**Naive Bayes Classifier And Data Cleaning**

**Assignment**

**Answer submission**

***Student ID – 10899486***

***Name - S.C.S.Sandanayake***

***Degree – Data science(Plymouth)***

Question

Employ the provided dataset to forecast the "final grade" column utilizing the Naive Bayes algorithm. Execute the task through two distinct methodologies: initially, without engaging in any preprocessing, and subsequently, after applying data preprocessing techniques. Develop a concise report encompassing code snippets with succinct explanations. Ultimately, analyze and contrast the variances in model accuracies between the instances with and without preprocessing.

* Answer 1 – R script for get accuracy of the model without data preprocessing and cleaning

The first R script is for get accuracy level without data preprocessing and data cleaning. I used 2 libraries in the 1st R script, library e1071 used for machine learning and statistical modeling tasks and caret library used for evaluating predictive models (Naive Bayes classification model training) etc...

1) First loaded the required libraries, then view and read the data set. After that I specify the target column and features. . I assume the target column as “final grade”. Then split data set into training set and testing sets .

This data set has 649 observations and 34 variables. Test data set includes 161 observations and 34 variables, train data set have 488 observations and 34 variables. The seed value is set to 600 because the table contain nearly 600 observations and the split index p value set to 7.5 (3/4) .

2) Next build the Naive bayes model using the train data set. After that made predictions on the test data set (y\_pred).

3) Then generate the summery and predicted results table on test data set.

4) Next step is handling categorical values before doing the confusion matrix calculations.

5) Then get the confusion matrix calculations with statistics.

6) Finally I evaluate the accuracy level for un-preprocessed instance.

The accuracy level is 0.2795031 ~ 0.2795 .

* Answer 2 – R script for get accuracy of the model with data preprocessing and cleaning

The second R script is for take the accuracy level with data preprocessing and data cleaning. I used 3 libraries in the 2st R script. Library e1071 , caret library , and mice library which is used for multiple imputations.

1) First load the required libraries and Then read the data set and check for the missing values, then view and read the data set. After that I specify the target column and features. . I assume the target column as “final grade”.

2) After that impute the missing values by using random forest method . The mice library was used for that purpose.

3) Then I split data set into training set and testing sets . The seed value is set to 600 that nearly equal to total table observations and the split index p value set to 7.5 (3/4) .

4) Next encode the categorical values into numerical values . caret package is used to create dummy variables for categorical predictors. It transforms categorical variables into a binary (0 or 1) format.

5) And after that I checked the categorical variables.

6) Now display head of process data. It displays the first few rows of the processed training data, which now includes the dummy variables for categorical predictors.

7) The next step is preprocessing the data .

8) Now we can initialize and train the naive bayes model.

9) In this step make predictions form the test data set.

10) Next step is handling categorical values before doing the confusion matrix calculations.

11) After that get the confusion matrix and statistics values.

12) Finally evaluate the accuracy level of the model.

The new accuracy level after the preprocessing data is 0.2091503 ~ 0.2092

* Summery and conclusion

Conclusion:

The accuracy level before preprocessing the data is = 0.2795031 ~ 0.2795 .

The new accuracy level after the preprocessing is = 0.2091503 ~ 0.2092.

According to calculations after preprocessing the data the accuracy has been dropped by, 0.0703528.

Summary:

The first R script was coded to train a Naive Bayes classifier model without using any preprocessing steps. It included the fundamental data segmentation, model training, and accuracy calculation steps. The model's accuracy on the test set was determined, offering an indication of performance without the need for data preparation. This accuracy score shows the Naive Bayes model's baseline performance on the given dataset.

The second R script extended the analysis by using data preprocessing steps. It has handling missing values using imputation, encoding categorical variables and scaling numeric predictors. After that the naïve bayes model was trained on preprocessed data. And the new accuracy calculated.