

**GROUP ASSIGNMENT**

**CT119-3-2-DMPM**

**DATA MINING AND PREDICTIVE MODELLING**

|  |  |  |
| --- | --- | --- |
| Name | : | MD Nowshad UI Alam (TP 057903)  Tarun Aitha (TP058015)  Law Pei Lin (TP051691) |
| Lecturer’s Name | : | Dr. Preethi Subramanian |
| Hand Out Date | : | 24th May 2021 |
| Hand in Date | : | 13th June 2021 |

Table of Contents

[1.0 Problem statement 3](#_Toc74486273)

[2.0 Objective and Scope of this project 3](#_Toc74486274)

[2.1 Objective 3](#_Toc74486275)

[3.0 Scope 3](#_Toc74486276)

[3.0 Methodology Selection 4](#_Toc74486277)

[References 6](#_Toc74486278)

# 1.0 Problem statement

In every country, water plays a very crucial role in development. Almost in every sector water is needed to use in producing their products including agriculture, industry and many more. From the history of Tanzania, it has a lot of water throughout the country, but they face a lot of challenges concerning the reliability and access of the water in common households.

The dataset is all about water pumps, where we will be using data from Taarifa and the Tanzanian Ministry of water. We need to predict the outcome like if the water pump is functional or not, if functional, which one’s needs repair, and which of them are fully damaged that we cannot fix, then we need to arrange a new pump. Thus in the data set, we have three outcomes which are functional, not functional and in some case functional but needs repair. Again, in the following data set if we notice we can see clearly that the data set is not perfectly balanced where more cases are functional followed by not functional and fewer cases related to functional but needs to repair. Through the finding and predictions, the Tanzanian Ministry of water would be able to improve the maintenance operation of water pumps in every location and which would help them serve their people clean and drinkable water (V defour, 2020).

# 2.0 Objective and Scope of this project

## 2.1 Objective

At first, we explore the data set through explorative data analysis, and we ensure that the data set is clean and get it ready for the data analysis and visualization. Some of the data mining techniques will be used to analyze patterns in the data set.

With the help of the decision tree, we will summarize the whole data set and help to speed up analysis.

Neural networks will help us to find the relationship and summarize the number of defective pumps in the data set.

With the help of cluster analysis, we will be able to easily find the pattern and investigate the defective pumps in Tanzania.

## 3.0 Scope

In this project we are using dataset name Pump it up which basically hold records about Taarifa and Tanzanian Ministry of Water and which is collected from an online platform named as Kaggle. In this dataset there are lots of features which are area or location of the pump, quality of water, extraction techniques and many more. The following dataset has almost 59,401 records which belongs to 40 attributes. Again, in the mentioned attributes there are 29 categorical variables including *source\_type*, *quality\_group*, *quantity*, *payment*, *payment\_type* etc., one data type variable and rest of them are numeric that is 10 attributes.

# 3.0 Methodology Selection

The selected methodology for this project is SEMMA. SEMMA is a data mining methodology that defined by SAS institute (SAS Institute Inc, 2017). There are five processes in SEMMA methodology which are Sample, Explore, Modify, Model and Assess.

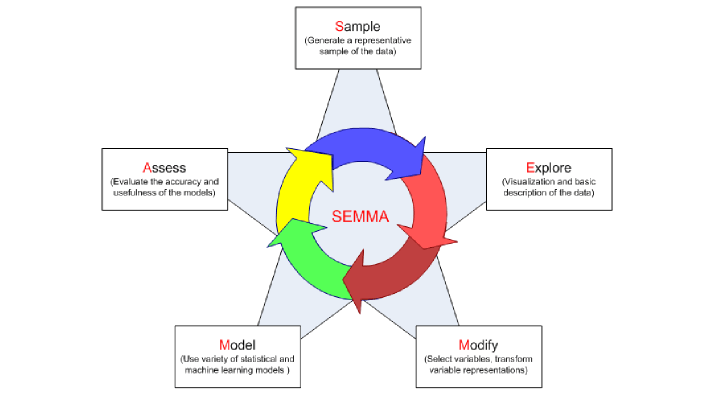


Figure 1: SEMMA Methodology Process (Quantum, 2019)

Figure above shows the stages of SEMMA Methodology. The first stage is start from Sample. This stage consists of sampling the data by extracting a portion of a large dataset which enough to contain the significant information, yet small enough to manipulate quickly (SAS Institute Inc, 2017). The selected dataset for this project was found and downloaded from Kaggle. Since the dataset that we found is store in two csv files, we are going to import the files to SAS Enterprise Miner and using merge node to do data integration. The metadata of this dataset have some changes such as the role of the attribute “*status\_group*” will change to target variable.

The second phase is Explore which is the data exploration process. In this phase we will explore the dataset by using Stat Explore node and Graph Explore node in SAS Enterprise Miner. By exploring the dataset, we can gain an understanding the dataset and identify the anomalies data, missing data, and noisy data. Besides, we can also get an idea on which attribute are more suitable to be used in modelling phase.

The next stage is Modify which is the data modification stage. In this stage, we will do data pre-processing such as transformation of variables or replace the missing values in the dataset. During data pre-processing, we will use the nodes in Modify tab in SAS Enterprise miner such as transform variables nodes and replacement nodes. This is an important stage as data pre-processing will affect the success rate of the project.

The fourth stage is Model which we will apply different modelling techniques in this phase. There are different modelling techniques offer by SAS Enterprise Miner such as Decision Tree Model, Regression, Neural Network, Clustering, Ensemble and so on. We will apply some of the modelling techniques in this phase to find and predict the status of the water pump.

The fifth stage is Assess. In this final stage, we will evaluate the reliability and the usefulness of findings and estimates the performance of the modelling techniques (SAS Institute Inc, 2017). The Assess process can be done in SAS Enterprise Miner by using the nodes in Assess tab such as model comparison node.

# References

Quantum, 2019. *Data Science project management methodologies.* [Online]   
Available at: https://medium.datadriveninvestor.com/data-science-project-management-methodologies-f6913c6b29eb  
[Accessed 10 June 2021].

SAS Institute Inc, 2017. SAS® Enterprise Miner™ 14.3: Reference Help.

V defour, 2020. *Kaggle.* [Online]   
Available at: https://www.kaggle.com/valentindefour/pump-it-up-data-mining-the-water-table  
[Accessed Monday June 2021].

Chen, J. (2019). *Neural Network Definition*. [online] Investopedia. Available at: https://www.investopedia.com/terms/n/neuralnetwork.asp [Accessed 10 Jun. 2021].

Sinharay (2010). *Cluster Analysis - an overview | ScienceDirect Topics*. [online] Sciencedirect.com. Available at: https://www.sciencedirect.com/topics/medicine-and-dentistry/cluster-analysis.

Lucidchat (2019). *What is a Decision Tree Diagram*. [online] Lucidchart. Available at: https://www.lucidchart.com/pages/decision-tree.