6. Vežbe

Data Mining

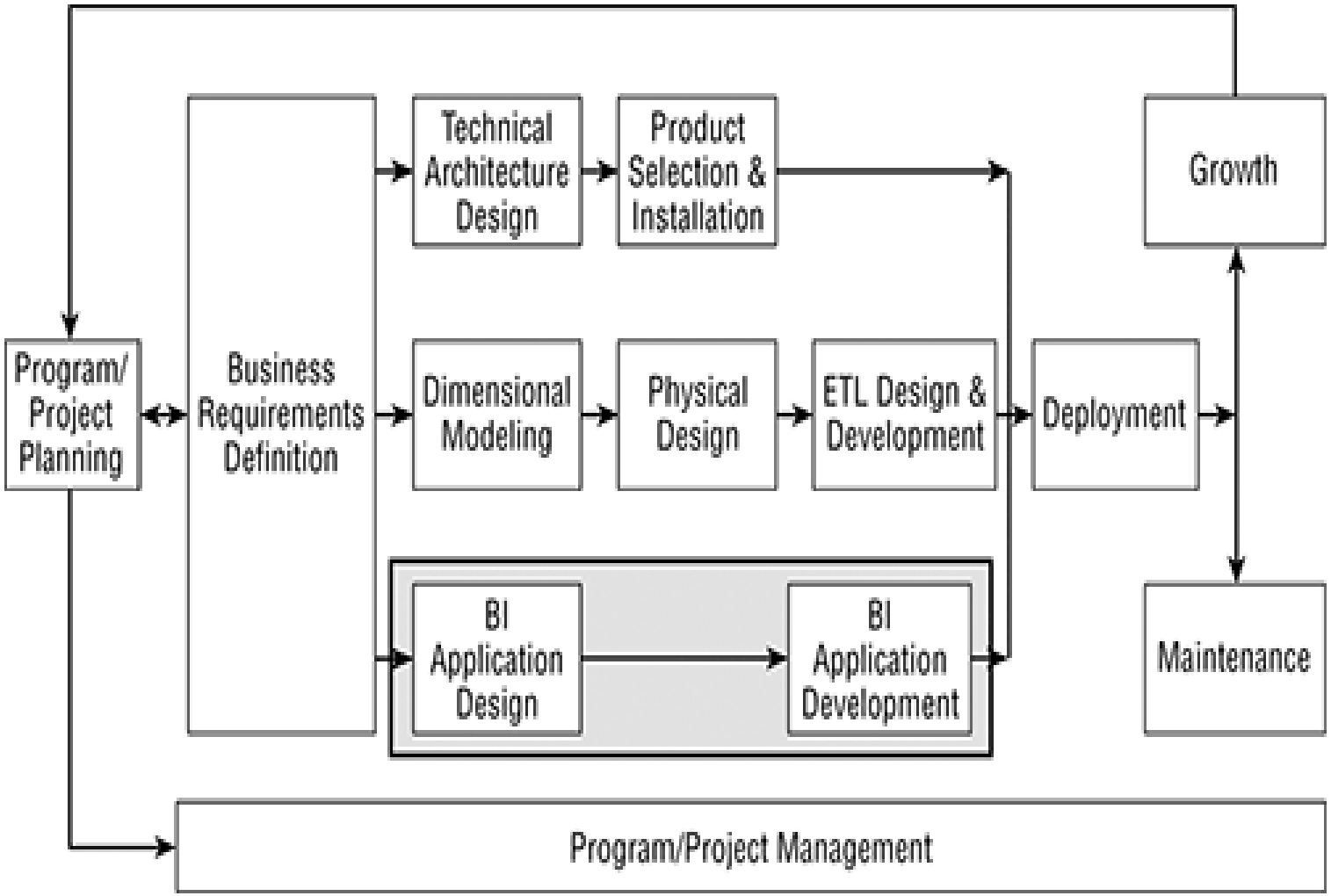
Izvođači laboratorijskih vežbi

- Marko Knežević (kancelarija TMD 9b)
- Nikola Obrenović

Termin konsultacija

- Marko Knežević: petak 15:00 TMD 9b
  marko.knezevic(AT)uns.ac.rs
- Nikola Obrenović
  nikob(AT)uns.ac.rs

Kimball's Lifecycle



Definicija data mining-a

- Proces istraživanja podataka u cilju pronalaženja šablona u podacima i korelacija između različitih skupova podataka koji imaju upotrebljivo domensko značenje
- Grupisanje pojava i predviđanje budućeg ponašanja
- Dva pristupa:
  - Istraživački (neusmereni) data mining
  - Usmereni data mining
- Kimbalova knjiga, poglavlje 13

Primeri data mining-a

- Detekcija prevara (krediti, platne kartice…)
- Grupisanje artikala koje se kupuju zajedno
- Identifikacije nepoželjne e-pošte
- Predviđanje trenda prodaje

Literatura

- Kimbalova knjiga, poglavlje 13
- http://www.sqlserverdatamining.com/ssdm/
- MS SSAS Data Mining Tutorial (MSDN): https://msdn.microsoft.com/en-us/library/bb677206%28v=sql.105%29.aspx

Osnovni DM zadaci

- Klasifikacija
- Estimacija
- Predikcija
- Asocijacija
- Klasterovanje
- Detekcija anomalija

Klasifikacija

- Dodeljivanje pojave u jedan od predefinisanih skupova (klasa) na osnovu osobina pojave
- Određivanje vrednosti diskretne promenjive (klase kojoj pojava pripada)
- Primer klasa: obični, srebrni i zlatni klijenti
- Metode u MS SSAS:
  - Microsoft Decision Trees,
  - Microsoft Neural Network i
  - Microsoft Naïve Bayes

Estimacija (regresija)

- Kontinualna verzije klasifikacije, tj. određivanje vrednosti kontinualne promenjive
- Primer: procena verovatnoće da će kupac prihvatiti promociju proizvoda na osnovu ranijih kupovina i promocija
- Metode u MS SSAS:
  - Microsoft Decision Trees i
  - Microsoft Neural Network

Predviđanje

- Estimacija ili klasifikacija za budući trenutak
- Primer: procena kretanja vrednosti nekretnine u budućnosti
- Metode u MS SSAS:
  - Microsoft Decision Trees,
  - Microsoft Neural Network i
  - Microsoft Time Series

Asocijacije

- Pronalaženje korelacija između grupa pojava
- Primer: Would you like to buy this, too?
- Metode u MS SSAS:
  - Microsoft Association i
  - Microsoft Decision Trees

Klasterovanje

- Grupisanje pojava u unapred nepoznate grupe
- Element jednog klastera treba da bude što sličniji elementima istog klastera i što više različit od elemenata ostalih klastera
- Klasterovanje sekvenci događaja. Npr. redosled posete stranica na web sajtu
- Metode u MS SSAS:
  - Microsoft Clustering i
  - Microsoft Sequence Clustering

Detekcija anomalija

- Detekcija pojava koje odstupaju od svih ostalih
- Drugi naziv: Outlier detection
- Koriste se već navedeni algoritmi:
  - klaster koji ima samo jedan element
  - Pojava koja ne može biti klasifikovana ni u jednu klasu

Potrebna znanja i veštine za DM

- Dobar osećaj posla i saradnja sa poslovnim ljudima
- Odlično poznavanje SSAS i SQL
- Dobro razumevanje statistike i verovatnoće
- Iskustvo u data miningu
- Veština programiranja

DM u SSAS

- Osnovni elementi DM projekta:
  - Izvor podataka
  - Pogled izvora podataka
  - Mining struktura
  - Mining model

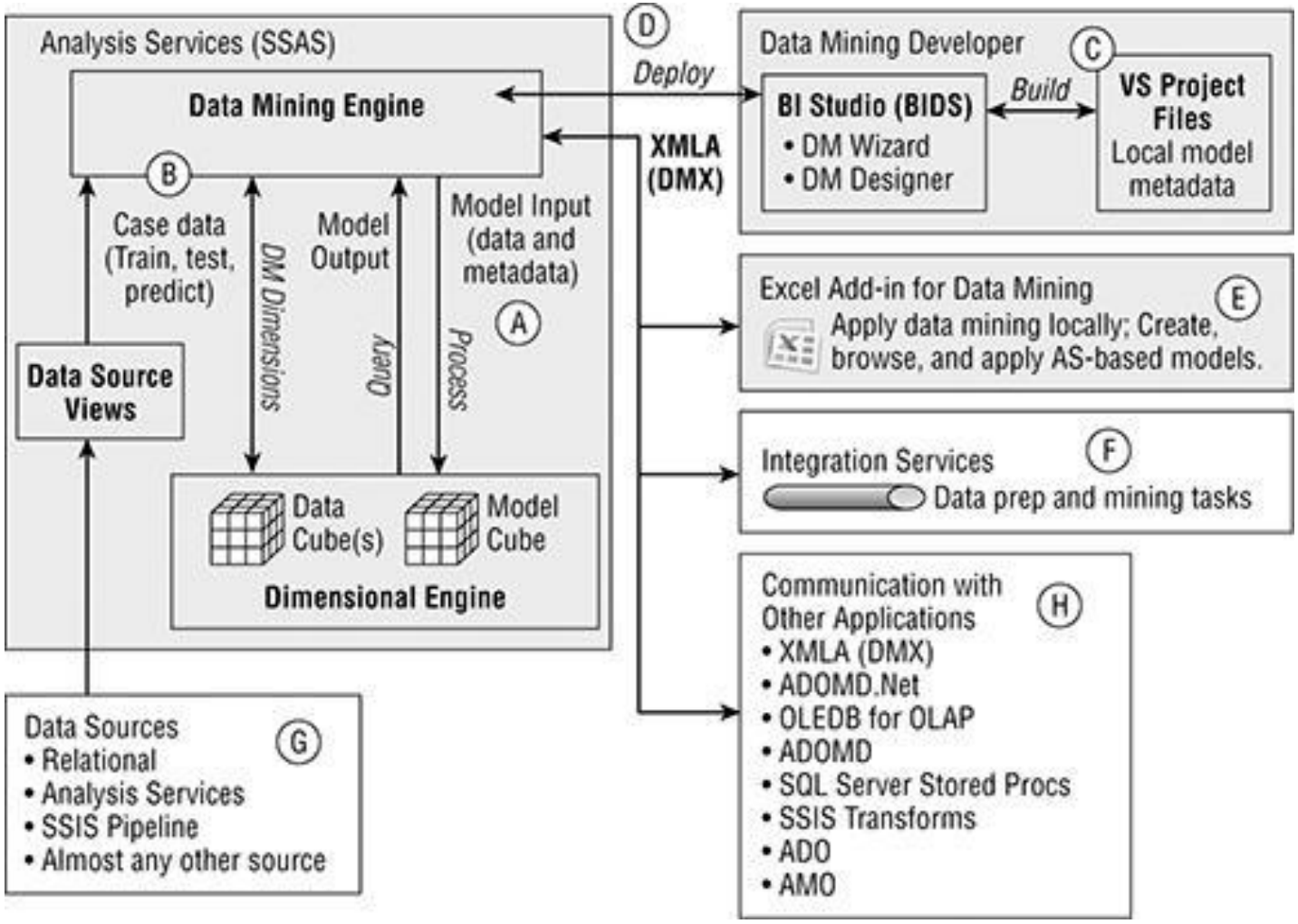
DM u SSAS

- Data Mining Wizzard služi da se kreira DM struktura i model
- Data Mining Designer
  - kreiranje i validacija modela
- Data Mining eXtensions to SQL language (DMX) je API za kreiranje, obuku, izmenu i postavljanje upita nad DM modelima

SSAS DM primeri

- AdventureWorksDW 2008 R2:
  - Opis: https://technet.microsoft.com/en-us/library/ms124623%28v=sql.105%29.aspx
  - Download AdventureWorksDW2008R2_Data.mdf: http://msftdbprodsamples.codeplex.com/releases/view/59211
- SSAS projekat:
  - AdventureWorks 2008R2 Analysis Services Project.zip
  - Skinuti sa ACS sajta

SSAS Arhitektura



DM terminologija u SSAS

- *Algorithm:* The programmatic technique used to identify the relationships or patterns in the data.
- *Model:* The definition of the relationship identified by the algorithm, which generally takes the form of a set of rules, a decision tree, a set of equations, or a set of associations.
- *Case:* The collection of attributes and relationships (variables) that are associated with an individual instance of the entity being modeled, usually a customer. The case is also known as an observation.
- *Case set*: A group of cases that share the same attributes. Think of a case set as a table with one row per unique object (like *customer*). It's possible to have a nested case set when one row in the parent table, like "customer," joins to multiple rows in the nested table, like "purchases." The case set is also known as an observation set.
- *Dependent variable(s)* (or predicted attribute or predict column): The variable the algorithm will build a model to predict or classify.

DM terminologija u SSAS

- *Independent variable(s)* (or predictive attribute or input column): The variables which provide the descriptive or behavior information used to build the model. The algorithm creates a model that uses combinations of independent variables to define a grouping or predict the dependent variable.
- *Discrete or continuous variables:* Numeric columns that contain continuous or discrete values.
  - Early data mining and statistical analysis tools required the conversion of strings to numeric values like the encoded salary ranges.
  - Most tools, including most of the SQL Server data mining algorithms, allow the use of character descriptions as discrete values. The string “0 to $25,000” is easier to understand than the number 1.
  - Discrete variables are also known as categorical. This distinction between discrete and continuous is important to the underlying algorithms in data mining, although its significance is less obvious to those of us who are not statisticians.

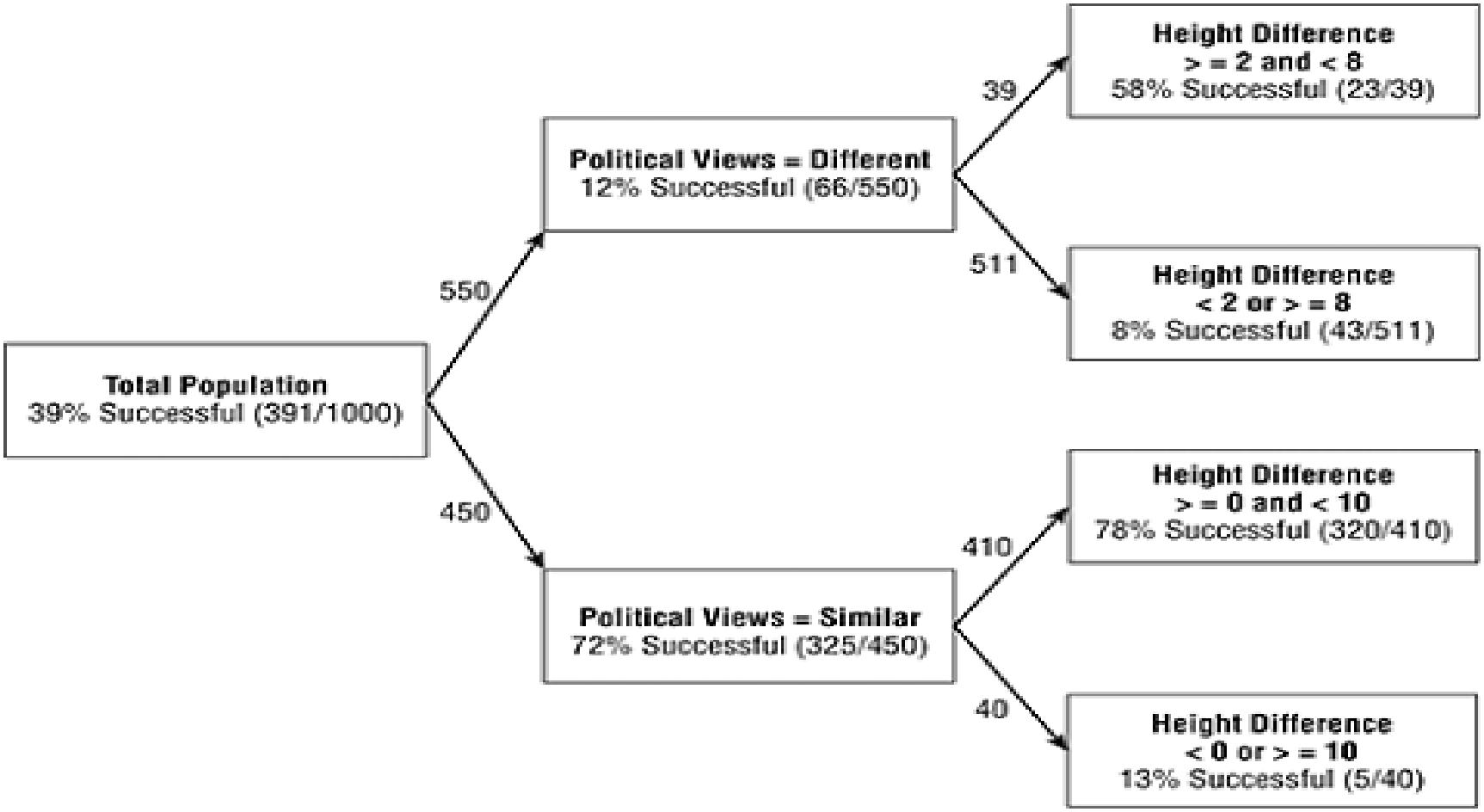
DM terminologija u SSAS

- *Regression:* A statistical technique that creates a best-fit formula based on a data set. The formula can be used to predict values based on new input variables. In linear regression, the formula is the equation for a line.
- *Deviation:* A measure of how well the regression formula fits the actual values in the data set from which it was created.
- *Mining structure:* A Microsoft data mining term used as a name for the definition of a case set in Analysis Services. The mining structure is essentially a metadata layer on top of a Data Source View that includes additional data mining–related flags and column properties, such as the field that identifies a column as input, predict, both, or ignore. A mining structure can be used as the basis for multiple mining models.
- *Mining model:* The specific application of an algorithm to a particular mining structure. You can build several mining models with different parameters or different algorithms from the same mining structure.

Microsoft Decision Trees

- Može se koristiti i za kontinualne i za diskretne promenjive – klasifikacija, estimacija i predviđanje
- Pronalazi korelacije između ulaznih i ocenjivanih atributa
- Gradi se binarno stablo
- U svakom čvoru bira se nezavisna promenjiva koja deli skup pojava na dva što različitija podskupa
- Za svaki ocenjivani atribut, algoritam pravi posebno stablo
- Primer: Targeted Mailing.dmm

Stablo odlučivanja



Microsoft Naive Bayes

- Za klasifikaciju i predviđanje diskretnih promenjivih
- Koristi samo diskretne atribute i podrazumeva da su nezavisni
- Zasniva se na uslovnim verovatnoćama i Bajesovoj teoremi

$$p(C|F_1,\ldots,F_n) = \frac{p(C)\ p(F_1,\ldots,F_n|C)}{p(F_1,\ldots,F_n)}.$$

Microsoft Naive Bayes

- Verovatnoća da će klasifikacioni atribut *C* uzeti neku vrednost *c*, pod uslovom da nezavisni atributi F_i imaju vrednosti f_i iznosi:

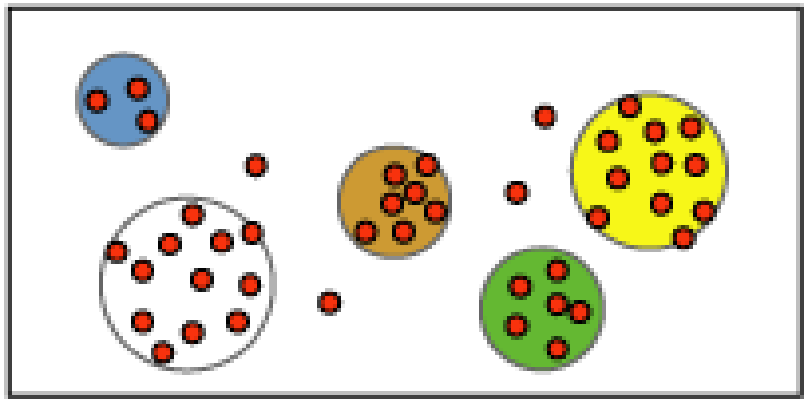
$$p(C|F_1,\ldots,F_n) = \frac{1}{Z}p(C)\prod_{i=1}^{n}p(F_i|C)$$

$$Z = p(F_1,\ldots,F_n)$$

- Primer: Targeted Mailing.dmm

Microsoft Clustering

- Otkrivanje grupa u podacima
- Algoritmi:
  - Expectation-Maximization – svaka tačka pripada svim klasterima sa određenom verovatnoćom, i
  - k-Means – tačka pripada samo jednom klasteru
- Primeri: Customer Mining.dmm



Microsoft Time Series

- Predikcija kontinualnih promenjivih tokom vremena
- Predikcija promenjive se vrši samo na osnovu ranijih vrednosti iste promenjive
- Kratkoročna predikcija: ARTxp algoritam
- Dugoročna predikcija: Podaci se predstavljaju putem ARIMA modela i koriste se algoritmi specijalizovani za takve modele
- Microsoft TS: kombinacija ARTxp i ARIMA algoritama
- Podržano je modelovanje sezona
- Algoritmi detektuju i koriste korelacije između atributa
- Primer: Forecasting.dmm

Microsoft Association

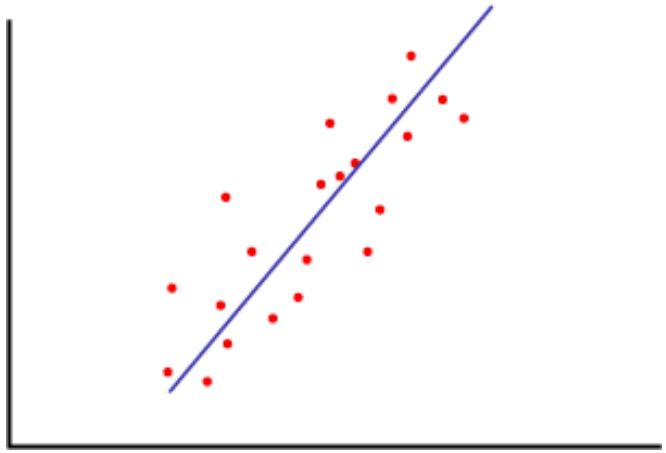
- Market basket analysis
- Asocijacije rade dobro sa ugnježdenim skupovima podataka, npr. transakcije i stavke transakcije
- Algoritam pronalazi skupova pojava koje često idu zajedno i pravi sledeća pravila:
  $A_1, \ldots, A_n \rightarrow C, P(C)$
- Parametri algoritma su:
  - *Minimal support*: minimalni broj pojava da bi pravilo bilo razmatrano
  - *Miminal probability*: minimalna verovatnoća da bi pravilo bilo usvojeno
- Primer: Market Basket.dmm

Microsoft Neural Network

- 3-nivovska mreža:
  - broj ulaznih čvorova je određen brojem ulaznih atributa
  - broj izlaznih čvorova je određen brojem atributa koji se klasifikuju ili predviđaju
- Parametri algoritma definišu broj čvorova u skrivenom sloju, veličinu uzorka za obuku mreže, procenat i način izbora pojava za evaluaciju modela, itd.

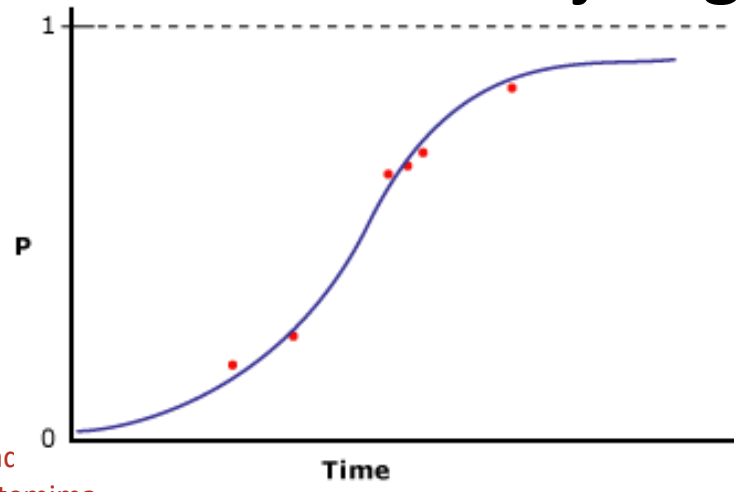
Microsoft Linear Regression

- Varijanta MS Decision Tree algoritma
- Pronalazi linearnu zavisnost između izlazne, zavisne promenjive i nezavisnih, ulaznih promenjivi
- Promenjive moraju biti kontinualne i numeričke

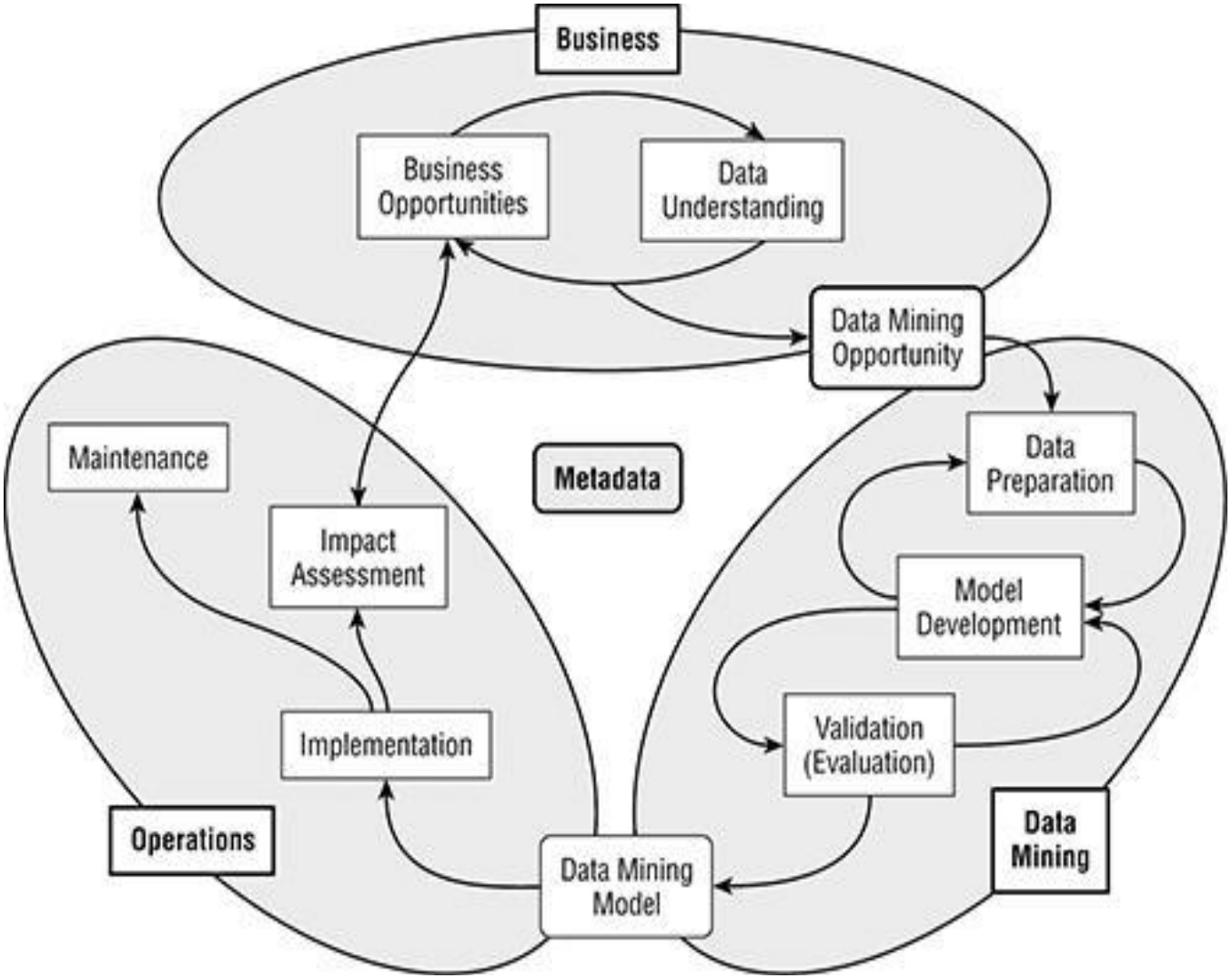


Microsoft Logistic Regression

- Varijanta MS Neural Network algoritma
- Predikcija promenjive sa dva stanja (npr. Bulova promenjiva)
- Vrednost izlazne promenjive je između 0 i 1
- Broj i tip ulaznih čvorova nije ograničen



Proces data mining-a



Poslovna faza

- Prikupljanje zahteva (*business opportunities*)
  - Precizna definicija i priotetizacija zahteva
  - Poslevne zahteve preslikati na podatke i algoritme
- Razumevanje podataka
  - Pronaći odgovarajuće izvore podataka za DM modele
  - Proceniti upotrebnu vrednost - da li su podaci potpuni i čisti

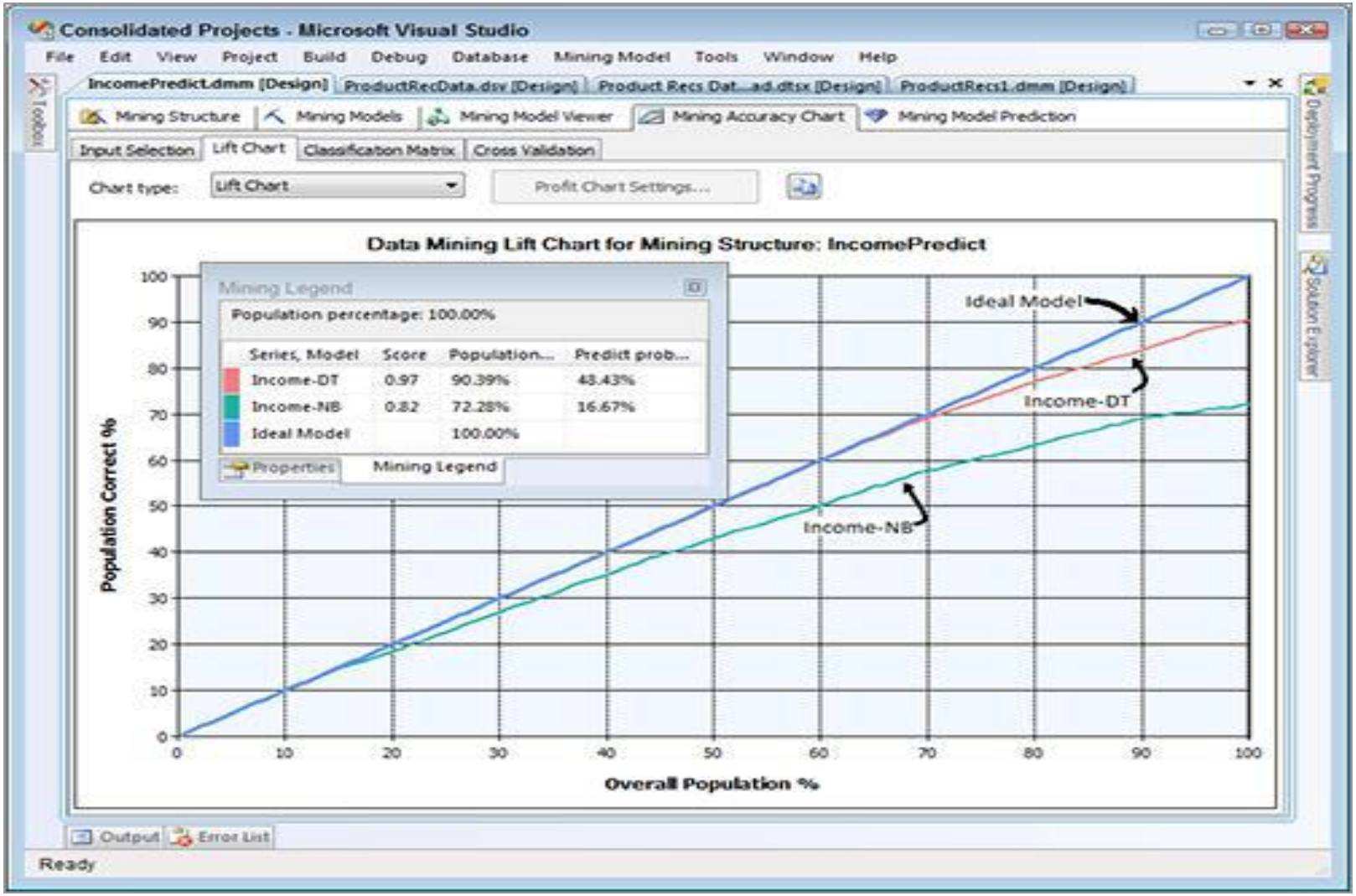
Data mining faza

- Priprema podataka
  - čišćenje nevalidnih podataka, konverzija podataka i generisanje nedostajućih podataka
  - proširenje postojećeg ETL procesa
  - pravljenje posebnog ETL procesa i posebnih tabela iz kojih će se graditi DM model
- Kreiranje trening i test skupa podataka
  - Držati ga odvojeno od tabela činjenica i dimenzija

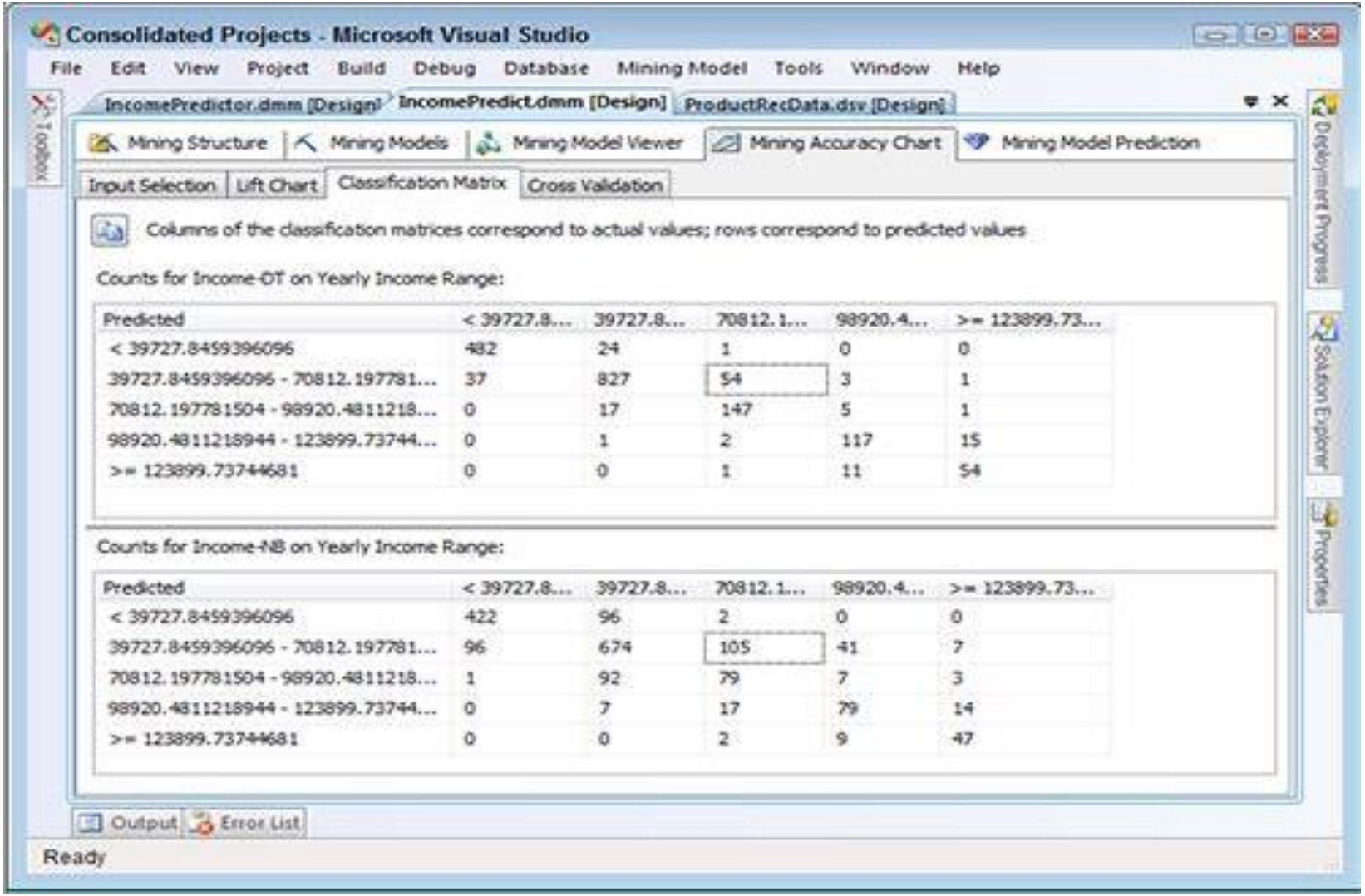
Data mining faza

- Razvoj DM modela
  - izbor DM algoritma i podešavanje njegovih parametara
- Evaluacija DM modela i upoređivanje kvaliteta rezultata različitih modela
  - lift chart
  - klasifikaciona matrica
  - unakrsna validacija

Lift chart


Consolidated Projects - Microsoft Visual Studio
File Edit View Project Build Debug Database Mining Model Tools Window Help
IncomePredict.dmm [Design]
ProductRecData.dsv [Design]
Product Recs Dat...ad.dtsx [Design]
ProductRecs1.dmm [Design]
Mining Structure
Mining Models
Mining Model Viewer
Mining Accuracy Chart
Mining Model Prediction
Input Selection
Lift Chart
Classification Matrix
Cross Validation
Chart type:
Lift Chart
Profit Chart Settings...
Data Mining Lift Chart for Mining Structure: IncomePredict
Mining Legend
Population percentage: 100.00%
Series, Model Score Population... Predict prob...
Income-DT 0.97 90.39% 48.43%
Income-NB 0.82 72.28% 16.67%
Ideal Model 100.00%
Properties
Mining Legend
Ideal Model
Income-DT
Income-NB
Population Correct %
Overall Population %
0 10 20 30 40 50 60 70 80 90 100
Output
Error List
Ready

Klasifikaciona matrica



Unakrsna validacija

- Ulazni skup skup podataka se podeli u disjunktne podskupove
- Za svaki podskup i svaki algoritam, generiše se poseban model
- Uporedjuje se preciznost različitih modela u okviru istog algoritma
- Za svaki algoritam postoji skup metrika kojima se meri preciznost modela
- Trenutno ograničeno na klasifikaciju/predviđanje samo jedne promenjive