Project Description

Dataset

The dataset used for the final project is 'concrete.csv'. The total raw data in the dataset is 1030. Data cleaning was done and an additional column for water-cement ration was added using the division formula in excel. The cleaned dataset 'concrete_clean.csv' having a total of 968 data and 10 columns *(cement, water, coarse aggregate, fine aggregate, age of testing, fly ash, superplasticizer, blast furnace slag, w/c ration & strength)* were used for executing the entire project.

Exploratory Data Analysis

In doing the exploratory data analysis the followings were found to analyze the data:

Missing values, the data type in each column, descriptive statistics summary, variable types count for numerical Variable, a boxplot of the entire dataset, a boxplot of Dependent Variable vs Discrete Independent Variable, histogram distribution of dependent variable vs discrete independent variable, histogram distribution plots of all independent continuous variables, swarm plot & violin plot for the dependent variable, skewness, and kurtosis of the variables, pair scatter Plot, joint plot (kernel density distribution included), empirical cumulative distribution function, correlation matrix and, model entropy

Models for Attribute Selection

2 models were performed for selecting the important features using model training:

- Random Forest Regression
- Gradient Boosting Regression

Model Training

For all the models the entire dataset was split into 75% training and 25% testing set. 5-fold cross-validation was done with the training data and 'negative mean square' error was taken as the accuracy metric to compare the models. Two methods were used for model training against the accuracy of the test set:

- Linear Regression
- Deep Neural Network

Metric Measurement

The following metrics were used as the base to compare different models:

- Mean Squared Error
- Mean Absolute Error
- Correlation
- Kling Gupta Efficiency Metric
- Cross-Validation Scoring

Dataset Description

The types of data included in the dataset are given below:

Table 01: Types of Features

Total Data	968
Output Feature (Continuous) – 1	Strength
Input Feature (Discrete) – 1	Age of Testing
Input Feature (Continuous) – 8	Cement, Water, Water / Cement Ration, Coarse Aggregate, Fine Aggregate, Fly Ash, Blast Furnace Slag, Superplasticizer

The strength of the cement is calculated and collected for different age days.

Age of Testing Collected – 1, 3, 7, 14, 28, 56, 90, 91, 100, 120, 180, 270, 360, 365

No 'NA' and missing values are further present in the dataset.

The boxplot of all the independent and dependent variables are provided below:

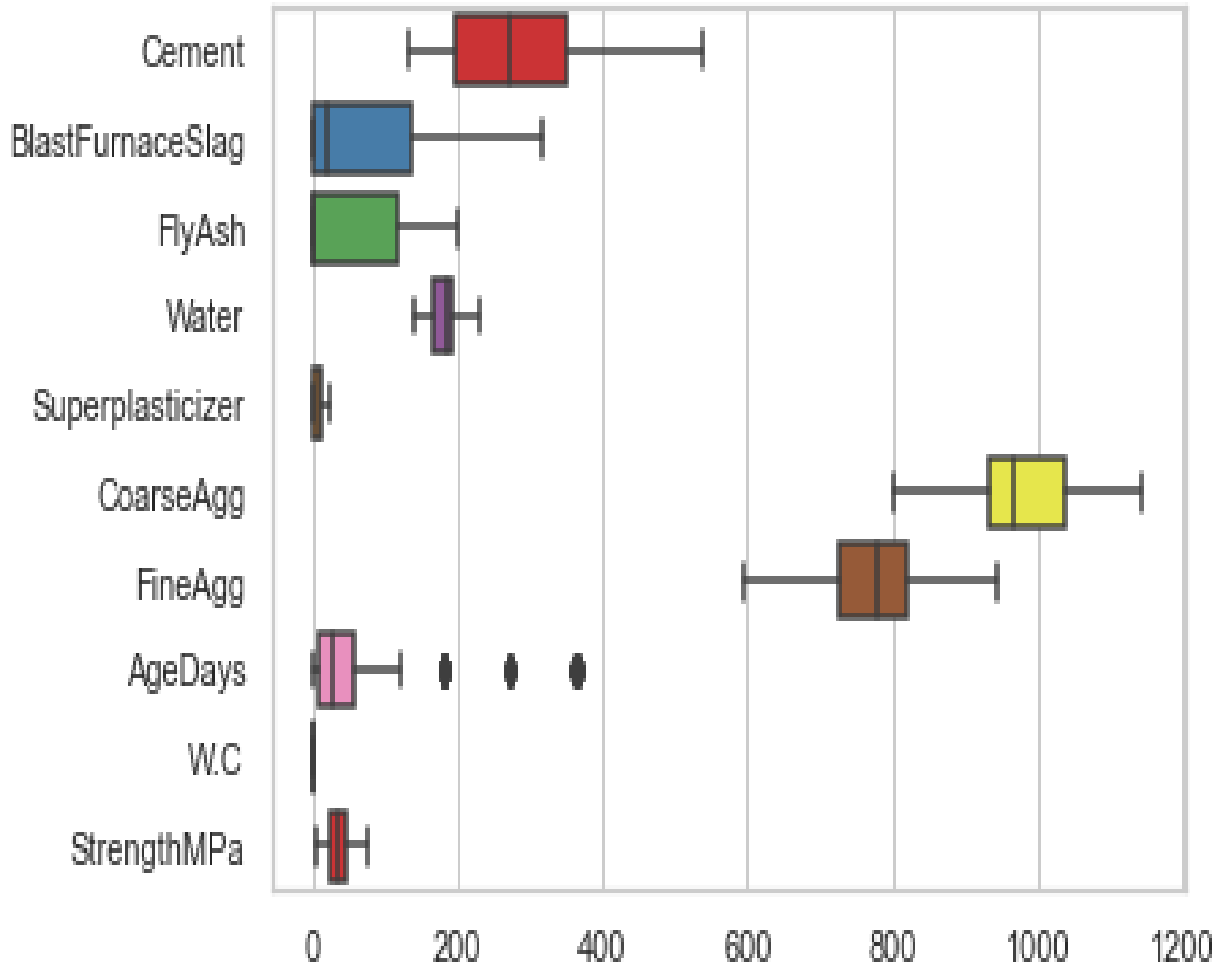

Figure 01: Distribution of the Data in Boxplot

Table 02: Statistics Summary

Description Summary	Cement	Water	W/C	Coarse Aggregate	Fine Aggregate	Fly Ash	Blast Furnace Slag	Superplasticizer	Age of Testing	Strength of the Concrete
Min.	132	140	0.2797	801	594	0	0	0	1	4.78
1st Qu.	194.7	166.7	0.5457	932	727.9	0	0	0	7	23.68
Median	272.8	185.7	0.6807	968	778.5	0	19	6.4	28	33.95
Mean	280.9	182.5	0.7414	975	771.1	56.29	69.53	5.892	46	33.26
3rd Qu.	349	192	0.9350	1038.5	821	118.3	135.7	10	56	45.08
Max	540	228	1.4805	1145	945	200.10	316.1	22	365	76.24
Std.	101.447	19.33	0.283	77.87	78.426	64.176	84.520	5.33	64.63	15.81

Description of the Output feature

The output feature used in the dataset is the strength of the concrete. The distribution of the strength of the concrete concerning the Age of testing is shown through the histogram distribution plot and boxplot.

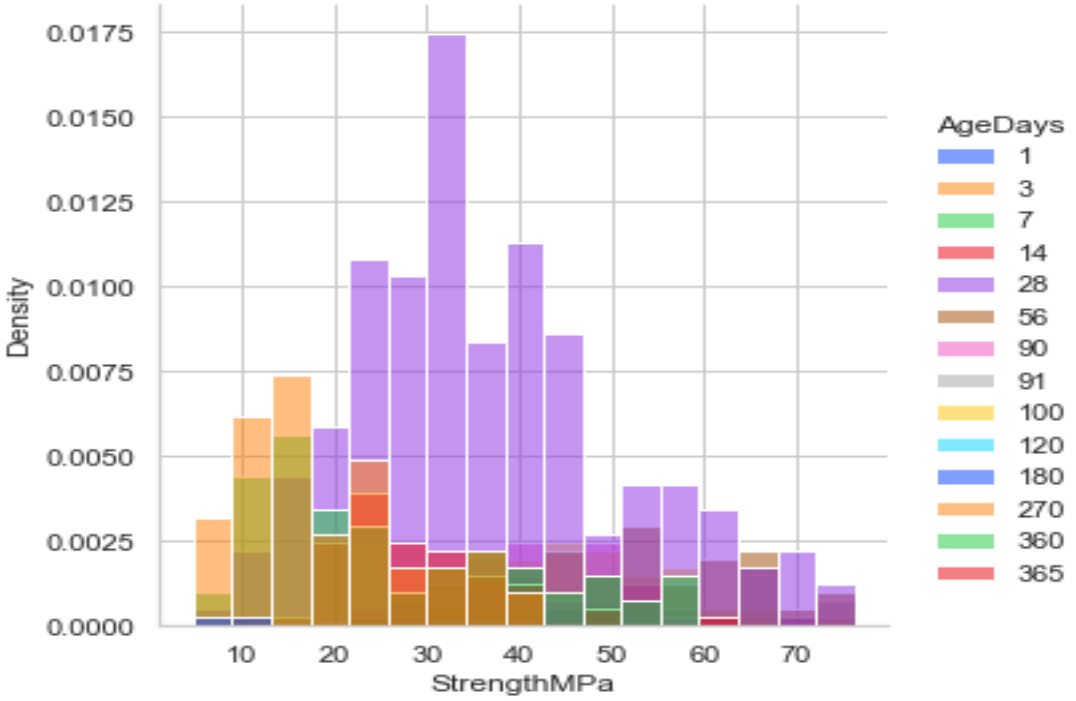

Figure 02: Distribution of Strength Vs Age of Testing

It is evident from the distribution plot that the concrete strength at Age Days = 28 is largely presented in the dataset. Age Days = 28, 100 are in large numbers while the concrete strength for Age Days = 1, 365, 120, 180 are low in numbers. So, the dataset will provide more information on the aforementioned age days having a high presence with a larger density than the ones having a lower density. Similarly, the strength of the concrete between the range 20 ~ 50 has a higher presence in the dataset concerning the other points among the data.

The distribution of the Strength of the concrete concerning the age of testing is shown in the boxplot below. The dividing median line states the symmetry of the strength data present in each set of the age of testing.

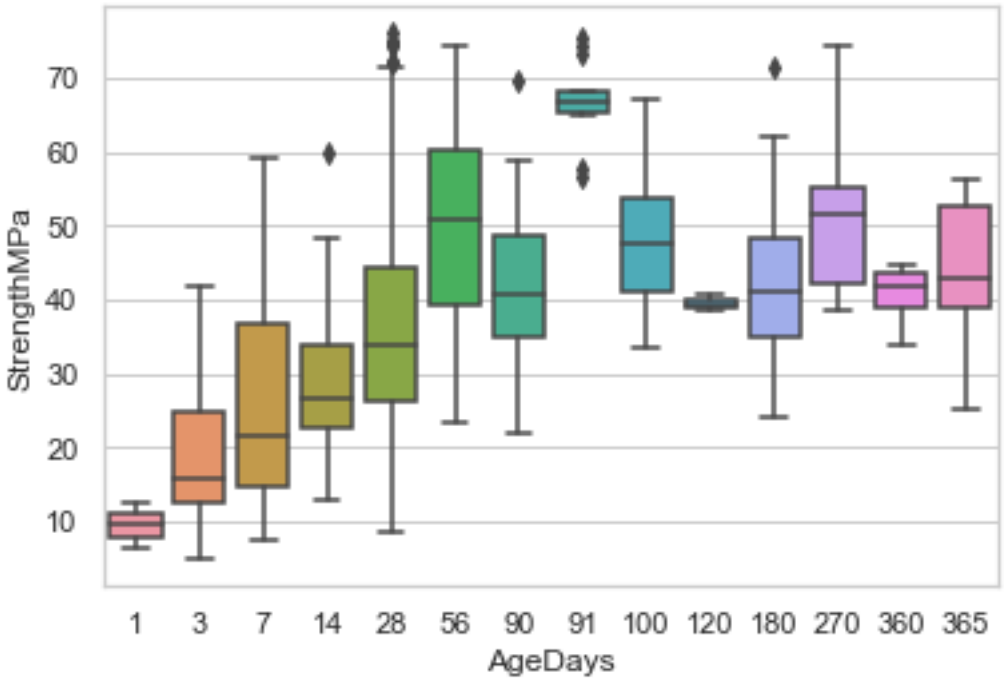

Figure 03: Distribution of the Concrete Strength Data in Boxplot

Normal Distribution & Probability Plot

The probability plot and the density distribution in the histogram are provided for all the features including the input features and output feature. Skewness and kurtosis values are provided to further study the skewness of the data and to what extent the normal distribution is followed by the data points.

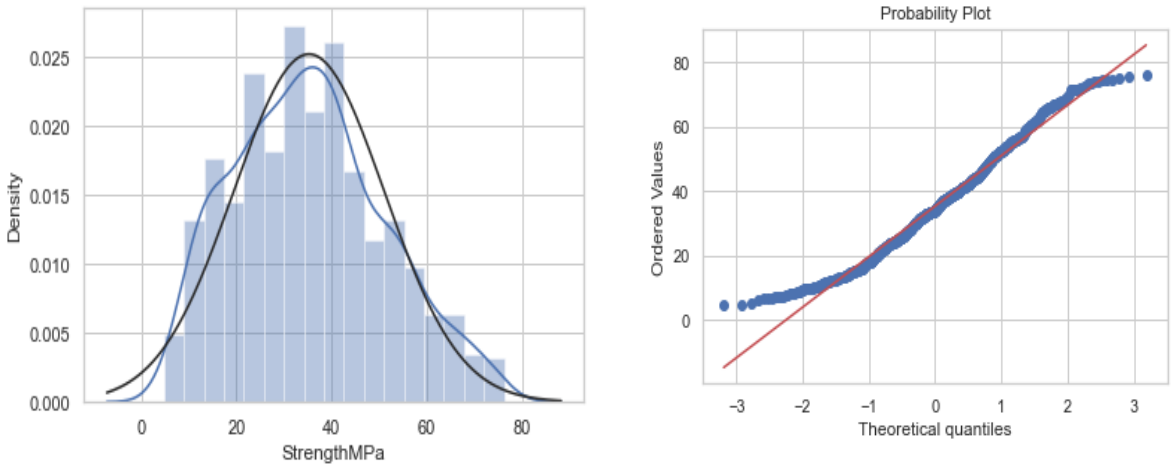

Figure 04: Distribution & Probability Plot for Concrete Strength

The skewness value = 0.326192 and the kurtosis value = -0.489804

Positive skewness is present so the data is positively skewed. The outlier data is less extreme and the kurtosis value tells that the data distribution is light-tailed.

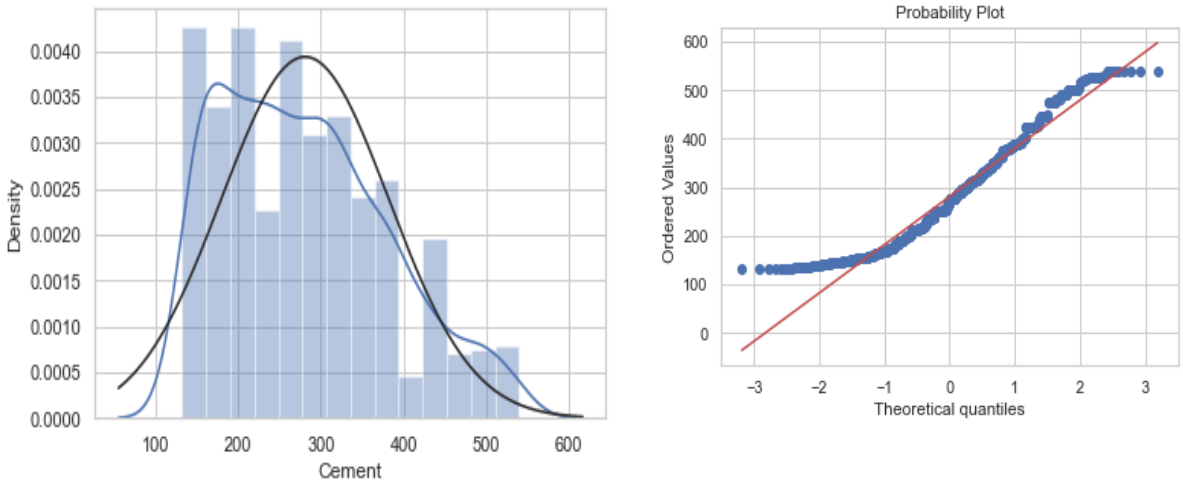

Figure 05: Distribution & Probability Plot for Cement

The skewness value = 0.548662 and the kurtosis value = -0.481464

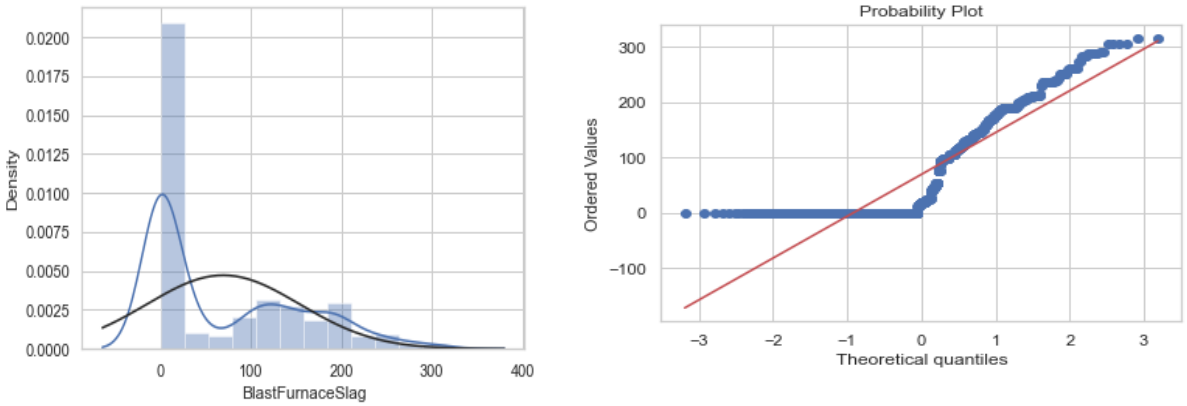

Figure 06: Distribution & Probability Plot for Blast Furnace Slag

The skewness value = 0.846263 and the kurtosis value = -0.558376

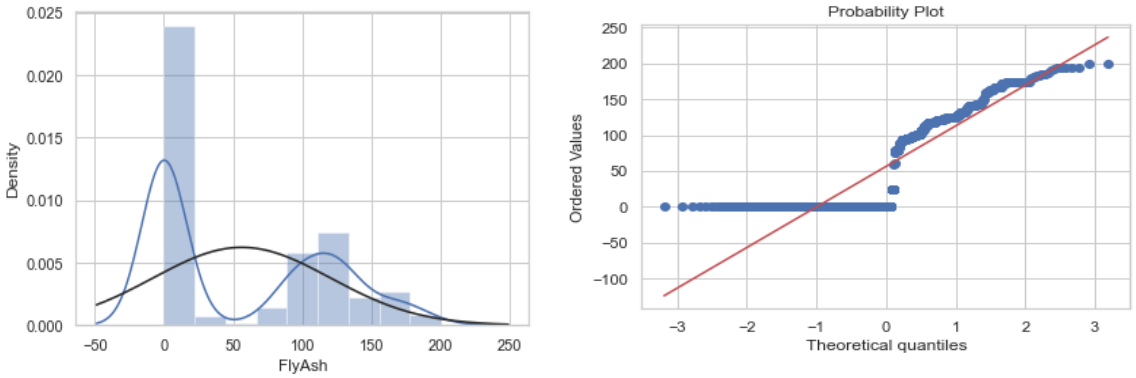

Figure 07: Distribution & Probability Plot for Fly Ash

The skewness value = 0.471492and the kurtosis value = -1.385251

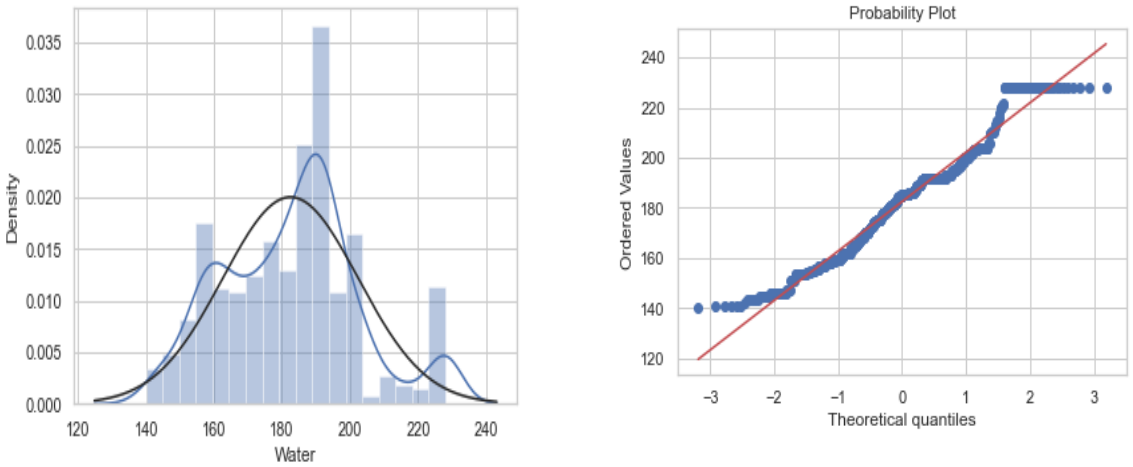

Figure 08: Distribution & Probability Plot for Water

The skewness value = 0.244322 and the kurtosis value = -0.147664

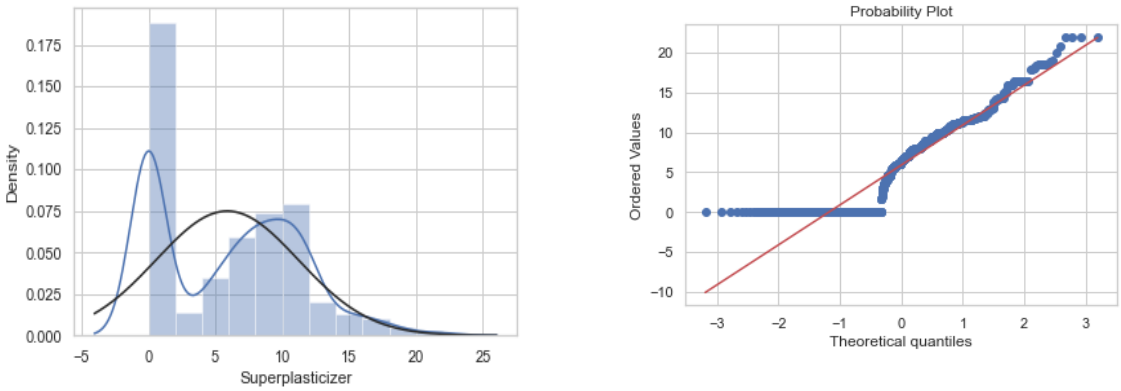

Figure 09: Distribution & Probability Plot for Superplasticizer

The skewness value = 0.337004and the kurtosis value = -0.839792

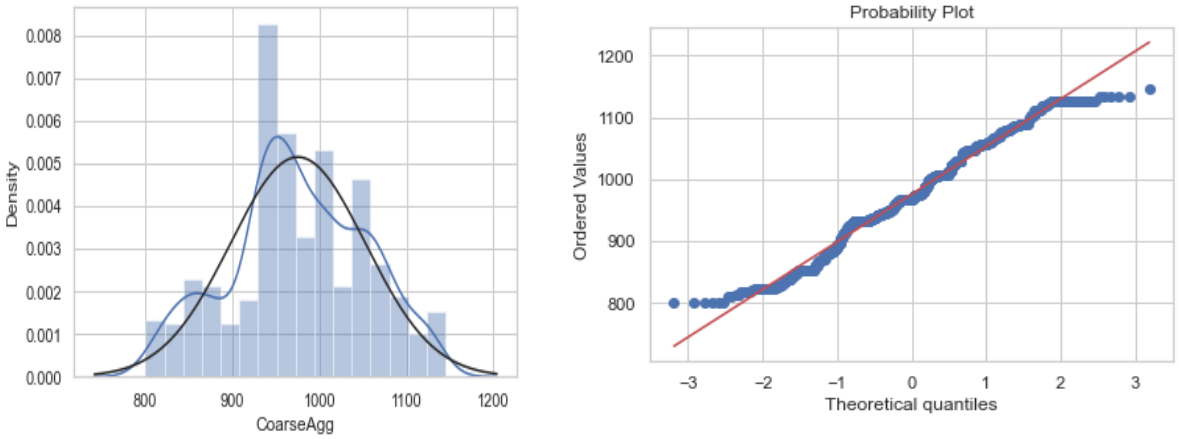

Figure 10: Distribution & Probability Plot for Coarse Aggregate

The skewness value = -0.088383 and the kurtosis value = -0.549960

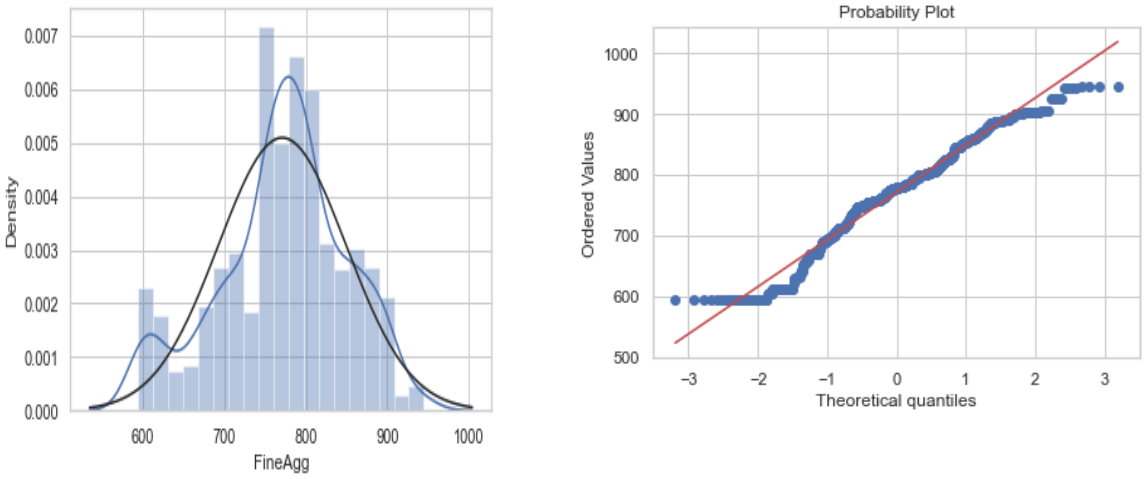

Figure 11: Distribution & Probability Plot for Fine Aggregate

The skewness value = -0.360099 and the kurtosis value = -0.181768

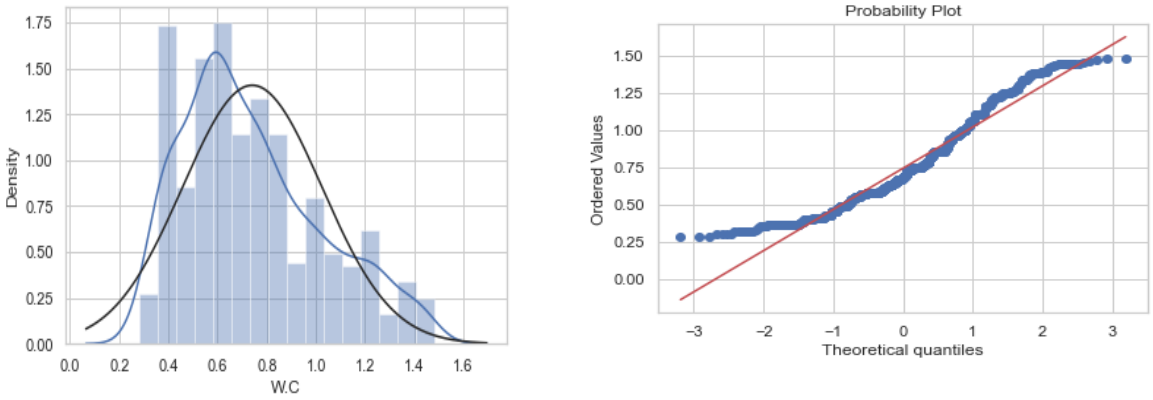

Figure 12: Distribution & Probability Plot for Water Cement Ratio

The skewness value = 0.665493 and the kurtosis value = -0.347714

Swarm Plot for Output Feature

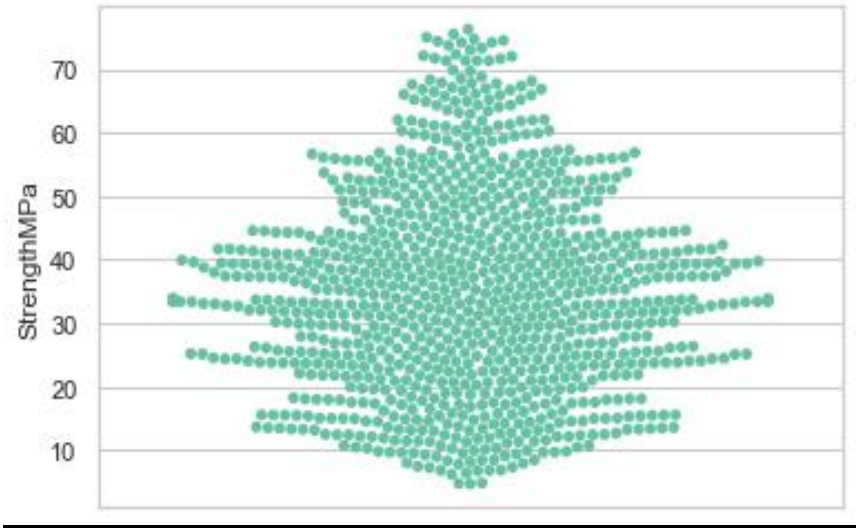

Figure 13: Swarm plot Distribution for Strength

From the swarm plot, it is evident that the concrete strength between 20 MPa and 50 MPa are widely present and collected in the dataset. The data is concentrated on this section. The swarm plot of the strength data is plotted against the age of testing and show on top of the violin plot.

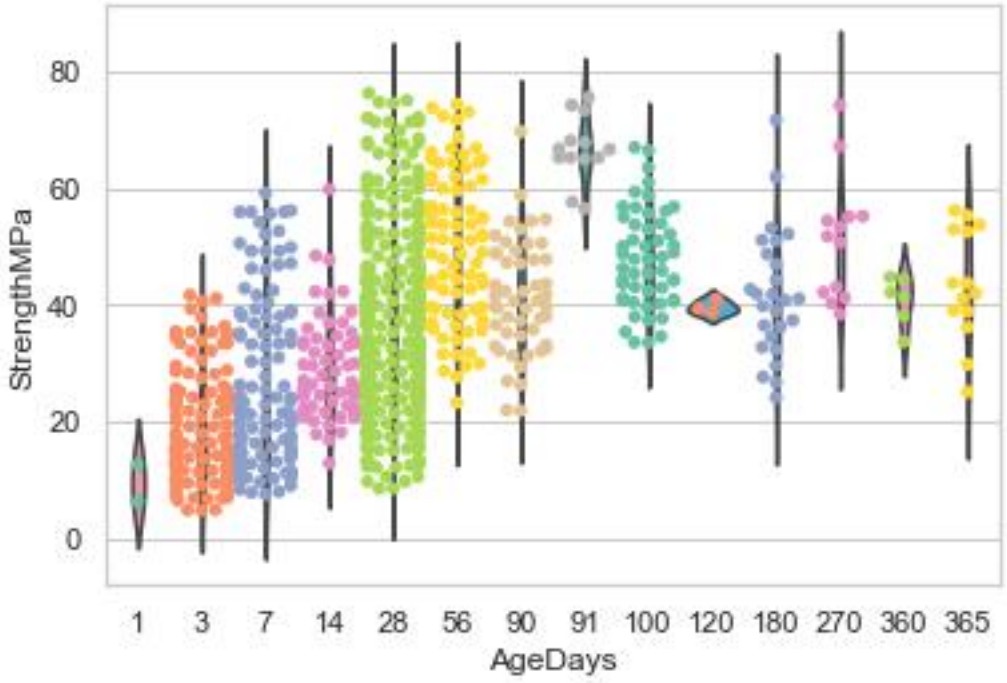

Figure 14: Swarm plot and Violin Plot Distribution for Strength

Scatter Plot

The scatter plot of all the dependent and independent variables is shown by pairing with each other. The pairing is divided into two parts and the scatter relations are shown among them using the pair plot command.

- Pair Plot 1 - Cement, Water, Coarse Aggregate, Fine Aggregate, Water-Cement Ratio and Strength of Concrete
- Pair Plot 2 – Blast Furnace Slag, Fly Ash, Superplasticizer, and Strength of Concrete

The scatter relationships among the variables are carefully observed and inferred how one is related to one another.

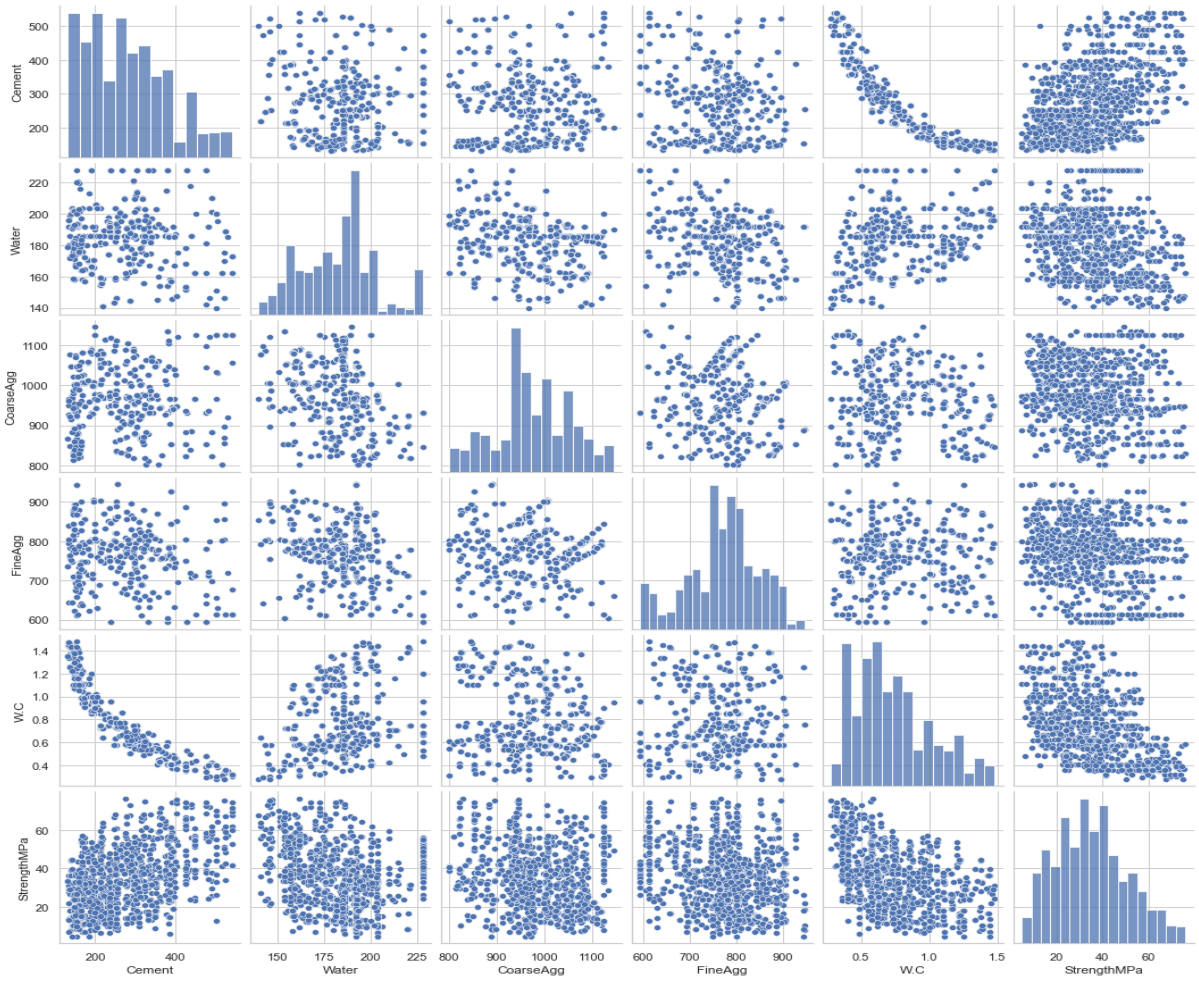

Figure 15: Scatter Plot among Cement, Water, Coarse Aggregate, Fine Aggregate, W/C, and Strength

From the pair plot above, the water-cement ratio is found to be correlated with the strength of the concrete and cement. Cement and water are also found to be correlated with the strength of the concrete. A slight relationship between the data of water and fine aggregate are found as well.

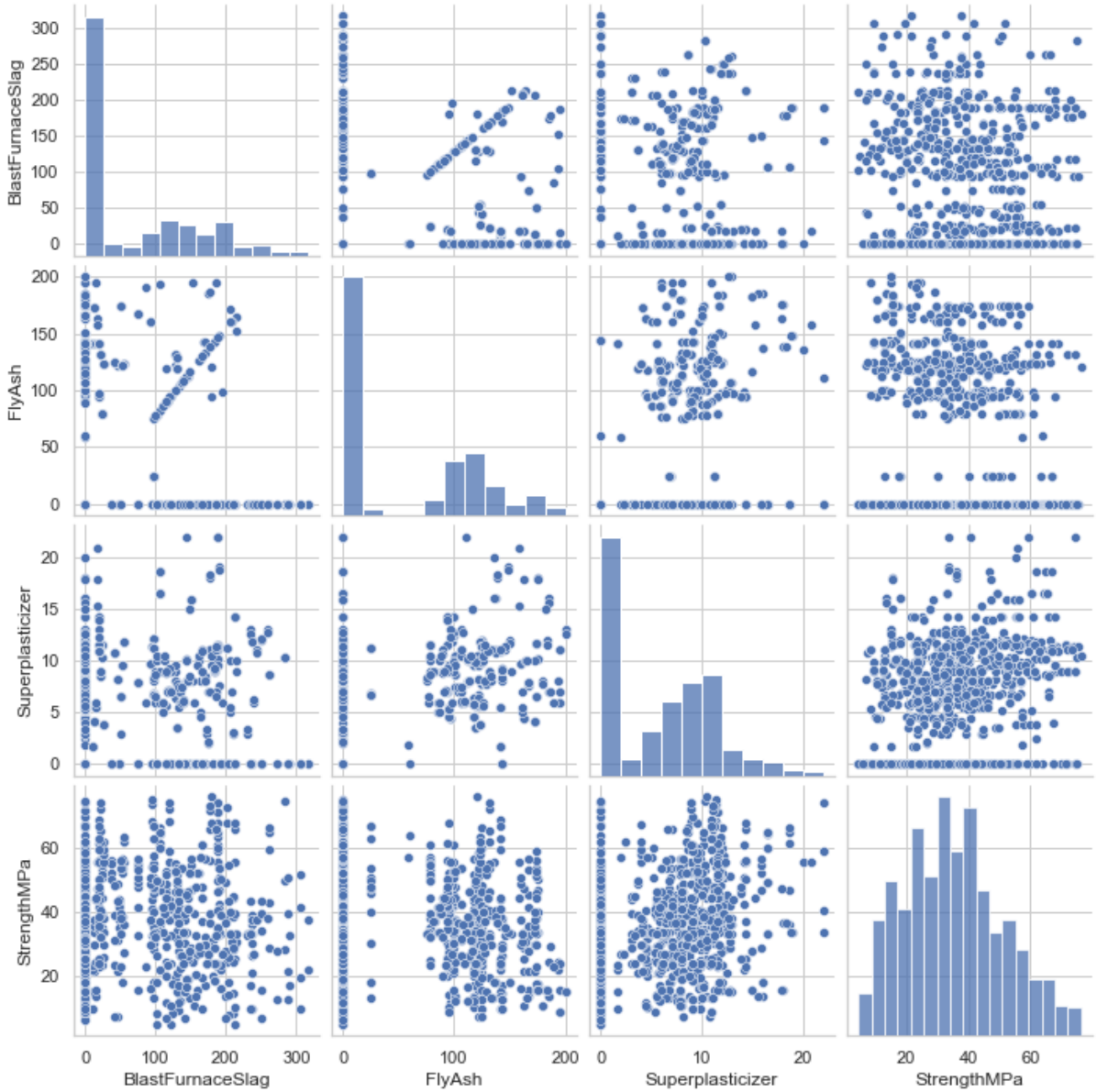

Figure 16: Scatter Plot among Blast Furnace Slag, Fly Ash, Superplasticizer, and Strength

From the scatter plot above it is found that all the sets of the Blast furnace, fly ash and superplasticizer have many 0s in their sets.

Joint Plot

The distribution of all the dependent and independent variables including kernel density estimation is shown by pairing with each other. The pairing is divided into two parts and the joint relations are shown among them using the pair grid command.

- Joint Plot 1 - Cement, Water, Coarse Aggregate, Fine Aggregate, Water-Cement Ratio and Strength of Concrete
- Joint Plot 2 – Blast Furnace Slag, Fly Ash, Superplasticizer, and Strength of Concrete

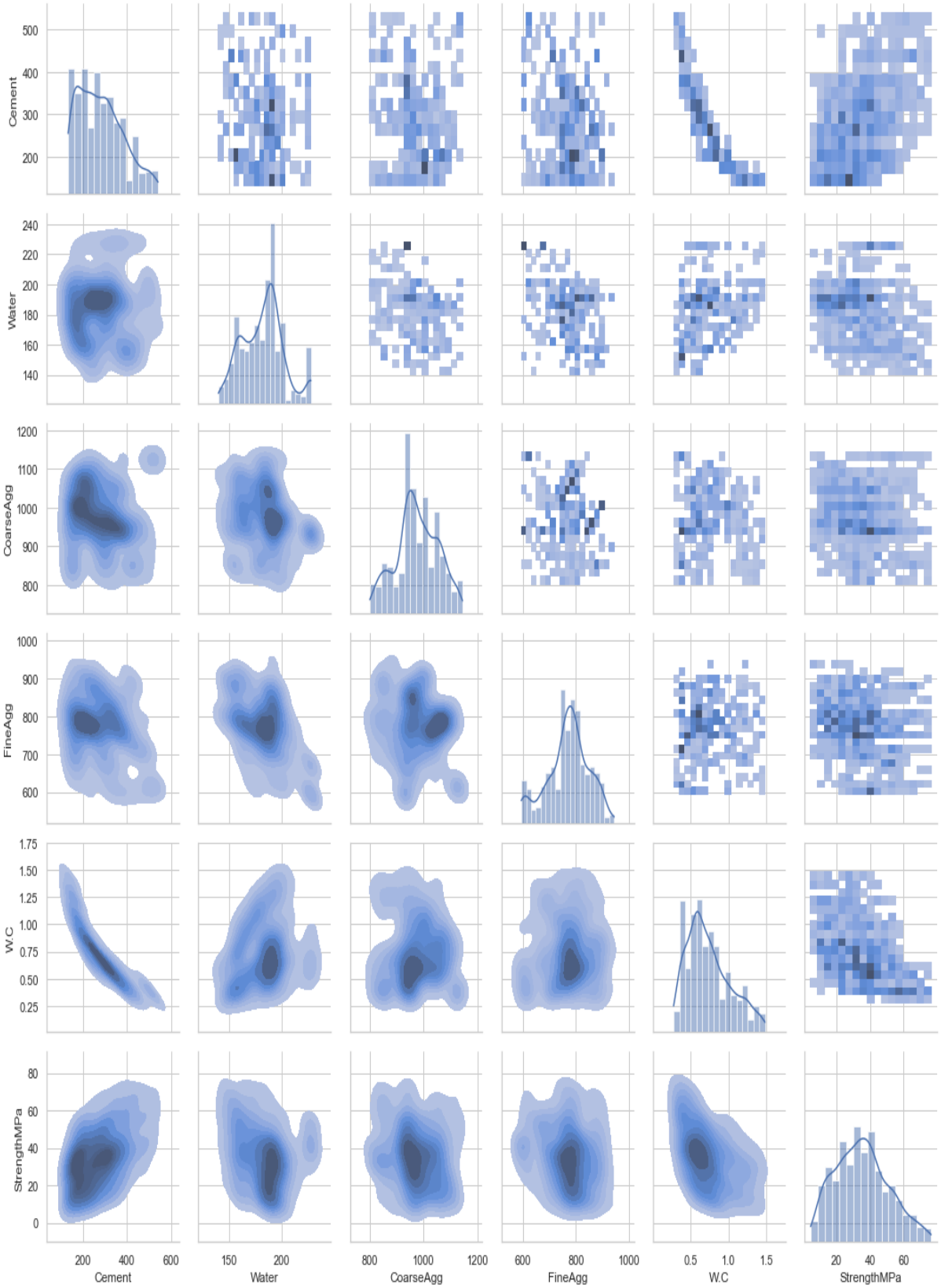

Figure 17: Bivariate Relationship among Blast Furnace Slag, Fly Ash, Superplasticizer and Strength

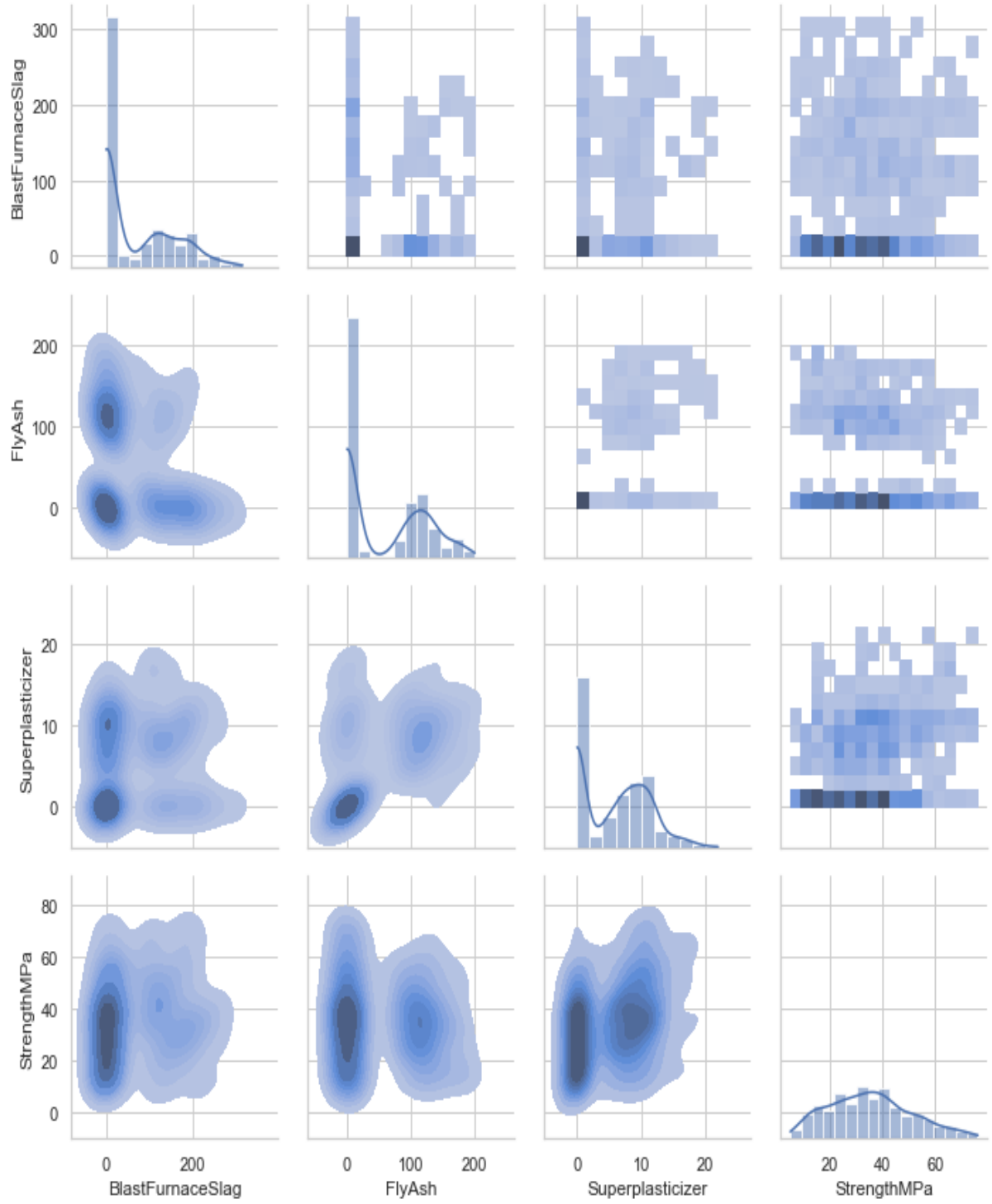

Figure 18: Bivariate Relationship among Blast Furnace Slag, Fly Ash, Superplasticizer and Strength

Empirical Cumulative Distribution of the Dependent Variable

The ECDF curve of the output variable (Strength of the concrete) is shown below.

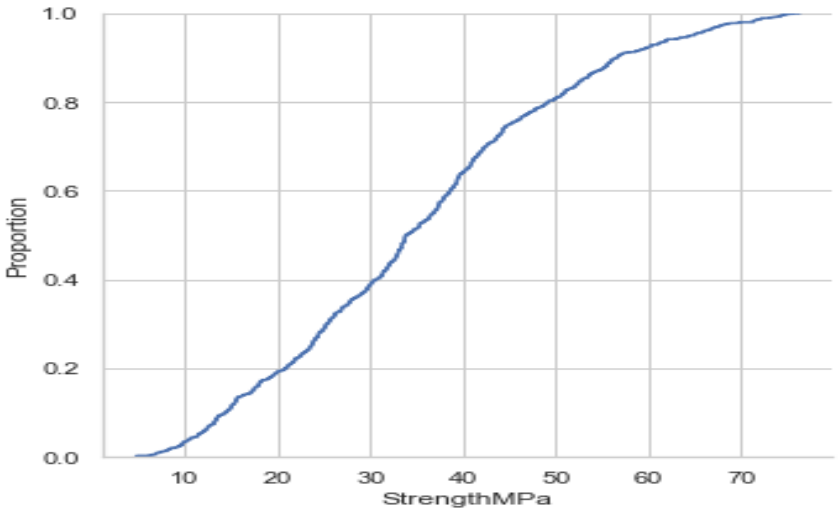

Figure 19: ECDF curve of concrete strength

Model Entropy & Correlation Matrix

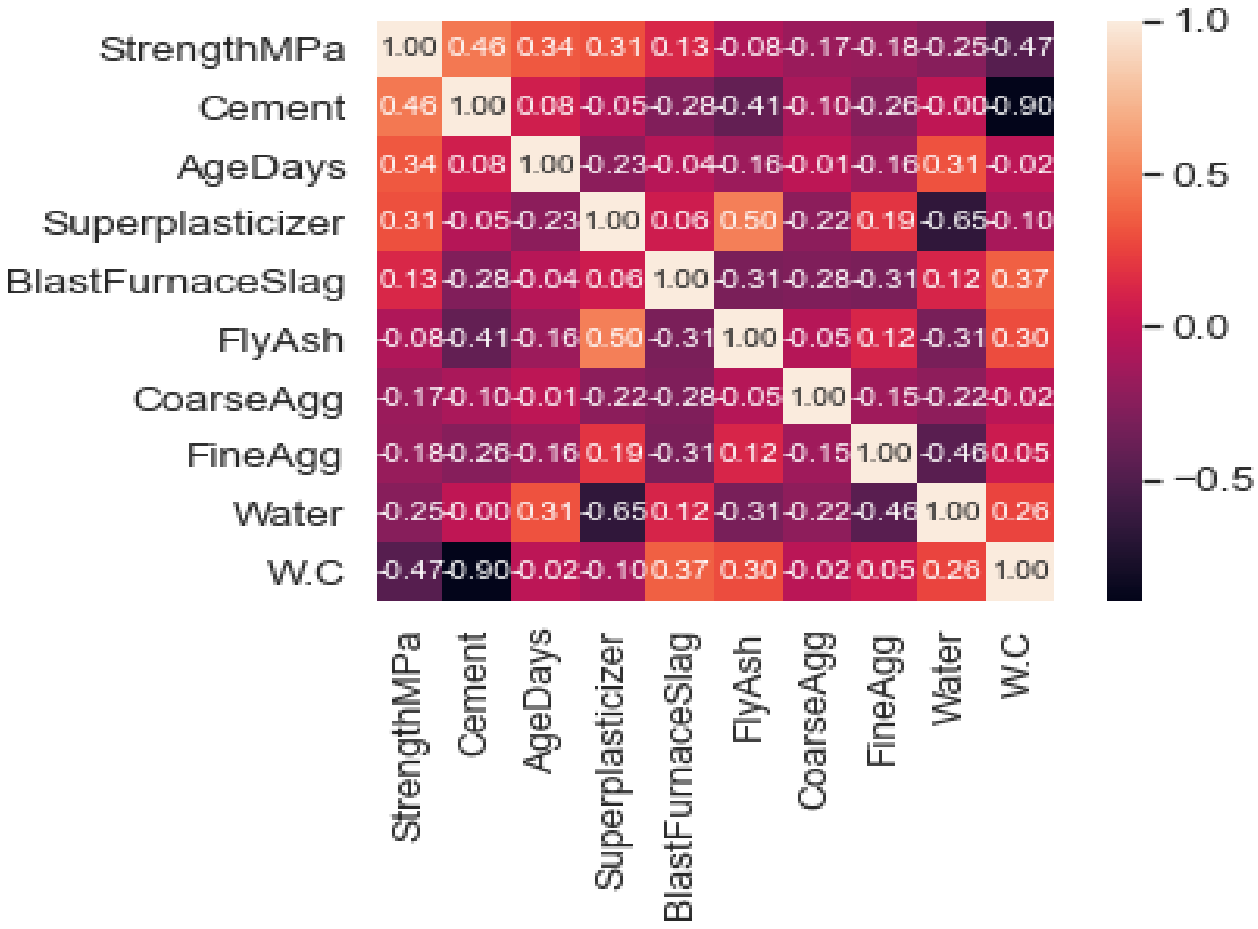

Figure 20: Correlation Matrix

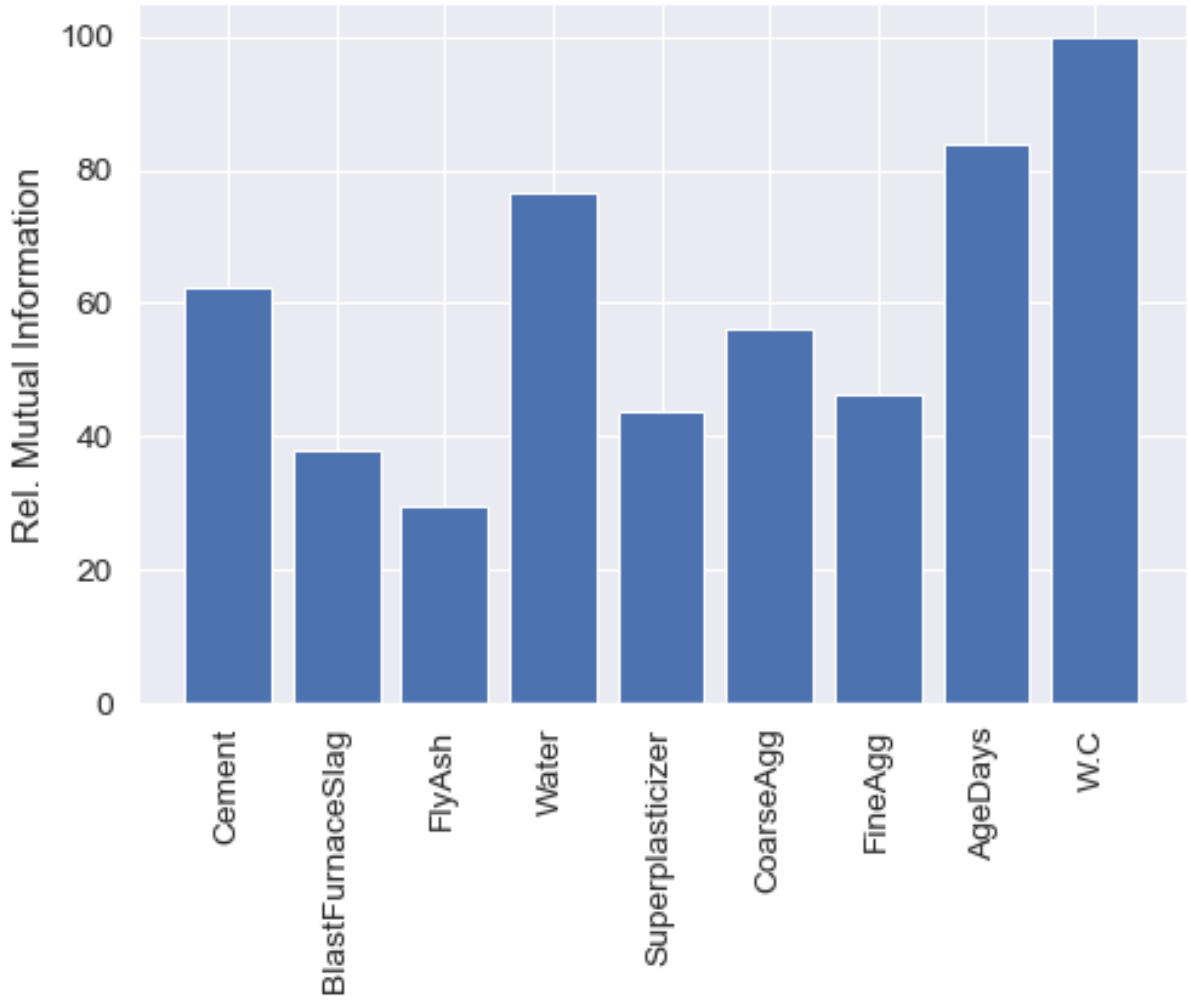

Figure 21: Model Entropy

None of the input variables shows a high correlation with each other in the above-mentioned correlation matrix. The highest correlation value against the dependent variable is found to be that of cement. From the model entropy distribution, the age of testing and water-cement ratio shows high relative importance. Water and cement both show a good degree of relative importance as well.

Model Building for Attribute Selection

For determining the relevant important features two models are used:

- Random Forest Regression
- Gradient Boosting Regression

For the random forest regression model, maximum depth with range (3,8) and trees size of 10, 50, 100, and 500 are used with a grid search and 5-folds cross-validation.

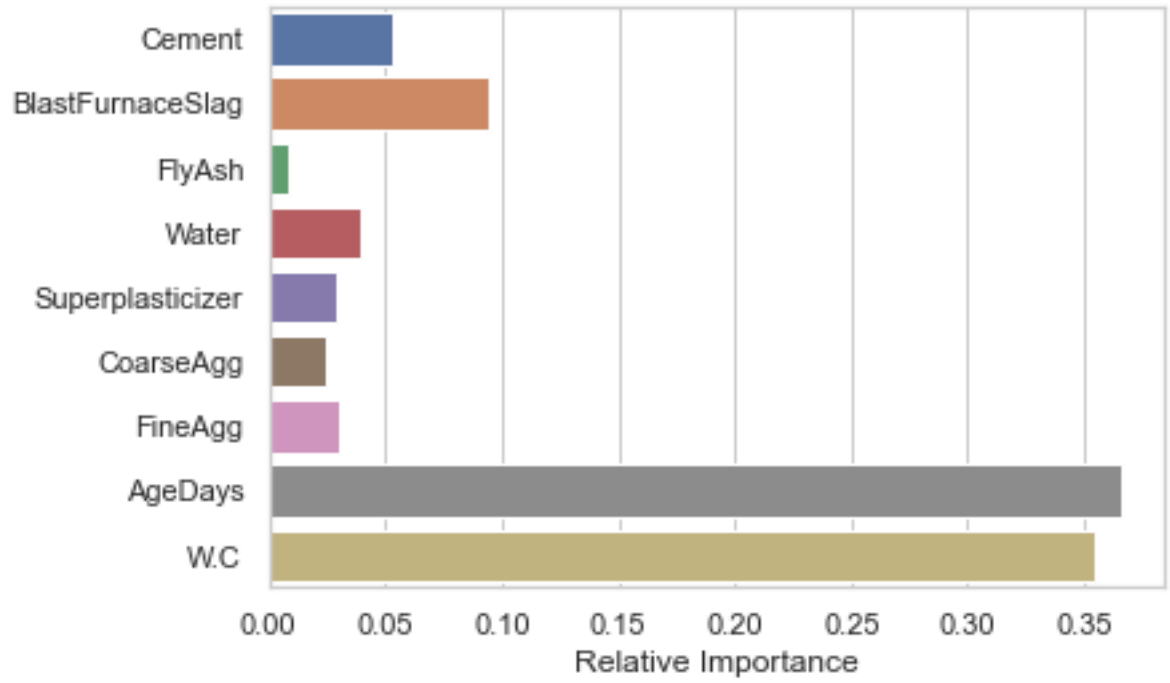

Figure 22: Relative Importance Plot from Random Forest

Important Features detected: Age of Testing, Water-Cement Ration & Blast Furnace Slag

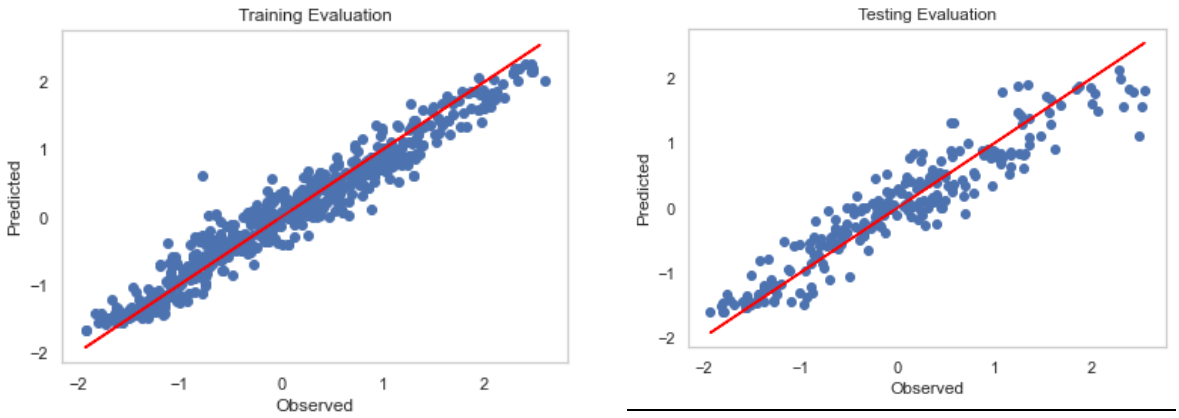

Figure 23: Model Evaluation for Random Forest Regression

Table 03: Model Evaluation Chart for Random Forest Regression

Metric Parameter	Train	Test
MSE	0.05374644	0.11675901
MAE	0.17477476	0.26844834
COR	0.97425295	0.94336995
Overall (KGE)	0.52664807	0.33417977
COR (KGE)	0.97425295	0.94336995
VAR (KGE)	0.67912883	0.58618214
BIAS (KGE)	1.34704586	1.51852136

For gradient boosting regression model, maximum depth with learning rate 0.1 and trees size of 10, 50, 100, and 500 are used with a grid search and 5-folds cross-validation.

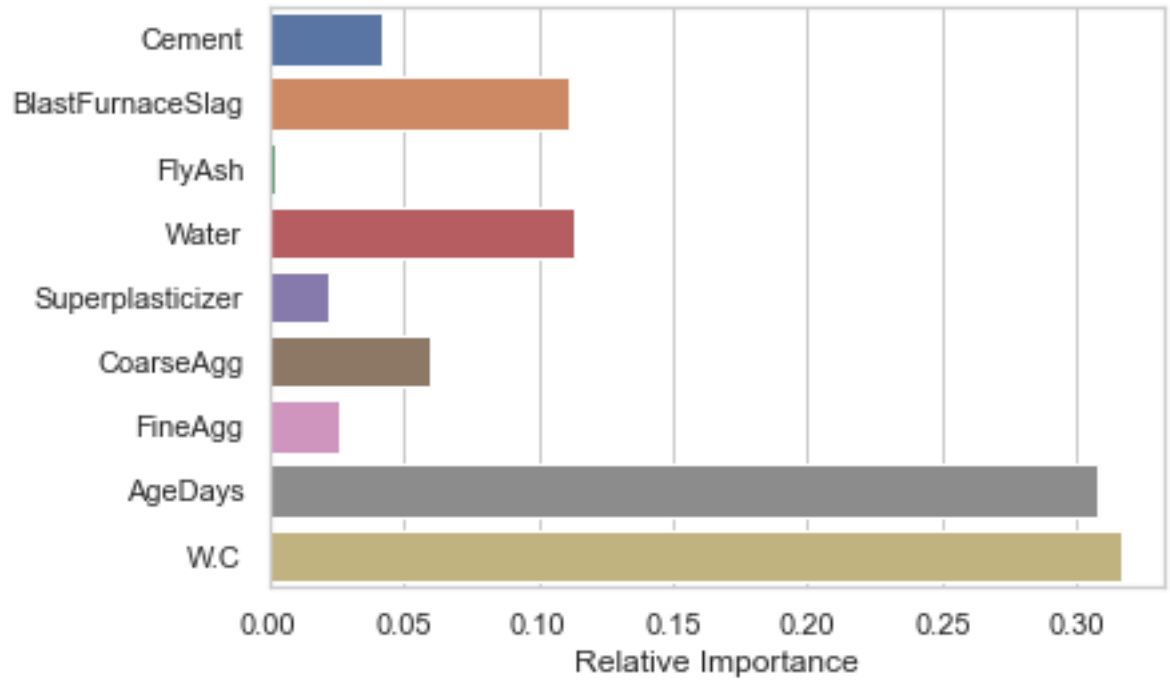

Figure 24: Relative Importance Plot from Gradient Boosting Regression

Important Features detected: Age of Testing, Water-Cement Ration, Water & Blast Furnace Slag

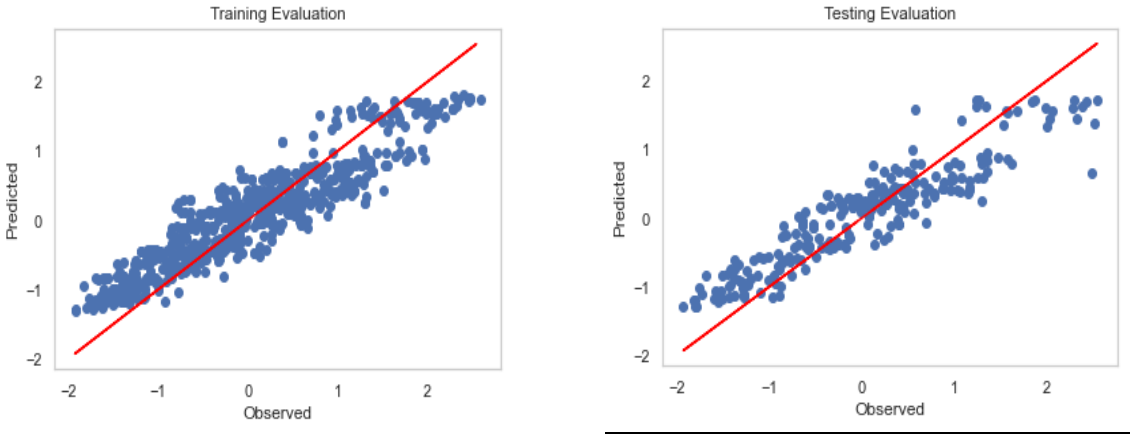

Figure 25: Model Evaluation for Gradient Boosting Regression

Table 04: Model Evaluation Chart for Gradient Boosting Regression

Metric Parameter	Train	Test
MSE	0.17407446	0.19418796
MAE	0.34754941	0.36258837
COR	0.92739989	0.92041282
Overall (KGE)	-14.49033626	-3.7261570
COR (KGE)	0.92739989	0.92041282
VAR (KGE)	-0.05165006	0.13122221
BIAS (KGE)	-14.45442586	5.64493822

Linear Regression

Standard Scaler was used to scale the data. The dataset was divided into 75% training and 25% testing set. All 9 input features were used to train the linear regression model. 5-folds cross-validation with negative mean squared error as a metric was used to calculate the scoring value of the training set.

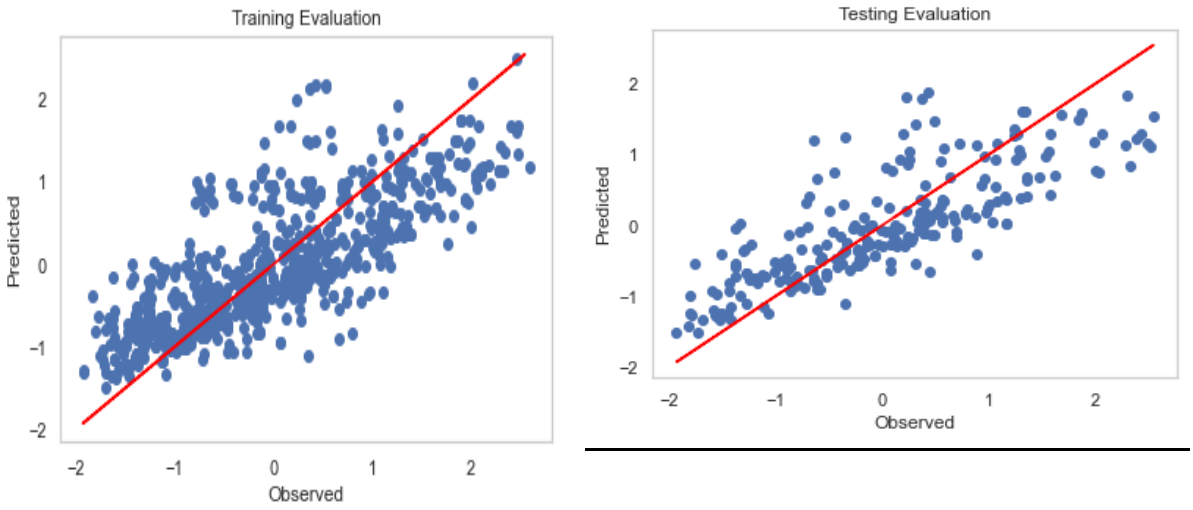

Figure 26: Model Evaluation for Linear Regression Model

Table 05: Model Evaluation Chart for Linear Regression

Metric Parameter	Train	Test
MSE	0.40082682	0.37724051
MAE	0.50191415	0.48206014
COR	0.77107496	0.79831873
Overall (KGE)	0.6762511	-1.73116804
COR (KGE)	0.77107496	0.79831873
VAR (KGE)	0.77107496	-0.65018309
BIAS (KGE)	1.0	-1.16691008

The scores of 5-folds cross-validation: -0.45570235, -0.3583923, -0.36762589, -0.43486129, -0.45009193

Mean cross validation value: -0.4133347540

Deep Neural Network Model

The compressive strength of concrete is a function of the following nine input features:

1. Water
2. Cement

3. Water – Cement Ratio
4. Coarse Aggregate
5. Fine Aggregate
6. Age of Testing
7. Blast Furnace Slag
8. Fly Ash
9. Superplasticizer

- **6 Models are used in this approach**

1. Model 1 – 9 Input Model (water, cement, age of testing, water-cement ratio, coarse aggregate, fine aggregate, blast furnace slag, fly ash, superplasticizer)
2. Model 2 – 8 Input Model (water, cement, age of testing, water-cement ratio, coarse aggregate, fine aggregate, blast furnace slag, fly ash)
3. Model 3 – 7 Input Model (water, cement, age of testing, water-cement ratio, coarse aggregate, fine aggregate, blast furnace slag)
4. Model 4 – 5 Input Model (water, cement, age of testing, water-cement ratio, blast furnace slag)
5. Model 5 – 4 Input Model (cement, age of testing, water-cement ratio, water)
6. Model 6 – 4 Input Model (water, age of testing, water-cement ratio, blast furnace slag)

- **Network Parameters**

  Number of Hidden Layers for each model = 2
  Number of neurons in the first hidden layer = 3n
  Number of neurons in the second hidden layer = n
  Activation function used for both layers = Relu
  Learning Cycles = 250
  Batch Size for each cycle = 15
  Regularizer used = L2

- **Model Evaluation**

  All the provided 6 models are evaluated using deep neural networks and compared concerning the given metrics:

  1. MSE
  2. MAE
  3. Kling – Gupta Efficiency Metrics
  4. Cross-Validation Scores

Table 06: Model Evaluation Chart for Neural Network

Metric Parameter	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
MSE	0.099011	0.1137156	0.1290255	0.155980	0.240961	0.18290452
MAE	0.237379	0.2441834	0.2755907	0.317846	0.395183	0.34363736
COR	0.952428	0.9439040	0.9362370	0.923223	0.876393	0.90791786
Overall (KGE)	-2.02634	-1.081621	-1.881087	-4.05346	-2.06601	-6.460660
COR (KGE)	0.95242	0.9439040	0.9362370	0.923223	0.876393	0.9079178
VAR (KGE)	-0.6493	0.3251373	0.2525994	0.156076	0.220167	-6.372876
BIAS (KGE)	-1.5369	2.9683911	3.7817243	5.981907	3.962603	-0.137390
CV Score (Mean)	-0.11777	-0.13159	-0.13649	-0.16007	-0.28492	-0.181573

From the six models provided above, Model 1 shows the greatest accuracy for the testing data concerning the provided metrics. The used parameters for Model 1:

- Input features = 9
- Number of Hidden Layers = 2
- Number of neurons in the first hidden layer = 27
- Number of neurons in the second hidden layer = 9
- Activation function used for both layers = Relu
- Learning Cycles = 250
- Batch Size for each cycle = 15
- Regularizer used = L2

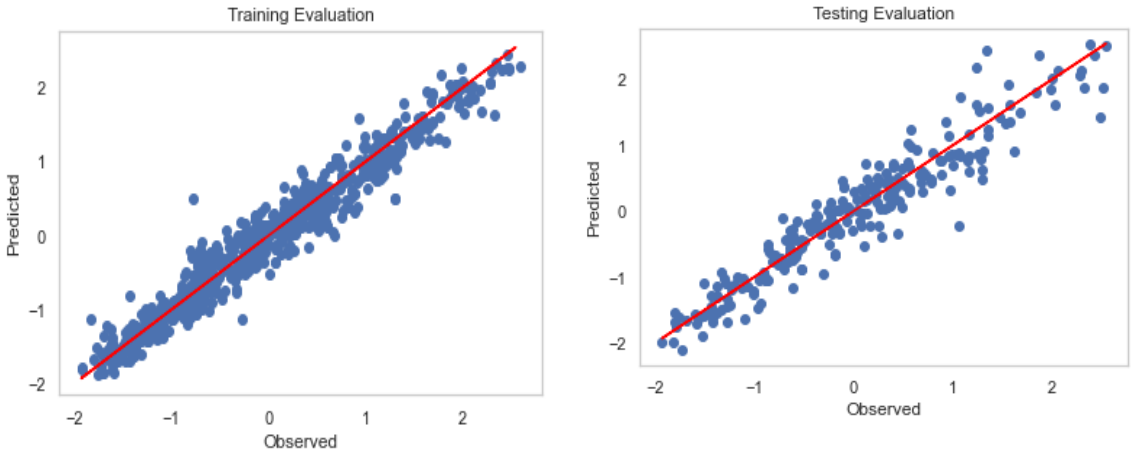

Figure 27: Model Evaluation for the neural network Model

The model over-fits the data to a small extent and there are few outliers in the testing evaluation curve. If the epoch and batch size are increase the model fits the training data to better accuracy. In that case, the over-fitting increases. As such, the learning cycles are reduced to properly fit the testing data

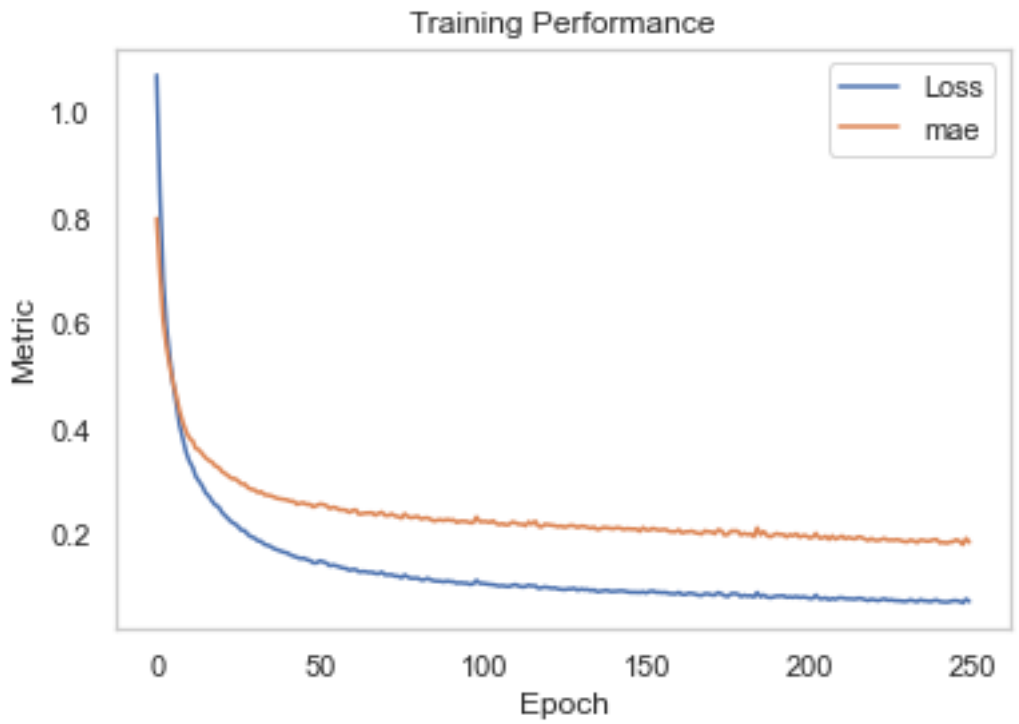

Figure 28: Model Performance evaluation vs Learning Cycle

Discussion & Model Comparison

Based on the metric evaluation of MSE, MAE, Kling-Gupta Efficiency, and Cross-Validation scoring the mentioned models are evaluated and compared. The metric values are provided in a table below for transparent comparison among the models.

Table 07: Model Evaluation Chart for Selected Models

Metric Parameter	Random Forest	Gradient Boosting Regression	Linear Regression	Deep Neural Network
MSE	0.11675	0.194187	0.377240	0.099011
MAE	0.26844	0.362588	0.482060	0.237379
COR	0.94336	0.920412	0.798318	0.952428
Overall (KGE)	0.33417	-3.72615	-1.73116	-2.02634
COR (KGE)	0.94336	0.920412	0.798318	0.95242
VAR (KGE)	0.58618	0.131222	-0.65018	-0.6493
BIAS (KGE)	1.51852	5.644938	-1.16691	-1.5369
CV Score (Mean)	-	-	-0.413334	-0.11777

Among all the models provided, the deep neural network tends to show more accuracy in terms of all the provided metrics.

Reference

Kling, H., Fuchs, M., and Paulin, M., 2012. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. Journal of Hydrology, 424, pp.264-277.

Yeh, I.C., 1998. Modeling concrete strength with augment-neuron networks. Journal of Materials in Civil Engineering, 10(4), pp.263-268.

Young, B.A., Hall, A., Pilon, L., Gupta, P., and Sant, G., 2019. Can the compressive strength of concrete be estimated from knowledge of the mixture proportions? New insights from statistical analysis and machine learning methods. Cement and Concrete Research, 115, pp.379-388.

Tukey, J., 1970. Exploratory Data Analysis. Pearson.

Panagiotis G. Asteris, Vaseilios G. Mokos1 (2019). "Concrete compressive strength using artificial neural networks." Neural Computing and Applications