Concrete Compressive Strength Project

Berkalp Altay

01/31/2022

Contents

1. Introduction

This project aims to build a machine learning model to predict the compressive strength of concrete using the Concrete Compressive Strength Data Set from the UCI Machine Learning Repository.

1.1. Dataset

The Concrete Compressive Strength data set was added to the UCI Machine Learning Repository by Prof. I-Cheng Yeh. As concrete is one of the most important materials in civil engineering, it is crucial to predict its compressive strength, a highly accepted measure to evaluate the performance of concrete admixtures. In this Concrete Compressive Strength data set, there are 1030 concrete admixture instances with 9 different attributes (i.e. columns) associated with each of these admixtures.

7 of the 9 attributes represent various components that go into making concrete. 1 of the other 2 is the age of the concrete since it was made. The last one is the compressive strength of concrete which will be predicted using machine learning models. In the UCI Machine Learning Repository, the details of these 9 attributes are given as follows:

- Cement (component 1) as kg in a m3 mixture
- Blast Furnace Slag (component 2) as kg in a m3 mixture
- Fly Ash (component 3) as kg in a m3 mixture
- Water (component 4) as kg in a m3 mixture
- Superplasticizer (component 5) as kg in a m3 mixture
- Coarse Aggregate (component 6) as kg in a m3 mixture
- Fine Aggregate (component 7) as kg in a m3 mixture
- Age as Day (1~365)
- Concrete compressive strength as megapascals (MPa)

The column names of the data set will be redefined to shorten them.

Table 1: First 5 Rows of Concrete Data Set with Renamed Columns

cement	slag	ash	water	superplasticizer	coarse_agg	fine_agg	age	strength
540.0	0.0	0	162	2.5	1040.0	676.0	28	79.98611
540.0	0.0	0	162	2.5	1055.0	676.0	28	61.88737
332.5	142.5	0	228	0.0	932.0	594.0	270	40.26954
332.5	142.5	0	228	0.0	932.0	594.0	365	41.05278
198.6	132.4	0	192	0.0	978.4	825.5	360	44.29608

The test set will be 20% of the overall data set. Since there are 1030 rows of data in the dataset, a higher test set size seemed highly likely to restrict the train set size (i.e. train set size = 1-(test set size)) for model development. A lower test set size could be employed but 20% seemed to be a reasonable size to strike a balance between data required for model development and for model evaluation.

1.2. Evaluation Criteria

In the next sections, the data will be explored and a machine learning model to predict concrete compressive strength will be built using the data. All machine learning models will be evaluated using the Root Mean Squared Error (RMSE).

The RMSE formula is as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} \hat{y}_i^2 - y_i^2}$$

In this formula, $\hat{y}_i$ represents the predicted compressive strength of concrete based on a machine learning model while y_i represents the actual compressive strengths from the dataset.
The lower the RMSE, the better the model is. The ultimate RMSE evaluation will be based on a test set that will be used only for model evaluation at the end.

The XGBoost Model with Cross-Validation and Grid Search, the final model in this project, reaches an RMSE of **4.164**.

2. Analysis

2.1. Initial Data Exploration

1030 rows of the original data set are divided into two datasets: *train_set* with 822 rows and *test_set* with 208 rows.
Also, both the *train_set* and *test_set* datasets have 9 columns.

Table 2: Columns

Column	Explanations
cement	Cement used in concrete making measured in kilograms (kg) in a cubic meter (m3) mixture
slag	Blast furnace slag, a calcium-silicate product used in concrete making, measured in kilograms (kg) in a cubic meter (m3) mixture
ash	Fly ash, a byproduct from burning coal used in concrete making, measured in kilograms (kg) in a cubic meter (m3) mixture
water	Water used in concrete making measured in kilograms (kg) in a cubic meter (m3) mixture
superplasticizer	Superplasticizer, an additive in concrete making to enhance workability and reduce water content, measured in kilograms (kg) in a cubic meter (m3) mixture
coarse_agg	Coarse Aggregate, irregular broken stones having a size of at least 4.75 mm or 3/16 inches used in concrete making, measured in kilograms (kg) in a cubic meter (m3) mixture
fine_agg	Fine Aggregate, sand or crushed stone particles used in concrete making, measured in kilograms (kg) in a cubic meter (m3) mixture
age	Age of concrete measured in days
strength	Concrete Compressive Strength, the capacity of concrete to withstand compressive weight, measured in megapascals (MPa)

The test set will be used as a final hold-out test set to evaluate the RMSE of the models. Therefore, only the train set will be used to build a model. Often, cross-validation will be used with models to try to improve the results.

Before moving to the next section, let's look at the first 5 columns of the train set and test set.

Table 3: First 5 Rows of Train Set

cement	slag	ash	water	superplasticizer	coarse_agg	fine_agg	age	strength
540.0	0.0	0	162	2.5	1040.0	676.0	28	79.98611
540.0	0.0	0	162	2.5	1055.0	676.0	28	61.88737
332.5	142.5	0	228	0.0	932.0	594.0	270	40.26954
332.5	142.5	0	228	0.0	932.0	594.0	365	41.05278
198.6	132.4	0	192	0.0	978.4	825.5	360	44.29608

Table 4: First 5 Rows of Test Set

cement	slag	ash	water	superplasticizer	coarse_agg	fine_agg	age	strength
380.0	95.0	0	228	0	932	594	365	43.69830
427.5	47.5	0	228	0	932	594	270	43.01296
475.0	0.0	0	228	0	932	594	7	38.60376
475.0	0.0	0	228	0	932	594	90	42.22903
237.5	237.5	0	228	0	932	594	365	38.99538

2.2. Data Visualization & Description

In the following sections, the entire data set composed of the train and test sets will be summarized and visualized. In `Section 2.3. Preprocessing` and `Section 2.4. Modeling`, the train and test sets will be used.

2.2.1. Summary Statistics

Before visualizing the distribution of variables in the concrete data set, we can get the summary statistics of the set. We look at the minimum, first quartile, median, mean, third quartile, and maximum values for each predictor.

Table 5: Summary Statistics for Concrete Data Set

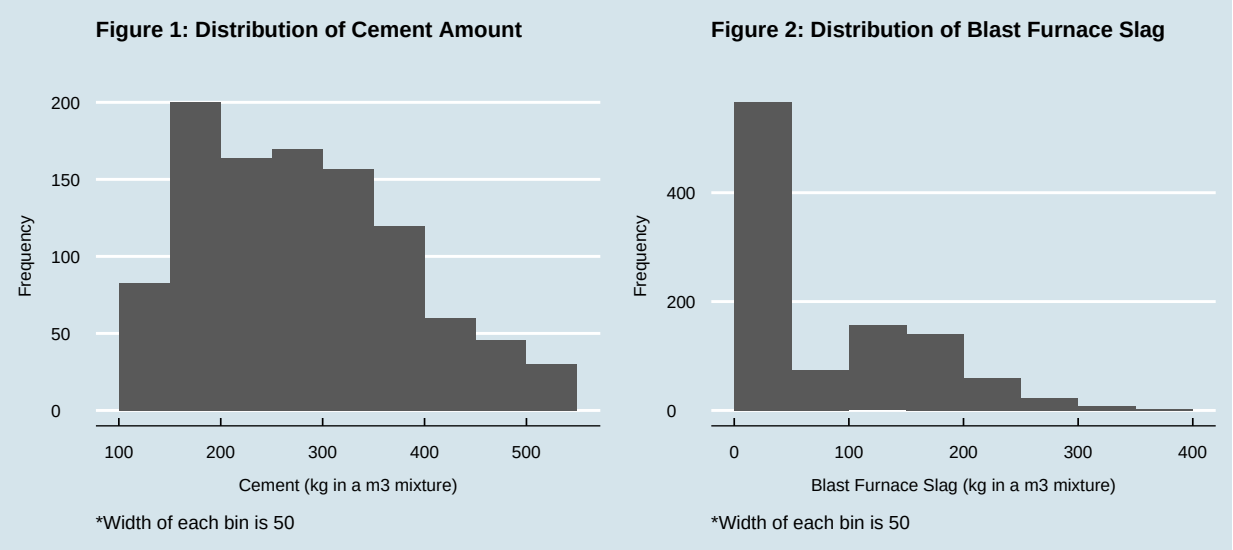
cement	slag	ash	water	superplasticizer
Min. :102.0	Min. : 0.0	Min. : 0.00	Min. :121.8	Min. : 0.000
1st Qu.:192.4	1st Qu.: 0.0	1st Qu.: 0.00	1st Qu.:164.9	1st Qu.: 0.000
Median :272.9	Median : 22.0	Median : 0.00	Median :185.0	Median : 6.350
Mean :281.2	Mean : 73.9	Mean : 54.19	Mean :181.6	Mean : 6.203
3rd Qu.:350.0	3rd Qu.:142.9	3rd Qu.:118.27	3rd Qu.:192.0	3rd Qu.:10.160
Max. :540.0	Max. :359.4	Max. :200.10	Max. :247.0	Max. :32.200

coarse_agg	fine_agg	age	strength
Min. : 801.0	Min. :594.0	Min. : 1.00	Min. : 2.332
1st Qu.: 932.0	1st Qu.:731.0	1st Qu.: 7.00	1st Qu.:23.707
Median : 968.0	Median :779.5	Median : 28.00	Median :34.443
Mean : 972.9	Mean :773.6	Mean : 45.66	Mean :35.818
3rd Qu.:1029.4	3rd Qu.:824.0	3rd Qu.: 56.00	3rd Qu.:46.136
Max. :1145.0	Max. :992.6	Max. :365.00	Max. :82.599

As seen in the table above, each attribute has a different range of values that it can take. This might be a problem for some machine learning algorithms such as the K-Nearest Neighbors (KNN). Therefore, 8 input variables that will be used to predict concrete compressive strength will be standardized after Section *2.2. Data Visualization & Description.*

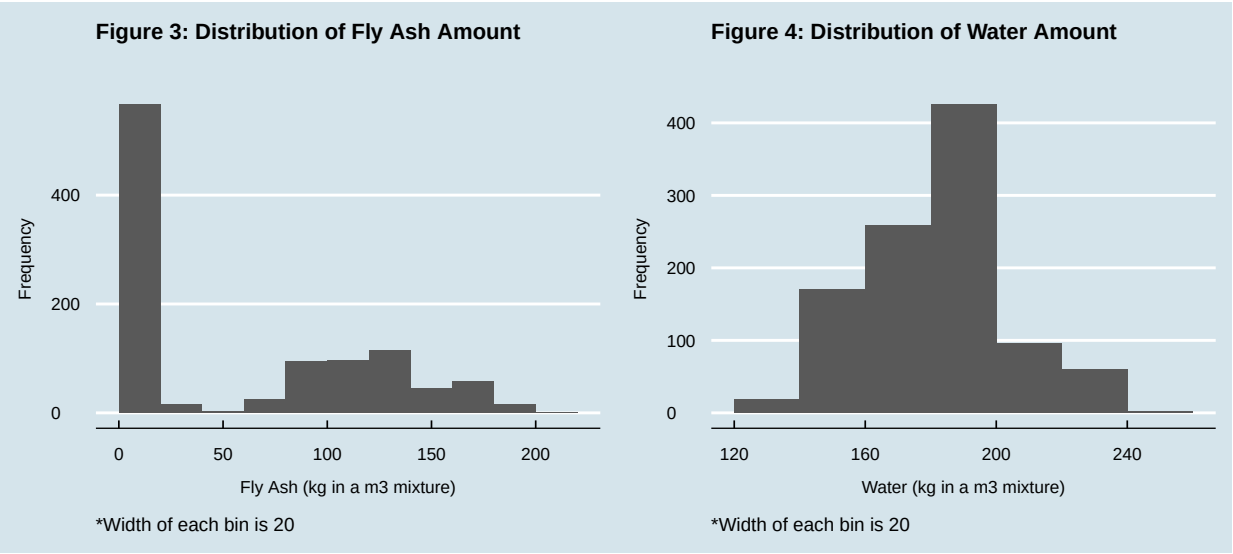
2.2.2. Distribution

First, we look at the cement and blast furnace slag columns.



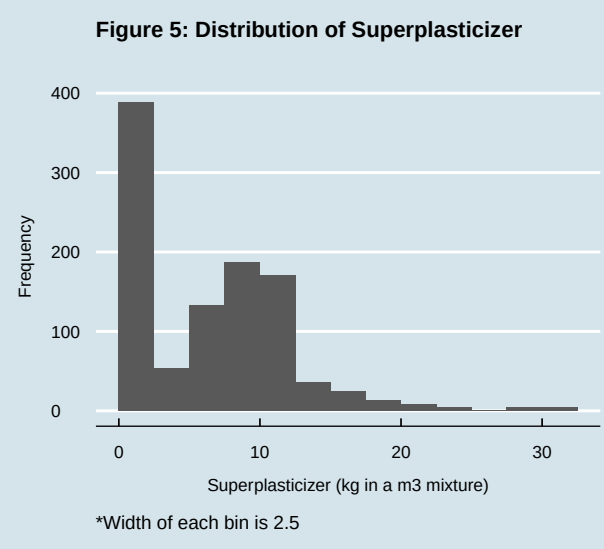
As seen in Figures 1 and 2 above, the distributions of the cement amount and blast furnace slag are right-skewed.

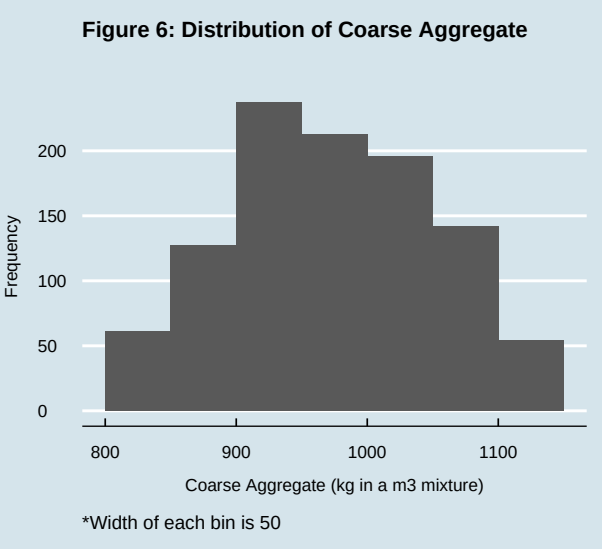
Second, we look at the ash and water columns.



As seen in Figures 3 and 4 above, most of the distribution of the fly ash amount is less than or equal to 20 kg in a m3 mixture while the distribution of water amount seems centered around 180 kg in a m3 mixture.

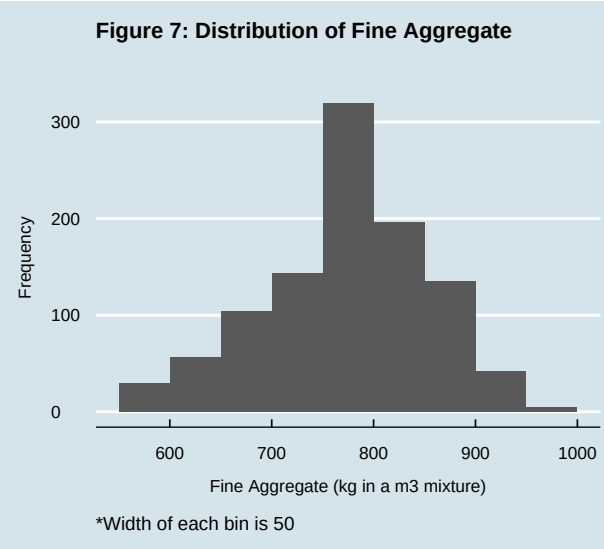
Third, we look at the superplasticizer and coarse aggregate columns.

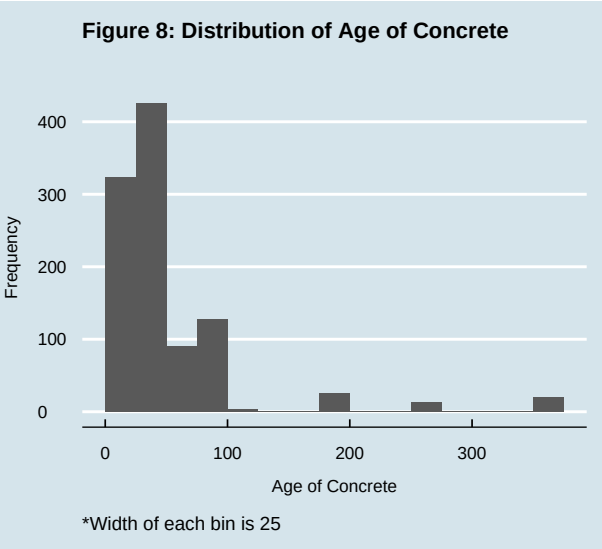




As seen in Figures 5 and 6 above, most of the distribution of the superplasticizer amount is less than or equal to 2.5 kg in a m3 mixture while the distribution of coarse aggregate amount might be considered approximately normal.

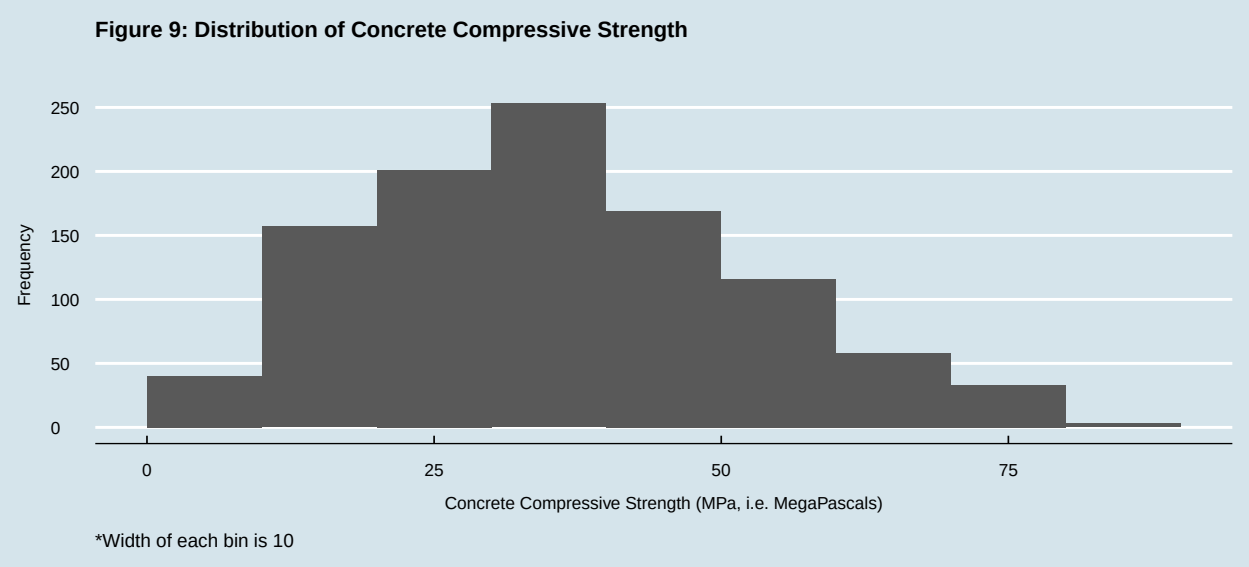
Fourth, we look at the fine aggregate and age columns.





As seen in Figures 7 and 8 above, most of the distribution of the fine aggregate amount can be considered approximately normal while most of the distribution of age of concrete is less than 100 days.

Last, we look at the distribution of the compressive strength of concrete.

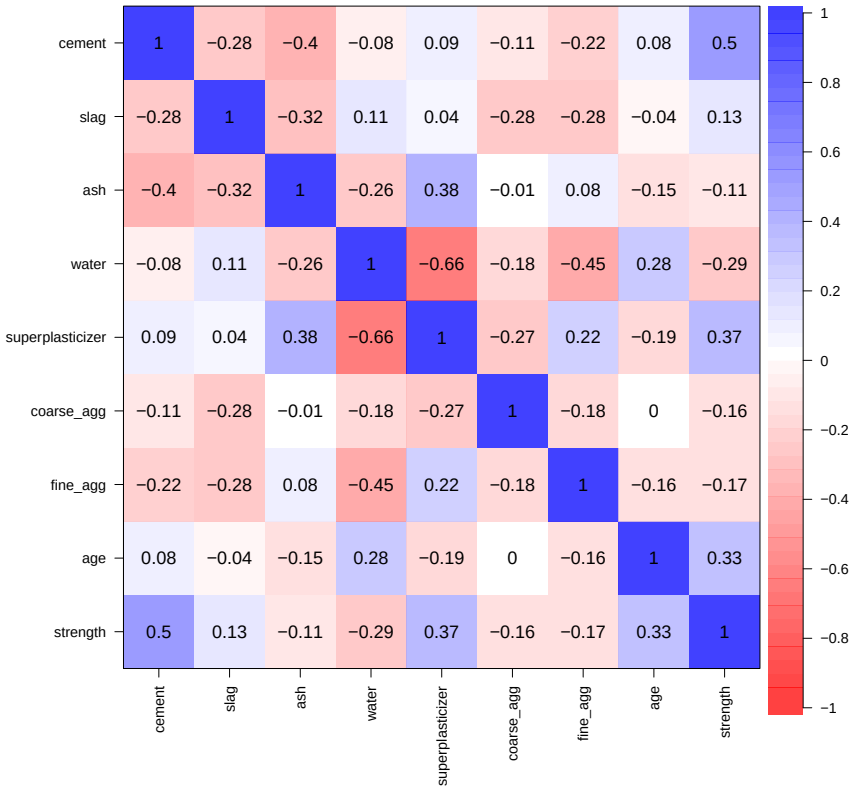


As seen in Figure 9 above, the distribution of concrete compressive strength can be considered approximately normal.

2.2.3. Correlations

The correlations among variables can also be explored. In Figure 10 below, we can see that most variables have some correlations among them with darker blues and reds indicating stronger positive and negative correlation, respectively. We notice a strong negative correlation between the superplasticizer content and the water content. This negative relationship makes sense because superplasticizers are a type of water reducers that significantly reduce the amount of water required to make concrete. Thus, as the superplasticizer content increases, the water content is expected to decrease.

Figure 10: Correlation Matrix



2.3. Preprocessing

First, all the predictors in the train set will be standardized with the mean and standard deviation of the predictors in the train set. Then, all the predictors in the test set will also be standardized with the mean and standard deviation of the predictors in the train set.

2.4. Modeling Approach

In this section, various models will be built. In all of these models, the test set will be used to evaluate the the models built using the train set.

2.4.1. Linear Regression

I start with the linear regression which will serve as the baseline model.

The baseline model of Linear Regression gives an RMSE of 10.891.

As seen in Table 6 below, the Linear Regression model shows that all variables except for Coarse Aggregate (coarse_agg) and Fine Aggregate (fine_agg) are statistically significant at the 5% and 1% significance levels (i.e. 95% and 99% confidence intervals) since their p-values (i.e. Pr(>|t|)) are less than 0.01. On the other hand, Coarse Aggregate and Fine Aggregate have p-values around 0.396 and 0.501, indicating that they are not statistically significant.

Regarding statistically significant predictors, cement has a regression coefficient of around 11.45. This

indicates that an increase of 1 kg per m3 mixture in cement is associated with an increase of about 11.45 MPa in concrete compressive strength.

On the other hand, water has a regression coefficient of about -3.94. This indicates that an increase of 1 kg per m3 mixture in water is associated with a decrease of 3.94 MPa in concrete compressive strength.

Table 6: Linear Regression Coefficients Table

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	35.9303619	0.3589386	100.1017026	0.0000000
cement	11.4464442	0.9574055	11.9556910	0.0000000
slag	8.2612449	0.9471971	8.7217801	0.0000000
ash	4.6783526	0.8817486	5.3057672	0.0000001
water	-3.9372260	0.9343616	-4.2138142	0.0000279
superplasticizer	1.8189578	0.6172160	2.9470360	0.0033001
coarse_agg	0.6723004	0.7910730	0.8498589	0.3956537
fine_agg	0.6172842	0.9273895	0.6656149	0.5058461
age	7.4367483	0.3804150	19.5490419	0.0000000

Given that the Linear Regression model gives an RMSE of 10.891, I will look at other models to improve the RMSE.

2.4.2. K-Nearest Neighbors (KNN)

Now, a KNN model is used with its default settings. In the next section, grid search and cross-validation will be used to improve the results.

The optimal number of nearest neighbors, i.e. k, is 7. The KNN model gives an RMSE of 9.152. This is an improvement over the baseline Linear Regression Model.

2.4.3. K-Nearest Neighbors (KNN) with Cross-Validation

With the KNN model, K-fold cross-validation is used. There are 10 folds, each test fold consisting of 10% of the train_set. Also, a grid search is employed for the parameter of the number of neighbors, i.e. k, which ranges from 1 to 51 in increments of 2. This means that the model tries each value of parameter k and determine which one gives the best result (i.e. lowest RMSE) when fitting the train set.

The KNN + Cross-Validation model gives an RMSE of 8.871. This is an improvement over the KNN model without Cross-Validation. The optimal number of nearest neighbors, i.e. k, for this model is 5. In the next part, a Random Forest model will be used to improve results further.

2.4.4. Random Forest

Now, a Random Forest model is developed. I use the "Rborist" package and proceed with its default parameters which will be changed when grid search and cross-validation are utilized in the next model.

The Random Forest model gives an RMSE of 4.886. This is a major improvement over the KNN model with Cross-Validation and grid search.

In Table 7 below, the relative variable importance of all the predictors for the Random Forest model is given. The variable importance gives the sum of decrease in error when a tree is split on that variable. This value when divided by the highest variable importance gives the relative variable importance.

As seen in Table 7, age has the highest variable importance whereas Fly Ash (ash) has the lowest variable importance with its relative variable importance as 7.5% of that of age as the predictor. When scaled against age, all other predictors have their variable importance ranging from 7.5% to 80.4% of that of age.

Table 7: Relative Variable Importance of Random Forest Model

	Overall
age	1.0000000
cement	0.8042676
water	0.3788214
slag	0.1946181
superplasticizer	0.1863015
fine_agg	0.1447846
coarse_agg	0.1076728
ash	0.0750405

Given that the Random Forest model gives an RMSE of 4.886, grid search with cross-validation will be used next along with the Random Forest model to improve results further.

2.4.5. Random Forest with Cross-Validation

Now, a Random Forest model with cross-validation and grid search are used to improve the results even further. The k-fold cross-validation is used with 10 folds. The Rborist package is utilized with the following search space:

- number of trial predictors for a split (predFixed) ranging from 2 to 8 in increments of 1
- minimum number of distinct row references to split a node (minNode) ranging from 2 to 10 in increments of 1

The Random Forest model with grid search and cross-validation (Random Forest+CV) gives an RMSE of 4.784. This is a slight improvement over the Random Forest model without Cross-Validation and grid search. The optimal values of predFixed and minNode for this model are 5 and 2, respectively.

In Table 8 below, the relative variable importance of all the predictors for the Random Forest + CV model is given.

As seen in Table 8, age still has the highest variable importance while Fly Ash (ash) has the lowest variable importance, again. Of particular interest is superplasticizer as a predictor. In the Random Forest model without cross-validation, superplasticizer was the fifth most important variable with 18.6% of the variable importance of age. With the Random Forest model with cross-validation and grid search, superplasticizer is the fourth most important variable with 26.6% of the variable importance of age.

Table 8: Relative Variable Importance of Random Forest+CV Model

	Overall
age	1.0000000
cement	0.7824915
water	0.3471144
superplasticizer	0.2658767
slag	0.2083180
fine_agg	0.1730177
coarse_agg	0.0898312
ash	0.0768247

While the Random Forest with grid search and cross-validation gives an RMSE of 4.784, the Extreme Gradient Boosting (XGBoost) model will be used next to improve results.

2.4.6. Extreme Gradient Boosting (XGBoost)

Now, the Extreme Gradient Boosting model is developed. The "xgboost" package is used with its default parameters which will be changed when I apply grid search and cross-validation in the next model.

The XGBoost model gives an RMSE of 5.01. This is slightly worse than the Random Forest model with cross-validation and grid search.

In Table 9 below, the relative variable importance of all the predictors for the XGBoost model is given.

As seen in Table 9, age has the highest variable importance in the XGBoost model as was the case in the Random Forest model with and without cross-validation. In this model, superplasticizer has the lowest variable importance, 12% of that of age as a predictor. Of particular note is the fact that the second most important variable in this model, i.e. water, has only 46% of the variable importance of age whereas, in the Random Forest models, cement as the second most important predictor had about 78 or 80% of the variable importance of age.

Table 9: Relative Variable Importance of XGBoost Model

	Overall
age	1.0000000
water	0.4599694
cement	0.2721699
coarse_agg	0.2098494
ash	0.2089966
fine_agg	0.1764845
slag	0.1406266
superplasticizer	0.1282448

While the XGBoost model gives an RMSE of 5.01, grid search and cross-validation will be used next along with the XGBoost model to improve results.

2.4.7. Extreme Gradient Boosting (XGBoost) with Cross-Validation

Now, an XGBoost model with cross-validation and grid search are used to improve the results even further.

The k-fold cross-validation is used with 10 folds. The xgboost package is utilized with the following search space:

- maximum number of iterations (nrounds) ranging from 100 to 500 in increments of 50
- step size shrinkage used in update (eta) ranging from 0.05 to 0.3 in increments of 0.05
- maximum depth of a tree (max_depth) ranging from 2 to 7 in increments of 1

The XGBoost model with grid search and cross-validation (XGBoost + CV) gives an RMSE of 4.164. This is an improvement over the next best model, the Random Forest model with cross-validation and grid search. The optimal values of nrounds, max_depth, and eta are 500, 4, and 0.15, respectively.

In Table 10 below, the relative variable importance of all the predictors for the XGBoost + CV model is given.

As seen in Table 10, age continues to have the highest variable importance in the XGBoost + CV model as was the case in previous XGBoost and Random Forest models. Cement is the second most important variable in this model as in the Random Forest models but unlike in the XGBoost model without cross-validation. Cement has 57.6% of the variable importance of age whereas, in the Random Forest models, it had about 78 or 80% of the variable importance of age.

Table 10: Relative Variable Importance of XGBoost+CV Model

	Overall
age	1.0000000
cement	0.5761578
water	0.3215923
superplasticizer	0.2212709
fine_agg	0.2110657
slag	0.2012803
ash	0.1701801
coarse_agg	0.1284964

3. Results

Overall, seven models have been built. The XGBoost Model with Cross-Validation and Grid Search (XGBoost + CV) has an RMSE of 4.164 MegaPascals (MPa), which is a major improvement of about 6.5 MPa over the baseline Linear Regression Model. For reference, the RMSE of all models found using the test set can be found in Table 11 below.

Table 11: RMSEs of All Models Using Test Set

Model	RMSE
Linear Regression	10.891350
KNN	9.151553
KNN+CV	8.870900
Random Forest	4.886325
Random Forest+CV	4.784191
XGBoost	5.010266
XGBoost+CV	4.163971

The ultimate model was built using different algorithms along with grid search and cross-validation. The performance improved from an RMSE of **10.891** (i.e. Linear Regression) to the final RMSE of **4.164** (i.e. XGBoost + CV).

4. Conclusion

In this project, the Concrete Compressive Strength Data Set from the UCI Machine Learning Repository was used to build a model to predict the compressive strength of concrete using 8 different predictors.

This report introduced the Concrete Compressive Strength dataset, explored its properties, and visualized the data.

After these steps, the dataset was standardized. Then, machine learning models were built, beginning from the baseline linear regression model to an XGBoost model with cross-validation and grid search that significantly reduced the root mean square error (RMSE) between the actual compressive strength of concrete and its predicted compressive strength.

The performance improved the most when the Random Forest and Extreme Gradient Boosting (XGBoost) models were used. Using cross-validation as well as grid search for hyperparameter tuning led to further improvements.

All in all, the final model achieved an RMSE of **4.164** MegaPascals (MPa).

Some limitations of this project are that it could not increase the number and range of hyperparameters for tuning a better model as such an increase would require greater computational power. Also, other models such as SVM and artificial neural networks were not used as this project already utilized several well-known and highly performant (e.g. XGBoost) models. Any future work might further improve on this report by employing different methods.