Predictive Model for Concrete Compressive Strength

This manuscript (permalink) was automatically generated from uiceds/cee-492-term-project-fall-2022-team-online@ecedd3b on November 17, 2022.

Authors

- **Ray Ausan**
  · rausan3
  Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign

- **Min Win Ye**
  · FrenchToastty
  Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign

- **Papa Ibrahima Mbodj**
  · pimbooo
  Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign

- **Dafar Obeidat**
  · dafarno2
  Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign

Proposal

The team plans to predict the 28th day compressive strength of the concrete given a proportion of water, cement, aggregates, and percentage of additives. The dataset has 8 input parameters and 1 output parameter. The model will predict the interactions between the concrete mix components to the compressive strength. The outputs being considered include a table of proportions for mix design and a formula for the compressive strength.

The model aims to predict the 28th day compressive strength of concrete based on the dataset. Traditionally, to compare the actual compressive strength of concrete (Ca) against designed for strength (Cd), 28 days would need to be passed before a cube sample can be crushed to check the compressive strength. This is usually not a problem, if the 28th day Ca matches the Cd. However, if there is a mismatch in Ca and Cd, there could be massive hacking of concrete or additionally structures put in place to further enhance the strength.

To minimize such errors, this model predicts the 28th day compressive strength instantaneously when a batch of concrete mix is created. Eventually, with enough confidence, it aims to change the default measurement of 28th day compressive strength from cube crushing to using this predictive model.

Dataset

Description

The dataset was retrieved from UCI Machine Learning Repository (Yeh, 2007). It has 1030 observations, 8 quantitative input variables, and 1 quantitative output variable.

Column A/ Component 1: Cement

Cement is an adhesive substance that acts as a binder for all the components in a concrete mix. Ordinary Portland Cement (OPC) is made up of limestone, clay, and iron ore; and it is most commonly used. According to the ASTM standard, there are five types of cement, the difference due to the chemical composition, altering the properties. In this dataset, Type 1 Ordinary Portland Cement will be used. The unit used is kg of cement per 1 m^3 of the concrete mixture (kg/m^3 of mixture).

Column B/ Component 2: Blast Furnace Slag

Blast furnace Ash is a nonmetallic co-product obtained in the production of iron, iron ore, iron scrap and fluxed. It is commonly used in cement production as a substitute for clinker and in concrete production as a substitute for aggregates. The use of slag cement improves performance and durability of concrete. The unit used is kg of per 1 m^3 of the concrete mixture (kg/m^3 of mixture).

Column C/ Component 3: Fly Ash

Fly Ash is byproduct of burning pulverized coal in electric generation. It is a fine powder used to improve the workability, the strength and the durability of Portland Cement Concrete. It also decreases the water demand of the concrete mix and reduces heat of hydration. The unit used is kg of per 1 m^3 of the concrete mixture (kg/m^3 of mixture).

Column D/ Component 4: Water

Water content is the most important factor affecting the consistency of fresh concrete. The higher the water content, the higher the workability but the lower the strenght of the concrete. The unit used is kg per 1 m^3 of the concrete mixture (kg/m^3 of mixture).

Column E/ Component 5: Superplasticizer

Superplasticizers are chemical compounds used to reduce the amount of water content in the concrete mixture to produce high-strength concrete while maintaining enough workability. The used unit is kg of the superplasticizer to 1 m^3 of the concrete mixture (kg/m^3 of mixture).

Column F/ Component 6: Coarse Aggregate

Coarse Aggregates are inert, granular, and inorganic material. Coarse Aggregates are aggregates that are larger or equal to the ASTM sieve size 4.75mm. Typical coarse aggregates are gravel, crushed stone or previously used concrete etc. They occupy a large volume in a concrete mix (~65-75%), as it acts as an economic filler for cement. The unit used is kg of coarse aggregate per 1 m^3 of the concrete mixture (kg/m^3 of mixture).

Column G/ Component 7: Fine Aggregate

Fine Aggregates are inert, granular, and inorganic material. Fine Aggregates are aggregates that are smaller than the ASTM sieve size 4.75mm. Typical fine aggregates are sand, crushed stone or burnt clays etc. The fine aggregates fill in the voids between coarse aggregates. It also provides resistance against shrinking and cracking. The unit used is kg of fine aggregate per 1 m^3 of the concrete mixture (kg/m^3 of mixture).

Column H/ Component 8: Age

This column represents the age of the concrete mixture after pouring. The concrete gains its strength gradually with time, and according to the ASTM, it reaches to 99% of the target compressive strength after 28 days. The strength will continue to increase after years and it can become larger than the target compressive strength (strength percent > 100%). The unit of this column data is in days.

Column I/ Output 1: Concrete compressive strength

It is the capacity of concrete to withstand compression load before failure. Again, based on the ASTM standards, this property reported at 28 days of curing time.

Exploratory Data Analysis

Summary Statistics

This section illustrates the general statistics of the dataset to show simple trends in the dataset.

Table 1: Summary Statistics

variable	mean	std	min	q25	median	q75	max
Cement (kg/m3)	281.168	104.506	102.000	192.375	272.900	350.000	540.000
Water (kg/m3)	181.567	21.354	121.800	164.900	185.000	192.000	247.000
Coarse Aggregate (kg/m3)	972.919	77.754	801.000	932.000	968.000	1029.400	1145.000
Fine Aggregate (kg/m3)	773.580	80.176	594.000	730.950	779.500	824.000	992.600
Blast Furnace Slag (kg/m3)	73.896	86.279	0.000	0.000	22.000	142.950	359.400
Fly Ash (kg/m3)	54.188	63.997	0.000	0.000	0.000	118.300	200.100
Superplasticizer (kg/m3)	6.205	5.974	0.000	0.000	6.400	10.200	32.200
Age (day)	45.662	63.170	1.000	7.000	28.000	56.000	365.000
Compressive strength (MPa)	35.818	16.706	2.330	23.710	34.445	46.135	82.600

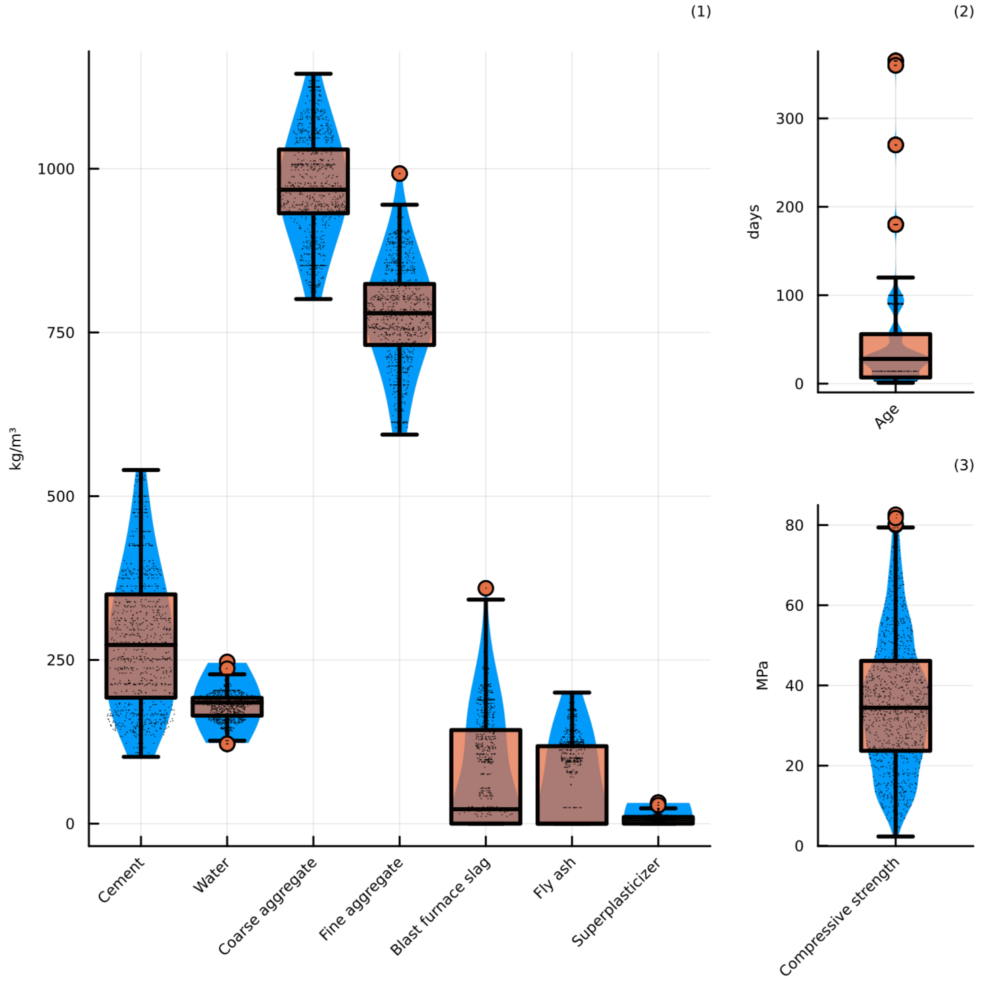

Figure 1: Violin, Box, and Dot Plots of Dataset. (1) Mass Axis (2) Days Axis (3) Strength Axis

Table 1 shows the mean, standard deviation, minimum & maximum, first quartile, median, and third quartile. Figure 1 shows a visual form of Table 1.

The main components, cement, water, and aggregates are present in all concrete mixes. The aggregates make a major portion of the concrete mix. The portion of water in the observations do not vary as much as the other main components.

Table 2: Secondary Component Observation Count

With Blast Furnace Slag	With Fly Ash	With Superplasticizer	Observations
false	false	false	209
false	false	true	23
false	true	false	6
false	true	true	233
true	false	false	164
true	false	true	170
true	true	true	225

Blast furnace slag, fly ash, and super plasticizer are not present in all observations. Table 2 shows that there are 209 observations without secondary components. Out of the secondary components, superplasticizer is the most prevalent with 651 total observations. However, superplasticizer has the least average mass in the concrete mix. There are no observations with both blast furnace slag and fly ash.

The median age of concrete strength measurement is at 28 days. Typical concrete testing in the industry is made on the 28th day. Some observations were measured after a year from casting.

The mean age of concrete strength for the dataset is 35.8 MPa. The minimum and maximum concrete strength observed is 2.3 MPa and 82.6 MPa respectively.

Correlation

In this section, the general correlation between various variables has been examined. The purpose is to understand the how each variable affects each other and as well as to understand how each variable affects the compressive strength of concrete.

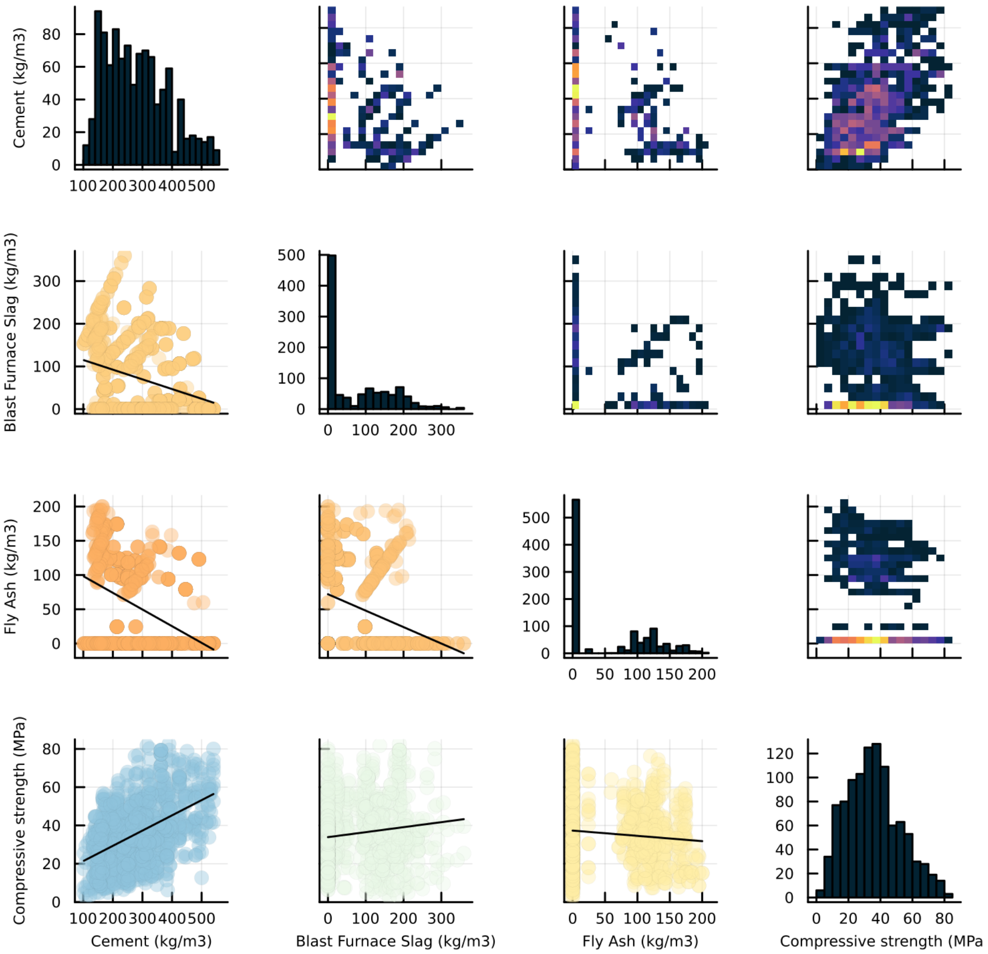

Figure 2: Correlation plot of Water, Superplasticizer & Compressive Strength

Table 3: Correlation table of Cement, Blast Furnace Slag, Fly Ash, Compressive Strength

	Cement (kg/m3)	Blast Furnace Slag (kg/m3)	Fly Ash (kg/m3)	Compressive strength (MPa)
Cement (kg/m3)	1.00	-0.28	-0.40	0.50
Blast Furnace Slag (kg/m3)	-0.28	1.00	-0.32	0.13
Fly Ash (kg/m3)	-0.40	-0.32	1.00	-0.11
Compressive strength (MPa)	0.50	0.13	-0.11	1.00

From Figure 2 and Table 3, is can be seen that apart from the positive correlation of compressive strength to cement, most other correlations have a neutral to negative correlation.

For the comparison of Compressive Strength against Cement, it is seen to have a moderate positive correlation of 0.50. This tells us that as cement content increases, the compressive strength of concrete also increases.

For Compressive Strength against Blast Furnace Slag and Fly Ash, which are commonly used as fillers for cement, it can be seen that they have a very weak correlation of 0.13 and -0.11. As such, it can be taken that Blast Furnace Slag and Fly Ash does not affect Compressive Strength when used as fillers for Cement.

For the correlation plots comparing Fly Ash against Cement, Fly Ash against Blast Furnace Slag, and Blast Furnace Slag against Cement, it can be seen that they have a weak correlation of -0.40, -0.32, -0.28. The negative correlations corelease with existing domain knowledge that Blast Furnace Slag and Fly Ash are substitutes of Cement. As when either the content of Blast Furnace Slag or Fly Ash goes up, the Cement content goes down.

From the histogram plots, it can be seen that Blast Furnace Slag and Fly Ash there is a high count of zero, showing that not all concrete mix designs require them. As such, it might be important to take into consideration the effects of the large amounts of zeros during preliminary modelling.

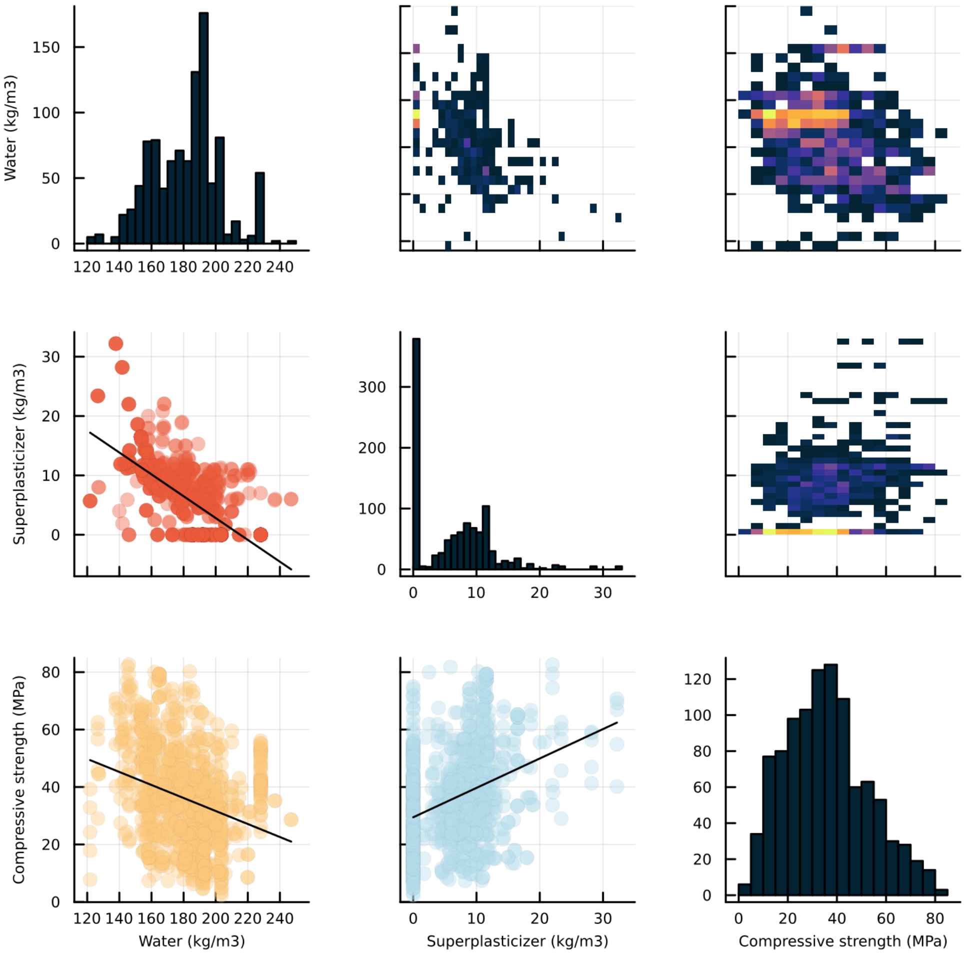

Figure 3: Correlation plot of water , superplasticizer and compressive strength

Table 4: Correlation coefficients of water , superplasticizer and compressive strength

	Water (kg/m3)	Superplasticizer (kg/m3)	Compressive strength (MPa)
Water (kg/m3)	1.00	-0.66	-0.29
Superplasticizer (kg/m3)	-0.66	1.00	0.37
Compressive strength (MPa)	-0.29	0.37	1.00

From Table 4, we see a weak positive correlation of 0.36 between the amount of superplasticizer in the mix and compressive strength. Table 4 also shows a weak negative correlation of -0.29 between Water content and compressive strength. As expected, the more water we have in the mix, the weaker the concrete and the more superplasticizer we have, the stronger the concrete.
We can also observe that Water content and superplasticizer amount have a strong negative correlation of -0.65. Overall the more superplasticizer we have, the less water there is in these concrete mixes.

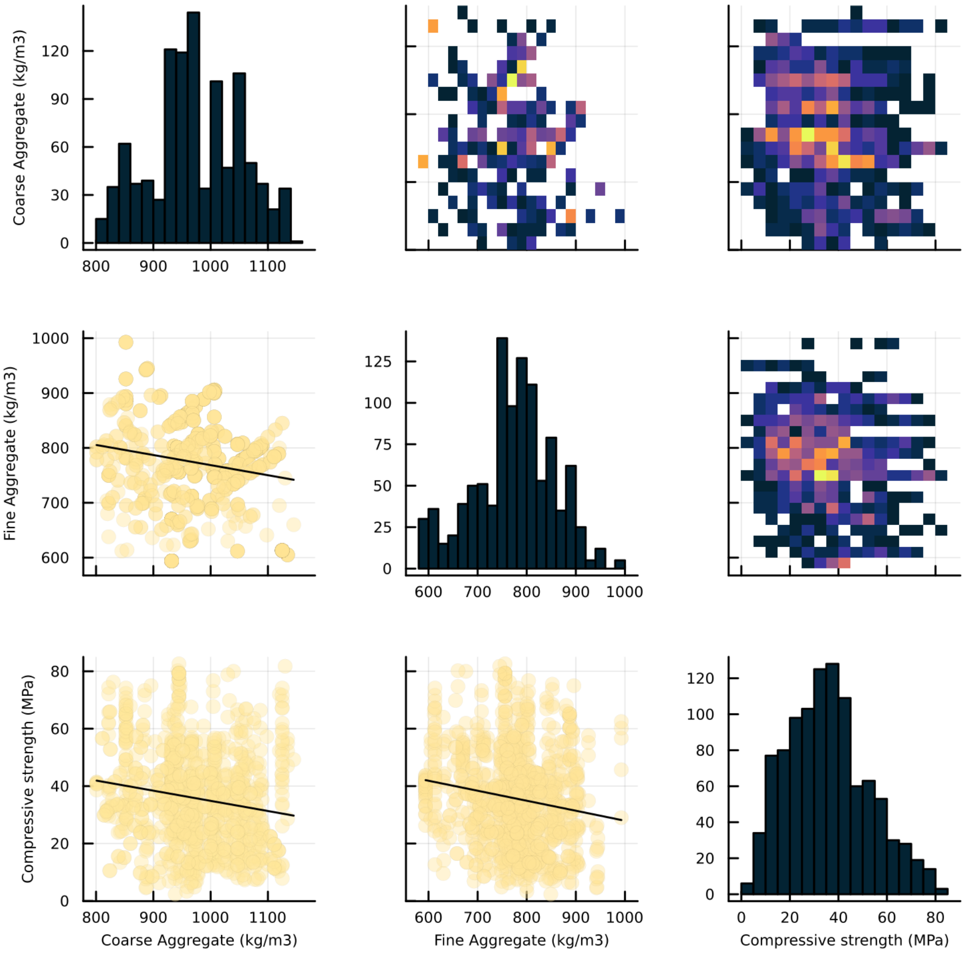

Figure 4: Correlation plot of Fine Aggregates , Coarse Aggregates and compressive strength

Table 5: Correlation coefficients of water , superplasticizer and compressive strength

	Coarse Aggregate (kg/m3)	Fine Aggregate (kg/m3)	Compressive strength (MPa)
Coarse Aggregate (kg/m3)	1.00	-0.66	-0.29
Fine Aggregate (kg/m3)	-0.66	1.00	0.37
Compressive strength (MPa)	-0.29	0.37	1.00

In Table 5, we can see the correlations between fine aggregates, coarse aggregates and compressive strength. We have a weak negative correlation of -0.16 between Fine aggregates and compressive strength and a weak negative correlation coefficient of -0.16 between coarse aggregates and compressive strength. We might gain more insight by taking these two elements together and comparing them to the compressive strength.

These correlations are in line with what is expected for a concrete mix.

Specific Correlation

In this section, the specific interactions between variables will be further discussed.

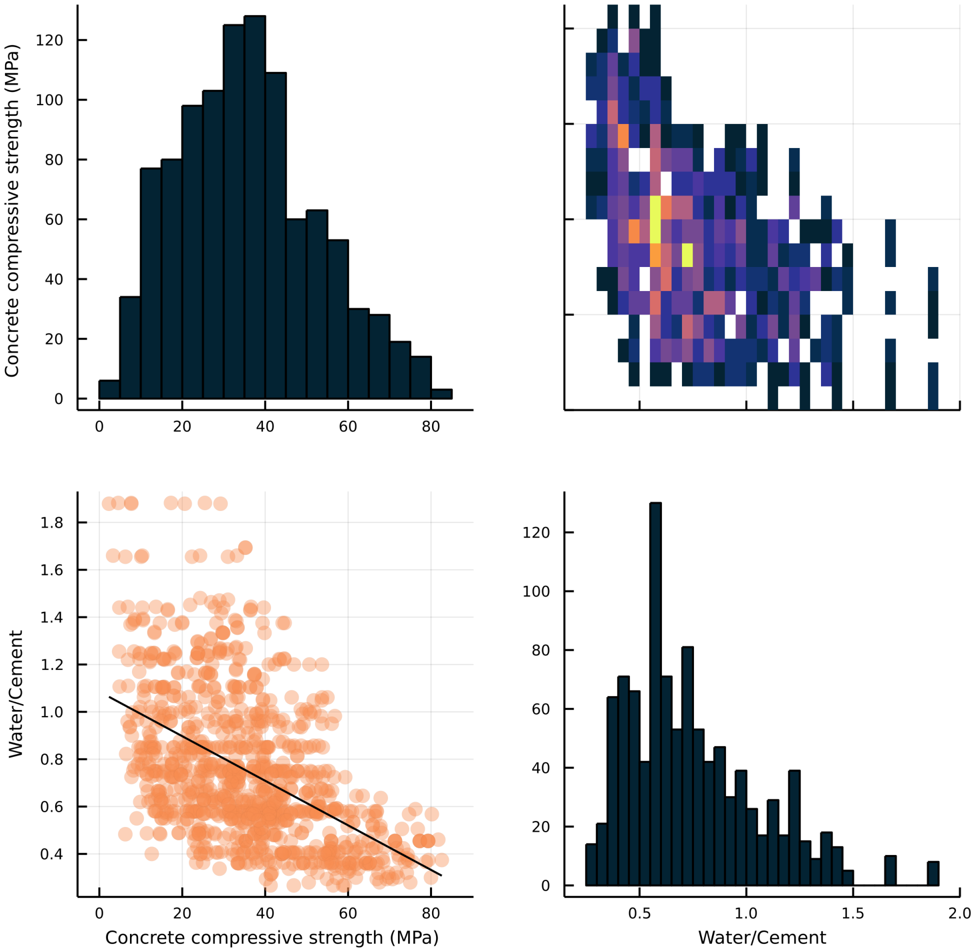

Figure 5: Correlation plot of Water/cement ratio against the concrete compressive strength

Figure 5 shows the relationship between the water to cement ratio (w/c) and concrete compressive strength. This ratio is considered the most significant influence on the final compressive strength of the concrete mixture. As we can see from Figure 5, the strength of concrete increases as the w/c decreases. The reason for that is as the amount of water increases in the mixture (for a specific amount of cement), the distance between the cement particles increases which leads to a more diluted and weaker paste. In practice, the w/c ranges from about 0.3 to over 0.8, which is the range where have the largest number of data as seen in the histogram in Figure 5. It is noticeable that there are significant amount of scatter in this relationship which can be attributed to two main reasons. First, this figure includes all the available concrete compressive strength data that have been measured at different concrete ages. The second, there are other factors that play a role in the final concrete compressive strength like the amount of fine and coarse aggregates, and these factors are not filtered within the data used to plot this figure.

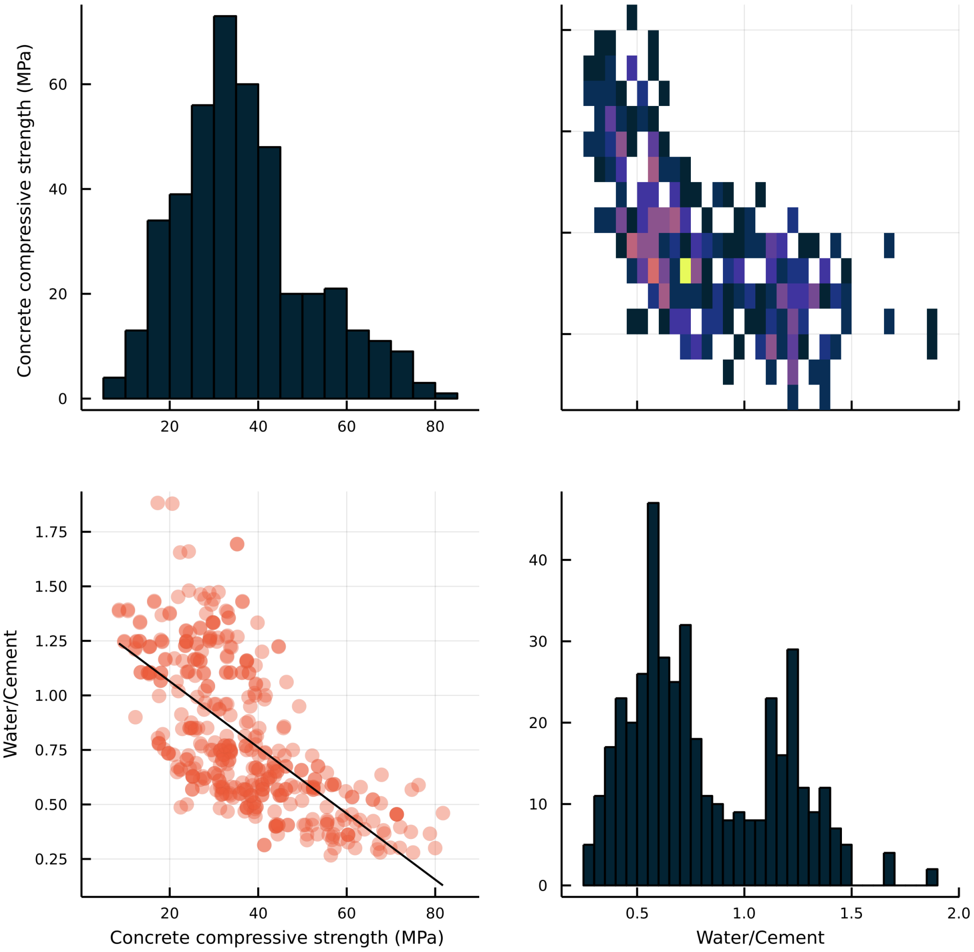

Figure 6: Correlation plot of Water/cement ratio against the concrete compressive strength at 28 days

To see how better the relationship can be, the factor of the concrete age was eliminated. In Figure 6, we are plotting the same relationship as in Figure 5 but just using the concrete compressive strength data at 28 days. It is clear that the scatter has been reduced and the effect of the w/c became more apparent.

The same trend can also be observed in the work that has been done by US Bureau of Reclamation (1975) as shown in Figure 7 below.

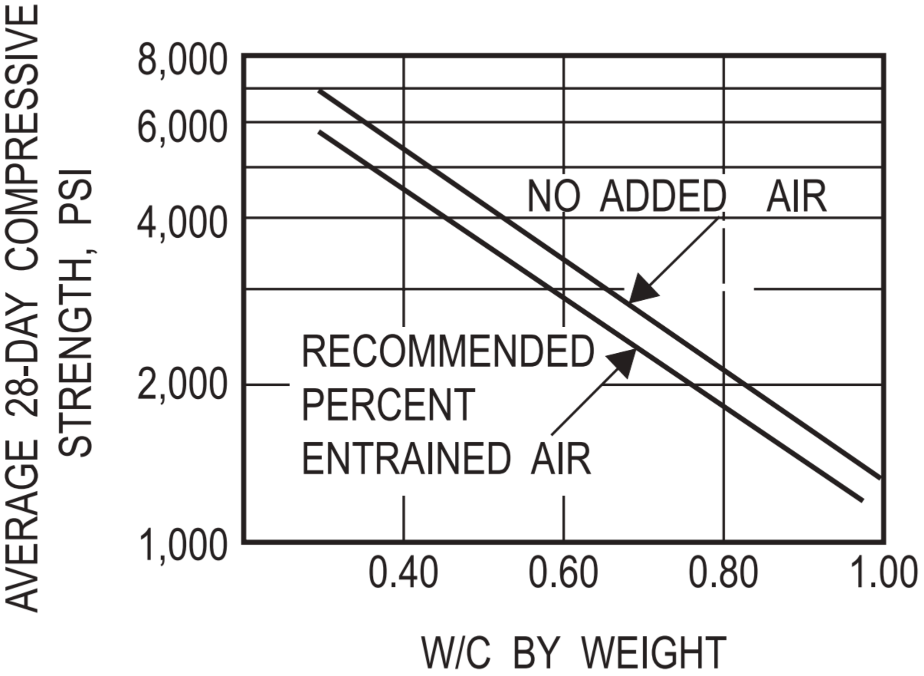

Figure 7: Compressive strength versus w/c (US Bureau of Reclamation, 1975)

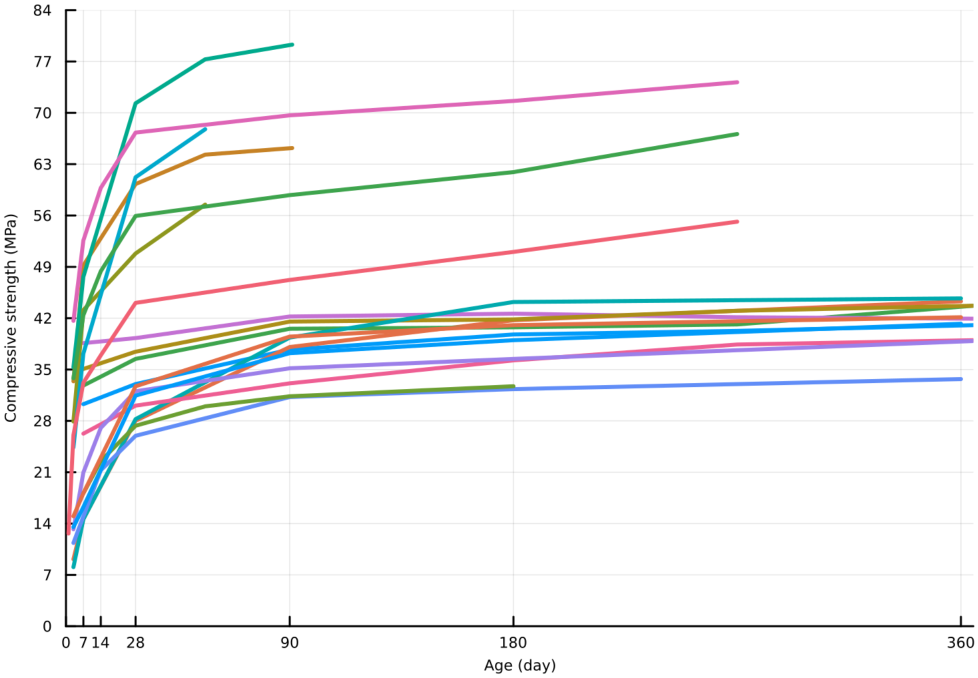

Figure 8: Average compressive strength versus age of observations without secondary components

Figure 8 shows averaged compressive strength from observations with the same primary components and more than five instances of measurement in time.

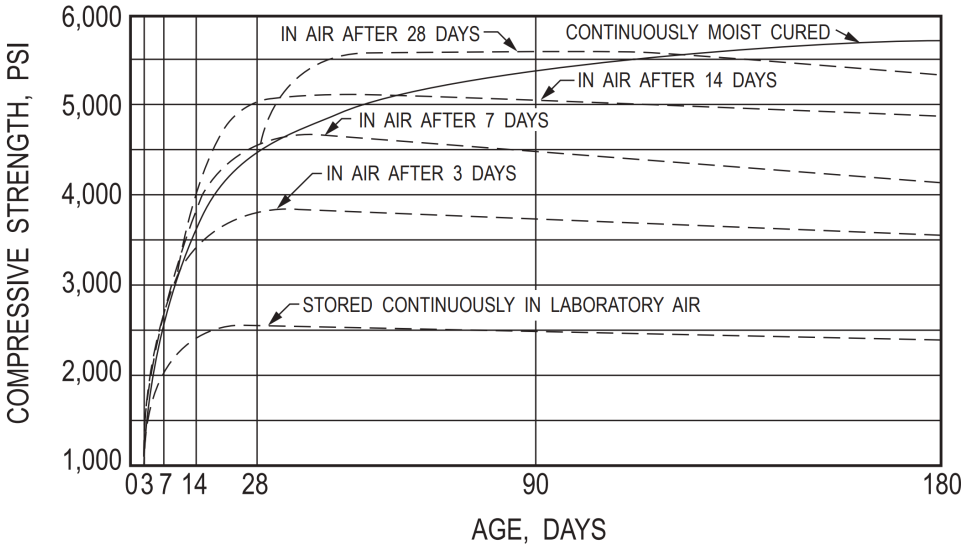

Figure 9: Compressive strength versus age (Merrit, 1983)

In addition to the w/c, the compressive strength against age can also be referenced to established standards. Figure 9 shows that concrete compressive strength increases significantly in the first four weeks from casting and increases gradually thereafter. The same trend illustrated in Figure 8 can be seen in Figure 9.

Predictive Modeling Plan

The predictive model for this project will be a supervised regression predictive model. The goal is to predict the 28th day concrete compressive strength, given cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate.

The dataset is a labelled dataset as such supervised learning method will be used train the model. The dataset will be split into training, testing and validation set using cross validation method. The training set will be used to train the model, the testing set will be used to optimize the model, and the validation set will be used to evaluate the performance of the model based on unseen data. The dataset will also be standardized to ensure there is no mismatch of the different scales for the variables.

The preliminary model will be a linear or polynomial model, using gradient descent and an error function such as MSE to train model. To ensure that the model does not overfit the training model, a regularization term either L1 or L2 will be used to optimize the model. L1 in the case of feature selection by reducing non-essential variables to zero, and L2 for the case of lowering the influence of non- essential variables.

Another preliminary predictive model is to use PCA for regression. By transforming the standardized training data into PCA coordinate systems, key variables can be selected while retraining confounding variables.

The output of the model will be able to predict the 28th day concrete compressive strength. The purpose is to use the model to achieve the instantaneous 28th day strength the moment a batch of concrete is mixed, as traditionally to achieve the 28th day strength, a cube sample will be crushed on the 28th day to find out the strength. By having instantaneous 28th day strength, faulty batches that do not meet the 28th day design strength requirements can be rectified immediately. Preventing additional cost from hacking or additional supporting structures.

Preliminary Models

This section covers four different models to predict the compressive strength of concrete. The root mean square error (rmse) was used to evaluate the performance of the models

Random Forest

The data was split into three datasets. 60% of the dataset was used for training and 30% of the dataset was used for testing.

No normalization was applied since it is a monotonic transformation that will not affect the decision trees.

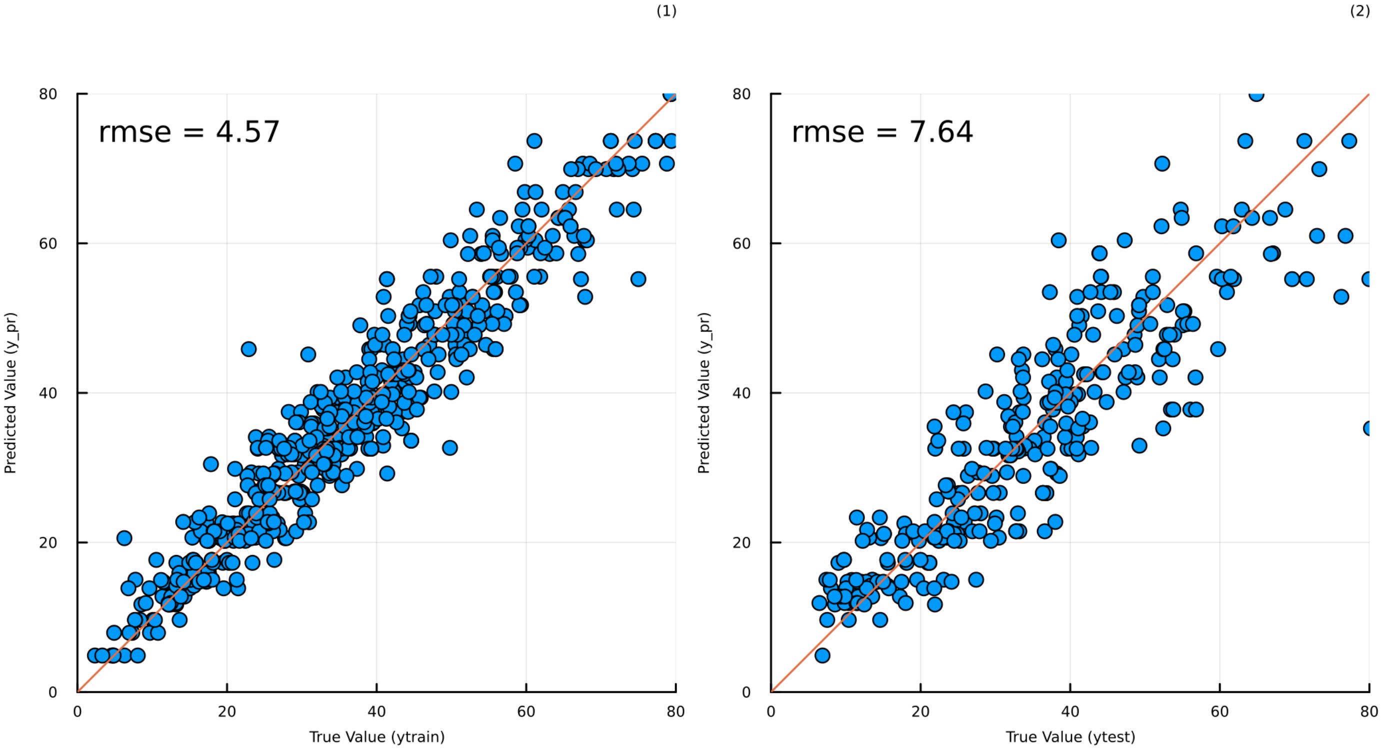

Figure 10: Predictive Model using Decision Tree. (1) Training Data (2) Testing Data

Figure 10 shows the performance of the preliminary decision tree model against the training and testing data. The preliminary model is overfitting to the testing data since the rmse for the testing data is smaller than the training data.

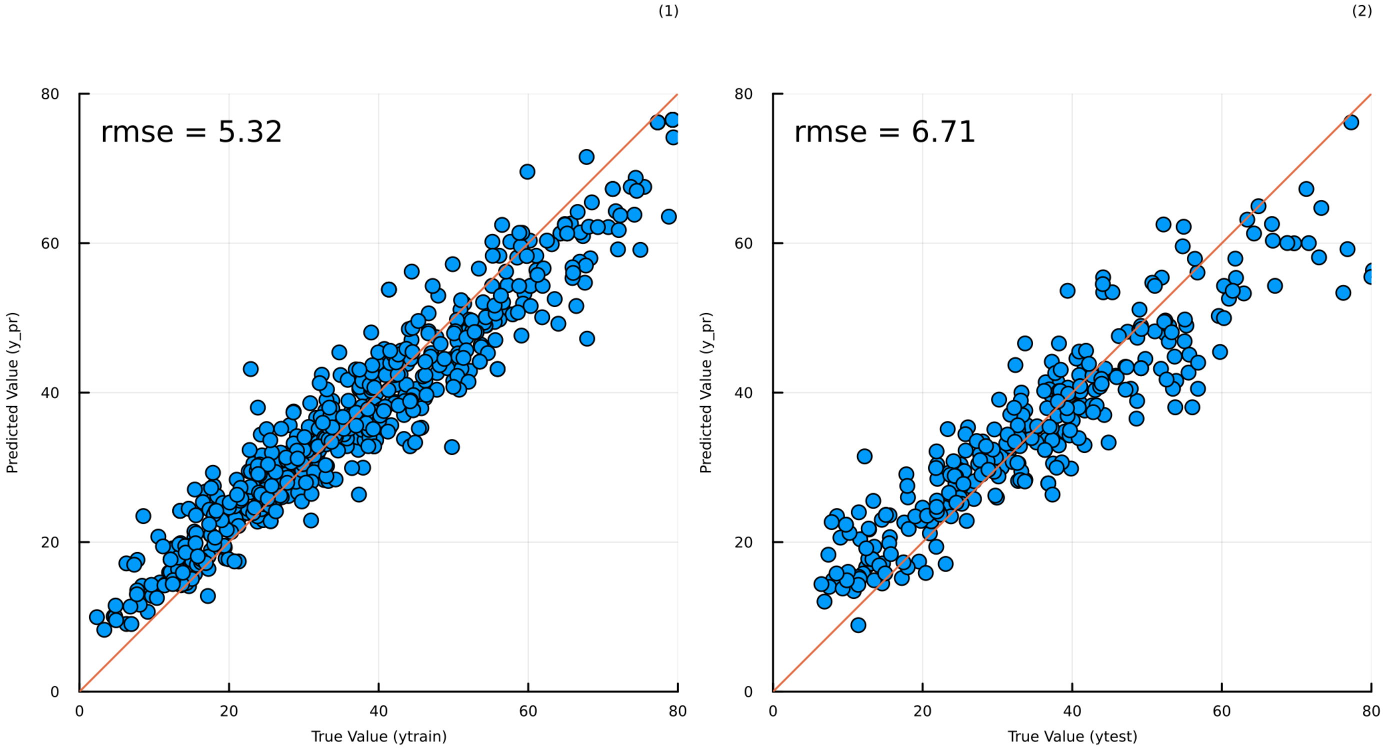

Figure 11: Predictive Model using Random Forest. (1) Training Data (2) Testing Data

Figure 11 shows the performance of the preliminary random forest model against the training and testing data. Like the decision tree model, the random forest model is overfitting to the test data.

Although the decision tree model performs better on the testing data, the random forest model is overfitting less. The next step is to optimize hyperparameters to reduce the difference in rmse of the predicted training and predicted test data. This in turn should also improve the performance of the models on the last 10% of validation data.

SVR

In this section, Support Vector Regression (SVR) with Radial Basis Function (RBF) kernel will be used to predict the concrete compression strength. This section will first cover the basics of SVR and RBF, followed by the model performance based on default hyperparameters. After which, the optimized model performance based on hyperparameter tuning using GridSearchCV will be evaluated. To further optimize the model, normalized will be introduced to the input variables, and the results of that will be discussed. Lastly, the model will be evaluated based on the performance using the validation set, which best mimics real world cases.

The data will also be split into three datasets. 60% of the data was used for training and 30% and 10 % of the data were used for testing and validating, respectively.

SVR & RBF

Support Vector Regression (SVR) is a supervised learning model that is used to predict discrete values. Support Vector Regression uses the same principle as the Support Vector Machine (SVM). The idea behind SVR is to find the best fit line. In SVR, the best fit is the hyperplane that has the maximum number of points. SVR tries to fit the best line within a threshold value. The threshold value epsilon (ε) is the distance between the hyperplane and boundary line. Radial Basis Function (RBF) in its application of Support Vector Machines (SVM), is a type of kernel method used to classify or regress data. it maps non-linear data into a higher dimensional space implicitly by computing the inner products between images of all pairs of data in the feature space (Theodoridis & Koutroumbas, 2009). RBF interpolation is an advanced approximation method, in which the interpolant is the weighted sum of radial basis functions, an example would be the gaussian distribution (Hardy, 1971). This method is popular due to its higher emphasis on radial distances closer to the center and it decreases as the radial distance expands.

Preliminary SVR model

Figure ?? below shows the preliminary SVR model with the training and testing data; with default hyperparameters.

Initial Model using SVR, with default hyperparameters. (1) Training Data (2) Testing Data “LR”){#fig:SVR_initial width="700px"}

From ??, it can be seen that the rmse of both training and testing data yielded a bad result of ~15 MPa. This is because the default hyperparameters C = 1.0, gamma = 1 / (n_features * X.var()) were used. To improve the performance of the model, the hyperparameters will be tuned.

Hyperparameter tuning & Optimized SVR Model

In Support Vector Regression using RBF kernel, there are two hyperparameters C and gamma. The parameter C is a regularization factor that controls the amount of misclassification each training data handles. A larger C will result in a smaller margin of the hyperplane, classifying all the points correctly. On the other hand, a smaller C will increase the margin for the hyperplane and increase the model's tolerance for misclassification. For parameter gamma, it controls the influence of training data. Low values of gamma will generate a flatter decision surface which corresponds to a simpler model. A larger gamma will increase the curvature of the model, fitting it more closely to the training data. There is no rule of thumb to choose C and gamma, as such tuning is required to find the optimal value (Wainer & Fonseca, 2021).

To tune these parameters, GridSearchCV from scikit learn will be used to perform a Grid Search. Grid Search uses different combinations of different values of the hyperparameters and evaluates the performance of a combination of these values, to find out the best value of the hyperparameter (Wainer & Fonseca, 2021).

Figure ?? below shows the optimized SVR model with the training and testing data; after tuned hyperparameters.

Optimized Model using SVR, with tuned hyperparameters C = 10, gamma = 1.1e-5. (1) Training Data (2) Testing Data
“LR”){#fig:SVR_Optimized width="700px"}

The model is trained with the focus of getting the lowest rmse, while ensuring the rmse of training and testing data is consistent. This is to prevent the case of overfitting where the model cannot generalize and fits too closely to the training dataset. From figure ??, it can be seen that after hyperparameter tuning, the model is optimized to yield a rmse of ~8.5 MPa. The hyperparameters used for this model are C=10, gamma = 1.1e-5.

Optimized SVR Model (with normalized input variables)

To further optimize the model, the input variables were normalized. Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change input variables to use a common scale, without distorting differences in the ranges of values.

Figure ?? below shows the SVR model based on normalized input variables with the training and testing data; after hyperparameter tuning.

Normalized input variable Model using SVR, with tuned hyperparameters C = 50, gamma = 0.02. (1) Training Data (2) Testing Data
“LR”){#fig:SVR_Normalized width=“700px”}

From figure ??, it can be seen that after normalizing the input variables (with tuned hyperparameters), the model did not yield a better rmse. It has a rmse of ~8 to 8.5 MPa which his comparable to the rmse without normalization of ~8.5 MPa.

Model Performance on Validation Set

To ensure that the model works well in the real world, the model performance will be evaluated using the validation dataset. The validation dataset is unseen/untested by the model. Because the data is unseen, the model has not been calibrated/optimized based on the validation set. Therefore, the validation set acts like the unseen real world use cases. In this section, the model performance for the Optimized SVR model with and without normalization will be evaluated.

Figure ?? below shows the comparison of the optimized SVR model without normalization and with normalization data, based on the validation set.

Model performance of Optimized SVR Model based on Validation Set (1) without Normalization (2) with Normalization
“LR”){#fig:SVR_Val width=“700px”}

From figure ??, it can be seen that based on the validation set, the Optimized SVR model without Normalization performed better than the Model with Normalization with a lower rmse of 8.87 MPa compared to 12.0 MPa

The Optimized SVR model (w/o normalization) also showed a consistent rmse value of ~8.5 to 8.9 MPa, which suggests that there is no overfitting. Without the evidence of overfitting, it shows that the model has a good indication of its performance to accurately predict real world concrete strength. However, this is not the case for the Optimized SVR model (with normalization), as it yielded a rmse of 12.0, it indicates that there might be overfitting and inaccurate predictions when exposed to real world cases. Although this was not seen in both the training and testing cases, it showed up in the validation case.

Therefore, the best performing model using Support Vector Regression is the Optimized SVR model (w/o normalization), as it yielded both a lower rmse and there is no signs of overfitting.

Linear Regression

Again, the data was split into three datasets. 60% of the data was used for training and 30% and 10 % of the data were used for testing and validating, respectively.

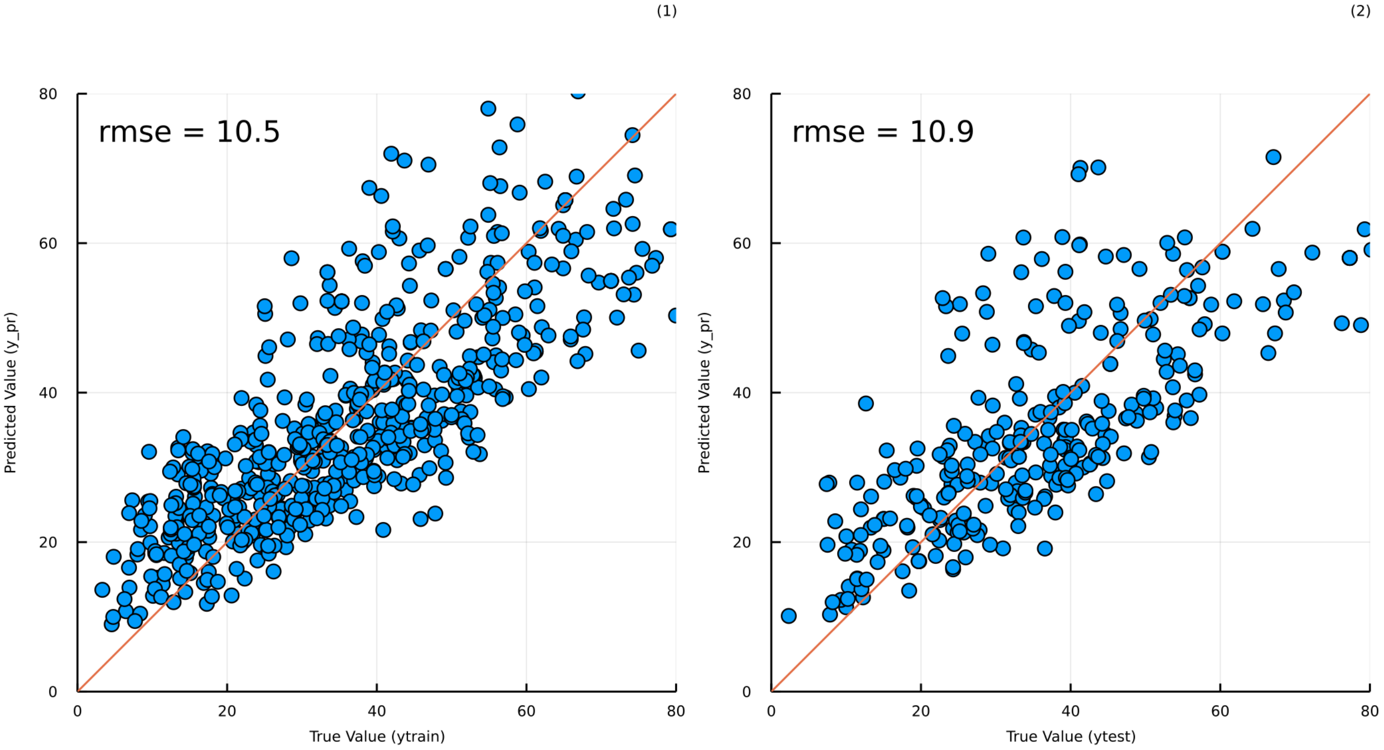

Figure 12: Predictive Model Using Linear Regression. (1) Training Data (2) Testing Data

As we can see from Figure 12, the performance of the linear regression is relatively good. The rmse was relatively low, and it is very similar for both the training and the testing datasets. This also suggests that there was no overfitting in our model.

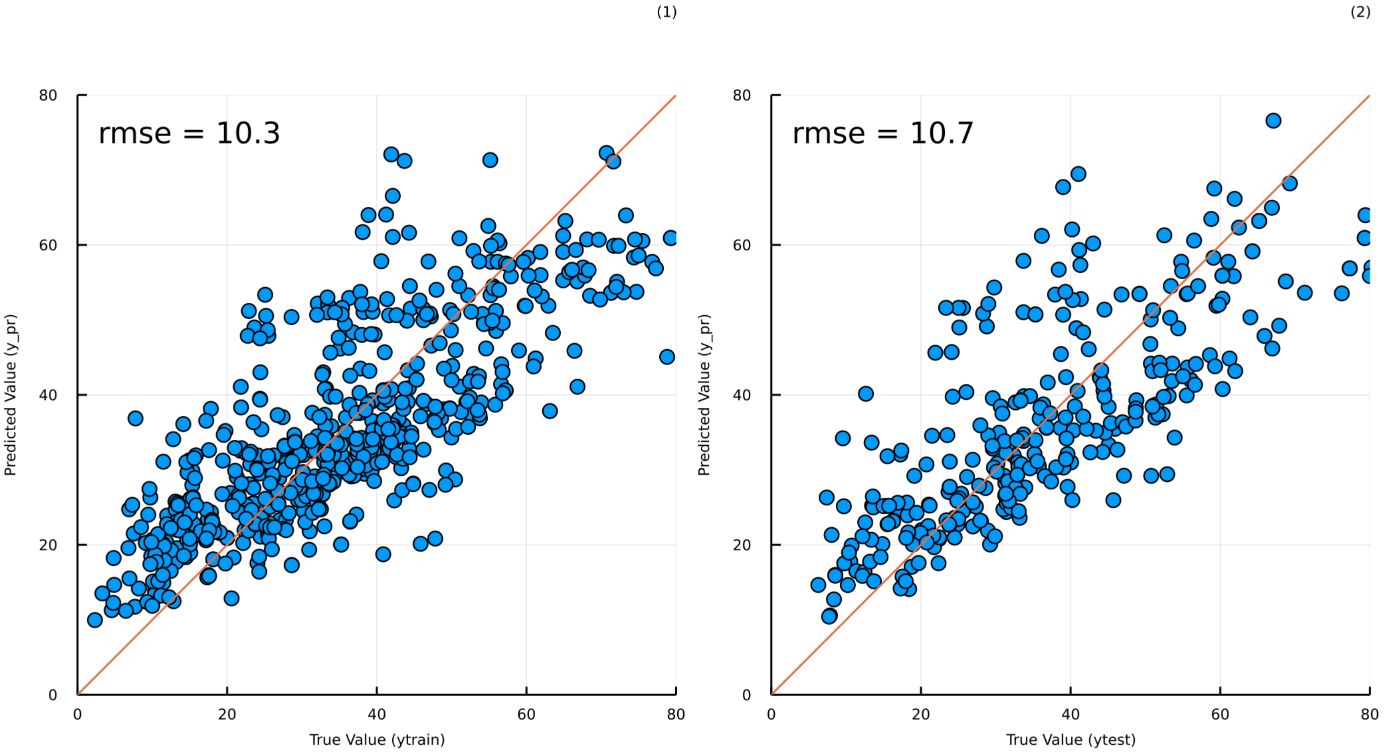

Figure 13: The Effect of Feature Engineering on the Linear Regression Model. (1) Training Data (2) Testing Data

Figure 13 shows the performance of the preliminary linear regression model after logically modifying the input data. The superplasticizer is a material that can be added to the concrete in order to increase its workability while maintaining the same strength. As a result, increasing or decreasing the amount of superplasticizer in the mixture would not affect, by itself, the compressive strength of concrete. And we have seen before from the data that there is no relationship between these two parameters.As a result, this column has been excluded from the datasets here. On the other hand, the most important parameter that affects the compressive strength of concrete is the water to cement ratio which has been included to further increase the accuracy of our linear regression model. We can see that the rmse decreased slightly from 10.9 and 10.5 to 10.7 and 10.3 for training and testing data, respectively.

In the next phase, further investigation of the effect of hyperparameters (e.g. regularization, feature engineering, etc) on the linear regression model will be performed. The purpose of that is to increase the accuracy of this model by reducing rmse for both the training and testing datasets, and with using the validating dataset as well.

Neural Network

Model Comparison

Table of rmse, pros and cons of each model

References

Concrete manual - A water resources technical publication. (8th ed.). (1975).

Merritt, F. S. (1983). Standard Handbook for Civil Engineers, 3 Ed. McGraw-Hill.

Yeh, I.-C. (2007). UCI Machine Learning Repository: Concrete Compressive Strength Data Set. Retrieved September 19, 2022, from https://archive.ics.uci.edu/ml/datasets/concrete+compressive+strength.