



## **HIGHER DIPLOMA IN SOFTWARE ENGINEERING (21.2)**

### **DATA WAREHOUSING AND DATA MINING**

#### **Course Work 2**

#### **Group members**

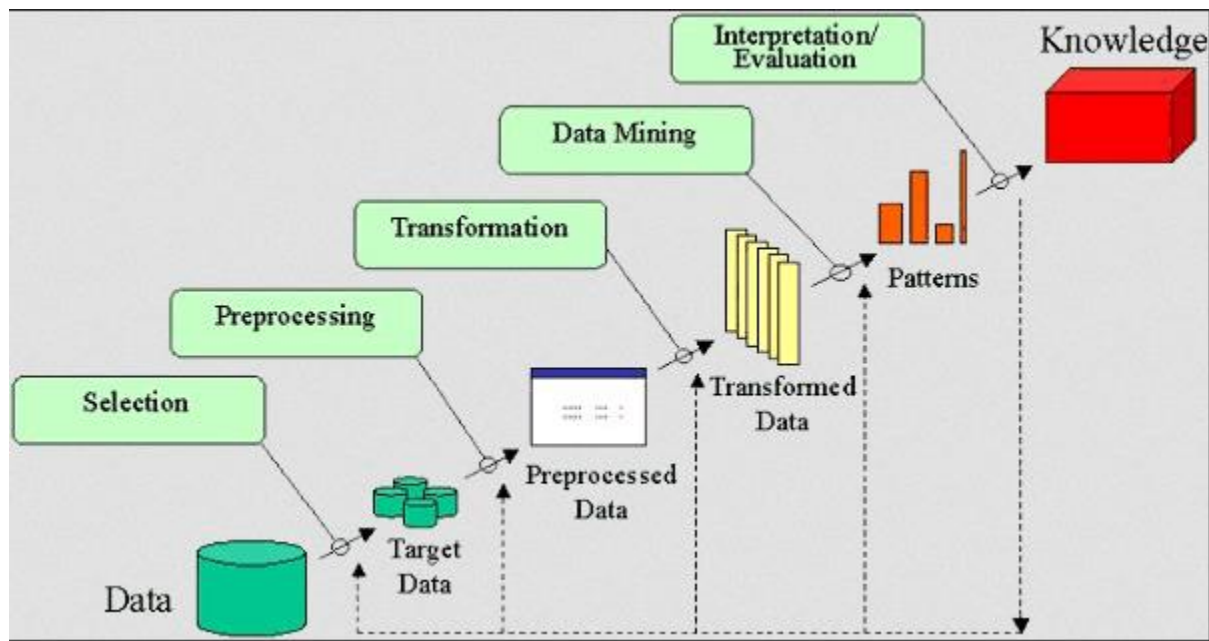
- COHDSE212F-034-A.M.Anfas
- COHDSE212F-051-H.N.M.D.Perera
- COHDSE212F-053-K.S.S.Kumara

## Question 01

Explain in detail the Knowledge Discovery Database (KDD) process. Clearly describe what tasks are carried out in each stage.

\*\*\*\*Answer\*\*\*\*

KDD process- The term Knowledge Discovery in Databases, or KDD for short, KDD is referred to as Knowledge Discovery in Database and is defined as a method of finding, transforming, and refining meaningful data and patterns to be used in a variety of applications or domains. the process starts with determining the KDD objectives and ends with the implementation of the discovered knowledge. this is composed of seven phases, the first four phases are used for data preprocessing that is data is prepared in a format for further use and the rest three are used to work on the data so formed to retrieve the hidden information.



## 1) Goal-Setting and Application Understanding

First understanding the problem we are going to face and propose to real solutions. this is important to know properties, limitations and rules of the data or information understudy and define the goal to be achieved. This is wherever we tend to decide how the transformed data and the patterns found by data mining are used to extract knowledge.

- The application domain
- The relevant prior knowledge
- The goals of the end-user

## 2) Creating a target data set

In this step, multiple data sources are combined. Selecting a data set and focusing on a subset of variables or data samples, on which discovery is to be performed. Once the goals and objectives are determined, it's necessary to select, sort, and categorize the data collected based on their availability, importance, accessibility, and quality into meaningful sets.

## 3) Data cleaning and preprocessing

In this step, the noise and inconsistent data is removed and data reliability is improved. The search and elimination of unwanted data are performed using certain algorithms, which are developed based on some attributes that are specific to each application.

- Strategies for handling missing data fields.
- Removal of noise or outliers.

- Accounting for time sequence information and known changes.
- Collecting necessary information to model or account for noise.

#### 4) Data Transformation

In this step, data is transformed data into the appropriate form required by the mining procedure by performing summary or aggregation operations. Therefore, the data need to be in an aggregated and consolidated form. based on functions, attributes, features, and other data characteristics, the data is consolidated, using dimensionality reduction or transformation methods.

#### 5) Data Mining

In this step, clever techniques are applied in order to extract data patterns. searching for patterns of interest in a particular representational form or a set of such representations as classification rules or regression, clustering, and so forth. this is the root or backbone process of the entire KDD.

#### 6) Pattern Evaluation

In this step, data patterns are evaluated and uses summarization and Visualization to make data understandable by user. Once the algorithms have been applied to the data set, we proceed to evaluate the patterns that were generated and the performance that was obtained to verify that it meets the goals set in the first phases.

#### 7) Knowledge representation

In this step, knowledge is represented. If all the steps are followed correctly and the results of the evaluation are satisfied, the last stage is simply to apply the knowledge found to the context and start to solve its problems. If otherwise, the results are not satisfactory then it is necessary to return to the previous stages to make some adjustments, analyzing from the selection of the data to the evaluation

stage. Finally generate reports, tables, discriminant rules, classification rules, characterization rules etc.

## **Question 02**

Clustering is a data mining function that helps marketers to find the distinct groups in their customer base and they can characterize their customer groups by using purchasing patterns.

Describe how your team does this task practically by using the KDD steps described in part(a) and recommendation to the top management. Clearly mention all assumptions, diagrams, and references (websites or textbooks) used for this task.

\*\*\*\*Answer\*\*\*\*

Data Mining helps the supermarket and retail sector owners to know the choices of the customers. Looking at the purchase history of the customers, the data mining tools show the buying preferences of the customers.

With the help of these results, the supermarkets design the placements of products on shelves and bring out offers on items such as coupons on matching products, and special discounts on some products.

### **Step 01: Goal-Setting and Application Understanding**

- Mainly, we need to identify user purchasing patterns and trends, and consumer purchasing habits.

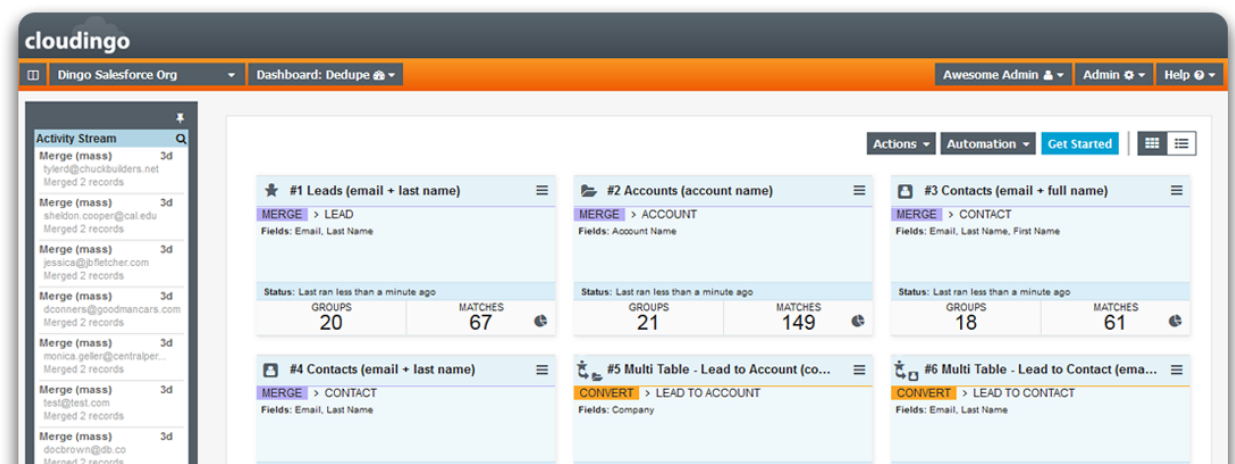
- we are able to collect data by conducting Surveys, transactional Data, Web Tracking, and using Marketing Analytics.
- Therefore, we can get a good understanding of the customers and their demographics, behaviors, attitudes, and actions. Retailers can use this data to tailor their purchasing, marketing, and pricing decisions to better meet their customer's needs and drive sales.

## Step 02: Creating a target data set

- supermarket and retail sector mainly target data set is information about your retail store performance as well as your customers and their demographics, behaviors, attitudes, and actions.

## Step 03: Data cleaning and preprocessing

- As part of this procedure, the data set is searched for missing data, and low quality, redundant, or noisy data is removed from it so as to improve the accuracy of the data, as well as the reliability of the data set overall.
- Next, we use cloudingo tool to is a one-stop-shop for importing, cleaning, and preparing Salesforce data. It is easily scalable and able to run on huge amounts of data.



### Step 03: Data Transformation

- This step is dedicated to converting data into an identical structure and format for later mining data. So, doing an analysis to identify useful features, dimensionality/variable reduction that help supermarket and retail sector owners the supermarkets design the placements of products on shelves and bring out offers on items such as coupons on matching products, and special discounts on some products.

### Step 05: Data Mining




- The importance of data mining is realized in the supermarket and retail industry, and it can be used to get a competitive advantage. data obtained from data mining can be used to provide customers' buying preferences and habits, product sales trends, seasonal variations, suppliers' lead time and delivery performance, customer peak traffic period, and other predictive data to make proactive decisions.

#### Market Basket Example



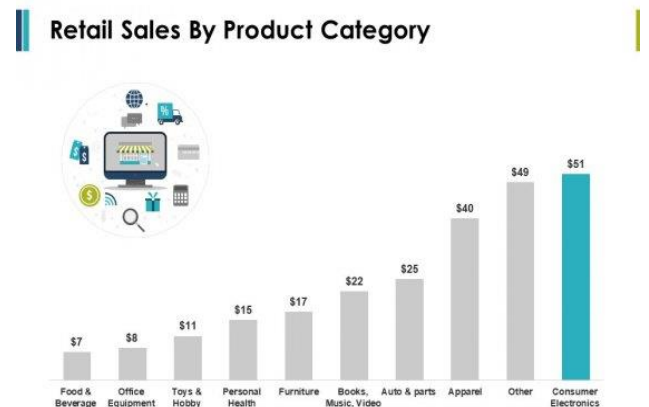
- As the example, Market Basket Analysis is used by many retailers as a marketing method to find out the optimum location to promote a particular product. this is a data mining technique used by retailers to increase sales through a better understanding of customer purchasing patterns. the adoption of market basket analysis was aided by the advent of electronic point-of-sale (POS) systems. Compared to handwritten records kept by owners, the digital records generated by POS systems made it easier for applications to process

and analyze large volumes of purchase data. So then deciding whether the goal of the KDD process is selecting techniques such as classification, regression, or clustering according to our business process.

	 Cluster A	 Cluster B	 Cluster C
Store format Demographic (age) Sales volume	Convenience store Age 18-35 Low sales volume	Grocery store Age 20-55 High sales volume	Speciality store Age 55-75 Moderate sales volume
Cluster insights	Small product range Moderate product prices Young, working consumers require convenience Occasional purchases	Large product range Low product prices Middle-income consumers make weekly-grocery shopping trips Regular purchases	Small-moderate product range High product prices Baby boomer consumers are willing to pay more for premium products Semi-regular purchases

## Step 06: Pattern Evaluation

- Once the trend and patterns are obtained from various data mining ways and iterations, these patterns got to be represented in distinct forms like bar graphs, pie charts, histograms, etc. to review the impact of data collected and transformed throughout previous steps. This helps in evaluating the effectiveness of a selected data model in view of the domain.



This slide is 100% editable. Adapt it to your needs and capture your audience's attention.



## Step 07: Knowledge representation

This is the final step in the process and requires the 'knowledge' extracted from the previous step to be applied to our supermarket and retail sector application or domain in a visualized format such as tables, reports, etc. this step drives the decision-making process. then we plan to make decisions to design the supermarket's product placements, and know the customer patterns or items correlations and buying preferences of the customers.

### References:

- <https://www.geeksforgeeks.org>
- <https://www.javatpoint.com/educational-data-mining>
- <https://www.eminenture.com/blog/a-simple-guide-to-kdd-process-in-data-mining/>
- <https://www.upgrad.com/blog/kdd-process-data-mining/>
- [https://www.tutorialspoint.com/data\\_mining/dm\\_knowledge\\_discovery.htm](https://www.tutorialspoint.com/data_mining/dm_knowledge_discovery.htm)