COS40007 - Artificial Intelligence for Engineering

Minh-Hieu Tran _ 104850021

Studio - Week 1

EDA Summary Report for Week 1

Objective: To clean the concrete dataset and perform an initial exploratory analysis to understand the data's structure, variable distributions, and key relationships before model building.

Data Cleaning Summary:

- **Duplicates:** Any duplicate rows found in the initial dataset were removed to prevent bias.
- **Outliers:** Outliers were identified using boxplots and removed via the IQR method to create a more robust dataset for analysis.

Exploratory Analysis Findings:

- **Key Predictors:** The correlation analysis confirmed that cement (correlation: 0.50) and age (correlation: 0.42) are the strongest positive predictors of concrete strength. This aligns perfectly with fundamental civil engineering principles.
- **Negative Predictor:** The water content showed a moderate negative correlation (-0.31) with strength, quantitatively confirming that a higher water-cement ratio leads to weaker concrete.
- **Distributions:** The distributions of variables like slag, ash, and age are heavily skewed, which is expected given their nature as additives or non-continuous measurements. The target variable, strength, has a favorable near-normal distribution.
- **Multicollinearity:** The heatmap showed no strong correlations between predictor variables, indicating that multicollinearity is not a significant issue and that all features can likely be used in initial model training without conflict.

Conclusion: The data has been successfully cleaned and analyzed. The initial EDA validates the dataset's quality and provides clear direction for the next step: using these features to build a predictive model for concrete strength.