

**Darko Krulj<sup>1</sup>, Milija Suknović<sup>2</sup>, Milutin Čupić<sup>2</sup>, Milan Martić<sup>2</sup>**

<sup>1</sup>Trizon Group, Beograd, Internacionalnih brigada 38, tel. 344-0068

<sup>2</sup>Fakultet organizacionih nauka, Beograd, Jove Ilića 154, tel. 3950-800

## **PRIMENA ALGORITAMA DATA MININGa U POSLOVNOM ODLUČIVANJU**

**Sadržaj:** *Jedan od čestih problema savremenih kompanija su velike količine podataka, prikupljene tokom poslovanja, a koje je potrebno obraditi i pretočiti u informacije u cilju boljeg upravljanja. Uvećanjem količine podataka, usložnjava se i problem njihovog skladištenja, obrade i analize. Sa druge strane postoji stalna potreba za informacijama radi dobrog i strukturiranog procesa odlučivanja. Upravo Data Mining (DM), iz velikih skladišta podataka omogućava korisnicima da otkriju skrivene šeme i važne poslovne informacije. U ovome radu je dat primer DMA koji je razvijen za potrebe studentske službe Fakulteta Organizacionih Nauka u Beogradu.*

**Ključne reči:** *FON, Data mining, OLAM, Data warehouse, OLAP, Microsoft Decision Trees (MDT) i Microsoft Clustering (MC), MS SQL SERVER 2000, Knowledge Discovery in Databases (KDD)*

### **1. UVOD**

Ovaj rad predstavlja logični nastavak rada: *Projektovanje i razvoj skladišta podataka studentske službe FONa*, od strane istih autora. U radu će biti predstavljene modeli data mininga (DM) i njihova primena u poslovnom odlučivanju.

Česta je konstatacija da se kompanije guše od podataka sa jedne strane, ali su žedne za informacijama sa druge. Kako baza podataka raste, postaje sve komplikovanije sa takvim podacima podržati odlučivanje u kompaniji. Za većinu organizacija, ciljevi DMA obuhvataju poboljšanje tržišnih sposobnosti, otkrivanje neuobičajenih obrazaca, predviđanje budućih trendova. Kroz rad će biti predstavljena implementacija DMA studentske službe FONa, u cilju podrške Kolegijumu Fakulteta u donošenju kvalitetnih poslovnih odluka.

### **2. ODLUČIVANJE I DATA MINING**

Proces vođenja i upravljanja preduzećem kao kontrolisanim sistemom, duboko se temelji na aktivnostima odlučivanja i donošenju adekvatnih i pravovremenih poslovnih odluka. Problemi sa kojima se susreću savremene kompanije odnose se pre svega na postojanje visokog nivoa globalne konkurencije, skraćenom životnom ciklusu proizvoda, naglom i brzom menjanju tržišnih uslova. Da bi se predvidele neočekivanih situacije i bar delimično umanjilo negativno dejstvo, potrebno je blagovremeno reagovati.

DM je proces kreiranja najrazličitijih upita i ekstrakcija korisnih informacija, uzora i trendova prethodno nepoznatih, sadržanih u velikim bazama podataka.

DM je skup tehnika za analizu podataka, čiji je cilj da u podacima pronađe određene zavisnosti, veze i pravila vezana za podatke i isti protumači u novi, viši nivo kvalitetne informacije. Za razliku od Data Warehousea (DW) koji ima jedinstven prilaz podacima, DM daje rezultate koji predstavljaju veze i zavisnosti između podataka, koje se ne bi mogle otkriti na drugi način, npr pomoću SQL upita ili prostim posmatranjem podataka.

### 3. DATA MINING

DM se često susreće pod različitim nazivima. Na Slici 1. prikazani su najčešći nazivi koji se koriste u literaturi, kao i jedna od najpotpunijih definicija data mininga[7].



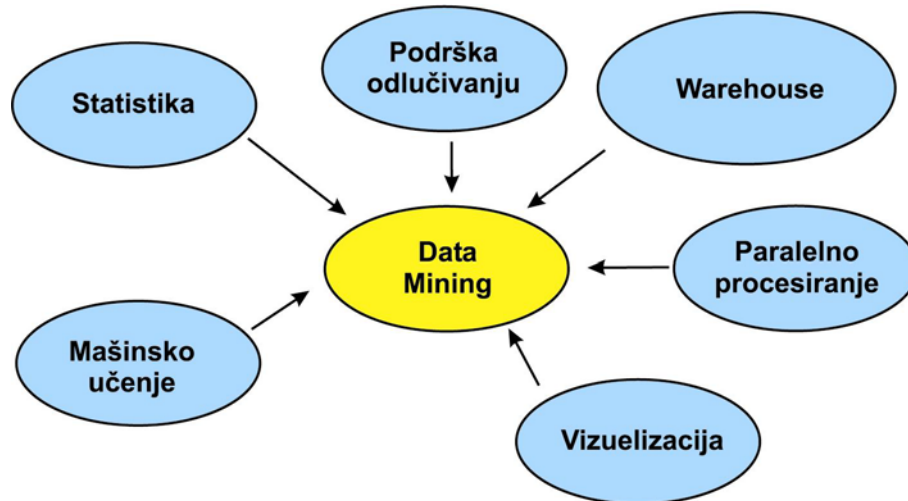
Slika 1. Definicija DMa

DM predstavlja integraciju više tehnologija kao što je prikazano na Slici 2. Njime je obuhvaćeno i upravljanje podacima kao što je upravljanje bazama, DW, statistika, podrška odlučivanju, mašinsko učenje, vizuelizacija, itd. U istraživanjima DMa se koriste znanja iz mnogih oblasti i disciplina. Tako npr. DW, kao jedna od ključnih tehnologija analize podataka, integriše različite izvore podataka i organizuje ih radi efikasnije analize (mininga).

Proces DMa se sastoji od nekoliko važnih koraka. Ti koraci obuhvataju organizovanje podataka za mining, određivanje željenog rezultata, izbor alata za mining, izvodjenje mininga nad podacima, selekciju rezultata kako bi se odvojili oni korisni, preduzimanje konkretnih akcija, i evaluacija akcija kako bi se izdvojilo ono što je korisno.

Postoji nekoliko tipova rezultata koji se dobijaju data miningom. Jedan od rezultata je klasifikacija, gde su slogovi grupisani u smislene podklase. Drugi "izlaz" iz DMa je "sequence detection", detektovanje sekvenci. Tako se, posmatrajući obrasce (paterne) u

podacima, utvrđuje njihova sekvenca. Sledeći oblik izlaza je “analiza zavisnosti podataka” gde se uočavaju potencijalno interesantne veze, zavisnosti ili asocijacije izmedju podataka. Analiza devijacija predstavlja još jedan oblik izlaza.



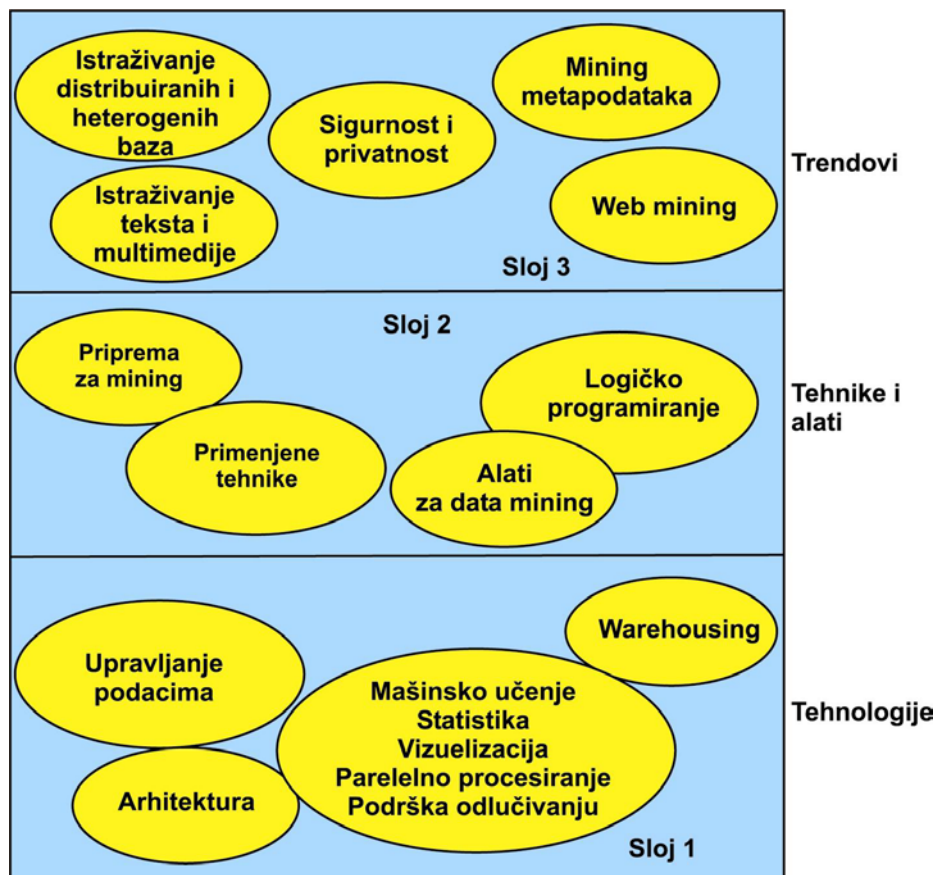
Slika 2 Uticaj raznih disciplina na data mining

Neki od sadašnjih pravaca data mininga su:

- Analize distribuiranih, heterogenih i starih baza podataka,
- Analize multimedijalnih podataka,
- Analiza podataka sa interneta,
- Sigurnost podataka u data miningu,
- Analize metapodataka, itd.

U većini slučajeva radi se o bazama koje sadrže distribuirane i heterogene podatke. Različite mining tehnike se koriste za sređivanje i struktuiranje tih podataka. Zatim se koriste različiti DM alati koji operišu sa ovako struktuiranih podacima. Ipak, znatan broj podataka je nestruktuiran. Takvi su na primer, podaci u multimedijalnim bazama. Za njih je potrebno razviti odgovarajuće data mining alate. Isto tako, podaci do kojih dolazimo preko interneta su mnogobrojni, pa je neophodno razviti i alate za ekstrakovanje bitnih podataka.

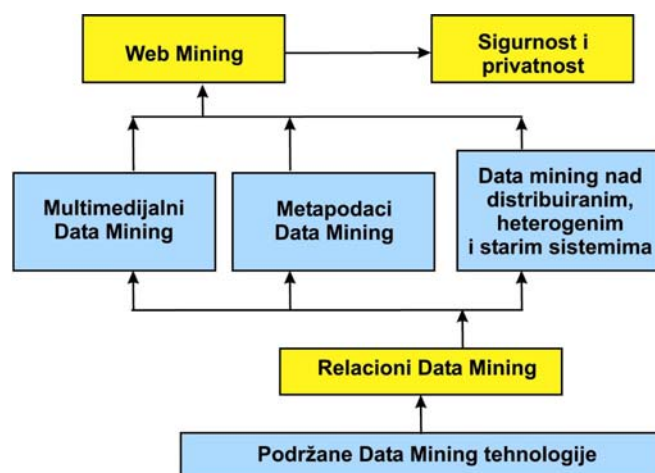
Na Slici 3 prikazana je evolucija data mining sistema. Na prvom nivou nalaze tehnologije koje se koriste u DMu: projektovanje baza i upravljanje podacima, DW, mašinsko učenje, statistika, vizualizacija, paralelno procesiranje i sistemi za podršku odlučivanju. Na drugom nivou predstavljene su tehnike i alati: priprema za DM, alati za DM, istraživanje podataka i logičko programiranje. Na trećem nivou prikazani su trendovi DMA: istraživanje distribuiranih i heterogenih baza, tekstulanih i multimedijalnih baza, sigurnosti, privatnosti i WWWa.



Slika 3. Evolucija DMA

Na Slici 4. prikazani su pravci razvoja. U početku DM je bio baziran isključivo na relacione izvore podataka. Kasnije se njegova upotreba proširila na DM nad mulitmedijalnim, metapodacima, distribuiranim, heterogenim i starim bazama podataka. Najnoviji trendovi su zasnovani na WEB miningu, sigurnosti i privatnosti na internetu.

Kako su podaci od ključnog značaja za DM, pre primene njegovih algoritama veoma je važno dobro strukturirati podatke. Zbog toga se veoma često DM analize primenjuju na DW sisteme, koji su već dobro strukturirani i ne poseduju nekonzistentnosti.



Slika 4. Pravci razvoja data mining sistema

#### 4. FAZE OTKRIVANJA ZNANJA

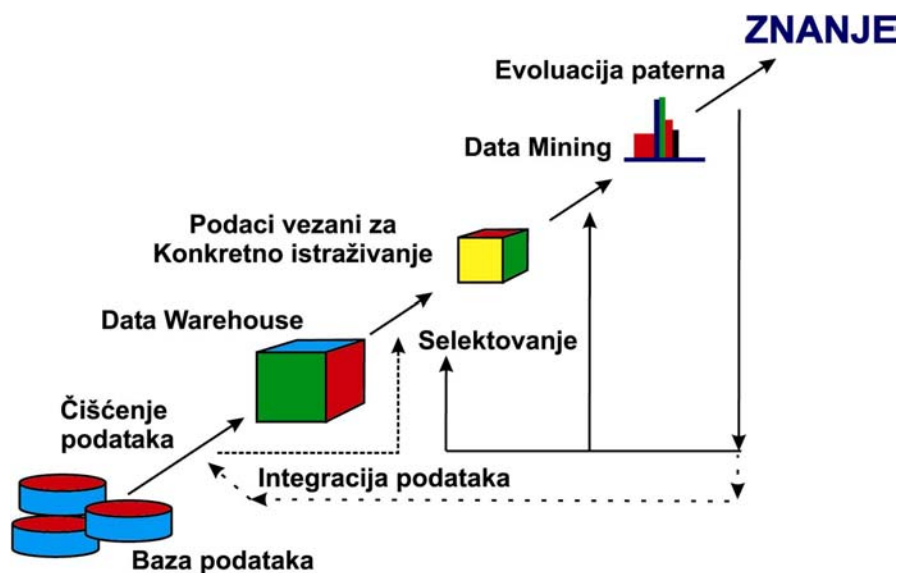
Otkrivanje znanja je dugotrajan i složen proces koji se sastoji od sedam osnovnih faza [2]:

1. Prečišćavanje podataka,
2. Integracija podataka,
3. Izbor podataka,
4. Transformacija podataka,
5. Data mining,
6. Evaluacija uzora,
7. Presentacija znanja.

Pomenute faze prikazane su na slici 5.

Veoma često baze podataka sadrže i podatke koji nisu tačni, potpuni, precizni ili ne poseduju dovoljan nivo konzistentnosti da bi bili upotrebljeni u svrhe analize. Prva faza u procesu data mininga ima ulogu da otkrije takve podatke i eliminiše ih iz dalje analize. Nije redak slučaj da kompanije koriste više različitih sistema za upravljanje bazama podataka. U tom slučaju poseban problem se javlja prilikom integracije podataka sadržanih u različitim sistemima. Potrebno je veoma mnogo pažnje i iskustva da bi se ovi podaci predstavili u odgovarajućem obliku. U fazi izbora podataka od “nepresušnog mora” podataka, izdavaju se “kapljice” koje sadrže podatke relevantne za posmatranu pojavu. U fazi transformacije izabrani podaci se prevode u najpogodniji oblik za primenu data mining analiza. Kada su podaci transformisani, primenjuju se data mining

tehnike koje generišu uzore, koji se kasnije procenjuju i tako se dolazi do dragocenih saznanja o posmatranoj pojavi.



Slika 5. Proces otkrivanja znanja

## 5. MICROSOFT DM ALGORITMI

Microsoft je u SQL Server 2000 integrisao DM alate kao deo njegovog Analysis servisa. Osim toga, Microsoft je implementirao i OLE DB za DM API. Ovaj API definiše Data Mining upitni jezik koji je baziran na SQL sintaksi. Sadrži dva algoritma za DM: Microsoft Decision Trees (MDT) i Microsoft Clustering (MC).

MDT algoritam je jedna od tehnika za modelovanje predviđanja i inspirisan je poznatim CART, CHAID i C4.5 algoritmima.

MC algoritam je baziran na Expectation and Maximization algoritmu (EM). EM algoritam vrši iteracije između dva koraka: u prvom koraku (Expectation) algoritam izračunava pripadnost klasteru za svaki od posmatranih slučajeva. Pripadanje klasteru predstavlja verovatnoću da slučaj pripada datom klasteru. U drugom koraku (Maximization) algoritam koristi te pripadnosti da bi odredio parametre modela, kao što su pozicija i parametri Gausove raspodele.

Klasteri su, sa druge strane, bazirani na statistici raznih atributa. Oni omogućavaju da se kreira model koji se ne koristi za predviđanje ali koji je veoma efikasan u pronalaženju slogova koji imaju zajedničke attribute pa se, prema tome, mogu svrstati u istu grupu.

Za razliku od većine klastering algoritama koji učitavaju sve podatke u memoriju, što predstavlja problem kod velikih količina podataka, MC algoritam podatke učitava selektivno, obrađuje ih i na osnovu datog Data Mining modela, analizira date slučajeve i

u jednoj iteraciji dolazi do rešenja što mu daje veoma dobre performanse u radu sa veoma velikim bazama podataka.

MDT je baziran na verovatnoći različitih atributa i koristi se kada je potrebno predviđanje slučajeva. MDT algoritam takođe generiše pravila. Svaki čvor u stablu se može izraziti kao set pravila koja ga opisuju kao i čvorove koji su doveli do njega.

Detaljniji prikaz navedenih algoritama može se naći na adresama [13] i [14].

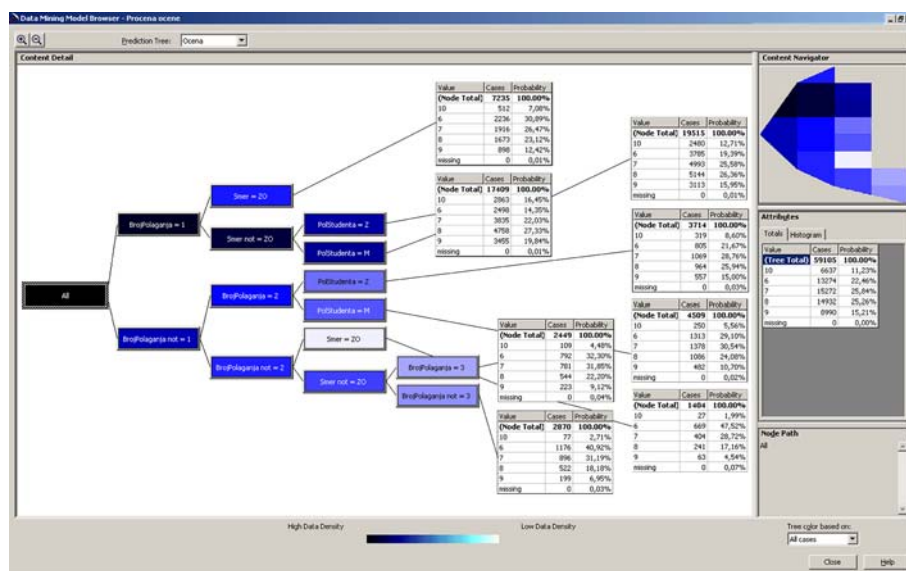
## 6. ILUSTRATIVNI PRIMERI DATA MININGA

U ovom projektu je implementirano preko sto stabala odlučivanja i klaster analiza, ali zbog ograničenog prostora izložićemo reprezentativne. Sve analize su podeljene u tri osnovne grupe:

1. Analize na nivou fakulteta,
2. Analize za diplomirane studente FONa,
3. Analize za izbor smerova, pri upisu treće godine studija.

### 6.1 ANALIZA NA NIVOU FAKULTETA

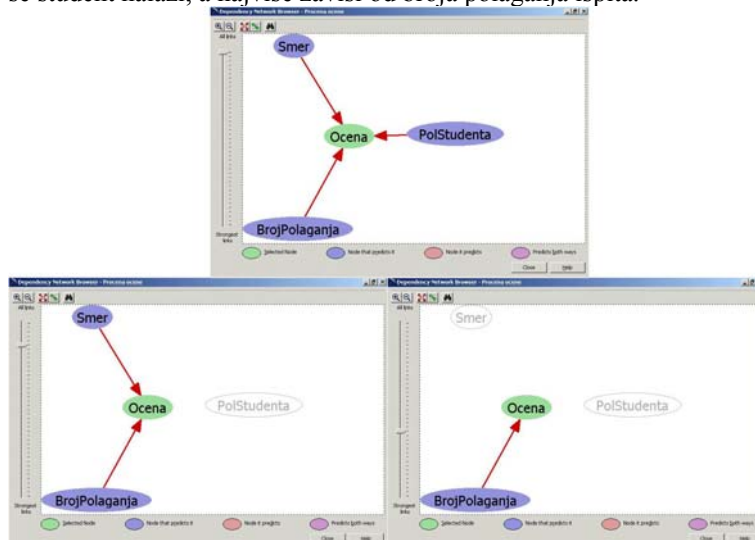
Na Slici 6. prikazan je MDT koji ima za cilj da na osnovu ranijih podataka o polaganjima ispita, predvidi ocenu koju će student dobiti prilikom izlaska na ispit. Na osnovu podataka sadržanih u stablu jasno se vidi da ukoliko student izlazi više puta na ispit, njegove šanse da će dobiti visoku ocenu se znatno smanjuju. Isto tako se može primetiti da studenti ženskog pola u opštem slučaju dobijaju veće ocene od studenata muškog pola i da studenti sa prve dve godine studija (smer ZO) dobijaju lošije ocene od studenta sa starijih godina.



Slika 6. MDT za procenu ocene

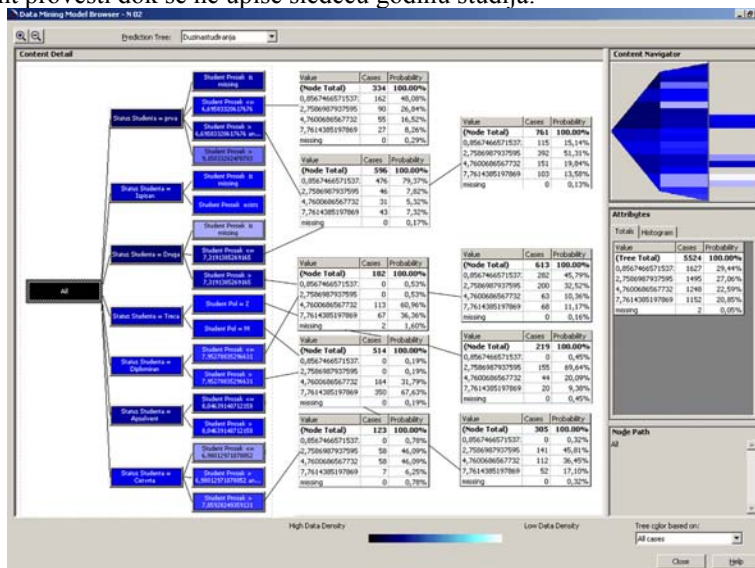


Na Slici 7. prikazana je međuzavisnost kriterijuma za procenu dobijene ocene. Pomeranjem klizača na levoj strani slika, određuje se jačina zavisnosti posmatranih kriterijuma. Dobijena ocena najmanje zavisi od pola studenta, nešto više zavisi od smera na kome se student nalazi, a najviše zavisi od broja polaganja ispita.



Slika 7. Analiza zavisnosti za dobijenu ocenu

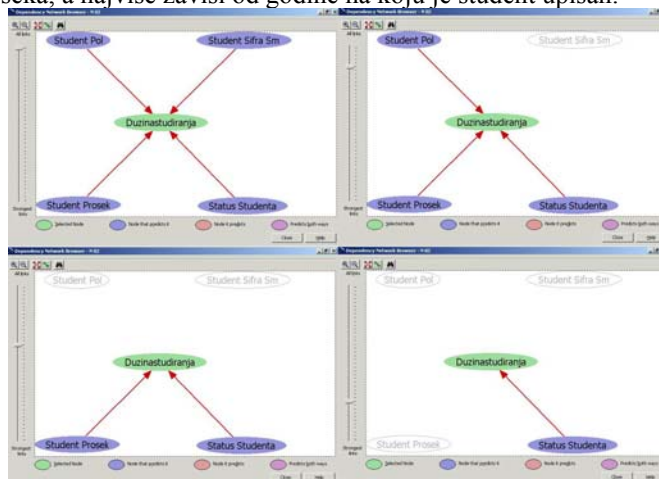
Na osnovu MDT za procenu dužine studiranja (Slika 8.) može se zaključiti koliko studenti provedu kalendarskih godina da bi upisali neku školsku godinu. Takođe na osnovu ovog stabla moguće je na osnovu proseka izračunati verovatnoću koliko godina će student provesti dok se ne upiše sledeću godinu studija.



Slika 8. MDT za procenu dužine studiranja

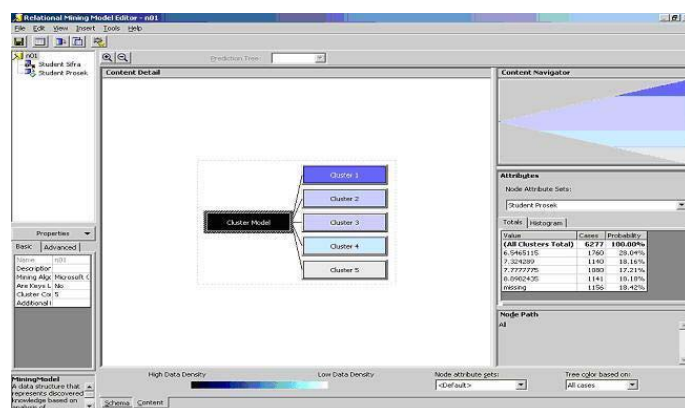


Na Slici 9. prikazana je međuzavisnost kriterijuma za procenu dužine studiranja. Pomeranjem klizača na levoj strani slika, određuju se jačina zavisnosti posmatranih kriterijuma. Dužina studiranja veoma malo zavisi od smera i pola studenta, nešto više zavisi od proseka, a najviše zavisi od godine na koju je student upisan.



Slika 9. Analiza zavisnosti za dužinu studiranja

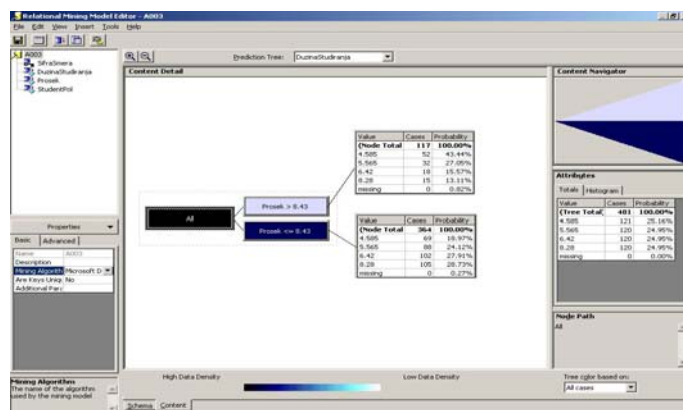
Na Slici 10. prikazani su rezultati klasterovanja za sve studente prema prosečnoj oceni. U analizi je učestvovalo 6277 studenata. U prvu grupu svrstani su studenti sa niskim prosekom njih 1760 sa prosečnom ocenom 6.54, interval od (6.00-7.093). U drugu grupu svrstani su prosečni studenti njih 1140 sa prosečnom ocenom 7.32, interval od (7.095-7.556). U trećoj grupi studenata nalazi se 1080 studenata sa prosečnom ocenom 7.77, interval od (7.558-8.000). Najbolju grupu predstavljaju 1141 student sa prosečnom ocenom 8.89, interval od (8.023-10.00). Poslednju grupu predstavljaju studenti koji nisu položili ni jedan ispit, njih 1156.



Slika 10. MC za studente fakulteta po prosečnoj oceni.

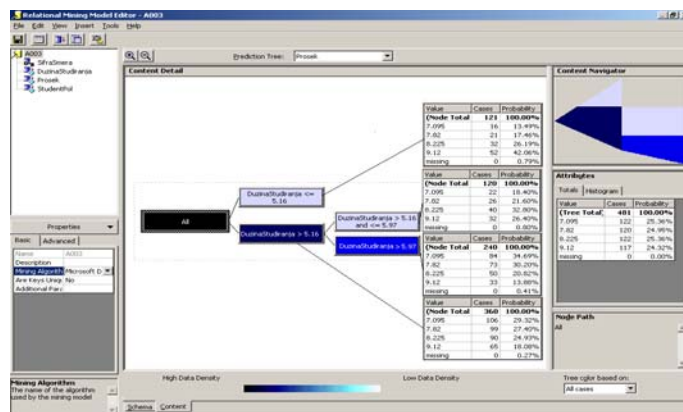
## 6.2 ANALIZA ZA DIPLOMIRANE STUDENTE

Na osnovu MDT za procenu dužine studiranja sa slike 11., može se zaključiti da studenti koji imaju veći prosek brže završavaju fakultet. Konkretno studenti koji imaju prosek veći od 8.43 u 43.44% slučajeva završavaju fakultet za 4.585 godina, dok verovatnoća da će studenata sa prosekom manjim od 8.43 završiti fakultet za 4.585 godina je tek 18.97%.



Slika 11. MDT za procenu dužine studiranja

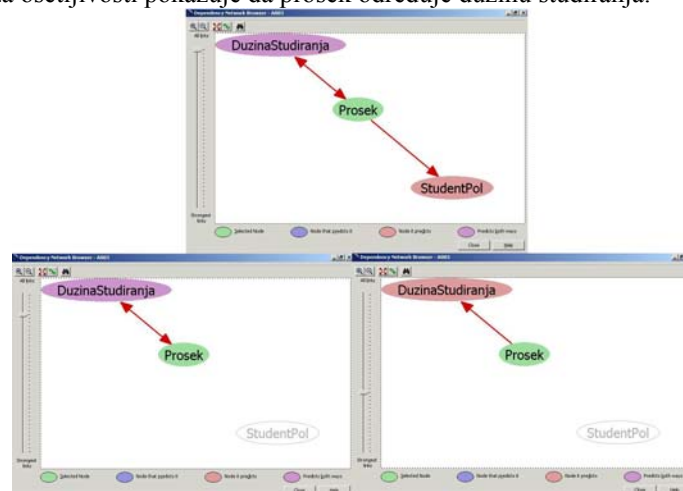
Na osnovu MDT za procenu proseka sa Slike 12., može se zaključiti da na prosek značajno utiče dužina studiranja i da studenti koji kraće studiraju imaju znatno veći prosek. Konkretno, ako je student diplomirao za manje od 5.16 godina verovatnoća da on ima prosek oko 9.12 je 42.06%, dok verovatnoća da student koji je studirao duže od 5.16 godina ima prosek oko 9.12 je 18.08%.



Slika 12. MDT za procenu proseka

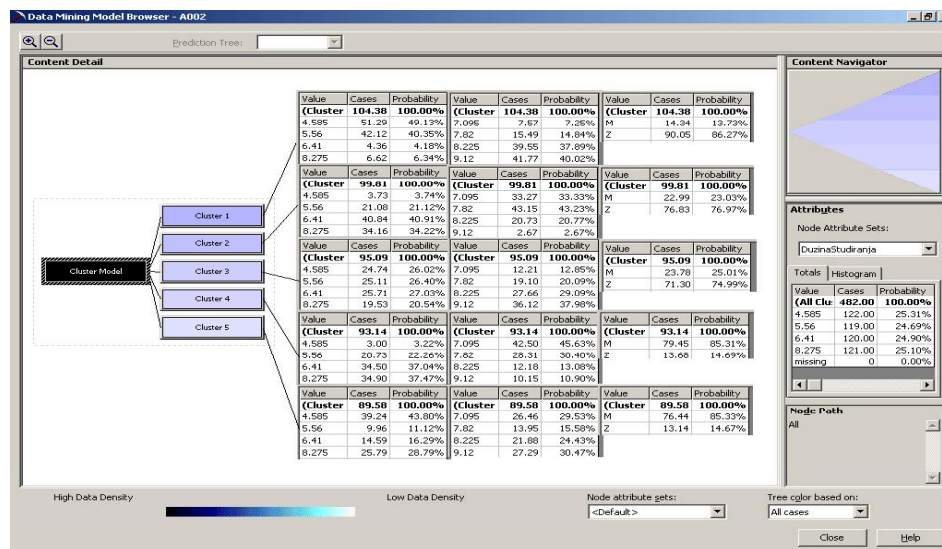
Na Slici 13. prikazana je međuzavisnost kriterijuma za procenu proseka. Pomeranjem klizača na levoj strani slika, određuje se jačina zavisnosti posmatranih kriterijuma.

Prosek veoma malo zavisi od pola studenta. Prosek i dužina studiranja su u međusobnoj vezi. Analiza osetljivosti pokazuje da prosek određuje dužinu studiranja.



Slika 13. Analiza zavisnosti za procenu proseka

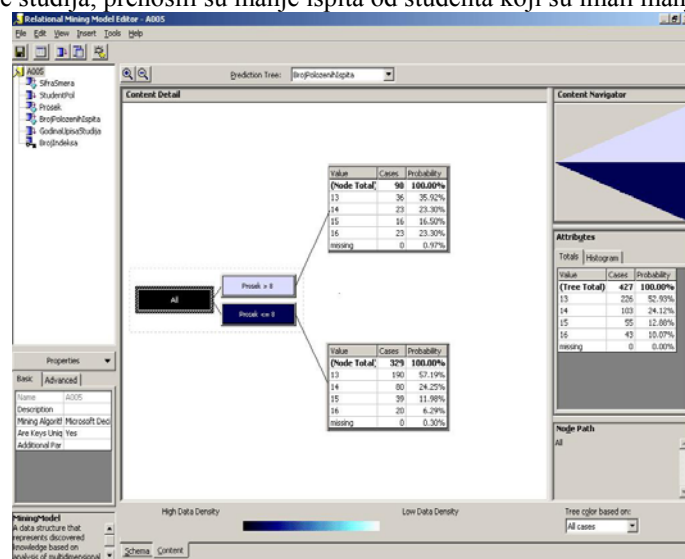
Na Slici 14. diplomirani studenti su podeljeni u pet klastera. Studenti su podeljeni u tzv. višedimenzionalne klustere. Prvu grupu studenata čine pretežno studenti ženskog pola koji su veoma brzo završili fakultet, i imaju veoma visok prosek. Drugu grupu čine studenti pretežno ženskog pola koji su dugo studirali i imaju relativno nizak prosek. Treću grupu čine studenti pretežno ženskog pola koji su prosečno dugo studirali i imaju prosečne ocene. Četvrtu grupu čine studenti pretežno muškog pola koji su dugo studirali i imali niske prosečne ocene. Konačno peta grupa je sačinjena od pretežno muških studenata koji su relativno brzo završili fakultet i imaju prosečne ocene.



Slika 14. MC za diplomirane studente

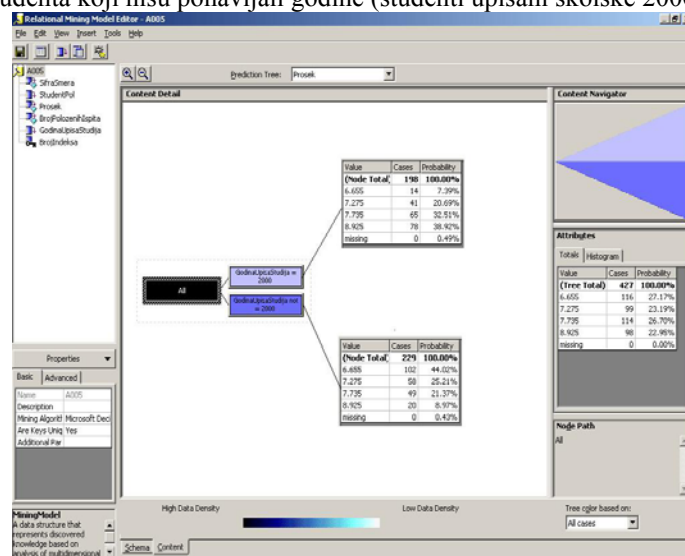
### 6.3. ANALIZA ZA IZBOR SMERA

Na osnovu MDT za procenu broja položenih ispita Slika 15. može se zaključiti da je najvažniji kriterijum za procenu broja položenih ispita, prosečna ocena studenta postignuta na polaganju ispita. Studenti koji imaju prosek preko 8.00, prilikom upisa treće godine studija, prenosili su manje ispita od studenta koji su imali manji prosek.



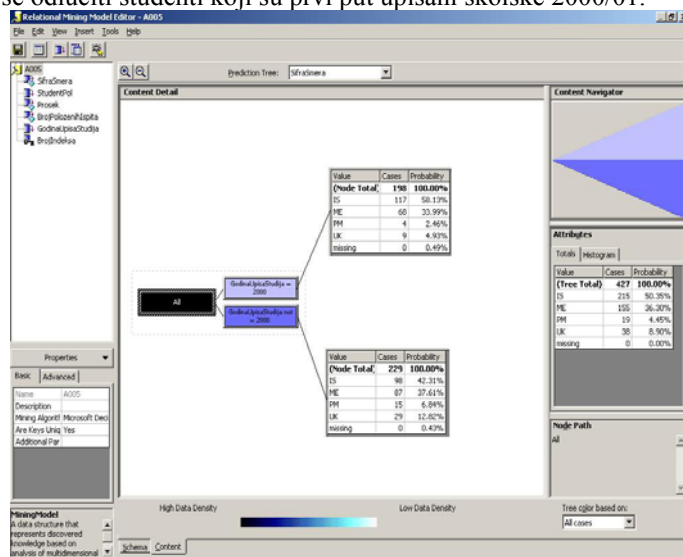
Slika 15. MDT za procenu broja položenih ispita

Na osnovu stabla prikazanog na Slici 16. može se zaključiti da godina upisa studija predstavlja najvažniji kriterijum za visinu proseka studenta. Naime studenti upisani pre školske 2000/01 su ponavljali godine, tako da je očekivati da je njihov prosek znatno manji od studenta koji nisu ponavljali godine (studenti upisani školske 2000/01)



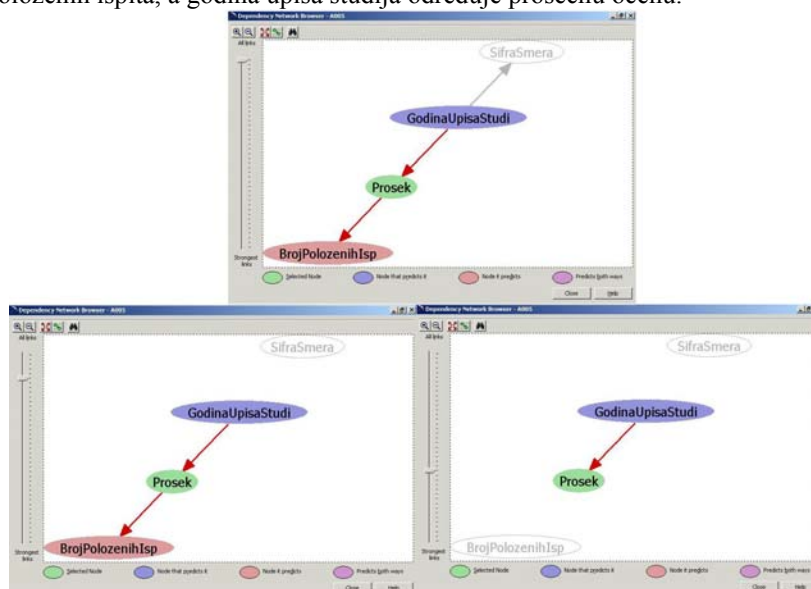
Slika 16. MDT za procenu proseka studenta

Godina upisa studija i kod procene smeru koji će upisati studenti igra najvažniju ulogu. Na osnovu stabla sa Slike 17., jasno se uočava da će iz obe grupe, najviše studenata upisati smerove informacijski sistemi i opšti menadžment, ali da će u većem procentu za te smerove se odlučiti studenti koji su prvi put upisani školske 2000/01.



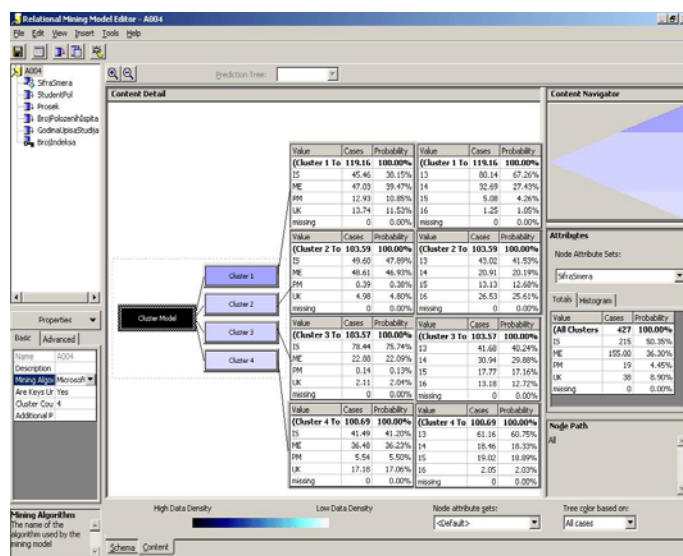
Slika 17. MDT za procenu smeru

Na Slici 18. prikazana je analiza međuzavisnosti kriterijuma za procenu smeru. Pomeranjem klizača na levoj strani slika, određuje se jačina zavisnosti posmatranih kriterijuma. Izbor smeru savisi samo od godine upisa studija, prosečna ocena određuje broj položenih ispita, a godina upisa studija određuje prosečnu ocenu.



Slika 18. Analiza zavisnosti za izbor smeru

Na slici 19. je prikazan MC za izbor smer. Studenti su podeljeni u četiri grupe. U prvoj grupi nalaze se studenti sa malim brojem položenih ispita, koji su se ravnomerno raspodelili za upis na sve smerove. U drugoj grupi nalaze se najbolji studenti koji su se odlučili da upišu informacione sisteme i opšti menadžment. U trećoj grupi nalaze se iznad prosečni studenti koji su se uglavnom odlučili za upis informacionih sistema i opšteg menadžmenta. U četvrtoj grupi nalaze se studenti koji su pokazali nešto veće interesovanje za upis na smer upravljanje kvalitetom i proizvodni menadžment, u ovu grupu spadaju studenti koji su položili najmanje ispita.



Slika 19. MC za izbor smer

## 7. ZAKLJUČAK

Na prezentiranim primerima jasno se uočava upotrebnost vrednosti MDT i MC algoritama. Oni na kvalitativno drugačiji način pristupaju analizi podataka u odnosu na konvencionalni pristup.

Sama mogućnost predviđanja i otkrivanja kriterijuma koji utiču na analizirane pojave daje im velike prednosti u odnosu na sve druge pristupe. Otkrivanje ciljnih grupa studenata koji žele da upišu neki od smerova mogu biti od velike koristi kolegijumu fakulteta, jer na osnovu rezultata upisa ove školske godine, lakše će se predvideti upisi u sledećim školskim godinama.

## 8. LITERATURA

- [1] Barry, D.: "Data Warehouse from Architecture to implementation", Addison-Wesley, 1997.
- [2] Jiwei, H., Micheline Kamber, "Data Mining: Concepts and Techniques", Simon Fraser University, 2001.
- [3] Seidman, C.: "Data Mining with Microsoft SQL Server 2000", Microsoft Press, 2001.
- [4] Gunderloy, M., Sneath, T.: "SQL SERVER Developer's Guide to OLAP with Analysis Services", Sybex, 2001.
- [5] Lory, O., Crandall, M.: "Programmers Guide for Microsoft SQL Server 2000", Microsoft Press, 2001.
- [6] Vidette Poe: Building a Data Warehouse for Decision Support , Prentice Hall, 1996.
- [7] Bhavani Thuraisingham: Data Mining: Technologies, Techniques, Tools and Trends.
- [8] M.J.A.Berry, G.Linoff: Mastering Data Mining: The Art and Science of Customer Relationship Management.
- [9] Milija Suknović, Darko Krulj, Milutin Čupić, Milan Martić: Projektovanje i razvoj skladišta podataka studentske službe FONa, SYMORG, Zlatibor, 2002.
- [10] Darko Krulj, Milija Suknović, Milutin Čupić, Milan Martić, Tihomir Vujnović, Projektovanje i razvoj OLAP sistema studentske službe FONa, INFOFEST, Budva, 2002.
- [11] Darko Krulj, Tihomir Vujnović, Milija Suknović, Milutin Čupić, Milan Martić, Algoritmi data mining-a, dobra osnova za poslovno odlučivanje, Tara, SYM-OP-IS 2002.
- [12] Milija Suknović, Milutin Čupić, Milan Martić, Darko Krulj: Projektovanje i razvoj skladišta podataka studentske službe FONa, SYM-OP-IS, Tara, 2002.
- [13] <http://citeseer.nj.nec.com/bradley98scaling.html>
- [14] [http://www.acm.org/sigmod/disc/p\\_scalableclassifsuj.htm](http://www.acm.org/sigmod/disc/p_scalableclassifsuj.htm)