Final Project

Mohamed Eraky: 20017275 | **Mahmoud Nassar:** 20008506 | **Anish Khilani:** 20008972

In this project, a linear regression model was constructed to predict the concrete strength given a span of features such as cement quantity, age, and water. First, the data frame has been cleaned from duplicates, and outliers are removed using Interquartile range (IQR). The data frame is then standardized/normalized. The correlation matrix is computed to identify the features that have a remarkable correlation with the concrete strength as a target. The standardized/ normalized data frame has been split between training and testing data sets. A linear regression model is then trained to fit the data to a linear model with coefficient matrix $w = [w_i], i = 0,1,2,..$ to minimize the residuals between the trained "strength-data set" and the strength predicted by linear fit. Model performance metrics such as RMS (that measures the average magnitude of errors) and accuracy have been reported while testing different features.
Further, a PCA model was built to reduce the dimensionality of the feature matrix, where the eigen values and their corresponding eigen vectors are computed for the covariance matrix and then reordered in a descending order to identify the features with highest weight. In this analysis, two features are considered, whereby linear fit model is trained using a similar procedure mentioned above.
The RMS and model Accuracy are reported in Table 1.

Table 1: Models Performance Metrics

	Linear Regression Model	PCA Model (k=2)
RMS	0.473	15.248
Accuracy	80%	22%

Design Choices

Based on the correlation matrix, the following features have been used to train the linear regression model, where model performance metric (RMS, Accuracy) have been computed.

- Testing three features (Cement, Superplasticizer, and Age) with the absolute highest correlation.
  - Normalized RMS = 0.126, Model accuracy = 65.7 %.
- Testing four features (Cement, Superplasticizer, Age, and Water) with the absolute highest correlation.
  - Normalized RMS = 0.123, Model accuracy = 67.74%.
- Testing all features, where the lowest correlated feature, namely "Fly Ash" has been removed.
  - Normalized RMS = 0.0961, Model accuracy = 80.01 % (Best model)

For the PCA model we find that reducing the dimensionality from 8 to 2 caused the model to deteriorate rapidly. The accuracy decreased by 60% and the RMS ramped up. Therefore, the PCA model is not effective in this case, otherwise the dimensionality must be increased. Based on the computed variance, a dimensionality of K= 6, gives somewhat acceptable 60 % accuracy.