|  |  |
| --- | --- |
| **Assignment A2: Text Mining + DT + Neural Nets + Optimisation** |  |
| **Student Name*(as per record)*** |  | **Student No** | *Student number* |
| **My other group members** | **A2Group No** | *As per CloudDeakin group number* |
| **Student Name*(as per record)*** |  | **Student Nos** | *Student number* |
|  |  |  | *Student number*  |
|  |  |  | *Student number*  |
|  |  |  |  |
|  | **Exceptional** | **Meets expectations** | **Issues noted** | **Improve** | **Unacceptable** |
| **Exec Report** |  |  |  |  |  |
| **Create****Models** |  |  |  |  |  |
| **Evaluate &****Improve** |  |  |  |  |  |
| **Provide****Solution** |  |  |  |  |  |
| **Research &****Extend** |  |  |  |  |  |
| **BriefComments** |  | **Total** |
|  |

## Include: Report and RMP files, with clear comments supplied to (easily) reproduce reported results.

|  |
| --- |
| **Executive summary (one page)** |

**Expectation**

The Australia Wine Importers gave us contract to come up with a better method, which the company can deploy to estimate the rating points of the imported wines based on the structure attributes and text. The firm provided data of one hundred and thirty thousands (130,000) of wines which were tested by different people to analysed and provide the best strategic method, which can be used in rating wines. The details provided include the wine tested, the country, province and region of where the wine originated description of the wine and their designation. The prices of each wine were also provided make the analysis easier and intensive. However, the name of the testers was not included in the data provided by the company. The Australia Wine Importers would like to use the information in the future to determine wine quality, based on the social media views. The purpose of the study is to determine the group of wines, which the new wines most similar to and also to get the reasons why the wines are similar. It is intents to establish the rating of the new wine in the Australia market.

**Business Problem**

The problem is to eliminate errors which occur when wines are being evaluated based on the taste.

**Solution to Business Problem**

The gradients boosted tree is one of the best methods which can be used to solve the problem. It is provides clear illustrations, which is very understandable. It is illustrated as indicated in the graph below.

**Succinctly describe the solution and justify it.**
Provide references to the supporting evidence, e.g. charts and plots from the following sections.

**Extension**

**The decision made by the company should be drawn as follows:**

* Identify the attributes which can be a predictor.
* Find out the relationship, which exist between the predictors.
* Find out if the text attributes are good enough in the prediction.
* It s also important to find out whether the polynomial and binomial attributes are good indicators of performance, which is achievable.

|  |
| --- |
|  **Create a Model(s) in RapidMiner (two pages / page 1)** |

**Expectation**

The classification of the model is done based on the gradient boosted tree techniques. The following are the steps, which are required to be followed in order to establish a better model:

* The initial and the most important steps are to read the data provided in ccv file.
* The wine taste is made as one of the label because the model is build to establish the relationship between the old and the new wines in the market.
* The value of the price and points each wine gets from the testing. It would be important to include the points because the model is built wine taste against the point each wine obtains. It is important to point that the purpose of the study to get the similarities of the old and new wines and this can be obtained by established how the wines taste both new and old.
* The text and designation of the wines would also be analysed since the classification of the wines is done based on the text and designation.
* The missing value operator was used to fill in the missing values.
* The data was also divided using stratified sampling techniques to ensure that 65% of the data is taken to the gradient booster and the rest to the model. In this case, it means that the model was built using 35% of the data.

|  |
| --- |
| **Create a Model(s) in RapidMiner (two pages / page 2)** |

**Extension**

To obtain the desire solution two models the gradients booster tree and K-NN Method. The K-NN method is illustrated as listed below:

* The initial step is the reading of the data saved in CSV files
* The wine taste is made as one of the label because the model is build to establish the relationship between the old and the new wines in the market.
* The value of the price and points each wine gets from the testing. It would be important to include the points because the model is built wine taste against the point each wine obtains. It is important to point that the purpose of the study to get the similarities of the old and new wines and this can be obtained by established how the wines taste both new and old.

|  |
| --- |
| **Evaluate and Improve the Model(s) in RapidMiner (two pages / page 1)** |

**Expectation**

In order to validate the model performance, the operator was applied to obtain most accurate information





As indicated in the above tables, it is evident that the accuracy to predict the value of label of the attributes was good. The Koppa value indicates 0.285, which means that there are high chances of error and therefore, the value is not accurate compared to accuracy, which shows 94.51%. It means that only a percentage of 1% of the data account for error when the evaluation of the wine taste is being data. It means that estimated 99.0% of data are utilized in the right way to determine the performance of the wine based on their tastes.

|  |
| --- |
| **valuate and Improve the Model(s) in RapidMiner (two pages / page 2)** |

**Extension**

When applying cross validation techniques, the best model to use for the analysis is the gradient booster tree model, which has proved to be the best of the entire model used for the attributes representation. It gives the Koppa value of 0.279, which indicates to be the highest among the other models which can be applied. And therefore, the application of gradient booster would be recommended for the analysis or the representation of the attributes. Though the model is not trustable, it is still much better than other models and therefore, it should be used. The representation of the data is therefore, listed below:





|  |
| --- |
| **Provide an Integrated Solution in RapidMiner (one page)** |

**Expectation**

Below are the list of steps, which should be utilized to develop the model and what we did to develop the model:

* It is important to first read the data provided in cvs files in order to develop the model.
* The data is then prepared and transform and in the process the missing values and nominal numeric are replaced.
* The changes made on the data are then applied back to the optimal parameter and then connected to the model through the parameter of the operator.
* In the optimal parameter the mode is then defined. At this point the cross validation is utilized with the assistant of gradient booster tree (Goslin & Hofmann, 2015).
* The model can then be loaded by with the Read Model and then swiftly connected to the information. However, the loading of the operator can be done using the Read Parameter and then connected to the default setting of the utilization and then linked to the administrator parameter.

|  |
| --- |
| **Further Research and Extensionsin RM (one page)** |

**Expectation**

The auto model features was utilized in building

**Expectation**

# Bibliography

Goslin, K., & Hofmann, M. (2015). Integrated Tutorial Tool for RapidMiner 5. *https://www.researchgate.net/publication/324606272\_Integrated\_Tutorial\_Tool\_for\_RapidMiner\_5* , 2-34.