e-Learning Product Sales Prediction using SMS Activation Data

V. S. W. Wickramasekara
189474B

Faculty of Information Technology
University of Moratuwa

July 2021

e-Learning Product Sales Prediction using SMS Activation Data

V. S. W. Wickramasekara
189474B

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for the partial fulfillment of Degree of Master of Science in Information Technology.

July 2021

Declaration

We declare that this thesis is our 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 Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.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
V. S. W. Wickramasekara

Signature of Student

Date

Supervised by
Name of Supervisor
Mr. S. C. Premaratne

Signature of Supervisor

Date

Acknowledgements

I am so grateful to my supervisor Mr. Saminda Premaratne, Senior Lecturer, Faculty of Information Technology, University of Moratuwa for academic guidance, advice, time, encouragement and patience that has seen me through, making this work success.

Also, I would like to give a special thanks to my beloved family members for their support, love, advice and encouragement throughout the period of my study. I lastly wish to extend my sincere gratitude toward my colleagues who have tirelessly guided and encouraged me at all times through my course.

.

Abstract

Most companies nowadays using data mining techniques and algorithms for the decision-making process and to make profitable adjustments in operation and production.

This helpful to identifying possibilities of a customer buying products after buying a particular product, identifying a group of customers who's in the difficulty in the activation process/ identifying customers who are going to drop from the company/ how it effects for the company, identifying connection between customer purchasing methods, find the group of customers to promote particular product/s and forecasting the product sales process (required quantity of each product for special annual events like exhibitions).

Understanding of consumers helps companies to sell more. Customers are a valuable source of information, collection of data that organizations identify different customers and observe how they behave. The more knowledge about consumers and their different needs, it is easier to identify market opportunities to sell and invent new products and target consumers with appropriate offers. Use of company past data, production and operations can predict using data mining techniques and algorithms.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

The Savi Publications (Pvt) Ltd is a leading e-publisher in Sri Lanka. The company develop, publish and market interactive e-learning software for the national curriculum in Sri Lanka. Products can be purchase as CD/DVD, USB pen drive or downloadable e-product. The content which available are grade 1 to 5 all subjects, grade 6 to 11 all compulsory subjects plus some optional subjects, for A/L students some subjects and some general knowledge products. When the government is updating the school syllabus then the company also modifying the subject contents accordingly. These products are only work in windows operating system.

They have several ways of publishing products. CD/ DVD version, Internet download version and Pen drive version. For CD/ DVD version and Internet download version, the products are activated via SMS. In Pen drive version product are activated in a different way.

In the Sri Lankan market, most of the products are selling in CD/ DVD version and Internet download version because the product prices are in reasonable rage. But the Pen drive version, buying a special range of customer because the price is high. Therefore, the demand is low.

Both CD/ DVD and Internet download version can be installed to two users (two computers) and can use as a lifetime product. The purpose of giving two users is if a customer installs a particular product to one computer then by any chance that computer is broken down then customers can get another chance to activate the same product to a different computer.

In modern Sri Lanka, most of the Students and parents are tend to learn with the use of technological way, but they are having a lack of knowledge in the technology field. Therefore, the company face various difficulties when selling products. This issue can be minimized by improving market demand and customer satisfaction. By analyzing past recodes the company can predict future trends. This can be utilized to the enhancement of the company sales and market demand while reducing the costs.

1.2 Problem Statement

There are several limitations in existing researches. There is no proper data mining prediction for the company in these areas.

- Difficulty in possibilities of a customer buying related other products after buying a particular product.
- Difficulty in identifying repeat buyers and drop customers.
- Difficulty in identifying connection between customer purchasing methods.
- Difficulty in finding a group of customers to promote product/s.
- Difficulty in forecasting product sales process.

To get decision support it needs a broad range of past data. By observing existing data, need to come up with a better solution.

1.3 Aim and Objectives

1.3.1 Aim

Provide a decision support that can be helpful for the decision-making process and identify future trends in the market and to make profitable adjustments in operation and production.

1.3.2 Objectives

Understanding of consumers helps companies to sell more. Customers are a valuable source of information, collection of data that organizations identify different customers and observe how they behave. The more knowledge about consumers and their different needs, it is easier to identify market opportunities to sell and invent new products and target consumers with appropriate offers.

- Possibilities of a customer buying related other products after buying a particular product.
- Identifying customers who are going to continuing purchasing or drop from the company and how it effects for the company.
- Identifying connection between customer purchasing methods.
- Find a group of customers to promote a particular product/s.
- Identifying product sales forecasting process (required quantity of each product for special annual events like exhibitions).

1.4 Scope of Research

This is helpful for the decision-making process and identify future trends in the market. This helpful to identifying possibilities of a customer buying products after buying a particular product, identifying a group of customers who's in the difficulty in the activation process/ identifying customers who are going to drop from the company/ how it effects for the company, Social media marketing impact of selling products, find the group of customers to promote particular product/s and forecasting the product sales process (required quantity of each product for special annual events like exhibitions). Use of company past data, production and operations can predict using data mining techniques.

1.5 Propose solution

- Area 1: Association Rule Mining is used as data mining technique. Apriori is used as data mining algorithm and identified set of rules to make decisions.
- Area 2: Classification is used as data mining technique. Trees J48 and Trees.RandomForest are used as data mining algorithms to make better decisions.
- Area 3: Association Rule Mining is used as data mining technique. Apriori is used as data mining algorithm and identified set of rules to make decisions.
- Area 4: Clustering is used as data mining technique. Simple EM (Expectation Maximization) class is used as data mining algorithm and identified better clusters to promote products.
- Area 5: Time series forecasting is used as data mining technique. Linear Regression and Multilayer Perceptron are used as data mining algorithm to forecast future products quantities.

1.6 Structure of the theses

The content of the overall thesis is structured as follows.

The First Chapter includes an introduction of project with the background and motivation, problem statement, aim and objectives, scope of research, propose solution and structure of the theses. The second chapter critically reviews the observation of objectives. The Third Chapter is for environment setup with observation on machine learning tools and packages. The Forth Chapter presents the research methodology on main phases with business understanding, data understanding, data preparation, modeling and evaluate algorithms in main 5 areas. Finally, Fifth Chapter concludes the solution with a note of further work.

Chapter 2

Literature Review

2.1 Introduction

This chapter includes all the observation for objectives such as Possibilities of a customer buying related other products after buying a particular product. Identifying customers who are going to continuing purchasing or drop from the company and how it effects for the company. Identifying connection between customer purchasing methods. Find a group of customers to promote a particular product/s. Identifying product sales forecasting process (required quantity of each product for special annual events like exhibitions).

2.2 The observation of objectives

- Possibilities of a customer buying other products after buying a particular product and Identifying a group of customers who's in the difficulty in the activation process/ identifying customers who are going to drop from the company/ how it effects for the organization.

Consideration for customer behavior is an essential success of corporate organizations. Customers personal features like interests and opinions, Costumers perceptions and attitudes of reaction and Social and legal factors are behind consumer behavior. It is important to study in how individual customer or group of customers selects, buys, uses and provides ideas for products or services. Understanding on buying pattern of a consumer is Personalized marketing. By analyzing behavior of individual customer, it is easy to predict future trends.

Identification of problems, search for information, evaluation of alternatives, decision taking to purchase and tend for after purchase are five phases of consumer's purchase decision process [1].

The work of organizations does not stop when the customer buys a product or service. Companies must be alert on consumer behavior or attitudes towards products or services. After buying and using the product, the customer will be satisfied or dissatisfied with the product and service. The satisfied customers are keeping the same product and more often to buy the same product or service again and again. Then those customers are become loyal and valuable customers to that companies. Keep satisfied consumer are greatest. These values and environment are benefit for all customers and organizations. Overall objective of the business of making a profit will be achieved in the company perspective [1].

It is important to identify the customers who are most likely to return for repeat purchases. The transaction history at the shopping cart level is generally used to predict the repeat purchase behavior of customers. With aggregated information and temporary information, various aspects of customer behavior can be recorded. Most of the existing models use only aggregate-level functions or independent window-based functions regardless of the order of transactions. [2].

- Identifying customers who are going to continuing purchasing or drop from the company and how it effects for the company.

The aim is to identify the popularity of continuing purchasing or drop from the company and explore its potential impact on marketing practices. Consumer behavior is all about observing consumers and the progressions of selection, consume and dispose of products. More than that it includes the emotional, mental and behavioral responses of consumers [3]. Below are the features that impact for the behavior of customers:

1. Identify customer personal features, such as interests and opinions.
2. Costumers' perceptions and attitudes of reaction to a marketing and to environment.
3. Social and legal factors like influence people, level of education, social networks, income and government rules and regulations.

In business's world most valuable assets is organizational loyal customers. They have the better environment values and its can helpful for creating brand promotions.

- Identifying connection between customer purchasing methods.

Understanding customers helps companies to sell more. As a short-term strategy sales promotion and special marketing offer that generates more profit [4]. With the correct information, can create a valuable outline of company related customers. There are different attributes are there, those are mention below:

1. Identify corporate individual customers by age and gender.
2. Identify customers interests and opinion about the organization products and services.
3. Identify customers buying behavior like what products are buying, places of buying products, when they buying products and how they pay for them [5].

There is a method to analyze Multiple Regression analysis with the use of SPSS application is a way of identifying connection between customer purchasing methods [6].

- Find a group of customers to promote a particular product/s.

Profiling corporate customers and group them into different segments to address separately to get maximum advantage. Companies can offer different promotions and strategies by observing customers really wants. When grouping customers to different market sections and then can attract clients into the company. Segmenting customers into different groups according to their needs and there are some advantages [7].

1. Most profitable and least profitable customers can identify.
2. Can improve customer service.
3. Focus on the customers who are most possible to buy a product or service.
4. Select markets that are profitable for the company.
5. Identify new products and services.
6. Offer new products and services to loyal customers.
7. Create competition in the market.
8. Use company resources cleverly.
9. Improve products and services to encounter customer needs.
10. Keeping costs low and reasonable prices for products and services to increase profit in the company.

- Identifying the sales forecasting process (required quantity of each product for special annual events like exhibitions).

Many studies indicate that past data influence product sales. To changes in historical demand, use regressions and some statistical [8]. To improve prediction accuracy, many researchers have improved the Bass model with providing explicit guidance in to the algorithms. There are some researchers are expanded the Bass and Norton models to grow market price and adding market factors to predict the use of wrapping technology. These models only can use to forecast product sales with historical sales and valuation data. [9].

2.3 Summary

As reported in the relevant literature, finding possibilities of a customer buying other products after buying a particular product and Identifying a group of customers who's in the difficulty in the activation process/ identifying customers who are going to drop from the company/ how it effects for the organization use existing models like aggregate-level functions or independent window-based functions regardless of the order of transactions methods. Customer personal interests and opinions with perceptions and attitudes of reaction and social and legal factors are there to identifying customers who are going to continuing purchasing or drop from the company and how it effects for the company.

For the area of identifying connection between customer purchasing methods use multiple Regression analysis with the use of SPSS application. identify most profitable and least profitable customers, improve customer service, identify new products and services, offer new products and services to loyal customers and there are so many advantages for find a group of customers to promote a particular product/s. For the area of identifying the sales forecasting process use Bass and Norton models to grow market price and adding market factors to predict the use of wrapping technology.

Chapter 3

Environment Setup

3.1 Introduction

In the previous chapters we discussed about different findings related areas and defined the research problem. This chapter explains how to use appropriate tools and machine learning tools for data preprocessing and implementing relevant model for prediction.

3.2 Machine Learning Tools

There are lots of machine learning tools such as Weka, RapidMinar etc., available to work with data. For this research I used Weka machine learning tool for predictions. Weka tool has visualization and collection of tools for predictive modeling and data analysis. This software has its own functions that are suitable for my research.

- Available GNU General Public License.
- Can run on any modern computing platform and its implemented in the Java programming language.
- Provides graphical user interfaces.
- Can perform modeling techniques and data preprocessing.

For data preprocessing, weka tool is provide several standards. Then can use different data mining techniques like clustering, classification, regression, etc. In data point it has fixed number of attributes and there are different techniques are based on the different assumptions [10].

3.2.1 Packages and Tools

There are several packages that have been used throughout this research in order to deal with prediction and machine learning tasks. Data set is extracting from the XAMPP and data preprocessing is done from the visual studio. Furthermore, for the prediction used WEKA machine learning tool.

3.3 Summary

This section represents the technologies and tools used. There are many other Technologies that can be used for do the same. In next chapter, it describes the research methodology of the research.

Chapter 4

Research Methodology

4.1 Introduction

Third chapter described about the techniques that used in this study. This chapter present the approach that applied with problem of 5 main area. Here it presents the approach by highlighting the business understanding, data understanding, data preparation, modeling and evaluation.

4.2 Main Phases

This methodology involves six phases, those are shown in Figure 1. First need to understand the business environment, then data understanding. In order to develop models, need to preparing data accordingly. Then after can evaluate the algorithm by performance wise and accuracy wise and etc.

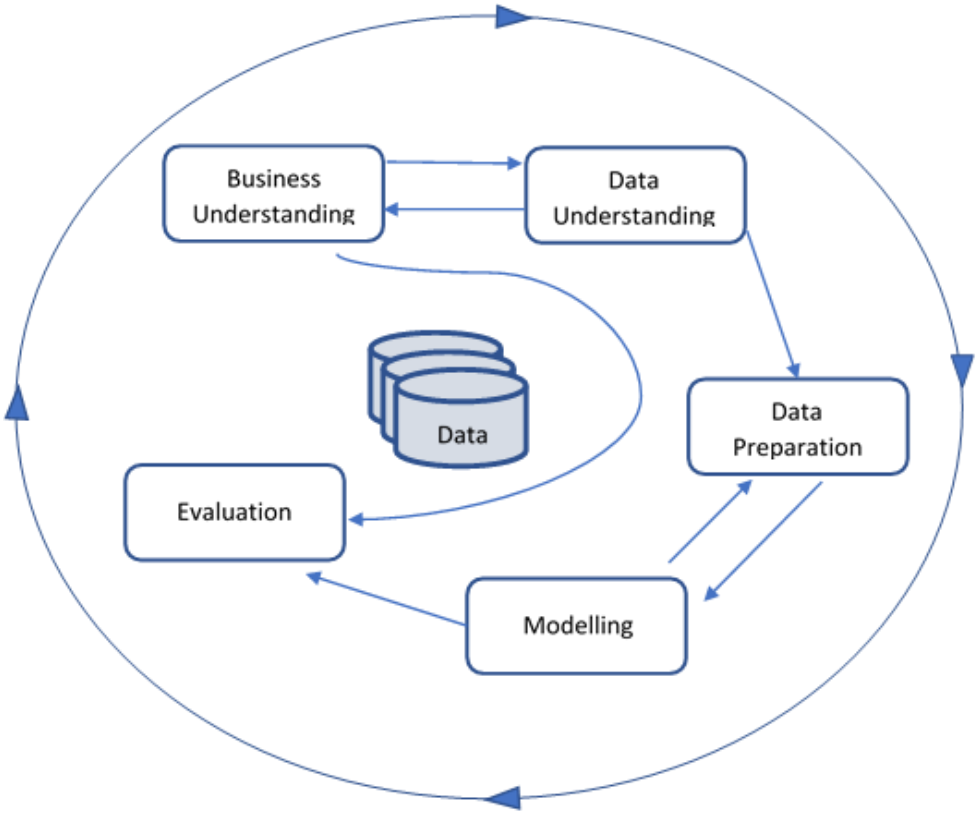

Figure 1: Main Phases

In this study, possible test areas are dependent on available data and detailed business objectives are identified after the data study. As a result, the phases must be carried out in parallel with understanding and preparing data values.

4.3 Business understanding

On the perspective of business environment focus is to understand research objectives and requirements to translate that in to valuable plan to achieve the research objectives. Literature reviews knowledge relevant work was identified form the research gap and potential difficulties. The research focused on data and assigned appropriate mining algorithms to each target.

4.4 Data understanding

Relevant data are collected from the company database and then needs to understand, test and verify. To play with the data, quality of data need to understanding clearly. To obtain information about the data need to detect subsets of data interest. A total of 33676 data was there in the database. There are set of different tables are there. In the database activated table is shown in Figure 2.

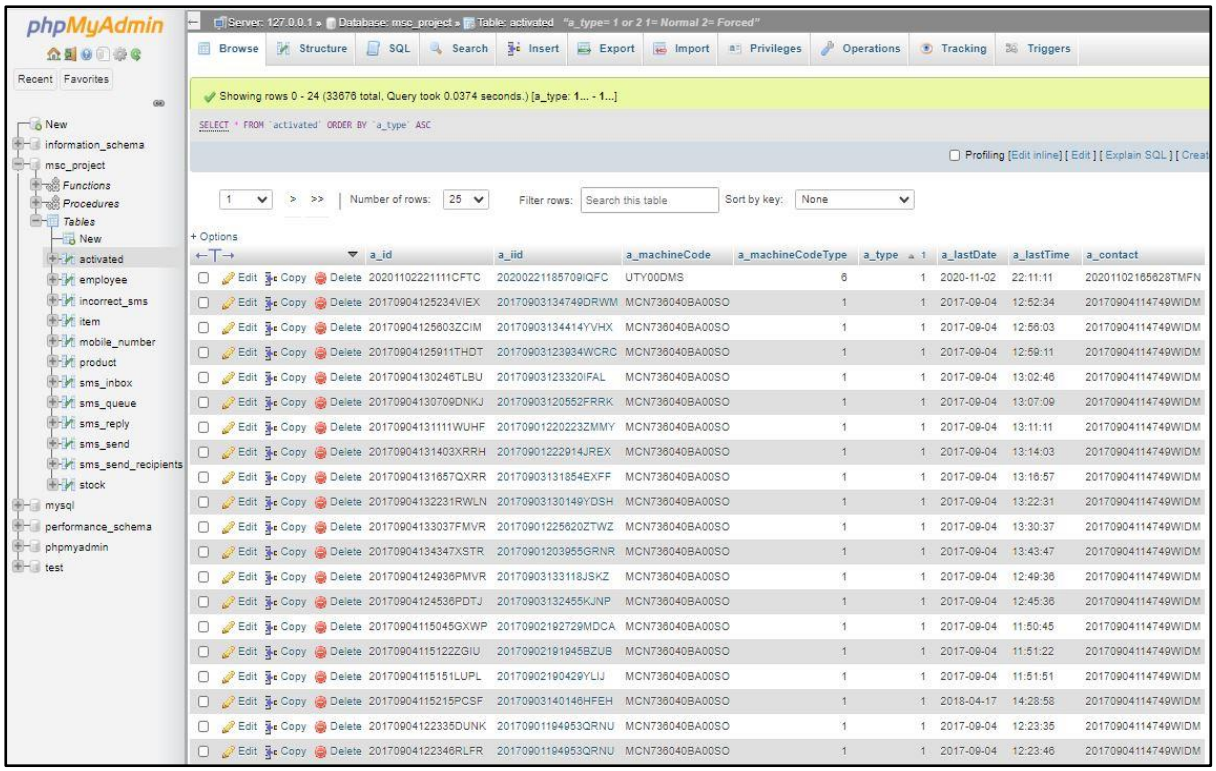

Figure 2: Database Activated Table

In preprocessing phase some data problem can be identified. Mainly data need to prepare for compatible to import to WEKA application.

4.5 Data Preparation, Modeling and Evaluate Algorithms

Preparation of data includes all the activities necessary to create the dataset for purpose of modeling. There are different manners to perform different tasks in multiple times. New problems and challenges are tended to reveal from the different data sets. When data is provided and it is very important to select correct data mining techniques and algorithms to expected to produce the best results.

Need to display the dependencies between different subsets of attributes. Some variables are used for analysis and modeling process. To identify appropriate attributes can use to predict and get target results.

4.5.1 Area 1

In order to identify the possibilities of a customer buying related other products after buying a particular product, first need to prepare the dataset. According to the company environment it has grade 1 to grade 13 school syllabuses. So here if a customer is buying a single product from the particular grade then there is a possibility if buying another product in the same grade, in order to identifying the buying behavior, it is more suitable to predict in grade wise.

In order to import the data set into the WEKA application dataset need to prepare. For that I have wrote a separate program to create the arff file. Activated table is used in order to predict the possibility of area 1. To identify the unique customers database, it has customers telephone numbers. To identify the unique product, it has product serial numbers. To prepare the arff file there is a format to provide.

Here I selected association rule mining for prediction. Analyze data pattern from the database can identify different machine learning techniques. If-then patterns can create and can search from association rules and it has support and confidence levels to identify important relationships [11].

Grade wise need to prepare the arff file to predict. Since this is a repeating, I automate the process by writing the C# program in visual studio. The program code is shown below.

```
0 references
static void CustomerBuying_Association()
{
    try
    {
        string connectionstring = "datasource=localhost;port=3306;username=root;password=";

        MySqlConnection conn = new MySqlConnection(connectionstring);

        conn.Open();

        StreamWriter File = new StreamWriter("Association.arff");

        string productListQuery = "select p_id, p_pid from msc_project.product;";

        MySqlCommand cmd = new MySqlCommand(productListQuery, conn);
        MySqlDataReader rdr = cmd.ExecuteReader();

        List<string> productList = new List<string>();

        File.WriteLine("@relation savi");
        File.WriteLine("@attribute 'customerContact' string");

        string grade = "G11";

        string[] compulsoryArr = {"G1011SLITV1D1", "G1011SLITV1D2", "G1011SLITV1D3" };

        List<string> compulsoryList = new List<string>(compulsoryArr);

        List<string> selectedProductSQLList = new List<string>();

        while (rdr.Read())
        {
            string prod = rdr["p_pid"].ToString();

            string gradeNextChar = prod.Substring((grade.Length), 1);

            if ((!prod.StartsWith(grade)
                || int.TryParse(gradeNextChar, out _))
                && !compulsoryList.Contains(prod))
```

Figure 3: Area 1-Program code part 1

```
            if ((!prod.StartsWith(grade)
                || int.TryParse(gradeNextChar, out _))
                && !compulsoryList.Contains(prod))
            {
                continue;
            }

            selectedProductSQLList.Add(prod);

            string mainProd = prod.Substring(0, prod.Length - 1);
            if (!productList.Contains(mainProd))
            {
                productList.Add(mainProd);
            }

        }
        rdr.Close();

        for (int i = 0; i < productList.Count; i++)
        {
            File.WriteLine("@attribute " + "'" + productList[i] + "'" + "  { t}");
        }

        File.WriteLine("");
        File.WriteLine("@data");

        string activatedCustomers = "SELECT DISTINCT(mobile_number.m_number), mobile_number.m_id " +
            "FROM msc_project.activated INNER JOIN msc_project.mobile_number ON activated.a_contact=mobile_number.m_id " +
                "INNER JOIN msc_project.item ON activated.a_iid = item.i_id " +
                "INNER JOIN msc_project.stock ON item.i_sid = stock.s_id " +
                "INNER JOIN msc_project.product ON stock.s_pid = product.p_id " +
            "WHERE product.p_pid IN ('" + string.Join("','", selectedProductSQLList) + "');";

        cmd = new MySqlCommand(activatedCustomers, conn);
        rdr = cmd.ExecuteReader();

        List<(string, string)> mobileList = new List<(string, string)>();
        while (rdr.Read())
        {
            mobileList.Add((rdr["m_number"].ToString(), rdr["m_id"].ToString()));
```

Figure 4: Area 1-Program code part 2

```
    {
        mobileList.Add((rdr["m_number"].ToString(), rdr["m_id"].ToString()));
    }
    rdr.Close();

    int count = 1;
    foreach (var mobileNum in mobileList)
    {
        string customerActivtProds = "select product.p_pid " +
            "from msc_project.activated INNER JOIN msc_project.item ON activated.a_iid = item.i_id " +
            "INNER JOIN msc_project.stock ON item.i_sid = stock.s_id " +
            "INNER JOIN msc_project.product ON stock.s_pid = product.p_id " +
            "where activated.a_contact='" + mobileNum.Item2 + "'" +
            "and product.p_pid IN ('"+ string.Join("','", selectedProductSQLList) +"');";

        cmd = new MySqlCommand(customerActivtProds, conn);
        rdr = cmd.ExecuteReader();

        List<string> customerActProds = new List<string>();
        while (rdr.Read())
        {
            string actProd = rdr["p_pid"].ToString();
            string mainActProd = actProd.Substring(0, actProd.Length - 1);
            customerActProds.Add(mainActProd);
        }
        rdr.Close();

        int activatedPrdCount = 0;
        string custProductStr = "";
        for (int i = 0; i < productList.Count; i++)
        {
            string bought = "?";
            if (customerActProds.Contains(productList[i]))
            {
                bought = "t";
                activatedPrdCount++;
            }
            custProductStr += "," + bought;
        }

        if(activatedPrdCount <= 0)
```

Figure 5: Area 1-Program code part 3

```
            if(activatedPrdCount <= 0)
            {
                continue;
            }
            Console.WriteLine("Customer : " + count);

            File.Write(mobileNum.Item1);
            File.Write(custProductStr + Environment.NewLine);

            if (count == 5000)
            {
                break;
            }

            count++;
        }

        File.Close();

        Console.WriteLine("==============End==================");
        Console.ReadKey();

    }
    catch (Exception ex)
    {
        Console.WriteLine(ex.Message);
    }
}
```

Figure 6: Area 1-Program code part 4

The format is in association rule mining is as follows,

```
@relation savi
@attribute 'customerContact' string
@attribute 'G1011SLITV1D'  { t}
@attribute 'G11BUDV1D'  { t}
@attribute 'G11GEOV1D'  { t}
@attribute 'G11HISV1D'  { t}
@attribute 'G11ICTV1D'  { t}
@attribute 'G11MATT1V1D'  { t}
@attribute 'G11MATT2V1D'  { t}
@attribute 'G11MATT3V1D'  { t}
@attribute 'G11SCIT1V1D'  { t}
@attribute 'G11SCIT2V1D'  { t}
@attribute 'G11SCIT3V1D'  { t}
@attribute 'G11SINV1D'  { t}

@data
+94703259900,t,t,t,t,t,t,t,t,t,t,t,t
+94717921390,t,t,t,?,t,t,t,t,t,?,t,?
+94718646877,t,t,t,t,?,t,t,?,t,?,t,t
+94771388605,t,?,t,t,?,?,?,?,t,t,t,t
+94715337201,t,t,t,t,?,t,t,?,t,?,?,t
+94711858386,t,t,?,t,?,t,?,?,t,?,?,t
+94723717568,t,?,t,?,t,?,t,t,?,t,?,t
+94719308059,t,?,?,t,?,?,?,?,t,?,t,?
+94718338987,t,t,t,t,t,t,t,t,?,t,?,t
+94763659583,t,?,?,t,?,?,?,?,?,?,?,?
+94786339912,t,t,?,t,?,t,?,t,?,t,?,t
+94718610492,t,t,?,t,t,?,t,?,t,?,t,?
+94777073435,t,t,t,?,t,t,t,?,t,?,t,t
+94702259302,t,?,?,?,?,t,t,?,?,?,?,?
+94773180664,t,t,t,t,?,t,t,t,?,t,?,t
+94718238148,t,t,?,t,t,t,t,t,t,t,?,?
+94783647364,t,t,t,t,t,?,t,?,t,?,t,t
+94756556399,t,t,t,t,?,t,?,t,?,t,?,t
+94753772533,t,t,?,t,?,t,?,?,t,?,?,?
```

Figure 7: Area 1 arff data file

Then after that need to import the arff file into WEKA application. As the preprocessing task remove the attribute" customerContact" column.

Aproiri algorithm is selected to predict in confidence 0.65 minMetric level. Following association rules are got as the output.

The Apriori algorithm uses as a property of large set of items of any subset of a large set. The elements of an element set are also assumed to be kept in lexicographical order.

Consideration of previously run large item sets are significantly reduce the large item sets in the candidates [12].

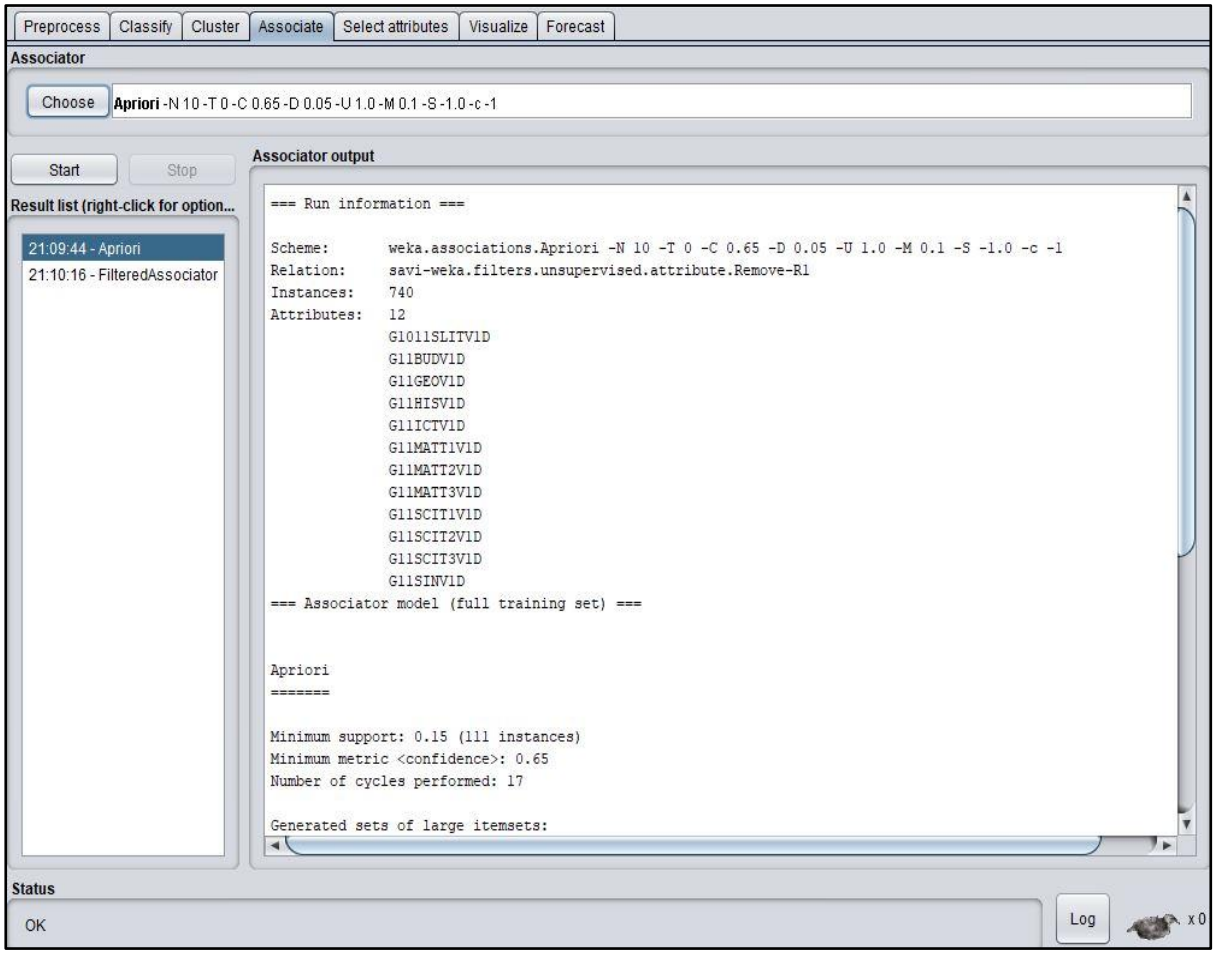

Figure 8: Apriori output part 1

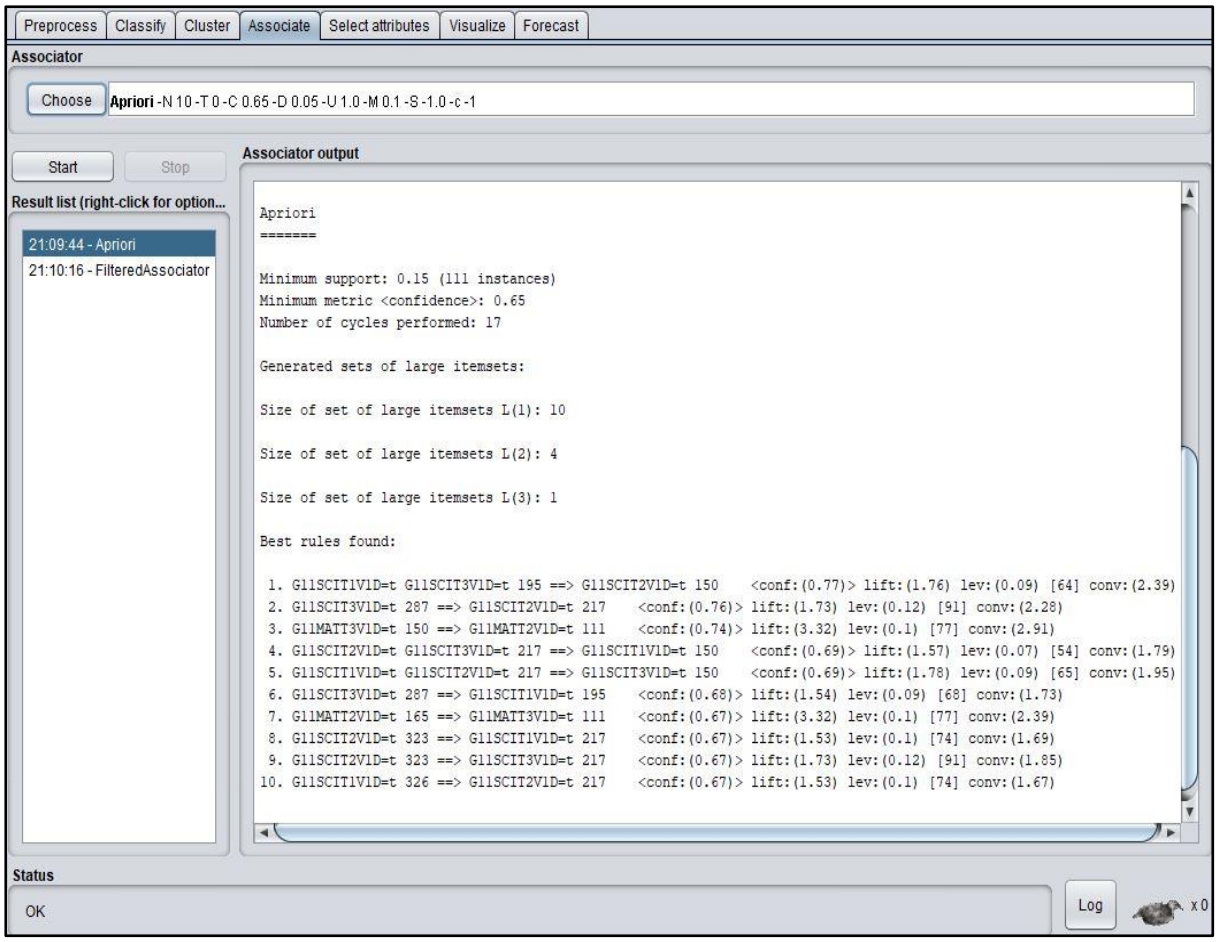

Figure 9: Apriori output part 2

The FilteredAssociator algorithm is also selected to predict with minimum confidence = 0.6. Following association rules are got as the output.

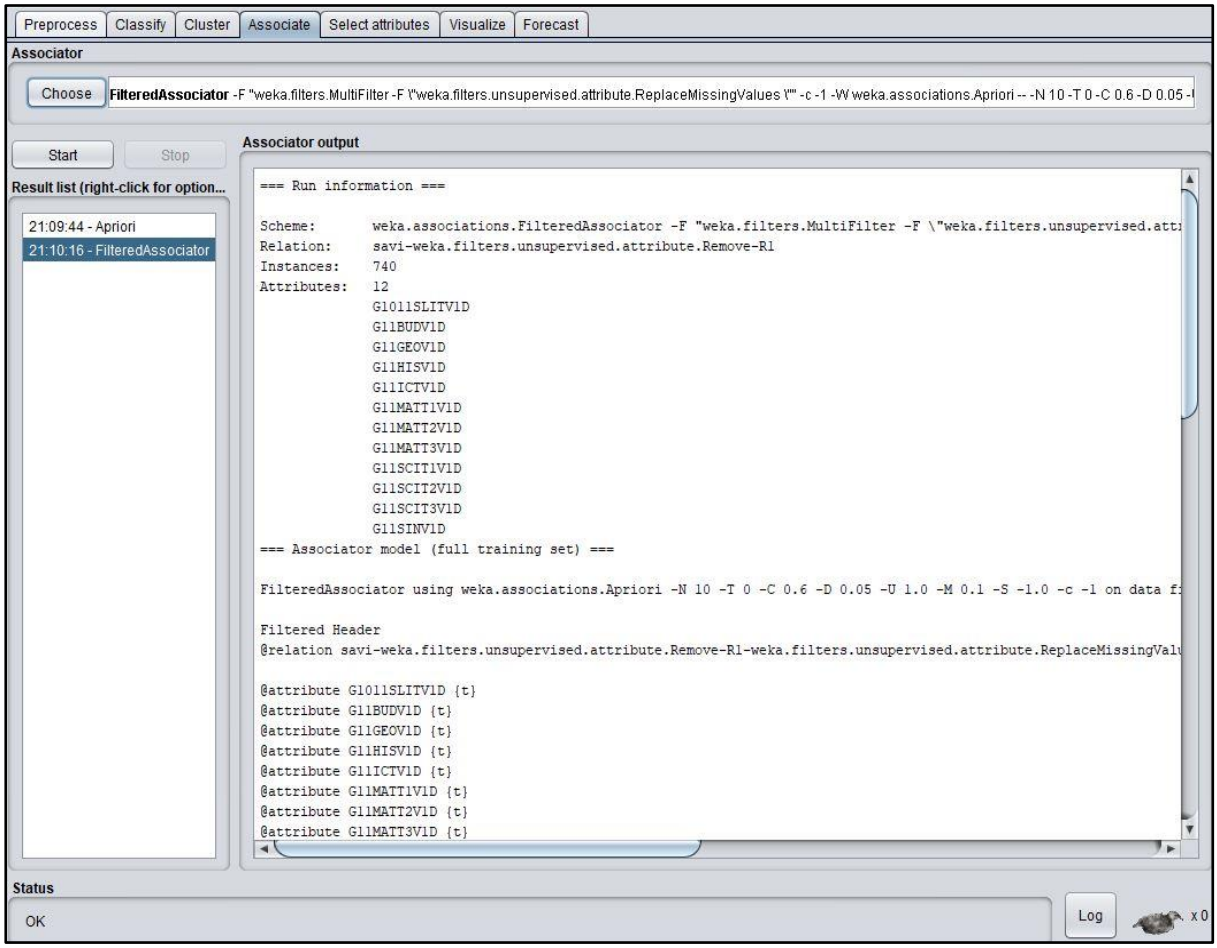

Figure 10: FilteredAssociator output part 1

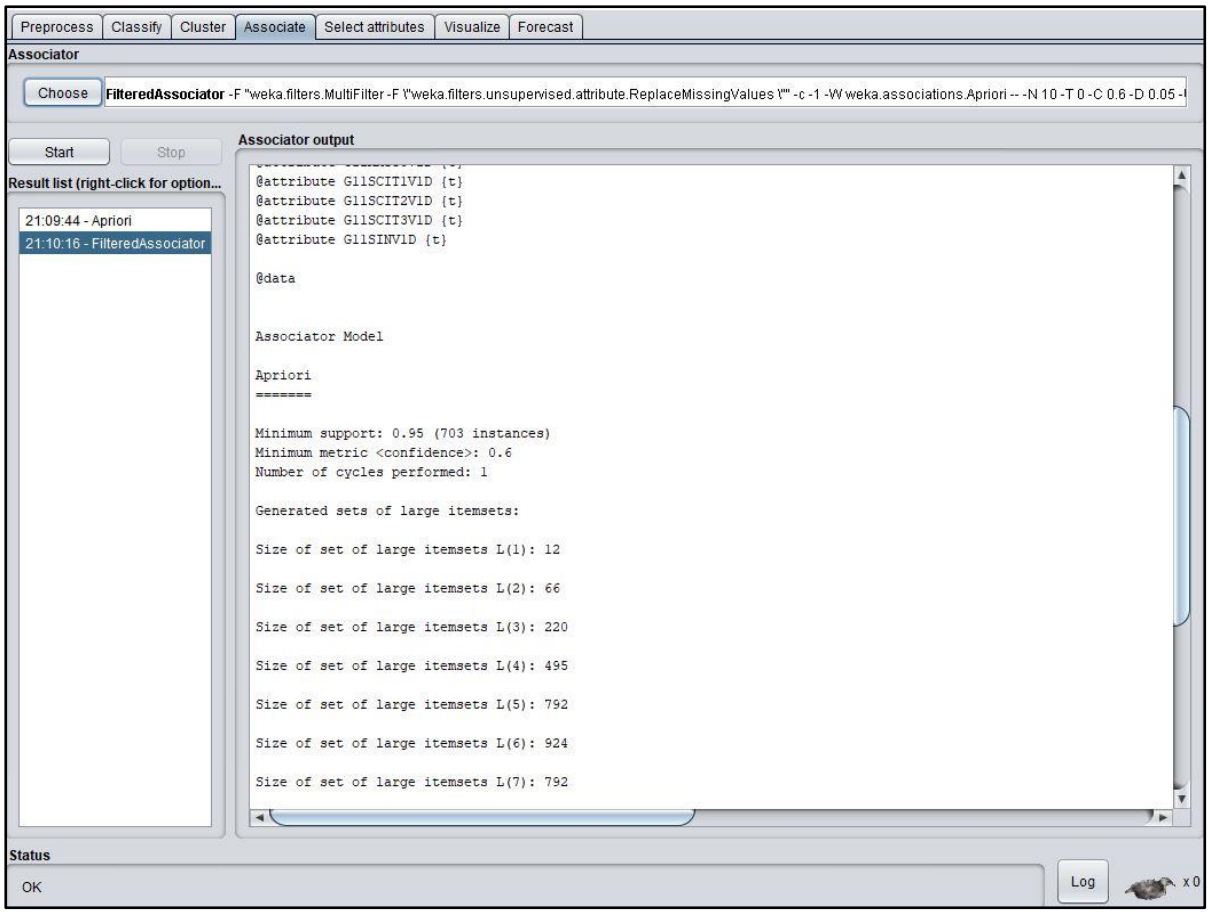

Figure 11: FilteredAssociator output part 2

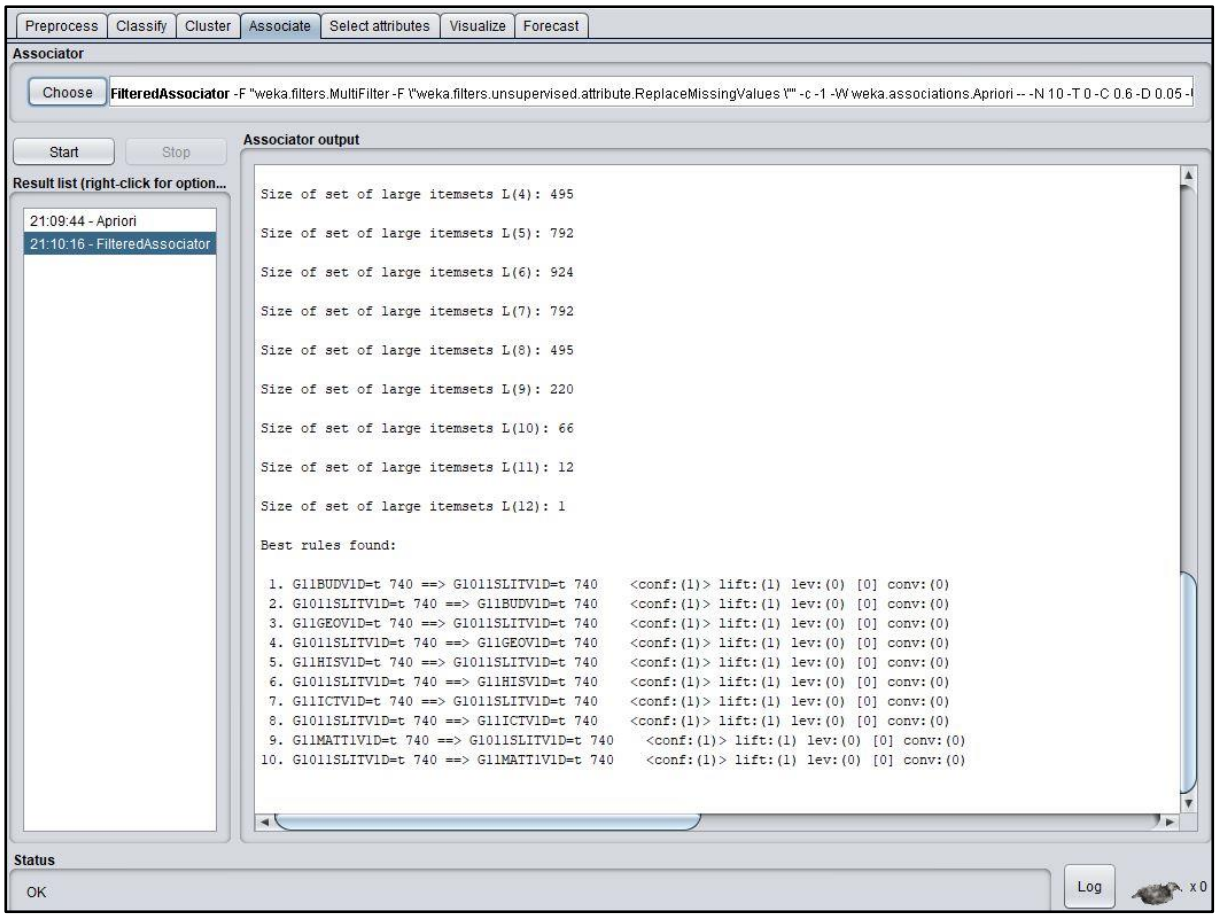

Figure 12: FilteredAssociator output part 3

Mentioned process has applied for each grade (1-13). Different rules are getting in each grade.

Evaluate Algorithms

According to above two algorithms should evaluate the algorithm by analyzing the confidence level. In Apriori algorithm got 0.65 confidence level and FilteredAssociator algorithm got 0.6 confidence level to generate the rules. Confidence level is near to 1.0 is better.

Best rules found:

```
Best rules found:

 1. G11SCIT1V1D=t G11SCIT3V1D=t 195 ==> G11SCIT2V1D=t 150    <conf:(0.77)> lift:(1.76) lev:(0.09) [64] conv:(2.39)
 2. G11SCIT3V1D=t 287 ==> G11SCIT2V1D=t 217    <conf:(0.76)> lift:(1.73) lev:(0.12) [91] conv:(2.28)
 3. G11MATT3V1D=t 150 ==> G11MATT2V1D=t 111    <conf:(0.74)> lift:(3.32) lev:(0.1) [77] conv:(2.91)
 4. G11SCIT2V1D=t G11SCIT3V1D=t 217 ==> G11SCIT1V1D=t 150    <conf:(0.69)> lift:(1.57) lev:(0.07) [54] conv:(1.79)
 5. G11SCIT1V1D=t G11SCIT2V1D=t 217 ==> G11SCIT3V1D=t 150    <conf:(0.69)> lift:(1.78) lev:(0.09) [65] conv:(1.95)
 6. G11SCIT3V1D=t 287 ==> G11SCIT1V1D=t 195    <conf:(0.68)> lift:(1.54) lev:(0.09) [68] conv:(1.73)
 7. G11MATT2V1D=t 165 ==> G11MATT3V1D=t 111    <conf:(0.67)> lift:(3.32) lev:(0.1) [77] conv:(2.39)
 8. G11SCIT2V1D=t 323 ==> G11SCIT1V1D=t 217    <conf:(0.67)> lift:(1.53) lev:(0.1) [74] conv:(1.69)
 9. G11SCIT2V1D=t 323 ==> G11SCIT3V1D=t 217    <conf:(0.67)> lift:(1.73) lev:(0.12) [91] conv:(1.85)
10. G11SCIT1V1D=t 326 ==> G11SCIT2V1D=t 217    <conf:(0.67)> lift:(1.53) lev:(0.1) [74] conv:(1.67)
```

Figure 13: Selected Best Rules from Apriori algorithm

So, the Apriori algorithm is selected for perdition. So that this is easy to make decisions for managers to identifying the market trends.

4.5.2 Area 2

Identifying customers who are going to continuing purchasing or drop from the company and how it effects for the company.

When identifying customers who are going to continuing purchasing or drop from the company and how it effects for the company, first need to prepare the dataset. According to the company environment it has grade 1 to grade 13 school syllabuses. If a customer is purchased a product or more than one product there is a possibility of purchasing new product from next year or it can be a not purchasing product from next year. As an example, Possibility of grade 10 customer is going to continuing or discontinuing purchase from grade 11.

For this prediction I selected classification data mining technique. This technique is used to obtain important and relevant information about data and metadata [13]. In

order to import the data set into the WEKA application dataset need to prepare. For that I have wrote a separate program to create the arff file. The program is shown below.

```
0 references
static void ContinuingPurchasing_Classification()
{
    try
    {
        string connectionstring = "datasource=localhost;port=3306;username=root;password=";

        MySqlConnection conn = new MySqlConnection(connectionstring);

        conn.Open();

        StreamWriter File = new StreamWriter("Classification.arff");

        string productListQuery = "select p_id, p_pid from msc_project.product;";

        MySqlCommand cmd = new MySqlCommand(productListQuery, conn);
        MySqlDataReader rdr = cmd.ExecuteReader();

        List<string> productList = new List<string>();

        File.WriteLine("@relation savi");
        File.WriteLine("@attribute 'customerContact' string");

        string grade = "G10";
        string targetGrade = "G11";

        string[] compArr = { "" };

        List<string> compulsoryList = new List<string>(compArr);

        List<string> selectedProductSQLList = new List<string>();
        List<string> targetProductSQLList = new List<string>();


        while (rdr.Read())
        {
            bool isSelectedProd = true;
            bool isTargetProd = true;
```

Figure 14: Area 2-Program code part 1

```
while (rdr.Read())
{
    bool isSelectedProd = true;
    bool isTargetProd = true;

    string prod = rdr["p_pid"].ToString();
    string gradeNextChar = prod.Substring((grade.Length), 1);

    if (!prod.StartsWith(grade)
        || int.TryParse(gradeNextChar, out _))
    {
        if(!compulsoryList.Contains(prod))
        {
            isSelectedProd = false;
        }
    }

    if (!prod.StartsWith(targetGrade)
        || int.TryParse(gradeNextChar, out _))
    {
        isTargetProd = false;
    }

    if (isSelectedProd)
    {
        selectedProductSQLList.Add(prod);

        string mainProd = prod.Substring(0, prod.Length - 1);
        if (!productList.Contains(mainProd))
        {
            productList.Add(mainProd);
        }
    }

    if (isTargetProd)
    {
        targetProductSQLList.Add(prod);
    }

}
```

Figure 15: Area 2-Program code part 2

```
    }
    rdr.Close();

    for (int i = 0; i < productList.Count; i++)
    {
        File.WriteLine("@attribute " + "'" + productList[i] + "'" + "  {y,n}");
    }

    File.WriteLine("@attribute purchaseNextYear {y,n}");

    File.WriteLine("");
    File.WriteLine("@data");

    string activatedCustomers = "SELECT DISTINCT(mobile_number.m_number), mobile_number.m_id " +
        "FROM msc_project.activated INNER JOIN msc_project.mobile_number ON activated.a_contact=mobile_number.m_id " +
            "INNER JOIN msc_project.item ON activated.a_iid = item.i_id " +
            "INNER JOIN msc_project.stock ON item.i_sid = stock.s_id " +
            "INNER JOIN msc_project.product ON stock.s_pid = product.p_id " +
        "WHERE product.p_pid IN ('" + string.Join("','", selectedProductSQLList) + "');";

    cmd = new MySqlCommand(activatedCustomers, conn);
    rdr = cmd.ExecuteReader();

    List<(string, string)> mobileList = new List<(string, string)>();
    while (rdr.Read())
    {
        mobileList.Add((rdr["m_number"].ToString(), rdr["m_id"].ToString()));
    }
    rdr.Close();

    int count = 1;
    foreach (var mobileNum in mobileList)
    {
        string customerActivtProds = "select product.p_pid " +
            "from msc_project.activated INNER JOIN msc_project.item ON activated.a_iid = item.i_id " +
            "INNER JOIN msc_project.stock ON item.i_sid = stock.s_id " +
            "INNER JOIN msc_project.product ON stock.s_pid = product.p_id " +
            "where activated.a_contact='" + mobileNum.Item2 + "'" +
            "and product.p_pid IN ('" + string.Join("','", selectedProductSQLList) + "');";
```

Figure 16: Area 2-Program code part 3

```
        cmd = new MySqlCommand(customerActivtProds, conn);
        rdr = cmd.ExecuteReader();

        List<string> customerActProds = new List<string>();
        while (rdr.Read())
        {
            string actProd = rdr["p_pid"].ToString();
            string mainActProd = actProd.Substring(0, actProd.Length - 1);
            customerActProds.Add(mainActProd);
        }
        rdr.Close();

        int activatedPrdCount = 0;
        string custProductStr = "";
        for (int i = 0; i < productList.Count; i++)
        {
            string bought = "n";
            if (customerActProds.Contains(productList[i]))
            {
                bought = "y";
                activatedPrdCount++;
            }
            custProductStr += "," + bought;
        }

        if (activatedPrdCount <= 1)
        {
            continue;
        }
        Console.WriteLine("Customer : " + count);

        string customerActivtTargetProds = "select product.p_pid " +
            "from msc_project.activated INNER JOIN msc_project.item ON activated.a_iid = item.i_id " +
            "INNER JOIN msc_project.stock ON item.i_sid = stock.s_id " +
            "INNER JOIN msc_project.product ON stock.s_pid = product.p_id " +
            "where activated.a_contact='" + mobileNum.Item2 + "'" +
            "and product.p_pid IN ('" + string.Join("','", targetProductSQLList) + "');";

        cmd = new MySqlCommand(customerActivtTargetProds, conn);
        rdr = cmd.ExecuteReader();
```

Figure 17: Area 2-Program code part 4

```
            cmd = new MySqlCommand(customerActivtTargetProds, conn);
            rdr = cmd.ExecuteReader();

            String purchasedTarget = "n";
            if (rdr.HasRows)
            {
                purchasedTarget = "y";
            }
            rdr.Close();
            custProductStr += "," + purchasedTarget;

            File.Write(mobileNum.Item1);
            File.Write(custProductStr + Environment.NewLine);


            if (count == 5000)
            {
                break;
            }

            count++;
        }

        File.Close();

        Console.WriteLine("==============End==================");
        Console.ReadKey();

    }
    catch (Exception ex)
    {
        Console.WriteLine(ex.Message);
    }
}
```

Figure 18: Area 2-Program code part 5

Activated table is used in order to predict the possibility of area 2. To identify the unique customers database, it has customers telephone numbers. To identify the unique product, it has product serial numbers. In order to prepare the arff file there is a format to provide. The format is in classification is as follows,

```
@relation savi
@attribute 'customerContact' string
@attribute 'G10BUDV1D'  {y,n}
@attribute 'G10GEOV1D'  {y,n}
@attribute 'G10HISV1D'  {y,n}
@attribute 'G10ICTV1D'  {y,n}
@attribute 'G10MATT1V1D'  {y,n}
@attribute 'G10MATT2V1D'  {y,n}
@attribute 'G10MATT3V1D'  {y,n}
@attribute 'G10SCIT1V1D'  {y,n}
@attribute 'G10SCIT2V1D'  {y,n}
@attribute 'G10SCIT3V1D'  {y,n}
@attribute 'G10SINV1D'  {y,n}
@attribute purchaseNextYear {y,n}

@data
+94703259900,y,y,y,y,y,y,y,y,y,y,y,y
+94716874245,y,y,y,n,y,y,y,y,y,y,n,y
+94718666790,y,y,y,y,n,y,y,n,y,n,y,y
+94719356151,y,y,y,y,y,y,n,y,y,y,y,y
+94779334004,y,y,y,y,n,y,y,n,y,n,y,y
+94776323826,y,y,y,y,y,y,n,y,y,y,y,y
+94715799280,y,y,y,y,y,y,y,y,y,n,n,y,n
+94771388605,y,y,y,n,y,y,n,y,y,y,y,y
+94773444611,y,y,y,y,y,n,y,n,y,n,y,y
+94773981378,y,y,y,y,y,y,y,n,y,y,n,y
+94719308059,y,n,y,n,y,y,y,y,n,y,y,y
+94778982617,y,y,n,y,n,y,n,y,y,n,y,n
+94714933792,y,y,y,y,y,n,y,y,y,n,y,n
+94714710096,y,n,y,n,y,y,y,y,n,y,y,y
+94771510600,y,y,y,y,n,y,n,y,y,n,y,y
+94772786274,y,y,y,n,y,n,y,y,y,y,y,n
+94717897275,y,y,n,y,n,y,y,n,y,n,y,y
+94766404365,y,y,y,y,y,y,y,y,y,y,n,y
+94719883360,y,y,y,y,y,y,y,n,y,y,y,n
```

Figure 19: Area 2 arff data file

Grade wise need to prepare the arff file to predict. Since this is a repeating, I automate the process by writing the C# program in visual studio.

Then after that need to import the arff file into WEKA application and done the preprocessing. As the preprocessing task remove the attribute" customerContact" column.

Bayes.NaiveBayes algorithm is selected to predict. Following classification results are got as the output.

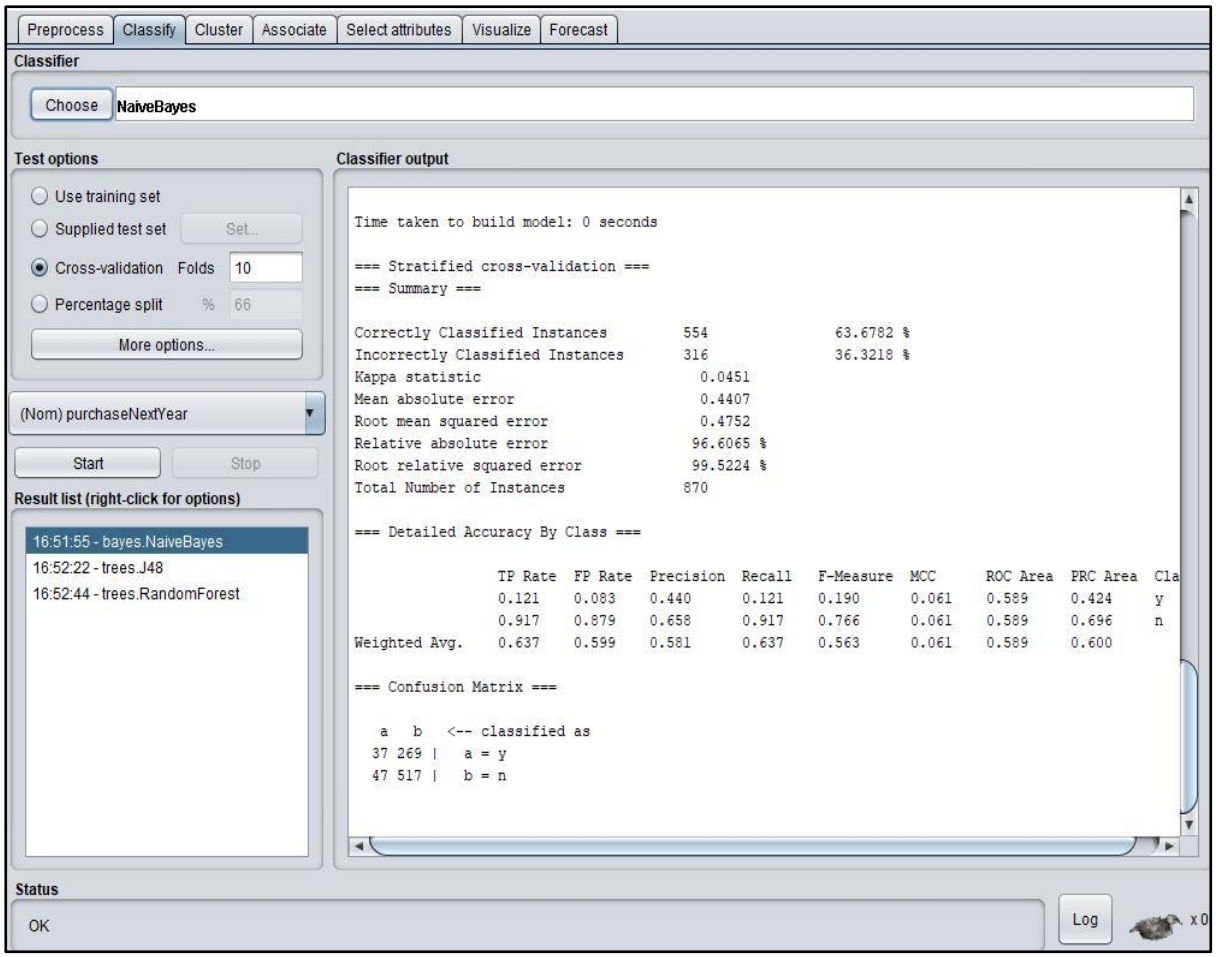

Figure 20: Bayes.NaiveBayes output

Trees J48 algorithm is selected to predict. Following classification results are got as the output.

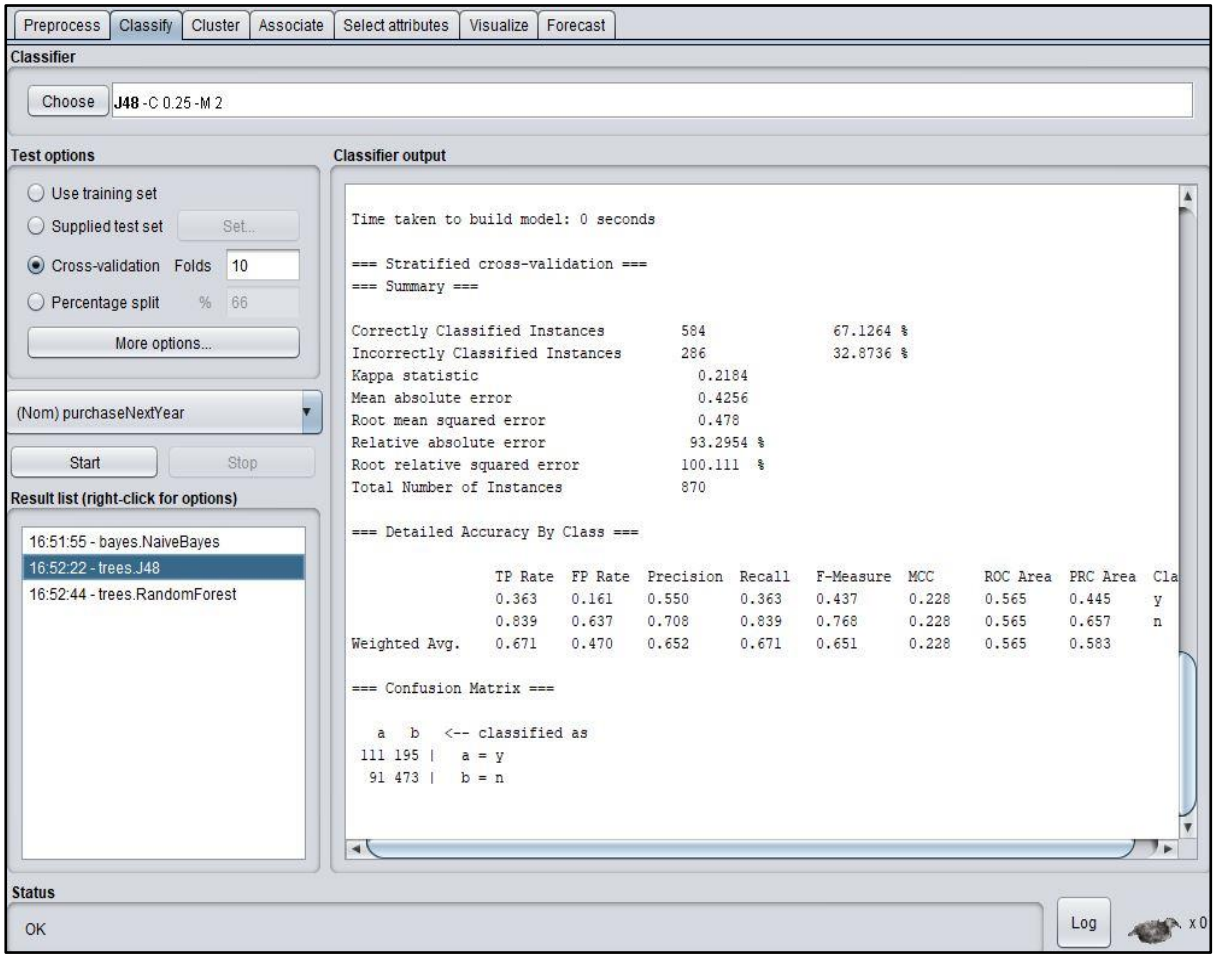

Figure 21: Trees J48 output

Tree visualizer is as follows.

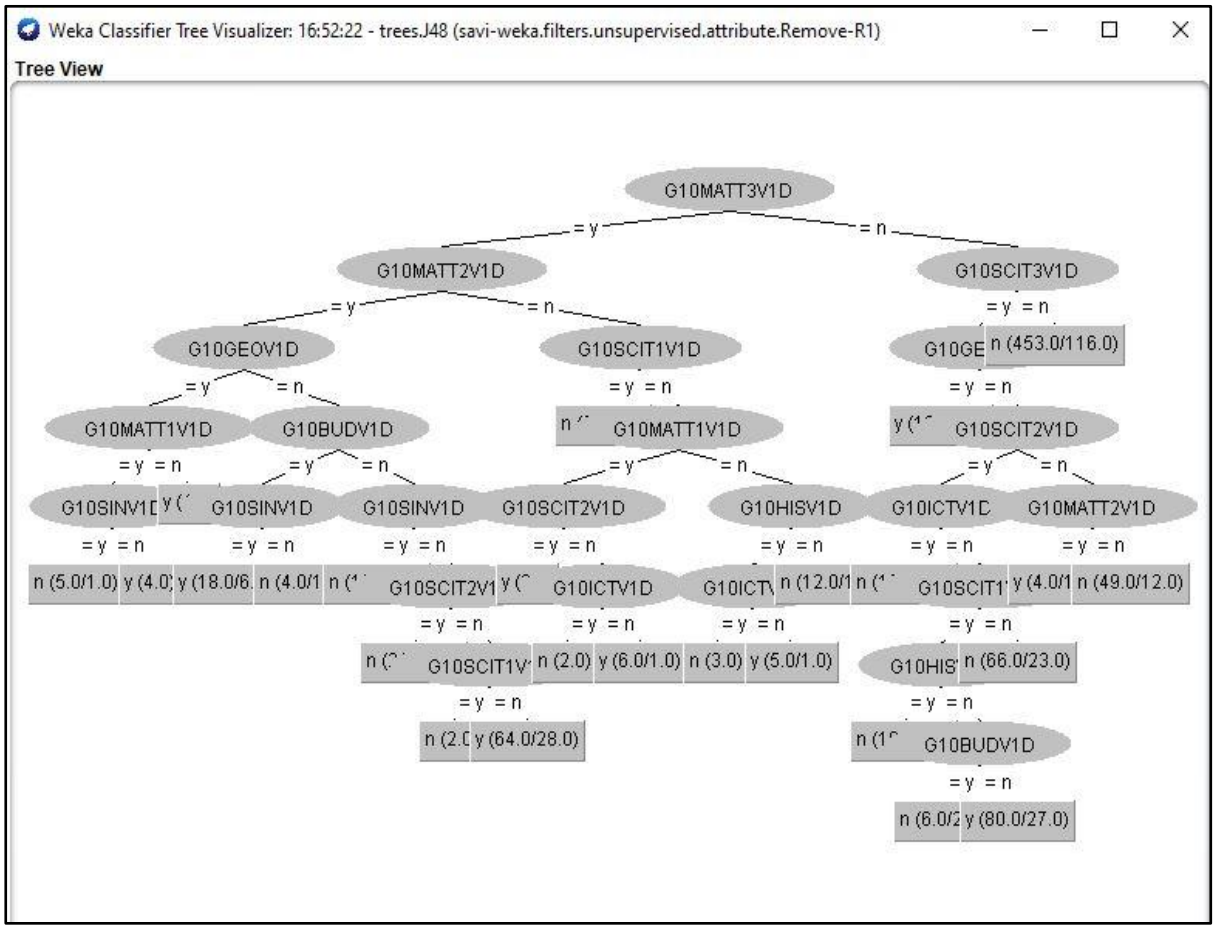

Figure 22: Trees J48 Visualizer output

Trees.RandomForest algorithm is selected to predict. Following classification results are got as the output.

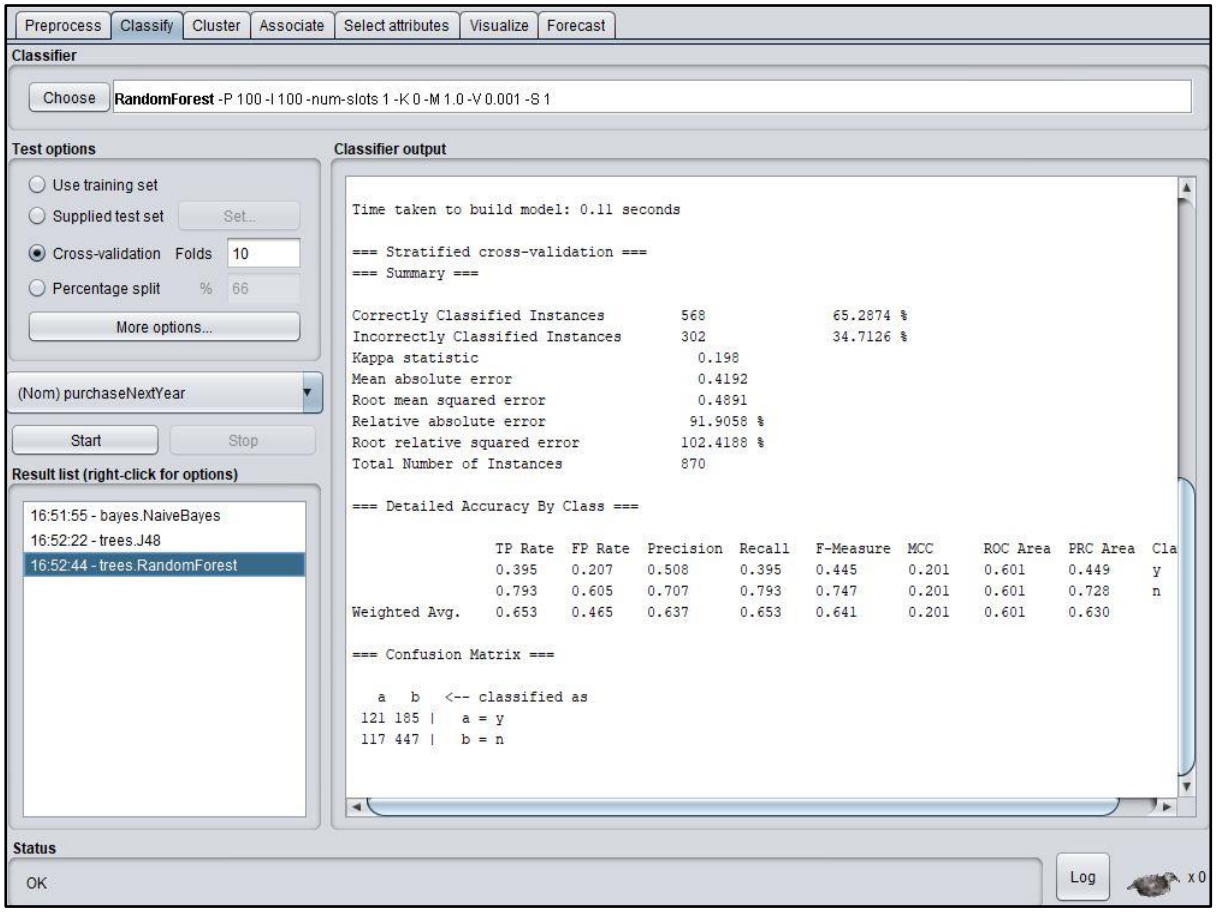

Figure 23: Trees.RandomForest output

Mentioned process has applied for each grade. Different results are getting in prediction.

Evaluate Algorithms

Among above algorithms need to select the best algorithm. By referring to the Multiple ROC (receiver operating characteristic) Curve performance wise can evaluate. All classification thresholds are shown in performance of a classification model in the graph of ROC curves. The curve plots two parameters, those are “True Positive Rate (TPR)” and “False Positive Rate (FPR)” [14].

$$TPR = TP / (TP+ FN)$$

$$FPR = FP / (FP + TN)$$

The knowledge flow is as follows,

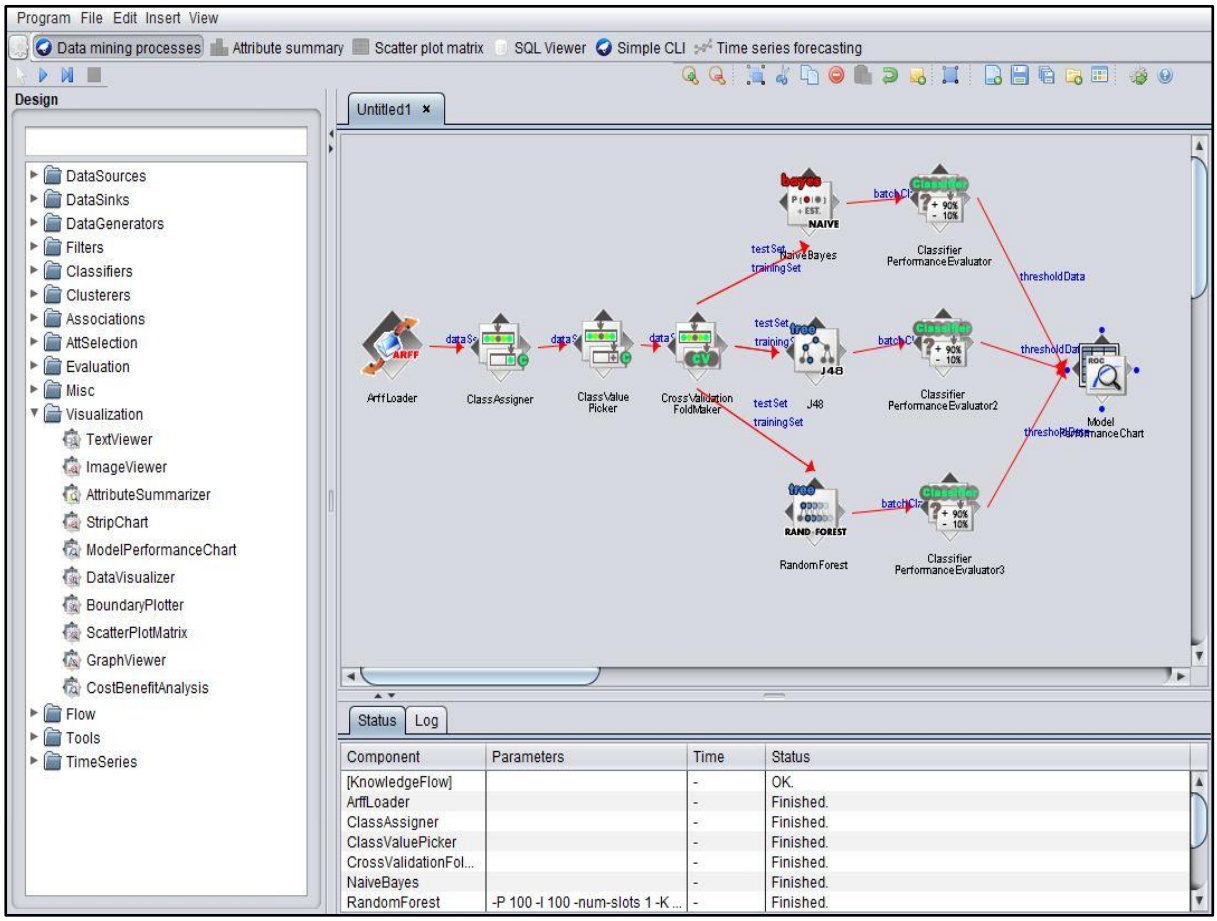

Figure 24: knowledge flow

Multiple ROC Curves with model performance chart is as follows,

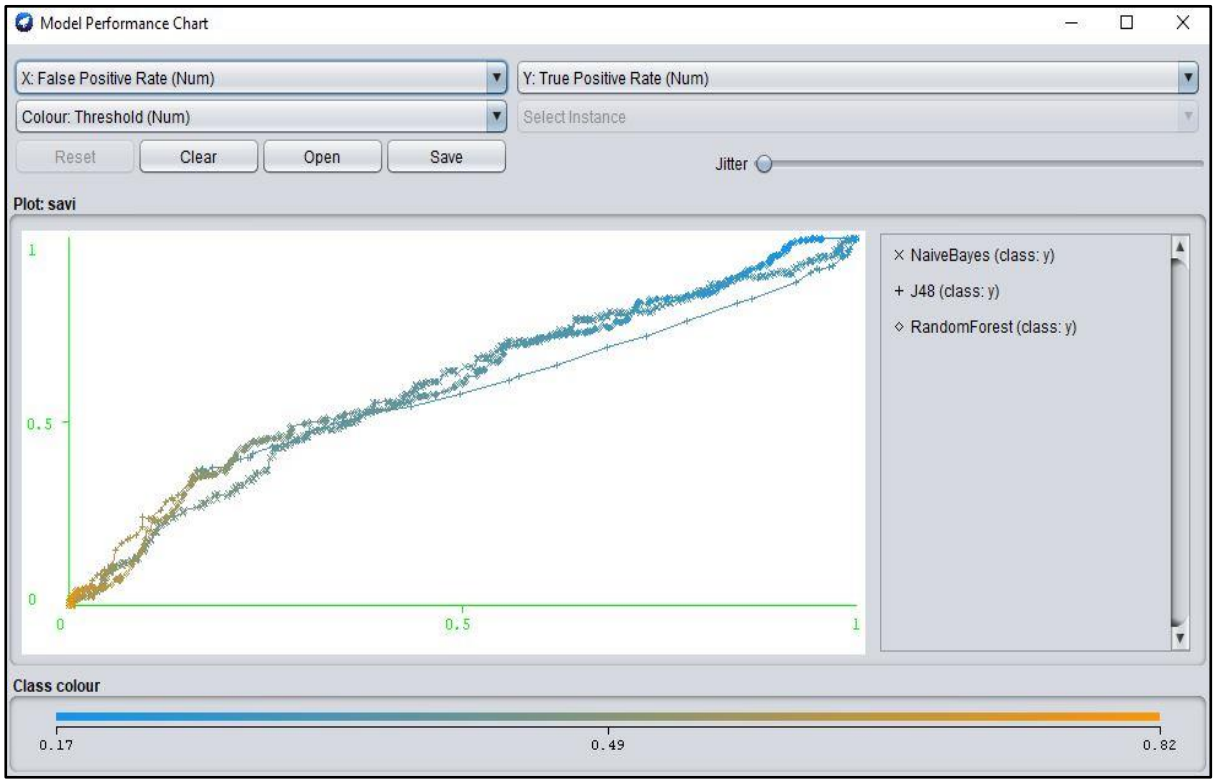

Figure 25: Multiple ROC Curves

There is another method to evaluate the above 3 algorithms. By referring to the F-Measure values provided by each algorithm.

Precision and recall measurements are offed from F-measure and its gives same weight. [15].

F-Measure = (2 * Precision * Recall) / (Precision + Recall)

Three algorithm results are as follows,

Bayes.NaiveBayes algorithm = 0.563

Trees J48 algorithm = 0.651

Trees.RandomForest algorithm = 0.641

The perfect F-measure is 1.0 and worst F-measure is 0.0 [15]. So as the conclusion Trees J48 algorithm and Trees.RandomForest algorithm values are got closer to 1.0 and ROC curve is also closer to 1. So as accuracy wise and performance wise Trees J48 and

Trees.RandomForest algorithms are better for decision making process. So that this is easy to identify how this effect for the company.

4.5.3 Area 3

In order to identify the connection between customer purchasing methods, first need to prepare the dataset. According to the company environment it has three main methods of releasing products. Those are releasing for Exhibition, Bookshops and Internet Download(online) method. Other than to these three main methods other few mechanisms are categorized to Other option.

Here I selected association rule mining for prediction. If-then patterns are created and search in association rule mining. Most important relationship criteria are support and confidence [16].

In order to import the data set into the WEKA application dataset need to prepare. For that I have wrote a separate program to create the arff file. The program code is shown below.

```
1 reference
static void PurchaseMethodRelationship_Association()
{
    try
    {
        string connectionstring = "datasource=localhost;port=3306;username=root;password=";

        MySqlConnection conn = new MySqlConnection(connectionstring);

        conn.Open();

        StreamWriter File = new StreamWriter("PurchMethAssociation.arff");

        string customerListQuery = "SELECT DISTINCT(mobile_number.m_number), mobile_number.m_id " +
            "FROM msc_project.activated INNER JOIN msc_project.mobile_number ON activated.a_contact=mobile_number.m_id;";

        MySqlCommand cmd = new MySqlCommand(customerListQuery, conn);
        MySqlDataReader rdr = cmd.ExecuteReader();

        File.WriteLine("@relation savi");
        File.WriteLine("@attribute 'customerContact' string");
        File.WriteLine("@attribute 'Exhibition' { t}");
        File.WriteLine("@attribute 'BookShop' { t}");
        File.WriteLine("@attribute 'Online' { t}");
        File.WriteLine("@attribute 'Other' { t}");

        File.WriteLine("");
        File.WriteLine("@data");

        string category_Exhibition = "stock.s_description LIKE '%Book Fair%' OR stock.s_description LIKE '%Exhibition%' " +
            "OR stock.s_description LIKE '%CIBF%'";
        string category_Bookshop = "stock.s_description LIKE '%Bookshop%' OR stock.s_description LIKE '%Book Shop%'";
        string category_Online = "stock.s_description LIKE '%Metropolitan%' OR stock.s_description LIKE '%+94%'" +
            "OR stock.s_description LIKE '%Office%' OR stock.s_description LIKE '%V 1%'" +
            "OR stock.s_description LIKE '%V 2%' OR stock.s_description LIKE '%V 3%'" +
            "OR stock.s_description LIKE '%V4%' OR stock.s_description LIKE '%computer%'" +
            "OR stock.s_description LIKE '%Abans%' OR stock.s_description LIKE '%Download%'" +
            "OR stock.s_description LIKE '%Daraz%' OR stock.s_description LIKE '%John Kells%'" +
            "OR stock.s_description LIKE '%School%'";
```

Figure 26: Area 3-Program code part 1

```
List<string> allCategories = new List<string>();
allCategories.Add(category_Exhibition);
allCategories.Add(category_Bookshop);
allCategories.Add(category_Online);


List<(string, string)> mobileList = new List<(string, string)>();
while (rdr.Read())
{
    mobileList.Add((rdr["m_number"].ToString(), rdr["m_id"].ToString()));
}
rdr.Close();


int count = 1;
foreach (var mobileNum in mobileList)
{
    List<string> customerPurchItems = new List<string>();
    string customerCategories = "";
    foreach (string category in allCategories)
    {
        string customerInPurchCategory = "SELECT activated.a_contact, item.i_id " +
            "FROM msc_project.activated INNER JOIN msc_project.item ON activated.a_iid = item.i_id " +
                "INNER JOIN msc_project.stock ON item.i_sid = stock.s_id " +
            "WHERE activated.a_contact = '" + mobileNum.Item2 + "' " +
                "AND (" + category + ");";

        cmd = new MySqlCommand(customerInPurchCategory, conn);
        rdr = cmd.ExecuteReader();

        if (rdr.HasRows)
        {
            customerCategories += ",t";

            while (rdr.Read())
            {
                customerPurchItems.Add(rdr["i_id"].ToString());
            }
        }
        else
```

Figure 27: Area 3-Program code part 2

```
            else
            {
                customerCategories += ",?";
            }
            rdr.Close();
        }

        // Customer other categories
        string customerInOtherCategory = "SELECT activated.a_contact " +
                "FROM msc_project.activated INNER JOIN msc_project.item ON activated.a_iid = item.i_id " +
                    "INNER JOIN msc_project.stock ON item.i_sid = stock.s_id " +
                "WHERE activated.a_contact = '" + mobileNum.Item2 + "' " +
                    "AND item.i_id NOT IN ('" + string.Join("','", customerPurchItems) + "');";

        cmd = new MySqlCommand(customerInOtherCategory, conn);
        rdr = cmd.ExecuteReader();

        if (rdr.HasRows)
        {
            customerCategories += ",t";
        }
        else
        {
            customerCategories += ",?";
        }
        rdr.Close();
        customerPurchItems.Clear();
        // End Customer other categories

        File.Write(mobileNum.Item1);
        File.Write(customerCategories + Environment.NewLine);

        if (count == 800)
        {
            break;
        }

        Console.WriteLine("Customer : " + count);
        count++;
    }
```

Figure 28: Area 3-Program code part 3

```
            if (count == 800)
            {
                break;
            }

            Console.WriteLine("Customer : " + count);
            count++;
        }

        File.Close();

        Console.WriteLine("==============End==================");
        Console.ReadKey();

    }
    catch (Exception ex)
    {
        Console.WriteLine(ex.Message);
    }
}
```

Figure 29: Area 3-Program code part 4

Activated and Stock tables are used in order to predict the possibility of area 3. To identify the unique customers database, it has customers telephone numbers. To

identify the unique product, it has product serial numbers. In order to prepare the arff file there is a format to provide. The format is in association rule mining is as follows,

```
@relation savi
@attribute 'customerContact' string
@attribute 'Exhibition' { t}
@attribute 'BookShop' { t}
@attribute 'Online' { t}
@attribute 'Other' { t}

@data
+94703259900,t,t,t,t
+94714482611,t,t,t,t
+94771696038,t,t,t,t
+94714001416,t,t,t,t
+94711363939,t,t,t,?
+94714392898,t,t,?,?
+94710807256,t,t,?,?
+94718361380,t,t,?,t
+94715474802,t,t,?,?
+94718482158,t,t,t,t
+94718135257,t,?,?,?
+94771716999,t,t,?,t
+94770385429,t,?,?,?
+94714124532,t,t,t,?
+94777224090,t,t,?,t
+94714247881,t,?,?,?
+94766056774,t,?,t,?
+94715736140,t,t,?,t
+94772974229,t,?,?,?
+94714404066,t,t,?,t
+94714302959,t,?,t,?
+94718082593,t,?,t,?
+94710749442,t,t,?,?
+94715203649,t,t,?,?
+94716918591,?,t,t,t
+94702729222,t,t,?,?
+94774674951,?,t,t,?
```

Figure 30: Area 3 arff data file

Then after that need to import the arff file into WEKA application and done the preprocessing. As the preprocessing task remove the attribute" customerContact" column.

Aproiri algorithm is selected to predict in confidence 0.66 minMetric level. Following association rules are got as the output.

Aproiri algorithm can use in large set of items in subset of a large set of items as a property [17].

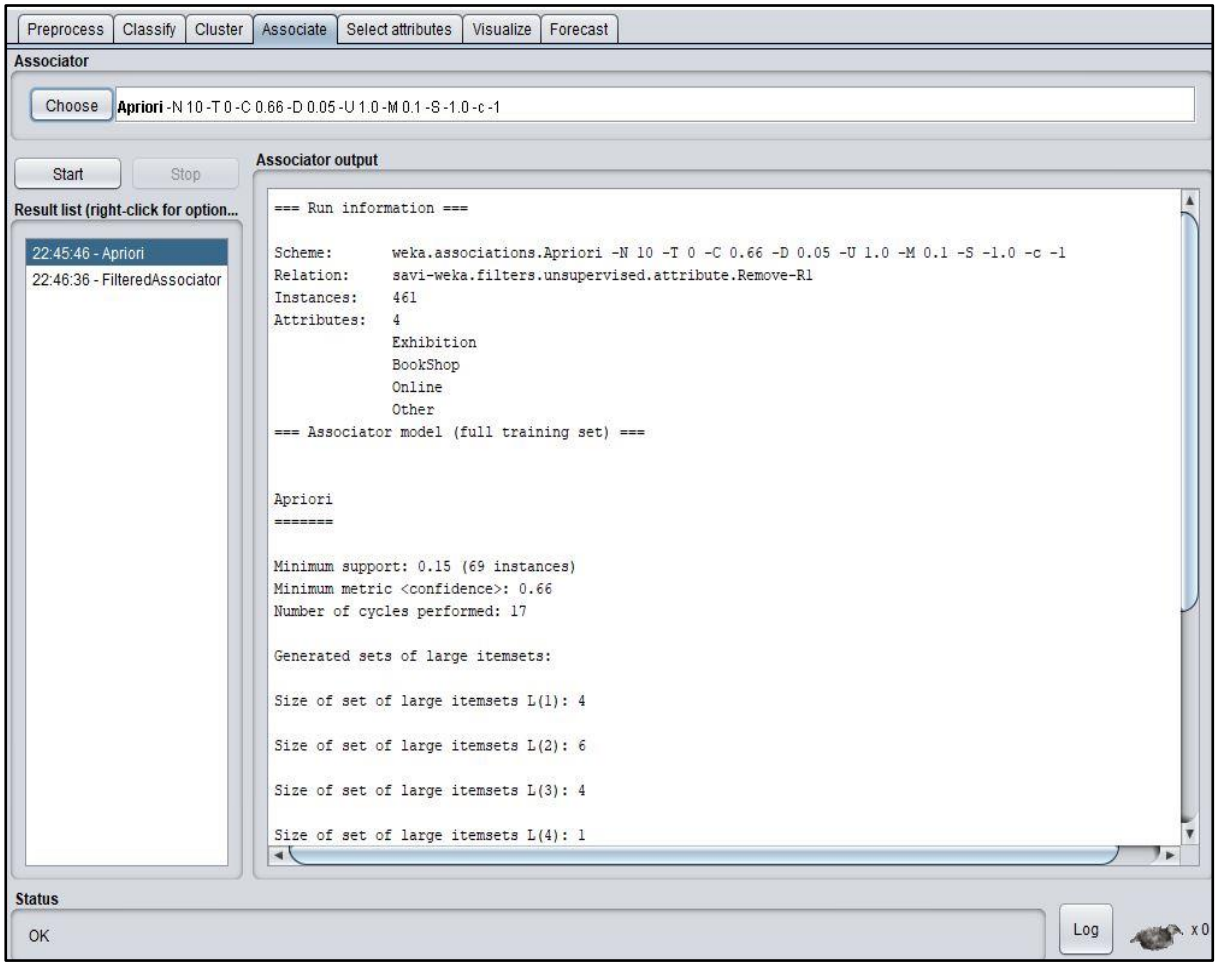

Figure 31: Apriori output part 1

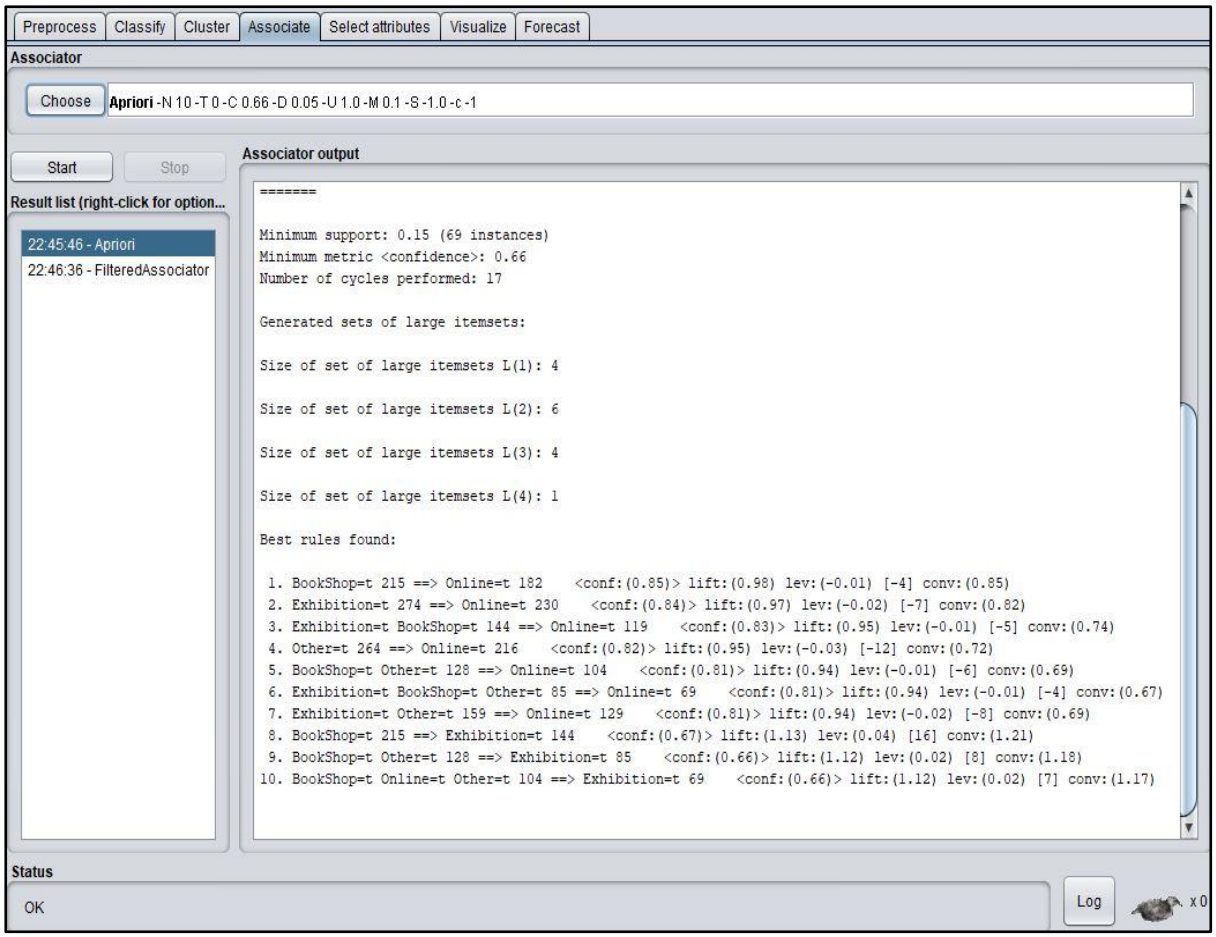

Figure 32: Apriori output part 2

The FilteredAssociator algorithm is also selected to predict with minimum confidence = 0.63. Following association rules are got as the output.

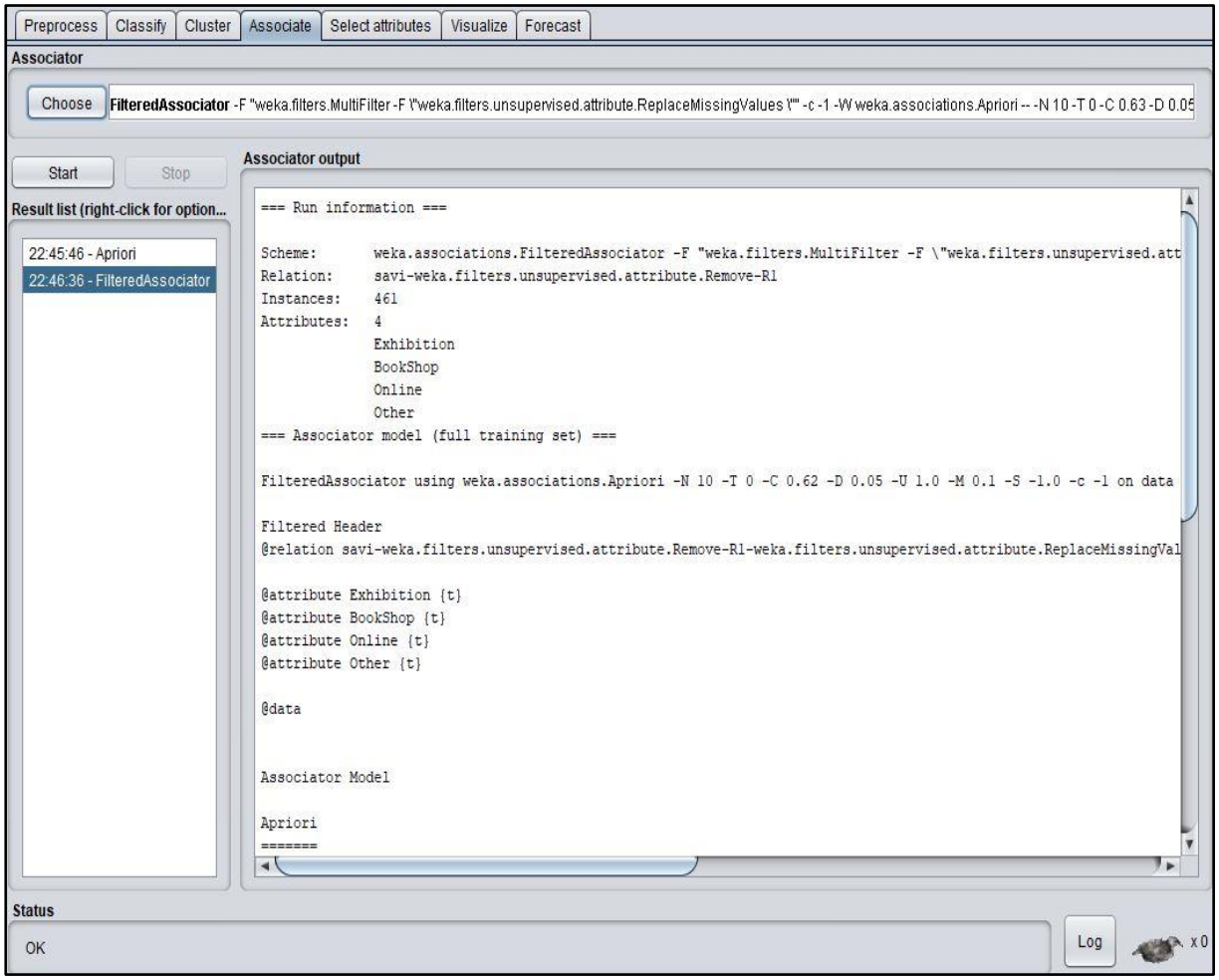

Figure 33: FilteredAssociator output part 1

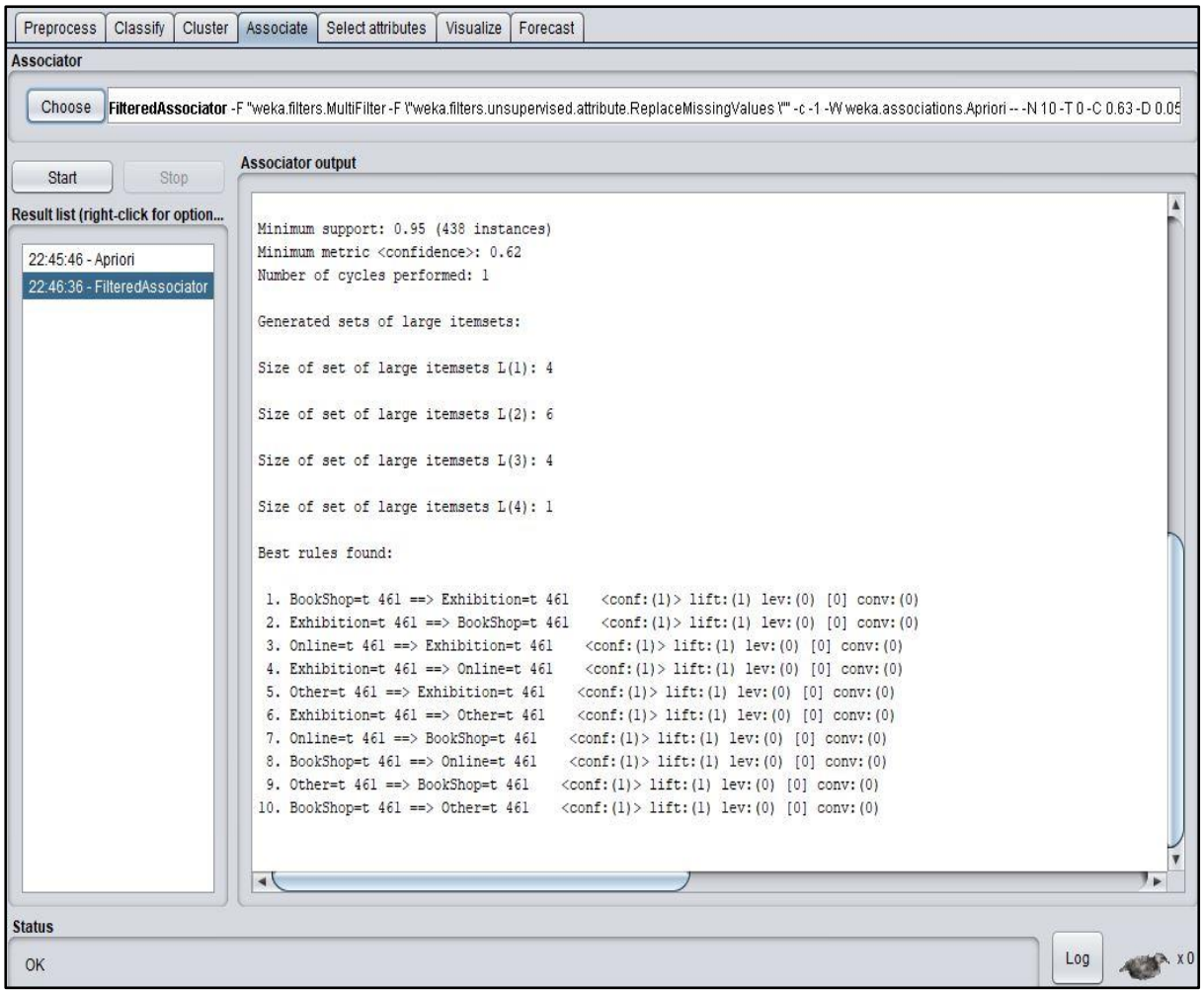

Figure 34: FilteredAssociator output part 2

Different rules are getting in each purchasing methods.

Evaluate Algorithms

According to above two algorithms should evaluate the algorithm by analyzing the confidence level. In Apriori algorithm got 0.66 confidence level and FilteredAssociator algorithm got 0.63 confidence level to generate the rules. Confidence level is near to 1.0 is better.

Best rules found:

```
Best rules found:

 1. BookShop=t 215 ==> Online=t 182    <conf:(0.85)> lift:(0.98) lev:(-0.01) [-4] conv:(0.85)
 2. Exhibition=t 274 ==> Online=t 230    <conf:(0.84)> lift:(0.97) lev:(-0.02) [-7] conv:(0.82)
 3. Exhibition=t BookShop=t 144 ==> Online=t 119    <conf:(0.83)> lift:(0.95) lev:(-0.01) [-5] conv:(0.74)
 4. Other=t 264 ==> Online=t 216    <conf:(0.82)> lift:(0.95) lev:(-0.03) [-12] conv:(0.72)
 5. BookShop=t Other=t 128 ==> Online=t 104    <conf:(0.81)> lift:(0.94) lev:(-0.01) [-6] conv:(0.69)
 6. Exhibition=t BookShop=t Other=t 85 ==> Online=t 69    <conf:(0.81)> lift:(0.94) lev:(-0.01) [-4] conv:(0.67)
 7. Exhibition=t Other=t 159 ==> Online=t 129    <conf:(0.81)> lift:(0.94) lev:(-0.02) [-8] conv:(0.69)
 8. BookShop=t 215 ==> Exhibition=t 144    <conf:(0.67)> lift:(1.13) lev:(0.04) [16] conv:(1.21)
 9. BookShop=t Other=t 128 ==> Exhibition=t 85    <conf:(0.66)> lift:(1.12) lev:(0.02) [8] conv:(1.18)
10. BookShop=t Online=t Other=t 104 ==> Exhibition=t 69    <conf:(0.66)> lift:(1.12) lev:(0.02) [7] conv:(1.17)
```

Figure 35: Selected Best Rules from Apriori algorithm

So, the Apriori algorithm is selected for perdition. So that this is easy to make decisions for managers to identifying the market trends. So that this is easy to make decisions for managers to identifying the market trends.

4.5.4 Area 4

In order to find a group of customers to promote a particular product/s, first need to prepare the dataset. There are different categories are form to identify group of customers.

Grade One (1) to Five (5) – Primary level

Grade Six (6) to Nine (9) – Secondary level

Grade Ten (10) and Eleven (11) – O/L

Grade Twelve (12) and Thirteen (13) – A/L

Other than to these four main group other few mechanisms are categorized to Other option. In order to import the data set into the WEKA application dataset need to

prepare. For that I have wrote a separate program to create the arff file. The program code is shown below.

```
static void FindGroupOfCutcomers_Clustering()
{
    try {

        string connectionstring = "datasource=localhost;port=3306;username=root;password=";

        MySqlConnection conn = new MySqlConnection(connectionstring);

        conn.Open();

        StreamWriter File = new StreamWriter("Clustring.arff");

        string productListQuery = "select p_id, p_pid from msc_project.product;";

        MySqlCommand cmd = new MySqlCommand(productListQuery, conn);
        MySqlDataReader rdr = cmd.ExecuteReader();

        List<string> productList = new List<string>();

        File.WriteLine("@relation savi");
        File.WriteLine("@attribute 'customerContact' string");

        string grade = "G";

        string[] compulsoryArr = { "Primary level", "Secondary level", "O / L", "A / L", "Other" };

        List<string> compulsoryList = new List<string>(compulsoryArr);

        List<string> selectedProductSQLList = new List<string>();

        while (rdr.Read())
        {
            string prod = rdr["p_pid"].ToString();

            string gradeNextChar = prod.Substring((grade.Length), 1);

            if ((!prod.StartsWith(grade)
                || int.TryParse(gradeNextChar, out _))
                && !compulsoryList.Contains(prod))
            {
                continue;
            }
```

Figure 36: Area 4-Program code part 1

```
            selectedProductSQLList.Add(prod);

            string mainProd = prod.Substring(0, prod.Length - 1);
            if (!productList.Contains(mainProd))
            {
                productList.Add(mainProd);
            }

        }
        rdr.Close();

        for (int i = 0; i < productList.Count; i++)
        {
            File.WriteLine("@attribute " + "'" + productList[i] + "'" + "  { t}");
        }

        File.WriteLine("");
        File.WriteLine("@data");

        string activatedCustomers = "SELECT DISTINCT(mobile_number.m_number), mobile_number.m_id " +
            "FROM msc_project.activated INNER JOIN msc_project.mobile_number ON activated.a_contact=mobile_number.m_id " +
                "INNER JOIN msc_project.item ON activated.a_iid = item.i_id " +
                "INNER JOIN msc_project.stock ON item.i_sid = stock.s_id " +
                "INNER JOIN msc_project.product ON stock.s_pid = product.p_id " +
            "WHERE product.p_pid IN ('" + string.Join("','", selectedProductSQLList) + "');";

        cmd = new MySqlCommand(activatedCustomers, conn);
        rdr = cmd.ExecuteReader();

        List<(string, string)> mobileList = new List<(string, string)>();
        while (rdr.Read())
        {
            mobileList.Add((rdr["m_number"].ToString(), rdr["m_id"].ToString()));
        }
        rdr.Close();

        int count = 1;
        foreach (var mobileNum in mobileList)
        {
```

Figure 37: Area 4-Program code part 2

```
        string customerActivtProds = "select product.p_pid " +
            "from msc_project.activated INNER JOIN msc_project.item ON activated.a_iid = item.i_id " +
            "INNER JOIN msc_project.stock ON item.i_sid = stock.s_id " +
            "INNER JOIN msc_project.product ON stock.s_pid = product.p_id " +
            "where activated.a_contact='" + mobileNum.Item2 + "'" +
            "and product.p_pid IN ('" + string.Join("','", selectedProductSQLList) + "');";

        cmd = new MySqlCommand(customerActivtProds, conn);
        rdr = cmd.ExecuteReader();
        List<string> customerActProds = new List<string>();
        while (rdr.Read())
        {
            string actProd = rdr["p_pid"].ToString();
            string mainActProd = actProd.Substring(0, actProd.Length - 1);
            customerActProds.Add(mainActProd);
        }
        rdr.Close();

        int activatedPrdCount = 0;
        string custProductStr = "";

        if (activatedPrdCount <= 0)
        {
            continue;
        }
        Console.WriteLine("Customer : " + count);

        File.Write(mobileNum.Item1);
        File.Write(custProductStr + Environment.NewLine);

        if (count == 2000)
        {
            break;
        }

        count++;
    }
    Console.WriteLine("==============End==================");
    Console.ReadKey();
}
catch (Exception ex)
{
    Console.WriteLine(ex.Message);
}
```

Figure 38: Area 4-Program code part 3

Activated table are used in order to predict the possibility of area 4. To identify the unique customers database, it has customers telephone numbers. To identify the unique product, it has product serial numbers. And then parentage value is taken for each group. Then indicate the relevant group that's associate with. In order to prepare the arff file there is a format to provide. The format is in association rule mining is as follows,

```
@RELATION dataClustring

@ATTRIBUTE 'customerContact'    String
@ATTRIBUTE G1-5 REAL
@ATTRIBUTE G6-9     REAL
@ATTRIBUTE G10-11   REAL
@ATTRIBUTE A/L  REAL
@ATTRIBUTE other    REAL
@ATTRIBUTE class    {Primary_level, Secondary_level, O/L, A/L, Other}

@DATA
+94703259716,5.1,3.3,1.3,0.2,0.1,Primary_level
+94716874245,6.1,2.3,1.3,0.2,0.1,Primary_level
+94718666790,7.1,1.3,1.3,0.2,0.1,Primary_level
+94714124532,5.1,2.3,2.3,0.2,0.1,Primary_level
+94719356151,5.1,2.3,2.3,0.2,0.1,Primary_level
+94779334004,6.1,1.3,2.3,0.2,0.1,Primary_level
+94711800304,5.1,2.3,1.3,1.2,0.1,Primary_level
+94776323826,7.4,1.2,1.2,0.1,0.1,Primary_level
+94715799280,6.1,1.3,2.3,0.2,0.1,Primary_level
+94771388605,4.1,2.3,1.3,1.2,1.1,Primary_level
+94773444611,5.3,2.2,2.2,0.2,0.1,Primary_level
+94718200086,6.1,1.3,2.3,0.2,0.1,Primary_level
+94773981378,4.1,2.3,1.3,1.2,1.1,Primary_level
+94777802453,7.2,1.2,1.3,0.2,0.1,Primary_level
+94719308059,6.1,2.3,1.3,0.2,0.1,Primary_level
+94778982617,4.4,2.2,1.3,1.1,1.1,Primary_level
+94763846264,5.1,2.3,2.3,0.2,0.1,Primary_level
+94714933792,5.1,2.3,1.3,1.2,0.1,Primary_level
+94714710096,6.1,2.3,1.3,0.2,0.1,Primary_level
+94777414566,4.5,2.1,1.1,1.1,1.2,Primary_level
+94771510600,6.1,2.3,1.3,0.2,0.1,Primary_level
+94702674169,7.1,1.3,1.3,0.2,0.1,Primary_level
+94769888600,5.1,2.3,1.3,1.2,0.1,Primary_level
+94767600145,7.1,1.3,1.3,0.2,0.1,Primary_level
+94718019268,7.2,1.2,1.2,0.2,0.2,Primary_level
+94766615386,4.5,2.1,1.1,1.1,1.2,Primary_level
+94714149204,5.3,2.2,1.2,1.1,0.2,Primary_level
+94772786274,7.2,1.2,1.2,0.2,0.2,Primary_level
```

Figure 39: Area 4 arff data file

Then after that need to import the arff file into WEKA application and done the preprocessing. As the preprocessing task remove the attribute" customerContact" column.

Simple EM (expectation maximization) class is used as data mining algorithm. This is an iterative method for estimates of parameters in statistical models to find maximum likelihood or maximum a posterior maximum (MAP) estimate. This model is depending on disregarded latent variables [18]. Following Clustering results are taken as the output.

For EM, use training set cluster model with 5 clusters the output shown below,

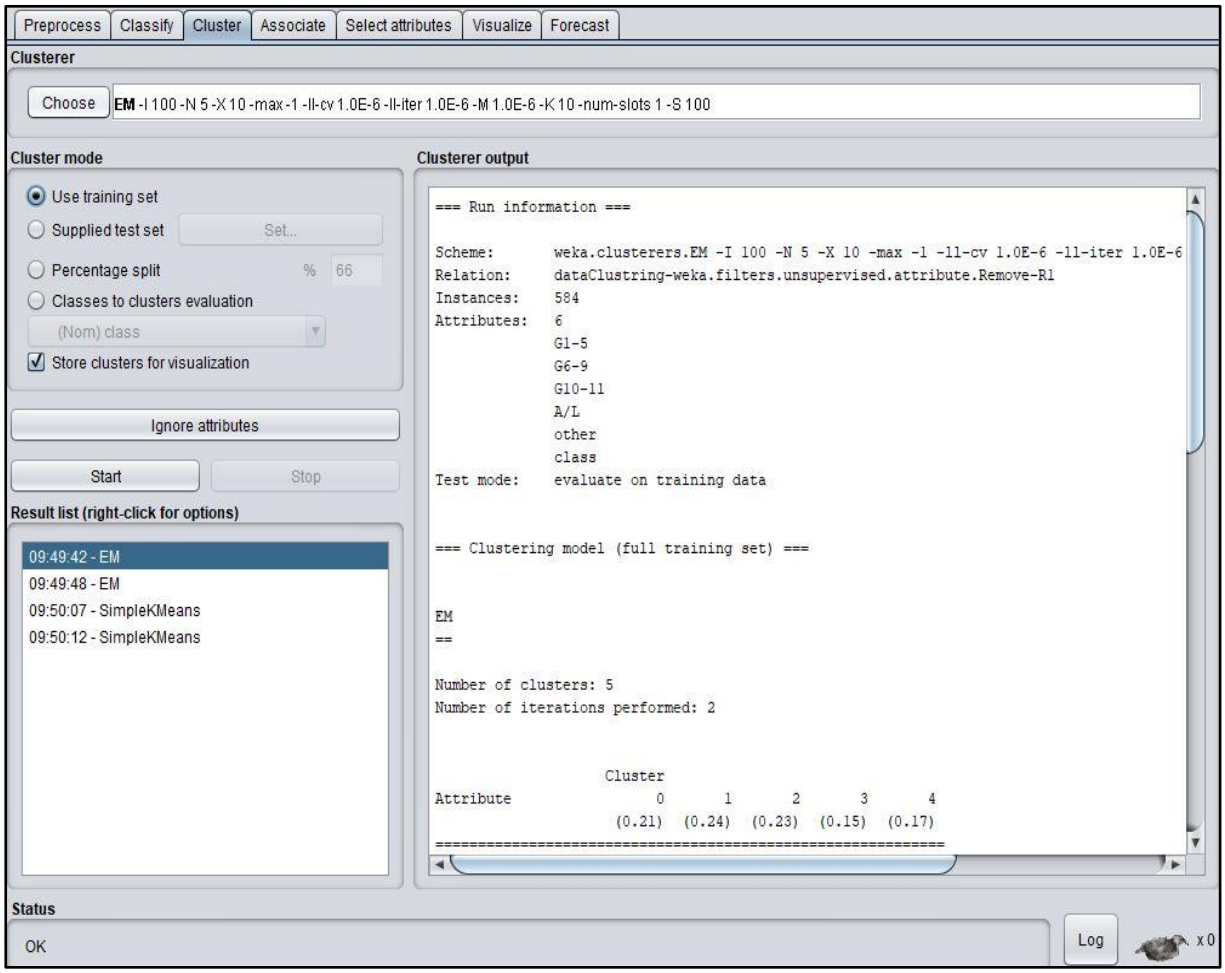

Figure 40: EM with training set cluster output part 1

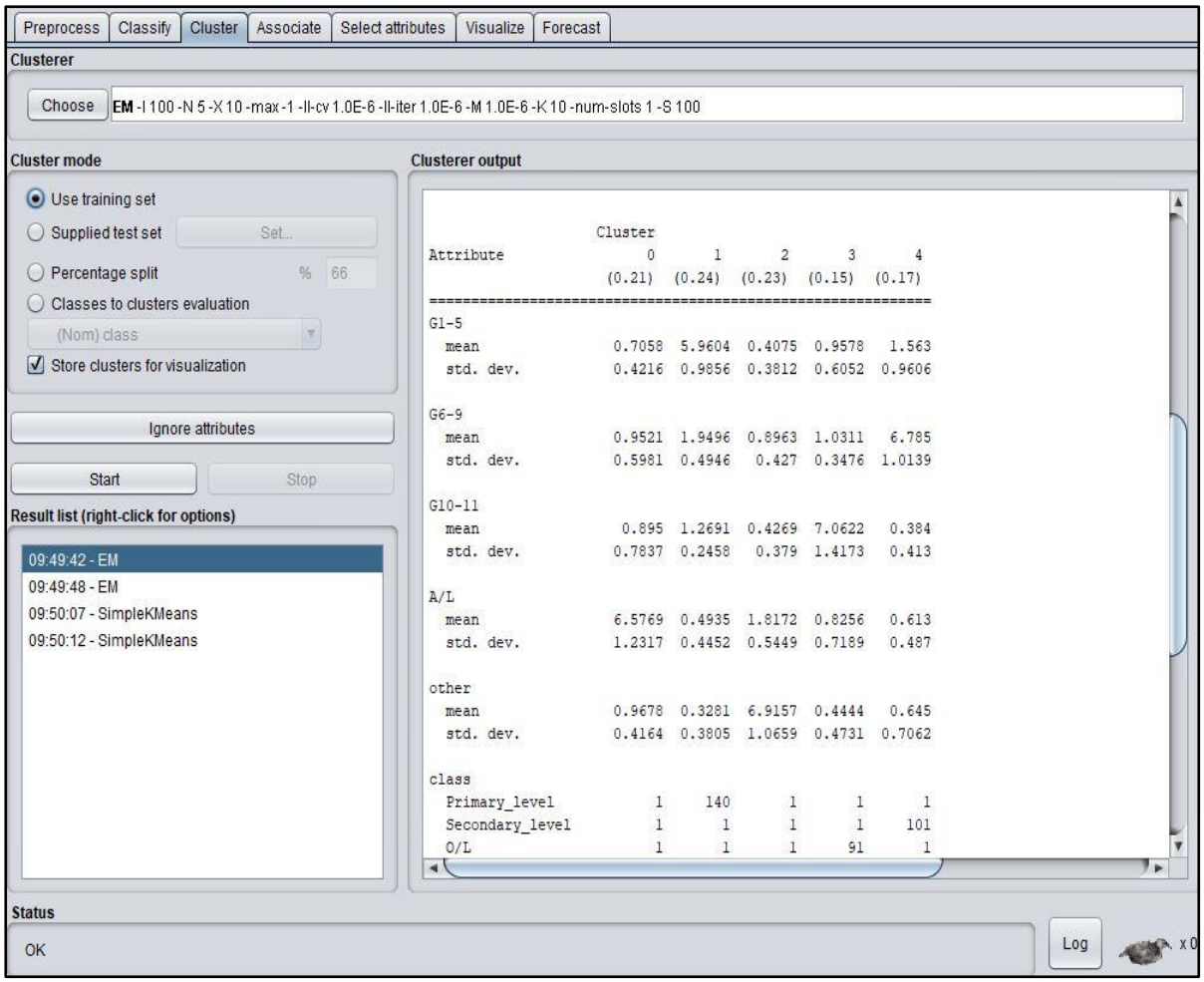

Figure 41: EM with training set cluster output part 2

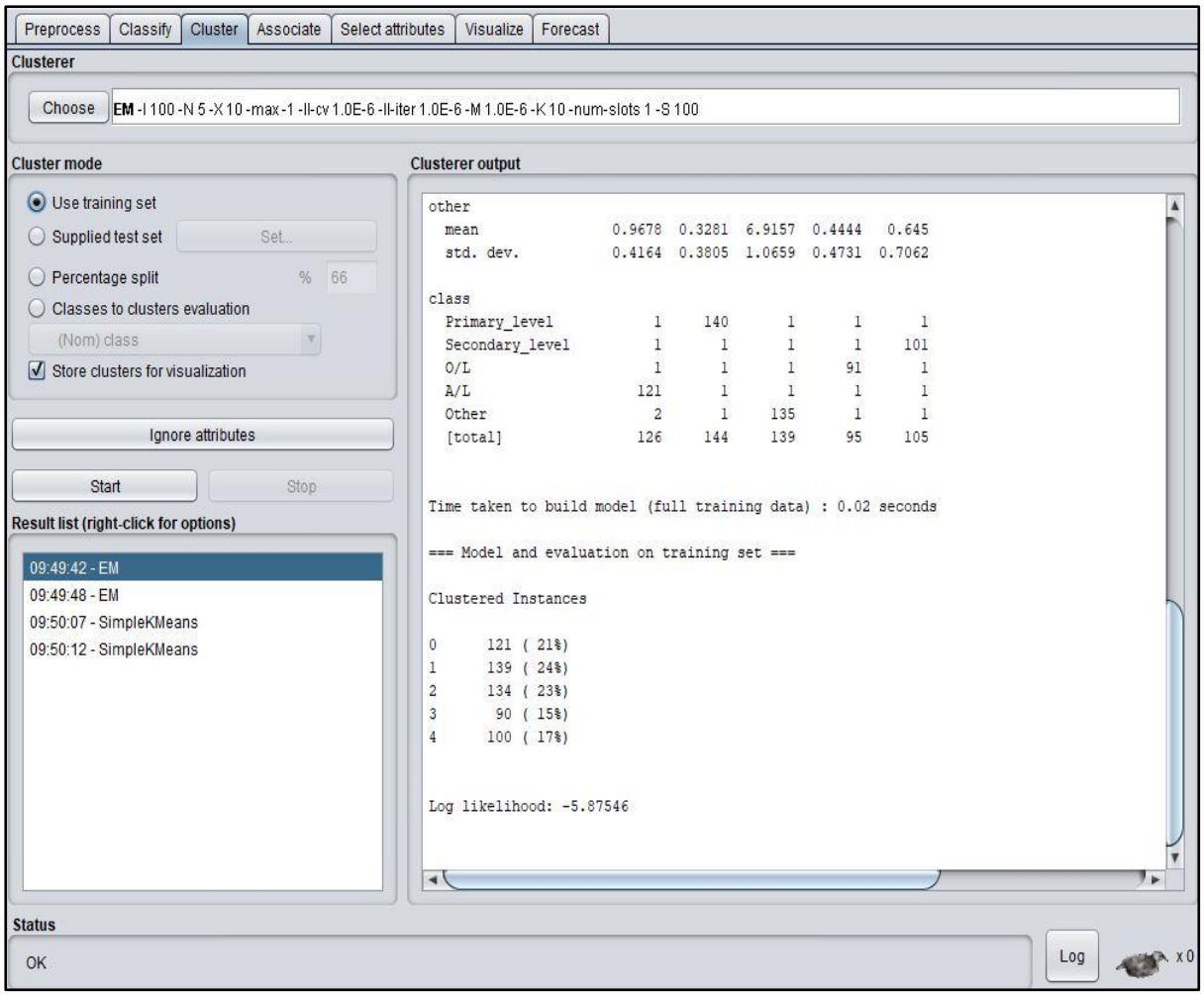

Figure 42: EM with training set cluster output part 3

For EM, use class attribute to cluster evaluation model with 5 clusters the output shown below,

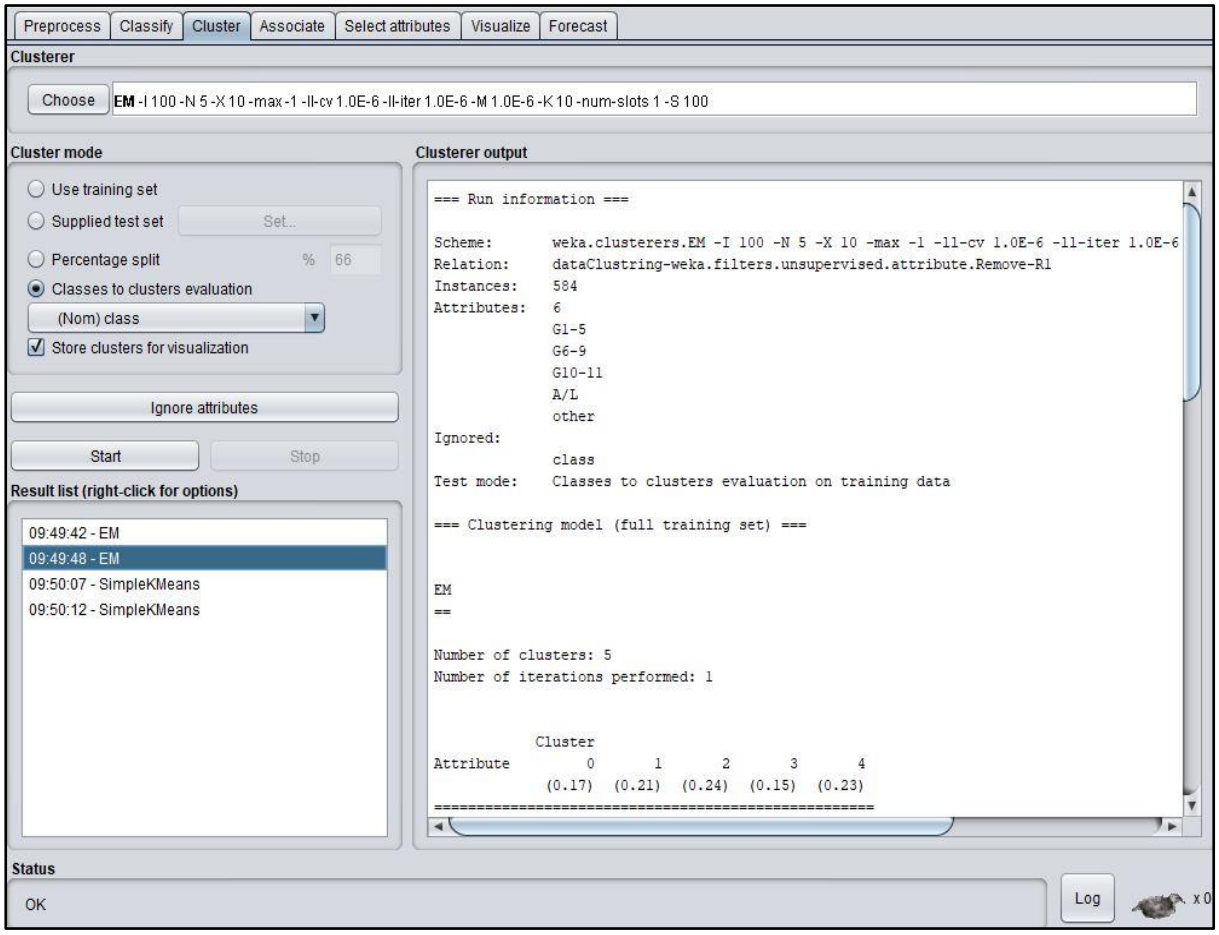

Figure 43: EM with class attribute cluster output part 1

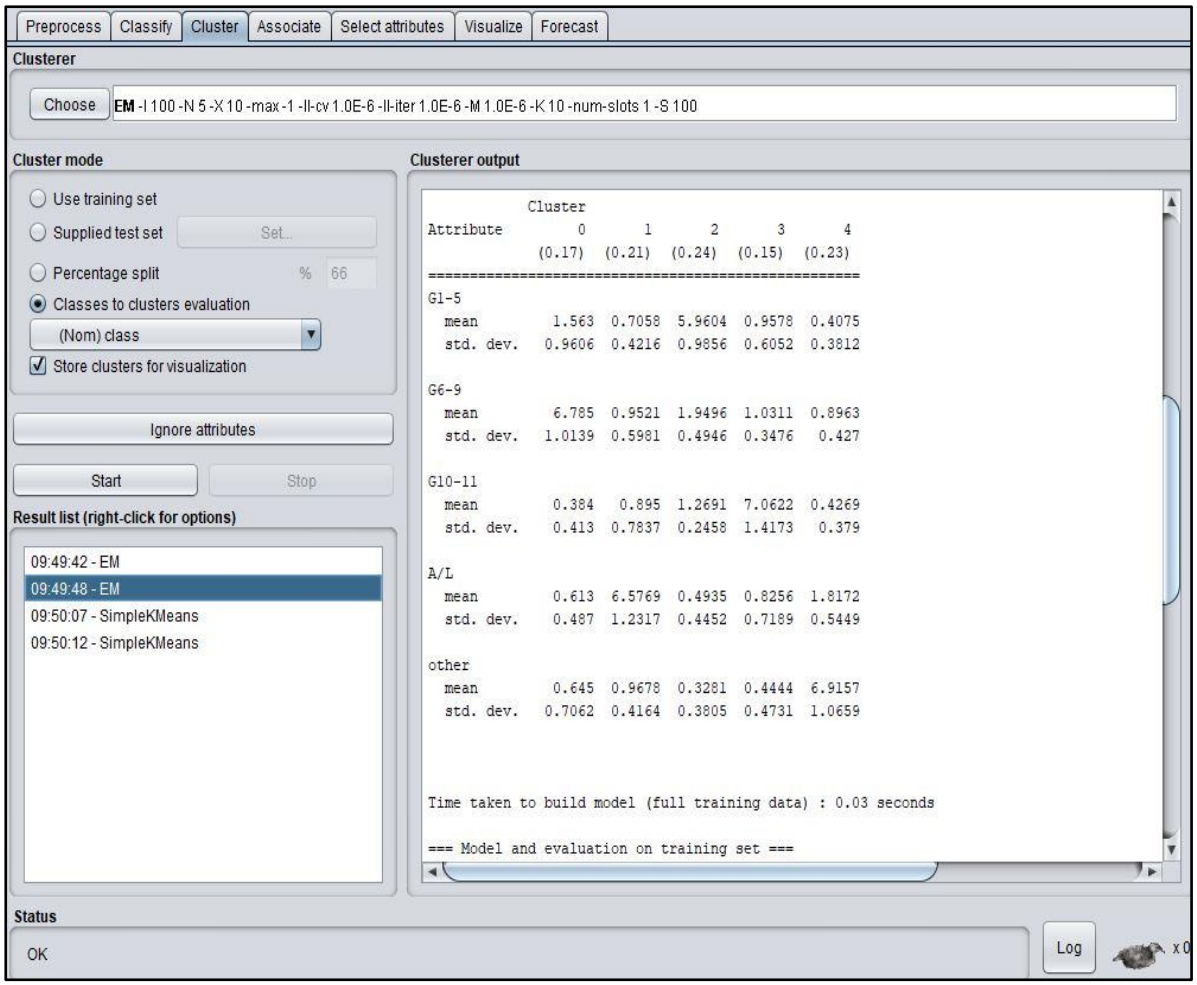

Figure 44: EM with class attribute cluster output part 2

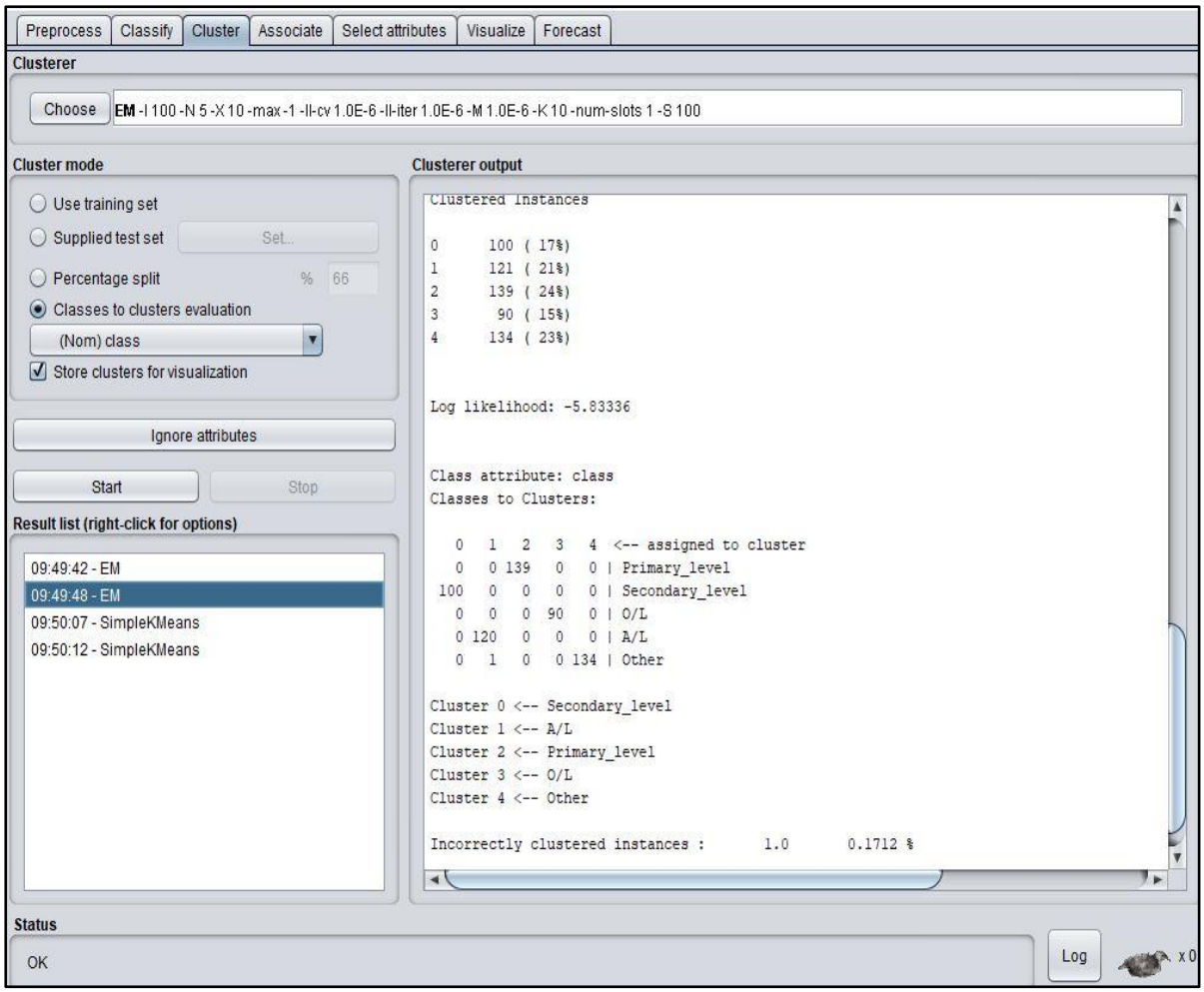

Figure 45: EM with class attribute cluster output part 3

As the another data mining algorithm, SimpleKMeans is used. This cluster refers to points aggregated together in collection of data [19]. Following Clustering results are taken as the output.

For SimpleKMeans, use training set cluster model with 5 clusters the output shown below,

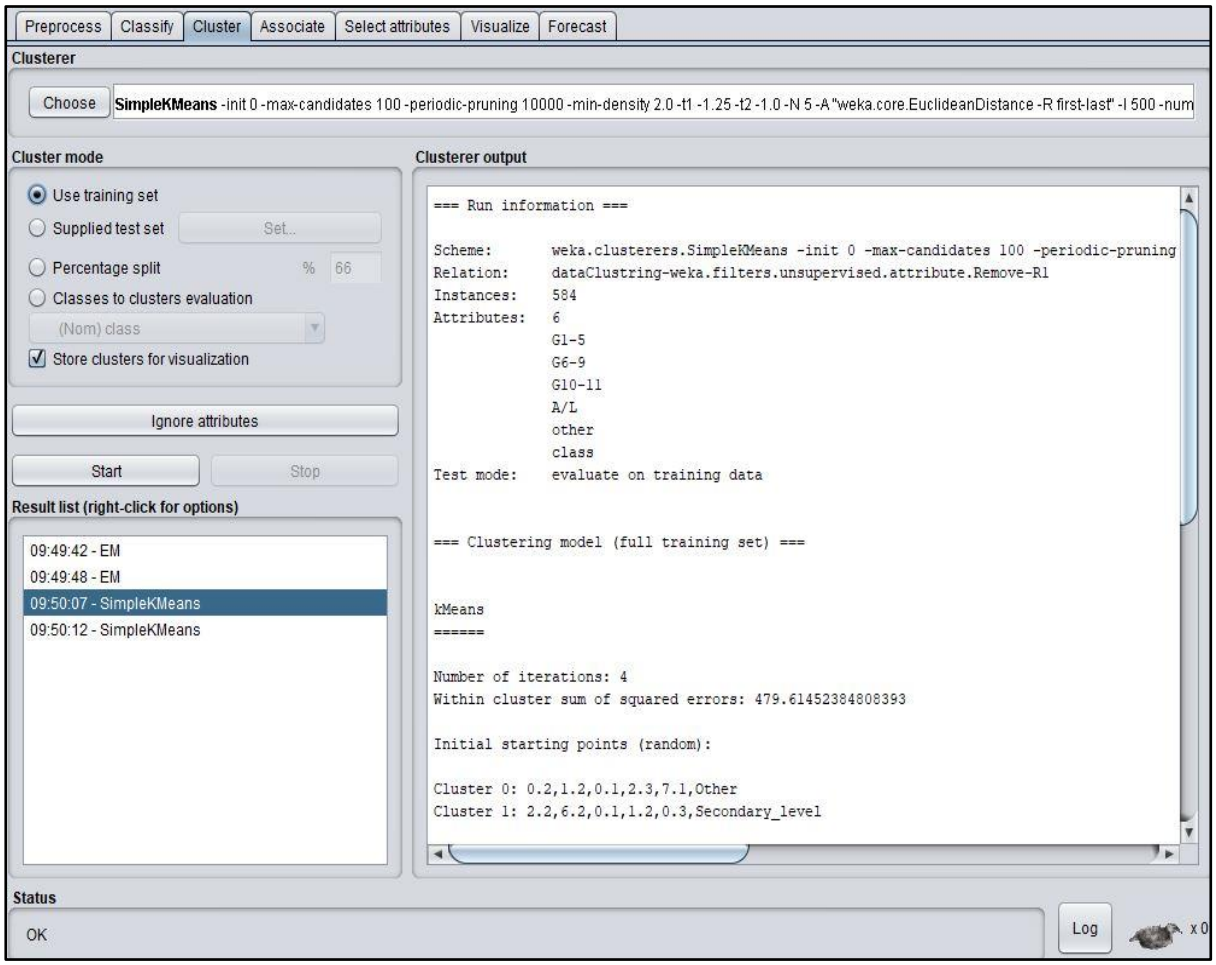

Figure 46: SimpleKMeans with training set cluster output part 1

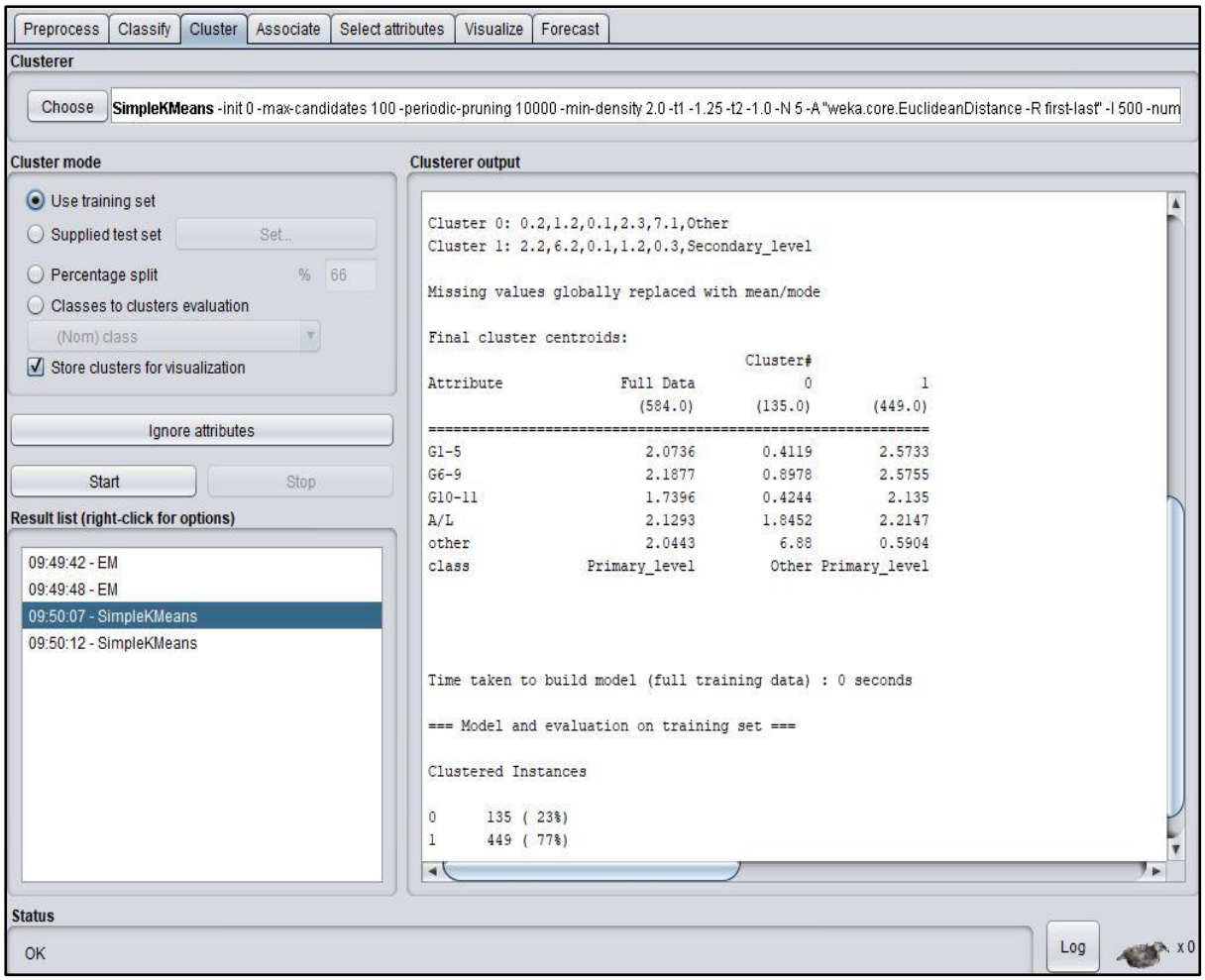

Figure 47: SimpleKMeans with training set cluster output part 2

For SimpleKMeans, use class attribute to cluster evaluation cluster model with 5 clusters the output shown below,

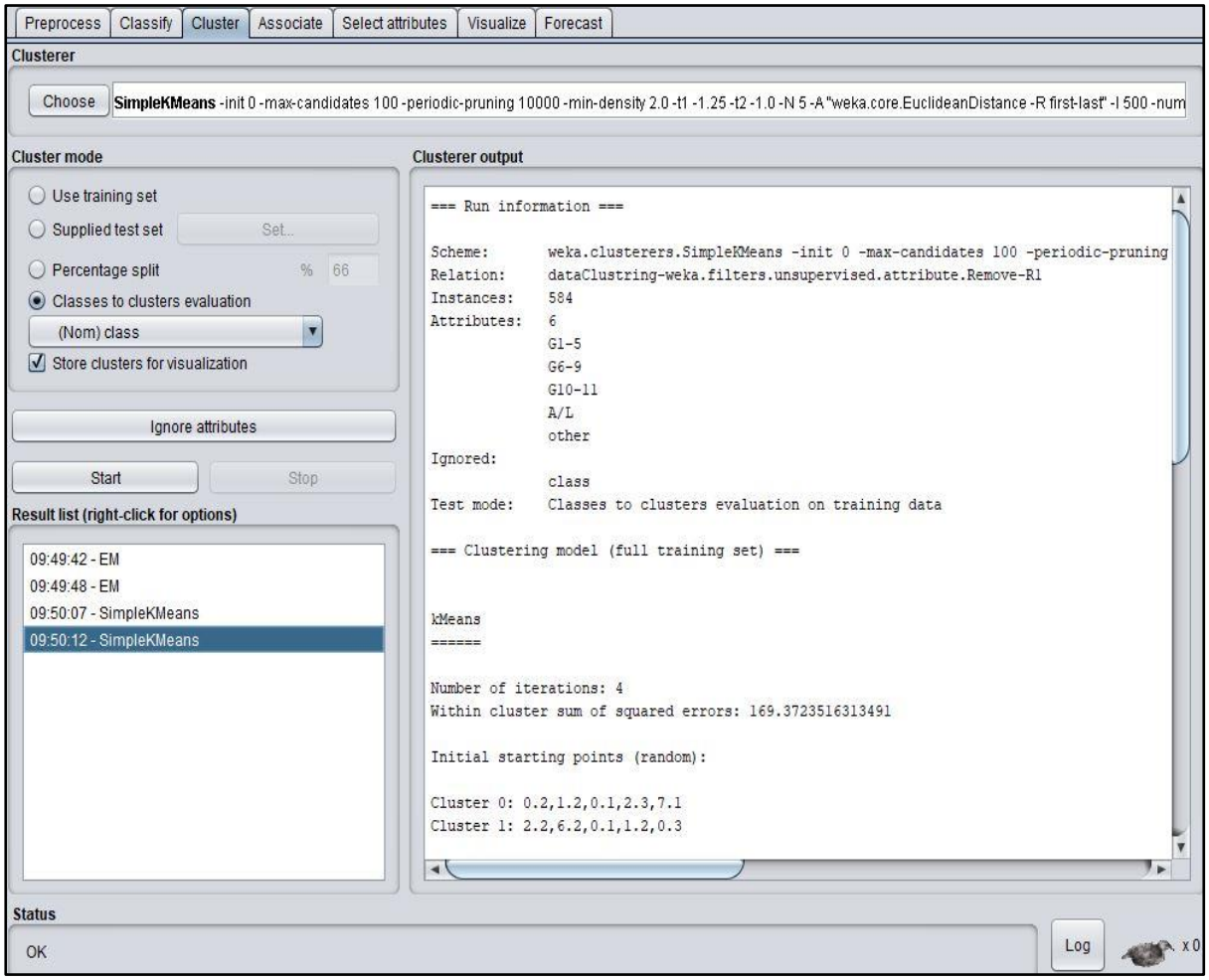

Figure 48: SimpleKMeans with class attribute cluster output part 1

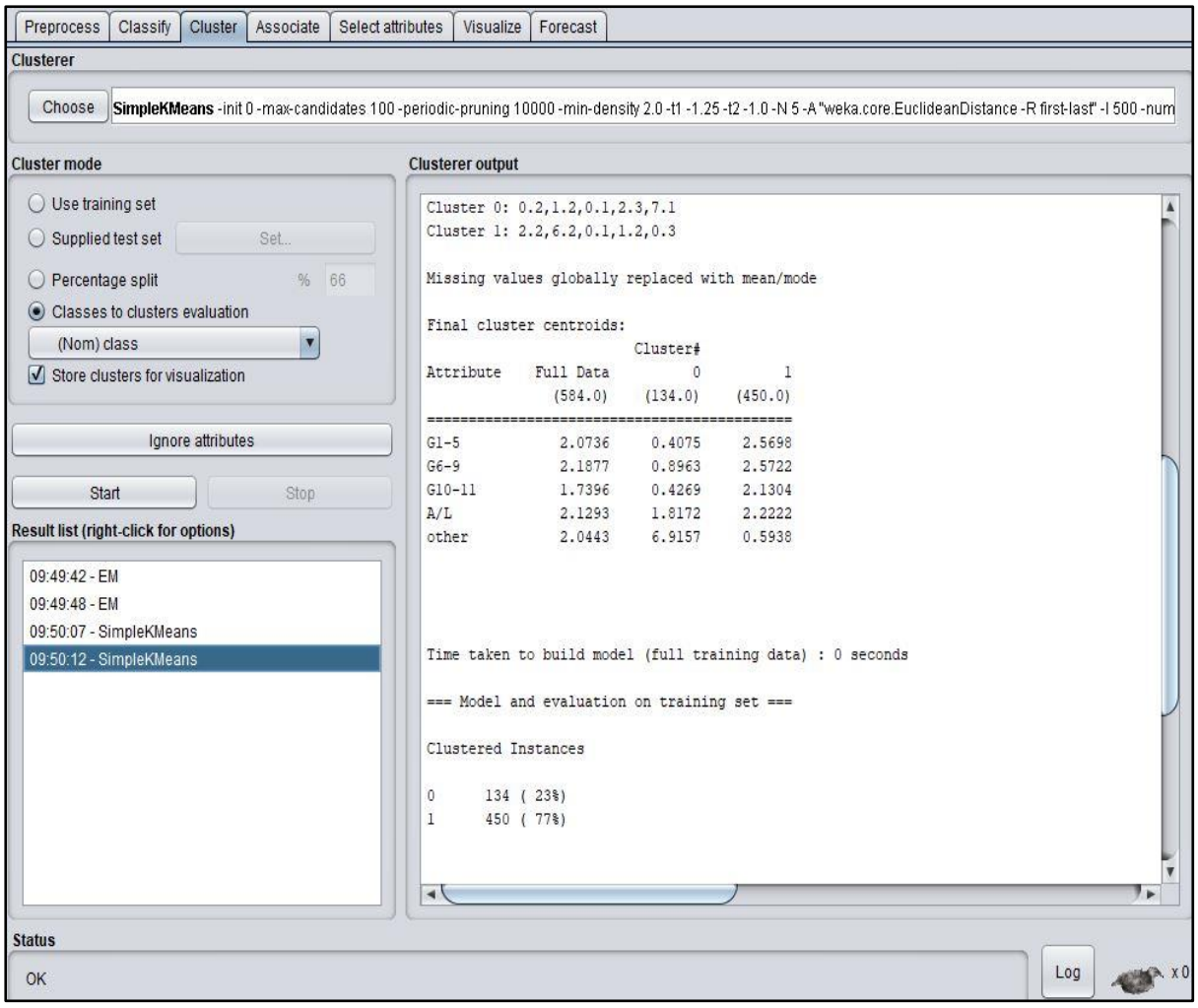

Figure 49: SimpleKMeans with class attribute cluster output part 2

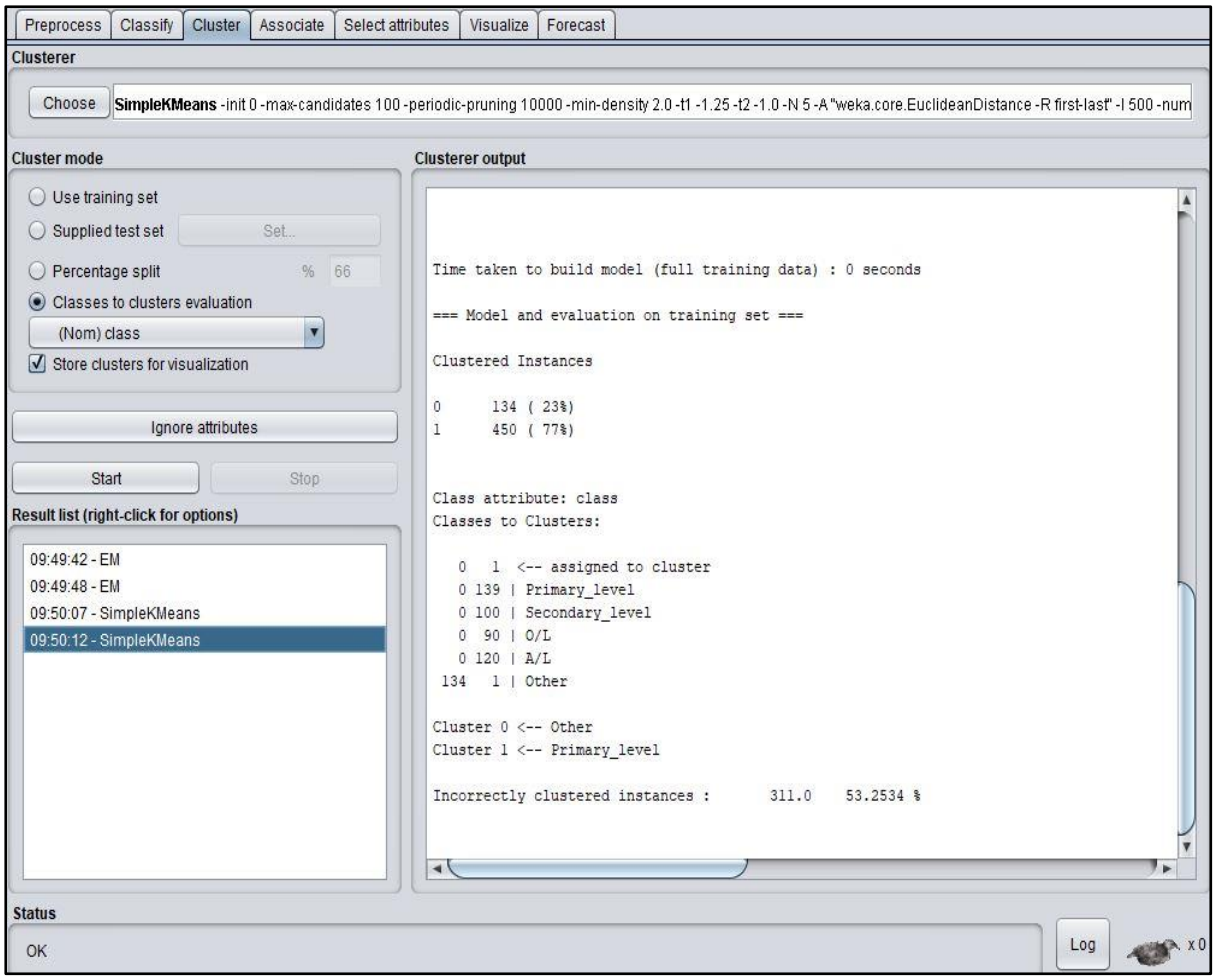

Figure 50: SimpleKMeans with class attribute cluster output part 3

Evaluate Algorithms

According to above two algorithms should evaluate by analyzing the Incorrectly clustered instances. In EM algorithm got 0.1 as percentage of 0.1712 % and SimpleKMeans algorithm got 311.0 as percentage of 53.2534 %. By referring to above two methods a smaller number of Incorrectly clustered instances are better.

So, EM algorithm is selected for data mining purposes. Different group of customers are getting to promote a particular product/s. So that this is easy to make decisions for managers to identifying the market trends.

4.5.5 Area 5

In order to identifying product sales forecasting process like required quantity of each product for special annual events like exhibitions (here took the Colombo International Book Fair), first need to prepare the dataset. In order to import the data set into the WEKA application arff dataset file need to prepare.

Separate data file is used in order to predict the possibility of area 4. Special event held year and relevant item quantity taken to prediction. In order to prepare the arff file there is a format to provide.

Forecasting takes models that fit historical data and uses them to predict future observations. Extrapolation is making predictions from time series data handling classical statistical way [20]. The data file format is in time series forecasting is as follows,

```
@relation savi
@attribute year numeric
@attribute Grade_1_Buddhism numeric
@attribute Grade_1_Mathematics numeric
@attribute Grade_1_Language numeric
@attribute Grade_1_Environment numeric
@attribute Grade_2_Buddhism numeric
@attribute Grade_2_Mathematics numeric
@attribute Grade_2_Language numeric
@attribute Grade_2_Environment numeric
@attribute Grade_3_Buddhism numeric
@attribute Grade_3_Mathematics numeric
@attribute Grade_3_Language numeric
@attribute Grade_3_Environment numeric
@attribute Grade_4_Buddhism numeric
@attribute Grade_4_Mathematics numeric
@attribute Grade_4_Language numeric
@attribute Grade_4_Environment numeric
@attribute Grade_5_Buddhism numeric
@attribute Grade_5_Mathematics numeric
@attribute Grade_5_Language numeric
@attribute Grade_5_Environment numeric
@attribute Grade_6_Science numeric
@attribute Grade_6_Mathematics numeric
@attribute Grade_6_History numeric

@data
2014,25,30,30,30,40,40,40,35,45,45,45,40,60,60,60,55,80,80,80,75,50,50,50,50,50,50,50,55,60,60,60,
2015,30,35,35,35,45,45,45,45,55,55,55,35,65,65,65,60,85,85,85,70,60,60,60,60,60,60,60,50,70,70,70,
2016,25,32,32,32,43,43,43,30,55,55,55,42,63,63,63,52,83,83,83,72,53,53,53,53,53,53,53,52,63,63,63,
2017,40,60,60,60,75,75,75,55,75,75,75,60,80,80,80,75,100,100,100,95,70,70,70,70,70,70,70,75,80,80,
2018,25,30,30,30,40,40,40,35,45,45,45,40,60,60,60,55,80,80,80,75,50,50,50,50,50,50,50,55,60,60,60,
2019,60,65,65,65,75,75,75,75,75,75,75,55,95,95,95,90,105,105,105,80,70,70,70,70,70,70,70,50,90,90,
2020,25,23,23,23,40,40,40,30,45,45,45,30,63,63,63,52,80,80,80,75,50,50,50,50,50,50,50,55,67,67,67,
```

Figure 51: Area 5 arff data file

Then after that need to import the arff file into WEKA application and done the preprocessing. As the preprocessing task remove the attribute" year" column.

One approach is to use Linear Regression data mining algorithm. From investigational or observational data can predict the behavior of dynamic systems [21]. Following results are taken as the output.

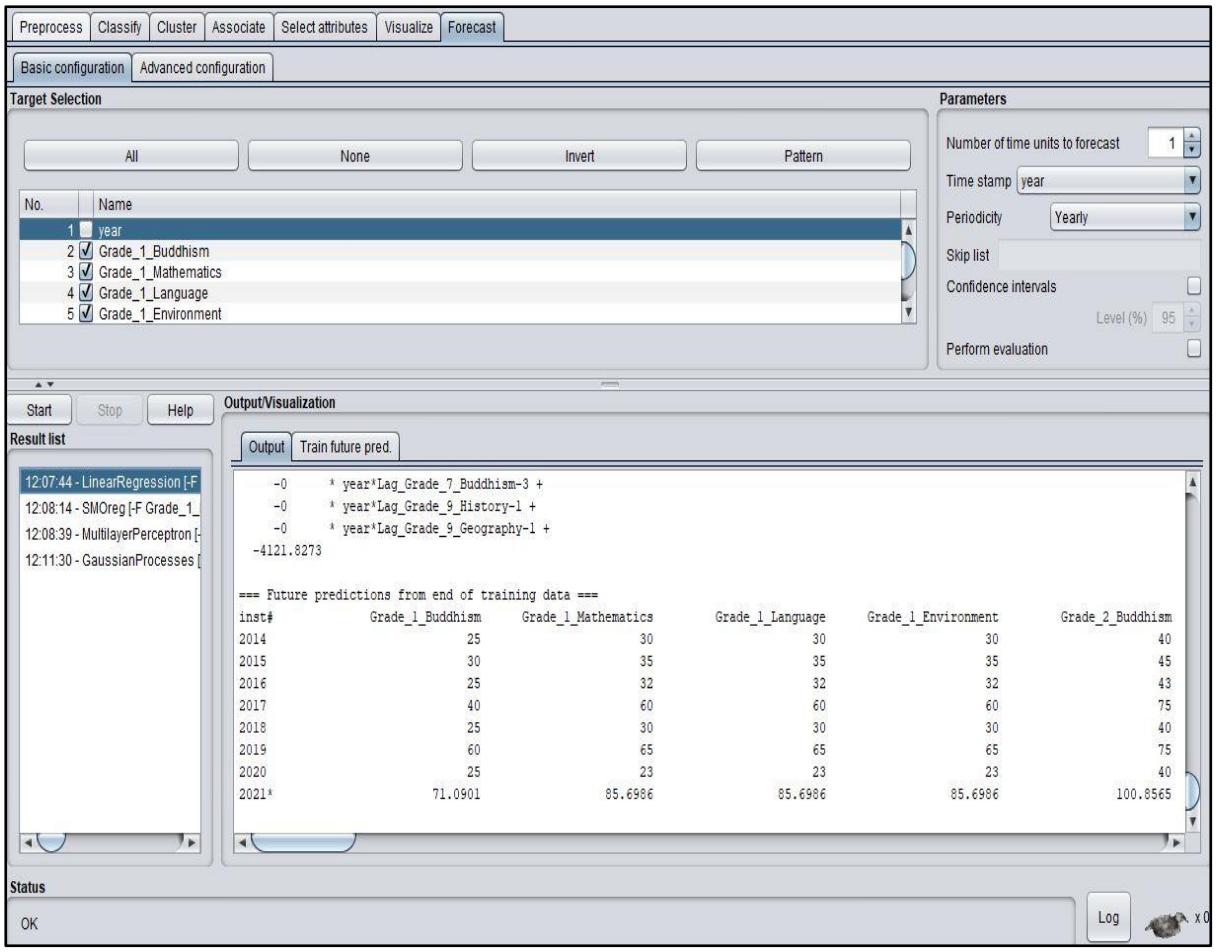

Figure 52: Linear Regression Prediction output

When Evaluation on training data there are two different errors, the mean values of absolute error and root mean value of root squared error values. The error rates are mention below,

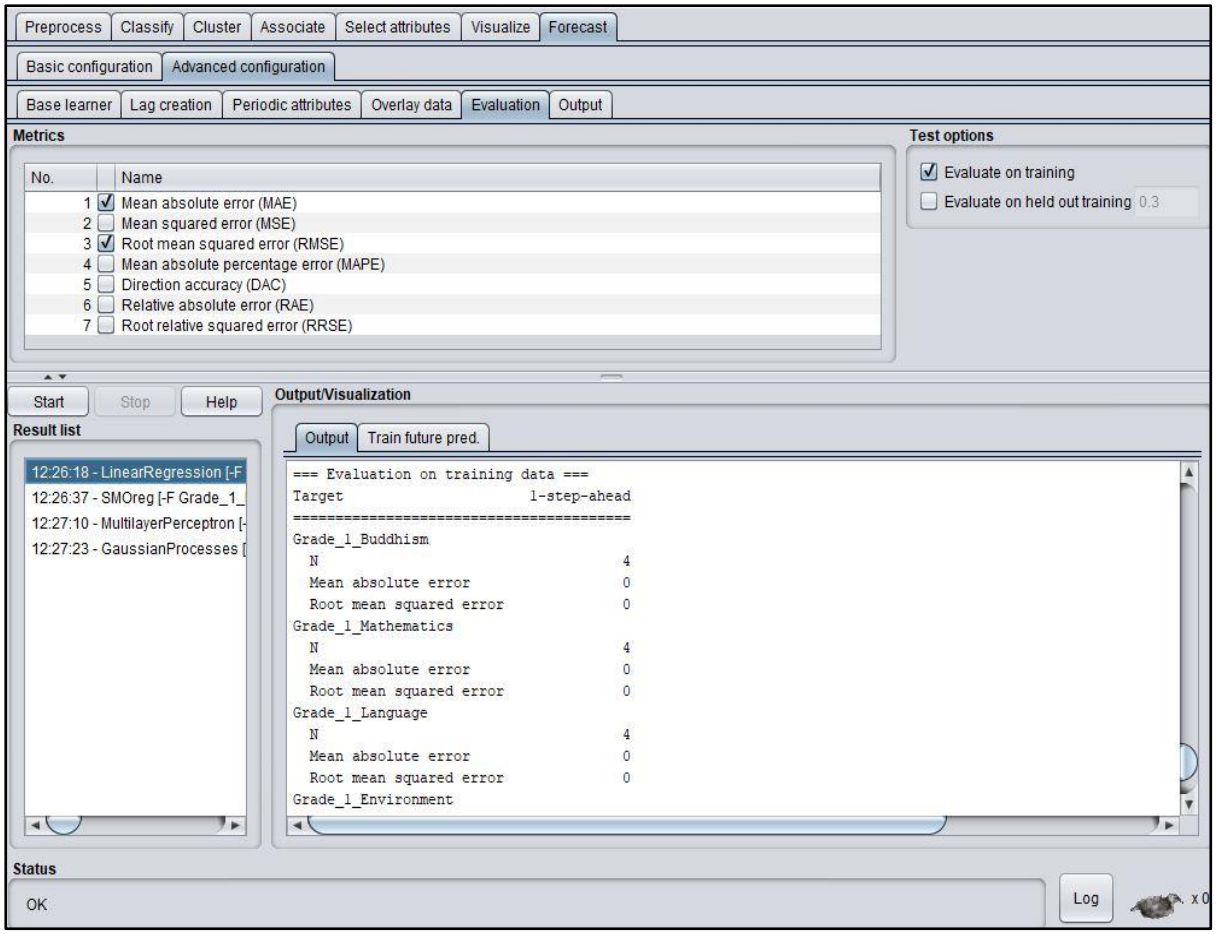

Figure 53: Linear Regression Error Rates evaluation

Another approach is to use SMOreg data mining algorithm. This implements regression in support vector machine. Using various algorithms can learn the parameters [22]. Following results are taken as the output.

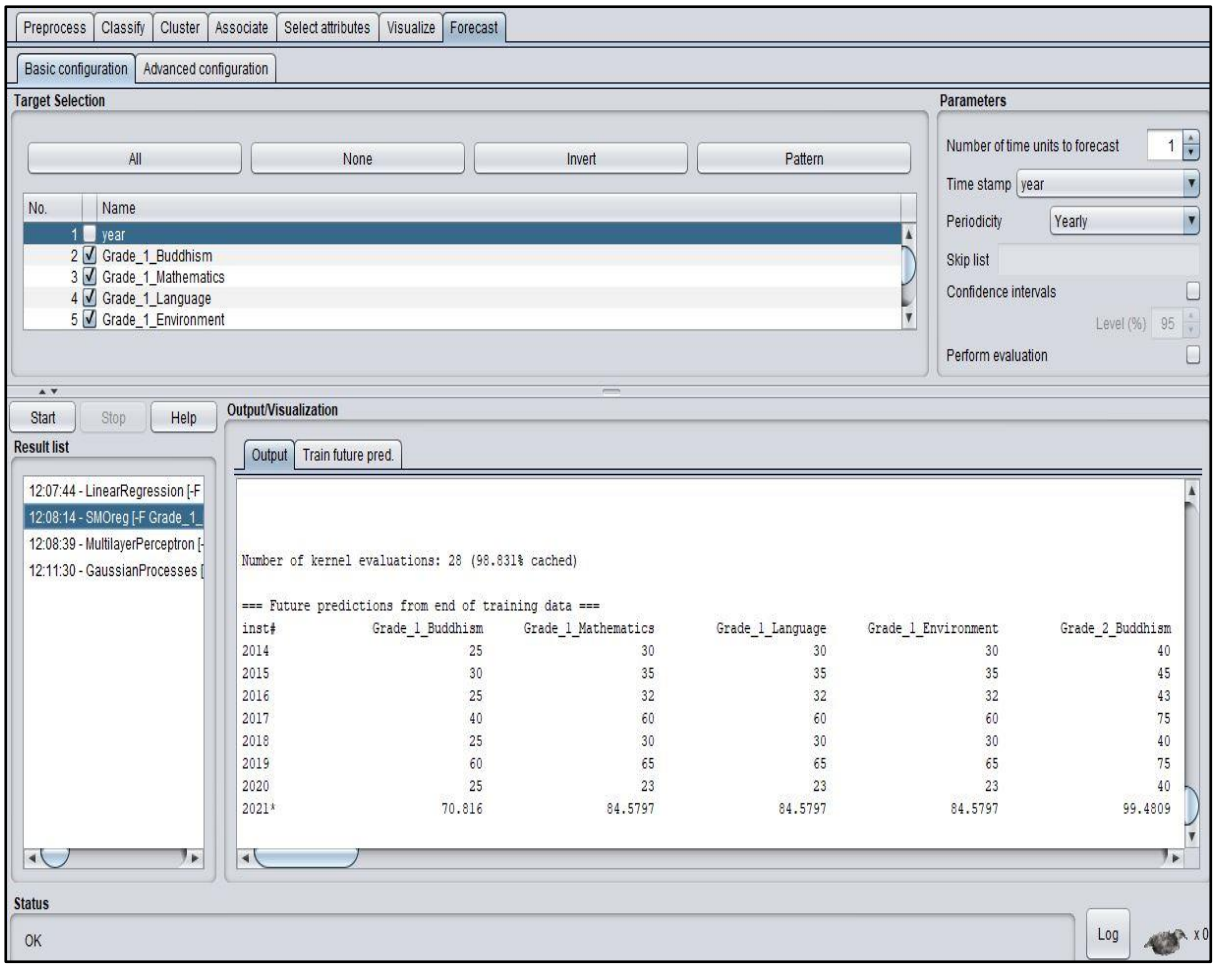

Figure 54: SMOreg Prediction output

When Evaluation on training data there are two different error values, the mean values of absolute error and root mean value of root squared error values. The error rates are mention below,

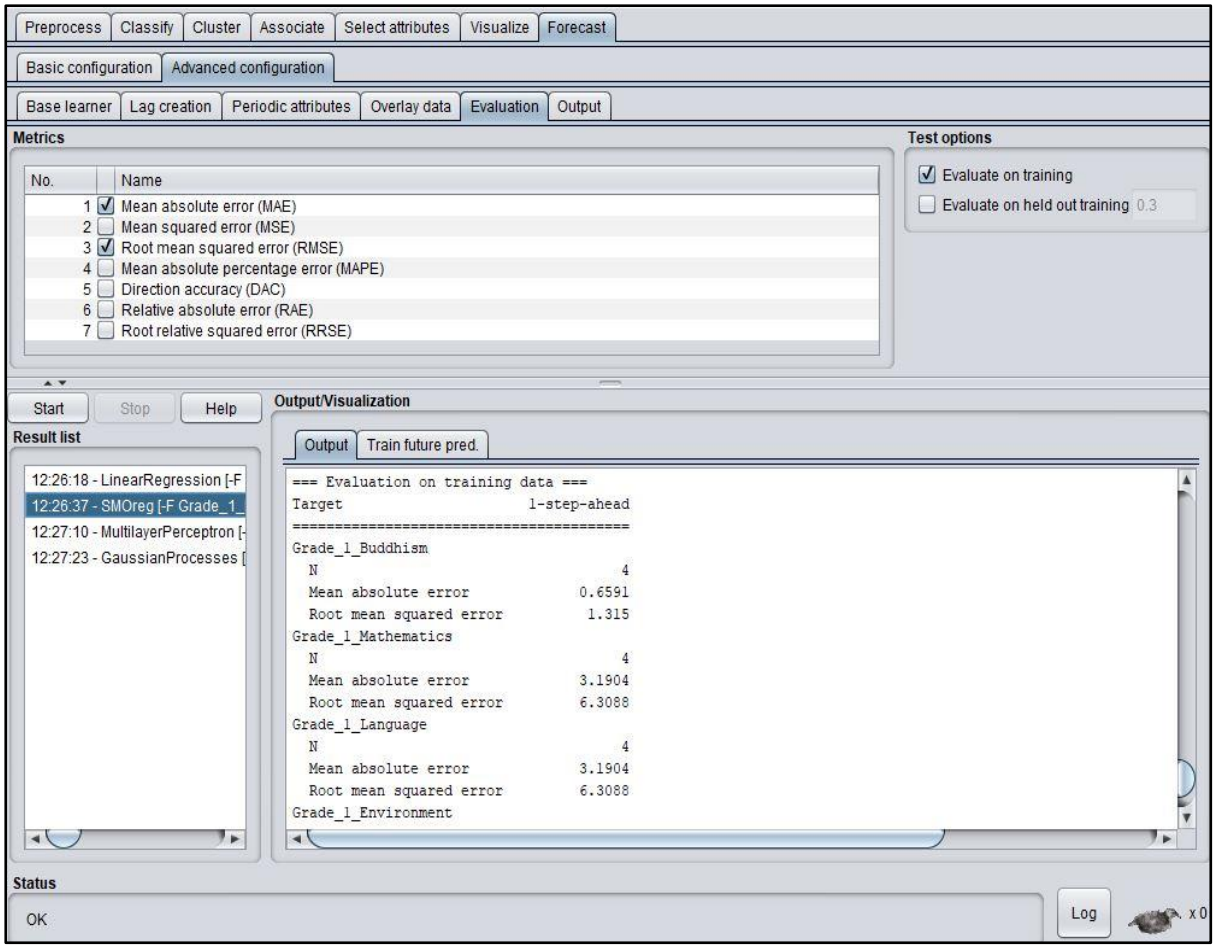

Figure 55: SMOreg Error Rates evaluation

Another approach is to use Multilayer Perceptron data mining algorithm. This is a subset of artificial feedback neural networks (ANN). For MLP training use backpropagation mechanism in supervised learning technique. In MLP linear perceptron has multiple layers and nonlinear activation separation. This cannot be separated linearly in distinguish data [23]. Following results are taken as the output.

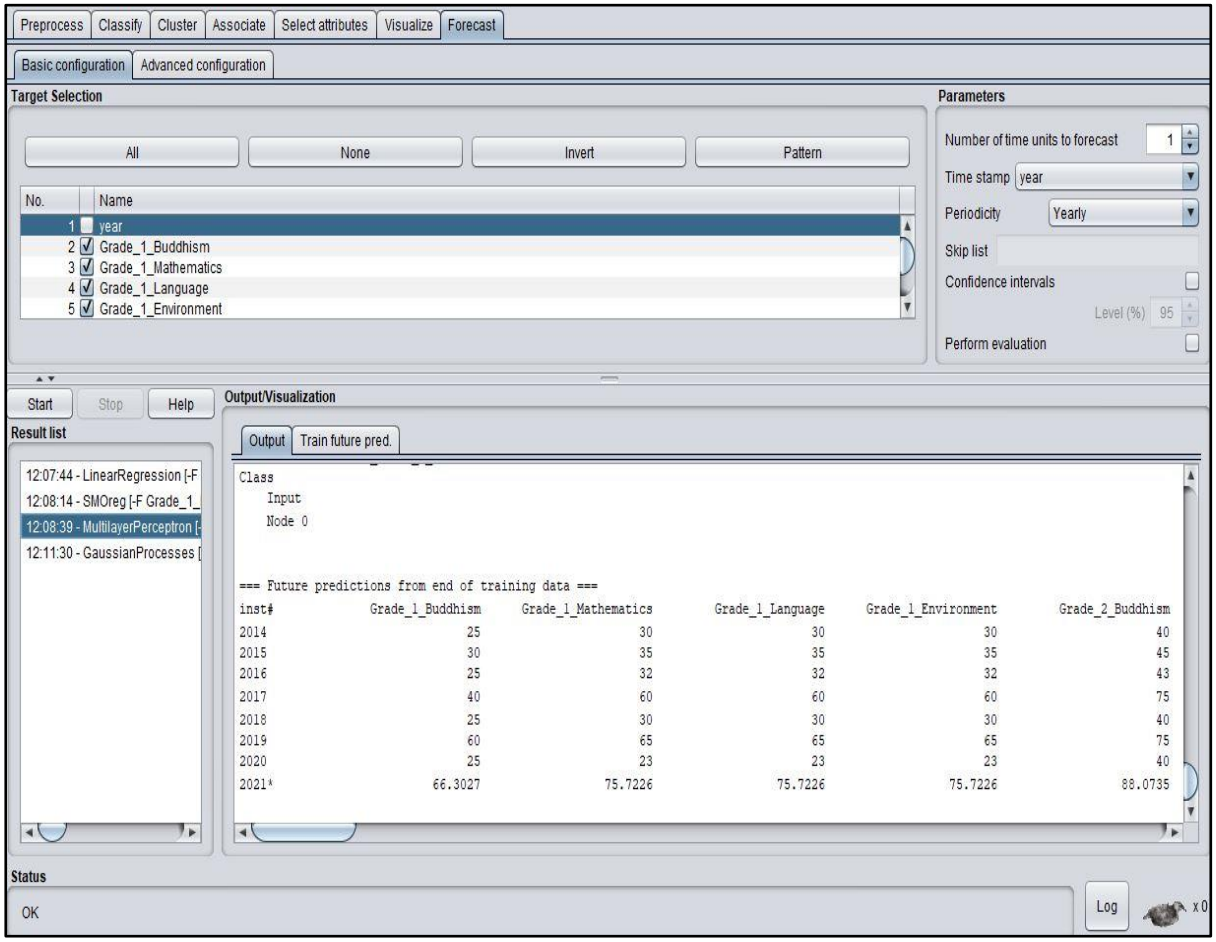

Figure 56: Multilayer Perceptron Prediction output

When Evaluation on training data there are two different errors, the mean values of absolute error and root mean value of root squared error values. The error rates are mention below,

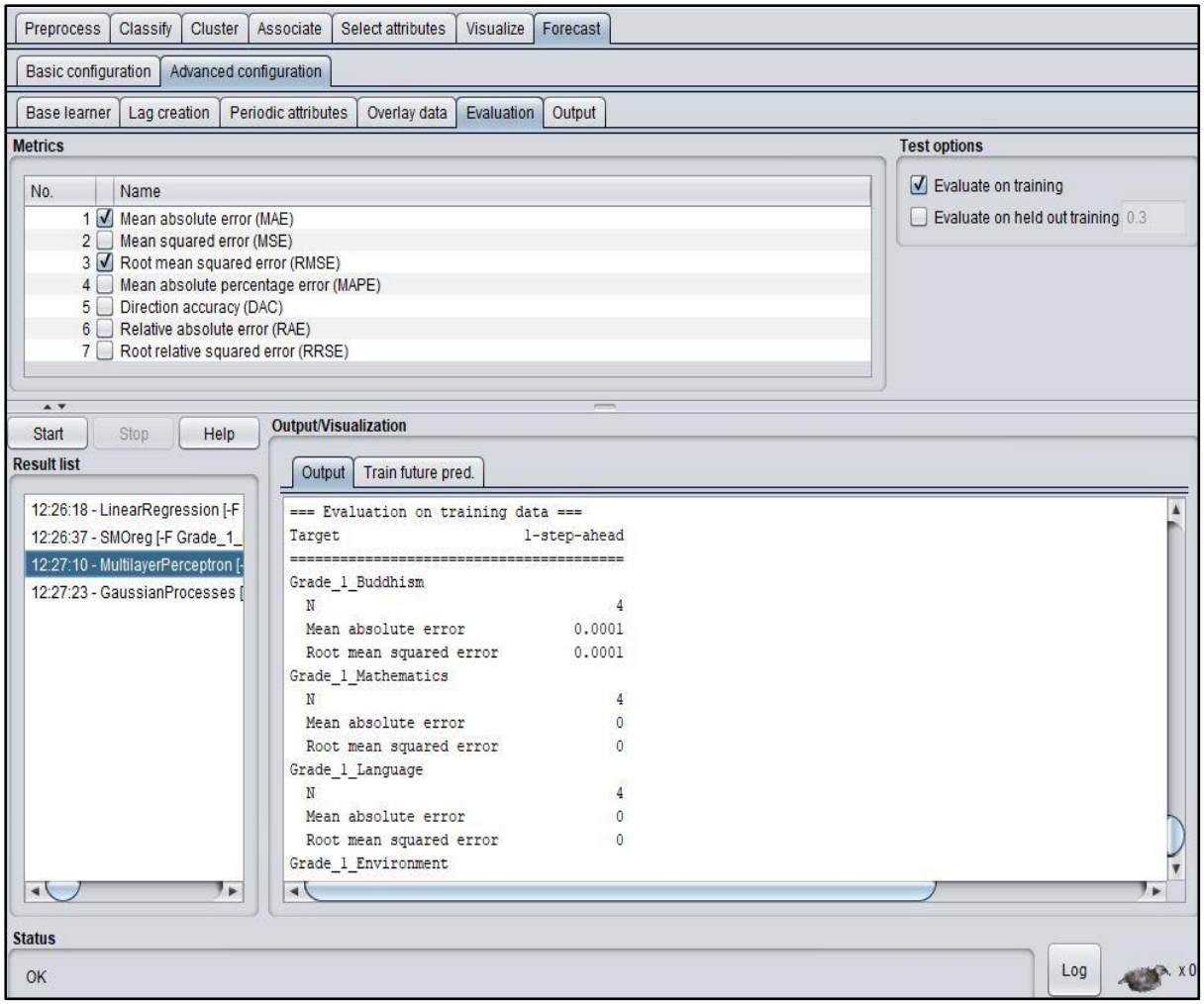

Figure 57: Multilayer Perceptron Error Rates evaluation

Another approach is to use Gaussian Processes data mining algorithm. The Gaussian process is coming under classification algorithm in the area of machine learning. Gaussian processes are a simplification of the Gaussian probability distribution process. It can use as machine learning sophisticated non-parametric algorithms in regression and classification. [24]. Following results are taken as the output.

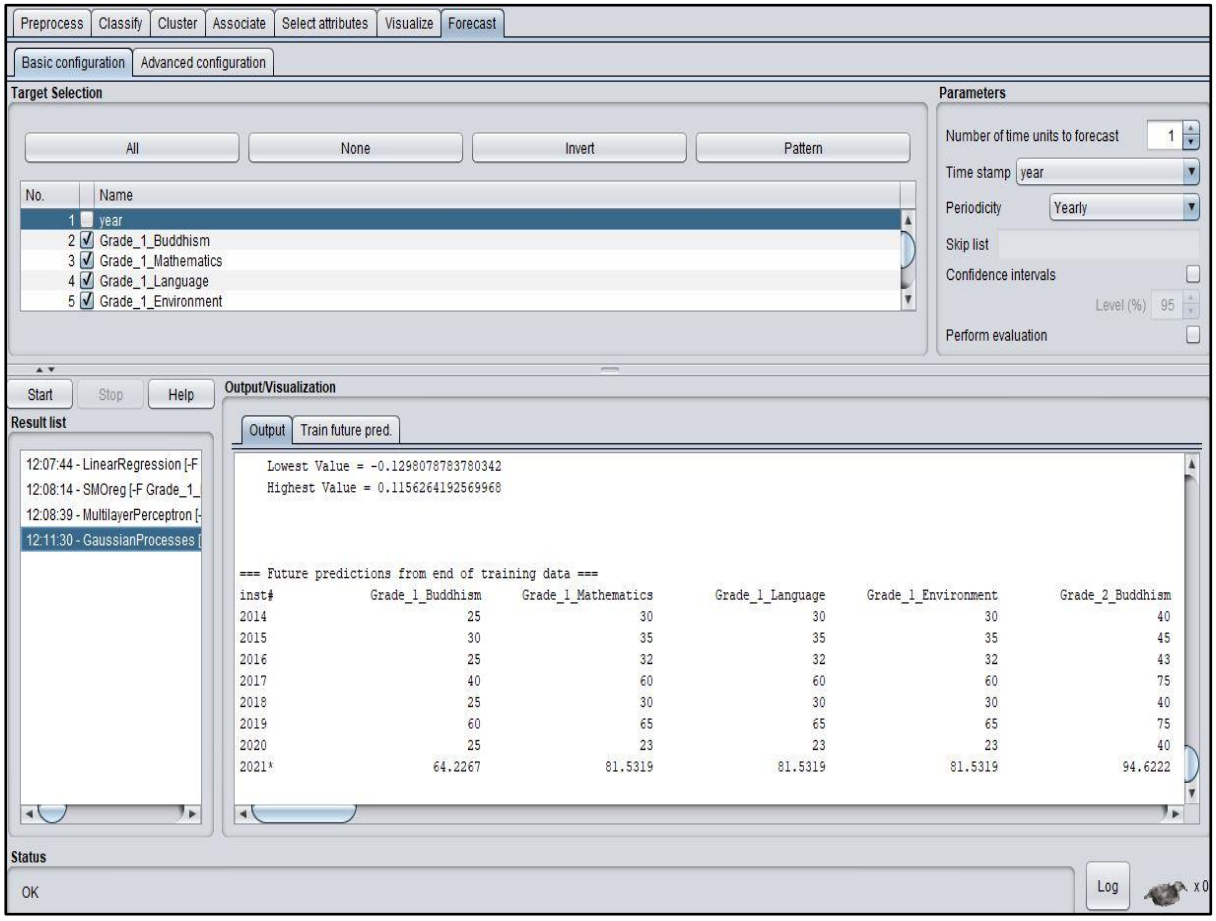

Figure 58: Gaussian Processes Prediction output

When Evaluation on training data there are two different errors are there, the mean values of absolute error and root mean value of root squared error values. The error rates are mention below,

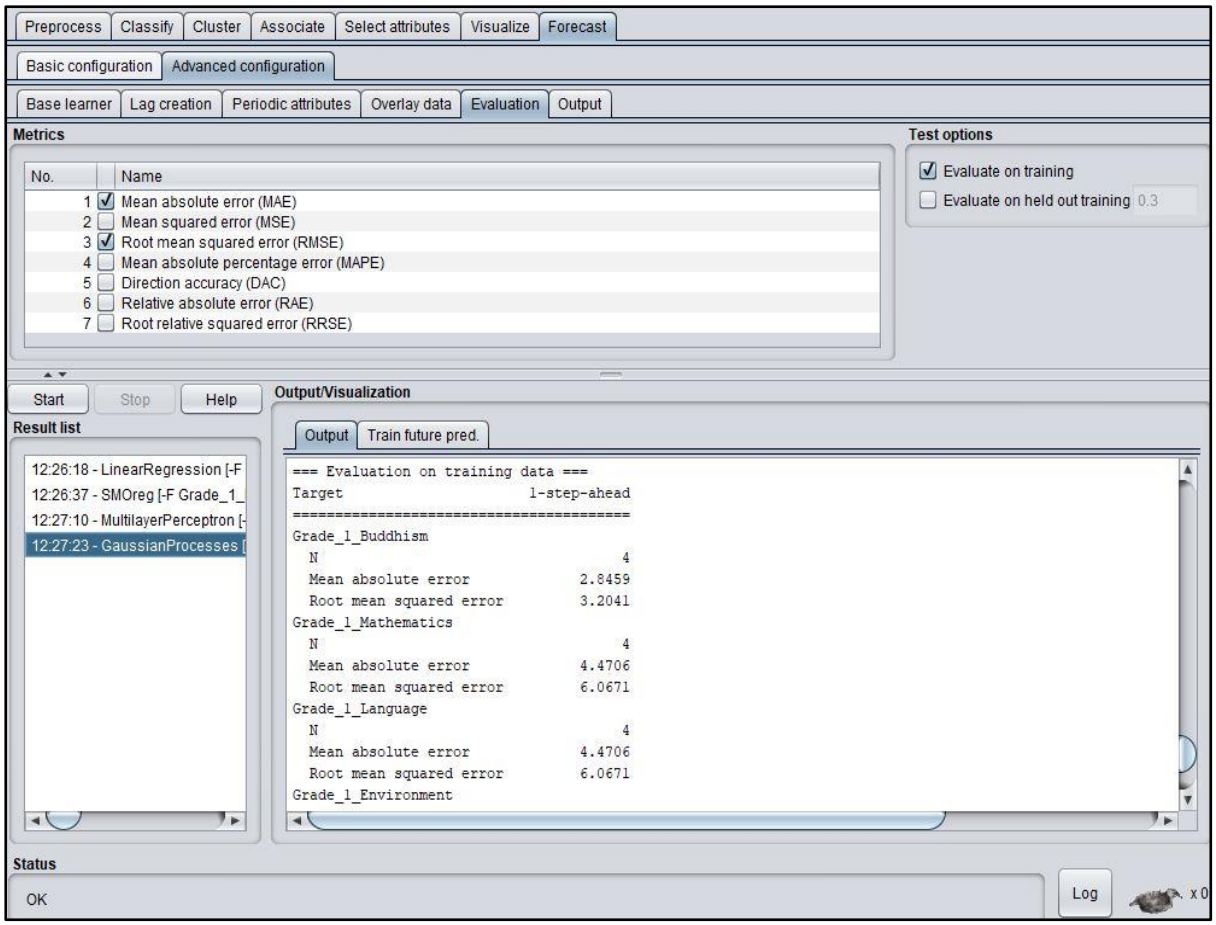

Figure 59: Gaussian Processes Error Rates evaluation

Evaluate Algorithms

According to above four algorithms should evaluate by analyzing the training data with error values. According to of mean values of absolute error and root mean value of root squared error values, these are results [Table 1] taken for one particular product (Grade_1_Buddhism).

Table 1: Error values for four algorithms

	Linear Regression	SMOreg	Multilayer Perceptron	Gaussian Processes
Mean absolute error	0	0.6591	0.0001	2.8459
Root mean squared error	0	1.315	0.0001	3.2041

By referring to above four methods a smaller number of mean values of absolute error and root mean value of root squared error results are better for forecasting. So, Linear Regression and Multilayer Perceptron algorithms are selected for forecasting purposes.

Table 2: Linear Regression and Multilayer Perceptron Prediction values

inst#	Linear Regression 2021*	Multilayer Perceptron 2021*
Grade_1_Buddhism	71.0901	70.3618
Grade_1_Mathematics	85.6986	76.5004
Grade_1_Language	85.6986	76.5004
Grade_1_Environment	85.6986	76.5004
Grade_2_Buddhism	100.8565	88.3001
Grade_2_Mathematics	100.8565	88.3001
Grade_2_Language	100.8565	88.3001
Grade_2_Environment	88.0675	84.7539

Different item quantities are getting as the results [Table 2]. So that this is easy to make decisions for managers to identifying the market trends.

4.6 Summary

This chapter identified the possible research process to solve the identified research problem. When evaluating each algorithm identified best algorithms for each research objective. The following chapter present the conclusion of the research.

Chapter 5

Conclusion and Future Work

5.1 Introduction

The earlier chapter described about the methodology and result of all 5 areas and this chapter will state the achievements of the conclusion and further developments.

5.2 Conclusion

In this research, the problem is to identifying proper data mining prediction for the company research areas. Those are difficulty in possibilities of a customer buying related other products after buying a particular product, difficulty in identifying repeat buyers and drop customers, difficulty in identifying connection between customer purchasing methods, difficulty in finding a group of customers to promote product/s. and difficulty in forecasting product sales process.

To achieving research objectives these are the findings for each research area. For area 1, Possibilities of a customer buying related other products after buying a particular product: Association Rule Mining is used as data mining technique. Apriori is used as data mining algorithm. For area 2, Identifying customers who are going to continuing purchasing or drop from the company and how it effects for the company: Classification is used as data mining technique. Trees J48 and Trees.RandomForest are used as data mining algorithms. For area 3, Identifying connection between customer purchasing methods: Association Rule Mining is used as data mining technique. Apriori is used as data mining algorithm. For area 4, Find a group of customers to promote a particular product/s: Clustering is used as data mining technique. Simple EM (Expectation Maximization) class is used as data mining algorithm. For area 5, Identifying product sales forecasting process (required quantity of each product for special annual events like exhibitions): Time series forecasting is used as data mining technique. Linear Regression and Multilayer Perceptron are used as data mining algorithm.

5.3 Future Developments

As a future work, the ability to increase the level of accuracy by increasing the size of the database can be tested. And for each research area can apply more accurate and efficient algorithms for modeling and evaluation.

5.4 Summary

This final chapter discussed the conclusion and future works of this research.

References

[1] Sheikh Qazzafi, "Consumer Buying Decision Process Toward Products", International Journal of Scientific Research and Engineering Development, Volume 2 Issue 5, 2019.

[2] Gaurangi Anand, Auon Haidar Kazmi, Pankaj Malhotra, Lovekesh Vig, Puneet Agarwal and Gautam Shroff, "Deep Temporal Features to Predict Repeat Buyers", TCS Research, New Delhi, India, 2015.

[3]“Omniconvert”,[Online]. Available: https://www.omniconvert.com/blog/consumer-behavior-in-marketing-patterns-types-segmentation/ [Accessed 02 5 2020].

[4] Mahsa Familmaleki, Alireza Aghighi and Kambiz Hamidi, "Analyzing the Influence of Sales Promotion on Customer Purchasing Behavior", Int J Econ Manag Sci, Volume 4, Issue 4, 2015.

[5] " Infoentrepreneurs ", [Online]. Available: https://www.infoentrepreneurs.org/en/guides/identify-and-sell-more-to-your-most-valuable-customers/ [Accessed 12 5 2020].

[6] C. Suresh, K. Anandanatarajan and R. Sritharan, "Effect of Sales Promotion Tools on Customer Purchase Decision with Special Reference to Specialty Product", Asia Pacific Journal of Research, Vol I, Issue XXI, 2015.

[7] “Infoentrepreneurs”, [Online]. Available: https://www.infoentrepreneurs.org/en/guides/group-your-customers-into-market-segments/ [Accessed 14 6 2020].

[8] “Forecast-total-market-demand”, [Online]. Available: https://hbr.org/1988/07/four-steps-to-forecast-total-market-demand/ [Accessed 14 6 2020].

[9] Zhi-Ping Fan, Yu-Jie Che and Zhen-Yu Chen, "Product sales forecasting using online reviews and historical sales data: A method combining the Bass model and sentiment analysis", Journal of Business Research 74, 2017.

[10] "GeeksforGeeks", [Online]. Available: https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/ [Accessed 18 9 2020].

[11]"Forecast-total-market-demand",[Online]. Available: https://hbr.org/1988/07/four-steps-to-forecast-total-market-demand/ [Accessed 14 10 2020].

[12] " Infoentrepreneurs ", [Online]. Available: https://www.infoentrepreneurs.org/en/guides/identify-and-sell-more-to-your-most-valuable-customers/ [Accessed 02 11 2020].

[13] "Javatpoint ", [Online]. Available: https://www.javatpoint.com/data-mining-techniques/ [Accessed 24 11 2020].

[14]"Developers.google.com",[Online].Available: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc/ [Accessed 07 12 2020].

[15] "Onlinehelp.explorance.com ", [Online]. Available: https://onlinehelp.explorance.com/blueml/Content/articles/getstarted/mlcalculations.htm?TocPath=Get%20started%7C_3/ [Accessed 26 12 2020].

[16] "Techtarget.com ", [Online]. Available: https://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining/ [Accessed 14 1 2021].

[17] A. K. Mishra, S. K. Pani and B. K. Ratha, " Association rule mining with apriori and fpgrowth using weka", International Journal of Advanced Technology in Engineering and Science, Vol 3, Issue 01, 2015.

[18] " Wikipedia.org ", [Online]. Available: https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm/ [Accessed 18 2 2021].

[19] " Towardsdatascience.com ", [Online]. Available: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1/ [Accessed 10 3 2021].

[20] " Machinelearningmastery.com ", [Online]. Available: https://machinelearningmastery.com/time-series-forecasting/ [Accessed 24 4 2021].

[21]"Mathworks.com",[Online].Available: https://www.mathworks.com/discovery/time-series-regression.html/ [Accessed 01 5 2021].

[22] " Weka.sourceforge.io ", [Online]. Available: https://weka.sourceforge.io/doc.dev/weka/classifiers/functions/SMOreg.html/ [Accessed 09 5 2021].

[23] " Wikipedia.org ", [Online]. Available: https://en.wikipedia.org/wiki/Multilayer_perceptron/ [Accessed 14 6 2021].

[24] " Machinelearningmastery.com ", [Online]. Available: https://machinelearningmastery.com/gaussian-processes-for-classification-with-python/ [Accessed 05 8 2021].