SHIVAN PRAMLALL

EDUV4945866@VOSSIE.NET

ITBDA4-14 - Big Data Analytics Techniques

PROJECT 1

2024

Table of Contents

[Question 1 2](#_Toc170668396)

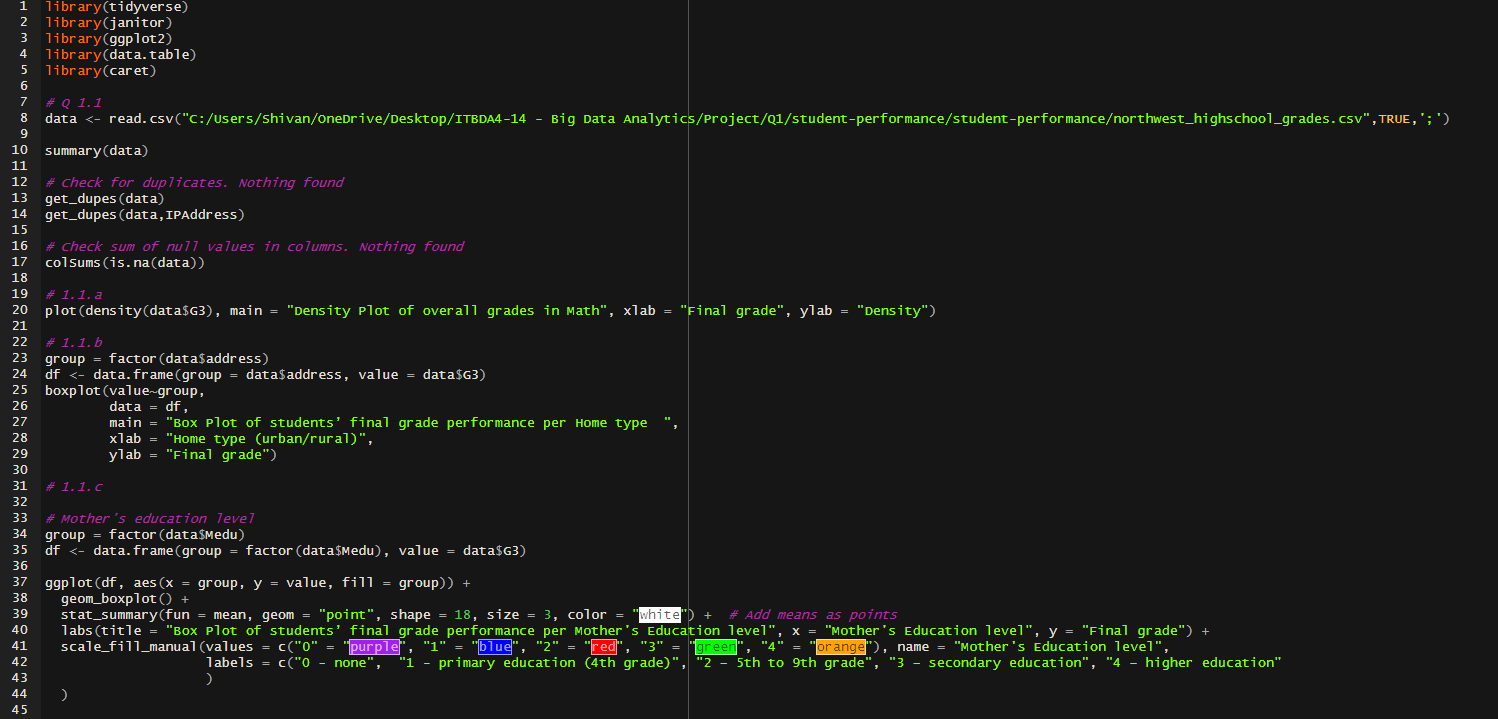
[Question 2 7](#_Toc170668397)

[Question 3 11](#_Toc170668398)

[Question 4 16](#_Toc170668399)

[References 23](#_Toc170668400)

**Question 1**

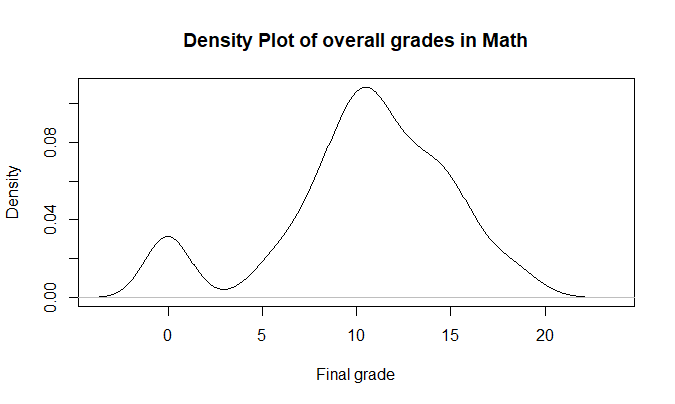


A screenshot of a computer screen

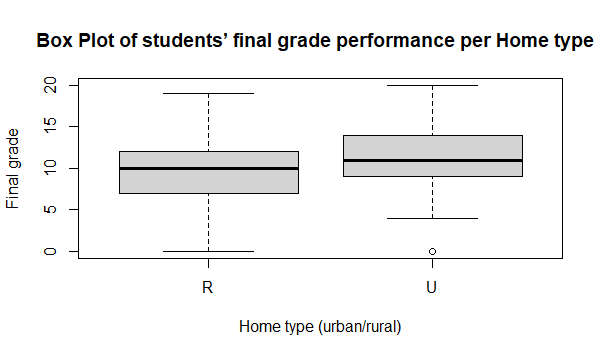
Description automatically generated

1.1.       Load the datasets into R, and provide the following graphs:

 a. a density function of the overall final grade of students            (2 marks)

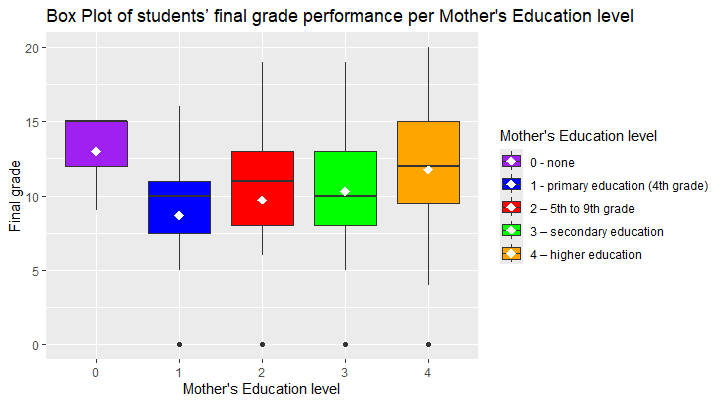


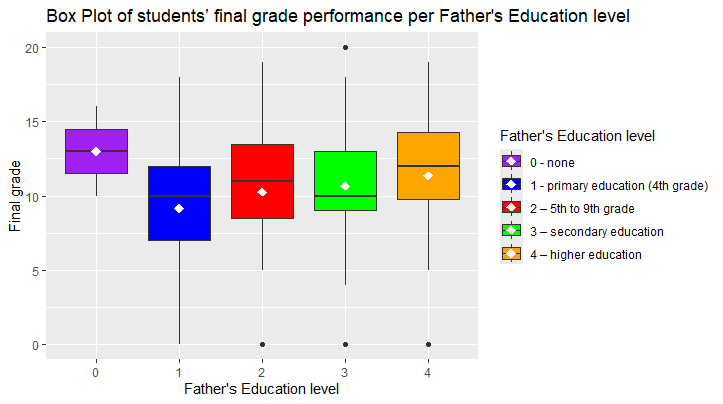
b. a box plot of students’ final grade performance per home type (urban/rural) (2 marks)



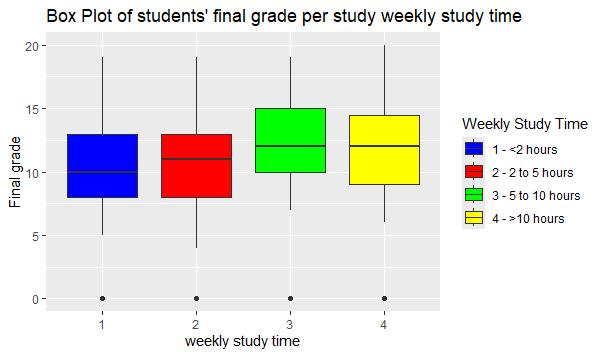
c. a box plot of students’ final grade performance per parents’ education. Average both parents’ education level (3 marks)

NOTE: white diamond represent the mean per education level"

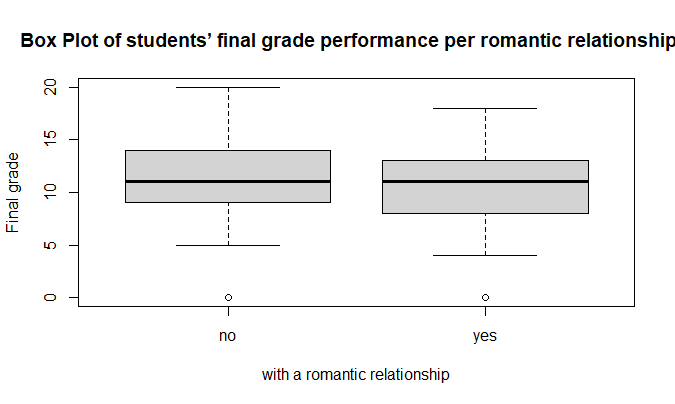




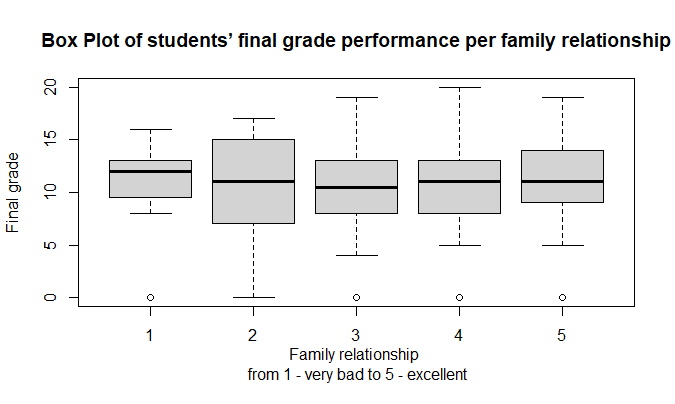
d. a box plot of students’ final grade performance per study time   (3 marks)



e. a box plot of students’ final grade performance per romantic status                   (2 marks)



f. a box plot of students’ final grade performance per quality of family relationships     (3 marks)



[Sub Total 15 Marks]

# **Question 2**

NOTE: student’s performance = dependent variable

2.1.a Student study time (3 marks)

A screenshot of a computer

Description automatically generated

Ho: There is no difference among group (different study times) means.

H1: There is at least one group differs significantly from the overall mean of the dependent variable.

Since P-Value = 0.161 > alpha = 0.05, we fail to reject Ho.

We conclude that there is not enough evidence to conclude that there is a significant difference between the group means.

2.2.b. The students home address type (3 marks)

A computer screen with white text

Description automatically generated

Ho: There is no difference among group (different addresses: Rural or Urban) means.

H1: There is at least one group differs significantly from the overall mean of the dependent variable.

Since P-Value = 0.0356 < alpha = 0.05, we reject Ho.

We conclude that at least one group differs significantly from the overall mean of the dependent variable.

2.2.c. Student romantic status (3 marks)

A computer screen with white text

Description automatically generated

Ho: There is no difference among group (different marital status: Yes or No) means.

H1: There is at least one group differs significantly from the overall mean of the dependent variable.

Since P-Value = 0.0.00971 < alpha = 0.05, we reject Ho.

We conclude that at least one group differs significantly from the overall mean of the dependent variable.

2.2.d. Quality of family relationship.  (3 marks)

A screenshot of a computer

Description automatically generated

Ho: There is no difference among group (different Quality of family relationship) means.

H1: There is at least one group differs significantly from the overall mean of the dependent variable.

Since P-Value = 081 < alpha = 0.05, we fail to reject Ho.

We conclude that there is not enough evidence to conclude that there is a significant difference between the group means.

2.2.e. The parents’ education level   (3 marks)

**Mother education level**

A screenshot of a computer

Description automatically generated

Ho: There is no difference among group (different Mother’s education level) means.

H1: There is at least one group differs significantly from the overall mean of the dependent variable.

Since P-Value < alpha = 0.05, we reject Ho.

We conclude that at least one group differs significantly from the overall mean of the dependent variable.

**Father’s education level**

A screenshot of a computer

Description automatically generated

Ho: There is no difference among group (different Father’s education level) means.

H1: There is at least one group differs significantly from the overall mean of the dependent variable.

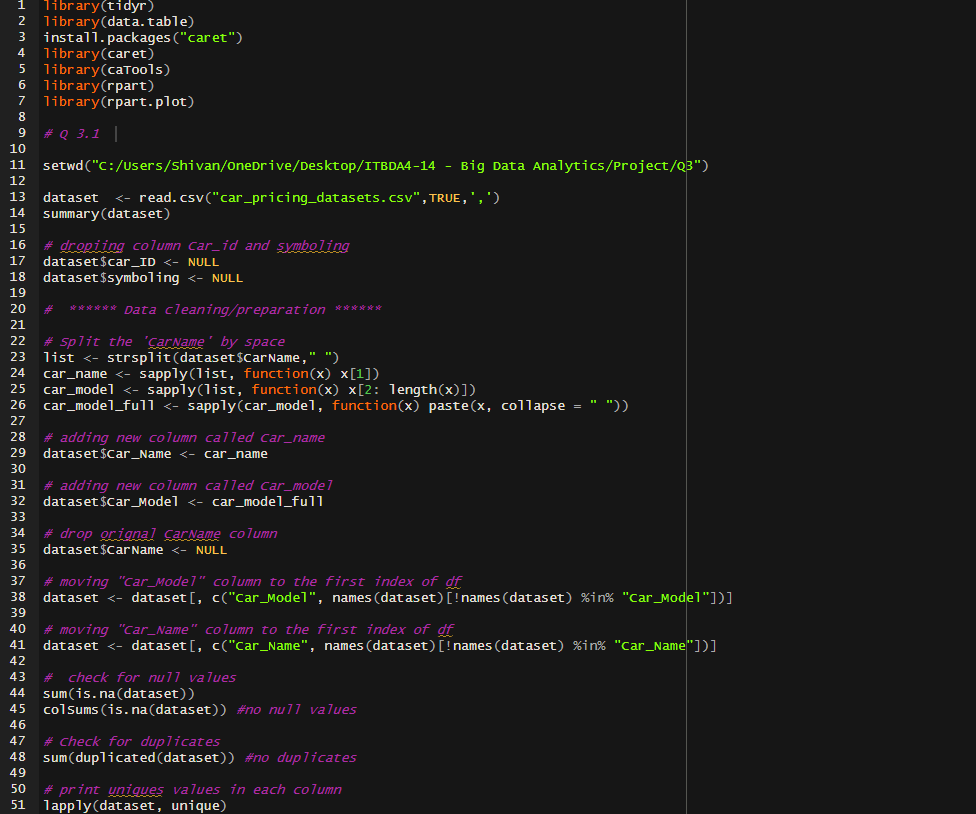
Since P-Value = 0.0222 < alpha = 0.05, we reject Ho.

We conclude that there is not enough evidence to conclude that there is a significant difference between the group means.

[Sub Total 15 Marks]

**Question 3**

3.1. Load the datasets into R, list all datasets and transform the CarName to extract the actual car name. Remove the car\_ID and symbolling columns as part of the datasets (4 marks)



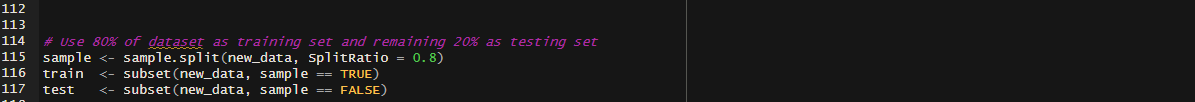
A screen shot of a computer code

Description automatically generated

3.2     Perform One Hot encoding to transform categorical variables into binary variables, normalise all feature variables and split the datasets into Training/Test with an 80/20 proportion.         (7 marks)

A screenshot of a computer program

Description automatically generated



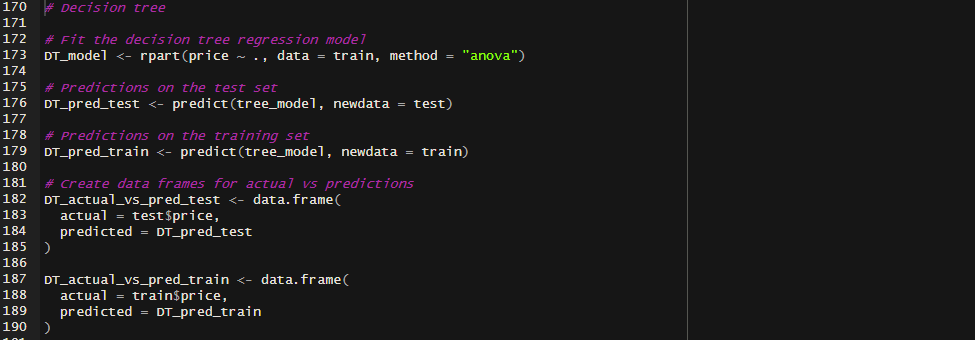
3.3     Build the following models:

a.         Multiple Linear Regression                                                (4 Marks)

A screen shot of a computer

Description automatically generated

b.         Decision Tress Regression                                                  (4 Marks)



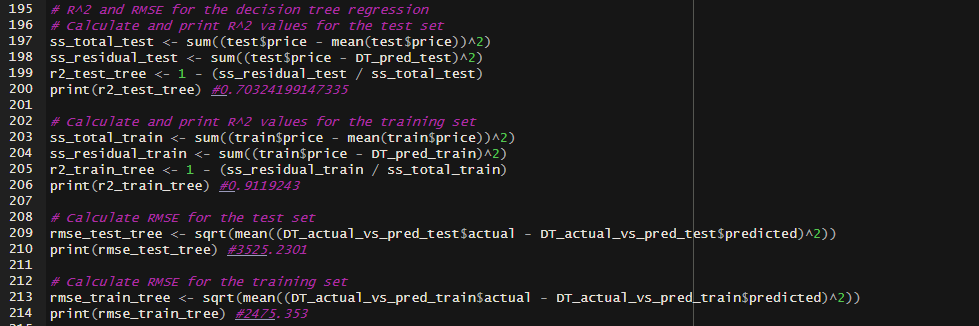
3.4 Compute the coefficient of determination and root mean square error for the two models on both the training and test set. (8 Marks)

Multiple Linear Regression model

A computer screen with text on it

Description automatically generated

Decision Tree Regression model

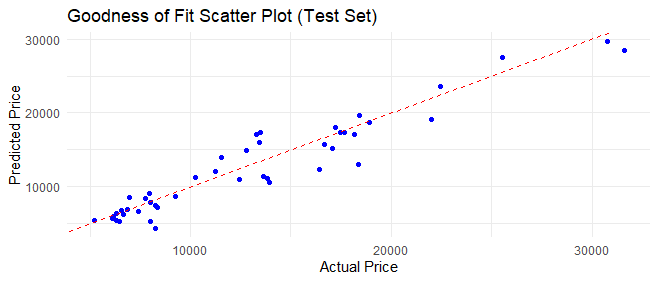


3.5 Draw the goodness of fit scatter plot on the test set for both models and comments on the model performances

(8 marks)

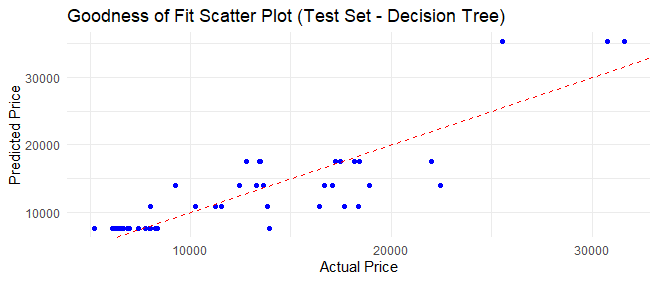
A computer screen shot of a computer code

Description automatically generated



A computer screen with text

Description automatically generated



3.6 Comment on the model performances, explain why one model performs better than the other.                                                                                                 (5 marks)

**Multiple Linear Regression (MLR) Model**

|  |  |
| --- | --- |
| **Training Set** | **Test Set** |
| R2 = 0.94 | R2 = 0.92 |
| RMSE = 2025.32 | RMSE = 2091.72 |

* From the training set, an R2 value of 0.94 indicates that 94% of the variance in the target variable is explained by the model. Similarly, From the test set, an R2 value of 0.92 indicates that 92% of the variance in the target variable is explained by the model. This value is lower than the R2 of the training set which is due to overfitting.
* A RMSE of 2025.32 indicates that the model's predictions are quite closer to the actual values compared to the RMSE on the test set.

**Decision Tree Regression Model**

|  |  |
| --- | --- |
| **Training Set** | **Test Set** |
| R2 = 0.70 | R2 = 0.91 |
| RMSE = 3525.23 | RMSE = 2475.35 |

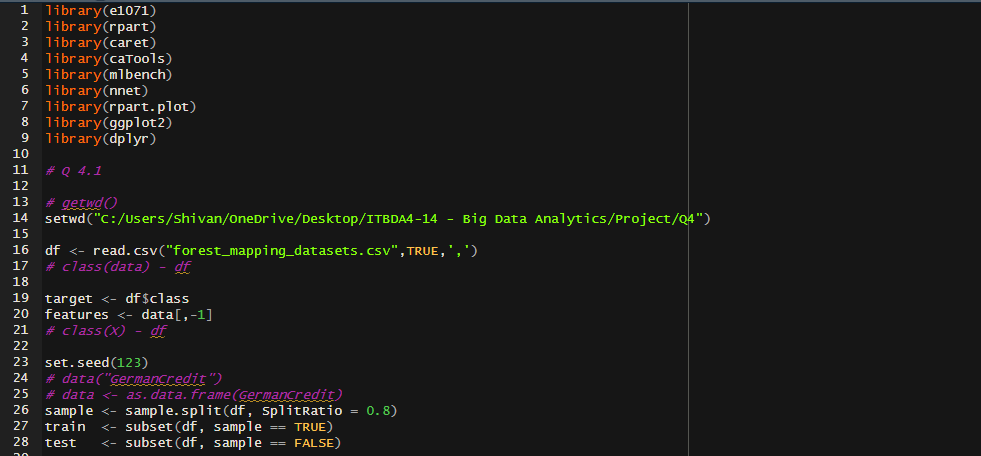
* From the training set, an R2 value of 0.70 indicates that 70% of the variance in the target variable is explained by the model. Similarly, From the test set, an R2 value of 0.91 indicates that 91% of the variance in the target variable is explained by the model. This indicates good explanatory power.
* A RMSE of 2475.35 indicates that the model's predictions are quite closer to the actual values compared to the RMSE on the training set.

The **Multiple Linear Regression** model performs better overall, with high R2 values and low RMSE values on both training and test sets, when compared to the decision tree regression model. This indicates that the MLR model captures the underlying patterns in the data well and generalizes better to unseen data.

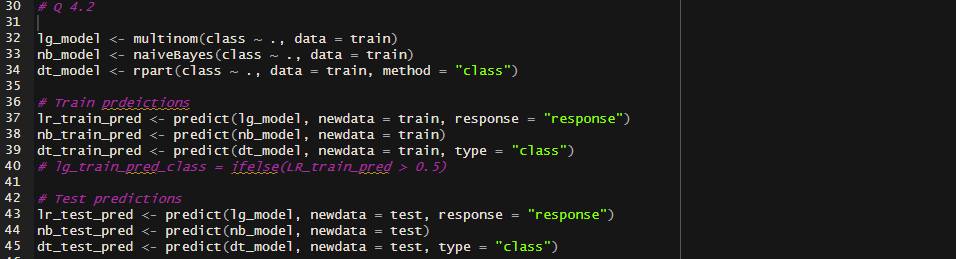
**[Sub Total 40 Marks]**

# **Question 4**

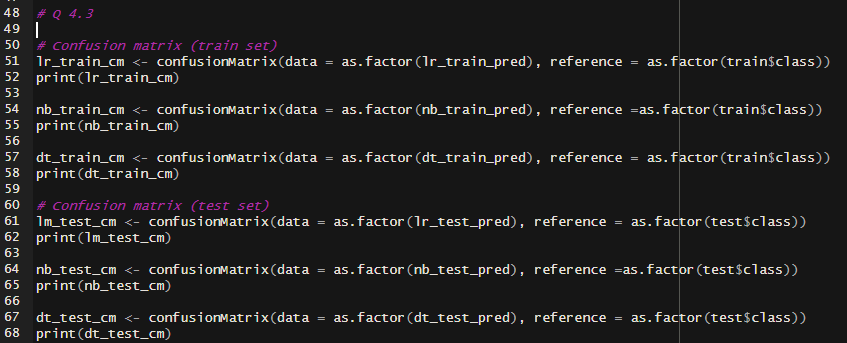
4.1 Load the datasets in R, extract the features and target variable and split the data in training/test set with an 80/20 proportion. (4 marks)



4.2 Build a logistic regression, a Naïve Bayes model, and a decision tree model to classify the types of forest in the Japanese landscape. (9 marks)



4.3 Compute the confusion matrices and classification accuracy for the three models on both the training and test set. Comment on the performance of the models. (10 marks)



Confusion matrices

Training sets

A screenshot of a computer

Description automatically generated

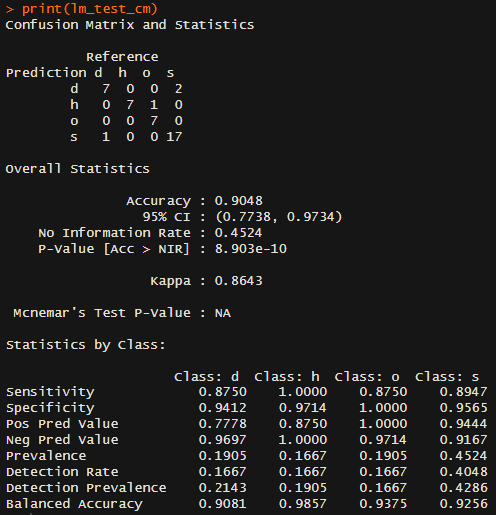
A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Test sets.



A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Logistic Regression model**

|  |  |
| --- | --- |
| **Training set** | **Test set** |
| Accuracy = 100% | Accuracy = 90% |

**Naïve Bayes model**

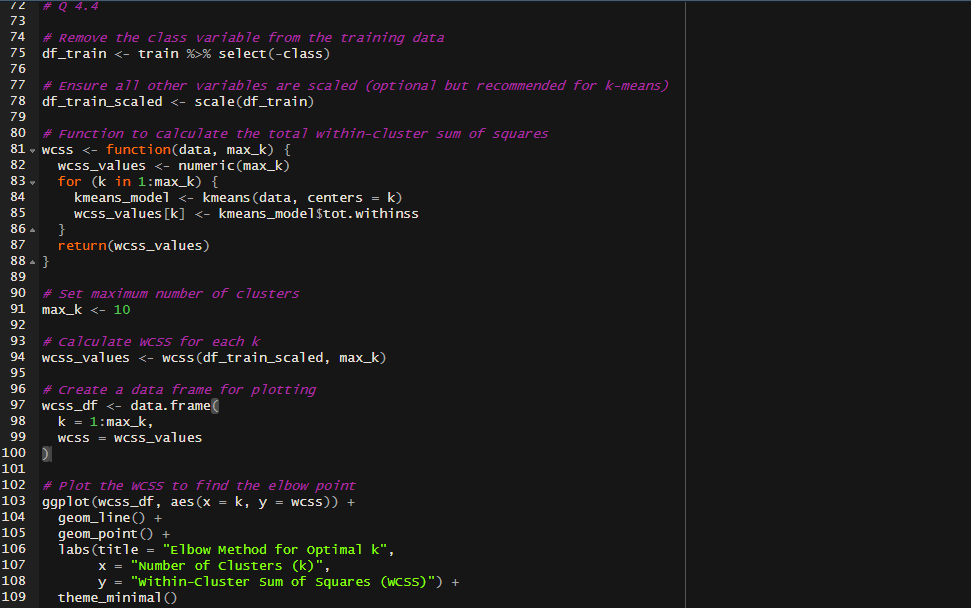
|  |  |
| --- | --- |
| **Training set** | **Test set** |
| Accuracy = 97% | Accuracy = 95% |

**Decision tree model**

|  |  |
| --- | --- |
| **Training set** | **Test set** |
| Accuracy = 98% | Accuracy = 93% |

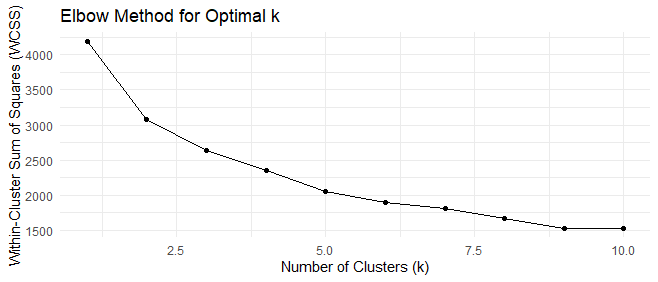
Naïve Bayes shows the best generalization with only a slight decrease in accuracy from the training set to the test set. Overall. The N**aïve Bayes** has the highest test set accuracy, suggesting it is the best performing model on unseen data.

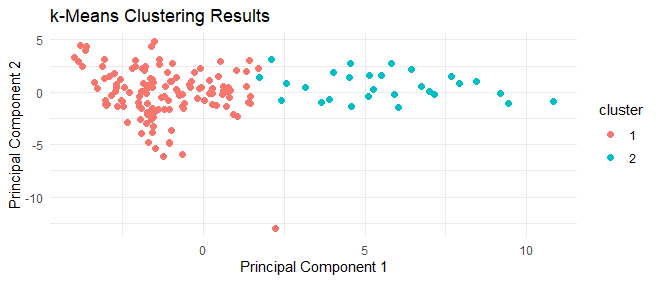
4.4 Using similarity information only (exclude the class variable) via optimal k-Means clustering distinguish the different types of forest with the Japanese landscape. Based your analysis on the training set only. Make use of the elbow method, and provide a within cluster sum of squares line graph (7 marks)



A computer screen with colorful text

Description automatically generated





**[Sub Total 30 Marks]**

# **References**

* <https://www.geeksforgeeks.org/box-plot-in-r-using-ggplot2/>
* <https://www.datacamp.com/tutorial/decision-trees-R>
* <https://stackoverflow.com/questions/47321665/finding-correlation-between-numeric-encoded-categorical-variables>
* <https://stats.stackexchange.com/questions/484299/how-to-check-the-correlation-between-categorical-and-numeric-independent-variabl>
* <https://www.youtube.com/watch?v=mno47Jn4gaU&t=2s>
* <https://medium.com/edureka/naive-bayes-in-r-37ca73f3e85c>
* <https://www.geeksforgeeks.org/root-mean-square-error-in-r-programming/>
* <https://stackoverflow.com/questions/50564111/how-to-only-print-adjusted-r-squared-of-regression-model>
* <https://www.r-bloggers.com/2021/07/how-to-calculate-root-mean-square-error-rmse-in-r/>
* <https://www.r-bloggers.com/2021/11/how-to-plot-observed-and-predicted-values-in-r/>
* <https://www.datacamp.com/tutorial/decision-trees-R>
* <https://www.statmethods.net/graphs/histograms-and-density.html>
* <https://www.statology.org/elbow-method-in-r/>
* <https://www.r-bloggers.com/2017/02/finding-optimal-number-of-clusters/>
* https://rpubs.com/markloessi/499223