

**NIT6160 Data Warehousing and Mining**

**Data pre-processing and Data mining on mushroom and groceries datasets**

**Submitted by:**

Anju Maharjan ()

Yasoda Rai(s4623250)

**Submitted to:**

Soheila Gheisari

June, 2020

Contents

[**List of figures:** 3](#_Toc42764069)

[**List of tables:** 3](#_Toc42764070)

[**1. Introduction:** 4](#_Toc42764071)

[**2. Groceries dataset analysis using association mining technique:** 5](#_Toc42764072)

[**3. Mushroom dataset analysis using classification and clustering technique:** 11](#_Toc42764073)

[**3.1 Classification:** 12](#_Toc42764074)

[**3.1.1 Tree-based Model (Decision Tree):** 13](#_Toc42764075)

[**3.1.2 Bayes Classifier (Naïve Bayes):** 15](#_Toc42764076)

[**3.1.3 Rule based Classifier (OneR)** 16](#_Toc42764077)

[**3.2 Clustering** 18](#_Toc42764078)

[**3.2.1 K-means** 18](#_Toc42764079)

[**3.2.2 Hierarchical Clustering** 19](#_Toc42764080)

[**Conclusion:** 21](#_Toc42764081)

[**Reference:** 22](#_Toc42764082)

# **List of figures:**

[Figure 1:Plot of top 20 item sets 6](#_Toc42764044)

[Figure 2: Scatter plot 9](#_Toc42764045)

[Figure 3:Grouped Matrix 9](#_Toc42764046)

[Figure 4: Graph for 4 rules 10](#_Toc42764047)

[Figure 5: Decision tree of mushroom dataset 14](#_Toc42764048)

[Figure 6: Decision tree shown in simple type 14](#_Toc42764049)

[Figure 7:Plot after k-means clustering 19](#_Toc42764050)

[Figure 8:Cluster Dendrogram after Hierarchical clustering 20](#_Toc42764051)

# **List of tables:**

[Table 1: Confusion matrix of decision tree 15](#_Toc42764056)

[Table 2:Confusion matrix of Naive Bayes 15](#_Toc42764057)

[Table 3:Confusion Matrix of OneR 17](#_Toc42764058)

# **1. Introduction:**

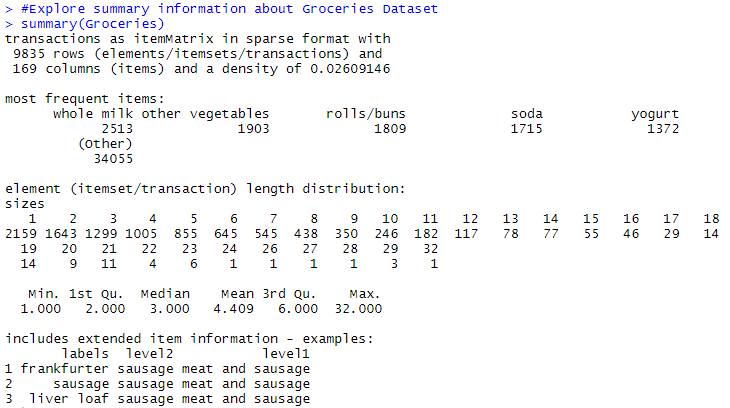
Data mining is one of the ideal solutions when we need to analyse the massive amount of data and extract the useful information. It is widely used in different fields like medical science, marketing, fraud detection, etc. Data mining is one of the steps of knowledge discovery database. It helps to discovered the unidentified properties or patterns from the large numbers of data and help to analyse the future trend. It is used by many organizations while making decision (Karapinar Senturk and Kara, 2014; Zand, 2015; Yue et al.,2018).

The main purpose of this assignment is to use data mining technique like clustering, classification, association rule mining on the datasets of groceries and Mushroom (extracted from UCI Machine Learning Repository) in RStudio. Association rule mining technique is used to find out the possible relationship between different attributes which are interesting from groceries dataset. Classification technique like decision tree, OneR and Naïve Bayes are used to construct model that can predict the class of unknown samples from the mushroom datasets. To construct the model, datasets are divided into two parts: training dataset and testing dataset in the ratio of 80 and 20 respectively. Clustering technique is also used in mushroom datasets.

# **2. Groceries dataset analysis using association mining technique:**

Association rule mining technique is used to analyse the groceries datasets. It is one of the data mining technique which is useful to find out the hidden patterns in the large amount of data. It helps to identify the relationship between features. They explore which item sets occurs together frequently and which item sets are correlated. They generate the association rules which helps to identify the relationship between different item sets (Sharma et al., 2014). For example, the rule {chocolate, coke}{chips} found from the groceries data would indicate that if customer buys chocolate and coke together, they are likely to also buy chips. This information can be used to make decision for marketing activities. Large datasets can have many association rules so interesting association rules can be selected by using confidence and support. Support show how frequently the item sets appears in the dataset and confidence tells how often the rules has been found to be true (Association rule learning, 2020).

At first, summary of Groceries dataset were viewed before using association rule mining.



We can see 9835 transactions and 169 item sets within the Groceries Dataset.

Then, top 20 itemsets of Groceries dataset are plotted as below:

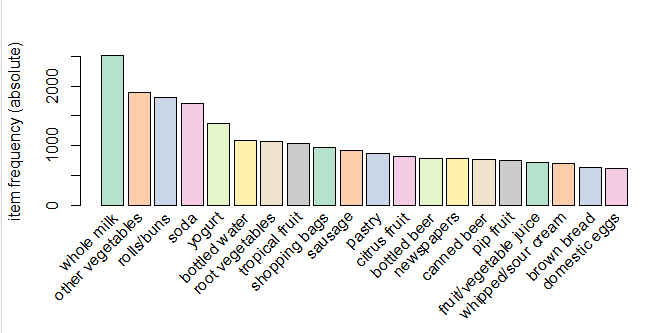
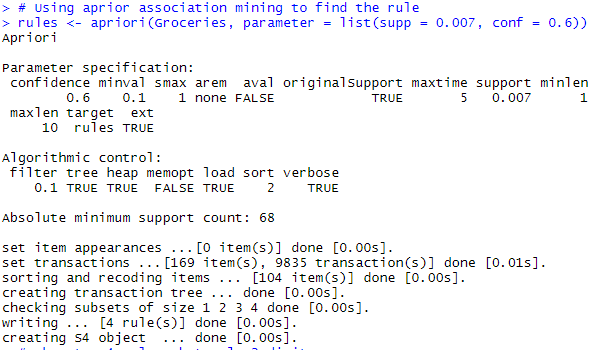
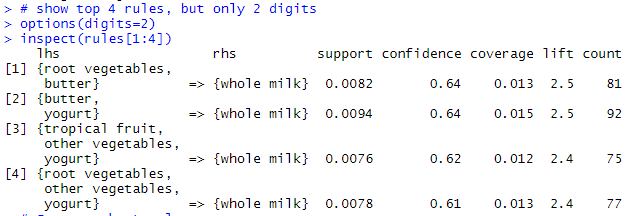


Figure 1:Plot of top 20 item sets

In order to find the interesting rules from this large dataset, we can apply apriori association mining rules. Some of the terms in association mining rules are support, confidence and lift. Suppose we have chocolate and coke, then support represents the frequency of item sets in the dataset. Confidence represents how likely the chocolate is bought when coke is bought. Lift check the confidence that chocolate will be bought if coke is bought. The below code is used to find the interesting rules with support=0.007 and confidence=0.6:

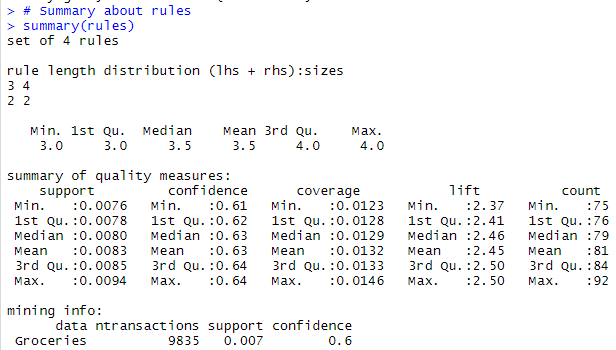


The following code was run to show top 4 rules:

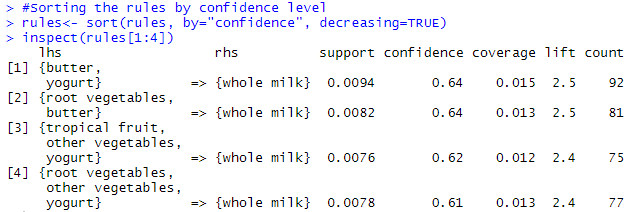


It shows in rule 1 that 0.0082% of all the order represent lhs and rhs combination where as 0.64% of the order that contain root vegetables and butter are likely to buy whole milk.

The below code gave the summary of 4 rules:



The following code were run to sort the rules on the basis of confidence:



It shows in rule 1 that 0.0094 % of all the order represent lhs and rhs combination where as 64% of the order that contain butter and yogurt also had order whole milk. Lift for this rule is 2.5 times increase in expectation that someone will buy whole milk, when we know that they bought butter and yogurt. Hence greater the lift, probability of occurrence of lhs and rhs are dependent of each other. High support and high confidence also lead to give good result. Hence, this rule is useful for making prediction. Rule 1 have high support, high confidence and high lift compare to others. Rule 2, 3 and 4 are also useful as they also have higher lift.

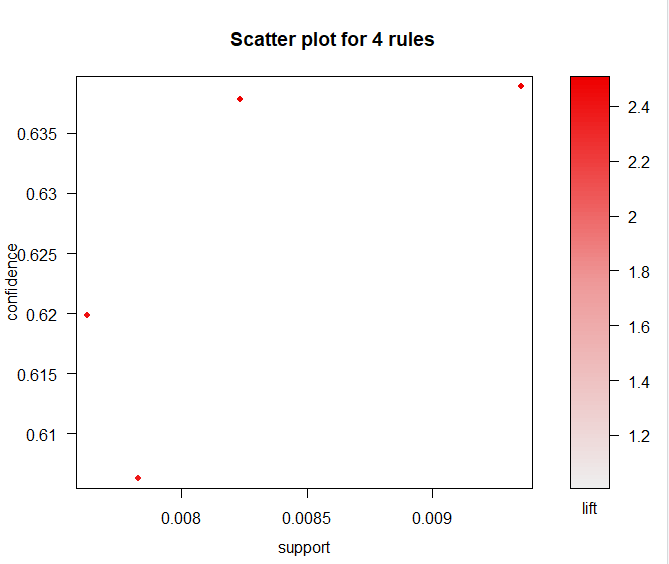


Figure 2: Scatter plot

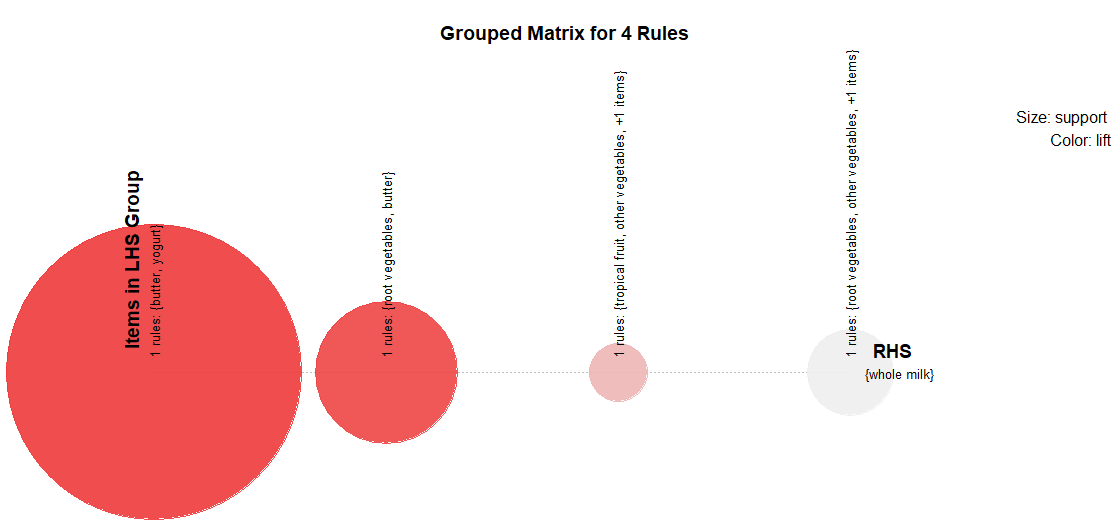


Figure 3:Grouped Matrix

Color represents the lift and size of the ball represent the support. This means ball should be big in size and should be darker in color to give interesting rules. It is shown in graph that butter and yogurt have big ball with darker color. It represents that person buying butter and yogurt are more likely to buy whole milk. So, this rule is good as compare to others three rules.

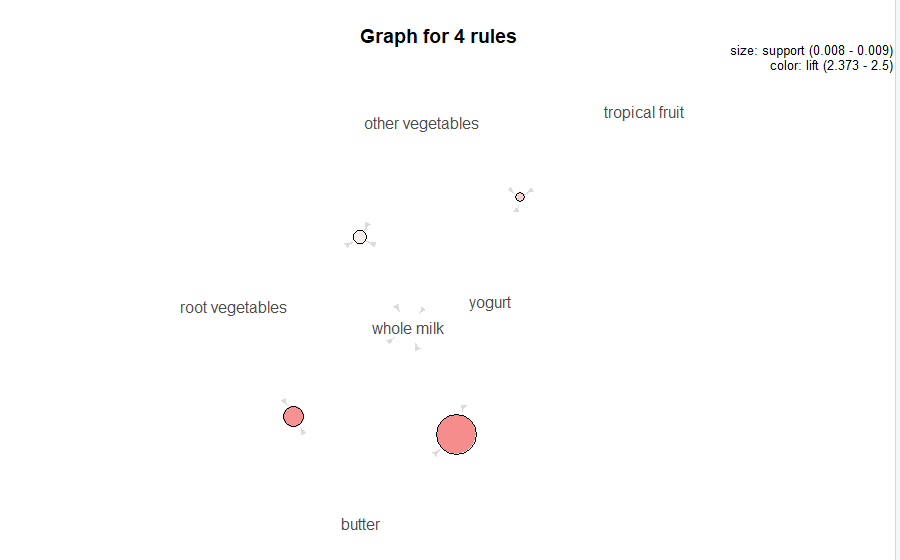
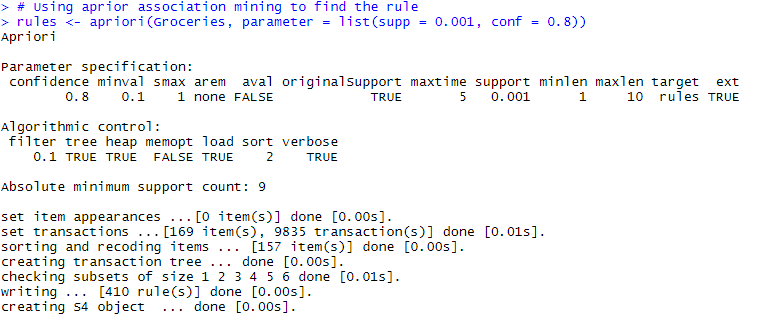
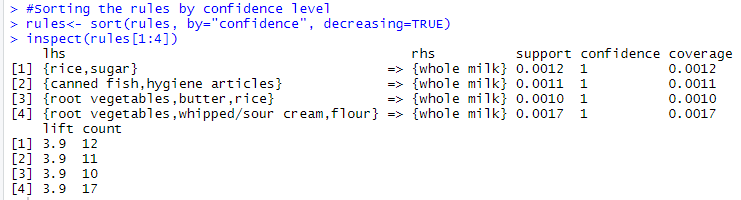
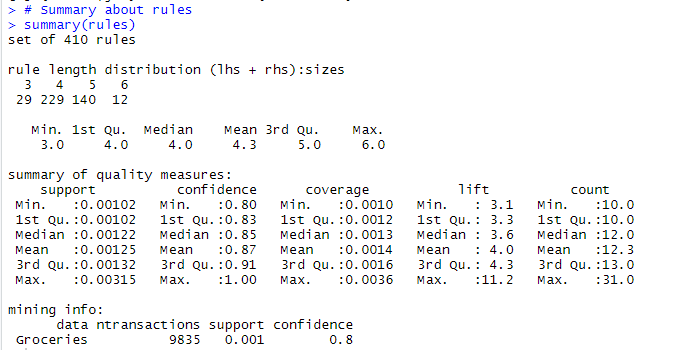
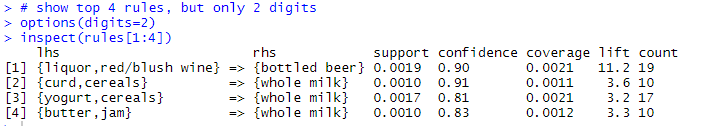


Figure 4: Graph for 4 rules

**Apriori algorithms with 0.001 support and 0.8 confidence:**





When confidence was increased to 0.8, it gave very interesting rules. As we can see in rule 1, 0.0012% support was seen with 100% of probability that person buying rice and sugar are more likely to buy whole milk. It also shows that lift was increased to 3.9 times more. The above rules are good to make prediction as they have high confidence and high lift.

When we compare the two result with different support and confidence, we found that different support and confidence can give different rules.

# **3. Mushroom dataset analysis using classification and clustering technique:**

First of all, the mushroom datasets structure is view in R software. Data of the mushroom datasets were character data type. The data types of the whole datasets were changed into factor data type.

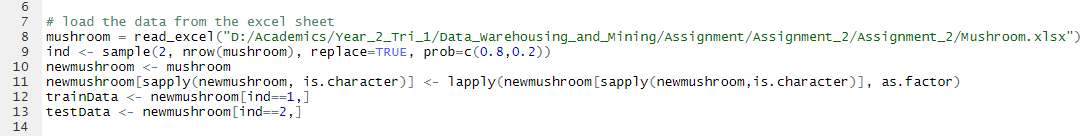
## **3.1 Classification:**

Classification is the supervised machine learning technique which means it learns from its experience. This technique is used to identify the class of unknown cases into predefined class label. Dataset are divided into two parts: training and testing datasets. Training dataset are used to build the model and testing dataset are used to identify the performance of model using different performance metrics (Sharma et al., 2013). Since, model can learn from training dataset so the numbers of training datasets should be more to get more accurate result. We had used mushroom dataset which has class label as edible and poisonous. Classification techniques like decision tree, Naïve Bayes and Rule based classifier (OneR) were used to identify the class of mushroom.

Steps for classification of mushroom dataset:

1. First, packages required for classification were installed.
2. Dataset were imported in RStudio.
3. Data pre-processing were done.
4. Dataset attributes were character data types which were converted to factor data types.
5. Dataset were divided into 80% training and 20% testing dataset. Training dataset were used to construct model and testing dataset were used to check the model validity.
6. Finally, accuracy of model was checked.

The following code were used to split the datasets into 80% training and 20% testing data and it also contain the code to change the attributes data types to factor data types:

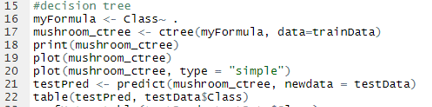


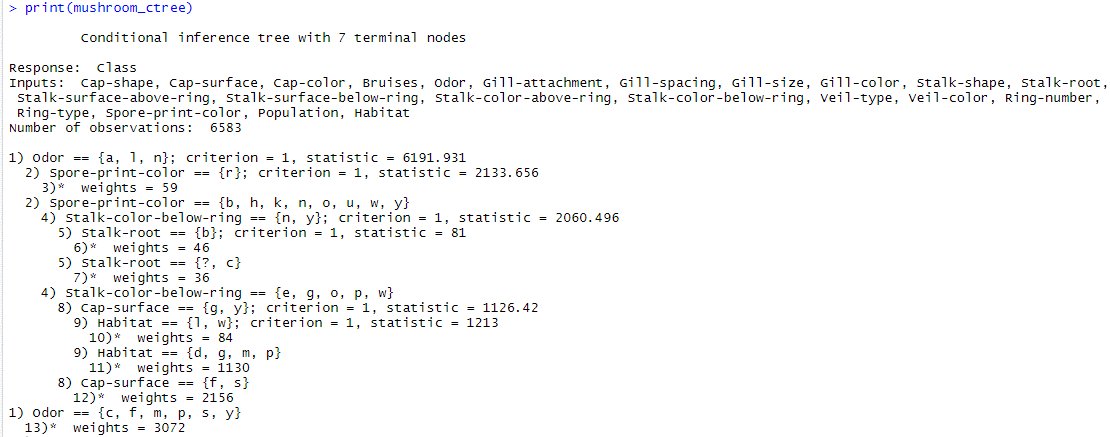
We had used ctree function to create a decision tree. It consists of formula and data. Formula represent the decision model used to predict. Data tells the function about the datasets from where the attributes need to be taken (Wong, 2020). Here in our case, we had used myFormula as fomula. We have trainData and testData as data.

The below mentioned code were used to create tree-based model. It also contain the code to print ctree of mushroom datasets, plot of ctree and the confusion matrix of the tree:

### **3.1.1 Tree-based Model (Decision Tree):**

Decision tree create model like a tree which has root, leaves and branch node (Neelamegam and Ramaraj, 2013).





When pirnt(mushroom\_ctree) code was runned, it showed the above result. It shows tree with 7 terminal nodes which weighted 3072.

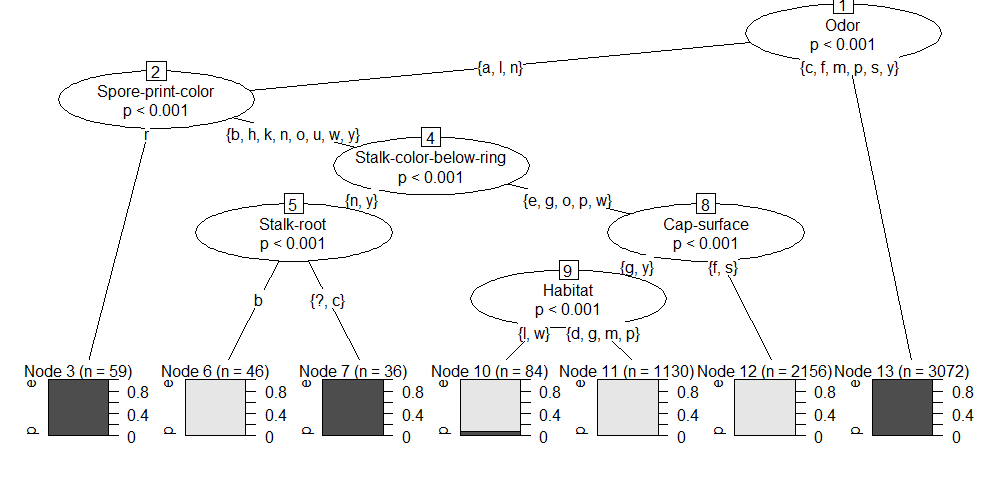


Figure 5: Decision tree of mushroom dataset

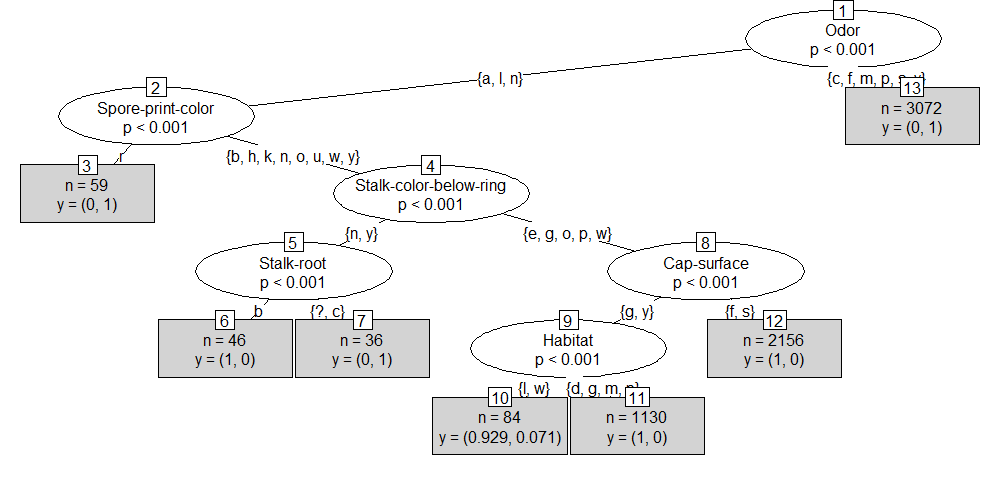


Figure 6: Decision tree shown in simple type

The below table shows the prediction made by decision tree.

|  |  |  |  |
| --- | --- | --- | --- |
| **Predicted** | **Actual** | |  |
| **e** | **P** | **Total** |
| **e** | ***798*** | ***2*** | **800** |
| **p** | ***0*** | ***741*** | **741** |
| **Total** | **798** | **743** |  |

Table 1: Confusion matrix of decision tree

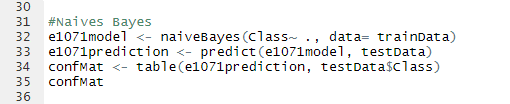
It shows that the 798 edible mushrooms were correctly identified whereas only 2 were incorrectly identified as edible. However, it had correctly identified all the poisonous mushroom correctly.

Decision tree gave the accuracy of 99.87%. We can conclude that decision tree technique is useful for predicting the poisonous class label of mushroom dataset.

### **3.1.2 Bayes Classifier (Naïve Bayes):**

Naïve Bayes identify the class by calculating the probability of the class. It assumed that the attributes are independent with each other (Rashmi, Lekha and Bawane, 2015).

The following code were used to apply Naïve Bayes algorithms to predict the class of the mushroom:



The below table is the confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
| **Predicted** | **Actual** | |  |
| **e** | **P** | **Total** |
| **e** | ***786*** | ***65*** | **851** |
| **p** | ***12*** | ***678*** | **690** |
| **Total** | **798** | **743** | **1,541** |

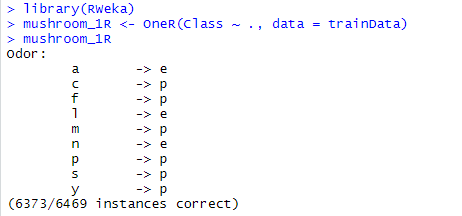
Table 2:Confusion matrix of Naive Bayes

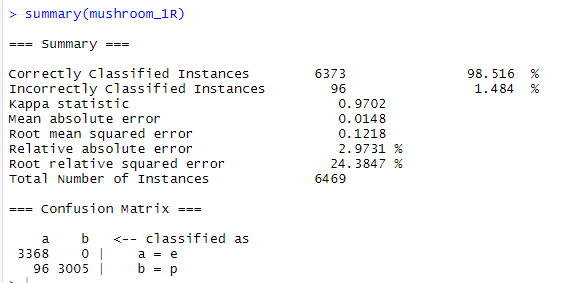
It shows that out of 851, 786 edible mushrooms were correctly identified where 65 mushrooms were incorrectly identified as edible. Out of 690 poisonous mushrooms, 678 were correctly identified whereas 12 were identified wrongly as edible. It was 95.003% accurate.

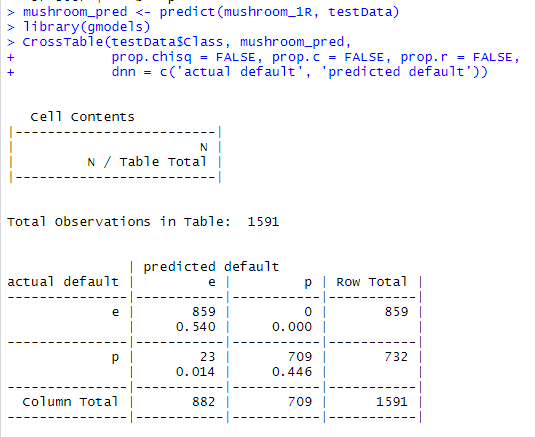
### **3.1.3 Rule based Classifier (OneR)**

It is the classification technique where numbers of rules are generated that test one specific features (OneR, 2020).

The following codes were used to find the confusion matrix of OneR using training data and testing data:







The below table is confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
| **Predicted** | **Actual** | |  |
| **e** | **P** | **Total** |
| **e** | **859** | **0** | **859** |
| **p** | **23** | **709** | **732** |
| **Total** | **882** | **709** | **1,591** |

Table 3:Confusion Matrix of OneR

It shows that 859 mushrooms were correctly identified as edible whereas identify 709 as poisonous mushrooms correctly. But the table shows that 23 poisonous mushrooms were identified as edible. It shows the accuracy of 98.516%.

From the above experiment, we found that decision tree was identify the edible and poisonous mushrooms more accurately than that of Naïve Bayes and OneR. It had 99.87% accuracy. Naïve Bayes and OneR had accuracy of 95.003% and 98.516% respectively.

## **3.2 Clustering**

Clustering is the process of splitting the data into several groups such that data points within the same groups have more similar properties. In contrast, data points in different groups have profoundly dissimilar features. If we see most of the data in real-world, they do not have any specific class defined. Clustering is useful in dealing with such data. Following are few of the popular applications of clustering:

* Medical imaging (for any disease detection such as breast cancer)
* Segmenting images
* Detecting any anomaly in the data
* Analysing social network

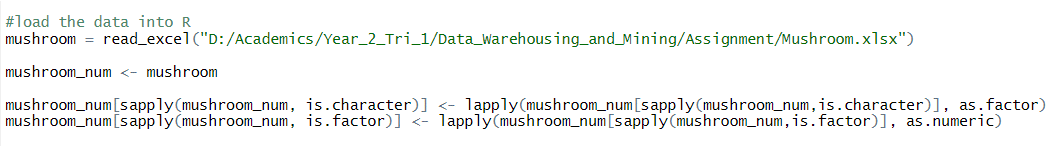
In this assignment, k-means and hierarchical clustering algorithms are applied for the analysis. The given dataset “mushroom” contains character as labels for defining the attributes. Hence before performing clustering technique, we need to convert the attributes label into numeric data.

### **3.2.1 K-means**

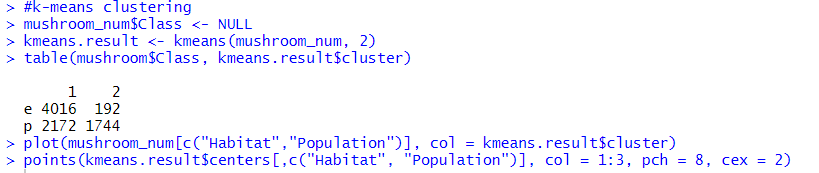
K-means is one of the most popular clustering algorithms. It is easy to understand and implement.

1. At first, select the desired number of class or cluster and randomly assign every data point to some cluster.
2. Based on the assigned points, recompute the centre of the group.
3. Reassign every point to the nearest group centre.
4. Repeat steps 2 and 3 until no significant changes are seen.

The following codes are used to load the data into R and convert all character values into numerical value and are repeated before implementing every clustering algorithm in the assignment.



After loading the data, we removed the Class of the clustering data set. However, later it is used to compare how accurate and effective this clustering algorithm is. Since the dataset contains only two groups, the desired number of clusters is also set to be 2. Additionally, we have plotted the result to have some visual observation.



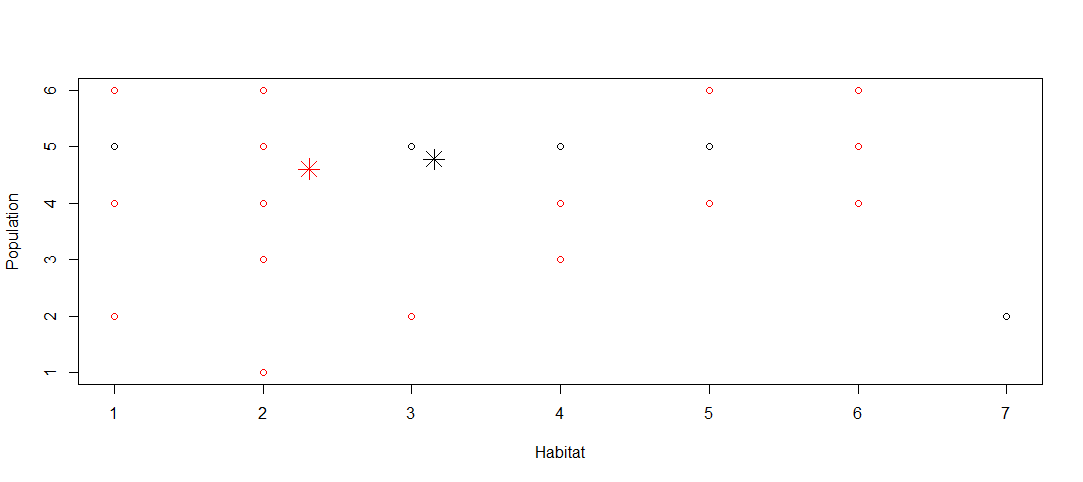


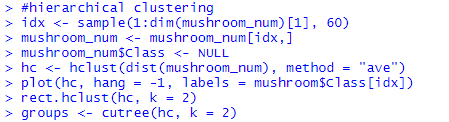
Figure 7:Plot after k-means clustering

As observed in the figure above, the data points are scattered and are not properly classified. The main aim of good clustering algorithm is to have similar data points in the same cluster. The one of the reasons for scattered data points could be the conversion of categorical data into numerical data.

### **3.2.2 Hierarchical Clustering**

Hierarchical clustering creates hierarchy or classes or clusters. All the data points are assigned to have to their own cluster. After that, two closest classes or groups are merged to form the single group. The process ends when a single cluster is formed.

As in K-means, the data loading and conversing character into numerical data is done at the beginning. Only 60 records are used so that the clustering plot will not be overcrowded. Similar to K-means, Classattribute is removed.



At the end, the cluster dendrogram is plotted as shown in the image below.

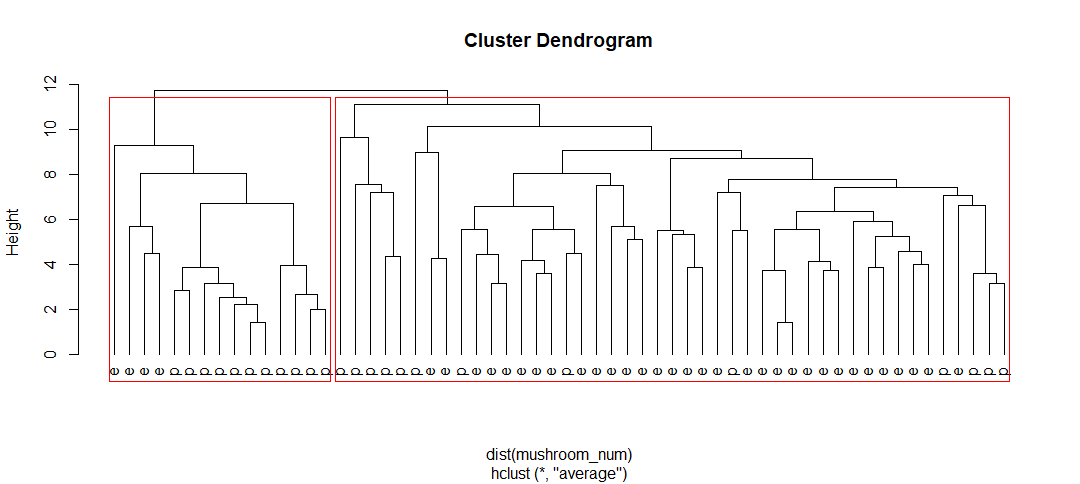


Figure 8:Cluster Dendrogram after Hierarchical clustering

As observed in the figure above, the data points are not properly clustered. Two classes “edible” and “poisonous” are overlapping in most of the cases.

# **Conclusion:**

RStudio is an open source software that uses R programming language. It allows us to use different data mining algorithms to find out the useful patterns from massive amount of datasets. We had used apriori association rule mining technique to find out the 4 interesting rules from the Groceries datasets. We had found that support and confidence can influence the result.

Classification algorithms namely decision tree, Naïve Bayes and OneR were used to analyze the mushroom datasets. It gave the result that decision tree works better than others.

Likewise, clustering like K-means and Hierarchical clustering technique were used to extract useful information from mushroom dataset.

This proposed study can be helpful for predicting future trends.

# **Reference:**

1. En.wikipedia.org. 2020, *Association Rule Learning*, viewed 10 June 2020, <<https://en.wikipedia.org/wiki/Association_rule_learning>>
2. George Seif, 2018, *The 5 Clustering Algorithms Data Scientists Need to Know,* M Towards Data Science, viewed on 08/06/2020, <https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68>
3. KARAPINAR SENTURK, Z. & KARA, R. 2014. Breast Cancer Diagnosis Via Data Mining: Performance Analysis of Seven Different Algorithms. *Computer Science & Engineering: An International Journal,* 4, 35-46.
4. NEELAMEGAM, S. & RAMARAJ, E. 2013. Classification algorithm in data mining: An overview. *International Journal of P2P Network Trends and Technology (IJPTT),* 4, 369-374.
5. RASHMI, G. D., LEKHA, A. & BAWANE, N. Analysis of efficiency of classification and prediction algorithms (Naïve Bayes) for Breast Cancer dataset. 2015. IEEE, 108-113.
6. Saedsayad.com 2020, *Oner,* viewed 11 June 2020, Available at: < <https://www.saedsayad.com/oner.htm#:~:text=OneR%2C%20short%20for%20%22One%20Rule,each%20predictor%20against%20the%20target.>>
7. Saurav Kaushik, 2016*,* *An Introduction to Clustering and different methods of clustering,*Analytics Vidhya, viewed on 08/06/2020, <https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/>
8. SHARMA, S., AGRAWAL, J., AGARWAL, S. & SHARMA, S. Machine learning techniques for data mining: A survey. 2013. IEEE, 1-6.
9. Wong, K. 2020, *Chapter 24: Decision Trees*, viewed 10 June 2020, <<https://ademos.people.uic.edu/Chapter24.html#:~:text=The%20function%20ctree()%20is,xtrafo%2C%20ytrafo%2C%20and%20scores.&text=formula%3A%20refers%20to%20the%20the,are%20using%20to%20make%20predicitions.>>
10. YUE, W., WANG, Z., CHEN, H., PAYNE, A. & LIU, X. 2018. Machine learning with applications in breast cancer diagnosis and prognosis. *Designs,* 2, 13.
11. ZAND, H. K. K. 2015. A comparative survey on data mining techniques for breast cancer diagnosis and prediction. *Indian Journal of Fundamental and Applied Life Sciences,* 5, 4330-4339.