

**TDS3301**

**Data Mining**

Group Assignment

|  |  |
| --- | --- |
| **Name** | **Student ID** |
| Jaiprashanth A/L Ramalingam | 1131122948 |
| Farhanuddin bin Hamdan | 1132702175 |
| Kevin Lee Xin Zhe | 1132700900 |

Table of Contents

[**Part 1 : Exploratory Data Analysis** 3](#_Toc473884890)

[**A.** **About The Dataset** 3](#_Toc473884891)

[**B.** **Possible Insights** 4](#_Toc473884892)

[**C.** **Data Mining Approaches** 5](#_Toc473884893)

[**D.** **Data Quality Issues** 6](#_Toc473884894)

[**E.** **Preprocessing Task** 7](#_Toc473884895)

[**Part 2: Association Rule Mining** 8](#_Toc473884896)

[**Objectives** 8](#_Toc473884897)

[**Data set description** 9](#_Toc473884898)

[**Rule mining process** 10](#_Toc473884899)

[**Rule Summary** 11](#_Toc473884900)

[**Recommendations** 12](#_Toc473884901)

# **Part 1 : Exploratory Data Analysis**

## **About The Dataset**

In general, the people in our group are not really movie-goers. However, when we do want to watch a movie, we would prefer it to be good and interesting. We cannot even rely on famous movie rating review websites like IMDB and Rotten Tomatoes because the users may generate biased reviews. The dataset that we have obtained for this assignment is the IMDB 5000 Movie Dataset.

The IMDB 5000 Movie Dataset was created by chuansun76 from the Kaggle website. This dataset comprises of 5, 043 different movies which is scraped from the IMDB website using a Python library called “scrapy”. Using this library, he managed to obtain 28 unique variables from the 5, 043 movies which spanned across 100 years from 66 different countries. In simpler words, there are 28 columns which describes 5, 043 movies, e.g. color, director’s name, duration and genres.

There are a few insightful columns such as the columns which hold values for the amount of Facebook “Likes” obtained by either the movie, the cast, actor, or director. We hope to identify an interesting pattern revolving the social media and movie rating.

## **Possible Insights**

For each year, several thousands of movies are produced. After movies are released, many reviewers on IMDB will give their reviews, opinions and critics so that the future viewer can use that as a guide on whether the movie is worth the time to be watched or not. So, is there a better way for us to identify a movie rating without relying on critics? With this dataset, we seek to find possible patterns. Based on the chosen dataset, we can acquire various information related to past movies. Then we intend to relate the past movies’ information for predicting the rating of future releases. For example, the new movie features Uvuvwevwevwe Onyetenyevwe Ugwemuhwem Osas as the cast for the main character. If that actor has a high amount Facebook “Likes”, it might be possible to assume that the movie will turn out good and the gross revenue generated will be high as well. Of course, other factors can affect the result like the involvement between the actors and directors. For example, by having “Stan Lee” as the director and “Dwayne ‘The Rock’ Johnson” as the main cast will likely to be a good movie due to their massive experience in the filming industry. Thus, we can assume that the new movie also is going to be good.

Aside from a movie rating, we want to uncover a possible tie with the duration of the movie and its success in the box office after being released. This can help set an accurate benchmark to predict how long a movie should be to ensure that it will get a good review or rating. Some viewers might dismiss unacceptably short or long movies because the content was either was not expanded properly or too detail was given on insignificant aspects.

## **Data Mining Approaches**

In our opinion, we find *Association rule mining* as the most relevant data mining technique to be used on our selected dataset. The reason is because we would like to analyze and predict the movie watcher’s reaction with the variables given in the IMDB dataset. As a possible end-result, a model can formed and it may be able to predict movie ratings before a movie is released in theatres. For example, we can associate the IMDB rating with the sum of Facebook Likes of every lead actors involved of the movie, the budget allocated for each movie and the country of origin or etc. . . . Another use of the model may be beneficial to the movie production team as the model might be able to predict the gross income.

However, clustering can be considered as another relevant data mining technique that we can consider to uncover new patterns. For example, we can map out a possible relationship between the number of users of IMDB with the number of Facebook likes for either the movie, the cast, director or individual actors. This could contribute to identify the authenticity or legitimacy of IMDB movie rating.

## **Data Quality Issues**

Within the dataset, we can assume the data is “dirty” and issues exist in most of the columns. This results in inconsistent evaluations or calculations. These dirty data plays the part of making the dataset becoming untidy. For example, there are Not Available(NA) data exist in the column named “Gross”, this means that some of the movies gross will be unknown to us and hence missing one important factor of determining the movie’s popularity and related attributes.

There were multiple values in a single observation. For example, the column “Genres” holds multiple values in most rows. For example, while there are rows with value just like ‘Action’, there are also rows with multiple ones such as ‘Action|Adventure|Fantasy|Sci-Fi’. Arrangement like this will make the dataset to be more difficult to sort which producing inaccurate information.

Another issue that was present is inconsistency in the data existed in the Color column. There are two labels for the Color variable, “Color” and “Black and White” but there are another label present which is “color”. This causes inconsistency in data.

## **Preprocessing Task**

Please kindly refer to the link below for the Github repository of our R scripts which we made to preprocess the dataset.

<http://bit.ly/2iuJaI5>

# **Part 2: Association Rule Mining**

## **Objectives**

What is the domain and what are the potential benefits to be derived from association rule mining? This is high level - not find patterns, but what would improve because of the use of the patterns?

The domain that can be derived from association rule mining is set of objects that are associated or correlated among themselves. The domain consists of patterns that regularly occur in it. In other words, it could actually be applied in a wide range of fields, such as advertising and medical diagnosis.

By applying association rule mining, one can predict future behavior based on analysis of the frequent patterns. For example, association rule mining is used in market basket analysis, it plays an important part of predicting the likelihood of a customer to purchase a certain product.

The use of patterns improves the development of technology fields that involves machine learning. If a program that is capable of machine learning is given a frequent pattern, it can proved to be efficient without being explicitly programmed.

## **Data set description**

What is in the data, and what preprocessing was done to make it amenable for association rule mining. Where choices were made (e.g., parameter settings for discretization, or decisions to ignore an attribute), describe your reasoning behind the choices.

We chose to use the Extended Bakery dataset of 1000 receipts. The CSV file we downloaded consisted of sparse vectors where each transaction has a record of what food items and drinks were purchased together in a single receipt at the bakery. E.g. The first transaction shows a customer purchasing Coffee Eclair, Blackberry Tart, Bottled Water and Single Espresso in one receipt.

Data preprocessing was necessary for the dataset we selected. We first dropped the first column which doesn’t provide any value to contribute to mining the possible relationship between food items and drinks. We also updated the headers to understand the different items we have but it doesn’t indicate the purchasing order of the item. Then, we substituted the codes given with the actual name of the item to get a better understanding.

For the parameter settings, we discretize the columns by changing the structure type to “Factor” and removing the first column containing the transaction number. The reason is to allow converting the data frame into transactions.

1. **Rule mining process**

Parameter settings, choice of algorithm, and the time required.

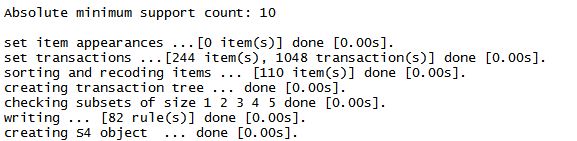
**Parameter setting:**

https://lh5.googleusercontent.com/U8vAcZ6fDAGnGRiNx-EfPoEADvYvuFD19fcdurI1tUa4f-hlI8-CRe4vF3-sb2DmHnUfv6S0eYwpgggOYInwJ1JnIziaMolhYodHfxszXmc4KC1p_UzXiIlyEwoN8lkjSJEVYkM4

**Choice of Algorithm:**

The Apriori algorithm.

**Time Required:**



We have demonstrated a simple example of applying the Association Rule Mining on the mentioned dataset. The source code is available at the following link:

<http://bit.ly/2iuJaI5>

## **Rule Summary**

Summary (number of rules, general description), and a selection of those you would show to a client.

In summary, after doing the preprocessing, some of the association rules that contains items that does not appear frequently were pruned. In the end, we were left with 23 rules.

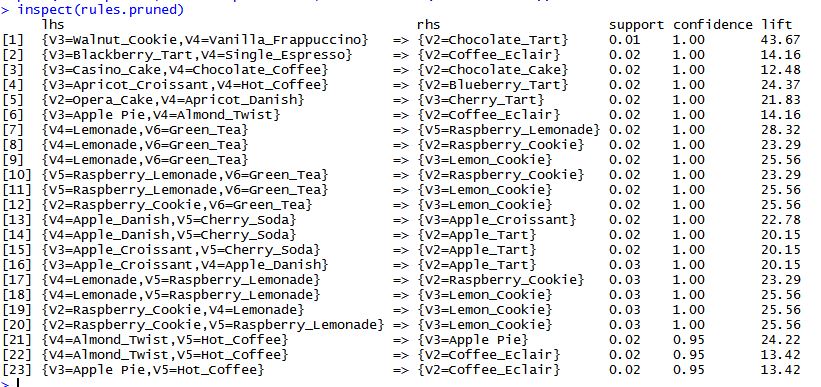


Figure 1

Within these 23 rules, only a selected few will be shown to our client. These few are the rules with a high lift. We choose to show the rules with a high lift because lift is a measure of a performance of an association rule at predicting cases as having an enhanced response. A rule is doing a great job if the response is much better than the average for the population as a whole. Lift is simply the ratio of this. This means that a rule with a high lift, tends to be the most reliable rule of all the rules. For e.g., I would tell my client from the bakery to sell Walnut Cookie and Vanilla Frappuccino as a combinational offer which includes a free Chocolate Tart to guarantee some profit.

In another example, the customer who bought Blackberry Tart and Single Espresso, tends to buy Coffee Eclair as well. So, by suggesting to the client, to put the Coffee Éclair next to the Blackberry Tart on the shelf will most likely to make customers at the bakery buy the food products. The given examples are simple but the application of Association Rules Mining can clearly shows that it can assist with product marketing for even small businesses.

## **Recommendations**

First recommendation on what our client should do about the rules discovered is pairing up the items shown in the LHS column, based on Figure 1 with the items shown in RHS column for weekly promotion. A bundle promotion of these pairs will most likely boost the customer satisfaction along with profits.

Second recommendation is to ask our client to act in reverse direction, the bakery can save money by reducing the production of a set of pastry and saving resources for more profitable products instead.

# **Part 3: Classification**

## **A. Exploratory data analysis**

About the dataset:

The dataset is related to the student achievement two Portuguese secondary schools. The data attributes include student grades, demographic, social and school related features. The two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por).  The attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1.

For Part 3, we will perform classification on the dataset for Mathematics subject to classify students who scored more than 10 out of 20 for G1, G2 and G3 consecutively. We can regard the

## **B. Pre-processing tasks**

For the preprocessing task, we will remove attributes which we assumed that doesn’t contribute to the classification task. We removed the following attributes:

* School
* Address
* Famsize
* Pstatus
* Medu
* Fedu
* Mjob
* Fjob
* Reason
* Guardian
* Traveltime
* Nursery
* Higher

## **C. Choice of performance measures**

In the case for decision tree classification technique, we used the K-fold cross validation performance measure. Decision tree have nodes which is the product of probability made with our algorithm. Decision tree is simple to use, does not cost a lot of resources to make one, extremely fast at classifying unknown records. The reason we choose to the mentioned performance measure was to measure the error of misclassification and this led us to prune the tree to reduce the errors.

In the case for Naïve Baiyes classification, the results should be produced in a form of confusion matrix, indicating the probability of each variable affecting the proficiency in Mathematics by observing the High column for each student. The confusion matrix describes the probability of being classified as Proficient with each of the factor type columns found in the dataset. It serves as a result of the Naïve Bayes as well as the performance measure for the classifier.

For ANN, we used a confusion matrix and a prediction plotting (Training data vs Test data). The reason of selecting those two measures to tabulate the classification results and visualize the difference between the training and test data for prediction purposes. As shown in Figure 4 below, the confusion matrix can inferred like so: 0 is equals to “No” and 1 is equals to “Yes”. This refers to the student’s proficiency in Mathematics (G3 > 10).

## **D. Performance of the 3 classifiers**

Decision Tree:

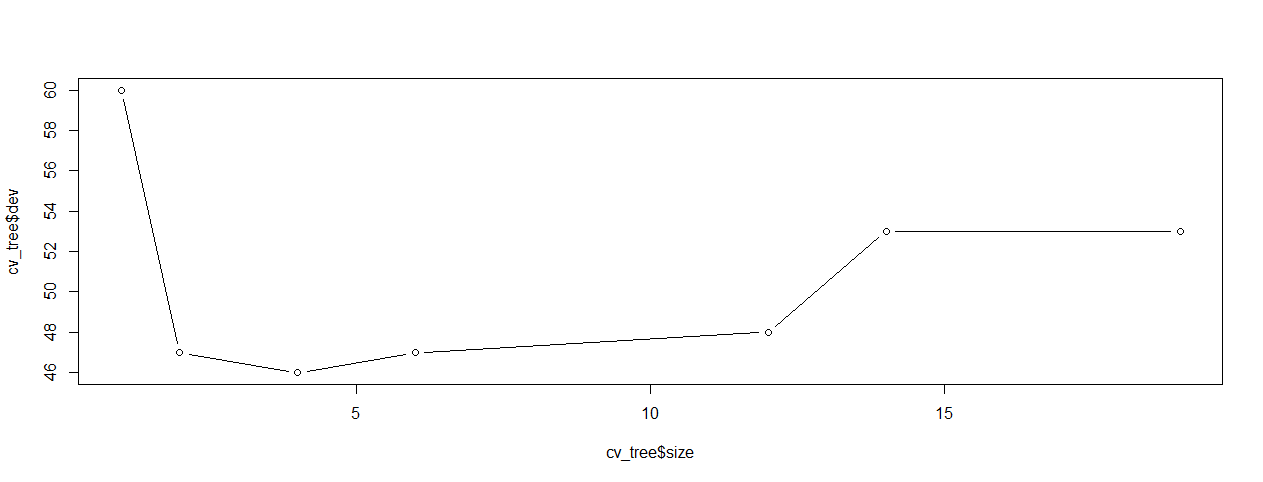


Figure 2

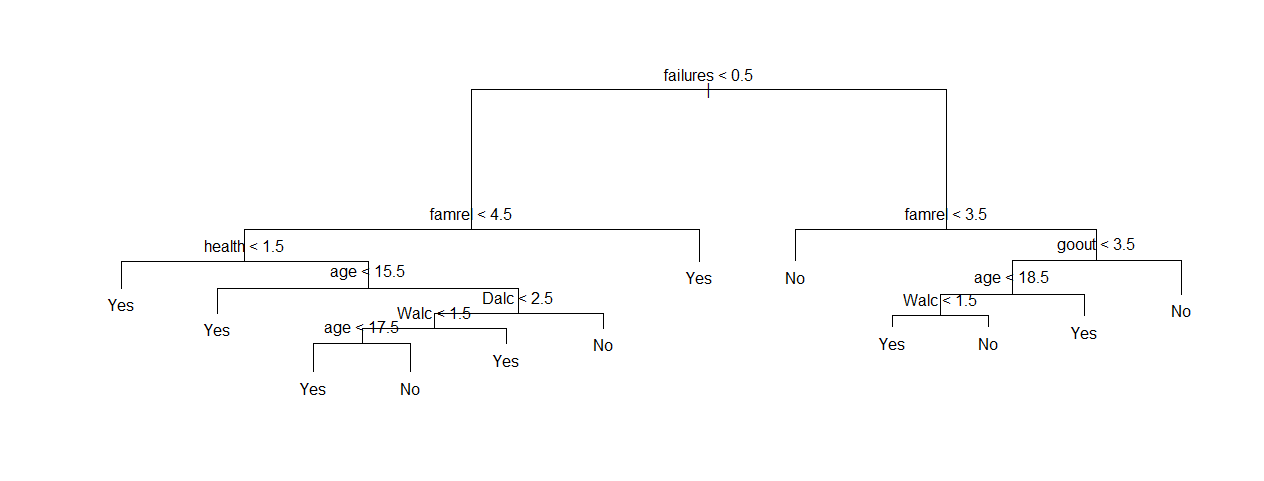


Figure 3

Naïve Bayes:

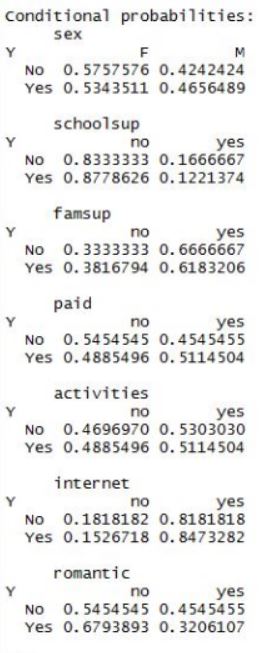


Figure 4

ANN:



Figure 5

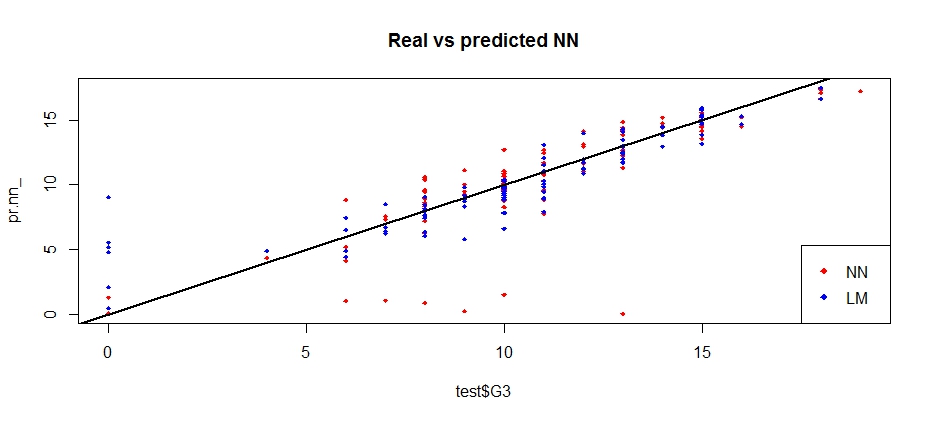
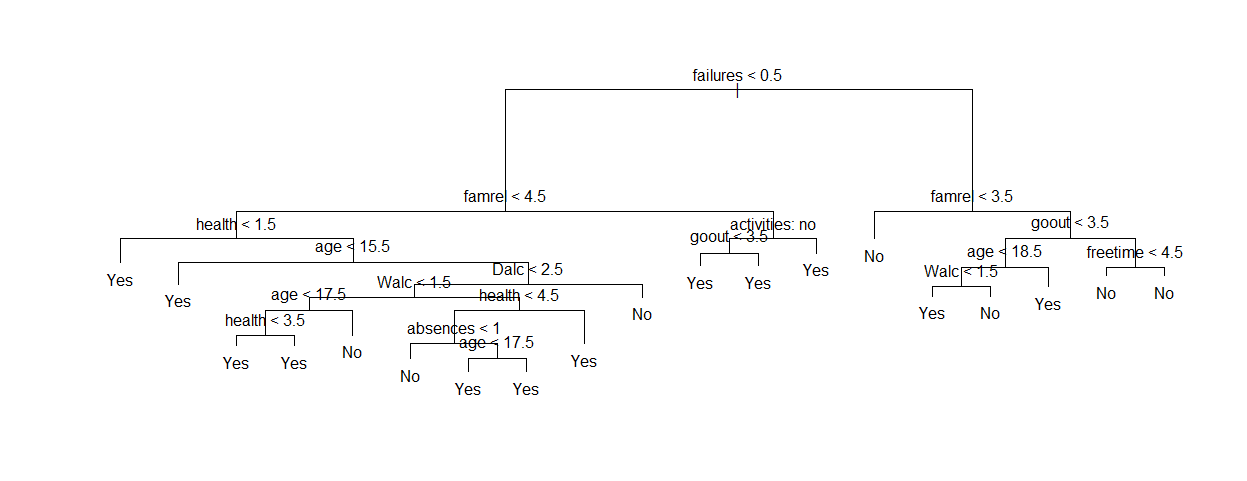


Figure 6

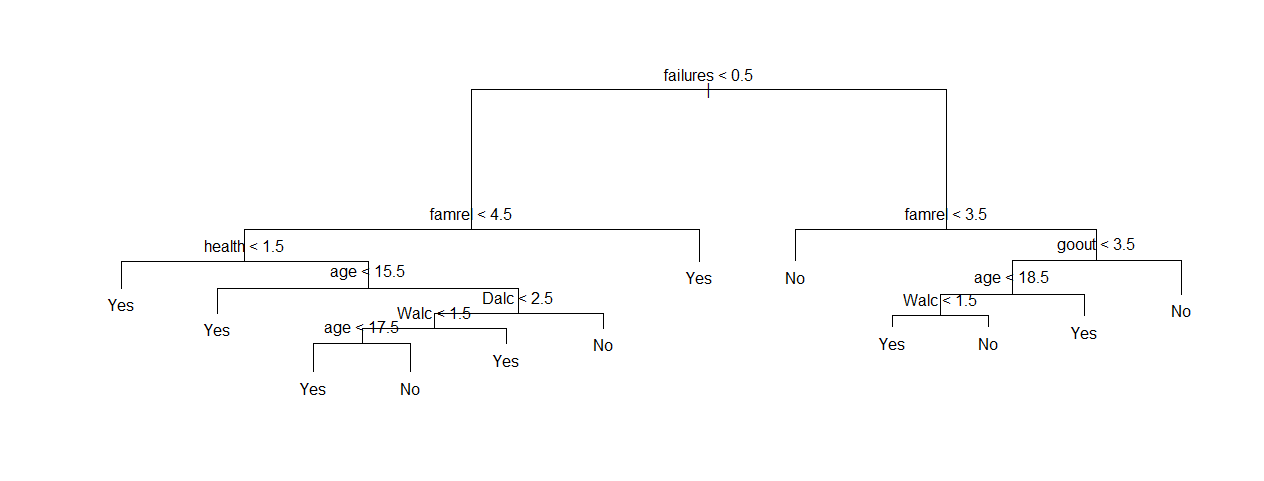
## **E. Suggestion as to why the classifiers behave differently**

Decision tree:

For this method, decision tree classifier requires the data type to be of Integer type. This enables the tree to be generated and for each column, the data split into different ranges according to the data and this results in the multi-level branching of the decision tree.



Picture above shows the result of our decision tree. Note that for our case, we take G3 (Final exam marks for Mathematics) more than 10 marks as the partition for the whole decision tree. The highest value for all categories will be chosen as the final nodes.



Picture above shows the result of our decision tree after cross validation and pruning of the decision tree.

Naive Bayes:

The classifier works well with columns of data type ‘Factor’. Thus, the numerical columns were excluded from this classification method unlike the Decision Tree and ANN classifiers which require numerical values to generate the classification output.  The image below shows the result of the Naive Bayes classification in the form of confusion matrices. The Naive Bayes uses “Yes” and “No” value from the “High” column which we made to form the classification according to each respective variables such as the sex, schoolsup and so on. For e.g., we can say there is 53.4% of a female student being proficient in Mathematics.  We can continue making more inferences for the remaining variables based on the result we obtained.

ANN :

ANN Classifier behaves differently because we require to train the classifier first with a training data so it can classify the data given for the assigned task. However, we approached the preprocessing for this classifier similar to the way we make a Decision Tree. We converted the factor type columns with two levels(Yes and No) into integer type (1 and 0) so we can pass it through the ANN for classification.