Data mining and Warehousing

Unit 2:

Data mining Description

Module Code: **CSC5901**

Olivier Angel Kevin ISHIMWE

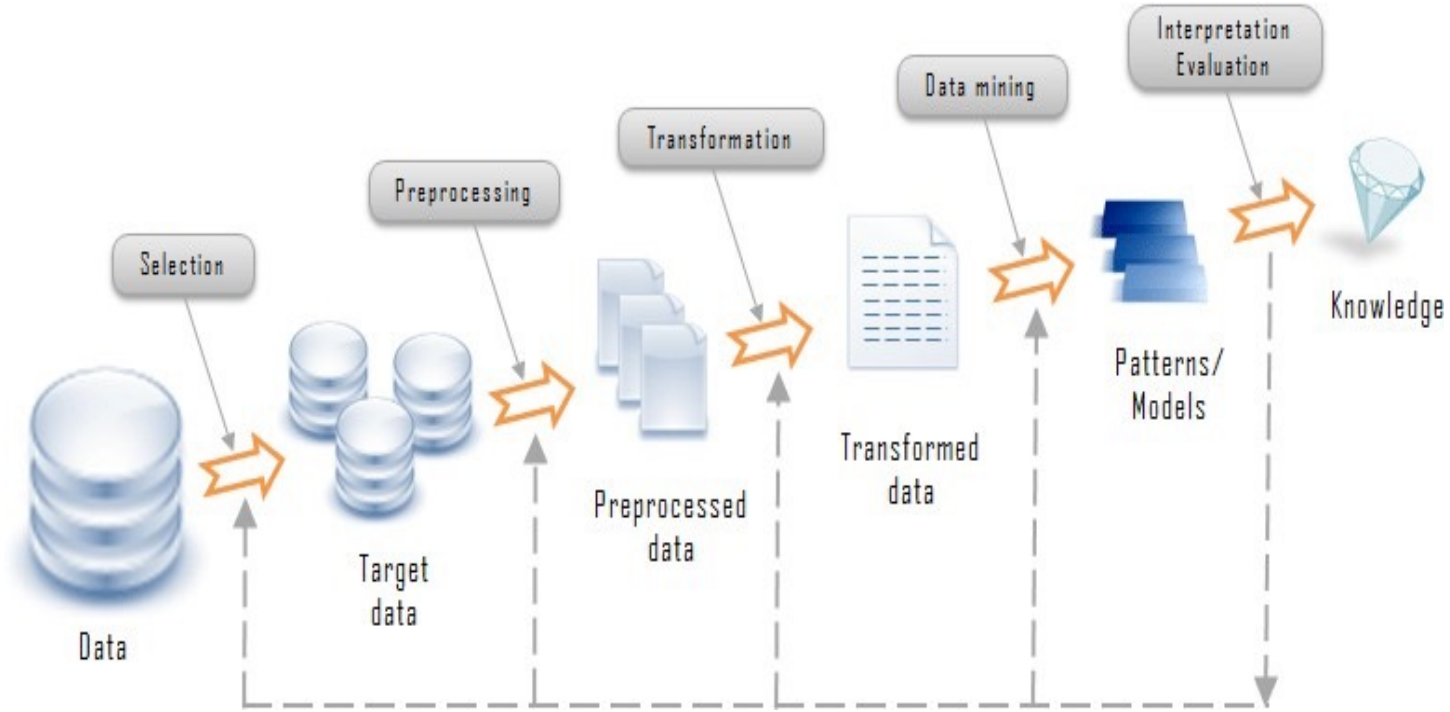
UNIVERSITY Of KIGALI

Steps involved in a data mining process

- Extract, transform and load data into a data warehouse
- Store and manage data in a multidimensional databases
- Provide data access to business analysts using application software
- Present analyzed data in easily understandable forms, such as graphs

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

Steps involved in a data mining process



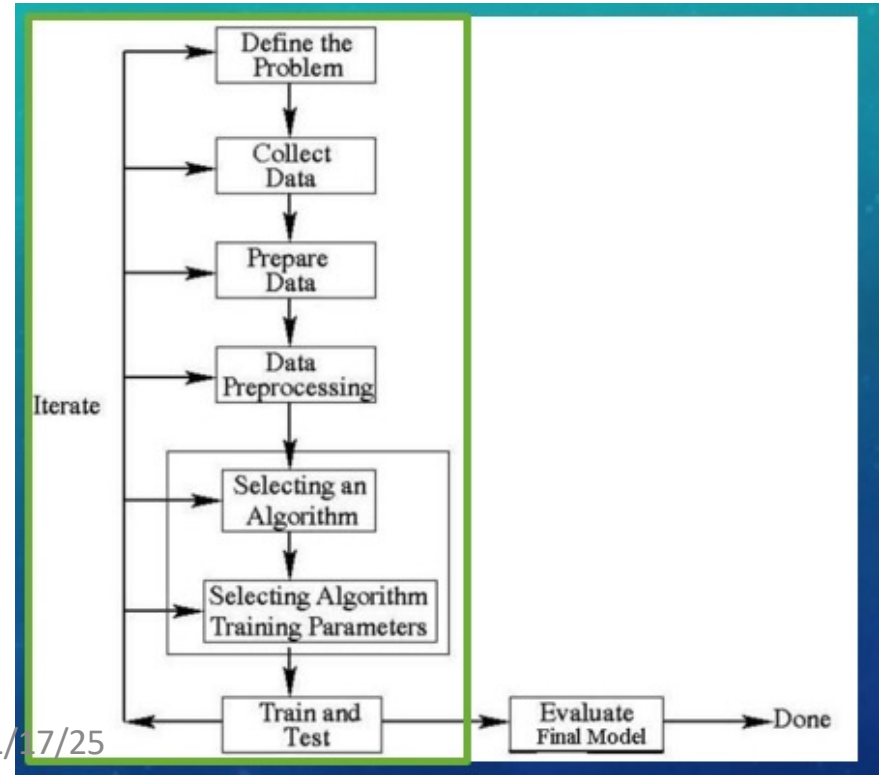
P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

THE DATA MINING PROCESS

- provide an overview of the eight steps in the data mining process.
  identify the issues involved in defining a data mining problem.
  determine when to use and not to use data mining
  explain how to conduct and experiment to determine whether more data is needed

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

8 STEP DATA MINING PROCESS



P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

8 STEP DATA MINING PROCESS

- **Defining the problem**

This defines the objective of the whole data mining process.

- **Collecting data**
- **Preparing data**
- **Pre-processing**

- The next three steps involve collecting the required data, preparing and pre-processing the data before a data mining technique could be applied.

- **Selecting an algorithm and training parameters**

- Select a model and perform training and testing to evaluate if the model is good.

- **Training and testing**

- Because one does not know how an algorithm will perform on a data set, one needs to try different models and compare each model against each other model.

- **Iterating to produce different models**

- Due to the previous set, many iterations may be required for all steps in order to determine the optimal model.

- **Evaluating the final model**

- The best model is selected based on the estimated accuracy. This model is then used for future predictions.

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

8 STEP DATA MINING PROCESS CONTINUED…

Defining the problem:

- One needs to determine which problems are suitable for data-driven modelling.
- How does one evaluate the results?
- Is it a classification or estimation problem?
- What are the inputs and outputs required for solving the problem?

When TO DO data mining:

- When **no good existing** solution exists & the problem has the following **characteristics**:
  - Lots of **data**
  - The problem is **not** well understood
  - The problem can be characterised as an **input-to-output** relationship
  - Existing models have **strong** and possibly **erroneous assumptions**

When NOT to do data mining:

- When the problem:
  - Has a **complete**, **closed-form** mathematical solution.
  - It is well understood and has a **good analytical** or **rule-based** solution

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

8 STEP DATA MINING PROCESS CONTINUED…

How do you evaluate the results?

- What level of accuracy would be considered successful?
- How will you benchmark the performance of a developed solution?
- What existing alternatives will you compare against?
- What kind of data will be used to evaluate the various models?
- What will the models be used for and how well do they support that purpose?

Classification or Estimation?

Discrete outputs = classification problem

Continuous outputs = estimation problem

Borderline outputs = can be either based upon the granularity of the outputs.

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

CLASSIFICATION VERSUS ESTIMATION

Classification:

In classification learning, the learning scheme is presented with a set of classified examples from which it is expected to learn a way of classifying unseen examples.

In plain English, the main idea behind classification is it uses 2 values, it is either yes or no, 1 or 0. It may belong to a single class not both. The algorithm will then learn how to predict future unseen data based upon the training data.

Estimation:

In estimation, the classification of the data is not based on an absolute value (IE 0 or 1) but rather on real numbers between 0 and 1. So what if the value sits between two classes? For example, 0,5 could fall into either class however if the classes state 0-0,49 and 0,5 to 1 then it would inevitably fall into the later class. When the model is first designed, a set of predefined classes are laid out in order to prevent a data point from lying between two or more categories.

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

WHAT ARE THE INPUTS AND OUTPUTS

An example: The inputs and outputs when classifying loan application at a bank. Outputs:

The outputs could possibly be “high risk” or “low risk”

Inputs:

The inputs may current salary, outstanding liabilities, bank account balances, other income, number of years employed.

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

IMPORTANT ISSUES

Causal and Non-Causal outputs:

Causal attributes occur when one attribute causes another. Meaning it is used to predict another attribute or it is included in the calculation. It is important to avoid using non-causal attributes as they produce a model not a representation of the future data.

Inputs:

The inputs must contain enough information to be able to generate the desired output. If the inputs contain insufficient data the accuracy of the final model will inevitably decrease.

Data set:

The data chosen must be an accurate representation of the future data set. If it is not the accuracy when trying to predict the future will produce undesired outputs.

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

HOW MUCH DATA IS ENOUGH

Key: The amount of data required depends on the problem complexity as well as the amount of noise in the data.

Each learning algorithm follows its own learning curve, the accuracy increases as the data size increases, however when the algorithm reaches its optimal performance further increase of the data set **CANNOT** improve performance.

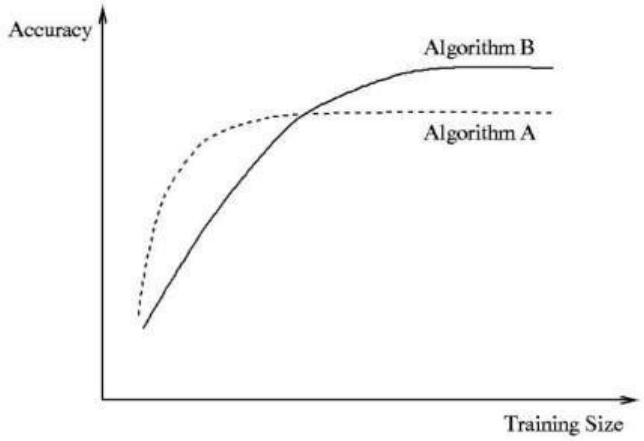


Figure 1:

Algorithm A reaches optimal performance sooner than Algorithm B, however the accuracy cannot increase after a certain point.

Note: One needs to experiment to determine when the algorithm reaches it's optimal performance. Can be determined by:

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

WHAT WE HAVE LEARNED:

- The 8 steps of data mining as well as the need for iteration in order to produce a satisfactory model.
- The difference between estimation and classification
- When to do data mining and when not to
- Important issues when defining a problem

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

REVIEW QUESTIONS

1. Explain how are you going to decide whether a given problem is suitable for a data mining solution.

Given a problem , I am going to check if there is no good existing solution and the problem has the following characteristics:

- Lots of data
- The problem is not well understood
- The problem can be characterised as an input-to-output relationship
- Existing models have strong and possibly erroneous assumptions

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

2. Suppose the data provided is the last promotional mail-out records which consist of information about each of the 100 customers (name, address, occupation, salary) and whether each individual customer responded to the mail (i.e., an attribute indicating "yes" or "no". You are asked to produce a data mining solution, that is, a model describing the characteristics of customers who are likely, as well as unlikely, to respond to the promotional mail-out. The company could then use this model to target customers who are likely to respond to the next promotional mail-out for the same product.

Discuss the following issues:

- Is this problem suitable for data mining solution? Yes

  This problem is suitable for data mining solution because there is no good existing solution and it has a lot of data, that can be easily characterised as an input-to-output relationship.

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

•Does the information above give us a classification or estimation problem? Justify your answer Classification problem, the model required is to be used to predict whether a customer is a repeat buyer or not.
i.e. two classes of customers.

- **What are the inputs and output?**

The inputs are the attributes: name, address, occupation, salary;

The outputs is an attribute called: “Repeat Buyer” with the labels “yes” or “no”

- **What is the alternative to producing a model?**

As the task is to select a subset of customers to send the promotional mail-out to, instead of building a model in order to identify the customers, one can simply perform a random selection of customers from all the available customers

- **How you will use the data for training a model and evaluating the model?**

In order to evaluate the trained models using test data, the given data set of 100000 customers can be split into two subsets; one for the training and one for the testing. There is no need to use more elaborate evaluation method such as 10-fold cross validation method, as the data set is big enough and the reserved for testing is unlikely to degrade the predictive accuracy of the trained model.

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

3. Let say you are given a set of training data with 50% class "positive" and 50% class "negative", and you have explored several models and selected the best model. You can now use the best model to do future prediction. Now, you are informed that the future data you are going to get is likely to have the following class distribution: 90% class "positive" and 10% class "negative".

Would you go ahead to use the best model to do prediction for all future data? Provide a reason for your answer. In the case that your answer is no, you shall also provide an alternative solution to do prediction for the future data.

The training data is not a representative of the testing data in terms of class distribution; the one has a 1:1 ratio and another has 9:1 ratio. This is why one should not use the model from a data set not representative of the testing data to do prediction for future data.

If one does, then it is likely to lead to poor performance i.e. high error rate.

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

The same model applied to 9:1 ratio testing data will perform a lot worse, as 80% of the time the model will predict a class “positive” regardless of the input. The simple model will have accuracy of 90% on the 9:1 ratio testing data.

One should be aware of how some models allow one to adjust outputs to match the class distribution of the testing data. Only with this adjustment, than one can use the model trained from different class distribution to be applied.

4. How does one decide whether to collect more data or not in a non-time series data mining task?

When there is a small portion of data to mine, one should collect more data.

Where the is a lot of data one should not collect more data is there is sufficient data to mine.

In both situations time is not a factor, everything is based on the quantity of data required to mine a data mining task.

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

Representation for visualizing the discovered patterns(Data)

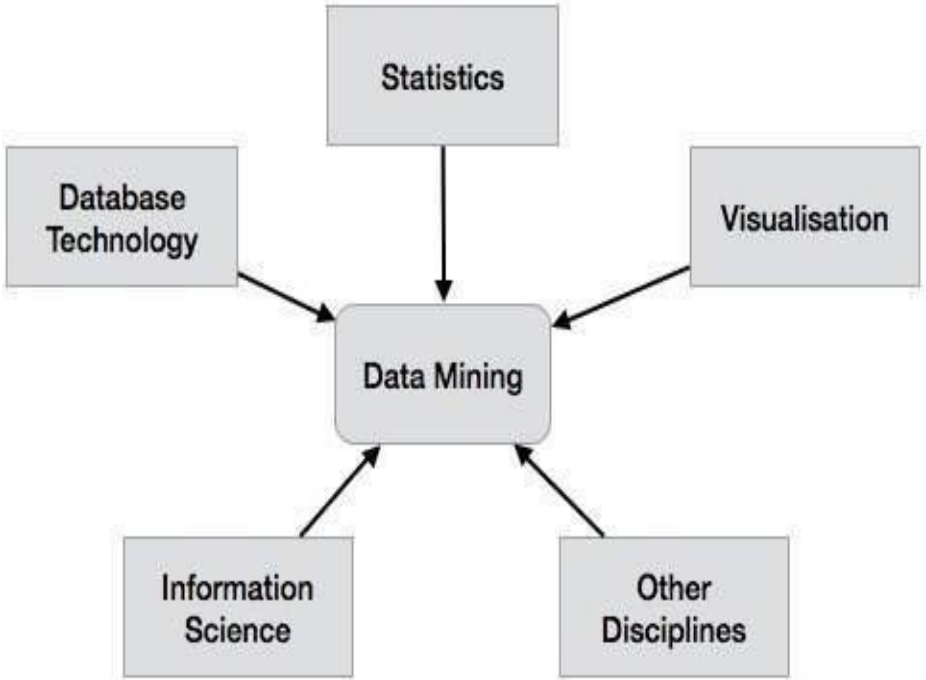
These representations may include the following. —

• Rules • Graphs	• Tables • Decision Trees	• Charts • Cubes

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

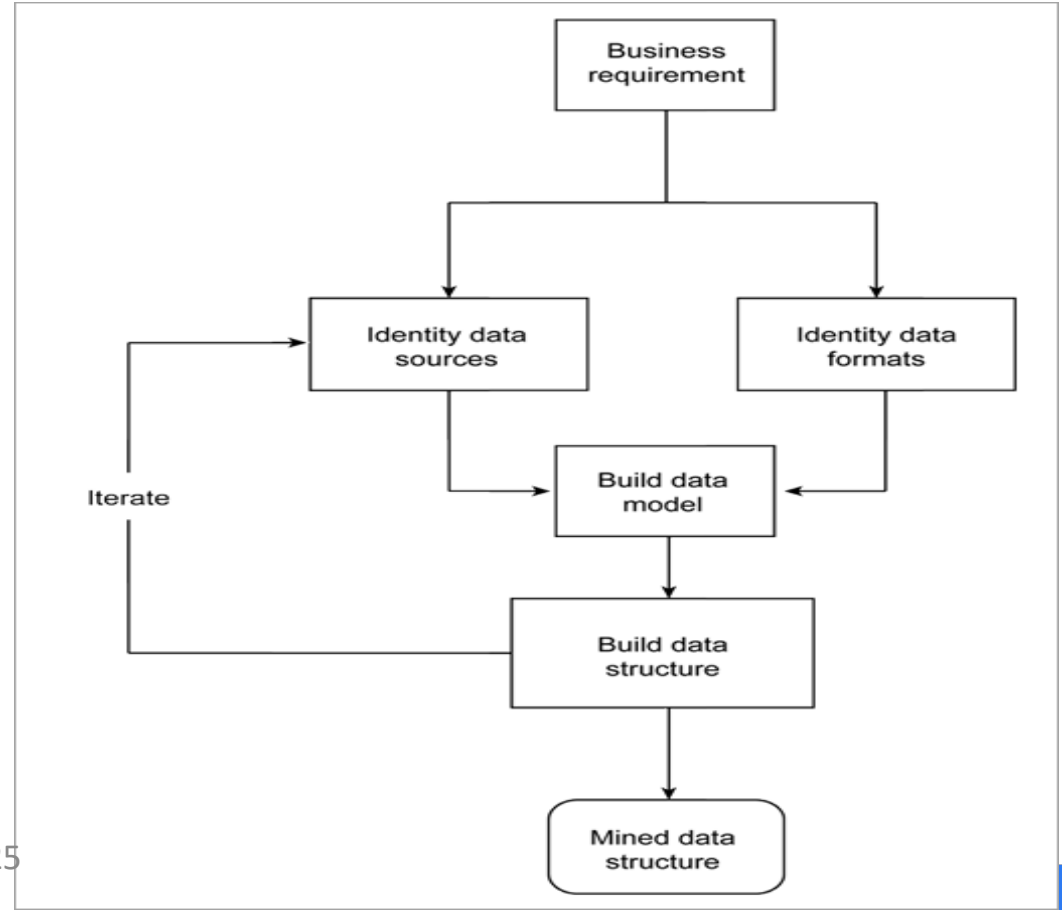
Data Mining System Classification

- A data mining system can be classified according to the following criteria –



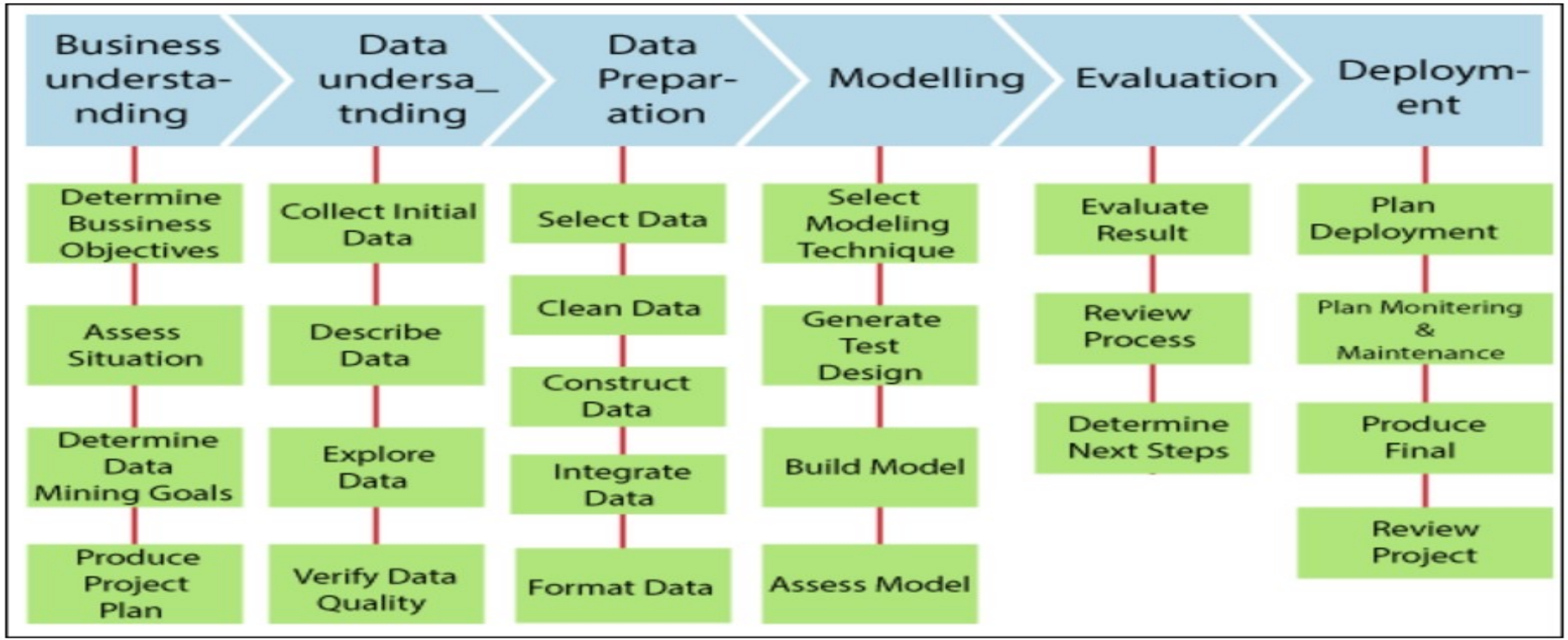
P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

Data Extraction Process



P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

Data mining Implementation Process



P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

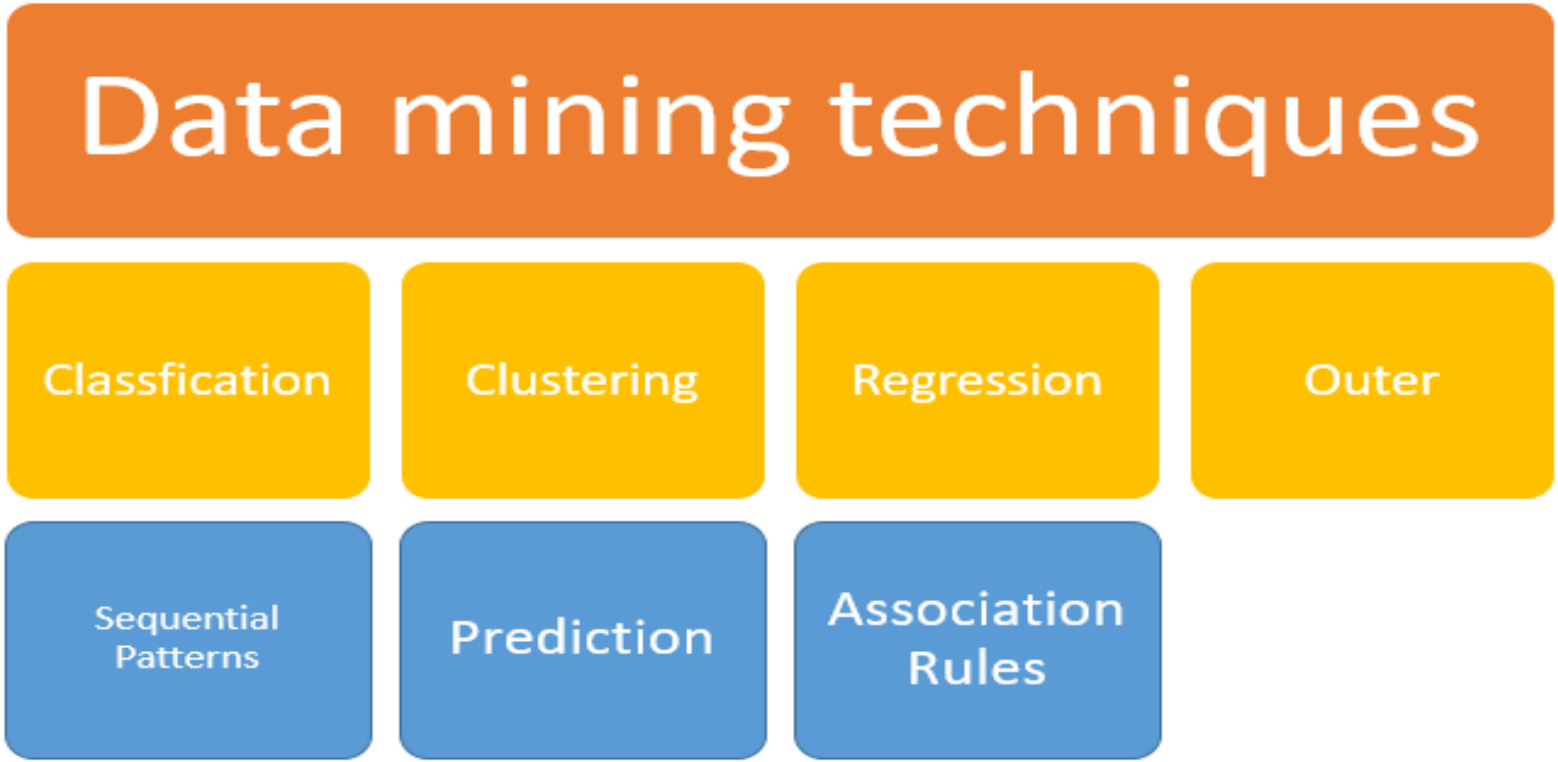
Data transformation:

Data transformation operations would contribute toward the success of the mining process.

- **Smoothing:** It helps to remove noise from the data.
- **Aggregation:** Summary or aggregation operations are applied to the data. I.e., the weekly sales data is aggregated to calculate the monthly and yearly total.
- **Generalization:** In this step, Low-level data is replaced by higher-level concepts with the help of concept hierarchies. For example, the city is replaced by the county.
- **Normalization:** Normalization performed when the attribute data are scaled up o scaled down. Example: Data should fall in the range -2.0 to 2.0 post-normalization.
- **Attribute construction**: these attributes are constructed and included the given set of attributes helpful for data mining.

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

Data Mining Techniques



P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

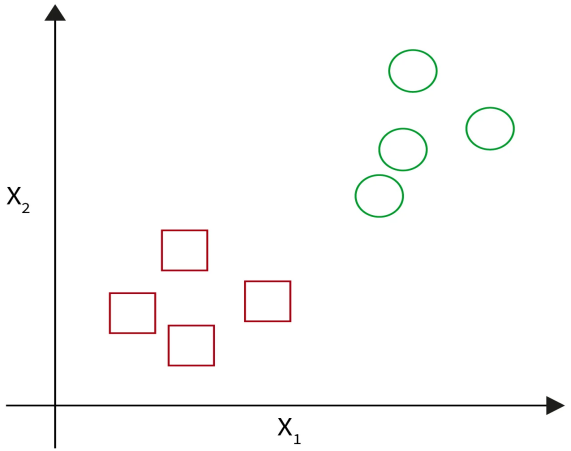
Data Mining Techniques(cont..)

1.Classification: This analysis is used to retrieve important and relevant information about data, and metadata. This data mining method helps to classify data in different classes.

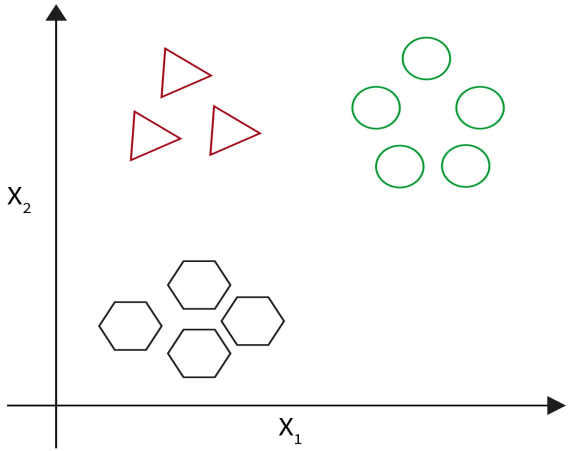
Classification techniques can be divided into categories - binary classification and multi-class classification. Binary classification assigns labels to instances into two classes, such as fraudulent or non-fraudulent. Multi-class classification assigns labels into more than two classes, such as happy, neutral, or sad.

P.O Box 2611 Kigali Info@uok.ac.rw ww.uok.ac.rw

Binary Classification



Multi-class Classification



SCALER Topics

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

Classification is a technique in data mining that involves categorizing or classifying data objects into predefined classes, categories, or groups based on their features or attributes.

It is a supervised learning technique that uses labeled data to build a model that can predict the class of new, unseen data.

It is an important task in data mining because it enables organizations to make informed decisions based on their data.

For example, a retailer may use data classification to group customers into different segments based on their purchase history and demographic data. This information can be used to target specific marketing campaigns for each segment and improve customer satisfaction.

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

Clustering: Clustering analysis is a data mining technique to identify data that are like each other. This process helps to understand the differences and similarities between the data. Clustering is very similar to the classification, but it involves grouping chunks of data together based on their similarities.

The given data is divided into different groups by combining similar objects into a group. This group is nothing but a cluster. A cluster is nothing but a collection of similar data which is grouped together.

For example, consider a dataset of vehicles given in which it contains information about different vehicles like cars, buses, bicycles, etc. As it is unsupervised learning there are no class labels like Cars, Bikes, etc for all the vehicles, all the data is combined and is not in a structured manner.

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

Data Mining Techniques(cont..)

3.Regression: Regression analysis is the data mining method of identifying and analyzing the relationship between variables. It is used to identify the likelihood of a specific variable, given the presence of other variables.

For example, we might use it to project certain costs, depending on other factors such as availability, consumer demand, and competition. Primarily it gives the exact relationship between two or more variables in the given data set.

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

Association Rules:

This data mining technique helps to find the association between two or more Items. It discovers a hidden pattern in the data set.

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently an item set occurs in a transaction.

A typical example is a Market – Based Analysis.

Market – Based Analysis is one of the key techniques used by large relations to show associations between items. It allows retailers to identify relationships between the items that people buy together frequently.

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

Data Mining Techniques(cont..)

5.Outer detection: This type of data mining technique refers to observation of data items in the dataset which do not match an expected pattern or expected behavior. Outer detection is also called Outlier Analysis or Outlier mining.

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

What Are The Different Types of Outliers?

Outlier detection, also known as anomaly detection, is a crucial task in data mining. It refers to the process of identifying data points that are significantly different from the rest of the data in a given dataset. Outliers can cause issues in data analysis, as they can skew results and mislead statistical models. Different types of outliers can exist in a dataset. A few of the most common types of outliers include the following -

- **Global Outliers** - These are data points that are significantly different from the rest of the dataset. They are often caused by measurement errors, incorrect data entry, or rare events.
- **Contextual Outliers** - These are data points that are considered outliers only in specific contexts. For example, a high income in a low-income neighborhood might be considered an outlier, but not in a high-income neighborhood.
- **Collective Outliers** - These are groups of data points that are collectively different from the rest of the dataset. They might indicate a subgroup or a different underlying distribution of data.

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

Sequential Patterns:

This data mining technique helps to discover or identify similar patterns or trends in transaction data for a certain period.

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

Data Mining Techniques(cont..)

7. Prediction:

Prediction has used a combination of the other techniques of data mining like trends, sequential patterns, clustering, classification, etc. It analyzes past events or instances in the right sequence for predicting a future event.

Consider the following scenario: A marketing manager needs to forecast how much a specific consumer will spend during a sale. In this scenario, we are bothered to forecast a numerical value. In this situation, a model or predictor that forecasts a continuous or ordered value function will be built.

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

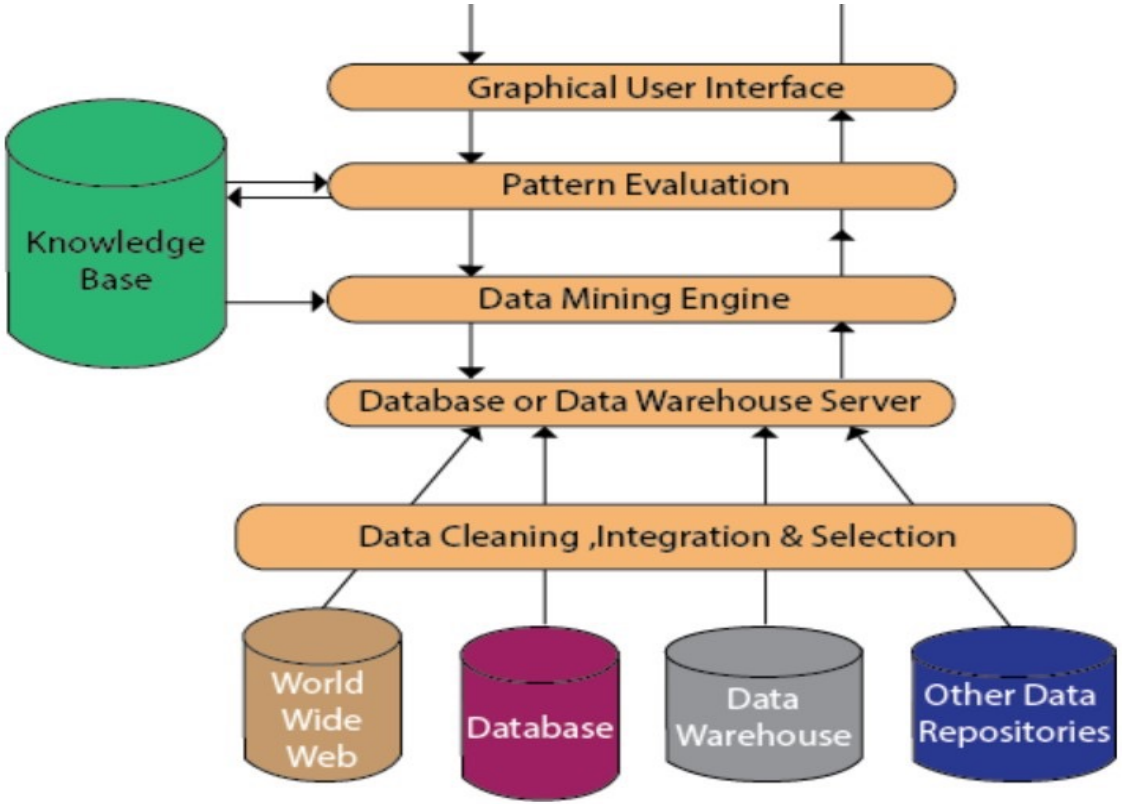
Data Mining Architecture

Data mining architecture is a system designed to extract valuable knowledge from large datasets by structuring the process into several key components: data sources, a data mining engine, pattern evaluation modules, a graphical user interface, and a knowledge base.

This architecture provides a blueprint for how data flows through the system, from initial collection and cleaning to the final interpretation of patterns and insights.

The specific architecture can range from no-coupling to tight-coupling, affecting how the system integrates with data storage systems.

Data Mining Architecture



P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

Data Mining Architecture (cont.….)

- **Data Source:** The actual source of data is the Database, data warehouse, World Wide Web (WWW), text files, and other documents. You need a huge amount of historical data for data mining to be successful.
- **Different processes:** Before passing the data to the database or data warehouse server, the data must be cleaned, integrated, and selected. As the information comes from various sources and in different formats, it can't be used directly for the data mining procedure because the data may not be complete and accurate. So, the first data requires to be cleaned and unified.

P.O Box 2611 Kigali Info@uok.ac.rw ww.uok.ac.rw

Data Mining Architecture(cont…)

Database or Data Warehouse Server: The database or data warehouse server consists of the original data that is ready to be processed. Hence, the server is cause for retrieving the relevant data that is based on data mining as per user request.

Data Mining Engine: The data mining engine is a major component of any data mining system. It contains several modules for operating data mining tasks, including association, characterization, classification, clustering, prediction, time-series analysis, etc.

In other words, we can say data mining is the root of our data mining architecture.

P.O Box 2611 Kigali Info@uok.ac.rw ww.uok.ac.rw

Data Mining Architecture(cont…)

- **Graphical User Interface:** The graphical user interface (GUI) module communicates between the data mining system and the user.
- **Knowledge Base:** The knowledge base is helpful in the entire process of data mining. It might be helpful to guide the search or evaluate the stake of the result patterns. The knowledge base may even contain user views and data from user experiences that might be helpful in the data mining process.

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

Types of Data Mining Architecture

- **No-coupling Data Mining:** In this architecture, data mining system does not use any functionality of a database, that is already very efficient in organizing, storing, accessing and retrieving data. It retrieves data from a particular data sources (eg email, Social media…). is considered a poor architecture for data mining system. But it is used for simple data mining processes.
- **Loose Coupling Data Mining:** In this architecture, data mining system retrieves data from a database and it stores the result in those systems.

P.O Box 2611 Kigali Info@uok.ac.rw ww.uok.ac.rw

Types of Data Mining Architecture

- **Semi-Tight Coupling Data Mining:** In semi-tight coupling, data mining system uses several features of data warehouse systems to perform some data mining tasks that includes sorting, indexing, aggregation. In this, some intermediate result can be stored in a database for better performance.
- **Tight Coupling Data Mining:** In tight coupling, a data warehouse is treated as an information retrieval component. All the features of database or data warehouse are used to perform data mining tasks. This architecture provides system scalability, high performance, and integrated information.

P.O Box 2611 Kigali
Info@uok.ac.rw
ww.uok.ac.rw

Thank you

UNIVERSITY
Of KIGALI