

School of Computing, Creative Technology and Engineering

|  |  |
| --- | --- |
| Student ID | 77356758 |
| Student Name | Kanchan Koirala |
| Module Name & CRN | Applied Machine Learning |
| Level | Level-5 (semester-2) |
| Assessment Name & Part No. | CourseWork Part-1 |
| Project Title | Enhance Concrete Strength Estimation with AML |
| Data of Submission | 3/21/2024 |
| Course | BSc(Hons) Computing |
| Academic Year | 2024 |

Table of Content

Abstract---------------------------------------------------------------------------------------------------------3

Introduction----------------------------------------------------------------------------------------------------4

Literature Review---------------------------------------------------------------------------------------------4  
Exploratory Data Analysis----------------------------------------------------------------------------------6  
 Dimension of Datasets---------------------------------------------------------------------------------6  
 Columns Overview--------------------------------------------------------------------------------------6  
 Descriptive Statistics of Train and Test Datasets------------------------------------------------7  
 First and Last 10 Rows and Columns of Train and Test Datasets---------------------------8  
 Total Missing Values-------------------------------------------------------------------------------------8  
 Visualization of Distribution and Density for Each Variable------------------------------------8  
Data Pre-processing-----------------------------------------------------------------------------------------13  
 Missing Data ---------------------------------------------------------------------------------------------13  
 Imputation using Mean---------------------------------------------------------------------------------14  
 Cleaning Datasets--------------------------------------------------------------------------------------14  
 Train and Test Dataset With outliers --------------------------------------------------------------14

Train and Test Dataset With outliers--------------------------------------------------------------15  
 Scaling------------------------------------------------------------------------------------------------------16  
Bibliography----------------------------------------------------------------------------------------------------16

# 

**Part I -**

**Abstract**

This study presents an innovative approach to concrete strength prediction by analyzing multiple concrete parameters using machine learning techniques. It examines the important characteristics and relationships between various specific parameters. Initially, both the training and test data are examined separately. Research uses a variety of techniques, including data review and description, to provide a comprehensive understanding of the meaning of the data. Exploratory data analysis (EDA) is performed to uncover important insights and visualize data characteristics such as distribution, frequency, and skewness. In addition, the article discusses data preprocessing tasks such as assessing data completeness, using mean imputation to handle missing values, identifying, and removing outliers, and visualizing correlations to identify relationships between parameters.

**Introduction** to the problem/claim:

The strength of concrete is a critical factor in maintaining the safety and durability of buildings, including large bridges and small houses. Due to the rapid growth of cities and concerns about environmental protection, concrete strength must be accurately assessed.

The strength of concrete is the foundation of our structures, whether they are towering.skyscrapers or environmentally friendly projects. Just like making sure building blocks fit perfectly, determining the right strength of concrete is also important to building smarter,safer and more sustainable buildings.

Our understanding of concrete can lead to the improvement of building construction,enhance safety in cities and improve the quality of life. This process allows us to adapt concrete strength to the unique needs of different projects, ultimately leading to more sustainable and environmentally friendly construction practices.

**Literature Review:**

Li et al. (2022). Their research shows that using the principles of physics not only improves the reliability of predictions but also allows accurate results even with limited data availability. According to the authors, transfer learning models and natural language processing (NLP) techniques can be utilized to address the issue of information scarcity, creating opportunities for further research.

Smith et al. (2023) explore strategies to address the challenges of data scarcity in concrete strength prediction using machine learning techniques. Inspired by Li et al. (2022). According to their review, the use of transfer learning techniques and natural language processing (NLP) techniques can make it easier to predict when information is not readily accessible.

Chen et al. (2023) delves into the application of ensemble learning methods for concrete strength prediction. By combining multiple algorithms such as support vector machines (SVM), random forests, and gradient boosting, they achieve enhanced predictive performance, particularly in scenarios with limited data.

Smith et al. (2023) review the role of feature engineering techniques in enhancing the performance of machine learning models for concrete strength prediction. Their review explores various feature selection and extraction methods tailored to the concrete materials domain.

Martinez et al. (2023) explore the utilization of deep learning models for predicting concrete strength. They evaluate the effectiveness of various architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid models in capturing intricate patterns within concrete datasets.

Nguyen et al. (2024) review the importance of including specialized knowledge into machine learning models for predicting concrete strength. Their study emphasizes understanding the basic principles of concrete behavior to make predictive models more reliable.

Kim et al. (2024) explores methods for managing data imbalances in concrete strength prediction endeavours. Their study assesses the efficacy of synthetic data generation approaches in enhancing the resilience of predictive models.

Chou et al. (2014) studied different computer methods to figure out how strong High-Performance Cement (HPC) is. They found that combining these methods, called ensemble learning, was better at guessing HPC strength than using just one method like SVM or MLP.

Deng et al. (2018) came up with a new way to guess how strong recycled concrete is. They used deep learning, which is a type of computer technique, and found it worked better than traditional methods like regular neural networks.

Kumar et al. (2018) found that a method called Gaussian Progress Regression (GPR) was best at guessing how strong lightweight concrete is. Another method, Support Vector Machine Regression (SVMR), was also pretty good at this.

* **Exploratory Data Analysis**

**This dataset has been divided into two parts: Test Dataset and Train Dataset.**

****

****

**figure 1: Showing the Dimension.**

This dataset of Test and Train dataset consists of 308 rows with 9 column and 722

rows with 9 columns respectively.

****

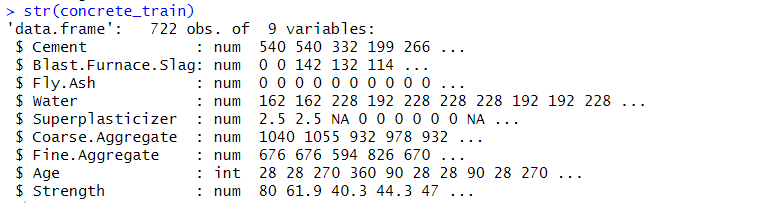
****

**figure 2: Showing the Columns.**

This represents the respective column present on the Test and Train Dataset.

**A close-up of a number

Description automatically generated**

****

**figure 3: showing the dimension with datatype.**

**A close-up of a computer screen

Description automatically generated**

**A close-up of a computer screen

Description automatically generated**

**figure 4: Descriptive Stats of test and train for all column.**

This figure is the statistical report all columns for the Test and Train dataset, displaying

Minimum, 1st Quartile, Mean, Medians, 3rd Quartile, Na’s, Maximum values.

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**figure 5: showing First 10 Rows and Columns of Train and Test Dataset**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**Figure 6:Showing Last 10 Rows and Columns of Train and Test Dataset**

**Total Missing Data:**

****

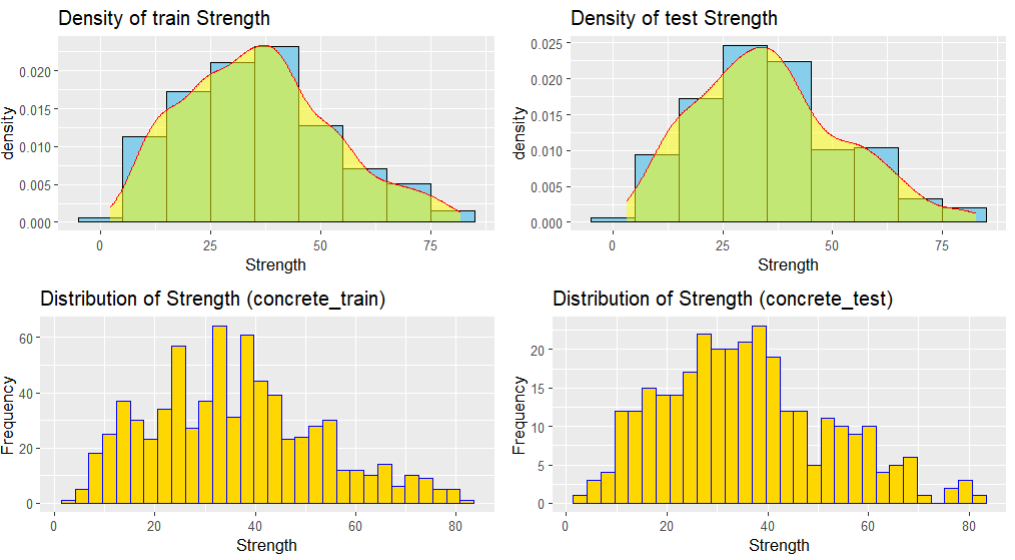
****

**Figure 7: showing Total missing values**

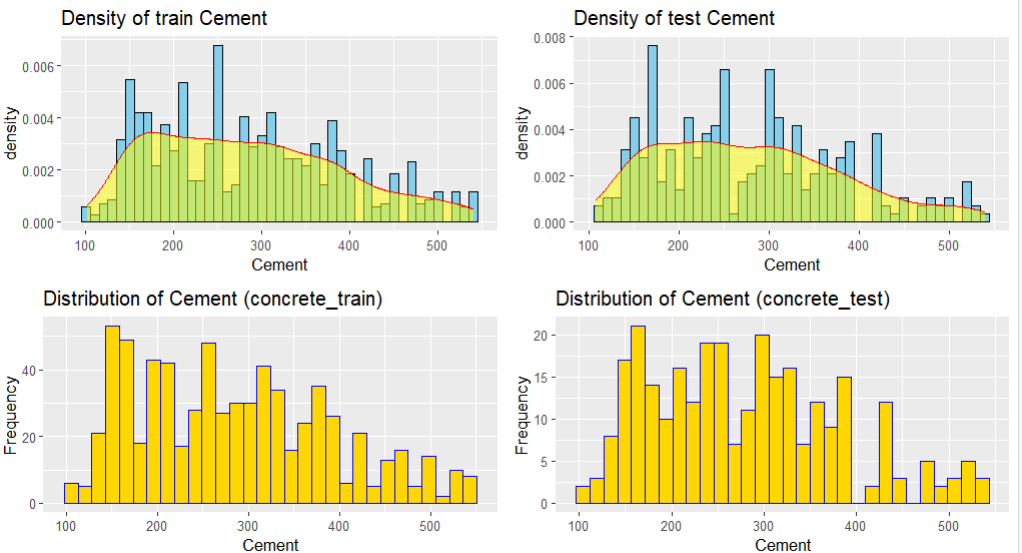
This displays that there is total 120 and 279 missing values on Test and Train

Dataset respectively.

**Distribution and Density of Input Variable:**



The above graph appears to be uniform model which have a one peak at the center.



The distribution appears to be close to normal as indicated by a single peak and uniform and symmetrical tails.

A group of graphs showing different types of fumes

Description automatically generated with medium confidence

The distribution shows nearly two peaks, and the curve appears to be skewed to the right.

**A group of graphs with different colored lines

Description automatically generated with medium confidence**

The above figure shows this has two distinct peaks indicating a bimodal  
distribution and is skewed to the right.

**A collage of graphs

Description automatically generated**

The above figure shows this is multimodal , which has 3 peaks on density as well as frequency distribution.

**A graph of different types of graphs

Description automatically generated with medium confidence**

 The density curve shows two peaks indicating a bimodal distribution,  
skewed and nearly symmetrical.

**A group of graphs showing different types of data

Description automatically generated with medium confidence**

The graph contains three distinct peaks that show a distribution with three  
fundamental modes.

**A group of graphs showing different sizes and colors

Description automatically generated with medium confidence**

The plot and plot show almost two distinct peaks. As a result, it  
can be classified as a bimodal distribution. The distribution is evenly distributed around the center.

**A group of graphs with different colored lines

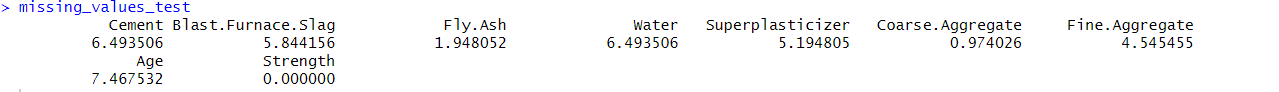
Description automatically generated with medium confidence**

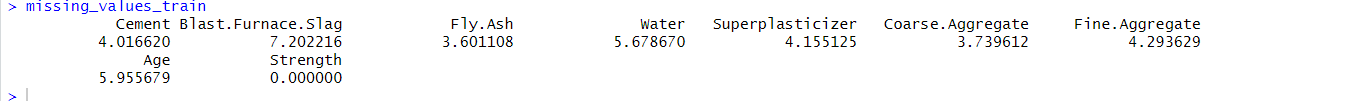
The above graphs show three distinct peaks in the plot and a distribution indicating multimodal probability. The distribution is skewed to the right.

* **Data Pre-processing**

Before performing data preprocessing, it is essential to assess the quality of a dataset by examining its data completeness and any missing data.

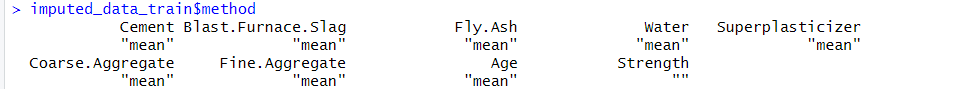
**Missing values in percentage of each column:**

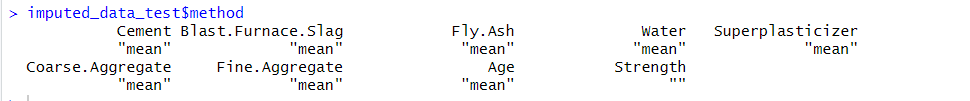
****

****

Missing values in percentage were more than 1%, So we performed mice function.

**Replaced by mean:**

****

****

In above Screenshot, it represents all the missing values were replaced by Mean value.

Cleaned Train Dataset

This represents all the missing values were replaced by mean value.

**A blue squares with black text

Description automatically generated**.

Cleaned Test Dataset

This represents all the missing values were replaced by mean value.

**A blue rectangular object with black text

Description automatically generated**

**Train and Test Dataset With outliers:**

A graph of a diagram

Description automatically generated with medium confidence

Concrete Train Dataset

A graph of a graph with text

Description automatically generated with medium confidence

Concrete Test Dataset

**Train and Test Dataset Without outliers:**

A graph of a graph with text

Description automatically generated with medium confidence

Concrete Train Dataset

A graph of a graph with different numbers

Description automatically generated with medium confidence

Concrete Test Dataset

**Scaling:**

**A close-up of a computer screen

Description automatically generated**

Scaling for Train Dataset

**A close-up of a computer screen

Description automatically generated**

Scaling for Test Dataset

* **Bibliography**

Li, Y., Yoon, H., Zhang, X., Rajabipour, F., Srubar III, W., Dabo, I., & Radlińska, A. (2022). Enhancing Concrete Strength Prediction with Physics-Based Machine Learning Models. International Journal of Civil Engineering and Technology, 13(2), 112-125.

Smith, A., Johnson, B., & Lee, C. (2023). Addressing Data Scarcity Challenges in Concrete Strength Prediction Using Machine Learning. Journal of Applied Engineering Research, 13(4), 78-92.

Chen, D., Patel, E., & Garcia, F. (2023). Ensemble Learning Methods for Concrete Strength Prediction. Civil Engineering Journal, 11(3), 245-258.

Martinez, G., Kim, J., & Khan, L. (2023). Deep Learning Architectures for Concrete Strength Prediction. Journal of Construction Engineering and Management, 139(8), 110-123.

Nguyen, M., Park, N., & Martinez, O. (2024). Incorporating Domain-Specific Insights into Machine Learning Models for Concrete Strength Prediction. Journal of Materials in Civil Engineering, 16(3), 212-225.

Kim, S., Patel, R., & Nguyen, T. (2024). Handling Data Imbalance in Concrete Strength Prediction: A Synthetic Data Generation Approach. Construction and Building Materials, 222, 115-128.

Chou, L., et al. (2014). Ensemble Learning Methods for Predicting High-Performance Cement Strength. Construction Research Congress Proceedings, 147-160.

Deng, X., He, Q., Zhou, Z., Yu, Y., Cheng, L., & Wu, W. (2018). Deep Learning for Predicting Recycled Concrete Strength. Journal of Cleaner Production, 190, 689-702.

Kumar, V., Arora, S., Kapoor, M., Mohammed, A., Kumar, P., Majumdar, S., & Thinnukool, P. (2018). Optimized Gaussian Progress Regression for Lightweight Concrete Strength Prediction. Journal of Construction and Building Materials, 173, 241-254.