Assignment A1-LP2: Classification			
Student Name	**Jose Arturo Gil Alonso**	**Student No**	*218 659 676*
My other group members		**A1 Group No**	*Group 1*
Team Names	**Omar Malaeb**	**Student Nos**	*218453505*
	Jocelyn Lin		*218189891*

Executive summary (one page)

Aim

To clearly articulate your understanding of the business problem and to present its solution to management.

The aim of this report is to analyse the opportunities of implementing two features, such as Booking a table and online meal ordering, in the restaurants of Bangalore, India. Nowadays due to the fact that the companies and businesses are taking advantage of the technology, those are improving their processes, getting more profits and getting more loyalty from the customers. Thus, to take advantage of the benefits of technology, it is necessary to analyse information related to the restaurants located in Bangalore and the customers opinions about their experiences in those businesses.

Expectation

For instance, a sample of 48.000 restaurants was gathered and provided by BFA, Bangalore Food Assist. In the dataset the following attributes have been provided:

- Restaurant name, time and phone number.
- Address, location and neighbourhood.
- Cuisine and meal types

Making a decision to successfully implement new processes in the organizations, require the application of a data mining method, building a model to forecast the projects revenue and visualise the information gathered to get enough insight to make an accurate decision.

Thus, there are some questions that want to be solved:

1. What are the neighbourhoods in which are located the restaurants not only with the cheapest food, but it also the best rated?
2. Based on the feedback from Customers, what could be the action plan for the restaurants?

Analysing this dataset requires the use of an Advanced Software known as RapidMiner. Some of the features that this Software offers, is the capability of analysing huge amounts of data from large datasets and visualise the information accurately.

Business Problem:

Increasing the customer satisfaction. Through applying these two alternatives, the customers would be more satisfied with the service that is provided by the Bangalore restaurants as it has plenty of benefits such as:

- The dishes liked by the customers
- Avoid wasting time in a queue or line.
- The customers will be allowed to reserve the perfect spot of the restaurant for them.
- When customers are booking a table, they can choose the day that they want to go to the restaurant and avoid busy places in which the food might not be well prepared.
- The customers will have time to decide the kind of food that they want to eat from the Menu and ask for a dietary requirement.
- The Customers will me more satisfied if they can eat in a comfortable place like their home from the restaurants that they like.
- The customers will now the estimate time that the food will arrive to the door of their house

Data exploration and preparation in RapidMiner (one page)

Expectation

The attributes used as labels are booking a table and online ordering meal, due to the fact that for this Business Problem was required to identify the performance of the restaurants that have implemented these two facilities, in order to know what are the features of the already established restaurants that made them to have the best rates and feedbacks, to be implemented in the already existing restaurants and new restaurants in Bangalore.

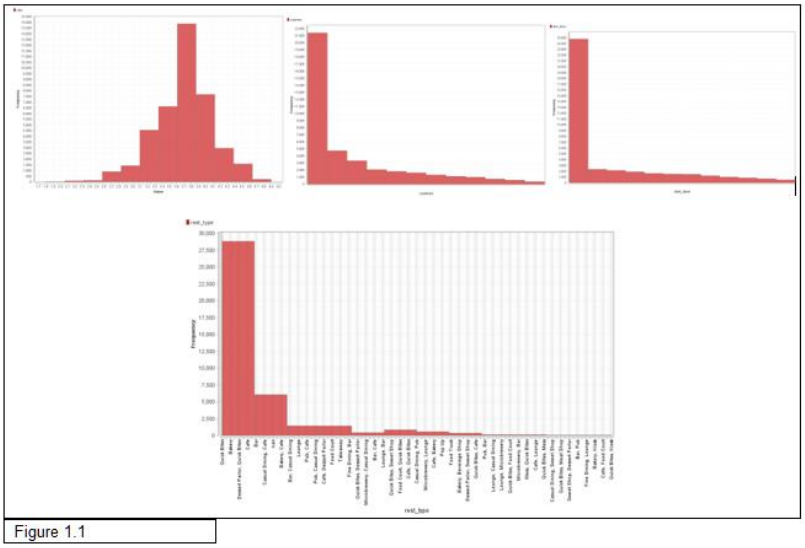


Figure 1.1

Moreover, as it can be identified after using the operator to measure the weight, called weight by information gain, the attributes with the highest weight that logically match with the labels are cuisine, dish liked, restaurant type and rate.

It is the distribution using histograms of the selected attributes:

Features of the selected attributes:

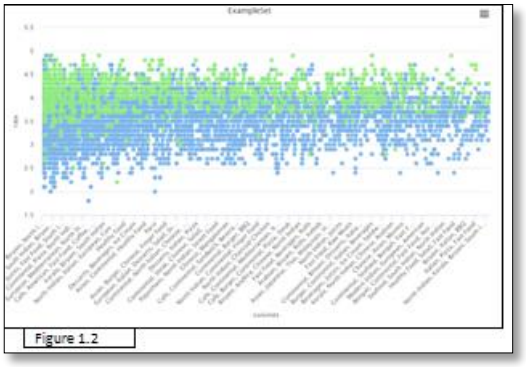


Figure 1.2

This scatter Plot shows that Cuisines, X Axis, is easy to correlate with Rate, Y Axis, depending on the booking table facility, as the last one is binomial, rate is continuous data and cuisines is polynomial data.

The table using Extract Statistics Operator, shows that the most suitable variables of those the numerical attributes are Minimum, Maximum, Average, Standard Deviation, and for categorical is the mode. Moreover, the type of data is also highly important.

The following table shows the Missing Values:

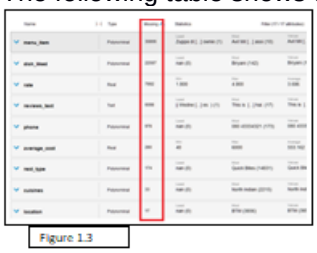


Figure 1.3

The operator Eliminate Attribute would be useful, however, there is still important information in this dataset as there are some attributes such as dish liked and rate that got a high weight and are highly important for the pre-processing and classification process. Therefore, replace missing values by average was used.

Extension:

To identify missing values and replace the values that were initially stored as nan, the Operator Declare Missing Values was used, in which the nan values were replaced for null values or with ?, to help in the data cleaning transforming the data more accurately.

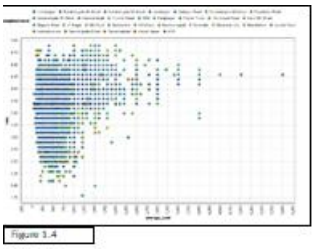


Figure 1.4

The Following scatter Plot represents that most attractive Neighbourhoods taking into account the average cost and the rate:

Discovering Relationships and Data Transformation in RapidMiner (one page)

Expectation

Figure 1.1: Data Exploration and Discovering Relationship Process

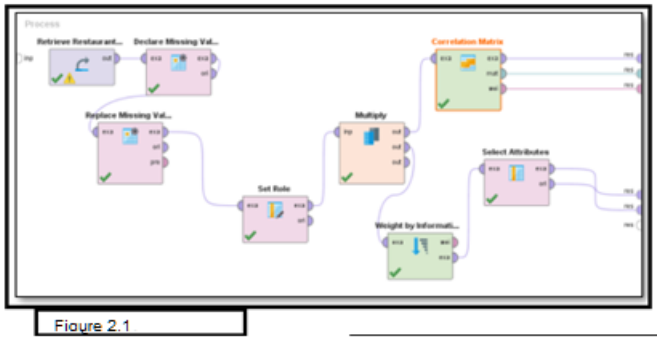


Figure 2.1

Attribut...	rate	votes	average...	id
rate	1	0.425	0.365	0.006
votes	0.425	1	0.380	-0.009
average...	0.365	0.380	1	-0.003
id	0.006	-0.009	-0.003	1

Figure 2.2

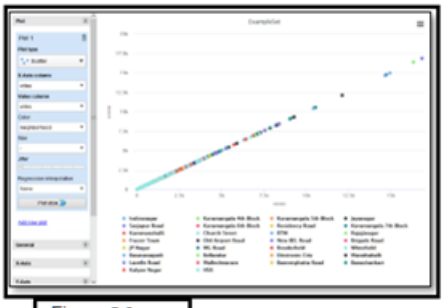


Figure 2.3

The Scatter Plot visualised shows a Linear Relationship between the Count of Votes and the Rate of restaurants per Neighborhood. In spite of the fact that Correlation Matrix shows the correlation coefficient, it can not be used as it is only useful for numerical attributes and either the labels or the meaninful attributes for the business problem contains categorical data.Thus, Weight byInformation Gain is the most suitable to determine the top attributes of the dataset.

The folowing are the attributes that have the highest weight with the labels of Online Ordering Meal and Booking Table:

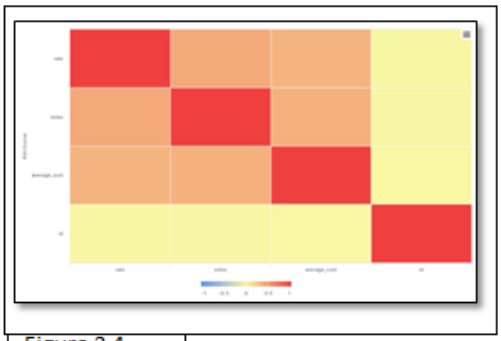


Figure 2.4

attribute	wei...
address	0.910
phone	0.894
name	0.840
reviews_...	0.840
dish_liked	0.405
cuisines	0.334

Extension

For instance, it can be determined that for numerical attributes, the attributes selected are mainly rate, votes and average cost, as those are highly important to identify the popularty of the restaurants. Moreover, using the weight by information gain Operator, it can be identify that the best attributes to be selected are address, phone, and name. The previous was identified taking into account that the labels used are the facilities, online ordering meal and booking a table.

Create a Model(s) in RapidMiner (one page limit)

Expectation

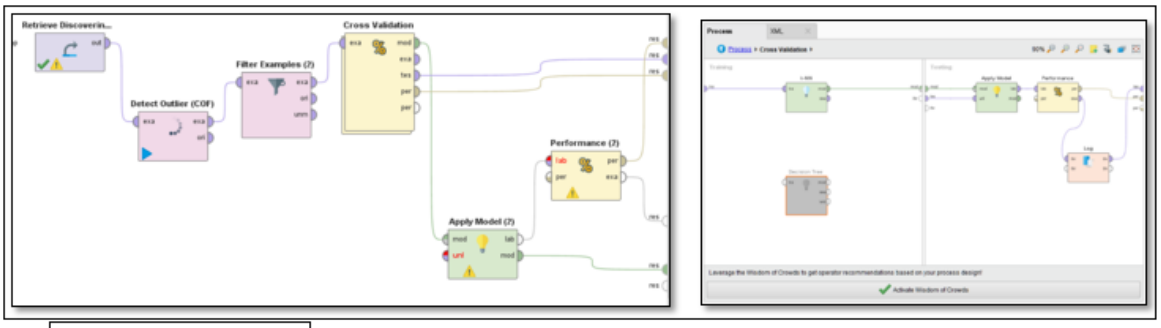
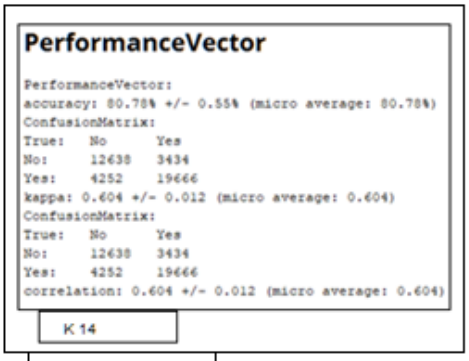


Figure 3.1



Fgure 3.2

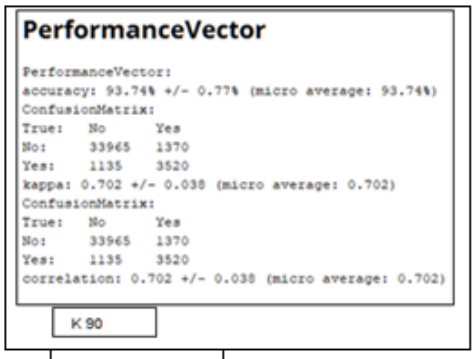


Figure 3.3

As it can be seen, Knn model was used for the data classification process. Thus, after cleaning the dataset as it was showed previously and filtering to avoid the outliers in the data, this regression model was implemented, and it can be seen in the Cross Validation.

After Retrieving the data, normalization was used, and filtering to avoid outliers in the dataset. Using Knn Model, which can be modified on the top right side in Parameters tab, that indicates the number of neighbours, the data shows the best performance as it is indicated in the Level of Accuracy, kappa and correlation parameters. Although, the model was tested 6 times for the label of booking a table, using k 20, k 5, k 70, k 30, k 90 and k 120, and for the label of online ordering meal, the model was tested 5 times using k 20, k 10, k 15, k 12 and k 14. Thus, the best k Parameters were chosen due to the fact that those showed a kappa over 0.61 and also a higher level of accuracy than 60%.

The operators apply model and Performance were also used, with the Parameters Level of Accuracy, Kappa and Correlation.

Extension

The decision tree model and logistic Regression were also applied. Even though, the decision tree has a good accuracy level, the kappa is lower 0.61 that is too low for the performance of the model. In spite of the fact that the logistic Regression got the best accuracy level and the best kappa, however, those values are too high to be implemented as the suitable classification model.

Evaluate and Improve the Model(s) in RapidMiner (one page)

Aim

To report and explain the performance of developed classification models.

Expectation

Normalize Operator was used with Z-Transformation to scale values and to help that they fit in the range. Moreover, Detect Outliers (COF) has been used to identify data that is highly spread from the mean of data points, using the parameters Outliers are equal to false, to bring only the outliers with infinite value.

Even though, the model was not tested changing the folds of the parameters or the random seeds as logically the model performance will change, and it is taking as cheating or not improving the performance fairly

Therefore, in the Cross Validation, a decision tree was also implemented and tested in order to improve the model:

Decision Tree

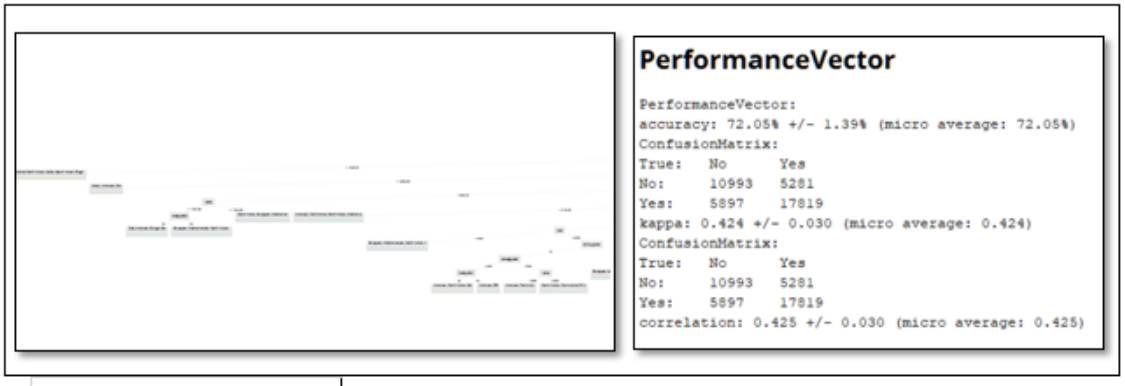


Figure 4.1

The Decision tree was used, in which the Accuracy level and the Kappa performed were good related to the other models developed

Logistic Regression

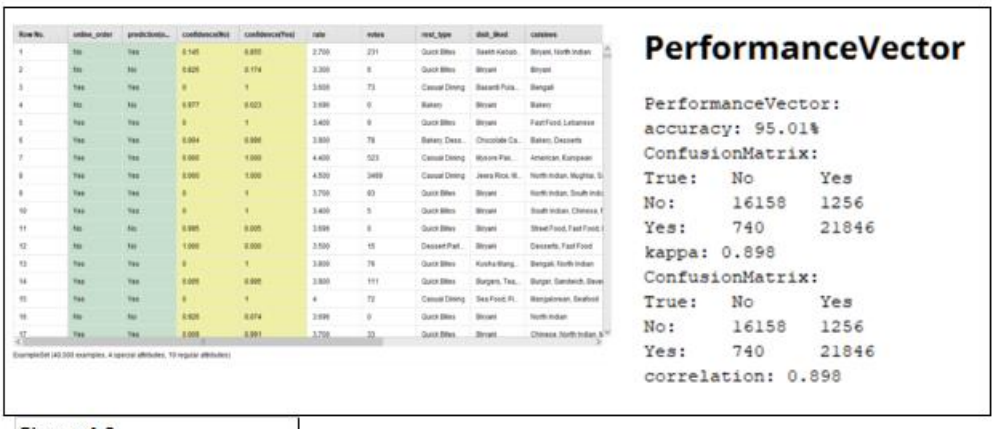


Figure 4.2

Logistic Regression has been also used to test the data, in which the values showed the highest accuracy level of the models with 95.01% and the highest kappa with 0.898.

Knn

- Label Booking a Table k90:

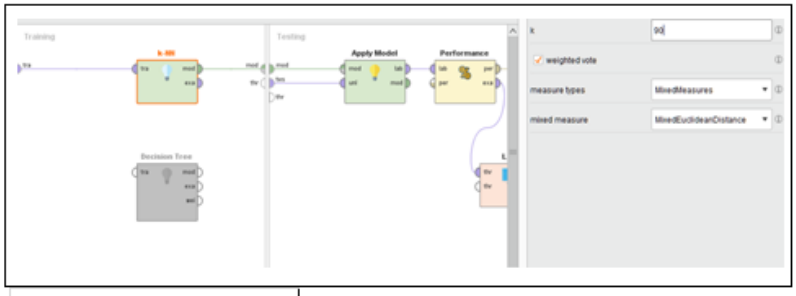


Figure 4.3

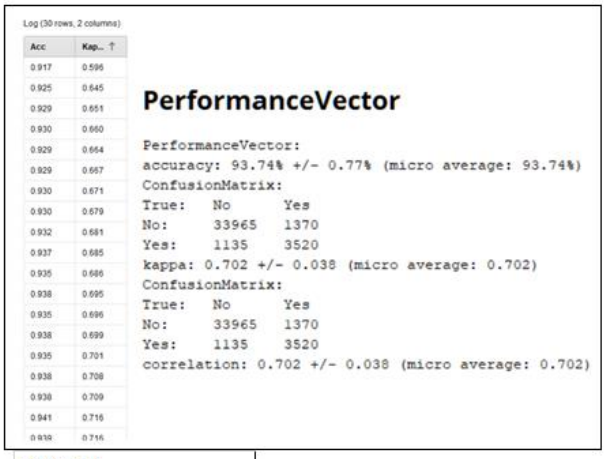


Figure 4.4

- Label Online Order K 14:

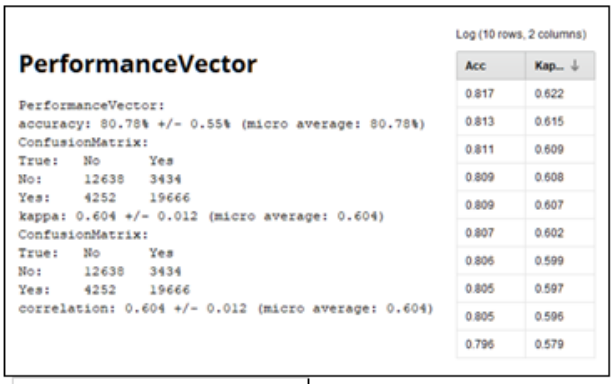


Figure 4.5

As it was previously mentioned, the data was tested with knn more than 7 times, in which, the best performance for the label Booking a table was with k90 and for the label Online Order was with the label K14.

Deployment in RapidMiner (one page)

Expectation

There were 5 processes involved in reach the point of performing the model with classification:

1. Convert the CSV File into a store:
2. Explore the Data and finding relationships using Graphs such as Scatter Plots and Tables
3. Pre Processing, in which the data is cleaned using Declare Missing Values, Replace Missing Values, Setting the Roles, in which the process is divided in the two labels, one for online ordering meal and another one for Booking a Table, this is to be able to weight the attributes depending on each label to find out the best predictors that are then selected using Select Attributes (Add Graph)
4. Classification Step, two processes applied depending on the label, in which detect outliers was used, followed by filter examples, in which the outliers equal to false were deselected from the dataset.
5. In the Cross Validation, the number of folds must remain the same, however the number of k in knn model can change and was tested several times. Thus, in Cross Validation it is find Knn Model Operator, Apply Model Operator and Performance Operator. In Performance Operator, accuracy level, kappa and correlation were used to describe the performance. Moreover, in the Cross Validation, decision tree was also implemented, with shows a lower performance than the best performance of each label. (Decision Tree does not need Set Role)

To open and load the File, follow the next process:

a) Download the Zip File, an unzip it in your computer Disk, it could be Disk (C:).
b) Create a Repository Folder in RapidMiner, in the Repository Tab located by Default on the top left side of RM, in the case that there was not a Repository already Created.
c) Type in the Operator Tab “Read CSV”, drag into Process. In Parameters Tab, click in “Import Configuration Wizard”
d) Select the File Location, click in Next, change the File Encoding to UTF-8 to be able to read the file properly, click Next, and explore the attribute types, such as Polynomial, Real, Integer, e.t.c. Click Finish. Link with the line the “Read CSV” operator to “res” port. Run and explore data in the “Results” tab.
e) In the Repository Tab, select the Repository that was created and click on the option “Store Process Here”, to save this process.

As it was previously mentioned, at the end there were 4 classification models developed to analyse the performance:

1. Knn Model for Established Restaurants with the label Online Ordering Meal, which is called Classification E Res Online Order.
2. Knn Model for Established Restaurants with the label Booking Table, which is called Classification E Res Booking Table.
3. Decision Tree, which is called Decision Tree File
4. Logistic Regression Model, which is called Logistic Regression
5. Follow the instructions previously mentioned to load this 4 Files located in the Zip File.

For answering question B, related to the strategy for table booking and online order for New Restaurants, as it can be seen, after analysing the information about New Restaurants that from the dataset are the restaurants without any rate, vote or feedback, there is not result generated. Thus, the restaurants with the best average cost and the highest rate, were selected to solve this question, as those are the example for the new restaurants.

Extension

For instance, the best strategy for the established restaurants would be consider the addition of different kind of cuisines in the Menu, such as Continental, North Indian, Italian, South Indian and Finger Food, which was the highest dish rated, with 4.9 and with the highest number of votes. Thus, in the analysis, other attributes were tested such as the impact of the Neighbourhood, the average cost, the name of the restaurants, e.t.c. It is highly recommended to consider this kind of cuisine implemented in the restaurant. To sum this up, the strategy for the new restaurants depends on the strategy for the established restaurants. Therefore, it is highly recommended that the New Restaurants have implement a similar cuisine, in order to be also successful in the implementation of the facilities of online ordering meal and booking a table.

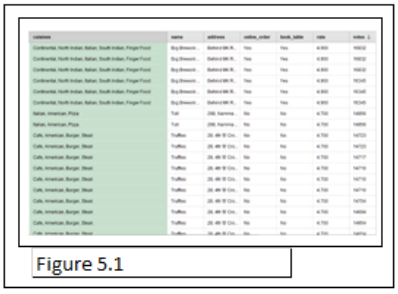
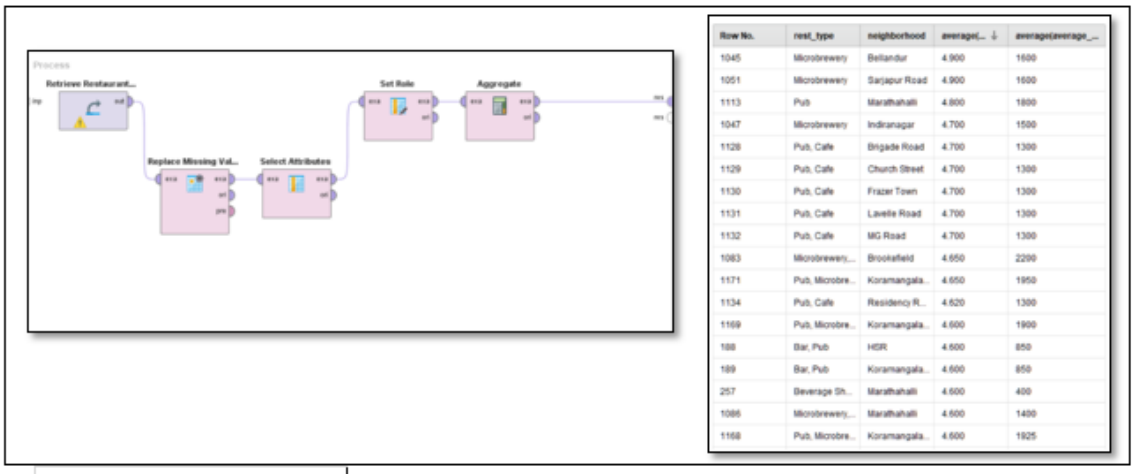


Figure 5.1

Further Research and Extensions in RM (one page)

Expectation

As it was mentioned, in order to improve the model, there were some other Operators used, as Aggregate Operator:



Row No.	rest_type	neighborhood	average(...	average(average_...
1045	Microbrewery	Bellandur	4.900	1600
1051	Microbrewery	Sarjapur Road	4.900	1600
1113	Pub	Marathahalli	4.800	1800
1047	Microbrewery	Indiranagar	4.700	1500
1128	Pub, Cafe	Brigade Road	4.700	1300
1129	Pub, Cafe	Church Street	4.700	1300
1130	Pub, Cafe	Frazer Town	4.700	1300
1131	Pub, Cafe	Lavelle Road	4.700	1300
1132	Pub, Cafe	MG Road	4.700	1300
1083	Microbrewery,...	Brookefield	4.650	2200
1171	Pub, Microbre...	Koramangala...	4.650	1950
1134	Pub, Cafe	Residency R...	4.620	1300
1169	Pub, Microbre...	Koramangala...	4.600	1900
188	Bar, Pub	HSR	4.600	850
189	Bar, Pub	Koramangala...	4.600	850
257	Beverage Sh...	Marathahalli	4.600	400
1086	Microbrewery,...	Marathahalli	4.600	1400
1168	Pub, Microbre...	Koramangala...	4.600	1925

Figure 6.1

Aggregate function was also used to test the data, in which the values are showed using a table ordered by average rate per restaurant type in the neighbours, also showing the average price. As it is known, normalization is also used to improve a model performance.

As it is known, there can be also a use of R and Python in Rapidminer, however, I am not very familiar with the implementation of language programming in Rapid miner due to the fact that this required plenty of training and mentoring to be develop that has not been acquired yet. Nevertheless, I am highly keen on learning in Predictive Analytics about the use of R and Python in Rapidminer.