SWINBURNE UNIVERSITY OF TECHNOLOGY

COS40007

Portfolio Week 2

(DUE 12/01/25 - 12:00 A.M)

NAME & STUDENTID : **Nguyen Gia Binh - 104219428**

CLASS : **COS40007 - AI for Engineering**

LECTURER : **Dr. Bui Ngoc Dung**

(dbui@swin.edu.au**)**

TUTOR : **Dr. Bui Ngoc Dung**

(dbui@swin.edu.au**)**

Hanoi, January 11, 2025

TABLE OF CONTENT

1) **DATASET**

For this week portfolio and studio i have chosen the Civil Engineering: Cement Manufacturing Dataset from Kaggle, the reason being Studio 1 doesnt exist for this unit in this semester. Overall, this dataset is about the strength of the cement respective to what is added into it. As I am interested in exploring real-world applications of machine learning, this dataset provides an opportunity to apply predictive modeling techniques to one of the vital problems in concrete manufacturing—optimizing concrete compressive strength and improving cement manufacturing processes.

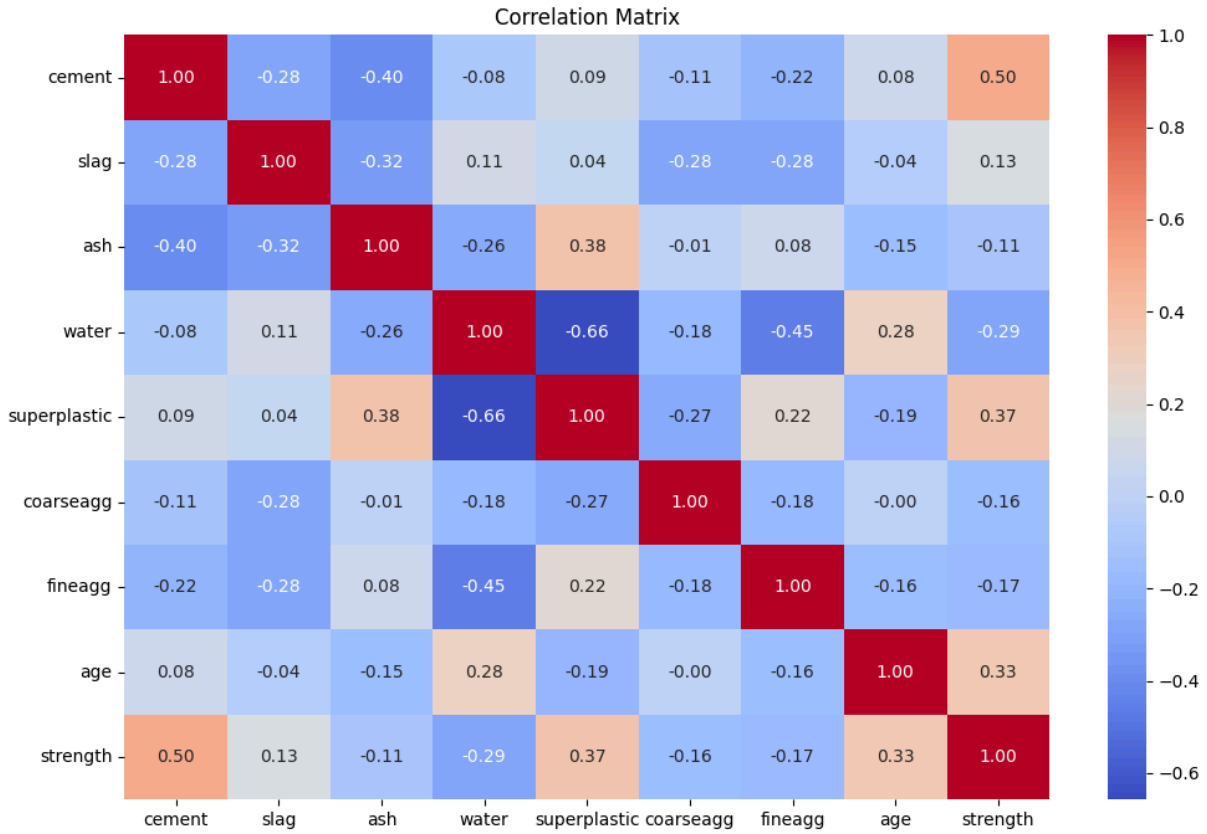
1.1) Fields

The dataset contains these field:

- **cement**: Amount of cement used in the concrete mix (measured in kg in a m3 mixture).
- **slag**: Amount of blast furnace slag used in the mix (measured in kg in a m3 mixture).
- **ash**: Amount of fly ash used in the mix (measured in kg in a m3 mixture).
- **water**: Amount of water used in the mix (measured in kg in a m3 mixture).
- **superplastic**: Amount of superplasticizer additive used (measured in kg in a m3 mixture).
- **coarseagg**: Amount of coarse aggregate used (measured in kg in a m3 mixture).
- **fineagg**: Amount of fine aggregate used (measured in kg in a m3 mixture).
- **age**: Age of the concrete sample at testing (1~365)).
- **strength**: Compressive strength of the concrete sample (measured in MPa).

1.2) EDA

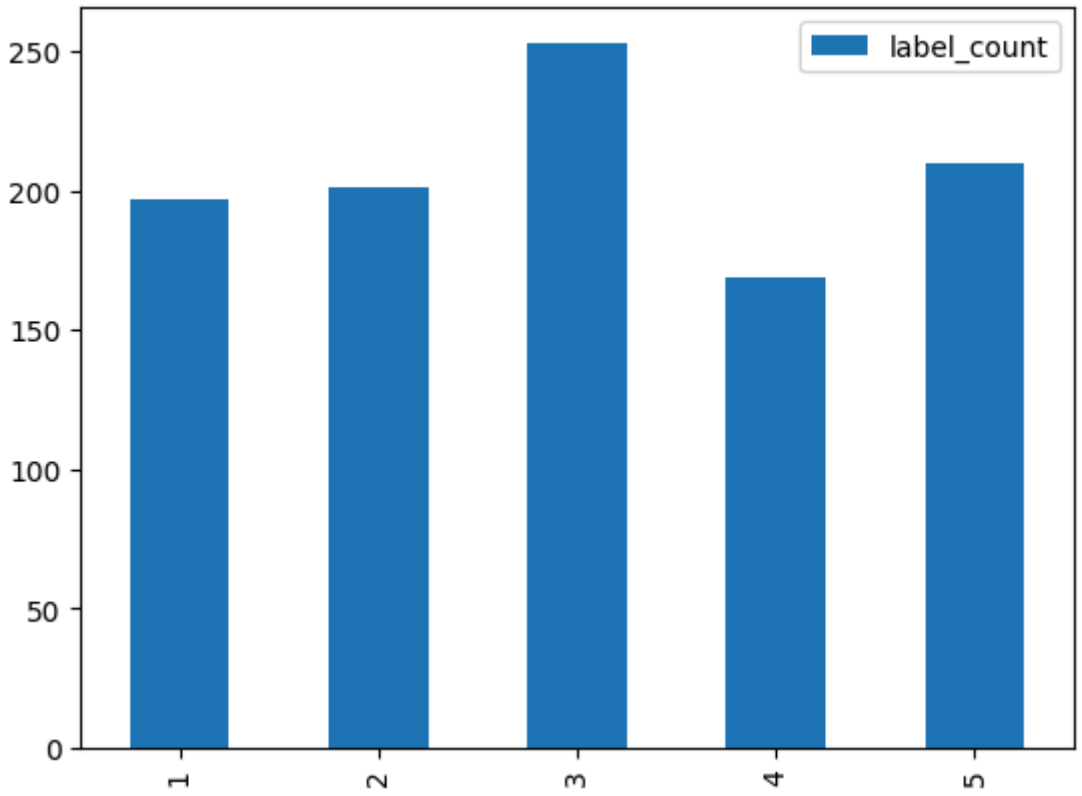
1.2.1) Correlation matrix



From the correlation matrix we can clearly see that:

- **Positive correlation**: cement, superplastic and age all have a high correlation to the strength of the cement
- **Negative correlation**: Water has the most Negative and the most meaningful correlation respective to the strength of the cement
- **Others**: the other features do have their own correlation value but they arent very meaningful and stay somewhere in the middle

1.2.2) Strength distribution



The distribution of the five strength classes (Very Low, Low, Moderate, Strong, Very Strong) represented by (1, 2, 3, 4, 5) is relatively balanced, with some minor variations.

- The "Moderate" class has the most instances, totaling nearly 250.
- The "Very Low" and "Low" classes have comparable counts, both slightly under 250.
- The "Strong" class has the fewest instances, with approximately 150 samples.
- The "Very Strong" class has over 200 instances but remains below the count for "Moderate."

I have identified an imbalance: The "Strong" class shows a clear disparity compared to the other classes, particularly the "Moderate" class, with a difference exceeding 100 instances. However, none of the differences between any two classes surpass 200 instances, so there isn’t a significant imbalance based on your provided criteria (more than 200 samples difference between two classes).

2) Feature Engineering

Simplify the age feature: The age feature, consisting of integer values, reveals only a limited number of unique values. To simplify model computations, the age feature was converted into categorical values (1, 2, 3, etc.). This transformation was based on its distribution, where unique age values such as 1, 3, and 7 were mapped to corresponding categorical levels (1 for 1, 2 for 3, 3 for 7, and so on).

Normalization of Numeric Features: To standardize the dataset, the numeric features (`cement`, `slag`, `ash`, `water`, `superplastic`, `coarseagg`, `fineagg`) were normalized using the Min-Max scaling technique. The resulting dataset, consisting of these normalized features alongside the categorical `age` feature and then saved as "normalised_concrete.csv".

Creation of Composite Features: Following the recommendations from EDA, four composite features were generated to capture relationships between key variables. The covariance of the normalized values was calculated for the following feature pairs:

- `cement` and `slag` → `cement_slag`
- `cement` and `ash` → `cement_ash`
- `water` and `fineagg` → `water_fineagg`
- `ash` and `superplastic` → `ash_superplastic`

3) Decision Tree model Comparison

```
Accuracy model 1:  0.7119741100323624
Accuracy model 2:  0.6925566343042071
Accuracy model 3:  0.7378640776699029
Accuracy model 4:  0.6084142394822006
Accuracy model 5:  0.6051779935275081
```

Through multiple re-runs, we can conclude that models 1 and 3 seem to perform the best. I have factory-reset and re-executed the full code on Kaggle several times, and models 1 and 3 consistently deliver the best performance, with model 3 slightly outperforming model 1 by a very small margin. Models 4 and 5 consistently struggle to compete with the top three, with no clear indication of which model consistently performs the worst.

4) Appendix

Here is the link to the kaggle notebook:
https://www.kaggle.com/code/binhswinburnehn/cos40007-week-2