan je kao pravougaoni grafik najveće visine. Na osnovu geometrijske ilustracije vrednosti obeležja *LAB* može se zaključiti da je većina studenata realizovala veći broj laboratorijskih vežbi.

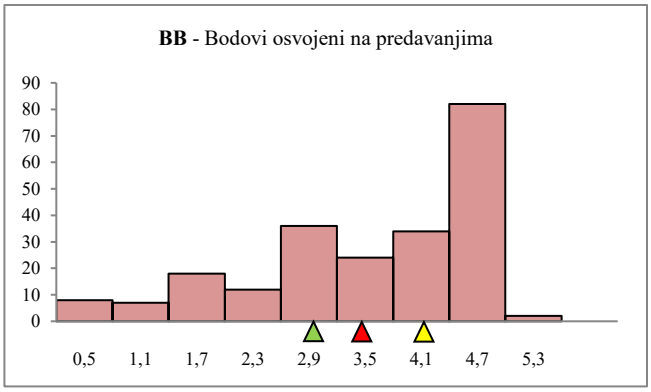
U slučaju obeležja *BB* histogram pokazuje uočljivo iskrivljenje na levo od centralne vrednosti median. Pravougaoni grafik najveće visine određen je za interval opsega od 4,2 do 4,8. Ovom intervalu pripada vrednost mere mod M_o =4,4 što predstavlja najfrekventniju vrednost analiziranog obeležja. Kako obeležje *BB* predstavlja aktivnost studenata na predavanjima, može se zaključiti da je veći broj studenata aktivno učestvovao na predavanjima.

Na slici 4.2 prikazana je vizuelna ilustracija histograma raspodele za obeležje *LAB* i *BB*.

Interval	*Gornja granica*	*Frekvencija*
8-9	8,9	1
9-10	9,9	9
10-11	10,9	30
11-12	11,9	43
12-13	12,9	41
13-14	13,9	52
14-15	14,9	47



Interval	*Gornja granica*	*Frekvencija*
0-0,6	0,5	8
0,6-1,2	1,1	7
1,2-1,8	1,7	18
1,8-2,4	2,3	12
2,4-3	2,9	36
3-3,6	3,5	24
3,6-4,2	4,1	34
4,2-4,8	4,7	82
4,8-5,4	5,3	2



Slika 4.2: Histogrami raspodele obeležja LAB,BB

Za obeležja koja označavaju prvi, drugi, treći i peti domaći zadatak (*DZ1, DZ2, DZ3, DZ5*), mera median je veća od aritmetičke sredine i vrednost mere mod je M_o =2. Apsolutna vrednost negativnog koeficijenta asimetrije, u slučaju ovih obeležja, je $|\alpha_3| > 0{,}5$ i označava jaku iskrivljenost u levo.

U slučaju obeležja *DZ4* vrednost mere mod M_o = 0 i manja je od aritmetičke sredine i mere median. Apsolutna vrednost negativnog koeficijenta asimetrije $|\alpha_3| > 0{,}5$ što kao i kod ostalih obeležja domaćih zadataka ukazuje na jaku iskrivljenost u levo. Vrednost median je neznatno veća od aritmetičke sredine $M_e = 1{,}35; \bar{X} = 1{,}13$. Obzirom na nisku vrednost mere mod izračunata je dodatno i mera takozvane doterane sredine (engl. *trimmed mean*). Vrednost *trimmed mean* uobičajno se nalazi između vrednosti aritmetičke sredine i median. Za obeležje

DZ4, vrednost *trimmed mean* sa procentom isključivanja od 50% je 1,3 i nalazi se u opsegu između vrednosti aritmetičke sredine i median.

Odnos iskrivljenja i mera aritmetičke sredine i median prikazana je izračunavanjem Pirsonove mere asimetrije.

$$S_k = \frac{3(\bar{X}-M_e)}{S} = \frac{3(1{,}14-1{,}35)}{0{,}734} = -0{,}281$$

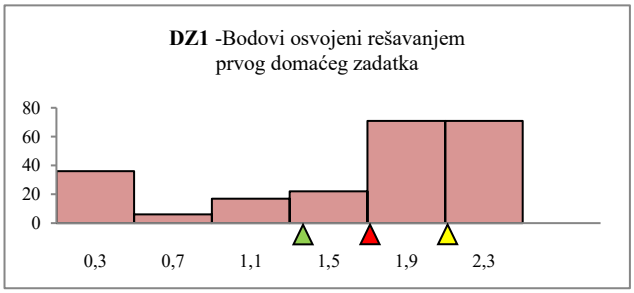
Kao i vrednost koeficijenta asimetrije $\alpha_3(DZ4) = -0{,}518$, negativna vrednost Pirsonove mere ukazuje na levostranu iskrivljenost vrednosti obeležja *DZ4*.

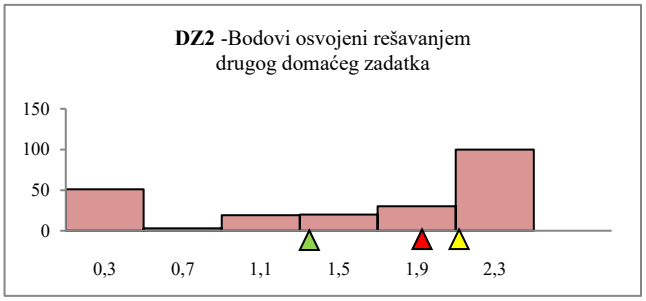
Na slici 4.3 prikazan je histogram raspodele vrednosti obeležja *DZ1, DZ2, DZ3, DZ4, DZ5*. Centar iskrivljenja je vrednost mere median, a rep koji se proteže na levo oslikava levostranu raspodelu vrednosti.

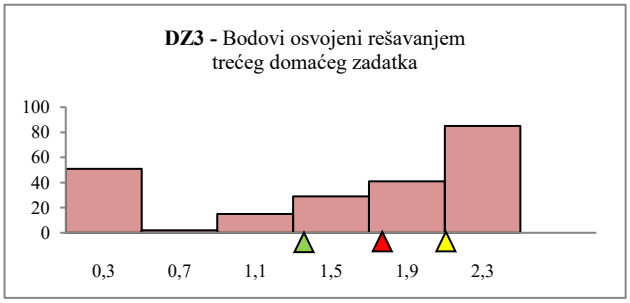
Interval	*Gornja granica*	*Frekvencija*
0-0,4	0,3	36
0,4-0,8	0,7	6
0,8-1,2	1,1	17
1,2-1,6	1,5	22
1,6-2	1,9	71
2-2,4	2,3	71

Interval	*Gornja granica*	*Frekvencija*
0-0,4	0,3	51
0,4-0,8	0,7	3
0,8-1,2	1,1	19
1,2-1,6	1,5	20
1,6-2	1,9	30
2-2,4	2,3	100

Interval	*Gornja granica*	*Frekvencija*
0-0,4	0,3	51
0,4-0,8	0,7	2
0,8-1,2	1,1	15
1,2-1,6	1,5	29
1,6-2	1,9	41
2-2,4	2,3	85

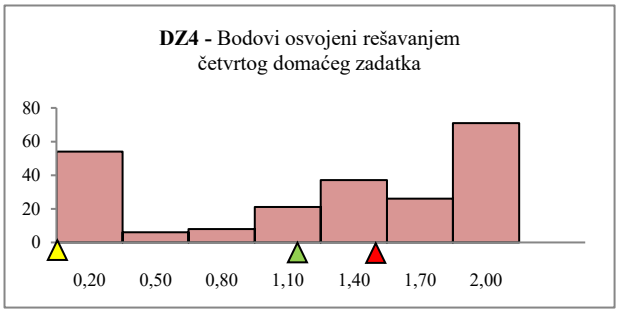


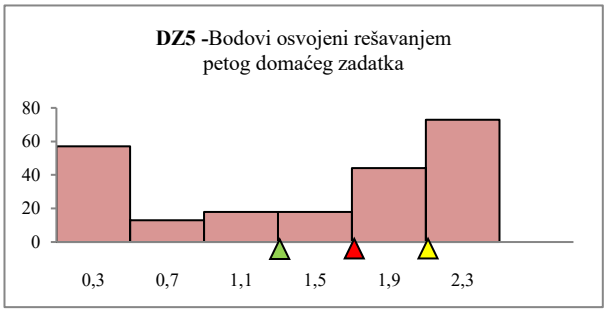




Interval	*Gornja granica*	*Frekvencija*
0-0,3	0,20	54
0,3-0,6	0,50	6
0,6-0,9	0,80	8
0,9-1,2	1,10	21
1,2-1,5	1,40	37
1,5-1,8	1,70	26
1,8-2,1	2,00	71

Interval	*Gornja granica*	*Frekvencija*
0-0,4	0,3	57
0,4-0,8	0,7	13
0,8-1,2	1,1	18
1,2-1,6	1,5	18
1,6-2	1,9	44
2-2,4	2,3	73





Slika 4.3: Histogrami raspodele vrednosti obeležja DZ1, DZ2, DZ3, DZ4

U slučaju *P2*, vrednost mere mod je nula i manja je od vrednosti aritmetičke sredine. Apsolutna vrednost negativnog koeficijenta asimetrije $|\alpha_3(\mathrm{P2})| = 0{,}165$ što ukazuje na umerenu iskrivljenost u levo.

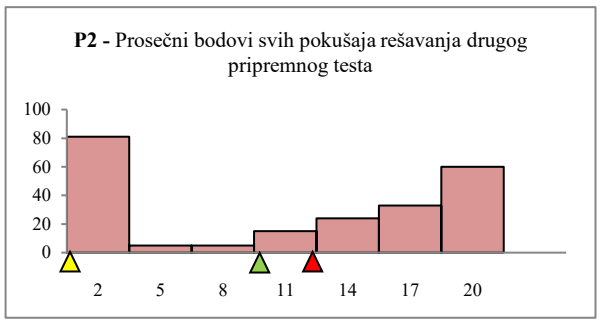
Vrednost mere median $M_e(P2) = 11{,}2$ je veća od aritmetičke sredine $\bar{X}(P2) = 9{,}85$. Kao i u slučaju obeležja DZ4 i za P2 je izračunata mera *trimmed mean* i Pirsonova mera asimetrije.

Vrednost *trimmed mean* sa procentom isključivanja od 50% je 10,05 i nalazi se u opsegu između vrednosti aritmetičke sredine i median. Odnos iskrivljenja i mera aritmetičke sredine, median prikazana je izračunavanjem Pirsonove mere asimetrije.

$$S_k = \frac{3(\bar{X} - M_e)}{S} = \frac{3(9{,}85 - 11{,}2)}{8{,}133} = -0{,}497$$

Na slici 4.4 prikazan je histogram raspodele vrednosti obeležja *P2*. Histogram pokazuje većinu rezultata u blizini centra iskrivljena (M_e) što ukazuje na činjenicu da je veći broj studenata rešavalo drugi pripremni test. Za interval opsega od 0 do 3 formiran je pravougaoni grafik najveće visine. Kako je najfrekvetnija vrednost domena M_o =0, može se zaključiti da jedan deo studenata nije rešavao drugi pripremni test.

Interval	*Gornja granica*	*Frekvencija*
0-3	2	81
3-6	5	5
6-9	8	5
9-12	11	15
12-15	14	24
15-18	17	33
18-21	20	60



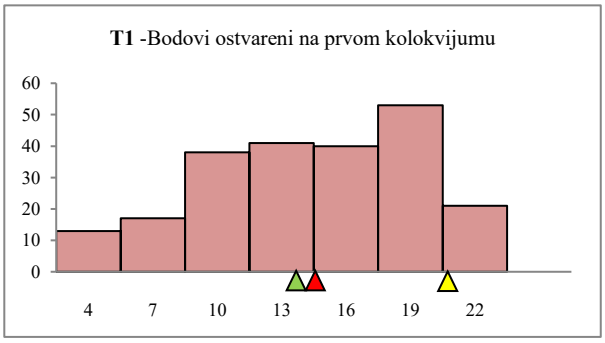
Slika 4.4: Histogram raspodele vrednosti obeležja P2

Histogrami raspodele za obeležja koja su označavala bodove osvojene na prvom, drugom kolokvijumu i ispitu su prikazani na slici 4.5. Vrednost mere mod, za sva tri obeležja, ukazuje da su najfrekventnije vrednosti maksimalan broj bodova po testu. Aritmetička sredina je neznatno manja od mere median. Distribucija vrednosti je većinski koncetrisana oko centralne tačke, odnosno mere median.

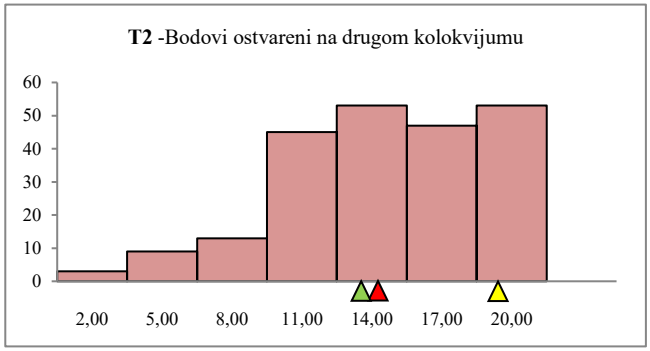
Za obeležje *T1*, pravougaoni grafik sa najvećim brojem frekvencija formiran je za interval opsega od 17 do 20 bodova i gornjom granicom 19. U slučaju obeležja *T2*, formirana su dva pravougaona grafika sa najvećim brojem frekvencija i to: za interval opsega od 12 do 15 bodova sa gornjom granicom 14 i interval opsega od 18 do 21 bodova sa gornjom granicom 20.

Histogram obeležja *ISPIT* obuhvata osam intervala podele. Koncetracija vrednosti je takođe oko centralne tačke. Pravougaoni grafik najveće visine formiran je za interval opsega od 24 do 28 sa gornjom granicom 27. Koeficijenat asimetrije sva tri obeležja je negativan i približno jednakih vrednosti $\alpha_3(T1) = -0{,}447; \alpha_3(T2) = -0{,}45; \alpha_3(ISPIT) = -0{,}414$. Apsolutna vrednost mere asimetrije je manja od 0,5 što ukazuje na umerenu iskrivljenost u levo za obeležja *T1, T2, ISPIT*.

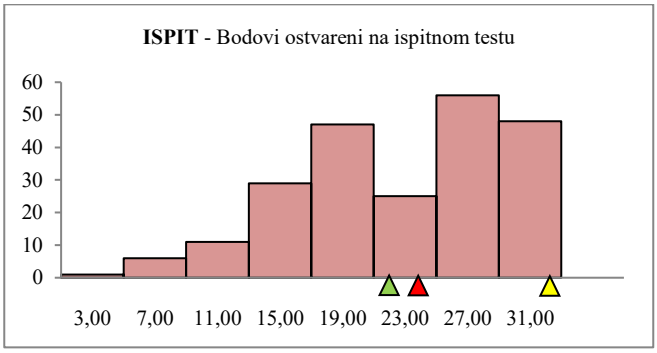
Interval	*Gornja granica*	*Frekvencija*
2-5	4	13
5-8	7	17
8-11	10	38
11-14	13	41
14-17	16	40
17-20	19	53
20-23	22	21



Interval	*Gornja granica*	*Frekvencija*
1-3	2	3
3-6	5	9
6-9	8	13
9-12	11	45
12-15	14	53
15-18	17	47
18-21	20	53



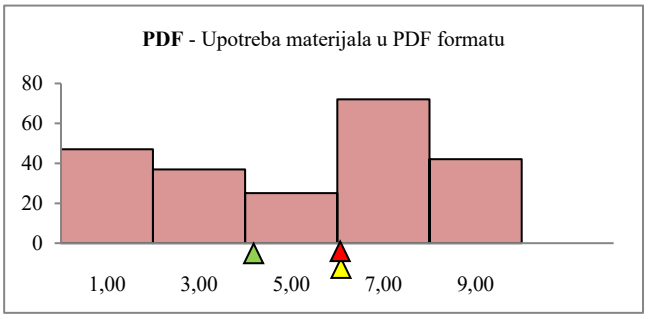
Interval	*Gornja granica*	*Frekvencija*
0-4	3	1
4-8	7	6
8-12	11	11
12-16	15	29
16-20	19	47
20-24	23	25
24-28	27	56
28-32	31	48



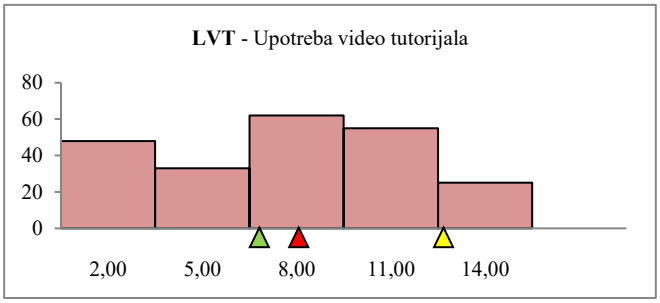
Slika 4.5: Histogrami raspodele vrednosti obeležja T1,T2,ISPIT

Histogrami obeležja koja su označavala upotrebu materijala u audio-video i *PDF* formatu prikazani su na slici 4.6.

Interval	*Gornja granica*	*Frekvencija*
0-2	1,00	47
2-4	3,00	37
4-6	5,00	25
6-8	7,00	72
8-10	9,00	42



Interval	*Gornja granica*	*Frekvencija*
0-3	2,00	48
3-6	5,00	33
6-9	8,00	62
9-12	11,00	55
12-15	14,00	25



Slika 4.6: Histogrami raspodele vrednosti obeležja PDF, LVT

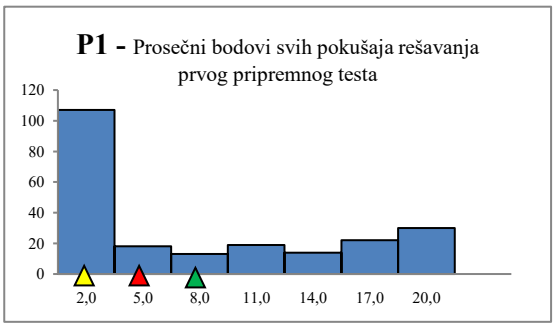
Za *PDF* obeležje najviši pravougaoni grafik je formiran za interval opsega od 6 do 8 sa gornjom granicom 7. Vrednosti mod i mere median su približno jednake $M_o \cong M_e = 6$ tako da se može zaključiti da je većina studenata koristila u proseku 6 od ponuđenih 8 datoteka PDF formata. Apsolutna vrednost negativnog koeficijenta iskrivljenja $|\alpha_3(\text{PDF})| = 0{,}338 < 0{,}5$ ukazuje na umerenu iskrivljenost u levo. Histogram ukazuje na veću koncentraciju vrednosti u blizini centra distribucije, sa manje rezultata na kraju iskrivljenja.

U slučaju obeležja *LVT* pravougaoni grafik sa najvećim brojem frekvencije formiran je za interval opsega od 6 do 9 sa gornjom granicom 8. Vrednost centralne mere distribucije je $M_e = 7$. Apsolutna vrednost negativnog koeficijenta iskrivljenja $|\alpha_3(\text{LVT})| = 0{,}23 < 0{,}5$ ukazuje na umerenu iskrivljenost u levo. Histogram pokazuje veći broj rezulata oko centra distribucije, a manji broj na kraju iskrivljenja.

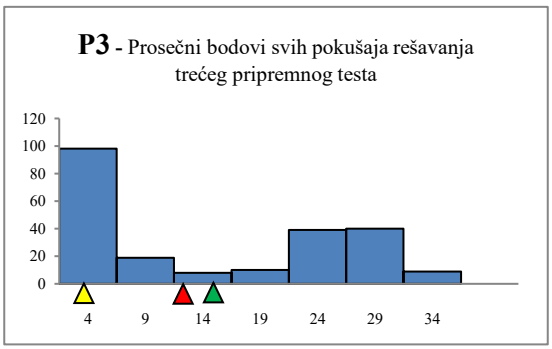
Analiza obeležja *PDF* i *LVT*, koji označavaju dve različite forme materijala, ukazuje na činjenicu da su studenti koristili oba tipa u procesu učenja. Za *P1*, *P3* mera aritmetičke sredine je veća od vrednosti median, median od mod, odnosno $\bar{X} > M_e > M_o$ i koeficijenata asimetrije je veći od nula za sva tri obeležja. Utvrđena je pojava pozitivne iskrivljenosti u podacima odnosno desnostrana asimetrična distribucija.

Histogrami obeležja *P1*, *P3* prikazuju pravougaone grafike sa najvećim brojem frekvencija formirane za slučajeve intervala sa najmanjim vrednostima (slika 4.7). Apsolutna vrednost koeficijenta asimetrije $|\alpha_3(\text{P1})| = 0{,}61$ ukazuje na jaku pozitivnu iskrivljenost u desno. U slučaju *P3* obeležja vrednost koeficijenta asimetrije je u opsegu od 0.25 do 0.5 što ukazuje na pojavu srednje asimetrije u distribuciji podataka.

Interval	*Gornja granica*	*Frekvencija*
0-3	2,0	107
3-6	5,0	18
6-9	8,0	13
9-12	11,0	19
12-15	14,0	14
15-18	17,0	22
18-21	20,0	30



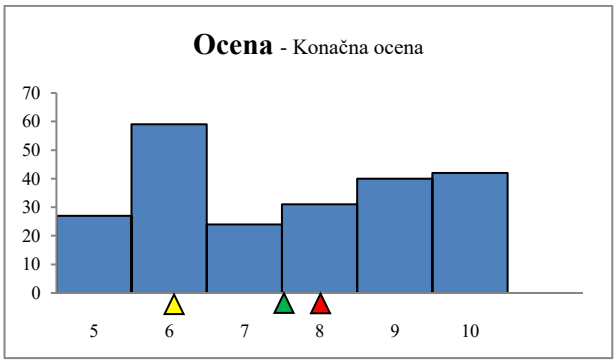
Interval	*Gornja granica*	*Frekvencija*
0-5	4	98
5-10	9	19
10-15	14	8
15-20	19	10
20-25	24	39
25-30	29	40
30-35	34	9



Slika 4.7: Histogrami raspodele vrednosti obeležja P1,P3

Histogram raspodele za obeležje Ocena koje predstavlja konačnu ostvarenu ocenu studenata dat je na slici 4.8. Za gornje granice intervala podele postavljene su vrednosti 5, 6, 7, 8, 9, 10 i kreirana je tabela frekvencija. Pravougaoni grafik najveće visine formiran je za interval gornje granice 6. Vrednost mere median je približno jednaka aritmetičkoj sredini, a vrednost mere mod odgovara gornjoj granici najfrekventnijeg intervala podele. Koeficijent asimetrije je pozitivan i manji od 0.1 što ukazuje na normalnu distribuciju vrednosti.

Gornja granica	*Frekvencija*
5	27
6	59
7	24
8	31
9	40
10	42



Slika 4.8: Histogram raspodele vrednosti obeležja Ocena

5. PREDPROCESIRANJE PODATAKA

Precizan model predviđanja zasniva se na efikasnom konceptu transformacije domena vrednosti. Osnovni zadatak faze predprocesiranja podrazumeva implementaciju metoda diskretizacije kako bi podaci bili spremni za proces otkrivanja znanja iz podataka. Istraživanje opisano u ovom poglavlju disertacije fokusirano je na utvrđivanju postupka diskretizacije kontinualnih obeležja obrazovnog skup mešanog okruženja učenja. Opisani eksperimenti zasnovani su na studiji slučaja prezentovanoj u radu [136]. Izvršena je komparativna analiza rezultata metode nenadgledanog učenja zasnovane na podeli domena na intervale jednake širine (engl. *equal-width binning*, *EWB*), metode diskretizacije zasnovane na histogramu i metode nadgledanog učenja zsnovane na meri entropije. Broj intervala podele u slučaju metode histogram diskretizacije utvrđen je primenom Skotovog pravila određivanja širine intervala. EWB metoda je implementirana sa dinamičkom potragom za optimalnom vrednošću k. Nadziranom metodom entropije, domeni numeričkih obeležja diskretizovani su različitim brojem diskretnih vrednosti. Efikasnost nenadzirane metode diskretizacije u slučaju modela predviđanja ostvarena je tehnikom ponovnog uzorkovanja (eng. *oversampling*) *SMOTE* (engl. *Synthetic Minority Oversampling Technique*) i *Randomize* filtra.

5.1. Ulazni skup podataka

Orginalni skup podataka mešanog okruženja učenja za slučaj kursa Računarska grafika opisan je u poglavlju četri ove disertacije. Definisana su numerička i kategorijska obeležja *LAB, BB, DZ1, DZ2, DZ3, DZ4, DZ5, P1, P2, P3, T1, T2, ISPIT, FD, MM, PDF, LVT, LESS_AC, OCENA*. Izvršeno je analiziranje domena vrednosti obeležja i čišćenje "sirovih" podataka. Numerička obeležja su imala neujednačene domene vrednosti, a kategorijski su bili binominalni sa dve moguće vrednosti i polinominalani sa više mogućih vrednosti. Čišćenje realnih "sirovih" podataka koji su uglavnom nekompletni, nedosledni i pogrešni, podrazumeva pokušaj popunjavanja vrednosti koje nedostaju, ispravljanje nedoslednosti i ublažavanje problema pogrešnih vrednosti identifikovanjem izuzetaka. Za analizirani skup podataka korišćene su globalne vrednosti za popunjavanje vrednosti koje nedostaju. U slučaju ulaznih obeležja, nedostajuće vrednosti popunjene su vrednošću nula koja je označavala da student nije

realizovao aktivnost. Numeričko obeležje *OCENA*, definisano je kao izlaz sa tendencijom transformacije u nominalan tip, a popunjavanje nedostajućih vrednosti izvršeno je sa vrednošću 3, što je označavalo da student nije polagao ispit.

Obzirom da su podaci izdvojeni iz baza podataka nije utvrđeno postojanje pogrešnih vrednosti i izuzetaka. Sprovedeni su eksperimenti diskretizacije i rangiranja ulaznih obeležja obrazovnog skupa podataka.

5.2. Diskretizacija obrazovnog skupa podataka

5.2.1. Entropija

Prvi eksperiment je predstavljao implementaciju metode entropije [58] sa kriterijumom zaustavljanja *MDL*, opisanim u poglavlju 2. Metoda entropije zasniva se na konceptu nadgledanog učenja. Uzima u obzir informaciju o klasi kandidata za podelu kako bi se odredile granice formiranja diskretnih vrednosti. Posmatra se domen svih vrednosti obeležja, a zatim vrši rekurzivna podela na podintervale dok se ne dostigne kriterijum zaustavljanja.

Diskretizacija numeričkog obeležja *OCENA* je podrazumevala transformaciju u nominalni tip i postavljanje za klasno obeležje. Numeričke vrednosti 3, 5, 6, 7, 8, 9, 10 diskretizovane su vrednostima {*nije_polagao*, *nije_polozio*, *sest*, *sedam*, *osam*, *devet*, *deset*}.

Ulazna numerička obeležja sortirana su u opadajućem redosledu. Podela njihovih domena izvršena je u tačkama sortirane liste gde se menjala vrednost klasnog obeležja. Za svaku tačku podele je izračunata vrednost entropije indukovanih particija, odnosno podskupova levo i desno od tačke podele.

Tačka podele T među svim kandidatima za $E(A, T; S)$ izabrana je na osnovu provere kriterijuma minimalne vrednosti entropije. Postupak se ponavljao rekurzivno sve dok oba podskupa nisu sadržala samo instance iste klase i nije postignut kriterijum za zaustavljanje.

Domeni numeričkih obeležja diskretizovani su različitim brojem diskretnih vrednosti. Broj diskretnih vrednosti ulaznih numeričkih obeležja prikazan je u tabeli 5.1.

Tabela 5.1. Diskretne vrednosti numeričkih obeležja

Obeležje	Broj diskretnih vrednosti
LAB	8
BB	5
DZ1	4
DZ2	3
DZ3	3
DZ4	3
DZ5	5
T1	11
T2	6
P1	1
P2	1
P3	3
ISPIT	6
PDF	1
LVT	1

Za obeležja *P1, P2, PDF, LVT* nisu utvrđene tačke podele, odnosno nije pronađen "čist" podskup koji sadrži samo slučajeve jedne klase. Efikasnost metode entropije utvrđena je test metodom procene klasifikatorskih modela Hidden Näive Bayes (*HNB*) i Random Forest (*RF*).

Obučavajući skup je obuhvatao dve trećine, a test skup jednu trećinu instanci. Klasifikatori su testirani na skupu sa svim obeležjima i na skupu bez obeležja *P1, P2, PDF* i *LVT*. *HNB* je generisao modele istih performansi za oba slučaja. *RF* klasifikator je generisao model boljih performansi u slučaju skupa sa svim obeležjima.

Ovim eksperimentom je ukazano na značaj *P1,P2, PDF* i *LVT* obeležja za model predviđanja i nepodobnost metode entropije za diskretizaciju numeričkih vrednosti obrazovnog skup podataka mešanog okruženja učenja.

5.2.2. Metoda jednakih intervala

U sledećem eksperimentu je primenjena nenadzirana metoda diskretizacije *EWB* [51] koja podrazumeva podelu domena vrednosti na intervale jednake širine. Izvršena je dinamička potraga za optimalnim brojem intervala podele [149] istovremenom diskretizacijom ulaznih numeričkih obeležja obučavajućeg skupa vrednostima $k = 2, 3,...10$.

Svaki diskretizovan skup testiran je procenom *RF* i *HNB* klasifikatora metodom test skupa. *RF* klasifikator kreirao je modele koji su ostvarili linerano poboljšanje performansi i smanjenje greške klasifikacije za vrednosti k={2,3,4,5,6,8,9,10}.

Podela domena vrednosti na k=7 intervala zabeležila je pad performansi formiranog modela u odnosu na prethodne, k=5, k=6. U slučaju *HNB* klasifikatora, linearno poboljšanje performansi modela uočeno je za podelu domena vrednosti na k={2,3,4,5,6} intervala. Podela domena na k={7,8,9,10} intervala uslovila je pojavu neujednačenosti u performansama.

Ostvarene tačnosti *HNB* i *RF* modela sa primenjenom istovremenom diskretizacijom ulaznih numeričkih obeležja prikazane su u tabeli 5.2. Izračunata je aritmetička sredina tačnosti kao i prosečna vrednost odstupanja. Aritmetička sredina tačnosti (*Acc_AS*) ukazuje na broj intervala podele k=6 za *HNB* model, a k=7 za *RF* model.

Tabela 5.2. Tačnost HNB i RF modela za različite intervale podele

Model	**k – broj intervala podele**									**Acc_AS**	**ST_Dev**
	2	3	4	5	6	7	8	9	10		
HNB	63.83	76.59	74.49	84.04	82.97	80.85	82.98	80.85	84.04	78.49	6.40
RF	69.15	79.78	80.85	87.23	89.36	85.10	88.3	86.17	91.49	84.04	6.78

5.2.3. Metoda diskretizacije zasnovana na histogramu

Poslednji eksperiment se odnosio na primenu histogram metode diskretizacije. Izvršeno je sortiranje, određivanje minimalne i maksimalne vrednosti. Broj intervala podele određen je na osnovu Skotovog pravila [56] izračunavanjem količnika razlike graničnih vrednosti domena obeležja i širine intervala *h*. Širina intervala zavisi od standardne devijacije i broja instanci analiziranog skupa.

$$k = \frac{a_{max} - a_{min}}{h} \tag{5.1}$$

a_{min}, a_{max} - minimalna i maksimalna vrednost posmatranog obeležja.

$$h = \frac{3.5 \times \sigma}{\sqrt[3]{n}}, \tag{5.2}$$

σ standardna devijacija, *n* je broj instanci skupa.

U tabeli 5.3 prikazane su vrednosti *h* i *k* obučavajućeg skupa podataka sa n=276 instanci.

Tabela 5.3. Određivanje broja intervala podele

Obeležje	min	max	Mean	StDev	h	(max-min)/h	k
LAB	0	14,9	10,919	4,329	2,33	6,40	7
BB	0	4,95	2,903	1,455	0,78	6,33	7
DZ1	0	2	1,168	0,796	0,43	4,67	5
DZ2	0	2	1,151	0,879	0,47	4,23	5
DZ3	0	2	1,131	0,877	0,47	4,24	5
DZ4	0	2	0,957	0,785	0,42	4,74	5
DZ5	0	2	1,004	0,867	0,47	4,29	5
PDF	0	8	4,420	2,694	1,45	5,52	6
LVT	0	12	6,630	3,739	2,01	5,97	6
P1	0	20	6,035	7,272	3,91	5,12	6
P2	0	20	9,363	8,170	4,39	4,55	5
P3	0	30	11,819	12,069	6,49	4,62	5
T1	0	20	11,405	6,170	3,32	6,03	7
T2	0	20	11,207	6,164	3,31	6,04	7
ISPIT	0	30	16,869	10,057	5,41	5,55	6

Broj intervala podele je dobijen zaokruživanjem na veću vrednost kako bi sve instance bile obuhvaćene. Određena je frekvencija, odnosno broj instanci čije se vrednosti nalaze u granicama posmatranog intervala (videti tabelu 5.4).

Table 5.4. Tabele frekvencija

LAB	
Interval	Frekvencija
(min-2.13]	34
(2.13-4,26]	0
(4.26-6.39]	0
(6.39-8.51]	1
(8.51-10.64]	27
(10.64-12.77]	92
(12.77-max)	122

BB	
Interval	Frekvencija
(min-0.71]	32
(0.71-1.41]	39
(1.41-2.12]	9
(2.12-2.83]	40
(2.83-3.54]	38
(3.54-4.24]	40
(4.24-max)	78

DZ1	
Interval	Frekvencija
(min-0.4]	79
(0.4-0.8]	9
(0.8-1.2]	37
(1.2-1.6]	62
(1.6-max]	89

DZ2	
Interval	Frekvencija
(min-0.4]	92
(0.4-0.8]	15
(0.8-1.2]	13
(1.2-1.6]	35
(1.6-max)	121

DZ3	
Interval	Frekvencija
(min-0.4]	97
(0.4-0.8]	3
(0.8-1.2]	21
(1.2-1.6]	29
(1.6-max)	126

DZ4	
Interval	Frekvencija
(min-0.4]	101
(0.4-0.8]	16
(0.8-1.2]	31
(1.2-1.6]	54
(1.6-max)	74

DZ5	
Interval	Frekvencija
(min-0.4]	109
(0.4-0.8]	10
(0.8-1.2]	29
(1.2-1.6]	22
(1.6-max)	106

P1	
Interval	Frekvencija
(min-0.33]	150
(0.33-6.67]	17
(6.67-10]	26
(10-13-33]	22
(13.33-16.67]	24
(16.67-max)	37

P2	
Interval	Frekvencija
(min-4]	112
(4-8]	14
(8-12]	24
(12-16]	41
(16-max)	85

P3	
Interval	Frekvencija
(min-6]	131
(6-12]	18
(12-18]	18
(18-24]	45
(24-max)	64

PDF	
Interval	Frekvencija
(min-1]	62
(1-2]	21
(2-4]	41
(4-5]	19
(5-6]	64
(6-max)	69

LVT	
Interval	Frekvencija
(min-2]	59
(2-4]	224
(4-6]	43
(6-8]	46
(8-10]	58
(10-max)	46

T1	
Interval	Frekvencija
(min-2.86]	37
(2.86-5.71]	27
(5.71-8.57]	15
(8.57-11.43]	41
(11.43-14.29]	51
(14.29-17.14]	37
(17.14-max)	68

T2	
Interval	Frekvencija
(min-2.86]	43
(2.86-5.71]	13
(5.71-8.57]	20
(8.57-11.43]	43
(11.43-14.29]	59
(14.29-17.14]	45
(17.14-max)	53

ISPIT	
Interval	Frekvencija
(min-5]	52
(5-10]	16
(10-15]	31
(15-20]	58
(20-25]	44
(25-max]	75

Za obeležja *LAB, BB, DZ1* i *DZ3*, histogram diskretizacijom su uspostavljeni intervali sa malim brojem pripadajućih instanci. Konkretno, za obeležje *LAB* utvrđena su dva intervala sa po nula instanci što je posledica podele domena na intervale jednakih širina. U cilju postizanja bolje raspodele diskretizovanih vrednosti izvršen je proces spajanja. Na taj način postignuta je ravnomernija raspodela za posmatrana obeležja. Rezultati spajanja intervala u slučaju obeležja LAB, BB, DZ1 i DZ3 prikazani su u Tabeli 5.5.

Tabela 5.5. Spajanje intervala podele za obeležja LAB, BB, DZ1,DZ3

LAB		
Interval	*F*	*Oznaka*
(min-6.39]	34	LAB1
(6.39-10.64]	28	LAB2
(10.64-12.77]	92	LAB3
(12.77-max)	122	LAB4

BB		
Interval	*F*	*Oznaka*
(min-0.71]	32	BB1
(0.71-2.12]	48	BB2
(2.12-2.83]	40	BB3
(2.83-3.54]	38	BB4
(3.54-4.24]	40	BB5
(4.24-max)	78	BB6

DZ1		
Interval	*F*	*Oznaka*
(min-0.8]	88	DZ11
(0.8-1.2]	37	DZ12
(1.2-1.6]	62	DZ13
(1.6-max]	89	DZ14

DZ3		
Interval	*F*	*Oznaka*
(min-0.8]	100	DZ31
(0.8-1.2]	21	DZ32
(1.2-1.6]	29	DZ33
(1.6-max)	126	DZ34

5.2.4. Uporedna analiza nenadziranih metoda diskretizacije

Završna faza podrazumevala je uporednu analizu klasifikatorskih modela kako bi se utvrdila efikasnost diskretizacije obrazovnog skupa podataka u slučaju mešanog okruženja učenja. Analiziranjem rezultata sprovedenih eksperimenata, metoda entropije je isključena obzirom da nije primenljiva za domene vrednosti obeležja *P1, P2, PDF, LVT*.

Utvrđivanje metode diskretizacije za ispitivanu studiju slučaja mešanog okruženja učenja podrazumevalo je poređenje performansi klasifikatora pogodnih za rad sa manjim obrazovnim skupom koji karakteriše multidimenzionalno klasno obeležje. Diskretizovani skupovi testirani su klasifikatorima zasnovanim na *Näive Bayes*, *Hidden Näive Bayes*, *J48*-stablom odluke i *Random Forest* algoritmima. Komparativna analiza zasnivala se na poređenju ostvarene tačnosti modela. U tabeli 5.6 dat je prikaz performansi kreiranih modela.

Tabela 5.6. Uporedna analiza kreiranih modela

Model	EWB (k=6)	EWB (k=7)	HD
NB	73.40	77.66	82.98
HNB	82.97	80.85	85.11
J48	82.97	78.72	82.98
RF	89.36	85.11	91.49

Rezultati završnog eksperimenta diskretizacije obrazovnog skupa studije slučaja ukazuju na zaključak da histogram diskretizacija sa Skotovim pravilom određivanja optimalnog broja intervala podele značajno utiče na kreiranje preciznijih modela predviđanja. *EWB* eksperimentom nije bilo moguće odrediti optimalan broj intervala obzirom da primena diskretizacije sa istom vrednošću parametra *k* za sva numerička obeležja može da rezultira potiskivanje pozitivnih efekata diskretizacije. Diskretizacija istim brojem intervala podele znači i da nije uzeta u obzir različita raspodela vrednosti domena svakog obeležja ponaosob.

Postavlja se pitanje efikasnosti u slučaju povećanja obučavajućeg obrazovnog skupa. U tom cilju implementirana je tehnika ponovnog uzorkovanja *SMOTE* [150] za modifikovanje neizbalansiranosti obučavajućeg skupa podataka tehnikom sintetičkog uzorkovanja. Izvršeno je podjednako raspoređivanje instanci sa manjinskom vrednošću klasnih obeležja. Kreiranjem sintetičkih instanci manjinske klase, poboljšala se prediktivna tačnost analiziranog skupa. Primena *SMOTE* algoritma je podrazumevala automatsko određivanje manjinske klase, postavljanje vrednosti parametra za izbor najbližih suseda na 5, a procenat *SMOTE* instanci koji će biti kreiran je postavljen na 100%.

Uočene su dve manjinske klase, jedna sa 24 instance i druga sa 27 instanci. Izvršene su dve iteracije primene *SMOTE* algoritma sa navedenim parametrima. Nakon druge iteracije ukupan broj instanci je bio n=327. Obzirom da su se sintetički kreirane instance manjinske klase koncentrisano generisale na kraju skupa, implementiran je filter *Randomize* i na taj način ostvaren nasumičan redosled instanci u obučavajućem skupu. Performanse generisanih modela prikazani su u tabeli 5.7.

Tabela 5.7. Performanse modela nakon primene Smote i Randomize

Model	HD_Smote_Randomize
NB	85.59
HNB	86.49
J48	83.79
RF	93.69

SMOTE metodom izvršeno je sintetičko kreiranje instanci manjiske klase kako bi se uspostavila bolja balansiranost u raspodeli vrednosti. Za posmatrani obrazovni skup koji je disktretizovan histogram metodom, uz primenu Skotovog pravila optimalnog izračuvanja, primenjeni klasifikatori su ostvarili poboljšanje tačnosti modela predviđanja. *Random forest* klasifikator se izdvojio ostvarenom tačnošću od 93.69% i greškom klasifikacije od samo 6.31%. Na osnovo sprovedenih eksperimenata, utvrđen je postupak diskretizacije obrazovnog skupa mešanog okruženja učenja. Nenadzirana histogram metoda diskretizacije zasnovana na Scotovom pravilu izračunavanja širine intervala omogućava kreiranje preciznih modela predviđanja uz minimalan gubitak informacija u tom procesu.

5.3. Određivanje značaja vektora obeležja obrazovnog skupa

Ulazni vektor, koji sadrži irelevantna obeležja, prouzrokuje složen model predviđanja lošijih performansi. Izbor metode selekcije zavisi od više faktora, a usmerena je ka postizanju što realnijih rezultata. Odluka o primeni odgovarajuće metode se zasnova na prethodnom utvrđivanju potrebnog vremena za izvršavanje, prikaza rezultujućeg izlaza, odnosa između očekivanog i ukupnog broja obeležja, dimenzionalnosti klasnog obeležja, tipu i kvalitetu podataka, odnosu između broja obeležja i ukupnog broja instanci u skupu [151]. Određivanje značaja ulaznih obeležja obrazovnog skupa mešanog okruženja učenja izvršeno je filter algoritmima *Information-Gain (IG)*, *Gain Ratio (GR)*, *SymmetricalUncertAttributeEval (SUAE)*, *Relief (RLF)*, *ChiSquaredAttributeEval (CHI)* i pretragom prostora obeležja metodom rangiranja (engl. *ranker search*). Izdvajanje vektora obeležja realizovano je algoritmom omotača *Wrapper Subset Evaluation (WSE*) sa metodom pohlepne postepene pretrage unapred (engl. *Greedy Step-wise Forward*). Sprovedena su tri postupka za svaki od klasifikatora *NB*, *HNB*, *J48* i *RF*.

Originalni vektor karakteristika sadržao je 18 ulaznih obeležja *LAB, BB, DZ1, DZ2, DZ3, DZ4, DZ5, P1, P2, P3, T1, T2, ISPIT, PDF, LVT, LESS_ACC, FD, MM*. Definisano višedimenzionalno obeležje *Ocena* označeno je sa sedam diskretnih klasnih oznaka: *nije_polagao, pao, sest, sedam, osam, devet, deset*. Izvršena je diskretizacija vrednosti domena prethodno opisanom nenadziranom metodom histogram diskretizacije. Rezultati primenjenih filter metoda i metode omotača navedeni su u tabeli 5.8 i tabeli 5.9 respektivno.

Tabela 5.8. Rezultat primene filter metoda

RANG	Prosečan značaj	IG	Prosečan značaj	GR	Prosečan značaj	SUAE	Prosečan značaj	RLF	Prosečan značaj	CHI
1	1.616	ISPIT	0.638	ISPIT	O.608	ISPIT	0.678	ISPIT	655.807	ISPIT
2	1.329	T2	O.492	T2	0.485	T2	0.568	T2	502.477	T2
3	1.149	T1	0.458	DZ5	0.417	T1	0.494	T1	396.188	T1
4	0.829	DZ5	O.422	T1	0.361	DZ5	0.379	DZ5	261.527	DZ5
5	0.742	DZ4	0.408	DZ3	0.31	DZ3	0.358	DZ4	247.299	LAB
6	0.694	DZ3	0.367	DZ2	0.305	DZ4	0.351	DZ1	221.138	DZ4
7	0.681	BB	0.353	LAB	0.291	DZ2	0.345	BB	215.434	BB
8	0.669	DZ2	0.356	DZ4	0.274	DZ1	0.322	DZ2	200.55	DZ3
9	0.642	DZ1	0.336	DZ1	0.27	LAB	0.321	DZ3	204.198	DZ2
10	0.608	LAB	0.269	BB	0.256	BB	0.308	LAB	192.734	DZ1
11	0.3	P3	0.15	P3	0.126	P3	0.193	P3	80.655	P3
12	0.25	P2	0.131	P1	0.107	P1	0.173	P2	75.701	P1
13	0.25	P1	0.13	P2	0.106	P2	0.13	P1	70.407	P2
14	0.185	LVT	0.092	MM	0.07	LVT	0.041	LVT	32.374	LVT
15	0.092	PDF	0.074	LVT	0.041	LESSAC	0.032	PDF	27.205	PDF
16	0.088	LESSAC	0.057	LESSAC	0.036	PDF	0.022	LESSAC	16.581	LESSAC
17	0.039	FD	0.038	PDF	0.021	FD	-0.003	MM	5.88	MM
18	0.23	MM	0.041	FD	0.015	MM	-0.007	FD	5.657	FD

Tabela 5.9. Rezultat primene wrapper metoda

Metode omotača Wrapper Subset Evaluator	Značaj (Merit)	Izdvojen vektor obeležja
RF	0.892	DZ1, DZ2, DZ4, DZ5, T1, T2,ISPIT, FD, MM
HNB	0.889	LAB, DZ1, DZ2, DZ4, DZ5, T1, P2,T2,ISPIT, FD
NB	0.872	LAB, BB, DZ1, DZ2, DZ5, P1 , T1,T2,ISPIT, LVT
J48	0.843	DZ4, DZ5,T2,P3,ISPIT

Primena četiri filter metode izdvojila je obeležja *ISPIT* i *T2*. Utvrđena vrednost ranga za obeležje *ISPIT*, RANG (*ISPIT*) =1 i za obeležje T2, RANG(*T2*)=2, ukazuje na njihov značaj i uticaj na klasno obeležje. Vrednosti ranga za *T1, DZ1, DZ2, DZ3, DZ4, DZ5, LAB* i *BB* takođe su ukazivale na uticaj i ovih obeležja na klasno obeležje. Niže vrednosti ranga utvrđene su u slučaju *P1, P2, P3, PDF, LVT, LESS_AC, FD, MM* obeležja. Metoda omotača, za sva četri klasifikatora, izdvojila je obeležja *DZ5, T2, ISPIT*. Iako je osnovni cilj metoda izbora izdvajanje podskupa minimalne kardinalnosti, od velikog je značaja i koji ća obeležja biti odbačena. Pogrešno odbačena ili izabrana obeležja u obrazovnom domenu mogu da prozrokuju slabije performanse u slučaju klasifikacije. Zbog toga je neophodno bilo utvrditi uticaj ulaznih vektora obeležja različite kardinalnosti na pouzdanost i preciznost modela predviđanja.

Izračunavanje mere zajedničke informacije (engl. *mutual information*, *MI*) izvršeno je za pojedinačne kombinacije ulaznih i klasnog obeležja *OCENA*. Mera zajedničke informacije [152] je informaciono-teorijska mera statističkih zavisnosti između dve varijable kako bi se utvrdila količina zajedničkih informacija koje dele. Ako se X i Y nezavisni, onda X ne sadrži informacije o Y i obrnuto, pri čemu je njihova zajednička informacija nula. U drugoj krajnosti, kada su X i Y identični, onda dele i sve informacije.

Tabela 5.10. Rezultat izračunavanja mere zajedničkih informacija - $MI(O_i, Ocena)$

ISPIT	T2	T1	DZ5	LAB	DZ3	DZ4	DZ1	DZ2	BB	P1	P2	P3
1,61	1,3	1,13	0,8	0,66	0,65	0,65	0,61	0,61	0,6	0,23	0,21	0,2

LVT	PDF	LESS AC	MM	FD
0,09	0,07	0,04	0,02	0,01

U tabeli 5.10. prikazani su rezultati određivanja postojanja značajnog uticaja između ulaznih obeležja i klasnog obeležja *Ocena* postupkom izračunavanja vrednosti mere zajedničke informacije. Postavljena je vrednost donjeg praga značajnosti na $MI(O_i, OCENA)_{min} = 0.1$. Za obeležja *LVT, PDF, LESS_AC, MM, FD* su utvrđene vrednosti mere zajedničke informacije manje od postavljenog praga MI_{min}. Na osnovu rezultata prikazanih u tabeli 5.1 izdvedena je i nejadnakost

$$MI(FD, OCENA) < MI(MM, OCENA) < MI(LESS_{AC}, OCENA) < MI(PDF, OCENA) < MI(LVT, OCENA) < MI(O_i, OCENA)_{min}$$

Zaključeno je da *LVT, PDF, LESS_AC, MM, FD* nemaju uticaj na klasno obeležje Ocena pa nisu uzeti u obzir pri formiranju optimalnog vektora obeležja.

Na osnovu rezultata filter algoritama, algoritma omotača i mere zajedničkih informacija, izdvojen je optimalan vektor značajnih obeležja *Student (LAB, BB, DZ1, DZ2, DZ3, DZ4, DZ5, P1,P2, ISPIT)* za obrazovni skup podataka mešanog okruženja učenja.

U tabeli 5.11 dat je prikaz pridruženih nominalnih oznaka novodobijenih diskretnih vrednosti.

Tabela 5.11. Prikaz pridruženih nominalnih oznaka ulaznih i klasnog obeležja

LAB		
Interval	F	Oznaka
(min-6.39]	34	LAB1
(6.39-10.64]	28	LAB2
(10.64-12.77]	92	LAB3
(12.77-max)	122	LAB4

BB		
Interval	F	Oznaka
(min-0.71]	32	BB1
(0.71-2.12]	48	BB2
(2.12-2.83]	40	BB3
(2.83-3.54]	38	BB4
(3.54-4.24]	40	BB5
(4.24-max)	78	BB6

DZ1		
Interval	F	Oznaka
(min-0.8]	88	DZ11
(0.8-1.2]	37	DZ12
(1.2-1.6]	62	DZ13
(1.6-max]	89	DZ14

DZ2		
Interval	F	Oznaka
(min-0.4]	92	DZ21
(0.4-0.8]	15	DZ22
(0.8-1.2]	13	DZ23
(1.2-1.6]	35	DZ24
(1.6-max)	121	DZ2

DZ3		
Interval	F	Oznaka
(min-0.8]	100	DZ31
(0.8-1.2]	21	DZ32
(1.2-1.6]	29	DZ33
(1.6-max)	126	DZ43

DZ4		
Interval	F	Oznaka
(min-0.4]	101	DZ41
(0.4-0.8]	16	DZ42
(0.8-1.2]	31	DZ43
(1.2-1.6]	54	DZ44
(1.6-max)	74	DZ45

DZ5		
Interval	F	Oznaka
(min-0.4]	109	DZ51
(0.4-0.8]	10	DZ52
(0.8-1.2]	29	DZ53
(1.2-1.6]	22	DZ54
(1.6-max)	106	DZ55

P3		
Interval	F	Oznaka
(min-6]	131	P31
(6-12]	18	P32
(12-18]	18	P33
(18-24]	45	P34
(24-max)	64	P35

P1		
Interval	F	Oznaka
(min-0.33]	150	P11
(0.33-6.67]	17	P12
(6.67-10]	26	P13
(10-13-33]	22	P14
(13.33-16.67]	24	P15
(16.67-max)	37	P16

T1		
Interval	F	Oznaka
(min-2.86]	37	T11
(2.86-5.71]	27	T12
(5.71-8.57]	15	T13
(8.57-11.43]	41	T14
(11.43-14.29]	51	T15
(14.29-17.14]	37	T16
(17.14-max)	68	T17

ISPIT		
Interval	F	Oznaka
(min-5]	52	ISPIT1
(5-10]	16	ISPIT2
(10-15]	31	ISPIT3
(15-20]	58	ISPIT4
(20-25]	44	ISPIT5
(25-max]	75	ISPIT6

P2		
Interval	F	Oznaka
(min-4]	112	P21
(4-8]	14	P22
(8-12]	24	P23
(12-16]	41	P24
(16-max)	85	P25

T2		
Interval	F	Oznaka
(min-2.86]	43	T21
(2.86-5.71]	13	T22
(5.71-8.57]	20	T23
(8.57-11.43]	43	T24
(11.43-14.29]	59	T25
(14.29-17.14]	45	T26
(17.14-max)	53	T27

6. ASOCIJATIVNA ANALIZA OBRAZOVNIH PODATAKA MEŠANOG OKRUŽENJA UČENJA

U ovom poglavlju disertacije prikazana je asocijativna deskriptivna analiza u mešanom okruženju učenja zasnovana na studiji slučaja prezentovanoj u radu [148]. Sprovedena su dva eksperimenta i to: otkrivanje korelacija od značaja između ulaznog vektora obeležja i unapređenje procesa e-testiranja. Procena značaja otkrivenih pravila izvršena je primenom objektivnog i subjektivnog pristupa. Asocijativna analiza zasnovana je na primeni algoritama za generisanje pravila udruživanja. Izdvojena značajna pravila formiraju obrasce korelacija kojima se definiše homogenost i predstavljaju suštinska svojstva u podacima. Predložen je koncept asocijativne deskriptivne analize mešanog okruženja učenja usmeren ka ostvarivanju poboljšanja peformansi studenata.

6.1. Pravila udruživanja

Pravila udruživanja se mogu opisati kao specijalan slučaj probablističkih pravila kojima se realizuje povezivanje jednog ili više obeležja skupa podataka sa drugim u AKO-ONDA (engl. *IF-THEN*) formu posmatranjem njihovih vrednosti. Generisanje *IF-THEN* pravila može se podeliti u dva zadatka. Prvi zadatak je pronalaženje onih skupova stavki čija verovatnoća dešavanja dostiže predhodno definisani prag podrške; ovakvi skupovi nazivaju se frekventni skupovi stavki. Drugi zadatak je generisanje pravila udruživanja iz predhodno dobijenih velikih skupova stavki, a koja su ograničena sa minimalnom podrškom. Apriori algoritam iterativno smanjuje minimalnu podršku dok ne pronađe pravila koja imaju poverenje jednako ili veće od postavljenog minimalnog poverenja. Interesantnost pravila zavisi i od samog problema koji treba da se reši. U mnogim slučajevima, korisniku je potrebno da pronađe pravila kod kojih postoji veza u slučaju realnog okruženja. Prediktivni Apriori algoritam (engl. *predicitive Apriori algorithm*) [93] generiše *n* pravila za koje može biti postignuta maksimalna tačnost predviđanja. Neophodno je unapred definisati broj pravila (*n*), ali nije potrebno postavljati vrednosti za minimalnu podršku i poverenje kao u slučaju Apriori algoritma. Značajnost

prediktivnog Aprirori algoritma je svakako u činjenici da je u pitanju dinamička tehnika određivanja pravila koja koriste gornje granice tačnosti svih pravila.

Interpretacija rezultata podrazumevala je analizu otkrivenih pravila sa ciljem obezbeđivanja povratnih informacija značajnih za sprovedenu studiju slučaja mešanog okruženja učenja. Postavljanjem ograničenja za poverenje i podršku Apriori algoritam je izdvojio pravila od značaja. Korisnost i interesantnost pravila su subjektivni koncepti koje je teško efektivno kvantifikovati. Procena otkrivenih pravila izvršena je primenom objektivnog i subjektivnog pristupa [153]. Objektivan pristup za procenu značajnosti pravila realizovan je lift metrikom. Pravila sa vrednošću lift parametra većom od jedan označena su kao interesantna. Subjektivan pristup procene pravila zasnivan je na znanju korisnika o domenu izučavanog problema.

6.2. Otkrivanje korelacija ulaznog vektora obeležja

Za istraživanje sprovedeno na Visokoj školi elektrotehnike i računarstva strukovnih studija u Beogradu izdvojena je populacija od 276 studenata Moodle kursa računarska grafika. Asocijativna analiza sprovedena je implementacijom Apriori i prediktivnog Apriori algoritma između ulaznog vektora $(DZ1, DZ2, DZ3, DZ4, DZ5, LAB, BB, P1, P2, P3, T1, T2, ISPIT)$ i klasnog obeležja *OCENA* obrazovnog skupa mešanog okruženja učenja. Pronađena pravila su u formi "*Ako prethodi onda i sledi*", gde su *"prethodi"* i *"sledi"* skupovi stavki. Apriori algoritam je implementiran postavljenjem minimalne vrednost poverenja na *min(Conf)=0.5* i minimalne vrednosti podrške na *min(Supp)=0.3*. Za broj transakcija postavljen je parametar *numRules=500*.

Eksperiment je sproveden u okruženju Weka sistema za analizu znanja, verzija 3.8.2. Formirani su skupovi stavki L(1), L(2), L(3), L(4) sa po 15, 14, 5, 1 stavki respektivno. Algoritam je izvršio 14 ciklusa, pri čemu je izdvojeno 83 instance sa podrškom većom ili jednakom od minimalne. Na osnovu skupova stavki, izdvojena su 72 pravila koja zadovoljavaju uslov unapred postavljenog minimalnog praga poverenja *min(Conf)=0.5 i min(Supp)=0.3*.

U nastavku su prikazani skupovi stavki i formirana pravila generisana u Weka okruženju

L(1) – 15 stavki
LAB=aktivan 122
LAB=dobar 92
DZ1='(1.6-inf)' 89
DZ1=merged 88
DZ2='(-inf-0.4]' 92
DZ2='(1.6-inf)' 121
DZ3='(-inf-0.4]' 97
DZ3='(1.6-inf)' 126
DZ4='(-inf-0.4]' 101
DZ5='(-inf-0.4]' 109
DZ5='(1.6-inf)' 106
P1='(-inf-3.333333]' 150
P2='(-inf-4]' 112
P2='(16-inf)' 85
P3='(-inf-6]' 131

L(2)-14 stavki
LAB=aktivan DZ2='(1.6-inf)' 85
LAB=aktivan DZ3='(1.6-inf)' 90
DZ2='(-inf-0.4]' DZ3='(-inf-0.4]' 85
DZ2='(-inf-0.4]' DZ4='(-inf-0.4]' 85
DZ2='(-inf-0.4]' DZ5='(-inf-0.4]' 86
DZ2='(1.6-inf)' DZ3='(1.6-inf)' 96
DZ2='(1.6-inf)' DZ5='(1.6-inf)' 89
DZ3='(-inf-0.4]' DZ4='(-inf-0.4]' 91
DZ3='(-inf-0.4]' DZ5='(-inf-0.4]' 97
DZ3='(1.6-inf)' DZ5='(1.6-inf)' 90
DZ4='(-inf-0.4]' DZ5='(-inf-0.4]' 93
LAB=aktivan DZ2='(1.6-inf)' 85
LAB=aktivan DZ3='(1.6-inf)' 90
DZ2='(-inf-0.4]' DZ3='(-inf-0.4]' 85

L(3)-5 stavki
DZ2='(-inf-0.4]' DZ3='(-inf-0.4]'
DZ4='(-inf-0.4]' 85
DZ2='(-inf-0.4]' DZ3='(-inf-0.4]'
DZ5='(-inf-0.4]' 85
DZ2='(-inf-0.4]' DZ4='(-inf-0.4]'
DZ5='(-inf-0.4]' 85
DZ3='(-inf-0.4]' DZ4='(-inf-0.4]'
DZ5='(-inf-0.4]' 91
P1='(-inf-3.333333]' P2='(-inf-4]'
P3='(-inf-6]' 84
L(4)-1 stavka
DZ2='(-inf-0.4]' DZ3='(-inf-0.4]'
DZ4='(-inf-0.4]' DZ5='(-inf-0.4]' 85

1. DZ3=DZ31 97 ==> DZ5=DZ51 97 <conf:(1)> lift:(2.53) lev:(0.21) [58] conv:(58.69)

2. DZ3=DZ31 DZ4=DZ41 91 ==> DZ5=DZ51 91 <conf:(1)> lift:(2.53) lev:(0.2) [55] conv:(55.06)

3. DZ2=DZ21 DZ4=DZ41 85 ==> DZ3=DZ31 85 <conf:(1)> lift:(2.85) lev:(0.2) [55] conv:(55.13)

4. DZ2=DZ21 DZ3=DZ31 85 ==> DZ4=DZ41 85 <conf:(1)> lift:(2.73) lev:(0.2) [53] conv:(53.89)

5. DZ2=DZ21 DZ3=DZ31 85 ==> DZ5=DZ51 85 <conf:(1)> lift:(2.53) lev:(0.19) [51] conv:(51.43)

6. DZ2=DZ21 DZ4=DZ41 85 ==> DZ5=DZ51 85 <conf:(1)> lift:(2.53) lev:(0.19) [51] conv:(51.43)

7. DZ2=DZ21 DZ4=DZ41 DZ5=DZ51 85 ==> DZ3=DZ31 85 <conf:(1)> lift:(2.85) lev:(0.2) [55] conv:(55.13)

8. DZ2=DZ21 DZ3=DZ31 DZ5=DZ51 85 ==> DZ4=DZ41 85 <conf:(1)> lift:(2.73) lev:(0.2) [53] conv:(53.89)

9. DZ2=DZ21 DZ3=DZ31 DZ4=DZ41 85 ==> DZ5=DZ51 85 <conf:(1)> lift:(2.53) lev:(0.19) [51] conv:(51.43)

10. DZ2=DZ21 DZ4=DZ41 85 ==> DZ3=DZ31 DZ5=DZ51 85 <conf:(1)> lift:(2.85) lev:(0.2) [55] conv:(55.13)

11. DZ2=DZ21 DZ3=DZ31 85 ==> DZ4=DZ41 DZ5=DZ51 85 <conf:(1)> lift:(2.97) lev:(0.2) [56] conv:(56.36)

12. DZ2=DZ21 DZ5=DZ51 86 ==> DZ3=DZ31 85 <conf:(0.99)> lift:(2.81) lev:(0.2) [54] conv:(27.89)

13. DZ2=DZ21 DZ5=DZ51 86 ==> DZ4=DZ41 85 <conf:(0.99)> lift:(2.7) lev:(0.19) [53] conv:(27.26)

14. DZ2=DZ21 DZ5=DZ51 86 ==> DZ3=DZ31 DZ4=DZ41 85 <conf:(0.99)> lift:(3) lev:(0.21) [56] conv:(28.82)

15. P2=P21 P3=P31 85 ==> P1=P11 84 <conf:(0.99)> lift:(1.82) lev:(0.14) [37] conv:(19.4)

16. DZ4=DZ41 DZ5=DZ51 93 ==> DZ3=DZ31 91 <conf:(0.98)> lift:(2.78) lev:(0.21) [58] conv:(20.11)

17. DZ3=DZ31 97 ==> DZ4=DZ41 91 <conf:(0.94)> lift:(2.56) lev:(0.2) [55] conv:(8.79)

18. DZ3=DZ31 DZ5=DZ51 97 ==> DZ4=DZ41 91 <conf:(0.94)> lift:(2.56) lev:(0.2) [55] conv:(8.79)

19. DZ3=DZ31 97 ==> DZ4=DZ41 DZ5=DZ51 91 <conf:(0.94)> lift:(2.78) lev:(0.21) [58] conv:(9.19)

20. DZ2=DZ21 92 ==> DZ5=DZ51 86 <conf:(0.93)> lift:(2.37) lev:(0.18) [49] conv:(7.95)

21. DZ3=DZ31 DZ4=DZ41 91 ==> DZ2=DZ21 85 <conf:(0.93)> lift:(2.8) lev:(0.2) [54] conv:(8.67)

22. DZ3=DZ31 DZ4=DZ41 DZ5=DZ51 91 ==> DZ2=DZ21 85 <conf:(0.93)> lift:(2.8) lev:(0.2) [54] conv:(8.67)

23. DZ3=DZ31 DZ4=DZ41 91 ==> DZ2=DZ21 DZ5=DZ51 85 <conf:(0.93)> lift:(3) lev:(0.21) [56] conv:(8.95)

24. DZ2=DZ21 92 ==> DZ3=DZ31 85 <conf:(0.92)> lift:(2.63) lev:(0.19) [52] conv:(7.46)

25. DZ2=DZ21 92 ==> DZ4=DZ41 85 <conf:(0.92)> lift:(2.52) lev:(0.19) [51] conv:(7.29)

26. DZ2=DZ21 92 ==> DZ3=DZ31 DZ4=DZ41 85 <conf:(0.92)> lift:(2.8) lev:(0.2) [54] conv:(7.71)

27. DZ2=DZ21 92 ==> DZ3=DZ31 DZ5=DZ51 85 <conf:(0.92)> lift:(2.63) lev:(0.19) [52] conv:(7.46)

28. DZ2=DZ21 92 ==> DZ4=DZ41 DZ5=DZ51 85 <conf:(0.92)> lift:(2.74) lev:(0.2) [53] conv:(7.63)

29. DZ2=DZ21 92 ==> DZ3=DZ31 DZ4=DZ41 DZ5=DZ51 85 <conf:(0.92)> lift:(2.8) lev:(0.2) [54] conv:(7.71)

30. DZ4=DZ41 101 ==> DZ5=DZ51 93 <conf:(0.92)> lift:(2.33) lev:(0.19) [53] conv:(6.79)

31. DZ4=DZ41 DZ5=DZ51 93 ==> DZ2=DZ21 85 <conf:(0.91)> lift:(2.74) lev:(0.2) [53] conv:(6.89)

32. DZ4=DZ41 DZ5=DZ51 93 ==> DZ2=DZ21 DZ3=DZ31 85 <conf:(0.91)> lift:(2.97) lev:(0.2) [56] conv:(7.15)

33. P2=P21 112 ==> P1=P11 101 <conf:(0.9)> lift:(1.66) lev:(0.15) [40] conv:(4.26)

34. DZ4=DZ41 101 ==> DZ3=DZ31 91 <conf:(0.9)> lift:(2.56) lev:(0.2) [55] conv:(5.95)

35. DZ4=DZ41 101 ==> DZ3=DZ31 DZ5=DZ51 91 <conf:(0.9)> lift:(2.56) lev:(0.2) [55] conv:(5.95)

36. DZ5=DZ51 109 ==> DZ3=DZ31 97 <conf:(0.89)> lift:(2.53) lev:(0.21) [58] conv:(5.44)

37. DZ3=DZ31 97 ==> DZ2=DZ21 85 <conf:(0.88)> lift:(2.63) lev:(0.19) [52] conv:(4.97)

38. DZ3=DZ31 97 ==> DZ2=DZ21 DZ4=DZ41 85 <conf:(0.88)> lift:(2.85) lev:(0.2) [55] conv:(5.16)

39. DZ3=DZ31 DZ5=DZ51 97 ==> DZ2=DZ21 85 <conf:(0.88)> lift:(2.63) lev:(0.19) [52] conv:(4.97)

40. DZ3=DZ31 97 ==> DZ2=DZ21 DZ5=DZ51 85 <conf:(0.88)> lift:(2.81) lev:(0.2) [54] conv:(5.14)

41. DZ3=DZ31 DZ5=DZ51 97 ==> DZ2=DZ21 DZ4=DZ41 85 <conf:(0.88)> lift:(2.85) lev:(0.2) [55] conv:(5.16)

42. DZ3=DZ31 97 ==> DZ2=DZ21 DZ4=DZ41 DZ5=DZ51 85 <conf:(0.88)> lift:(2.85) lev:(0.2) [55] conv:(5.16)

43. DZ5=DZ51 109 ==> DZ4=DZ41 93 <conf:(0.85)> lift:(2.33) lev:(0.19) [53] conv:(4.07)

44. DZ5=DZ55 106 ==> DZ3=DZ35 90 <conf:(0.85)> lift:(1.86) lev:(0.15) [41] conv:(3.39)

45. DZ4=DZ41 101 ==> DZ2=DZ21 85 <conf:(0.84)> lift:(2.52) lev:(0.19) [51] conv:(3.96)

46. DZ4=DZ41 101 ==> DZ2=DZ21 DZ3=DZ31 85 <conf:(0.84)> lift:(2.73) lev:(0.2) [53] conv:(4.11)

47. DZ4=DZ41 101 ==> DZ2=DZ21 DZ5=DZ51 85 <conf:(0.84)> lift:(2.7) lev:(0.19) [53] conv:(4.09)

48. DZ4=DZ41 101 ==> DZ2=DZ21 DZ3=DZ31 DZ5=DZ51 85 <conf:(0.84)> lift:(2.73) lev:(0.2) [53] conv:(4.11)

49. DZ5=DZ55 106 ==> DZ2=DZ25 89 <conf:(0.84)> lift:(1.92) lev:(0.15) [42] conv:(3.31)

50. DZ5=DZ51 109 ==> DZ3=DZ31 DZ4=DZ41 91 <conf:(0.83)> lift:(2.53) lev:(0.2) [55] conv:(3.85)

51. P1=P11 P2=P21 101 ==> P3=P31 84 <conf:(0.83)> lift:(1.75) lev:(0.13) [36] conv:(2.95)

52. P3=P31 131 ==> P1=P11 105 <conf:(0.8)> lift:(1.47) lev:(0.12) [33] conv:(2.21)

53. P1=P11 P3=P31 105 ==> P2=P21 84 <conf:(0.8)> lift:(1.97) lev:(0.15) [41] conv:(2.84)

54. DZ2=DZ25 121 ==> DZ3=DZ35 96 <conf:(0.79)> lift:(1.74) lev:(0.15) [40] conv:(2.53)

55. DZ5=DZ51 109 ==> DZ2=DZ21 86 <conf:(0.79)> lift:(2.37) lev:(0.18) [49] conv:(3.03)

56. DZ5=DZ51 109 ==> DZ2=DZ21 DZ3=DZ31 85 <conf:(0.78)> lift:(2.53) lev:(0.19) [51] conv:(3.02)

57. DZ5=DZ51 109 ==> DZ2=DZ21 DZ4=DZ41 85 <conf:(0.78)> lift:(2.53) lev:(0.19) [51] conv:(3.02)

58. DZ5=DZ51 109 ==> DZ2=DZ21 DZ3=DZ31 DZ4=DZ41 85 <conf:(0.78)> lift:(2.53) lev:(0.19) [51] conv:(3.02)

59. DZ3=DZ35 126 ==> DZ2=DZ25 96 <conf:(0.76)> lift:(1.74) lev:(0.15) [40] conv:(2.28)

60. P2=P21 112 ==> P3=P31 85 <conf:(0.76)> lift:(1.6) lev:(0.12) [31] conv:(2.1)

61. P2=P21 112 ==> P1=P11 P3=P31 84 <conf:(0.75)> lift:(1.97) lev:(0.15) [41] conv:(2.39)

62. LAB=LAB4 122 ==> DZ3=DZ35 90 <conf:(0.74)> lift:(1.62) lev:(0.12) [34] conv:(2.01)

63. DZ2=DZ25 121 ==> DZ5=DZ55 89 <conf:(0.74)> lift:(1.92) lev:(0.15) [42] conv:(2.26)

64. DZ3=DZ35 126 ==> LAB=LAB4 90 <conf:(0.71)> lift:(1.62) lev:(0.12) [34] conv:(1.9)

65. DZ3=DZ35 126 ==> DZ5=DZ55 90 <conf:(0.71)> lift:(1.86) lev:(0.15) [41] conv:(2.1)

66. DZ2=DZ25 121 ==> LAB=LAB4 85 <conf:(0.7)> lift:(1.59) lev:(0.11) [31] conv:(1.82)

67. P1=P11 150 ==> P3=P31 105 <conf:(0.7)> lift:(1.47) lev:(0.12) [33] conv:(1.71)

68. LAB=LAB4 122 ==> DZ2=DZ25 85 <conf:(0.7)> lift:(1.59) lev:(0.11) [31] conv:(1.8)

69. P1=P11 150 ==> P2=P21 101 <conf:(0.67)> lift:(1.66) lev:(0.15) [40] conv:(1.78)

70. P3=P31 131 ==> P2=P21 85 <conf:(0.65)> lift:(1.6) lev:(0.12) [31] conv:(1.66)

71. P3=P31 131 ==> P1=P11 P2=P21 84 <conf:(0.64)> lift:(1.75) lev:(0.13) [36] conv:(1.73)

72. P1=P11 150 ==> P2=P21 P3=P31 84 <conf:(0.56)> lift:(1.82) lev:(0.14) [37] conv:(1.55)

Subjektivan pristup podrazumevao je identifikaciju opštih obrazaca i veza sa ciljem izdvajanja pravila koji zadovoljavaju utvrđeni šablon. Izdvojena pravila označena kao značajna na osnovu lift metrike klasifikovana su prema povratnim informacijama koje obezbeđuju. Utvrđeno je da pravila pokazuju jake korelacije između resursa Moodle kursa označenih kao domaći zadaci. Subjektivnim pristupom uočeno je pet korisnih obrazaca korelacija:

- *Obrazac1*: *IF* student osvoji na drugom, trećem i četvrtom domaćem zadatku manje od od 50% *THEN* student će na petom domaćem zadatku osvojiti manje od 50%.
- *Obrazac2*: *IF* student osvoji na drugom domaćem zadatku više od 80% *THEN* student će i na petom osvojiti više od 80%.
- *Obrazac3*: *IF* student osvoji na drugom domaćem zadatku više od 80% *THEN* student će i na trećem osvojiti više od 80%.
- *Obrazac4*: *IF* student loše odradi prvi i drugi pripremni test *THEN* student će loše odraditi i treći pripremni test.
- *Obrazac5*: IF student je student aktivan i efikasno realizuje zadatke na laboratorijskim
- vežbama THEN student će dobro uraditi drugi i treći domaći zadatak.

Uočeni odnosi predstavljaju osnovu koncepta asocijativnog modela mešanog okruženja učenja koji obezbeđuje povratne informacije za detekciju delova gradiva koji su usko povezani. Nastavnik može proveriti da li postoji mogućnost spajanja usko povezanih delova gradiva u jedan širi koncept i izvršiti potrebne modifikacije u cilju poboljšanja procesa učenja. S druge strane, ovi odnosi ukazuju nastavniku na koncepte gradiva koji imaju najviše uticaja na ostvaren dobar rezultat kao i na koncepte koji imaju slabiji uticaj tako da nastavnik može proveriti njihov sadržaj i po potrebi ga proširiti i poboljšati.

6.3. Asocijativna analiza Moodle testova

Drugi eksperiment je podrazumevao asocijativnu analizu probnih i zvaničnih testova u elektronskom obliku realizovanih na Moodle sistemu. U okviru klasičnog oblika nastave organizuju se tri časa predavanja i dva časa laboratorijskih vežbi nedeljno. Na predavanjima su objašnjavani pojmovi i koncepti iz oblasti računarske grafike, a na vežbama studenti samostalno realizuju praktične zadatke.

U toku semestra, studenti su polagali dva kolokvijuma preko kojih je vršena provera znanja iz oblasti tehnologija ulaza, izlaza, geometrijskih transformacija, osnovnih rasterskih algoritama. Završni test je obuhvatao praktične zadatke iz oblasti interfejsa OpenGL (*Open Graphics Library*) i održavao se u terminu ispita. Provere znanja nisu bile eliminatornog

karaktera, već su korišćene kako bi studenti prikupili poene na predispitnim obavezama. Pored redovnih testova, bodovana je aktivnost studenata na laboratorijskim vežbama kao i učešće u aktivnostima na Moodle kursu.

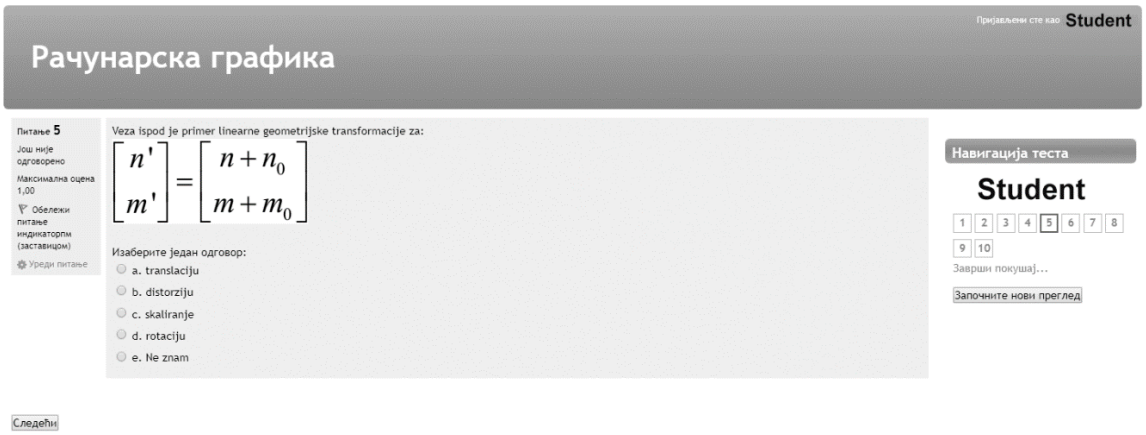
U okviru Moodle kursa studentima su bili dostupni i probni testovi za samotestiranje. Zvanične provere znanja su realizovane kroz tri Moodle testa. Semestralni testovi (prvi i drugi kolokvijum) obuhvatali su proveru znanja iz gradiva obrađenog do momenta testiranja. Prvi kolokvijum je održan u aprilu, a drugi u maju mesecu. Sadržali su po 10 pitanja (5 teoretskih i 5 praktičnih zadataka), maksimalno je bilo moguće osvojiti 20 bodova po testu, netačan odgovor je označen sa nula bodova, vremensko trajanje je postavljeno na 40 minuta. Teorijska pitanja su bodovana sa jednim poenom, a zadaci sa tri poena.

Završni test je sadržao 6 pitanja u vidu zadataka, pri čemu je student mogao osvojiti najviše 5 bodova po pitanju, a trajanje testa je ograničeno na 40 minuta. Korisničko okruženje i navigaciona struktura svih testova je podešena tako da se svako pitanje prikazivalo na posebnoj stranici, a student je pomoću navigacije sa desne strane mogao nasumično da pristupi pitanjima. Pitanje na koje nije dat odgovor kao i pitanja sa označenim odgovorom „ne znam“ su evidentirana sa nula bodova.

Student je testove radio iz laboratorije uz unos odgovarajuće pristupne lozinke. Pored toga, pristup testu je bio ograničen IP adresama i omogućen samo iz određenih laboratorija škole. Postavljena su podešavanja za otvaranje testa u posebnom prozoru pri čemu je isključena navigaciona kontrola browsera.

Pre kreiranja testova, napravljene su kategorije za svaku oblast što je omogućavalo organizacinu strukturu banke pitanja. Na osnovu toga, korišćena je opcija slučajnog izbora pitanja iz određene kategorije. Podešeno je mešanje odgovora u okviru pitanja čime je izbegnuta opcija kreiranja testova po grupama kako bi se sprečilo prepisivanje. Slučajnim generisanjem pitanja iz kategorija, studenti su dobijali različite testove.

Na slici 6.1 prikazan je primer jednog pitanja iz testa.



Slika 6.1: Korisničko okruženje Moodle testa za prvi kolokvijum iz predmeta Računarska grafika

Probni testovi su kreirani po uzoru na kolokvijume i ispit sa ciljem simulacije okruženja zvaničnog testiranja. Ovi testovi su bili dostupni tokom održavanja kursa, a studenti su mogli da im pristupe najviše 10 puta. Moodle resursi za učenje su bili preporučeni studentima u toku klasične nastave, ali su studenti pristupali ponuđenom elektronskom materijalu prema sopstvenom nahođenju, kao i samotestiranju. Na kraju semestra zabeleženo je 415 pokušaja rešavanja prvog probnog testa, pri čemu je 350 završenih, 55 nikada predatih, 236 završenih pokušaja za drugi probni test i 30 nikada predatih, 470 pokušaja za probni završni test, od toga 437 završenih, a 33 nikada predatih. Prilikom zvaničnih provera znanja student je imao samo jedan pokušaj rešavanja testa. Ukupan broj studenata koji su radili prvi semestralni test je bio 133, drugi 122, a završni test je 113.

U tabelama 6.1 i 6.2 dati su ostvareni rezulati probnih i zvaničnih testova. U tabeli 6.1. sa *K1_P, K2_P* označeni su probni testovi za prvi i drugi kolokvijum, sa *ZT_P* probni test za završni ispit, broj završenih testova označen je sa *PK_U*, broj nikad predatih pokušaja sa *NP*, minimalan broj ostvarenih bodova sa Min, maksimalan broj ostvarenih bodova sa Max, prosečan broj bodova sa *Avg(PK_U)*, broj pokušaja sa najbolje ostvarenim rezultatima sa *BNT*, prosečan broj bodova ostvaren na izdvojenim najboljim pokušajima rešavanja testova sa *Avg(BNT)*.

Tabela 6.1. Ostvareni rezultati na probnim testovima

Probni testovi	Ukupno pokušaja	PK_U	NP	Min	Max	Avg(PK_U)	BNT	Avg(BNT)
K1_P	415	350	55	0	20	12,93	110	17,13
K2_P	266	236	30	0	20	12,89	106	16,16
ZT_P	470	437	33	0	30	18,66	107	22,85

U tabeli 6.2. prvi i drugi kolokvijum označeni su sa *K1, K2* respektivno, završni test sa *ZT,* broj ukupnih pokušaja rešavanja zvaničnih testova označen je sa *U*, *minU* označava minimalan broj ostvarenih bodova u odnosu na sve pokušaje, sa *maxU* maksimalan broj ostvarenih bodova, a sa *AVG(U)* prosečan broj ostvarenih bodova.

Tabela 6.2. Ostvareni rezultati na zvaničnim testovima

Zvanični testovi	U	minU	maxU	Avg(U)
K1	133	3,12	20	14,2
K2	122	4,06	20	14,68
ZT	113	9,9	30	25,16

Asocijativna analiza podataka usmerena je na procenu efikasnosti pripreme studenata za zvanične provere znanja upotrebom probnih testova za samotestiranje dostupnih u toku procesa učenja. Za studiju slučaja opisanu u radu kreirane su tri matrice rezultata (engl. *Score matrix*) [147] koje su predstavljale bodove osvojene na pitanjima i ukupan ostvaren broj bodova za svakog studenta. Dodate su kolone za bodove ostvarene u najboljem pokušaju rešavanja probnog testa.

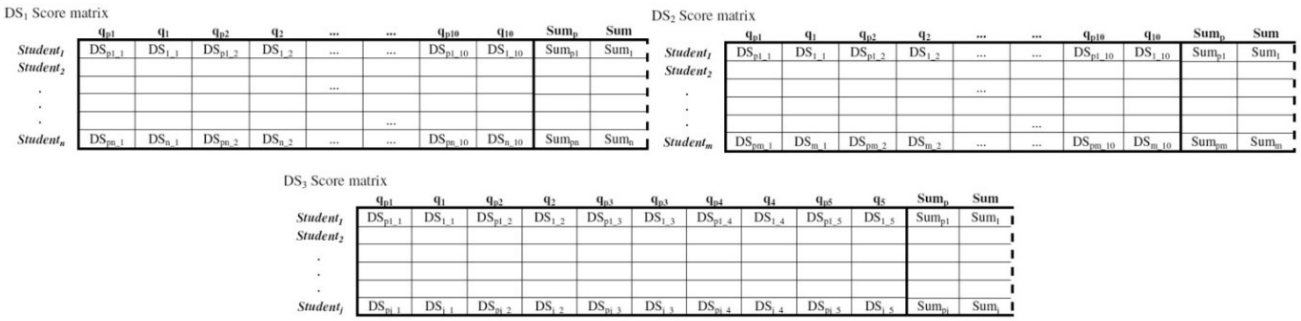
Za potrebe kreiranja matrica rezultata, generisani su Moodle izveštaji po pitanjima, preuzeti kao excel dokumenti, a zatim uveženi kao tabele u bazu podataka kreiranu pod PostgresSQL serverom. Nad kreiranim tabelama izvršeni su upiti u cilju izdvajanja skupova podataka sa potrebnim zapisima.

Kreirana su tri skupa podataka tako da je *DS1* predstavljao podatke iz prvog semestralnog i njegovog probnog testa, *DS2* podatke iz drugog semestralnog i njegovog probnog testa, a *DS3* podatke iz završnog i njegovog probnog testa. Redovi u izdvojenim skupovima predstavljali su zapis za svakog studenta o osvojenim bodovima na zvaničnom i probnom testu.

Kolone su predstavljale broj pitanja u testovima kao i ukupan broj osvojenih bodova po testu. Skupovi *DS1* i *DS2* imali su po 22 karakteristike (kolone), a skup *DS3* 12 karakteristika (kolona). Izdvojeno je 107 instanci studenata koji su radili probni test pre prvog kolokvijuma, 97 instanci studenata koji su radili probni test pre drugog kolokvijuma i 94 instance koje su predstavljale zapise o studentima koji su radili probni test pred završni ispit.

Elementi prve dve matrice su predstavljali bodove od 0 do 1 za pitanja, a za zadatke od 0 do 3. Ukupan broj ostvarenih bodova, od 0 do 20, na prva dva testa predstavljala je realnu vrednost koju je Moodle sistem automatski računao. Treća matrica je predstavljala osvojene bodove na zadacima (od 0 do 5) i ukupan broj ostvarenih poena na probnom i završnom testu.

Na slici 6.2 prikazane su kreirane matrice rezultata.



Slika 6.2: Matrice rezultata Moodle testa

Utvrđeno je da su u proseku, za prvi probni test ostvarena 3 završena pokušaja po studentu, za drugi probni test 2 pokušaja, a za probni završni 4 pokušaja. Analiza pokušaja probnih i zvaničnih testova data je u tabeli 6.3.

Tabela 6.3. Analiza pokušaja probnih i zvaničnih testova

Probni testovi	Ukupno pokušaja	PK_U	NP	BNT	Zvanični testovi	Broj studenta koji su radili probni i zvanični
K1	415	350	55	110	133	107
K2	266	236	30	106	122	97
ZT	470	437	33	107	113	94

Iz tabele 6.3 može se zaključiti da je prosečno 80% studenta koji su tokom semestra radili kolokvijume i izašli na ispit koristilo probne testove za samotestiranje i pripremu.

Postavljena su sledeća istraživačka pitanja:

- Da li rešavanje probnog testa utiče na ostvarivanje rezultata odgovarajućeg testa za proveru znanja?
- Da li rezultati probnih testova mogu da ukažu na lakše i teže koncepte gradiva?

Na tačno odgovoreno pitanje prvog i drugog kolokvijuma moguće je bilo osvojiti maksimalno 1 ili 3 boda, a na završnom do 5 bodova. Ukoliko je na pitanje odgovoreno netačno ili nije označan ni jedan od ponuđenih odgovora, student je osvojio nula bodova. Na pitanja koja su imala više tačnih odgovora, bodovi za delimično tačan odgovor su se izračunavali u zavisnosti od broja tačnih odgovora.

Transformacija osvojenih bodova po pitanjima izvršena je uvođenjem četri diskretne vrednosti na sledeći način: tačan odgovor označen je oznakom *True*, netačan odgovor oznakom *False*, neodgovoreno pitanje oznakom *No_Answer*, a delimično tačan odgovor oznakom *Partially_True*.

Na testovima su studenti mogli da osvoje od nula do 20, odnosno do 30 bodova na završnom testu. Testovi nisu bili eliminatornog karaketera, student nije mogao da padne test. Statistički parametri za karakteristike koje su označavale ukupno ostvarene bodove u izdvojenim skupovima podataka dati su u tabeli 6.4.

Tabela 6.4. Statistički rezultati probnih i zvaničnih testova

Skupovi	Testovi	Min	Max	Mean	StDev
DS1	K1_P	1	20	17.36	3.84
	K1	3.12	20	14.879	4.247
DS2	K2_P	0	20	16.68	4.37
	K2	4.06	20	15.658	3.727
DS3	ZT_P	0	30	23.876	7.864
	ZT	10.1	30	25.967	4.609

Matrice rezultata izvedene su kao tri različite datoteke u *CSV* formatu (*Comma-Separated Values*) a zatim transformisane u *ASCII* tekstualne datoteke u *ARFF* formatu (*Attribute-Relation File Format*) koji opisuje listu instanci deljenih među skupom karakteristika i koristi se u Weka okruženju za otkrivanje znanja.

6.3.1.Otkrivanje pravila iz matrica rezultata

Asocijativna analiza podataka u formi matrica rezultata DS1, DS2, DS3 sprovedena je implementacijom Apriori [92] i prediktivnog Apriori algoritma [93]. Pronađena pravila su "*Ako prethodi onda i sledi*", gde su *"prethodi"* i *"sledi"* skupovi frekventnih stavki. Apriori algoritam je implementiran sa lift metrikom postavljenjem minimalne vrednosti parametra *lift* na 1.

Prediktivni Apriori algoritam implementiran je na matrice rezultata sa postavkom vrednosti za broj pravila n =100. Razmatrana su pravila sa ostvarenom prediktivnom tačnošću većom od 0.5 i minimalnim brojem stavki "prethodi" (*minant=10*) i "sledi" (*mincont=10*) u okviru kreiranog pravila. Pravila kreirana sa manjim brojem stavki "prethodi" i " sledi" nisu uzeta u razmatranje. Za otkrivena pravila određene su vrednosti mere poverenje (*Conf*) i podrške (*Sup*), lift parametar (*Lift*), tačnost (*Acc*). Pravila koja zadovoljavaju postavljena ograničenja prikazana su u tabeli 5.

Tabela 6.5. Izdvojena pravila

DS1				
Pravilo	**Conf**	**Lift**	**Sup**	**Acc.**
q_{p1}= TRUE ==> q_1=TRUE	0.63	1.2	0.53	0.62666
q_{p1}=PARTLY_TRUE ==> q_1=TRUE	0.73	1.13	0.12	0.71602
q_{p2}=TRUE ==> q_2=TRUE	0.78	1.03	0.67	0.78459
q_{p2}=PARTLY_TRUE ==> q_2=TRUE	0.79	1	0.14	0.79312
q_{p3}=TRUE==> q_3=TRUE	0.76	1.06	0.45	0.72629
q_{p4}=TRUE ==> q_4=TRUE	0.77	1.03	0.62	0.76312
q_5=TRUE ==> q_5=PARTLY_TRUE	0.52	1.04	0.43	0.52525
q_{p6}=PARTLY_TRUE ==> q_6=PARTLY_TRUE	0.59	1.35	0.12	0.57429
q_{p6}=TRUE ==> q_6=TRUE	0.6	1.15	0.49	0.59512
q_{p7}=PARTLY_TRUE ==> q_7=PARTLY_TRUE	0.52	1.47	0.13	0.52347
q_{p7}= TRUE ==> q_7= TRUE	0.65	1.11	0.52	0.62678
q_{p8}=TRUE==> q_8=TRUE	0.58	1.01	0.56	0.57114
q_{p9}=TRUE ==> q_9=TRUE	0.74	1.01	0.63	0.72758
q_{p10}=TRUE ==> q_{10}=TRUE	0.53	1.06	0.46	0.51868
DS2				
Pravilo	**Conf**	**Lift**	**Sup**	**Acc.**
q_{p1}=TRUE==> q_1= PARTLY_TRUE	0.65	1.02	0.47	0.66527
q_{p2}=TRUE==> q_2=TRUE	0.67	1.04	0.52	0.67316
q_{p3}=TRUE ==> q_3=TRUE	0.84	1.03	0.64	0.80791
q_{p4}=TRUE ==> q_4=TRUE	0.56	1.07	0.38	0.53323
q_{p5}=TRUE ==> q_5=FALSE	0.51	1.06	0.32	0.50847
q_{p6}=TRUE ==> q_6=FALSE	0.58	1.03	0.43	0.58293
q_{p7}= TRUE ==> q_7=FALSE	0.54	1.16	0.41	0.67945
q_{p8}=TRUE ==> q_8=TRUE	0.54	1.04	0.36	0.52368

q_{p9}=TRUE ==> q_9=TRUE	0.66	1.08	0.55	0.61707
q_{p10}=TRUE ==> q_{10}=TRUE	0.73	1.03	0.41	0.69656
q_{p10}=PARTLY_TRUE ==> q_{10}=TRUE	0.7	1.01	0.23	0.67945

DS3

Pravilo	Conf	Lift	Sup	Acc.
q_{p1} =PARTLY_TRUE ==> q_1=PARTLY_TRUE	0.76	1.42	0.27	0.71821
q_{p1} = TRUE ==> q_1= TRUE	0.57	1.28	0.33	0.54756
q_{p2}= PARTLY_TRUE ==> q_2 = PARTLY_TRUE	0.54	1.64	0.14	0.55264
q_{p2}=TRUE ==> q_2 =TRUE	0.78	1.18	0.52	0.77272
q_{p3} =TRUE ==> q_3 =TRUE	0.82	1.07	0.52	0.80099
q_{p4}=PARTLY_TRUE ==> q_4=PARTLY_TRUE	0.71	1.95	0.12	0.71975
q_{p4} =TRUE ==> q_4 =TRUE	0.73	1.14	0.54	0.67875
q_{p5} =TRUE ==> q_5 =TRUE	0.92	1.19	0.37	0.89596

Matrice rezultata *DS1*, *DS2*, *DS3* su transformisane redukovanjem broja karakteristika tako da su kreirani skupovi *DS11*, *DS22* i *DS33* sadržali podatke koji su se odnosili samo na zvanične testove.

Za svaki skup, generisano je 100 pravila sa prediktivnom tačnošću od 0.5 do 0.99. Izdvojena su pravila sa stavkama "sledi" iz skupa vrednosti {*Sum=VERY_GOOD, Sum=EXCELLENT}*. Za skup *DS11* izdvojeno je 23 pravila, 21 pravila za *DS22*, a 14 pravila za *DS33*. U stavkama "prethodi" uočena je različita frekventnost karakteristika koje su označavale ostvarene bodove po pitanjima, dok je većina stavki "sledi" imala vrednost *Sum=EXCELLENT*. Frekventnost karakteristika za izdvojeni skup pravila prikazana je u tabeli 6.6.

Tabela 6.6. Frekventnost karakteristika

Pitanje	K1	K2	Z
p_1	12	15	5
p_2	1	15	6
p_3	12	6	6
p_4	6	7	6
p_5	6	6	9
p_6	5	6	
p_7	11	7	
p_8	11	7	
p_9	12	8	
p_{10}	7	3	

Na osnovu rezultata prikazanih u tabeli 5 zaključeno je da postoji pozitivna korelacija između odgovora studenata na pitanja probnih i odgovarajućih zvaničnih testova. Vrednost lift parametra veća od jedan, ukazuje da najbolji pokušaj rešavanja probnog testa može da pokaže kako će student uraditi i odgovarajući zvanični test, a vrednosti očekivane prediktivne tačnosti potvrdile su prethodni zaključak. Pravila q_{p10}=*FALSE* ==> q_{10}=*PARTLY_TRUE,* q_{p6}=*FALSE* ==> q_6=*PARTLY_TRUE,* q_{p7}=*NO_ANSWER*==> q_7=*PARTLY_TRUE* pokazuju odnose između stavki koje sadrže logički nepovezane vrednosti. Na primer, ako student ne zna odgovor ili odgovori pogrešno na pitanje u probnom testu, onda će odgovoriti delimično tačno u zvaničnom testu. Ove vrste odnosa ne obezbeđuju značajne informacije za poboljšanje testova, tako da su ova pravila isključena.

Rezultati takođe ukazuju da su studenti neka pitanja lakše rešavali, a neka teže (videti tabelu 6.5). Sproveden je i subjektivni pristup za utvrđivanje značaja izdvojenih pravila koji je podrazumevao formiranje razumljivih obrazaca.

Obrazac1 pokazuje odnos između odgovora studenta na pitanja prvog probnog testa i odgovarajućeg koncepta prvog kolokvijuma.

> *IF* student odgovori tačno na pitanja iz koncepta *Interaktivna računarska grafika, Tehnologije ulaza, Tehnologije izlaza* probnog testa *THEN* student će odgovoriti tačno i na pitanje iz odgovarajućeg koncepta prvog kolokvijuma.

Nastavnik može zaključiti da koncepti koji se testiraju u prvom kolokvijalnom testu predstavljaju jednostavniji deo gradiva za studente. U skladu sa tim može izvršiti određene izmene određenih koncepta gradiva, izmeniti formu i sadržaj pitanja kako bi se ostvario viši nivo težine u testu.

Obrazac2 pokazuje odnos između odgovora studenta na pitanja drugog probnog testa i odgovarajućeg koncepta drugog kolokvijuma.

> *IF* student odgovori tačno na pitanje iz koncepta *Algoritam popunjavanja i odsecanja, Bresenhamov algoritam za liniju i krug, Fraktali* probnog testa *THEN* student će odgovoriti netačno na pitanje odgovarajućeg koncepta drugog kolokvijuma.

Nastavnik može izdvojiti teža pitanja koja mogu predstavljati diskriminatore testiranja znanja drugog kolokvijuma. Studenti koji su tačno odgovorili na ova pitanja imaju tendenciju ostvarivanja boljeg uspeha.

Obrazac3 pokazuje odnos između odgovora studenta na pitanja i zadatke trećeg probnog testa i odgovarajućeg koncepta ispitnog testa.

> *IF* student odgovori delimično tačno na pitanje iz koncepta *Matematičko modeliranje projekcija, BSP stabla* probnog testa *THEN* student će odgovoriti delimično tačno ili netačno na pitanje odgovarajućeg koncepta ispitnog testa.
> *IF* student delimično reši zadatak iz koncepta *OpenGL* probnog testa *THEN* student će rešti delimično tačno ili netačno zadatake odgovarajućeg koncepta ispitnog testa.

Na osnovu navedenih pravila, nastavnik može izdvojiti koncepte gradiva koji su nerazumljivi studentima ili neadekvatno objašnjeni u dostupnim materijalima za učenje. Takođe, nastavnik treba da proveri sadržaj gradiva u kome su objašnjavani koncepti koji su testirani na ispitu sa ciljem da ih modifikuje i poboljša.

7. KLASIFIKACIJA OBRAZOVNIH PODATAKA MEŠANOG OKRUŽENJA UČENJA

Postupak klasifikacije u obrazovnim okruženjima zasnovan je na predviđanju uspešnosti studenata na osnovu njihovog znanja, ponašanja, motivacije, aktivnosti u toku procesa učenja. Generalno, izbor efikasnog klasifikatora zavisi od domena vrednosti i problematike zadatka. U istraživanju opisanom u disertaciji, prikazan je postupak izbora kandidata klasifikatora na osnovu principa relevantnih za obrazovni skup mešanog okruženja učenja. Sprovedena je komparativna analiza kreiranih modela poređenjem vrednosti učestalosti stvarno pozitivnih (engl. *true positive*, TP), lažno pozitivnih (engl. *false positive*, FP), preciznosti i F-mere. Primena koncepta kombinovanja klasifikatora u ansambal realizovana *Vote* meta klasifikatorom metodom glasanja većine (engl. *majority vote*). Generisan je pouzdan i stabilan model predviđanja mešanog okruženja učenja u slučaju više-dimenzionalnog klasnog obeležja.

7.1 Opšti principi izbora klasifikatora

Algoritam koji realizuje klasifikaciju neviđenih instanci naziva se klasifikator. Klasifikator se može posmatrati kao model ili funkcija *M* koja predviđa oznaku klase y za datu instancu x tako da je $\hat{y} = M(x)$, x $= (x_1, x_{2,\dots,} x_d)^T \in R_d$ predstavlja instancu u d-dimenzionalnom prostoru, a $\hat{y} \in \{c_1, c_2, \dots, c_k\}$ predviđajuću klasu.

Izgradnja modela klasifikacije podrazumeva obučavanje modela skupom instanci sa poznatim vrednostima klasnih oznaka. Testiranje obučenog klasifikatora i procena performansi kreiranog modela realizuje se na test skupu sa nepoznatim vrednostima klasnih oznaka. Nakon utvrđivanja pouzdanosti, model se može implementirati za klasifikaciju instanci predviđanjem oznaka klasa. Razmatrani su principi koji utiču na izbor klasifikatora i ostvarena tačnost klasifikacije. Definisani su diskriminativni i probabilistički klasifikatori, metode procene performansi klasifikacije, pojmovi prenaučenosti (engl. *overfitting)* i nedovoljne naučenosti (engl. *underfitting*).

Diskriminativnim klasifikatorima određuje se samo jedna klasna vrednost (oznaka) za svaku instancu skupa podataka. Neka je *M* klasifikator, *C* klasa sa klasnim oznakama $C=\{c_1, \ldots, c_k\}$ i *t* instanca u skupu podataka. Predviđena klasna oznaka instance *t* će biti $M(t) = c_i$ za samo jedno *i*.

Probabilistički klasifikator definiše verovatnoću klase za sve klasifikovane redove tako da je $M(t) = [P(C = c_1|t), \ldots, P(C = c_k|t)]$, gde je $P(C = c_i|t)$ verovatnoća da instanca *t* pripada klasi c_i.

Prenaučenost (engl. *overfitting*) predstavlja problem koji je vezan za meru koja označava tačnost. Ukazuje na suviše naučen model koji se ne može generalizovati na buduće podatke. Prenaučen model prikazuje specijalne slučajeve i greške u podacima. Prenaučenost se javlja kao posledica suviše složenog formiranog modela u odnosu na veličinu skupa podataka. Složeni modeli imaju veću reprezentativnu snagu i mogu predstavljati sve karakteristike podataka, uključujući greške. Sa druge strane, jednostavni modeli imaju manju reprezentativnu snagu, ali mogu da generalizuju buduće podatke. U slučaju suviše jednostavnog modela, nije moguće pronaći značajne obrasce u podacima što se javlja kao posledica nedovoljne naučenosti (engl. *underfitting)*. To znači da je model označen kao loš i slab ili ne postoji pravi model.

U obrazovnim okruženjima, prenaučenost predstavlja kritičan problem. Obrazovni skup podataka može da sadrži veliki broj obeležja za formiranje kompleksnih modela, ali mali broj podataka za obučavanje. Na osnovu pravila navedenog u literaturi [154,155], za svako obeležje skupa trebalo bi imati od 5 do 10 redova (instanci) podataka. Što je model jednostavniji, to znači da je potreban manji broj obeležja. Za skup sa *k* binarnih obeležja, Näive Bayes klasifikator sadrži *O(k)* parametara. To znači da ulazni skup treba da sadrži $n > 5k$ instanci podataka. Za slučaj obeležja koji nisu binarnog tipa, potrebno je više podataka.

Problem prenaučenosti moguće je izbeći primenom jednostavnijih klasifikatora koji zahtevaju manje parametara za formiranje modela i redukcijom dimenzionalnosti obučavajućeg skupa uvrđivanjem optimalnog vektora obeležja.

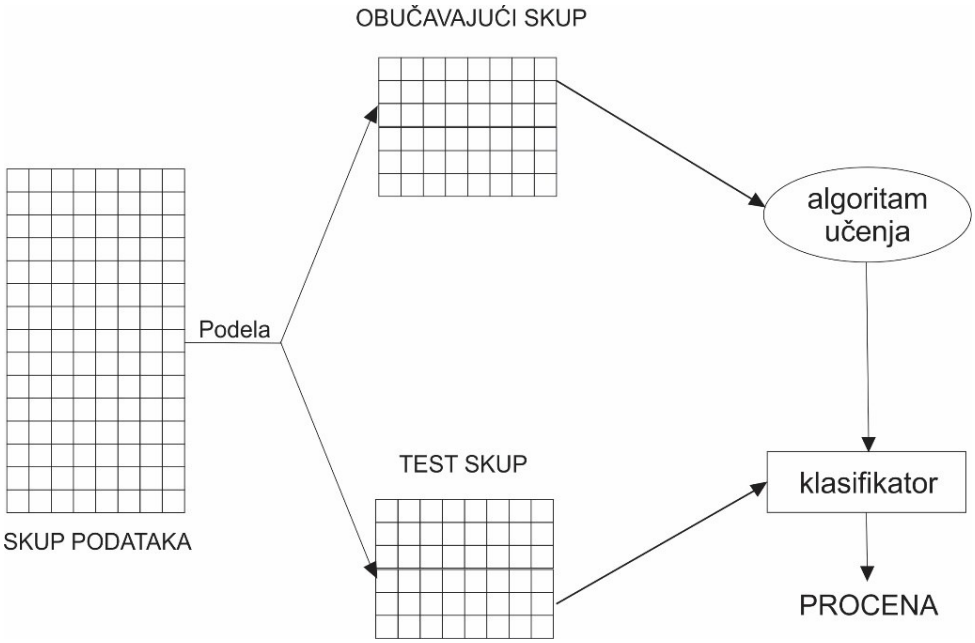
7.1.1. Metode za procenu modela klasifikacije

Najvažniji kriterijum za procenu performansi klasifikatora je tačnost predviđanja, tj. procenat tačno klasifikovanih nepoznatih instanci. Preciznost implementiranog klasifikatora *M* utvrđuje se primenom određene mere učinka θ .

Ulazni skup podataka *D* je nasumično podeljen u skup za obučavanje i skup za testiranje. Skup obuke se koristi za učenje modela *M*, a skup testiranja se koristi za procenu mere θ. Ukoliko bi se podela skupa D realizovala nasumično, slučajnim razdvajanjem, može se formirati isuviše jednostavan ili kompleksan test skup što bi ukazalo na suviše dobre ili loše performanse klasifikatora.

Dobra strategija za procenu performansi klasifikatora zasnovana je na karakteristikama ulaznog skupa podataka i kombinovanju primene različitih metoda estimacije.

Metod test skupa (slika 7.1) podrazumeva podelu ulaznog skupa na dva međusobno nezavisna podskupa – obučavajući i test skup. Uobičajna podela ulaznog skupa u dva podskupa moguća je u proporcijama 1:1, 2:1, 70%:30% ili 60%:40%.



Slika 7.1: Metod test skupa

Obučavajući skup sa poznatim vrednostima klasnog obeležja se koristi za obučavanje primenjenog klasifikatora. Obučen klasifikator se zatim koristi za predviđanje klasa instanci

test skupa gde su vrednosti klasnog obeležja nepoznate. Ako test skup sadrži *N* instanci od kojih je *C* pravilno klasifikovano, tačnost predviđanja klasifikatora se određuje kao p= $C/_N$. Važno je imati na umu da opšti cilj nije samo klasifikacija instanci u test skupu veći i procena tačnosti predviđanja svih mogućih instanci gde su vrednosti klasnog obeležja nepoznate, a koje obično predstavljaju sve instance test skupa. Ako je *p* predviđena tačnost izračunata za test skup i klasifikator se koristi za klasifikaciju instanci u drugom test skupu, vrlo je verovatno da će se dobiti drugačija vrednost za prediktivnu tačnost. Može se reći da je *p* procena stvarne prediktivne tačnosti klasifikatora za sve moguće nepozanate instance. Za pronalazak niza vrednosti unutar kojih leži prava vrednost prediktivne tačnosti, sa datom verovatnoćom ili "stepenom pouzdanosti" određuje se standardna greška povezana sa procenjenom vrednošću predviđene tačnosti. Ako je p izračunata verovatnoća za test skup od N instanci, vrednost njegove standardne greške je $\sqrt{p(1-p)/N}$.

Značaj standardne greške se odnosi na mogućnost da je verovatnoća istinske prediktivne tačnosti klasifikatora u okviru standardnih grešaka iznad ili ispod procenjene vrednosti. Verovatnoća istinske prediktivne tačnosti označava nivo pouzdanosti klasifikatora (engl. *confidence level*).

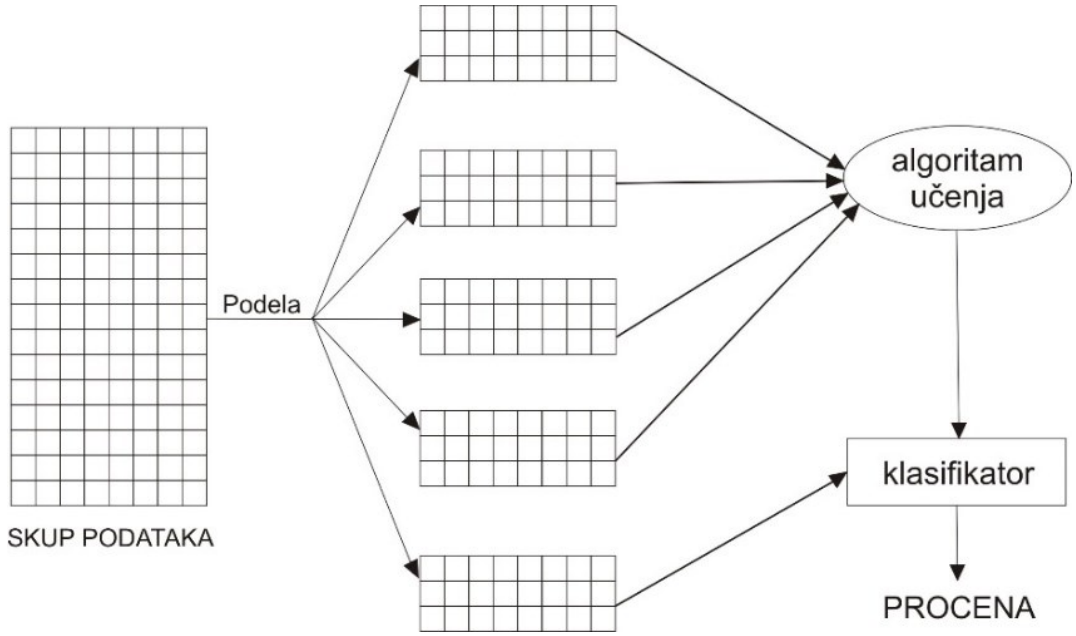
Metod unakrsnog ocenjivanja (engl. *cross-validation*) poznat je i kao rotaciona estimacija. Podrazumeva podelu ulaznog skupa *D* u *K* jednakih delova (engl. *folds*) imenovanih sa D_1 , D_2, … , D_k. Deo D_i se tretira kao test skup tako da sa preostalima obrazuje obučavajući skup, odnosno važi $D \setminus D_i = \bigcup_{j \neq i} D_j$. Nakon obučavanja M_i na $D \setminus D_i$, procena performansi klasifikatora M_i se realizuje na test skupu D_i kako bi se dobila *i-ta* procena θ_i. Vrednost merenja performansi klasifikatora definiše se kao aritmetička sredina očekivanih vrednosti svih procena.

$$\hat{\mu}_\theta = E[\theta] = \frac{1}{K}\sum_{i=1}^{K} \theta_i \tag{7.1}$$

a njegova varijansa kao

$$\hat{\sigma}_\theta^2 = \frac{1}{K}\sum_{i=1}^{K} (\theta_i - \hat{\mu}_\theta)^2 \tag{7.2}$$

Na slici 7.2 prikazan je postupak procene klasifikatora primenom metode unakrsnog ocenjivanja.



Slika 7.2: Metod unakrsnog ocenjivanja

Algoritam metode unakrsnog ocenjivanja može se prikazati sledećim pseudo-kodom datim u nastavku.

Algoritam K-fold cross validation

1. **Random shuffle D**
   Nasumično pomeranje skupa D
2. $\{\boldsymbol{D_1}, \boldsymbol{D_2}, \dots, \boldsymbol{D_K}\}$:
   Podela skupa D u *K* jednakih particija izuzev, eventualno, poslednje**.**
   (3-5) Svaki podskup $\boldsymbol{D_i}$ se koristi kao test skup za procenu performansi $\boldsymbol{\theta_i}$ klasifikatora $\boldsymbol{M_i}$ na skupu $\boldsymbol{D} \setminus \boldsymbol{D_i}$.
3. **For each** $\boldsymbol{i} \in [\boldsymbol{1}, \boldsymbol{K}]$ **do**
4. $\boldsymbol{M_i}$ - obučavajući klasifikator za $\boldsymbol{D} \setminus \boldsymbol{D_i}$
5. $\boldsymbol{\theta_i}$ – procena $\boldsymbol{M_i}$ na $\boldsymbol{D_i}$
   (6-7) Izračunava se vrednost aritmetičke sredine i varijanse procene performansi
6. $\widehat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} = \frac{\boldsymbol{1}}{\boldsymbol{K}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{K}} \boldsymbol{\theta_i}$
7. $\widehat{\boldsymbol{\sigma}}_{\boldsymbol{\theta}}^{\boldsymbol{2}} = \frac{\boldsymbol{1}}{\boldsymbol{K}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{K}} (\boldsymbol{\theta_i} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{\theta}})^{\boldsymbol{2}}$

8. return $\widehat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}, \widehat{\boldsymbol{\sigma}}_{\boldsymbol{\theta}}^{\boldsymbol{2}}$

Unakrsna validacija se ponavlja više puta, pri čemu inicijalno nasumično pomeranje skupa *D* iz Koraka 1 obezbeđuje da se delovi razlikuju svaki put.

Uobičajna vrednost *K* je 5 ili 10. Specijalan slučaj *K=n* označava izostavi - jedan unakrsnu validaciju (engl. *leave-one-out cross validation*), kada test skup sadrži jednu instancu, a ostali podaci se koriste za obučavanje klasifikatora.

Treći pristup za procenu očekivanih performansi klasifikatora predstavlja metod ponovnog uzimanja uzoraka (engl. *bootstrap resampling*). Umesto podele skupa *D* u *K* međusobno različitih particija, ovaj metod je zasnovan na izdvajanju *K* slučajnih uzoraka veličine *n* koji zamenjuju skup *D*. Svaki uzorak D_i je iste veličine kao D i ima nekoliko ponovljenih instanci. Uzorci se koriste za obučavanje klasifikatora, a testiranje se vrši na kompletnom skupu podataka *D*. Primenom ovog pristupa estimacije klasifikatora treba imati na umu da će rezultati biti poprilično optimistični s obzirom na postojanje preklapanja u instancama između obučavajućeg i test skupa.

7.1.1. Mere performansi klasifikacije

Neka je sa D označen test skup koji sadrži *n* instanci definisanih u *d* - dimenzionalnom prostoru; $\{c_1, c_2, \dots, c_k\}$ označava klasno obeležje sa k klasnih oznaka, a *M* klasifikator. Za instancu $x_i \in D$, y_i označava tačnu klasu, a $\hat{y}_i = M(x_i)$ predviđajuću klasu.

Stopa greške (engl. *error rate*) predstavlja osnov za procenu kreiranih modela. Stopa greške označava procenat pogrešno klasifikovanih instanci i ukazuje na kreiran loš i nestabilan model i definiše se jednačinom (7.3).

$$\textit{Error rate} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i) \tag{7.3}$$

I označava indikator funkcije i ima vrednost 1 kada je argument tačan, a 0 u suprotnom. Stopa greške predstavlja procenat pogrešne klasifikacije, što znači da što je ova vrednost manja što je klasifikator precizniji.

Tačnost klasifikacije (engl. *accuracy*) ukazuje na deo pravilno klasifikovanih novih instanci korišćenjem obučenog klasifikatora na test skupu i definiše se jednačinom (7.4).

$$Accuracy=\frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i) = 1\text{- } Error\ rate \tag{7.4}$$

Tačnost klasifikacije smatra se osnovnim pokazateljem performansi kreiranog klasifikacionog modela. Ukazuje na procenu verovatnoće tačnog predviđanja – što je veća tačnost, to je bolji klasifikator. S obzirom da eksplicitno ne razmatraju klasne oznake koje doprinose grešci, stopa greške i tačnost predstavljaju globalne mere. Informativnije mere o slaganju i neslaganju stvarnih i predviđajućih klasnih oznaka nad test skupom mogu se predstaviti tabelarnim prikazom.

Neka $D = \{D_1, D_2, \dots, D_k\}$ označava podelu test skupa na osnovu tačnih klasnih oznaka tako da je $D_j = \{x_i \in D | y_i = c_j\}$ i $n_i = |D_i|$ veličina stvarne klase c_i . Neka $R = \{R_1, R_2, \dots, R_k\}$ označava podelu test skupa na osnovu predviđajućih klasnih oznaka, tako da je $R_j = \{x_i \in D | \hat{y}_i = c_j\}$ i $m_j = |R_j|$ veličina predviđajuće klase c_j. Skupovi R i D indukuju matricu grešaka (engl. *confusion matrix*) definisanu sa:

$$N(i,j) = n_{ij} = |R_i \cap D_j| = |\{x_a \in D | \hat{y}_a = c_i \, i \, y_a = c_j\}|, 1 \le i,j \le k. \tag{7.5}$$

Vrednost n_{ij} označava broj instanci sa predviđajućom klasom c_i čija je prava oznaka c_j , a vrednost n_{ii} $(1 \le i \le k)$ broj instanci za koje se rezultat klasifikatora slaže sa stvarnom klasnom oznakom c_i . Preostale vrednosti n_{ij} $(i \ne j)$ predstavljaju instance za koje se rezultati klasifikatora i stvarne klasne oznake ne slažu. Matrica grešaka prikazuje različite tipove grešaka klasifikacije, pri čemu redovi matrice sadrže broj predviđajućih instanci, a kolone broj stvarnih instanci za odgovarajuću klasnu oznaku. Broj pravilno klasifikovanih instanci predstavlja zbir instanci po dijagonali matrice. Ostali elementi matrice označavaju pogrešno klasifikovane instance, odnosno, instance sa neodgovarajućim klasnim oznakama. U slučaju klase sa više klasnih oznaka, formira se niz matrica sa dve klasne oznake tako što se za svaki član niza izdvaja po jedna ciljna oznaka.

Opoziv (engl. *recall)* i preciznost (engl. *precision*) predstavljaju drugi par mera za evaluaciju klasifikacijskih modela. Preciznost klasifikatora *M* za specifičnu klasu c_i prikazana je jednačinom (7.6) kao deo tačnih predviđanja svih instanci tačkama koje su predviđene da budu u klasi c_i .

$$precc_i = \frac{n_{ii}}{m_i} \tag{7.6}$$

gde je m_i broj instanci klasifikovanih sa previđajućom klasnom oznakom c_i.

Ukupna preciznost klasifikatora može se predstaviti jednačinom (7.7) kao procenjen prosek specifičnih preciznosti klasa :

$$precc = \sum_{i=1}^{k} \left(\frac{m_i}{n}\right) precc_i = \frac{1}{n}\sum_{i=1}^{k} n_{ii} \quad \text{..... (7.7)}$$

Opoziv klasifikatora *M* za klasu c_i je deo ispravnih predviđanja svih instanci u klasi c_i .

$$recall_i = \frac{n_{ii}}{n_i} \tag{7.8}$$

gde je n_i broj instanci u klasi c_i .

Stabilan model podrazumeva postojanje kompromisa između preciznosti i opoziva klasifikatora. F-mera (engl. *F-measure*) pokazuje balans između vrednosti preciznosti i opoziva i definiše se kao harmonijska sredina klase c_i. Veća vrednost ove mere ukazuje na bolji klasifikator.

$$F_i = \frac{2}{\frac{1}{precc_i}+\frac{1}{recall_i}} = \frac{2 \cdot precc_i \cdot recall_i}{precc_i + recall_i} \tag{7.9}$$

Ukupna vrednost F-mere za klasifikator M predstavlja srednju vrednost mera odgovarajućih klasa.

$$F = \frac{1}{k}\sum_{i=1}^{r} F_i \tag{7.10}$$

7.1.1.1. Binarna klasifikacija: pozitivna i negativna klasa

U slučaju kada klasno obeležje ima samo dve oznake, klasa $c_1$označava pozitivnu, a klasa c_2 negativnu klasu. Osnovni tip matrice grešaka predstavlja slučaj klase sa dve klasne oznake *(k=2)*. Dvodimenzionalna matrica grešaka prikazana tabelom 7.1 sadrži ćelije sa oznakama *TP, FP, FN, TN, P, N*.

Tabela 7.1. Dvodimenzionalna matrica grešaka

Predviđajuća klasa	Stvarna klasa		Ukupno instanci
	Pozitivna (c_1)	Negativna (c_2)	
Pozitivna (c_1)	$TP = n_{11}$	$FP = n_{12}$	m_1
Negativna (c_2)	$FN = n_{21}$	$TN = n_{22}$	m_2
	n_1	n_2	

TP (engl. *True Positive rate*) predstavlja broj instanci koje klasifikator pravilno klasifikuje kao pozitivne.

$TP = n_{11}$=$|\{x_i|\hat{y}_i = y_i = c_1\}|$

FP (engl. *False Positive rate*) ukazuje na broj instanci koje klasifikator predviđa kao pozitivne, a zapravo treba da budu klasifikovane kao negativne.

$FP = n_{12}$=$|\{x_i|\hat{y}_i = c_1\ i\ y_i = c_2\}|$

FN (engl. *False Negative rate*) predstavlja broj instanci koje klasifikator predviđa da budu klasifikovane u negativnoj klasi, a zapravo pripadaju pozitivnoj klasi.

$FN = n_{21}$=$|\{x_i|\hat{y}_i = c_2\ i\ y_i = c_1\}|$

TN (engl. *True Negative rate*) ukazuje na broj instanci pravilno klasifikovane kao negativne.

$TN = n_{22}$=$|\{x_i|\hat{y}_i = y_i = c_2\}|$

Vrednost m_1 predstavlja ukupan broj instanci klasifikovanih kao pozitivne m_1=*TP+FP*, a m_2 ukupan broj instanci klasifikovanih kao negativne m_2=*FN+TN*.

Korisno je razlikovati dva tipa greški klasifikacije: grešku tipa 1 i grešku tipa 2. Greška tipa 1 (*FP*) označava lažno pozitivne instance. Javlja se u slučaju kada su negativne instance

klasifikovane kao pozitivne. Greška tipa 2 (*FN*) označava lažno negativne instance. Javlja se kada su primeri pozitivnih instanci klasifikovani kao negativni.

Stopa greške za slučaj binarne klasifikacije definisana je kao deo grešaka (ili lažnih predviđanja):

$$Error\ rate = \frac{FP+FN}{n} \tag{7.11}$$

Tačnost za slučaj binarne klasifikacije definisana je jednačinom (7.12) kao udeo tačnih predviđanja:

$$Accuracy = \frac{TP+TN}{n} \tag{7.12}$$

Preciznost za pozitivnu i negativnu klasu data je jednačinama (7.13) i (7.14):

$$precc_P = \frac{TP}{TP+FP} = \frac{TP}{m_1} \tag{7.13}$$

$$precc_N = \frac{TN}{TN+FN} = \frac{TN}{m_2} \tag{7.14}$$

gde je m_i broj instanci koje klasifikator M klasifikuje u klasu c_i.

Učestalost stvarno pozitivnih, tj. osetljivost (engl. *sensitivity*), označava udeo tačnih predviđanja u odnosu na sve instance u pozitivnoj klasi.

$$TPR = recall_P = \frac{TP}{TP+FN} = \frac{TP}{n_1} \tag{7.15}$$

gde je $n_1$veličina pozitivne klase.

Učestalost lažno negativnih, tj. specifičnost (engl. *specificity*), predstavlja opoziv negativne klase

$$TNR = recall_N = \frac{TN}{FP+TN} = \frac{TN}{n_2} \tag{7.16}$$

gde je $n_2$veličina negativne klase.

7.2 Primena klasifikatora na obrazovni skup mešanog okruženja učenja

U doktorskoj disertaciji prikazana je studija slučaja primene klasifikatora na mešano okruženje učenja. Sprovedeno istraživanje fokusirano je na proces detaljne klasifikacije ulaznog skupa koji je kreiran integrisanjem podataka izdvojenih iz različitih okruženja učenja. Generisan klasifikatorski model omogućio je predviđanje konačne ocene studenata na osnovu realizovanih aktivnosti u okviru Moodle kursa i klasične nastave. Obeležje Ocena je definisano kao izlazna predviđajuća varijabla sa sedam klasnih oznaka koje su ukazale na nelinearnost granica i više-dimenzionalnost domena.

Prilikom izbora kandidata izvršeno je metodološko izdvajanje klasifikatora podesnih za rad sa malim uzorcima skupova i kategorijskim tipom podataka, neosetljivih na više-dimenzionalnost klasnog obeležja, postojanje izuzetaka i nedostajućih vrednosti. Ulazni skup podataka sadržao je 276 instanci.

U fazi predprocesiranja izvršena je detaljna priprema i obrada nepotpunih podataka i izuzetaka. Transformacija numeričkih u kategorijalan tip podataka realizovana je na osnovu sprovedene deskriptivne statističke analize sirovih podataka.

Odluka o izboru klasifikatora zasnovana je na karakteristikama mešanog okruženja učenja, više-dimenzionalnom klasnom obeležju i kategorijalnom tipu podataka. Pored navedenih principa odlučivanja, razmatran je odnos veličine obučavajućeg skupa i broja ulaznih obeležja koji može da dovede do pojave prenaučenosti modela. Dimenzionalnost klasnog obeležja uslovila je isključenje klasifikatora linearne regresije, višestruke linearne regresije i metode vektora oslonca. Usmerenost ka kreiranju što jednostavnijih modela isključila je upotrebu veštačkih neuronskih mreža.

Na osnovu prethodno navedenih principa i analiziranjem referenci navedenih u pregledu literature, zaključeno je da su bajesovski modeli i stabla odlučivanja tehnike najpodesnije za primenu u opisanoj studiji slučaja. Testirani su bajesovski algoritmi Naïve Bayes (*NB*), Hidden Naïve Bayes (*HNB*) i tehnike stabla odlučivanja *J48* i Random Forest (*RF*) metodom estimacije test skupa. Ulazni skup je podeljen na osnovu postavljenog koeficijenta podele (engl. *split ratio*) na vrednost 0.66. Kreirana su dva podskupa, obučavajući skup koji je obuhvatao 182 instance i test skup sa 94 instance podataka. Provera klasifikatora ostvarena

je na različitim test skupovima primenom deset ciklusa nasumične podele tako da svaki sledeći put obučavajući i test skup sadrže različite kombinacije instanci. Izvršena je analiza matrice grešaka (engl. *confusion matrix*) i poređenje vrednosti: preciznost, opoziv i F-mera. Klasne oznake obeležja *Ocena* označene su slovima *A= nije_polagao, B=pao, C=sest, D=sedam, E=osam, F=devet, G=deset*.

7.2.1. Model Näive Bayes

U prvom eksperimentu primenjen je Näive Bayes (NB) klasifikator slabijih varijansi, pogodan za mali obučavajući skup. Pretpostavka o uslovnoj nezavisnosti obeležja znatno pojednostavljuje proračune potrebne za kreiranje NB modela.

Kreiran je model sa izlaznim čvorom Ocena. Verovatnoća predviđanja zasnovana je na ulaznim čvorovima koji su označavali realizovane aktivnosti. Koncept mešanog okruženja učenja podrazumevao je postojanje interakcija između određenih resursa. Studenti koji su radili prvi i drugi domaći zadatak na Moodle kursu ostvarili su bolje rezultate na prvom kolokvijumu.

Kreiran NB model predviđanja ostvario je tačnosti od 71,28% sa greškom klasifikacije od 28,72%. Formirana je višedimenzionalna matrica grešaka. Elementi na dijagonali ukazivali su na broj pravilno klasifikovanih instanci, kada su stvarna i predviđajuća klasa jednake. Ostali elementi matrice označavali su instance klasifikovane u pogrešne predviđajuće klase, za slučaj kada se stvarna i predviđajuća klasa razlikuju. Višedimenzionalna matrica grešaka data je u tabeli 7.2a. Performanse *NB* modela prikazane su vrednostima *True Positive Rate* označeno sa TP, *False Positive Rate* označeno sa FP, preciznošću označeno sa P, opozivom sa R i F-merom označenom sa FM (tabela 7.2b).

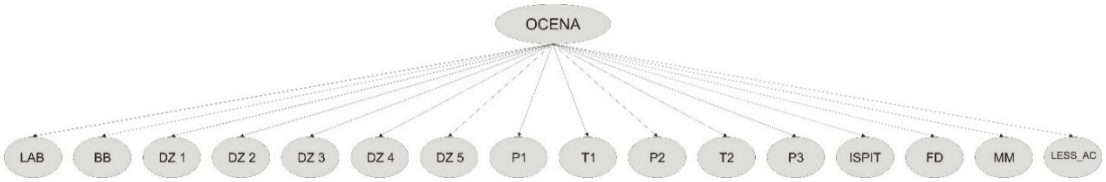
Tabela 7.2a. Matrica grešaka NB modela

Predviđajuće klase							**Stvarne klase**
A	B	C	D	E	F	G	
12	0	1	0	1	0	0	A
0	7	0	0	0	0	0	B
0	1	13	6	4	0	0	C
0	0	2	7	1	0	0	D
1	0	0	1	8	6	0	E
0	0	0	0	0	9	1	F
0	0	0	0	0	2	11	G

Tabela 7.2b. Performanse NB modela

Stvarne klase	TP	FP	P	R	FM
A	0,86	0,01	0,92	0,86	0,89
B	1,00	0,01	0,86	1,00	0,93
C	0,54	0,04	0,81	0,54	0,65
D	0,70	0,08	0,50	0,70	0,58
E	0,50	0,08	0,57	0,50	0,53
F	0,90	0,09	0,53	0,90	0,67
G	0,85	0,01	0,92	0,85	0,89
NB model	0,71	0,05	0,74	0,71	0,71

Analizirane su greške klasifikacije ostvarene po klasama ponaosob. Najveći broj pogrešno klasifikovanih instanci uočen je u slučaju klase C. Jedna instanca stvarne klase C pogrešno je klasifikovana u predviđajuću klasu B, šest u klasu D, a četri u klasu E. Najbolja vrednost opoziva klase B (R=1,00) ukazuje na najmanju grešku klasifikacije, sve stvarne instance su klasifikovane u odgovarajuću predviđajuću klasu. Najveća preciznost P=0,92 ostvarena je za klase A i G, a najmanja P=0,50 u slučaju klase D. Uočena neusaglašenost između greške klasifikacije, preciznosti i opoziva po klasama ukazivala je na postojanje disbalansa u podacima. Sa ciljem ostvarivanja bolje stabilnosti NB modela predviđanja sa karakteristikom više-dimenzionalnog klasnog obeležja izračunate su pojedinačne i ukupna vrednost F-mere. Balansiranjem performansi generisan je NB model najveće tačnosti od 71%. Struktura NB model mešanog okruženja učenja prikazana je na slici 7.3.



Slika 7.3: Struktura NB modela

7.2.2. Model Hideen Näive Bayes

Poboljšanje NB modela podrazumevalo je izračunavanje korelacija između ulaznih obeležja. Implementiran je *Hidden Naïve Bayes (HNB)* klasifikator i definisane težine skrivenih čvorova roditelja svakog obeležja. Na taj način, prilikom formiranja modela uzeto je u obzir postojanje međusobnih uticaja obeležja. Utvrđivanje korelacija izvršeno je primenom metrike određivanja zajedničke informacije grešaka. Mera zajedničke međusobne informacije *MI(X,Y)* predstavlja informaciono-teorijsku meru statističkih zavisnosti između X i Y. Izračunavanje se svodi na utvrđivanje količine informacija koju dele promenljive. Ako su međusobno nezavisne, onda promenljiva X ne sadrži informacije o Y i obrnuto pa je njihova

zajednička informacija nula. U drugoj krajnosti, kada su X i Y identični, onda dele i sve informacije Y [152].

Mera međusobne informacije ukazuje na količinu neizvesnosti u X nakon poznavanja Y. Na taj način se potvrđuje značenje međusobne informacije u smislu količine informacija koje jedna promenljiva obezbeđuje o ostalim promenljivima. Međusobna informacija se definiše jednačinom (7.17)

$$MI(X,Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{7.17}$$

gde p(x,y) predstavlja zajedničku funkciju raspodele verovatnoće X i Y, a p(x) i p(y) marginalne funkcije distribucije verovatnoće za X i Y.

Za obučavajući skup opisanog istraživanja, težine skrivenih čvorova roditelja svakog obeležja utvrđene su direktno iz podataka. Izračunata je mera zajedničkih informacija za sve parove ulaznih obeležja. Na osnovu jednačine (7.18), težina skrivenog čvora roditelja svakog obeležja, hp_i, predstavljena je kao suma mere zajedničke informacije između posmatranog i ostalih obeležja.

$$hp_i = \sum_{j=1}^{15} I(O_i, O_j),\ i \neq j \tag{7.18}$$

HNB klasifikator je generisao model tačnosti predviđanja od 76,59% sa greškom klasifikacije od 23,40%. Formirana je višedimenzionalna matrica grešaka i izračunate su performanse *HNB* modela (videti tabele 7.3a i 7.3b).

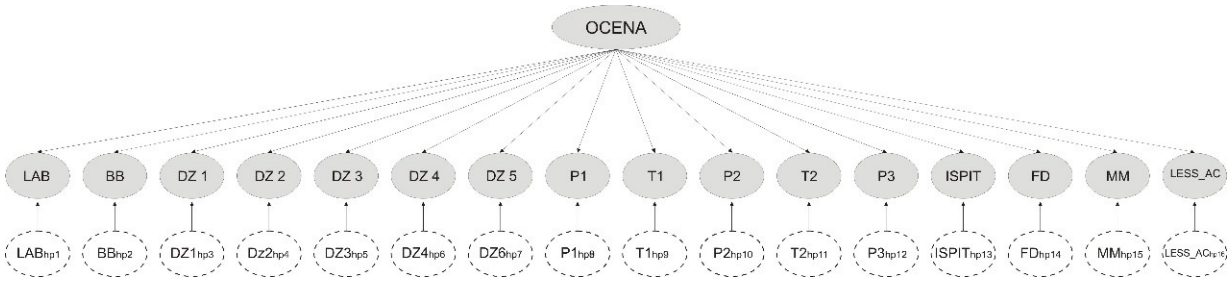
Tabela 7.3a. Matrica grešaka HNB modela

Predvidjaće klase							**Stvarne klase**
A	B	C	D	E	F	G	
13	0	1	0	1	0	0	A
0	6	1	0	0	0	0	B
0	2	17	7	0	0	1	C
0	0	3	7	0	0	0	D
1	0	0	1	9	5	0	E
0	0	0	0	1	9	0	F
0	0	0	0	0	2	11	G

Tabela 7.3b. Performanse HNB modela

Stvarne klase	TP	FP	P	R	FM
A	0,93	0,01	0,93	0,93	0,93
B	0,86	0,02	0,75	0,86	0,80
C	0,71	0,07	0,77	0,71	0,74
D	0,70	0,06	0,58	0,70	0,64
E	0,56	0,01	0,90	0,56	0,69
F	0,90	0,08	0,56	0,90	0,69
G	0,85	0,01	0,92	0,85	0,88
HNB model	0,77	0,04	0,79	0,77	0,77

Najveći broj pogrešno klasifikovanih instanci zabeležen je kao i kod *NB* modela u slučaju klase C. Dve instance stvarne klase pogrešno su klasifikovane u predviđajuću klasu B, a sedam u klasu D. Najbolje vrednosti preciznosti i opoziva (P=R=0,93) uočene su kod klase A. Najmanja preciznost P=0,56 ostvarena je za klasu F, a opoziv R=0,56 u slučaju klase E. Balansiranje performansi, kompromis preciznosti i opoziva je postignuto F-merom od 77%. *HNB* klasifikator je ostvario poboljšanje modela predviđanja mešanog okruženja učenja u odnosu na *NB* klasifikator primenjen u prethodnom eksperimentu. Proširena struktura *HNB* modela prikazana je na slici 7.4.



Slika 7.4: Struktura HNB modela

7.2.3. Model stabla odluke – J48

J48 model potkresanog stabla odlučivanja ostvario je tačnost od 73,40% sa greškom klasifikacije od 26,59%. Grananje stabla zasnovano je na određivanju značaja obeležja kriterijumom odnosa dobiti (engl. *Gain Ratio*). Obeležje *ISPIT* sa najvećom vrednošću GRatio=0,68 definisano je kao koreni čvor. Testiranjem obeležja u čvorovima formirano je stablo sa 68 pravila i 52 lista.

Matrica grešaka i performanse kreiranog J48 modela date su u tabelama 7.4a i 7.4b respektivno.

Tabela 7.4a. Matrica grešaka J48 modela

Predvidjaće klase							Stvarne klase
A	B	C	D	E	F	G	
14	0	0	0	0	0	0	A
1	6	0	0	0	0	0	B
0	2	17	5	0	0	0	C
0	1	1	7	1	0	0	D
1	0	1	3	5	7	0	E
0	0	0	0	1	8	1	F
0	0	0	0	0	1	12	G

Tabela 7.4b. Performanse J48 modela

Stvarne klase	TP	FP	P	R	FM
A	1,00	0,01	0,93	1,00	0,96
B	0,86	0,03	0,67	0,86	0,75
C	0,71	0,03	0,89	0,71	0,79
D	0,70	0,09	0,47	0,70	0,56
E	0,31	0,03	0,71	0,31	0,44
F	0,80	0,09	0,50	0,80	0,62
G	0,92	0,01	0,92	0,92	0,92
J48 model	0,73	0,04	0,77	0,73	0,73

Najveća greška klasifikacija uočena je u slučaju klase E na šta je ukazala najmanja vrednost opoziva R= 0,31. Najveća vrednost opoziva R=1,00 u slučaju klase A ukazuje na najmanju grešku klasifikacije. Najbolja balansiranost između mere opoziva i preciznosti postignuta je u slučaju klasa A i G na šta ukazuju veće vrednosti F-mere.

J48 klasifikator kreirao je model sa F-merom od 73%, preciznošću od 77% i opozivom od 73%. U poređenju sa prethodnim eksperimentima, može se uočiti da je klasifikator stabla odlučivanja – *J48* kreirao bolji model od *NB,* a lošiji u odnosu na *HNB* klasifikator.

7.2.4. Model Random Forest

U svakom čvoru su izabrana obeležja pretraživanja kako bi se realizovala najbolja podela. Važnost obeležja zasnovana je na smanjivanju prosečene nečistoće obeležja u odnosu na pripadajuću klasu. Izračunate su vrednosti značaja obeležja primenom kriterijuma *Gini Index*. Kao i u slučaju *J48* klasifikatora, obeležje *ISPIT* je označeno kao najznačajnije. *Random Forest* (*RF*) klasifikator je kreirao model stabla zasnovan na kolekciji od 100 stabala i je ostvario tačnost od 81,91% sa greškom klasifikacije od 18,09%. Matrica grešaka prikazana je u tabeli 7.5a, a performanse *RF* modela u tabeli 7.5b.

Tabela 7.5a. Matrica grešaka RF modela

Predvidjaće klase							Stvarne klase
A	B	C	D	E	F	G	
18	0	0	0	0	0	0	A
0	7	0	0	0	0	0	B
0	1	20	0	2	0	0	C
0	0	2	4	0	1	0	D
1	0	0	1	6	3	0	E
0	0	0	0	0	10	5	F
0	0	0	0	0	1	12	G

Tabela 7.5b. Performanse RF modela

Stvarne klase	**TP**	**FP**	**P**	**R**	**FM**
A	1,00	0,01	0,95	1,00	0,97
B	1,00	0,01	0,88	1,00	0,93
C	0,87	0,03	0,91	0,87	0,89
D	0,57	0,01	0,80	0,57	0,67
E	0,55	0,02	0,75	0,55	0,63
F	0,67	0,06	0,67	0,67	0,67
G	0,92	0,06	0,70	0,92	0,80
RF model	0,82	0,03	0,82	0,82	0,81

Najveća greška klasifikacije uočena je u slučaju klase E, na šta je ukazala najmanja vrednost opoziva R=0,55. Najveća vrednost opoziva R=1,00 u slučaju klase A i B ukazuje na najmanju grešku klasifikacije. Najbolja balansiranost između mere opoziva i preciznosti postignuta je u slučaju klase A na šta ukazuju najveća vrednost F-mere. *RF* klasifikator kreirao je balansiran model predviđanja sa ostvarenom tačnošću od 81%.

7.2.5. Uporedna analiza klasifikatora

Uporedna analiza klasifikacionih modela kreiranih u prethodno opisanim eksperimentima je zasnovana na poređenju ostvarene preciznosti, opoziva i F-mere (videti tabelu 7.6).

Tabela 7.6. Uporedna analiza formiranih modela

	P	**R**	**FM**
Naive Bayes - NB	0,74	0,71	0,71
HNB	0,79	0,77	0,77
J48	0,77	0,73	0,73
RANDOM FOREST	0,82	0,82	0,81

Na osnovu prikazanih vrednosti u tabeli 7.6, može se zaključiti da *RF* klasifikator stabla odlučivanja formira model predviđanja najboljih performansi. Najlošiji model studije slučaja mešanog okruženja učenja kreiran je sa *NB* klasifikatorom.

S obzirom da je klasno obeležje ulaznog skupa podataka više-dimenzionalno sa definisanih sedam klasnih oznaka, neophodno je bilo izvršiti detaljniju analizu. U tabeli 7.7 prikazano je poređenje mere preciznosti, opoziva i F-mere pojedinačno za svaku klasu formiranih modela.

Tabela 7.7. Uporedna analiza performansi P, R, FM formiranih modela po klasama

	Naïve Bayes			**Hidden Naïve Bayes**			**J48**			**Random Forest**		
	P	**R**	**FM**	**P**	**R**	**FM**	**P**	**R**	**FM**	**P**	**R**	**FM**
A	0,92	0,86	0,89	0,93	0,93	0,93	0,93	1,00	0,96	0,95	1,00	0,97
B	0,86	1,00	0,93	0,75	0,86	0,80	0,67	0,86	0,75	0,88	1,00	0,93
C	0,81	0,54	0,65	0,77	0,71	0,74	0,89	0,71	0,79	0,91	0,87	0,89
D	0,50	0,70	0,58	0,58	0,70	0,64	0,47	0,70	0,56	0,80	0,57	0,67
E	0,57	0,50	0,53	0,90	0,56	0,69	0,71	0,31	0,44	0,75	0,55	0,63
F	0,53	0,90	0,67	0,56	0,90	0,69	0,50	0,80	0,62	0,67	0,67	0,67
G	0,92	0,85	0,89	0,92	0,85	0,88	0,92	0,92	0,92	0,70	0,92	0,80

Uporednom analizom modela izvršeno je izdvajanje rezultata klasifikatora pojedinačno po klasama.

U slučaju klase A, izdvojeni su klasifikatori *J48* i Random forest. *RF* klasifikator ostvario je preciznost od 0,95. Oba klasifikatora su postigla opoziv od 1,00, a najveća vrednost F-mere, FM=0,97, utvrđena je za Random forest klasifikator.

U slučaju klase B, izdvojeni su klasifikatori *NB* i Random forest. *RF* klasifikator ostvario je preciznost od 0,88. Oba klasifikatora su ostvarila opoziv od 1,00 i najveću vrednost F-mere, FM=0,93.

Za klasu C istakao se *RF* klasifikator sa ostvarenom preciznošću od 0,91, opozivom od 0,87 i FM merom od 0,89.

U slučaju klase D, najveća preciznost, P=0,80, ostvarena je sa *RF* klasifikatorom. Opoziv od 0,70 ostvaren je sa *NB*, *HNB* i *J48* klasifikatorom, a najveću vrednost F-mere, FM= 0,67 ostvario je *RF* klasifikator.

Za klasu E istakao se *HNB* klasifikator sa ostvarenom preciznošću od 0,90, opozivom od 0,56 i FM merom od 0,69.

U slučaju klase F najbolji rezultati su postigli *RF* (P=0,67) i *HNB* (R=0,90, FM=0,69).

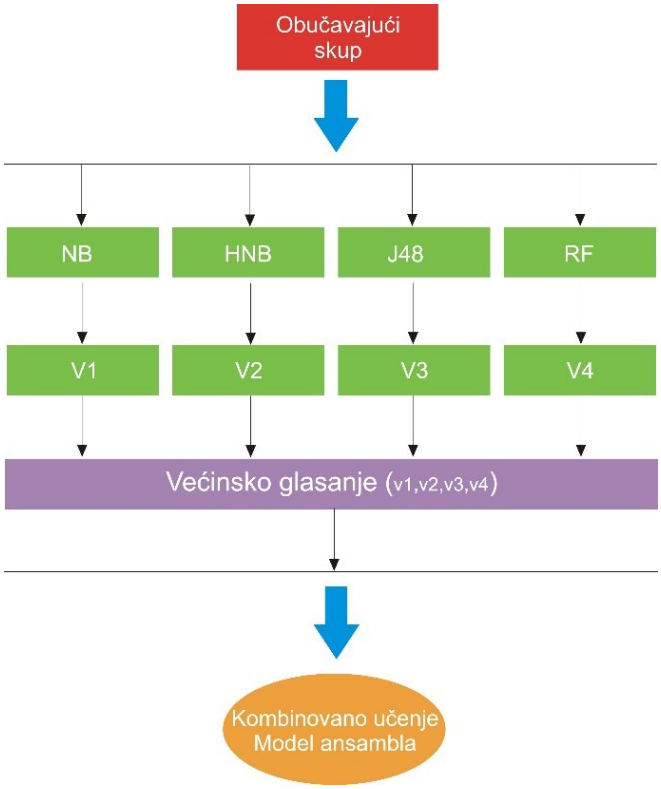
Kod klase G, klasifikatori *NB*, *HNB* i *J48* su ostvarili istu vrednost preciznosti od 0,92, najbolji opoziv zabeležen je sa *RF*, a FM mera sa *J48* klasifikatorom.

Za studiju slučaja mešanog okruženja učenja, rezultati sprovedene analiza ukazuju na neuaglašenost klasifikatora po klasama više-dimenzionalnog klasnog obeležja. U cilju izgradnje stabilnog i preciznog modela predviđanja primenjen je mehanizam kombinovanja klasifikatora u ansambl implementacijom glasanja većine.

7.3. Model predviđanja za mešano okruženje učenja primenom glasanja većine

Eksperement kombinovanja klasifikatora sproveden je izgradnjom ansambla primenom meta algoritma *Vote* [156] u okruženju *Weka*.

Uzimajući u obzir da su prethodno sprovedeni eksperimenti ukazali na značaj primenjenih klasifikatora za klase više-dimenzionalnog obeležja Ocena, predložen je model glasanja većine mešanog okruženja učenja. Predloženi predviđajući model integrisanja *NB*, *HNB*, *J48* i *RF* klasifikatora u ansambl prikazan je na slici 7.5.



Slika 7.5: Struktura predloženog Vote modela

Osnovni koncept zasnovan je na utvrđivanju značaja svakog od klasifikatora (v1,v2,v3,v4) pojedinačno za klase obeležja *Ocena*. Svaki od kandidata klasifikatora ostvario je dobre rezultate za neku od klasa i predložen je model kombinovanjem sva četiri kandidata klasifikatora u ansambl. Agregacija performansi klasifikatora omogućava da se nedostaci jedne metode mogu kompenzovaniti prednostima drugih.

Utvrđivanje najboljeg klasifikatora pravilom većinskog glasanja zasniva se na smanjivanju greške klasifikacije, balansiranjem preciznosti, opoziva i FM mere. Formirani model ansambla ostvario je tačnost predviđanja od 90,43%. Više-dimenzionalna matrica grešaka i performanse predloženog modela prikazane su u tabelama 7.8a i 7.8b respektivno.

Tabela 7.8a. Matrica grešaka Vote modela

Predvidajuće klase							**Stvarne klase**
A	B	C	D	E	F	G	
22	0	0	0	0	0	0	A
1	9	1	0	0	0	0	B
0	1	17	1	0	0	0	C
0	0	0	4	0	0	0	D
1	0	0	1	6	1	0	E
0	0	0	0	2	11	1	F
0	0	0	0	0	0	16	G

Tabela 7.8b. Performanse Vote modela

Klase	**TP**	**FP**	**P**	**R**	**FM**
A	1,00	0,01	0,96	1,00	0,99
B	0,82	0,01	0,90	0,82	0,86
C	0,89	0,01	0,94	0,89	0,92
D	1,00	0,02	0,67	1,00	0,80
E	0,75	0,02	0,75	0,75	0,75
F	0,77	0,01	0,92	0,79	0,85
G	1,00	0,01	0,94	1,00	0,97
VOTE model	0,90	0,01	0,91	0,90	0,90

Kreirani su ansambli kombinovanjem po jednog klasifikatora stabla odlučivanja i bajesovskih metoda. *Vote* algoritam je generisao modele koji su pokazali smanjenje performansi u odnosu na prvobitno kreiran ansambl kombinovanjem sva četri klasifikatora (videti tabelu 7.9).

Tabela 7.9. Performanse različitih modela anasambla

Ansambli	P	R	FM
NB+J48	0,79	0,72	0,72
NB+Random Forest	0,82	0,79	0,79
HNB+J48	0,79	0,73	0,73
HNB+Random Forest	0,80	0,78	0,78
HNB+NB+J48+RF	**0,91**	**0,90**	**0,90**

Rezultati prikazani u tabeli ukazuju da kombinovanje *RF* klasifikatora sa jednim od bajesovskih metoda utiče na ostvarenje boljih performansi modela predviđanja. Integrisanjem *J48* sa *NB* ili *HNB* klasifikatorom formirani su modeli približno istih performansi.

U cilju definisanja stabilnog i pouzdanog modela predviđanja mešanog okruženja učenja, izvršeno je poređenje FM mere po klasama predloženog ansambla *NB+HNB+J48+RF* sa vrednostima formiranih modela u prethodnim eksperimentima (tabela 7.10).

Tabela 7.10. Poređenje FM mere različitih formiranih modela i predloženog modela ansambla

	FM				
	NB	**HNB**	**J48**	**RF**	**Vote**
A	0,89	0,93	0,96	0,97	0,99
B	0,93	0,80	0,75	0,93	0,86
C	0,65	0,74	0,79	0,89	0,92
D	0,58	0,64	0,56	0,67	0,80
E	0,53	0,69	0,44	0,63	0,75
F	0,67	0,69	0,62	0,67	0,85
G	0,89	0,88	0,92	0,80	0,97

Iz rezulata prikazanih u tabeli 7.10 evidentno je da predloženi model ansambla ostvaruje značajno poboljšanje osim u slučaju klase B. Ova nedoslednost modela može se objasniti postojanjem neizbalansiranosti u skupu podataka. Za rešenje ovog problema korišćena je funkcija *Resample* [157] koja omogućava kreiranje slučajnog poduzorka skupa podataka koristeći uzorkovanje sa mogućnošću zamene. Na taj način se poboljšava balansiranost vrednosti podataka, što utiče na rezultate primenjenih klasifikatora. Funkcija *Resample* utiče na uspostavljanje poboljšane distribucije klasa u poduzorku, a time i raspodela vrednosti koja je približna uniformnoj distribuciji. Implementacija opisane funkcije slučajnog

uzorkovanja podataka realizovana je primenom *Resample* filtera za nadgledano učenje koji je dostupan u *Weka* okruženju.

Kreiran je novi model ansambla predviđanja ali sada za ulazni skup reuzorkovanih (engl. *resampled*) podataka. Novi model je ostvario značajno poboljšanje tačnosti od 98,94% i smanjenje greške klasifikacije na 1,06%. Više-dimenzionalna matrica grešaka i FM mera poboljšanog modela prikazani su u tabelama 7.11a i 7.11b respektivno.

Tabela 7.11a. Matrica grešaka Vote Resample modela

Predviđajuće klase							**Stvarne klase**
A	B	C	D	E	F	G	
14	0	0	0	0	0	0	A
0	13	0	0	0	0	0	B
0	0	24	0	0	0	0	C
0	0	0	6	0	0	0	D
0	0	0	0	8	0	0	E
0	0	0	0	0	13	0	F
0	0	0	0	0	1	15	G

Tabela 7.11b. Performanse Vote Resample modela

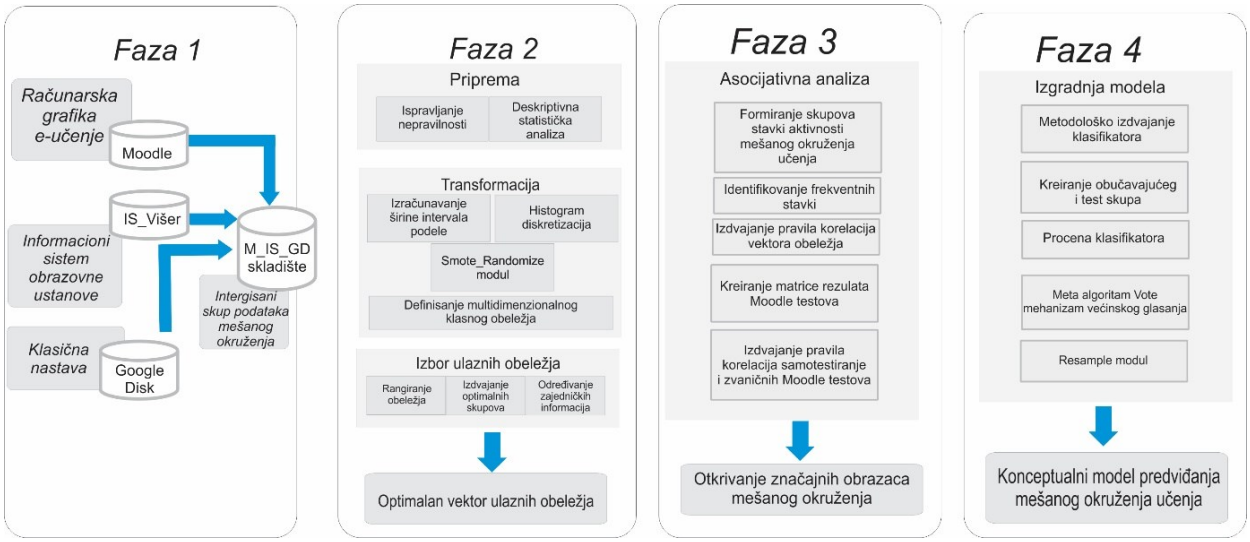
Klase	**FM**
A	1,00
B	1,00
C	1,00
D	1,00
E	1,00
F	0,96
G	0,97
***VOTE Resample* model**	0,99

Model ansambla generisan kombinovanjem izabranih klasifikatora omogućio je značajno poboljšanje tačnosti i preciznosti predviđanja više-dimenzionalnog klasnog obeležja rešavanjem problema neuravnoteženosti podataka.

8. ZAKLJUČAK

Predviđanje performansi studenata predstavlja značajan zadatak integrisan u proces izgradnje modela obrazovnog okruženja za učenje. Stabilan i precizan model bi trebalo da omogući praćenje procesa učenja i tačno predviđanje uspeha studenata na osnovu realizovanih aktivnosti u mešanom okruženju učenja. Efikasnost modela zasnovana je na detaljnoj dubinskoj analizi podataka izdvojenih iz Moodle sistema za upravljanje učenjem, informacionog sistema obrazovne ustanove i zapisa o klasičnom načinu održavanja nastave. Na osnovu rezultata eksperimenata sprovedenog istraživanja, identifikovana je metodologija za otkrivanje znanja iz Moodle sistema što ukazuje na ostvaren doprinos disertacije u oblasti otkrivanja znanja, sistema za upravljanje učenjem i mešanog okruženja učenja. Razvijena metodologija za mešano okruženje učenja *(MMOU)* podrazumeva dubinsku analizu podataka Moodle kursa kombinovanog sa klasičnim načinom realizacije nastave. Implementacija predloženog procesa smanjuje vreme izgradnje modela predviđanja. Uzimajući u obzir prirodu analiziranog skupa podataka, rad u ovoj disertaciji takođe je od pomoći za identifikaciju važnih mera i kriterijuma efikasne procene i vrednovanja modela.

Na slici 8.1 je dat šematski prikaz predložene metodologije koji obuhvata četri osnovne faze.



Slika 8.1: Predloženo okruženje sistema za izgradnju modela predviđanja mešanog okruženja učenja

Cilj prve faze se odnosi na formiranje obrazovnog skupa integrisanjem podataka distribuiranih izvora: baze podataka Moodle sistema, informacionog sistema obrazovne ustanove i zapisa vođenih na Google disku o realizaciji klasične nastave. Druga faza obuhvata detaljnu deskriptivnu statističku pripremu i obradu podataka kreiranog obrazovnog skupa mešanog okruženja učenja u formi pogodnoj za otkrivanje znanja. Treća faza istražuje i identifikuje postojanje interesantnih značajnih obrazaca za performanse učenja koji ukazuju na korelacije između realizovanih aktivnosti studenata u kombinovanim okruženjima mešanog učenja. Četvrta faza razvija konceptualni model predviđanja implementacijom identifikovanih metoda dubinske analize i otkrivanja znanja pogodnih za mešano okruženje učenja.

Prva faza predložene metodologije obuhvata zadatke prikupljanja podataka pogodnih za istraživačke aktivnosti. Neophodno je bilo izdvojiti podatke iz tri različita izvora i integrisati u jedno skladište. Postupak je zahtevao znatnu količinu vremena s obzirom na stepen neusklađenosti strukture podataka distribuiranih izvora. Informacije o aktivnostima studenata na online kursu Računarska grafika izdvojeni su kreiranjem SQL pogleda nad bazom podataka Moodle sistema. Generisanjem elektronskih izveštaja, iz informacionog sistema obrazovne ustanove, izdvojeni su podaci o ostvarenim ocenama za predmetni kurs. Zapisi o evidenciji, realizovanih aktivnosti studenata na laboratorijskim vežbama i predavanjima, preuzeti su sa Google diska. Integrisanjem izdvojenih podataka u tabelarnu formu formirano je jedinstveno skladište mešanog okruženja u okviru sistema za upravljanje bazama podataka.

Druga faza predstavlja proširenu fazu predprocesiranja i obuhvata tri zadatka: pripremu, transformaciju i izbor optimalnog vektora ulaznih obeležja. Zadatak pripreme odnosio se na utvrđivanje postojanja nepravilnosti u podacima. Identifikovane su vrednosti koje nedostaju i zamenje vrednošću nula, a izuzeci izdvojeni u poseban skup. Sprovedena deskriptivna statistička analiza identifikovala je neravnomerenu raspodelu i varijabilnost podataka. Zadatak transformacije obuhvatao je diskretizaciju numeričkih obeležja histogram metodom kombinovanom sa Skotovim pravilom za izračunavanje širine intervala diskretnih vrednosti i definisanje više-dimenzionalnog klasnog obeležja utvrđivanjem sedam nominalnih klasnih oznaka. SMOTE i Randomize filter integrisani su u modul za rešavanje problema neravnomerne raspodele i poboljšanja efikasnosti histogram diskretizacije. Zadatak izbora optimalnog ulaznog vektora obeležja realizovan je kombinovanom primenom filter metoda, metoda omotača i zajedničke mere informacija.

Treća faza zasnovana je na konceptu asocijativne analize mešanog okruženja učenja Otkrivene su korelacija od značaja između obeležja ulaznog vektora i Moodle testova. Izdvojeni su značajni, razumljivi obrasci koji definišu homogenost i označavaju suštinska svojstva u podacima.

U okviru četvrte faze kreiran je konceptualni model predviđanja mešanog okruženja učenja. Detaljna klasifikacija studenata realizovana je definisanjem više-dimenzionalnog klasnog obeležja. Ostvarena je personalizacija i praćenje procesa učenja. Izvršeno je metodološko izdvajanje klasifikatora podesnih za sveobuhvatni vizualni način prikaza, rad sa malim uzorcima skupova i kategorijskim tipom podataka, neosetljivih na više-dimenzionalnost klasnog obeležja, postojanje izuzetaka i nedostajućih vrednosti. Za kreiranje modela predviđanja predložen je meta algoritam Vote. Mehnizmom kombinovanja Naïve Bayes, Hidden Naïve Bayes, J48 i Random Forest klasifikatora formiran je ansambl. Osnovni koncept ansambl modela zasnovan je na utvrđivanju značaja svakog od klasifikatora pojedinačno za sve klase obeležja Ocena. Model ansambla proširen je primenom *Resample* filtera čime je ostvareno značajno poboljšanje tačnosti predviđanja i rešen problem postojanja nebalansiranosti domena vrednosti klasnog obeležja. Predloženi model predviđanja mešanog okruženja učenja je ostvario značajno poboljšanje tačnosti od 98,94% i smanjenje greške klasifikacije na 1,06%.

8.1. Rezime doprinosa

U ovoj disertaciji ostvareni su sledeći doprinosi:

- Potvrđen je značaj sprovođenja deskriptivne statističke analize originalnog ulaznog skupa mešanog okruženja kojom se omogućava uvid u raspodelu domena vrednosti i određivanja načina transformacije numeričkih obeležja.
- U predloženom modelu predviđanja mešanog okruženja učenja, ostvarena je personalizacija procesa učenja studenata zasnovana na klasnom više-dimenzionalnom obeležju *Ocena*.
- U sistemu zasnovanom na razvijenoj metodologiji za mešano okruženje učenja identifikovani su i izdvojeni obrasci od značaja Moodle kursa i obrasci korelacija između određenih aktivnosti studenata i ostvarenog uspeha.

8.2. Smernice za dalja istraživanja

Dalji nastavak istraživanja opisanog u disertaciji biće usmeren u nekoliko sledećih pravaca otvorenih za dalji rad. Validacija modela, testiranjem predložene metodologije na podacima iz drugih oblasti kako bi se proverilo da li predložene metode mogu raditi na njima. Integracija zadataka druge faze u jedan modul predloženog okruženja uslovila bi smanjenje vremena potrebnog za obradu ulaznog skupa. Razvoj funkcije za identifikovanje i zamenu vrednosti koje označavaju aktivnosti koje student nije radio omogućila bi efikasniji način rada i predviđanje izračunavanjem korelacionih zavisnosti između nerealizovanih zadataka.

Realizacija fizičkog modela zasnovanog na razvijenoj metodologiji i konceptualnom modelu omogućila bi izgradnju sistema za generisanje povratnih informacija o efikasnosti procesa mešanog okruženja učenja. Integracija ovog sistema podrške sa socijalnim mrežama omogućila bi studentima pronalazak personalizovanijih resursa za učenje kao i mogućnost da dele svoje znanje. Podaci koji se dobiju iz socijalnih sistema podrške i preporuke mogu se koristiti za ostvarivanje boljih rezulata u predviđanju performansi studenata.

9. LITERATURA

[1] W.Horton, K.Horton, *E-learning Tools and Technologies. A consumer's guide for trainers, teachers, educators, and instructional designers*.Wiley Publishing Inc., 2003.

[2] R. Lacurezeanu et al., *A Comparative Analysis of the Opportunity and the Possibility of Implementing some E-Learning Components at FSEGA*, WSEAS Transactions on Business and Economics Issue 8, Volume 3, August 2006.

[3] R.Cobcroft, S.Towers, J.Smith, A.Bruns, Mobile learning in review: *Opportunities and challenges for learners, teachers, and institutions*, Proceedings of Online Learning and Teaching (OLT) Conference, Queensland University of Technology, Brisbane, pp. 21-30, September 2006.

[4] P.Ramsden, *Learning to teach in higher education*, London: Routledge Falmer, 2003.

[5] M.Oliver, K.Trigwell, *Can'Blended Learning' Be Redeemed?*, E-Learning and Digital Media, 2(1), 17-26., 2005.

[6] J.Hofmann, *Why blended learning hasn't (yet) fulfilled its promises: Answers to those questions that keep you up at night.* In C. J. Bonk & C. R. Graham (Eds.), Handbook of blended learning: Global perspectives, local designs. San Fancisco, CA: Pfeiffer, 2006.

[7] J.Mostow, J.Beck, *Some useful tactics to modify, map and mine data from intelligent tutors*, Natural Language Engineering. 12(2), 195–208, 2006.

[8] W.H.Rice*, Moodle E-learning Course Development. A complete guide to succeful learning using Moodle*, Packt publishing. 2006.

[9] U. Fayyad, G.P. Shapiro, P.Smyth, *From data mining to knowledge discovery in databases*, AI Magazine, 17(3):37–54, 1996.

[10] J.F.William, G.P.Shapiro, C.J. Matheus, *Knowledge discovery in databases: An overview*, AI Magazine, 13(3):57–70, 1992.

[11] C.Romero, S.Ventura*, Data Mining in Education*, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 3.10.1002/widm.1075, 2013.

[12] C. Romero, and S. Ventura, *Educational data mining: a survey from 1995 to 2005.* In Expert Systems with Applications, volume 33(1), pages 135-146, 2007.

[13] R.Baker, K.Yacef, *The State of Educational Data Mining in 2009: A Review and Future Visions.* Journal of Educational Data Mining, 1, 1, 3-17., 2009.

[14] A.Ortigosa, R.M.Carro, *The Continuous Empirical Evaluation Approach: Evaluating Adaptive Web-based Courses*, International Conf. User Modeling, Canada (pp. 163-167). 2003.

[15] Moodle, http://moodle.org/, *Free open source course management system for online learning,* 2006.

[16] E.Gaudioso, L.Talavera, *Data mining to support tutoring in virtual learning communities: experiences and challenges*, 2006

[17] L. Tsantis, J.Castellani, *Enhancing learning environments through solution-based knowledge discovery tools: Forecasting for self-perpetuating systemic reform,* Journal of Special Education Technology, 16(4), 39–52. 2001.

[18] U.Fayyad, G.Piatetsky-Shapiro, P. Smyth, *From data mining to knowledge discovery: An overview*, Advances in Knowledge Discovery and DataMining, MIT Press, pages 1-36, 1996.

[19] J.Han, M.Kamber, *Data Mining: Concepts and Techniques*, 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor. 2006.

[20] A.Y.K. Chan, K.O. Chow, K.S. Cheung, *Online Course Refinement through Association Rule Mining*, Journal of Educational Technology Systems Volume 36, Number 4 / 2007-2008, pp 433 – 44, 2008

[21] *http://www.educationaldatamining.org/index.html*

[22] J.Hom, *The Usability Methods ToolBox*, 1998

[23] O.Zaiane, J.Han, *Mining for E-Learning*, Gold. 2001.

[24] C.Romero, S.Ventura, E.García, *Data mining in course management systems: Moodle case study and tutorial*, Computers & Education,Volume 51, Issue 1, Pages 368-384, ISSN 0360-1315, 2008, https://doi.org/10.1016/j.compedu.2007.05.016.

[25] A.L. Blum, P. Langley*, Selection of Relevant Features and Examples in Machine Learning*, Artificial Intelligence, vol. 97, pp. 245-271, 1997.

[26] K. Kira, L. Rendell, *A practical approach to feature selection*, in Proceedings of the Ninth International Conference on Machine Learning. 1992, pp. 249-256, Morgan Kaufmann.

[27] M. Dash, H. Liu, *Feature Selection for Classification*, Intelligent Data Analysis: An Int'l J., vol. 1, no. 3, pp. 131-156, 1997.

[28] H. Liu, H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*, Boston: Kluwer Academic, 1998.

[29] M. Ben-Bassat, *Pattern Recognition and Reduction of Dimensionality*, Handbook of Statistics-II, P.R. Krishnaiah and L.N. Kanal, eds., pp. 773-791, North Holland, 1982.

[30] P. Mitra, C.A. Murthy, S.K. Pal, *Unsupervised Feature Selection Using Feature Similarity*, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 3, pp. 301-312, Mar. 2002.

[31] K. Kira, L.A. Rendell, *The Feature Selection Problem: Traditional Methods and a New Algorithm*, Proc. 10th Nat'l Conf. Artificial Intelligence, pp. 129-134, 1992.

[32] R. Kohavi, G.H. John, *Wrappers for Feature Subset Selection*, Artificial Intelligence, vol. 97, nos. 1-2, pp. 273-324, 1997.

[33] E. Leopold, J. Kindermann, *Text Categorization with Support Vector Machines. How to Represent Texts in Input Space?*, Machine Learning, vol. 46, pp. 423-444, 2002.

[34] K. Nigam, A.K. Mccallum, S. Thrun, and T. Mitchell, *Text Classification from Labeled and Unlabeled Documents Using EM*, Machine Learning, vol. 39, 103-134, 2000.

[35] Y. Rui, T.S. Huang, S. Chang, *Image Retrieval: Current Techniques*, Promising Directions and Open Issues, Visual Comm. and Image Representation, vol. 10, no. 4, pp. 39-62, 1999

[36] W. Lee, S.J. Stolfo, K.W. Mok, *Adaptive Intrusion Detection: A Data Mining Approach*, AI Rev., vol. 14, no. 6, pp. 533-567, 2000.

[37] S. Das, Filters, *Wrappers and a boosting-based hybrid for feature selection*, Proceedings of the Eighteenth International Conference on Machine Learning, pp. 74–81, 2001.

[38] M.A. Hall, *Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning*, Proc. 17th Int'l Conf. Machine Learning, pp. 359-366, 2000.

[39] H. Liu and R. Setiono, *A Probabilistic Approach to Feature Selection-A Filter Solution*, Proc. 13th Int'l Conf. Machine Learning,pp. 319-327, 1996.

[40] L. Yu and H. Liu, *Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution*, Proc. 20th Int'l Conf.Machine Learning, pp. 856-863, 2003.

[41] R. Kohavi, G.H. John, *Wrappers for Feature Subset Selection*, Artificial Intelligence, vol. 97, No. 1-2, pp. 273-324, 1997.

[42] S. Das, Filters, *Wrappers and a Boosting-Based Hybrid for Feature Selection*, Proc. 18th Int'l Conf. Machine Learning, pp. 74-81, 2001.

[43] E.Xing, M.Jordan, R.Karp, *Feature selection for high-dimensional genomic microarray data*, Proceedings of the Eighteenth International Conference on Machine Learning, pp. 601–608, 2001.

[44] M.Dash, H.Liu, H. Motoda, *Consistency based feature selection*, Proceedings of the Fourth Pacific Asia Conference on Knowledge Discovery and Data Mining, pp. 98–109, Springer-Verlag, 2000.

[45] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA., 1993.

[46] I. Kononenko, *Estimating attributes: Analysis and extensions of relief*, in Proceedings of the Seventh European Conference on Machine Learning, pp. 171-182, Springer-Verlag, 1994.

[47] L. Hu, L. Zhang, *Real-time Internet Traffic Identification Based on Decision Tree*, World Automation Congress (WAC). IEEE. 1-3, 2012.

[48] K.Cios,W.Pedrycz, R.Swiniarski, L.Kurgan, *Data Mining A Knowledge Discovery Approach*, Springer , 2007.

[49] H.Liu, F.Hussain, C.Tan, M.Dash, *Discretization: An Enabling Technique,* Data Mining and Knowledge Discovery, Vol.6(4), pp.393-423, 2002.

[50] E Xu, S.Liangshan, R.Yongchang, W.Hao, Q. Feng, *A new discretization approach of continuous attributes*, Asia-Pacific Conference on Wearable Computing Systems, vol. 5, no. 2, pp. 136-138, 2010.

[51] J.Dougherty, R.Kohavi, M.Sahami, *Supervised and unsupervised discretization discretization of continuous feature*, In Proc. 12th International Conference on Machine Learning, Los Altos, CA:Morgan Kaufman,pp 194-202, 1995.

[52] R. Kerber, *Discretization of Numeric Attributes*, Proceedings of the 10th National Conference on Artificial Intelligence, MIT Press, Cambridge, MA, pp.123-128. 1992.

[53] M.R Chmielevski, J.W. Grzymala-Busse, *Global discretization of continuous attributes on preprocessing for machine learning*. In Third International Workshop on Rough Sets and Soft Computing, pp. 294-301, 1994.

[54] H. Sturges. *The Choice of a Class Interval*, J. American Statistical Association: 1926, pp. 65–66.

[55] D. Freedman, P. Diaconis. *On the Histogram as a Density estimator: L2 Theory, Probability Theory and Related Fields* (Heidelberg: Springer Berlin) 57 (4), pp. 453–476, December 1981

[56] D. Scott, *On Optimal and Data-based Histograms*, Biometrika 66 (3), pp. 605–610, 1979.

[57] I. H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, San Mateo, CA: Morgan Kaufman, 1999.

[58] U. M. Fayyad, K. B. Irani, *Multi-interval discretisation of continuous-valued attributes for classification learning*, in Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, pp. 1022-1027, Morgan Kaufmann, 1993.

[59] T.Pang-Ning, M.Steinbach, V.Kumar, *Introduction to Data Mining,*University of Minnesota, 2006.

[60] J.R.Quinlan, *Induction of Decision Tree*, Journal of Machine learning, Morgan Kaufmann Vol.1, pp.81-106, 1986.

[61] B.Nithyassik, Nandhini, E.Chandra, *Classification Techniques in Education Domain*, International Journal on Computer Science and Engineering, Vol. 2, No.5, pp.1647-1684, 2010.

[62] *WEKA*, http://www.cs.waikato.ac.nz/ml/weka/, University of Waikato, New Zealand.

[63] T.M Mitchell, *Machine Learning*, McGraw-Hill, New York, 1997.

[64] G. Valentini, F. Masulli, *Ensembles of learning machines*, Volume 2486 of Lecture Notes in Computer Science, pp. 3–19. Springer-Verlag, 2002. Invited Review

[65] L. Breiman,. *Random forest*, Machine Learning, 45(1):5–32, 2001.

[66] E. Bauer, R. Kohavi, *An empirical comparison of voting classification algorithms,* Machine Learning, 36(1/2):105–139, 1999.

[67] T. Dietterich, *An experimental comparison on three methods for constructing ensembles of decision trees: bagging, boosting, and randomization,* Machine Learning, 40(2):139–157, 2000.

[68] L. Breiman, *Bagging predictors*, Machine Learning, Vol. 24, No. 2, pp. 123–140, 1996.

[69] L.Breiman, J.H. Friedman, R.A.Olshen, C.J Stone*, Classification and Regression Trees*, Wadsworth, Monterey, CA., 1984.

[70] M.Pal, P. M.Mather, *An assessment of the effectiveness of decision tree methods for land cover classification,* Remote Sensing of Environment, Vol.86, pp.554-565. 2003.

[71] W.Feller*, An Introduction to Probability Theory and Its Application*, Vol.1, 3rd edition, New York: Wiley,1968.

[72] Y. Amit, D. Geman, *Shape quantization and recognition with randomized trees*, Neural Computation,Vol.9, pp.1545–1588, 1997.

[73] A.Darwiche, Bayesian Networks, In: *Handbook of Knowledge Representation, Foundations of Artificial Intelligence*, Frank Van Harmelen, Vladimir Lifschitz and Bruce Porter (Eds.), Elsevier, Amsterdam, ISBN-10: 0080557023, pp: 1034-1034. 2008.

[74] H. Zhang. *The optimality of Näive Bayes*, In Proceedings of the Seventeenth Florida Artificial Intelligence Research Society Conference, pages 562–567. The AAAI Press, 2004.

[75] P. Domingos, M. Pazzani, *On the optimality of the simple Bayesian classifier under zero-one loss*, Machine Learning, 29:103–130, 1997.

[76] H. Zhang, L. Jiang, J. Su, *Hidden Naive Bayes*, In: Twentieth National Conference on Artificial Intelligence, 919-924, 2005.

[77] N. Friedman, D. Geiger, M. Goldszmidt*, Bayesian network classifiers*, Machine Learning, 29(2-3):131–163, 1997.

[78] G. Webb, J. Boughton, Z. Wang, *Not So Naive Bayes: Aggregating One-Dependence Estimators*, Machine Learning. 58(1):5-24. 2005

[79] Larose, D. T. *k-Nearest Neighbor Algorithm, in Discovering Knowledge in Data: An Introduction to Data Mining*, John Wiley & Sons, Inc., Hoboken, NJ, USA. doi:10.1002/0471687545.ch5, 2004.

[80] A. H. Peterson and T. R. Martinez, *Estimating the potential for combining learning models*, Proceedings of the ICML Workshop on Meta-Learning, pp. 68–75, 2005.

[81] L. I. Kuncheva and C. J. Whitaker, *Measures of diversity in classifier ensembles*, Machine Learning, Vol. 51, pp. 181–207, 2003

[82] D. Ruta and B. Gabrys, *Analysis of the correlation between majority voting error and the diversity measures in multiple classifier systems*, Proceedings of the Fourth International Symposium on Soft Computing, 2001

[83] G. Giacinto and F. Roli, *A theoretical framework for dynamic classifier selection*, Proceedings of the Fifteenth International Conference on Pattern Recognition, 2000.

[84] *Classifier selection for majority voting*, Information Fusion, Vol. 6, No. 1, pp. 63–81, 2005.

[85] L. Hall, K. Bowyer, W. Kegelmeyer, T. Moore, C. Chao, *Distributed learning on very large data sets*. In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 79–84. 2000

[86] S. Kotsiantis,C. Pierrakeas, P.Pintelas, *Prediction of Student's Performance in Distance Learning Using Machine Learning Techniques*, Applied Artificial Intelligence,Volume 18, 2004 - Issue 5, 411-426

[87] S. Kotsiantis, P. Pintelas, *Local voting of weak classifiers*, International Journal of Knowledge-based and Intelligent Engineering Systems 9:pp. 239-248, 2005.

[88] W. Zang,F. Lin*, Investigation of web-based teaching and learning by boosting algorithms*. In Proceedings of IEEE International Conference on Information Technology: Research and Education (ITRE 2003), pp. 445–449, 2003.

[89] R. Agrawal, T. Imielinski, A. Swami, *Mining association rules between sets of items in large databases*, in: Proceedings of the Association for Computing Machinery—Special Interest Group on Management of Data, ACM-SIGMOD, May, pp. 207–216, 1993.

[90] R. Agrawal, R. Srikant, *Fast algorithms for mining association rules*, in: Proceedings of the 20th International Conference on Very Large Databases, VLDB, September, pp. 487–491, 1994.

[91] A.Merceron, K. Yacef, *Interestingness Measures for Association Rules in Educational Data*, In International Conference on Educational Data Mining, Montreal, Canada, 57-66., 2008.

[92] http://www3.cs.stonybrook.edu/~cse634/lecture_notes/07apriori.pdf

[93] T. Scheffer, *Finding Association Rules That Trade Support Optimally against Confidence*. In: 5th European Conference on Principles of Data Mining and Knowledge Discovery, pp.424-435, 2001.

[94] R.Baker, A.Merceron, P.I.Pavilk, *Educational Data Mining* , 3st International Conference on Educational Data Mining, Proceedings, Pittsburgh, USA, 2010.

[95] B. K. Baradwaj, S. Pal, *Mining Educational Data to Analyze Student Performance*, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol.2(6), pp.63-69, 2011.

[96] A.Srivastava, J. Srivastava, *Data Mining in Education Sector: A Review, 2015*.

[97] O. Maimon and L. Rokach, *Handbook of data mining and knowledge discovery*, Mlc, Publisher: Springer US, volume: 2, ed., 2002.

[98] C.Romero, S.Ventura, *Educational Data Mining: A Review of the State of the Art*, IEEE Transactions on Systems, Man, and Cybernetics, Part C, Volume: 40, Issue: 6, Nov. 2010.

[99] C. Silva, J.Fonseca, *Educational Data Mining: A Literature Review*, Europe and MENA Cooperation Advances in Information and Communication Technologies. Advances in Intelligent Systems and Computing, Vol 520. pp. 87-94 Springer, 2017.

[100] A. Dutt et al., *Systematic Review on Educational Data Mining*, IEEE Access, Volume 7, DOI 10.1109/ACCESS.2017.2654247,2017.

[101] M.Rahkila, M. Karjalainen, *Evaluation of learning in computer based education using log systems*. In ASEE/IEEE frontiers in education conference, San Juan, Puerto Rico, 16–21, 1999.

[102] Beck, J., & Woolf, B. *High-level student modeling with machine learning*, Proceedings of the 5th International Conference on Intelligent Tutoring Systems, pp.584–593, 2000.

[103] J.Sheard, J.Ceddia, J.Hurst, J. Tuovinen, *Inferring Student Learning Behaviour from Website Interactions: A Usage Analysis*. In Journal Education and Information Technologies, Vol.8(3), pp.245-266. 2003.

[104] E.Heathcote, S.Dawson, *Data Mining for Evaluation, Benchmarking and Reflective Practice in a LMS*, In World conference on E-learning in corporate, government, healthcare & higher education, Vancouver, Canada, pp.326-333. 2005.

[105] A. Merceron, K.Yacef, *Educational Data Mining: a Case Study*, Artificial Intelligence in Education (AIED2005)

[106] A.A. Ramli, *Web usage mining using apriori algorithm: uum learning care,* Proceedings of the International Conference on Knowledge Management 2005

[107] Á.Horváth, E.Jókai, J. Horváth, *Analysis of learning aspects of students by Moodle based data mining methods,* JAMPAPER 4./II./2007.

[108] B. Ribeiro and A. Cardoso, *Behavior pattern mining during the evaluation phase in an e-learning course,* International Conference on Engineering Education – ICEE 2007.

[109] F. Castro, A.Vellido, A.Nebot, F.Mugica, *Applying Data Mining Techniques to e-Learning Problems*, Evolution of Teaching and Learning Paradigms in Intelligent Environment. Studies in Computational Intelligence, 62, Springer-Verlag, 183-221. 2007.

[110] G.J.Hwang, P.S. Tsai, C.C. Tsai, J.C.R.Tseng, *A novel approach for assisting teachers in analyzing student web-searching behaviors*. In Computer & Education Journal, 51, 926-938., 2008.

[111] T.D. Gedeon, H.S.Turner, *Explaining student grades predicted by a neural network.* In International conference on Neural Networks, Nagoya, 609-612. 1993

[112] M. Delgado, E. Gibaja, M.C. Pegalajar, O Pérez,. *Predicting Students' Marks from. Moodle Logs using Neural Network Models*. In International Conference on Current Developments in Technology-Assisted Education, Sevilla, Spain, 586-590, 2006

[113] T. Want, A. Mitrovic*, Using Neural Networks to Predict Student's Performance*. In International Conference on Computers in Education, Washington, DC, 1-5. 2002

[114] Z. Pardos, N. Heffernan, B. Anderson, C.Heffernan*, The Effect of Model Granularity on Student Performance Prediction Using Bayesian Networks*. In International Conference on User Modeling, Corfu, Greece, 435-439. 2007

[115] N.T.N. Hien, P. Haddawy*, A decision support system for evaluating international student applications.* In Frontiers In Education Conference, Milwaukee, 1-6. 2007

[116] R. Stevens, A. Giordani, M. Cooper, A. Soller, L. Gerosa, C.Cox, *Developing a Framework for Integrating Prior Problem Solving and Knowledge Sharing Histories of a Group to Predict Future Group Performance.* In International Conference on Collaborative Computing: Networking, Applications and Worksharing, Boston, pp.1-9., 2005

[117] G.Dimić, D.Prokin, K.Kuk, P.Spalević, *The use of data mining methods for analyzing and evaluating course quality in the Moodle system,* Unitech 2010, International scientific conference, Gabrovo, pp.309-315., 2010.

[118] C.C. Chan, *A Framework for Assessing Usage of Web-Based e- Learning Systems.* In International Conference on innovative Computing, Information and Control, Washington, DC, 147- 151., 2007

[119] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, *Dropout prediction in e-learning courses through the combination of machine learning techniques*, Computers & Education, vol. 53, pp. 950–965, Nov. 2009

[120] K .Shyamala, *Data Mining Model for a Better Higher Educational System*, Information Techology Journal, Vol. 5 2006

[121] G. Dekker, M. Pechenizkiy, and J. Vleeshouwers, *Predicting students drop out: a case study*, in 2nd International Educational Data Mining Conference (EDM09), pp. 41–50, 2009.

[122] G.Dimić, K.Kuk,P.Spalević, Z.Trajčevski, Z.Todorović,. *Accuracy analysis of the classification model evaluation in the e-learning environment*, Journal Technics Technologies Education Management, Volume 8, Number 2, pp. 667-673, 2013

[123] Z. Kovačić, *Early prediction of student success: Mining students enrolment data*, in Informing Science & IT Education Conference (InSITE) 2010, 2010.

[124] Y. Zhang, S. Oussena, T. Clark, and H. Kim, *Use Data Mining To Improve Student Retention in Higher Education - A Case Study*, in 12th International Conference on Enterprise Information Systems (ICEIS), 2010.

[125] D. Kabakchieva, *Student Performance Prediction by Using Data Mining Classification Algorithms*, International Journal of Computer Science and Management Research, vol. 1, no. 4, pp. 686–690, 2012.

[126] W.Hämäläinen, M. Vinni, *Comparison of machine learning methods for intelligent tutoring systems. Conference Intelligent Tutoring Systems,* Taiwan, pp. 525–534., 2006.

[127] M. Cocea, S.Weibelzahl. *Cross-system validation of engagement prediction from log files.* In EC-TEL, pages 14–25, 2007

[128] C. Romero, S.Ventura, P. G. Espejo and C.Hervás. *Data mining algorithms to classify students.* In International Conference on Educational Data Mining, Montreal,Canada, pp. 8-17., 2008.

[129] Z.Pardos, N.Heffernan, B. Anderson, C.Heffernan, *The Effect of Model Granularity on Student Performance Prediction Using Bayesian Networks,* In International Conference on User Modeling, Corfu, Greece, pp.435-439, 2007.

[130] Diego Garcia-Saiz, Marta Zorrilla; *Comparing classification methods for predicting distance students' performance,* Proceedings of the Second Workshop on Applications of Pattern Analysis, PMLR 17:26-32, 2011.

[131] W. Zang, and F. Lin, *Investigation of web-based teaching and learning by boosting algorithms.* In Proceedings of IEEE International Conference on Information Technology: Research and Education (ITRE 2003), pp. 445–449, 2003.

[132] B. Ribeiro and A. Cardoso, *Behavior pattern mining during the evaluation phase in an e-learning course,* International Conference on Engineering Education – ICEE 2007,

[133] G.Dimić, J.Kajević, D. Rančić, P.Spalević, D. Milić, *Implementation of features selection methods and oversampling technique in blended learning environment,* Proceedings of the 26th International Electrotechnical and Computer Science Conference, ERK 2017, Portorož, Slovenija, pp. 403-407, September 2017

[134] N. Thai-Nghe, ,A. Busche, L.Schmidt-Thieme, *Improving academic performance prediction by dealing with class imbalance.* In Proceeding of 9th IEEE International Conference on Intelligent Systems Design and Applications (ISDA'09). Pisa, Italy, IEEE Computer Society, pp.878-883,2009.

[135] M. A. Santana, E. B. Costa, B. Fonseca, F. F. De Araujo, J. Rego, *Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses*, Comput. Human Behav. Vol.73, pp.247–256. 2017, https://doi.org/10.1016/j.chb.2017.01.047

[136] G. Dimić, D.Rančić, I.Milentijević, P. Spalević, *Improvement the accuracy of prediction using unsupervised discretization method: educational data set case study*, Technical Gazette, Vol. 25/No. 2, April 2018, Print: ISSN 1330-3651, Online: ISSN 1848-6339, 10.17559/TV-20170220135853.

[137] L. Ge, H. Tang, Q. Zhou, Y. Tang, and J. Lang.. *Classification algorithms to predict students'extraversion-introversion traits*. In Cyberworlds (CW), International Conference, pp.135–138. IEEE. 2016

[138] S.B.Kotsiantis, C.J.Pierrakeas, P.E.Pintelas, *Preventing Student Dropout in Distance Learning Using Machine Learning Techniques*. In: Palade V.,Howlett R.J., Jain L. (eds) Knowledge-Based Intelligent Information and Engineering Systems. Lecture Notes in Computer Science, vol 2774. Springer, Berlin, Heidelberg, KES 2003.

[139] G. Dimić , D. Rančić , I.Milentijević , P. Spalević , K. Plećić, *Comparative study: feature selection methods in the blended learning environment*, Facta Universitatis Series: Automatic Control and Robotics Vol. 16, No 2, 2017, pp. 95 - 116

[140] B.Minaei-Bidgoli, P.Tan, W.Punch, *Mining interesting contrast rules for a web-based educational system*, In International Conference on Machine Learning Applications, Los Angeles, California, pp. 1-8, 2004.

[141] A.Merceron, A., K.Yacef, *Mining student data captured from a web-based tutoring tool: Initial exploration and results*, Journal of Interactive Learning Research, Vol.15(4), pp.319–346, 2004.

[142] Ramli, A.A., *Web usage mining using apriori algorithm: UUM learning care portal case*, In International Conference on Knowledge Management, Malaysia, pp. 1-19. 2005.

[143] C.Romero, P.Gonzalez, S.Ventura, M.J.del Jesus, F.Herrera, *Evolutionary algorithms for subgroup discovery in e-learning: A practical application using Moodle data*. In Expert System with Application Journal, Vol.36, Issue 2, Part 1, pp.1632-1644, 2009.

[144] E.García, C. Romero, S.Ventura, C.d. Castro*, A collaborative educational association rule mining tool*, In The Internet and Higher Education, Volume 14, Issue 2, pp. 77-88, ISSN 1096-7516, https://doi.org/10.1016/j.iheduc.2010.07.006., 2011.

[145] Y. Psaromiligkos et al., *Mining log data for the analysis of learners' behaviour in web-based learning management systems*, Oper Res Int Journal, Volume 11, Issue 2,pp 187–200, 2011.

[146] G. Dimić, D. Prokin, K. Kuk, P.Spalević, *Applying educational data mining in e-learning environment*, International Symposium Infoteh-Jahorina, Vol. 10, Ref. E-V-6, p. 775-779, March 2011.

[147] C.Romero, A. Zafra, J. M. Luna, S.Ventura, *Association rule mining using genetic programming to provide feedback to instructors from multiple-choice quiz data*, Expert Systems, Vol. 30, No. 2, pp.162-172, May 2013.

[148] Dimić G, Predić B, Rančić D, Petrović V, Maček N, Spalević P. *Association analysis of moodle e-tests in blended learning educational environment.* Comput Appl Eng Educ. 2017;1–14.https://doi.org/10.1002/cae.218

[149] F.Kayah, F, *Discretizing Continuous Features for Naive Bayes and C4. 5 Classifiers*, University of Maryland publications: College Park, MD, USA, 2008.

[150] Jishan et al., *Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique*, Decision Analytics, 2:1, DOI 10.1186/s40165-014- 0010-2, Springer, Open Journal, 2015.

[151] H.Liu, L.Yu, *Toward Integrating Feature Selection Algorithms for Classification and Clustering*, IEEE Transactions on knowledge and data engineering, Vol. 17, No. 4, April 2005.

[152] R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Classification*, Wiley-Interscience Publication, New Yor, 2nd edition, page.632, 2000.

[153] L.Geng, H.J. Hamilton, *Interestingness measures for data mining: a survey*, ACM Computing Surveys, 38, 1–32, 2006.

[154] A.K.Jain, P.W. Duin, J. Mao, *Statistical pattern recognition: a review*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(1):4–37, 2000.

[155] R. Duin, *Learned from neural networks*, In Proceedings of the 6th annual conference of the advanced school for computing and imaging (asci-2000), 9–13. Advanced School for Computing and Imaging (ASCI), 2000.

[156] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, John Wiley and Sons, Inc., 2004.

[157] A.Gionis, H.Mannila, T.Mielikainen, P. Tsaparas, *Assessing data mining results via swap randomization*, ACM Transactions on Knowledge Discovery from Data, 1 (3): 14, 2007.

SKRAĆENICE

ARFF	–	*Attribute-Relation File Format*
AODE	–	*Aggregating One-Dependence Estimators*
Avg(PK_U)	–	*Prosečan broj bodova ostvaren rešavanjem Moodle testova*
	–	
Avg(BNT)	–	*Prosečan broj bodova ostvaren na izdvojenim najboljim pokušajima rešavanja Moodle testova*
BB	–	*Bodovi za aktivnosti na predavanjima*
BLE	–	*Blended Learning Enivirоment*
BN	–	*Bayesian Network*
BNT	–	*Broj pokušaja sa najbolje ostvarenim rezultatima na Moodle testovima*
CART	–	*Classification and Regression Tree Algorithm*
CSV	–	*Comma-Separated Values*
CFS	–	*Correlation-Based Feature Selection*
CHI	–	*ChiSquaredAttributeEval*
DM	–	*Data Mining*
DZ	–	*Bodovi osvojeni rešavanjem domaćih zadataka u okviru Moodle kursa*
DT	–	*Decision Tree*
DS1, DS2, DS3	–	*Matrice rezulata rešavanja Moodle testova*
EDM	–	*Educational Data Mining*
EWB	–	*Equal-Width Binning*
FD	–	*Učešće u diskusijama na forumima Moodle kursa*
FP	–	*False positive rate*
FN	–	*False negative rate*
FM	–	*F- measeure*
FS	–	*Feature Selection*
GR	–	*Gain Ratio*
GI	–	*Gini Index*
HNB	–	*Hidden Näive Bayes*

HTML – *Hyper Text Markup Language*
IG – *Information Gain*
ID3 – *Iterative Dichotomiser 3 algorithm*
J48 – *J48 algorithm*
KD – *Knowledge Discovery*
K1_P, K2_P – *Probni testovi za prvi i drugi kolokvijum*
LAB – *Bodovi osvojeni na laboratorijskim vežbama*
LESS_AC – *Akcije u lekcijama kojima je student pristupio*
LMS – *Learning Management System*
LVT – *Upotreba video tutorijala*
MOODLE – *Modular Object Dynamic Learning Environment*
ML – *Machine Learning*
MLR – *Multiple Linear Regression*
MSE – *Mean Squared Error*
MDL – *Minimal Description Length Principle*
MI – *Mutual Information*
MV – *Majority Vote*
MM – *Upotreba Moodle poruka za elektronske konsultacije*
MMOU – *Metodologija mešanog okruženja učenja*
NB – *Näive Bayes*
NP – *Broj nezavršenih predatih pokušaja rešavanja Moodle testova*
OpenGL – *Open Graphics Library*
PDF – *Upotreba PDF materijala*
P1, P2, P3 – *Prosečni bodovi svih pokušaja rešavanja pripremnih testova*
P – *Precision*
PK_U – *Broj završenih Moodle testova*
R – *Recall*
RF – *Random forest*
RLF – *Relief filter method*
SummU – *Symmetrical Uncertainty Attribute Evaluation*
SMOTE – *Synthetic Minority Oversampling Technique*
SUAE – *SymmetricalUncertAttributeEval*
SSE – *Error Sum of Squares*

SVM – *Support Vector Machines*

TAN NB – *Tree Augmented Näive Bayes*

T1, T2, T3 – *Bodovi ostvareni na prvom i drugom kolokvijumu*

TP – *True positive rate*

TN – *True negative rate*

ZT_P – *Probni Moodle test za završni ispit*

WSE – *Wrapper Subset Evaluation*

SPISAK SLIKA

SPISAK TABELA

SPISAK RADOVA

a) Radovi štampani u međunarodnim časopisima sa SCI liste

[a1] **G. Dimić**, D. Rančić, I. Milentijević, P. Spalević, *Improvement of the Accuracy of Prediction Using Unsupervised Discretization Method: Educational Data Set Case Study*, Technical Gazette, Vol.25 No.2, str.407-414, April 2018, (SCIe, IF=0,686)

[a2] **G.Dimić**, B.Predić, D. Rančić, V.Petrović, N. Maček, P. Spalević, *Association analysis of moodle e-tests in blended learning educational environment*, Computer Applications in Engineering Education 26(3), Volume 26, Issue 3, December 2017, Pages 417-430, (SCIe, IF=1,153)

[a3] B. Predić, **G. Dimić**, D. Rančić, P. Štrbac, N.Maček, P. Spalević, *Improving final grade prediction accuracy in blended learning environment using voting ensembles*, Computer Applications in Engineering Education, July 2018 DOI:10.1002/cae.22042, (SCIe, IF=1,153)

[a4] **G. Dimić**, K.Kuk, P.Spalevic, Z. Trajcevski, Z. Todorovic, *Accuracy analysis of the classification model evaluation in the e-learning environment*, Journal of society for development of teaching and buisness processes in new net enviroment, BIH, ISSN 1840-1503, TTEM 2013, Vol.8, No.2, 2013, str.667-676 (SCIe, IF=0,351)

[a5] **G. Dimić**, K. Kuk, M. Zahorjanski, *Mining Student'S Data for Analyze Electronic Learning Materials Available on the Moodle Course*, REVISTA METALURGIA INTERNATIONAL, Romania, ISSN 1582-2214, (2011), vol.12 br.12, str. 78-82. (SCIe, IF=0,06)

[a6] K.Kuk, D. Prokin, **G. Dimić**, B. Stanojević, *Interactive Tasks as a Supplement to Educational Material in the Field of Programmable Logic Devices* - ELEKTRONIKA IR ELEKTROTECHNIKA, KTU Litvania, ISSN 1392 – 1215 , 2010. Vol 2(98), T120, pp 63-66, (SCIe, IF=0,659)

[a7] K.Kuk, D. Rančić, P.Spalević, **G. Dimić**, *The System for Keeping Records of Radio and TV Receivers based on the Java J2ME Platform*- ELEKTRONIKA IR ELEKTROTECHNIKA, KTU Litvania, ISSN 1392 – 1215 , 2010. Vol 10(106), T180, pp 121-124, (SCIe, IF=0,659)

b) Radovi štampani u časopisima nacionalnog značaja

[b1] **G.Dimić**, D. Rančić, I. Milentijević, P. Spalević, K. Plečić, *Comparative Study: Feature Selection Methods in the Blended Learning Environment*, Facta Universitatis, Series: Automatic Control and Robotics Vol. 16, No 2, 2017, pp. 95 – 116 , DOI: 10.22190/FUACR1702095D

[b2] K.Kuk, D. Prokin, **G. Dimić**, B. Stanojević, *New Approach in Realization of Laboratory Exercises in the Subject Programmable Logic Devices in the System for Electronic Learning – Moodle*, Facta Univ. Ser.: Elec. Energ., vol. 24, No. 3, April 2011, pp. 131-140.

c) Radovi prezentovani na konferencijama međunarodnog značaja

[c1] **G.Dimić**, J. Kajević, D. Rančić, P. Spalević, D. Milić, *Implementation of features selection methods and oversampling technique in blended learning environment*, Proceedings of the 26th International Electrotechnical and Computer Science Conference ERK'2017, Portoroz, Slovenija, 25.- 26. september 2017,pp.403-406

[c2] P.Milić, **G.Dimić**, D.Rančić, K.Kuk, *Linking e-learning systems with Youtube*, UNITECH 2015, Gabrovo, Bulgaria, 2015.

[c3] **G.Dimić**, D.Rančić, P.Spalević, K.Kuk, *Case study - Application of feature selection methods on a educational data set extracted from Moodle LMS system*, UNITECH 2014, Gabrovo, Bulgaria, 2014.

[c4] **G. Dimić**, D. Prokin, K. Kuk, P. Spalević, *The use of data mining methods for analyzing and evaluating course quality in the Moodle system*, UNITECH 2010, Bulgaria, Gabrovo, 2010.

[c5] K. Kuk, D. Prokin, **G.Dimić**, P. Spalević, *Learning unary logical operations through the modern interactive educational application - ARHICOMP*, UNITECH 2010, Gabrovo, Bulagria, 2010.

[c6] **G.Dimić**, D. Prokin, K. Kuk, P. Spalević. *Prediction of student's success analyzing their activities on the Moodle course*, Međunarodna konferencija"Matematičke i informacione tehnologije - MIT 2011, 27.avgust - 31. avgust 2011, V. Banja.

[c7] K. Kuk, **G.Dimić**, D. Prokin, P. Spalević. *Model for assessment of students' knowledge in an educational environment based on the game*, Međunarodna konferencija"Matematičke i informacione tehnologije" - MIT 2011, 27.avgust - 31. avgust 2011, V. Banja.

[c8] K. Kuk, D. Prokin, **G. Dimić**, P. Spalevic, *Interactive Tasks in Support to Modern Pedagogical Approach in Implementation of Teaching in the Subject Programmable Logic Devices*, UNITECH 2009, Gabrovo, Bulgaria, 2009.

[c9] P. Spalević, B. Milosević, K. Kuk, **G.Dimić**, *The Roles of Colours in the Multimedia Presentation Building*, ICEST '07, Ohrid, Macedonia. Jul 2007.

d) Radovi saopšteni na naučnim skupovima nacionalnog značaja

[d1] **G.Dimić**, D. Prokin, K. Kuk, B.Bogojević. Izbor klasifikatora za mali obučavajući skup obrazovnih podataka, Međunarodni naučno-stručni simpozijum INFOTEH 2013, 20. mart - 22. mart 2013, Jahorina, Vol. 12, March 2013.

[d2] D. Prokin**, G.Dimić**, K. Kuk, M.Prokin. Moodle kao platforma za realizaciju nastavnih aktivnosti iz predmeta Arhitektura i organizacija računara 1, Međunarodni naučno-stručni simpozijum INFOTEH 2012, 21. mart - 23. mart 2012, Jahorina, Vol. 11, March 2012.

[d3] **G.Dimić**, D. Prokin, K. Kuk, M.Micalović. Primena Decision Trees i Naive Bayes klasifikatora na skup podataka izdvojen iz Moodle kursa, Međunarodni naučno-stručni simpozijum INFOTEH 2012, 21. mart - 23. mart 2012, Jahorina, Vol. 11, March 2012.

[d4] **G. Dimić**, D. Prokin, K. Kuk, P. Spalević. Obrazovni data mining u sistemima za e-učenje, Međunarodni naučno-stručni simpozijum INFOTEH 2011, 16. mart - 18. mart 2011, Jahorina.

[d5] **G.Dimić**, P. Spalević, K. Kuk. Mining Student Data Using Clustering Expectation-Maximization Algorithm, ICEST 2011, Niš, Serbia. June 29 –July, 1 2011.

[d6] K. Kuk, D. Prokin, **G.Dimić**, P. Spalević. Primena interaktivnih zadataka u moodle okruženju na predmetu programabilna logička kola, Međunarodni naučno-stručni simpozijum INFOTEH 2011, 16. mart - 18. mart 2011, Jahorina.

[d7] **G.Dimić**, D. Prokin, K. Kuk, P. Spalević. Primena data mininga metoda za analizu i procenu kvaliteta kursa u Moodle sistemu, Elektronsko učenje na putu ka društvu znanja, Univerzitet Metropolitan SANU, Beograd, 16-17. Septembar 2010.

[d8] D. Prokin, K. Kuk, **G. Dimić**. Primena sistema Moodle u nastavi iz predmeta Programabilna logička kola, INFOTEH 2010, V.Banja, Međunarodna Konferencija i Izložba, 01-03. Jun 2010.

[d9] K. Kuk, D. Prokin, **G.Dimić**, P. Spalević. ARHICOMP – Interaktivna obrazovna igra za učenje unarnih logičkih operacija, Elektronsko učenje na putu ka društvu znanja, Univerzitet Metropolitan SANU, Beograd,16-17. Septembar 2010.

[d10] D. Prokin, K. Kuk, **G.Dimić**, Multimedijalne laboratorijske vežbe iz programabilnih logičkih kola, 53. Konferencija ETRAN 2009, V.Banja, 15-18 jun 2009.

[d11] P. Spalevic, Lj.Lazic, K. Kuk, **G. Dimic**, Ajax technologies of web site of higher school of electrical engineering in Belgrade, Naučno stručni skup - SNTPI ' 09, FIT, Beograd, 19-20 jun 2009.

[d12] **G. Dimić**, K. Kuk, I.Petrović, Unapređenje nastavnog procesa kao jedna od varijanti e-learning modela, Međunarodni naučno-stručni simpozijum INFOTEH 2008, 26-28 mart, Jahorina

[d13] K. Kuk, Lj.Lazić, **G.Dimić**, Poređenje metrike za veličinu softvera u modelima za procenu troškova i napora izrade web aplikacija (sajta), Međunarodni naučno-stručni simpozijum INFOTEH 2008, 26-28 mart 2008, Jahorina

[d14] K. Kuk, **G.Dimić**, M.Petrović, D.Jokanović, Poređenje tehnologija za izradu video tutorijala korišćenih u nastavi - iskustva autora, XIV naučno-stručna konferencija, YUINFO 2008, 9-12 mart 2008, Kopaonik

[d15] V. Stojanović, K. Kuk, **G.Dimić**, I.Petrović, Psihološki aspekti procesa učenja u multimedijalnim udžbenicima, Međunarodni naučno - stručni simpozijum INFOTEH 2007, 28-30 mart 2007, Jahorina.

[d16] Spalević P., Petrović M., **Dimić G**., Kuk K., Multimedijalna aplikacija za testiranje znanja studenata na Višoj elektrotehničkoj školi u Beogradu, 51. Konferencija ETRAN 2007, Herceg Novi – Igalo, Crna Gora, 4-8. jun 2007.

e) Pomoćni udžbenici

[e1] D.Prokin, M.Mijalković, G.Dimić, V.Korać, B.Bogojević, D.Popović, Priručnik za laboratorijske vežbe iz Arhitekture i organizacije računara, Visoka škola elektrotehnike i računarstva strukovnih studija, Beograd, 2018.

[e2] S. Obradović, B. Pavić, V. Petković, G. Dimić,MS Access- 2013 Projektovanje baza podataka i aplikacija, VIŠER, Beograd, 2015.

[e3] D.Starčević, K.Kuk, G.Dimić, M.Stupar, N.Vučković, Praktikum za laboratorijske vežbe iz Računarske grafike, Visoka škola elektrotehnike i računarstva strukovnih studija, Beograd, 2008.

[e4] D.Jokanović, G.Dimić, K.Kuk, Priručnik iz predmeta Digitalne multimedije 1, Viša elektrotehnička škola, Beograd, 2006.

BIOGRAFIJA

Ime i prezime: **Gabrijela Dimić**

Datum rođenja: 25.06.1971.

Adresa: Kaplara Momčila Gavrića 10, 11010 Beograd.

Kontakt telefon: 011/ 630-3176, 063 / 200-5134

E-mail**:** gabrijela.dimic@viser.edu.rs

Godina upisa osnovnih studija fakultet, smer, trajanje studija:

2003/2004, Tehničkom fakultetu „Mihajlo Pupin“, Zrenjanin, Univerzitet u Novom Sadu, Diplomirani inženjer informatike - petogodišnje akademske studije.

Godina završetka osnovnih studija:

2006/2007, Diplomirani inženjer informatike

Godina upisa doktorskih studija, fakultet, smer:

2007/2008, Elektronski fakultet Univerziteta u Nišu, Računarstvo i informatika

Znanje svetskih jezika: Engleski --- čita i piše.

Profesionalna orjentacija (oblast, uža oblast i uska orjentacija): Računarstvo i informatika

Pedagoško i radno iskustvo: Anagažovana u realizaciji nastave na Visokoj školi elektrotehnike i računarstva strukovnih studija u Beogradu kao asistent na predmetima Arhitektura i organizacija računara, Računarska grafika, Baze podataka, Relacione baze podataka, Standardni korisnički interfejsi, Meko računarstvo, Big data infrastrukture i sistemi.

Pored anagažovanja u nastavi, obavlja i poslove administratora informacionog sistema i radi kao tehnička podrška sajta visokoškolske ustanove u kojoj je zaposlena.

Učešća na projektima Ministarstva za nauku i zaštitu životne sredine:

Učesnik međunarodnog Tempus projekta "*Innovation and Implementation of the Curriculum Vocational Studies in the Field of Digital Television and Multimedia*", No. 517002-1-2011-1-RS-JPCR.

Ukupan broj naučnih radova: 34

Broj radova sa SCI liste: 7

Broj radova štampan u časopisima nacionalnog značaja: 2

Broj radova na međunarodnim konferencijama: 9

Broj radova na konferencijama nacionalnog značaja: 16