Level 4

DATA MINING TECHNIQUES TO IDENTIFY FRAUDS IN WATER BOTTLE DELIVERY AND PREDICT THE FUTURE DEMAND FOR SALES TRENDS

D.A.S.D Kalansuriya

169314T

Supervised by:

Mr. Saminda Premaratne

(Senior Lecturer)

Department of Information Technology

University of Moratuwa

2018

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Name of Student	Signature of Student
D.A.S.D Kalansuriya	Date: ……………………

Supervised by

Name of Supervisor	Signature of Supervisor
S. C. Premaratne	Date: ……………………

Acknowledgment

I would like to express my deepest Gratitude to my project supervisor Mr. S.C. Premaratne for his patient guidance, enthusiastic encouragement and useful reviews to success this project. Furthermore, my next big thank goes to Dr. Mohamed Firdhous who taught us Research Methodology and Dr Mohamed Firdhous.

Moreover, I would also like to acknowledge with much appreciation the help of the all the lecturers in M.Sc. in Information Technology degree program of Faculty of IT, who gave their fullest support success this program, by sharpen our knowledge and ideas throughout these two years as they were the illumination which lit up our pathways to success.

My special thanks should go to Director at American Premium Water System (Pvt) Ltd Mr. Fayaz Fazal for giving me details, which are helpful to complete the report.

Apart from the people who were directly involved, many more helped to make this project a success.

Finally, I wish to thank to my family and friends for their support and encouragement throughout my study

Abstract

Data mining is a subset of databases management and it mainly applicable to large and complex databases to eliminate the randomness and discover the hidden pattern. Fraud detection in data mining is the process of identifying fraudulent acts by analyzing the dataset. Research is based on identifying fraudulent acts of water bottle delivery process. The research study focusses on to manage the invoicing process with the water delivery process. Due inefficacies in the water delivering process bottle lost cost in the last six months is Rs 213,070.00 approx. Through detecting fraudulent acts, the institutes can save resources and cost [3], for this study a sample data set has been used to identify how the fraudulent activities are occurring. Sample dataset has been selected from where data entry person had found physical evidence that the bottle had been sold for outsiders.

Data mining tools which used to detect frauds are Naïve Bayes, Decision Trees, and neural networks. By developing predictive models can be generated to estimate things such as the probability of fraudulent behavior. ROC curves have deployed for model assessment to provide a more intuitive analysis of the models and confusion matrix is has used to describe the performance of a classification model on the test data for which the true values are known.

Contents

List of Figures

List of Tables

List of Equation

List of Abbreviation

ANN	Neural Networks
ROC	Receiver operating characteristic

Chapter 1

1 Introduction

This Research is based on Business Data mining. Data mining is a subset of databases management, and it mainly applicable to large and complex databases to eliminate the randomness and discover the hidden pattern.

By using modern technologies of computers, networks, and sensors have made data collection, therefore data collection and storing has become very easy. However, the captured data need to be converted into information and knowledge from recorded data to become useful. Traditionally, analysts have performed the task of extracting useful information from the recorded data, But the increasing volume of data in modern business and science calls for computer-based approaches. Data mining is the entire process of applying the computer-based methodology, including new techniques for knowledge discovery, from data [1]. The research is based on applying of data mining techniques to discover the possible frauds in the water bottle delivering process and water consumption pattern changes. Once the water consumption is dropped below the satisfactory levels, it impacts to the revenue/sale of the company.

Bottle Water Industry is one of the growing Industry in Sri Lanka with a growth rate of 10% per annum[2] .In Sri Lanka, Bottle water Industry was started in 1980[3]. Per the analysis was done by the Ministry of Health, Nutrition and Indigenous Medicine, currently, there are 118 Market competitors in bottle water Industry[4]. Therefore, there is a huge competition only for bottle water excluding the completion of packaged food. Hence that losing market share due to operational inefficiencies directly affects the future of the company.

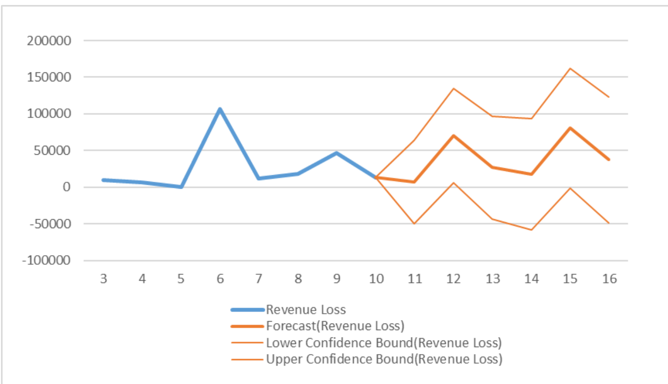


Figure 2.1.1-1: Revenue Loss Forecast

The research study focusses on to manage the invoicing process with the water delivery process. Due inefficacies in the water delivering process bottle lost cost in last six months is Rs. 213,070.00 approx. Figure 1.1-1. The research was selected because manual investigation to detect irregularities in inefficacies in the water delivering process is hard and slow; hence by using data mining techniques can increase the speed and accuracy of the investigation. Whereas, it is very important to detect the methods of suspicious bottle deliveries and irregularities to identify the possible frauds.

In the selected entity, the bottled water delivering process is one of a main process which enables to serve the customer directly. But since it is a manual process,it has ineffiencienies, the objective of this research is to reduce the inefficiencies by identifying impacts to reduce the water consumption. Since in the worst-case scenario when the planned customer is not catering on time, the customer gets infuriated with the company, where as it leads to close their accounts which leads to an increase in the expenses of the company in numerous ways. Therefore, customer services are very important to service organization as a selected entity.

By doing this research can identify how operational frauds are happening and inefficiencies water bottle Delivering Process and by what means its effects to the final invoicing value. Moreover future demand can be identified.

Negative approaches of bottle delivering process, where it affects to the inefficiencies

1. If bottle could not be delivered to the customer, then the bottle is recording under house closed, stock available and missed deliveries. Nevertheless most of the time it is not recording under missed deliveries because it affects to the incentives hence, it records as house closed or stock available.
2. Bottle are delivering without keeping proper records via manual tickets. Therefore, when customer details could not identify, it considers as a dispute ticket, so that, those delivery tickets could not be invoiced and sometimes those deliveries affect for the bottle lost as well. Once delivered bottles did not invoice, it reduces the projected revenue of the company and then it leads to increases the operation cost of the company
3. Bottles are delivered on a cash basis by route men to outsiders, thus that those bottles are can be invoiced.

4. Deliver less quantity than customer needed to balance the time of delivery and to balance unloaded and loaded bottle details in the store, consequently that customer get less than required amount

1.1 Aim

Apply a rapid, intelligent model (data mining models) to detect possible fraudulent behavior of the delivery process

1.2 Objectives

- Identify patterns of water bottles consumption of Customer, and to identify a future trend
- Review the related works and the mining methods to detect frauds in the selected field.
- Apply a model to measure the changes in water consumption based on one appropriate technique.
- Evaluate the applied model accuracy.
- Selecting the best attributes that improve the accuracy

By predicting consumption pattern can get alerts, to take necessary actions to keep the average consumption pattern of the company constant. Whereas, the average consumption pattern of the company must equal to the contracted value of the customer at the time of account acquisition. When average consumption is less than the contracted value, it reduces the projected revenue from the customer.

1.3 Assumption

The data which has taken for the model has been updated on a regular basis.

1.4 Thesis structure

- Abstract- Briefly summary of
    - Research problem/Methodology/ Results/ Conclusion.
- Table of contents. List the key headings and subheadings with their page numbers.
- List of figures. Include the figure numbers, figure titles, and page numbers.
- List of tables. Include the table numbers, table titles, and page numbers.
- Introduction
- Literature Review
- Technology Adopted

- Analysis and the Design
- Implementation of the Model
- Discussion
- Reference
- Appendix

Chapter 2

2 Literature Review

2.1 Introduction

This section presents the background and theoretical concepts of the data mining techniques applied in this research study.

Data Mining is a discipline, combination of statistics, machine learning, data management and databases, pattern recognition, artificial intelligence to generate information on the data. This method use converts data into knowledge and actionable information [5]. Data mining techniques are used to the analyzed large dataset to identify new unknown patterns; these patterns support to identify bottle delivery patterns and consumption patterns of the customer in a scientific manner. When suspicious customer patterns are identified using data mining techniques, then it is easy to narrow down the circle of investigation to get fraudulent issues very fast.

Areas of Data Mining Applying[6]

- Financial Data Analysis
- Retail Industry
- Telecommunication Industry
- Biological Data Analysis
- Other Scientific Applications
- Intrusion Detection

2.1.1 Data Mining Methodology

The data mining techniques (knowledge discovery) has perspective steps.

1. Data cleaning (to remove noise and inconsistent data).
2. Data integration (where multiple data sources may be combined).
3. Data selection (where data relevant to the analysis task are retrieved from the database).
4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance). Data mining (an essential process where intelligent methods are applied to extract data patterns).

5. Pattern evaluation (to identify the truly interesting patterns representing knowledge Based on some interesting measures).
6. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user via graphs) [5]

2.1.2 Standard Data Mining Process

The cross-industry standard process for data mining, (C**RISP-DM**) is a data mining process model that describes approaches that data mining experts use to tackle problems[7]. This process has six set steps[8]

1. Business Understanding

   Focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition and a preliminary plan.

2. Data Understanding

   Starts with an initial data collection and proceeds with activities to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

3. Data Preparation

   The data preparation phase covers all activities to construct the final dataset from the initial raw data.

   Example: Linear Regression function could be applied only to numerical data

   Neural Networks/Naïve bays/Decision Trees could be applied to nominal data

4. Modeling

   Modeling techniques are selected and applied.

   Ex: When it wants to identify a trend in numerical, a linear regression is used

5. Evaluation

   Once one or more models have been built that appear to have a high quality based on a percentage of truly positive values, these need to be tested to ensure they generalize against unseen data and that all key business issues have been sufficiently considered

6. Deployment

Generally, this will mean deploying a code representation of the model into an operating system to score or categorize new unseen data as it arises and to create a mechanism for the use of that new information in the solution of the original business problem. Importantly, the code representation must also include all the data prep steps leading up to modeling so that the model will treat new raw data in the same manner as during model development.

2.1.3 Data Mining Methods

The mining method can be segregated into main categories

A **Descriptive model** presents the data in a concise form which is essentially a summary of the data points, finds patterns in the data and understands the relationships between attributes represented by the data. This includes tasks such as Clustering, Association Rules, Summarizations, and Sequence Discovery.

The **predictive model** works by making a prediction about values of data for future though existing datasets. The Predictive data mining model includes classification, regression, Choice modeling, Rule Induction, Network/Link Analysis, Clustering/Ensembles, Neural networks, Memory-based/Case-based reasoning, Decision trees and Uplift modeling [9]

Patten recognition can do to identify the relationship between input and output values. Data mining methods include Neural networks, fuzzy logic

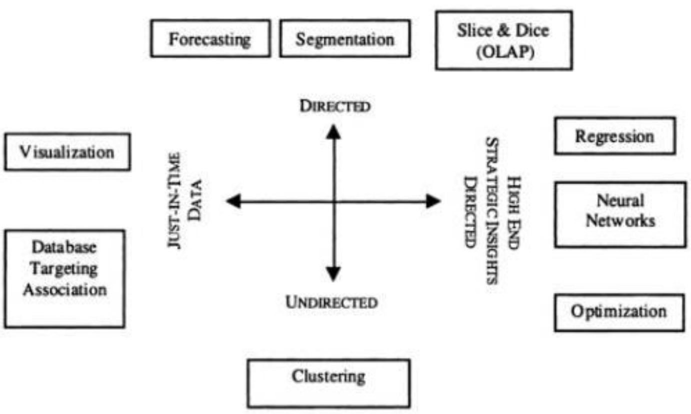


Figure: 2.1-1:Data Mining Techniques

Source: Data Mining and Business Intelligence, “By Stephan Kudyba, Richard Hoptroff"

2.1.4 Data Mining Techniques

- Naïve Bayes-This can be used to classify and predict. This calculates the probabilities for each possible state of the input attribute given each state of predictable attribute[10].
- Layered, feed-forward neural networks are suited to the analysis of non-linear and multivariate data[11]
- Decision Trees-This classify each case to one of a few (discrete) board categories of selected attribute(variables) and explain the classification with few selected input variables[10]
- Neural Networks - A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. Neural networks can adapt to changing input so the network generates the best possible result without needing to redesign the output criteria.[12]
- Association Analysis is done to identify the connection between two actions or two objects [13]
- Slice and dice enable to get summary data easily [13]
- Segmentation algorithm/Clustering algorithm enables to group the data as per to similar attribute[13]
- Regression and Neural algorithm enables to fit the data into a curve[13]
- Optimization algorithm enables to identify the best option out of others[13]
- Visualization enables the reader to identify the data more easily
- Support Vector Machines is the technique of machine learning, SVMs (SVMs, also support vector networks) is supervised learning model with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier [5]
- K-Nearest-Neighbour this supports non-linear problem[5]

- Prediction discovers the relationship between independent variables and dependent variables. For instance, the prediction analysis technique can be used in the sale to predict profit for the future[14]

2.2 Background to Frauds

Frauds is a wrongful or criminal action done intentionally to get financial or personal gain

In the case of National Health Service, the allocated budget to National Health Service (NHS) in 2007/08 is £104 billion, however, due to frauds, per annum loss is reaching to £76.35 million to £118 million whereas it is a huge cost to the company [13]therefore as per the author it is important to have fraud-free society and if you could find it will benefit to company and the society

However, Researcher has measured the cost of fraud: to get a competitive advantage by identifying the frauds. This research has been done to analyzed 132 fraud risk measurement exercises from nine countries in a range of different sectors. During the study, it has identified 32 different types of expenditure due to frauds totalling almost £800 billion, in 44 organizations from nine countries. The paper shows fraud and error can be measured, and also if fraud could be regularly measured ,the company is could reap financial benefits to the organization.[15]

Frauds can be detected as per fraud frequency rate (FFR) or the percentage of expenditure lost due to fraud or by using both techniques or separating and frauds and errors and calculating effect of fraud and error. The research has been done by using statistical confidence. Selected Methodology is able to use in the sample when there is a large population[1][15].However, sampling methods have disadvantages such as chances of biases and difficulty in identifying truly problem as well [16].

Moreover, researchers have identified how human characteristics effects on frauds. The selected variables are the perpetrator’s position (i.e. employee, manager, executive/owner), gender, education level and the presence of assistants (i.e. collusion). The analysis has been done on multivariate regression analysis, by doing multivariate regression analysis the researcher has been able to estimates a single regression model

[1] “Statistics How to”, http://www.statisticshowto.com/confidence-level/

with multiple outcome variables and one or more predictor variables. [2]However, only the perpetrator's position and collusion are statistically significant when controlling for the potential correlation among explanatory factors. As per author, This study is useful to regulatory agencies and anti-fraud professionals to reduce frauds.[17]

2.2.1 Related Works

The researcher has done different types of Data mining method to detect frauds. The selected methods are a statistics-based algorithm, decision tree based algorithm and rule-based algorithm, Bayesian classification, Naïve Bayesian visualization is selected to analyze and interpret the classifier predictions. [18]

2.2.1.1 <u>Electricity Consumption Identification</u>

Frauds could be identified with the behavioral patterns of the data set. The study on to identify the "Anomalies in School Electricity Consumption Data" has been done based on outlier analysis or Anomaly detection. In this study, irregular behavior has been identified by detecting patterns in a given data set that do not conform to an established normal behavior [19].

As per the author there are Three types of anomaly detection techniques has used:

- Supervised techniques build models for both anomalous data and normal data. An unseen data instance can be classified as normal or anomaly by comparing which model it belongs to.
- Semi-supervised techniques only build a model for normal data in the training data set. An unseen data instance can be classified as normal if it can fit the model sufficiently well. Otherwise, the data instance will be classified as anomalies.
- Unsupervised techniques do not need any training data. These approaches assume that anomalies are much rarer than normal data in the data set

As per the author Outliers can be easily identified ones the data is visualized as well, the author has proposed a new outlier detection algorithm is which combines the image processing method with the data processing method. In this algorithm, a measure in image processing, the degree of sharpness, is adopted to detect the

[2]"MULTIVARIATE REGRESSION ANALYSIS | SAS DATA ANALYSIS EXAMPLES." https://stats.idre.ucla.edu/sas/dae/multivariate-regression-analysis/

outliers for the first time. The proposed algorithm can be easily applied to the applications of data pre-processing, equipment fault diagnosis, credit fraud detection, traffic incident detection etc.[13]. But this method cannot be used in pattern recognition

2.2.1.2 Credit Card Fraud Detection

In the research of "Credit Card Fraud Detection with a Neural Network," the author has used a neural network to identify the frauds. As per the author, a neural network-based fraud detection system has been shown to provide substantial improvements in both the accuracy and timeliness of fraud detection. The frauds loss is able to reduce from 40% to 20%.[20].Whereas the neural network captures knowledge through learning, and it can explore more possible data relationship than other algorithms[10]

As per the researcher, Neural Network methods can be used for data classification, clustering, feature mining, prediction and pattern recognition. It uses the idea of non-liner mapping, the method of parallel processing. The structure of the neural networks itself to express the associated knowledge of input and output data. This was not used at the beginning due to the fact it had defects in large complex structures, poor interpretability and long training time. however, at present, it had been widely used in the medical, finance and marketing research because it had the predictive power than statistical techniques using real data sets and power-full ability in pattern recognition. Additionally, neural networks have ability to afford noisy data, low error rate and continuously advancing and optimization of various network training algorithms[21]

The use of neural networks(data-driven) approach is ideal for real world data mining problems where data are plentiful but the meaningful patterns or underlying data structure are yet to be discovered and impossible to be pre-specified.[21]

2.2.1.3 Non-Technical Loss (NTL) identification in Electricity Consumption

The Research is based on identifying of electricity consumption that is not billed. The researcher has said it identified by detecting and inspecting the customers that have null consumption during a certain period. The author has divided the into two modules first module is based on text mining and artificial neural network. The second module, developed from a data mining process, contains a Classification & Regression tree and a Self-Organizing Map neural network. As per the suggestions of the researcher the results of analysis were limited due to the lack information in each location gathered, therefore to do a better analysis it need more information[22]. Non-technical loss could be also identified more efficiently by using Optimum-Path Forest method.

2.2.1.4 Accounting-Fraud Detection

The financial data could be detected by Logistic regression, Neural Networks, Induction trees, Bayesian trees, Statistical Clustering and Association rule[23] [24]

2.2.2 Fraud Detecting Methods

According to the article, one way to approach the issue of fraud detection is to consider it a predictive modeling problem. If historical data are available where fraud or opportunities for preventing loss have been identified and verified, then the typical useful predictive modeling workflow can be directed at increasing the chances to capture those opportunities.[14]

Use of Machine Learning Techniques for fraud detection, there are seven methods have introduced by Neural Networks, Multilayer perceptron, Radial basis functions, Support vector machines, Naïve Bayes, k-nearest neighbors, Geospatial Predictive Modelling [22]

2.3 Summary

This research is based on the use of data mining techniques to identify possible cases of fraud. The research is done using binning techniques, pattern recognition techniques, clustering techniques and statistical approaches. These methods were selected based on the suggestions of the literature reviews

Chapter 3

3 Technology Adopted

3.1 Introduction

As discussed in the previous chapter the standard data mining process is **Cross-industry standard process for data mining, (CRISP-DM)**[10] had used to the research**.** This process has six steps to reach to result. As this research is based on identifying possible fraud, the prediction methods were selected methodology is to do research. This chapter highlights the effectiveness of selected technology that distinguishes it from the technologies applied in the existing literature.

3.2 Selected methods or techniques

- Binning techniques

Binning or discretization is the process of transforming numerical variables into categorical counterparts. Moreover, binning may improve the accuracy of the predictive models by reducing the noise or non-linearity. Finally, binning allows easy identification of outliers, invalid and missing values of numerical variables. There are two types of binning, unsupervised and supervised.

In this research have used the supervised method

Supervised binning methods transform numerical variables into categorical counterparts and refer to the target (class) information when selecting discretization cut points. Entropy-based binning is an example of a supervised binning method

– Entropy-based Binning

Entropy-based method uses a split approach. The entropy (or the information content) is calculated based on the class label. Intuitively, it finds the best split so that the bins are as pure as possible that is many of the values in a bin corresponding to have the same class label. Formally, it is characterized by finding the split with the maximal information gain.

Calculate "Entropy" for the target

$$E(s) = \sum_{i=1}^{c} -pilog2pi$$

Equation 3.2-1:Entropy

Calculate "Entropy" for the target given a bin.

$$E(AS) = \sum_{v \epsilon A} \frac{|Sv|}{|S|} E(s)$$

Equation 3.2-2:"Entropy" for the target given a bin

Calculate "Information Gain" given a bin.

$$Information\ Gain = E(S) - E(S, A)$$

Equation 3.2-3:Information Gain

- Naïve Bayes

The Naive Bayesian classifier is based on Bayes' theorem with the independence assumptions between predictors.

The calculation process Naïve Bayes Algorithm

Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors.[25]

$$P(c|x) = P(x|c)P(c)/P(x)$$

Equation 3.2-4:Naive Bayes

– P(c|x) is the posterior probability of class (target) given predictor (attribute).

– P(c) is the prior probability of class.

– P(x|c) is the likelihood which is the probability of predictor given class.

– P(x) is the prior probability of predictor.

- Decision Tree

A decision tree is a structure that includes a root node, branches, and leaf nodes[6] .Decision tree algorithm is used to predict the model and the attribute selection attribute is used to evaluate best suitable attribute

– Basic algorithm (a greedy algorithm)[26]

1. Tree is constructed in a top-down recursive divide-and-conquer manner
2. At the start, all the training examples are at the root

3. Attributes are categorical (if continuous-valued, they are discretized in advance)
4. Examples are partitioned recursively based on selected attributes
5. Test attributes are selected based on a heuristic or statistical measure (e.g., information gain)
   – Conditions for stopping partitioning

6. All samples for a given node belong to the same class
7. There are no remaining attributes for further partitioning – the majority voting is employed for classifying the leaf
   – Method of finding which attribute have the highest priority

$$Information\ Gain\ =\ Entropy(BS)\ -\ Entrop(AS)$$

Equation 3.2-5: Decision Tree Algorithm

BS- Before Selecting

AS- After Selecting

– Attribute Selection Measure: Information Gain
  - Information needed (after using A to split D into (partitions) to classify D:

$$InfoA(D) = \sum_{j=1}^{v} \frac{|Dj|}{|D|} xI(Dj)$$

Equation 3.2-6:Information needed

  - Information gained by branching on attribute A

$$Gain(A) = \ Info(D) - \ InfoA(D)$$

Equation 3.2-7:Information gained

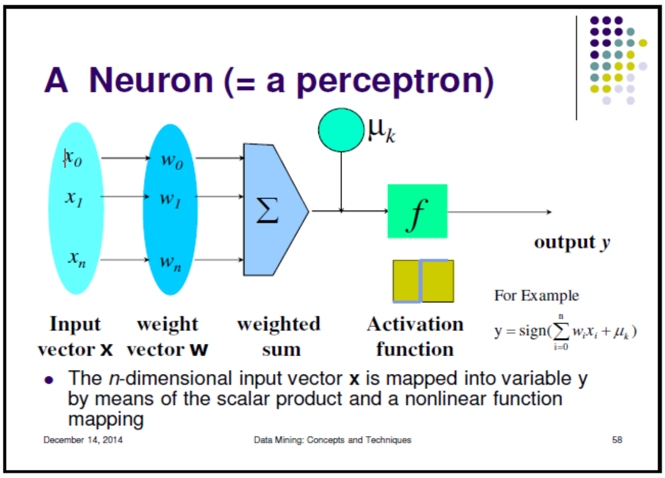
- Neural Network (ANN)

A neural network is a series of algorithms that used to recognize underlying relationships in a set of data through a process the way the human brain operates. Neural networks can adapt to changing input, so the network generates the best possible result without needing to redesign the output criteria.[12]

Data Processing in ANN developed under three building blocks[27] .

- Network Topology
- Adjustments of Weights or Learning
- Activation Functions

Selected Method is under Network Topology Single layer feedforward network is used for research design. In Single layer feedforward having only one weighted layer. Whereas, input layer is fully connected to the output layer.



Equation 3.2-8:Backpropagation: A neural network learning algorithm[26]

- Statistics Approaches and Data Visualization

The Statistical Approaches were used for data summarization of the data as per the theories and hypothesis testing

3.3 Tools using for a data mining

- R studio

R studio is used for data visualization and for data summarization. R Studio is a free, open source IDE (integrated development environment) for R. the interface of the is organized so that the user can clearly view graphs, data tables, R code, and output all at the same time. It also offers an Import-Wizard-like feature that allows users to import CSV, Excel, SAS (*.sas7bdat), SPSS (*.sav), and Stata (*.dta) files into R without having to write the code to do so.[28] therefore it is easy when analyzing the data

- Weka

Weka tool is used for data analyzing.by using Weka tool can identify what is most suitable prediction type for the research and it enables to identify the most reliable variable for the prediction

Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization

Weka is open source software issued under the GNU General Public License.[29][30]

- Excel

Excel was used to data pre-processing, whereas Data Pre-processing is a technique that is used to convert the raw data into a clean data set. By using Excel was able to remove null values and to remove the missing values and to rescale the data.

Microsoft Excel is a spreadsheet developed by Microsoft for Windows, Mac OS, Android, and iOS. It features calculation, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications[31]

3.4 Summary

This chapter is about technology proposed to analyze Frauds in the Water bottle delivery process. For this Excel, R Studio and WEKA can be used to data prepossessing, to data modeling and to analyze the data

Chapter 4

4 Analysis and the Design

4.1 Introduction

This chapter includes attributes involve for analysis and methodologies using to identify the fraudulent acts. The research is done by taking sample dataset from the population. The population is entire customer bases, and for this research, a selected customer bases has been taken. The selection was based on the physical evidence where the data operators had found the bottles has been sold for an outsider with the manual tickets

4.2 Attributes of the analysis

The proposed model is about to detect fraudulent deliveries of bottles

The following data is collected to develop the model

1. Customer Description
2. Consumption Levels
3. Customer Complaints
4. Stock Available
5. Missed Delivery
6. Housed Closed
7. Instances of manual tickets
8. Instances of manual Invoices

The research study is based on, the detection approach illustrated in Figure 3.2, by using historical data transforms the data into the required format for the classifier. Customers are represented by their consumption profiles over a period of 12 months. These profiles are characterized by means of patterns, which significantly represent their general behavior, and it is possible to evaluate the similarity measure between each customer and their consumption patterns. This creates a global similarity measure between normal and possible fraud bottle deliveries.

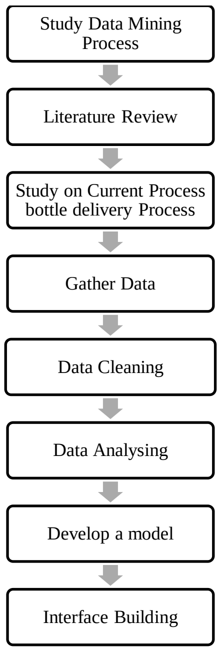


Figure 2.2.2-1:Research Methodology

4.3 Sample Selection Process

In order to evaluate the consumption pattern one-year data has been taken (Nov 2017-October 2018). In here the data be segmented as per the delivery location. Data set had selected, where there is the highest percentage on irregularities in consumption patterns as per the customer count

Selected Data set size is: 27501

Ranking of the location as per the consumption problem

Table 2.2.2-1:Ranking table of the Consumption Problem

Location Code	**Percenatge_of_Worst Cases**
Colombo 10 Store	50%
Colombo 05 Store	47%
Kalutara	44%
Factory	44%
Kandy	43%
Negombo	41%

Anuradhapura	40%
Kurunegala	38%
Hambantota	28%
Galle	27%
Distributors	17%
Tangalle	0%

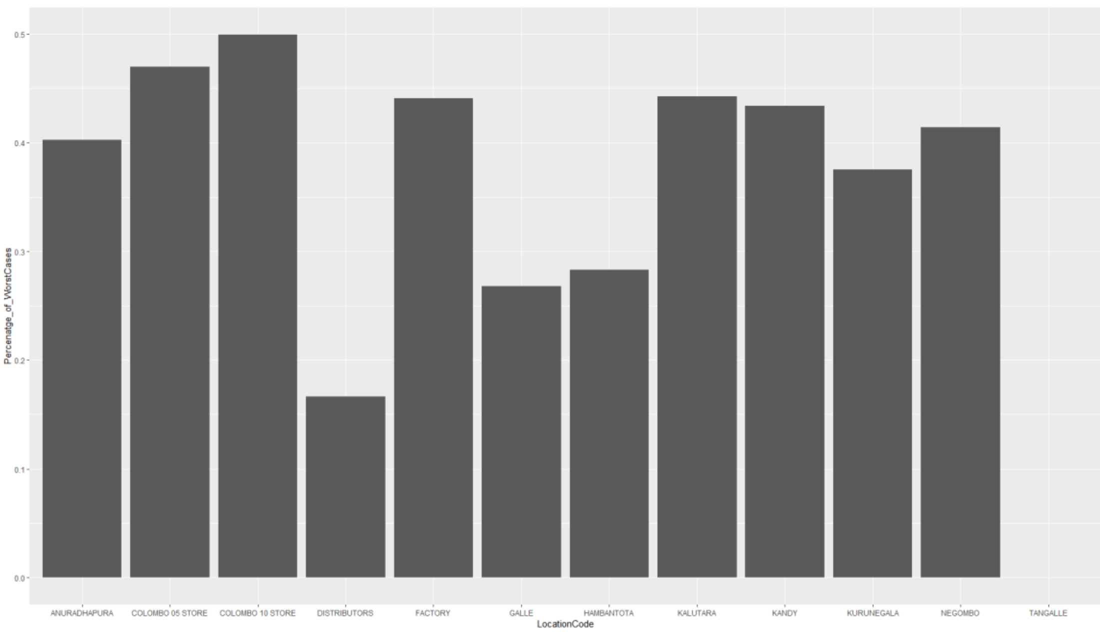


Figure 2.2.2-1:Location wise summary

By ranking the location where the problem was able to limit the data size. Furthermore, by talking with experts who have the experience and found physical that the water bottle is selling for cash is able to narrow down the data set to 2272.

4.4 Summary

The chapter highlights the approach of reducing the data set to identify possible frauds and customers have irregularities in the consumption pattern the frauds and to identify the consumption pattern of the customers

Chapter 5

5 Implementation

5.1 Introduction

This chapter depict about the use of different algorithms to predict consumption prediction and fraud detection

5.2 Data collection

The data has been collected from four routes in which in Colombo 10 operations from where the data operators had found that the manual ticket has been used to use to sold-out bottles. the data has been taken for the last 6 months January 2018 to June 2018.the selected sample size 13% percent from the total bases.

Collected data

1. Customer Description
2. Consumption Levels
3. Customer Complaints
4. Stock available
5. Missed delivery
6. Housed closed
7. Instances of manual tickets
8. Manual Invoices

5.3 Data Preparation

Data has been taken from the delivery ticket data. When preparing data if there were any null values it is considered as the customer has not consumed water for that period and missing data of contracted values have been calculated based on inventory value and the frequency.

5.3.1 Customer Selection

Customer were categorized as per the consumption level and as per the nature of business

Table 5.3.1-1:Customer Categorization

Categorization	**Bottle consumption**	**Category**	**Description**
Consumption Level	Over 100	Corporate	In here the group level consumption is considered
	Less than 100	Household	
		SME	
Nature of the Business	Government		

5.3.2 Consumption levels

By Binning technique categorise the data into classes as “Overconsumed”,” Good”, “Better”,” Worst”

Class labels were identified based on the ROI Sheet, considering overall consumption of the entire customer base and the projected revenue loss

Table 5.3.2-1:Class Lables of Water Consumption

level	Class	Description
>100%	Overconsumption	More than the contracted water bottles. In normal circumstance a customer can be consumed only up to the contracted bottles, but increase and if it is happening in continually, it is a problem. because with the increase of consumption the contracted value is adjusted.
100%-95%	Good	Prescribe level of consumption. When accruing a customer keep a threshold of 5%, to drop the consumption

95%-80%	Normal	When Considering the overall consumption of the entire base, the consumption is in 80%
<80%	Worst	If the consumption level is below 80%, it is a problem for the company because it leads to the loss of projected revenue.

5.3.3 Customer Complaints

- Complaint Data (6 months' data)

Table 5.3.3-1:Count of Complaints

Complaint Category	Count of Complain Number
Missed Delivery Calls	10715
Request Before the Scheduled Date	4124
Delivery Pending	3438
Same Day Delivery	3377
House closed	2942
Stock Available	2383
Pending Call Over 3 days	1135
Customer Was Not at Home	854
Call On Delivery	757
Invoice Dispute	661
Delivered Full Inventory	300
CSD Inactive	204
Missed Delivery	60
Visited - After office hours	41
OB – Stock Available	28
Inactive for Over 30 Days	2
Route planning Issue	1
Stop Delivery	1

- Data Categorization

The data has collected for six months; no. of instance a customer can complain has add up.

Condition of the data categorization:

Table 5.3.3-2:Complaint data selection

No. of instance	Class Label	Description
<=0	Nonproblem	Customer is satisfied with the delivery
<=3	Okay	In six months', time a customer can complaint for delivery issue
>3	Not okay	Above 3 months means the customer has complaint >=4, it means the customer is complaining regularly for delivery issues

5.3.4 Stock available

In here consider the stock availability at the time delivering the bottle. If the customer has the water bottle he/she won't take water.

The data has collected for six months. Therefore no. of instance a customer can refuse of bottle taking has added up.

The condition of the data categorization:

Table 5.3.4-1:Stock data selection

No. of instance	Class Label	Description
<=0	Able to deliver	The customer is taking the bottled water at every time it delivers

<=2	Okay	In six months', time a customer can refuse taking a bottle
>2	Not okay	Above 2 months means the customer has complaint >=2, it means the customer is rejecting in very frequently

5.3.5 Missed delivery

In here consider about the times that have missed customer by not delivering

The data has collected for six months. Therefore no. of instance a customer can refuse of bottle taking has added up.

The condition of the data categorization:

Table 5.3.5-1:Missed Delivery data selection

No. of instance	Class Label	Description
<=0	AbletoDeliver	Customer is taking the bottle water at every time it delivers
<=2	Okay	In six months', time a no. of times customer is missed
<2	Not okay	Above 2 months means the company has missed two deliveries

5.3.6 Housed closed

In here consider about the times about the no. of the time is not present at the time delivering

The data has collected for six months, therefore no. of instance a customer can refuse of bottle taking has add up.

Condition of the data categorization:

Table 5.3.6-1:House closed data selection

No. of instance	Class Label	Description
<=0	Abletodeliver	Customer is taking the bottle water at every time it delivers
<=2	Okay	In six months', time the no. of instance customer is not available
<2	Not okay	Above 2 months means customer is not present often

5.3.7 Instances of manual tickets

In here consider about the times about the no. of the times customer has taken a bottle more than planned delivery or prior to planned dates

The data has collected for six months. Therefore no. of instance a customer can refuse of bottle taking has added up.

The condition of the data categorization:

Table 5.3.7-1:Manual tickets data selection

No. of instance	Class Label	Description
<=0	DelivedviaPrintedTicket	The customer is taking the bottled water at every time it delivers
<=3	Okay	In six months', time a customer can refuse the

<3	Not okay	Above 3 months means the customer has taken from the manual ticket. If the customer wants more bottle customer can contact the company can adjust the delivery, if it is done then it won't go through the manual tickets

5.3.8 Manual Invoices

In here consider about the times about the no. of the times Customer invoice adjusted and also it is adjusted mostly if there is a case that the customer is not accepting with the final value

The data has collected for six months, therefore no. of instance a customer can refuse of bottle taking has added up.

Condition of the data categorization:

Table 5.3.8-1:Manual Invoice data selection

No. of instance	Class Label	Description
<=0	No error	The customer is taking the bottled water at every time it delivers
<=2	Okay	In six months can adjust the bill if there is a mistake of Price
<2	Not okay	Above 2 months means Customer invoice is adjusted frequently Where the customer is not accepting with final price values

5.4 Consumption Predication Methods

5.4.1 Naïve Bayes

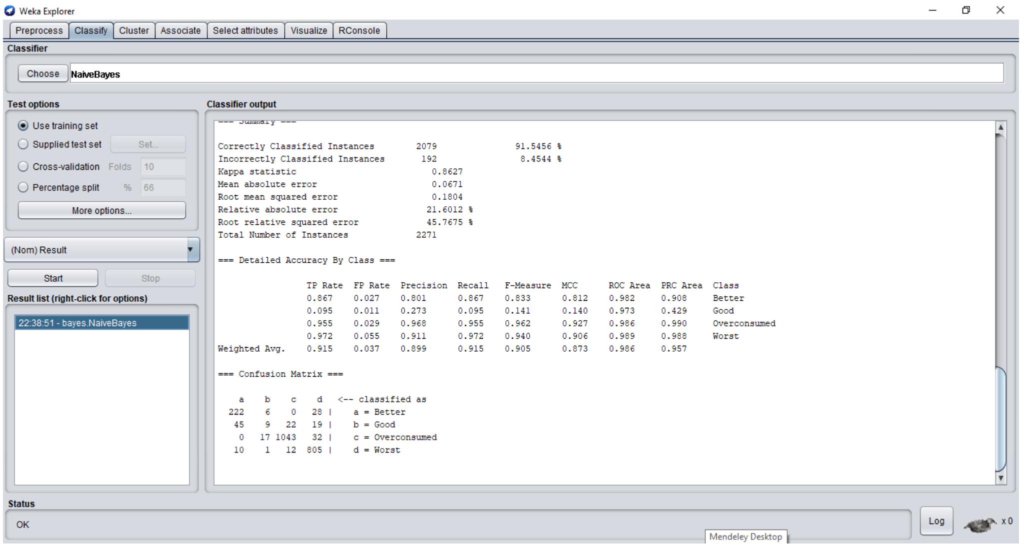


Figure 5.4.1-1:Naïve Bayes Summary Window

Result

Correctly Classified Instances 2079 91.5456 %

Incorrectly Classified Instances 192 8.4544 %

<u>Model</u>

```
=== Classifier model ===

Naive Bayes Classifier

                     Class
Attribute               Better      Good    Overconsumed      Worst
                        (0.11)     (0.04)    (0.48)           (0.36)
=============================================================
==============
Customercategory
  Corporate              42.0       14.0      111.0      43.0
  Government              9.0        6.0       20.0      15.0
  Household             113.0       48.0      485.0     437.0
  SME                    96.0       31.0      480.0     337.0
```

[total]	260.0	99.0	1096.0	832.0
LatestFrequncy				
Call On Delivery	13.0	7.0	15.0	44.0
Fortnight	101.0	29.0	306.0	321.0
Monthly	68.0	31.0	141.0	339.0
Weekly	78.0	32.0	634.0	128.0
[total]	260.0	99.0	1096.0	832.0
Contracted				
mean	14.1925	27.1234	10.2975	12.7604
std. dev.	32.7717	102.6365	40.6824	30.119
weight sum	256	95	1092	828
precision	10.7813	10.7813	10.7813	10.7813
AverageConsumption				
mean	13.2489	27.74	18.6313	6.1581
std. dev.	28.6648	100.4021	44.1921	13.0983
weight sum	256	95	1092	828
precision	1.6197	1.6197	1.6197	1.6197
Percentage				
mean	0.8732	0.989	2.064	0.536
std. dev.	0.0476	0.0282	1.697	0.1687
weight sum	256	95	1092	828
precision	0.0565	0.0565	0.0565	0.0565

Figure 5.4.1-2: Naive Bayes Model

5.4.2 Decision Tree

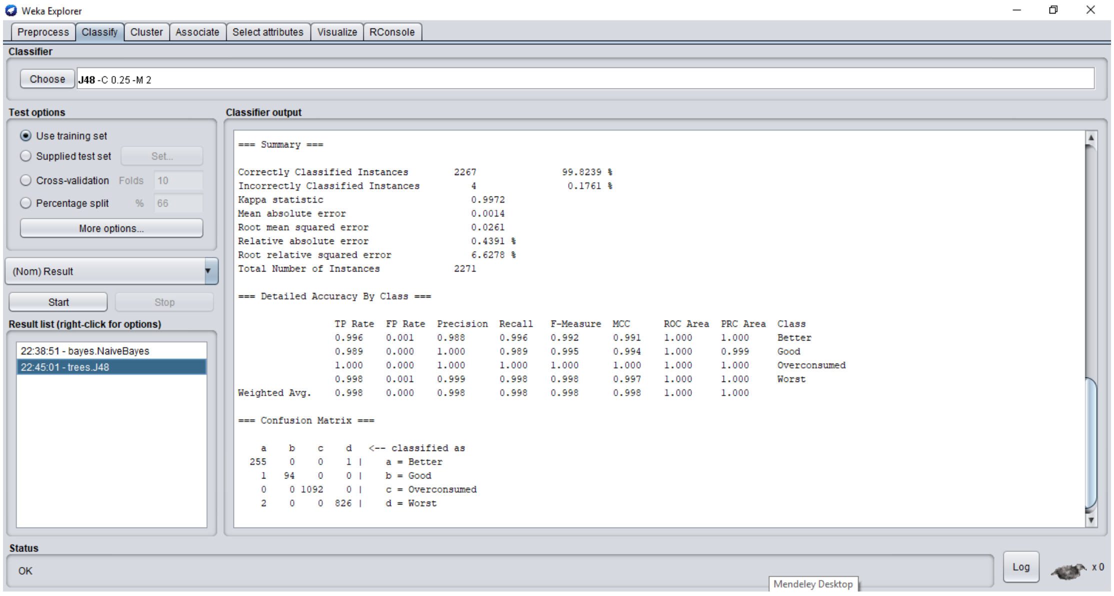


Figure 5.4.2-1:Decsion Tree Summary indow

Result

Correctly Classified Instances 2267 99.8239 %

Incorrectly Classified Instances 4 0.1761 %

<u>Model</u>

```
Percentage <= 1
|  Percentage <= 0.79: Worst (808.0)
|  Percentage > 0.79
|  |  Percentage <= 0.95
|  |  |  Percentage <= 0.8
|  |  |  |  Contracted <= 4: Worst (8.0)
|  |  |  |  Contracted > 4
|  |  |  |  |  Contracted <= 5: Better (16.0)
|  |  |  |  |  Contracted > 5
|  |  |  |  |  |  AverageConsumption <= 7.916667: Worst (11.0/1.0)
|  |  |  |  |  |  AverageConsumption > 7.916667: Better (6.0/2.0)
|  |  |  Percentage > 0.8
|  |  |  |  Percentage <= 0.94: Better (225.0)
|  |  |  |  Percentage > 0.94
|  |  |  |  |  Contracted <= 5
|  |  |  |  |  |  AverageConsumption <= 4.727273: Better (4.0)
|  |  |  |  |  |  AverageConsumption > 4.727273: Good (3.0)
|  |  |  |  |  Contracted > 5: Better (7.0/1.0)
|  |  Percentage > 0.95: Good (91.0)
Percentage > 1: Overconsumed (1092.0)


Number of Leaves  :   11
Size of the tree :    21
```

Figure 5.4.2-2: Desicion Tree model

Tree view

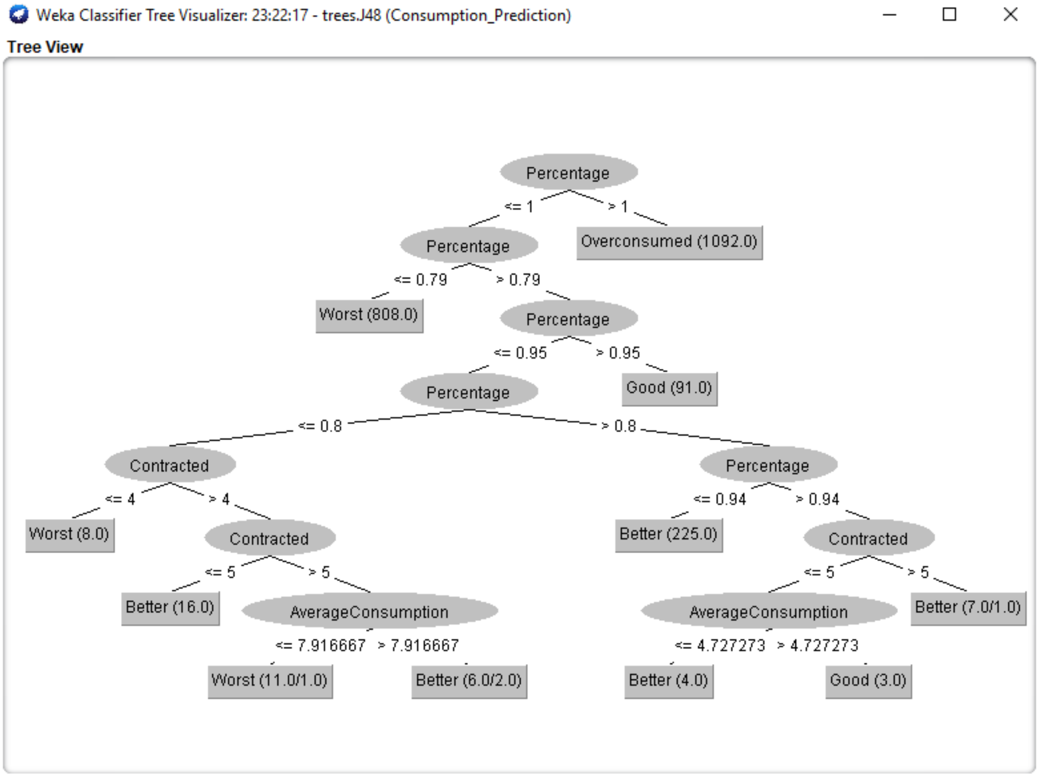


Figure 5.4.2-3:Tree View

5.4.3 ANN (Neural Networks)

By adding different counts of neurons the accurate percentage has evaluated

Model

=== Classifier model ===

Sigmoid Node 0

Inputs Weights

Threshold 2.0932258075898376

Node 4 -8.693816139406872

Node 5 -0.4028261922605185

Node 6 -7.586194766459652

Node 7 -5.1867453929484615

Sigmoid Node 1

Inputs Weights

Threshold 0.354330732086679
Node 4 -13.428851058620845
Node 5 1.2436613521422377
Node 6 -1.9427026402267342
Node 7 -3.5278788231584484
Sigmoid Node 2
Inputs Weights
Threshold -3.688608013552241
Node 4 -22.320012944706196
Node 5 -1.773346137970906
Node 6 11.485222109833174
Node 7 8.547069426143745
Sigmoid Node 3
Inputs Weights
Threshold 0.8165950986680904
Node 4 6.797480740385635
Node 5 -5.062207002905253
Node 6 -27.192280823175555
Node 7 -20.943896210919814
Sigmoid Node 4
Inputs Weights
Threshold -33.734279423148344
Attrib Customercategory=Corporate 16.836143178790255
Attrib Customercategory=Government 16.51671539832366
Attrib Customercategory=Household 17.26898125968492
Attrib Customercategory=SME 16.925729551378645
Attrib LatestFrequncy=Call On Delivery 16.844475984053375
Attrib LatestFrequncy=Fortnight 17.293202606173374
Attrib LatestFrequncy=Monthly 16.465305103150286
Attrib LatestFrequncy=Weekly 16.91842184342466
Attrib Contracted 13.205025451336056
Attrib AverageConsumption -16.492228210741825
Attrib Percentage -104.4039857907635

Sigmoid Node 5
    Inputs   Weights
    Threshold   5.500285063929485
    Attrib Customercategory=Corporate   -0.41710966811693617
    Attrib Customercategory=Government   -2.7537518738074707
    Attrib Customercategory=Household   -5.973365064057261
    Attrib Customercategory=SME   -1.8493299664616865
    Attrib LatestFrequncy=Call On Delivery   -0.9658465928973093
    Attrib LatestFrequncy=Fortnight   -4.680363859148938
    Attrib LatestFrequncy=Monthly   -0.5150080421328419
    Attrib LatestFrequncy=Weekly   -4.825788321078447
    Attrib Contracted   7.285628730301476
    Attrib AverageConsumption   12.459378421583468
    Attrib Percentage   5.8255943652856494
Sigmoid Node 6
    Inputs   Weights
    Threshold   31.613813241591842
    Attrib Customercategory=Corporate   -15.891276478900782
    Attrib Customercategory=Government   -15.734420719878702
    Attrib Customercategory=Household   -15.8580848651698
    Attrib Customercategory=SME   -15.716016438850296
    Attrib LatestFrequncy=Call On Delivery   -15.850790184461895
    Attrib LatestFrequncy=Fortnight   -15.896269044221432
    Attrib LatestFrequncy=Monthly   -15.81486667053061
    Attrib LatestFrequncy=Weekly   -15.659072438571986
    Attrib Contracted   -12.56489350711633
    Attrib AverageConsumption   13.333567411815652
    Attrib Percentage   104.83045232374707
Sigmoid Node 7
    Inputs   Weights
    Threshold   23.183934020894544
    Attrib Customercategory=Corporate   -11.29529774552557
    Attrib Customercategory=Government   -11.840248726309031

Attrib Customercategory=Household   -11.723356619565775

Attrib Customercategory=SME   -11.293355588475976

Attrib LatestFrequncy=Call On Delivery   -11.80521507180794

Attrib LatestFrequncy=Fortnight   -11.448766771354578

Attrib LatestFrequncy=Monthly   -11.578256811031224

Attrib LatestFrequncy=Weekly   -11.3869493132088

Attrib Contracted   -12.404331691380085

Attrib AverageConsumption   13.40868764171589

Attrib Percentage   77.35960461060596

Class Better

Input

Node 0

Class Good

Input

Node 1

Class Overconsumed

Input

Node 2

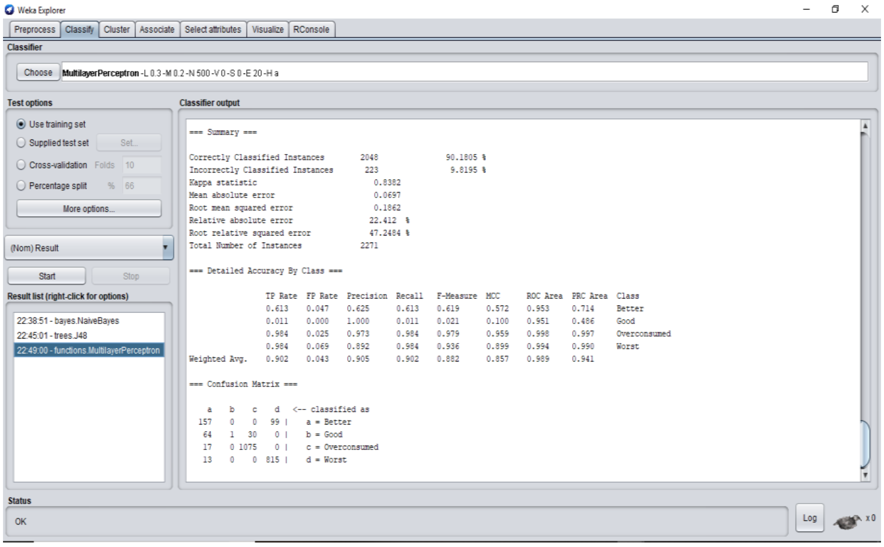
Class Worst

Input

Node 3

Figure 5.4.3-1:Neural Network model

- **Default node default Layer**

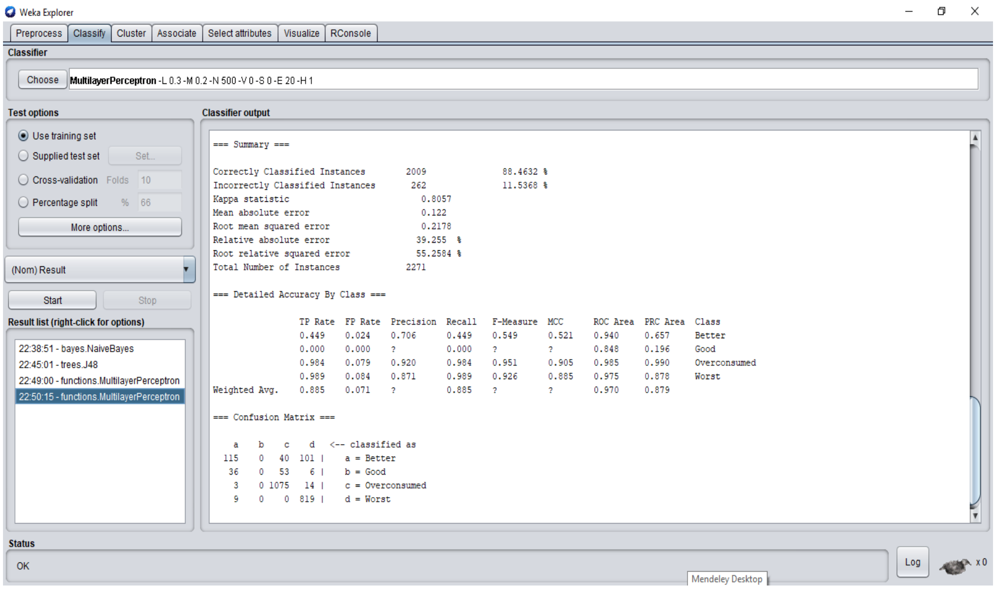


Result

Correctly Classified Instances 2048 90.1805 %

Incorrectly Classified Instances 223 9.8195 %

- **One node One Layer**

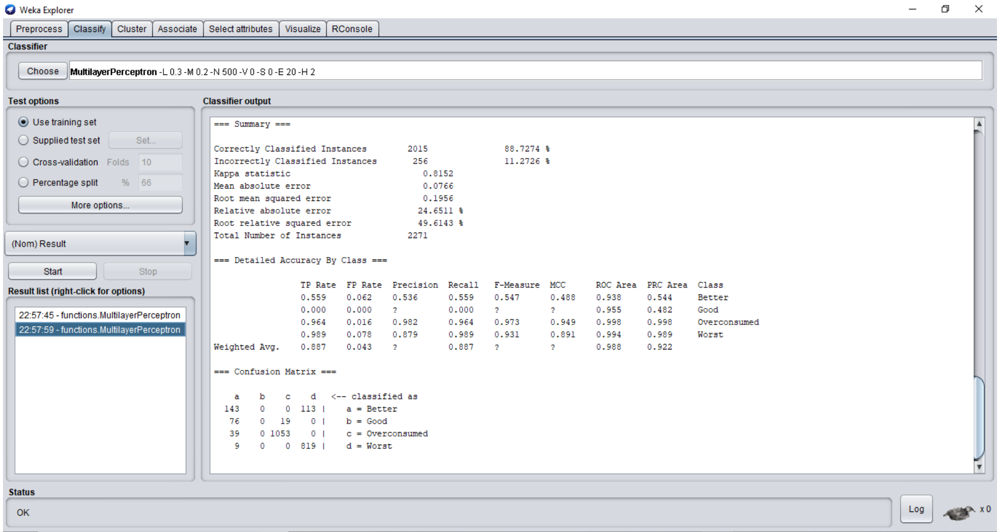


Result

Correctly Classified Instances 2009 88.4632 %

Incorrectly Classified Instances 262 11.5368 %

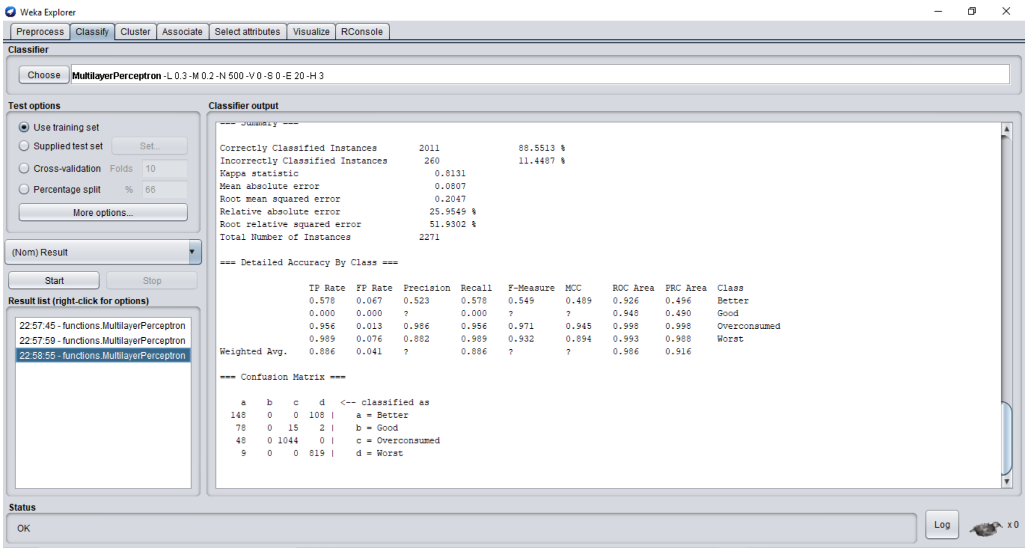
- **Two node one Layer**



Result

Correctly Classified Instances	2015	88.7274 %
Incorrectly Classified Instances	256	11.2726 %

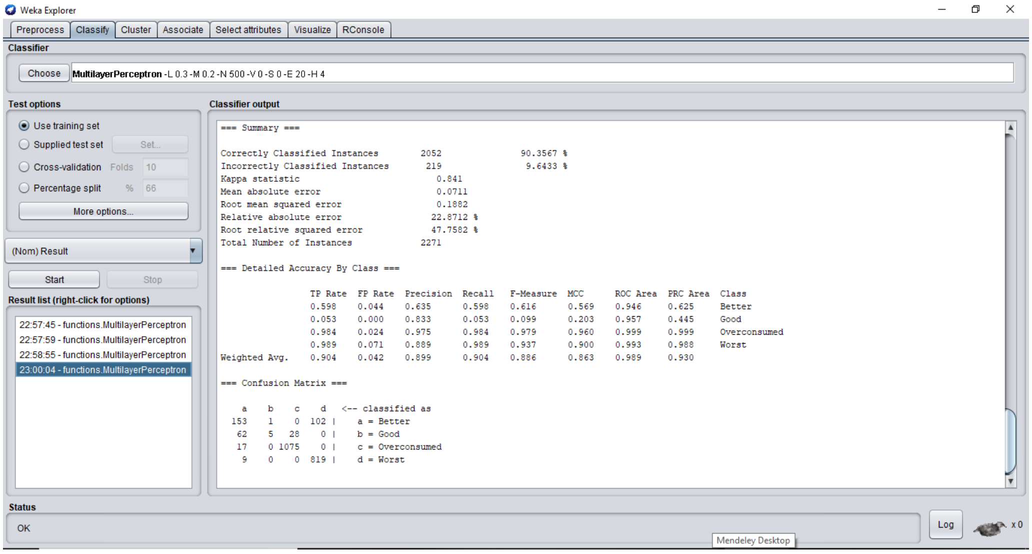
- **Three node one layer**



Result

Correctly Classified Instances	2011	88.5513 %
Incorrectly Classified Instances	260	11.4487 %

- **Four node one layer**

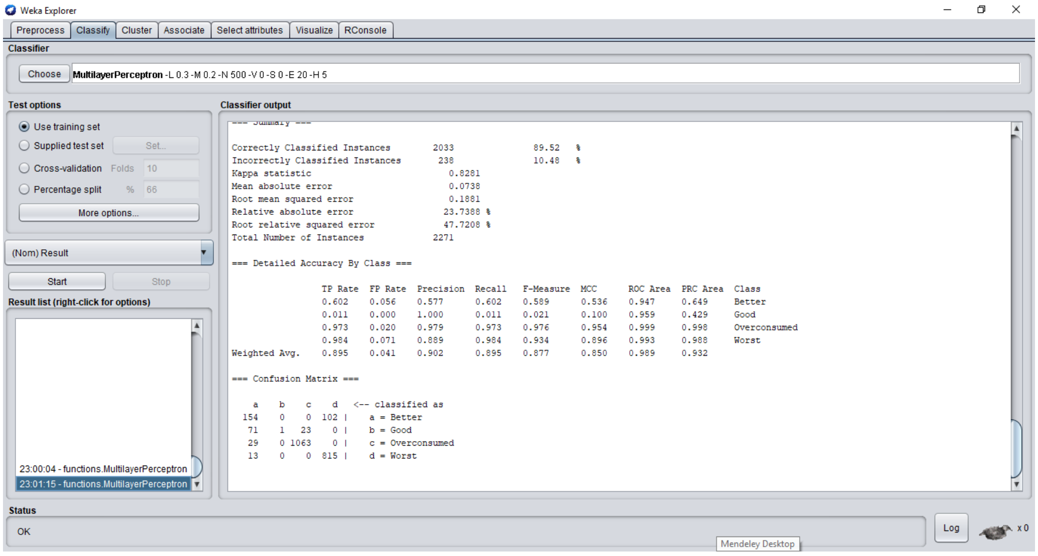


Result

Correctly Classified Instances 2052 90.3567 %

Incorrectly Classified Instances 219 9.6433 %

- **Five node one layer**

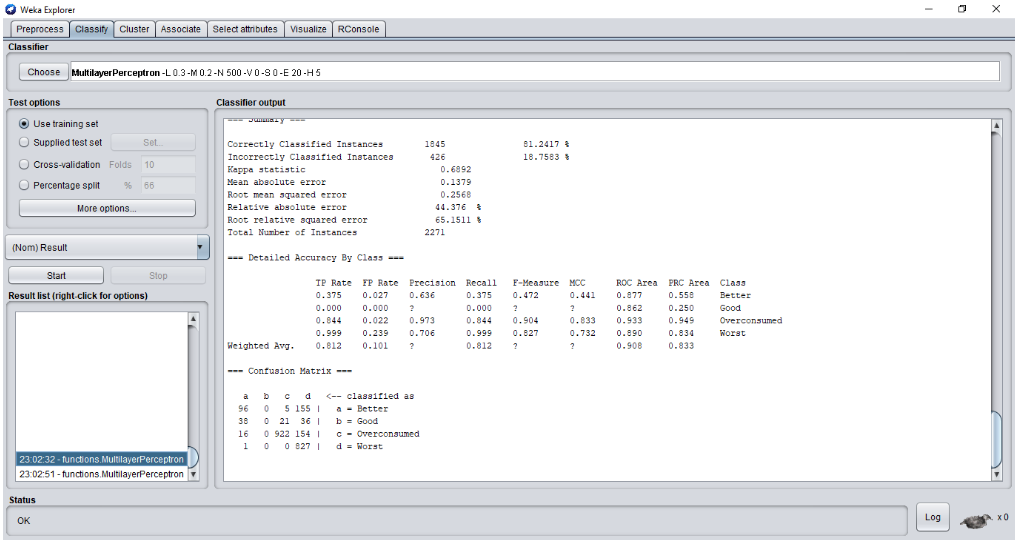


Result

Correctly Classified Instances 1547 89.52 %

Incorrectly Classified Instances 724 10.48 %

- **Four, Three nodes two Layer**

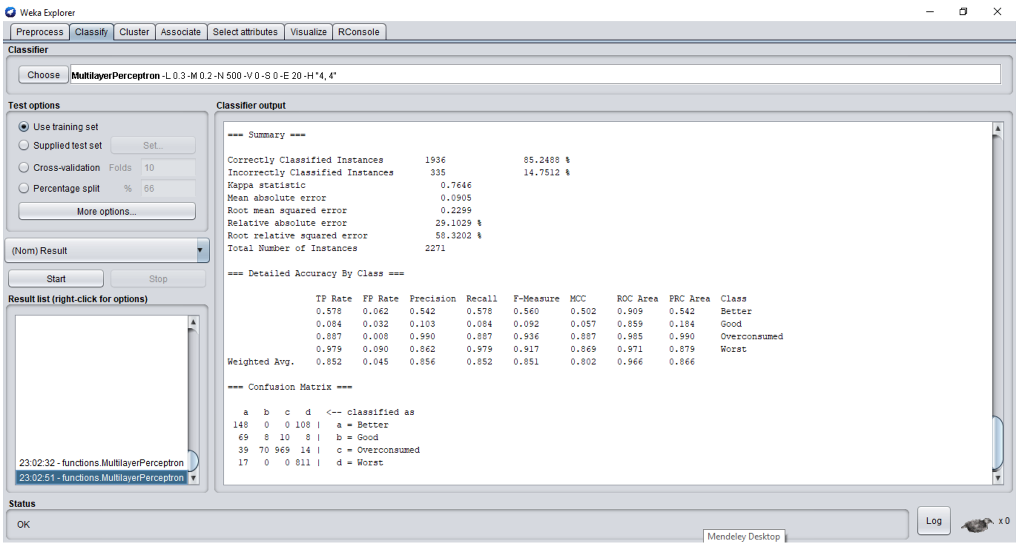


Result

Correctly Classified Instances 1845 81.2417 %

Incorrectly Classified Instances 426 18.7583 %

- **Four, Four node two layers**

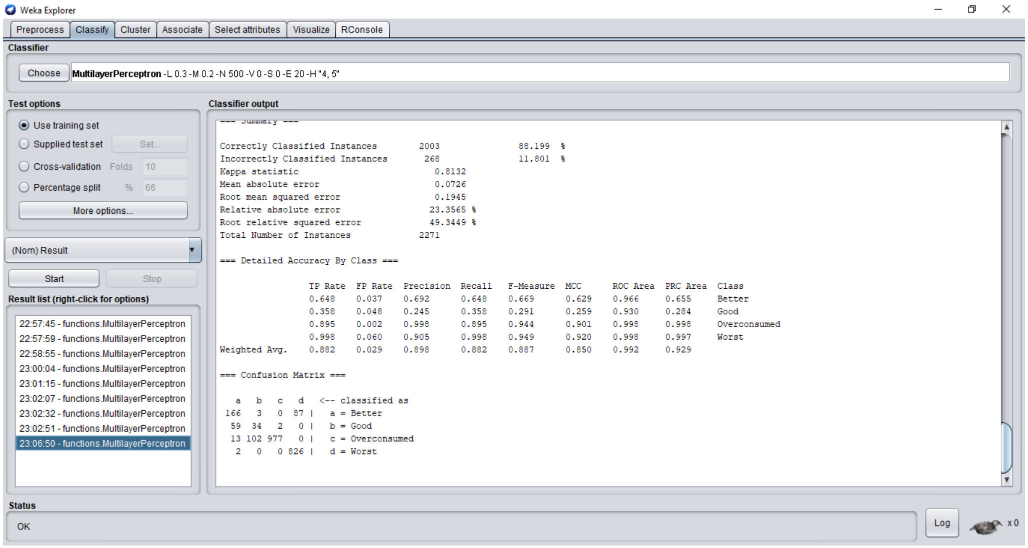


Result

Correctly Classified Instances 1936 85.2488 %

Incorrectly Classified Instances 335 14.7512 %

- **Four, Five node two layers**



Result

Correctly Classified Instances 2003 88.199 %

Incorrectly Classified Instances 268 11.801 %

- Conclusion

Result Analysis of ANN (Neural Networks)

Table5.4.3-1: Neral Natwork Result

Case	Correctly Classified Instances	Incorrectly Classified Instances
Default node default Layer	90.1805 %	9.8195 %
One node One Layer	88.4632 %	11.5368 %
Two node one Layer	88.7274 %	11.2726 %
Three node one layer	88.5513 %	11.4487 %
Four node one layer	90.3567 %	9.6433 %
Five node one layer	89.52 %	10.48 %
Four, Three nodes two Layer	85.2488 %	14.7512 %
Four, Five node two layers	90.3567 %	9.6433 %
Four, Five node two layers	88.199 %	11.801 %

Therefore, the most suitable results Four, Four node two layer

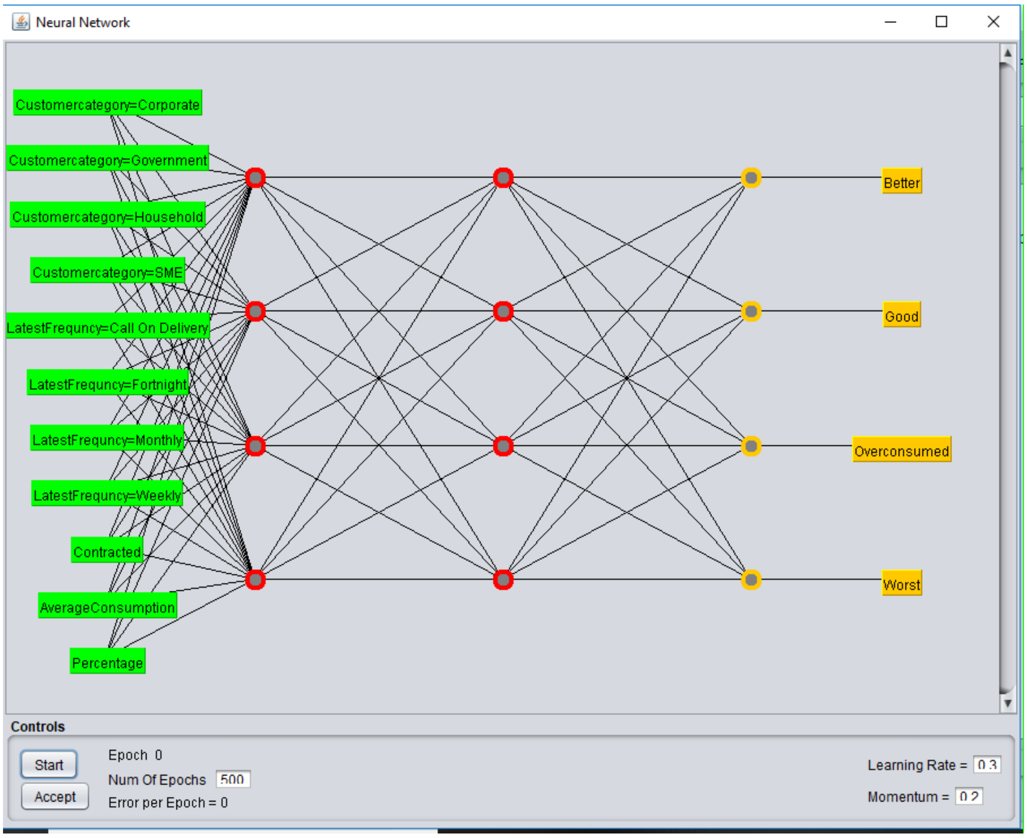


Figure 5.4.3-2:Four, Four Two Layer

5.4.4 Accuracy of the model

Table 5.4.4-1:Acuracy table

Algorithms	(1) Naïve Bayes	2) Neural Networks	(3) Decision Trees
Accuracy	90.86%	90.75%	99.60% v

v[i]- Standard for most suitable algorithms for the consumption prediction

As per the results the selected algorithms are Decision Trees

5.5 Possible Fraud detection

To identify the instances where fraud can happen, a rule set has been introduced. By applying of these rules set to data, the dataset has been learned.

Rules set

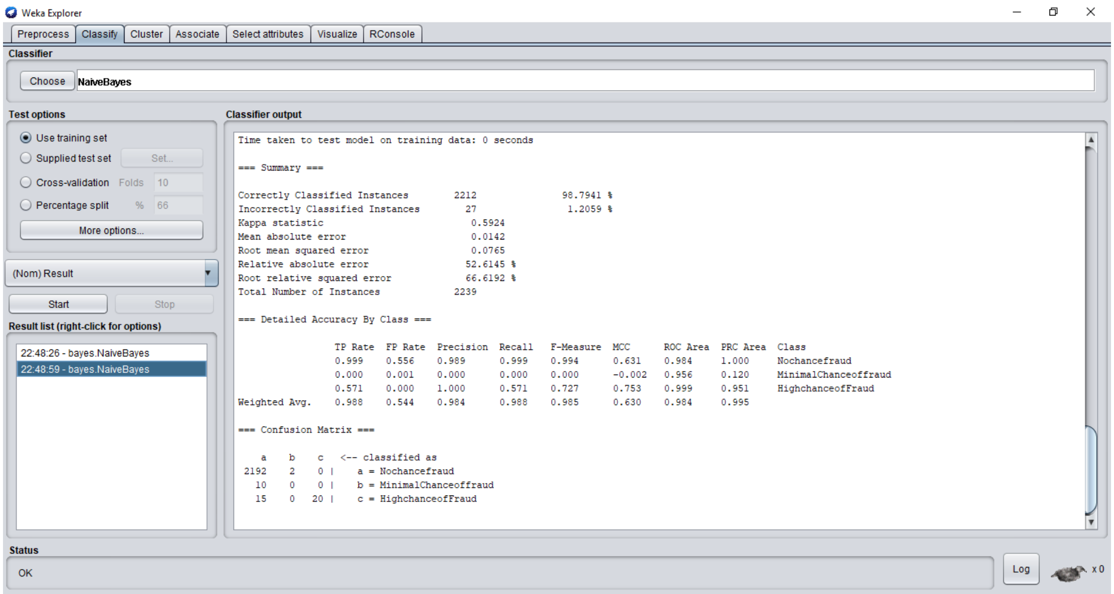
Table 5.4.4-1:Rule Based to classify

Consumption	Houseclosed	StockAvaible	MissedDelivery	WaterComplaint	Manaulticket	Manaulti	Result
Overconsumed	Okay	Okay	Okay	Anything	okay	okay	MinimalchanceofFraud
Overconsumed	Notokay	Notokay	Notokay	Anything	DelivereviaPrintetedtikets	Notokay	Highchanceoffruad
Worst	Abletodeliver	Notokay	Abletodeliver	Anything	Notokay	Anything	Highchanceoffruad
Worst	Abletodeliver	Okay	Abletodeliver	Anything	okay	Anything	MinimalchanceofFraud
Better	Abletodeliver	Notokay	Abletodeliver	Anything	Notokay	Anything	MinimalchanceofFraud

The rules have been evaluated by using tree algorithms

- Naïve Bayes
- Decision Tree
- Neural Networks

5.5.1 Naïve Bayes



Result

Correctly Classified Instances 2212 98.7941%

Incorrectly Classified Instances 27 1.2059%

Model

=== Classifier model (full training set) ===

Naive Bayes Classifier

Attribute	Class Nochancefraud (0.98)	MinimalChanceoffraud (0)	HighchanceofFraud (0.02)
Customercategory			
Corporate	203.0	2.0	4.0
Government	47.0	1.0	1.0
Household	1053.0	8.0	21.0
SME	895.0	3.0	13.0
[total]	2198.0	14.0	39.0
LatestFrequncy			
Call On Delivery	75.0	2.0	1.0
Fortnight	742.0	4.0	10.0
Monthly	549.0	6.0	23.0
Weekly	832.0	2.0	5.0
[total]	2198.0	14.0	39.0
Consumption			
Better	251.0	5.0	1.0
Good	96.0	1.0	1.0
Overconsumed	1092.0	2.0	1.0
Worst	759.0	6.0	36.0
[total]	2198.0	14.0	39.0
StockAvaible			
Abletodeliver	776.0	1.0	1.0
Okay	558.0	6.0	36.0
Notokay	863.0	6.0	1.0
[total]	2197.0	13.0	38.0
Manual ticket			
Notokay	2021.0	6.0	36.0
Okay	161.0	6.0	1.0

DelivedviaPrintedTicket	15.0	1.0	1.0
[total]	2197.0	13.0	38.0
House closed			
Okay	725.0	4.0	1.0
NotOkay	350.0	2.0	1.0
Abletodeliver	1122.0	7.0	36.0
[total]	2197.0	13.0	38.0
missed delivery			
Abletodeliver	856.0	10.0	36.0
Okay	1028.0	2.0	1.0
Notokay	313.0	1.0	1.0
[total]	2197.0	13.0	38.0
WaterComplaint			
NotComplaint	1160.0	5.0	27.0
Not oaky	292.0	5.0	2.0
Okay	744.0	3.0	9.0
NoMaunalInvoice	2.0	1.0	1.0
[total]	2198.0	14.0	39.0
ManaulInvoice			
okay	96.0	1.0	2.0
Noerror	2083.0	10.0	35.0
Not okay	18.0	2.0	1.0
[total]	2197.0	13.0	38.0

Figure 5.5.1-1:Naive Bayes model

5.5.2 Decision Tree

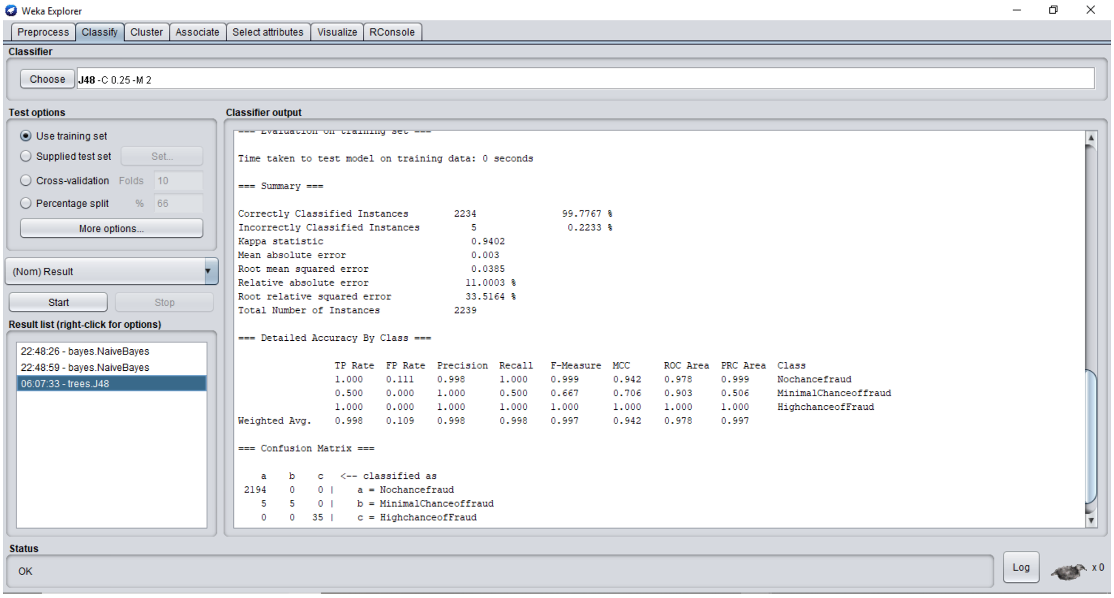


Figure 5.5.2-1:Decision tree result window

Result

Correctly Classified Instances 2234 99.7767 %

Incorrectly Classified Instances 5 0.2233 %

Model

```
=== Classifier model (full training set) ===


J48 pruned tree
------------------
StockAvaible = Abletodeliver: Nochancefraud (775.0)
StockAvaible = Okay
|   Consummption = Better: Nochancefraud (65.0)
|   Consummption = Good: Nochancefraud (19.0)
|   Consummption = Overconsumed: Nochancefraud (303.0)
|   Consummption = Worst
|   |   MissedDelivery = Abletodeliver
|   |   |   Houseclosed = Okay: Nochancefraud (26.0)
|   |   |   Houseclosed = NotOkay: Nochancefraud (20.0)
|   |   |   Houseclosed = Abletodeliver
|   |   |   |   Manaulticket = Notokay: HighchanceofFraud (35.0)
|   |   |   |   Manaulticket = Okay: MinimalChanceoffraud (5.0)
|   |   |   |   Manaulticket = DelivedviaPrintedTicket: HighchanceofFraud (0.0)
|   |   MissedDelivery = Okay: Nochancefraud (100.0)
|   |   MissedDelivery = Notokay: Nochancefraud (24.0)
StockAvaible = Notokay: Nochancefraud (867.0/5.0)


Number of Leaves:   12
Size of the tree:   17
```

Figure 5.5.2-2:Desion Modeler

Tree View

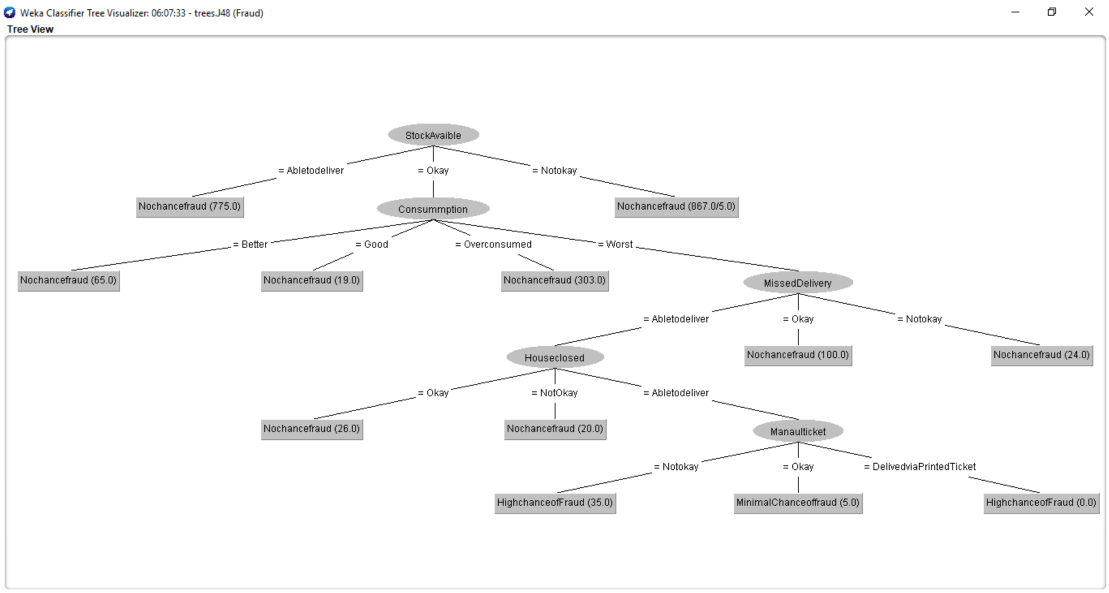
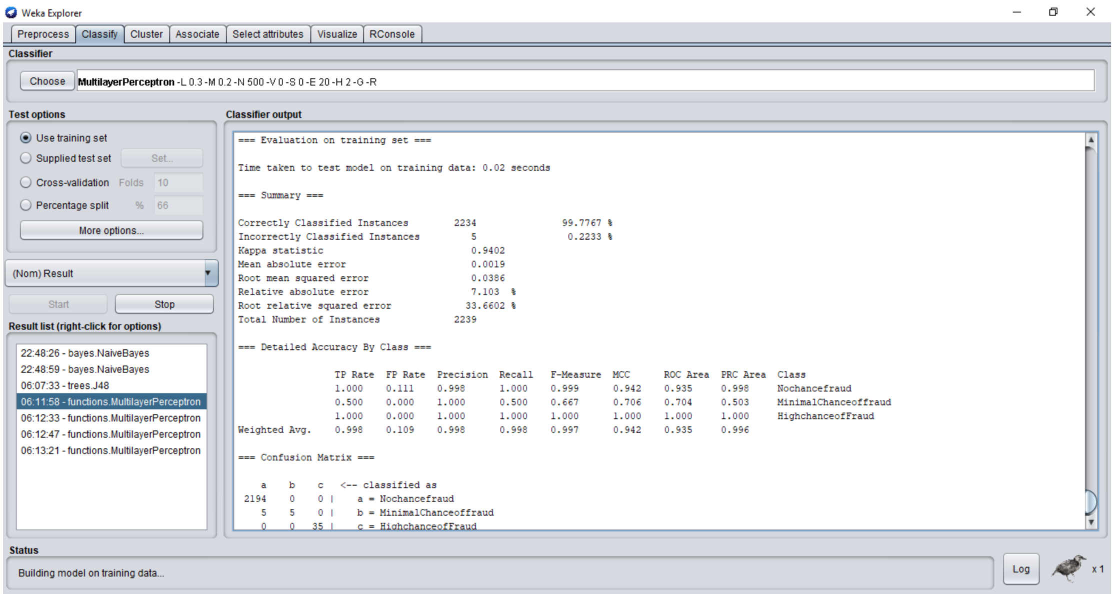


Figure 5.5.2-3:Decision tree view

5.5.3 Neural Networks

- **Default nodes**

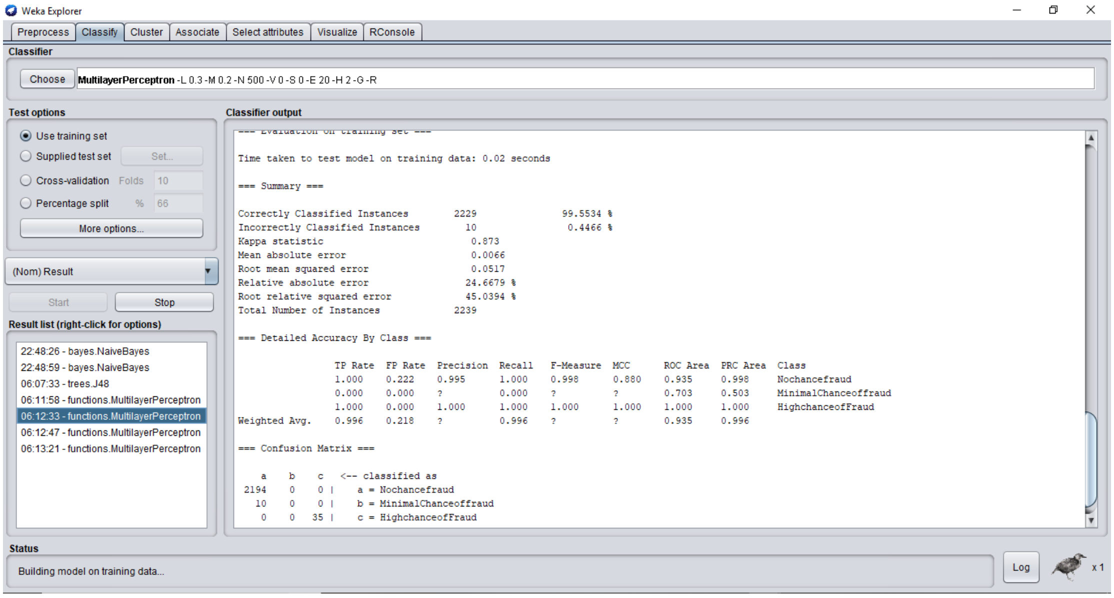


Result

Correctly Classified Instances 2234 99.7767 %

Incorrectly Classified Instances 5 0.2233 %

- **one node**

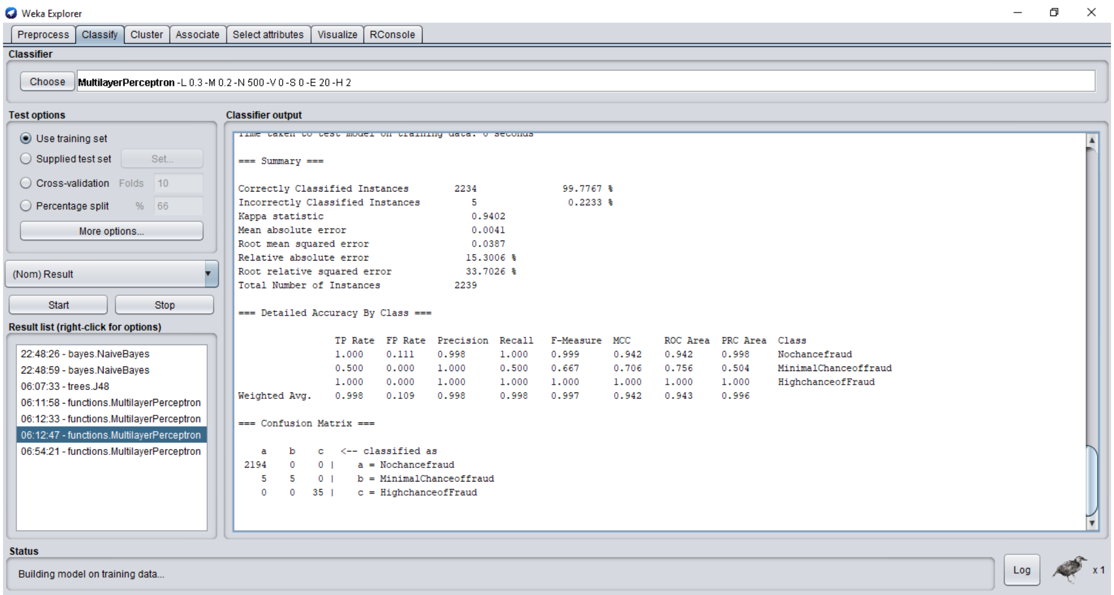


Results

Correctly Classified Instances 2229 99.5534 %

Incorrectly Classified Instances 10 0.4466 %

- **Two nodes**



Results

Correctly Classified Instances 2234 99.7767 %

Incorrectly Classified Instances 5 0.2233 %

- For three/four nodes also the same result is deriving

Results

Correctly Classified Instances 2234 99.7767 %

Incorrectly Classified Instances 5 0.2233 %

Therefore no. of nodes to evaluate the problem is 2 and above

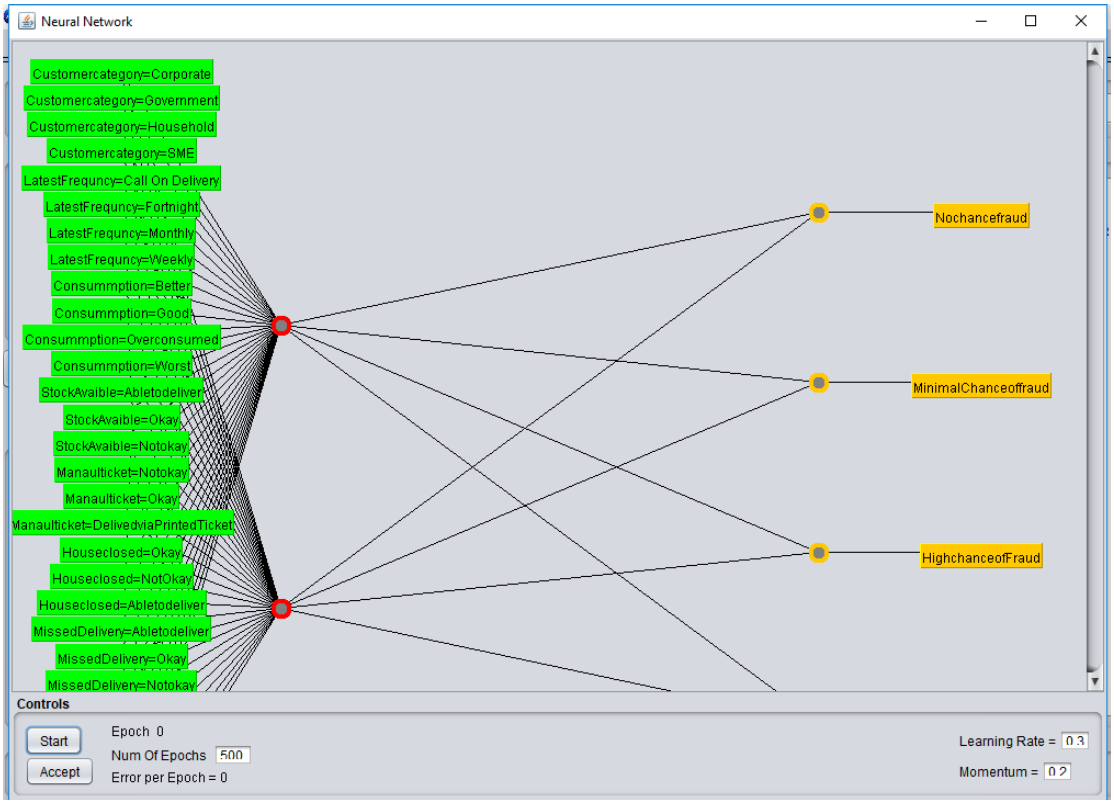


Figure 5.5.3-1:Two nodes, one layer

5.5.4 Accuracy of the Algorithms

Table 5.5.4-1:Acuracy table

Algorithms	(1) Naïve Bayes	2) Neural Networks	(3) Decision Tree
Accuracy	98.61%	99.78 v	99.78 v

v[ii]- Standard for most suitable algorithms for the prediction possible fraud detection

Since Neural Network and Tree is having same accuracy, do the research the selected method is Neural Networks.

5.6 Summary

The summary illustrates the use of different classification methods to predict the consumption and detect the possible frauds in the water delivery process.

5.6.1 Consumption Prediction

The data set has evaluated using three algorithms as Naïve Bayes/ Decision Trees/ Neural Networks. Out of their the decision tree produces the most accurate percentage.

5.6.2 Fraud Detection

The data set has evaluated by using three algorithms as Naïve Bayes/ Decision Trees/ Neural Networks. Out of them, Neural network produce the most accurate percentage

By using the following techniques identify whether there is a pattern of putting manual tickets, House closed and missed delivery for certain customers and whether there is a link between instance average consumption below contracted value and percentage number of instances below average consumption

Chapter 6

6 Implementation of the Model

6.1 Introduction

In chapter 6, the design of the solution has been described in terms of what and how each component does. This chapter described implementation of each models

6.2 Result evaluation

6.2.1 Consumption Prediction Model Evaluation

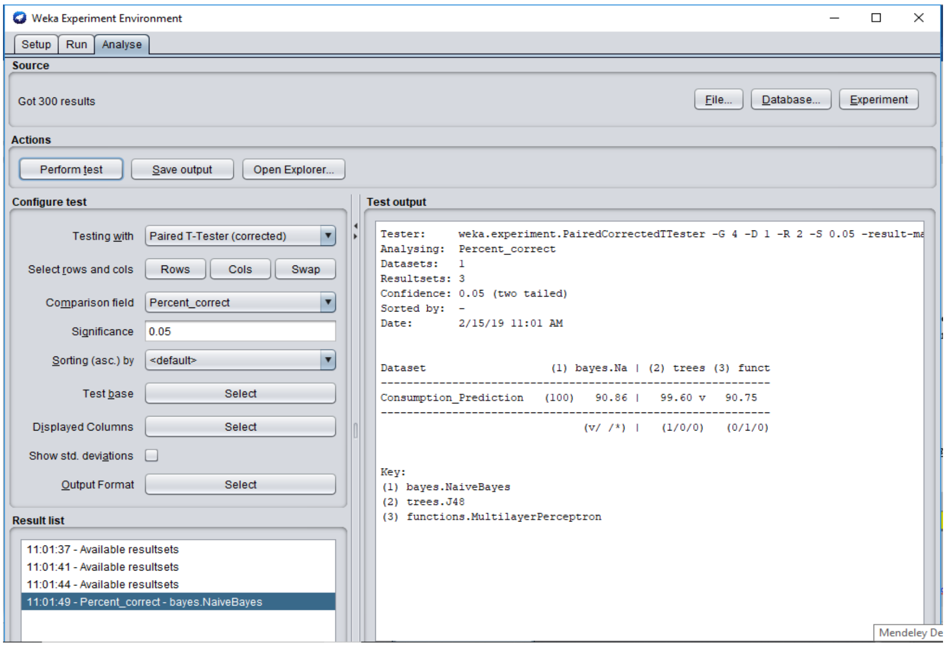


Figure 6.2.1-1:Results of the model selection

Selected model for evaluation is Decision Tree

6.2.1.1 <u>Result of Consumption category prediction</u>

- Knowledge flow process:

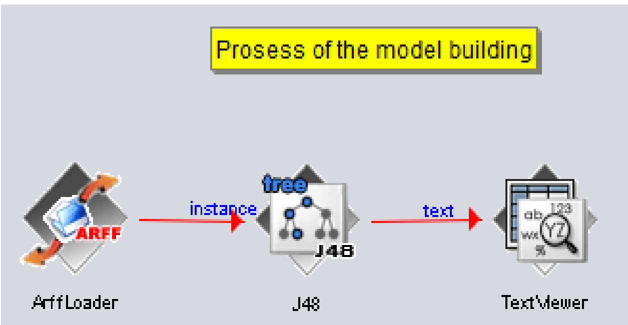


Figure 6.2.1-2:Knowledge flow Steps

- Summary of Prediction list:

Selected model: Decision Tree

Model Predicted Accuracy: 99.8239 %

Actual Result Accuaracy:100%

Table 6.2.1-1:Result table of Consumption predicion

Model Result

Class Label	Count of Class labels correctly classified
Better	2
Worst	40
Grand Total	42

Actual Result

Class Label	Count of Actual Result
Better	2
Worst	40
Grand Total	42

6.2.1.2 Attribute Selection for consumption prediction

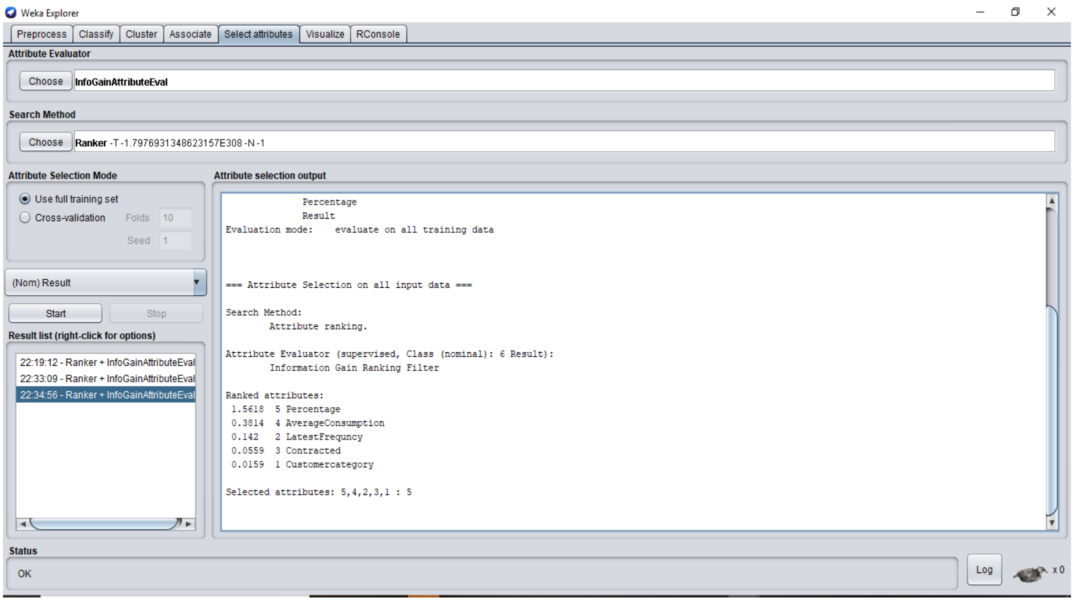


Figure 6.2.1-3:Attribute selection window

Table 6.2.1-2:Attribute raninking table

Infor Gain	Ranked	Attributes:
1.5618	5	Percentage
0.3814	4	AverageConsumption

0.142	2	LatestFrequncy
0.0559	3	Contracted
0.0159	1	Customercategory

Selected attributes: 5,4,2,3,1: 5

6.2.2 *Fraud Detection Model*

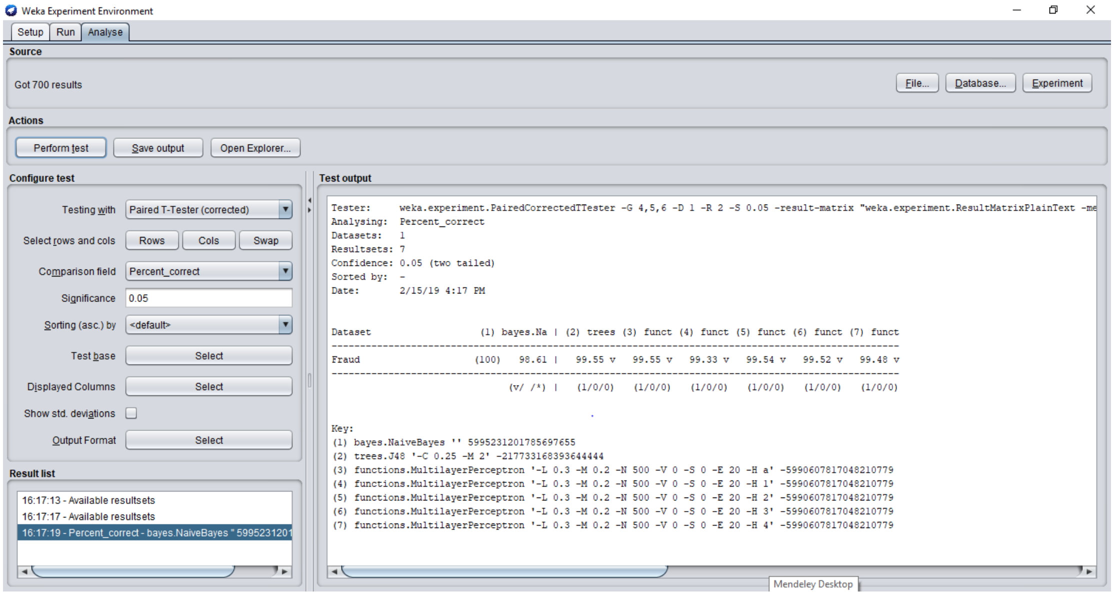
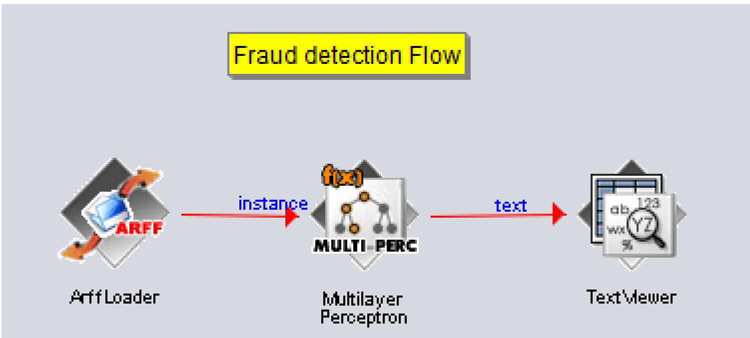


Figure 6.2.2-1:Fraud detection Model

Selected model for evaluation is Neural Networks

6.2.2.1 <u>*Result of Fraud detection prediction*</u>

- Knowledge flow process:



- Summary of possible fraud prediction

Selected model: Neural Networks

Model Predicted Accuracy: 99 99.732 %

Actual Result Accuaracy:100%

Table 6.2.2-1:Result table of Fraud detection

Model Result

Class Label	Count of Class labels correctly classified
HighchanceofFraud	1
Nochancefraud	31
Grand Total	32

Actual Result

Class Label	Count of Actual Result
HighchanceofFraud	1
Nochancefraud	31
Grand Total	32

6.2.2.2 Attribute Selection for consumption prediction

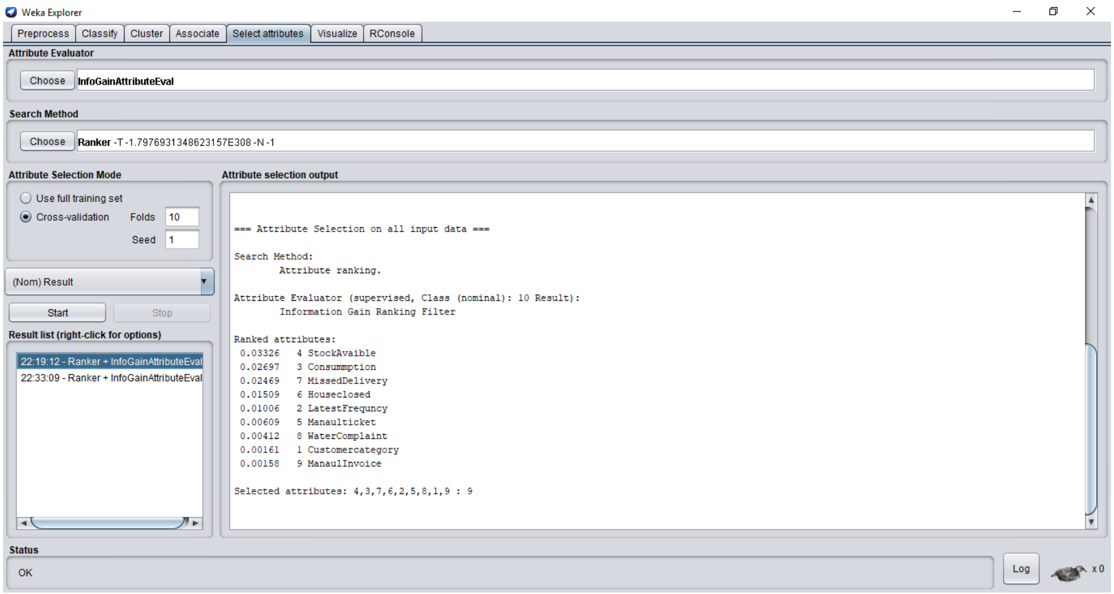


Figure 6.2.2-2:Attribute selection window

Information Gain	Ranked	Attributes:
0.03326	4	StockAvaible
0.02697	3	Consumption
0.02469	7	Missed Delivery
0.01509	6	House closed
0.01006	2	LatestFrequncy
0.00609	5	Manaulticket

0.00412	8	WaterComplaint
0.00161	1	Customercategory
0.00158	9	ManaulInvoice

Selected attributes: 4,3,7,6,2,5,8,1,9: 9

6.3 Summary

This chapter includes results of the models, Attribute ranking and the tables of accuracy of the models.

Chapter 7

7 Discussion

7.1 Introduction

This chapter illustrates the accuracy of the model by using confusion matrix and ROC Curve.

A confusion matrix is a table that is used to **describe the performance of a classification model** on a set of test data for which the true values are known

A ROC curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR) where it is commonly used graph the summarizes of the performance of a classifier over all possible thresholds.[32]

By considering the instances of correctly classified, the most suitable model has been selected.

7.2 Importance of the research

- Consumption Prediction

  As per the table

True positives (TP): The cases in which we predicted which it will fall under better/Good/Overconsumed/Worst categories.

False positives (FP): We predicted it would fall under better/ Good/ Overconsumed/ Worst, but where they don't actually fall under those categories.

Since this model is correctly classified the data TPR is equal to 1 where ROC is 1

Table 6.2.2-1:Classifications table

	TP Rate	**FP Rate**	**ROC Area**	**Class**
	0.996	0.001	1.000	Better
	0.989	0.000	1.000	Good
	1.000	0.000	1.000	Overconsumed
	0.998	0.001	1.000	Worst
Weighted Avg.	0.998	0.998	0.998	1.000

- Confusion matrix

a	b	c	d	<-- classified as
252	0	0	4	a = Better
4	91	0	0	b = Good
0	0	1092	0	c = Overconsumed
2	0	0	826	d = Worst

There are four possible predicted classes: better/Good/Overconsumed/Worst categories If we were predicting Water consumption for example, "Good" would mean they have the consumed water in Average water consumption

- Possible fraud Detection

True positives (TP): The cases in which we predicted which it will fall under HighchanceofFraud, MinimalChanceoffraud and Nochancefraudcategories.

False positives (FP): We predicted it will fall under HighchanceofFraud, MinimalChanceoffraud and Nochancefraud but where they don't actually fall under those categories.

Since this model is correctly classified the data TPR is equal to 1 where ROC is 1

Table 6.2.2-2:Detailed Accuracy by Class (Nureal Networks)

	TP Rate	FP Rate	ROC Area	Class
	1.00	0.11	0.942	HighchanceofFraud
	0.50	0.00	0.756	MinimalChanceoffraud
	1.00	0.00	1.00	Nochancefraud
Weighted Avg.	0.998	0.109	0.942	

Confusion Matrix

a	b	c	<-- classified as
2194	0	0	a = Nochancefraud

5	5	0	b = MinimalChanceoffraud
0	0	35	c = HighchanceofFraud

There are three possible predicted classes: HighchanceofFraud, MinimalChanceoffraud and Nochancefraudcategories. categories If we were predicted Minimal chance of fraud example, it will be “MinimalChanceoffraud " Where it means water delivery of this customer is suspicious

Summary of the Customer base distribution as per HighchanceofFraud, MinimalChanceoffraud and Nochancefraudcategories

Class	**Count of Unique code**	**Sum of average**	Percentage of Customer Count	Percentage Delivery
HighchanceofFraud	36	178.98	1.6%	1%
MinimalChanceoffraud	10	75.22	0.4%	0%
Nochancefraud	2225	31,234.92	98.0%	99%
Grand Total	**2271**	**31489.12**		

Average loss to the company from losing the inventory (bottle)

Average Active Customer Base Count	18,000
Selected bases	2,271
Percentage of the selected base	13%
Average Water Bottle delivery (Per Month)	300,000
Average Water Bottle delivery (Per Month) in Selected bases	31,902
Possible fraud in Entire	2390.457274
Average Price per bottle in 2018 (Rs.130)	310,759.45
Revenue lost per month	310,759.45

7.3 Future Works

Due to this fraudulent act, there is a huge lost incurring to the company in many ways. Main lost is revenue lost. To mitigate the problem the company had taken security measures by balancing stock report daily, but in order to increase efficiency of identifying the fraudulent acts the data mining techniques could be used as proved in the research

7.3.1 Areas of future study

- Identify whether there is actual fraud with selected base and by doing further
- Study on customer rent ion

8 Reference

[1] A. Tagaris, P. Mnimatidis, D. Koutsouris, and S. Member, “Implementation of a Prescription Fraud Detection Software Using RDBMS Tools and ATC Coding,” no. November, pp. 5–7, 2009.

[2] C. Ellawala, B. V. Y. Bopetta, and C. Warnasinghe, “Bottled Drinking Water : Evaluation of Consumption Tendencies and Quality,” pp. 3–4, 2015.

[3] D. D. M. N. P. Koggalage R, “Consumption of Bottled Water and its Impact on Bottled Water Market in Sri Lanka,” *Annu. Res. Symp.*, no. 13th Annual Research Symposium, University of Kelaniya, 2012.

[4] “FOOD CONTROL ADMINISTRATION UNIT ,” 2015.

[5] E. Humaid, “A Data Mining Based Fraud Detection Model for Water Consumption Billing System in MOG,” p. 79, 2012.

[6] TutorialPoint, “Data Mining Applications & Trends.” .

[7] Wikipedia, “Cross Industry Standard Process for Data Mining.” .

[8] “CRISP-DM – a Standard Methodology to Ensure a Good Outcome - Data Science Central.” [Online]. Available: https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome. [Accessed: 02-Dec-2017].

[9] B. L. Patterson and V. Marketing, “Nine Common Types of Data Mining Techniques Used in Predictive Analytics,” pp. 1–5.

[10] S. Dejan, “Data Mining Algorithms In R,” 2011. [Online]. Available: http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R. [Accessed: 02-Dec-2017].

[11] T. B. Blank and S. D. Brown, “Data processing using neural networks,” *Analytica Chimica Acta*, vol. 277, no. 2. pp. 273–287, 1993.

[12] Dotdash publishing Family, “Neural Network.” [Online]. Available: https://www.investopedia.com/terms/n/neuralnetwork.asp.

[13] Z. Wang and X. Huang, “An Outlier Detection Algorithm Based on the Degree of Sharpness and Its Applications on Traffic Big Data Preprocessing,” pp. 478–

482, 2017.

[14] E. Turban, R. Sharda, and D. Delen, “Data Mining Methods,” *Decision Support and Business Intelligence Systems*. pp. 216–228, 2011.

[15] J. S. Randhawa and I. S. Ahuja, “Examining the role of 5S practices as a facilitator of business excellence in manufacturing organizations,” *Meas. Bus. Excell.*, vol. 21, no. 2, pp. 191–206, 2017.

[16] “Advantages and Disadvantages of Sampling.” [Online]. Available: https://accountlearning.com/advantages-and-disadvantages-of-sampling/. [Accessed: 01-Dec-2017].

[17] D. Peltier-Rivest and N. Lanoue, “Thieves from within: occupational fraud in Canada,” *J. Financ. Crime*, vol. 19, no. 1, pp. 54–64, 2011.

[18] R. Bhowmik, “Data Mining Techniques in Fraud Detection,” *Proc. Conf. Digit. Forensics, Secur. Law*, vol. 3, no. 2, pp. 57–72, 2008.

[19] A. I. Rana, G. Estrada, M. Sole, and V. Muntes, “Anomaly Detection Guidelines for Data Streams in Big Data,” *2016 3rd Int. Conf. Soft Comput. Mach. Intell.*, pp. 94–98, 2016.

[20] Ghosh and Reilly, “Credit card fraud detection with a neural-network,” *1994 Proc. Twenty-Seventh Hawaii Int. Conf. Syst. Sci.*, vol. 3, pp. 621–630, 1994.

[21] prasana sarma Venkatraam Balasubramanian, “(18) Neural Network in Data Mining - YouTube,” 2015. [Online]. Available: https://www.youtube.com/watch?v=HrdrXQxw3oo&t=305s. [Accessed: 19-Nov-2017].

[22] L. Swani and P. Tyagi, “Predictive Modelling Analytics through Data Mining,” pp. 5–11, 2017.

[23] D. Yue, X. Wu, Y. Wang, Y. Li, and C. H. Chu, “A review of data mining-based financial fraud detection research,” *2007 Int. Conf. Wirel. Commun. Netw. Mob. Comput. WiCOM 2007*, pp. 5514–5517, 2007.

[24] S. Wang, “A comprehensive survey of data mining-based accounting-fraud detection research,” *2010 Int. Conf. Intell. Comput. Technol. Autom. ICICTA*

2010, vol. 1, pp. 50–53, 2010.

[25] S. Sayad, “Naive Bayesian,” 2019. [Online]. Available: https://www.saedsayad.com/naive_bayesian.htm.

[26] S. Premaratne, “Chapter 6 : Classification and Prediction Classification — A Two-Step Process Examples of Classification Task,” 2014.

[27] “Artificial Neural Network Building Blocks,” *Tutorialspoint.com*. [Online]. Available: https://www.tutorialspoint.com/artificial_neural_network/artificial_neural_network_building_blocks.htm.

[28] Kent State University, “About R and RStudio - Statistical & Qualitative Data Analysis Software - LibGuides at Kent State University.” .

[29] The University of Waikato, “Weka 3 - Data Mining with Open Source Machine Learning Software in Java.” .

[30] R. R. Bouckaert *et al.*, “WEKA Manual for Version 3-8-1,” 2016.

[31] Microsoft Excel - Wikipedia, “Microsoft Excel - Wikipedia.” .

[32] Kevin Markham, “Simple guide to confusion matrix terminology,” *Data School*. 2014.

[33] P. Sonego, A. Kocsor, and S. Pongor, “ROC analysis: Applications to the classification of biological sequences and 3D structures,” *Briefings in Bioinformatics*, vol. 9, no. 3. pp. 198–209, 2008.

Appendix A

9 Appendix

9.1 Code snippet to generate the summary from R studio

> CONSUM_recovered_1 <- read_excel ("F:/MYFiles/Semester2/Project Data set/CONSUM_recovered_1.xlsx", +   sheet = "Loctaion Code Selection")

> View(CONSUM_recovered_1)

>summary(CONSUM_recovered_1)

9.2 Code Snippet for data visualize as Bar Chart

> CONSUM_recovered_1 <- read_excel ("F:/MYFiles/Semester2/Project Data set/CONSUM_recovered_1.xlsx”, sheet = "Loctaion Code Selection")

> view(CONSUM_recovered_1)

>ggplot(Check2, aes(x = LocationCode, y = Percenatge_of_WorstCases)) + geom_col ()

Appendix B

9.3 Model Development Process

9.3.1 Model Selection

Weka Manual(3.8.1)[30]

Steps:

1. Select the tab: Classify
2. Choose → “Selected classify name”
3. Test options: Select “Use training set “
4. Click “Start”

9.3.2 Saving Model

Steps:

1. Right click on “Result list (Right click for options)”
2. Save the model in the computer
   (When needed to load the model)
3. Right click on “Result list (Right click for options)”
4. Select Load model
   (Upload the testing data set)
5. Test option →Supplied test set
6. Select data set which need to evaluate
7. In “Result list (Right click for options)” select “Revaluate the model

9.3.3 Model Evaluation

Steps:

1. Under “Experimenter “in software opening window
2. Select the Tab: Setup
3. Click “New” button
4. Add dataset to evaluate from “datasets”
5. Select the algorithms to evaluate from “Algorithms”
   To Evaluates the algorithms
6. Select the Tab: Run
7. Press “Start”
   Once the it stopped. To analyze the algorithms

8. Select "Experiment" button
9. Select "Row" as Dataset
10. "Cols" as Scheme
11. Press the button "Perform test"

9.3.4 Attribute selection

Steps:

1. Select the tab: Selection Attribute
2. Choose → InfoGainAttributeEval:
3. Attribute Selection Mode: Select "Cross Validation Folds:10 Seed=1"
4. Click "Start"

9.3.5 Confusion Matrix

True Positive Rate (TPR) and False Positive Rate FPR

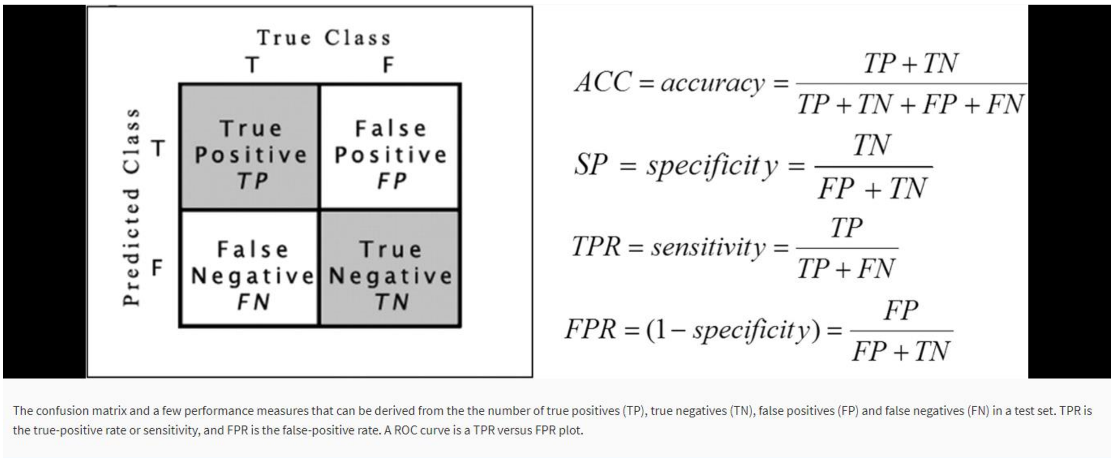


Figure 9.3.5-1:TRP and FPR[33]

---

[i] V is designed by the weak tool to indicate the most suitable algorithms