Data Analytics - Final Project

Name: Viren Chhabria **ID:** 18301780 **Course:** CS7DS1

December 31, 2018

1 Introduction

In this project, the following analysis was performed on the given dataset:

- Finding the relation between different independent variables with the target label
- Visualizing and imputing missing data
- Assessing the impact of data imputation on the association between variables
- Finding the most important predictors of the target label

A heatmap was used for visualizing the correlation between variables and to check the missingness in the data. k-nearest neighbours (kNN) algorithm was used for data imputation. Decision Trees and Random Forest were used to compare the effects of different variables on the target label and to find the variable importance.

2 Methodology

2.1 Dataset

The dataset consists of 17 attributes and 296 samples. The first attribute was "ID", which was an index to the rows in the dataset. The target label was "Response" which either had class "0" or class "1". X1 - X7 were 7 attributes with continuous values. Y1 - Y7 were 7 categorical attributes, with each class of Y dependent on the respective X attribute having values in a specific range.
Table 1 shows the attributes in the dataset along with the respective data types. Table 2 shows a descriptive overview of the dataset.

Table 1: Dataset Attributes

Attribute	Type
Response	Categorical
Group	Categorical
X1 - X7	Continuous
Y1 - Y7	Categorical

Table 2: Description of the dataset

ID	Name of the Feature	Type	If Numerical			If Factor	Number of NA (Missing Data)
			Min	Max	Mean	Values	
1.	ID	Numeric	1.00	296.00	148.50		0
2.	Group	Factor				0,1	0
3.	X1	Numeric	5.00	9743.00	301.30		4
4.	X2	Numeric	0.00	80919.00	2908.60		130
5.	X3	Numeric	0.00	143856.00	5015.00		131
6.	X4	Numeric	21.82	6864.00	233.34		0
7.	X5	Numeric	0.10	99.80	35.32		4
8.	X6	Numeric	0.90	9.70	3.84		63
9.	X7	Numeric	110.30	8491.10	1353.10		24
10.	Y1	Factor				0,1	4
11.	Y2	Factor				0,1	130
12.	Y3	Factor				0,1	131
13.	Y4	Factor				0,1	0
14.	Y5	Factor				0,1	4
15.	Y6	Factor				0,1,2	63
16.	Y7	Factor				0,1	24
17.	Response(Target)	Factor				0,1	0

2.2 Tools

R programming language was used for analyzing the dataset, implementing the algorithms and creating the visualizations(default plot and GGPlot2).

2.3 Algorithms

2.3.1 K-nearest neighbours (kNN)

kNN was used in this project for data imputation. It can be used for both, regression and classification. kNN works by finding k existing samples closest to the new given sample, and by majority vote, deciding the value of the given sample. In R, the "VIM" library was used for the implementation of the kNN algorithm.

2.3.2 Decision Trees

Decision trees are popular and powerful tools used for classification and prediction. They represent rules that can be understood by humans and used in knowledge systems such as a database. Decision tree classifier was used to build a model with the given dataset. In R, the "rpart" library was used for the implementation of decision tree classifier.

2.3.3 Random Forest

Random forest is simply an ensemble of decision trees. It uses Bagging and Random Subspace Sampling to build a model. In this project, the random forest was used to determine variable importance and to find the classification error rate trends with respect to the number of trees in the random forest. In R, the "randomForest" library was used for the implementation of random forest classifier.

2.4 Preprocessing

2.4.1 Missing Data

Figure 1 shows the visual representation of the densities of missing values across each attribute in the dataset. Segments represent the sub sections of the dataset. The dataset here was split into 10 segments. "All" indicates the average missing value rate for entire dataset by variables. Overall 15.03% of data was missing in the dataset.

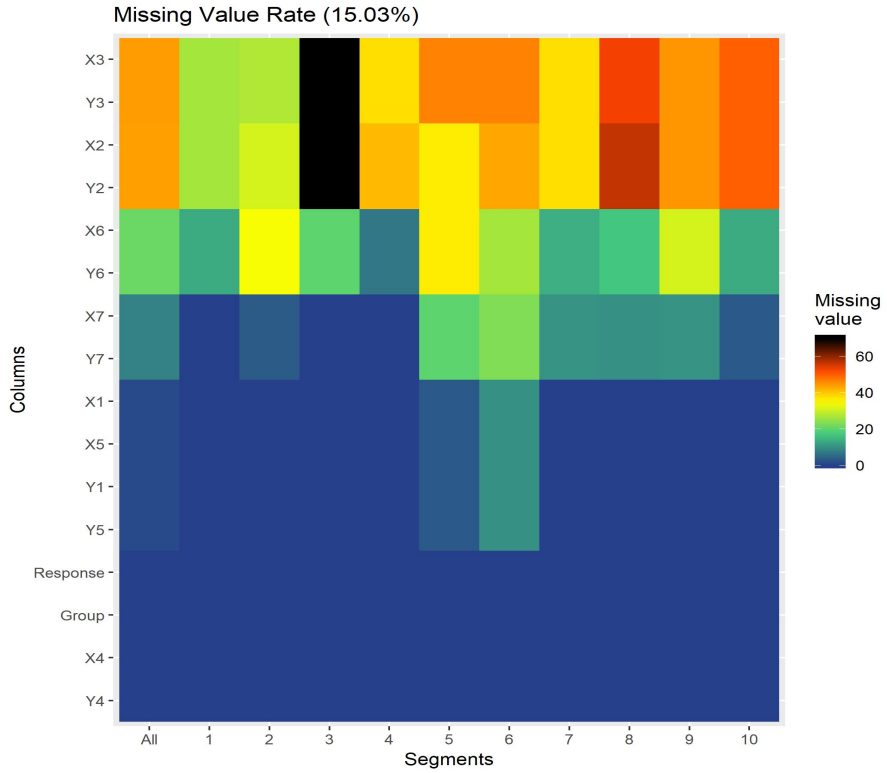

Figure 1: Missing Data heatmap for all attributes

2.4.2 Data Transformation

It is known that the Y variables were derived based on the corresponding X variables. For examples, Y1 was derived from X1, Y2 was derived from X2 and so on, though the exact criteria which were used to get the respective Y values from X are not known. Qhile loading the dataset in R, the default datatype of the Y variables in the dataset was taken to be numeric, however, in reality, it was categorical. Therefore data transformation was performed on the Y variables where they were converted from numeric to factors.

2.4.3 Feature Removal

The first attribute in the dataset, "ID" was an index to the rows in the dataset. It was removed from the dataset, due to its irrelevance to the target label.

2.4.4 Data Imputation

The missing data in the given dataset was imputed using K-nearest-neighbours (kNN) algorithm, with the value of k being 5. Figure 2 shows the summary of the dataset before imputation, while Figure 3 shows the summary of the dataset after imputation. Both the figures show that the data imputation didn't alter the mean of the variables too much, except for X2 and X3. Figure 4 is a heatmap that shows that all the missing data was imputed successfully.

```
Response  Group          X1                X2
 0:154    0: 96   Min.   :   5.0   Min.   :    0.00
 1:142    1:200   1st Qu.:  16.0   1st Qu.:   30.25
                  Median :  38.0   Median :  126.00
                  Mean   : 301.3   Mean   : 2908.60
                  3rd Qu.: 186.0   3rd Qu.:  558.75
                  Max.   :9743.0   Max.   :80919.00
                  NA's   :4        NA's   :130

       X3               X4                X5
 Min.   :     0   Min.   :  21.82   Min.   : 0.100
 1st Qu.:    40   1st Qu.:  50.61   1st Qu.: 9.057
 Median :   192   Median :  71.83   Median :19.300
 Mean   :  5015   Mean   : 233.34   Mean   :35.317
 3rd Qu.:   880   3rd Qu.: 132.38   3rd Qu.:61.970
 Max.   :143856   Max.   :6864.00   Max.   :99.800
 NA's   :131                        NA's   :4

       X6               X7              Y1          Y2
 Min.   :0.900   Min.   : 110.3   0   :143   0   :105
 1st Qu.:3.100   1st Qu.: 368.1   1   :149   1   : 61
 Median :3.600   Median : 653.2   NA's:  4   NA's:130
 Mean   :3.836   Mean   :1353.1
 3rd Qu.:4.300   3rd Qu.:1519.2
 Max.   :9.700   Max.   :8491.1
 NA's   :63      NA's   :24

    Y3          Y4          Y5          Y6          Y7
 0   : 89   0:144   0   :157   0   :  4   0   :112
 1   : 76   1:152   1   :135   1   :104   1   :160
 NA's:131           NA's:  4   2   :125   NA's: 24
                               NA's: 63
```

Figure 2: Dataset Summary Before Imputation

```
Response  Group          X1                X2
 0:154    0: 96   Min.   :   5.0   Min.   :    0.0
 1:142    1:200   1st Qu.:  16.0   1st Qu.:   33.0
                  Median :  39.0   Median :  114.0
                  Mean   : 299.2   Mean   : 1921.8
                  3rd Qu.: 186.0   3rd Qu.:  531.8
                  Max.   :9743.0   Max.   :80919.0

       X3               X4                X5
 Min.   :     0   Min.   :  21.82   Min.   : 0.100
 1st Qu.:    44   1st Qu.:  50.61   1st Qu.: 9.057
 Median :   192   Median :  71.83   Median :19.610
 Mean   :  3419   Mean   : 233.34   Mean   :35.347
 3rd Qu.:   944   3rd Qu.: 132.38   3rd Qu.:61.970
 Max.   :143856   Max.   :6864.00   Max.   :99.800

       X6               X7              Y1      Y2
 Min.   :0.900   Min.   : 110.3   0:144   0:186
 1st Qu.:3.200   1st Qu.: 374.4   1:152   1:110
 Median :3.700   Median : 671.9
 Mean   :3.815   Mean   :1396.5
 3rd Qu.:4.200   3rd Qu.:1644.7
 Max.   :9.700   Max.   :8491.1

   Y3      Y4      Y5      Y6      Y7
 0:163   0:144   0:158   0:  4   0:115
 1:133   1:152   1:138   1:128   1:181
                         2:164
```

Figure 3: Dataset Summary After Imputation

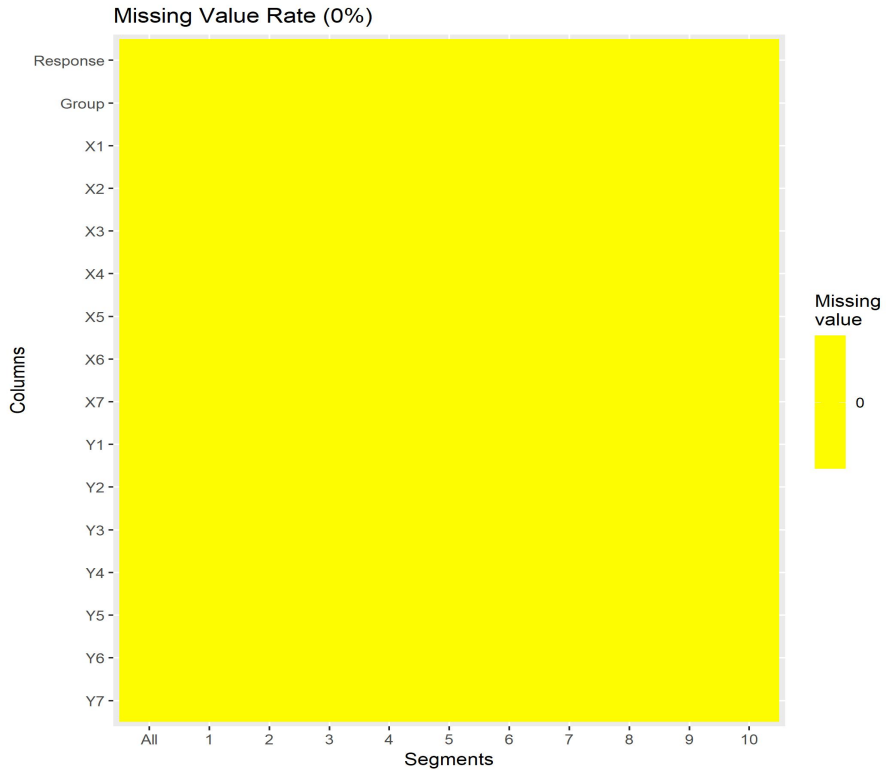

Figure 4: Missing Data heatmap for all attributes after imputation

2.5 Evaluation

The relationship between the different predictors with the target label "Response" and their importance was assessed by first using different combinations of independent variables to build decision tree's using the original dataset as is, without any data imputation and then repeating the same analysis using the imputed dataset. The combinations of attributes used were:

1. All X variables
2. All X variables plus Group
3. All Y variables
4. All Y variables plus Group
5. All variables

Decision trees can grow extremely large and fit each and every sample in the dataset. This would lead to an overfitted model (high variance). To avoid this scenario, a technique called pruning was applied to control the number of splits in the decision tree by removing sections of the tree that provide little power to classify the instances. All the decision tree's in this project were pruned to the complexity parameter (CP) corresponding to the minimum cross-validation error (xerror).

3 Results and Discussion

3.1 Correlation

Correlation matrix can be derived for numerical values, to understand the relationship between 2 variables. Heatmap's were created to visualize the relation between the various X values. Figure 5 shows the correlation heatmap before data imputation and Figure 6 shows the correlation heatmap after data imputation. There is no visible changes in the correlation between the X variables after data imputation. X2 and X3 are highly correlated in both the cases, hence one of them can be removed while building the model.

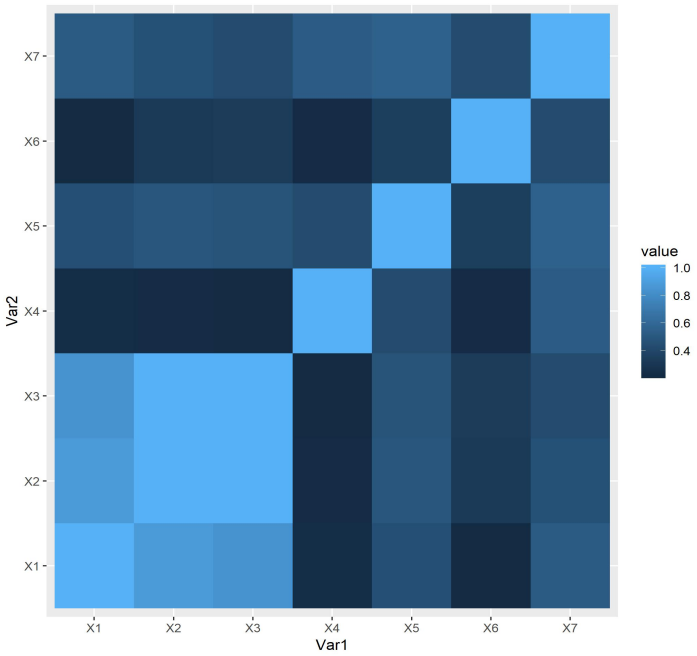

Figure 5: Coorelation heatmap for X1 - X7 attributes before data imputation

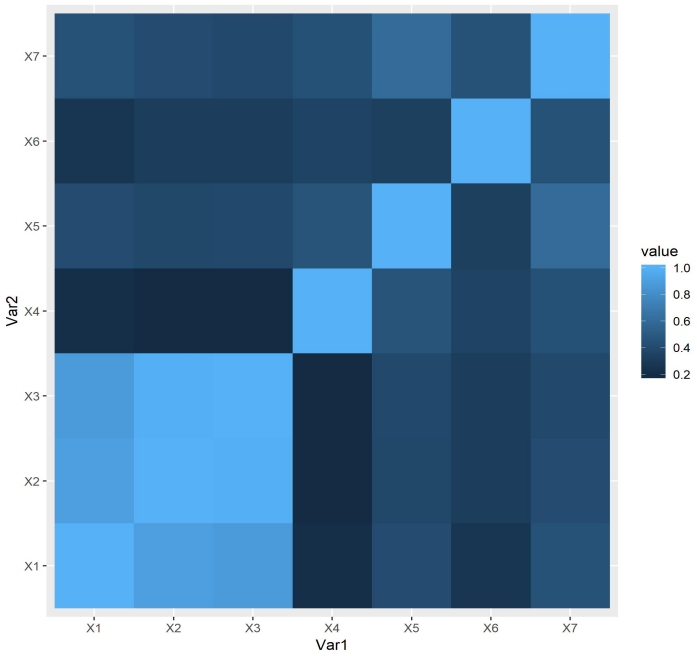

Figure 6: Coorelation heatmap for X1 - X7 attributes after data imputation

3.2 Decision Tree

1. **All X Variables**

Decision tree with only X variables, before data imputation had only 2 splits, however, post data imputation, 5 splits was found to have the lowest cross validation error. X1 and X4 remained significant splits in both the trees as shown in Figure 7 and Figure 8.

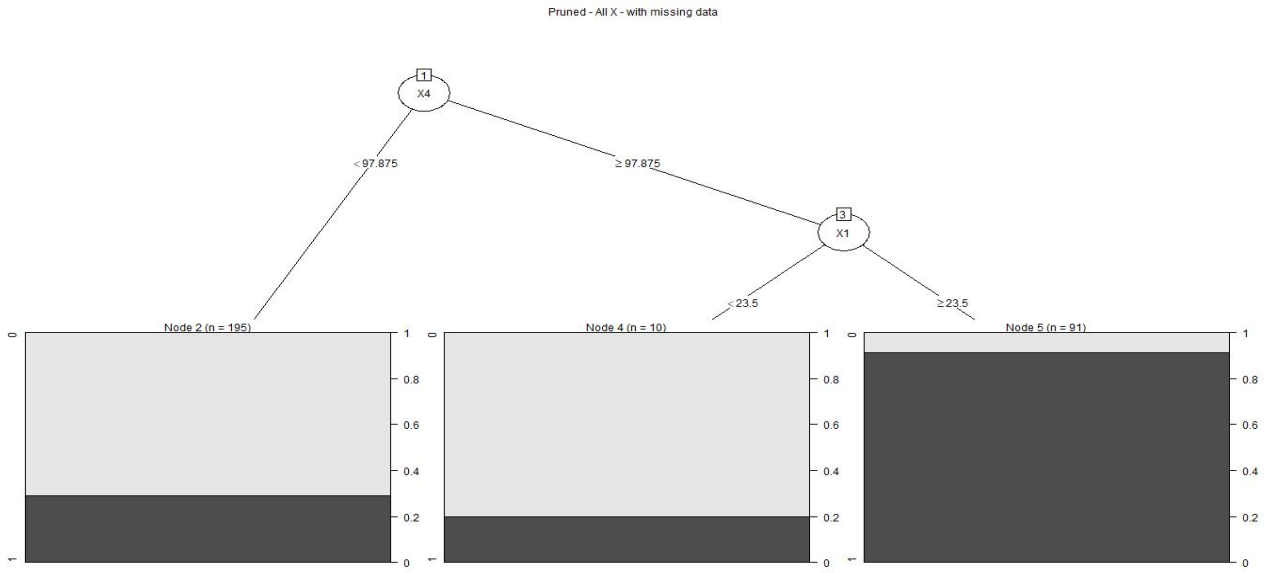

Figure 7: Pruned Decision Tree - All X variables as predictors before data imputation

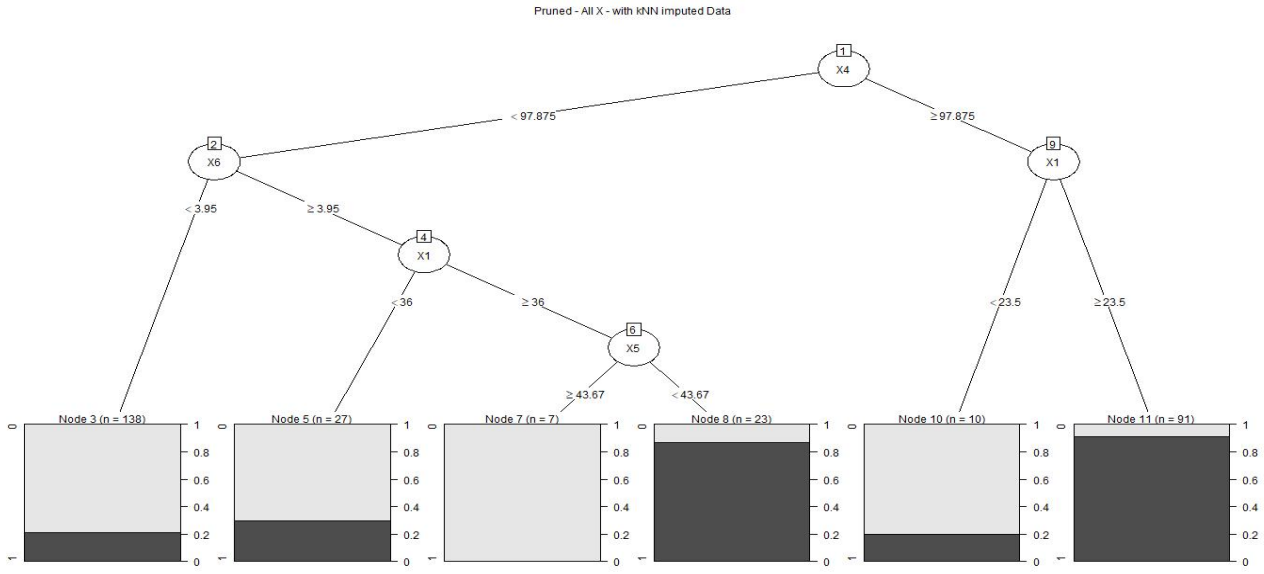

Figure 8: Pruned Decision Tree - All X variables as predictors after data imputation

2. **All X variables plus Group**

Considering "Group" along with the X variables had some impact on the decision tree using the dataset with missing values. Figure 9 shows that the best number of splits went up from 2 in the previous case to 9 in this case (overfitted model) for the decision tree before data imputation. After data imputation the best split decision tree remained the same as the previous case as shown in Figure 10. "Group" variable was not considered for a split in either tree, and X4 remained the root node in both cases.

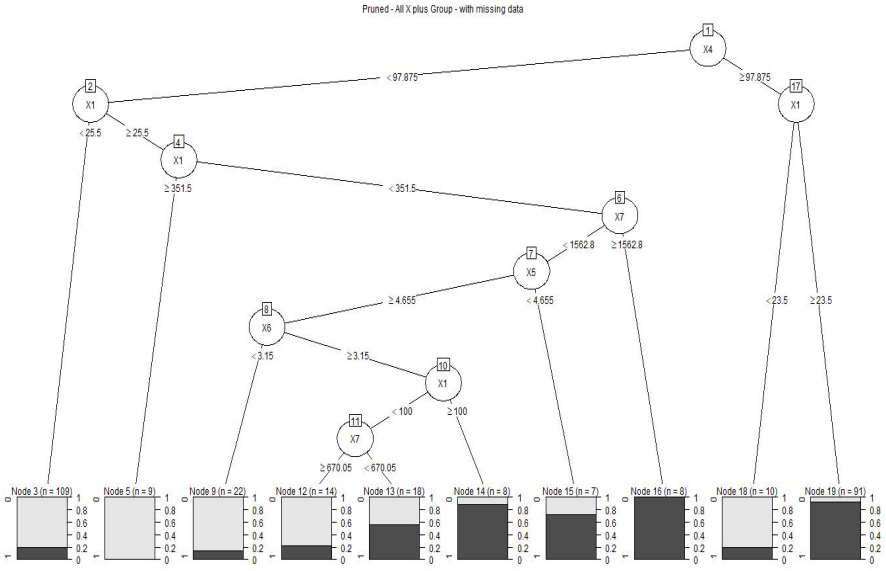

Figure 9: Pruned Decision Tree - All X plus Group variable as predictors before data imputation

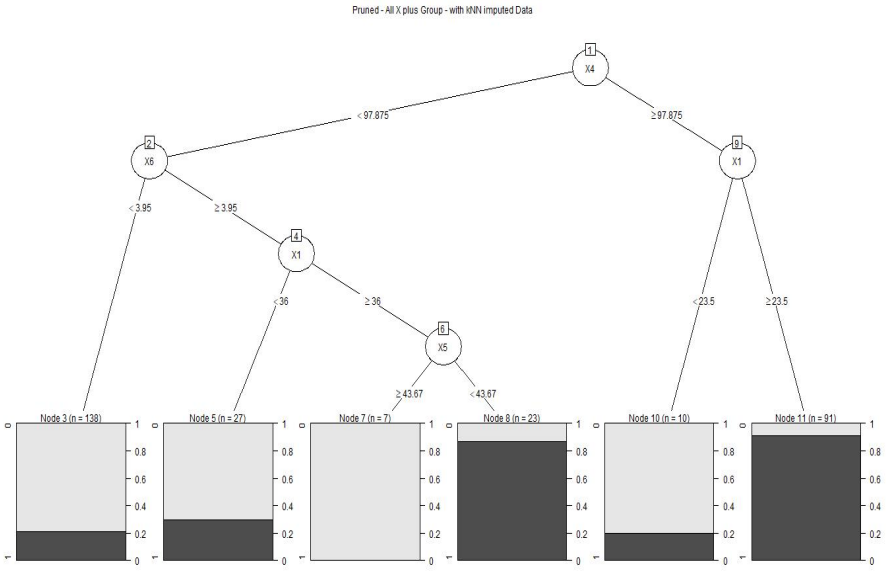

Figure 10: Pruned Decision Tree - All X plus Group variable as predictors after data imputation

3. **All Y variables**

Data imputation had an impact on the decision tree, when built with only Y variables as the predictors. Without imputation, the root node was Y5 and the best number of splits were 4, while after imputation, Y2 was the root node and the best number of splits were 3 as shown in Figure 11 and Figure 12.

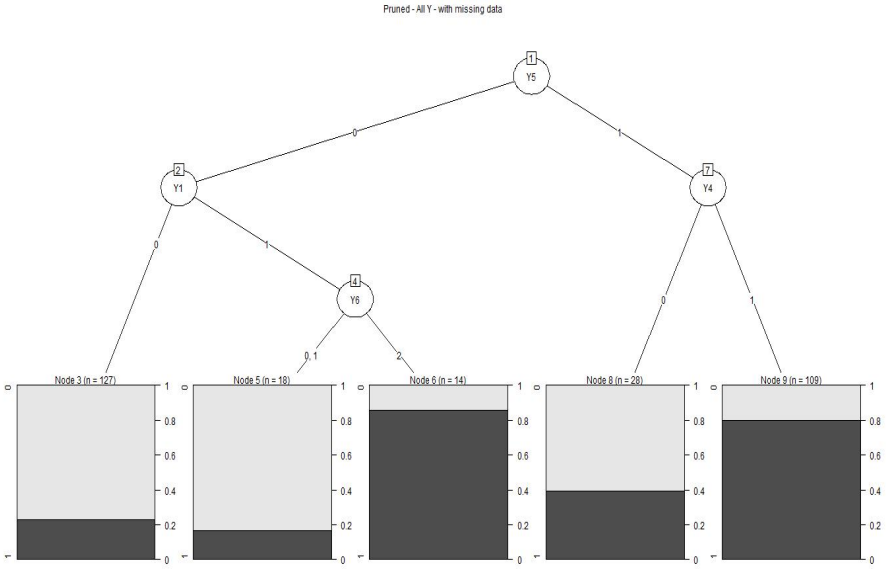

Figure 11: Pruned Decision Tree - All Y variables as predictors before data imputation

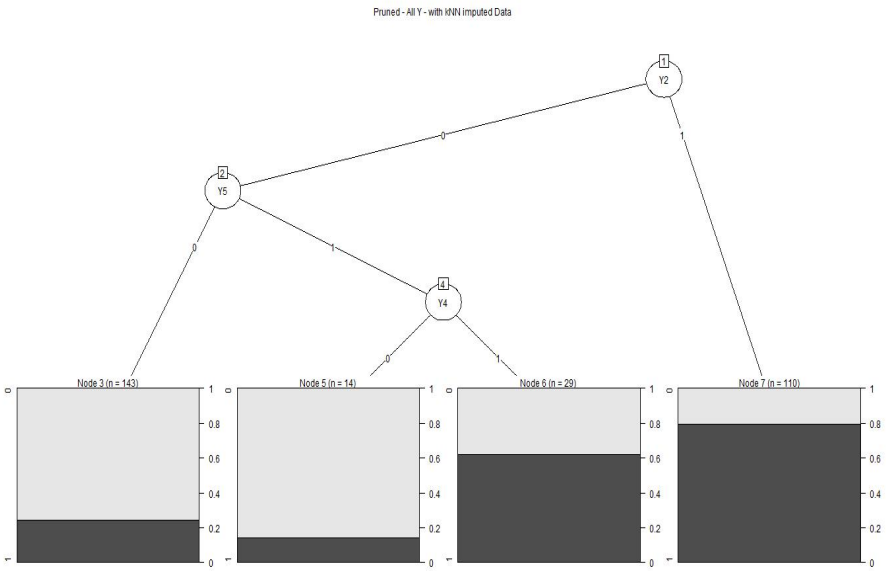

Figure 12: Pruned Decision Tree - All Y variables as predictors after data imputation

4. **All Y variables plus Group**

Considering "Group" along with the Y variables had some impact on the decision trees. The decision tree without imputation in Figure 13 remained similar to the pre-imputation decision tree with only Y variables, with Y5 as the root node and best number of splits were 4, while after imputation, Y2 was the root node and the only split, leading to an underfitted model. Though "Group" wasn't considered for a split, its introduction reduced the number of splits to 1, while using the imputed dataset as shown in Figure 14.

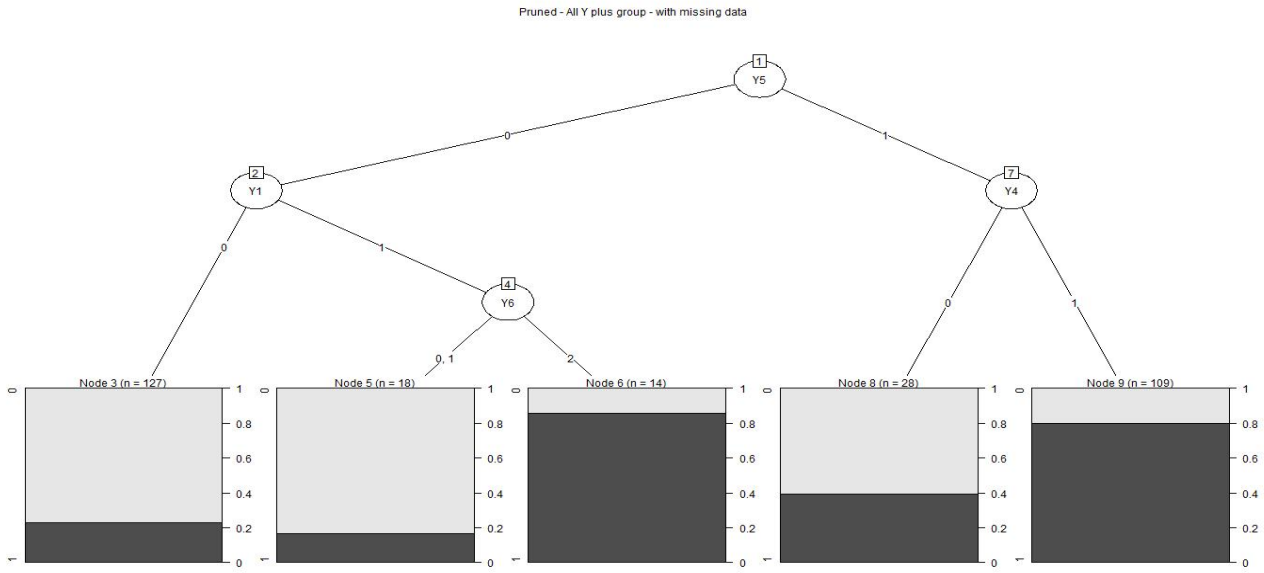

Figure 13: Pruned Decision Tree - All Y plus Group variable as predictors before data imputation

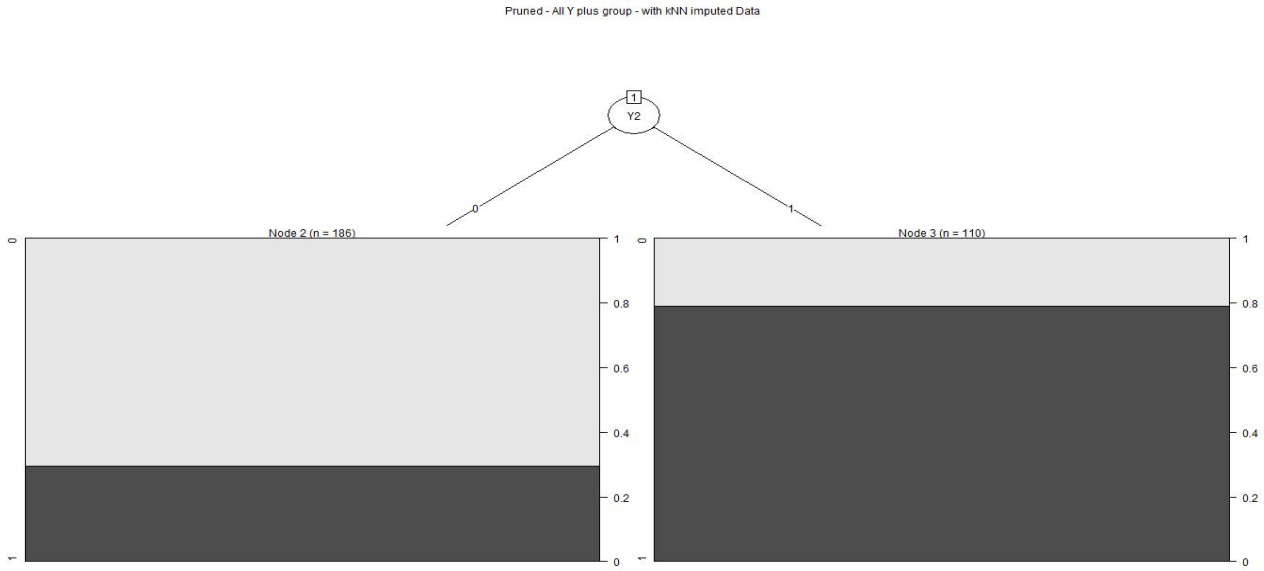

Figure 14: Pruned Decision Tree - All Y plus Group variable as predictors after data imputation

5. **All variables**

When all the variables were considered for building a decision tree, pre-imputation, only 2 splits with X4 and X1 were found to be the best as shown in Figure 15. Post-imputation, X4 remained as the primary split, with the total number of splits as 5. Only 1 Y variable, Y1 was used for a split, the remaining 4 were all X variables - X4, X1, X5 and X6 as shown in Figure 16.

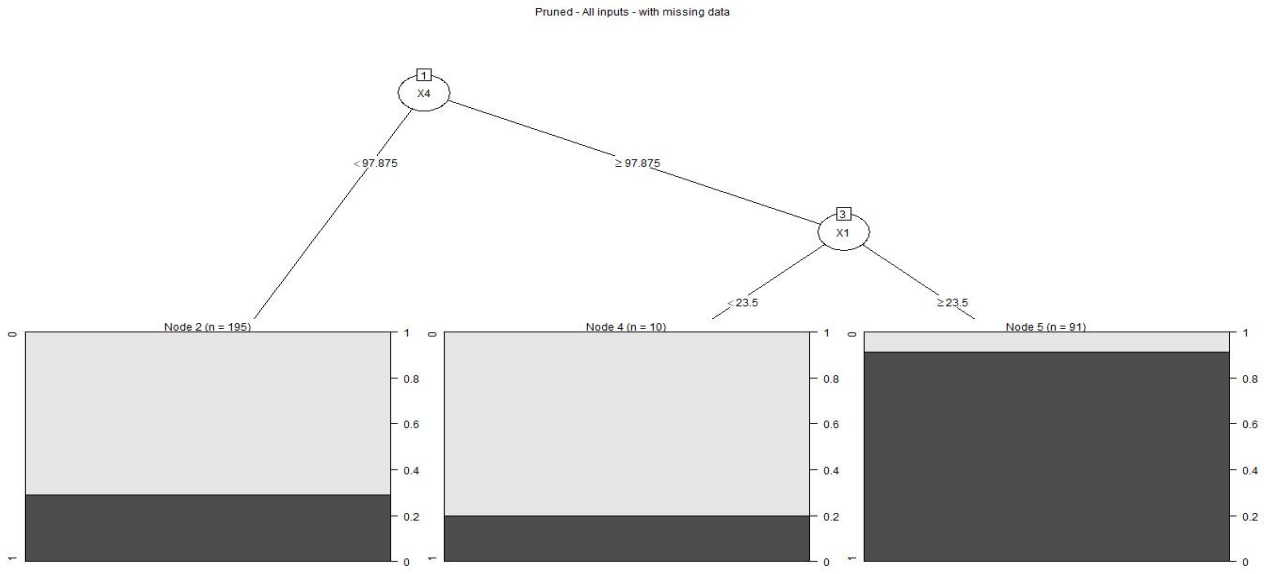

Figure 15: Pruned Decision Tree - All variables as predictors before data imputation

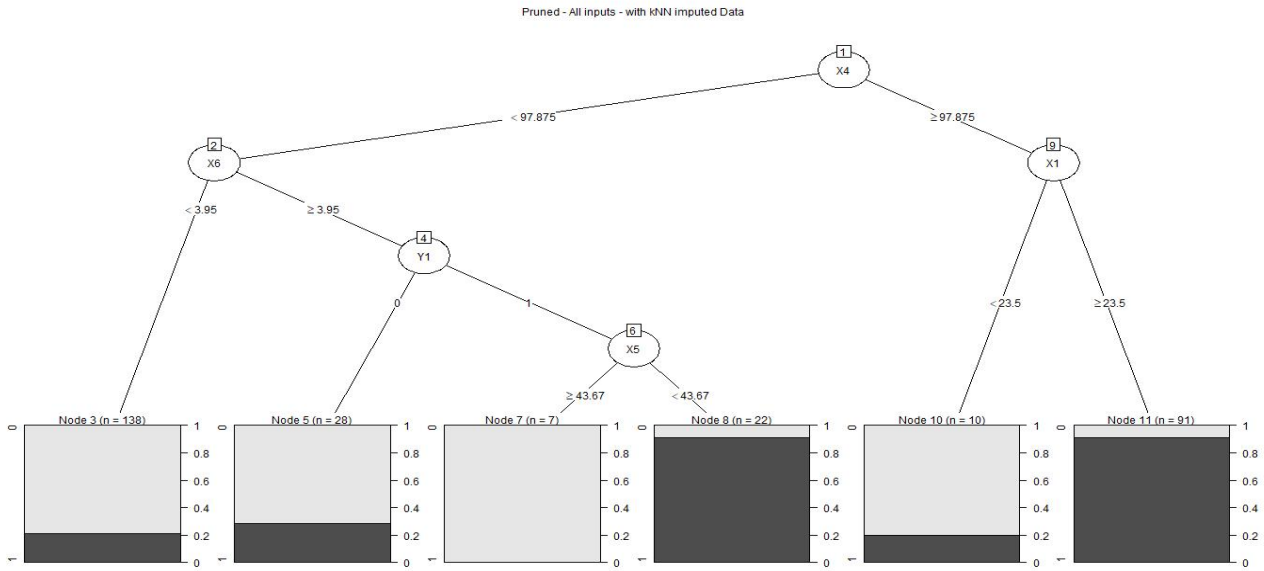

Figure 16: Pruned Decision Tree - All variables as predictors after data imputation

3.3 Random Forest

Random forest models were created using all the variables in the dataset in both the cases, with and without data imputation. From the variable importance plots in Figure 17 and Figure 18 it is evident that the X variables are significantly more important to predicting "Response", than the Y variables or the "Group" variable. Before and after data imputation, X4 was the most important variable with the lowest impurity. All the X variables were the better predictors in both cases over the Y variables and the "Group" variable.

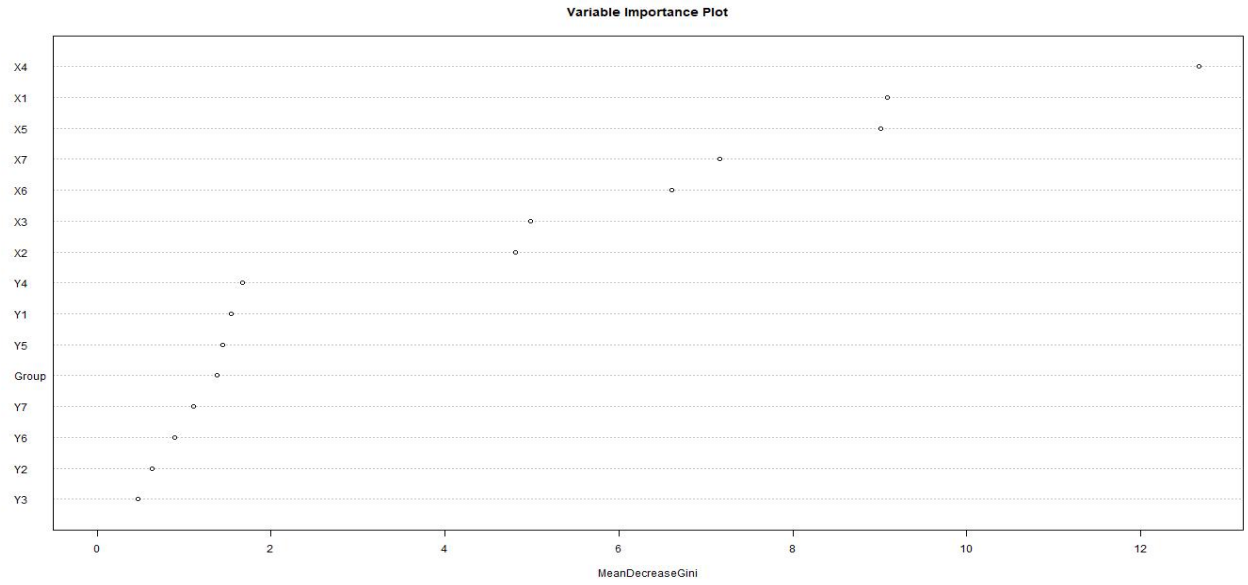

Figure 17: Variable Importance Plot before data imputation

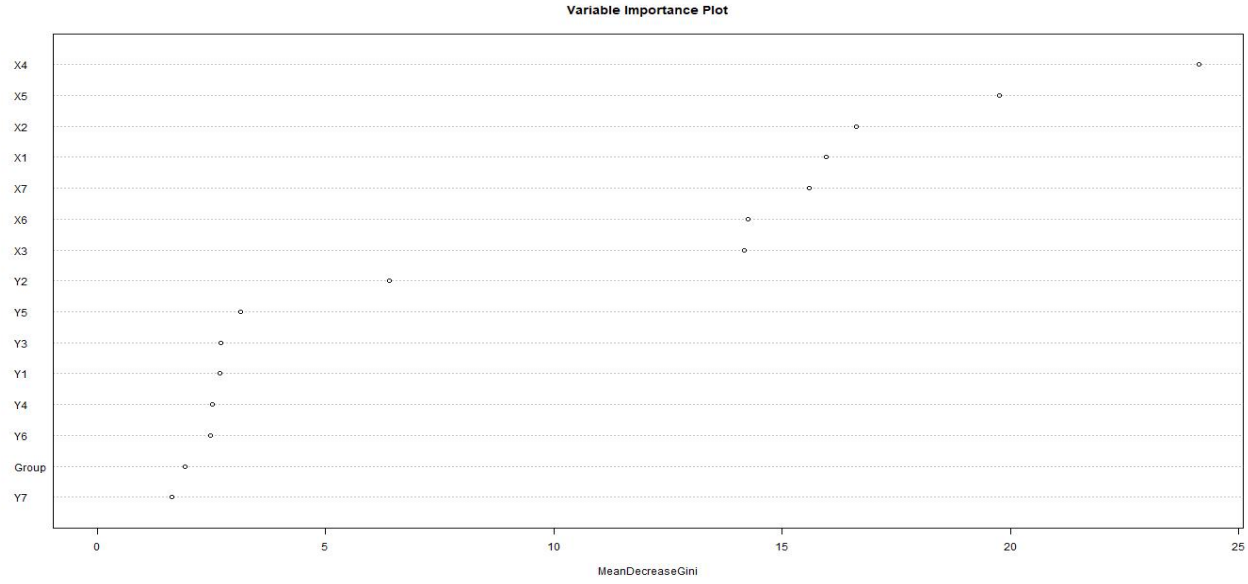

Figure 18: Variable Importance Plot after data imputation

The error plots in Figure 19 and Figure 20 show decreasing trends in prediction errors with an increase in the number of decision tree classifiers. The green line represents classification error for class "1", black line represents out-of-bag error and the red line represents classification error for class "0". The misclassification and out-of-bag error rate is slightly lower when using the imputed dataset.

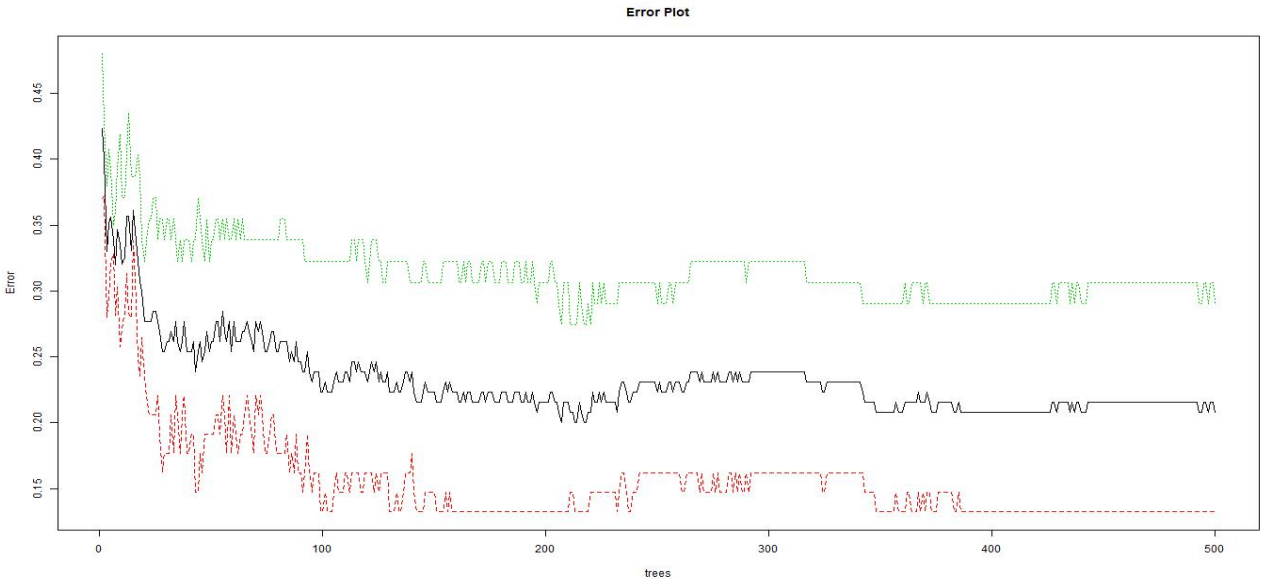

Figure 19: Error Plot before data imputation

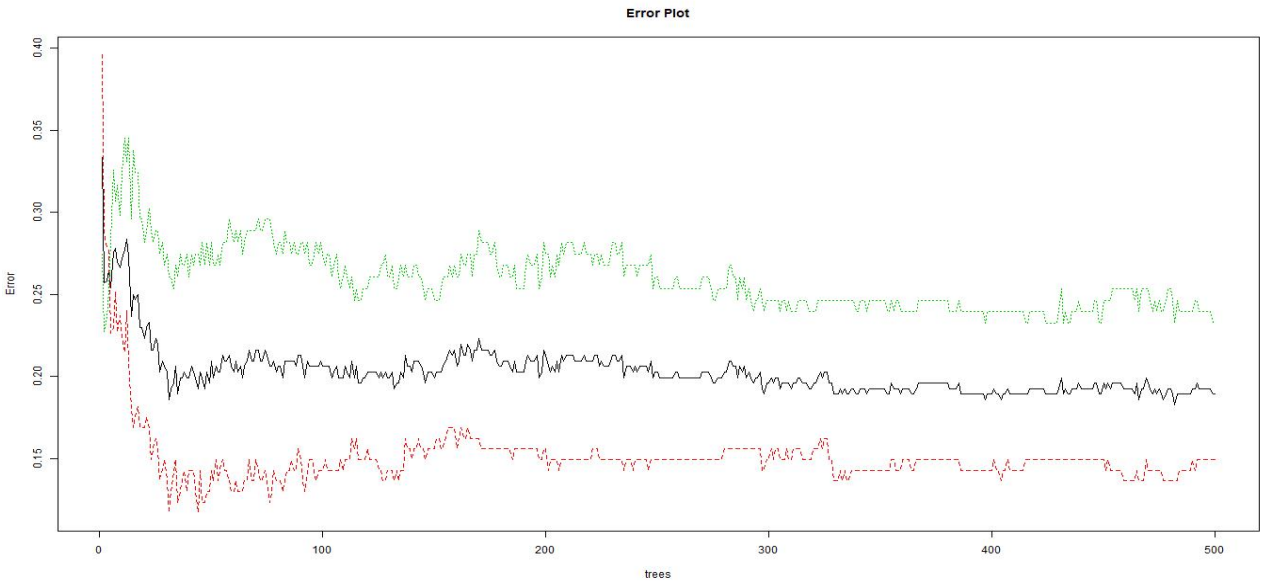

Figure 20: Error Plot after data imputation

4 Conclusion

The analysis of the given dataset in this project showed that X4 is the most important predictor of the "Response" label. In general, all the X variables provided a better understanding of the target label than the Y variables or the "Group" variable. The above conclusion can be attributed to 2 reasons as per the findings of this project:

1. The high importance of X4 can be attributed to the fact that X4 was the only variable (consequently Y4 also, since it is derived from X4) that did not have any missing data originally as shown in Figure 1. This meant that X4 had accurate values and hence gave the purest information about the target label

2. All the X variables were better predictors than the Y variables since Y variables were derived from X. Also, X variables were continuous and provided more specific information than Y variables, which were categorical, and the class of each Y variable was determined by a range of values of the respective X attribute