UDC: 004.82 INFO M: str. 26-34

DATA MINING U SISTEMU ELEKTRONSKOG OBRAZOVANJA
DATA MINING IN E-EDUCATION SYSTEM

Zorica Bogdanović, Marijana Despotović, Božidar Radenković

REZIME: U ovom radu se razmatra data mining kao deo procesa otkrivanja znanja u aplikacijama elektronskog poslovanja. Opisane su mogućnosti primene data mininga u e-poslovanju, problemi koji tom prilikom mogu nastati, kao i trendovi razvoja. Data mining se može primeniti u svim oblastima elektronskog poslovanja. Detaljno su opisane mogućnosti primene u e-obrazovanju. Dat je primer primene data mining algoritama u sistemu obrazovanja na daljinu na poslediplomskim studijama Fakulteta organizacionih nauka.
KLJUČNE REČI: data mining, e-obrazovanje, e-poslovanje, poslovna inteligencija
ABSTRACT: This paper presents data mining as a part of a knowledge discovery process in e-business applications. Possibilities, challenges and problems of using data mining in e-business applications are described as well as trends of future development. Data mining can be applied in various areas of e-business. Application in e-education is described in detail. An example of data mining application in distance education system on postgraduate studies on Faculty of Organizational Sciences is given, too.
KEY WORDS: data mining, e-education, e-business, business intelligence

1. UVOD

Elektronsko poslovanje stvara sve više podataka koji se tiču ponašanja klijenata, kao i njihovih navika pri krstarenju Internetom. Prikupljeni podaci mogu se obraditi kako bi se dobile neke korisne informacije. Na primer, u sistemu obrazovanja na daljunu primena data mining-a može obezbediti informacije o ponašanju studenata. Dalje, moguće je kreirati profil korisnika, radi boljeg upravljanja odnosima sa klijentima ili usmerenog reklamiranja na Web-u. Kompanije koje se bave elektronskim poslovanjem mogu poboljšati prodaju ili kvalitet proizvoda, tako što će naslutiti probleme pre nego što oni nastanu. Data mining predstavlja dobijanje implicitnih, prethodno nepoznatih, važnih i potencijalno korisnih informacija iz prikupljenih podataka.

2. PROCES DATA MINING-A U WEB OKRUŽENJU

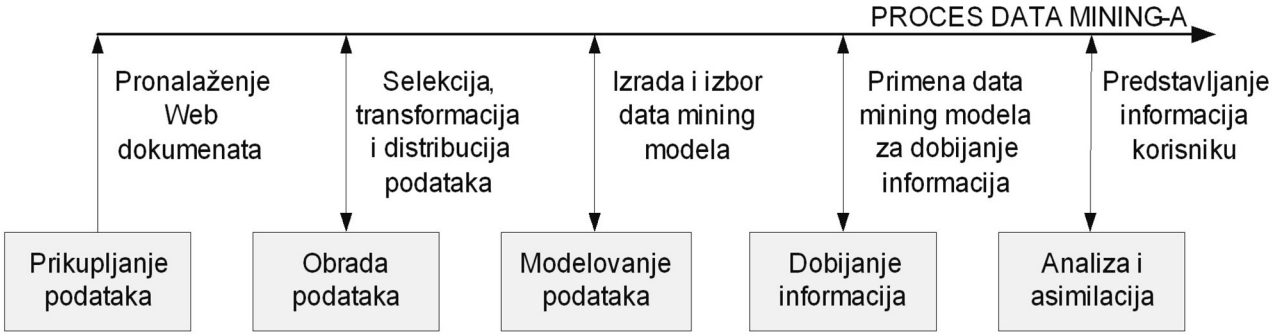
Data mining proces se može podeliti na mnoge podprocese u slučaju Web okruženja. Proces počinje pronalaženjem i preuzimanjem željenih Web dokumenata ili logova. Sledeći i najvažniji korak je analiza podataka dobijenih iz Web dokumenata. Ovo uključuje predprocesiranje, sam data mining proces i asimilaciju znanja. Na kraju, dobijene informacije se prikazuju korisniku u željenom obliku. Na primer, iz informacija se može videti koliko je Web sajt kompanije bitan pri donošenju odluka i da li ga je potrebno menjati. Analiza može ukazati na strategije za privlačenje novih kupaca i zadržavanje starih. Proces data mininga prikazan je na slici 1 [4].

Zavisno od ciljeva krajnjeg korisnika, proces data mining-a može se obavljati sa tri različita postupka: modeliranje predviđanjem, klasterovanje i link analiza [2].

Modeliranje predviđanjem je postupak u kome se izvode predviđanja na osnovu glavnih karakteristika podataka. Ovo se postiže klasifikacionim i regresionim postupcima data mining-a. Klasifikacijom se izgrađuje model kojim se mapiraju (klasifikuju) podaci u unapred definisane klase. Regresijom se za svaki podatak vezuje po jedna realna predviđena promenljiva. Glavni cilj oba postupka je da se napravi predviđanje važnih promenljivih.

Klasterovanje. Cilj klasterovanja je pronalaženje elemenata sa sličnim karakteristikama i pravljenje hijerarhije klasa od postojećeg skupa. Set podataka je particionisan na grupe elemenata koji imaju neka zajednička svojstva. Elementi unutar jedne grupe (klastera) su slični jedan drugom, dok su elementi iz različitih klastera međusobno različiti.

Link analiza. Cilj ovog postupka je da uspostavi interne odnose između elemenata u konkretnom setu, što se postiže otkrivanjem sličnosti i sekvencnih šablona. Otkrivanje aso-



Slika 1. – *Proces data mining-a*

cijacija se svodi na izgradnju modela koji traži elemente koji ukazuju na prisustvo drugih elemenata u setu sa izvesnim stepenom pouzdanosti. Na ovaj način otkriva se skrivena povezanost elemenata.

3. DATA MINING U OBLASTI ELEKTRONSKOG WEB POSLOVANJA

Jedan od glavnih izazova u elektronskom poslovanju je kako naći način da se razume ponašanje kupaca na osnovu podataka sa Web sajta. Posmatranje ponašanja klijenata će omogućiti da se predvidi njihovo ponašanje u budućnosti. Analiza podataka može se vršiti klasičnim statističkim pristupom, koji je loš zbog velike količine podataka. Sa druge strane, data mining proces se obavlja na čitavom skupu podataka, a informacije se dobijaju sa bogatim, detaljnim opisima koji omogućavaju da se pronađu veze koje nisu očigledne.

U suštini, podaci koji se dobijaju iz elektronskog poslovanja su:

1. primarni podaci koji uključuju konkretan Web sadržaj
2. sekundarni podaci poput logova sa servera ili browe-ser-a, podataka o registraciji i ako postoje, korisničkih sesija, upita, kolačića, itd.

Data mining primarnih Web podataka

Obrada primarnih podataka može pozitivno uticati na povećanje preciznosti povratnih informacija. Osnovna kategorizacija, klasterovanje, asocijaciona analiza i tehnike za predviđanje trendova, mogu da se na prikupljenim informacijama za bolju organizaciju. Neke od pogodnih aplikacija za takve tipove podataka su:

- Predviđanje nekih veličina u budućnosti na osnovu trendova u prikupljenim podacima.
- Primena klasifikacije teksta na prikupljenim informacijama, radi boljeg razumevanja.
- Primena asocijacione analize kod posmatranja Web sajtova konkurenata.
- Otkrivanje sličnosti i odnosa između različitih Web sajtova da bi se klasifikovale Web strane.
- Korišćenje Web upitnih jezika za bolju organizaciju nestrukturiranih podataka sa Web-a.

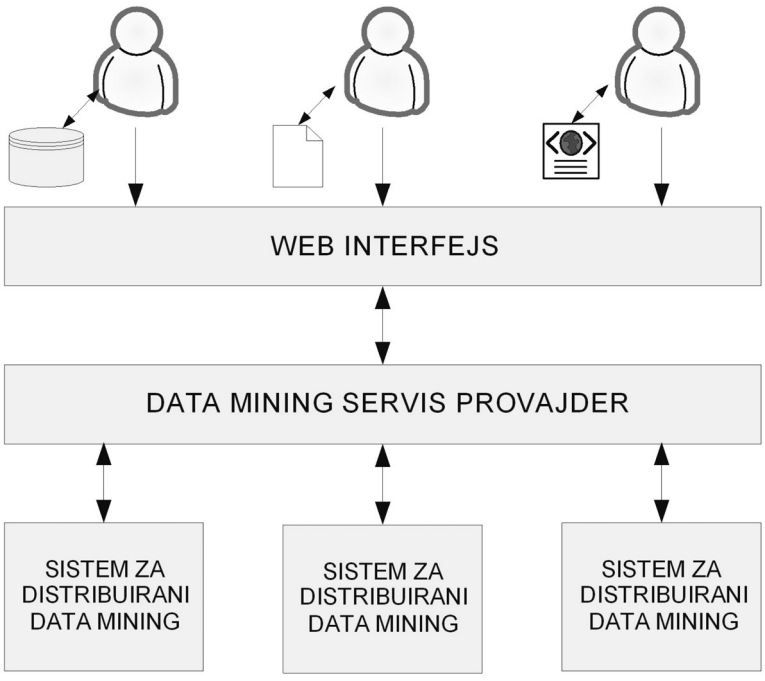
Data mining sekundarnih Web podataka

Sekundarni podaci su podaci o transakcijama koji se vade iz Web logova. Cilj je da se pronikne u kupovne navike potrošača u Web okruženju. Bilo koji metod prepoznavanja šablona, poput klasifikacije ili klasterovanja, može da se koristi i ovde, posle predprocesiranja podataka. Neke od aplikacija u ovoj grupi su:

- Unapređenje usluga analizom logova , što može pokazati navike klijenata.
- Održavanje i izmena Web sajtova kako bi bolje služili potrebama klijenata
- Personalizacija Web sajtova kako bi se prilagodili svačijem ukusu.

3.1. Distribuirani data mining

Okruženje u kome se obavlja data mining se satoji od korisnika, podataka, hardvera i softvera (ovde se misli na algoritme i sve neophodne programe). Distribuirani data mining (DDM) se opisuje kao proces obrade podataka koji se nalaze na više fizički, tj. geografski razdvojenih lokacija. Drugim rečima, u pitanju je proces koji se odvija nad distribuiranom bazom podataka (mada se ovde termin distribuiran koristi u širem smislu, kako bi se obuhvatili svi oblici homogenosti i heterogenosti (slika 2) [3].



Slika 2. – *Model distribuiranog data mining-a*

Karakteristike i ciljevi DDM-a čine ga veoma pogodnim iz nekoliko razloga:

- Distribuiranost podataka je posledica distribuiranosti klijentskih organizacija.
- Transfer velikih količina podataka rezultuje visokim cenama komunikacije.
- Potreba za jedinstvenim rezultatima sa distribuiranih i heterogenih izvora podataka uslovljava integraciju znanja
- DDM obezbeđuje okruženje koje omogućava da se veliki setovi podataka podele na manje setove, za koje je potrebno manje resursa i manje vremena za obradu.

3.2. Servisno orijentisani distribuirani data mining

Većina algoritama data mining-a pretpostavlja da će analitičari podataka agregirati podatke izvedene iz sistema proizvodnje u serveru, za računarski intenzivne procese istraživanja podataka. Međutim, problemi kao što su briga o privatnosti podataka i ograničenja u širini prenosa podataka pokazuju da agregiranje podataka za centralizovan mining jednostavno nije moguće u sve više slučajeva. Zahtevi kao što su očuvanje privatnosti i dobijanje prave informacije u pravom trenutku nameću dodatne zahteve za DDM, ukčljučujući obezbeđenje servisa „prema tražnji“ [1].

Ovi problemi mogu se rešiti usvajanjem servisno orijentisane arhitekture (SOA) za DDM. SOA kao infrastruktura može promeniti fokus procesa razvoja DDM-a sa implementacije algoritma na otkrivanje i sastavljanje algoritma za narednu generaciju DDM aplikacija.

4. TEŠKOĆE U PRIMENI DATA MINING-A

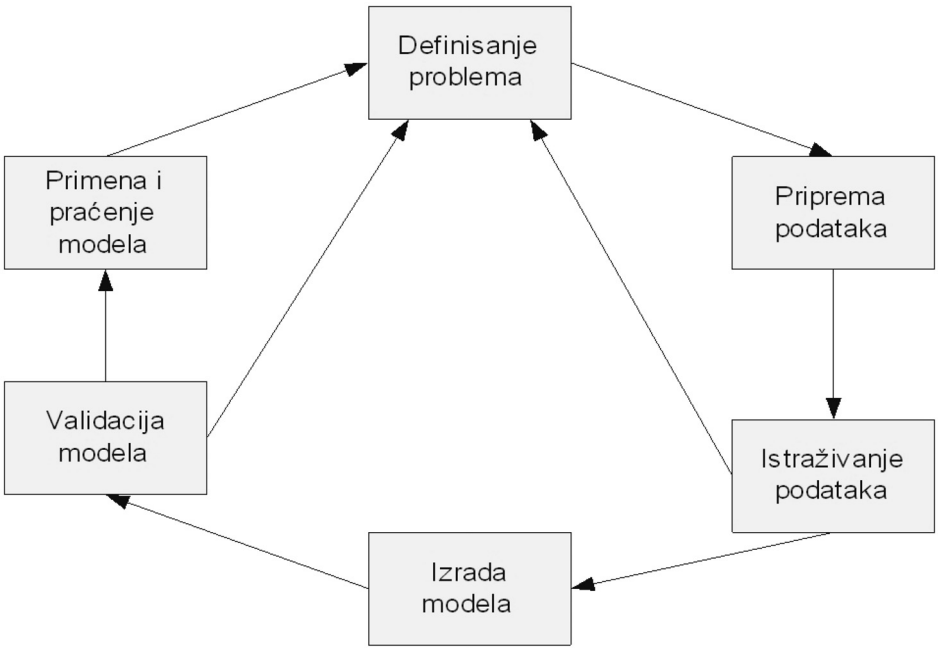
Ideja otkrivanja znanja u velikoj količini podataka je privlačna i intuitivna, ali je istovremeno izazovna i teška za realizaciju u tehničkom smislu. Moraju postojati strategije koje je potrebno implementirati za bolje iskorišćavanje podataka prikupljenih u e-poslovanju. Osnovni problemi u primeni data mining-a su:

- Format podataka - podaci su najčešće malo ili nimalo struktuirani, ne pridržavaju se nikakvih definisanih šema i kao takvi, mogu dovesti do neregularnih ili nekompletnih informacija.
- Količina podataka - setovi podataka prikupljenih u aplikacijama elektronskog poslovanja su veliki, ali tradicionalne data mining metode podržavaju rad sa velikim setovima podataka.
- Kvalitet podataka - web server logovi ne moraju da sadrže sve potrebne podatke, a podaci mogu i da budu oštećeni, što će otežati predviđanje.
- Prilagodljivost podataka - podaci na Web-u se stalno menjaju, pa je potrebno je omogućiti da data mining algoritmi i modeli rade sa podacima u realnom vremenu.
- XML podaci - XML dokumenti nisu uvek u istom formatu, što rezultuje gubljenjem nekih vrednosti.
- Problemi sa privatnošću - problem je u nalaženju ravnoteže između želje kompanije da koristi lične podatke klijenata i njihove želje da ih zaštite.

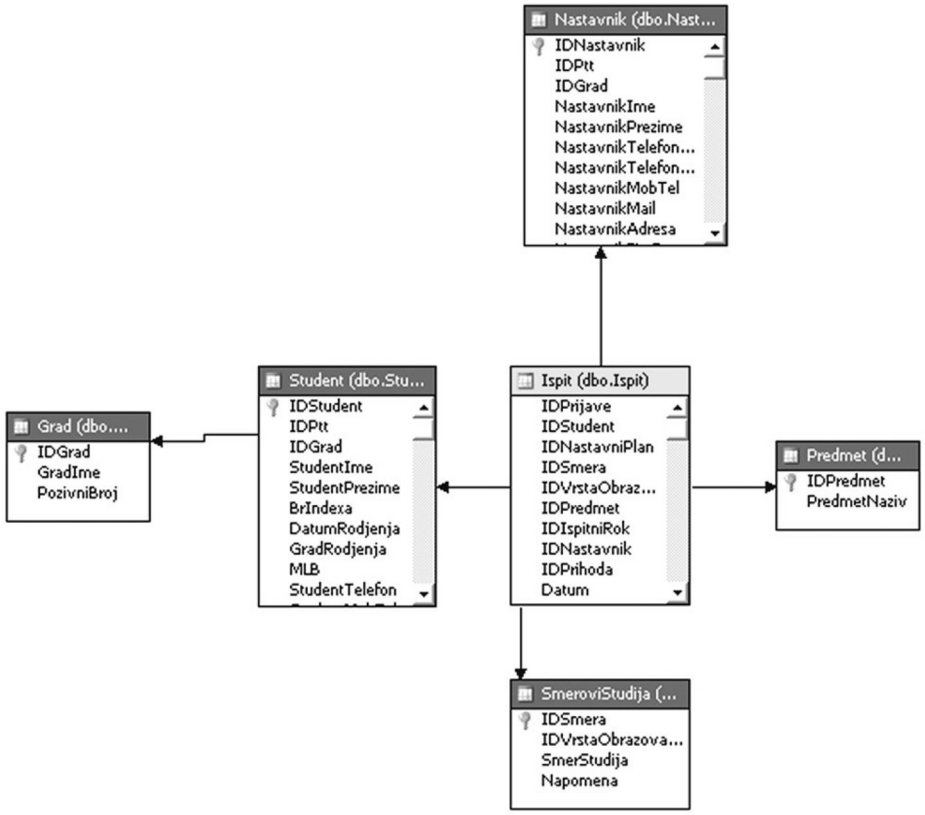
5. PRIMER PRIMENE DATA MINING-A U ELEKTRONSKOM OBRAZOVANJU

U daljem tekstu opisan je primer primene data mininga u sistemu za obrazovanje na daljinu na poslediplomskim studijama Fakultata organizacionih nauka u Beogradu. Za analizu korišćeni su podaci iz baze podataka poslovnog informacionog sistema, kao i podaci prikupljeni korišćenjem aplikacije za učenje na daljinu Moodle. Osnovna pretpostavka istraživanja je da ocena koju student dobija na polaganju ispita zavisi od većeg broja parametara. Potrebno je utvrditi koji su to parametri i kao i stepen zavisnosti.

Prikaz korišćene metodologije dat je na slici 3 [5].



Slika 3. – *Data mining metodologija*



Slika 4. – *Pahuljičasta šema kocke podataka*

Definisanje problema

Polazna osnova za primenu data mining-a je dobro definisanje poslovnog problema. Cilj data mining-a ne sme se definisati opšte i preopširno. Prilikom definisanja problema potrebno je dati odgovore na pitanja kao što su:

- Šta je to što treba pronaći u podacima?
- Da li je cilj data miniga predviđanje ili pronalaženje veza između podataka?
- Koja je ciljna promenljiva?
- Koji tip veze između podataka treba pronaći?

Na primer, problem je moguće definisati na sledeći način: potrebno je ispitati faktore koji utiču na ocenu studenata, kao i predvideti ocene studenata u budućnoti na osnovu tih faktora.

Priprema podataka

Pri izboru podataka koji će se koristiti za formiranje kocke podataka i data mining modele potrebno je izdvojiti podatke o studentima, ispitima, profesorima, smerovima i gradovima iz kojih studenti dolaze. Ostali podaci nisu predmet interesovanja u ovom primeru tako da se mogu zanemariti. Od mnoštva tabela u bazi za potrebe primera uzeto je šest: Student, SmeroviStudija, Grad, Predmet, Nastavnik, Ispit. U cilju jednostavnijeg i efikasnijeg otkrivanja raznovrsnih skrivenih obrazaca i zaključaka kreirani su određeni pogledi (views). Kroz njih su izdvojene i preformulisane odgovarajuće kolone, ali i pridodate neke nove. Svaki pogled je upotrebljen kao izvor podataka u nekom od kasnijie kreiranih modela data mining-a.

Istraživanje podataka

Pre kreiranja modela potrebno je dobro istražiti i razumeti podatke, što se najčešće vrši primenom različitih statističkih analiza ili analizom vizuelnih prikaza podataka. Takođe, u ovoj fazi neophodno je izvršiti i prečišćavanje podataka. Neke vrednosti mogu nedostajati, ili se u podacima mogu naći vrednosti unete greškom korisnika, ili vrednosti koje su posledica specifičnih uslova poslovanja u kratkom vremenskom periodu, pa kao takvi nisu pogodni za analizu.

Kako bi se kreirala kocka podataka potrebno je definisati tabelu činjenica i tabele dimenzija. Obzirom da se za tabelu činjenica uzima ona tabela koja sadrži najdetaljnije podatke, za potrebe ovog primera biće uzeta tabela Ispit kao tabela činjenica. Ona povezuje ostale tabele, pa se tabele koje su u direktnoj vezi sa njom, odnosno tabele Student, Grad, SmeroviStudija, Nastavnik i Predmet mogu posmatrati kao dimenzione tabele. Pahuljičasta šema prikazana je na slici 4.

Izrada modela

Pre nego što se pristupi izradi modela, potrebno je sve podatke podeliti u dve grupe: grupu za „trening“ modela i grupu za testiranje modela. Prva grupa podataka služi za kreiranje samog modela, dok se druga koristi za proveru njegove tačnosti. Najčešće se za „trening“ podatke uzimaju podaci iz dalje prošlosti, dok se podaci iz bliske prošlosti koriste za procenu modela. Procena se vrši tako što se test podaci analiziraju pomoću modela, a zatim vrednosti koje je model predvideo uporede sa stvarnim vrednostima promenljivih.

Validacija modela

U koraku validacije proverava se kolika je tačnost modela, koliko model dobro opisuje (objašnjava) posmatrane podatke, sa kojom verovatnoćom (tačnošću) model vrši predviđanje, koliko je model razumljiv, i sl. Neophodno je razumeti greške koje model pravi jer se među njima može napraviti gradacija. Tačnost estimatora se izražava razlikom između predviđenog rezultata i stvarnog rezultata, i standardno se izražava računanjem standardne greške (varijanse).

Primena i praćenje modela

U poslednjem koraku najbolji modeli postaju operativni i ugrađuju se u produkciono okruženje. Modeli se mogu koirsti-ti nezavisno, ili se ugrađivati u postojeće poslovne aplikacije. Najčešća primena dobijenih modela je za predviđanja ciljnih promenljivih, kreiranje različitih izveštaja u skladu sa potrebama korisnika, upravljanje novim podacima, i sl.

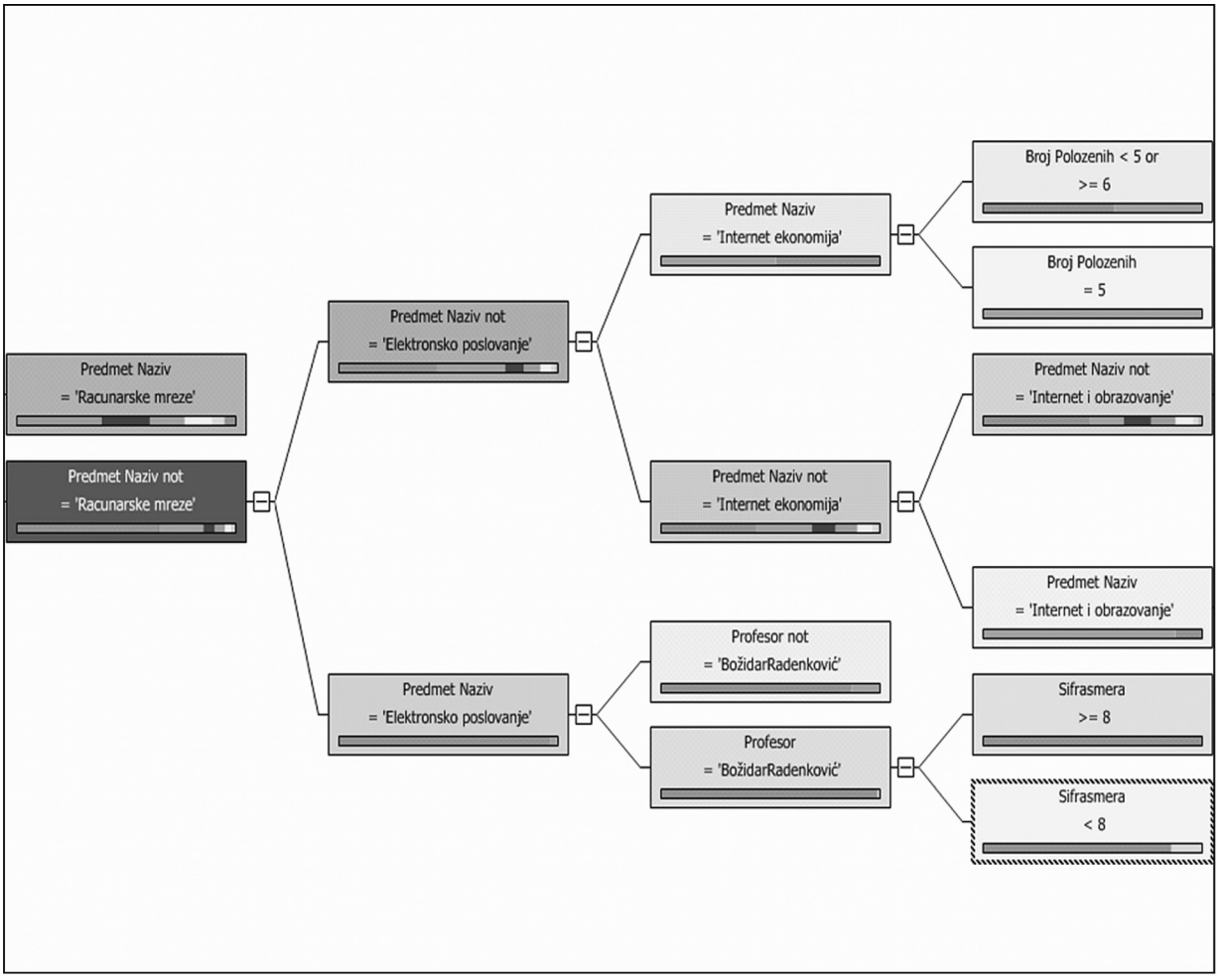
U slučaju primene modela u e-obrazovanju moguće je pratiti uspeh studenata, predviđati ispite koje će studenti polagati i ocene koje će dobiti, proceniti interesovanje za smerove studija, izvršiti uspešnu personalizaciju obrazovnog portala na osnovu uočenog ponašanja studenta, itd.

Rezultati i analiza

Za analizu izabranih podataka korišćena su tri algoritma: model stabla odlučivanja, NaiveBayes algoritam i clustering algoritam.

Koristeći kreiranu kocku podataka kao izvor ulaznih podataka, izborom Decision Tree algoritma, i označavanjem atributa "ocena" kao predviđajuće promenljive, kreira se data mining model koji ima za cilj da predvidi ocene koje studenti na poslediplomskim studijama dobijaju u zavisnosti od različitih parametra (slika 5). Na osnovu dobijenih podataka može se zaključiti sledeće:

- Postoji jaka zavisnost između ocene koju studenti dobijaju i broja ispita koje su položili. Naime, može se zaključiti da studenti koji su položili manje ispita, odnosno oni koji su na početku studija obično dobijaju veće ocene od onih koji su pri kraju studija.
- Postoji slabija korelacija između starosne strukture studenata i ocene koju dobijaju. Na osnovu ovoga se može zaključiti da studenti srednjih godina postižu bolje rezultate i od starijih i od mlađih studenata.
- Postoji vrlo slaba povezanost između mesta rođenja i mesta stanovanja studenata i ocena koje dobijaju. Ovaj rezultat je posledica toga što je većina studenata prijavila Beograd kao mesto prebivališta bez obzira na stvarno mesto rođenja i stanovanja.



Slika 5. – *Data mining model zasnovan na stablu odlučivanja*

Rezultati pokazuju da je najčešće stanje promenljive ocena 10 (tabela 1.)

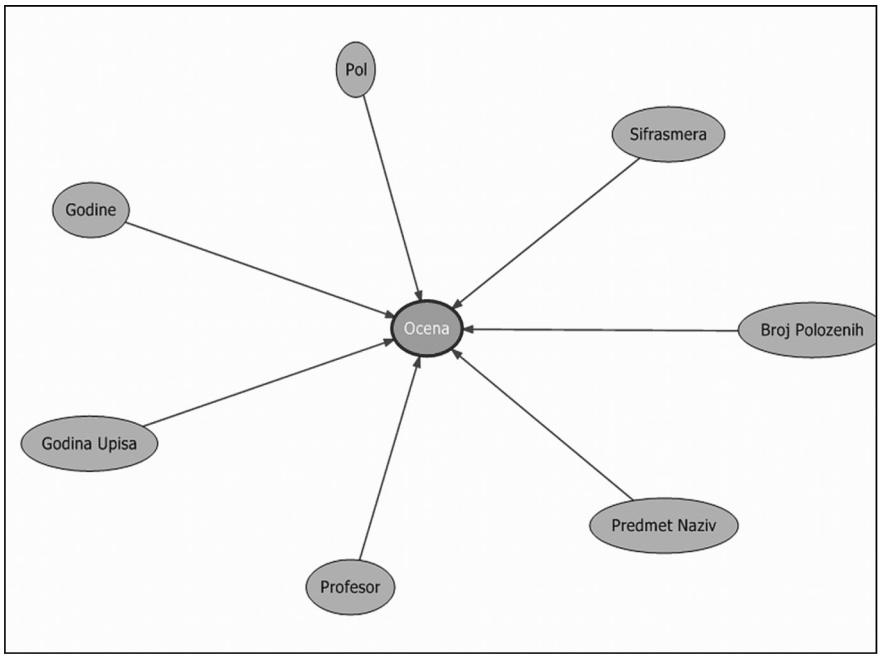
Ocena	Broj pojavljivanja	Verovatnoća
10	163	0,424899008
9	73	0,193536541
8	61	0,162688212
7	41	0,11127433
6	25	0,070143224
5	10	0,031582813

Tabela 1. – *Stanja promenljive ocena, broj pojavljivanja i verovatnoća pojavljivanja*

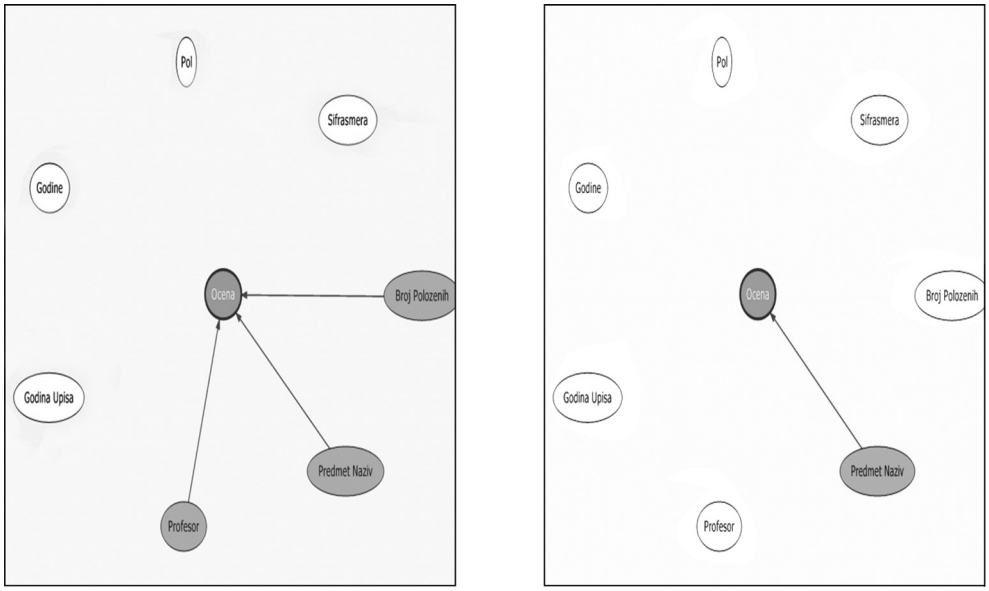
Pored analize stabla odlučivanja moguće je na osnovu datog modela izvršiti i analizu zavisnosti za dobijenu ocenu. Analiza pokazuje zavisnost ocene od određenih parametara za koje je utvrđeno da imaju izuzetan uticaj na predviđajuću promenljivu, u ovom slučaju ocenu. Analizom zavisnosti utvrđeno je da ocena zavisi od više parametra, kao što su broj položenih ispita, predmet koji student polaže, profesor, i drugi (slika 6).

Nakon pomeranja klizača ka jačim vezama ostaju samo Profesor, PredmetNaziv i BrojPoloženih (ispita), a na kraju samo PredmetNaziv. Dakle, dolazi se do očekivanog zaključka da naziv predmeta ima najveći uticaj na ocenu (slika 7).

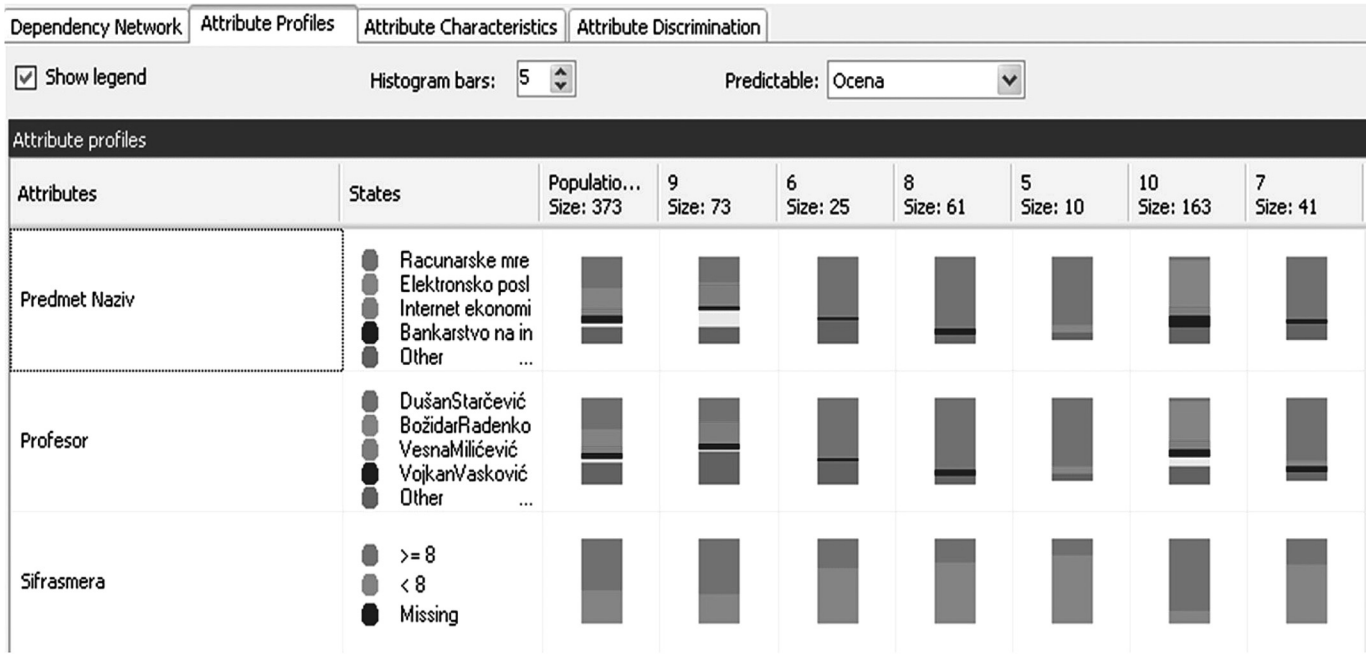
Primenom NaiveBayes algoritma utvrđeno je kako različita stanja ulaznih promenljivih utiču na izlaz atributa odluke, odnosno u ovom slučaju ocene. Sa slike 8. moguće je zaključiti kako pojedine vrednosti atributa PredmetNaziv i Profesor utiču na vrednost ocene.



Slika 6. – *Uticaj ulaznih atributa na ocenu*



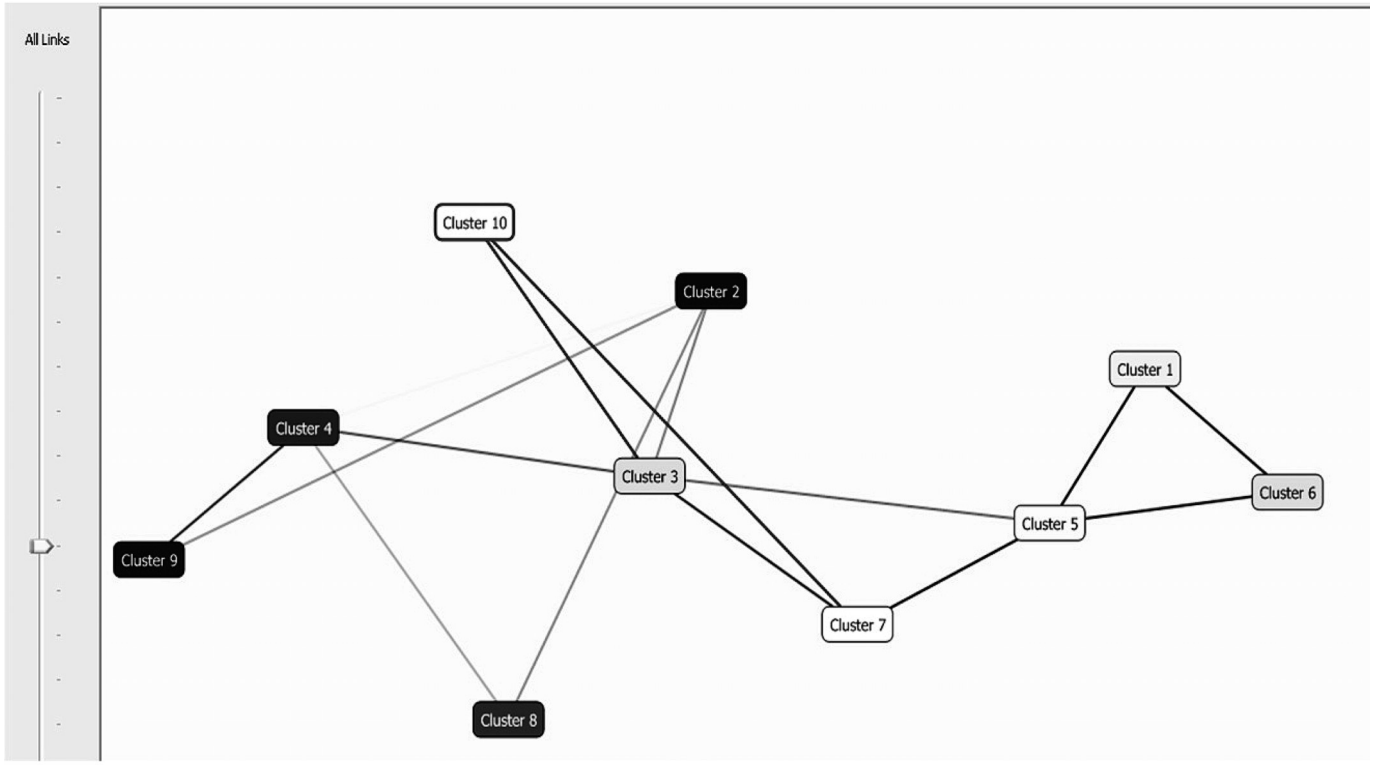
Slika 7. – *Jačina zavisnosti ocene od parametara*



Slika 8. – *Zavisnost ocene od pojedinih atributa*

Primenom clustering algoritma moguće je pronaći prirodna grupisanja među podacima, onda kada grupe nisu uočljive. Na slici 9 se nalazi model dobijen primenom modela klasterovanja na podatke iz kreirane mining strukture.

Na osnovu dijagrama, mogu se prikazati veze između klastera koje su otkrivene algoritmom. Linije između klastera predstavljaju bliskost. One su različito osenčene, u zavisno od jačine veza koja postoji među klasterima. Boja klastera predstavlja frekvenciju promenljive. Može se zaključiti da klasteri 2 i 5 imaju najveću frekvenciju, odnosno, najveći je broj ocena deset, a najjača veza je između klastera 1 i 5. Cluster Discrimation daje mogućnost analize ključnih razlika između klastera. Na slici 10 su prikazane razlike između klastera 3 i 7. Zaključuje se sa se u klasteru 7 nalaze studenti sa većim brojem položenih ispita, a koji su položili predmete Računarske mreže i Osnove organizacije, dok se u klasteru 3 nalaze studenti sa manjim brojem položenih ispita, koji su položili ispite Internet i obrazovanje i Bankarstvo na Internetu.



Slika 9. – *Klaster dijagram*

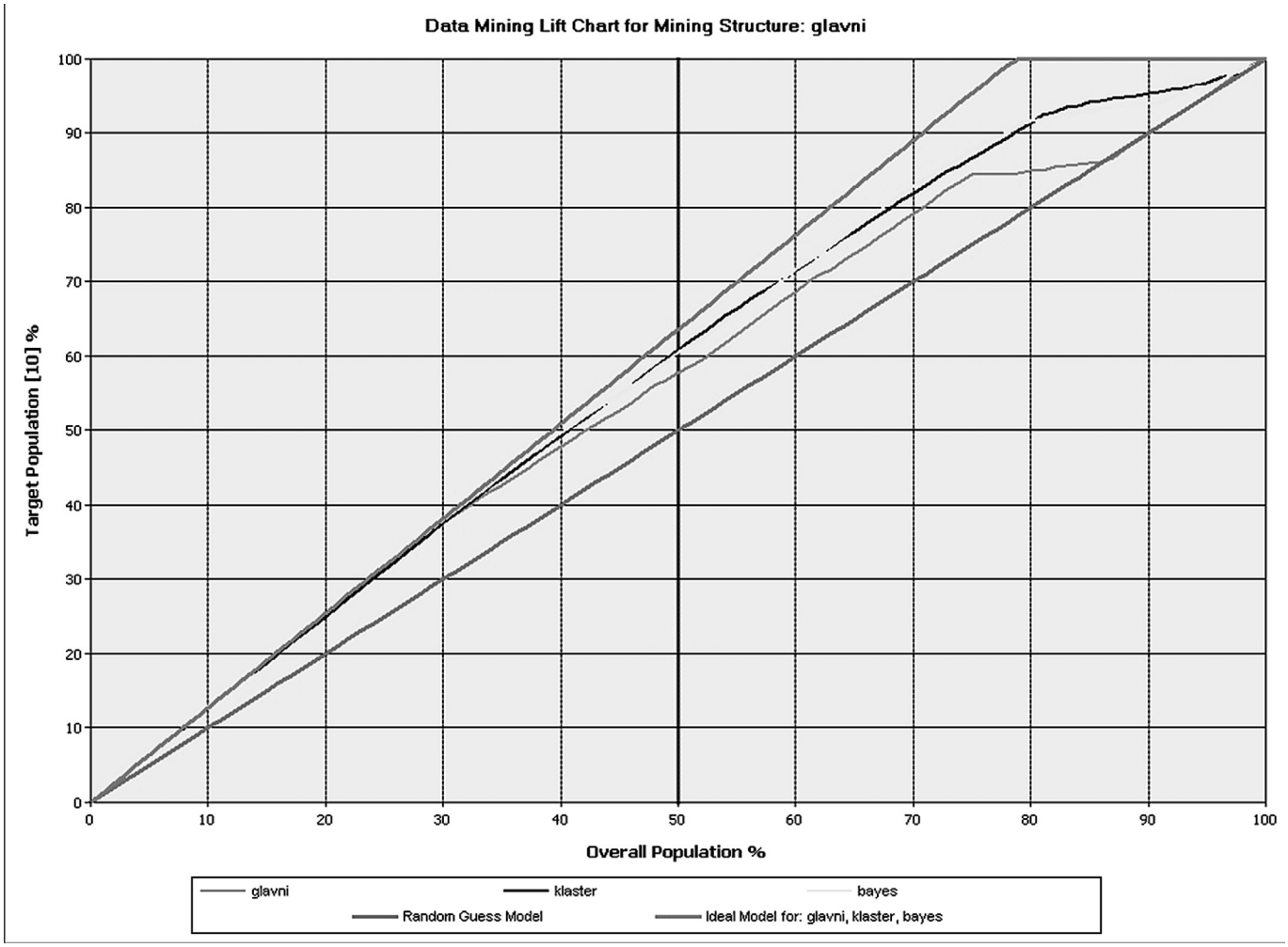
Cluster Diagram | Cluster Profiles | Cluster Characteristics | Cluster Discrimination

Cluster 1: Cluster 3 Cluster 2: Cluster 7

Discrimination scores for Cluster 3 and Cluster 7

Variables	Values	Favors Cluster 3	Favors Cluster 7
Predmet Naziv	Racunarske mreze		
Profesor	DušanStarčević		
Predmet Naziv	Osnovi organizacije		
Profesor	ŽivkoDulanović		
Profesor	VojkanVasković		
Profesor	GordanaMilosavljević		
Predmet Naziv	Internet i obrazovanje		
Predmet Naziv	Bankarstvo na internetu		
Broj Polozenih	>= 6		
Ocena	8		
Broj Polozenih	3 - 5		
Ocena	9		
Profesor	BožidarRadenković		

Slika 10. – *Osnovne razlike između klastera 3 i 7*



Slika 11. – *Lift model predviđanja za vrednost predviđajuće promenljive ocena 10*

Nakon izgradnje modela neophodno je izvršiti evaluaciju u cilju provere koliko dobro funkcioniše model koji je kreiran, ili ukoliko je izrađeno više različitih modela, koji od njih pokazuje najbolje performanse. Ako se utvrdi da model ne postiže zadovoljavajuće rezultate, potrebno je vratiti se na prethodne korake data mining procesa i izvršiti odgovarajuće korekcije.

Na slici 11. prikazan je lift model predviđanja. Koordinate na vertikalnoj osi pokazuju koji stepen ciljne populacije će biti uhvaćen ako se primeni odgovarajući model na onom procentu populacije definisanom na horizontalnoj osi. Može se uočiti da je za manje delove populacije najbolji klaster model (zelena boja), a za veće uzorke optimalan je model stabla odlučivanja (crvena boja). Uočava se da je idealni model obuhvatio 100% ciljne populacije (Ocena=10) koristeći 78% ukupnih podataka. Modeli na ovoj slici pokazuju skoro podjednake performanse, a najbitnije je da su iznad "random guess" linije (plava boja). Koristeći date modele, oko 80% ciljne populacije se može uhvatiti koristeći 70% raspoloživih podataka.

9. ZAKLJUČAK

Savremeni data mining alati podržavaju različite metode obrade, poznati su široj javnosti i veoma su uspešni. Postoji nekoliko važnih aspekata elektronskog poslovanja gde data mining može biti koristan, kao što su analiza ponašanja korisnika koja pokazuje koliko su zadovoljni Web sajtom, korelaciona analiza između Web sadržaja, bili to proizvodi ili dokumenta, analiza Web podataka kako bi se omogućila personalizacija u realnom vremenu i izmena strategije.

Elektronsko obrazovanje je samo jedna od oblasti elektronskog poslovanja u kojoj se data mining može primenti. U radu je dat jedan od primera primene data mining-a za predviđanje ocena i uspeha studenata na poslediplomskim studijama Fakultata organizacionih nauka. Primenom data mininga u elektronskom obrazovanju moguće je bolje uočiti potrebe studenata i unaprediti obrazovni proces u celini.

Literatura

[1] Cheung W., Zhang X., Wong H., Liu J., Luo Z., Tong F.: Service-Oriented Distributed Data Mining, Internet Computing, Volume 10, Number 4, July-August 2006, p.44-54

[2] Ćirić B., Poslovna inteligencija, Beograd, Data status, 2006

[3] Loke S.W.: Internet Delivery of Distributed Data Mining Services: Architectures, Issues and Prospects, Architectural Issues of Web–Enabled Electronic Business, Idea Group Publishing, 2003

[4] Nayak R.: Data Mining for Web–Enabled Electronic Business Applications, Architectural Issues of Web–Enabled Electronic Business, Idea Group Publishing, 2003

[5] Data Mining Concepts, SQL Server 2005 Books Online, http://msdn2.microsoft.com/en-us/library/ms174949.aspx

Zorica Bogdanović, Fakultet organizacionih nauka, Beograd
Oblast interesovanja: elektronsko poslovanje, poslovna inteligencija, internet tehnologije mr

Marijana Despotović, Fakultet organizacionih nauka, Beograd.
Oblast interesovanja: informacioni sistemi, internet tehnologije, elektronsko poslovanje.

dr Božidar Radenković, Fakultet organizacionih nauka, Beograd.
Oblast interesovanja: informacioni sistemi, internet tehnologije, elektronsko poslovanje.