Primena Decision Trees i Naive Bayes klasifikatora na skup podataka izdvojen iz Moodle kursa

Gabrijela Dimić, Dragana Prokin, Kristijan Kuk
Visoka škola elektrotehnike i računarstva strukovnih studija
Beograd, Srbija
gdimic@viser.edu.rs, dprokin@viser.edu.rs,
kkristijan@viser.edu.rs

Marko Micalović
Univerzitet u Nišu
Elektronski fakultet
Niš, Srbija

Sadržaj— **Educational data mining obuhvata data mining metode koje se mogu primeniti u različitim oblastima obrazovanja za otkrivanje upotrebljivih obrazaca informacija iz postojećih podataka. Osnovna primena klasifikacijskih metoda u obrazovnom okruženja odnosi se na predviđanje uspešnosti studenata na završnom ispitu. U ovom radu analizirani su modeli predviđanja dva klasifikatora (Decision Tree i Naive Bayes). Izvršena je procena relevantnih atributa primenom filtera InfoGain i GainRatio, a estimacija klasifikatora realizovana je primenom metode unakrsne validacije. Za otklanjanje neizbalansirane distribucije vrednosti klasnih varijabli u generisanim modelima predviđanja korišćena je funkcija Resample i analizirani su modeli generisani nad re-sampled podacima.**

Ključne riječi: Klasifikacija; Data mining; Naive Bayes; Decision Tree; Predviđanje

I. UVOD

Educational Data Mining predstavlja noviju oblast naučnog istraživanja usmerenu na razvoj data mining metoda [1] za istraživanje obrazovnih podataka. Izdvajanje korisnih, upotrebljivih saznanja iz različitih izvora podataka obrazovnih okruženja, omogućava efikasnije razumevanje ponašanja studenata u procesu učenja [2]. Ključna primena *Educational Data Mining* metoda je kreiranje što preciznijih modela predviđanja karakteristika i akademskih performansi studenata. Veća preciznost omogućava identifikaciju studenata koji uče sporije i studenata koji će imati lošije rezultate na završnom ispitu.

Generalno, cilj predviđanja je kreiranje modela koji na osnovu kombinacije nezavisnih varijabli donosi zaključke o jedinstvenom aspektu zavisne varijable. Predviđanje zahteva postojanje oznaka za izlaznu varijablu na ograničenom skupu podataka, pri čemu oznake predstavljaju informacije o vrednosti izlazne promenljive u specifičnim slučajevima. Izbor varijabli iz raspoloživog skupa podataka utiče na preciznost i tačnost generisanih modela predviđanja. Zbog toga se u fazi pripreme podataka primenjuju različite tehnike za selekciju relevantnih varijabli i procena njihovog značaja na izlaznu varijablu predviđanja.

Predviđanje uspešnosti studenata u e-learning okruženju opisana je u radovima [3], [4], [5], [6], [7]. Analizirani su kursevi implementirani u e-learning okruženju, a izdvojeni podaci odnosili su se uglavnom na jednu generaciju studenata. Kao glavni atributi razmatrani su zadaci za vežbanje, aktivnosti studenata na kursu, demografski i anketni podaci. U [8] je prikazano poređenje dva klasifikatora (*Decision Trees* i *Bayes networks*) za predviđanje prosečne ocene studenata. Rezultati su pokazali da je klasifikator *Decision Trees* bio bolji od *Bayes networks* u svim klasama. Tačnost je dodatno poboljšana upotrebom tehnika ponovnog uzimanja uzorka posebno u slučaju primene *Decision Trees* za sve slučajeve klase. Pored toga smanjena je i pogrešna klasifikacija posebno u manjinskim klasama nebalansiranih skupova podataka jer *Decision Trees* algoritam teži da se fokusira na lokalni optimum. [9] prikazuje poređenje šest klasifikatora (*Naive Bayes*, *Decision Trees*, *Neural networks*, *Support vector machine, K – nearest neighbors, Logistic regression*) za predviđanje studenta koji odustaju u toku kursa. Izdvojeni atributi se odnose na demografske podatke, rezultate zadataka i učešće u grupnim sastancima. Najbolji klasifikatori, *Naive Bayes* i *Neural networks*, bili su u stanju da predvide oko 80% slučajeva odustajanja. Rezultati su pokazali da jednostavan *Naive Bayes* model dobro generalizuje mali skup podataka u odnosu na druge metode kao što su *Decision Trees* i *K – nearest neighbors* koji zahtevaju mnogo veći skup podataka za analizu.

U ovom radu opisani su klasifikatori *Naive Bayes* i *Decision Trees*. Klasifikatori su implementirani za kreiranje modela predviđanja uspeha studenata na prvom kolokvijumu e-learning kursa realizovanog u LMS (*Learning Management System*) Moodle [10] sistemu. Analiza i procena generisanih modela izvršena je primenom metode estimacije. Za primenu izabranih klasifikatora korišćeno je okruženje Weka [11] *data mining* alata.

II. *NAIVE BAYES* I *DECISION TREES* KLASIFIKATORI

Klasifikacija je zadatak *data mininga* kojim se vrši razdvajanje primeraka skupa podataka u unapred utvrđene klase izlazne varijable, a na osnovu vrednosti ulaznih varijabli [12].

Postupak klasifikacije podrazumeva sledeće korake:

- Izbor klasifikatora za primenu algoritma klasifikacije

Ovaj rad je realizovan pod delimičnim pokroviteljstvom Ministarstva za obrazovanje i nauku Republike Srbije, projekti br. TR32039 i III47016

- Izbor klasnog atributa (izlazna varijabla)
- Podela skupa podataka na dva dela: trenirajući (obučavajući) i test skup.
- Obučavanje klasifikatora na obučavajućem skupu kod koga su poznate vrednosti klasnog atributa
- Testiranje klasifikatora na test skupu kod koga su sakrivene vrednosti klasnog atributa

Testiranjem, klasifikator vrši klasifikaciju primeraka test skupa u unapred utvrđene klase klasnog atributa. Ukoliko klasifikator napravi previše grešaka u test podacima tj. ako je veći procenat nepravilno klasifikovanih primeraka, može se zaključiti da je kreiran loš i nestabilan model. U tom slučaju neophodno je izvršiti poboljšanje modela modifikacijom primenjenog procesa klasifikacije. Dosadašnja istraživanja pokazuju da najčešće primenjivani klasifikatori uključuju *Bayes networks*, *Decision Trees*, *Neural networks*, S*upport vector machine, K – nearest neighbors* [13] .

Za razliku od *Bayes networks* [14], klasifikator *Naive Bayes* [15] generiše model predviđanja sa snažnom nezavisnošću od pretpostavke, Predstavlja semantički jasan i jednostavan pristup prikazu, upotrebi i indukciji probabilističkog znanja. Nazvan je *"naivnim"*, jer pojednostavljuje problematiku oslanjajući se na dve važne pretpostavke: pretpostavlja da su atributi koji se koriste u predviđanju uslovno nezavisni uz poznatu klasifikaciju i da ne postoje sakriveni atributi koji bi mogli uticati na proces predikcije. Ove pretpostavke omogućavaju vrlo efikasne algoritme, kako za klasifikaciju tako i za mašinsko učenje. Za uslovno nezavisne attribute A_1, ..., A_k verovatnoća za klasni atribut C se izračunava po pravilu:

$$P(A_1,...A_k \mid C) = \prod_{i=1}^{k} P(A_i \mid C) \qquad (1)$$

Jednostavanost, efikasanost, laka interpretacija i pogodnost za male skupove podataka osnovne su prednosti *Naive Bayes* modela.. Međutim, u praksi, snažna nezavisnost od pretpostavke može da naruši međusobnu zavisnost atributa.

Decision Trees [12] su najpoznatije klasifikacijske tehnike, budući da uključuju više načina izgradnje lako interpretirajućih stabala koji služe za klasifikaciju kategorijskih i numeričkih vrednosti atributa. Ove metode klasifikacije vrše podelu podataka na čvorove i listove sve dok ceo skup podataka ne bude analiziran. Najpoznatiji algoritmi su *ID3* [16] i *C4.5* [17]. Osnovna ideja u ovim algoritmima je podela prostora atributa sve dok u svakom listu nije ispunjen kriterijum prekida, odnosno da sve tačke u listu pripadaju jednoj klasi. Ukoliko su podaci neusaglašeni, ispunjenje kriterijuma nije moguće. Rešenje ovog problema je u izboru najčešće klase među tačkama podataka u listu. Prednosti *Decision Trees* klasifikatora su: jednostavnost i lakoća razumevanja, mogućnost rada sa varijablama numeričkih i kategorijskih vrednosti, mogućnost brze klasifikacije novih primera i fleksibilnost.

III. METODE I MERE ESTIMACIJE KLASIFIKATORA

Estimacija klasifikatora omogućava predviđanje performansi, izbor najboljeg i ocene ponašanja više različitih klasifikatora. Za ocenu klasifikacijske tačnosti koriste se sledeće metode estimacije [18]:

- Metod testnog uzorka (*holdout method*)
- Metod unakrsnog ocenjivanja (*cross validation*)
- Metod početnog punjenja (*bootstramp*)

Metod testnog uzorka deli skup podataka na dva međusobno nezavisna podskupa – obučavajući i testni skup primeraka. Skup podataka se deli tako da se 2/3 primeraka koristi za učenje a 1/3 za ocenu tačnosti predviđanja. Estimacija tačnosti predviđanja je slučajan broj koji zavisi od podele primeraka na obučavajući i testni skup. Postupak se ponavlja *k* puta i tačnost se može dobiti kao srednja vrednost. Standardna devijacija se dobija kao standardna devijacija pojedinačnih ocena. Tačnost predviđanja naučenog skupa pravila povećava se sa povećavanjem broja primeraka u obučavajućem skupu. Izdvajanjem većeg broja primeraka za testiranje, povećava se pristrasnost estimacije, dok se smanjivanjem skupa primeraka proširuju granice intervala poverenja.

Metod unakrsnog ocenjivanja deli skup primeraka na *k* međusobno isključivih skupova približno iste veličine. Postupak učenja i ocenjivanja se ponavlja *k* puta , svaki put koristeći jedan podskup kao testni skup. Estimacija tačnosti predviđanja unakrsnim ocenjivanjem je slučajan broj koji zavisi od podele primeraka na podskupove.

Osnovni pojam pri proceni tačnosti klasifikacijskih modela je pojam greške. Greškom se smatra pogrešna klasifikacija: primena klasifikacijskog modela na odabrani primerak rezultira prognozom pripadnosti klasi koja je različita od stvarne klase primerka. Ukoliko se svakoj grešci pridaje jednak značaj, tada je ukupan broj grešaka na posmatranom skupu primeraka dobar indikator uspeha/neuspeha klasifikacijskog modela. Tačnost predviđanja je osnovni pokazatelj performansi kreiranog klasifikacijskog modela, koji predstavlja procenat pravilno klasifikovanih, novih primeraka korišćenjem naučenih pravila. Dva osnovna nedostatka mere tačnosti odnose se na zanemarivanje razlika između tipova grešaka i zavisnosti od distribucije klasa u skupu podataka, a ne od karakteristika primeraka. Naime, u većini praktičnih situacija vrlo je važno razlikovati određene tipove grešaka.

U slučajevima kada je potrebno razlikovati više tipova grešaka rezultat klasifikacije prikazuje se u obliku dvodimenzionalne matrice grešaka. Svaki red matrice odgovara jednoj klasi i sadrži broj primeraka kojima je to prognozirana klasa. Svaka kolona matrice označena je sa jednom klasom i sadrži broj primeraka kojima je to stvarna klasa. Broj tačno klasifikovanih instanci nalazi se na dijagonali matrice. Svi ostali elementi matrice označavaju broj primeraka koji su neispravno klasifikovani kao neka od preostalih klasa.

A. Standardne mere za analizu tačnosti

Najveći broj mera za procenu klasifikacijskih modela odnosi se na klasifikacijske probleme s dve klase. To ne

predstavlja posebno ograničenje za upotrebljivost tih mera, budući da se problemi s većim brojem klasa mogu prikazati u obliku niza problema s dve klase. Svaki od njih posebno izdvaja jednu od klasa kao ciljnu, a skup podataka se deli na pozitivne i negativne primerke ciljne klase, s tim da u negativne spadaju primerci svih ostalih klasa.

Postoje četiri moguća rezultata predviđanja:

- Stvarno pozitivni (*TP –True Positive Rate*) i stvarno negativni (*TN -True Negative Rate*) ishodi predstavljaju ispravnu klasifikaciju.
- Lažno pozitivni (*FP-False Positive Rate*) i lažno negativni (*FN-False Negative Rate*) ishodi predstavljaju dva moguća tipa greške.

Senzitivnost i specifičnost spadaju u mere za evaluaciju klasifikacijskih modela koje razlikuju dva spomenuta tipa greške. Zasnivaju se na istom principu kao i tačnost: senzitivnost meri tačnost u pozitivnim primercima, a specifičnost predstavlja tačnost u negativnim primercima.

$$Senzitivnost = \frac{TP}{TP + FN} \tag{2}$$

$$Specifičnost = \frac{TN}{TN + FP} \tag{3}$$

Odziv i preciznost su drugi par mera za evaluaciju klasifikacijskih modela koji je osjetljiv na različite tipove grešaka. Najčešće se koristi u problemima kod kojih je broj stvarno pozitivnih dosta mali u odnosu na broj stvarno negativnih primeraka tj. $TP << TN$. Odziv i preciznost se definišu izrazima:

$$Odziv = \frac{TP}{TP + FN} \tag{4}$$

$$\Pr eciznost = \frac{TP}{TP + FP} \tag{5}$$

Odziv je definisan kao tačnost u pozitivnim primercima, preciznost je tačnost u pozitivnoj prognozi ciljne klase. U kontekstu problema pronalaženja informacija, odziv predstavlja odnos pronađenih relevantnih dokumenata u odnosu na ukupan broj relevantnih dokumenata, dok je preciznost udio relevantnih dokumenata u ukupnom broju pronađenih dokumenata. Pozitivna i negativna prediktivna vrednost su još jedan par mera koji uvažava različite tipove klasifikacijskih grešaka. Zasnivaju se na principima preciznosti tj. pozitivna prediktivna vrednost predstavlja preciznost u pozitivnim primercima, dok je negativna prediktivna vrednost preciznost u negativnim primercima. Osim navedenih mera, postoje i mere koje se ne zasnivaju na fiksiranju jedne komponente mera.

U takve spada i F-mera koja se definiše sa:

$$F - mera = \frac{2 \times odziv \times preciznost}{odziv + preciznost} \tag{6}$$

B. ROC analiza

Kvalitet klasifikacijskog modela izražen jednim brojem ili parom brojeva ne može u potpunosti ilustrovati varijabilnost rešenja koja se mogu izvesti upotrebom odabrane tehnike modeliranja. Međusobne zavisnosti različitih mera i/ili parametara bolje se mogu opisati upotrebom grafova kojima je moguće prikazati velik raspon različitih mogućnosti. Jedna od grafički orijentisanih mera klasifikacijskog modela je *ROC* analiza [19] preuzeta iz oblasti detekcije signala (*Receiver Operating Characteristic*). Za razumevanje *ROC* analize najvažniji je pojam graničnih vrednosti pouzdanosti određenog klasifikatora. Senzitivnost i specifičnost su dve mere koje karakterišu specifičnu tačnost klasifikatora u pozitivnim i negativnim primercima. Kada je granična vrednost pouzdanosti postavljena visoko, tj. naglašen kriterijum pouzdanosti klasifikatora, senzitivnost klasifikatora će biti niska, a specifičnost visoka. Ako se kriterijum pouzdanosti snizi, senzitivnost će porasti, a specifičnost opadati. Na taj način, dva klasifikacijska modela mogu se uporediti preko širokog spektra pouzdanosti, generišući krivu koja opisuje zavisnost broja stvarno pozitivnih primeraka od broja lažno pozitivnih primeraka detektovanih modelom. Ukoliko se sa T = (x, y) označi tačka na *ROC* krivoj posmatranog klasifikatora, vrednosti koordinata x i y izračunavaju se pomoću izraza :

$$x = \frac{FP}{FP + TN} \tag{7}$$

$$y = \frac{TP}{TP + FN} \tag{8}$$

Dakle, vertikalna osa *ROC* krive označava broj stvarno pozitivnih primeraka izražen kao udeo u ukupnom broju pozitivnih primeraka. S druge strane, horizontalna osa *ROC* krive označava odgovarajući broj lažno pozitivnih primeraka izražen kao udeo u ukupnom broju negativnih primeraka. Oblik *ROC* krive nekog klasifikatora zavisi od skupa primeraka koji je analiziran.

IV. STUDIJA SLUČAJA

U zimskom semestru školske 2011/12. - te godine na Visokoj školi elektrotehnike i računarstva strukovnih studija u Beogradu, tradicionalan način realizacije nastave (predavanja i laboratorijske vežbe) iz predmeta Arhitektura i organizacija računara 1 pratio je *on-line* kurs na *open-source* LMS (*Learning Management System*) sistemu. Kurs je kreiran na Moodle platformi i korišćen je za obezbeđivanje različitih interaktivnih resursa za učenje, lakše komunikacije sa studentima putem foruma, kao i za realizaciju ispitne provere znanja. Kurs nije kreiran za koncept daljinskog učenja, već u cilju poboljšavanja efikasnosti klasičnog načina održavanja nastave. Pristup kursu, izbor aktivnosti, upotreba resursa, vreme pristupa, trajanje pristupa i ostale akcije u potpunosti su bile indivudualno određene od strane svakog studenata kursa. Opisani način organizacije *on-line* kursa, nametnuo je potrebu za otkrivanjem povezanosti između karakterističnih aktivnosti studenata i ostvarenog uspeha na ispitu.

Razvijen je model predviđanja uspešnosti studenata na prvom kolokvijumu uzimajući u obzir promenljive koje se

odnose na aktivnosti studenata u okviru posmatranog kursa. Modeli su kreirani implementacijom *Naive Bayes* i *Decision Trees* klasifikatora i analiziran je uticaj ulaznih atributa na performanse modela. Primenom odgovarajućih tehnika postignuta je veća tačnost generisanih modela.

V. IZDVAJANJE I PRIPREMA PODATAKA

Iz baze podataka LMS Moodle sistema, izdvojeno je 222 zapisa za analizu. Odbačeni su zapisi za studente koji su samo upisani na kurs a nisu imali nikakve registrovane aktivnosti. Oiginalnog vektora je sadržio 51 karakteristiku aktivnosti studenata, od kojih su izdvojeni relevantni ulazni prediktivni atributi i kreirane smislene grupe u okviru njih. Primenom filtera diskretizacije [20], numeričke vrednosti atributa *Prvi kolokvijum* transformisane su u nominalne i utvrđena su četri intervala (*pali, slabi, dobri, vrlo_dobri).* Atribut *Prvi kolokvijum* označen je kao klasni odnosno kao odzivna promenljiva. Oznake i opis atributa, izdvojenog skupa podataka, prikazani su u Tabeli 1.

TABELA I. ATRIBUTI IZDVOJENOG SKUPA PODATAKA

Atribut	Opis
Lekcije	Moodle lekcije – Nisu bodovane
DZ1 DZ2 DZ3	Testovi sa slučajnim izborom pitanja i zadataka koji su bodovani od 0 do 2 boda.
Pripremni	Test za pripremu pred kolokvijum. Bodovi od 0 do 35.
Prvi kolokvijum	Test se radi za računarom u laboratoriji škole. Bodovi od 0 do 35.

Nakon što su podaci prikupljeni i izdvojeni u jedinstvenu relaciju, potrebno je bilo steći uvid u njihovu strukturu i informativnu vrednost kako bi se što bolje pripremili za primenu data mining algoritama i metoda. U tu svrhu, koriste se jednostavne statističke tehnike i alati za vizualizaciju. Izrađuju se histogrami distribucije vrednosti nominalnih atributa, ispituje distribucija numeričkih atributa, grafički prikazuju vrednosti pojedinih atributa u odnosu na klasu ili neki drugi atribut. Pomoću ovakvih tehnika mogu se uočiti neočekivani podaci koji se razlikuju od obrasca koji se provlači kroz ostatak vrednosti, neodređene vrednosti atributa, duplikati u podacima itd.

U fazi pripreme podataka izvršena je procena vrednosti ulaznih atributa u odnosu na klasni atribut. Cilj postupka procene i izbora atributa je izdvajanje nevažnih i redundantnih atributa iz skupa podataka za učenje. Kako bi se u skupu podataka za učenje zadržali samo poželjni atributi, potrebna je neka mera vrednosti atributa u odnosu na klasifikacijski problem. Metode filtera obuhvataju tehnike za ocenjivanje vrednosti atributa oslanjajući se na heuristike zasnovane na opštim svojstvima podataka [12]. Za inteligentnu analizu podataka metode filtera su praktičnije rešenje iz sledećih razloga: izbor i ocenjivanje atributa znatno kraće, nezavisnost od algoritma mašinskog učenja omogućava primenu u kombinaciji s bilo kojom tehnikom modeliranja podataka.

Primenjene su metode filtera *InfoGain* i *GainRatio* sa *Ranker* metodom pretrage. *InfoGain* predstavlja procenu vrednosti atributa merenjem njegove informativnosti u odnosu na klasu. *GainRatio* predstavlja procenu vrednosti atributa merenjem relativne informativnosti u odnosu na klasu. Atributi sa procenom manjom od 0.01 treba da budu isključeni iz analiziranog skupa podataka. Rezultati procene i rangiranja atributa na osnovu njihovih individualnih vrednosti prikazani su u Tabeli 2.

TABELA II. REZULTATI PROCENE I RANGIRANJA ATRIBUTA

Atributi	*Information Gain*	*Gain Ratio*
Lekcije	0.0635	0.0718
DZ1	**0.4769**	0.1526
DZ2	0.4514	**0.1595**
DZ3	0.3687	0.134
Pripremni	**1.2914**	**0.2066**

Iz Tabele 2 može se videti da su atributi sa najvećim rangom *Pripremni, DZ1 i DZ2*, a sa najmanjim rangom atribut *Lekcije*. S obzirom da su vrednosti najmanjeg ranga veće od vrednosti 0.01 (0.0635 > 0.01, 0.0718 > 0.01) atribut *Lekcije* nije isključen iz analiziranog skupa podataka.

VI. ANALIZA KLASIFIKACIJSKIH MODELA ZA PREDVIĐANJE USPEHA STUDENATA

U ovoj studiji slučaja implementirani su klasifikatori *Naive Bayes* i *J48 Decision Trees* za izgradnju modela predviđanja. Metoda unakrsnog ocenjivanja je primenjena za estimaciju modela. Iz Tabele 3 može se videti da je primenom *Naive Bayes* klasifikatora kreiran tačniji model predviđanja.

TABELA III. POREĐENJE KLASIFIKATORA

	CCI (%)	*ICI (%)*	*TP*	*FP*	*P*	*R*	*FM*	*ROC*
NB	43.24	56.76	0.43	0.20	0.39	0.43	0.39	0.66
DT	35.59	64.41	0.36	0.35	0.13	0.36	0.19	0.48

Naive Bayes klasifikator generisao je model sa 43.24% pravilno klasifikovanih primera (*CCI*), preciznošću od 39% (0.39) i klasifikacijom iznad oblasti *ROC* krive (0.66 > 0.5). Za *Naive Bayes* klasifikator generisana je matrica grešaka (Tabela 4). Četri slučaja nominalnog klasnog atributa *Prvi kolokvijum* označeni su slovima *a=pali*, *b=slabi*, *c=dobri*, *d=vrlo dobri*. Broj tačno klasifikovanih primera nalazi se na dijagonali matrice, a ostali elementi matrice označavaju broj nepravilno klasifikovanih primera kao neka od preostalih klasa.

Iz Tabela 4 može se videti da postoji disbalans u raspodeli vrednosti izlazne klase i da je tačnost kod manjih klasa manja nego tačnost kod većih klasa. Tako na primer, broj studenata predviđenih da urade slabo kolokvijum (*klasa b*) je čak 52 puta manji od broja studenta predviđenih da urade vrlo dobro kolokvijum (*klasa d*), a tačnost za ove klase je 2.6% i 65.8% respektivno.

TABELA IV. MATRICA GREŠAKA NAIVE BAYES KLASIFIKATORA

Predviđajuće klase				
a	***b***	***c***	***d***	***Stvarne klase***
29	2	5	3	***a***
18	**1**	10	10	***b***
18	6	**14**	27	***c***
15	4	8	**52**	***d***
74.4	**2.6**	**21.5**	**65.8**	%

U slučaju nebalansiranog skupa podataka, primeri malih klasa se teže obučavaju (uče). Za rešenje ovog problema korišćena je funkcija *Resample* za ponovno uzimanja uzoraka podataka. Na taj način poboljšana je balansiranost u disitribuciji podataka, što je uticalo na rezultate primenjenih algoritama. Tabela 5 prikazuje tačnost predviđanja klasifikatora na *re-sampled* podacima.

TABELA V. TAČNOST PREDVIĐANJA NA RE-SAMPLED PODACIMA

	CCI (%)	***ICI (%)***	***TP***	***FP***	***P***	***R***	***FM***	***ROC***
NB	63.51	36.49	0.64	0.14	0.67	0.64	0.63	0.86
DT	67.57	32.43	0.68	0.15	0.7	0.68	0.67	0.87

Analizom podataka može se utvrditi da je tačnost predviđanja na izbalanisiranim podacima, u slučaju primene oba klasifikatora, značajno poboljšana, a da je klasifikator stabla odluke kreirao precizniji model. Detaljnija analiza distribucije *re-sampled* podataka prikazana je u matrici grešaka (Tabela 6).

TABELA VI. MATRICA GREŠAKA RE-SAMPLED PODATAKA

Predviđajuće klase								
J48 Decision Trees				***Naive Bayes***				
a	***b***	***c***	***d***	***a***	***b***	***c***	***d***	***Stvarne klase***
20	2	5	9	**28**	0	6	2	***a***
2	**19**	4	14	7	**16**	8	8	***b***
1	1	**44**	22	13	1	**37**	17	***c***
2	3	7	**67**	11	2	6	**60**	***d***
55.6	**48.7**	**64.7**	**84.8**	**77.8**	**41**	**54.4**	**75.9**	%

Iz tabele 6 se vidi da je u slučaju klasifikatora *J48* veća tačnost predviđanja po klasama, osim za *klasu a.* Pored toga, manjim klasama odgovara i manja tačnost predviđanja. Za klasifikator *Naive Bayes* tabela 6 prikazuje neusaglašenost između brojnosti klase i njene tačnosti predviđanja. Na primer, u *klasi a* klasifikovano je 28 primeraka stvarne *klase a* sa tačnošću predviđanja od 77.8%, dok je u *klasi d* klasifikovano 60 primeraka stvarne *klase d* ali je tačnost predviđanja 75.9% što je manje od tačnosti manje brojne *klase a.* Takođe, može se uočiti da je najveća tačnost u slučaju klasifikatora *J48* postignuta kod *klase d,* a u slučaju *Naive Bayes* klasifikatora kod *klase a*.

VII. ZAKLJUČAK

Izbor klasifikatora za generisanje preciznih modela za predviđanja uspešnosti studenata u okruženju e-learning sistema predstavlja složen zadatak. Pre izbora tehnike modelovanja treba znati kakvi su podaci na raspolaganju. U slučaju realnih sistema česta je pojava neočekivanih podataka (outliers) koji značajno odstupaju od većine i ne uklapaju se u isti model sa ostalim vrednostima. Uzroci lošeg kvaliteta podataka je pojava duplikata u zapisima, tipografske grešake i neažurirannje podataka. Stvarni podaci često sadrže neodređene vrednosti atributa (noise) ili zapise kod kojih vrednosti nekih atributa nedostaju (missing values). Obrazovni podaci su poprilično čisti, ne sadrže pogrešne vrednosti u velikoj meri, jer se sakupljaju automatski (log fajlovi, baze podataka studentskih ocena, baze podataka sistema za učenje). Vrednosti ovih podataka mogu biti numeričke, kategorijske, diskretne (bodovi, ocene, opisne skale ocenjivanja, binarne vrednosti). Pojava neočekivanih podataka, kod obrazovnih podataka je posledica šuma. Konkretno, uvek postoje izuzeci među studentima koji ostvare dobre rezultate na ispitu sa manje napora i rada ili oni koji ne polože ispit uprkos svim očekivanjima.

U radu je opisana analiza upotrebe klasifikatora *Decision Trees* i *Naive Bayes* za izgradnju modela predviđanja uspeha studenata na prvom kolokvijumu realizovanog u okruženju Moodle kursa. Razmatrane su standardne mere procene tačnosti modela (broj pravilno i nepravilno klasifikovanih primeraka, TP, FP, Preciznost, Odziv, F-mera, oblast *ROC* krive).

Za detaljniju analizu distribucije primeraka klasnog atributa korišćena je matrica grešaka. Prvobitno generisani modeli pokazali su neizbalansiranu raspodelu primera po klasnim varijablama što nam je ukazalo da podaci nisu dobro pripremljeni. S obzirom da su analizirani podaci bili nominalnog tipa primenjen je *Resample* Weka filter za učenje pod nadzorom. Izvršeno je ponovno uzimanje uzoraka iz skupa podataka sa ili bez zamene instanci. Rezultat primene ovog filtera uslovio je da se različitost u vrednostima nominalnih atributa (*DZ1, DZ2; DZ3, Pripremni*) smanjila. Pre primene *Resample* filtera atribut *DZ1* sadržio je 33 različite vrednosti, *DZ2* - 31 različitu vrednost, *DZ3* – 32 različite vrednosti, a atribut *Pripremni* čak 135 različitih vrednosti. Kada su podaci uzorkovani broj različitih vrednosti po atributima je smanjen (*DZ1* – 23, *DZ2* -21, *DZ3*- 22, *Pripremni* – 95). Generisani su modeli sa *re-sampled* podacima koji su omogućili veću preciznost predviđanja. Poređenje klasifikatora (Tabela 5) ukazuje da u ovom slučaju model veće tačnosti kreira klasifikator *J48 Decision Trees*. Međutim, uzimajući u obzir analizu distribucije podataka po klasama u matrici grešaka (Tabela 6) možemo zaključiti da i model *Naive Bayes* klasifikatora nije dao loše rezultate.

U cilju budućeg istraživanja planirano je detaljnije istraživanje postupka izdvajanja, selekcije, transformacije atributa kao i mogućnost generisanja modela predviđanja kombinovanjem najbolje ocenjenih klasifikatora.

LITERATURA

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

[2] E. Baker, "International Encyclopedia of Education", 3rd ed., Oxford, UK: Elsevier.

[3] 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, pp. 445–449, 2003.

[4] V.P.Bresfelean, M.Bresfelean, N.Ghisoiu and C.A.Comes, "Determining students academic failure profile founded on data mining methods", In Proceedings of the 30th international conference on information technology interfaces, pp.317–322, 2008.,

[5] B.Minaei-Bidgoli, D.A.Kashy, G.Kortemeyer and W. Punch, "Predicting student performance: an application of data mining methods with an educational web-based system", In Proceedings of 33rd frontiers in education conference, T2A–13–T2A18, 2003.

[6] C. Romero, S. Ventura, P.G.Espejo and C.Hervas, "Data mining algorithms to classify students", Proceedings of the 1st international conference on Educational data mining, pp.8–17, 2008.

[7] M. Mühlenbrock, "Automatic action analysis in an interactive learning environment", In Proceedings of the workshop on usage analysis in learning systems, pp.73–80, 2005.

[8] T. N.Nguyen, P.Janecek and P.Haddawy, "A comparative analysis of techniques for predicting academic performance" ASEE/IEEE Frontiers in Education Conference, pp. T2G7-T2G12, 2007.

[9] S. B. Kotsiantis, C. J. Pierrakeas and P.E.Pintelas, "Preventing student dropout in distance learning using machine learning tachniques", In proceedings of 7th International Conference on Knowledge-Based Intelligent Information and Engineering Systems, pp. 267- 274, 2003.

[10] Moodle, http://moodle.org/

[11] Weka, University of Waikato, New Zealand, http://www.cs.waikato.ac.nz/ml/weka

[12] H. Witten and F.Eibe, "Data mining: Practical machine learning tools and techniques", 2nd ed., Morgan Kaufmann, 2005.

[13] G. Benoît, "Data Mining", Annual Review of Information Science and Technology, Vol. 36, pp. 265-310, 2002.

[14] J. Pearl, "Probabilistic reasoning in intelligent systems: networks of plausible inference", Morgan Kaufman Publishers, 1988.

[15] Z.Harry, "The Optimality of Naive Bayes", FLAIRS conference, 2004.

[16] J. R.Quinlan, "Induction of decision trees-Machine Learning", pp.81-106. , 1986.

[17] J. R. Quinlan, "C4.5: Programs for Machine Learning", San Mateo, Morgan Kaufmann, 1993.

[18] R. Kohavi, "A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection", In.Proc.International Joint Conference on Artificial Intelligence, 1995.

[19] F. Provost and T. Fawcett, "Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions", In Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, AAAI Press, 1997.

[20] G. Dimic, D. Prokin, K. Kuk, P. Spalevic, " The Use of Data Mining Methods for Analyzing and Evaluating Course Quality in the Moodle System", Международна научна конференция "УНИТЕХ'10" – Габрово, P. 309-315, 2010.

ABSTRACT

Educational data mining involves data mining methods that can be applied in different areas for discover usable patterns from existing data. The main application of classification methods in the educational environment refers to the prediction of student performance on the final exam. This paper describes analysis of the process for creating model predictions two classifiers (Decision Tree and Naive Bayes). For estimation of the relevant attributes we were used filters GainRatio and InfoGain. Estimation of classifiers was carried out using the method of cross validation. For elimination an unbalanced distribution values of class variables we were used Resample function and generated new prediction models.

APPLICATION OF DECISION TREES AND NAÏVE BAYES CLASSIFIER ON THE DATA SET EXTRACTED FROM THE MOODLE COURSE

Gabrijela Dimić, Dragana Prokin,
Kristijan Kuk, Marko Micalović