CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

FACULTY OF ECONOMICS AND MANAGEMENT

DEPARTMENT OF INFORMATION TECHNOLOGIES

ČESKÁ
ZEMĚDĚLSKÁ
UNIVERZITA V PRAZE

BACHELOR THESIS

DATA MINING

Author: Arezoo Noushiravan

The Bachelor Thesis Supervisor: Ing. Martin Havránek, Ph.D.

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

BACHELOR THESIS ASSIGNMENT

Arezoo Noushiravan

Informatics

Thesis title

Data Mining

Objectives of thesis

The bachelor thesis is thematically focused on an in-depth coverage of the principles of data mining.
The main goal of this bachelor thesis is to understanding the concepts of data mining and data mining techniques, the process of managing and extracting data, analyzing and establishing hidden connections and patterns in a group of data sets from different perspectives and summarizing it into useful information.
Partial goals of the thesis are:
- to provide an introduction to the multidisciplinary field of data mining
- to describe data mining tasks
- to describe data mining techniques
- to analyze practical applications of data mining

Methodology

Methodology of the thesis is based on study and specialized information resources. The practical part aim is to analyze practical applications of data mining which will be used in business intelligence, industries, computer security, etc. Based on synthesis of theoretical knowledge and results of author's own work, the conclusions of the thesis will be formulated.

The proposed extent of the thesis

40 – 60 pages

Keywords

Data Minig, Data Mining Techniques, Data Mining Classification, Data Mining Clustering, Association Rule Learning, Sequential Pattern Mining, Regression Analysis, Deviation Detection, K-Nearest Neighbors, Automatic Text Summarization

Recommended information sources

1. Jiawei Han, Micheline Kamber, Jian Pei, Data Mining: Concepts and Techniques, 3rd Edition, USA: Morgan Kaufmann Publishers, 2011, ISBN 978-0-12-381479-1.
2. Anand Rajaraman, Jure Leskovec, Jeffrey D. Ullman, Mining of Massive Datasets, [Online], Available at WWW:<http://infolab.stanford.edu/˜ullman/mmds.html>.
3. David J. Hand, Heikki Mannila, Padhraic Smyth, Principles of Data Mining, USA: The MIT Press, 2001, ISBN: 978-0262082907.
4. Ian H. Witten , Eibe Frank , Mark A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition, USA: Morgan Kaufmann Publishers 2011, ISBN: 978-0123748560.
5. Sang C. Suh, Practical Applications Of Data Mining, USA: Jones & Bartlett Learning, LLC, 2011, ISBN: 978-0-7637-8587-1

Expected date of thesis defence

2015/16 WS – FEM

The Bachelor Thesis Supervisor

Ing. Martin Havránek, Ph.D.

Supervising department

Department of Information Technologies

Electronic approval: 31. 10. 2014

Ing. Jiří Vaněk, Ph.D.

Head of department

Electronic approval: 11. 11. 2014

Ing. Martin Pelikán, Ph.D.

Dean

Prague on 30. 11. 2015

Declaration

I declare that I have worked on my bachelor thesis titled "Data Mining" by myself and I have used only the sources mentioned at the end of the thesis. As the author of the bachelor thesis, I declare that the thesis does not break copyrights of any third person.

In Prague on 30.11.2015

Arezoo Noushiravan

Acknowledgement

I would like to thank my Bachelor Thesis supervisor Ing. Martin Havránek, Ph.D., for his advice and support during my work on this thesis.

Dedicated to my husband Hamid Hashemi and my brother Behnam Noushiravan, for all of their support, inspiration, and love.

DATA MINING

Abstract

This thesis is concerned with Data Mining concepts, generally data mining (sometimes called knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information so that extracted information or discovered knowledge can be used to increase revenue, cuts costs, or both. Nowadays Data mining is primarily used by companies with a strong consumer focus - retail, financial, communication, and marketing organizations, Data mining enables companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics, data mining also enables companies to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data. The initial phase of a data mining project focuses on understanding the project objectives and requirements. Business perspective any interesting problem can be formulated as a data mining problem and develop a preliminary implementation plan. For example, a problem from business perspective might be: "How can I sell more of my product to customers?" this problem might be translated into a data mining problem such as: "Which customers are most likely to purchase the product?" A model that predicts who is most likely to purchase the product must be built using analyzing a specific data which may be included the customers statistics who have purchased the product in the past.

Keywords

Data Mining, Data Mining Techniques, Data Mining Tasks, Data Mining Prediction, Data Mining Methods, Data Mining Classification, Clustering Technique, Decision Tree, K-nearest Neighbors algorithm, Regression algorithm, K-Means as a Centroid-Based Technique, C4.5 algorithm, Association Rule, Apriority algorithm.

Abstrakt

Tato bakalářská práce je zaměřená na data miningu. Data mining je analytická metodologie získávání netriviálních skrytých a potenciálně užitečných informací z dat. Klasifikace a Clustering (Shluková analýza) jsou dvě techniky data miningu, Kde klasifikace je ve strojovém učení a statistice druh problému, když máme určit, do které z kategorií dat dané pozorování patří, a shluková analýza je vícerozměrná statistická metoda, která se používá ke klasifikaci objektů. Slouží k třídění jednotek do skupin (shluků) tak, aby si jednotky náležící do stejné skupiny byly podobnější než objekty z ostatních skupin.

Data mining se používá v komerční sféře (například v marketingu při rozhodování, které klienty oslovit dopisem s nabídkou produktu), ve vědeckém výzkumu (například při analýze genetické informace) i v jiných oblastech (například při monitorování aktivit na internetu s cílem odhalit činnost potenciálních škůdců).

Úvod je zaměřen na vymezení základních pojmů z teorie data miningu a popsány jeho základní techniky. V druhé části jsou za pouţití rozhodovacích stromů.

Klíčová Slova

Data Mining, Techniky Dolování Dat, Data Mining Tasks, Data Mining Prediction, Metody Data Mining, Klasifikace Data Miningu, Clusterová Analýza, Rozhodovacích Stromů,, K-nearest Neighbors algorithm, Regression algorithm, K-Means as a Centroid-Based Technique, C4.5 algorithm, Association Rule, Apriority algorithm.

Table of Contents

What is Data? 5

Data mining definition 6

Data mining model 7

Data mining tasks 13

Data Mining Algorithms 15

- C4.5 algorithm 15
- K-Means: A Centroid-Based Technique 19
- Apriority algorithm 22
- Regression algorithm 26
- K-nearest neighbors algorithm 27

Practical Example of the Decision Trees 29

- Practical Example 29

Conclusion 41

References 42

What is Data?

Data is a pieces of information which are stored in a variety of form such as number, text.

Data in many industry *can* generate the most revenues, so data analyses is a challenge to companies. The huge volume of data in a different formats are collected during organization activities and these data are analyzed to extract used information, therefore these useful information can be caused to drive organization with better business decision.

Data is typically analyzed using a variety of data mining software and its results allows the company applies these results in order to develop their business. For example France's orange as a biggest mobile carries in the world has collected data anonymously about 2.5 billion records, data such as call details and text messages from 5 million users, the France's orange uses data in order to develop companies project to keep public health and security or improve public security requirement, as well as these data also can be used to predict business requirement to find answers for business question to improve company operation to develop products.

Another benefits and services of data mining can be decoded DNA in minutes, predict where terrorists plan to attack, indicate which products has more profit to the company, which gene has effect for certain diseases, to analyses customer behavior, find out customers interests.

As far as data is increasingly collected useful information is extracted through data mining software, these data mining would be more effective way to utilize data by markets. (1)

Types of Variable

There are many types of variable that can be used to measure the properties of an object. These without having any knowledge about various types of variable can lead to problems with any form of data analysis. Main types of variable are in following:

Nominal Variables: Value of nominal variable is contained from a specific category and it means object is belongs to the specific category, category can be text value or numerical value and numerical values have no mathematical interpretation. For example level of education can be stored as numbers between 1 and 5.

Binary Variables: A binary variable (a special type of a nominal variable) can be contained only two possible values, yes or no, true and false, 1, 0 and etc.

Ordinal Variables: similar to nominal variables, ordinal values are contained from a specific meaningful order category, for example level of education can be stored such as excellent, very good, and good and failed values as a ordinal values.

Integer or numerical Variables: Integer variables contains only numbers, for example amount of credit cart.

Interval-scaled Variables: The interval type is the degree of difference between objects and takes only integer values, but the value cannot mean the ratio between the object. The most common example for Interval-scaled Variables **is** temperature with the Celsius scale, temperature that is separated into 100 intervals (point from the freezing and boiling point of water). (2)

Data mining definition

Easy to save data on inexpensive disks or CDs, easy to record data online causes the world to be overwhelmed with data and seems ever-increasing and there's no end in sight. So data mining is invented to make this data useful. Generally, Data mining is the process of Knowledge discovery within the collected data to analyze data from different dimensions or angels. Let's make it clear.

Data mining is able to answer many business questions, allows companies drive its business with a better business decisions while without data mining making decision can be time consuming for companies. For instance more often in marketing relation between goods can be caused to produce more profit to the company, so finding correlations or patterns between goods from data is requested from analyzer. For example 'Which products have been sold together' can be requested from analyzer. Products name and its brand from list of products which have been bought by customer is analyzed.

Fertilization is contained combination of woman's ovaries with sperms, after fertilization several embryos are produced in woman's uterus. All these operation result a live child in woman's uterus so that the "best" embryos is selected based on around 60 recorded features of embryos and is causes a live child. Embryos features can be like their morphology, follicle, sperm sample and the oocyte.

Each year, near the end of the milking season when the feed is used up dairy Farmers have to select cows based on breeding and milk production history in order to sell to the abattoir. Factors include age (its productive life at eight years), health problems, and history of difficult calving, undesirable temperament traits, and not being pregnant with calf for the following season are also recorded as the attributes for each of million cows over the years.

Data mining tools and techniques are used in a wide range of industries including retail, finance, heath care, manufacturing transportation to take advantage of historical data.

So data mining or knowledge discovery is a process to predict behaviors and future trends, exceptions, significant facts, relationships, patterns from historical data in database. (3)

Specific uses of data mining include:

Fraud detection – helps to identify which transactions are most likely to be fraudulent.

Market basket analysis – identify which products or services are commonly sold together. For instance bread and butter.

Trend analysis - demonstrate future movement of typical customers for instance in consecutive two month (what has happened in the past what will happen in the future).

Interactive marketing - Predicts which product to each individual customer is interested that are most likely in seeing company's website.

Direct marketing - Identify which advantages should be included in a mailing list to obtain the highest response rate from the customers.

Forecasting: Estimating sales

Risk and probability: Choosing the best customers for targeted mailings, for example choosing customers who have high probability to sell a specific product or determining the customers how have high probability of risk to the targeted mailings.

Recommendations: Determining which products are likely to be sold together, generating recommendations to the customers.

Finding sequences: Analyzing customer selections in a shopping cart to predict next likely product so that by finding sequences the company can offer a specific product with a sale on products which the costumers likely have plan to buy it.

Grouping: Separating customers into cluster based on a specific product in order to offer new company technologies. (4)

Data mining model

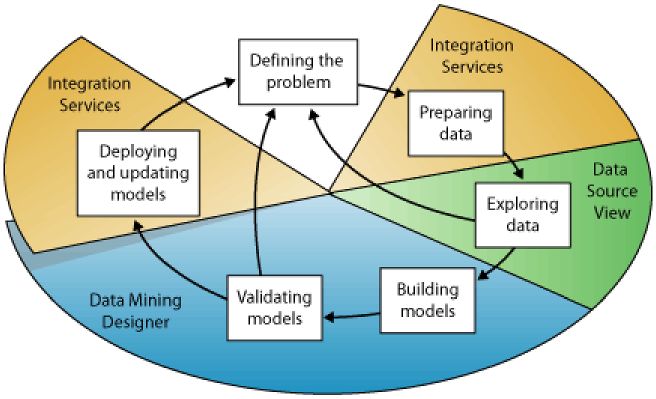
Building a mining model is part of a data mining process that starts from defining the problem to applying the model into a working environment. (5)

The following six basic steps demonstrate how to build a data mining model:

1. Defining the Problem
2. Preparing Data
3. Exploring Data
4. Building Models

5. Exploring and Validating Models
6. Deploying and Updating Models

In each step the diagram demonstrates the relationships between each step to nest and previous step.



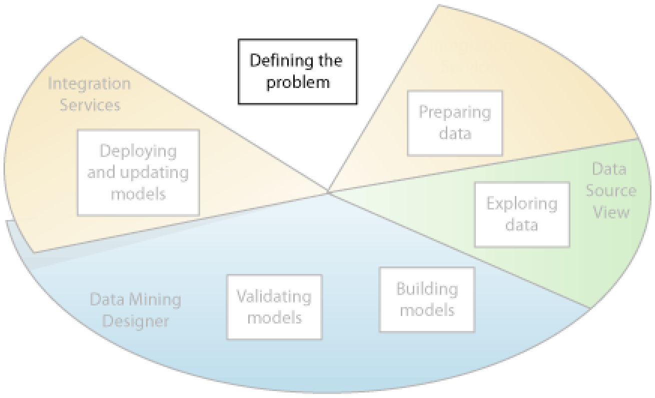
As above diagram illustrates, creating a data mining model is a dynamic and iterative process, for instance after preparing the data, analyzer finds out data is not sufficient to create the appropriate mining models, and analyzer needs to choose data column that can be used therefore analyzer needs to look for more data, or analyzer may build several models and then realize that the models do not answer the problem have been defined, and therefore analyzer must redefine the problem. Data mining model needs to update after deploying because data is increasing all the time. Each step in the process might need to be repeated many times in order to create a best model. (5)

Defining the Problem

The first step is to define the problem and its scope quite clearly, analyzer considers how data can be used to provide an appropriate answer to the problem. Analyzer tries to analyze business requirement and define and purpose of data mining model, specifies the metrics that evaluates the data mining model.

These above tasks transform into questions such as:

- What is must find out? What types of relationships is needed to find?
- Does the problem was defined reflects the policies or processes of the business?
- Predictions from the model or interesting patterns is needed to extract?
- Which outcome or attribute is needed to predict?
- What kind of data do you have and what kind of information is in each column? If there are multiple tables, how are the tables related? Do you need to perform any cleansing, aggregation, or processing to make the data usable? (5)

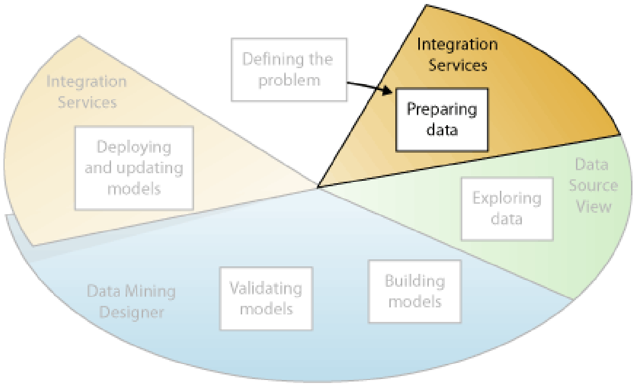


Preparing Data

The second step in the data mining process, is to clean the data that was identified in the Defining the Problem step.

Data can be stored in different formats or can be included with incorrect or missing values. For example, the data might show that a customer bought a product before the product was offered on the market or value of income column have not been stored for some of customers.

In addition to removing incorrect or missing values, determining columns are the most appropriate for use in analysis and finding data that are most accurate, for example total cost of the product or price of the product? Total price or a discounted price influences the sales? Sometimes incomplete data also can influence the results of the model in ways that is not normally expected. (5)

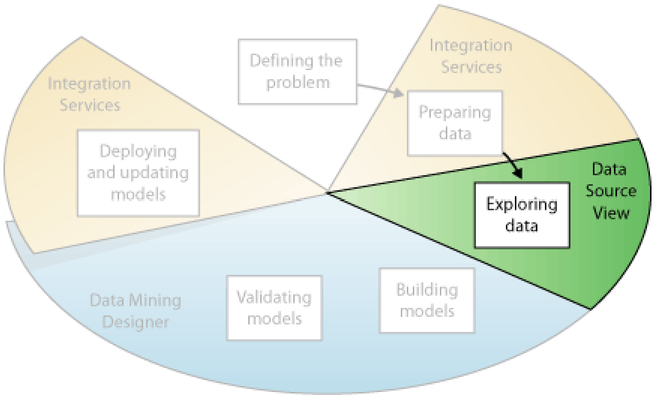


Exploring Data

The third step is to explore the data which has been extracted from database, statistical tools are used to explore data, for instance statistical summaries such as calculating mean and mode, median, standard deviations, and exploration of the distribution of the data, calculating the minimum and maximum values provide more useful information about dataset. So if the maximum, minimum, and mean values that were determined does not represent business processes, analyzer have to balance dataset to build the model.

Let’s make it clear, standard deviation might produce useful information about the consistency and accuracy of the exploration. For example a large standard deviation might illustrate adding more data might result better model. By exploring dataset analyzer devise a strategy for fixing if the dataset contains flawed data or obtain a figure of the behaviors of business in company.

Therefore exploration helps to analyzer analyze the distribution of data and repair issues such as wrong or missing data if there is. (5)



Building Models

The fourth step is to build the mining model. The knowledge that have been obtained in the Exploring Data step to help define and create the models.

Analyzer defines the columns of data that is used to create a mining structure. This structure is linked to the source of data, but actually contain data when the model is applied. When the mining structure is applied, Analysis Services provide statistical information that can be used for analysis. This information can be used by any mining model that is based on the structure.

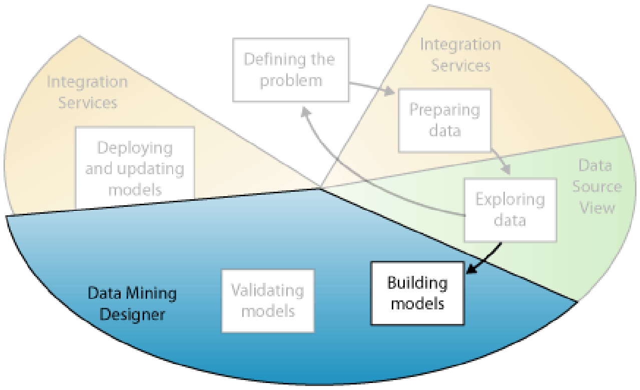
Before the mining structure is applied, the structure includes the columns that is used for entering data, the attribute that should be predicted, and parameters that the algorithm uses in order to specify how data should be processed, in other word to adjust each algorithm.

Processing a model is called training. Training is applying a specific mathematical algorithm to the dataset in the structure in order to discover useful patterns, the patterns that is provided during training process depend on:
How analyst has specified or defined the algorithm,
Type of training data, the algorithm have been chosen by analyst.

As well as analyst apply filters to the training dataset to use just a subset of the data to create different results. After passing data through the model, summaries and patterns will be queried or used to predict future trend.

There are many different algorithms, each algorithm is applied to a different type of task, and each create a different type of data mining model. For instance Clustering algorithm is used to analyze customers based on searching and selling patterns, or Decision Trees algorithm is used to identify customers for suggesting extra products for purchase.

As the data changes, the mining structure is updated, the mining structure retrieves dataset from the source. If the models is based on the structure, the models is updated based on retrained on the new data, the models stays as is. (5)



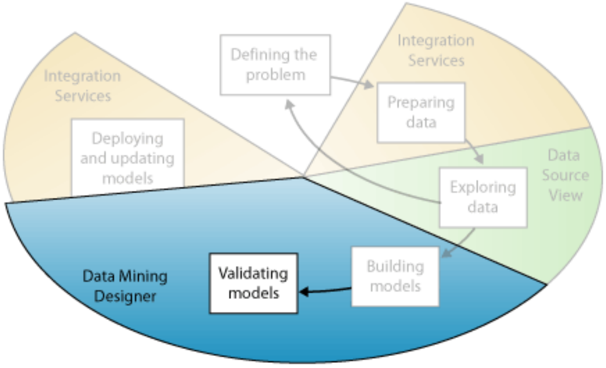
Validating Models

The fifth step is to explore the mining models that have been built and is to test their effectiveness.

Before expanding and applying the model into a production environment, analyst might test how well the model works. Typically multiple models with different configurations are created and the only thing that is must done is to test all models to find out which provides the best results for defined problem and dataset.

There are two type of dataset which are called “training dataset” and “testing dataset”, training dataset is used to build the mining model and testing dataset is used to test accuracy of the extracted model by constructing prediction queries

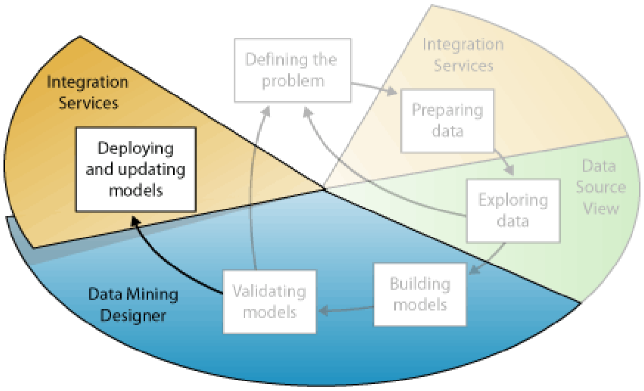
If the models that are constructed in this step does not work well, analyzer needs to return to a first step and redefine the problem or reconsider the data in the original data. (5)



Deploying and Updating Models

The last step in the data mining process, is to apply the models to provide any prediction or extract any pattern to a production environment. The following are some of the tasks that is performed to update the mining model or for validating the mining model.

- Models is used to create predictions in order to make business decisions.
- Update the models after review and analysis. Any update requires that you reprocess the models.
- Update the models dynamically, as more data comes into the organization, and making constant changes to improve the effectiveness of the solution should be part of the deployment strategy. (5)



Data mining tasks

There are several type of data mining task for different purpose in business.
Main data mining tasks are described in following:

Sequence mining (categorical): The sequence mining task is to extract sequences of cases that commonly occur *frequently*, in sequence mining task a set of attributes among a large number of objects in a database that are recorded over a period of time are discovered.

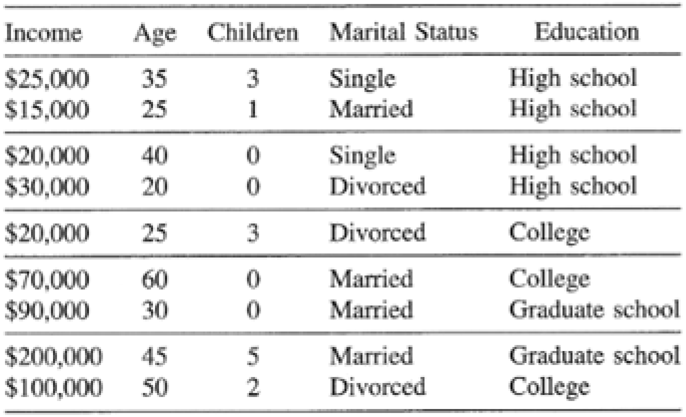
E.g. let's assume that the sales database for a bookstore customers are recorded as an object and the authors are recorded as an attributes in database columns. The database records the books has been bought by each customer. The sequences of books most frequently has been bought by the customers are the extracted patterns. An example could be that, "those who buy Harry Potter and the Deathly Hallows also buy Harry Potter and the Philosopher's Stone within a month by 65%" these patterns can be used for placement of books on the shelfs, catalog design, promotions and etc. (6)

Clustering: Clustering is another data mining task that the objects are divided into groups so that members of a group share a common set of features.

Objects in a database are usually without any predetermined classes. The group's similarity between members of the same group is high and the similarity between members of different groups is low.

E.g. A casting company does not have any predetermined classes or labels for its members, so the company wants to divide its customers into groups based on common properties, in this case company using clustering algorithm classify members to groups with a same properties, after applying clustering method to the data and getting clustering result, the company will be mailing new target to advertise their products based on properties of each groups. Customer's information such as total income, birth place, birth date, their age, and number of children, height, weight, marital status, and education are stored in company database.

Let's make clustering method more clearly, imagine that the company announced sale advertising on children's clothes, we could target the advertising only to the persons with children. The first group of people have young children and high school degree, while the second group is similar but no children, the third group has both children and college degree. The last two groups have higher incomes and at least a college degree. The very last group has children. Different clustering would have been found by examining age and marital status. (7)



Income	Age	Children	Marital Status	Education
$25,000	35	3	Single	High school
$15,000	25	1	Married	High school
$20,000	40	0	Single	High school
$30,000	20	0	Divorced	High school
$20,000	25	3	Divorced	College
$70,000	60	0	Married	College
$90,000	30	0	Married	Graduate school
$200,000	45	5	Married	Graduate school
$100,000	50	2	Divorced	College

Figure 1 Different clustering attributes

As illustrated in Figure1 a given set of data clustered on age and marital status attributes.

Deviation detection: Deviation detection (anomaly detection or outlier detection) is a data mining task to identify observations which are not unusual object that do not represent an expected pattern. Typically the Deviation can be found as a bank fraud, a medical problems or finding errors in text. Deviation are also referred to as outliers or exceptions.

Deviation detection is applicable in a variety of fields, such as intrusion detection, fraud detection, fault detection, system health monitoring, event detection in sensor networks, and detecting Eco-system disturbances. It is often used in preprocessing to remove anomalous data from the dataset. In supervised learning, removing the anomalous data from the dataset often results in a statistically significant increase in accuracy. (7)

Association: Association is one of data mining task that correlations among items in a dataset is major subject. In business it helps retailers in planning marketing strategies, catalog design and store layout by finding the association between the different items purchased by the customer. E.g. If retailer keeps bread with butter then the chances of sale will be increased because customer who buys bread, buy butter as well. (7)

Classification and Prediction: Classification is data mining task that tries to extract a mining model using an appropriate classification method, classification method looks for rules that is used to predict the class of an unseen instance, therefore is one form of prediction classes or labels for its objects. Numerical prediction (that is called regression) is another type of prediction that in this case the purpose is to predict numerical value, such as a company's profits or consumption of products.

Another type of classification is to find many association rules that can be found from any given dataset, this model is built based on training data and testing data set. Sometimes a training set is used to find any relationship among the values of variables, in these case the objects are classified to many groups, generally extracted rule are known as association rules. A common form of this type of association rules application is called 'market basket analyses. (7)

The produced model of classification methods and its functions that are done by classification technique are in the following:

The Classification and Prediction model:

- Classification (IF-THEN) Rules
- Decision Trees
- Mathematical Formulae

The Classification and Prediction functions:

Classification - The purpose is to predict labels or classes of objects based on analyzing set of training data, objects which its label is not known.

Prediction - Numerical data values are predicted instead of class labels. Regression Analysis is generally used for prediction. Prediction can also be used for identification of distribution trends based on collected data in database. (7)

Outlier Analysis - The Outliers are data objects that represents different behavior in compare to other objects. (7)

Data Mining Algorithms

C4.5 algorithm

C4.5 one of the most effective classification method that constructs a classifier in the form of decision tree.
Decision tree is used specifically in decision analysis to identify a strategy that likely to be reached a desirable goal. Decision tree is easy to read and understand. (8)

The decision tree has three types of nodes:

- A root node that has only zero or more outgoing edges.
- Internal nodes, each of which has exactly one incoming edge and two or more outgoing edges, Internal nodes does test on training set attributes
- Leaf or terminal nodes, each of which has exactly one incoming edge and no outgoing edges, is assigned a class label

Except leaf or terminal nodes, root node and internal nodes contain attribute test conditions to separate records that contains different characteristics.

Typical example of decision tree is shown in figure 1, illustrates decision tree looks like flowchart structure, the bold text indicate decision tree node (internal node) and the colorful shapes indicate leaf of nodes which is hold class label of objects. (7)

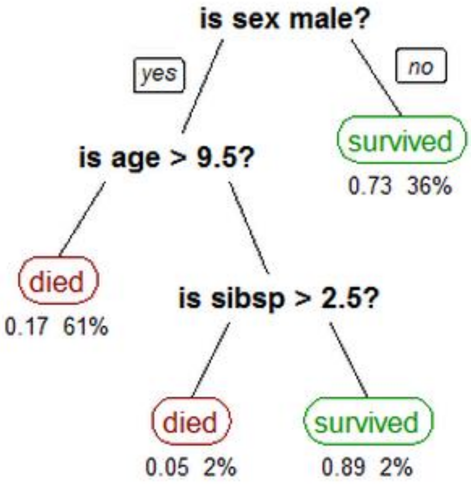


Figure 2Classification by decision tree

Decision tree for the class of titanic passengers, indicate died or survived of male gender. Each internal node represents a decision based on the value of corresponding attribute, also each leaf node represents a class (the value of labeled-attribute =died or survived). After this model has been built, it is possible to predict survived or died people based on a new data.

Therefore decision tree is a tool to predict class or label of objects in the new dataset. Decision trees is created using a set of training set with concept of information entropy. Training data set is a dataset which are already classified and information entropy refers to the degree of randomness. In Information theory if something is completely predictable, it means it is completely certain and its entropy is zero, as well as when something is quite uncertainly and unpredictability, it means the outcome cannot be predicted perfectly so its entropy will increase and entropy is measured in terms of bits (binary digit). Concept of entropy is used in attribute selection methods that is used as a parameter to define the splitting attribute in order to create decision tree nodes by C4.5 algorithms. (7)

Attribute selection methods includes Information gain and gain ratio that are used to select the best attribute as an internal node to predicts label of objects.

Information gain measures the level of impurity in a group of example, entropy means a common way to measure impurity. Information gain when I have smaller numbers of distinct values (mostly nominal and ordinal values that are not continuous

Expected information (entropy):

D indicate attribute with labeled value

$$Info(D) = -\sum_{i=1}^{m} p_i \ \log_2(p_i)$$

A indicate training set attribute,

$$Info_A(D) = \sum_{j=1}^{y} \frac{|D_J|}{|D|} * I\ (D_j)$$

Information gained by branching on a specific attribute

$$\text{Gain (A)} = \text{Info (D)} - Info_A \text{ (D)}$$

Information gain selects attributes with highest information gain.

Gain ratio splits attributes using information gain so that gain ratio adjusts information gain by the entropy of partitioning, higher entropy of attributes (large number of small partitions, attributes that have more partitions with less members) is penalized in gain ratio which was a problem in information gain method (normalization of information gain). (7)
Information gain method in compare to gain ratio method tends to select multi-valued attributes while gain ratio

Advantage of decision tree model:

- Decision tree model is an inexpensive way to create
- It is enough easy to interpret and explain for small-size trees
- Decision tree model is very fast to classify objects using class label
- Can handle continuous and categorical variables
- Most important for prediction or classification (7)

Disadvantage:

- Suffers from overfitting (happens when selection of an attribute is not optimal for prediction). (7)

C4.5 algorithm continues to grow a tree until it does not make any error ahead of time, so in this case overfitting can be happened. Since training set may include noise in dataset, Lack of representative samples also can be as a reason for overfitting, so Pruning method using confidence interval estimates is used to reduce overfitting. By this technique allows all of the available labeled data to be used for training. Whilst Pruning reduce the accuracy on the training data, but increases the accuracy on new dataset. (8)

Classification is stopped when attributes for further partitioning doesn’t exist anymore, no sample left or all samples for a given node belong to the same class.

Pruning technique calculate confidence interval for the error rate. In brief, one starts from an observed error rate f measured over the set of N training objects. The observed error rate is analogous to the observed fraction of heads in N tosses of a (usually loaded) coin. One wishes to estimate the true error rate p that would be observed if one could test over all possible data instances that will ever become available. The true error rate p is analogous to the actual probability of heads of the given coin. Note the difference between f and p: f is the fraction of errors (or heads) that were observed in one particular sample of size N, while p is the fraction of errors (heads) that will be observed over the infinitely large set of all instances, past, present, and future. (8)

The confidence interval is calculated as follows:

$$p = f + z * \mathrm{sqrt}(f * \frac{1 - f}{N})$$

Here, f, p, and N are as described above, z is a factor that depends on the desired level of confidence. Values of z for selected confidence levels appear below. (8)

Pruning based on confidence intervals

In order to decide whether to replace a near-leaf node and its child leaves by a single leaf node, C4.5 compares the upper limits of the error confidence intervals for the two trees. For the unpruned tree, the upper error estimate is calculated as a weighted average over its child leaves. Whichever tree has a lower estimated upper limit on the error rate "wins" and is selected. (7)

Decision Trees application might be used to Predict

- Forecast next year's sales.
- Predict site visitors given past historical and seasonal trends.
- Generate a risk score given demographics.
- Flag the customers in a prospective buyers list as good or poor prospects.
- Calculate the probability that a server will fail within the next 6 months.
- Categorize patient outcomes and explore related factors. (7)

K-Means: A Centroid-Based Technique

Clustering definition

Clustering is to classify objects into groups without any prior knowledge of relationships between objects. After doing clustering, objects in the same group are “similar” and are “dissimilar” to the objects which are belongs to other groups or clusters. (7)

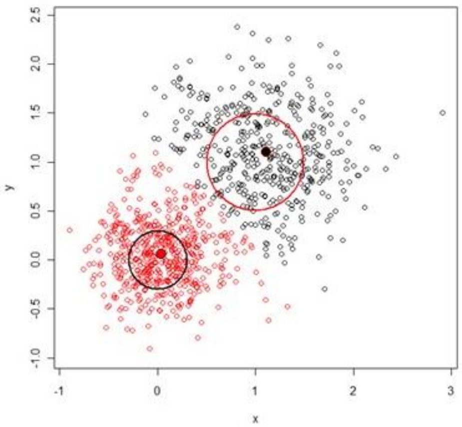


Figure 3 k-means clustering k=2, the centroids which define the partitions

For example, a dataset of patients are classified according to patient’s measure of blood pressure, various information about each patient such as age, sex, blood pressure has been recorded in hospital database.

Although clustering is the process of grouping a set of objects into group of similar but classification is an effective means for distinguishing groups or classes of objects due to classification using decision tree requires costly and labeled training tuples or patterns, which the classifier uses these labeled training tuples to predict class of objects.

Clustering contains various algorithms that each algorithm constitutes clusters in different way. K-Means algorithm (A Centroid-Based Technique) is one of clustering technique that classifies objects into k number of clusters. (7)

The goal of k-means algorithm is to divide n observation to k clusters as initial centroid. Choosing the number of k is important and must match with the amount of data. Incorrect choice of the number of k definitely destroy the whole process. Centre-based clusters is a cluster that an object in a cluster is nearest (more similar) to the “Centre” of a cluster, than to the Centre of any other cluster. These centers should be placed in a cunning way because of different location generate different result. So, the better choice is to set Centres as much as possible far away from each other. The next step is to set objects to the nearest center through Calculation of the distance between object with cluster centers, therefore when all objects placed in a specific cluster with iterative method, the k number of clusters have been created. The second step is to redo specifying k number (number of clusters or centers) and selected objects becomes as an initial center for the new centroid for the cluster and to replace objects to the new clusters according to the nearest center or minimum distance from the all the cluster centers, Since the centroid has changed cluster membership has been changed and now each of previous group members might be closer to different centroids, as result cluster membership changes within a cluster or group.

This loop continues until location of centers do not change anymore and don’t move to other location (stable clusters). (7)

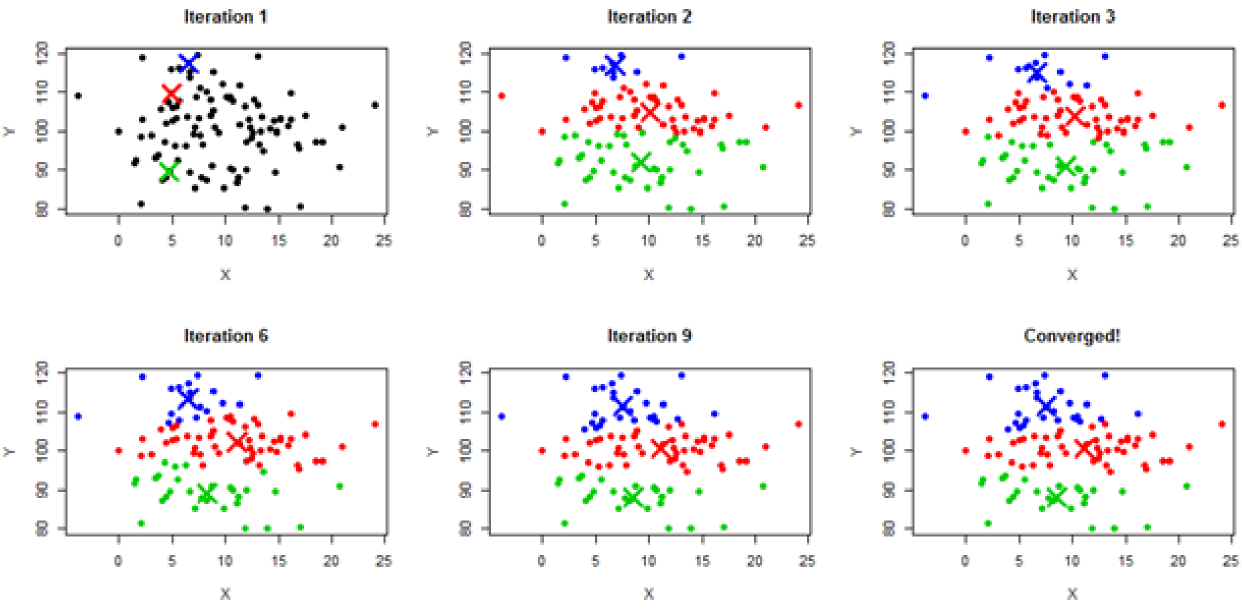


Figure 4 how K-Means Clustering works

Objects which are selected as a cluster center can be chosen in different way:

The first is Dynamically Chosen, it is applicable when the amount of data is expected to grow. In a simple way, the first few objects might set as the initial cluster means. For instance, if the objects will be classified into 5 clusters, then the first 5 objects are located as the initial cluster means. The second way is to choose objects as the initial cluster means randomly.

The third way is to choose objects from Upper and Lower Bounds so that objects which are defined as the initial cluster means are selected through the highest and lowest value of the data based on type of data.

The clustering is the process of partitioning by minimizing the sum of squared distances (Euclidean distances) between items and the corresponding centroid. (7)

Euclidean distances is calculated by following formula: (9)

J objective function

K number of clusters or Initialize the center of the clusters (means)
N Number of objects
c_j Centroid for cluster j
$x_i^{(j)}$ Object i
$x_i^{(j)} - c_j$ Distance function

$$j = \sum_{j=i}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Advantages

- Centroid-Based Technique is easy to understand.
- Often stop iteration in a local optimum

Disadvantages

- Not suitable to cluster for nonlinear data set.
- Applicable for numerical variables due to defining k is not possible for categorical variables.
- This method needs to specify appropriate k number in advance since if number of k is selected equally to number of objects it might decrease the optimal solution, as well as choosing randomly objects as initial cluster means cannot lead to the acceptable results. (9)

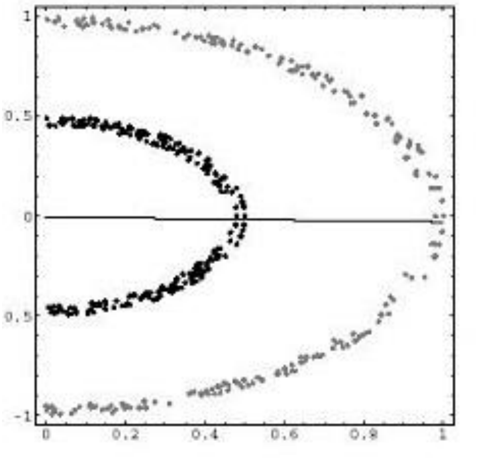


Figure 5 K-means algorithms fails for non-linear dataset

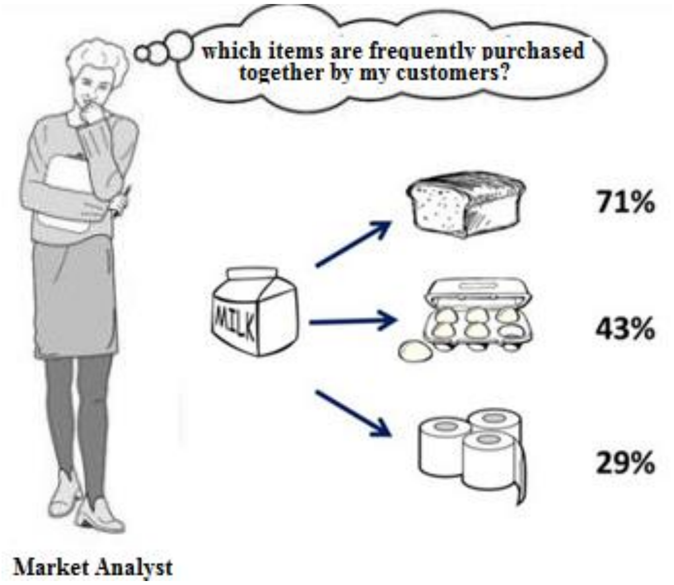
Unable to handle outliers, outliers may change the centroid away from its true position.

K-means algorithm Will not able to classify objects into two groups when data set is highly overlapping. (9)

Apriority algorithm

One of data mining algorithm for mining frequent objects for Boolean association rule, is applied to a database containing a large number of transactions. Apriority algorithm tries to find correlations and relations among objects in a database. Since Apriority algorithm used to discover interesting patterns and relationships generally considered an unsupervised learning approach.

For example imagine a market transactions, where each row represents a customer transaction and every column represents different sold items by market to the customer, apriority algorithm finds the items which are frequently purchased together, according to the customer transaction the analyst find outs that customers who buy milk, by 71% purchase bread, customers who buy milk, by 43% purchases eggs, as well as customers who buy milk, by 29% purchases toilet paper. (10)


which items are frequently purchased together by my customers?
MILK
71%
43%
29%
Market Analyst

71% included bread
43% included eggs
29% included toilet paper

Retailer uses such information as the basis for decisions about advertising, designing catalog, marketing activities such as promotional pricing, product placements, product clustering, also uses to predict customer behavior.

Association rules helps to find out hidden correlation pattern between item sets that seems data are unrelated to each other and does not have any relation between them in a relational database. An example for association rule can be "If a customer buys a bread, she is 90% likely to purchase milk."

Association algorithm contains rules that analyzes data using the criteria support and confidence so that these criteria tries to extract the most important relationships between item sets. Support is an indication of how frequently the items have been appeared in the database. Confidence indicates the number of times the if/then statements have been found to be true.

Frequent item sets are those that its support is equal or greater than minimum support threshold, as well as its confidence is equal or greater than minimum confidence threshold. (10)

Support count σ
Support s
Confidence c

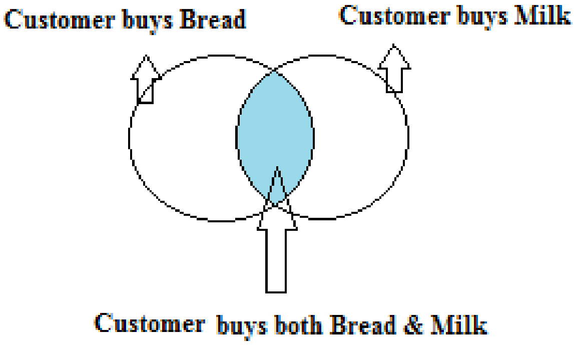
$$s \geq minsup$$
$$c \geq minconf$$

Let consider below example, imagine shopping cart transections as following:

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Implication of x=>y (association rule) where x and y both are item, indicates transection which includes item x, may include item y too, therefore items y and x could be as a frequent item sets. (10)

For example in given transection if we consider Bread =>Milk



Support count: 3
Support: 3/8 = 0.38
Confidence: 3/4 = 0.75

If we set minimum support = 2, the association rule (Bread =>Milk) can be frequent item sets due to less its support than minimum support.

There are two approach for mining association rule from the transection as following:

1. Frequent items generation
2. Rule generation

Frequent items generation: generates all item sets which satisfies $s \geq minsup$

Brute-force approach:

- Each item set is a candidate frequent item sets
- Count support of each candidate
- Match each transection against every candidate
- Complexity, number of possible candidates $M = 2^d$
- Total number of possible association rules $R = 3^d - 2^{d+1} + 1$

One of the frequent items generation's strategy is to reduce number of candidates **(M),** so that after generating completely candidates according to the **Brute-force approach**, using apriority principle tries to reduce number of candidates.
Apriority principle says that subset of a frequent item sets must be frequent too when support of an item set never exceeds the support of its subsets (this is called anti-monotone property of support).

Customer Shopping cart transaction

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

By minimum support = 3 and according to the Apriori Algorithm , if support of an item sets exceeds the support of its subsets, subsets will be removed from the association rules, so no Eggs and Coke are removed and cannot be involved to generate.

If minimum support is set too high, item sets involving interesting rare items for example expensive products may be missed, also if minimum support is set too low, it is computationally expensive and the number of item sets is very large. (10)
Apriori Algorithm

- Let k=1

- Generate frequent item sets of length 1
- Repeat until no new frequent item sets are identified
    - Generate length (k+1) candidate item sets from length k frequent item sets
    - Prune candidate item sets containing subsets of length k that are infrequent
    - Count the support of each candidate by scanning the DB
    - Eliminate candidates that are infrequent, leaving only those that are frequent

Apriori Algorithm

C_K: Candidate item set of size k
L_K: Frequent item set of size k

```
$L_1$= {frequent item};
For (k=1, $L_K$ ! =$\emptyset$, k++) do begin
    $C_{K+1}$= candidate generated from$L_K$;
    For each transaction t in database do
        Increment the count of all candidates in $C_{K+1}$ that are contained in t
    $L_{K+1}$ = candidates in $C_{K+1}$ with min-support
    End
    Return $U_k L_K$
```

Items = 1-itemsets

Item	Count
Bread	4
Milk	4
Diaper	4
Beer	3
Eggs	1
Coke	2

Pairs = 2-itemsets

Item	Count
{Bread, Milk}	3
{Bread, Beer}	2
{Bread, Diaper}	3
{Milk, Beer}	2
{Milk, Diaper}	3
{Beer, Diaper}	3

{Bread, Beer} and {Milk, Beer} are removed.

Triplets = 3-itemsets

Item	Count
{Bread, Milk, Diaper}	3

Finally the frequent item sets {Diaper, Bread, Milk} are extracted from customers transaction.

As the above example shows, association Rules is a data mining method commonly used in retail business to find out the customer behavior of purchasing, which one of products consumers will to purchase together? This information can be used to perform appropriate sales promotions or to organize the products in a supermarket.

Another example of association rules for Workers Compensation insurance company, Association Rules is applied to analyze historical Workers Compensation insurance to find co-occurring features of the claim, for example fractures result a large part of leg injuries result and also large claims result. This insight can be used to improve the safety program to take preventive measures, for example leg safety gear, better training etc. (10)

Regression algorithm

Linear Regression algorithm as a data mining technique and is a statistical tool is used to extract interesting relationship between a dependent and independent variable, and then use that relationship for prediction. (11)

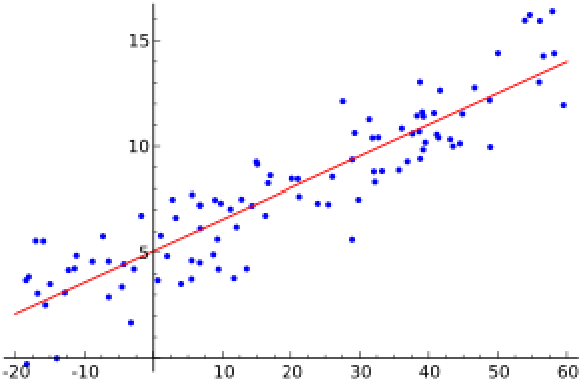


Figure 6Illustration of linear regression on a data set

There are two type linear regression, simple and multiple linear regression. Simple linear regression conducts a relationship between an outcome variable and an independent variable. The relationship is typically expressed in terms of a mathematical equation as following:

γ_i Parameter
x_i Independent variables

y Dependent variable

Simple linear regression $y = \gamma_1 x_{1t} + \gamma_2\, x_{2t} + \varepsilon$
Multiple linear regression $y = \gamma_1 x_{1t} + \gamma_2\, x_{2t} + \gamma_3\, x_{3t} + \gamma_4\, x_{4t} + \gamma_5\, x_{5t} + \varepsilon$

The multiple linear regression model that is constructed using several attributes is more effective and can predict better in comparison to the simple linear regression model.

Regression analysis is used for understanding on how the outcome variable changes if one of predictor variable change, therefore Regression analysis is widely used for prediction and forecasting, forecasting is to predict future trends based on past and present data and analysis of trends, consumption of products. For example multiple linear regression is applicable to numerous data mining situations such as estimation of profit, customer behavior on the credit cart, prediction of manufacturing expenditures. (11)

K-nearest neighbors algorithm

K nearest neighbors is one of data mining algorithm that classifies new object based on a specific parameters such as distance functions and k parameter. The goal of K-nearest neighbor algorithm is to find a class label for the unknown object. Requirement for this data mining method is k number of nearest objects and training dataset and the metric which is used to measure distance between nearest objects and unknown object. (12)

Following three distance measures only valid for continuous variables. (12)

Euclidean function

$$\sqrt{\sum_{i-1}^{k}(x_i - y_i)^2}$$

Manhattan function

$$\sum_{i=1}^{k}|x_i - y_i|$$

Murkowski function

$$\left[\sum_{i=1}^{k}(|x_i - y_i|)^q\right]^{1/q}$$

Following Humming Distance function is valid for categorical variables:

$$D_H = \sum_{i=1}^{k} |x_i - y_i|$$

For example the below figure illustrates the k number objects and the unknown object (x_u) and the three group of training dataset objects w_1 , w_2 , w_3 that have the same label. The black arrows illustrate nearest objects in size of k number. According to the distance measurement new object will be labeled. (12)

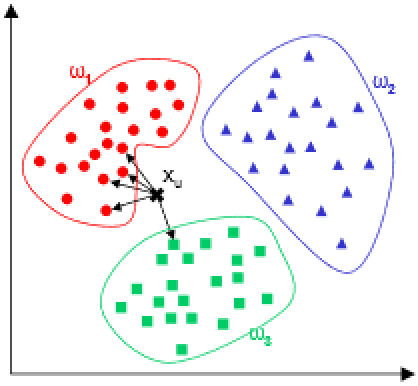


Figure 7 the K nearest Neighbor Rule

Practical Example of the Decision Trees

Practical Example – a task for a sales manager of the retailer company is about classification of customers, or to label customers according to Innovator, Early adopter, Early Majority and Late Majority.

Suppose in a company that is online retailer, is going to launch next generation of Notebook in few days, so the company wants to maximize the effectiveness of their marketing.

Owner of company in order to drive the sales of its new generation of Notebook is going to offer specific products and services for the generation of Notebook through its web site so that customers who will purchased the new generation of Notebook will be able to buy a new version of Antivirus, wireless mouse and keyboard, Microsoft office software and so forth. The company also sells tens of other type of products such as External Hard Drives, cooling Fan, Flash Memory and every kind of products in related to Notebooks.

It is good to be noticed that company only sells the products through its website and customers in order to buy any products need to offer their needs through the company's website.

The company already have information of customers who have purchased one of the company's previous generation of Notebook while the sales manager of company noticed from the information of customers some of customers are most interested to buy new generation of Notebook as soon as product comes to the market, some of them has less driven to have the product.

The company also has set of data for the current customers including the personal information of customers, items they have browsed the website of company for, items has actually purchased from the company by the customers. In other hand the Sales manager believes in the classic diffusion theories which is written by sociologist Everett Rogers in the 1960s. **Diffusion of innovations** is a theory that seeks to explain how, why, and at what rate new **ideas** and **technology** spread through cultures.

According to **Diffusion of innovations** theory, Innovations are not adopted by all individuals in a social system at the same time. Instead, they tend to adopt in a time sequence, and can be classified into adopter categories based upon how long it takes for them to begin using the new idea.

Rogers believes that the adoption of a new technology or innovation tends to follow an 'S' shaped curve, with a smaller group of the most enterprising and innovative customers adopting the technology first, followed by larger groups of middle majority adopters, followed by smaller groups of late adopters. (Figure 12).

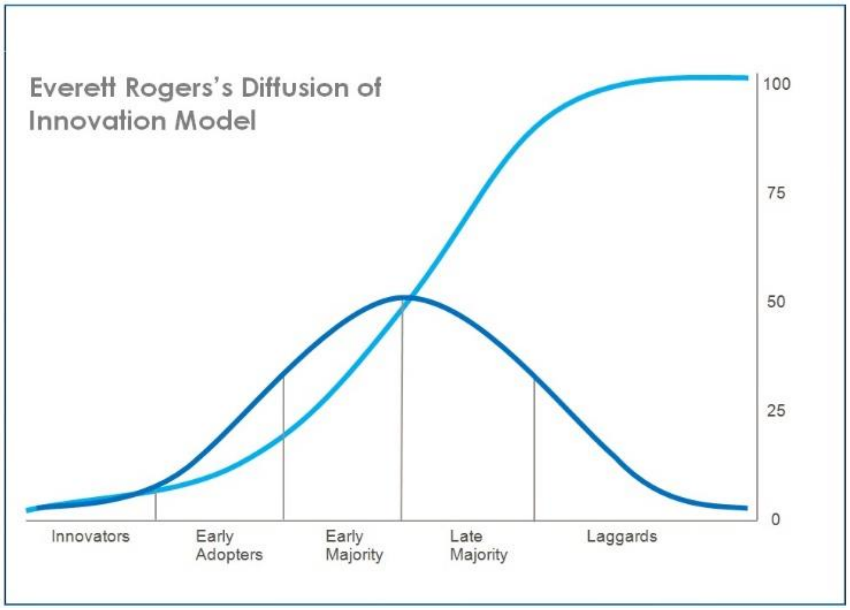


Figure 12 Everett Rogers' theory of adoption of new innovations.

Those at the front of the dark blue curve are the smaller group that are first to want and buy the technology. Most of us, the masses, fall within the middle 70-80% of people who eventually acquire the technology. The low end tail on the right side of the dark blue curve are the laggards, the ones who eventually adopt.

Sale manager according to the Rogers' theory categorize the company's clients into one of four groups that will eventually buy the new Digital camera generation:

Innovators: Innovators are willing to take risks, have the highest social status, have financial liquidity, are social and have closest contact to scientific sources and interaction with other innovators. Their risk tolerance allows them to adopt technologies that may ultimately fail. Financial resources help absorb these failures

Early Adopters: Early adopters have a higher social status, financial liquidity, advanced education and are more socially forward than late adopters. They are more discreet in adoption choices than innovators.

Early Majority: They adopt an innovation after a varying degree of time that is significantly longer than the innovators and early adopters. Early Majority have above average social status, contact with early adopters.

Late Majority: They adopt an innovation after the average participant. These individuals approach an innovation with a high degree of skepticism and after the majority of society has adopted the innovation. Late Majority are typically skeptical about an innovation, have below average social status, little financial liquidity, in contact with others in late majority and early majority.

Laggards: They are the last to adopt an innovation. Laggards typically tend to be focused on "traditions", lowest social status, lowest financial liquidity, oldest among adopters, and in contact with only family and close friends.

Now the company in order to be able to predict the timing of buying behaviors of customers is going to divide all customers according to Rogers Groups as has been explained into four groups including Innovators, Early Adopters, Early Majority or Late Majority.

So a task for sales manager to figure out when and which customers are most interested to purchase the new generation of Notebook as soon as the product comes out, which customers and when will purchase the product next, which customers and when will purchase later on from the information of current customers activities or regarding general consumer behaviors, also the company believes that will be able to time his target marketing to the people most ready to respond to advertisements and promotions as mentioned above.

In this case data mining can help the sales manager to figure out which activities are the best predictors of which category a customer will fall into. Knowing this, he can time his marketing to each customer to coincide with their likelihood of buying.

Decision tree model in order to find good early predictors of buying behavior. The sales manager using two data sets Training set and Scoring set, so that the first one is contained the web site activities of customers who has bought previses generation of Notebook and the timing with which they bought their previses generation of Notebook. The second is contained attributes of current customers which Sales manager hopes will buy the new generation of Notebook. He hopes to figure out which category of adopter each person in the scoring data set will fall into based on the profiles and buying timing of those people in the training data set.

In analyzing data set, The Sales manager has found that customers' activity in the areas of Antivirus and wireless mouse and keyboard, and their general activity with other type of products for sale on his company's site, seem to have a lot in common with when a person buys a Notebook. With this in mind, we have worked with data sets is included the following attributes:

User-ID: A numeric, unique identifier assigned to each person who has an account on the company's web site.

Gender: The customer's gender, as identified in their customer account. It is recorded an 'M' for male and 'F' for Female.

Age: The customer's age.

Marital-Status: The person's marital status.

Website-Activity: This attribute is an indication of how active each customer is on the company's web site, which records the duration of each customer visits to the web site to calculate how frequently, and for how long each time, the customers use the web site. This is then translated into one of three categories: Seldom, Regular, or Frequent.

Browsed-Product: Yes/No column indicates whether or not the person browsed for products on the company's web site in the past year

Bought-Product: Yes/No column indicates whether or not they purchased an item through company's web site in the past year.

Bought-Wireless-Mouse-Keyboard: Yes/No indicates whether or not the person has purchased some form of Wireless mouse and keyboard in the past year and a half. This attribute does not include Antivirus purchases.

Bought-Antivirus: The Sales manager believes that as an indicator of buying behavior relative to the company's new generation of Notebook, this attribute will likely be the best indicator. Thus, this attribute has been set apart from the purchase of other types of Wireless mouse and keyboard. Further, this attribute indicates whether or not the customer has *ever* bought a digital book, not just in the past year or so.

Payment-Method:

- Bank Transfer
- Website Account
- Credit Card
- Monthly Billing

Notebook-Adoption: This attribute exists only in the training data set. It consists of data for customers who purchased the previous generation **Notebook**. Those who purchased within a week of the product's release are recorded in this attribute as **'Innovator'**. Those who purchased after the first week but within the second or third weeks are entered as **'Early Adopter'**. Those who purchased after three weeks but within the first two months are **'Early Majority'**. Those who purchased after the first two months are **'Late Majority'**. This attribute will serve as our label when we apply our training data to our scoring data.

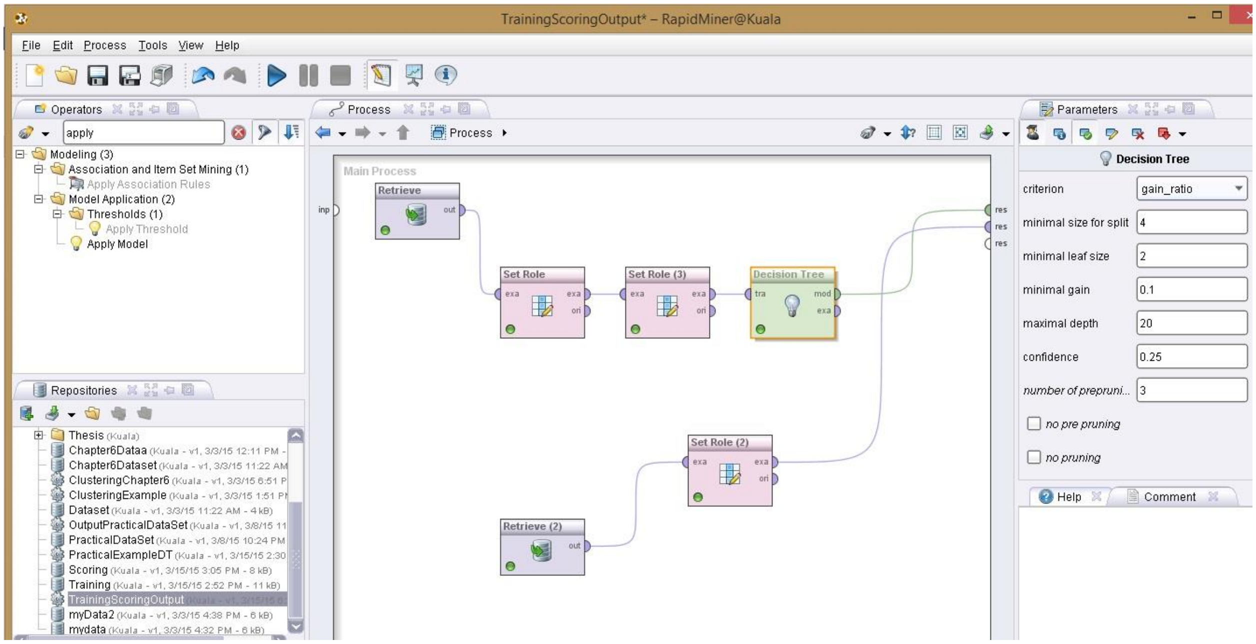
Data Preparation - using RapidMiner Software with data sets and an understanding of what it means, we can now induct the decision tree and analyze or interpret its results.

First of all, we import the two data sets, the User-ID is an arbitrarily assigned value for each customer. The customer doesn't use this value for anything, it is simply a way to uniquely identify each customer in the data set. It is not something that relates to each person in any way that would be predictive of their buying and technology adoption tendencies. As such, it should not be included in the model as an independent variable.

We can handle the User-ID attribute using a Select Attributes operator as a non-predictive attribute, so the target role for the User-ID attribute should be id to identify each customers. Thus the software won't consider the User-ID attribute as a predictor for the label attribute (Notebook-Adoption). We will do this for both the training and scoring data sets, since the User-ID attribute is found in both of them.

Before adding the Decision Tree operator which expects the training stream to supply a 'label' attribute, we need to do another data preparation step so that we add another Set Role operator and we set the Notebook-Adoption as a label for indicating of Adaptor groups of customers.

After all we will add the decision tree operator to the main process area and set its criteria to **gain_ratio**.



TrainingScoringOutput* – RapidMiner@Kuala
File Edit Process Tools View Help
Operators
apply
Modeling (3)
Association and Item Set Mining (1)
Apply Association Rules
Model Application (2)
Thresholds (1)
Apply Threshold
Apply Model
Repositories
Thesis (Kuala)
Chapter6Dataa (Kuala - v1, 3/3/15 12:11 PM -
Chapter6Dataset (Kuala - v1, 3/3/15 11:22 AM
ClusteringChapter6 (Kuala - v1, 3/3/15 6:51 P
ClusteringExample (Kuala - v1, 3/3/15 1:51 PM
Dataset (Kuala - v1, 3/3/15 11:22 AM - 4 kB)
OutputPracticalDataSet (Kuala - v1, 3/8/15 11
PracticalDataSet (Kuala - v1, 3/8/15 10:24 PM
PracticalExampleDT (Kuala - v1, 3/15/15 2:30
Scoring (Kuala - v1, 3/15/15 3:05 PM - 8 kB)
Training (Kuala - v1, 3/15/15 2:52 PM - 11 kB)
TrainingScoringOutput (Kuala - v1, 3/15/15 6:
myData2 (Kuala - v1, 3/3/15 4:38 PM - 6 kB)
mydata (Kuala - v1, 3/3/15 4:32 PM - 6 kB)
Process
Main Process
inp
Retrieve
out
Set Role
exa
exa
ori
Set Role (3)
exa
exa
ori
Decision Tree
tra
mod
exa
res
res
res
Set Role (2)
exa
exa
ori
Retrieve (2)
out
Parameters
Help
Comment

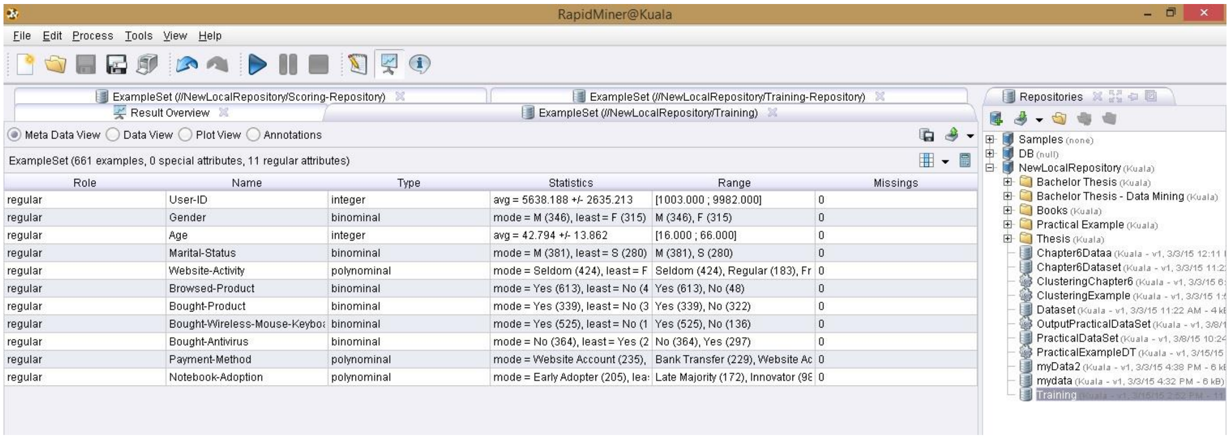
Decision Tree

criterion: gain_ratio
minimal size for split: 4
minimal leaf size: 2
minimal gain: 0.1
maximal depth: 20
confidence: 0.25
number of prepruni...: 3
□ no pre pruning
□ no pruning

Figure 8

Figure 9 Process Design Model by Decision Tree

And now we run the process and the following graphs and results are generated by RapidMiner:

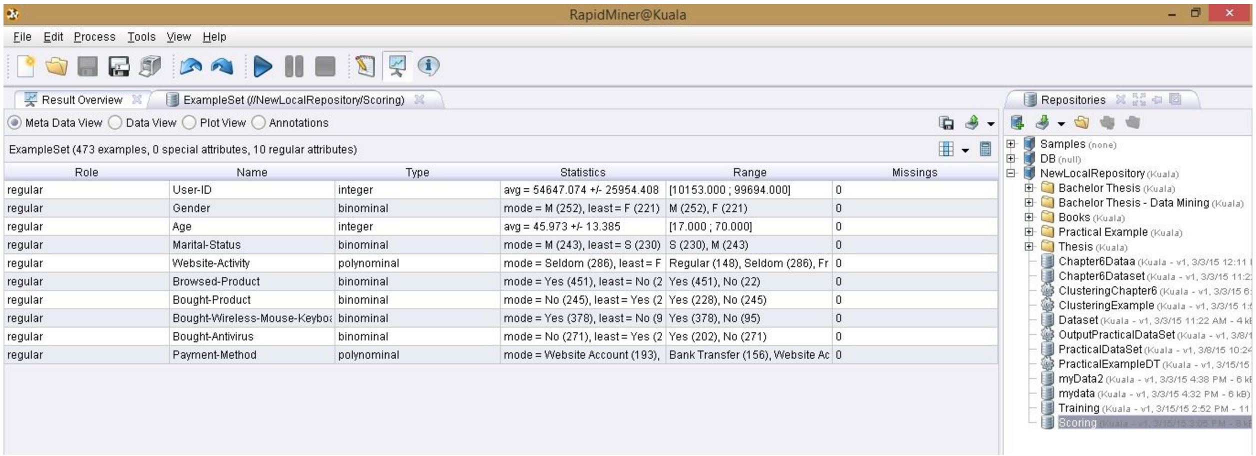


RapidMiner@Kuala
File Edit Process Tools View Help
ExampleSet (//NewLocalRepository/Scoring-Repository)
ExampleSet (//NewLocalRepository/Training-Repository)
Result Overview
ExampleSet (//NewLocalRepository/Training)
● Meta Data View ○ Data View ○ Plot View ○ Annotations
Repositories
Samples (none)
DB (null)
NewLocalRepository (Kuala)
Bachelor Thesis (Kuala)
Bachelor Thesis - Data Mining (Kuala)
Books (Kuala)
Practical Example (Kuala)
Thesis (Kuala)
Chapter6Dataa (Kuala - v1, 3/3/15 12:11
Chapter6Dataset (Kuala - v1, 3/3/15 11:2
ClusteringChapter6 (Kuala - v1, 3/3/15 6:
ClusteringExample (Kuala - v1, 3/3/15 1:
Dataset (Kuala - v1, 3/3/15 11:22 AM - 4 k
OutputPracticalDataSet (Kuala - v1, 3/8/1
PracticalDataSet (Kuala - v1, 3/8/15 10:24
PracticalExampleDT (Kuala - v1, 3/15/15
myData2 (Kuala - v1, 3/3/15 4:38 PM - 6 k
mydata (Kuala - v1, 3/3/15 4:32 PM - 6 kB)
Training (Kuala - v1, 3/15/15 2:52 PM - 11

ExampleSet (661 examples, 0 special attributes, 11 regular attributes)

Role	Name	Type	Statistics	Range	Missings
regular	User-ID	integer	avg = 5638.188 +/- 2635.213	[1003.000 ; 9982.000]	0
regular	Gender	binominal	mode = M (346), least = F (315)	M (346), F (315)	0
regular	Age	integer	avg = 42.794 +/- 13.862	[16.000 ; 66.000]	0
regular	Marital-Status	binominal	mode = M (381), least = S (280)	M (381), S (280)	0
regular	Website-Activity	polynominal	mode = Seldom (424), least = F	Seldom (424), Regular (183), Fr	0
regular	Browsed-Product	binominal	mode = Yes (613), least = No (4	Yes (613), No (48)	0
regular	Bought-Product	binominal	mode = Yes (339), least = No (3	Yes (339), No (322)	0
regular	Bought-Wireless-Mouse-Keybo	binominal	mode = Yes (525), least = No (1	Yes (525), No (136)	0
regular	Bought-Antivirus	binominal	mode = No (364), least = Yes (2	No (364), Yes (297)	0
regular	Payment-Method	polynominal	mode = Website Account (235),	Bank Transfer (229), Website Ac	0
regular	Notebook-Adoption	polynominal	mode = Early Adopter (205), lea	Late Majority (172), Innovator (9	0

Figure 10 Meta data for the Training data set.



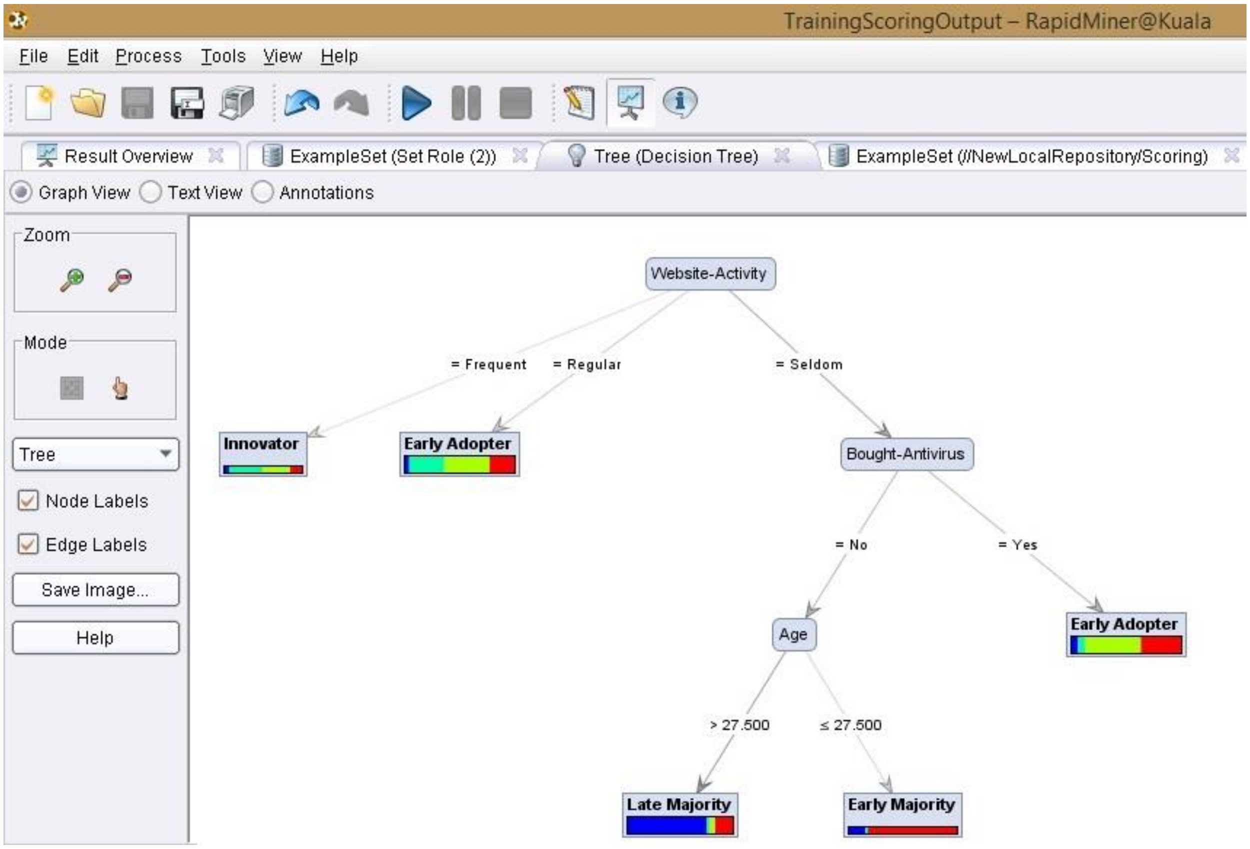
RapidMiner@Kuala
File Edit Process Tools View Help
Result Overview
ExampleSet (//NewLocalRepository/Scoring)
● Meta Data View ○ Data View ○ Plot View ○ Annotations
Repositories
Samples (none)
DB (null)
NewLocalRepository (Kuala)
Bachelor Thesis (Kuala)
Bachelor Thesis - Data Mining (Kuala)
Books (Kuala)
Practical Example (Kuala)
Thesis (Kuala)
Chapter6Dataa (Kuala - v1, 3/3/15 12:11
Chapter6Dataset (Kuala - v1, 3/3/15 11:2
ClusteringChapter6 (Kuala - v1, 3/3/15 6:
ClusteringExample (Kuala - v1, 3/3/15 1:
Dataset (Kuala - v1, 3/3/15 11:22 AM - 4 k
OutputPracticalDataSet (Kuala - v1, 3/8/1
PracticalDataSet (Kuala - v1, 3/8/15 10:24
PracticalExampleDT (Kuala - v1, 3/15/15
myData2 (Kuala - v1, 3/3/15 4:38 PM - 6 k
mydata (Kuala - v1, 3/3/15 4:32 PM - 6 kB)
Training (Kuala - v1, 3/15/15 2:52 PM - 11
Scoring (Kuala - v1, 3/15/15 3:05 PM - 8 k

ExampleSet (473 examples, 0 special attributes, 10 regular attributes)

Role	Name	Type	Statistics	Range	Missings
regular	User-ID	integer	avg = 54647.074 +/- 25954.408	[10153.000 ; 99694.000]	0
regular	Gender	binominal	mode = M (252), least = F (221)	M (252), F (221)	0
regular	Age	integer	avg = 45.973 +/- 13.385	[17.000 ; 70.000]	0
regular	Marital-Status	binominal	mode = M (243), least = S (230)	S (230), M (243)	0
regular	Website-Activity	polynominal	mode = Seldom (286), least = F	Regular (148), Seldom (286), Fr	0
regular	Browsed-Product	binominal	mode = Yes (451), least = No (2	Yes (451), No (22)	0
regular	Bought-Product	binominal	mode = No (245), least = Yes (2	Yes (228), No (245)	0
regular	Bought-Wireless-Mouse-Keybo	binominal	mode = Yes (378), least = No (9	Yes (378), No (95)	0
regular	Bought-Antivirus	binominal	mode = No (271), least = Yes (2	Yes (202), No (271)	0
regular	Payment-Method	polynominal	mode = Website Account (193),	Bank Transfer (156), Website Ac	0

Figure 15 Meta data for the Scoring data set.

In Figure 16, the decision tree is showed with its **nodes** and **leaves**. The nodes are the gray oval shapes. They are attributes which serve as good predictors for our label attribute. The leaves are the multicolored end points that show us the distribution of categories from our label attribute that follow the branch of the tree to the point of that leaf.



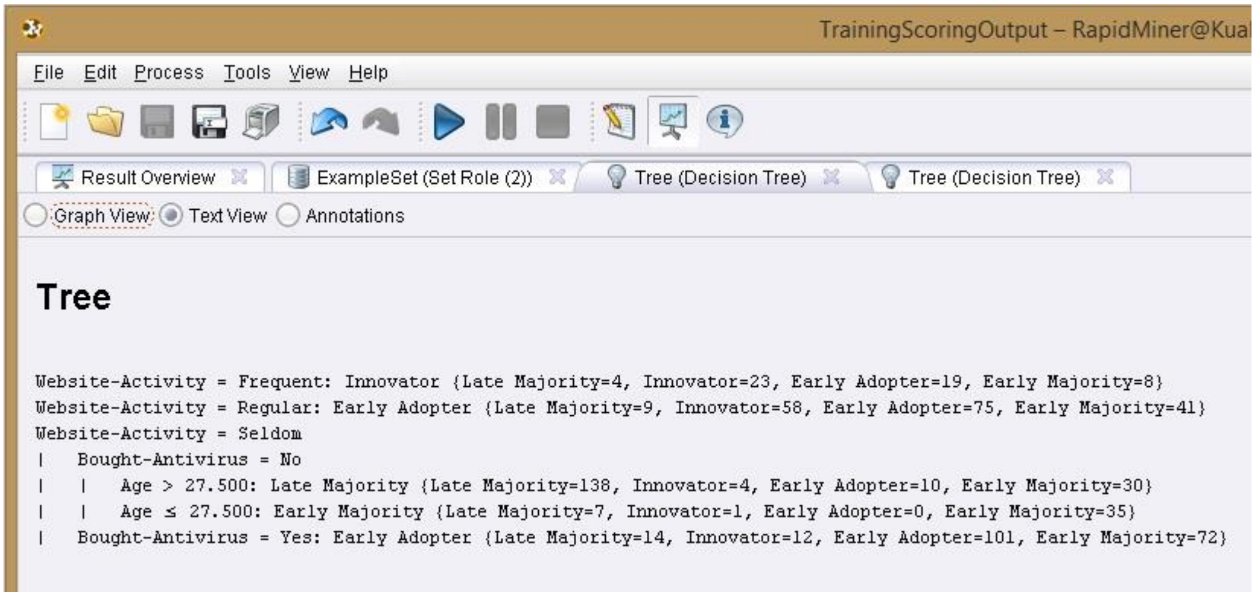
● Graph View ○ Text View ○ Annotations

☑ Node Labels

☑ Edge Labels

Figure 16 Decision tree results (gain ratio)

In this tree that **Website-Activity** is the best predictor of whether or not a customer is going to adopt (buy) the company's new generation of Notebook. If the person's activity is frequent or regular, so they are likely to be an Innovator or Early Adopter, respectively. If however, they seldom use the web site, then whether or not they've bought Antivirus becomes the next best predictor of their Notebook adoption category. If they have not bought digital books through the web site in the past, Age is another predictive attribute which forms a node, with younger folks adopting sooner than older ones. This is seen on the branches for the two leaves coming from the Age node in Figure 16. Those who seldom use the company's website, have never bought digital books on the site, and are older than 27 ½ are most likely to land in the Late Majority category, while those with the same profile but are under 27 ½ are bumped to the Early Majority prediction.


TrainingScoringOutput – RapidMiner@Kual
File Edit Process Tools View Help
Result Overview
ExampleSet (Set Role (2))
Tree (Decision Tree)
Tree (Decision Tree)
○ Graph View ● Text View ○ Annotations

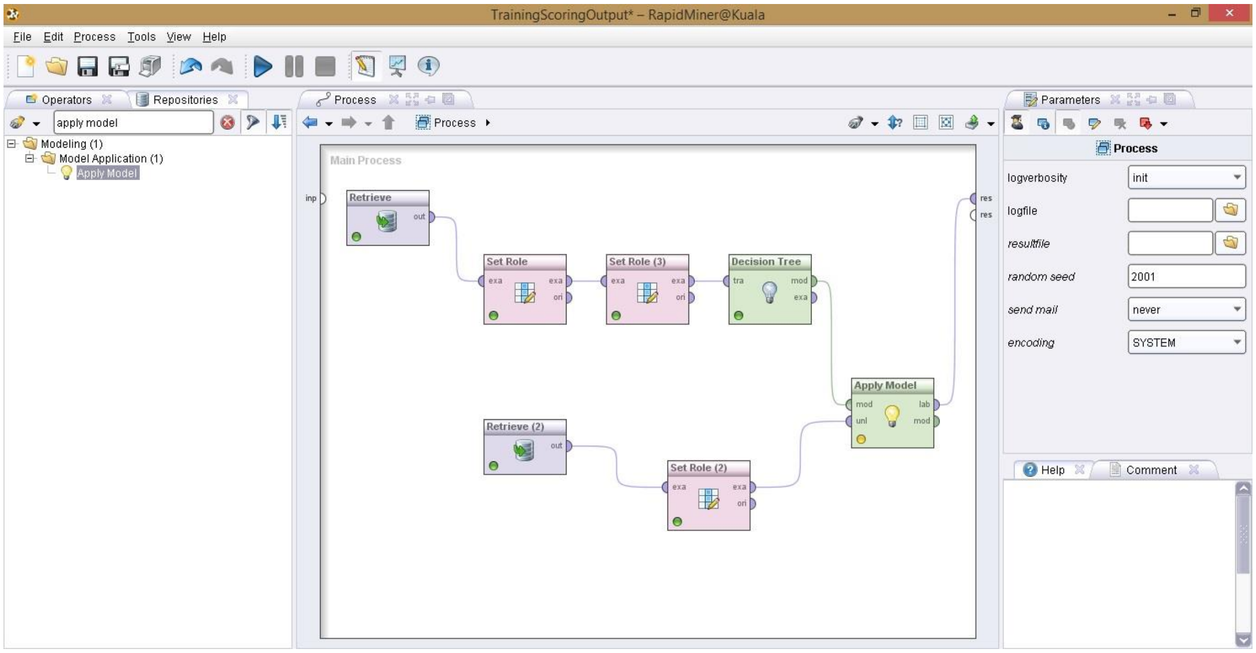
Tree

```
Website-Activity = Frequent: Innovator {Late Majority=4, Innovator=23, Early Adopter=19, Early Majority=8}
Website-Activity = Regular: Early Adopter {Late Majority=9, Innovator=58, Early Adopter=75, Early Majority=41}
Website-Activity = Seldom
|   Bought-Antivirus = No
|   |   Age > 27.500: Late Majority {Late Majority=138, Innovator=4, Early Adopter=10, Early Majority=30}
|   |   Age ≤ 27.500: Early Majority {Late Majority=7, Innovator=1, Early Adopter=0, Early Majority=35}
|   Bought-Antivirus = Yes: Early Adopter {Late Majority=14, Innovator=12, Early Adopter=101, Early Majority=72}
```

Figure 17 Text View of expanded leaf detail in our tree.

Although the training data is going to predict that ‘regular’ web site users are going to be Early Adopters, the model is not 100% based on that prediction. As we read that in the training data set, 9 people who fit this profile are Late Adopters, 58 are Innovators, 75 are Early Adopters and 41 are Early Majority. When we get to Evaluation phase, we will see that this uncertainty in our data will translate into confidence percentages.

Now the sales manager applies the model to the second data set which is contains current customers to figure out adaptor categories. Thus an Apply Model operator is added to the process area and after connecting the operators properly the process is ready to run.

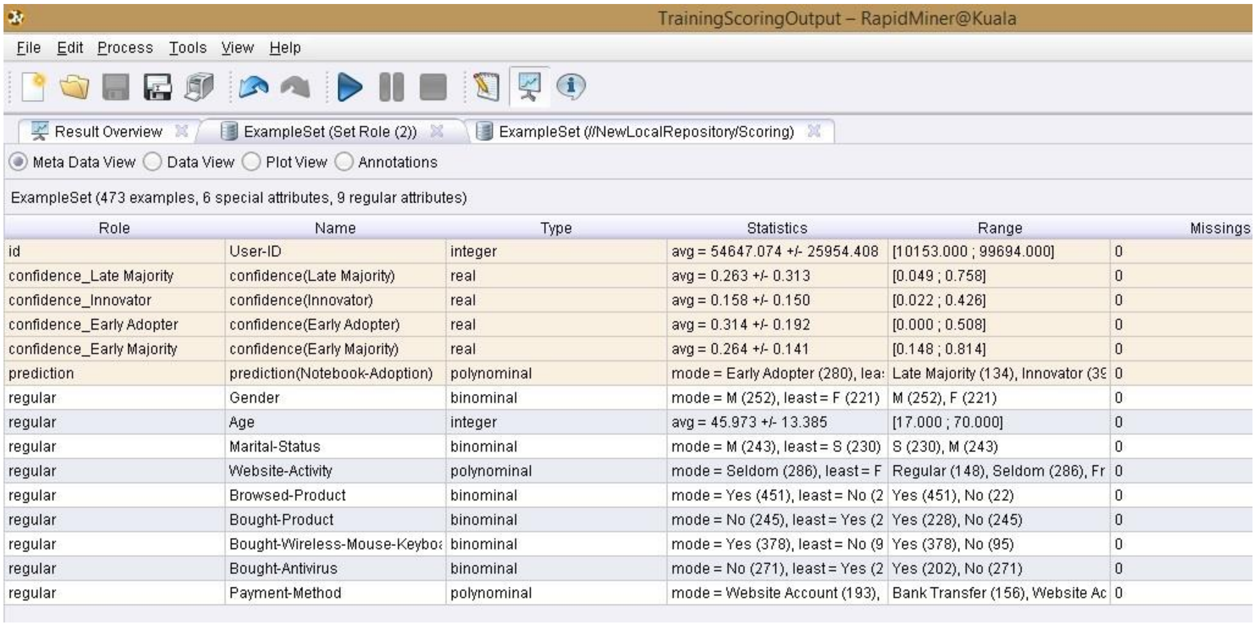


Process

logverbosity	init
logfile	
resultfile	
random seed	2001
send mail	never
encoding	SYSTEM

*Figure 18 applying the model to our scoring data, and outputting label predictions (*lab*) and a decision tree model (*mod*).*

As the figure 19 shows the familiar result was generated, also confidence attributes have been created by RapidMiner, along with a prediction attribute.

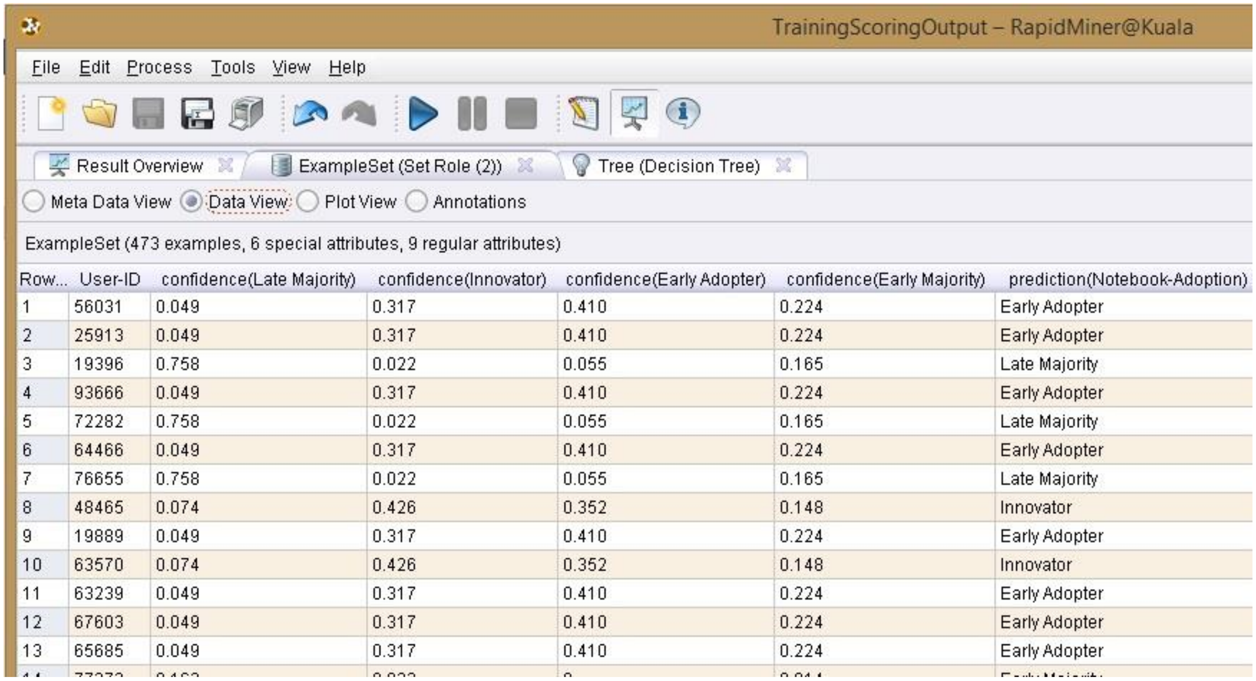

TrainingScoringOutput – RapidMiner@Kuala
File Edit Process Tools View Help
Result Overview
ExampleSet (Set Role (2))
ExampleSet (//NewLocalRepository/Scoring)
● Meta Data View ○ Data View ○ Plot View ○ Annotations

ExampleSet (473 examples, 6 special attributes, 9 regular attributes)

Role	Name	Type	Statistics	Range	Missings
id	User-ID	integer	avg = 54647.074 +/- 25954.408	[10153.000 ; 99694.000]	0
confidence_Late Majority	confidence(Late Majority)	real	avg = 0.263 +/- 0.313	[0.049 ; 0.758]	0
confidence_Innovator	confidence(Innovator)	real	avg = 0.158 +/- 0.150	[0.022 ; 0.426]	0
confidence_Early Adopter	confidence(Early Adopter)	real	avg = 0.314 +/- 0.192	[0.000 ; 0.508]	0
confidence_Early Majority	confidence(Early Majority)	real	avg = 0.264 +/- 0.141	[0.148 ; 0.814]	0
prediction	prediction(Notebook-Adoption)	polynominal	mode = Early Adopter (280), lea:	Late Majority (134), Innovator (39	0
regular	Gender	binominal	mode = M (252), least = F (221)	M (252), F (221)	0
regular	Age	integer	avg = 45.973 +/- 13.385	[17.000 ; 70.000]	0
regular	Marital-Status	binominal	mode = M (243), least = S (230)	S (230), M (243)	0
regular	Website-Activity	polynominal	mode = Seldom (286), least = F	Regular (148), Seldom (286), Fr	0
regular	Browsed-Product	binominal	mode = Yes (451), least = No (2	Yes (451), No (22)	0
regular	Bought-Product	binominal	mode = No (245), least = Yes (2	Yes (228), No (245)	0
regular	Bought-Wireless-Mouse-Keyboa	binominal	mode = Yes (378), least = No (9	Yes (378), No (95)	0
regular	Bought-Antivirus	binominal	mode = No (271), least = Yes (2	Yes (202), No (271)	0
regular	Payment-Method	polynominal	mode = Website Account (193),	Bank Transfer (156), Website Ac	0

Figure 19 Meta data for scoring data set predictions.

In Figure 20 the prediction for each customer's adoption group, along with confidence percentages for each prediction are showed. According to the four possible values in the label (Notebook-Adoption) there are four confidence attributes. The prediction is whichever category yielded the highest confidence percentage. RapidMiner is very (but not 100%) convinced that person **64466** (Row 6, Figure 20) is going to be a member of the early adaptor (41%). Despite some uncertainty, RapidMiner is completely sure that this person is *not* going to be a late majority (4.9%).


TrainingScoringOutput – RapidMiner@Kuala
File Edit Process Tools View Help
Result Overview
ExampleSet (Set Role (2))
Tree (Decision Tree)
○ Meta Data View ● Data View ○ Plot View ○ Annotations

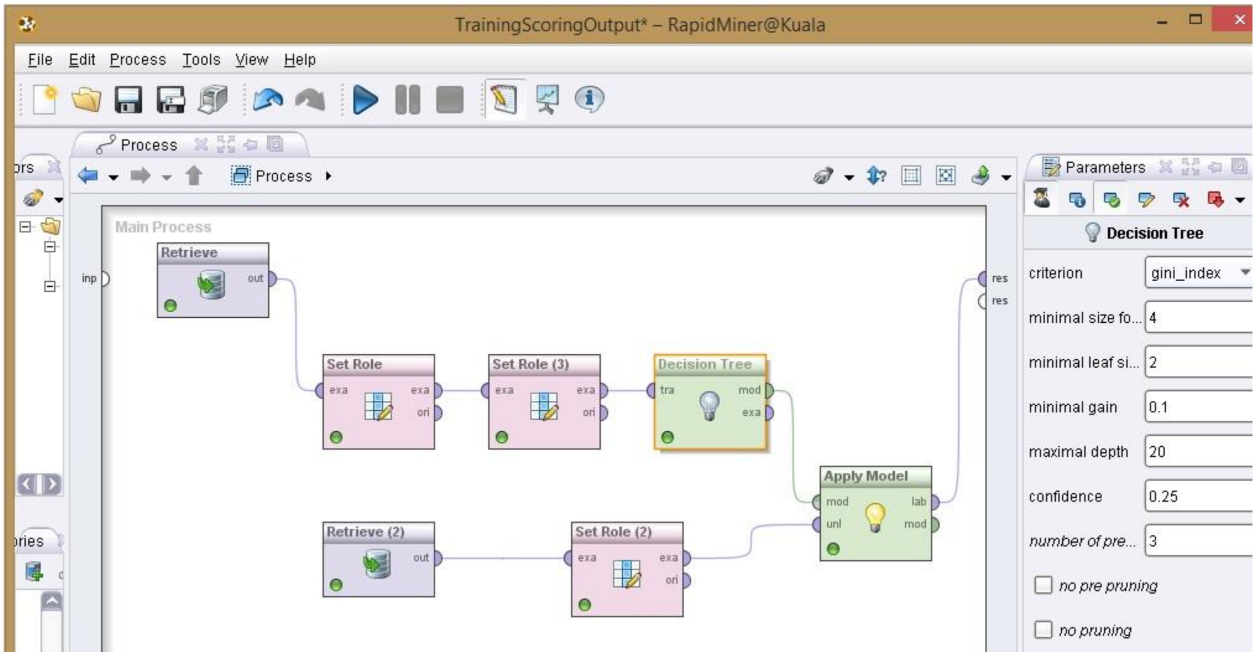
ExampleSet (473 examples, 6 special attributes, 9 regular attributes)

Row...	User-ID	confidence(Late Majority)	confidence(Innovator)	confidence(Early Adopter)	confidence(Early Majority)	prediction(Notebook-Adoption)
1	56031	0.049	0.317	0.410	0.224	Early Adopter
2	25913	0.049	0.317	0.410	0.224	Early Adopter
3	19396	0.758	0.022	0.055	0.165	Late Majority
4	93666	0.049	0.317	0.410	0.224	Early Adopter
5	72282	0.758	0.022	0.055	0.165	Late Majority
6	64466	0.049	0.317	0.410	0.224	Early Adopter
7	76655	0.758	0.022	0.055	0.165	Late Majority
8	48465	0.074	0.426	0.352	0.148	Innovator
9	19889	0.049	0.317	0.410	0.224	Early Adopter
10	63570	0.074	0.426	0.352	0.148	Innovator
11	63239	0.049	0.317	0.410	0.224	Early Adopter
12	67603	0.049	0.317	0.410	0.224	Early Adopter
13	65685	0.049	0.317	0.410	0.224	Early Adopter

Figure 20 Predictions and their associated confidence percentages using our decision tree.

The sales manager is going to induct the decision tree in order to get results with greater detail, or granularity in model change the **'criterion'** parameter to '**gini_index'** and then he re-runs the model, finally he generates the result according to the figure 23.

The Gini algorithm alone is much more sensitive than is the Gain Ratio algorithm in identifying nodes and leaves. Other independent variables (predictor attributes) are now being used, and the granularity with which the sales manager can identify each customer's likely adoption category is much greater. Website-Activity is still the single best predictor, but gender, and multiple levels of age have now also come into the decision tree, also single attribute is sometimes used more than once in a single branch of the tree.



criterion: gini_index

minimal size fo...: 4

minimal leaf si...: 2

minimal gain: 0.1

maximal depth: 20

confidence: 0.25

number of pre...: 3

□ no pre pruning

□ no pruning

Figure 21 constructing our decision tree model using the gini_index algorithm rather than the **gain ratio** algorithm.

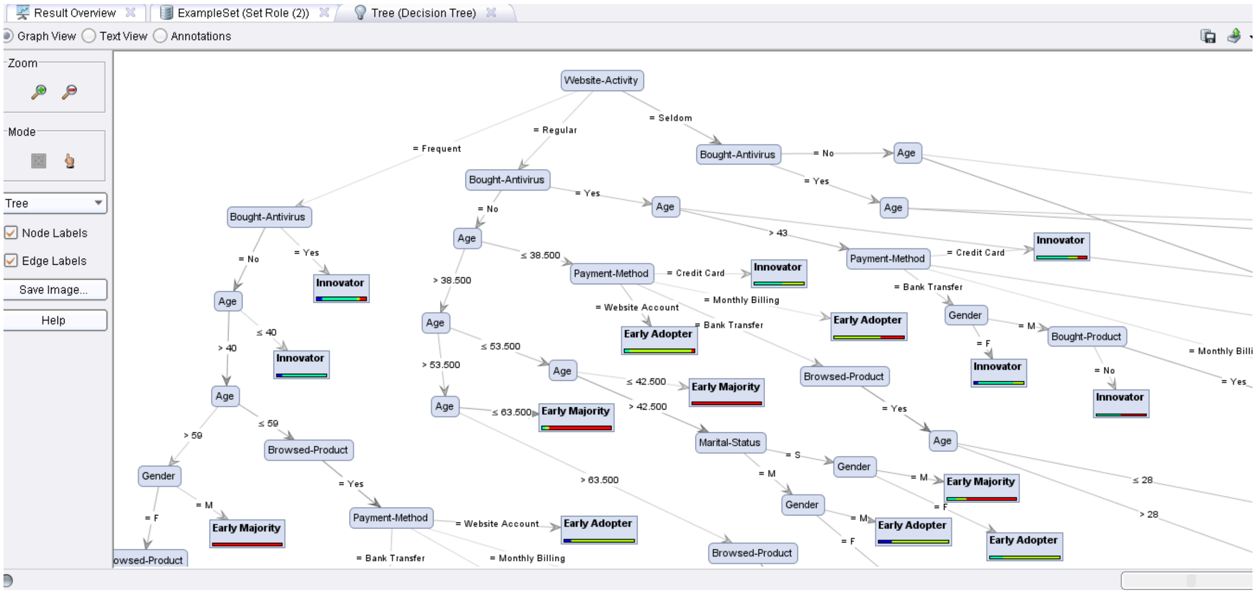
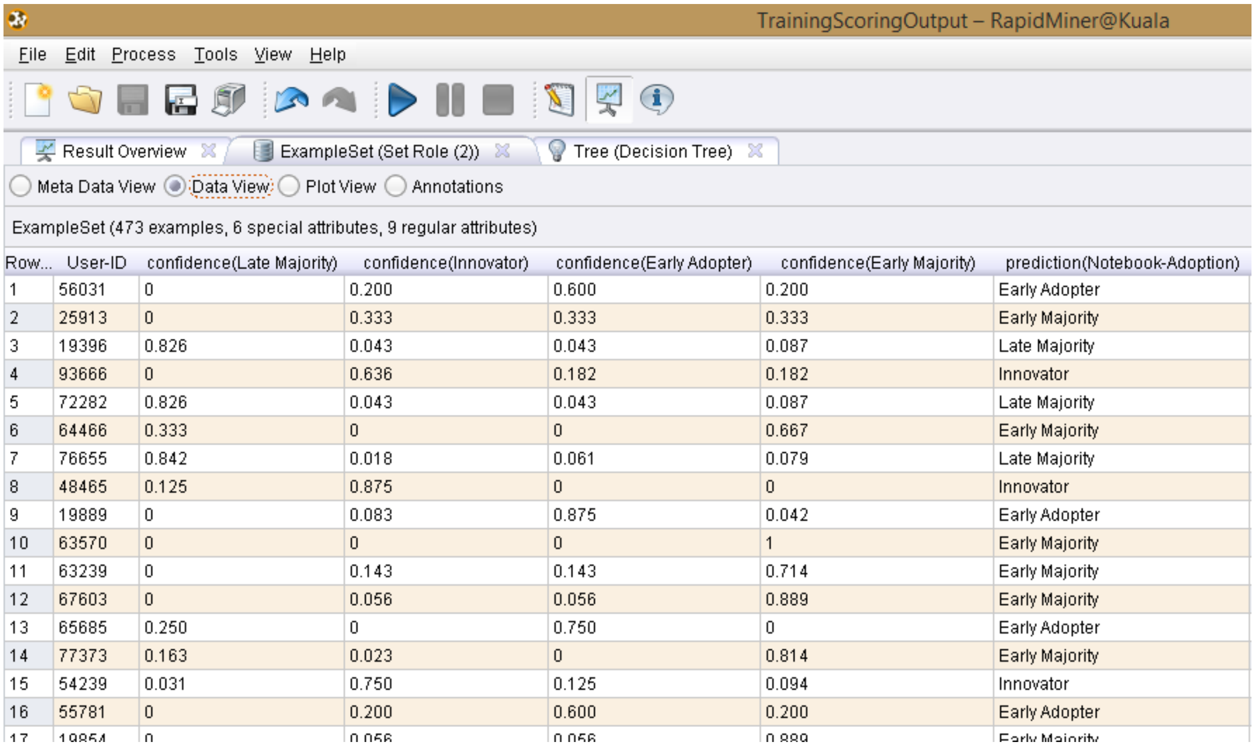


Figure 22 Tree resulting from a gini_index algorithm.

As the figure shows, confidence percentages of each person has changed in the prediction using Gini. For example now RapidMiner is very convinced that person **64466** (Row 6, Figure 23) is going to be a member of the early majority (66.7%). Despite some uncertainty, RapidMiner is completely sure that this person is *not* going to be an early adaptor (0%).



ExampleSet (473 examples, 6 special attributes, 9 regular attributes)

Row...	User-ID	confidence(Late Majority)	confidence(Innovator)	confidence(Early Adopter)	confidence(Early Majority)	prediction(Notebook-Adoption)
1	56031	0	0.200	0.600	0.200	Early Adopter
2	25913	0	0.333	0.333	0.333	Early Majority
3	19396	0.826	0.043	0.043	0.087	Late Majority
4	93666	0	0.636	0.182	0.182	Innovator
5	72282	0.826	0.043	0.043	0.087	Late Majority
6	64466	0.333	0	0	0.667	Early Majority
7	76655	0.842	0.018	0.061	0.079	Late Majority
8	48465	0.125	0.875	0	0	Innovator
9	19889	0	0.083	0.875	0.042	Early Adopter
10	63570	0	0	0	1	Early Majority
11	63239	0	0.143	0.143	0.714	Early Majority
12	67603	0	0.056	0.056	0.889	Early Majority
13	65685	0.250	0	0.750	0	Early Adopter
14	77373	0.163	0.023	0	0.814	Early Majority
15	54239	0.031	0.750	0.125	0.094	Innovator
16	55781	0	0.200	0.600	0.200	Early Adopter
17	19854	0	0.056	0.056	0.889	Early Majority

Figure 23 New predictions and confidence percentages using Gini.

In Figure 20, this person ID 64466 was calculated as having at least some percentage chance of landing in any one of the four adopter categories. Under the Gain Ratio algorithm, we were 31.7% sure he will be an innovator, but almost 41% sure he might also turn out to be an early adopter. In other words, the sales manager has confident he'll buy the new generation of Notebook early on, but we're not sure how early. The sales manager will have to decide during the deployment phase. But using Gini, the sales manager will be able to decide. In Figure 23, this same man is now shown to have a 66.7% chance of being an early majority and only a 33.3% chance of being late majority. The odds of him becoming part of the innovator and early adaptor crowd under the Gini model have dropped to zero. Now the sales manager knows that the person will be early majority and he is *predicting* with 100% confidence that he will adopt after 1 month. Note that while Gini has changed some of our predictions, it hasn't affected all of them.

Conclusion

Databases today may contains more than 1000^4 Terabyte (TB) of data and human analysts is unable to discover importance information or potentially useful information that can be hidden within these huge of data, therefore scientists with a many kind of data analysis tools tries to extract valid model and relationships in data in order to predict future trend or valid predictions.
Particularly innovative organization worldwide already tend to use data mining methods to find higher value customers in order to reconfigure their product offerings, to increase level of sales which can be lead to more profit, to minimize the cost of products, losses due to a specific error and also to increase revenues in company lifetime.

Scientists in order to utilize Data mining to solve their problem at first they should prepare and clean data from any error or missing values according their needs and also they need to know the exact problem in order to choose appropriate datamining method since Data mining has two different methods Prediction and Description Methods that both is used in a specific area, and according to the problem and their interesting goal apply appropriate data mining method.

Scientists or analyzers in prediction Methods such as Classification, Deviation Detection, and Regression use an interesting variables to predict unknown class of objects or value of a specific variable and in description method such as Association Rule Discovery, Sequential Pattern Discovery, Clustering analyzers extract interpretable patterns in order to describe the data.

Based on my practical example, RapidMinder software is used as a data mining tool that provides various technique and algorithms to be applied on data sets. In practical part decision tree had been chosen to classify customers according to type of class due to many reasons. Decision tree excellent predictive models in data mining, Decision tree is also able to handle both continuous and discrete variables, it is easy to interpret for small-sized trees and extremely fast at classifying unknown objects, to help predict the future trends and is easy to understand, also work more efficiently with discrete attributes. Briefly Decision tree generates predictions for the scoring observations.

Also decision tree is a popular technique for supervised classification, especially when the results are interpreted by human. And as I noticed, in Decision tree method Irrelevant attributes may affect badly the construction of a decision tree (E.g. ID numbers).

Most of retailers can use the decision tree approach to deal in marketing situations and identify important components of the decision process.

References

1. ***wikipedia.* [Online] https://en.wikipedia.org/wiki/Data.**

2. What is the difference between categorical, ordinal and interval variables? *UCLA (Institute for Digital Research and Education).* [Online] http://www.unesco.org/webworld/idams/advguide/Chapt1_3.htm.

3. Jiawei Han and Micheline Kamber. *Western Michigan University.* [Online] 2000. https://cs.wmich.edu/~yang/teach/cs595/han/ch01.pdf.

4. S. Sumathi, S.N. Sivanandam. . *Introduction to Data Mining and Its Applications.* Oct 12, 2006.

5. Data Mining Concepts. *microsoft.com.* [Online] https://msdn.microsoft.com/en-us/library/ms174949.aspx.

6. ZAKI, MOHAMMED J. SPADE: An Efficient Algorithm for Minings. *philippe Fournier-Viger.* [Online] 2001. http://www.philippe-fournier-viger.com/spmf/SPADE.pdf.

7. Aggarwal, Charu C. *Data Mining.* New York, 2015.

8. Building Classification Models: ID3 and C4.5. [Online] http://cis-linux1.temple.edu/~giorgio/cis587/readings/id3-c45.html.

9. k-means clustering algorithm. *Data Clustering Algorithms.* [Online] https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm.

10. Rakesh Agrawal - Ramakrishnan Srikant. Rakesh Agrawal. [Online] http://rakesh.agrawal-family.com/papers/vldb94apriori.pdf.

11. wikipedia. [Online] https://en.wikipedia.org/wiki/Regression_analysis.

12. Dr. Saed Sayad. KNN - Classification. *Data Mining.* [Online] http://www.saedsayad.com/k_nearest_neighbors.htm.

13. *Wikipedia.* [Online] http://en.wikipedia.org/wiki/Anomaly_detection.

14. Zaki, Mohammed J. Efficient Enumeration of Frequent Sequences. *Ca' Foscari University of Venice.* [Online] Ca' Foscari University. http://www.dsi.unive.it/~dm/CIKM98_ps.pdf.

15. Microsoft. Data Mining Algorithms . *Microsoft.* [Online] https://msdn.microsoft.com/en-us/library/ms175595.aspx.