Power Plant Energy Output Prediction and Analysis

The project involved analyzing the Combined Cycle Power Plant Data Set, which contains data points collected over 6 years to predict the hourly electrical energy output of the plant.
The following tasks were performed:

(a) Data Exploration:
The dataset was downloaded, and the number of rows and columns were determined. Pairwise scatterplots were created for all variables, and the mean, median, range, and interquartile ranges of each variable were summarized in a table.

(b) Regression Models:
Univariate and multiple linear regression models were built, and the association between the predictor and response variables was determined. Outliers were removed, and a full linear regression model with all pairwise interaction terms was run. The possibility of improving the model using interaction terms or nonlinear associations was explored, and the regression model was tested on a randomly selected 70% subset of the data.

(c) KNN Regression:
K-nearest neighbor regression was performed on the dataset using normalized and raw features, and the value of k that gave the best fit was determined. The results were compared with the linear regression model that had the smallest test error.

Overall, the project was successful in predicting the hourly electrical energy output of the plant and reducing testing error of 20% by removing non-significant variables.