Academic Year	Module	Assessment Number	Assessment Type
2025	Concept and Technology of AI	III	Individual Report

Regression Analysis Report

Student Id : 2407947

Student Name. : Shirish Pandit Chhetri

Section : L5CG2

Module Leader. : Siman Giri

Tutor : Sunita Parajuli

Submitted on. : 11 Feb 2025

Abstract

This report includes a thorough analysis and modeling procedure for machine learning-based concrete compressive strength prediction. The research includes data preparation, model creation, hyperparameter tuning, feature selection, and comparison of two regression models: Random Forest and Ridge Regression. The best model is chosen using the performance metric, and the outcomes are examined to show how predictive the models are.

Introduction

In construction engineering, concrete compressive strength is one of the most important parameters that define the capacity of a structure to resist loads. Prediction of compressive strength enhances quality control and structure durability and safety. The goal of this study is to develop prediction models that forecast the compressive strength of concrete from a range of input parameters using machine learning algorithms.

Data Processing

This research uses the "Concrete Compressive Strength" dataset, which includes information on cement, fly ash, blast furnace slag, water, superplasticizer, coarse and fine aggregate, and age. After loading, rudimentary data exploration is done to comprehend the structure and properties of the dataset. Duplicate entries are eliminated to guarantee data quality, and missing values are located and filled up using the proper techniques. To improve model performance, the characteristics are normalized to have a unit variance and zero mean. Using matlablib , I use various graph to visualize the data.

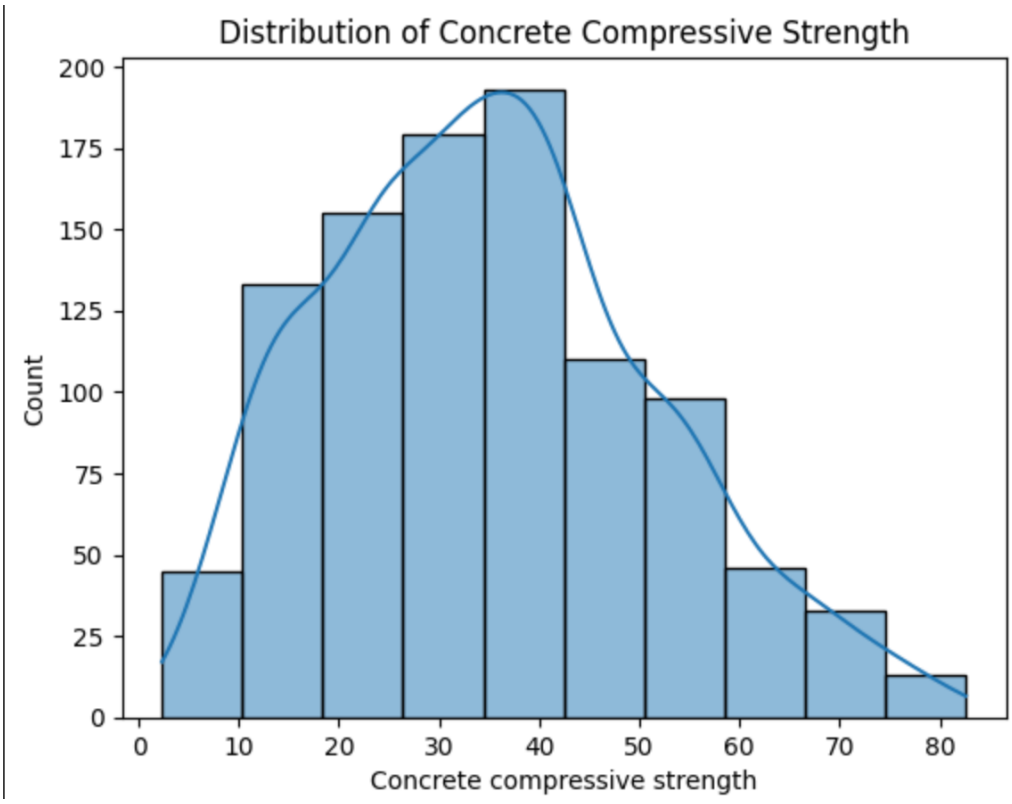

Fig : Histograph of Concrete compressive strength relative to count.

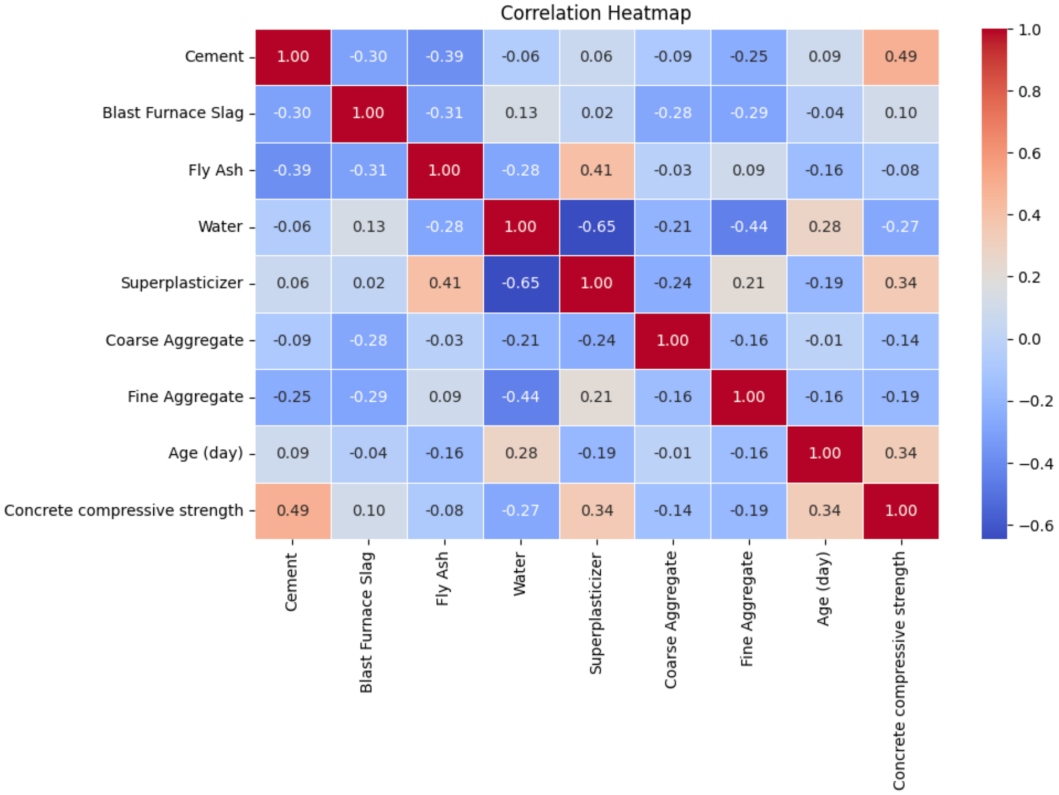

Fig: Correlation heatmap of features/ columns.

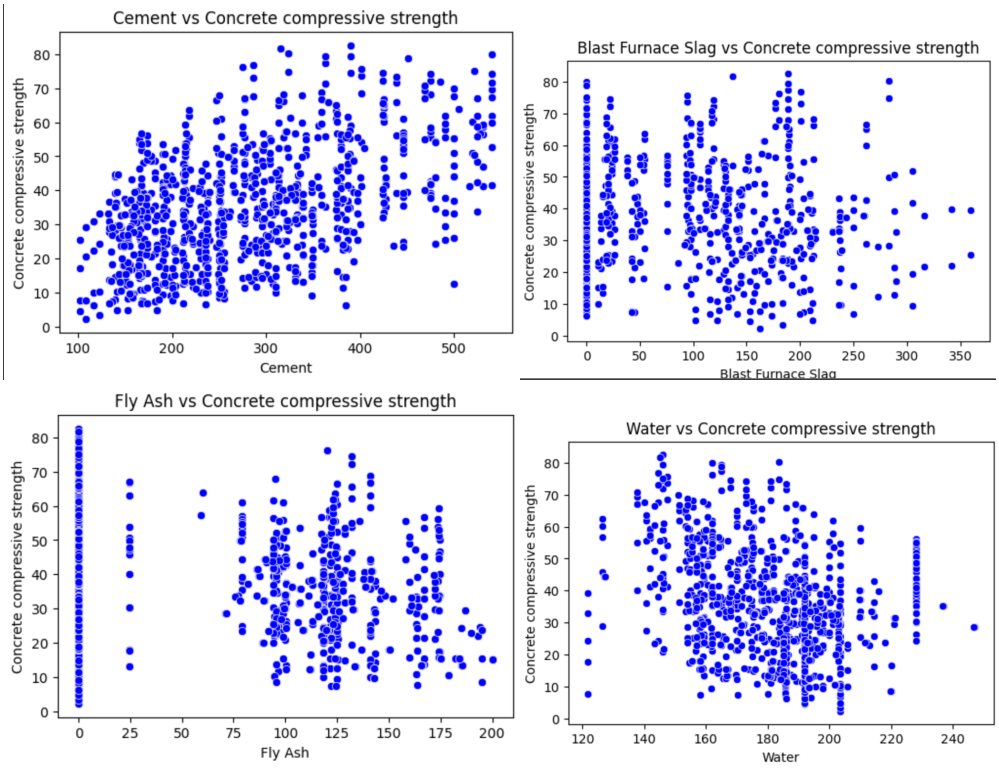

Fig: Scatterplot of different features vs target.

Model Building:

Ridge Regression and Random Forest are two different regression models developed in scikit-learn. In general, Random Forest uses hundreds of trees-in-oaks models and then aggregates their collective predictions to improve the accuracy of a single tree. It is important to note that Ridge Regression is not ensemble, but a linear model with L2 regularization to control overfitting. The metrics employed for the assessment of the models' first performance are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2), which come into play after their training on standardized training data.

The dataset was divided into 2 parts, one to test the models and another to train the models.

After the model were evaluated we noticed that Random Forest Regression was far superior to Ridge Regression, with R^2 at 0.8558 as compared to Ridge regressions 0.5799.

Hyper-parameter Optimization

We refined the performance of the models through hyper-parameter tuning. A grid search is performed varying alpha values to select the best regularization parameter for Ridge Regression. A randomized search is conducted for Random Forest on the hyperparameters: number of estimators, maximum depth, minimum samples split, and minimum samples leaf. The best hyper-parameters were selected from cross-validation scores.

Feature Selection

In order to find out which features are most important for concrete compressive strength prediction, feature selection is performed. For Ridge Regression, the five most important features are selected by Recursive Feature Elimination (RFE), and for Random Forest, features are selected by SelectFromModel based on importance scores. The models are trained again with the selected features, and their performance is again verified.

Conclusion

The final model is chosen based on the performance metrics of the models with optimized hyperparameters and selected features. With carefully chosen features and optimized hyperparameters, the Random Forest model performed better than the Ridge Regression model, obtaining higher MAE, RMSE, and R2 values. Performance indicators are supplied once the final model is assessed using the test data.

Final model

The final model selected is the Random Forest model with the best hyperparameters and selected features. The model achieved an MAE of 4.7498, RMSE of 6.5598, and R^2 of 0.8558 on the data, indicating high predictive accuracy and robustness.

Challenges:

We encountered several challenges during the study of dataset, including handling missing values, selecting appropriate hyperparameters, and identifying the most relevant features. Balancing the trade-off between model complexity and performance was also a critical aspect of the modeling process.

Future work:

To increase predicted accuracy even more, future research can concentrate on investigating other machine learning algorithms like neural networks and gradient boosting. The model's performance can also be improved by adding new features and using domain knowledge. It is possible to improve cross-validation methods to guarantee reliable model assessment.

Model performance:

The performance of the models was evaluated using MAE, RMSE, and R^2 metrics. The Random Forest model with optimized hyperparameters and selected features was able to achieve the best performance, with an MAE of 4.7498, RMSE of 6.5598, and R^2 of 0.8558. These metrics indicate that the model can accurately predict concrete compressive strength with minimal error.

Interpretation of Results:

The result demonstrates how accurately the Random Forest model forecasts the compressive strength of the concrete. The most significant factors influencing compressive strength are the selected properties: cement, blast furnace slag, water, superplasticizer, coarse aggregate, and age. The model can explain a sizable portion of the variance in the target variable, according to its high R^2 score.

Limitations and Suggestions:

The study contains certain limitations even if the results were generally good. The model's generalizability may be impacted by the dataset's modest size. Furthermore, the study only looks at two regression models; alternative algorithms might produce better outcomes. Future research ought to examine a greater variety of machine learning approaches and take into account bigger datasets. Model performance can also be improved by using feature engineering and domain-specific information.

Summary

In brief, this presentation lays down a complete process for predicting compressive strength of concrete by application of machine learning techniques. The final model is selected by performance measures after the elaborate discussion of data preprocessing, modeling, hyperparameter tuning, and feature selection processes. With carefully selected features and well-tuned hyperparameters, Random Forest model fared the best in predictions of concrete compressive strength. The very importance of feature selection and hyperparameter tuning in developing a good predictive model is clear from this study. These findings are open for future extension for achieving better accuracy and applicability in the model.