Data Mining Analysis Report

Name: Dang Anh Duc

Student ID: 1387562

Subject: CP3404-Data Mining

Assignment 2

Menu

[Abstract 3](#_Toc515246604)

[Introduction 3](#_Toc515246605)

[Related Work 4](#_Toc515246606)

[Data Mining Area Justification 5](#_Toc515246607)

[Methods 6](#_Toc515246608)

[Algorithm Selection 6](#_Toc515246609)

[Decision Tree(Classification) 6](#_Toc515246610)

[Naïve Bayes Classifier(Classification) 6](#_Toc515246611)

[OneR(Classification) 6](#_Toc515246612)

[DBSCAN (Clustering) 7](#_Toc515246613)

[K-Means(Clustering) 7](#_Toc515246614)

[Data Mining Processing 7](#_Toc515246615)

[Description of Data 7](#_Toc515246616)

[Pre-processing 7](#_Toc515246617)

[File Import 8](#_Toc515246618)

[Data Mining Strategy 10](#_Toc515246619)

[Classification 10](#_Toc515246620)

[Clustering 11](#_Toc515246621)

[Results 12](#_Toc515246622)

[Classification 12](#_Toc515246623)

[Clustering 13](#_Toc515246624)

[Pattern 13](#_Toc515246625)

[Discussion 14](#_Toc515246626)

[Classification 14](#_Toc515246627)

[Clustering 15](#_Toc515246628)

[Issue 15](#_Toc515246629)

[Conclusion 16](#_Toc515246630)

[Future Work 16](#_Toc515246631)

[References 16](#_Toc515246632)

# Abstract

In this data mining report, different methods of pre-processing and data mining techniques will be using to archive more knowledge about two main areas of data mining (clustering and classification). Moreover, it also aims to discover the bank client in a Portuguese banking institution for the marketing campaigns.

The dataset of bank marketing which included information about bank client in Portuguese banking institution will be use in this report. The data set was cleaned with discretisation and normalisation pre-processing methods. Clustering techniques such as K-Means, K-Medoid and DBScan will be applied to find any interesting patterns. Classification algorithms like decision tree, Naive Bayes and Neural Network were also used in separate tests to find the algorithm and settings that the highest bank term deposit of the client. At the end, the Induction Rule miming and Association Rule mining techniques were applied to get the associations

There are many results based on the methods we used. The classification gave the most useable information while clustering creates confusing result which difficult to interpret. Naïve Bayes was the most accurate method to predict which customer will subscribe a term deposit. Moreover, decision tree is the method that help us to discovered that the contract duration is the most important attributes in determining the subscribe term deposit of each customer.

This report is an excellent way to understand the purpose, advantages and disadvantages of main data mining areas, as well as learning how to apply suitable mining algorithms for each dataset. We also know that with the same dataset but using different pre-processing methods will affect the accuracy of results.

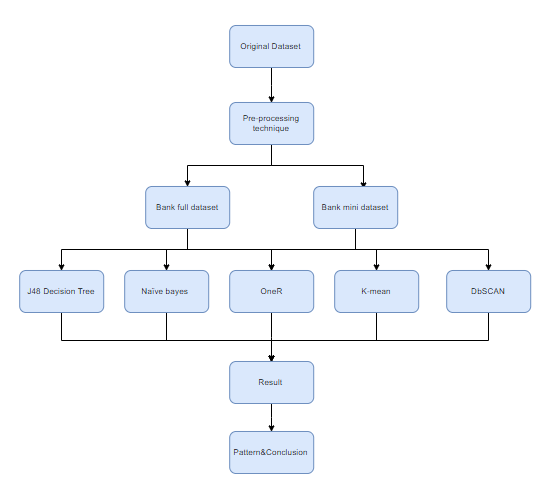
# Introduction

In this research will use many type of pre-processing techniques, classification methods and clustering algorithms to compare the techniques about accuracy an efficiency. Based on the processing performance and results, it is show that the mining algorithms with high processing time cost will retrieve more accurate and fruitful while the faster algorithms prove to be less accurate and reliable. Various attributes and parameters will be discussed and customize to analyse the outcomes.

After the research, it can prove the techniques hold the highest value for return on time spent processing the algorithms. Moreover, it also shows the techniques with higher processing cost but giveback worthy information for making business decisions.

In the first scenario, the research finds which customer will subscribe the long-term deposit based on their age, job, education, housing, loan, etc. Classification like decision tree, naïve bayes and neutral network will be used to produce the models based on the pre-classified Bank dataset with the customer subscribe the long-term deposit or not. After that, the models will be test on the set of Bank dataset that has not been given the variable y to analyse each algorithms accuracy. Various pre-processing techniques also will be used to see how they affect the accuracy of the model.

The second scenario will apply methods of clustering to set the customer with characteristic in the same group called a cluster, these customers are more similar in one way than others in different groups. This is the important task of data mining and a common technique used in statistical data analysis, machine learning and pattern recognition. With the clustering, the Bank can find out the characteristic of customer and use it for business decisions. The K-Mean is one of important algorithms have been used in this research because of the accuracy and time taken is not so long to retrieve critical information.



# Related Work

There is some previous work in this area has focused on mining the dataset and predict what type of customer will subscribe the term deposit.

Sérgio Moro, Paulo Cortez and Paulo Rita (2013): The authors of the dataset have using data mining technique to predict the success of telemarketing calls for selling bank long-term deposit (Sérgio Moro, 2013). This research represents the implementation of Data Mining project based on the CRISP-DM methodology with many DM model such as: logistic regression, decision trees (DTs), neural network (NN) and support vector machine. To estimate the accuracy of results, area of the receiver operating characteristic curve (AUC) and area of the LIFT cumulative curve (ALIFT) have been used as two main metrics. The research finds out that the neutral network algorithm brings the best results which is credible and valuable for telemarketing campaign managers.

Ali Keles and Ayturk Keles (2015): The research figures out the exhibit poor performance and drawback of direct marketing in bank marketing campaigns. Over many years, a large amount of data that banks have been collected and have been apply the Data mining technique to analyse this data set. By using the DM to find out the patterns and trends, bank manager can get more information about customer, such as which customers will likely accept new product offers and which customer will have higher risk of not paying a loan, etc. With all results and information from the dataset, it can help to develop an Intelligent Bank Market Management Systems (IBMMS) for bank manager to deploy efficient marketing campaigns (Ali KELES, 2015).

Kim Yong Jin, Park Jang Won and Jeen Vern Liew (2016): This project is using both SQL Server Data Tools and Data Mining technique to analysing the data and find out which attributes affects customer long-term deposit and which is the number of months for each term. With the SQL Server Data Tools, all the information from classification are visualize by chart and accuracy parameter to bring the specific view to the dataset. Moreover, all cluster in clustering technique also presented clearly and

# Data Mining Area Justification

In this research, classification and clustering are 2 main mining algorithms that will be used.

Classification is chosen to learn more about the supervised learning. One of the aims of this report is to predict if the client will subscribe a term deposit with a high enough degree of accuracy without a large cost to the timeliness of the information. One of the ways of measuring the performance of a classification algorithm is to evaluate its accuracy, which is the total number of predictions that were correct using the following formula:

Accuracy = (TP + TN) / (TP + FP + FN + T N)

TP = correctly identified  
FP = incorrectly identified  
TN = correctly rejected  
FN = incorrectly rejected

The classification algorithms will be used in this analysis report is the decision tree, Naïve Bayes and Neutral Network algorithms with the use of WEKA which is a popular data mining tool.

To get a well-rounded view on the different data mining method, a contrast unsupervised method to classification is clustering has also been chosen. Clustering will find patterns of all attributes within the dataset and then divide into groups of similar data objects called cluster. This method is suitable for the report because it very useful in dividing a dataset into more manageable groups that we can control and understand. Based on the characteristic of the dataset, clustering method will divide the customer base into groups that contain similar value of attribute and easy to find out which type of customer will subscribe a term deposit.

# Methods

## Algorithm Selection

### Decision Tree(Classification)

Decisions trees provide the user a set of rules which will be apply to the dataset to predict the outcome. In the decisions trees structure, each branch node represent as a choice and leaf will represents the classification. In the other way, the branch node is like the question, the branch is the rule and the leaf is the statement. The attribute that is to be predicted in Decision Tree is the dependent variable and it is decided by the value of all the other attributes. The attributes which help to predict the value of dependent variable are independent variable in the dataset. When classify a new item, the Decision Tree will be created based on the existing values of attributes. Because of that reason, if user set the wrong depth of the tree can take the overfitting consequence in that the model will be too rigid to fit any other dataset. The Decision Tree is also good at dealing with noisy, incomplete data and it can handle continuous and discrete.

### Naïve Bayes Classifier(Classification)

Naïve Bayes is a classification method which is based on the Bayes rule of conditional probability and has comparable performance to the Decision Tree. It uses all the attributes contained in the data, analyses them by applying Bayes theorem with strong independence assumptions. The Naïve Bayes will consider each attribute separately when classify a new instance which lead to the lower computation time but loss of the accuracy in the result. This classification algorithms are chosen in many datasets because it is very easy to implement, needs a small amount of training data and can handle large number of attributes but still retrieve good results.

### OneR(Classification)

OneR is a popular classification algorithm which generate one rule for each predictor in the dataset and then choose the smallest total errors. The rule has been defined as the following step:

1. Count how often each value of the target appears.
2. Find the most frequent class.
3. Make the rule assign that class to this value of the predictor.
4. Calculate the total error of the rules of each predictor and choose the predictor with the smallest total error.

The OneR has been chose because it is a simple algorithm but works very well for most of the dataset.

### DBSCAN (Clustering)

DBSCAN is one of most important clustering which is bases on density. By using the density parameters as termination conditions, DBSCAN can help us to find out the arbitrary shapes, handle noise, etc. DBSCAN use one point and retrieves all points which is density-reachable from giving parameters and then group all satisfied points to be a cluster. Changing supplied parameters of DBSCAN can greatly affect the result of the scan.

### K-Means(Clustering)

This is a partitioning method of clustering by constructs partitioning from various partitions. Then, the k-clustering will be given to evaluates the partitions or self-determining when k is not given. Each cluster is displayed by the centre of it and objects are partitioned into k, non-empty subsets based on the nearest mean point of the cluster. But the K-Means have some drawbacks such as the outliers, the inability to find out non-convex shapes. Moreover, it cannot analyse nominal data and need to specify the number of clusters.

# Data Mining Processing

## Description of Data

The “Bank” dataset from the UCI Machine Learning Repository which chosen for this Data Mining report as it has many record for mining purpose. The “Bank” dataset has 17 attributes such as age, jo, marital, education, default, housing, loan, contract, etc., with most of them are categorical and numeric. There are 45211 instances in this dataset and no instance have missing values. Without missing values, this research is easier to apply pre-processing techniques, mining algorithms to retrieve accuracy results.

## Pre-processing

By using WEKA as the data mining tool so we know that WEKA only accept the Attribute Relation File-Format (. arff) and the Comma Separated Values(.csv) file. The original data set is the Comma Separated Values so it suitable to work with WEKA and no need to convert the file format. But the original dataset has error datatype so it need to change to xls file to fix the error before change back to CSV file.

Then, the dataset was imported into WEKA where all attributes was analysed. WEKA will give the basic information about each attribute in the dataset such as the type, number of distinct, maximum, minimum, mean and let we know many missing value in each attribute which could affect the results of mining algorithms. The “Bank\_additonal\_full “dataset have not contained any missing or incomplete data so it let the data mining algorithms work more efficiency and correctly to provide reliability outcomes for our experiment.

For easier to compare and conduct the best results, the dataset with be split into 2 main records and we will apply multiple mining algorithms to each record. From the original dataset record with 21 attributes, we will apply some pre-processing techniques and remove unnecessary attributes to get the first “Bank full dataset”. After research with the dataset to get more information and knowledge, there are 5 attributes have been decided to remove including “employment variation rate”,” consumer price index”,” consumer confidence index”, “euribor 3-month rate “and “number of employees”. The second records are the mini dataset which contain only 10% of the first record with 17 inputs and selected randomly. The second record will help us to compare the result of different size of dataset.

## File Import

The “Bank” dataset will be imported directly to WEKA by using Explorer application from WEKA user interface (Figure 2). In “pre-process” tab, we choose open file and go to the folder that we contain the dataset (Figure 3). The dataset will be visible when we choose “CSV data files(\*.CSV) or “ARFF data files” in the feature” files of type” (Figure 4). The “AttributeSelection” filter (Figure 5) and apply to the dataset. The unnecessary attributes can be remove with button to make the progress running faster.



Figure 2. WEKA initialize screen with the explorer applications.



Figure 3. WEKA Explorer application’s screen in the “Pre-process” tab.

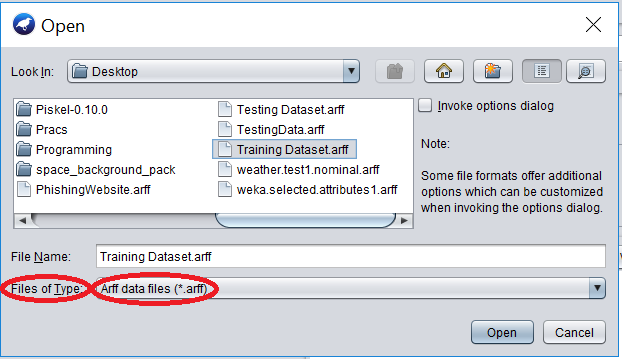


Figure 4. WEKA’s open file option shows the selecting file type method.

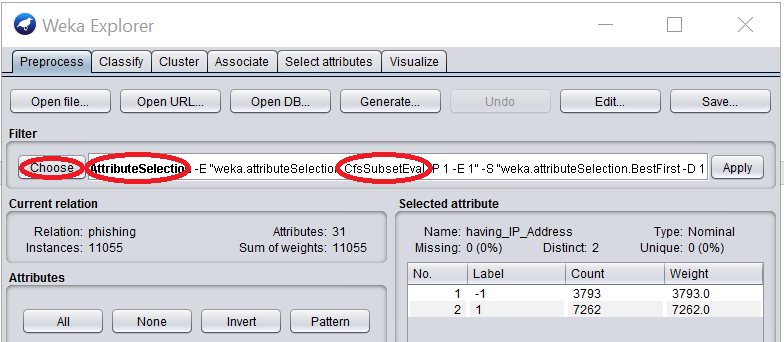


Figure 5. WEKA Pre-process screen showing where to choose a filter, the chosen filter in here is “AttributeSelection” and the filter’s algorithm is “CfsSubsetEval”.

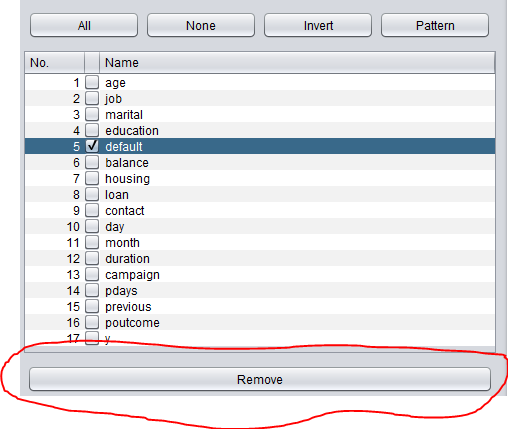
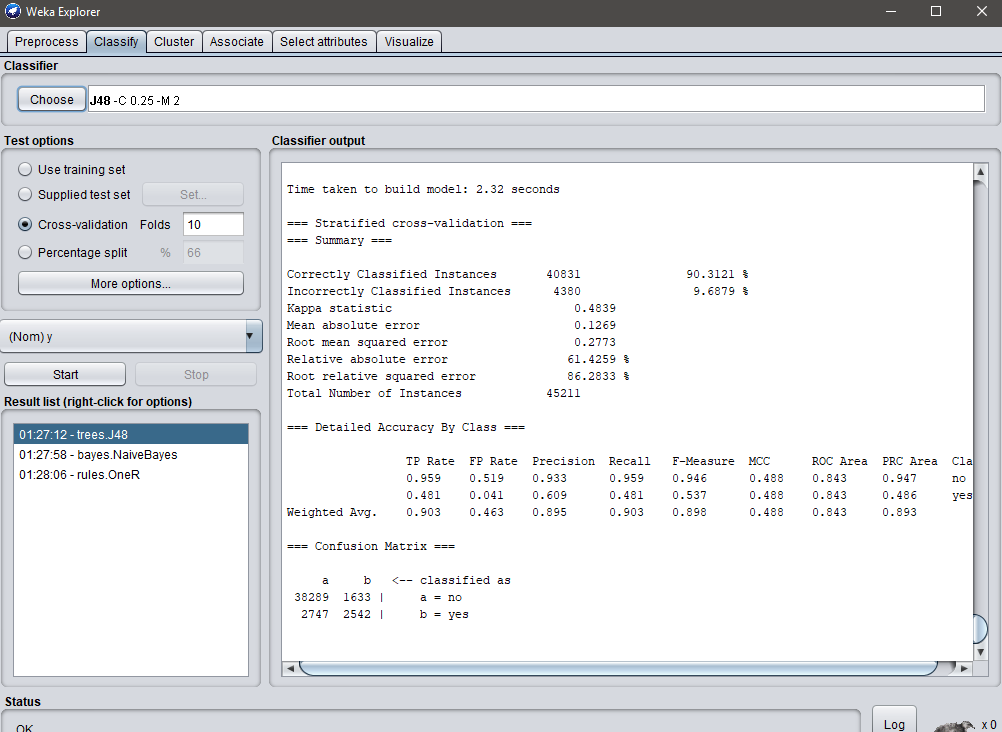


Figure 6. Remove button in WEKA Pre-processing screen.

## Data Mining Strategy

### Classification

The main purpose of the classification testing is to measure the prediction accuracy and the processing time of each classification algorithm with the Bank dataset. Furthermore, the testing will demonstrate the effects of the pre-processing techniques to the results of classification. The classification will be doing through classify tab in Weka. After choosing the algorithms and set the option for the test, we click start and the result will be displayed in the right hand-side.

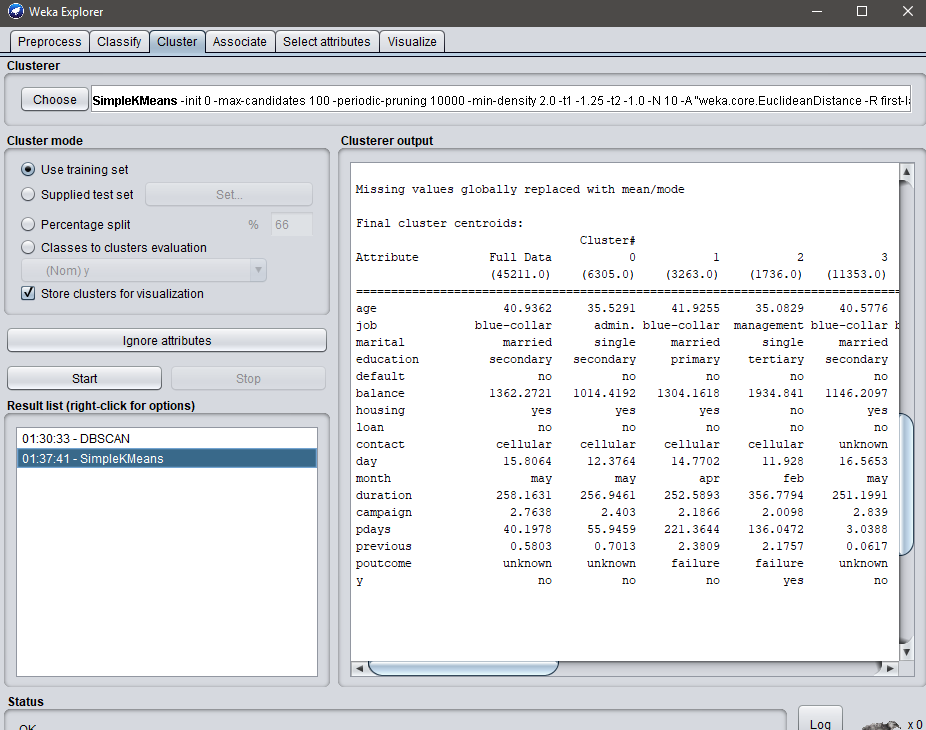


### Clustering

The clustering is used to discover the patterns within the dataset and find out the similarities of customer characteristic to divide into groups of customers for business decision. The test using DBSCAN and K-Mean as two main clustering algorithms with varying parameters to gauge the patterns.

The DBSCAN use the parameters: epsilon, min pts and measure type

The K-Mean use the parameters: k, max runs, measure type, divergence and max optimisation steps.



# Results

## Classification

The different algorithms retrieve different accuracy but most of them are over 88% accuracy in both dataset in test. The Decision Tree bring back the highest accuracy (90.31%) while Naïve bayes and OneR have lower accuracy.

* J48 Decision Tree

|  |  |  |
| --- | --- | --- |
| Predicted | Yes (Actual) | No (Actual0 |
| Yes | 38289 | 1633 |
| No | 2747 | 2542 |

DT Accuracy = (38289+2542)/ (38289+1633+2747+2542)

= 90.3121 %

* Naïve bayes

|  |  |  |
| --- | --- | --- |
| Predicted | Yes (Actual) | No (Actual0 |
| Yes | 36998 | 2924 |
| No | 2498 | 2791 |

NB Accuracy = (36998+2791)/ (36998+2924+2489+2791)

= 88.0073 %

* OneR

|  |  |  |
| --- | --- | --- |
| Predicted | Yes (Actual) | No (Actual0 |
| Yes | 39185 | 737 |
| No | 4434 | 855 |

OneR Accuracy = (39185+855)/ (39185+737+4434+855)

= 88.5625%

With the accuracy of each algorithms, the processing time of decision tree is longest and the Naïve bayes is the most efficiency algorithms because of the fastest processing time and high accuracy. There is no significant change for J48 Decision Tree and Naïve Bayes when using data with continuous attributes and cross validation.

While test with the mini dataset, the decision tree and Naïve bayes retrieve lower accuracy than expected, 88.9847 % and 86.8834 % appropriately. The OneR algorithm still have the same accuracy with 88.5645 %.

## Clustering

With the K-Mean and DBSCAN, we can find out 10 main group of customers with characteristics that will affect the variable y.

Cluster 0: 34, admin., divorced, secondary, no,151, yes, no, cellular,12, may,131,1,361,1, other, no

Cluster 1: 57, admin., divorced, primary, no,207, yes, no, cellular,16, apr,284,1,317,4, failure, no

Cluster 2: 39, management, single, tertiary, no,763, no, no, cellular,22, feb,543,1,189,1, failure, yes

Cluster 3: 32, technician, married, secondary, no,68, yes, no, cellular,9, jul,224,1, -1,0, unknown, no

Cluster 4: 33, blue-collar, married, tertiary, no,870, no, yes, cellular,8, jul,545,3, -1,0, unknown, no

Cluster 5: 38, admin., divorced, secondary, no,1314, no, no, cellular,29, aug,156,2, -1,0, unknown, no

Cluster 6: 33, technician, married, tertiary, no,1699, no, no, cellular,4, aug,600,1, -1,0, unknown, yes

Cluster 7: 42, management, married, tertiary, no,251, no, no, cellular,20, aug,413,2, -1,0, unknown, no

Cluster 8: 37, services, single, secondary, no,105, no, yes, cellular,21, nov,197,2,157,4, failure, no

Cluster 9: 41, technician, married, tertiary, no,354, no, no, cellular,29, jul,95,5, -1,0, unknown, no

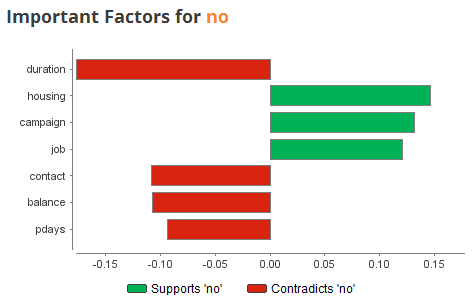
The clustering algorithms is very quick and bring back useful information for further analysis.

# Pattern

By using mining techniques, some pattern of customer has been find. The following characteristic of customer will have higher rate of “yes” in subscribe long-term deposit.

* Age: From 25 to 39
* Job: Management and Technician
* Marital: Married and Single
* Education: Tertiary and Secondary
* Default: No
* Balance: $0- $6667
* Housing: No
* Loan: No
* Contact: Cellular
* Month: February, May, August
* Duration: 132-263 and 263-395
* Campaign: 2.4351
* Pdays: 61.0971
* Previous: 1.1897

The “Duration”- Contact Duration was the most powerful factor when it comes to predict the result



# Discussion

## Classification

From the classification tests performed, it showed that Naïve bayes gives the best overall of accuracy and processing time. Decision Tree generally retrieve the highest prediction accuracy, the different between accuracy using decision tree and actual accuracy is only around 2%.

In the second dataset, it is unknow why the result for decision tree and naïve bayes have decreased in the accuracy and the accuracy of OneR remained unchanged. The test in mini dataset is also make sure that no errors in the test setting.

During the test, it was showed that the attributes will support or contradict the decision to subscribe the term deposit. The duration, housing, campaign, loan, contact, balance and pdays have the strong effect on the result and the duration was the most powerful factor when it comes to predict the result.

Surprisingly, it was found that adjusting the classification algorithm setting only produced slight improvement to the results. In the other ways, the changing in pre-processing techniques will affect strongly to the outcome of each algorithms.

## Clustering

While DBSCAN result were very confusing and no real intelligence could be gathered from the experiments, the K-Means retrieve a lot of critical information to define group of customers have similar characteristic and help to predict the outcome. By changing setting in K-Mean algorithms, many interesting patterns have been found.

# Issue

Through this project, I have meet many issues with the dataset and the mining algorithms. The original dataset was created with many data type error in attributes so it unable to import and processed in Weka. For that reason, the original dataset has been converted to the xls file to fix the error before change back to the csv file. When multiple attributes were used with large original dataset, the classification usually have been stopped. This problem was overcome by using the computer with higher configuration.



# Conclusion

The prediction of decision to subscribe long-term deposit of the customer can be done by using data mining techniques with the dataset from the bank institution. Considering the accuracy and the processing time taken to build model for each algorithm, J48 is the algorithms that provided the highest accuracy in the prediction. With the high accuracy, the J48 Decision Tree taken the longer time than other technique. The Naïve Bayes and One R have more positive result in time taken but lower accuracy. OneR is the only algorithm that retrieve the same accuracy percentage in different dataset

The clustering algorithm is the best way to find out the pattern in characteristic of customer who likely to subscribe the deposit. This is an important algorithm for Manager to design the future for marketing.

Through many tests with Weka algorithms and pre-processing technique, the project has showed the most efficient model and rule for the prediction. The result of this project can be reused for the future work.

# Future Work

This report is only focused on the accuracy and processing time of the (S. Moro, n.d.)classification algorithms in analysis the subscribe the term deposit of the customer. The clustering is using for finding out the potential characteristic of customer and help for business decisions. This report is not enough to draw any serious conclusions but the model will be applied to another dataset to compare the result.

# References

(n.d.). Retrieved from Mining Method: http://www.d.umn.edu/~padhy005/Chapter5.html

Ali KELES, A. K. (2015). *IBMMS DECISION SUPPORT TOOL FOR MANAGEMENT OF BANK TELEMARKETING CAMPAIGNS .*

Research, I. (n.d.). Retrieved from Knowledge Discovery and Data Mining: http://researcher.ibm.com/researcher/view\_group.php?id=144

S. Moro, P. C. (n.d.). *UCI Machine Learning Repository*. Retrieved from Bank Marketing Data Set : https://archive.ics.uci.edu/ml/datasets/bank+marketing

Saedsayad. (n.d.). Retrieved from OneR: http://www.saedsayad.com/oner.htm

Sérgio Moro, P. R. (2013). A data-driven approach to predict the success of bank telemarketing.

# Appendix

## Appendix 1



## Appendix 2



## Appendix 3

