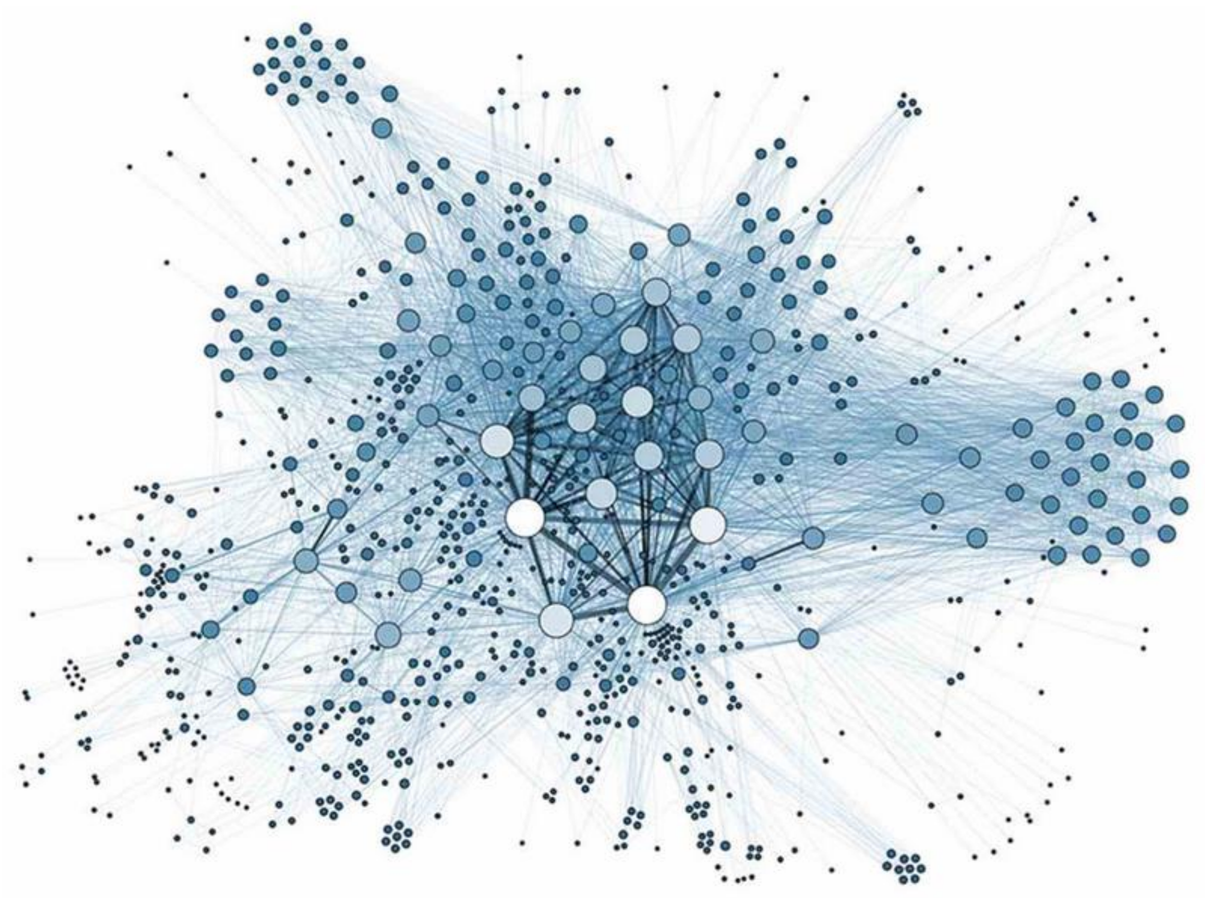


ASDM Assignment: Data Mining using SAS and R

Carlos Beltrán

2018/19

INDEX

TABLE OF CONTENTS

APPLY CLASSIFICATION ON DATASET USING R & SAS

INTRODUCTION

Since I was a teenager I have been interested in the financial industry and how It impacts the rest of society. The credit's market plays a significant role in how families, business and public sector plan their projects. Data Science is increasing the capability of financial institutions to assess correctly the risk implied on granting of credit. It will impact on the markets of credits since financial institutions will be able to assign more appropriate interests rates depending on the customer's profile and It will cause the interest rate to lower because of the significant efficiency.

AIM AND OBJECTIVE OF THE TASK

This task aims to present the classification approaches transparently and apply Random forest, one of the essential techniques within the classification's methods, to a loan Dataset.

This work will explain how to implement Random Forest on a dataset for classification purposes, how to predict whether a customer will pay back the loan or not and an assessment whether the model has correctly predicted the outcome.

BRIEF LITERATURE REVIEW

Classification [1] is a statistical technique used for predicting, classifying and categorizing to which of a set of categories a new observation belongs.

The classification models [2] need a collection of records (Training set), which each record contains a set of attributes, one of the attributes is the class. A model will be built according to the classification technique chosen, and It will find a model for the class attribute as a function of the values of the other attributes. The model [3] built should assign a class value as accurate as possible to the unseen records. Finally, the model will be validated in Test set in order to determine the accuracy of the model.

There are different techniques of classification that could be used for predicting, classifying and categorizing. The most important are as follows [2]:

- Decision Tree-based methods
- Rule-based methods
- Memory-based reasoning
- Neural Networks
- Support Vector Machines

Random Forest is a supervised learning algorithm that can be used for both regression and classification tasks, and It belongs to the Decision Tree-based methods [2][7]. The Decision trees [4] is a tree in which each internal (non-leaf) node is labelled with an input feature. The arcs coming from a node labelled with an input feature are labelled with each of the possible values of the target or output feature, or the arc leads to a subordinate decision node on a different input feature. Each leaf of the tree is labelled with a class or a probability distribution over the classes.

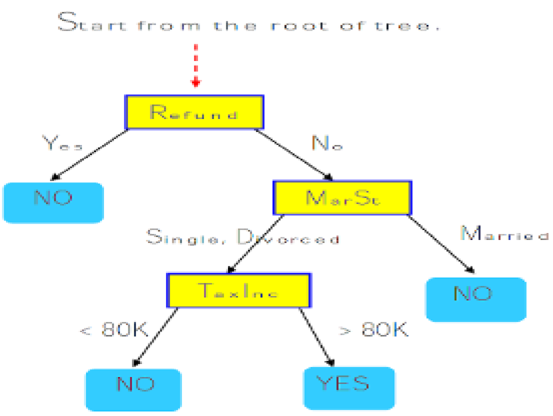


Figure 1. Example of decision tree. Source: MSc Data science notes [2].

Random Forest [5] operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

Note that most of the time Random Forest models have been trained with the “Bagging” [6] method. The general idea of the bagging method is that a combination of learning models

increases the overall result.

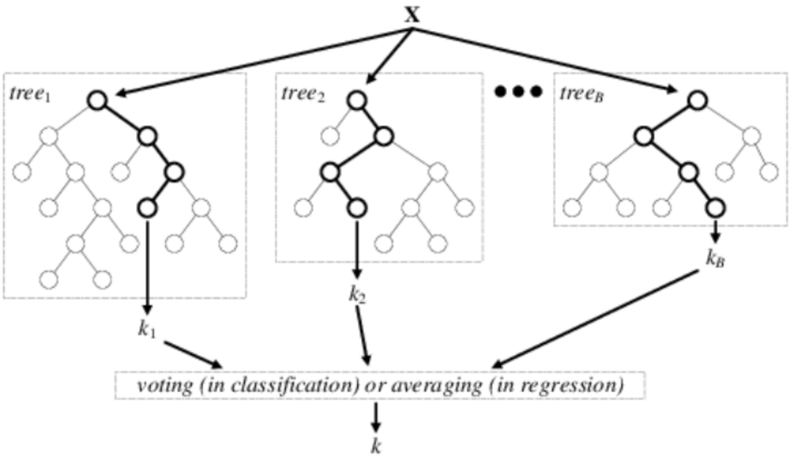


Figure 2. Example of Random Forest. Source: www.researchgate.net [7].

One of the advantages [7] of using the Random Forest model is that the algorithm is simple, and It uses default hyperparameters, which produces a good prediction, and is easy to understand. Besides, Random Forest prevents to incur in overfitting since It uses enough trees to add additional randomness to the model.

On the other hand, the main limitation [7] of Random Forest is that a large number of trees can make the algorithm and ineffective for real-time prediction.

DATA SEARCH STRATEGY

The data selected is a dataset of customer eligibility for a loan. The dataset used was found on https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/. The reasons why I liked the dataset was because of the topic and because it contained lots of missing values and attributes with outlier values.

EXPLANATION AND PREPARATION OF DATASETS

The dataset is made up of 614 rows and 13 columns or attributes. The 13 attributes are as follow:

- Loan ID
- Gender
- Married
- Dependents
- Education
- Self Employed
- Applicant Income
- Co-applicant Income
- Loan Amount
- Loan Amount Term
- Credit History
- Property Area
- Loan Status

All the variables are categorized as factors but Applicant Income, Loan Amount, Loan Amount Term and, Credit History which is integers, and Co-applicant Income which is considered numerical.

The dependent variable is Loan status, the rest of the attributes are independent.

```
'data.frame':   614 obs. of  13 variables:
 $ Loan_ID          : Factor w/ 614 levels "LP001002","LP001003",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Gender           : Factor w/ 3 levels "","Female","Male": 3 3 3 3 3 3 3 3 3 3 ...
 $ Married          : Factor w/ 3 levels "","No","Yes": 2 3 3 3 2 3 3 3 3 3 ...
 $ Dependents       : Factor w/ 5 levels "","0","1","2",..: 2 3 2 2 2 4 2 5 4 3 ...
 $ Education        : Factor w/ 2 levels "Graduate","Not Graduate": 1 1 1 2 1 1 2 1 1 1 ...
 $ Self_Employed    : Factor w/ 3 levels "","No","Yes": 2 2 3 2 2 3 2 2 2 2 ...
 $ ApplicantIncome  : int  5849 4583 3000 2583 6000 5417 2333 3036 4006 12841 ...
 $ CoapplicantIncome: num  0 1508 0 2358 0 ...
 $ LoanAmount       : int  NA 128 66 120 141 267 95 158 168 349 ...
 $ Loan_Amount_Term : int  360 360 360 360 360 360 360 360 360 360 ...
 $ Credit_History   : int  1 1 1 1 1 1 1 0 1 1 ...
 $ Property_Area    : Factor w/ 3 levels "Rural","Semiurban",..: 3 1 3 3 3 3 3 2 3 2 ...
 $ Loan_Status      : Factor w/ 2 levels "N","Y": 2 1 2 2 2 2 2 1 2 1 ...
```

Figure 3. Screen Shot of dataset structure. Source: RStudio customer eligibility for loan.

The dataset summary of the training dataset shows some missing values and possible outliers.

```
     Loan_ID       Gender     Married    Dependents        Education    Self_Employed ApplicantIncome CoapplicantIncome
 LP001002:  1        : 13        :  3      : 15     Graduate    :480     : 32        Min.   :  150   Min.   :    0
 LP001003:  1    Female:112   No :213     0 :345    Not Graduate:134   No :500       1st Qu.: 2878   1st Qu.:    0
 LP001005:  1    Male  :489   Yes:398     1 :102                       Yes: 82       Median : 3812   Median : 1188
 LP001006:  1                             2 :101                                     Mean   : 5403   Mean   : 1621
 LP001008:  1                             3+: 51                                     3rd Qu.: 5795   3rd Qu.: 2297
 LP001011:  1                                                                        Max.   :81000   Max.   :41667
 (Other) :608
   LoanAmount     Loan_Amount_Term Credit_History     Property_Area Loan_Status
 Min.   :  9.0   Min.   : 12       Min.   :0.0000   Rural    :179   N:192
 1st Qu.:100.0   1st Qu.:360       1st Qu.:1.0000   Semiurban:233   Y:422
 Median :128.0   Median :360       Median :1.0000   Urban    :202
 Mean   :146.4   Mean   :342       Mean   :0.8422
 3rd Qu.:168.0   3rd Qu.:360       3rd Qu.:1.0000
 Max.   :700.0   Max.   :480       Max.   :1.0000
 NA's   :22      NA's   :14        NA's   :50
```

Figure 4. Screen Shot of dataset summary. Source: RStudio customer eligibility for loan.

The data pre-processing performed has consisted on replacing [3][8] the missing values (NA) for central tendency measures such a mode and mean, and on subsequent stage outliers' detections [3][8] and treatment [3][8].

The missing values on the dataset were found on all independent all attributes, but Applicant Income, Co-applicant Income and Property Area, as you can appreciate on figure 4.

Note that the central tendency measures [8] were applied for Its simplicity and because It was not biasing the information since the number of the missing values on every single attribute was not high. Besides, it is a simple and powerful technique for cleaning data. However, It suffers from arbitrarity, and It may lead to data corruption. The central tendency measures applied to the missing values by attribute is described as follow:

Attribute	Central Tendency Method Applied
Gender	Mode
Dependents	Mode
Loan Amount	Mean
Loan Amount Term	Mean
Self-employed	Mode

Figure 7. Central tendency method applied by attribute. Source Self-Made

The missing values of the Credit History were removed since I considered Credit History a critical attribute which is better not having the information than inferring a value. It was done this way to prevent biasing results

A Boxplot Diagram detected the outliers [9] as It shows the picture below:

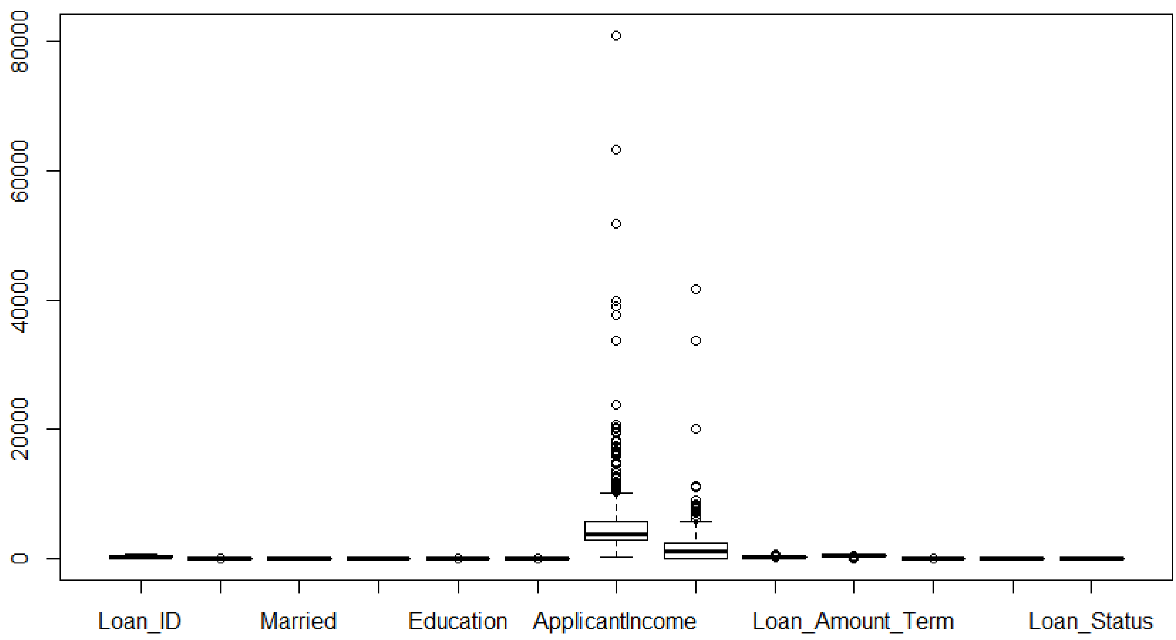


Figure 8. Boxplot diagram of all the attributes of the dataset pre-processing. Source RStudio customer eligibility for loan.

All the data that was above of the maximum was replaced by the value of the maximum value of the boxplot. It has been made in order to smooth the data and prevent the data to be skewed for the outliers.

Please find the summary and the boxplot diagram of the training dataset post-processing as per below:

```
   Gender     Married     Dependents        Education    Self_Employed ApplicantIncome CoapplicantIncome
 Female:101   No  :199   0 :331     Graduate    :443   No :489       Min.   :  150   Min.   :    0
 Male  :463   Yes :362   1 : 90     Not Graduate:121   Yes: 75       1st Qu.: 2893   1st Qu.:    0
              NA's:  3   2 : 95                                      Median : 3815   Median :1106
                         3+: 48                                      Mean   : 4649   Mean   :1378
                                                                     3rd Qu.: 5804   3rd Qu.:2250
                                                                     Max.   :10170   Max.   :5625
  LoanAmount    Loan_Amount_Term Credit_History     Property_Area Loan_Status Credit_History_f
 Min.   :  9.0   Min.   : 36.0   Min.   :0.0000   Rural    :165   N:179       0: 89
 1st Qu.:101.8   1st Qu.:360.0   1st Qu.:1.0000   Semiurban:217   Y:385       1:475
 Median :128.5   Median :360.0   Median :1.0000   Urban    :182
 Mean   :136.7   Mean   :342.1   Mean   :0.8422
 3rd Qu.:162.0   3rd Qu.:360.0   3rd Qu.:1.0000
 Max.   :252.4   Max.   :480.0   Max.   :1.0000
```

Figure 9. Screen Shot of the training dataset summary post-processing. Source: RStudio customer eligibility for loan.

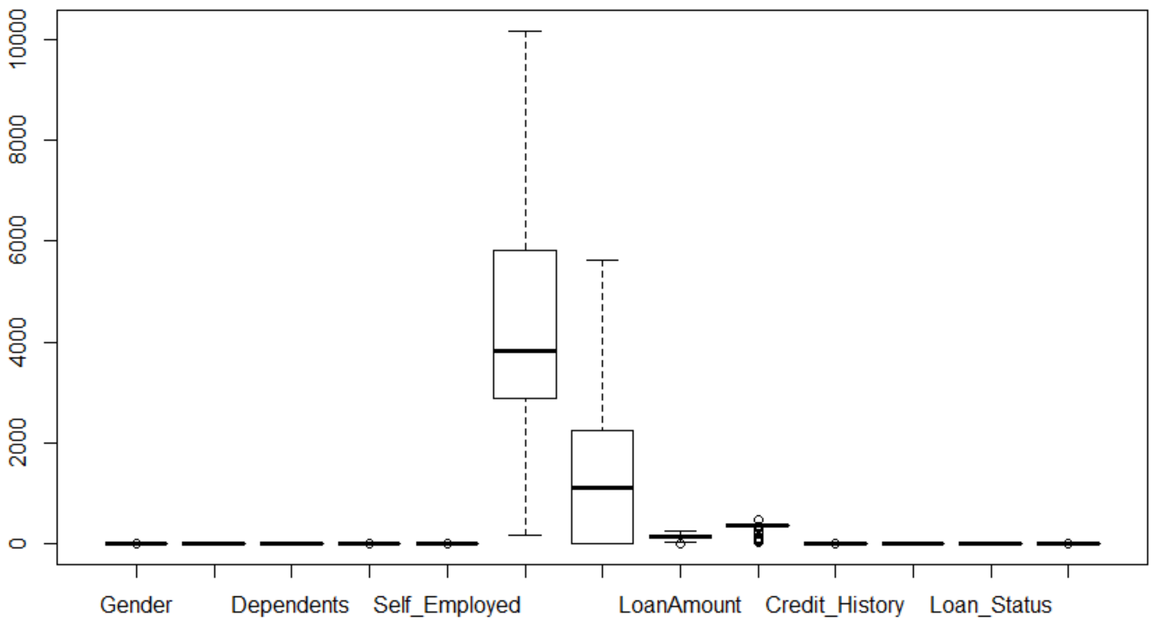


Figure 10. Boxplot diagram of all the attributes of the training dataset post-processing. Source RStudio customer eligibility for loan.

TASK: CLASSIFICATION

In this section will be performing a classification analysis using Random Forest one of the Decision Tree-based methods. The method aims to decorrelate the several trees which are generated by the different bootstrapped samples of the training dataset. It reduces the variance and of the trees by averaging them, improve the performance on the test dataset and avoid overfitting.

The analysis will use two software such as an R programming language and SAS miner. We will perform first the analysis with R programming language followed by SAS miner, and then we will compare the results.

- RANDOM FOREST IMPLEMENTATION IN R

R programming language is a useful tool when comes to analyse data. One of the most significant advantages is that is open source and make possible for many R programmers to upload their work and share with the rest of the community. On this assignment will use a black-box approach making use of other's packages to analyse the information. The black-box approach has been chosen for its simplicity but needs to be noted that this approach entails a great peril of not understanding what happens within the function and end up with wrong results. In this case, since it is only an academic work, we are more interested in the analysis of the results rather than the actual result. For this reason, we can allow certain privileges like delegate the arduous task of coding the functions to the R programmers community.

The model construction has needed of 2 packages installation:

1. Package 'randomForest' [10].
2. Package 'e1071'[11]

Package ‘randomForest’ allows to use Random forest algorithm to train the model to be able to validate with the test dataset, while Package ‘e1071’ allows to train support vector machine (SVM), predictions from the model, as well as decision values from the binary classifiers Using this method obtains predictions from the model, decision values from the binary classifiers, data visualization and perform a grid research over specified parameter ranges.

```
library(randomForest)
library(e1071)
```

Figure 10. R screen shoot of packages. Source: Self-made R

A dataset partition has been performed, which the training set was made of 80 % of the data, and the test set was made of 20 %. The % has been chosen arbitrary but taking into consideration that over half of the data needed to be on the training side for the model to allow the model to be as trained as possible but leaving enough data to test the trained model.

```
pd <-sample(2,nrow(loan_train),replace=TRUE,prob=c(0.8,0.2))
train <-loan_train[pd==1,]
validate <-loan_train[pd==2,]
```

Figure 11. R screen shoot of data partition. Source: Self-made R

The Random Forest was applied by using the in-built function RandomForest()[10]. Parameter formula request of the target attribute which is Loan_Status and the independent attributes which are Gender, Married, Dependents, Education, Employed, Applicant, Co-applicant Income, Loan Amount Term, Credit History, and Property Area expressed by ‘~.’ Parameter data stands for training set, Ntree for number of trees, Mtry for number of variables randomly sampled as candidates at each split, importance for predictor assessment and proximity for the calculation of proximity of the rows.

```
rf<-randomForest( formula = Loan_Status ~ .,data=train, ntree=500, mtry=5,importance= T,proximity=T)
```

Figure 12. R screen shoot of RandomForest function. Source: Self-made R

The results of the trained Random Forest model are an out of bag error of 16.59%, which means that 16.59 % of the classifications made by the model are wrong. 16% is a pretty good number since for many industry projects over 25% Out-of-bag error (OOB) would be considered not good enough.

The confusion matrix shows a different picture. The True are well predicted but with a 0.037 class error, however, the negatives are poorly assessed, and the class error is at 0.473.

```
Call:
 randomForest(formula = Loan_Status ~ ., data = train, ntree = 500,      mtry = 5, importance = T, proximity = T)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 5

        OOB estimate of  error rate: 16.59%
Confusion matrix:
   N   Y class.error
N 70  63  0.47368421
Y 12 307  0.03761755
```

Figure 13. R screen shoot of RandomForest results. Source: Self-made R

On the figure 14 are drawn three lines; Red line that represent YES class error, Green line that represents NO class error and Black line that represents OOB estimate error rate. It illustrates the rate of the three error regarding the number of trees used in the model.

We see from figure 14 that the errors achieve their highest value around the tree number 10 and from the to the tree number 100 the values decrease progressively. From tree number 100 onwards the value does not change significantly.

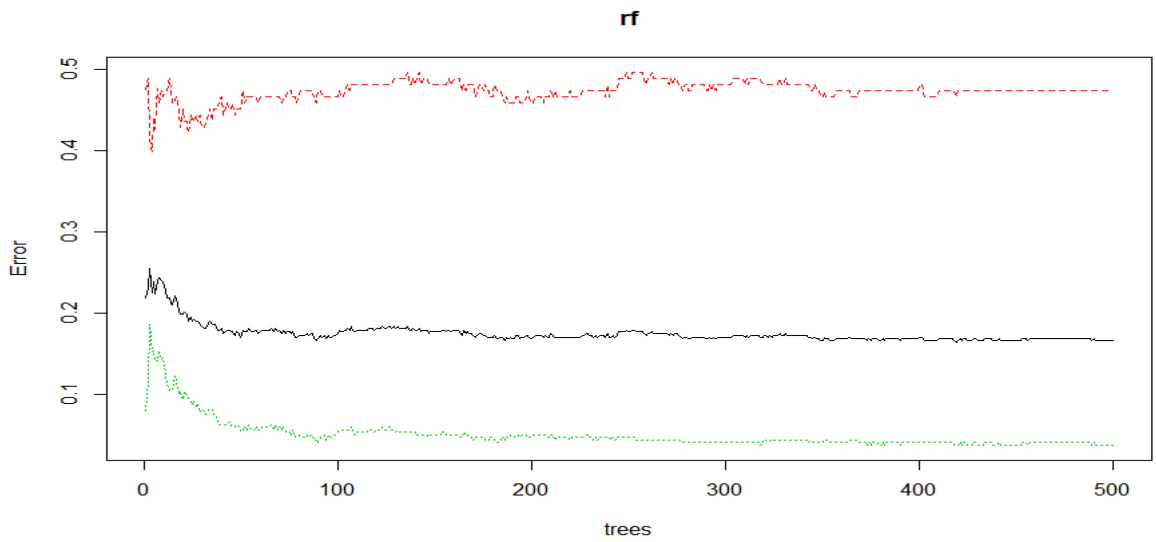


Figure 14. R screen shoot of model errors. Source: Self-made R

Once the model has been trained, it needs to be tested with data that It has not been seen yet.

```
p1<-predict(rf,validate)
confusionMatrix(p1,validate$Loan_Status)
```

Figure 15. R screen shoot of trained model predicting test set. Source: Self-made R

The model trained shows a result [13] of its performance on the test a bit poor with an Accuracy of 0.72 which means that the only 72% of the results were predicted correctly. The confidence interval of 95 % the model explains between a 63% and 80 % of the data on the test set. The reason why the model performs poorly is explained by its Sensitivity which is at 39% and means that the model only predicts correctly YES 39% of the times, while the Specificity is at 0.95 that means that No is predicted correctly 95% of the time.

```
Confusion Matrix and Statistics

          Reference
Prediction  N  Y
         N 18  3
         Y 28 63

               Accuracy : 0.7232
                 95% CI : (0.6307, 0.8036)
    No Information Rate : 0.5893
    P-Value [Acc > NIR] : 0.002246

                  Kappa : 0.3769
 Mcnemar's Test P-Value : 1.629e-05

            Sensitivity : 0.3913
            Specificity : 0.9545
         Pos Pred Value : 0.8571
         Neg Pred Value : 0.6923
             Prevalence : 0.4107
         Detection Rate : 0.1607
   Detection Prevalence : 0.1875
      Balanced Accuracy : 0.6729

       'Positive' Class : N
```

Figure 16. R screen shoot of results of test set with model trained. Source: Self-made R

Another important information that we can find on the model is the importance of the variable and the number of the nodes for the tree.

In figure 17, Mean Decrease accuracy shows that the most critical variable is credit History accounting over 80% of explanation of the model and followed by a significant difference by applicants' income with around 20 %. The Mean decreases Gini also show that credit history is the most significant variable with over 60 % of significance, while gender is the least important.

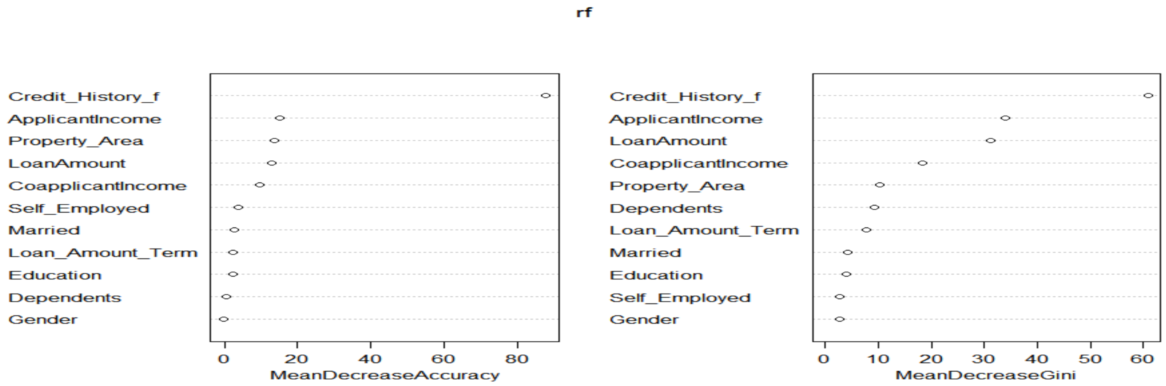


Figure 17. R screen shoot of results of Mean Decrease accuracy and Mean Decrease Gini. Source: Self-made R

In figure 18, the histogram of Number of nodes for the trees show that most common number of nodes in tree were between 70 and 75.

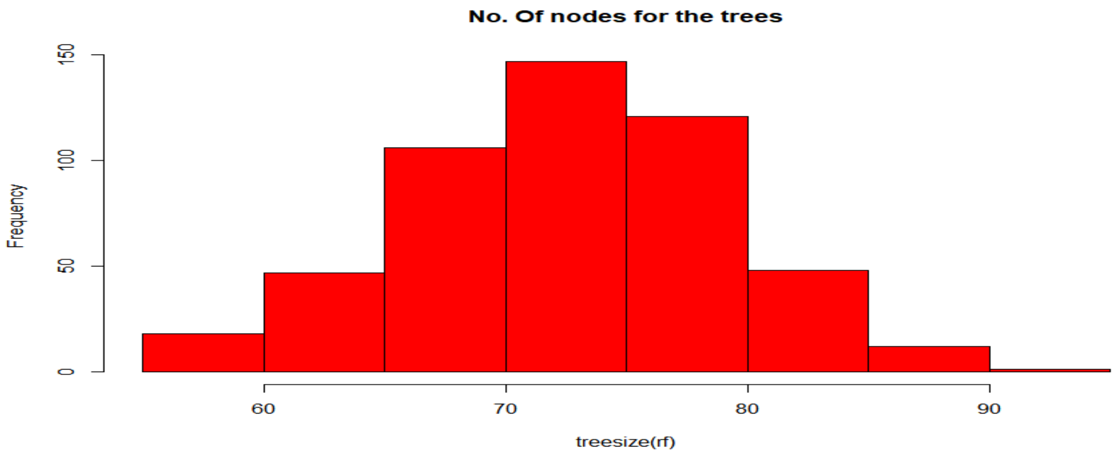


Figure 18. R screen shoot of the number of nodes for tree. Source: Self-made R

In order to see whether the Random forest model can be improved, we will use the function tuneRF()[14]. X request for the variable of the dataset, but the target value, while Y is the target value. Stepfactor increases or decreases the Mtry at each iteration. The plot is whether to plot the OOB error as a function of Mtry. Ntreetry is the number of the number of trees used at the tuning step. Trace is whether to print the progress of the search and Improve the (relative) improvement in OOB error must be by this much for the search to continue.

The values were assigned randomly initially, and they have tweaked until which I have considered the right values were found.

```
tuneRF(x=subset(train,select = -Loan_Status),y = train$Loan_Status,stepFactor = 0.5,
       plot= T,ntreeTry = 100,trace = T,improve = 0.05)
```

Figure 19. R screen shoot of the tune function. Source: Self-made R

Along with the TuneRF() function, we will use the figure 14 to tune the model and try to improve its performance. The figure 14 shows that OOB achieves a steady value around 150 trees, while figure 19, extracted from TuneRf(), shows that OOB error is optimal at 3 Mtry.

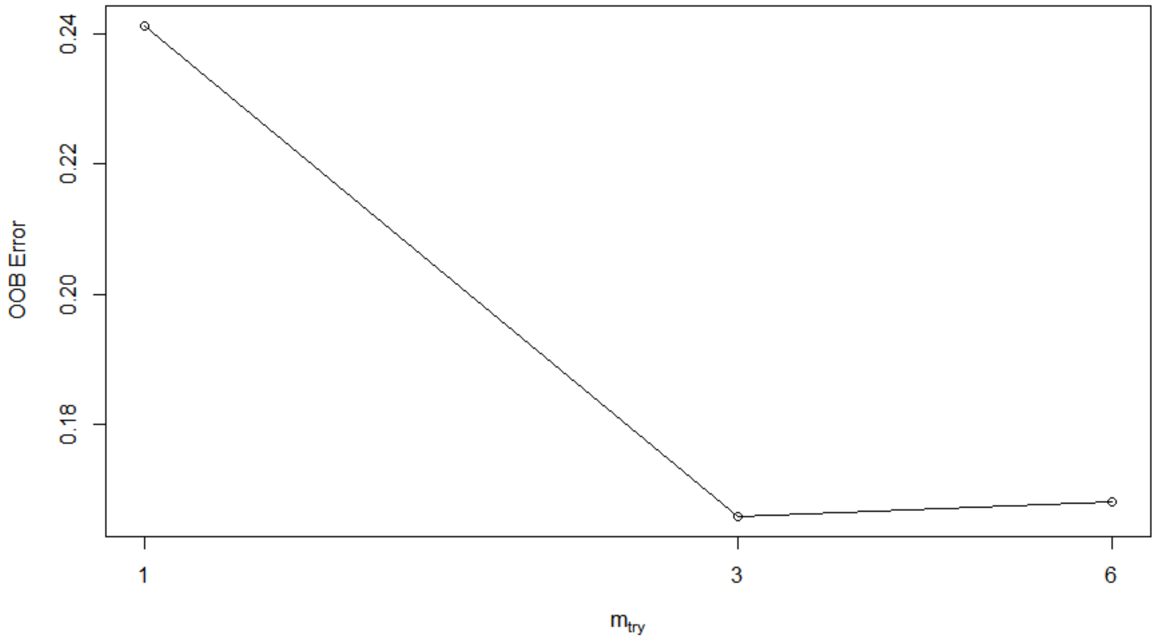


Figure 20. R screen shoot of the tune OOB error in regards of Mtry . Source: Self-made R

The Random Forest model is re-run with the new parameter Ntree= 100 and Mtry=3

```
rf<-randomForest( formula = Loan_Status ~ .,data=train, ntree=150, mtry=3,importance= T,proximity=T)
```

Figure 21. R screen shoot of RandomForest function tuned. Source: Self-made R

The results of the trained Random Forest model are an out of bag error of 17.48 %, which is higher than the original model 16.59%, Although, it still a good result it has got worse with the tune. The same has happened with the classification error that has worse it performance with NO at 49% and YES 4%, respectively.

```
Call:
 randomForest(formula = Loan_Status ~ ., data = train, ntree = 150,      mtry = 3, importance = T, proximity = T)
               Type of random forest: classification
                     Number of trees: 150
No. of variables tried at each split: 3

        OOB estimate of  error rate: 17.48%
Confusion matrix:
   N   Y class.error
N 67  66  0.49624060
Y 13 306  0.04075235
```

Figure 22. R screen shoot of RandomForest tuned results. Source: Self-made R

However, the tuned model has performed better with the test data set than the original Radom Forest. The accuracy is slightly better, and the 95 % CI has increased a bit too. The Sensitivity has worsened a bit, and the Specificity has been perfect this time.

```
Confusion Matrix and Statistics

          Reference
Prediction  N  Y
         N 16  0
         Y 30 66

               Accuracy : 0.7321
                 95% CI : (0.6402, 0.8114)
    No Information Rate : 0.5893
    P-Value [Acc > NIR] : 0.001166

                  Kappa : 0.386
 Mcnemar's Test P-Value : 1.192e-07

            Sensitivity : 0.3478
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.6875
             Prevalence : 0.4107
         Detection Rate : 0.1429
   Detection Prevalence : 0.1429
      Balanced Accuracy : 0.6739

       'Positive' Class : N
```

Figure 23. R screen shoot of results of test set with model trained and tuned. Source: Self-made R

On the overall, the tune has been useful to improve slightly the model. Even though It still suffering to predict correctly NO and It significantly impacts Its accuracy.

- RANDOM FOREST IMPLEMENTATION IN SAS

SAS Miner is software made with a friendly interface to be able to do data science with no coding experience. The benefits are clear; you can analyse data quickly and intuitively. Contrary, as all the models are made for you to use it incurs on the black-box approach, and It entails the same problems.

The model was built partitioning the data on three. Training data set was 80%, validation was 10% and test set was 10%. This partitioning percentage has been done following the same It has been done in the previous Random Forest implementation for R.

Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	0.8
Validation	0.1
Test	0.1

Figure 24. SAS screen shoot of data partitioning. Source: Self-made in SAS

Once the data was partitioned, we have created the Random Forest model by linking the nodes.

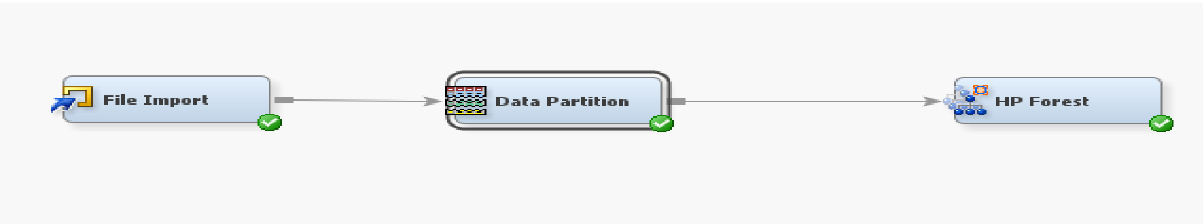


Figure 25. SAS screen shoot of the Random Forest nodes. Source: Self-made in SAS

Find the Random Forest model information below, where the most significant parameter change is variable to try that by default is 3.

Data Access Information

Data	Engine	Role	Path
WORK.HPDMFOREST_TRAINDATA	V9	Input	On Clien

Model Information

Parameter	Value	
Variables to Try	3	(Default)
Maximum Trees	500	
Inbag Fraction	0.5	
Prune Fraction	0	(Default)
Prune Threshold	0.1	(Default)
Leaf Fraction	0.00001	(Default)
Leaf Size Setting	1	(Default)
Leaf Size Used	1	
Category Bins	30	
Interval Bins	100	
Minimum Category Size	5	
Node Size	100000	(Default)
Maximum Depth	50	
Alpha	0.05	
Exhaustive	5000	
Rows of Sequence to Skip	5	(Default)
Split Criterion	.	Gini
Preselection Method	.	Loh
Missing Value Handling	.	Valid value

Figure 26. SAS screen shoot of the Random Forest model information. Source: Self-made in SAS

The model has performed relatively good, the Average square error on the train set 15%, on the validation 13% and on the Test set 14%, respectively.

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Loan_Status		_ASE_	Average Sq...	0.155213	0.132036	0.141972
Loan_Status		Target Label	Divisor for A...	898	114	116
Loan_Status		_MAX_	Maximum A...	0.788963	0.779582	0.775754
Loan_Status		_NOBS_	Sum of Fre...	449	57	58
Loan_Status		_RASE_	Root Avera...	0.393971	0.363367	0.376791
Loan_Status		_SSE_	Sum of Squ...	139.3813	15.05206	16.46872
Loan_Status		_DISF_	Frequency ...	449	57	58
Loan_Status		_MISC_	Misclassific...	0.193764	0.140351	0.155172
Loan_Status		_WRONG_	Number of ...	87	8	9

Figure 27. SAS screen shoot of the Random Forest error results. Source: Self-made in SAS

The Error tends to achieve its lowest around the tree 150 as figure 28 shows.

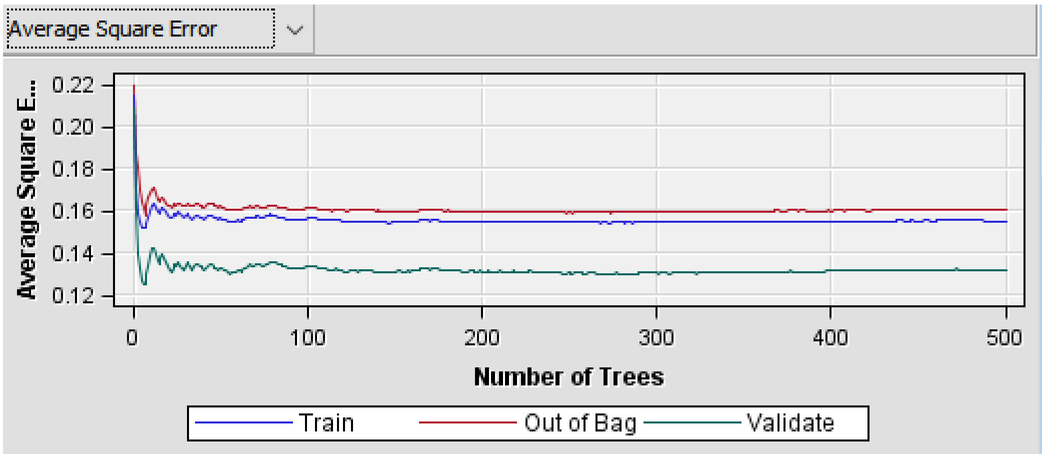


Figure 28. SAS screen shoot of the Random Forest error results regarding number of trees. Source: Self-made in SAS

The most significant variable is Credit-History, followed by Loan amount, while the least important is Dependents.

Variable Name	Number of Splitting Rules	Train: Gini Reduction	Train: Margin Reduction	OOB: Gini Reduction	OOB: Margin Reduction	Valid: Gini Reduction	Valid: Margin Reduction	Label
Credit_Hist...	410	0.081448	0.162895	0.08115	0.16238	0.12734	0.21272	
Loan_Amo...	144	0.003301	0.006601	-0.00188	0.00151	-0.00020	0.00308	
Property_Ar...	124	0.003931	0.007862	-0.00084	0.00313	0.00441	0.01209	
Education	52	0.001200	0.002399	-0.00034	0.00088	-0.00127	0.00035	
Married	50	0.001030	0.002059	-0.00093	0.00011	0.00121	0.00223	
LoanAmount	45	0.000944	0.001889	-0.00124	-0.00045	0.00017	0.00162	
Gender	42	0.001026	0.002052	-0.00094	0.00016	-0.00288	-0.00160	
Coapplican...	29	0.000810	0.001621	-0.00072	0.00004	-0.00061	0.00014	
ApplicantIn...	16	0.000393	0.000785	-0.00056	-0.00023	-0.00019	0.00001	
Self_Emplo...	13	0.000257	0.000515	-0.00047	-0.00020	-0.00047	-0.00023	
Dependents	5	0.000094	0.000187	-0.00021	-0.00016	-0.00007	0.00005	
VAR1	0	0.000000	0.000000	0.00000	0.00000	0.00000	0.00000	

Figure 29. SAS screen shoot of the Random Forest of variable significance. Source: Self-made in SAS

As seen, on the results the model could be improved by selecting 150 trees instead of 500. If we re-run the model, it shows slightly better performance on the percentage of error on the train set and test set, but not significant.

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Loan_Status		_ASE_	Average Squa...	0.155647	0.133463	0.141542
Loan_Status		_DIV_	Divisor for ASE	898	114	116
Loan_Status		_MAX_	Maximum Abs...	0.788973	0.777792	0.777766
Loan_Status		_NOBS_	Sum of Frequ...	449	57	58
Loan_Status		_RASE_	Root Average ...	0.394522	0.365326	0.376221
Loan_Status		_SSE_	Sum of Squar...	139.7713	15.21482	16.41889
Loan_Status		_DISF_	Frequency of ...	449	57	58
Loan_Status		_MISC_	Misclassificati...	0.193764	0.140351	0.155172
Loan_Status		_WRONG_	Number of Wr...	87	8	9

Figure 30. SAS screen shoot of the Random Forest error results post-tune. Source: Self-made in SAS

CONCLUSION

The Random Forest is one of the decision Tree-based methods that help us with classification task. The benefits of using Random Forest is that reduces the variance of the model due to the decorrelation of the trees and averaging the results, as well as prevention of model overfitting.

The data selected was data of loan approvals, and the purpose of the task was classifying whether the loan would be approved using the Random Forest model. The dataset was cleaned of missing values and removed outliers that could distortion the results.

The implementation with R and SAS Miner has achieved its purpose of building a model to classify customer for loan approval. However, the results have been different because of the size of the dataset.

On the train set, both R and SAS have performed more on less the same. Contrary, on the test set in which the Random Forest in R has not been satisfactory, while the Random Forest in SAS Miner has achieved a satisfactory error.

The tuning part has affected more the model made by R than the one made by SAS miner, and It is a clear example that the R is more customizable than SAS Miner.

On the overall, I consider the result obtained has been satisfactory since the purpose of the assignment was to make a Random Forest model and analyse the results with R and SAS Miner. However, the results could be improved in order to achieve a sound model able to be implemented. In that sense, a bigger dataset will be needed to train more the model.

REFERENCES

1- https://en.wikipedia.org/wiki/Statistical_classification
2- MSc Data science notes, Salford University. Classification: Decision trees
3- Han, Kamber, and Pei. Data Mining: Concepts and Techniques,3rd Edition, 2012.
4- https://en.wikipedia.org/wiki/Decision_tree_learning
5- https://en.wikipedia.org/wiki/Random_forest
6- https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd
7- https://www.researchgate.net/figure/Architecture-of-the-random-forest-model_fig1_301638643
8- MSc Data science notes, Salford University. Data preparation
9- MSc Data science notes, Salford University. ASDM Workshop: Week1
10- https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf
11- https://cran.r-project.org/web/packages/e1071/e1071.pdf
12- https://data-flair.training/blogs/e1071-in-r/
13- https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/
14- http://math.furman.edu/~dcs/courses/math47/R/library/randomForest/html/tuneRF.html
15- https://cran.r-project.org/web/packages/reshape2/reshape2.pdf
16- https://cran.r-project.org/web/packages/randomForest/randomForest.pdf
17- https://cran.r-project.org/web/packages/caret/caret.pdf
18- https://cran.r-project.org/web/packages/e1071/e1071.pdf

APPLY ASSOCIATION RULES MINING ON DATASET USING R & SAS

INTRODUCTION

The business environment is in a constant change due to the changing necessity of the customers. To succeed on the business world is not enough with exploiting a business model that has proven successful in the past, also, business needs to continually update its offer to strive and be competitive among the market competitor.

How business used to assess the new products and service have changed dramatically. In the past, the businessmen or the CEO of the company had to take decision-based on intuition, while nowadays it is taken based on the information.

Among all the techniques that modern business use to assess their customer are the association rules. This technique helps businesses to detect trends and patterns on customer purchases and give useful information on how the market is evolving.

AIM AND OBJECTIVE OF THE TASK

This task aims to present the Association Rules and apply it to a supermarket dataset.

This work will explain how to implement Association Rules on a dataset for Association purposes, and how to detect patterns on customer transactions.

BRIEF LITERATURE REVIEW

Association rules [1] are the result of searching data for patterns using metrics such as support, confidence and lift to detect the most important relationships.

$$Rule:\ X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Figure 1. Association rules formulas: Source www.saedsayad.com[2].

The association rules can show Novel and actionable associations. The interestingness of an association is measured by Support, Confidence and lift. A significant confidence and support threshold may show 'Folklores' or known facts, while a small support and confidence threshold may show too many association rules that are not interesting.

The most common techniques used to search for patter within the dataset is the Apriori technique [1]. The Apriori [3] technique for frequent itemset mining and association rules learning. It aims to identify individual items in the dataset and extending them to larger sets.

$$
\begin{array}{l}
\mathrm{Apriori}(T, \epsilon) \\
\quad L_1 \leftarrow \{\mathrm{large}\ 1 - \mathrm{itemsets}\} \\
\quad k \leftarrow 2 \\
\quad \mathbf{while}\ L_{k-1} \neq \emptyset \\
\quad\quad C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \land b \notin a\} - \{c \mid \{s \mid s \subseteq c \land |s| = k-1\} \nsubseteq L_{k-1}\} \\
\quad\quad \mathbf{for\ transactions}\ t \in T \\
\quad\quad\quad D_t \leftarrow \{c \mid c \in C_k \land c \subseteq t\} \\
\quad\quad\quad \mathbf{for\ candidates}\ c \in D_t \\
\quad\quad\quad\quad count[c] \leftarrow count[c] + 1 \\
\quad\quad L_k \leftarrow \{c \mid c \in C_k \land\ count[c] \geq \epsilon\} \\
\quad\quad k \leftarrow k + 1 \\
\quad \mathbf{return}\ \bigcup_k L_k
\end{array}
$$

Figure 2. Apriori pseudo code. Source: en.wikipedia.org/wiki/Apriori_algorithm[3]

Even though the Apriori is the most used method, we need to consider its advantages and disadvantages.

Its advantages are that it uses large items property, it is easy parallelized, and it is easy to implement. Contrary, It assumes transaction database is memory resident and requires up to m database scans.

DATA SEARCH STRATEGY

The data selected is a supermarket basket. The dataset used was found on https://www.kaggle.com/.The reasons why I liked the dataset was because of the topic and because it needs to use wrangling techniques to make it work with SAS and R.

EXPLANATION AND PREPARATION OF DATASETS

The data selected a supermarket basket transaction dataset, which is made of 1499 rows and 35 columns. The rows are the transaction, and the columns correspond to the items purchased.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	1/1/2000yogurt	pork	sandwich	lunch me	all- purpc	flour	soda	butter	vegetable	beef	aluminun	all- purpc	dinner ro	shampoo	all- purpc	mixes	soap	laundry d	ice cream	dinner ro					
2	1/1/2000toilet paper	shampoo	hand soa	waffles	vegetable	cheeses	mixes	milk	sandwich	laundry d	dishwash	waffles	individua	hand soa	vegetable	individua	yogurt	cereals	shampoo	vegetable	aluminun	tortillas	mixes		
3	2/1/2000soda	pork	soap	ice cream	toilet pap	dinner ro	hand soa	spaghetti	milk	ketchup	sandwich	poultry	toilet pap	ice cream	ketchup	vegetable	laundry d	spaghetti	bagels	soap	ice cream	shampoo	lunch me	cereals	spaghetti
4	2/1/2000cereals	juice	lunch me	soda	toilet pap	all- purpc																			
5	2/1/2000sandwich loaves	pasta	tortillas	mixes	hand soa	toilet pap	vegetable	vegetable	paper tov	vegetable	flour	vegetable	pork	poultry	eggs	vegetable	pork	spaghetti	vegetable	milk	waffles	individua	vegetable	dinner ro	all- purpc
6	2/1/2000laundry detergent	toilet pap	eggs	toilet pap	vegetable	bagels	dishwash	cereals	paper tov	laundry d	butter	cereals	bagels	paper tov	shampoo	toilet pap	soap	soap	pasta	coffee/te	poultry	bagels	aluminun	butter	spaghetti
7	3/1/2000individual meals	paper tov	tortillas	vegetable	milk	ice cream	juice	dishwash	soap	sandwich	pasta	ketchup	all- purpc	yogurt	mixes	mixes	toilet pap	vegetable	beef	sandwich	eggs	spaghetti	fruits	toilet pap	
8	4/1/2000ice cream	juice	paper tov	waffles	soda	cheeses	poultry	toilet pap	vegetable																
9	4/1/2000juice	poultry	coffee/te	coffee/te	dishwash																				
10	5/1/2000ketchup	coffee/te	toilet pap	pork	flour	milk	soda	dishwash	eggs	tortillas	tortillas	vegetable	fruits	sugar	soap	dishwash	juice	sandwich	poultry	eggs	flour	bagels	soda	bagels	soda
11	5/1/2000sandwich loaves	ice cream	soda	bagels	dishwash	eggs	sugar	waffles	individua	toilet pap	pasta	sandwich	cheeses	vegetable	hand soa	vegetable	vegetable	ketchup	poultry	dinner ro	dishwash	coffee/te	milk	shampoo	ketchup
12	6/1/2000pork	tortillas	pork	shampoo	lunch me	pasta	juice	bagels	vegetable	bagels	vegetable	laundry d	yogurt	sugar	waffles	hand soa	hand soa	butter	juice	spaghetti	vegetable	flour			
13	7/1/2000sugar	fruits	all- purpc	aluminun	laundry d	individua	flour	pork	shampoo	sugar	aluminun	shampoo	soap	bagels	toilet pap	juice	ice cream	ice cream	sandwich	soap	sugar	paper tov	vegetable		
14	7/1/2000fruits	dinner ro	individua	shampoo	ketchup	cereals	sandwich	laundry d	vegetable	sandwich	cheeses	vegetable													
15	7/1/2000individual meals	ice cream	cereals	paper tov	bagels	mixes	lunch me	juice	toilet pap	cheeses	paper tov	paper tov	juice	aluminun	fruits	butter	sandwich	tortillas	spaghetti	hand soa					
16	8/1/2000sugar	sandwich	flour	juice	milk	paper tov	cereals	sandwich	pasta	soda	dishwash	ice cream	soap	cheeses	beef	flour	dinner ro								
17	8/1/2000milk	hand soa	pasta	individua	spaghetti	cereals	sandwich	hand soa	individua																
18	8/1/2000sandwich bags	toilet pap	bagels	shampoo	coffee/te																				
19	9/1/2000individual meals	laundry d	coffee/te	eggs	aluminun	beef	juice	flour	sugar	individua	waffles	bagels	vegetable	coffee/te	spaghetti	butter	beef								
20	10/1/2000shampoo	dishwash	yogurt	juice	sugar	soap	sandwich	butter	sandwich	coffee/te	ketchup	aluminun	dishwash	mixes	laundry d	shampoo	cheeses	tortillas	toilet pap	hand soa	paper tov	beef	individua	toilet pap	soda
21	11/1/2000waffles	fruits	all- purpc	pork	juice	bagels	mixes																		
22	11/1/2000cheeses	vegetable	cereals	sugar	bagels	soda																			
23	11/1/2000vegetables	vegetable	aluminun	bagels	vegetable	shampoo	shampoo	vegetable	dishwash	cereals	cheeses	flour	lunch me	pasta	soda	vegetable	toilet pap	poultry	coffee/te	flour	all- purpc	waffles	coffee/te	milk	yogurt
24	11/1/2000fruits	all- purpc	pasta	cheeses	juice	sandwich	vegetable	sandwich	coffee/te	juice	lunch me	spaghetti	paper tov	hand soa	soap	ice cream	individua	ketchup	butter	poultry	cheeses	bagels	all- purpc	ketchup	waffles
25	11/1/2000bagels	sugar	pork	sandwich	tortillas	ice cream	all- purpc	yogurt	fruits	pasta	lunch me	dishwash	juice	vegetable	vegetable										
26	12/1/2000fruits	sandwich	vegetable	coffee/te	aluminun	vegetable	shampoo	cereals	dinner ro	coffee/te	poultry	butter	juice	paper tov	vegetable	beef	mixes	lunch me	bagels	tortillas	aluminun	individua	beef	juice	soda
27	13/1/2000laundry detergent	pork	pasta	cheeses	fruits	sugar	lunch me	laundry d	paper tov	butter	sandwich	ice cream	waffles	all- purpc	sandwich	lunch me	sandwich	poultry	vegetable	shampoo	dinner ro	milk	soda	mixes	
28	13/1/2000pork	bagels	poultry	pasta	butter	all- purpc	pasta	shampoo	sugar	ketchup	eggs	soda	tortillas	soap	individua	ketchup	butter	mixes	juice	lunch me	cereals	waffles	shampoo	shampoo	toilet pap
29	13/1/2000pasta	butter	sandwich	spaghetti	juice	dinner ro	all- purpc	pork	yogurt	vegetable	spaghetti	vegetable													
30	14/1/2000flour	bagels	cheeses	sandwich	toilet pap																				
31	15/1/2000aluminum foil	eggs	ice cream	pasta	juice	waffles	shampoo	dinner ro	soda	vegetable	laundry d	hand soa	waffles	cereals	vegetable	toilet pap	individua	yogurt	spaghetti	ice cream	soda	lunch me	paper tov		
32	15/1/2000soap	paper tov	individua	dinner ro	vegetable	lunch me	sugar	soap	eggs	beef	yogurt	soda	sandwich	aluminun	dinner ro	soap	cheeses	sandwich	ice cream	ketchup					
33	15/1/2000sandwich loaves	pork	sandwich	ketchup	coffee/te	vegetable	soda	poultry	pasta	all- purpc	coffee/te	paper tov	individua	vegetable	poultry	flour	aluminun	soap	yogurt	mixes	soda				
34	16/1/2000lunch meat	mixes	vegetable	soap	hand soa	tortillas	coffee/te	cheeses	tortillas	cheeses	ice cream	tortillas	beef	yogurt	spaghetti	soda	eggs								
35	16/1/2000aluminum foil	cheeses	cereals	mixes	laundry d	juice	pork	bagels	sandwich	yogurt	milk	sandwich	spaghetti	cereals	individua	eggs									
36	17/1/2000soap	vegetable	lunch me	flour	juice	yogurt	shampoo	spaghetti	ketchup	laundry d	dinner ro	ketchup	vegetable	shampoo	sugar	flour	sandwich	paper tov	pork	vegetable	butter	dinner ro	pork	all- purpc	hand soa
37	17/1/2000fruits	lunch me	dinner ro	shampoo	vegetable	hand soa	mixes	vegetable	pasta	sugar	vegetable														
38	17/1/2000eggs	hand soa	dinner ro	hand soa	lunch me	laundry d	pasta	aluminun	toilet pap	cereals	coffee/te	toilet pap	vegetable	juice	lunch me	soda	pork	fruits	spaghetti	laundry d	butter	tortillas	individua	aluminun	toilet pap

Figure 3. supermarket basket transaction pre-processing dataset Source: Self-made Excel.

In order to process the data in R, the first column had to fix the first column to separate the date from the item.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	01/01/2000	yogurt	sandwich	lunch me	all- purpc	flour	soda	butter	vegetabl	beef	aluminun	all- purpc	dinner ro	shampoo	all- purpc	mixes	soap	laundry d	ice cream	dinner ro								
2	01/01/2000	toilet pap	hand soa	waffles	vegetabl	cheeses	mixes	milk	sandwich	laundry d	dishwash	waffles	individua	hand soa	vegetabl	individua	yogurt	cereals	shampoo	vegetabl	aluminun	tortillas	mixes					
3	02/01/2000	soda	soap	ice cream	toilet pap	dinner ro	hand soa	spaghetti	milk	ketchup	sandwich	poultry	toilet pap	ice cream	ketchup	vegetabl	laundry d	spaghetti	bagels	soap	ice cream	shampoo	lunch me	cereals	spaghetti	pork	vegetabl	ch
4	02/01/2000	cereals	lunch me	soda	toilet pap	all- purpc																						
5	02/01/2000	sandwich	tortillas	mixes	hand soa	toilet pap	vegetabl	vegetabl	paper tov	vegetabl	flour	vegetabl	pork	poultry	eggs	vegetabl	pork	spaghetti	vegetabl	milk	waffles	individua	vegetabl	dinner ro	all- purpc	soda	yogurt	
6	02/01/2000	laundry d	eggs	toilet pap	vegetabl	bagels	dishwash	cereals	paper tov	laundry d	butter	cereals	bagels	paper tov	shampoo	toilet pap	soap	soap	pasta	coffee/te	poultry	bagels	aluminun	butter	spaghetti	ketchup	all- purpc	mi
7	03/01/2000	individual	tortillas	vegetabl	milk	ice cream	juice	dishwash	soap	sandwich	pasta	ketchup	all- purpc	yogurt	mixes	mixes	toilet pap	vegetabl	beef	sandwich	eggs	spaghetti	fruits	toilet pap				
8	04/01/2000	ice cream	paper tov	waffles	soda	cheeses	poultry	toilet pap	vegetabl																			
9	04/01/2000	juice	coffee/te	coffee/te	dishwash																							
10	05/01/2000	ketchup	toilet pap	pork	flour	milk	soda	dishwash	eggs	tortillas	tortillas	vegetabl	fruits	sugar	soap	dishwash	juice	sandwich	poultry	eggs	flour	bagels	soda	bagels	soda	beef	coffee/te	di
11	05/01/2000	sandwich	soda	bagels	dishwash	eggs	sugar	waffles	individua	toilet pap	pasta	sandwich	cheeses	vegetabl	hand soa	vegetabl	vegetabl	ketchup	poultry	dinner ro	dishwash	coffee/te	milk	shampoo	ketchup	sandwich	hand soa	ba
12	06/01/2000	pork	pork	shampoo	lunch me	pasta	juice	bagels	vegetabl	bagels	vegetabl	laundry d	yogurt	sugar	waffles	hand soa	hand soa	butter	juice	spaghetti	vegetabl	flour						
13	07/01/2000	sugar	all- purpc	aluminun	laundry d	individua	flour	pork	shampoo	sugar	aluminun	shampoo	soap	bagels	toilet pap	juice	ice cream	ice cream	sandwich	soap	sugar	paper tov	vegetabl					
14	07/01/2000	fruits	individua	shampoo	ketchup	cereals	sandwich	laundry d	vegetabl	sandwich	cheeses	vegetabl																
15	07/01/2000	individual	cereals	paper tov	bagels	mixes	lunch me	juice	toilet pap	cheeses	paper tov	paper tov	juice	aluminun	fruits	butter	sandwich	tortillas	spaghetti	hand soa								
16	08/01/2000	sugar	flour	juice	milk	paper tov	cereals	sandwich	pasta	soda	dishwash	ice cream	soap	cheeses	beef	flour	dinner ro											
17	08/01/2000	milk	pasta	individua	spaghetti	cereals	sandwich	hand soa	individua																			
18	08/01/2000	sandwich	bagels	shampoo	coffee/te																							
19	09/01/2000	individual	coffee/te	eggs	aluminun	beef	juice	flour	sugar	individua	waffles	bagels	vegetabl	coffee/te	spaghetti	butter	beef											
20	10/01/2000	shampoo	yogurt	juice	sugar	soap	sandwich	butter	sandwich	coffee/te	ketchup	aluminun	dishwash	mixes	laundry d	shampoo	cheeses	tortillas	toilet pap	hand soa	paper tov	beef	individua	toilet pap	soda	cheeses	vegetabl	al
21	11/01/2000	waffles	all- purpc	pork	juice	bagels	mixes																					
22	11/01/2000	cheeses	cereals	sugar	bagels	soda																						
23	11/01/2000	vegetable	aluminun	bagels	vegetabl	shampoo	shampoo	vegetabl	dishwash	cereals	cheeses	flour	lunch me	pasta	soda	vegetabl	toilet pap	poultry	coffee/te	flour	all- purpc	waffles	coffee/te	milk	yogurt	vegetabl	mixes	pa
24	11/01/2000	fruits	pasta	cheeses	juice	sandwich	vegetabl	sandwich	coffee/te	juice	lunch me	spaghetti	paper tov	hand soa	soap	ice cream	individua	ketchup	butter	poultry	cheeses	bagels	all- purpc	ketchup	waffles	soap		
25	11/01/2000	bagels	pork	sandwich	tortillas	ice cream	all- purpc	yogurt	fruits	pasta	lunch me	dishwash	juice	vegetabl	vegetabl													
26	12/01/2000	fruits	vegetabl	coffee/te	aluminun	vegetabl	shampoo	cereals	dinner ro	coffee/te	poultry	butter	juice	paper tov	vegetabl	beef	mixes	lunch me	bagels	tortillas	aluminun	individua	beef	juice	soda	hand soa	dinner ro	flc
27	13/01/2000	laundry d	pasta	cheeses	fruits	sugar	lunch me	laundry d	paper tov	butter	sandwich	ice cream	waffles	all- purpc	sandwich	lunch me	sandwich	poultry	vegetabl	shampoo	dinner ro	milk	soda	mixes				
28	13/01/2000	pork	poultry	pasta	butter	all- purpc	pasta	shampoo	sugar	ketchup	eggs	soda	tortillas	soap	individua	ketchup	butter	mixes	juice	lunch me	cereals	waffles	shampoo	shampoo	toilet pap	poultry	lunch me	
29	13/01/2000	pasta	sandwich	spaghetti	juice	dinner ro	all- purpc	pork	yogurt	vegetabl	spaghetti	vegetabl																
30	14/01/2000	flour	cheeses	sandwich	toilet pap																							
31	15/01/2000	aluminum	ice cream	pasta	juice	waffles	shampoo	dinner ro	soda	vegetabl	laundry d	hand soa	waffles	cereals	vegetabl	toilet pap	individua	yogurt	spaghetti	ice cream	soda	lunch me	paper tov					
32	15/01/2000	soap	individua	dinner ro	vegetabl	lunch me	sugar	soap	eggs	beef	yogurt	soda	sandwich	aluminun	dinner ro	soap	cheeses	sandwich	ice cream	ketchup								
33	15/01/2000	sandwich	sandwich	ketchup	coffee/te	vegetabl	soda	poultry	pasta	all- purpc	coffee/te	paper tov	individua	vegetabl	poultry	flour	aluminun	soap	yogurt	mixes	soda							
34	16/01/2000	lunch mea	vegetabl	soap	hand soa	tortillas	coffee/te	cheeses	tortillas	cheeses	ice cream	tortillas	beef	yogurt	spaghetti	soda	eggs											
35	16/01/2000	aluminum	cereals	mixes	laundry d	juice	pork	bagels	sandwich	yogurt	milk	sandwich	spaghetti	cereals	individua	eggs												
36	17/01/2000	soap	lunch me	flour	juice	yogurt	shampoo	spaghetti	ketchup	laundry d	dinner ro	ketchup	vegetabl	shampoo	sugar	flour	sandwich	paper tov	pork	vegetabl	butter	dinner ro	pork	all- purpc	hand soa	sandwich	all- purpc	
37	17/01/2000	fruits	dinner ro	shampoo	vegetabl	hand soa	mixes	vegetabl	pasta	sugar	vegetabl																	
38	17/01/2000	eggs	dinner ro	hand soa	lunch me	laundry d	pasta	aluminun	toilet pap	cereals	coffee/te	toilet pap	vegetabl	juice	lunch me	soda	pork	fruits	spaghetti	laundry d	butter	tortillas	individua	aluminun	toilet pap	poultry	toilet pap	to

Figure 4. supermarket basket transaction dataset post-processing for R Source: Self-made Excel.

For SAS the requirement was different, and the data had to be gathered by the transactions.

	A	B
1	char	num
2	01/01/2000	yogurt
3	01/01/2000	toilet paper
4	02/01/2000	soda
5	02/01/2000	cereals
6	02/01/2000	sandwich loaves
7	02/01/2000	laundry detergent
8	03/01/2000	individual meals
9	04/01/2000	ice cream
10	04/01/2000	juice
11	05/01/2000	ketchup
12	05/01/2000	sandwich loaves
13	06/01/2000	pork
14	07/01/2000	sugar
15	07/01/2000	fruits
16	07/01/2000	individual meals
17	08/01/2000	sugar
18	08/01/2000	milk
19	08/01/2000	sandwich bags
20	09/01/2000	individual meals
21	10/01/2000	shampoo
22	11/01/2000	waffles
23	11/01/2000	cheeses
24	11/01/2000	vegetables
25	11/01/2000	fruits
26	11/01/2000	bagels
27	12/01/2000	fruits
28	13/01/2000	laundry detergent
29	13/01/2000	pork
30	13/01/2000	pasta
31	14/01/2000	flour
32	15/01/2000	aluminum foil
33	15/01/2000	soap
34	15/01/2000	sandwich loaves
35	16/01/2000	lunch meat
36	16/01/2000	aluminum foil
37	17/01/2000	soap

Figure 5. supermarket basket transaction dataset post-processing for SAS Source: Self-made Excel.

Note that the code for the cleaning has been added on the appendix.

TASK: ASSOCIATION RULES

In this section will be performing association rules on the dataset using the Apriori algorithm This technique aims to find 'interesting'[1] relationship within the dataset. In order to detect this association rules will perform an analysis with R and with SAS.

- ASSOCIATION RULES IMPLEMENTATION IN R

An initial exploration of the data we can see that the most purchased item are Vegetables, followed by Poultry, while the least is Bagels.

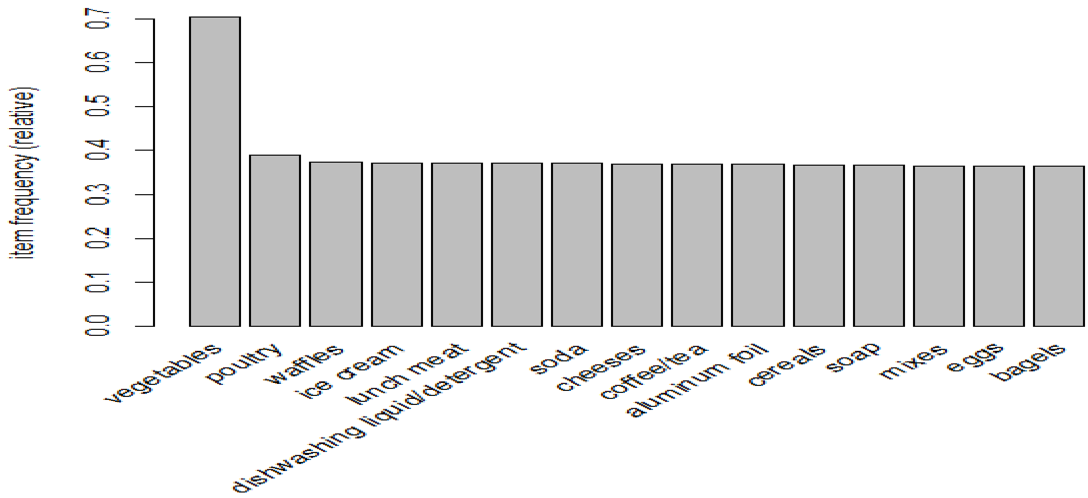


Figure 6. The most frequent item purchased. Source: Self-made R.

The Apriori algorithm was used to detect association rules. The thresholds were set very low in order to have as many rules as possible and have a better picture of the associations.

```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.1    0.1    1 none FALSE            TRUE       5     0.1      1      2  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 150

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[1508 item(s), 1500 transaction(s)] done [0.02s].
sorting and recoding items ... [38 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 done [0.00s].
writing ... [1444 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

Figure 7. Association rules overview. Source: Self-made R.

As we can appreciate on figure 7 that sets the support and confidence as low as 10% the Apriori algorithm returns 1947 rules.

Inspecting the top 10 rules, the Apriori returns the associations rules along with its support, confidence and lift ordered by lift.

	lhs		rhs	support	confidence	lift	count
[1]	{sandwich bags}	=>	{cheeses}	0.1573333	0.4618395	1.2527293	236
[2]	{cheeses}	=>	{sandwich bags}	0.1573333	0.4267631	1.2527293	236
[3]	{toilet paper}	=>	{juice}	0.1573333	0.4402985	1.2391140	236
[4]	{juice}	=>	{toilet paper}	0.1573333	0.4427767	1.2391140	236
[5]	{shampoo}	=>	{juice}	0.1500000	0.4385965	1.2343241	225
[6]	{juice}	=>	{shampoo}	0.1500000	0.4221388	1.2343241	225
[7]	{juice}	=>	{yogurt}	0.1573333	0.4427767	1.2299354	236
[8]	{yogurt}	=>	{juice}	0.1573333	0.4370370	1.2299354	236
[9]	{shampoo}	=>	{dinner rolls}	0.1513333	0.4424951	1.2246175	227
[10]	{dinner rolls}	=>	{shampoo}	0.1513333	0.4188192	1.2246175	227

Figure 8. Association rules overview. Source: Self-made R.

In order to be able to inspect the data, the Apriori algorithm has rerun setting the confidence at 85% to have fewer association rules and to be able to explore best rules and to be able to visualize the graphs.

An interesting graph that shows us the rules distribution is the scatter plot. It maps the relation between Confidence and Support.

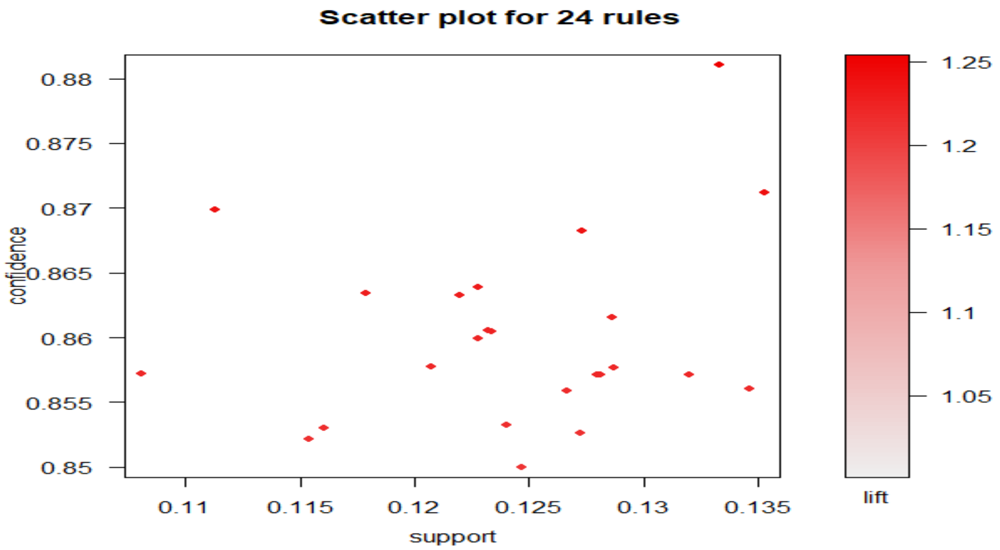


Figure 9. Association rules scatter plot. Source: Self-made R.

The groups of Matrix show the association's rules found order by lifts. The colour of the lift bubble represents the interestingness of the rule.

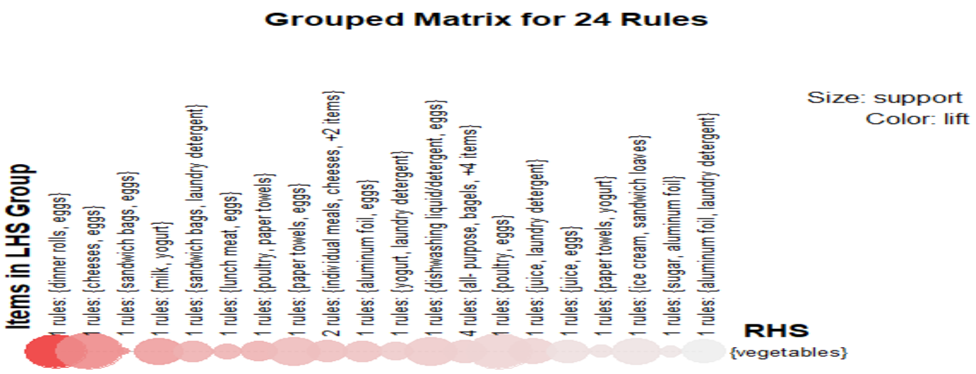


Figure 10. Association rules group matrix. Source: Self-made R.

The parallel coordinates [5] allow the visualization the in a high-dimensional geometry and analysing multivariate data.

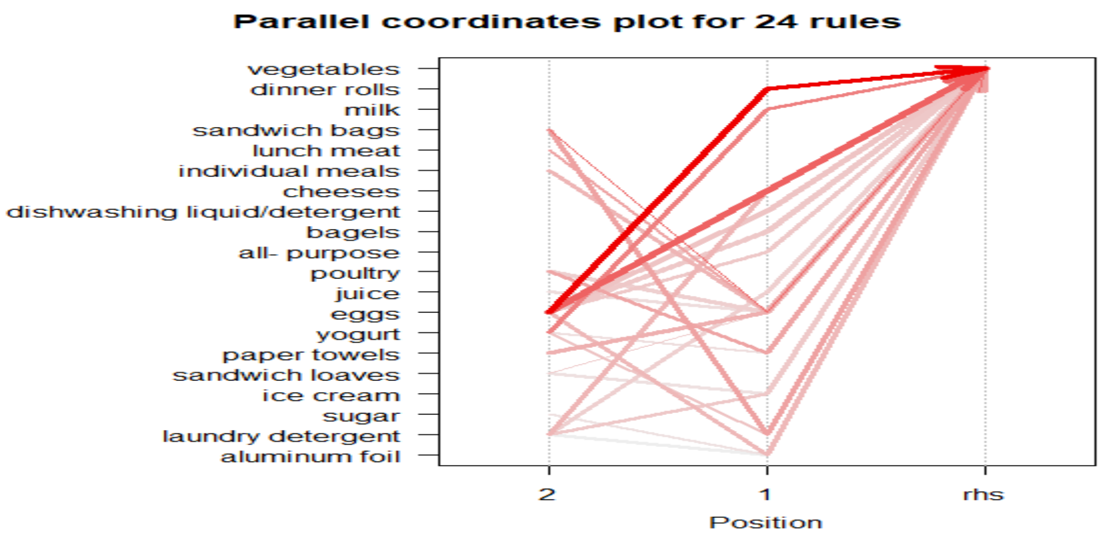


Figure 11. Association rules parallel coordinates. Source: Self-made R.

The association's rules parameters matrix shows an overview of the relationship among all the parameters such as Support, Confidence lift and count.

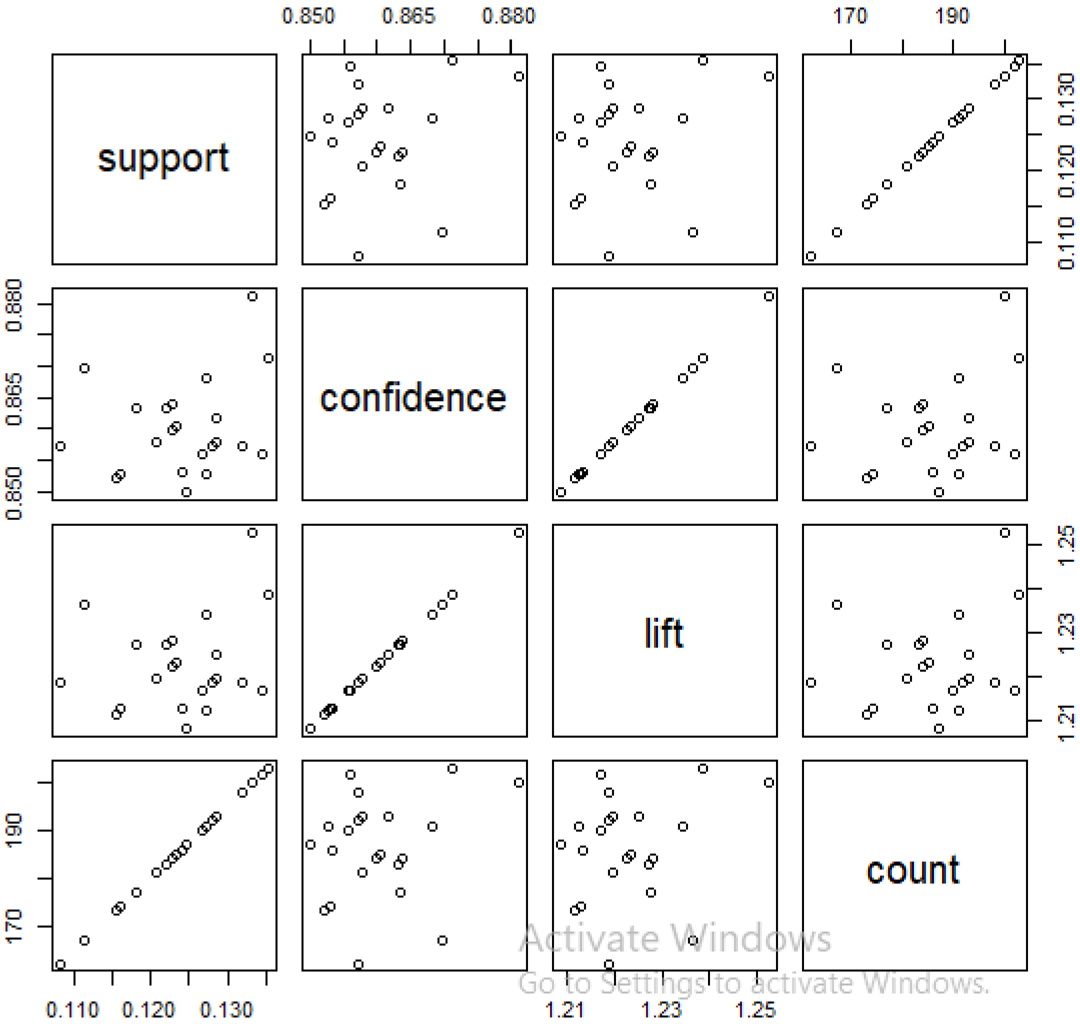


Figure 12. Association rule parameters matrix coordinates. Source: Self-made R.

Another exciting feature that R uses for Data exploration is the rule Explorer () function.

Figure 13. Association rules Ruler explorer. Source: Self-made R.

- ASSOCIATION RULES IMPLEMENTATION IN SAS.

Since SAS Miner is very user-friendly, there is no need for condign to visualize the association rules. It just needs to import the information and link it with the Association rules method for the data.

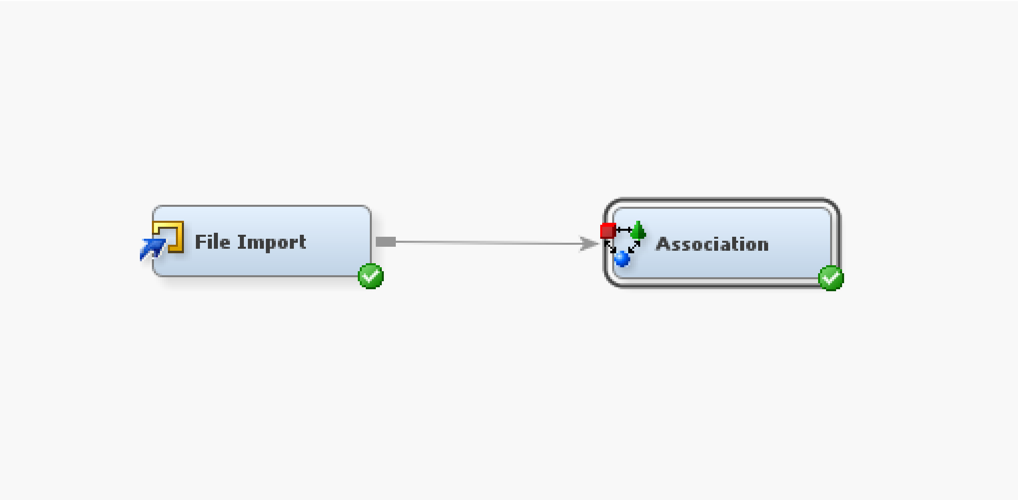


Figure 14. Association rules SAS. Source: Self-made SAS.

It needs for selecting the ID variable and the Target variable. In this case are data and item, respectively.

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Date	ID	Nominal	No		No	.	.
Item	Target	Nominal	No		No	.	.

Figure 15. Association rules variables SAS. Source: Self-made SAS.

The Apriori algorithm's parameters need to be set. As in the R case, we will set the parameter very low in order to be able to visualize the data.

General	
Node ID	Assoc
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Maximum Number of Items to	100000
Rules	...
Association	
Maximum Items	2
Minimum Confidence Level	10
Support Type	Percent
Support Count	.
Support Percentage	5.0
Sequence	
Chain Count	3
Consolidate Time	0.0
Maximum Transaction Duratio	0.0
Support Type	Percent
Support Count	.
Support Percentage	2.0
Rules	
Number to Keep	200
Sort Criterion	Default
Number to Transpose	200
Export Rule by ID	No
Recommendation	No
Status	
Create Time	07/12/18 17:46
Run ID	220886ca-499a-45fe-82b3-2
Last Error	
Last Status	Complete
Last Run Time	10/12/18 18:48
Run Duration	0 Hr. 0 Min. 5.15 Sec.
Grid Host	
User-Added Node	No

Figure 16. Association rules setting rules for Apriori algorithm. Source: Self-made SAS.

The Apriori algorithm returns the association rules along with the parameter confidence, support and lift.

Rules Table

Relations	Expected Confidence(%)	Confidence(%)	Support(%)	Lift	Transaction Count	Rule	Left Hand of Rule	Right Hand of Rule	Rule Item 1	Rule Item 2	Rule Item 3	Rule Index	Transpose Rule
2	4.96	17.65	0.76	3.56	6.00	yogurt ==> pape...	yogurt	paper towels	yogurt	============...	paper towels	1	1
2	4.32	15.38	0.76	3.56	6.00	paper towels ==...	paper towels	yogurt	paper towels	============...	yogurt	2	1
2	4.96	16.13	0.64	3.25	5.00	ketchup ==> ch...	ketchup	cheeses	ketchup	============...	cheeses	3	1
2	3.94	12.82	0.64	3.25	5.00	cheeses ==> ke...	cheeses	ketchup	cheeses	============...	ketchup	4	1
2	5.08	15.38	0.76	3.03	6.00	paper towels ==...	paper towels	aluminum foil	paper towels	============...	aluminum foil	5	1
2	4.96	15.00	0.76	3.03	6.00	aluminum foil =...	aluminum foil	paper towels	aluminum foil	============...	paper towels	6	1
2	4.96	14.71	0.64	2.97	5.00	yogurt ==> lunc...	yogurt	lunch meat	yogurt	============...	lunch meat	7	1
2	4.32	12.82	0.64	2.97	5.00	lunch meat ==> ...	lunch meat	yogurt	lunch meat	============...	yogurt	8	1
2	4.57	13.51	0.64	2.95	5.00	eggs ==> all- pu...	eggs	all- purpose	eggs	============...	all- purpose	9	1
2	4.70	13.89	0.64	2.95	5.00	all- purpose ==...	all- purpose	eggs	all- purpose	============...	eggs	10	1

Figure 17. Association rules results table algorithm. Source: Self-made SAS.

The statistic plot returns a plot where it shows the support regarding the confidence, and it helps us to understand the association rules.

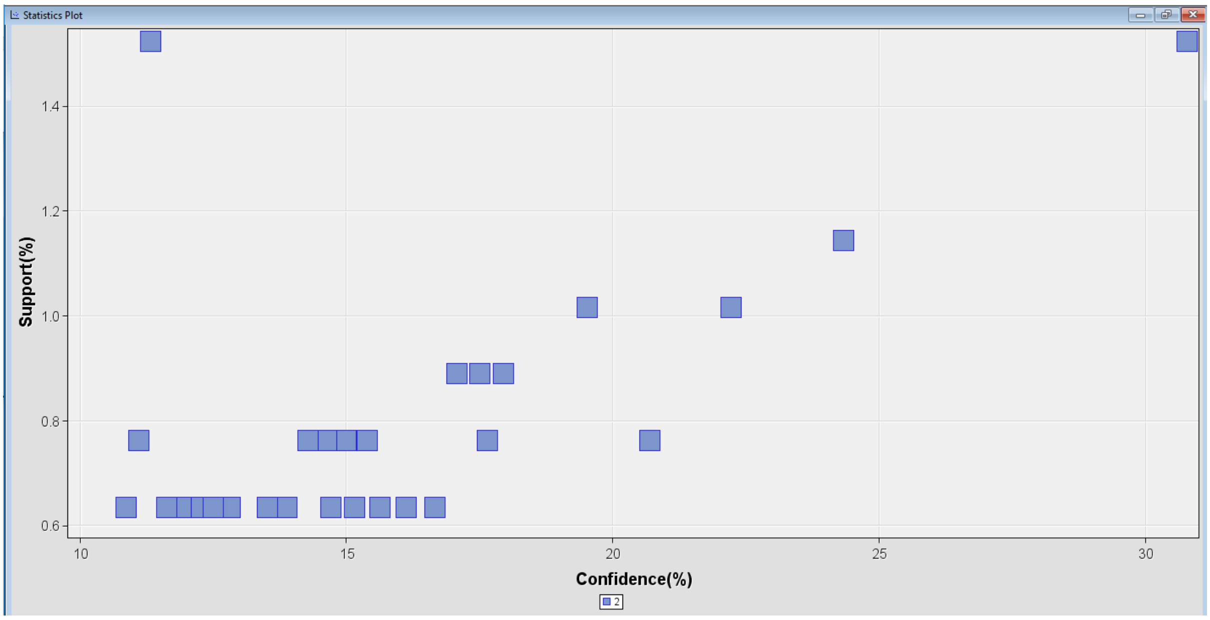


Figure 18. Association rules statistic plot. Source: Self-made SAS.

The Rule matrix returns the relation between left hand of rules and right hand of rules.

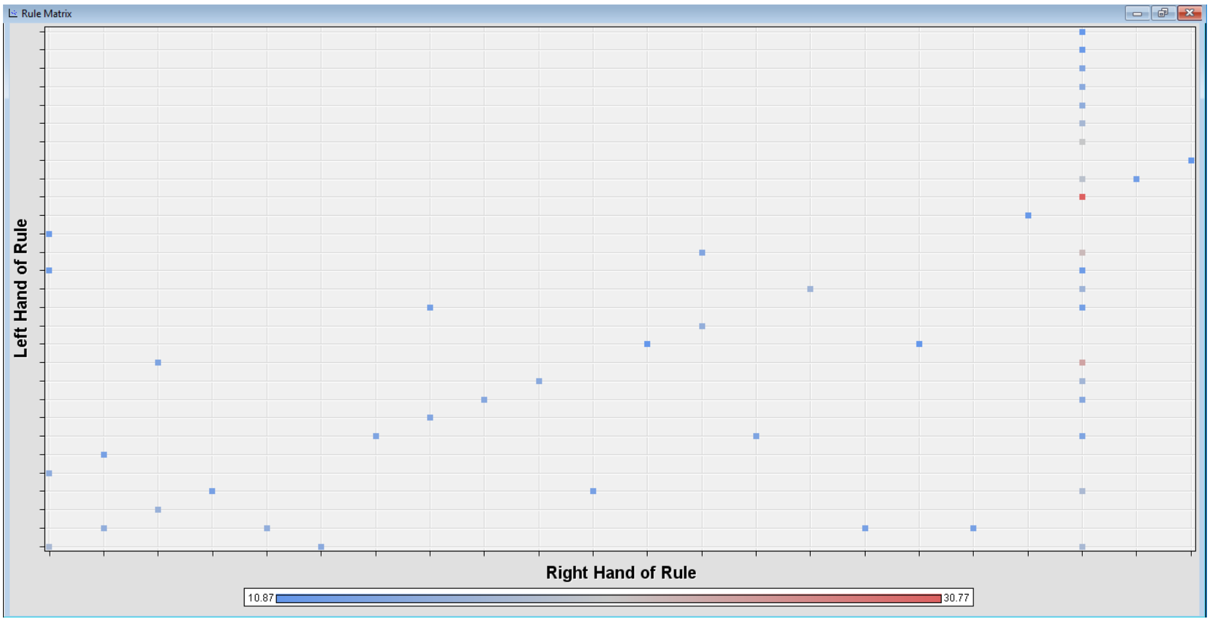


Figure 19. Association rules Rule matrix. Source: Self-made SAS.

The statistic Use plot shows the relation between paraments and rules.

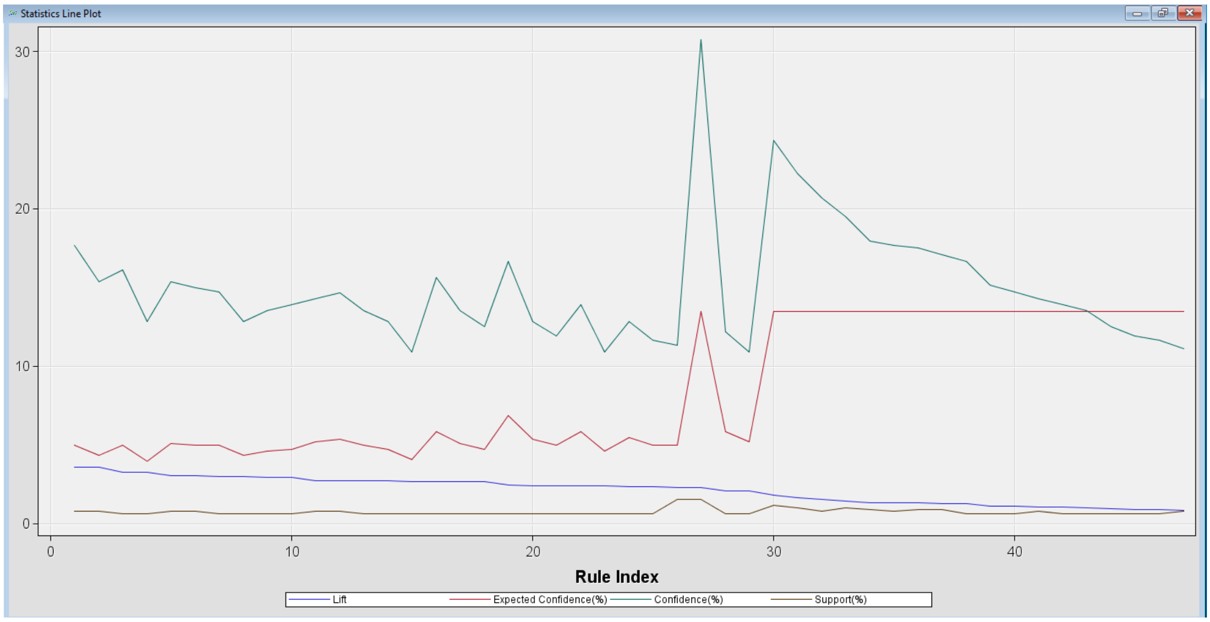


Figure 20. Association rules statistic use pot. Source: Self-made SAS.

CONCLUSION

The association rules are one of the most advanced techniques to find associations among items in a dataset.

Parameters such as confidence, support evaluate the rules found on a dataset and lift. However, not a high value on this parameter means that the rules are useful. The interestingness on this rule falls on their actionability and their newness.

On this assignment has been focused on showing the process of how the association rule works and how we can explore the data to find fascinating insight, but it did not aim to find the most newness and actionable association on the dataset.

The association rules are an excellent methodology to explore data and gain insight, and for this purpose, in my opinion, R is much better since It makes the research much more customizable for the use. On the other hand, SAS makes the process fast and easy, but It is a bit harder to explore data and customize the search.

In conclusion, Association Rules are great to search for insight in a database. It needs from an expert to find interestingness on the associations since the parameters are suitable for filtering but need a quality assessment that only a human can do.

REFERENCES

1- Associations Rules notes, Salford University. Dr.M Saraee.
2- https://www.saedsayad.com/association_rules.htm
3- https://en.wikipedia.org/wiki/Apriori_algorithm
4- https://www.rdocumentation.org/packages/tidyr/versions/0.8.2/topics/separate
5- https://en.wikipedia.org/wiki/Parallel_coordinates
6- https://cran.r-project.org/web/packages/arules/arules.pdf
7- https://blackboard.salford.ac.uk/bbcswebdav/pid-3341961-dt-content-rid-7430977_1/courses/SG-G500-M0141-T1-M-19/arulesViz.pdf

APPLY TEXT MINING ON DATASET USING R & SAS

INTRODUCTION

On today's' world the information has a significant role in how society takes decisions. The numeric decision seems to control the explosion of the new economy based on data, but nothing could be further from the truth than this. Approximately, 90 % of the data is not structured [1]. It means that there is substantial amount information not being utilized yet and this data could turn out to be essential to understand the way society works, as human mainly use qualitative data to make decisions.

Unstructured data is one of the most promising fields nowadays in Data Science, and many institutions and companies are investing in developing techniques to understand better the data and make it actionable.

The success of this research could make our economies more efficient and lead our societies to a different stage and make possible to achieve better welfare for everyone, as we have experienced in the last decades with the rest of the technological breakthroughs.

AIM AND OBJECTIVE OF THE TASK

This assignment aims to perform a text mining analysis to retrieve information from the Hotel reviews database and turn it into text categorization and trend topic discovery.

BRIEF LITERATURE REVIEW

Text mining [1] is one of the branches of data mining, but instead of working with structured data It works with unstructured such as Word files, PDF files, XML files and so forth.

Text mining aims to extract information, tack topics, summarize, categorize, clustering, linking concepts and answer questions.

Text mining process is made up of 3 steps:

Step 1: Establish the corpus

- Collect all relevant unstructured data

Step 2: Create the Term–by–Document Matrix (TDM).

Step 2: Create the Term–by–Document Matrix

Documents \ Terms	investment risk	project management	software engineering	development	SAP	...
Document 1	1			1		
Document 2		1				
Document 3			3		1	
Document 4		1				
Document 5			2	1		
Document 6	1			1		
...						

Figure 1. Term- by-Document Matrix. Source: Text Mining notes [1].

Step 3: Extract patterns/knowledge

- Classification
- Clustering
- Association
- Trend Analysis

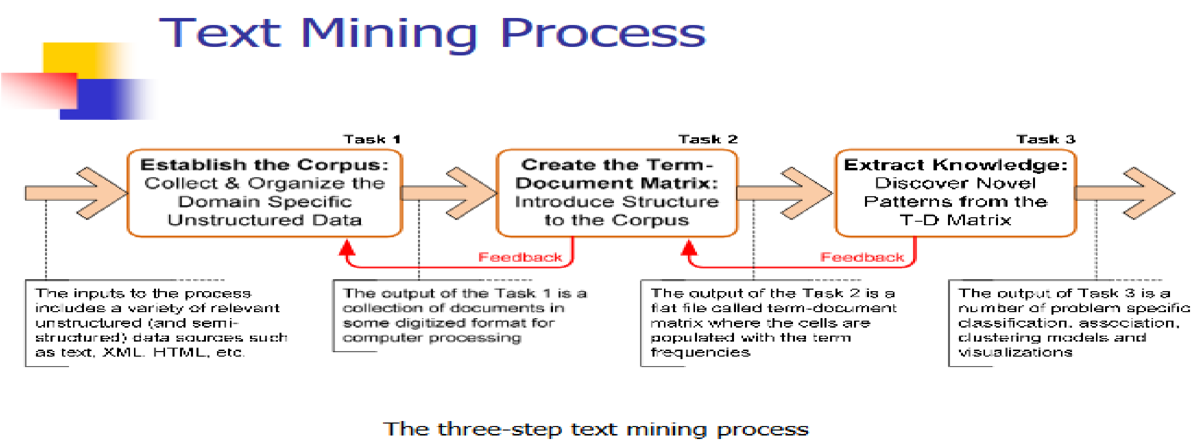


Figure 2. Text mining process. Source: Text Mining notes [1].

Even though the Text mining is a brilliant tool to analyse unstructured data it has disadvantages such as very high number of possible “dimensions,” unlike data mining, complex relationships between concepts in text, ambiguity and context sensitivity, word ambiguity, noisy data and not well-structured text.

DATA SEARCH STRATEGY

The data selected is a hotels review dataset. The dataset used is on https://blackboa.salford.ac.uk. It is a mandatory task in order to summit the assignment.

EXPLANATION AND PREPARATION OF DATASETS

Hotel reviews is a dataset made up of 21094 rows and 8 columns. The attributes are Review iD, Hotel name, Travel review account, Review Date, Review via Mobile, Guest Location, Review Heading and review.

	A	B	C	D	E	F	G	
1	Review id	Hotel Name	Total review count	Review Date	Review_viaMobile	Guest Location	Review Heading	Review
2	rn629306037	Holiday Inn Resort Kandooma Maldives	1679	October 29, 2018	via mobile	Waikiki, Hawaii	Amazing!!!	This is one of my most favorite resort that I have ever visited in Maldives. The
3	rn623195151	Holiday Inn Resort Kandooma Maldives	1679	October 8, 2018	via mobile	Berkshire, United Kingdom	Stunning, wonderful and peaceful	One of the more affordable resorts in the Maldives, but still absolutely unreal!
4	rn627907077	Holiday Inn Resort Kandooma Maldives	1679	October 24, 2018	via mobile	Edinburgh, United Kingdom	We don't want to leave!	We arrived here after an unfortunate experience at another hotel. Right from t
5	rn623362603	Holiday Inn Resort Kandooma Maldives	1679	October 8, 2018	NA	Melbourne, Australia	Great Madlivian Experience	Firstly, the Maldivian staff were all very friendly and wonderful throughout the
6	rn623255673	Holiday Inn Resort Kandooma Maldives	1679	October 8, 2018	via mobile	Dubai, United Arab Emirates	Good hotel friendly staff but make sure to see the room first before taking it	We loved the resort and the friendly staff and the beautiful view. The only thi
7	rn615351040	Holiday Inn Resort Kandooma Maldives	1679	September 10, 2018	NA	Copenhagen, Denmark	Great Holiday!	We were met by great personnel at the airport and swiftly taken to the boat.
8	rn614605990	Holiday Inn Resort Kandooma Maldives	1679	September 7, 2018	via mobile	Singapore, Singapore	Maldivian getaway	Anywhere in the maldives is beautiful and holiday inn is a slightly more afford
9	rn613803037	Holiday Inn Resort Kandooma Maldives	1679	September 4, 2018	via mobile	NA	Amazing getaway!	We went for a holiday along with our 11 month old kid.
10	rn613651262	Holiday Inn Resort Kandooma Maldives	1679	September 3, 2018	NA	NA	sublime....	We stayed in this resort last week of July 2018, thank goodness the last storm
11	rn613530618	Holiday Inn Resort Kandooma Maldives	1679	September 3, 2018	NA	London, United Kingdom	Paradise	Just returned from family trip to Kandooma, absolutely fabulous, pure paradis
12	rn611296877	Holiday Inn Resort Kandooma Maldives	1679	August 27, 2018	via mobile	NA	Magical 8 night at holiday in Kandooma	The hotel was amazing, we took beach house which was one bedroom on the
13	rn610856718	Holiday Inn Resort Kandooma Maldives	1679	August 27, 2018	NA	Singapore, Singapore	Efficient personalised service provided	Hotels in Maldives charge by the number of people staying in the hotel, unlike
14	rn608993575	Holiday Inn Resort Kandooma Maldives	1679	August 21, 2018	NA	London, United Kingdom	This place is amazing!!	I spent 5 glorious days with my wife on this resort. It caters for coupes and is fa
15	rn608834228	Holiday Inn Resort Kandooma Maldives	1679	August 21, 2018	NA	London, United Kingdom	Unforgettable experience that is so family friendly	This was a magical experience for the whole family. We arrived at Male airport
16	rn607488241	Holiday Inn Resort Kandooma Maldives	1679	August 17, 2018	via mobile	NA	Great place for family and kids	A wonderful beautiful place to be. Ideal for 5 nights stay. Food 10/10. Service
17	rn599774144	Holiday Inn Resort Kandooma Maldives	1679	July 26, 2018	NA	Muscat, Oman	Amazing resort to visit if you're thinking of going to the Maldives!	Holiday Inn Kadooma is in beautiful, yet convenient location! After landing in
18	rn599069586	Holiday Inn Resort Kandooma Maldives	1679	July 23, 2018	via mobile	Charlotte, North Carolina	Great Service and Staff!!	Highly recommended!!! Great for Families. Multiple kids pools with a small sp
19	rn597171551	Holiday Inn Resort Kandooma Maldives	1679	July 17, 2018	via mobile	Singapore, Singapore	Paradise	Ah where do I start?
20	rn596563222	Holiday Inn Resort Kandooma Maldives	1679	July 15, 2018	NA	NA	Family Trip in June	Our family of 4 stayed in in a over water villa and highly recommend it. Room s

Figure 3. Hotel Reviews screenshot. Source: Text Mining notes [1].

As my laptop cannot support more than 8 GB of memory RAM. I had to remove a few attributes that were redundant and a few hotels reviews in order to be able to run the software.

```
> termFrequency<-rowSums(as.matrix(dtm))
Error: cannot allocate vector of size 8.6 Gb
```

Figure 4. R screenshot of RAM issue. Source: Self-made in R

The new data is made up of 7294 rows and 3 attributes such as a Hotel name, Review Heading and review

	A	B	C
1	Hotel Name	Review Heading	Review
2	Bandos Maldives	Amazing resort for a honeymoon	Overall an amazing resort. It made our honeymoon unforgettable. The island was beautiful andyou can make nice pictures around the island. Food was delicious and we enjoyed our private dineron th
3	Bandos Maldives	excellent stay at bandos	Well, readers ignore all reviews of bandos Maldives written before this date, so I will start with my arrival at the airport. As I arrived at the airport with my family, and immediately I boarded the speed
4	Bandos Maldives	Very good service level and excellent snorkeling opportunities	Very good service level and excellent snorkeling. All the surroundings gives a perfect nature experience. To me this is paradise. We enjoyed our stay very much. I consider this as one of our best family
5	Bandos Maldives	Bandos - my favourite Maldivian island.	Bandos is the best Maldivian island I visited til now. Perfect facilities, very good choice of food, very friendlystaff. Very fine for snorkeling. I didn't see such a variety of fishes / turtles before.More
6	Bandos Maldives	Absolutely perfect!	Went to Bandos for our first time in the Maldives from the 23rd of October to the 1st of November. From the second we stepped off the speedboat at Bandos our trip was absolutely amazing. Food was
7	Bandos Maldives	Bandos vacation	Great for scuba,but the beach is a litle short and not enough sun beds . The rooms are cleaned too times a day and they provide one large bottle of water for each person every day. The buffet was goo
8	Bandos Maldives	Awesome abode of serenity	The island of Bandos is a perfect setting for that romantic getaway or my one week of bliss away from work.Lovely place/Turqouise blue water/supportive staff/clean/smooth check in /fast speedboat
9	Bandos Maldives	All was perfect	The island is very nice and the houseriff is so beautifull. The Service is absolutly perfect and the meal was delicious. The room was always clean and it was so great that the room was cleaned two times
10	Bandos Maldives	Amazing stay, thank you very much!	Stayed for 6 nights (20th to 26th of Oct 2018) and absolutely loved this place. We were picked from the airport and the transfer by boat takes only 10 minutes. We were welcomed with an iced tea and
11	Bandos Maldives	Fantastic resort island with amazing snorkelling!	This was my first trip to the Maldives, and I can't recommend Bandos enough for how welcoming they were and the level of service we received. The transfer from the Airport was efficient, and the sta
12	Bandos Maldives	Perfect holiday	It was great expirience on Maldives. Beautiful place for holiday, and beautiful place for diving. Thereis so many shark and tropical fish i can imagen how it was great. Stuff are great and very nice and fo
13	Bandos Maldives	An Awesome Stay	Stay - perfectService - awesomeBuffet - Good - Need some improvement / addition here. Try add more local dishes as well as Asian food as well.Cleanliness - Room was cleaned so as surroundings but
14	Bandos Maldives	Good rooms but poor water sports	Good food, nice rooms and excellent value for money considering only the above. But water sports prices and organization of the packages, rules and polices reeks of 'take it or leave it' attitude-Bad.M
15	Bandos Maldives	Amazing	It was amazing, very helpful staff/brilliant food/ the place and Beach are amazing. weather <U+0001F609> was good as well evening entertainments top class. Definitely place to recommend (close to m
16	Bandos Maldives	Nice island, nice stay	We really had great days there. Comfortable and nice rooms. Everything clean.The breakfast and dinner were absolutely awesome. We also had great dives on our boat trip with the diving centre.Highl
17	Bandos Maldives	Beautiful swimming beach	You
18	Bandos Maldives	Recommended	I was very satisfied with everything on the Bandos. The staff was very friendly, ready to help, they treat you like you are the most important person on the world. The villa on the beach was very good,
19	Bandos Maldives	Fun	Fantastic beautiful place and spa clean tidy friendly atmosphere food lots of variety worldwide value for money pools great staff brilliant in all areas room clean and tidy entertainment brilliant and fu
20	Bandos Maldives	Perfect resort	This resort was perfect for our honeymoon! My wife and I cant wait to come back. Every employee was very nice And always asked if we needed assistance. Scuba diving was beautiful. I greatly recomm
21	Bandos Maldives	Superb & Awesome	It an
22	Bandos Maldives	Beautiful island, room equipment is in need of improvement	Really beautiful island, the food tastes great and is very varied, frendly employees, but the room equipment is in need of improvement. The furniture is outdated, the couch unusable. We had a Jacuzz
23	Bandos Maldives	Overall satisfied :-)	Nice island with very good services, tasty and varying meals in restaurant, espresso machines everywhere, well established dive center, big swimming pool with bar in the middle of island and sunrise
24	Bandos Maldives	Thanks	Wonderful island and it will be one of our best memories My daughter enjoyed all facilities there and she pushed us to do another visit The weather was amazing and the night life alsoThanks for ever
25	Bandos Maldives	Heavenly island with kindest staff members	The Island of Bandos is a jewel within the Maldives. The whole resort is absolutely recommendable, we will travel there again, for sure. One of the best things were the reef sharks. Upon arrival, we w
26	Bandos Maldives	First time to Maldives and Bandos will be my come back resort	Over all a great experience, in resort prices are bit pricey but you get service in return. Stuff are very friendly and helpful. Make sure to get flu medications and painkillers, etc since the in-resort clinic
27	Bandos Maldives	Stay at Bandos-Mid September.	Arrived at Maldives on third week of September. The weather was just perfect. It drizzled slightly on the first day. The hotel was good the the staff were attentive. The transfer from the airport was or
28	Bandos Maldives	NOT RECOMMENDED	Expect disappointments after seeing your roomBook sports in advance which is expensive even for non motorized. Need to wait and slog until your turn due to limited availability.Overall unhappy, dis
29	Bandos Maldives	Awesome stay at Hand is.	We used the Jacuzzi beach villa. It was awesome. It was a 10sec walk to the beach from the villa. Beautifully landscaped surrounding. We had an infant travelled with us, all staff were very friendly and
30	Bandos Maldives	A magical holiday	We had booked the pool villa & a spa room. The pool villa had a sundeck plus a wooden pergola next to the infinite pool & the Jacuzzi tub was indoors in the adjoining large bathroom that had large gla
31	Bandos Maldives	Honeymoon in Bandos	The Bandos offers great beachside villa's with Jacuzzi, we were offered free wine in our package, the island walk over whitish sands makes it a memorable stay. The scuba diving crew was also superb.
32	Bandos Maldives	grate	its ideal resort to stay in Maldives. few minutes from the airport. grater foods and can enjoy the stay. its recommended to stay bandos. very friendly helpful staffs, serve with smile. one of the recomm
33	Bandos Maldives	Perfect stay!	The Location, the Island, the Rooms and Food are perfect; For us is was "the perfect" stay for our family. During our travel time (August and September), the island was not too busy. We will come back
34	Bandos Maldives	Amazing Honeymoon Location	We stayed for 7 days at Bandos Maldives and it was absolutely incredible. We arrived by speedboat which was about 15 minutes away from Male airport. We were greeted by amazing staff who made
35	Bandos Maldives	Excellent Value for Money	Co-operative Staff Caring Chef Amit We had excellent time during our 3 days stay whenever we go to Maldives this will be our destination place.The Bar Area also is well managed Thanks Bandos for g
36	Bandos Maldives	Its is an average hotel near to the airport	The hotel sells ocean view rooms which does not have any view of the ocean at all. The service of the staff were excellent though and the price of the food is reasonable. The airport transfer is relative
37	Bandos Maldives	Very Good	Everything was perfect.but buggy vehicle driver was very rude...very disappointed only about that,and we called reception at the check out time and asked just send some buggy vehicle.and they send

Figure 5. Hotel Reviews screenshot after information removal. Source: Text Mining notes [1].

The file before cleaning is very messy and full of inconvenient such as a capital letter, punctuation, numbers, stop words and white spaces.

```
> inspect(mycorpus[3])
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 1

[1] Bandos Maldives,excellent stay at bandos,"Well, readers ignore all reviews of bandos Maldives written before this date, so I will start with my arrival at
 the airport. As I arrived at the airport with my family, and immediately I boarded the speed boat to the island, barely I could steal...More"
```

Figure 6. Hotel Reviews pre-processing. Source: self-made R.

To clean the text, we will use the package tm[2], which is specially made to deal with text mining problems.

After performing the removing of capital letter, punctuation, numbers, stop words and white spaces as well as a few words such as for instance the name of the hotels, the text looks as per below:

```
> inspect(mycorpus[3])
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 1

[1] bandos maldivesexcellent stay at bandoswell readers ignore all reviews of bandos maldives written before this date so i will start with my arrival at the
airport as i arrived at the airport with my family and immediately i boarded the speed boat to the island barely i could stealmore
```

Figure 7. Hotel Reviews post-processing. Source: self-made R.

TASK: TEXT MINING

- TEXT MINING IMPLEMENTATION IN R

As seen on the explanation and exploration section, the data set had to be cleaned in order to be able to deal with the text and perform the text mining analysis.

Once the establish corpus is set, the next step is to make up the term-document Matrix for later to be able to extract information. DTM [2] is a matrix that lists all occurrences of words in the corpus. In DTM, documents are represented by rows and the terms (or words) by columns. If a word occurs in a particular document n time, then the matrix entry for corresponding to that row and column is n if it does not occur at all, the entry is 0.

```
> dtm
<<TermDocumentMatrix (terms: 31033, documents: 37361)>>
Non-/sparse entries: 514978/1158908935
Sparsity           : 100%
Maximal term length: 983
Weighting          : term frequency (tf)
```

Figure 8. Term-document Matrix. Source: self-made R.

Once the DTM is built, we can see that 31033 terms are found over 37361 documents. To filter the information, we will search words that are repeated at least 1000 times to see the most important topics on the reviews.

```
> findFreqTerms(dtm,lowfreq = 1000)
 [1] "ever"        "holiday"     "inn"         "kandooma"    "one"         "staff"       "boat"        "every"       "nights"      "wonderful"   "airport"
[12] "arrived"     "excellent"   "experience"  "service"     "beach"       "stayed"      "best"        "everything"  "friendly"    "stay"        "beautiful"
[23] "first"       "loved"       "room"        "view"        "great"       "water"       "good"        "many"        "went"        "clean"       "day"
[34] "fantastic"   "week"        "family"      "food"        "helpful"     "paradise"    "special"     "trip"        "amazing"     "themore"     "really"
[45] "days"        "kids"        "place"       "spent"       "like"        "male"        "can"         "location"    "visit"       "also"        "small"
[56] "time"        "rooms"       "perfect"     "back"        "vacation"    "honeymoon"   "much"        "nice"        "pool"        "well"        "just"
[67] "get"         "made"        "lovely"      "sea"         "ocean"       "will"        "spa"         "shangrilas"  "villingili"
```

Figure 9. Most frequent terms on DTM. Source: self-made R.

Exploring further in the DTM we can find how many times the most frequent terms occurred.

```
> termFrequency <-subset(termFrequency,termFrequency>=1000)
> termFrequency
  amazing     beach beautiful      food      stay      best      good   service     great      time     water      room     place  friendly     staff
     1649      1313      1350      1674      1364      1100      1348      1211      1588      1267      1590      1105      1420      1042      2439
   stayed
     1328
```

Figure 10. Most frequent terms on DTM with number of occurrences. Source: self-made R.

Visualize the terms will make us understand better what the most popular topics on the reviews are.

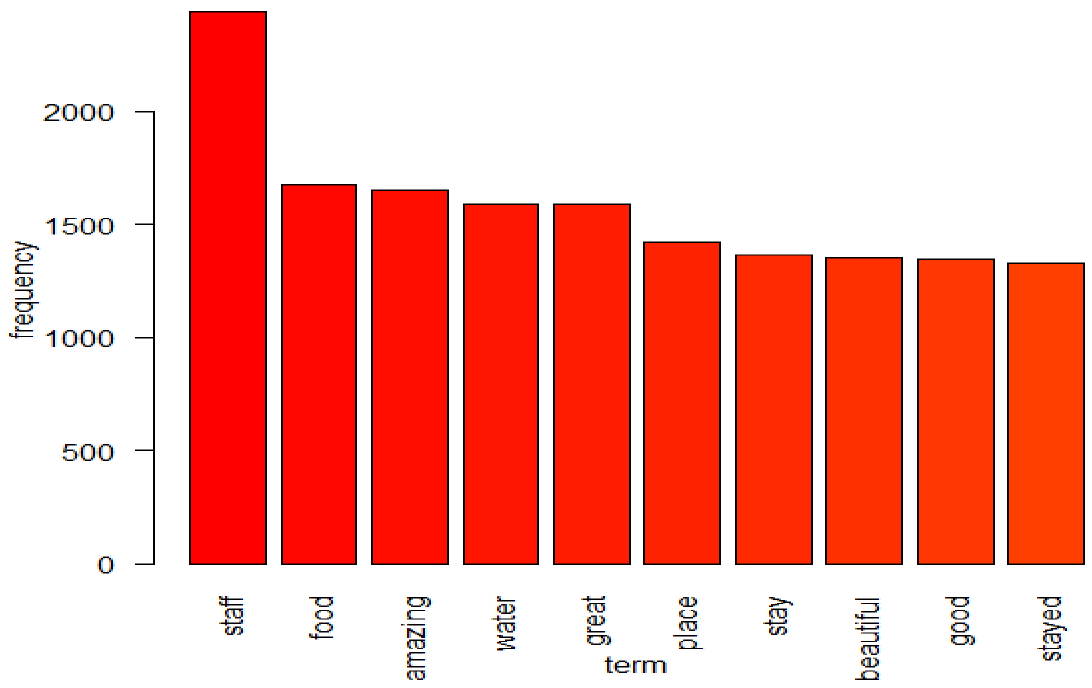


Figure 11. Most frequent terms ordered by frequency. Source: self-made R.

Figure 12. Word cloud of the most frequent terms. Source: self-made R

- TEXT MINING IMPLEMENTATION IN SAS

Since the file could not run properly on R, I had to use the data with fewer rows and columns as mentioned previously.

The process with SAS was importing the file using the function File Import, then using the Text Parsing and Text Filter functions to make up the DTM and remove the capital letter, punctuation, numbers, stop words and white spaces as well as a few words such for instance the name of the hotel.

The final step was to link everything with the function Text topic to be able to perform the text mining Analysis.

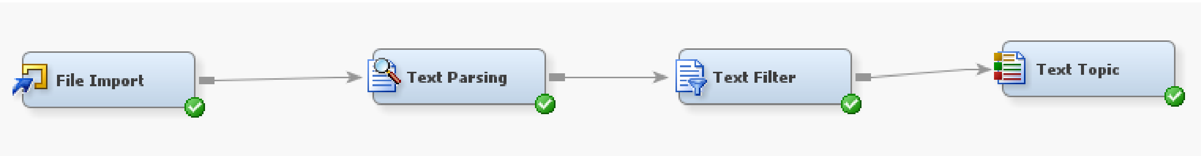
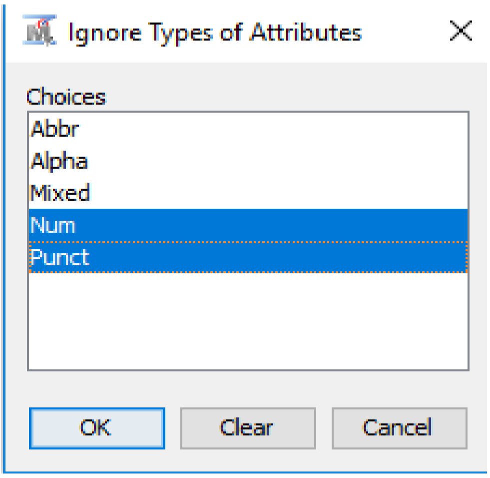
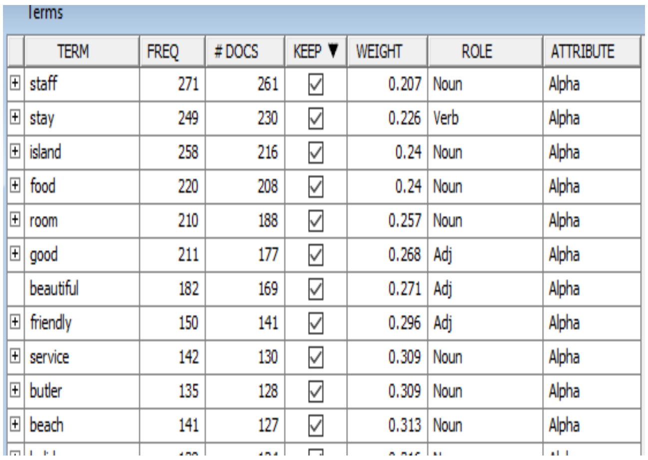


Figure 13. SAS workflow for SAS. Source: self-made SAS





Terms

TERM	FREQ	# DOCS	KEEP ▼	WEIGHT	ROLE	ATTRIBUTE
staff	271	261	☑	0.207	Noun	Alpha
stay	249	230	☑	0.226	Verb	Alpha
island	258	216	☑	0.24	Noun	Alpha
food	220	208	☑	0.24	Noun	Alpha
room	210	188	☑	0.257	Noun	Alpha
good	211	177	☑	0.268	Adj	Alpha
beautiful	182	169	☑	0.271	Adj	Alpha
friendly	150	141	☑	0.296	Adj	Alpha
service	142	130	☑	0.309	Noun	Alpha
butler	135	128	☑	0.309	Noun	Alpha
beach	141	127	☑	0.313	Noun	Alpha

Figure 14. Text mining Cleaning. Source: self-made SAS Figure 14. SAS Text mining Cleaning. Source: self-made SAS

The final step was to link everything with the function Text topic to be able to perform the text mining Analysis.

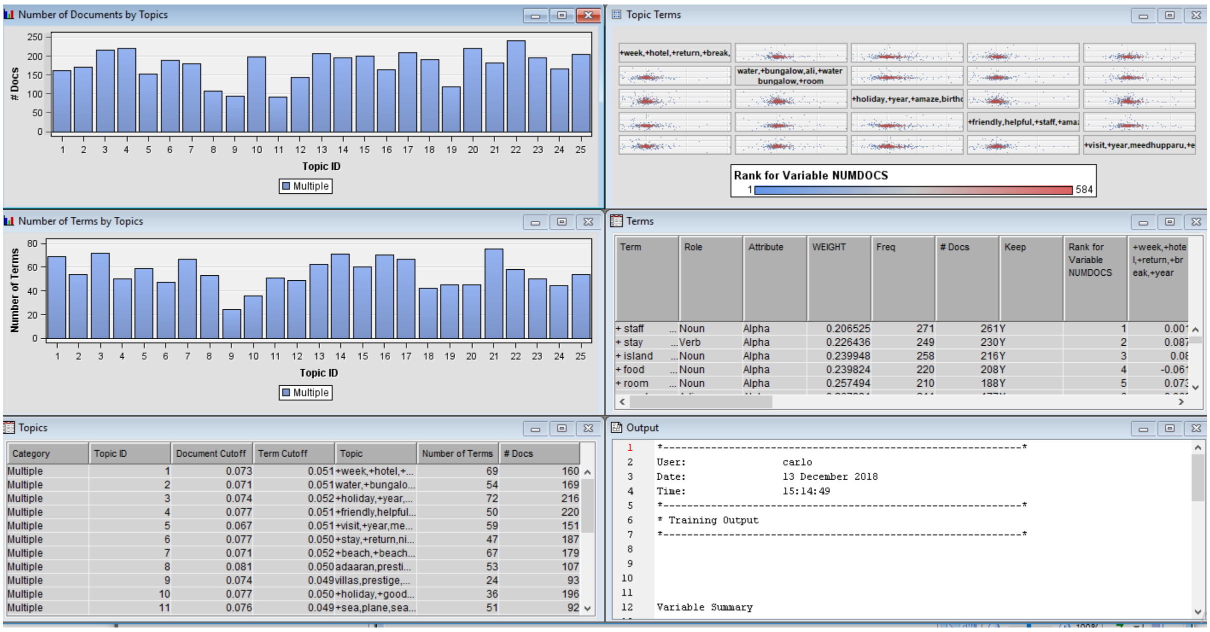


Figure 15. Text mining Results overview. Source: self-made SAS

Term	Role	Attribute	WEIGHT	Freq	# Docs	Keep	Rank for Variable NUMDOCS	+week,+hotel,+return,+break,+year	water,+bungalow,ali,+water bungalow,+room	+holiday,+year,+amaze,birthday,+book	+friendly,helpful,+staff,+amaze,fantastic	+visit,+year,meedhupparu,+experience,second	+stay,+return,night,meedhupparu,perfect	+beach,+beach villa,+amaze,+nice,villa	adaaran,prestige,vadoo,+adaaran prestige ocean villa,+butler	villas,prestige,water,ocean,+experience	+holiday,+good,+best holiday,first,+want
+ staff	... Noun	Alpha	0.206525	271	261	Y	1	0.001	0.071	0.08	0.279	0.009	-0.015	-0.023	-0.024	-0.023	0.029
+ stay	...Verb	Alpha	0.226436	249	230	Y	2	0.087	0.057	-0.02	-0.008	-0.046	-0.001	0.071	0.037	0.071	-0.054
+ island	...Noun	Alpha	0.239948	258	216	Y	3	0.08	-0.102	-0.012	-0.093	0.095	0.026	0.116	-0.048	-0.009	0.025
+ food	... Noun	Alpha	0.239824	220	208	Y	4	-0.061	0.006	-0.035	0.16	0.033	0.015	0.074	0.014	0.012	0.063
+ room	... Noun	Alpha	0.257494	210	188	Y	5	0.073	0.21	-0.115	0.059	0.045	-0.004	-0.005	0.081	-0.046	-0.002
+ good	...Adj	Alpha	0.267981	211	177	Y	6	-0.069	-0.058	-0.113	0.048	0.156	0	0.004	0.008	0.064	0.434
beautiful	...Adj	Alpha	0.270686	182	169	Y	7	0.139	-0.038	-0.025	-0.123	0.009	-0.078	-0.054	-0.005	-0.081	0.077
+ friendly	...Adj	Alpha	0.29615	150	141	Y	8	0.003	0.134	0.122	0.35	0.014	-0.021	-0.032	-0.098	-0.022	-0.027

Category	Topic ID	Document Cutoff	Term Cutoff	Topic ▲	Number of Terms
Multiple	16	0.071	0.051	+airport,male,seaplane,+transfer,+experience	70
Multiple	7	0.071	0.052	+beach,+beach villa,+amaze,+nice,villa	67
Multiple	4	0.077	0.051	+friendly,helpful,+staff,+amaze,fantastic	50
Multiple	15	0.080	0.051	+great,+recommend,+stay,highly,+friend	60
Multiple	10	0.077	0.050	+holiday,+good,+best holiday,first,+want	36
Multiple	3	0.074	0.052	+holiday,+year,+amaze,birthday,+book	72
Multiple	13	0.074	0.051	+island,lovely,+visit,beautiful,+staff	62
Multiple	14	0.074	0.051	+location,+time,meedhupparu,wonderful,snorkelling	71
Multiple	22	0.083	0.051	+nice,+good,good,+view,awesome	58
Multiple	18	0.077	0.050	+night,+stay,+water villa,+time,+couple	42

Figure 16. Text mining Results overview. Source: self-made SAS

As we can see on figure 14 and 15, SAS shows us the trendiest word from the hotel's reviews, as well as information about the clusters.

CONCLUSION

Data mining is one of the most powerful tools we have to analyse unstructured data. Text mining is one of the techniques within Data mining that aims to extract information, tack topics, summarize, categorize, clustering, linking concepts and answer questions.

The text mining technique is made up of 3 steps such a Establish the corpus, Create the Term–by–Document Matrix and Extract patterns/knowledge.

The assignment shows all three steps in detail and discusses all the different aspects of the process from stabilizing the corpus and creating the Term–by–Document Matrix to extracting knowledge from the information in the text.

The results after processing the data and applying the text mining technique with its filter such a capital letter, punctuation, numbers, stop words and white spaces as well as a few words such as for instance the name of the hotels, are on both R and SAS that the most frequent word is staff, followed by food, amazing and water.

This assignment only aims to explain the technique and how it can be used to extract valuable information, but as most of the methods, the data retrieved needs from an expert to make it useful and applicable.

REFERENCES

1- Text Mining notes, Salford University. Dr.M Saraee
2- Text Mining worksop, Salford University. Dr.M Saraee and Charith Silva
3- https://cran.r-project.org/web/packages/tm/tm.pdf by Ingo Feinerer
4- https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf by Ian Fellows
5- https://en.wikipedia.org/wiki/Text_mining

APPENDIX CLASSIFICATION CODE IN R AND SAS

- R CODE

```
###READ FILE###

library('reshape2')

setwd('C:/Users/carlo/Desktop/LOAN PREDICCTION PROJECT/Exercise 1')

data<-read.csv('Loan_Dataset.csv',header=T,na.strings=c("", "NA"))

###CLEANING MISSING VALUES###

summary(data)# Summary of data pior cleaning

data$Married[which(is.na(data$Married))]<-'Yes'

data$Gender[which(is.na(data$Gender))]<-'Male'# Used mode to fill in the missing values. Central tendency
measure

data$Dependents[which(is.na(data$Dependents))]<-'0'#Used mode to fill in the missing values. Central
tendency measure

data$LoanAmount[which(is.na(data$LoanAmount))]<-mean(data$LoanAmount,na.rm = TRUE) #Used mean to
fill in the missing values. Central tendency measure

data$Loan_Amount_Term[which(is.na(data$Loan_Amount_Term))]<-mean(data$Loan_Amount_Term,na.rm =
TRUE) #Used mean to fill in the missing values.Central tendency measure

data$Self_Employed[which(is.na(data$Self_Employed))]<-'No'

new_data<-data[-c(which(is.na(data$Credit_History))),]

rownames(new_data)<-new_data[,1]

new_data[,1]<-NULL

new_data$Credit_History_f<-as.factor(new_data$Credit_History)# Factor Target attribute Credit_History

new_data[,10]<-NULL

###OUTLIERS###

IQR_ApplicantIncome=quantile(new_data$ApplicantIncome, 0.75)-quantile(new_data$ApplicantIncome, 0.25)

value_IQR_ApplicantIncome=quantile(new_data$ApplicantIncome, 0.75)+1.5*IQR_ApplicantIncome

new_data$ApplicantIncome[which(new_data$ApplicantIncome>value_IQR_ApplicantIncome)]<-
value_IQR_ApplicantIncome

IQR_CoapplicantIncome=quantile(new_data$CoapplicantIncome, 0.75)-
quantile(new_data$CoapplicantIncome, 0.25)

value_IQR_CoapplicantIncome=quantile(new_data$CoapplicantIncome, 0.75)+1.5*IQR_CoapplicantIncome
```

```
new_data$CoapplicantIncome[which(new_data$CoapplicantIncome>value_IQR_CoapplicantIncome)]<-
value_IQR_CoapplicantIncome

IQR_LoanAmount=quantile(new_data$LoanAmount, 0.75)-quantile(new_data$LoanAmount, 0.25)

value_IQR_LoanAmounte=quantile(new_data$LoanAmount, 0.75)+1.5*IQR_LoanAmount

new_data$LoanAmount[which(new_data$LoanAmount>value_IQR_LoanAmounte)]<-
value_IQR_LoanAmounte

write_New_loan_dataset<-write.csv(new_data,'Loan_dataet_cleaned.csv')

### PACKAGES LIBRARY ###

library(randomForest)

library(caret)

library(e1071)

### OPEN FILE ###

setwd('C:/Users/carlo/Desktop/LOAN PREDICCTION PROJECT/Exercise 1')

loan_train<-read.csv('New_loan_dataset_cleaned_train.csv',header = T)

loan_train$Credit_History_f<-as.factor(loan_train$Credit_History_f)

### DATA PARTITION ###

set.seed(1234)

pd <-sample(2,nrow(loan_train),replace=TRUE,prob=c(0.8,0.2))

train <-loan_train[pd==1,]

summary(train)

validate <-loan_train[pd==2,]

rownames(train)<-train[,1]

train[,1]<-NULL

rownames(validate)<-validate[,1]

validate[,1]<-NULL

### RANDOM FOREST ###

set.seed(222)

rf<-randomForest( formula = Loan_Status ~ .,data=train, ntree=145, mtry=5,importance= T,proximity=T)

print(rf)

plot(rf)
```

```
p1<-predict(rf,train)

p2<-predict(rf,validate)

### CONFUSION MATRIX ###

confusionMatrix(p2,validate$Loan_Status)

### TUNING OF MODEL ###

tuneRF(x=subset(train,select = -Loan_Status),y = train$Loan_Status,stepFactor = 0.5, plot= T,ntreeTry =
100,trace = T,improve = 0.05)

### GRAPHS ###

hist(treesize(rf),main='No. Of nodes for the trees',col='Red')

varImpPlot(rf)

importance(rf)

varUsed(rf)
```

- SAS CODE

```
%macro em_hpfst_score;


 %if %symexist(hpfst_score_input)=0 %then %let hpfst_score_input=&em_score_output;

 %if %symexist(hpfst_score_output)=0 %then %let hpfst_score_output=&em_score_output;

 %if %symexist(hpfst_id_vars)=0 %then %let hpfst_id_vars = _ALL_;


 %let hpvvn= %sysfunc(getoption(VALIDVARNAME));

 options validvarname=V7;

 proc hp4score data=&hpfst_score_input;

 id &hpfst_id_vars;

 %if %symexist(EM_USER_OUTMDLFILE)=0 %then %do;

   score file="C:\Users\carlo\Desktop\ASDM\Exercise 2\Association
rules\Workspaces\EMWS2\HPDMForest\OUTMDLFILE.bin" out=&hpfst_score_output;

 %end;

 %else %do;

   score file="&EM_USER_OUTMDLFILE" out=&hpfst_score_output;

 %end;
```

```
    PERFORMANCE  DETAILS;
  run;

  options validvarname=&hpvvn;

  data &hpfst_score_output;
    set &hpfst_score_output;
%mend;

%em_hpfst_score;
*------------------------------------------------------------*;
*Computing Classification Vars: Loan_Status;
*------------------------------------------------------------*;
length _format200 $200;
drop _format200;
_format200= ' ' ;
length _p_ 8;
_p_= 0 ;
drop _p_ ;
if P_Loan_StatusY - _p_ > 1e-8 then do ;
   _p_= P_Loan_StatusY ;
   _format200='Y';
end;
if P_Loan_StatusN - _p_ > 1e-8 then do ;
   _p_= P_Loan_StatusN ;
   _format200='N';
end;
I_Loan_Status=dmnorm(_format200,32); ;
length U_Loan_Status $3;
```

```
label U_Loan_Status = 'Unnormalized Into: Loan_Status';

format U_Loan_Status $3.;

if I_Loan_Status='Y' then

U_Loan_Status='Y';

if I_Loan_Status='N' then

U_Loan_Status='N';
```

APPENDIX ASSOCIATION RULES MINING

- R CODE

```
### PACKAGES ###

library(arules)

library(arulesViz)

### OPEN FILE ###

setwd('C:/Users/carlo/Desktop/ASDM/Exercise 2')

retail<-read.transactions('market basket.csv',format = 'basket', sep=',')

### DATA INSPECTION ###

itemFrequencyPlot(retail,topN=15)

### ASSOCIATION RULES ###

rules<-apriori(retail,parameter=list(minlen=1,maxlen=2,supp= 0.1,conf = 0.1))

rules <- sort(rules, by='lift', decreasing = TRUE)

inspect(rules)

### GRAPHS ###

rules1<-apriori(retail,parameter = list(minlen=2, maxlen=3,conf = 0.85))

inspect(rules1)

plot(rules1)

plot(rules1,method = 'grouped')

plot(rules1,method = 'paracoord')

plot(rules1@quality)
```

```
ruleExplorer(rules1)

rules1<-apriori(retail,parameter = list(minlen=2, maxlen=3,conf =
0.50),appearance=list(rhs=c('bagels'),default="lhs"))
```

- SAS CODE

```
*------------------------------------------------------------*;

* Assoc: Score Code;

* To run this score code as stand alone uncomment the code below and set the ASSOCDATA and
EM_SCORE_OUTPUT macro variables:;

*;

* %let EM_SCORE_OUTPUT=;

* %let ASSOCDATA =;

* data &EM_SCORE_OUTPUT;

* set &ASSOCDATA;

* run;

*------------------------------------------------------------*;

*------------------------------------------------------------*;

* &nodeid: Creating RULES data set;

*------------------------------------------------------------*;

data WORK.RULEID;

 length   SET_SIZE                 8

        EXP_CONF                 8

        CONF                     8

        SUPPORT                  8

        LIFT                     8

        COUNT                    8

        RULE                   $ 61

        _LHAND                 $ 28

        _RHAND                 $ 28

        ITEM1                  $ 28
```

```
ITEM2 $ 28
ITEM3 $ 28
index 8
ruleid 8
;

label SET_SIZE="Relations"
EXP_CONF="Expected Confidence(%)"
CONF="Confidence(%)"
SUPPORT="Support(%)"
LIFT="Lift"
COUNT="Transaction Count"
RULE="Rule"
_LHAND="Left Hand of Rule"
_RHAND="Right Hand of Rule"
ITEM1="Rule Item 1"
ITEM2="Rule Item 2"
ITEM3="Rule Item 3"
index="Rule Index"
;
format SET_SIZE 6.
EXP_CONF 6.2
CONF 6.2
SUPPORT 6.2
LIFT 6.2
COUNT 6.2
;
SET_SIZE=2; EXP_CONF=4.95552731893265; CONF=17.6470588235294; SUPPORT=0.76238881829733;
LIFT=3.56108597285067; COUNT=6; RULE="yogurt ==> paper towels"; _LHAND="yogurt"; _RHAND="paper
```

towels"; ITEM1="yogurt"; ITEM2="===========================>"; ITEM3="paper towels"; index=1; ruleid=1;

output;

SET_SIZE=2; EXP_CONF=4.32020330368488; CONF=15.3846153846153; SUPPORT=0.76238881829733; LIFT=3.56108597285067; COUNT=6; RULE="paper towels ==> yogurt"; _LHAND="paper towels"; _RHAND="yogurt"; ITEM1="paper towels"; ITEM2="===========================>"; ITEM3="yogurt"; index=2; ruleid=2;

output;

SET_SIZE=2; EXP_CONF=4.95552731893265; CONF=16.1290322580645; SUPPORT=0.63532401524777; LIFT=3.25475599669148; COUNT=5; RULE="ketchup ==> cheeses"; _LHAND="ketchup"; _RHAND="cheeses"; ITEM1="ketchup"; ITEM2="===========================>"; ITEM3="cheeses"; index=3; ruleid=3;

output;

SET_SIZE=2; EXP_CONF=3.93900889453621; CONF=12.8205128205128; SUPPORT=0.63532401524777; LIFT=3.25475599669148; COUNT=5; RULE="cheeses ==> ketchup"; _LHAND="cheeses"; _RHAND="ketchup"; ITEM1="cheeses"; ITEM2="===========================>"; ITEM3="ketchup"; index=4; ruleid=4;

output;

SET_SIZE=2; EXP_CONF=5.08259212198221; CONF=15.3846153846153; SUPPORT=0.76238881829733; LIFT=3.02692307692307; COUNT=6; RULE="paper towels ==> aluminum foil"; _LHAND="paper towels"; _RHAND="aluminum foil"; ITEM1="paper towels"; ITEM2="===========================>"; ITEM3="aluminum foil"; index=5;

ruleid=5;

output;

SET_SIZE=2; EXP_CONF=4.95552731893265; CONF=15; SUPPORT=0.76238881829733; LIFT=3.02692307692307; COUNT=6; RULE="aluminum foil ==> paper towels"; _LHAND="aluminum foil"; _RHAND="paper towels"; ITEM1="aluminum foil"; ITEM2="===========================>"; ITEM3="paper towels"; index=6; ruleid=6;

output;

SET_SIZE=2; EXP_CONF=4.95552731893265; CONF=14.7058823529411; SUPPORT=0.63532401524777; LIFT=2.96757164404223; COUNT=5; RULE="yogurt ==> lunch meat"; _LHAND="yogurt"; _RHAND="lunch meat"; ITEM1="yogurt"; ITEM2="===========================>"; ITEM3="lunch meat"; index=7; ruleid=7;

output;

SET_SIZE=2; EXP_CONF=4.32020330368488; CONF=12.8205128205128; SUPPORT=0.63532401524777; LIFT=2.96757164404223; COUNT=5; RULE="lunch meat ==> yogurt"; _LHAND="lunch meat"; _RHAND="yogurt"; ITEM1="lunch meat"; ITEM2="===========================>"; ITEM3="yogurt"; index=8; ruleid=8;

output;

SET_SIZE=2; EXP_CONF=4.57433290978399; CONF=13.5135135135135; SUPPORT=0.63532401524777; LIFT=2.9542042042042; COUNT=5; RULE="eggs ==> all- purpose"; _LHAND="eggs"; _RHAND="all- purpose"; ITEM1="eggs"; ITEM2="==========================>"; ITEM3="all- purpose"; index=9; ruleid=9;

output;

SET_SIZE=2; EXP_CONF=4.70139771283354; CONF=13.8888888888888; SUPPORT=0.63532401524777; LIFT=2.9542042042042; COUNT=5; RULE="all- purpose ==> eggs"; _LHAND="all- purpose"; _RHAND="eggs"; ITEM1="all- purpose"; ITEM2="==========================>"; ITEM3="eggs"; index=10; ruleid=10;

output;

SET_SIZE=2; EXP_CONF=5.20965692503176; CONF=14.2857142857142; SUPPORT=0.76238881829733; LIFT=2.74216027874564; COUNT=6; RULE="toilet paper ==> bagels"; _LHAND="toilet paper"; _RHAND="bagels"; ITEM1="toilet paper"; ITEM2="==========================>"; ITEM3="bagels"; index=11; ruleid=11;

output;

SET_SIZE=2; EXP_CONF=5.33672172808132; CONF=14.6341463414634; SUPPORT=0.76238881829733; LIFT=2.74216027874564; COUNT=6; RULE="bagels ==> toilet paper"; _LHAND="bagels"; _RHAND="toilet paper"; ITEM1="bagels"; ITEM2="==========================>"; ITEM3="toilet paper"; index=12; ruleid=12;

output;

SET_SIZE=2; EXP_CONF=4.95552731893265; CONF=13.5135135135135; SUPPORT=0.63532401524777; LIFT=2.72695772695772; COUNT=5; RULE="waffles ==> cheeses"; _LHAND="waffles"; _RHAND="cheeses"; ITEM1="waffles"; ITEM2="==========================>"; ITEM3="cheeses"; index=13; ruleid=13;

output;

SET_SIZE=2; EXP_CONF=4.70139771283354; CONF=12.8205128205128; SUPPORT=0.63532401524777; LIFT=2.72695772695772; COUNT=5; RULE="cheeses ==> waffles"; _LHAND="cheeses"; _RHAND="waffles"; ITEM1="cheeses"; ITEM2="==========================>"; ITEM3="waffles"; index=14; ruleid=14;

output;

SET_SIZE=2; EXP_CONF=4.06607369758576; CONF=10.8695652173913; SUPPORT=0.63532401524777; LIFT=2.67323369565217; COUNT=5; RULE="sandwich loaves ==> pork"; _LHAND="sandwich loaves"; _RHAND="pork"; ITEM1="sandwich loaves"; ITEM2="==========================>"; ITEM3="pork"; index=15; ruleid=15;

output;

SET_SIZE=2; EXP_CONF=5.84498094027954; CONF=15.625; SUPPORT=0.63532401524777; LIFT=2.67323369565217; COUNT=5; RULE="pork ==> sandwich loaves"; _LHAND="pork"; _RHAND="sandwich loaves"; ITEM1="pork"; ITEM2="==========================>"; ITEM3="sandwich loaves"; index=16; ruleid=16;

output;

SET_SIZE=2; EXP_CONF=5.08259212198221; CONF=13.5135135135135; SUPPORT=0.63532401524777; LIFT=2.65878378378378; COUNT=5; RULE="eggs ==> dinner rolls"; _LHAND="eggs"; _RHAND="dinner rolls"; ITEM1="eggs"; ITEM2="==========================>"; ITEM3="dinner rolls"; index=17; ruleid=17;

output;

SET_SIZE=2; EXP_CONF=4.70139771283354; CONF=12.5; SUPPORT=0.63532401524777; LIFT=2.65878378378378; COUNT=5; RULE="dinner rolls ==> eggs"; _LHAND="dinner rolls"; _RHAND="eggs"; ITEM1="dinner rolls"; ITEM2="==========================>"; ITEM3="eggs"; index=18; ruleid=18;

output;

SET_SIZE=2; EXP_CONF=6.86149936467598; CONF=16.6666666666666; SUPPORT=0.63532401524777; LIFT=2.42901234567901; COUNT=5; RULE="mixes ==> tortillas"; _LHAND="mixes"; _RHAND="tortillas"; ITEM1="mixes"; ITEM2="==========================>"; ITEM3="tortillas"; index=19; ruleid=19;

output;

SET_SIZE=2; EXP_CONF=5.33672172808132; CONF=12.8205128205128; SUPPORT=0.63532401524777; LIFT=2.4023199023199; COUNT=5; RULE="paper towels ==> dishwashing liquid/detergent"; _LHAND="paper towels"; _RHAND="dishwashing liquid/detergent"; ITEM1="paper towels"; ITEM2="==========================>";

ITEM3="dishwashing liquid/detergent"; index=20; ruleid=20;

output;

SET_SIZE=2; EXP_CONF=4.95552731893265; CONF=11.9047619047619; SUPPORT=0.63532401524777; LIFT=2.4023199023199; COUNT=5; RULE="dishwashing liquid/detergent ==> paper towels"; _LHAND="dishwashing liquid/detergent"; _RHAND="paper towels"; ITEM1="dishwashing liquid/detergent";

ITEM2="==========================>"; ITEM3="paper towels"; index=21; ruleid=21;

output;

SET_SIZE=2; EXP_CONF=5.84498094027954; CONF=13.8888888888888; SUPPORT=0.63532401524777; LIFT=2.37620772946859; COUNT=5; RULE="shampoo ==> sandwich loaves"; _LHAND="shampoo"; _RHAND="sandwich loaves"; ITEM1="shampoo"; ITEM2="==========================>"; ITEM3="sandwich loaves"; index=22;

ruleid=22;

output;

SET_SIZE=2; EXP_CONF=4.57433290978399; CONF=10.8695652173913; SUPPORT=0.63532401524777; LIFT=2.37620772946859; COUNT=5; RULE="sandwich loaves ==> shampoo"; _LHAND="sandwich loaves"; _RHAND="shampoo"; ITEM1="sandwich loaves"; ITEM2="==========================>"; ITEM3="shampoo"; index=23;

ruleid=23;

output;

SET_SIZE=2; EXP_CONF=5.46378653113087; CONF=12.8205128205128; SUPPORT=0.63532401524777; LIFT=2.34645199761478; COUNT=5; RULE="paper towels ==> juice"; _LHAND="paper towels";

_RHAND="juice"; ITEM1="paper towels"; ITEM2="===========================>"; ITEM3="juice"; index=24; ruleid=24;

output;

SET_SIZE=2; EXP_CONF=4.95552731893265; CONF=11.6279069767441; SUPPORT=0.63532401524777; LIFT=2.34645199761478; COUNT=5; RULE="juice ==> paper towels"; _LHAND="juice"; _RHAND="paper towels"; ITEM1="juice"; ITEM2="===========================>"; ITEM3="paper towels"; index=25; ruleid=25;

output;

SET_SIZE=2; EXP_CONF=4.95552731893265; CONF=11.3207547169811; SUPPORT=1.52477763659466; LIFT=2.28447024673439; COUNT=12; RULE="vegetables ==> pasta"; _LHAND="vegetables"; _RHAND="pasta"; ITEM1="vegetables"; ITEM2="===========================>"; ITEM3="pasta"; index=26; ruleid=26;

output;

SET_SIZE=2; EXP_CONF=13.4688691232528; CONF=30.7692307692307; SUPPORT=1.52477763659466; LIFT=2.28447024673439; COUNT=12; RULE="pasta ==> vegetables"; _LHAND="pasta"; _RHAND="vegetables"; ITEM1="pasta"; ITEM2="===========================>"; ITEM3="vegetables"; index=27; ruleid=27;

output;

SET_SIZE=2; EXP_CONF=5.84498094027954; CONF=12.1951219512195; SUPPORT=0.63532401524777; LIFT=2.08642629904559; COUNT=5; RULE="sandwich bags ==> milk"; _LHAND="sandwich bags"; _RHAND="milk"; ITEM1="sandwich bags"; ITEM2="===========================>"; ITEM3="milk"; index=28; ruleid=28;

output;

SET_SIZE=2; EXP_CONF=5.20965692503176; CONF=10.8695652173913; SUPPORT=0.63532401524777; LIFT=2.08642629904559; COUNT=5; RULE="milk ==> sandwich bags"; _LHAND="milk"; _RHAND="sandwich bags"; ITEM1="milk"; ITEM2="===========================>"; ITEM3="sandwich bags"; index=29; ruleid=29;

output;

SET_SIZE=2; EXP_CONF=13.4688691232528; CONF=24.3243243243243; SUPPORT=1.14358322744599; LIFT=1.80596634370219; COUNT=9; RULE="waffles ==> vegetables"; _LHAND="waffles"; _RHAND="vegetables"; ITEM1="waffles"; ITEM2="===========================>"; ITEM3="vegetables"; index=30; ruleid=30;

output;

SET_SIZE=2; EXP_CONF=13.4688691232528; CONF=22.2222222222222; SUPPORT=1.01651842439644; LIFT=1.64989517819706; COUNT=8; RULE="shampoo ==> vegetables"; _LHAND="shampoo"; _RHAND="vegetables"; ITEM1="shampoo"; ITEM2="===========================>"; ITEM3="vegetables"; index=31; ruleid=31;

output;

SET_SIZE=2; EXP_CONF=13.4688691232528; CONF=20.6896551724137; SUPPORT=0.76238881829733; LIFT=1.53610930383864; COUNT=6; RULE="sugar ==> vegetables"; _LHAND="sugar"; _RHAND="vegetables"; ITEM1="sugar"; ITEM2="==========================>"; ITEM3="vegetables"; index=32; ruleid=32;

output;

SET_SIZE=2; EXP_CONF=13.4688691232528; CONF=19.5121951219512; SUPPORT=1.01651842439644; LIFT=1.44868844914864; COUNT=8; RULE="sandwich bags ==> vegetables"; _LHAND="sandwich bags"; _RHAND="vegetables"; ITEM1="sandwich bags"; ITEM2="==========================>"; ITEM3="vegetables"; index=33;

ruleid=33;

output;

SET_SIZE=2; EXP_CONF=13.4688691232528; CONF=17.9487179487179; SUPPORT=0.88945362134688; LIFT=1.33260764392839; COUNT=7; RULE="cheeses ==> vegetables"; _LHAND="cheeses"; _RHAND="vegetables"; ITEM1="cheeses"; ITEM2="==========================>"; ITEM3="vegetables"; index=34; ruleid=34;

output;

SET_SIZE=2; EXP_CONF=13.4688691232528; CONF=17.6470588235294; SUPPORT=0.76238881829733; LIFT=1.31021087680355; COUNT=6; RULE="yogurt ==> vegetables"; _LHAND="yogurt"; _RHAND="vegetables"; ITEM1="yogurt"; ITEM2="==========================>"; ITEM3="vegetables"; index=35; ruleid=35;

output;

SET_SIZE=2; EXP_CONF=13.4688691232528; CONF=17.5; SUPPORT=0.88945362134688; LIFT=1.29929245283018; COUNT=7; RULE="poultry ==> vegetables"; _LHAND="poultry"; _RHAND="vegetables"; ITEM1="poultry"; ITEM2="==========================>"; ITEM3="vegetables"; index=36; ruleid=36;

output;

SET_SIZE=2; EXP_CONF=13.4688691232528; CONF=17.0731707317073; SUPPORT=0.88945362134688; LIFT=1.26760239300506; COUNT=7; RULE="bagels ==> vegetables"; _LHAND="bagels"; _RHAND="vegetables"; ITEM1="bagels"; ITEM2="==========================>"; ITEM3="vegetables"; index=37; ruleid=37;

output;

SET_SIZE=2; EXP_CONF=13.4688691232528; CONF=16.6666666666666; SUPPORT=0.63532401524777; LIFT=1.23742138364779; COUNT=5; RULE="mixes ==> vegetables"; _LHAND="mixes"; _RHAND="vegetables"; ITEM1="mixes"; ITEM2="==========================>"; ITEM3="vegetables"; index=38; ruleid=38;

output;

SET_SIZE=2; EXP_CONF=13.4688691232528; CONF=15.1515151515151; SUPPORT=0.63532401524777; LIFT=1.1249285305889; COUNT=5; RULE="spaghetti sauce ==> vegetables"; _LHAND="spaghetti sauce"; _RHAND="vegetables"; ITEM1="spaghetti sauce"; ITEM2="==========================>"; ITEM3="vegetables"; index=39;

ruleid=39;

output;

SET_SIZE=2; EXP_CONF=13.4688691232528; CONF=14.7058823529411; SUPPORT=0.63532401524777; LIFT=1.09184239733629; COUNT=5; RULE="ice cream ==> vegetables"; _LHAND="ice cream"; _RHAND="vegetables"; ITEM1="ice cream"; ITEM2="===========================>"; ITEM3="vegetables"; index=40; ruleid=40;

output;

SET_SIZE=2; EXP_CONF=13.4688691232528; CONF=14.2857142857142; SUPPORT=0.76238881829733; LIFT=1.06064690026954; COUNT=6; RULE="toilet paper ==> vegetables"; _LHAND="toilet paper"; _RHAND="vegetables"; ITEM1="toilet paper"; ITEM2="===========================>"; ITEM3="vegetables"; index=41;

ruleid=41;

output;

SET_SIZE=2; EXP_CONF=13.4688691232528; CONF=13.8888888888888; SUPPORT=0.63532401524777; LIFT=1.03118448637316; COUNT=5; RULE="butter ==> vegetables"; _LHAND="butter"; _RHAND="vegetables"; ITEM1="butter"; ITEM2="===========================>"; ITEM3="vegetables"; index=42; ruleid=42;

output;

SET_SIZE=2; EXP_CONF=13.4688691232528; CONF=13.5135135135135; SUPPORT=0.63532401524777; LIFT=1.0033146353901; COUNT=5; RULE="eggs ==> vegetables"; _LHAND="eggs"; _RHAND="vegetables"; ITEM1="eggs"; ITEM2="===========================>"; ITEM3="vegetables"; index=43; ruleid=43;

output;

SET_SIZE=2; EXP_CONF=13.4688691232528; CONF=12.5; SUPPORT=0.63532401524777; LIFT=0.92806603773584; COUNT=5; RULE="dinner rolls ==> vegetables"; _LHAND="dinner rolls"; _RHAND="vegetables"; ITEM1="dinner rolls"; ITEM2="===========================>"; ITEM3="vegetables"; index=44; ruleid=44;

output;

SET_SIZE=2; EXP_CONF=13.4688691232528; CONF=11.9047619047619; SUPPORT=0.63532401524777; LIFT=0.88387241689128; COUNT=5; RULE="dishwashing liquid/detergent ==> vegetables"; _LHAND="dishwashing liquid/detergent"; _RHAND="vegetables"; ITEM1="dishwashing liquid/detergent";

ITEM2="===========================>"; ITEM3="vegetables"; index=45; ruleid=45;

output;

SET_SIZE=2; EXP_CONF=13.4688691232528; CONF=11.6279069767441; SUPPORT=0.63532401524777; LIFT=0.86331724440544; COUNT=5; RULE="flour ==> vegetables"; _LHAND="flour"; _RHAND="vegetables"; ITEM1="flour"; ITEM2="===========================>"; ITEM3="vegetables"; index=46; ruleid=46;

output;

SET_SIZE=2; EXP_CONF=13.4688691232528; CONF=11.1111111111111; SUPPORT=0.76238881829733; LIFT=0.82494758909853; COUNT=6; RULE="tortillas ==> vegetables"; _LHAND="tortillas";

```
_RHAND="vegetables"; ITEM1="tortillas"; ITEM2="=========================>"; ITEM3="vegetables";
index=47; ruleid=47;

output;

;

run;

*------------------------------------------------------------*;

* Assoc: Creating RULEMAP and Output data sets;

*------------------------------------------------------------*;

%let _scoreDs = &EM_SCORE_OUTPUT;

proc sort data=&_scoreDs;

by Date;

run;

proc mbscore data=&_scoreDs out=score_ruleid INCLUDE ALL_ID

;

customer Date;

target Item;

;

rules data=work.ruleid;

run;

data &_scoreDs;

set score_ruleid;

array _r{47} _r1-_r47 (47*0);

by Date;

if first.Date then do;

do i=1 to 47;

_r[i]=0;

end;

end;

if ruleid ne . then _r[ruleid]=1;
```

```
if last.Date then output;

drop i ruleid;

run;

%let _lib=%str();

%let _ds=%str();

%macro _dsname;

%let _lib =%scan(&EM_SCORE_OUTPUT, 1, .);

%let _ds =%scan(&EM_SCORE_OUTPUT, 2, .);

%if "&_ds" = "" %then %do;

%let _lib=WORK;

%let _ds=%scan(&EM_SCORE_OUTPUT, 1, .);

%end;

%mend _dsname;

%_dsname;

data _null_;

set ruleid end = eof;

if _N_=1 then do;

call execute("proc datasets lib=&_lib nolist;");

call execute("   modify &_ds;");

end;

call execute("   rename _r"!!strip(put(_N_, best.))!!"= RULE"!!strip(put(INDEX, best.))!!";");

call execute("   label  RULE"!!strip(put(INDEX, best.))!!'='!!quote(RULE)!!";");

if eof then do;

call execute("run;");

call execute("quit;");

end;

run;

proc datasets lib=work nolist;

delete score_ruleid ruleid;
```

```
run;

quit;
```

APPENDIX TEXT MINING

- R CODE

```
setwd('C:/Users/carlo/Desktop/ASDM/Exercice 3')

library(tm)

library(wordcloud)

library(cluster)

library(factoextra)

dataset<-readLines("Hotels review_cleaned2.csv")

mycorpus <-Corpus(VectorSource(dataset))

mycorpus <-tm_map(mycorpus,tolower)# Lower case

mycorpus <-tm_map(mycorpus,removePunctuation)# Puntuation

mycorpus <-tm_map(mycorpus,removeNumbers)# Numbers

dataclean <-tm_map(mycorpus,stripWhitespace)#White space

dataclean <-
tm_map(dataclean,removeWords,c('hotel','biyadhoo','kihavah','anantara','cocoon','dhigu','fushi','cinnamon','fi
litheyo','thani',

        'dhonveli','filitheyo','dhonveli','thani','villa','bungalow','embudu','gangehi',

        'gangehi','angaga','amari','angaga','resorts','villas','maldives','island','resort','adaaran'))#Stop words

dataclean1 <-tm_map(dataclean,removeWords,stopwords("english"))

inspect(mycorpus[3])

dtm <-TermDocumentMatrix(dataclean1,control = list(minWordLength=c(1,Inf)))# Document Matrix

findFreqTerms(dtm,lowfreq = 2)

termFrequency<-rowSums(as.matrix(dtm))

termFrequency

termFrequency <-subset(termFrequency,termFrequency>=1000)

termFrequency

barplot(termFrequency,las=2,col=rainbow(20))

wordfreq<-sort(termFrequency,decreasing = TRUE)

wordcloud(words = names(wordfreq),freq=wordfreq,max.words=100,min.freq = 5,random.order = F,colors =
rainbow(20))
```

```
barplot(wordfreq[1:50],xlab = "term",ylab = "frequency",las=2,col=heat.colors(50))
```

- SAS CODE

```
/* First we create a Weighted TMOUT Data Set based on weighted terms*/

proc tmutil data=work.TextFilter_out key=termloc.TextFilter_filtterms;

control init release;

weight cellwgt=LOG in_weight=termloc.TextFilter_filtterms (keep=key weight);

output out=work._weighted_tmout;


%row_pivot_normalize(transds=work._weighted_tmout, outtransds=WORK.TMOUTNORM,

    col_sumds=work._termsumds,row=_document_,col=_termnum_,entry=_count_,

    pivot=0.7,tmt_config=termloc.TextFilter_tmconfig,tmt_train=0,prefix=TextTopic);


/*initialize topics and termtopics datasets in case they do not exist (0 topics case)*/

%macro tmt_check_topics_exist;

%if(^%sysfunc(exist(termloc.TextTopic_topics))) %then %do;

  proc sql noprint; create table termloc.TextTopic_topics

  (_topicid decimal, _docCutoff decimal, _termCutoff decimal, _name char(1024), _cat char(4), /* _apply
char(1), */ _numterms decimal, _numdocs decimal, _displayCat char(200) );

  quit;

%end;

%if(^%sysfunc(exist(termloc.TextTopic_termtopics))) %then %do;

  proc sql noprint; create table termloc.TextTopic_termtopics

  (_topicid decimal, _weight decimal, _termid decimal);

  quit;

%end;

%mend tmt_check_topics_exist;

%tmt_check_topics_exist;

data work.TextTopic_termtopics; set termloc.TextTopic_termtopics; run;

data work.TextTopic_topics; set termloc.TextTopic_topics; run;
```

```
%tmt_doc_score(termtopds=work.TextTopic_termtopics, docds=&em_score_output,
outds=WORK.TMOUTNORM, topicds=work.TextTopic_topics, newdocds=work._newdocds, scoring=yes,
termsumds=work._termsumds, prefix=TextTopic_,pivot=0.7);
data &em_score_output; set work._newdocds;
```