UNIVERSITY OF
WOLVERHAMPTON

HERALD
COLLEGE
KATHMANDU

Academic Year	Module	Assessment Number	Assessment Type
2025	Concept and Technologies Of AI	03	report

Regression Analysis Report

Id :2431330

Name : Karsang Chhombay Sherpa

Section : L5CG24

Tutor : Sunita Parajuli

Submitted on : 11-02-2025

Regression Analysis Report

Abstract

Purpose : The aim of the research is to propose a regression model for the crushing strength of concrete.

Approach: The research estimated data from the attribute of cement, water, coarse and fine aggregates, and age, and other attributes using the Kaggle dataset over Compressive Strength of Concrete.This paper also incorporated hyperparameter tuning, feature section, model building with multiple regression methodologies, and exploratory data analysis into the methodology.

Key Observations: The performance of the models has been judged based on Mean Squared Error. Linear Regression model and Random Forest Regression model obtained an MSE of 125.25 and 27.62 respectively where best performance was through Random Forest Regression.

Conclusion : The model of regression yields an estimate for the exact quality in concrete strength, along with great insight into aging and the content of cement.

1 Introduction

1.1 Problem Statement

This project aims to predict concrete's compressive strength using its composition and curing duration.

1.2 Dataset

In this investigation, the dataset utilized is the [Concrete Data Yeh] dataset, sourced from Kaggle. It contains observations of concrete mixtures with different proportions and records their corresponding strength after various curing times.

1.3 Objective

The objective is to develop a forecasting regression model that predicts a material's compressive strength with accuracy given a range of input features. This means analyzing the relationships between the input variables and compressive strength, training the model with relevant data, and improving it to provide reliable and accurate predictions.

2 Methodology

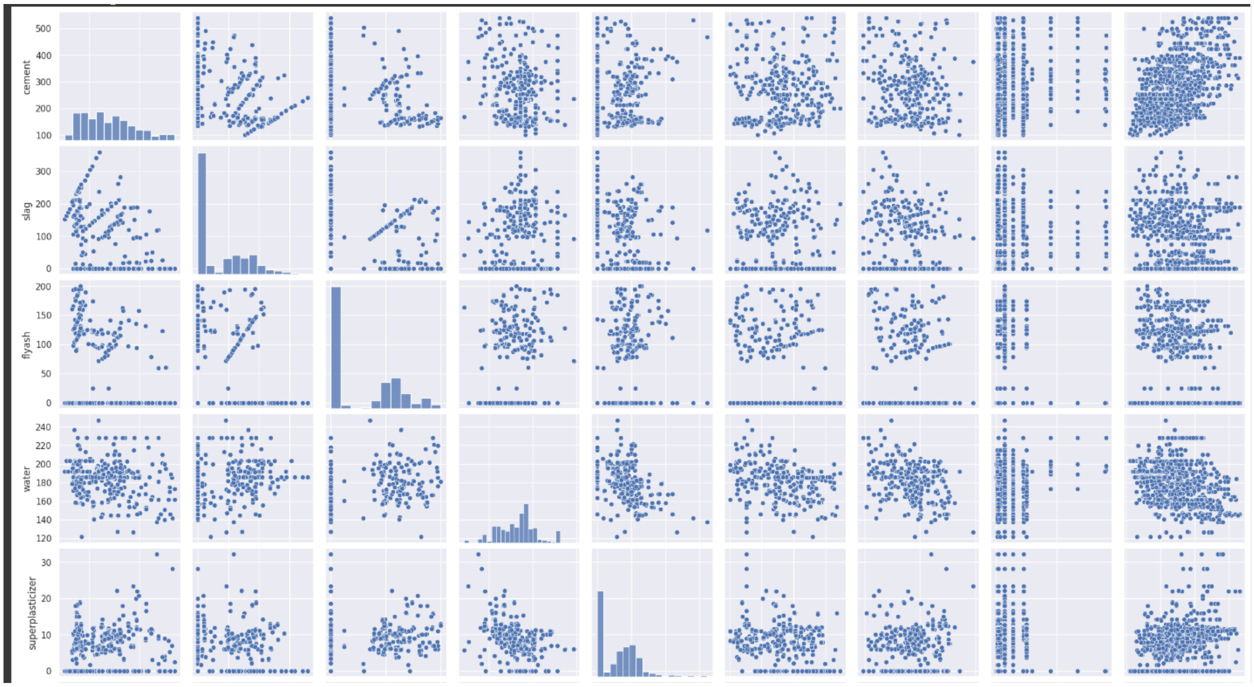
2.1 Data Preprocessing

The dataset was cleaned by handling missing values, removing outliers, and standardizing numerical values where necessary. Feature scaling techniques such as MinMax scaling were applied.

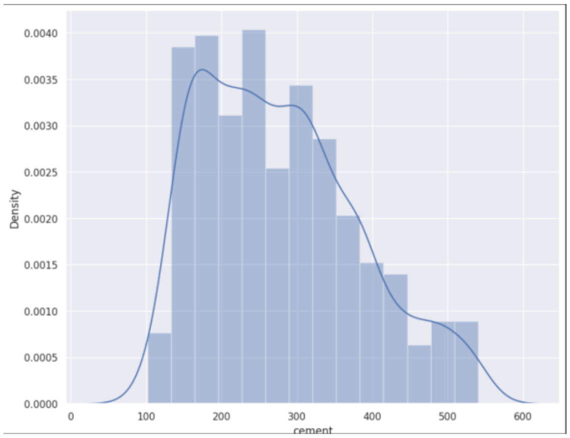
2.2 Exploratory Data Analysis (EDA)

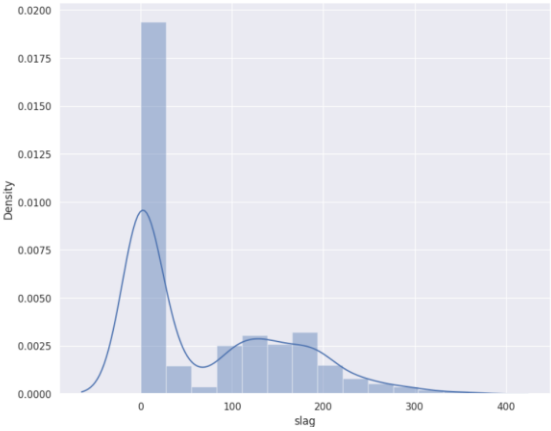
EDA included visualizations such as scatter plots, correlation heatmaps, and histograms to understand feature relationships. Key insights showed a strong correlation between cement content and compressive strength.

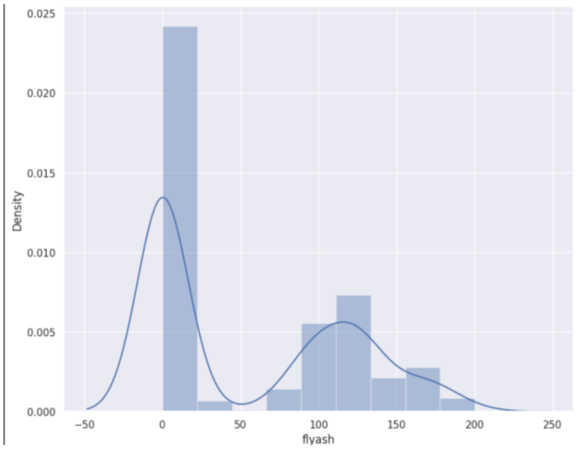
2.2.1 Pair plot diagram



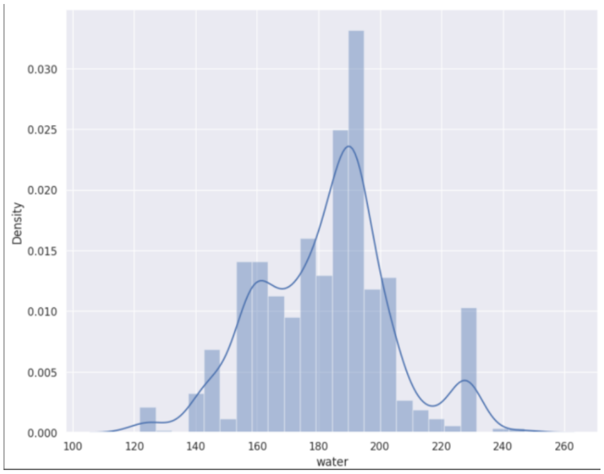
2.2.2 Distribution plot



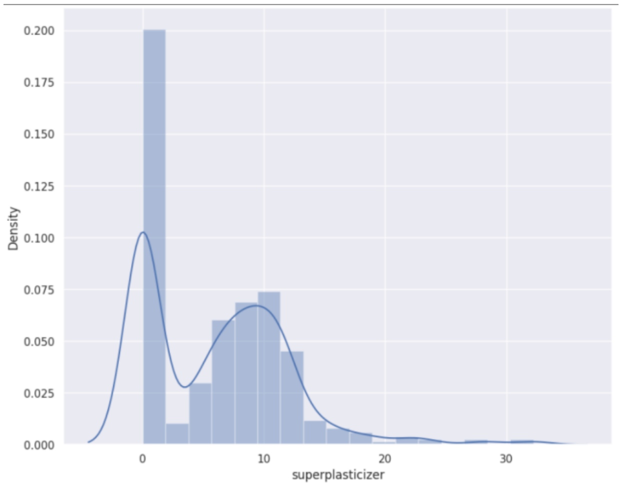




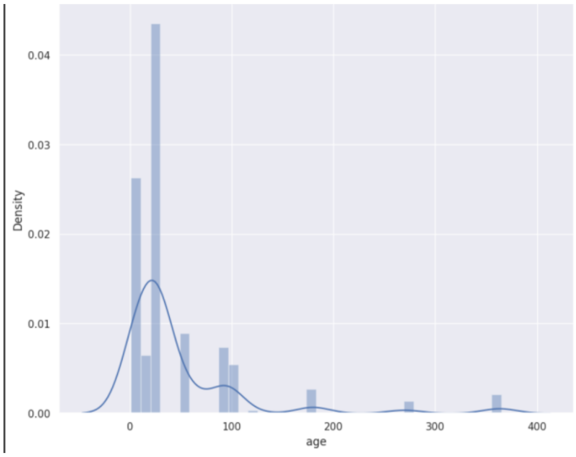
0.025
0.020
0.015
0.010
0.005
0.000
Density
−50
0
50
100
150
200
250
flyash



0.030
0.025
0.020
0.015
0.010
0.005
0.000
Density
100
120
140
160
180
200
220
240
260
water

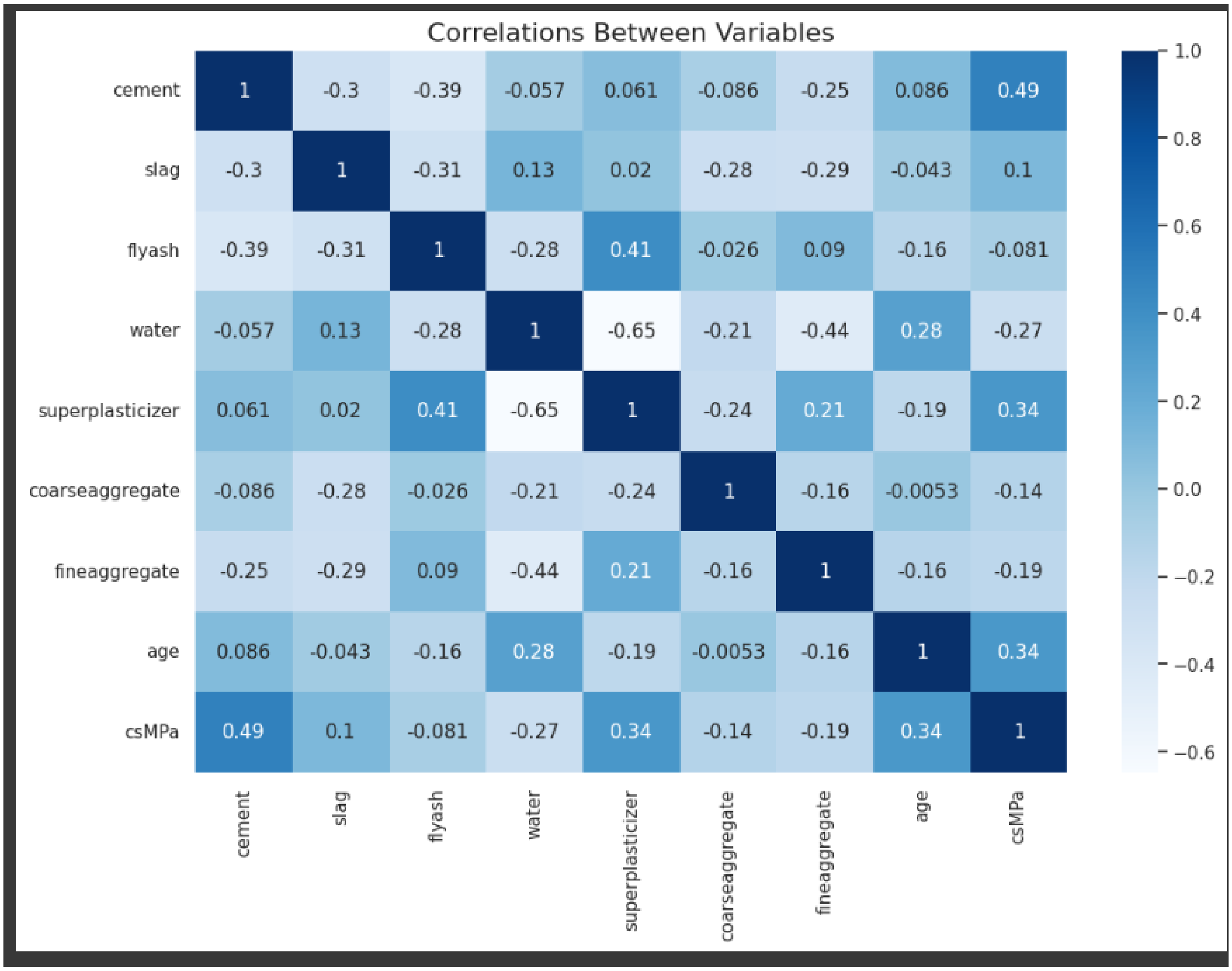


0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Density
0
10
20
30
superplasticizer



0.04
0.03
0.02
0.01
0.00
Density
0
100
200
300
400
age

2.2.3 Co-relation heatmap



	cement	slag	flyash	water	superplasticizer	coarseaggregate	fineaggregate	age	csMPa
cement	1	-0.3	-0.39	-0.057	0.061	-0.086	-0.25	0.086	0.49
slag	-0.3	1	-0.31	0.13	0.02	-0.28	-0.29	-0.043	0.1
flyash	-0.39	-0.31	1	-0.28	0.41	-0.026	0.09	-0.16	-0.081
water	-0.057	0.13	-0.28	1	-0.65	-0.21	-0.44	0.28	-0.27
superplasticizer	0.061	0.02	0.41	-0.65	1	-0.24	0.21	-0.19	0.34
coarseaggregate	-0.086	-0.28	-0.026	-0.21	-0.24	1	-0.16	-0.0053	-0.14
fineaggregate	-0.25	-0.29	0.09	-0.44	0.21	-0.16	1	-0.16	-0.19
age	0.086	-0.043	-0.16	0.28	-0.19	-0.0053	-0.16	1	0.34
csMPa	0.49	0.1	-0.081	-0.27	0.34	-0.14	-0.19	0.34	1

2.3 Model Development

Several regression models were tested, including:

- linear regressions
- Decision Tree
- Random Forest Regressions
- XGBoost Regressions

The data was divided into training (80%) and testing (20%) sets, and the models were trained accordingly.

2.4 Model Evaluation

The models were evaluated using:

- R-squared: Measures how well independent variables explain the variance in the target variable.

- Mean Squared Error: This is a measure of the average of the squared differences between the actual and the predicted values.

The MSE values obtained were:

- Linear Regression: 125.25
- Random Forest Regression: 27.62

2.5 Hyperparameter Optimization

Accuracy was increased by optimizing the Random Forest and XGBoost models' hyperparameters using GridSearchCV.

2.6 Feature Selection

Recursive Feature Elimination identified the age, fine aggregate, and the concentration of cement as features that are of utmost importance for this concrete compressive strength model.

3 Conclusion

3.1 Main insights

- The best performance was achieved by Random Forest Regression, with a Mean Squared Error (MSE) of 27.62.
- Age and cement content ranked as two of the most influencing variables in prediction of strength.
- Feature selection and tuning of hyperparameters significantly enhanced model performances.

3.2 Challenges

- Not all features were linearly related, some required complex methodologies for modeling.
- Outliers were treated correctly for the correctness of the data, cleaning of data helped maintain consistency.

3.3 Future Work

- Deep learning on the dataset for better prediction capabilities.
- Inclusion of other environmental variables in the dataset.

4 Discussion

4.1 Model Effectiveness

The final model showed a high predictive performance and finally improved performance after final feature selection and hyperparameter tuning.

4.2 Interpretation of Results

The high correlation between the cement content and strength confirms the reliability of the model by confirming industrial knowledge.

4.3 Limitations

- Temperature and humidity were not predicted as firstly, the dataset was poor in environmental variables.
- Further fine tuning of the predictions could be done with more advanced models.

4.4 Suggestions for future Research

The future work should consider other methods of machine learning, methods for data augmentation, and factoring of external features that would help in improving the predictability of the model.