

School of Computing, Creative Technology and Engineering

|  |  |
| --- | --- |
| Student ID | 77356725 |
| Student Name | Sadikshya Duwadi |
| Module Name & CRN | Applied Machine Learning-18910 |
| Level | 5 |
| Assessment Name & Part No. | Case Study-3: Concrete Strength, PART-II |
| Project Title | Data Analysis on Concrete Strength |
| Date of Submission | 10 May, 2024 |
| Course | BSc. (Hons) Computing |
| Academic Year | 2024 |

Table of Contents

Summary……………………………………………………………………………………………3

Modelling………………………………………………………………………………………..3-16

Linear Regression…………………………………………………………………………………4-6

Random Forest….…………………………………………………………………………………6-8

Decision Tree…………………………………………………………………………………….9-11

KNN…………………………………………………………………………………………….11-13

KNN [=4]……………………………………………………………………………...………..13-14

Model Interpretation……………………………………………………………………………….16

Conclusion...……………………………………………………………………………………….16

Bibliography………………………………………………………………………………………17

“Data Analysis on Concrete Strength Part -2”

# PART II

Sadikshya Duwadi

BSc (hons) Computing, 2024

**Summary**

The first part of the report was started by combining the two concrete strength datasets ‘test’ and ‘train’ followed by executing preliminary tasks. These tasks encompassed the following: Exploratory Data Analysis (EDA) to uncover hidden patterns and trends in the data, Data Visualization and Principal Component Analysis (PCA) to reduce dimensionality. Data pre-processing was done to find missing values and overall missingness in the data and the missing values were handled using mean imputation. Finally, the outliers were detected and removed. The respective means of the variables were used to replace any missing values and the variables were scaled. The final dataset contained 1030 rows and 10 columns.

**Modelling**

Equipped with the pre-processed data, five different machine learning algorithms were developed to predict the concrete strength.

Data frame of combined scaled numeric data with non-numeric data

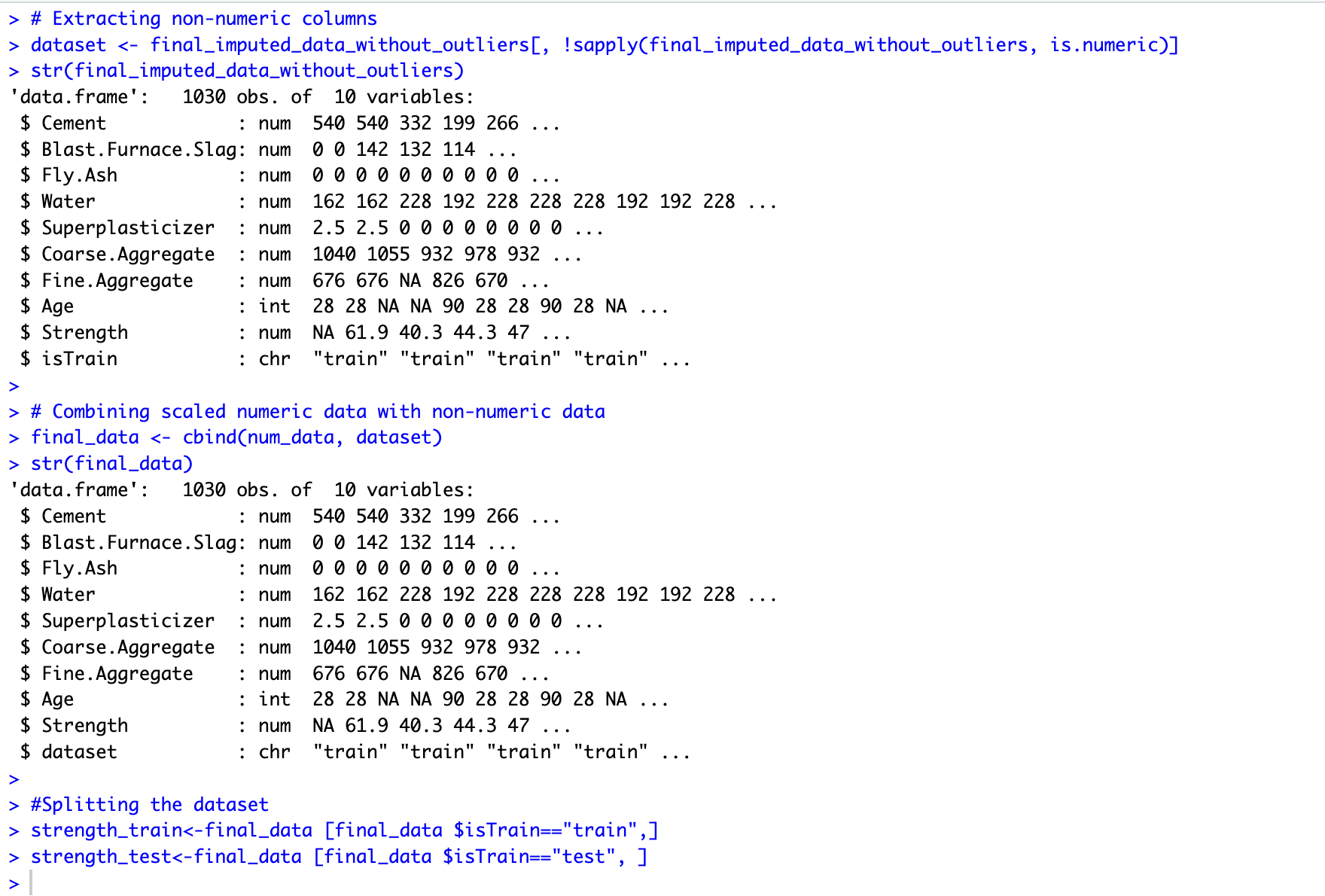
****

Figure 1.1 Structure of Combined numeric data and dataset

Data frame of test and train data after removing the column is\_Train.

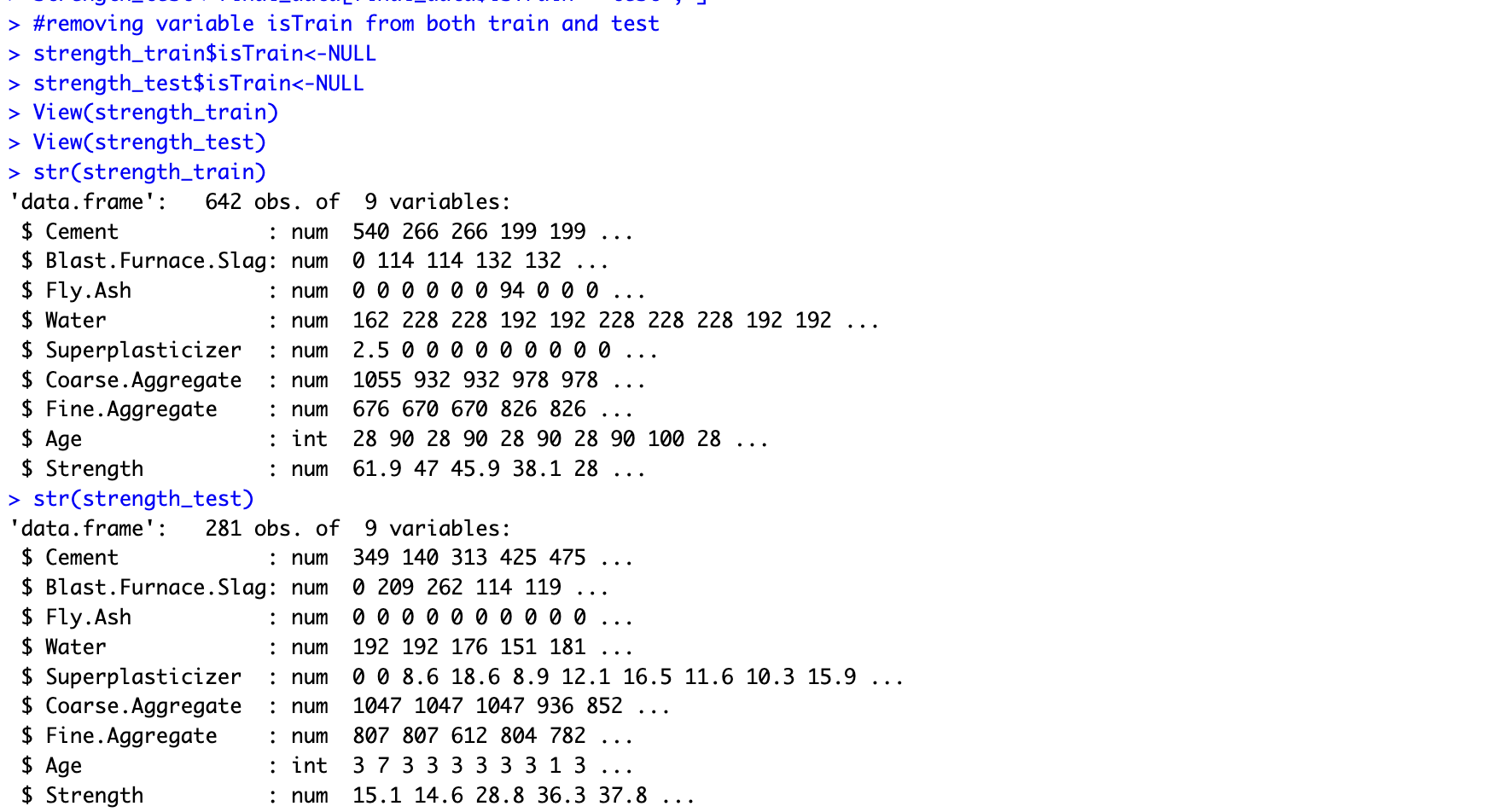


Figure 1.2 Structure of Train and Test Datasets after removing is\_Train

1. **Linear Regression**

The value of an independent variable is used to predict the value of a dependent variable using linear regression analysis. A straight line or surface that minimises the differences between the expected and actual output values is fitted using linear regression.

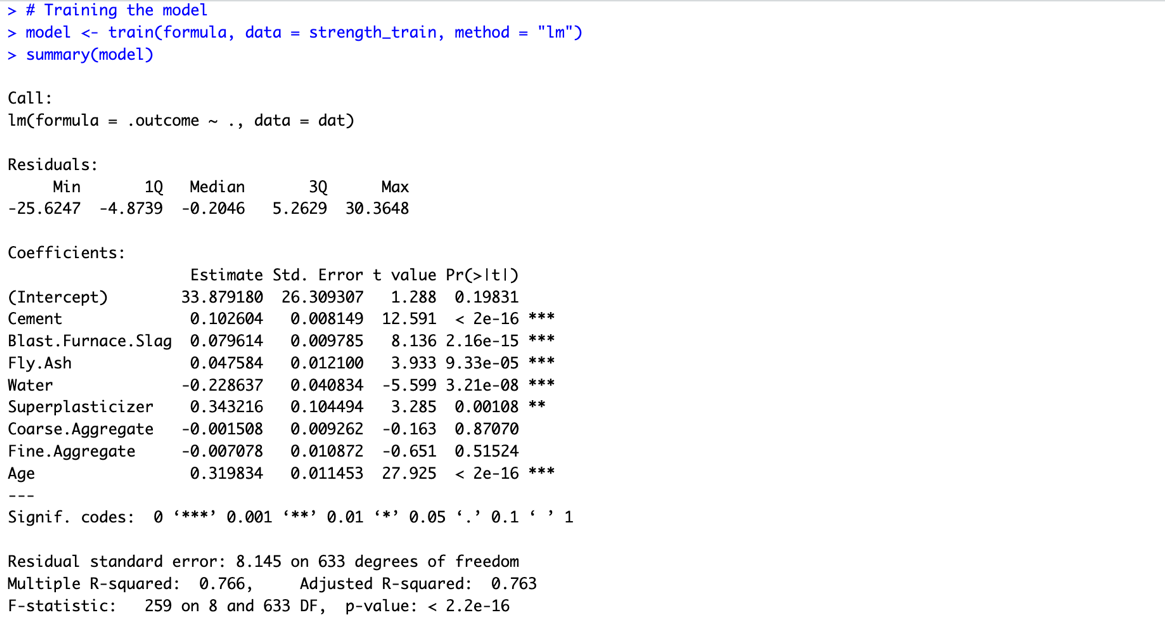


Figure 1.1.1 Summary

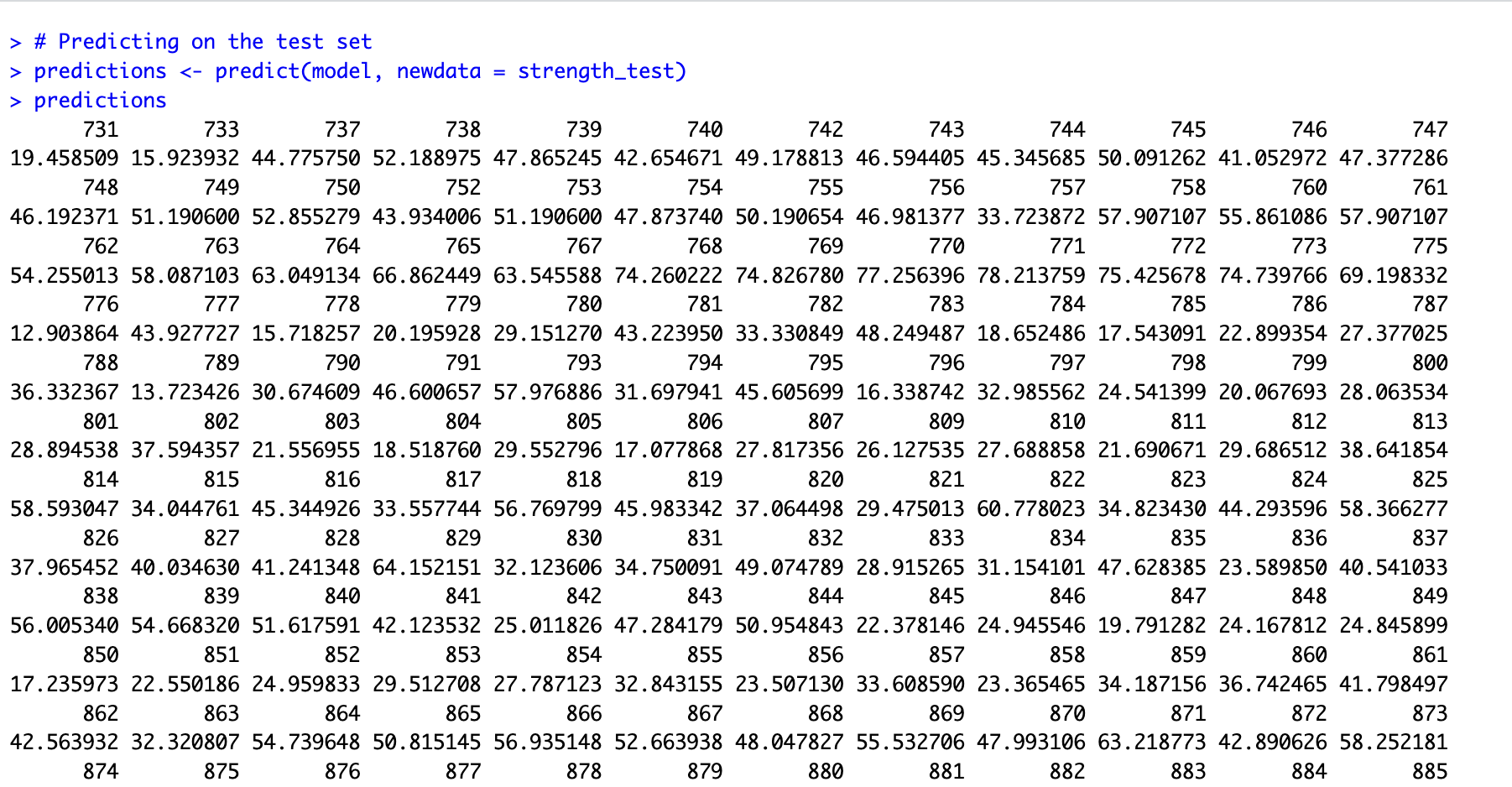


Figure 1.1.2 Predictions of Linear Regression

Above Figure 1.1.1 shows the summary of the linear regression model and Figure 1.1.2 shows the prediction made by linear regression model in test dataset.

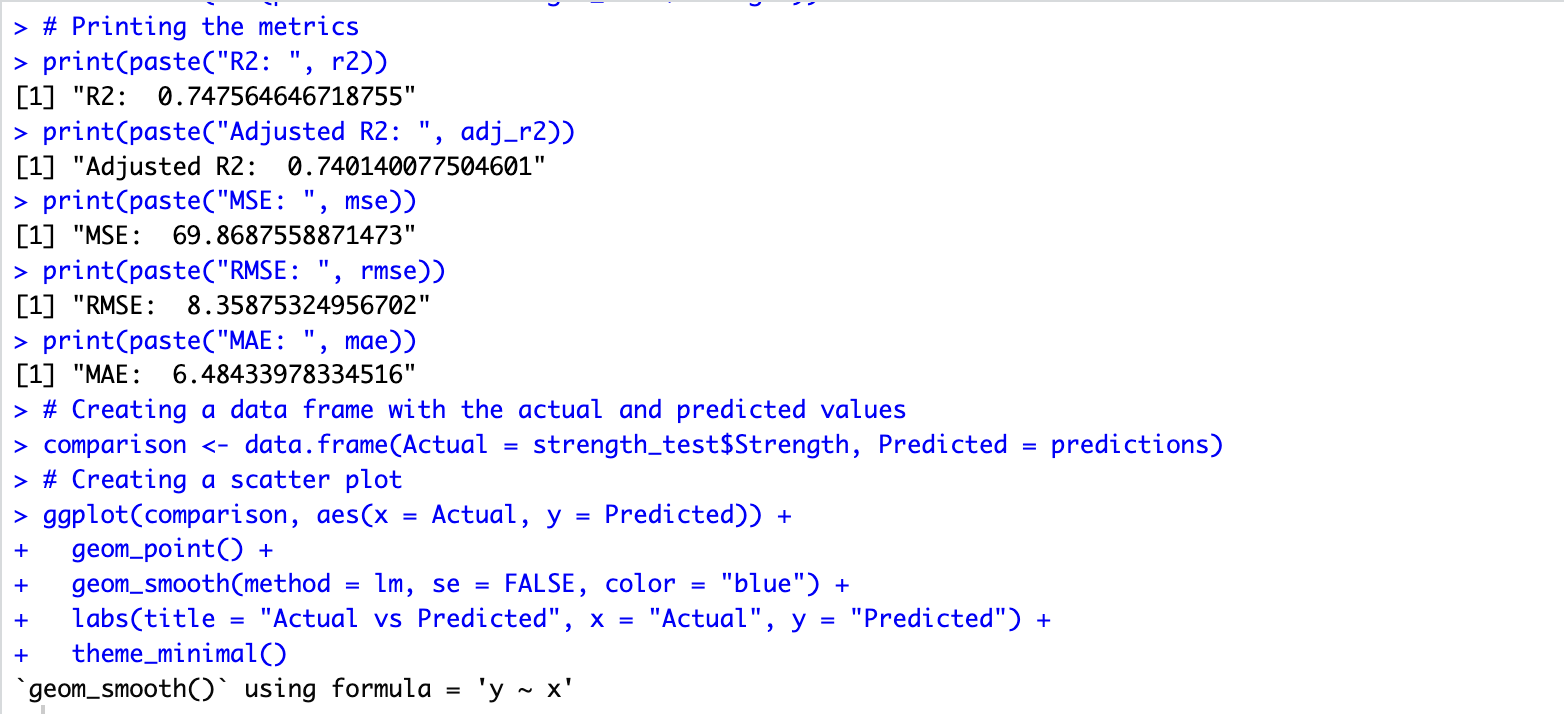
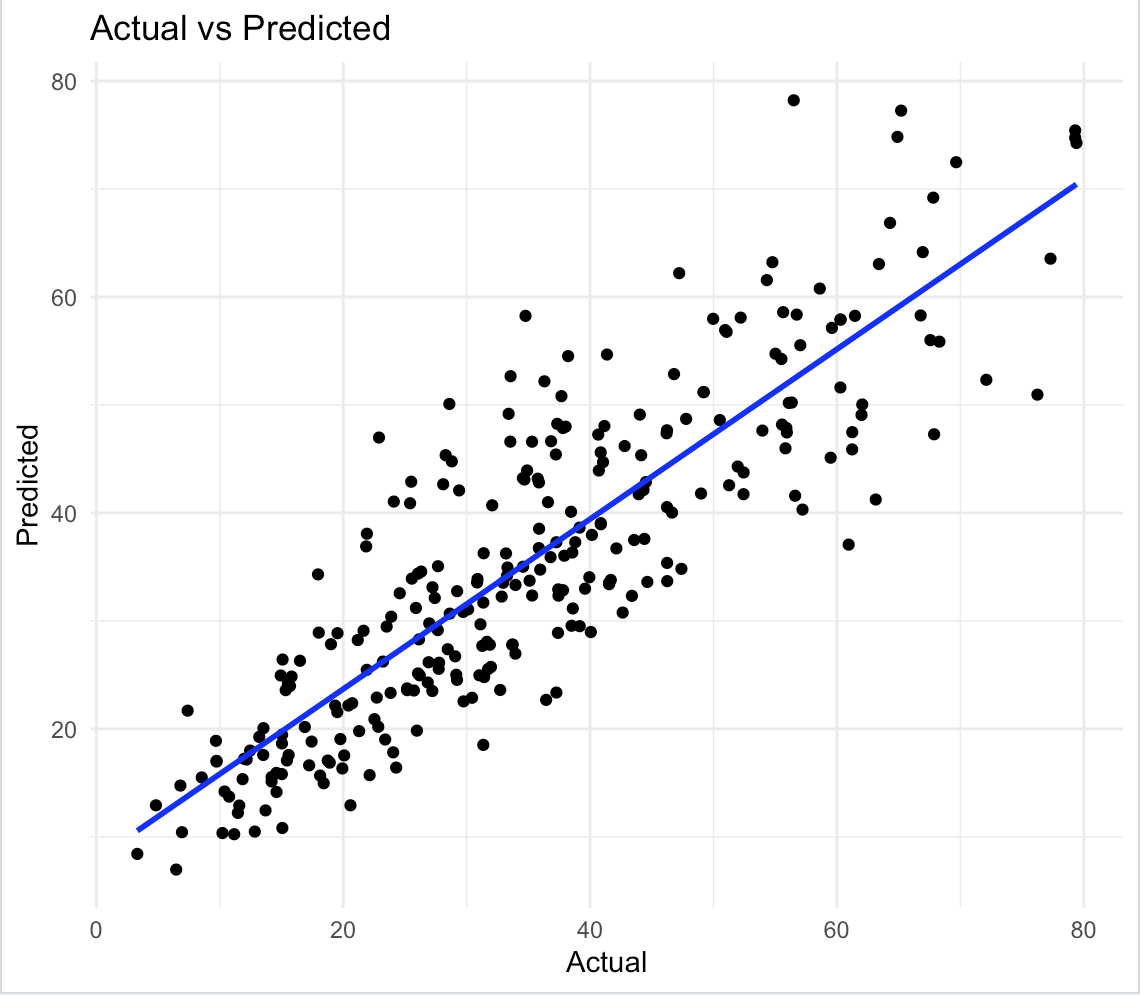


Figure 1.1.3 Evaluation Metrics

The value for evaluating the metrics is printed in the above fig.1.1.3. Different performance metrics: R-squared (R2), Adjusted R-squared (Adj. R2), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) values are obtained for linear regression model.



*Figure 1.1.4 Scatter Plot for Linear Regression with actual vs predicted values*

The predicted values of a linear regression model is shown against the actual values in the above scatter plot.

1. **Random Forest**

A popular machine learning approach called Random Forest mixes the outputs of several decision trees to get a single outcome. Its versatility and ease of use, combined with its ability to handle both regression and classification issues, makes it more flexible to use.

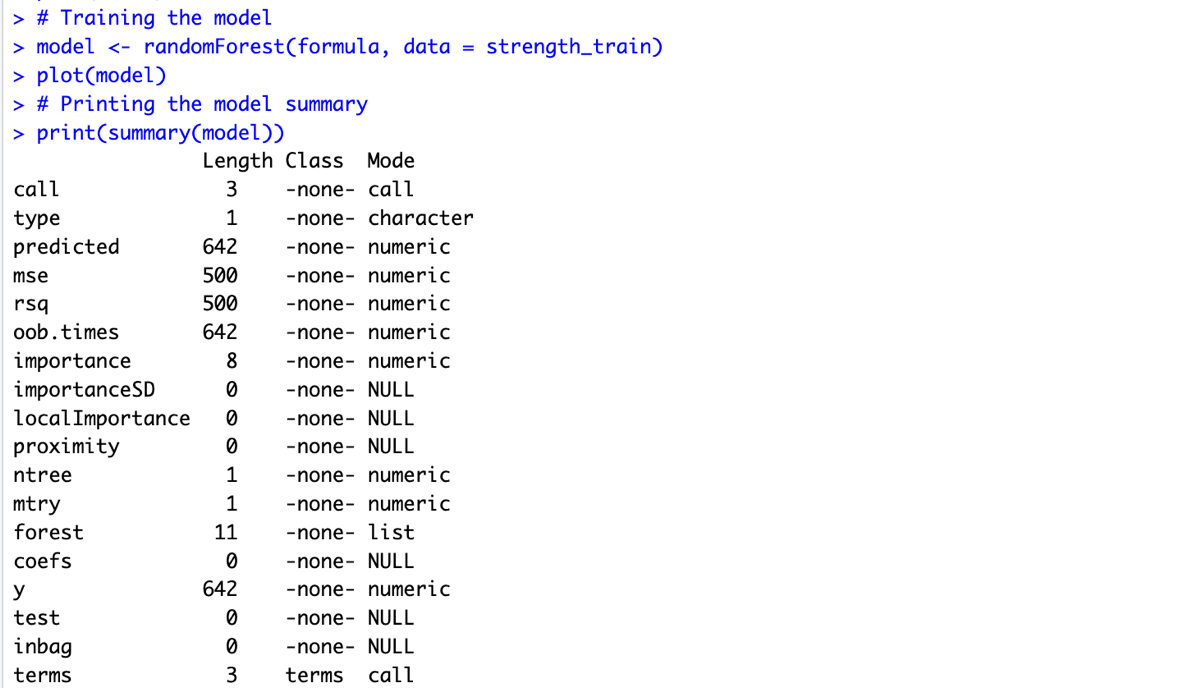


Figure 1.2.1 Summary of Random Forest

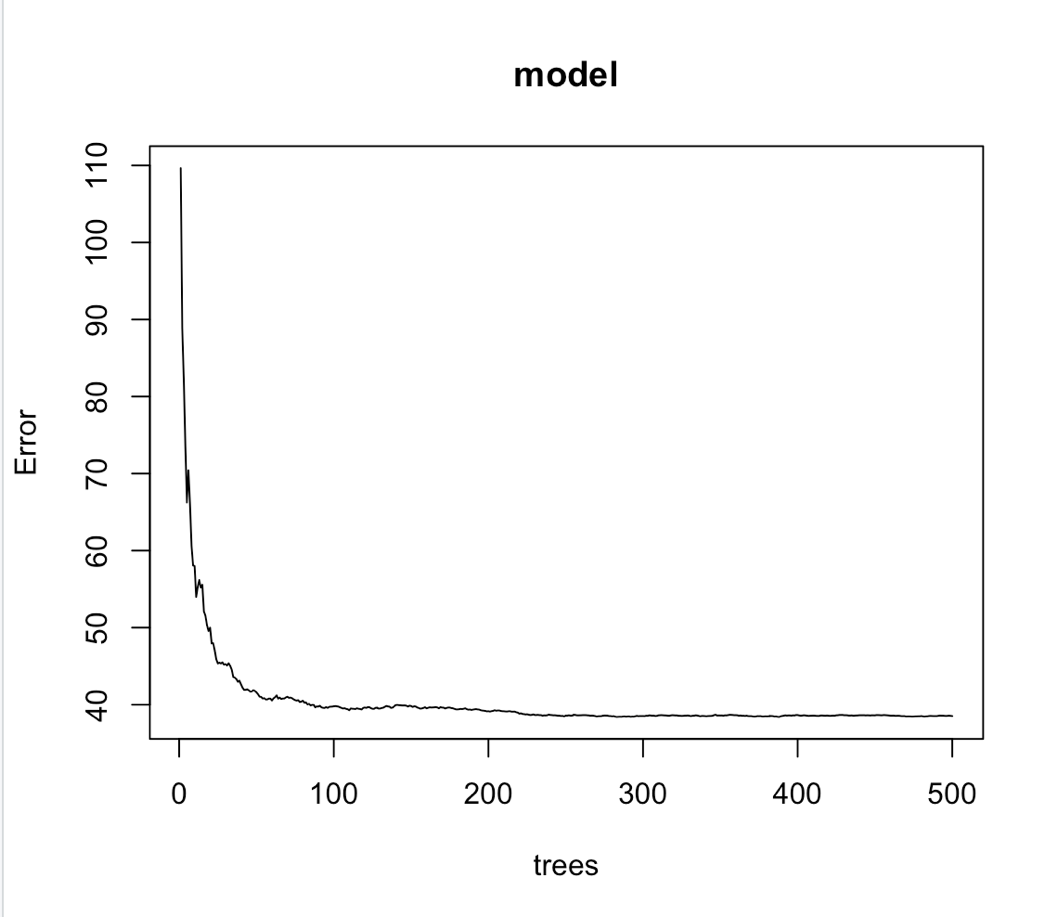


Figure 1.2.2 Plotting the Model value

The model value after using randomForest() on Strength\_Train data is plotted and is shown in fig.1.2.2, where the plot of the errors in y-axis is from 0-110 as the number of trees rises from 0 to 500 in x-axis using a random forest model trained on train data.

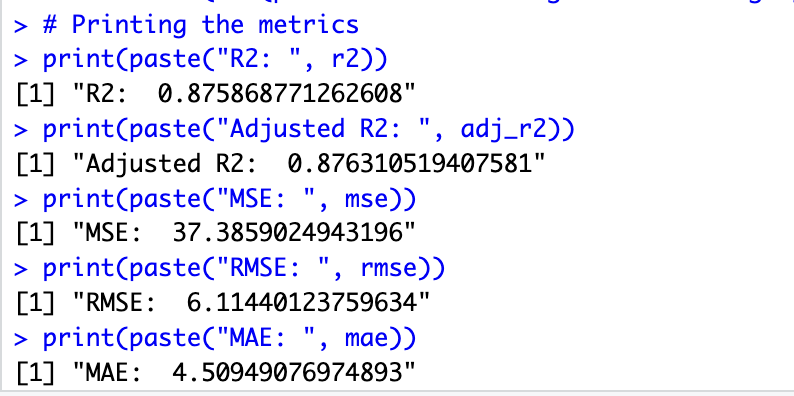
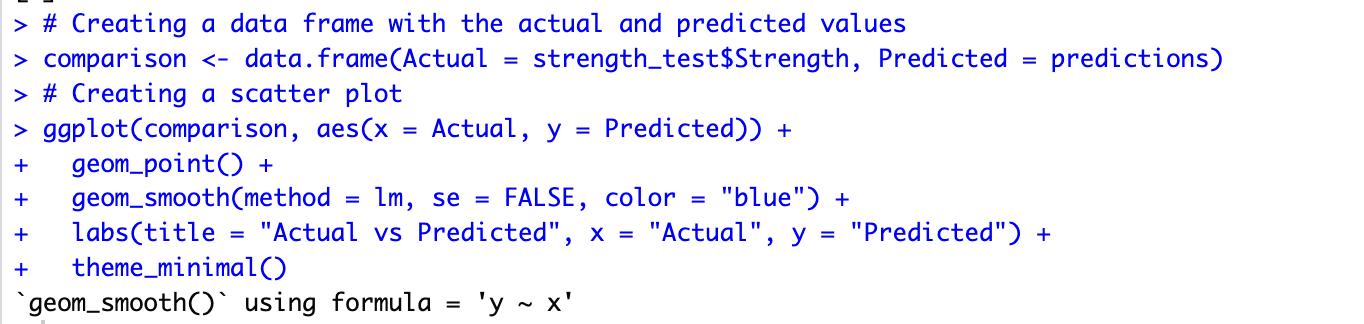
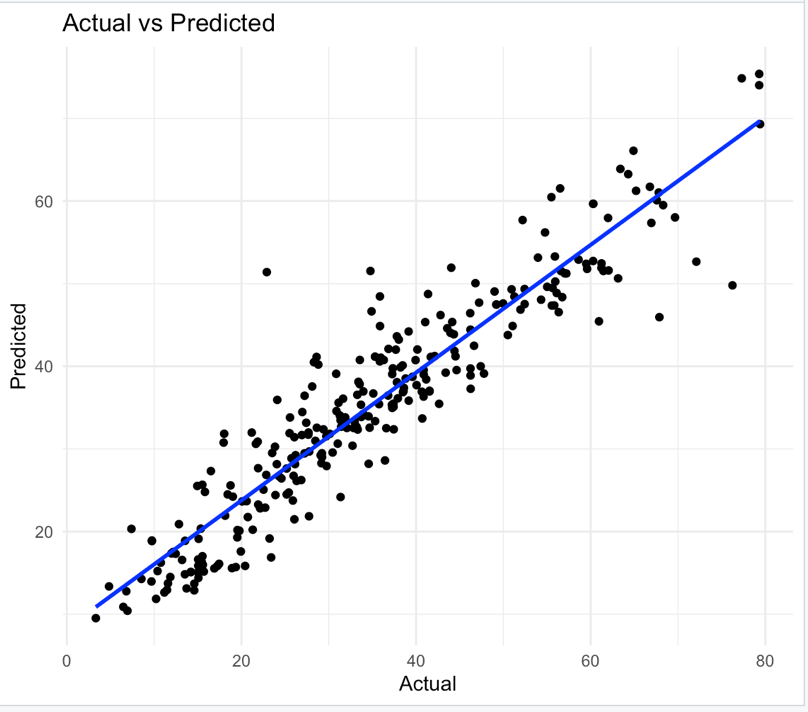


Figure 1.2.3 Evaluation Metrics

The value for evaluating the metrics is printed in the above fig.1.2.3. Different performance metrics: R-squared (R2), Adjusted R-squared (Adj. R2), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) values are obtained for Random Forest model.



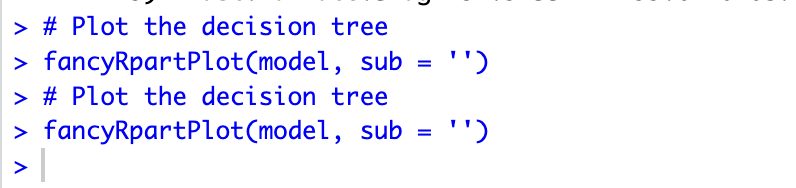


*Figure 1.2.4 Scatter Plot for Random Forest with actual vs predicted values*

The predicted of a *Random Forest* model is shown against the actual values in the above scatter plot. Here we can see the data seem to be more closely packed around the diagonal line than in Linear Regression Model suggesting that the *Random Forest* model would offer a much better fit to the data.

1. **Decision Tree**

In a decision tree, each attribute test is represented by each internal node, the test result is represented by each branch, and the class label is held by each leaf node, or terminal node, which resembles a flowchart. The decision tree works by splitting the data into smaller and smaller subsets based on certain features.



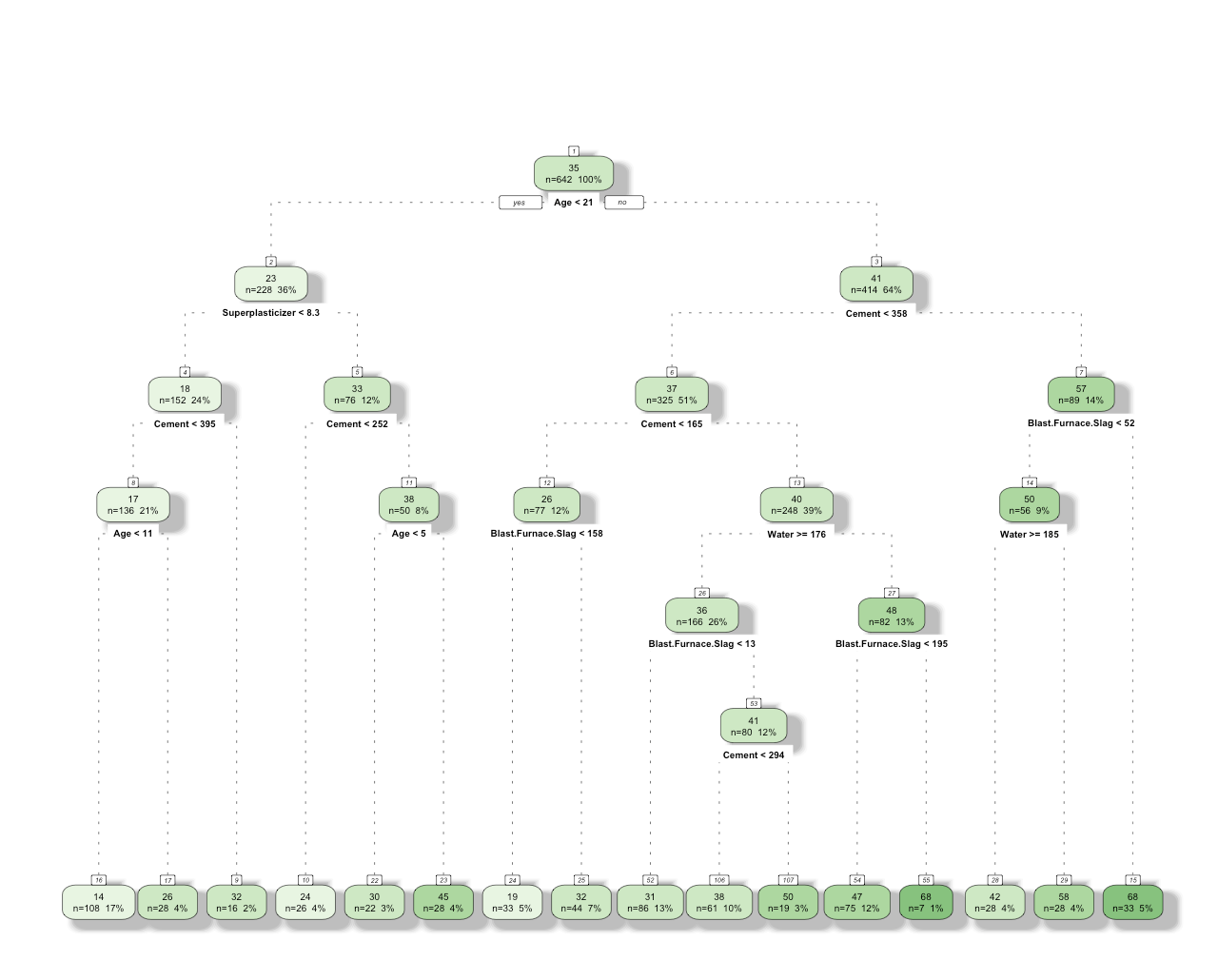


Figure 1.3.1 Decision Tree Model Plot

In fig.1.3.1, we can see a decision tree visualisation produced by using fancyRpartPlot() function to forecast the target variable. The tree divides the data into multiple categories trying to predict the compressive strength of concrete. Initially, the decision tree divides the data into two categories according on whether the concrete is older than 11 years. Next, it divides the data once again according to whether or not there is less than 158 blast furnace slag. The procedure carries on until the decision tree reaches a leaf node, which has a forecast regarding the concrete's compressive strength.

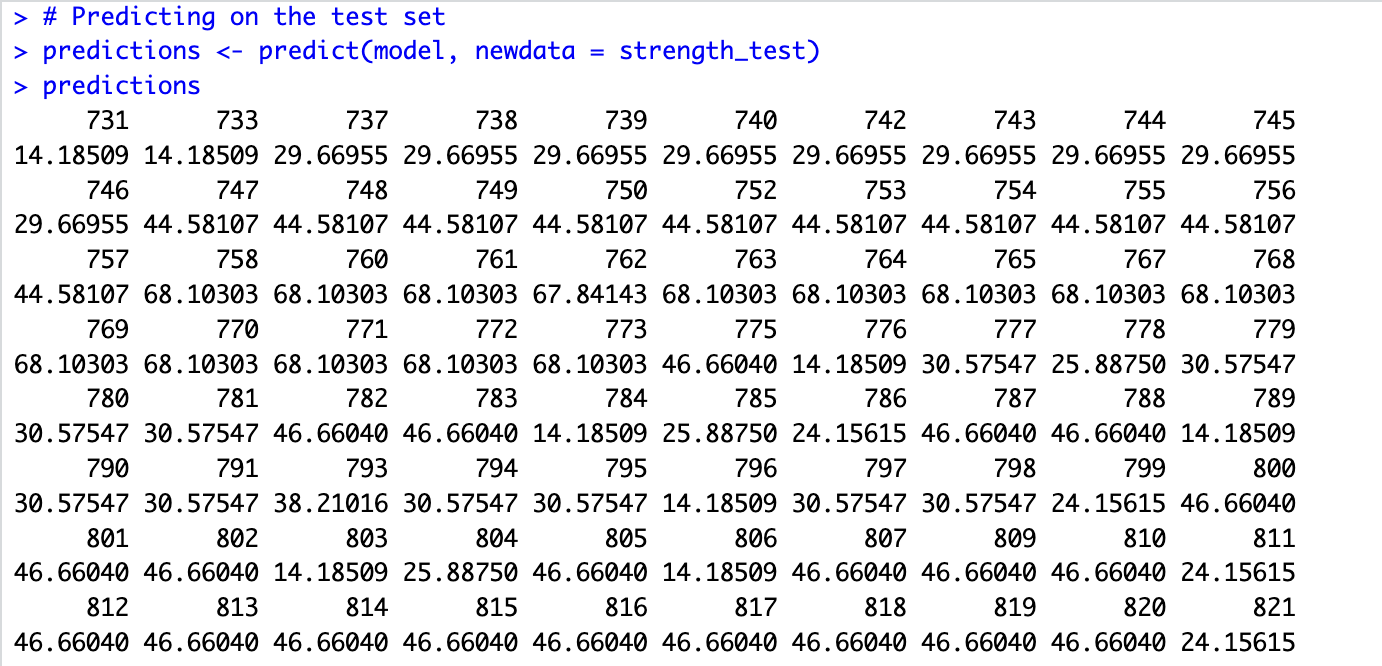


Figure 1.3.2 Predictions of test dataset.

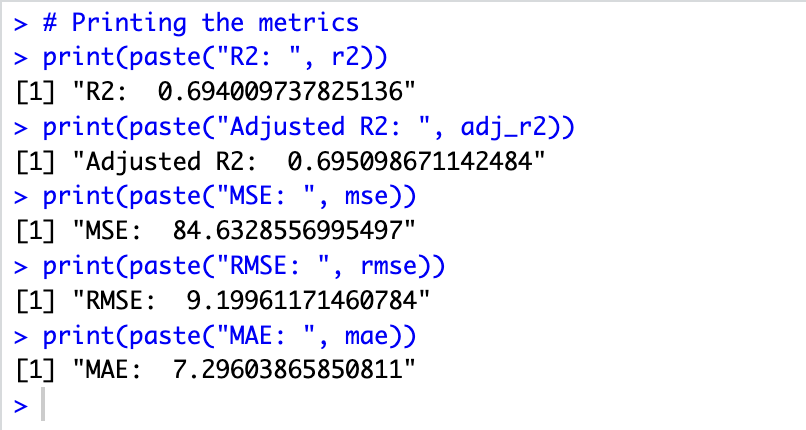
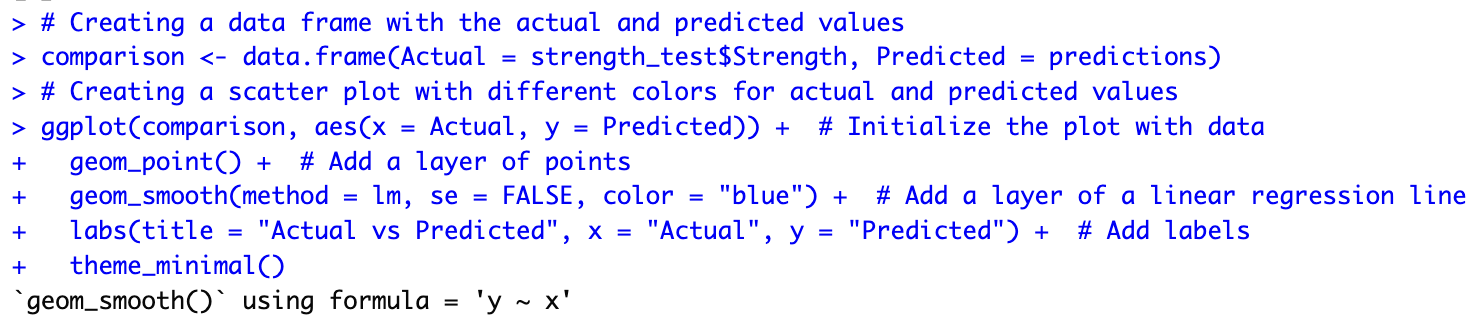


Figure 1.3.3 Evaluation Metrics

In the above fig.1.3.3, the data probability is displayed for each decision: R2, Adjusted R2, MSE, RMSE, and MAE.



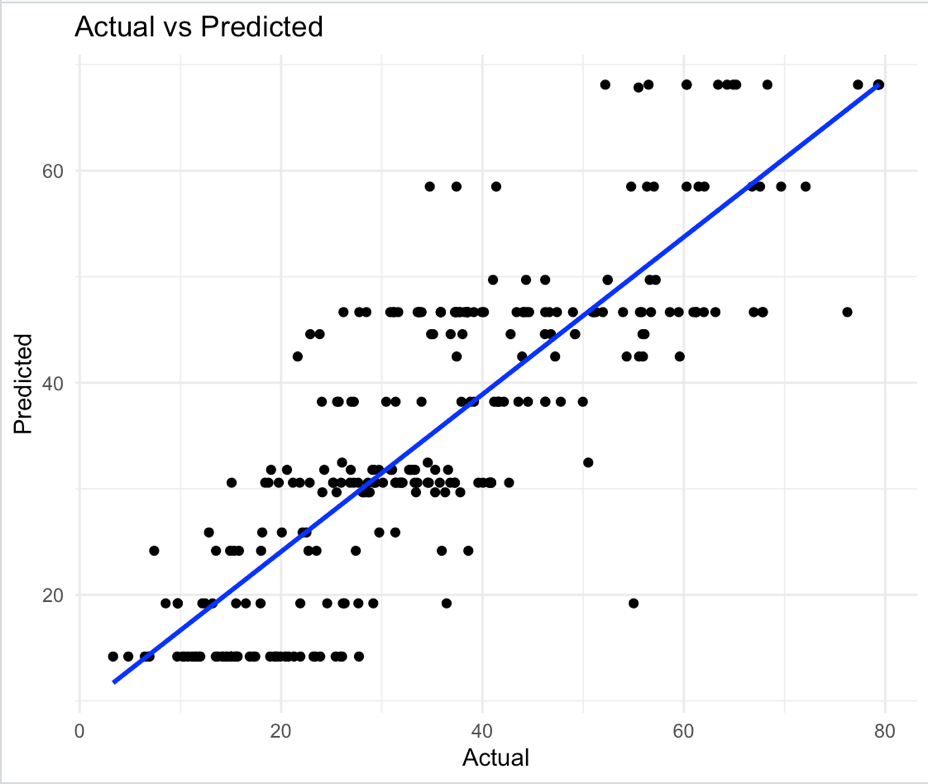


Figure 1.3.4 Scatter Plot for Decision Tree with actual vs predicted values

In the fig.1.3.4, scatter plot for Decision Tree with actual vs predicted values is shown to evaluate the model's accuracy and decision-making abilities.

1. **k-Nearest Neighbours (KNN)**

KNN is a member of the supervised learning domain. It is simple to grasp, nonparametric as it does not make any underlying assumptions about the distribution of the data and lazy. However, for huge datasets, it may be computationally costly.

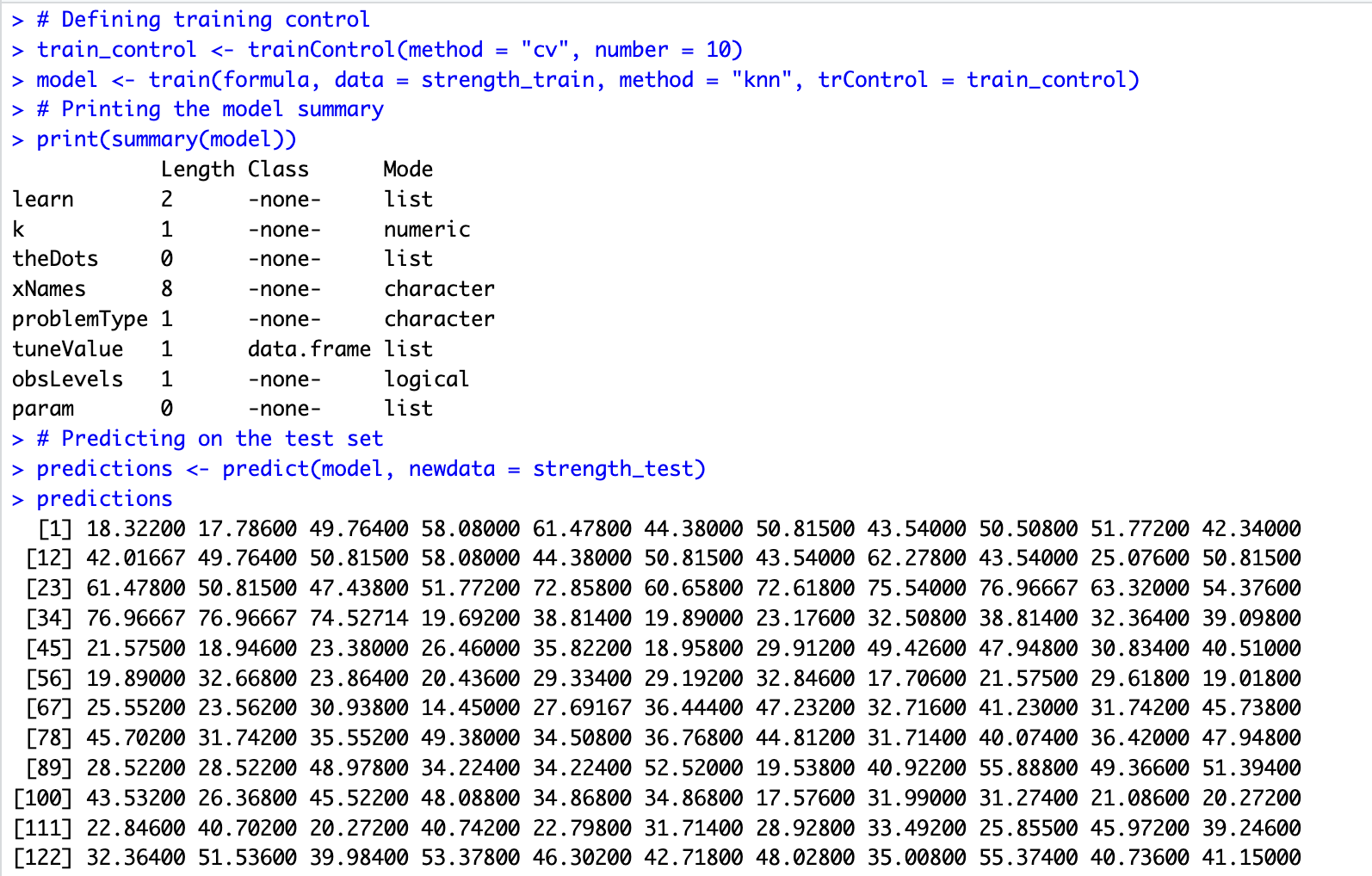


Figure 1.4.1 Summary and Predictions

Above fig.1.4.1 prints the summary of the KNN model and gives the prediction on the test set data.

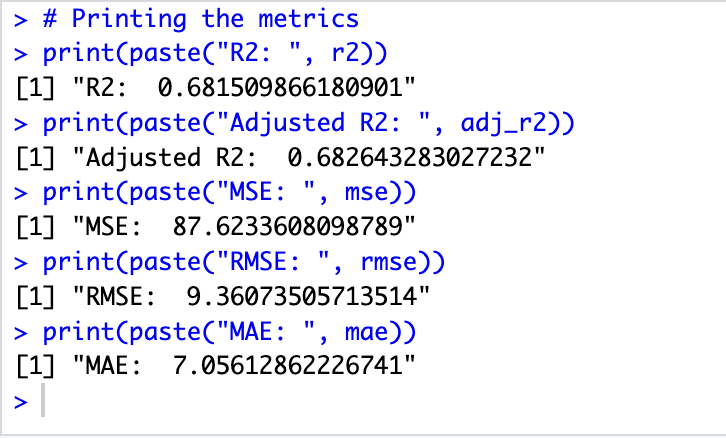
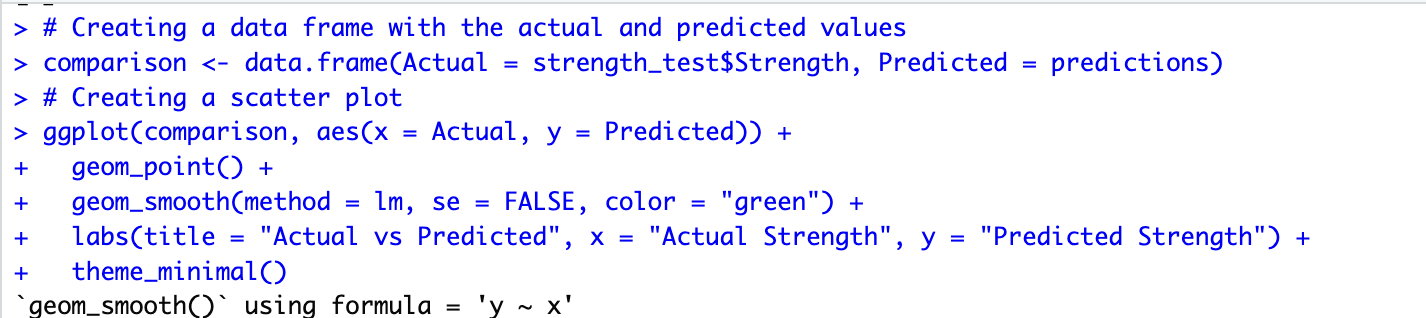


Figure 1.4.2 Evaluation Metrics on

The value for evaluating the metrics is printed in the above fig.1.2.3. Different performance metrics: R-squared (R2), Adjusted R-squared (Adj. R2), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) values are obtained for KNN model.



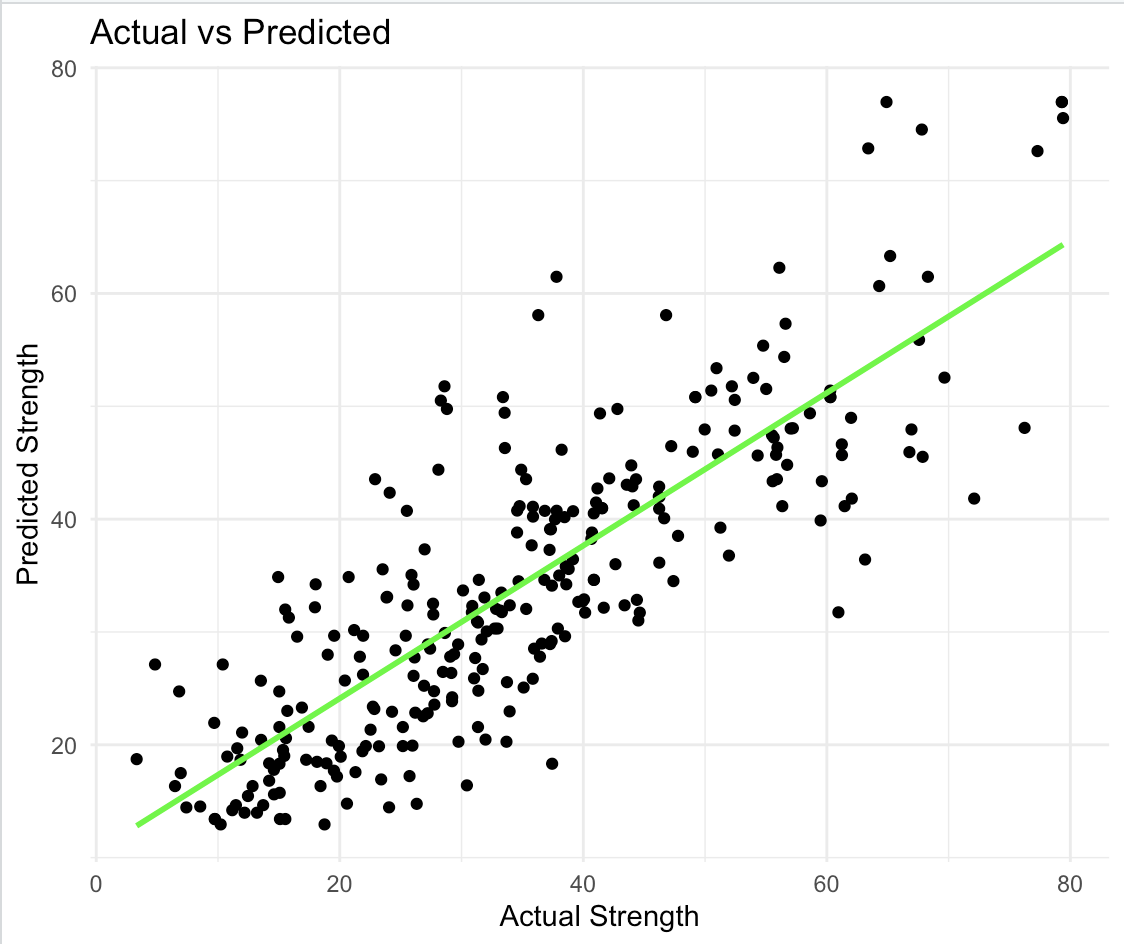


Figure 1.4.3 Scatter Plot for KNN with actual vs predicted strength

The above scatter plot showing the relationship between actual strength on the x-axis and predicted strength on the y-axis.

1. **KNN Model with optimal value of k = 4**

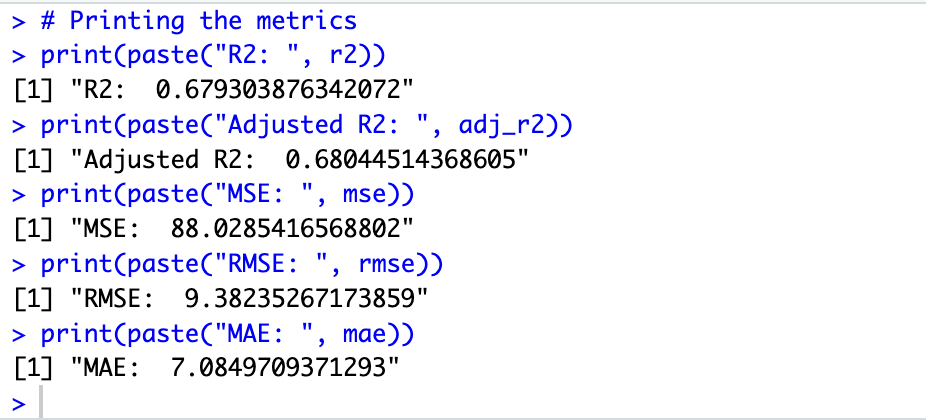
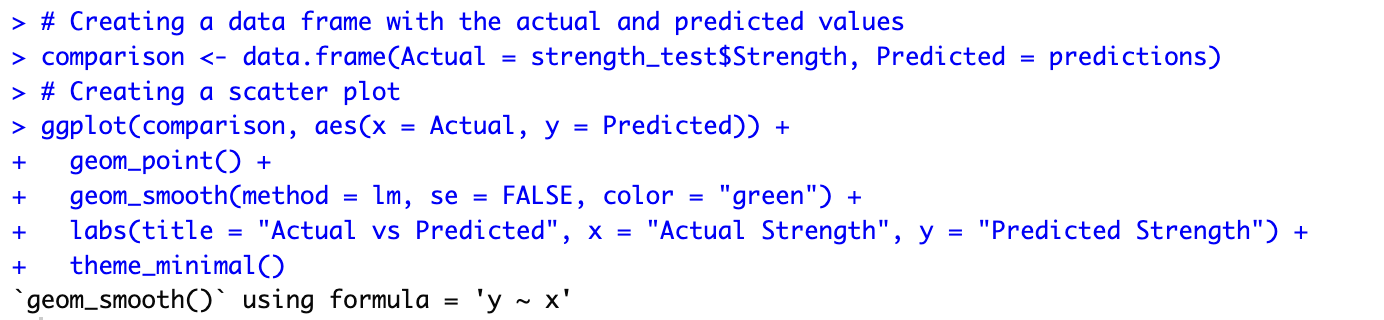
****

Figure 1.5.1 Evaluation Metrics



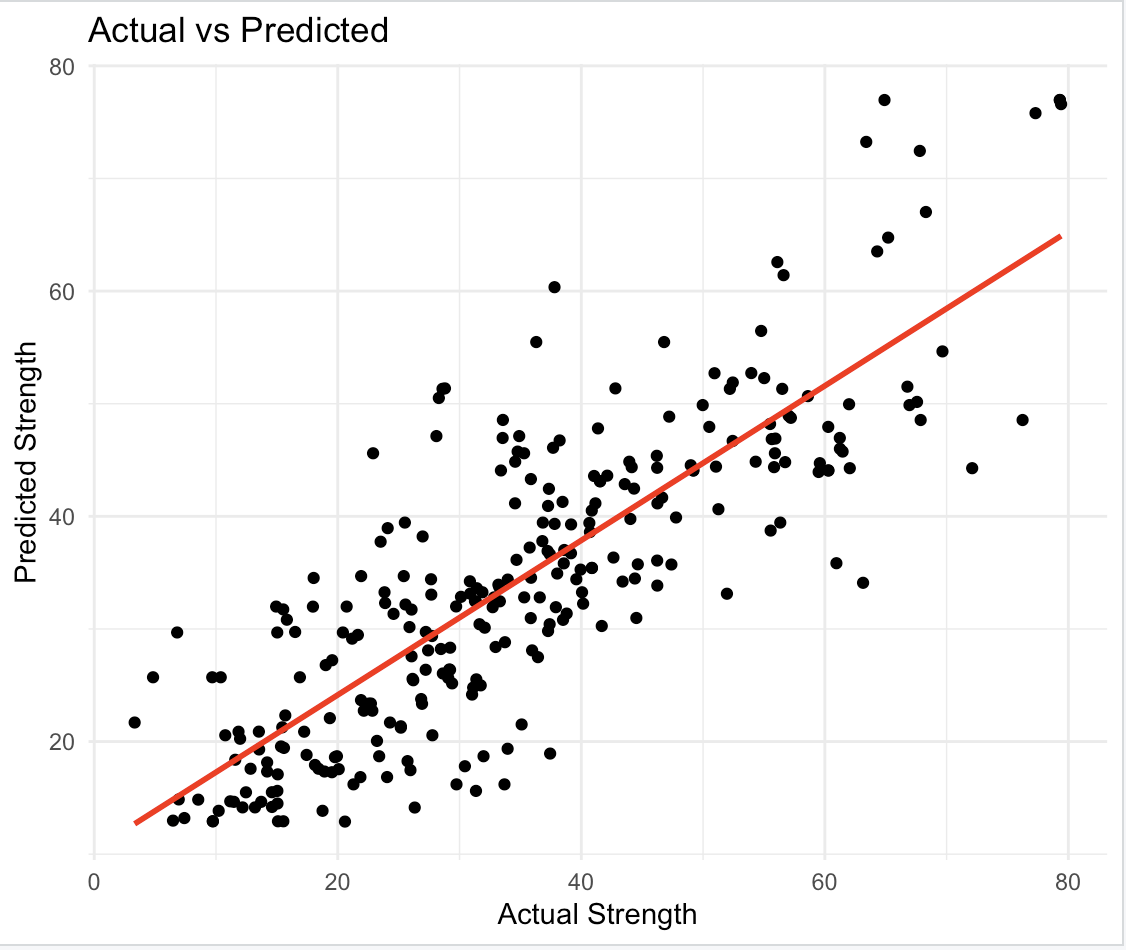


Figure 1.5.2 Scatter Plot for KNN=4 with actual vs predicted values

The above scatter plot showing the relationship between actual strength on the x-axis and predicted strength) on the y-axis with different KNN value i.e. KNN=4.

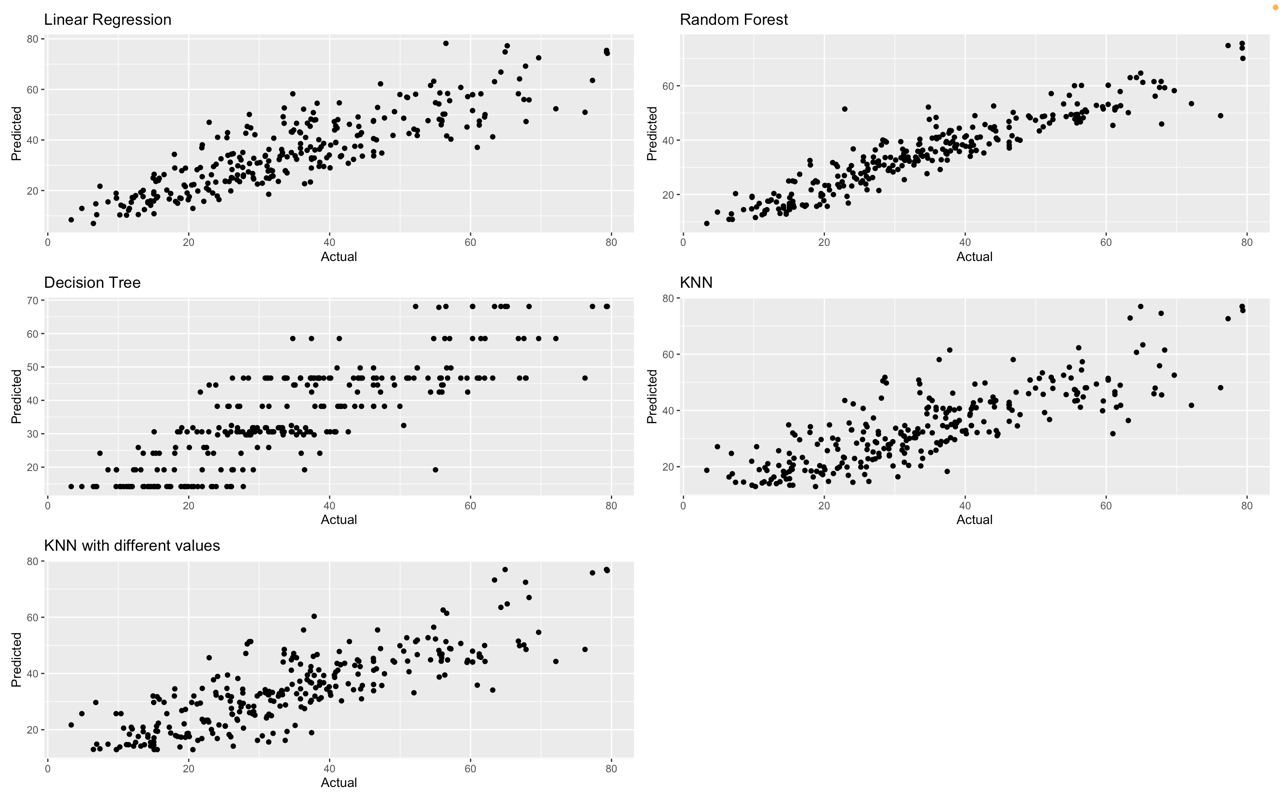


Figure 1.6 Scatter Plot for all five Models

The scatter plot for all five different models that were developed is shown to compare in the above fig.1.6. The scatter plot of the Random Forest and Linear Regression seems to be much more compacted and fit for the data than other models.

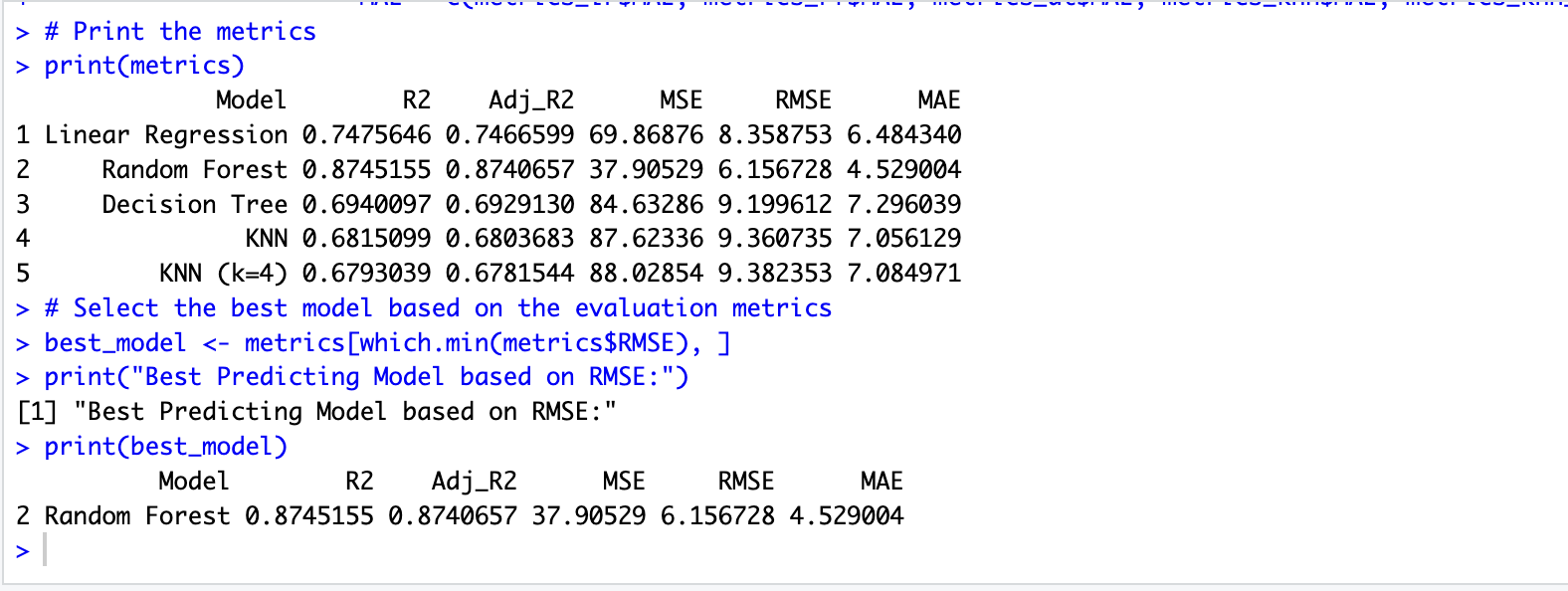


Figure 1.7 Finding the Best Model

Above, the metrics for all five models are printed and the best predicting model based on evaluation metrics is shown in fig.1.7.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **R2** | **Adj. R2** | **MSE** | **RMSE** | **MAE** |
| LR | 0.7474646 | 0.7401400 | 69.8687558 | 8.3587532 | 6.4843397 |
| RF | 0.8758687 | 0.8763105 | 37.3859024 | 6.1144012 | 4.5094907 |
| DECISION | 0.6940097 | 0.6950986 | 84.6328556 | 9.1996117 | 7.2960386 |
| KNN | 0.6815098 | 0.6826432 | 87.6233608 | 9.3607350 | 7.0561286 |
| KNN=4 | 0.6793038 | 0.6804451 | 88.0285416 | 9.3823526 | 7.0849709 |

An overview of the many regression models assessed using a range of performance metrics is given in this table which are: R-squared (R2), Adjusted R-squared (Adj. R2), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The five models which were assessed are: Linear Regression Model, Random Forest Model, k-Nearest Neighbours Model (KNN), Decision Tree Model, and a Modified KNN model (KNN=4). We may compare and choose the best model for the task at hand by observing each model's performance across these measures based on this summary.

**Model Interpretation**

We can observe that, in comparison to the other models, the Random Forest (RF) model has the highest R-squared (0.8758687) and the lowest values for MSE, RMSE, and MAE indicating a superior fit to the data with lower error and high accuracy rate. When extracting and plotting the importance of random forest in the code we can see the following:

|  |  |
| --- | --- |
|  | IncNodePurity |
| Blast.Furnace.Slag | 11066.801 |
| Water | 24434.640 |
| Coarse.Aggregate | 10528.946 |
| Age | 50511.355 |
| Cement | 34792.586 |
| Fly.Ash | 9079.506 |
| Superplasticizer | 17187.622 |
| Fine.Aggregate | 12678.759 |

Based on the test results, the Random Forest model seems to be the most accurate predictor. A more thorough understanding of the variables affecting concrete strength may result from this knowledge, which might enhance decision-making and optimise the design and formulation of concrete mixes.

**Conclusions**

In conclusion, Among the five different models that was used in the regression problems, the most accurate one is the Random Forest as the model is consistent. In case of the regression model, MAE, MSE, EMSE, R2 and adjusted R2 we, R2 gave more variance between predictor variable and the target variable. The variance measured by R2 is the ratio of two different variables one being dependent variable which is predictable from the independent variable. Evaluation metrices help find or explain the variability in the data.

**Bibliography**

* Ibm.com. (2024). *What Is Linear Regression? | IBM*. [online] Available at: https://www.ibm.com/topics/linear-regression#:~:text=IBM-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable. [Accessed 10 May 2024].
* Ibm.com. (2024). *What Is Random Forest? | IBM*. [online] Available at: https://www.ibm.com/topics/random-forest [Accessed 10 May 2024].
* GeeksforGeeks (2017). *Decision Tree*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/decision-tree/ [Accessed 10 May 2024].

‌

‌

‌