BAIT3013
Business Intelligence

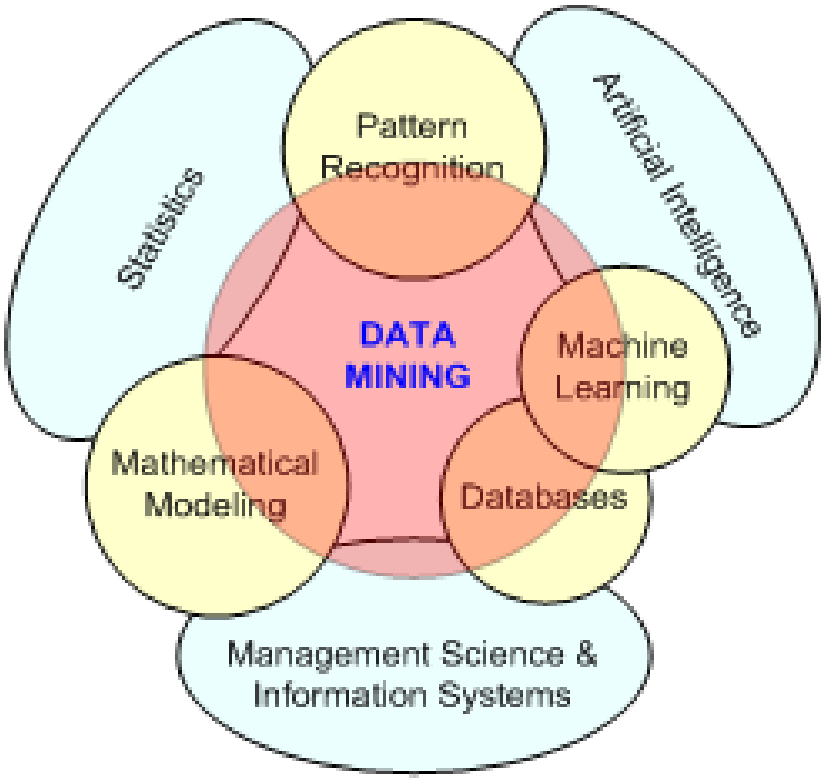
Lecture 4

Data Mining For Business Intelligence

What is Data Mining ?

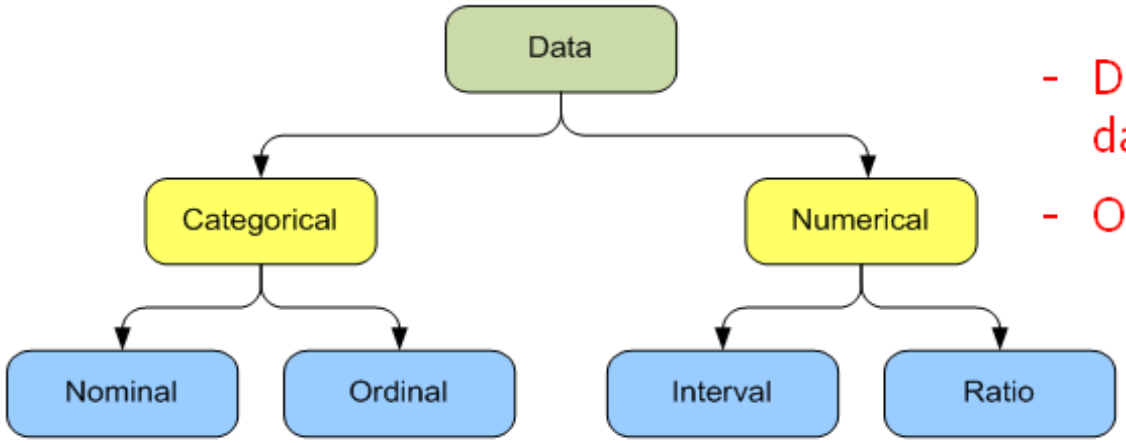
- Used to discover the unknown patterns.
- Also known as knowledge discovery, knowledge extraction, pattern analysis

Data Mining at the Intersection of Many Disciplines



Data in Data Mining

- Data: a collection of facts usually obtained as the result of experiences, observations, or experiments
- Data may consist of numbers, words, and images
- Data: lowest level of abstraction (from which information and knowledge are derived)



- DM with different data types?
- Other data types?

What Does DM Do?
How Does it Work?

- DM extracts patterns from data
  - Pattern?

    A mathematical (numeric and/or symbolic) relationship among data items
- Types of patterns
  - Association
  - Prediction
  - Cluster (segmentation)
  - Sequential (or time series) relationships

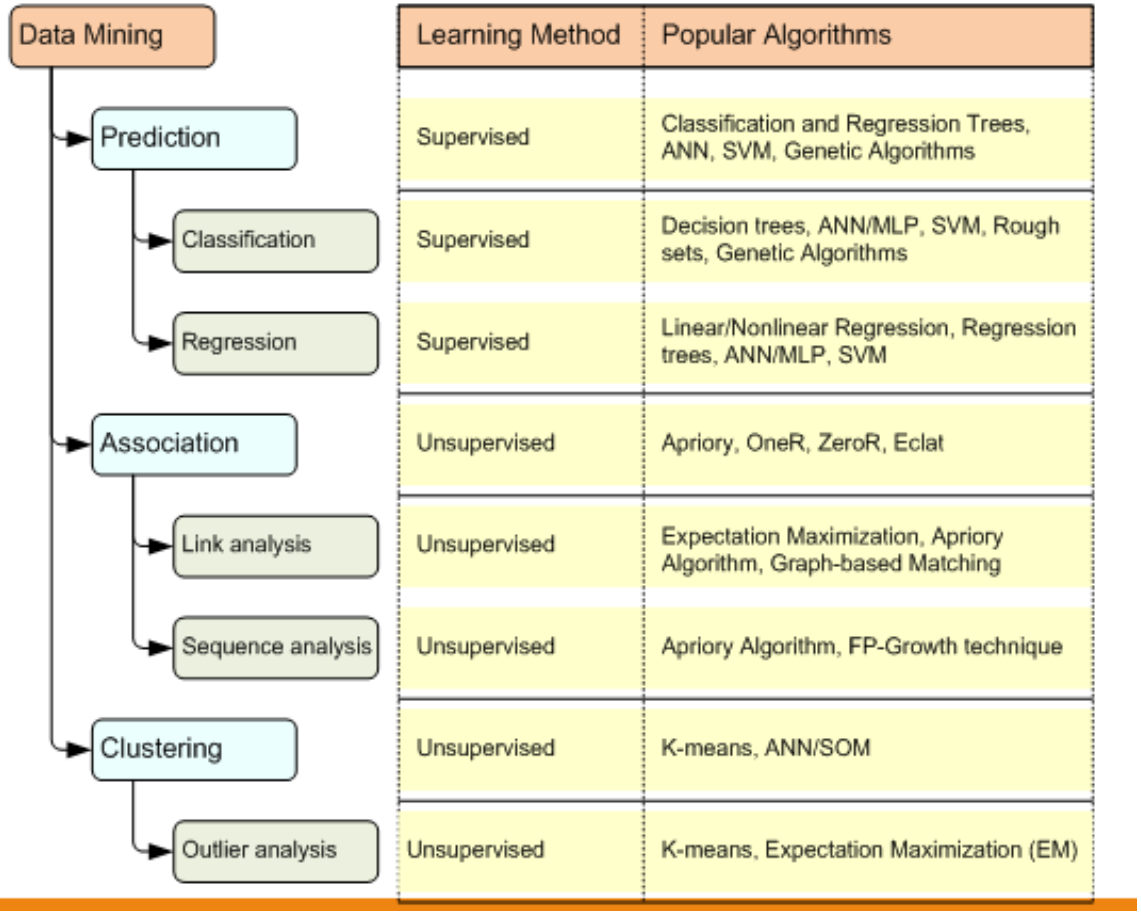
Types of patterns

- Association
  - Basket Market Analysis
    - Given a database of purchase transactions and for each transaction a list of purchased items
    - Find rules that correlate a set of items occurring in a list with another set of item
    - For example:
      - 98% of people who purchase tires and auto accessories also get automotive services done
      - 60% of people who buy diapers also buy a beer

Types of patterns

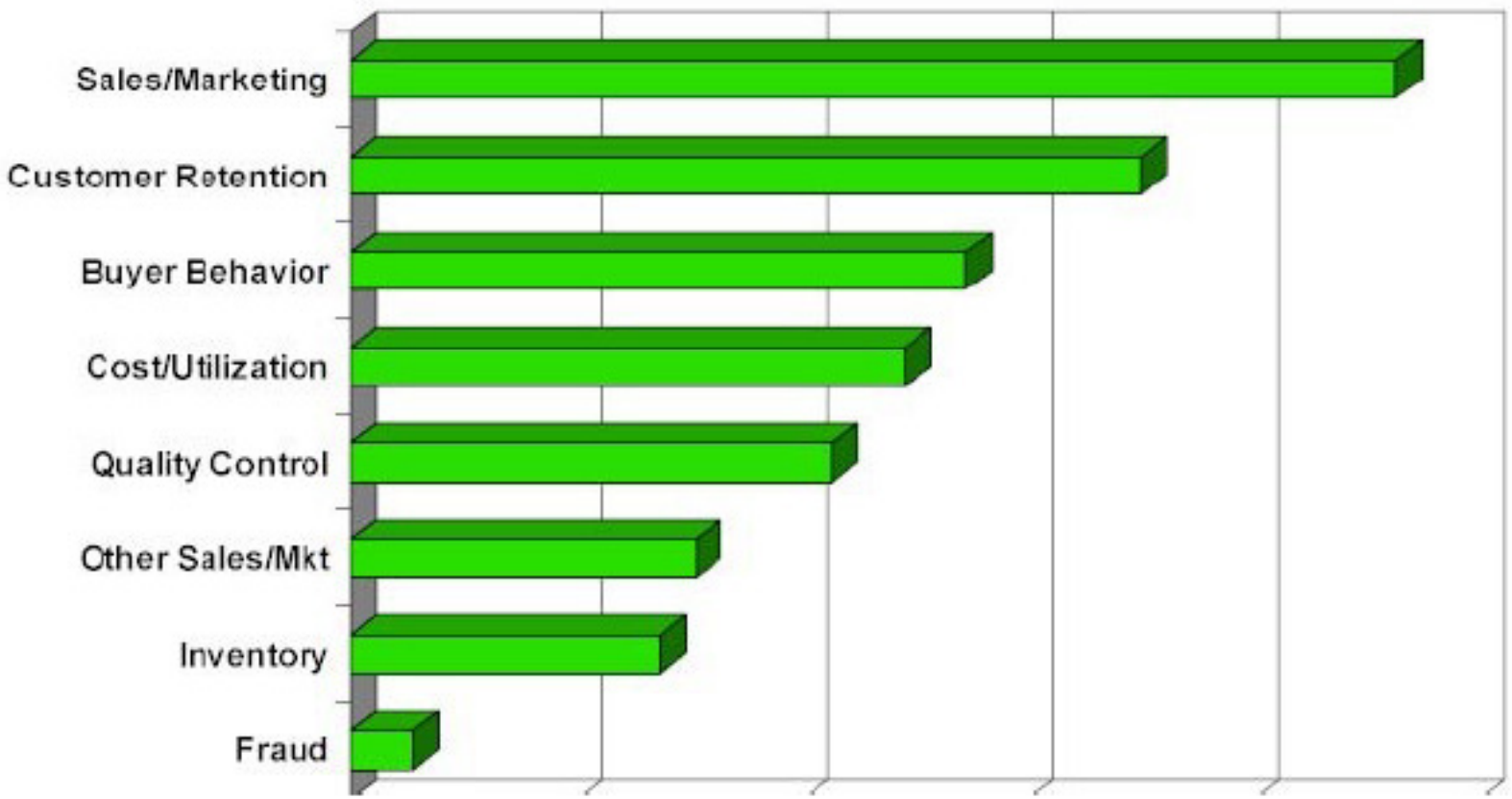
- Association
  - Classification
    - classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
  - Prediction
    - models continuous-valued functions, for example : predicts unknown or missing values

A Taxonomy for Data Mining Tasks



Learning Method	Popular Algorithms
Supervised	Classification and Regression Trees, ANN, SVM, Genetic Algorithms
Supervised	Decision trees, ANN/MLP, SVM, Rough sets, Genetic Algorithms
Supervised	Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM
Unsupervised	Apriory, OneR, ZeroR, Eclat
Unsupervised	Expectation Maximization, Apriory Algorithm, Graph-based Matching
Unsupervised	Apriory Algorithm, FP-Growth technique
Unsupervised	K-means, ANN/SOM
Unsupervised	K-means, Expectation Maximization (EM)

Data Mining Applications



Data Mining Applications

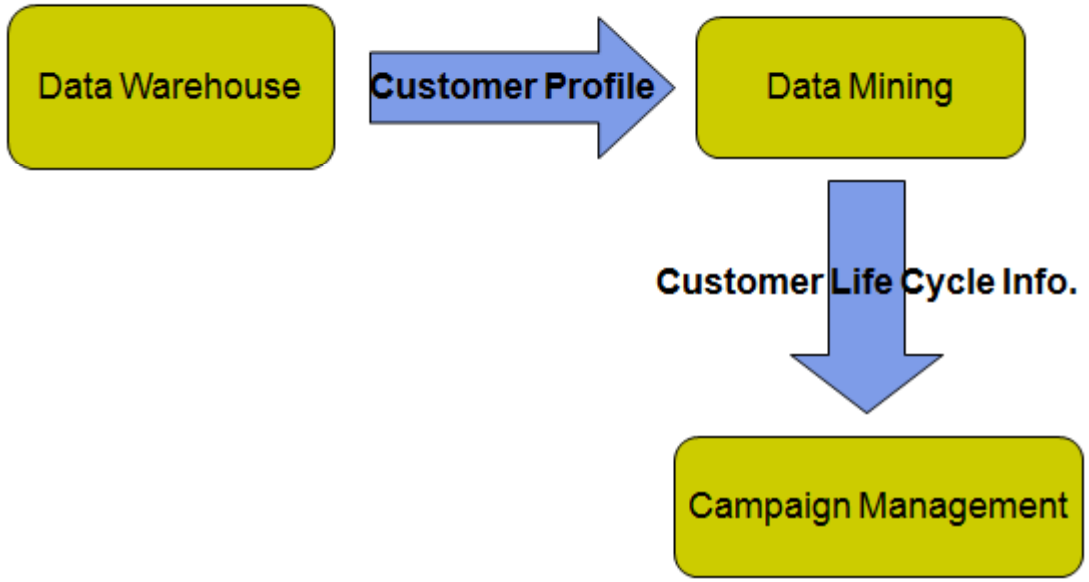
- Banking & Other Financial

Data Mining Applications

- Retail
  - What determines the best product mix to sell on a regional level?
  - What are the latest product trends?
  - When is a merchandise department saturated?
  - What are the times when a customer is most likely to buy?
  - What types of products can be sold together?

Data Mining Applications

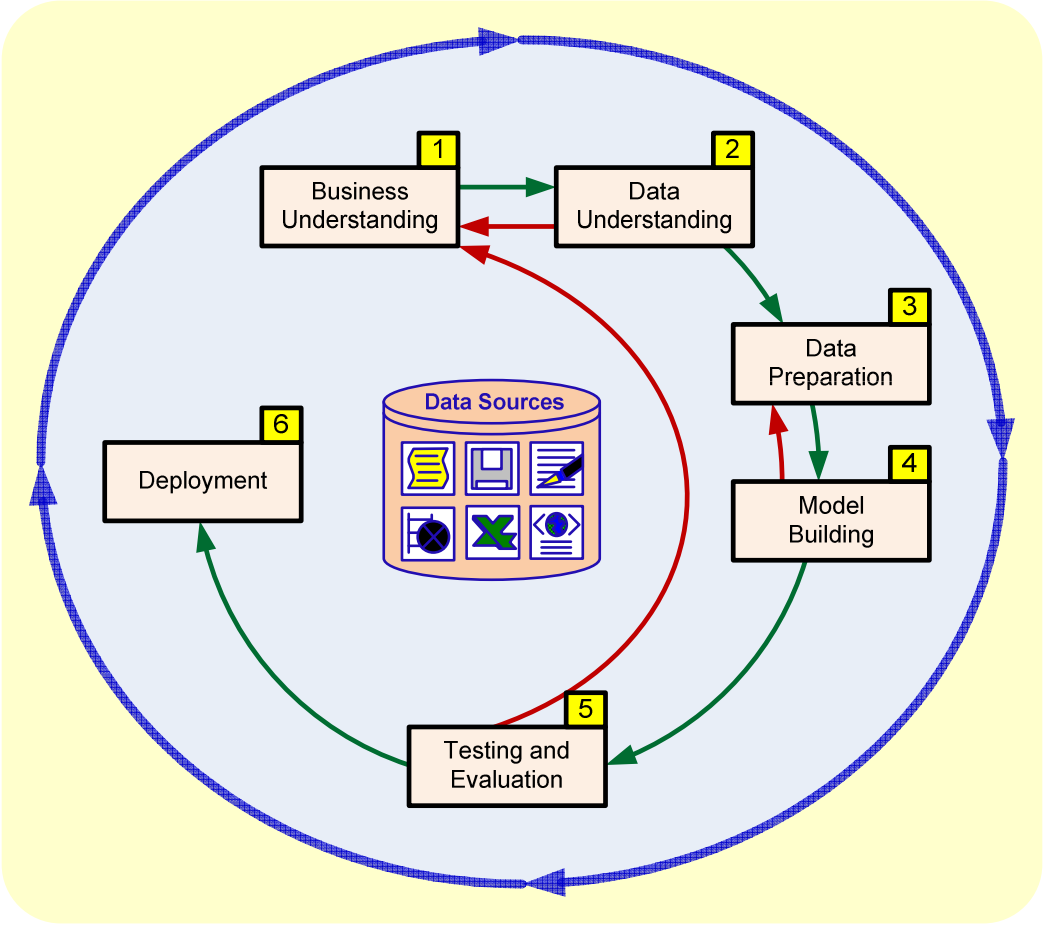
- Customer Relationship Management



Data Mining Process

- Most common standard processes:
  - CRISP-DM (Cross-Industry Standard Process for Data Mining)
  - SEMMA (Sample, Explore, Modify, Model, and Assess)
  - KDD (Knowledge Discovery in Databases)

Data Mining Process: CRISP-DM



Data Mining Process: CRISP-DM

- Step 1: Business Understanding
- Step 2: Data Understanding
- Step 3: Data Preparation
- Step 4: Model Building
- Step 5: Testing and Evaluation
- Step 6: Deployment

Phase 1 : Business Understanding

- Focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives

Phase 1 : Business Understanding

- Determine business objectives
  - Thoroughly understand, from a business perspective, what the client really wants to accomplish
  - Uncover important factors, at the beginning, that can influence the outcome of the project
- Assess situation
  - More detailed fact-finding about all of the resources, constraints, assumptions and other factors that should be considered

Phase 1 : Business Understanding

- Determine data mining goals
  - For example : “Predict how many products a customer will buy given their purchases over the past three years”.

Phase 1 : Business Understanding

- Produce project plan
  - Describe the intended plan for achieving the data mining goals and the business goals

Phase 2 : Data Understanding

- Collect initial data
  - Acquire within the project the data listed in project resources
- Describe data
  - Examine the “gross” or “surface” properties of the acquired data

Phase 2 : Data Understanding

- Explore data
  - tackles the data mining questions, which can be addressed using querying, visualization and reporting
- Verify data quality
  - examine the quality of the data, addressing questions such as:
  - “Is the data complete?”, Are there missing values in the data?”

Phase 3 : Data Preparation

- Select data
- Clean data
- Construct data
- Integrate data
- Format data

Phase 4 : Modeling

- Select modeling techniques (ANN, decision tree)
- Build model
- Access model

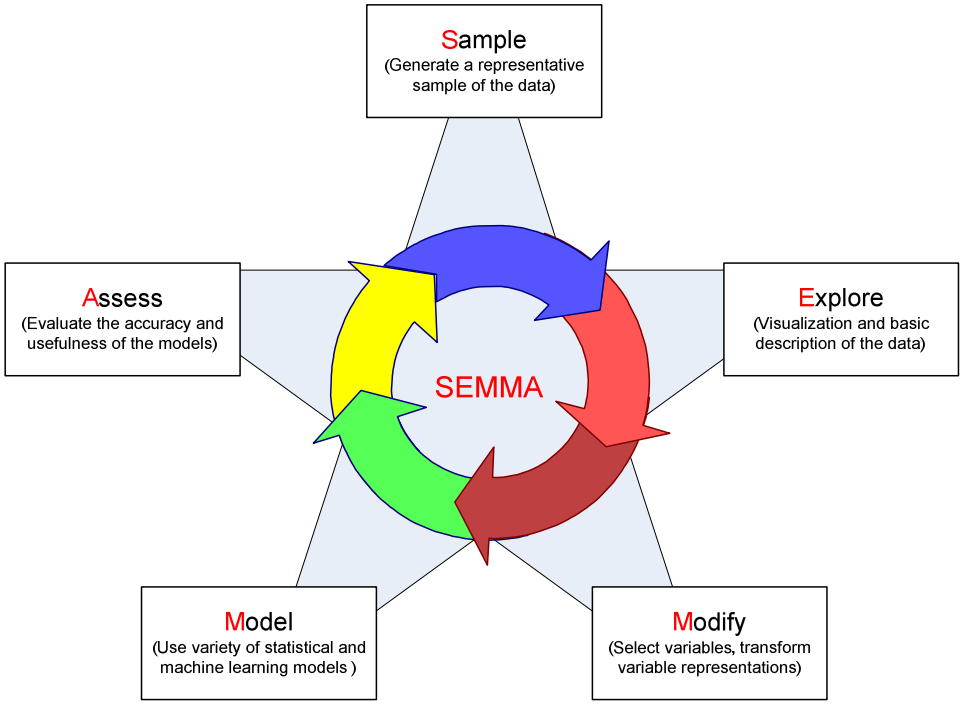
Phase 5 : Evaluation

- Evaluation of model
  - How well it performed on test data
- Evaluate results
  - Access the degree to which the model meets the business objectives

Phase 6 : Deployment

- Determine how the results need to be utilized
- Who needs to use them ?
- How often do they need to be used

Data Mining Process: SEMMA



Phase 1 : Sample

- This stage consists of sampling the data by extracting a portion of large data set big enough to contain the significant information

Phase 2 : Explore

- This stage consists on the exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas

Phase 3: Modify

- This stage consists of modification of the data by creating, selecting and transforming the variables to focus the model selection process.

Phase 4 : Model

- This stage consists on modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.

Phase 5 : Access

- This stage consists on assessing the data by evaluating the usefulness and reliability of the findings from the DM process and estimate how well it performs.

KDD

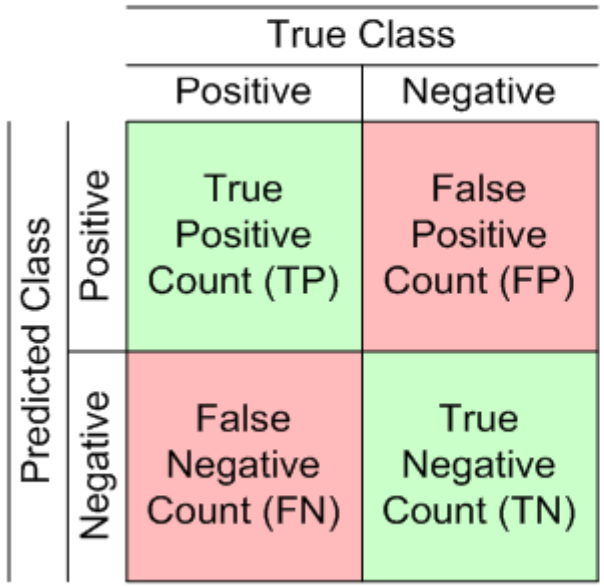
- Phase 1 : Selection
- Phase 2: Pre-processing
- Phase 3 : Transforming
- Phase 4 : Data mining
- Phase 5 : Evaluating

Data Mining Methods: Classification

- Most frequently used DM method
- Part of the machine-learning family
- Learn from past data, classify new data
- The output variable is categorical (nominal or ordinal) in nature

Accuracy of Classification Models

- In classification problems, the primary source for accuracy estimation is the confusion matrix.



$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$True\ Positive\ Rate = \frac{TP}{TP + FN}$$

$$True\ Negative\ Rate = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

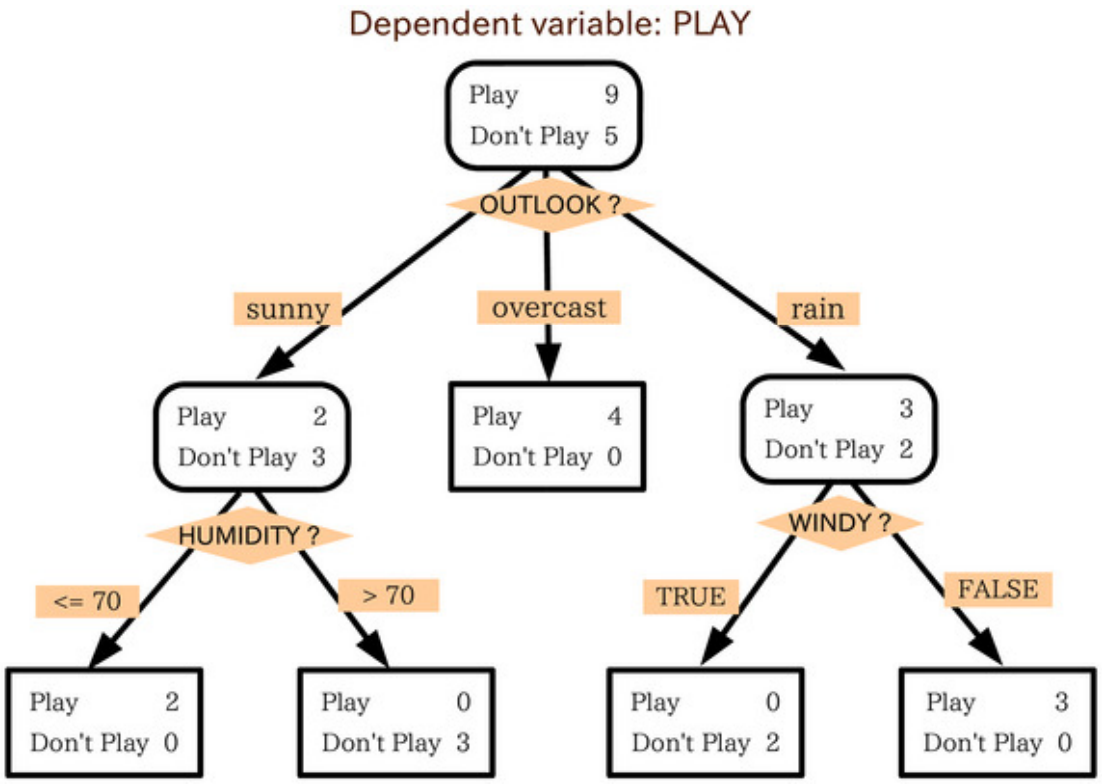
Classification Techniques

- Decision tree analysis
- Statistical analysis
- Neural networks

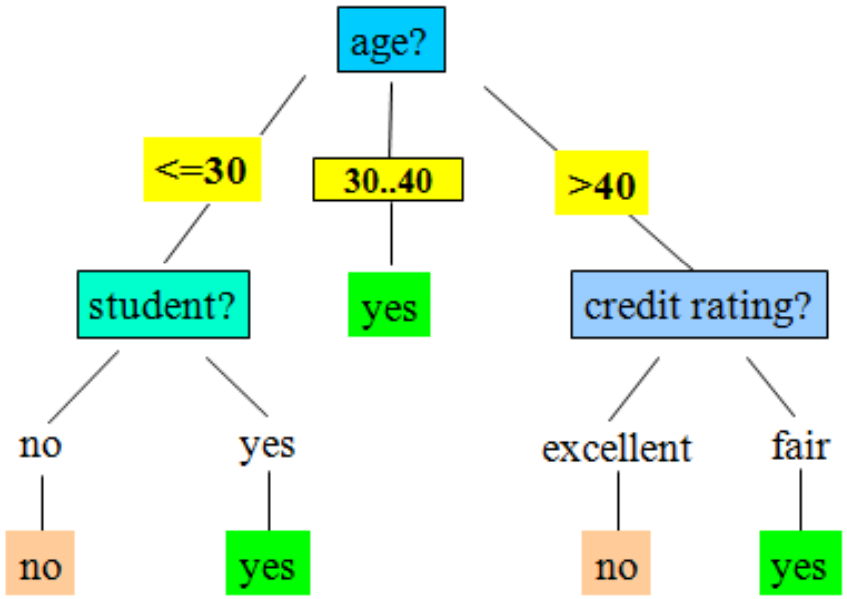
Decision Trees

- Employs the divide and conquer method
- Recursively divides a training set until each division consists of examples from one class
  - Create a root node
  - Select the best splitting attribute.
  - Add a branch to the root node for each value of the split. Split the data into mutually exclusive subsets along the lines of the specific split.
  - Repeat the steps 2 and 3 for each and every leaf node until the stopping criteria is reached.

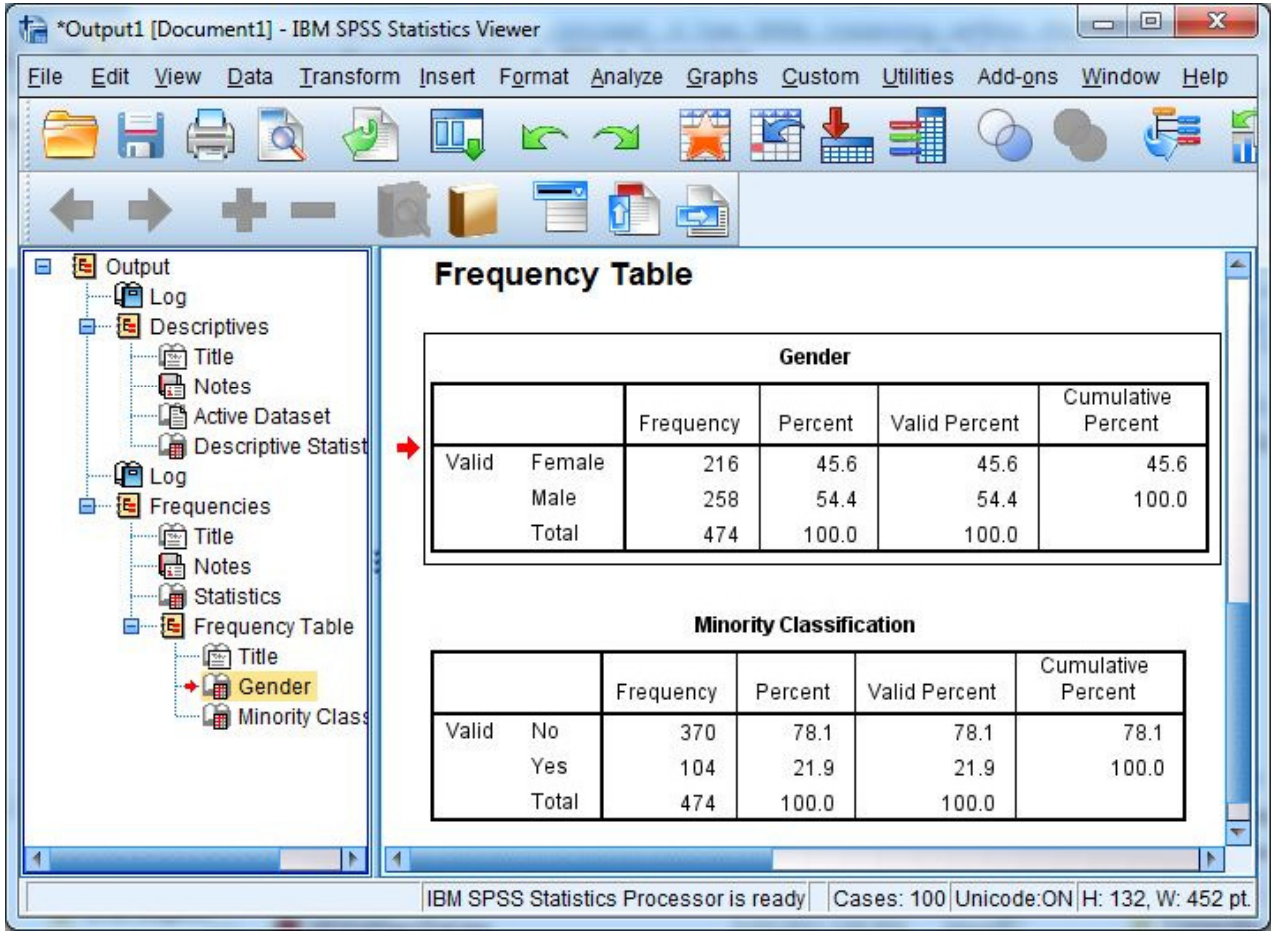
Decision Trees



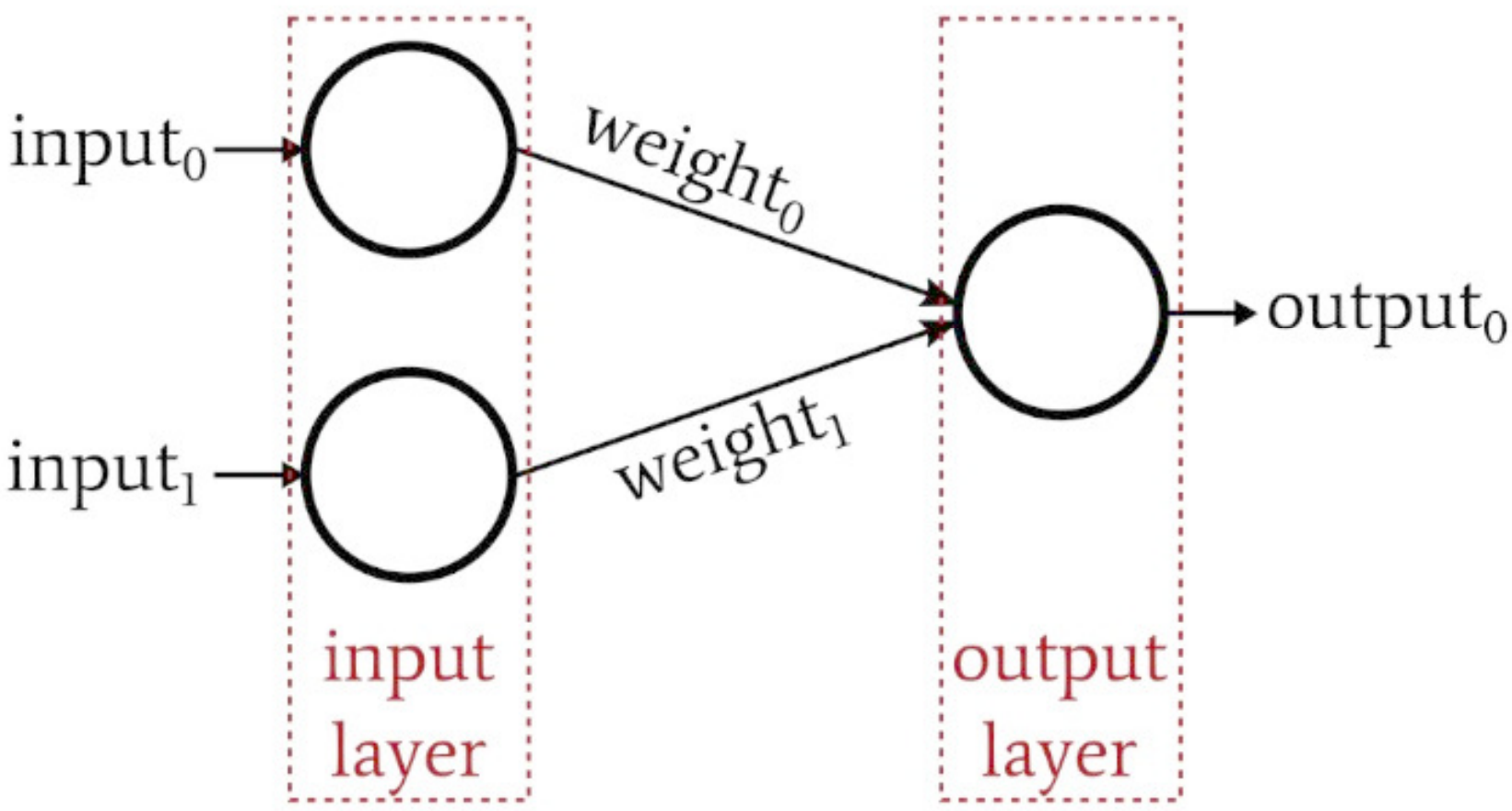
Decision Trees



Statistical analysis



Neural Network



References

- Sharda, R., Delen, D., Turban, E., 2018. Business Intelligence, Analytics, and Data Science : A Managerial Perspective. Pearson.
- Samaddar, S. and Nargundkar, S.. 2019. Data Analytics : Effective Methods for Presenting Results. CRC Press.
- Verbeke, W., Baesens, B. and Bravo, C.. 2018. Profit Drive Business Analytics : A Practitioner's Guide to Transforming Big Data into Added Value. John Wiley & Sons.