COS40007: Artificial Intelligence for Engineering

Portfolio Assessment: "Hello Machine Learning for Engineering"

Name: Minh Hieu Tran
Student ID: 104850021

September 20, 2025

1. Dataset Selection and Justification

The dataset selected for this assessment is the **Concrete Compressive Strength** dataset. This dataset contains 1030 instances with 8 quantitative independent variables and 1 quantitative dependent variable, which is the concrete compressive strength.

Reason for Choice: As this unit focuses on AI for Engineering, I chose a dataset with direct application to a core engineering. Concrete is a fundamental material in civil and structural engineering. The ability to predict its strength based on its components is a classic, high-value engineering problem. This dataset provides a practical and relevant context to apply machine learning techniques to predict material properties, which is a critical task in engineering design, quality control, and material science.

2. Summary of Exploratory Data Analysis (EDA)

The initial EDA involved loading the dataset, checking for missing values, and analyzing the statistical properties of each feature. It was confirmed that the dataset contains **no missing values**, which streamlines the pre-processing phase.

The primary focus of the EDA was to understand the relationship between the input features and the target variable, strength. A correlation matrix was generated to quantify these relationships. The heatmap of this matrix is shown in Figure 1.

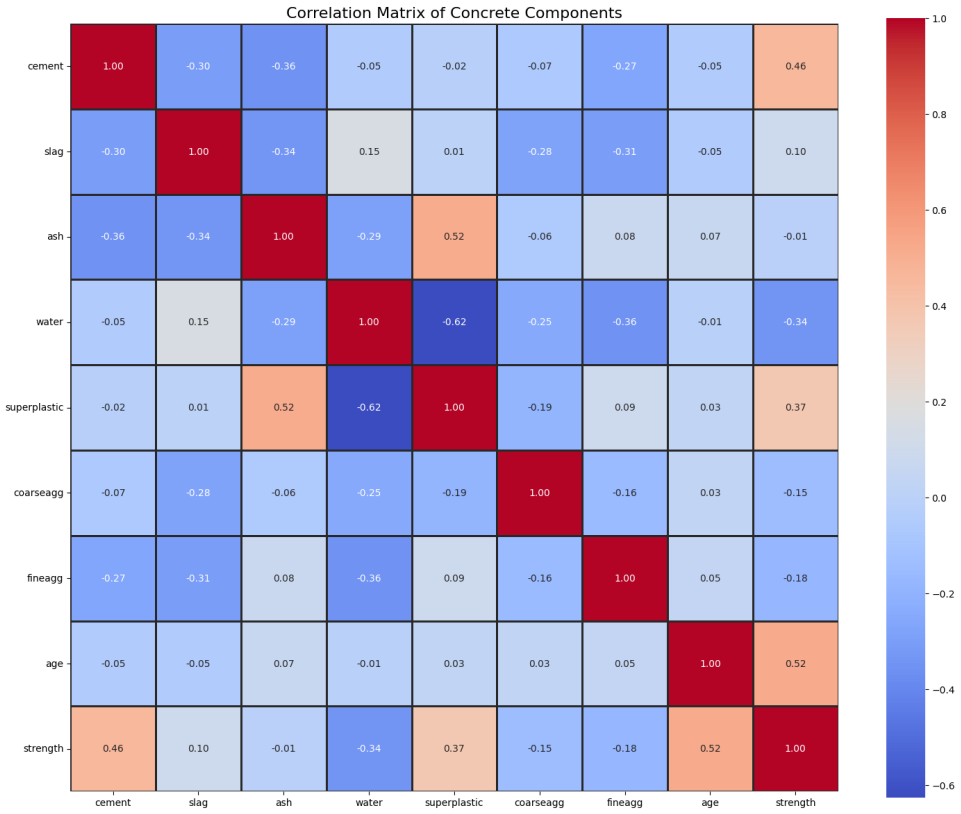

Figure 1: Correlation matrix of all features.

Key findings from the correlation analysis include:

- **Strong Positive Correlation:** The features cement, superplastic, and age showed the most significant positive correlations with concrete compressive strength. This is high quality following engineering principles, where cement content and curing time (age) are primary drivers of strength.
- **Strong Negative Correlation:** water showed a strong negative correlation, which is also expected, as a higher water-to-cement ratio typically weakens the concrete.

- **Weak Correlations:** Other features like slag, ash, coarseagg, and fineagg showed weaker correlations.

3. Class Labelling for Target Variable

The target variable, strength, is a continuous numerical value. For this to be a classification problem, the target variable was converted into discrete classes. Following the assessment guidelines to create a balanced class distribution, the strength values were binned into four distinct categories which divides the data into equal-sized bins.

The defined classes are: 'Low', 'Medium', 'High', and 'Very High'. The distribution of instances across these classes is shown in Figure 2, confirming a nearly balanced distribution.

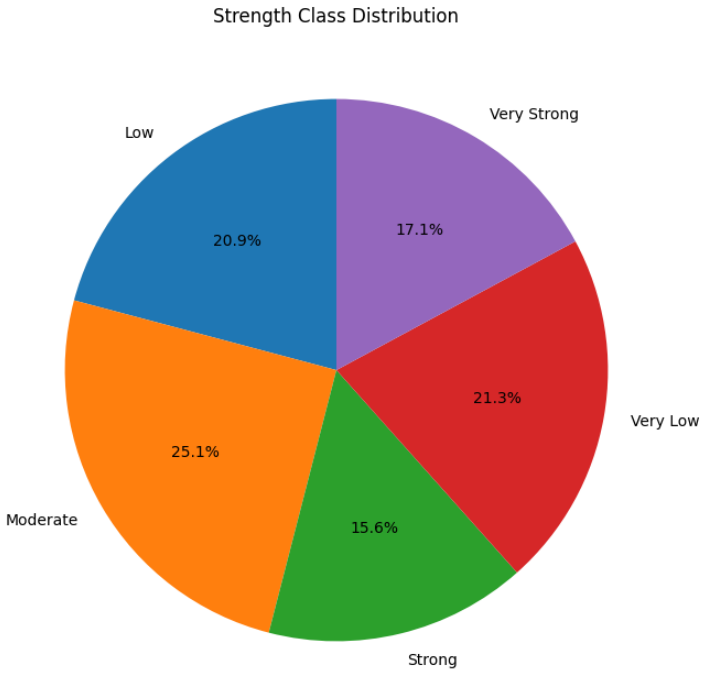

Figure 2: Distribution of instances across the four defined strength classes.

4. Feature Engineering and Feature Selection

Based on the EDA and engineering domain knowledge, five distinct feature sets were engineered to evaluate the performance of the Decision Tree models. Normalization was not performed as Decision Tree algorithms are not sensitive to the scale of input features.

- **Feature Set 1 (All Original Features):** A baseline model using all eight original numerical input features.
- **Feature Set 2 (Highest Correlation Features):** A reduced set including only the features with the highest correlation to the target: cement, superplastic, and age.
- **Feature Set 3 (All Features with Age Categorical):** An engineered feature set where the continuous age variable was categorized into bins. This tests if the model benefits from viewing age as discrete stages of curing.
- **Feature Set 4 (Reduced Features - No Slag/Ash):** Based on engineering judgment, slag and ash were removed as they often contain zero values and can be considered additives rather than core components.

- **Feature Set 5 (Selected Features from Feature Selection):** This feature set was created by selecting a specific subset of features: cement, water, superplastic, and age. This tests the model's performance on a targeted, reduced set of features.

5. Training and Decision Tree Model Development

For each of the five feature sets, a Decision Tree classification model was developed using the DecisionTreeClassifier from the scikit-learn library. The methodology was consistent for each model:

1. The data was split into a training set (70%) and a testing set (30%).
2. A Decision Tree model was initialized with random_state=42 for reproducibility.
3. The model was trained on the training data.
4. The trained model was used to make predictions on the unseen test data.
5. The performance of the model was evaluated using the accuracy metric, which measures the proportion of correctly classified instances.

6. Final Comparison Table

The performance of the five models was compiled into a single comparison table to facilitate a clear evaluation of the different feature engineering strategies. The results are presented in Table .

Feature Set	**Accuracy**	**Number of Features**
All Original Features	0.624088	8
Highest Correlation Features	0.609489	3
All Features with Age Categorical	0.624088	8
Reduced Features (No Slag/Ash)	0.587591	6
Selected Features from Feature Selection	0.558394	4

Table 1: Comparison of Decision Tree Model Performance with Different Feature Sets.

7. Brief Summary of Your Observation

The final comparison table reveals key insights into the model's performance. The highest accuracy was achieved by both the **"All Original Features"** and the **"All Features with Age Categorical"** (Feature Set 3) sets, each with an accuracy of 0.624088. This suggests that a comprehensive feature set is crucial for the best predictive performance.

In contrast, the model built using the **"Selected Features from Feature Selection"** (Feature Set 5) performed the worst, with an accuracy of 0.558394. This set included the features *cement*, *water*, *superplastic*, and *age*. The poor performance indicates that the feature selection process used for this set was not effective in capturing the essential information needed for accurate prediction, resulting in a model with significantly reduced predictive power.

Appendix

The complete Python source code for this assessment is available in the accompanying Jupyter Notebook. A shared link to the notebook is provided below.

Shared Link: Link to Google Colab Notebook