

# Potruga za zlatom



spremljeni u skladištima podataka. Pritom se upotrebljavaju automatizirane metodologije dubinskog pristupa spremljenim podacima i identificiraju uzorci i odnosi preko kojih se potom grade prediktivni modeli. To je ujedno i multidisciplinarno područje koje uključuje baze podataka, statistiku i umjetnu inteligenciju te zahtijeva visokobrazovane i iskusne stručnjake (i u tehničkom i u poslovnom smislu).

SQL Server 2008 pruža zaokruženu platformu za razvoj najčešćih procesa rudarenja podataka implementirajući 10 ključnih algoritama koji pokrivaju najveći dio poslovnih procesa. Time i poslovni analitičari mogu sudjelovati u izradi prediktivnih analiza koje su prije bile namijenjene samo uskom krugu visokospecijaliziranih stručnjaka iz područja statistike i rudarenja podataka.

Ukratko, glavni je cilj rudarenja podataka izdvajanje znanja iz dostupnih podataka, izrada trendova, analiza i opisa postojećeg skrivenog znanja koje je često ključno za donošenje strateških odluka te predstavlja odlučujući faktor na vrlo konkurentnom tržištu.

## Poslovni problemi rudarenja podataka

Rudarenje podataka moguće je koristiti u skoro svim poslovnim aktivnostima. Ono ujedno pomaže odgovoriti na prediktivna pitanja koja sežu od predviđajućeg do neregularnog ponašanja određenog promatranog objekta. Rezultat rudarenja podataka kasnije je primjenjiv u scenarijima stvaranja preporuka u odlučivanju za poduzeće ili njegovog klijenta, analizi anomalija u bankarskom poslovanju, segmentaciji korisnika u grupe prema unaprijed definiranim parametrima, analizi kreditnog rizika, predviđanju budućeg ponašanja kao i u mnogim drugim poslovnim i znanstvenim analizama.

Već od SQL Servera 2005 rudarenje podataka postaje standardni modul unutar Visual Studija koji omogućava stvaranje projekata rudarenja podataka neovisno o vrsti izvora podataka (baze podataka, OLAP, tekstualne datoteke...). Osnovne mogućnosti rudarenja podataka dostupne su u Standard verziji SQL Servera, dočim je za punu funkcionalnost potrebno koristiti Developer ili Enterprise licencu.

## Integracija s drugim sustavima

Jedna je od glavnih prednosti SQL Server 2008 rudarenja podataka integracija u postojeće sustave koja omogućava direktno povezivanje rudarenja podataka u poslovne aplikacije. Na primjer, CRM aplikacije mogu posjedovati funkcionalnost grupiranja korisnika u segmente ili omogućiti izbor kontakata za koje postoji najveća vjerojatnost da će postati korisnik vaše usluge. ERP aplikacije, s druge strane, mogu koristiti ru-

Rudarenje podataka (engl. data mining) sastavni je i ključni dio poslovne inteligencije koji značajno povećava poslovnu vrijednost već postojećeg analitičkog sustava. SQL Server 2008 nudi zaokruženu platformu koja pruža i povezuje sve potrebne alate koji, osim ETL-a, OLAP-a i izvještavanja, nude i podršku za rudarenje podataka

### SQL Server 2008 Data Mining

Proizvođač	Microsoft
Tip	Alat za rudarenje podataka
Minimalna konfiguracija	Pentium III procesor 1 GHz ili brži, 512 MB RAM-a
Preporučena konfiguracija	procesor 1,6 GHz (preporučeno 2,2 GHz), 384 MB RPentium IV procesor 2 GHz ili brži, 2 GB RAM-a
Cijena (bez PDV-a)	13.969 USD sa 25 CAL-ova (Enterprise verzija, retail cijena)

+

Potpuna integracija u SQL Server, poznato razvojno okruženje, kratko vrijeme prilagodbe, velike mogućnosti, dostupna u svim fazama projekata poslovne inteligencije, dokazana tehnologija

-

Proizvod je relativno nov i još treba zauzeti svoje tržišno mjesto u odnosu na konkurenciju koja je tu puno duže

### DOJAM

Vrhunski proizvod za rudarenje podataka potpuno integriran u postojeću SQL Server 2008 platformu

### USTUPIO

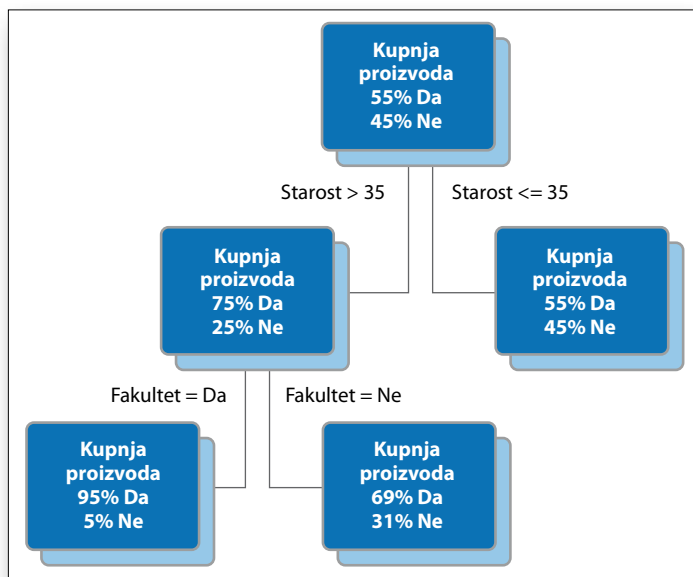
Dobiveno preko akademske licence (MSDN-AA)

### ■ JOSIP ŠABAN

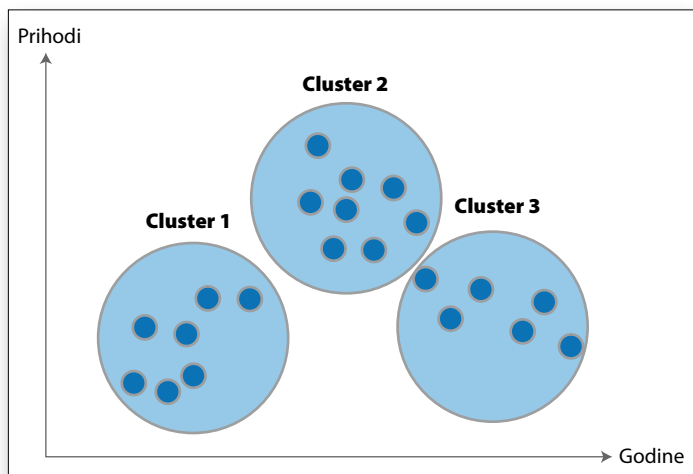
U zadnjih nekoliko godina računalna snaga eksponencijalno se povećavala prema poznatom Mooreovom zakonu, ali možda je još značajnije i većini ljudi nepoznato da se kapacitet čvrstih diskova povećavao za red veličine brže od procesorske snage. Točnije, mogućnost spremanja podataka značajno nadilazi procesorsku snagu. Rezultat je pohrana velikih količina podataka u bazama podataka gdje većina dolazi iz poslovnog softvera kao što su financijske aplikacije, CRM i ERP sustavi, logovi web servera i slično. To rezultira posjedovanjem velike količine spremljenih podataka iz kojih se najčešće ne izvodi novo znanje koje bi nam potom moglo biti iskorišteno u svrhu poslovnog odlučivanja.

Rudarenje podataka je proces pronalaženja novog i potencijalno korisnog znanja iz dostupnih podataka koji su danas najčešće

# SQL Server 2008 Data Mining



**Primjer klasifikacije na temelju stabla odlučivanja za kupnju određenog proizvoda temeljenog na ulaznim varijablama starosti i obrazovanja**



**Grupiranje uzoraka korištenjem clusteringa na primjeru varijabli o prihodima i starosti**

darenje podataka za predviđanje proizvodnje i stanja skladišta. Proizvodne aplikacije koje mogu predviđati greške na proizvodima i određivati uzroke tih grešaka.

Integracija rudarenja podataka u druge aplikacije (koristeći API-je za SQL Server Data Mining ugrađene u .NET jezike) stvara sustav koji se može automatski osvježavati i biti prilagođen svakom korisniku ili scenariju upotrebe.

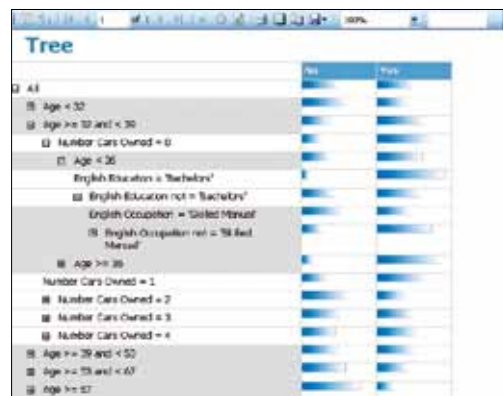
No kakva god primjena sustava rudarenja podataka, uvijek se temelji na tri osnovna principa: (1) Dati više podataka krajnjim korisnicima - npr. pružajući trenutne rezultate predviđanja u Office aplikacijama, čime im daju više informacija i bolju mogućnost odlučivanja ili provjeravaju ispravnost unosa u web forme; (2) Pružiti dodatne mogućnosti programerima preko dostupnih analitičkih alata; (3) Unaprijediti aplikacije kako bi krajnjem korisniku prikazali rezultate prediktivnih analiza - jedan od najčešćih primjera na webu su stranice koje vam daju prijedloge za kupnju analizirajući podatke o vašim prijašnjim aktivnostima.

## SQL Server 2008 kao alat za rudarenje podataka

Rudarenje u SQL Serveru 2008 nastavlja se na promjene nastale u verziji 2005 te korištenjem napredne funkcionalnosti rudarenja podataka, alata i API-ja koji dolaze s bazom podataka - i, nimalo zanemarivo, za koje se ne mora ništa dodatno platiti - SQL Server 2008 ujedno nastavlja sa svojom filozofijom koja odudara od filozofije konkurentskih alata za rudarenje podataka.

SQL Server 2008 Data Mining modul moguće je jednostavno povezati s novim ili postojećim poslovnim aplikacijama. Time se može značajno povećati i njihove prediktivne mogućnosti te ih i na taj način na tržištu izdvojiti od konkurencije. Upotrebom DMX-a, jezika vrlo sličnog SQL-u, programeri i administratori podataka mogu izvoditi složene prediktivne upite nad postojećim podacima bez potrebe za dugim učenjem novog jezika (što je slučaj kod konkurentskih alata). Time je rudarenje podataka još za jedan korak približeno poslovnim analitičarima.

Rijetko se, naime, problemi u rudarenju podataka mogu riješiti samo korištenjem alata. Ako na trenutak zanemarimo ključno ra-



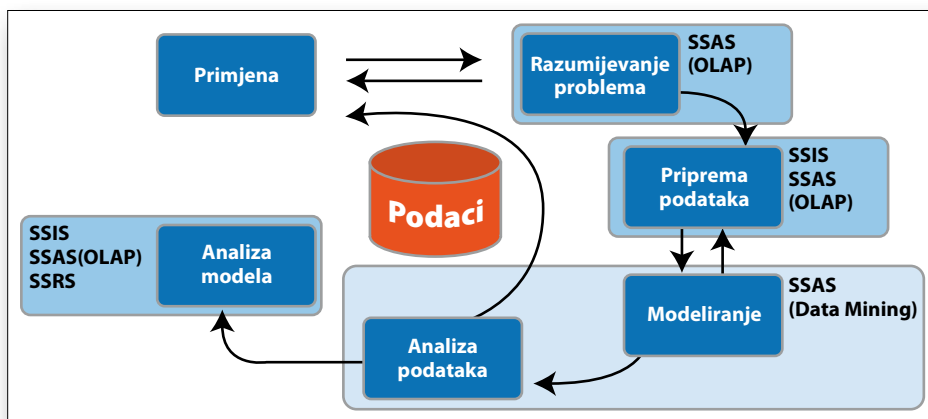
**Prikaz stabla odluke u Reporting Services**

zumijevanje podataka i poslovnih procesa, što je apsolutni preduvjet u bilo kojem ozbiljnijem projektu rudarenja podataka, SQL Server 2008 Data Mining, koristeći poznati BI Studio, predstavlja nerazdvojni dio skupa alata za razvoj aplikacija poslovne inteligencije.

Neki od primjera su (1) Integracija s Office paketom - integracija s Excelom i Visiom omogućuje rudarenje podataka krajnjim poslovnim korisnicima; (2) SSIS integracija - transparentna integracija rudarenja podataka u operativne tokove podataka u ETL procesima; (3) OLAP integracija - integracija s OLAP kockama dopušta rudarenje podataka nad složenim višedimenzijalnim izračunima i korištenje rezultata u OLAP klijentima kao što Excel Pivot Tables, te (4) SSRS integracija - integracija s Reporting Services pruža jednostavno okruženje u kojem prikazujemo rezultate rudarenja podataka krajnjim korisnicima.

Rudarenje podataka danas je najbrže rastuća grana u području poslovne inteligencije. Iako matematičke postavke postoje već dugo vremena, razvojem navedenih alata i njihovom integracijom s drugim sustavima ono izlazi iz domene naprednih korisnika i znanstvenog okruženja i približava se poslovnim ili ne-tehničkim korisnicima. Upravo ta zatvorenost i orijentiranost isključivo na napredne korisnike kočila je u prošlosti širu primjenu tehnologija rudarenja podataka.

Govoriti o nekom alatu, a ne dati nikakav primjer najčešće nema smisla, no napraviti "Hello, world" u rudarenju podataka je s druge strane nemoguće unutar ograničenja koja postavlja jedan ovakav članak. Iako se



**Poslovni proces rudarenja podataka korištenjem SQL Server 2008 Data Mininga**

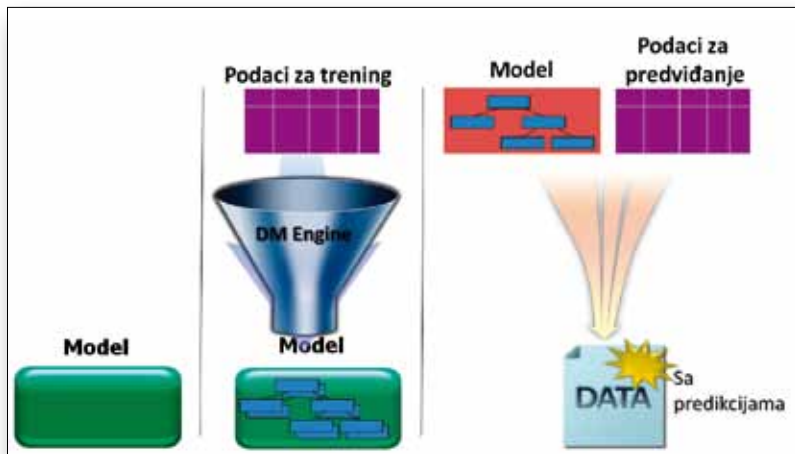
## ZADACI RUDARENJA PODATAKA

Prije nego što počnemo razmišljati o projektu implementacije rudarenja podataka, prvo treba stvoriti temeljni skup pitanja nužan za provedbu poslovne analize. Neka od tih pitanja mogu biti: je li navedeni klijent kvalitetan za odobravanje kredita (*credit scoring*), ukazuje li provedena transakcija kreditnom karticom na moguću krađu kartice (detekcija anomalija) ili je li prijava štete osiguravateljskoj kući pokušaj prevare (*fraud detection*).

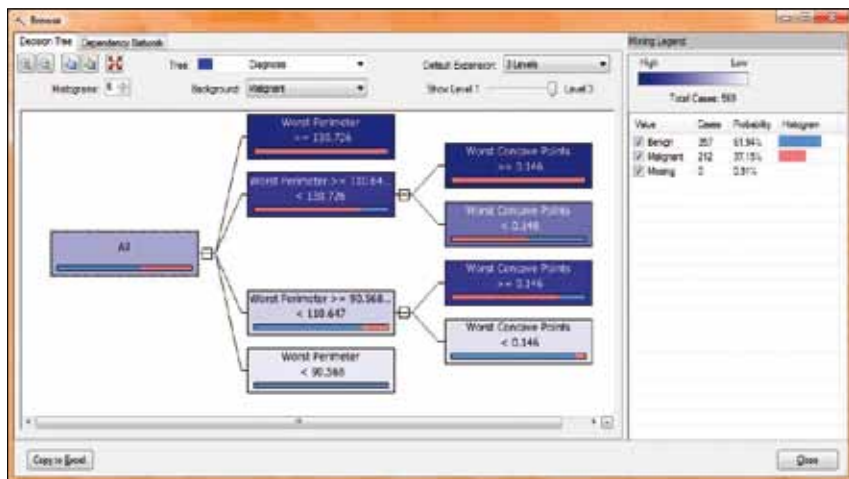
Za svako pitanje koje postavimo sustavu rudarenja podataka može se izvršiti više zadataka te tako doći do željenog odgovora. Iako je ponekad dovoljno upotrijebiti samo jedan zadatak, najčešće pak moramo kombinirati više njih da bismo došli do cilja. Kako je ovo iznimno opširna tema, u nastavku će biti dan pregled dvaju zadataka: klasifikacije i klasteriranja podataka. Svi ostali zadaci, kao i detalji vezani za njih, mogu se pronaći u literaturi navedenoj u zasebnom okviru.

Klasifikacija je vjerojatno najčešći zadatak koji se upotrebljava za analizu i sprečavanje odljeva postojećih korisnika, upravljanje rizicima i ciljano oglašavanje. Ona se sastoji od dodjeljivanja kategorija svakom promatranom slučaju. Svaki slučaj (npr. u bazama podataka jedan red u tablici) sastoji se od niza atributa (npr. u bazama podataka jedan stupac u tablici). Jedan atribut predstavlja klasni atribut, tj. onaj koji promatramo. Modeliranje klasifikacijom zahtijeva pronalaženje modela koji opisuje klasni atribut kao funkciju ulaznih atributa (npr. u bazama podataka drugih stupaca u tablici). Standardni klasiifikacijski algoritmi su stabla odluke, neuronske mreže i Naive-Bayes.

Klasteriranje - negdje zvano i segmentacija - koristi se za identificiranje prirodnih grupiranja slučajeva baziranih na skupu atributa. Slučajevi u istoj grupi imaju više ili manje slične vrijednosti atributa. Klasteriranje, za razliku od klasifikacije, spada u grupu nenadziranih zadataka rudarenja podataka gdje nema jedinstvenog atributa koji rabimo za izradu modela, već se sve ulazne varijable smatraju jednakima. Većina algoritama klasteriranja gradi model kroz više iteracija i zaustavljaju se tek kada model konvergira, tj. kada se stabiliziraju granice grupa.



Proces rudarenja podataka



Stablo odlučivanja izrađeno klasiifikacijskim alatom iz Data Mining klijenta za Excel 2007

to možda naizgled čini čudnim, u ovom je slučaju to više nego opravdano jer bismo pritom zanemarili neke od najbitnijih postavki rudarenja podataka: (1) pripremu i razumijevanje podataka, (2) postupak stvaranja poslovnih pitanja, (3) interpretaciju rezultata te još puno toga što čini jedan projekt rudarenja podataka. Za sve zainteresirane čitatelje i razvojne inženjere na stranicama SQL Server Data Mininga (<http://www.sqlserverdatami->

[ning.com/ssdm/](http://www.sqlserverdatami-ning.com/ssdm/)) moguće je pronaći gotove primjere koji pokazuje sve najvažnije aspekte korištenja ovog alata.

Stoga, ako ste se barem malo zainteresirali za ovo područje, nastavite s literaturom navedenom u okviru u kojoj ćete naći sve što vas zanima o ovom zanimljivom, ali i iznimno zahtjevnom području koje danas koriste mnoga velika poduzeća u ključnim područjima svog djelovanja te, vrlo vjerojatno, predstavlja

budućnost razvoja softvera. Imate li već SQL Server 2008, ili ga možete nabaviti, odvojite si vremena i proučite što vam on pruža na području rudarenja podataka. Mogli biste se ugodno iznenaditi onime što otkrijete. @

## Popis literature i korisnih web

<http://www.kdnuggets.com/>

Stranice koje okupljaju ljude koji se bave rudarenjem podataka

<http://www.microsoft.com/sqlserver/2008/en/us/data-mining.aspx>

Microsoftova glavna stranica za SQL Server 2008 Data Mining

<http://www.sqlserverdatamining.com/ssdm/>

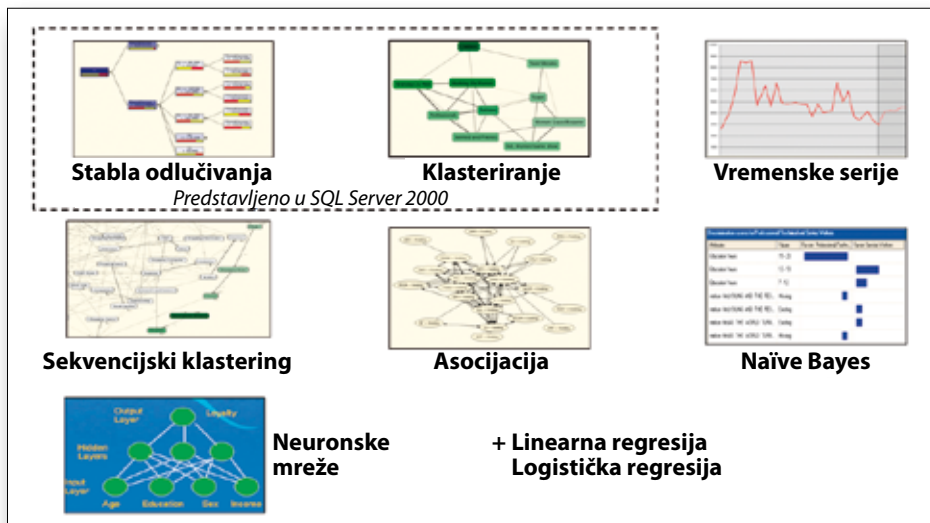
Stranica koja okuplja korisnike SQL Server 2008 Data Mininga i na kojoj se nalazi mnoštvo materijala i primjera

Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, 2nd edition, Morgan Kaufmann, 2006

Vjerojatno najbolja knjiga za metodološko proučavanje rudarenje podataka

Wiley & Sons - Data Mining With Microsoft SQL Server 2008

Vrhunska knjiga koja opisuje sve aspekte rudarenja podataka sa SQL Serverom 2008



Algoritmi implementirani u SQL Server 2008 Data Miningu