School of Computing, Creative Technology and Engineering

Student ID	77356785
Student Name	Pooja Pantha
Module Name & CRN	Applied Machine Learning 18909
Level	5
Assessment Name & Part No.	Coursework - II
Project Title	Case Study on Concrete Strength Analysis
Data of Submission	10/05/2024
Course	Computing
Academic Year	2024

Table of Contents

PART II 3

Splitting the Concrete data 3

Modelling 3

Linear Regression: 3

K-Nearest Neighbour (KNN): 4

Decision Tree: 5

Random Forest: 6

K-Nearest Neighbour with different parameters: 7

Results for all the regression models 7

Model Interpretation 8

Conclusion 10

Bibliography 11

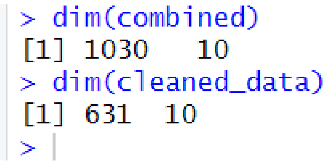
Table of Figures

Figure 1: Plot for Linear Regression 4
Figure 2: Plot for Linear Regression 4
Figure 3: Plot for KNN 5
Figure 4: Plot for KNN 5
Figure 5: Plot for Decision Tree 5
Figure 6: Plot for Decision Tree 6
Figure 7: Plot for Random Forest 6
Figure 8: Plot for Random Forest 6
Figure 9: Plot for KNN with different Parameters 7
Figure 10: Plot for KNN with different Parameters 7
Figure 11: Variable Importance in Concrete Strength based on Random Forest Model 9

PART II

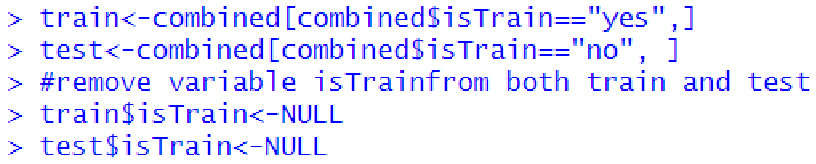
Case Study on Concrete Strength Analysis
Pooja Pantha, Computing 2024

The dataset has 1030 rows and 10 columns after combining data at the first instance. After preprocessing, the dataset consists of 631 rows and 10 columns as data was cleansed. The preprocessing steps involved identifying variables with missing values and imputing them using the mice() function. The author examined column missingness to determine variables with missing values. The average percentage of missing values in the dataset was calculated to be 3.87%.



```
> dim(combined)
[1] 1030   10
> dim(cleaned_data)
[1] 631  10
> 
```

Splitting the Concrete data



```
> train<-combined[combined$isTrain=="yes",]
> test<-combined[combined$isTrain=="no", ]
> #remove variable isTrainfrom both train and test
> train$isTrain<-NULL
> test$isTrain<-NULL
```

Modelling

The author proceeds on to the development of predictive models for evaluating concrete strength after completing the exploratory data analysis (EDA) stage. The main goal is to develop a model that, given the input features (cement content, blast furnace slag, etc.), accurately forecasts the strength of concrete. Both Type 1 (false positives) and Type 2 (false negatives) mistakes should be reduced by this methodology. Five different models—Linear Regression, Decision Tree, Random Forest, K-Nearest Neighbor (KNN), and KNN with varying parameters—will be investigated and put into action. The author seeks to determine which model performs best for predictive task by conducting thorough parameter tuning, cross-validation, and performance evaluation.

The residuals vs. observed and anticipated vs. observed are the two crucial diagnostic graphs that the author then created. By highlighting possible trends in the errors and the distribution of projected values in contrast to the actual observations, these visualizations would aid in evaluating the effectiveness of the model.

Linear Regression:

Linear regression establishes a linear relationship between features and the target variable, making it interpretable but potentially limited in capturing complex patterns.

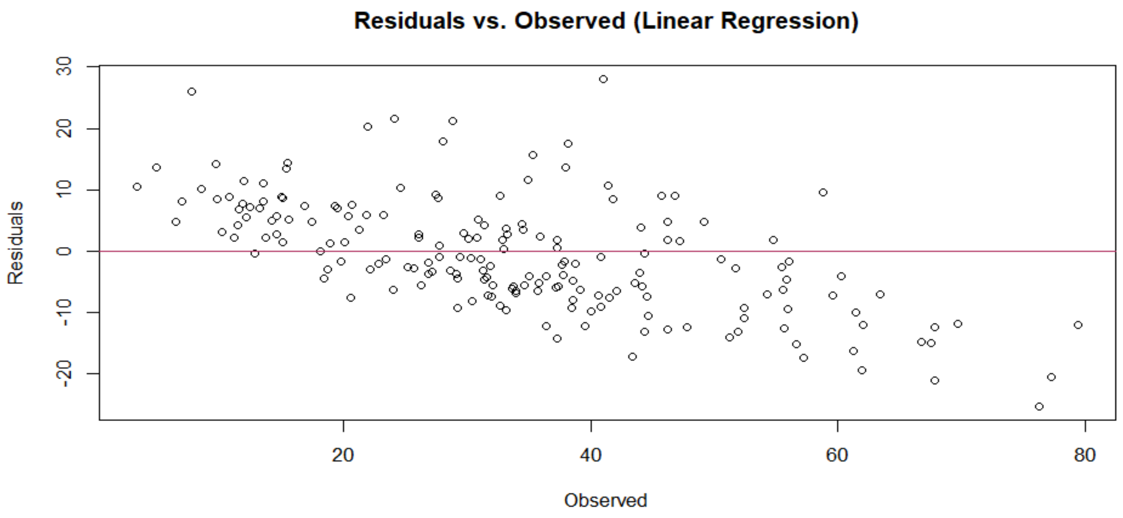


Figure 1: Plot for Linear Regression

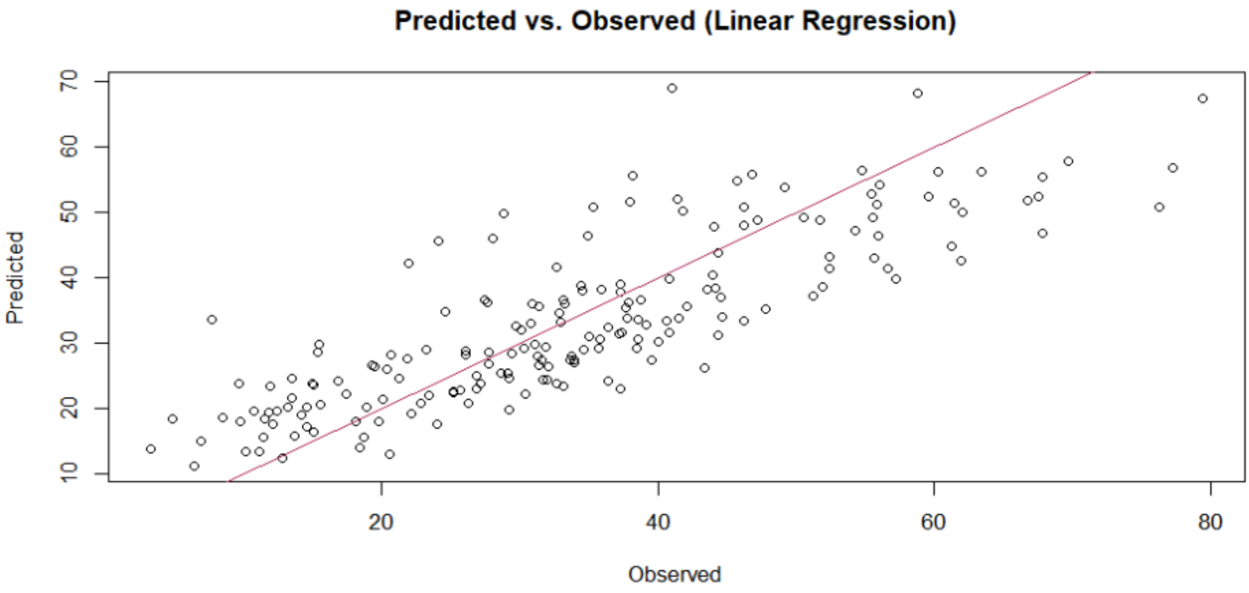


Figure 2: Plot for Linear Regression

K-Nearest Neighbour (KNN):

It determines which k training set data points are most similar to a new data point, and then uses the mean value (regression) of those neighbors to forecast the value of the target variable. The smoothness and noise sensitivity of the model can be changed by adjusting the k parameter.

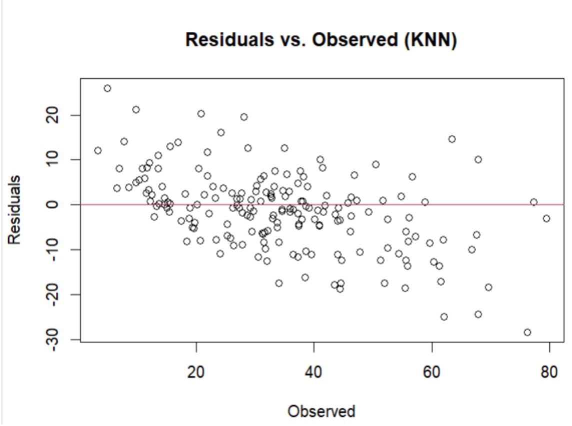


Figure 3: Plot for KNN

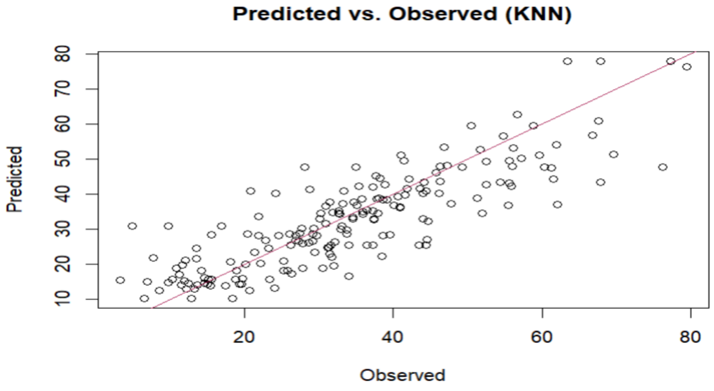


Figure 4: Plot for KNN

Decision Tree:

It provides interpretability and a predictability structure like to a flowchart by dividing the data into groups based on a sequence of yes/no questions regarding features.

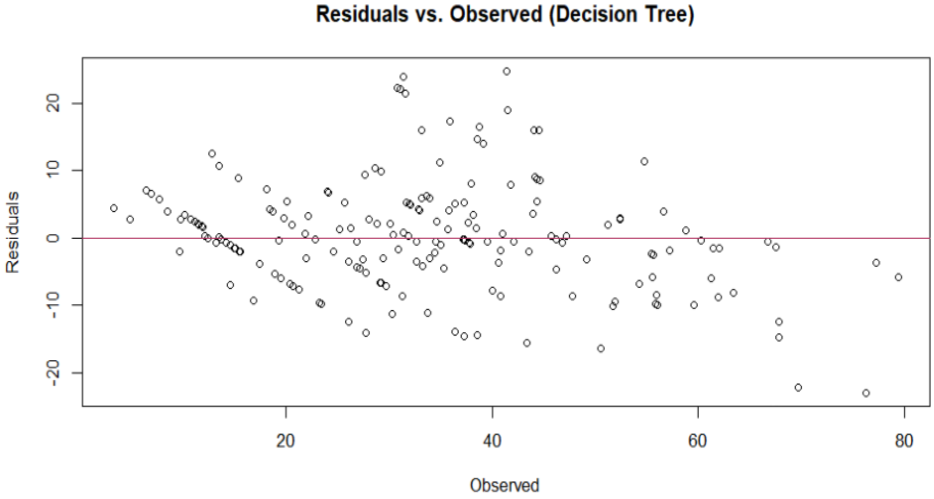


Figure 5: Plot for Decision Tree

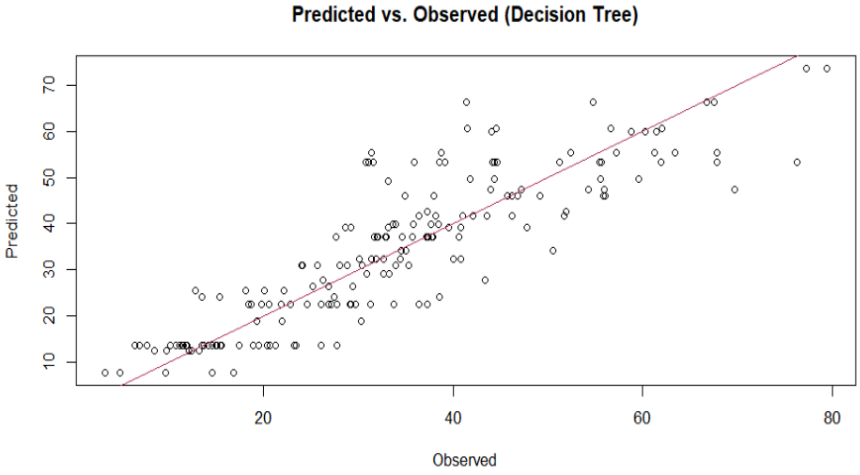


Figure 6: Plot for Decision Tree

Random Forest:

Several decision trees are combined in a random forest, leading to improved accuracy and robustness by averaging predictions from each tree, making it less prone to overfitting.

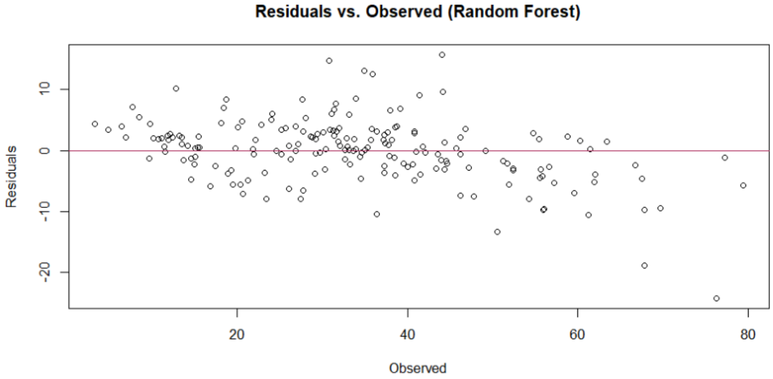


Figure 7: Plot for Random Forest

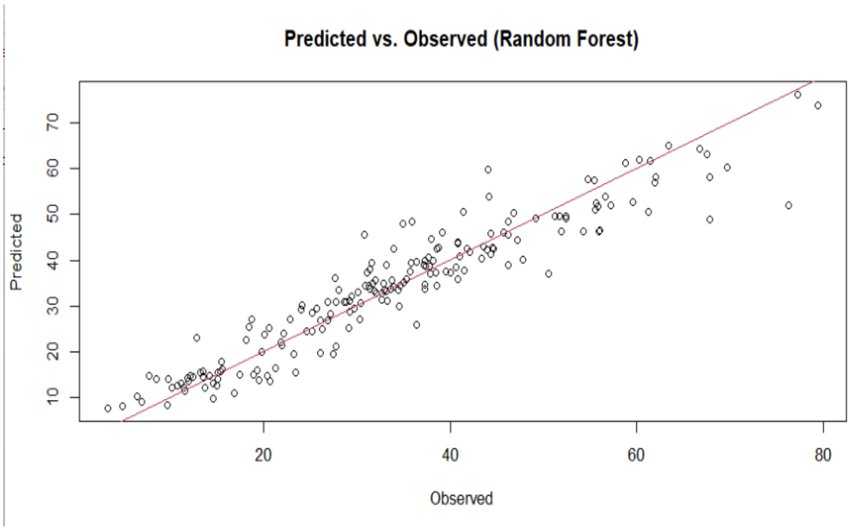


Figure 8: Plot for Random Forest

K-Nearest Neighbour with different parameters:

Though this model is similar to a regular KNN, it investigates how different K values affect the overall performance. The K value that produces the most favorable performance metrics is chosen after testing a few different values.

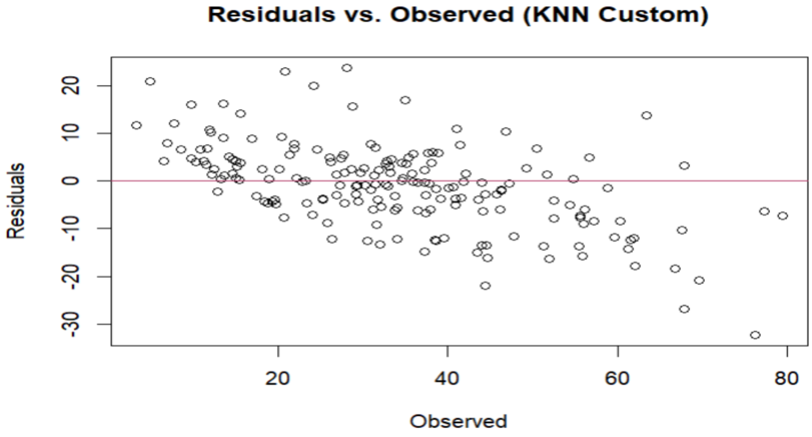


Figure 9: Plot for KNN with different Parameters

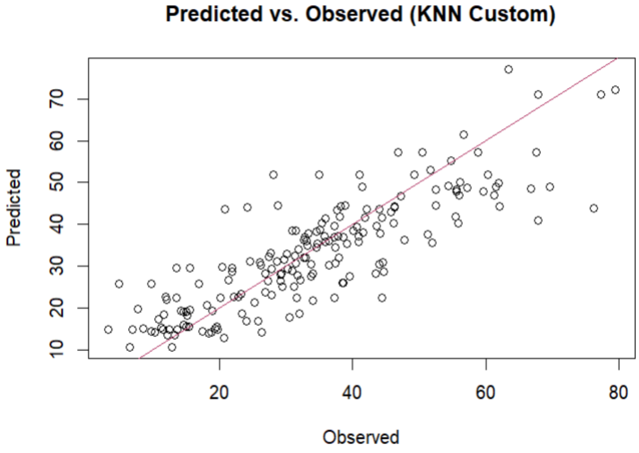
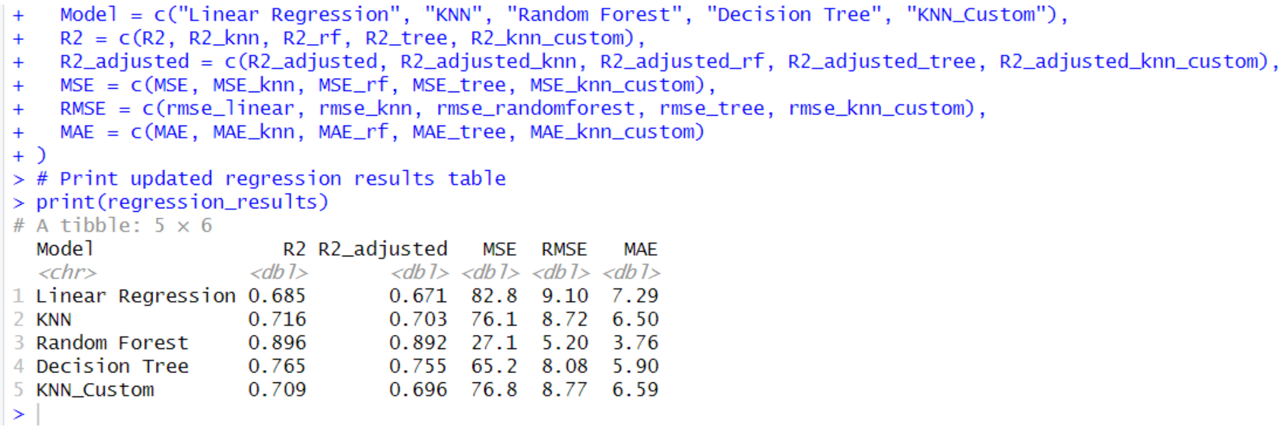


Figure 10: Plot for KNN with different Parameters

Results for all the regression models

The author assessed how well five potential models performed in terms of predicting the strength of concrete. To identify the best model for precise concrete strength prediction, the findings will be compared and reported.



```
+   Model = c("Linear Regression", "KNN", "Random Forest", "Decision Tree", "KNN_Custom"),
+   R2 = c(R2, R2_knn, R2_rf, R2_tree, R2_knn_custom),
+   R2_adjusted = c(R2_adjusted, R2_adjusted_knn, R2_adjusted_rf, R2_adjusted_tree, R2_adjusted_knn_custom),
+   MSE = c(MSE, MSE_knn, MSE_rf, MSE_tree, MSE_knn_custom),
+   RMSE = c(rmse_linear, rmse_knn, rmse_randomforest, rmse_tree, rmse_knn_custom),
+   MAE = c(MAE, MAE_knn, MAE_rf, MAE_tree, MAE_knn_custom)
+ )
> # Print updated regression results table
> print(regression_results)
# A tibble: 5 × 6
  Model                R2 R2_adjusted   MSE  RMSE   MAE
  <chr>             <dbl>       <dbl> <dbl> <dbl> <dbl>
1 Linear Regression 0.685       0.671  82.8  9.10  7.29
2 KNN               0.716       0.703  76.1  8.72  6.50
3 Random Forest     0.896       0.892  27.1  5.20  3.76
4 Decision Tree     0.765       0.755  65.2  8.08  5.90
5 KNN_Custom        0.709       0.696  76.8  8.77  6.59
> |
```

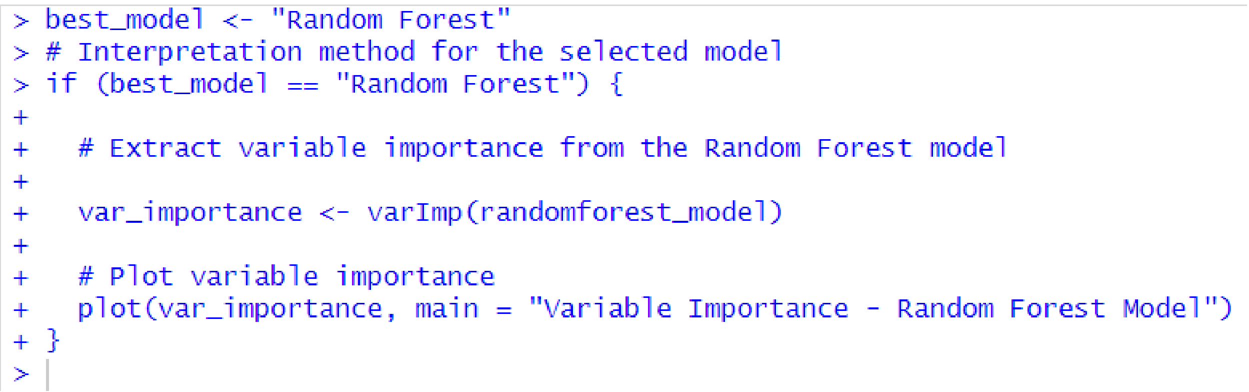
Model	R2	Adjust R2	MSE	RMSE	MAE
Linear Regression	0.685	0.671	82.8	9.10	7.29
KNN	0.716	0.703	76.1	8.72	6.50
Random Forest	0.896	0.892	27.1	5.20	3.76
Decision Tree	0.765	0.755	65.2	8.08	5.90
KNN with different parameters	0.709	0.696	76.8	8.77	6.59

Model Interpretation

When looking at R-squared, MSE, RMSE, and MAE values collectively, Random Forest comes out on top. The most exact estimates of concrete strength are produced, showcasing its exceptional capacity for understanding complex connections. Linear regression, K-Nearest Neighbors (KNN), Random Forest, Decision Tree, and KNN with various parameters were the five models that were assessed. R-squared, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) were used to compare their performance (Wohlwend, 2023).

With the greatest R-squared value of 0.896, Random Forest is the best match between the model and the data, explaining nearly 90% of the variance in concrete strength. In comparison to other models, it also showed the lowest mean absolute error (MAE) of 3.76 and root mean squared error (RMSE) of 5.20, demonstrating better predicted accuracy.

Meanwhile, compared to Random Forest, the The performance of the linear regression model was mediocre, with an R-squared value of 0.685 and substantially larger errors. With similar error metrics to Linear Regression and R-squared values ranging from 0.709 to 0.716, both KNN models also functioned rather well. Having a 0.765 R-squared value and errors in between those of the KNN and linear regression models, the Decision Tree model performed in an intermediate manner.



```
> best_model <- "Random Forest"
> # Interpretation method for the selected model
> if (best_model == "Random Forest") {
+
+   # Extract variable importance from the Random Forest model
+
+   var_importance <- varImp(randomforest_model)
+
+   # Plot variable importance
+   plot(var_importance, main = "Variable Importance - Random Forest Model")
+ }
>
```

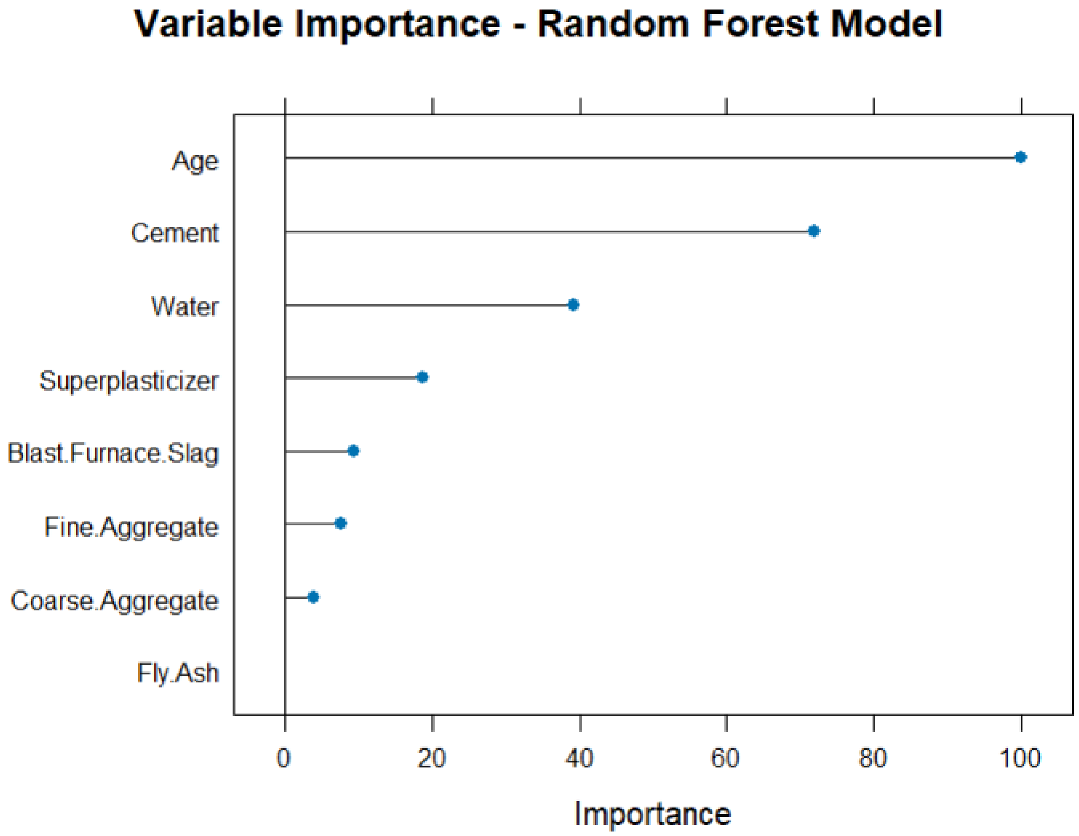


Figure 11: Variable Importance in Concrete Strength based on Random Forest Model

The figure depicts the relative significance of various factors in predicting concrete strength using a Random Forest model. According to the plot, 'Age' emerges as the most critical feature for predicting concrete strength. This suggests that the model prioritizes the age of the concrete as the most influential factor in determining its strength. Following 'Age' in importance 'Cement,' and 'Water' likely play an important part in the concrete mix design and its subsequent strength development. The remaining features, 'Superplasticizer,' 'Blast Furnace Slag,' 'Fine Aggregate,' 'Coarse Aggregate,' and 'Fly Ash,' appear to have a relatively lower impact on the model's predictions compared to the aforementioned factors. However, it's essential to acknowledge that even these features contribute to the overall strength of the concrete.

Conclusion

The results of the produced models is compared critically to the chosen machine learning (ML) techniques, revealing important information about the predictive power of each strategy. We emphasized in the introduction the effectiveness of individual and group machine learning methods, including Random Forest (RF), AdaBoost, and Support Vector Regression (SVR), in predicting the compressive strength of concrete. Five models were assessed by our analysis.

We reviewed the knowledge gained during the data interpretation and preprocessing phases after examining the key variables for the chosen Random Forest model. In estimating concrete compressive strength, the Random Forest model performed better than other ML techniques, indicating that it can handle complex interactions in the dataset. The three main variables that were found to be important—"Age," "Cement," and "Water" in particular—offer important insights into the variables that affect concrete strength and can help improve the accuracy of mix designs and building techniques to maximize structural performance and longevity.

All things considered, this analytical study provides professionals in the field with insightful knowledge and useful resources to improve concrete performance, project outcomes, and eventually the global advancement of resilient and sustainable infrastructure.

Bibliography

Garg A (2017) *Machine Learning with Iris Dataset, RPubs* [Online]. Available from: <https://rpubs.com/Aakansha_garg/261616> [Accessed 8/5/2024].

Tavoosi S (2019) *A beginner's Guide to Machine Learning with R*, *Kaggle* [Online]. Available from: <https://www.kaggle.com/code/tavoosi/a-beginner-s-guide-to-machine-learning-with-r> [Accessed 8/5/2024].

Wohlwend, B. (2023) *Regression model evaluation metrics: R-squared, adjusted R-squared, MSE, RMSE, and Mae*, *Medium*. Available from: <https://medium.com/@brandon93.w/regression-model-evaluation-metrics-r-squared-adjusted-r-squared-mse-rmse-and-mae-24dcc0e4cbd3> [Accessed: 10/5/2024].