WQD7005 AA1 – Nur Aisyah Yusof (22072845)

Reflection and Learning Outcomes

Alternative Assessment 1 of WQD7005 access data mining skill from how we find a reliable dataset according to our objective until the prediction modelling. From the assessment, I have gained a significant insight into how to perform data integration for Talend Data Integration tool, Data Profiling for data quality check in Talend Data Preparation tool and Data preprocessing and modelling in SAS e-Miner.

For the data integration, I need to know how the ETL process works. First, I need to export the data into the tool, then transform the model according to the objective of the study, and lastly, load the data into one dataset. As for Data Profiling, the tool facilitates the data quality check progress as I can get the basic statistics and visualization for each variable column. I also can detect null value and format pattern if necessary. With the tool, I managed to perform basic data cleaning such as remove the null value of MCAR and change a different format into one uniform format to improve the data quality. For the data preprocessing, as there is missing value at random, which might affect the results, I decided to input the missing value with mean value of the variable by using the SAS EM impute node. As for Data Modelling, I chose two models which are Decision Tree and Random Forest Predictive Model. The understanding of how the model works is crucial in this part as to evaluate the performance of the model. For this case study, it is found that Random Forest outperforms Decision Tree but with small significant differences.

From this study, I have found that a quality dataset will affect the entire modelling later. Thus, to obtain a quality dataset is meticulous work especially in corporate world that use data as their business model. The process of obtaining a quality model includes data integration, data profiling, and data preprocessing. As for modelling, it is important to understand each modelling from the root as it will be used for decision-making later. This is to decide which modelling gives the best performance for the prediction. For conclusion, data mining methodology progress involves a rigorous works to obtain insight from the data and achieve the objective of the study.