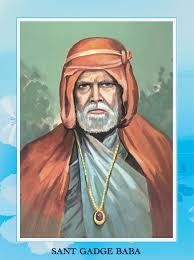
“EDUCATION THROUGH SELF-HELP IS OUR MOTTO”

**RAYAT SHIKSHAN SANSTHA**



 **SADGURU GADAGE MAHARAJ**

**COLLEGE, KARAD**

(An Autonomous College)



**DEPARTMENT OF STATISTICS**

Project report on,

“To Study Concrete Compressive Strength”

Submitted by,

Mr. Kumbhar Ravindranath Raghunath (M. Sc. II)

Roll No. 19

**CERTIFICATE**

This is to certify that the Project Report entitled “To Study Concrete Compressive Strength.” being submitted by Mr. Ravindranath Raghunath Kumbhar as partial fulfillment for the award of degree of M.Sc.-II in Statistics of Sadguru Gadage Maharaj College, Karad is a record of bonafied work carried out by her under supervision and guidance.

To the best of my knowledge and belief, the matter presented in the project report is original and has not been submitted elsewhere for any other purpose.

Place: Karad Date:

Teacher in-charge Examiner PG Co-ordinator Head

Department of Statistics

**ACKNOWLEDGEMENT**

I have great pleasure in presenting this report of successful completion of my project viz. Prediction of Employee Attrition using Techniques of Data Mining.

I take this opportunity to express my great sense of gratitude to my Guide Dr. Mrs. Chavan R.V. of Statistics Department, S.G.M. College, Karad. For granting me permission to undertake this project for their constant encouragement, guidance and inspiration without which I could not have completed this task.

I would like to extend my sincere thanks to Mrs. Mahajan S.V. (Head Department of Statistics), Dr. Mrs. Patil S.P. and Miss. Patil R.D. for their guidance and kind operation in this project study.

Yours Sincerely,

Kumbhar R.R

M.Sc-II

Department of Statistics

**INDEX**

|  |  |  |
| --- | --- | --- |
| Sr. No. | Title | Page No. |
| 1 | Introduction | 5 |
| 2 | Objectives | 6 |
| 3 | Data Source | 7 |
| 4 | Statistical Tools and Techniques | 8 |
| 5 | Research Methodology | 9 |
| 6 | Data Preprocessing | 10 |
| 7 | Exploratory Data Analysis | 11 |
| 8 | Standardization | 19 |
| 9 | Train Test Split | 19 |
| 10 | Fitting Of Models   * Linear Regression * Lasso Regression * Decision Tree * Random Forest * Support Vector Regress * K Nearest Neighbor * Ada Boost Regressor | 20 |
| 11 | Conclusion | 30 |
| 12 | Reference | 32 |
|  |  |  |

**INTRODUCTION**

The Compressive Strength of Concrete determines the quality of Concrete. This is generally determined by a standard crushing test on a concrete cylinder. This requires engineers to build small concrete cylinders with different combinations of raw materials and test these cylinders for strength variations with a change in each raw material. The recommended wait time for testing the cylinder is 28 days to ensure correct results. This consumes a lot of time and requires a lot of labour to prepare different prototypes and test them. Also, this method is prone to human error and one small mistake can cause the wait time to drastically increase.

One way of reducing the wait time and reducing the number of combinations to try is to make use of digital simulations, where we can provide information to the computer about what we know and the computer tries different combinations to predict the compressive strength. This way we can reduce the number of combinations we can try physically and reduce the amount of time for experimentation. But, to design such software we have to know the relations between all the raw materials and how one material affects the strength. It is possible to derive mathematical equations and run simulations based on these equations, but we cannot expect the relations to be same in real-world. Also, these tests have been performed for many numbers of times now and we have enough real-world data that can be used for predictive modelling.

As a data scientist, we generally faced many real-life problems, like- some social issue, construction, etc. we have to solve this type of problem using machine learning techniques, let’s we take an example of construction, what will you refer the word construction? it is the art and science to form objects, systems, or organizations. what will you imagine from the word construction is that mega buildings, machines, material, etc… but you know that what is used to build these mega buildings, for construction we use material, cement, iron rods, etc. where the material is a most important part of building making.

**OBJECTIVES**

* To predict concrete compressive strength by using machine learning algorithm.
* To estimate how much quantity of raw material we need for acceptable compressive strength.
* To study the different relation between the components.
* To find the best effective model for this problem.

**DATA SOURCE**

For academic project the Concrete Compressive Strength dataset which is the secondary dataset has been collected from the ineuron.ai website. ineuron is a reputed institution in the data science training field from India. The Dataset contains 1030 rows and 09 columns. All the columns in this data set are numeric such as Cement, blast furnace slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, Age, Compressive Strength.

* Cement: It is the input variable.
* Blast furnace slag: Blast furnace slag is a non-metallic coproduct which is coming from different blasts. It provides strength to the concrete.
* Fly ash: It is the fine particles of the coal.
* Water: It is the input variable.
* Superplasticizer: It is also known as high range water reducer. Their addition to concrete allows reduction of water to the cement ratio without negatively affecting.
* Coarse Aggregate: Coarse aggregate are the medium grained particulate material used in construction, including sand, gravel, crushed stone.
* Fine Aggregate: Fine aggregate are similar to the coarse aggregate and they are much finer.
* Age: Age in days.
* Compressive strength: Compressive strength of cement block.

**STATISTICAL TOOLS AND TECHNIQUES**

Technique:

* Linear Regression
* Ridge Regression
* Lasso Regression
* Decision Tree
* Random Forest
* Support Vector Regressor
* K Nearest Neighbour
* Ada Boost Regressor

Tools:

* Jupyter Notebook

**RESEARCH METHODOLOGY**

* Data Preprocessing:

Data preprocessing is undertaken in following steps,

Data understanding and Description:

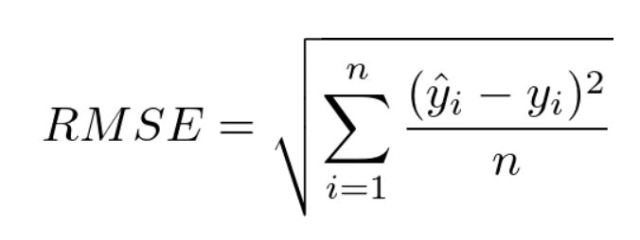
To predict the Concrete Compressive Strength based on the variables, describe the variables before analyzing them so that have a clear understanding. The dataset consists of 1030 instance with 9 attributes. There are 8 input variables and 1 output variable. Seven input variables represent the amount of raw material (measured in Kg/m^2) and one represents age (in days). The target variable is Concrete Compressive Strength measured in (MPa — Mega Pascal). We shall explore the data to see how input features are affecting compressive strength.

Data Cleaning and Reduction:

In Data Cleaning, to check for the missing values and observed that the dataset has no any missing value and it is complete. Also, to check if dataset has any duplicated values and observed no any duplicated values.

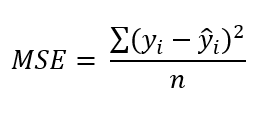
RMSE:

Root mean square error is one of the most commonly used measure for evaluating the quality of predictions. It shows how far predictions fall from measured true values using Euclidean distance. Lower value of RMSE indicates better fit.



MSE:

Mean square error measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values.

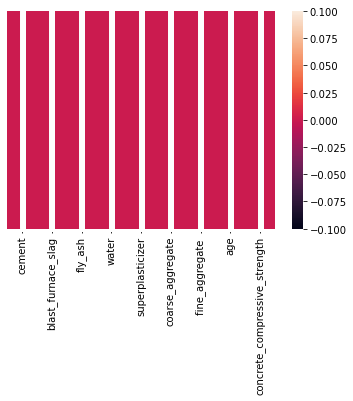


**EXPLORATORY DATA ANALYSIS**

This includes checking for any missing values, plotting the features with respect to the target variable, observing the distributions and relationship of all the features and so on. It is an important step to build any machine learning project, it is an approach of analyzing datasets to summarize their main characteristics.

These are some visualization.

Check missing values by heatmap:



Fig(A): Heatmap Of Missing Values

From Fig(A),

we can observe that there are no any missing present in data set.

Heatmap:

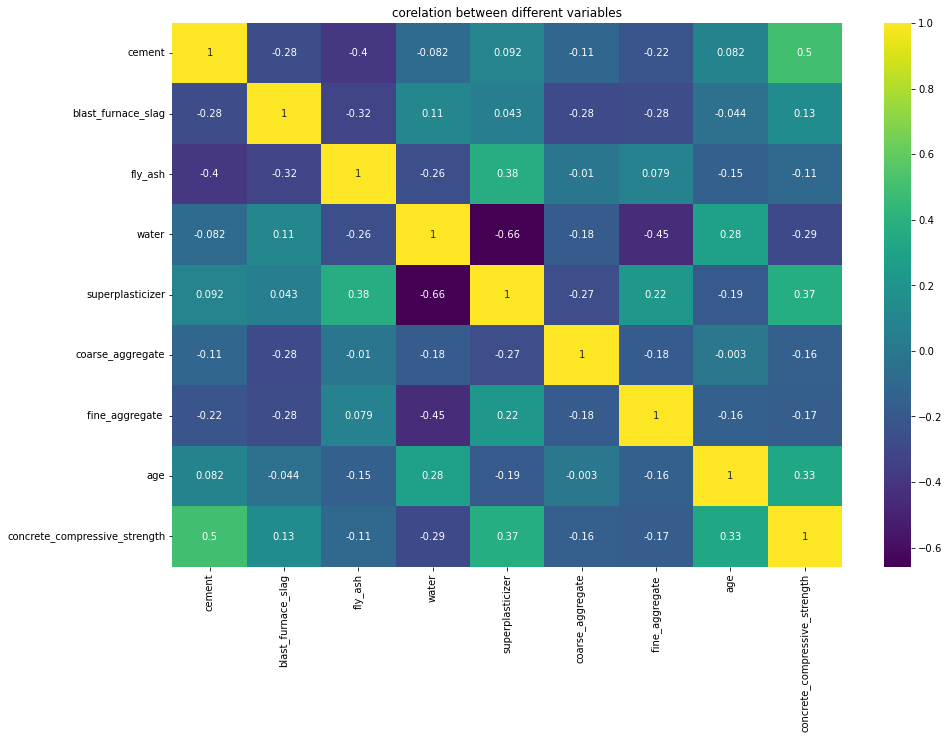
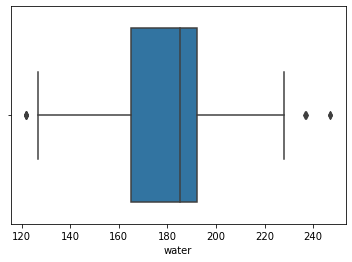


Fig (B): Correlation between variables

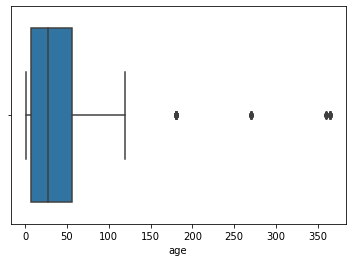
From Fig(B), we can observe that there is no collinearity between the feature column. Only the Cement and Compressive Strength have the strong positive correlation. This is true because strength concrete indeed increases with an increase in the amount of cement used in preparing it.

* A negative correlation between Superplasticizer and Water.
* Positive correlation between Superplasticizer and Fly ash, Fine aggregate.

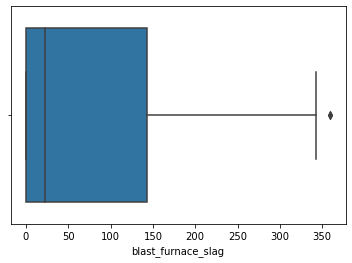
Outlier Detection:



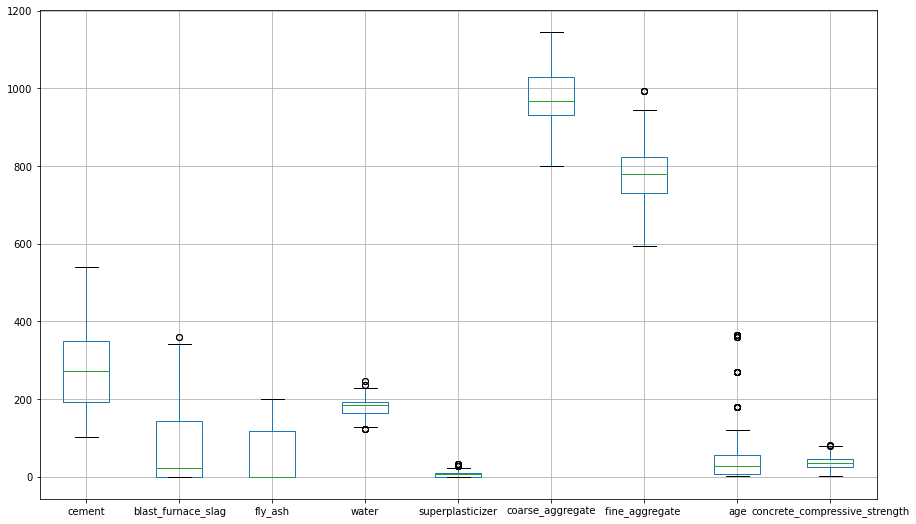
Fig(C): Water Outliers



Fig(D): Age Outlier

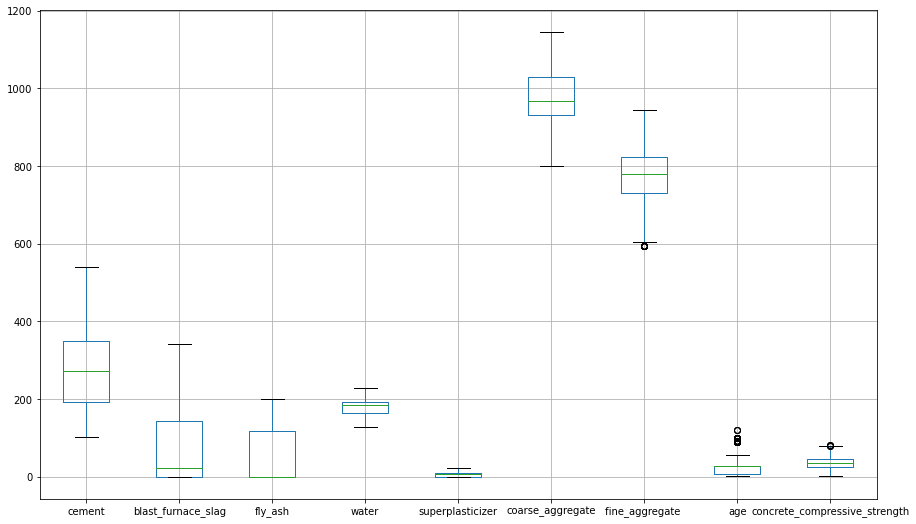


Fig(E)



Fig(F):

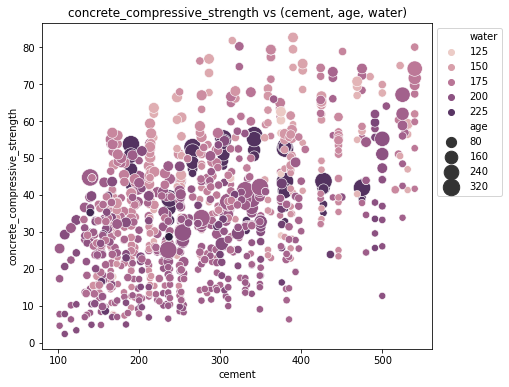
From the figures C,D,E,F we can observe that the columns blast furnace slag, water, fine aggregate and age are contains outliers. We can find the outliers by using Inter Quartile Range method.



Fig(G)

As we see in Fig(G), Outliers are replaced by the median.

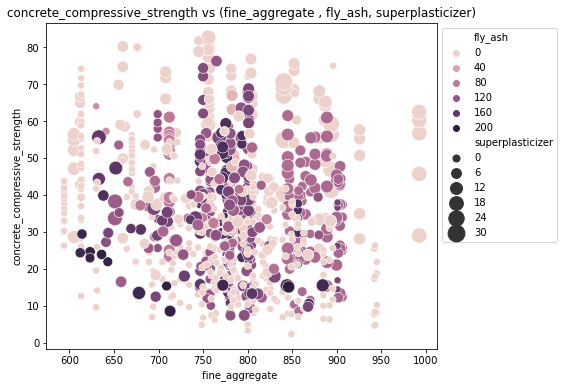
Scatterplots:



Fig(H)

From the fig(H), we observe that,

* Compressive strength is increases when amount of Cement is increases. as the dots move up when we move towards right on the x-axis.
* Compressive strength is increases with Age.
* **Concrete strength increases when less water is used** in preparing it since the dots on the lower side (y-axis) are darker and the dots on higher-end (y-axis) are brighter.

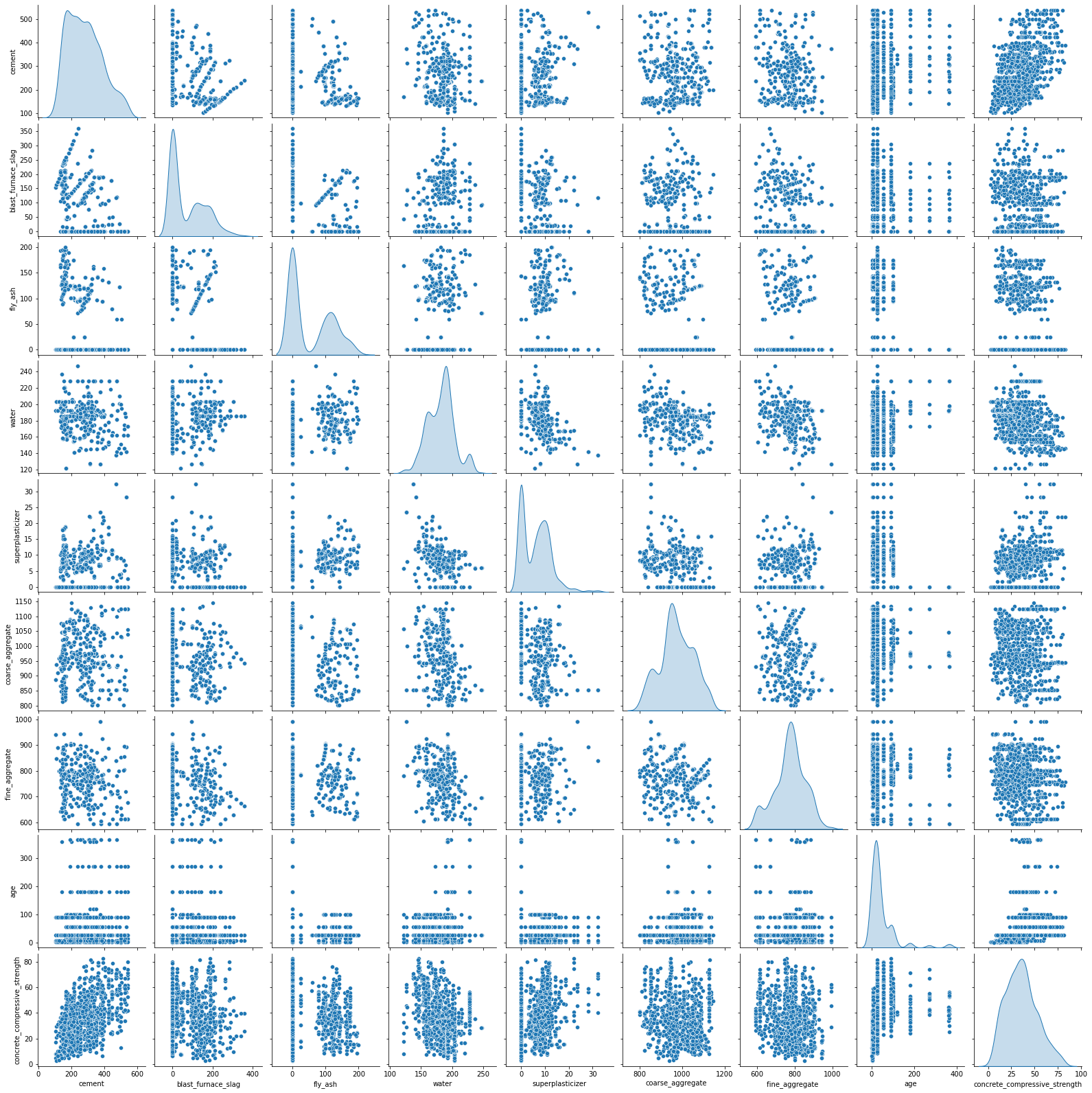


Fig(I)

From Fig(I),

* **Compressive strength decreases Fly ash increases**, as darker dots are concentrated in the region representing low compressive strength.
* **Compressive strength increases with Superplasticizer** since larger the dot the higher they are in the plot.

Pair plot:

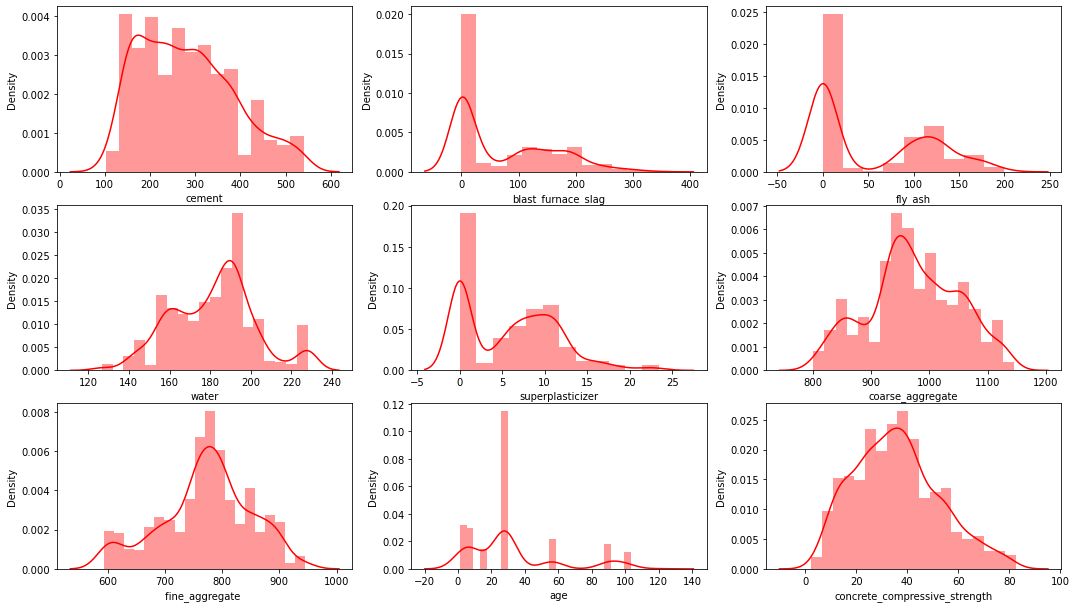


Fig(J)

From the above Fig(J),

We can observe that there is no any correlation between the independent variables.

Distribution Plot:



Fig(K)

From the Fig(K),

* Cement looks almost normal.
* Blast furnace slag and Fly ash has right skewed.
* Water is a slightly left skewed.
* Superplasticizer has two gaussians and is rightly skewed.
* Coarse aggregate has three gaussians and is rightly skewed.
* Fine aggregate looks almost normal.
* Age has multiple gaussians and is rightly skewed.

Standardization:

Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data

We use Standard scalar() for standardize the data.

Train Test Split:

The train-test split is a technique for evaluating the performance of a machine learning algorithm. It is used for classification or regression problems for any Supervised machine Learning Algorithm.

The procedure involves taking a dataset and dividing it into two subsets. The first one is the Training set which is used to fit the model and the second is not used to train but to make predictions known as the testing set.

* Train Dataset: It is used to fit the machine learning model.
* Test Dataset: It is used to evaluate the fitted machine learning model.

x train size: 824

x test size: 206

y train size: 824

y test size: 206

**MODEL BUILDING**

The target variable is Concrete Compressive Strength that is the dataset is labelled so Supervised Machine Learning Algorithms are used for the prediction of the Target Variable.

Following techniques are used,

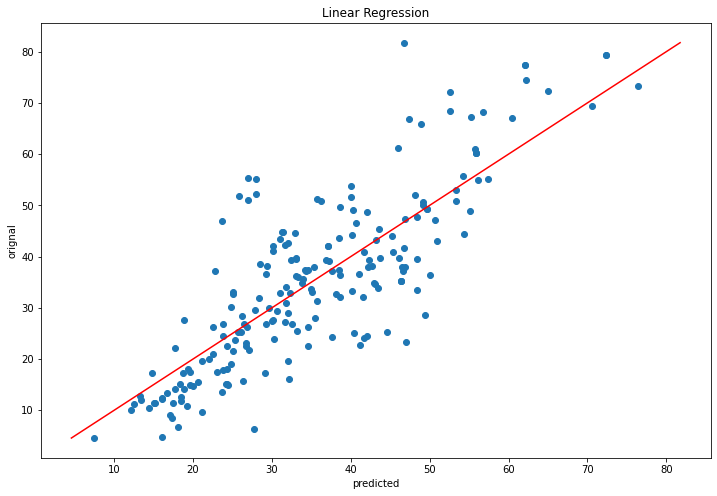
Linear Regression:

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

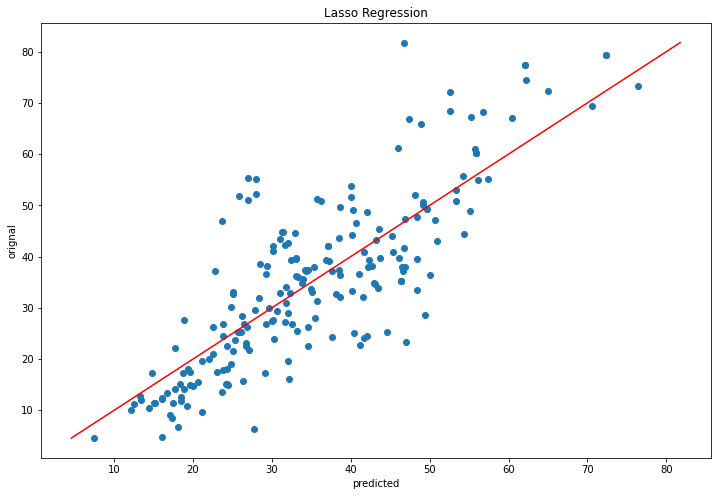
Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

Lasso Regression:

Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean.



Fig(L)



Fig(M)

Conclusion:

From the above Fig(L) and Fig(M) we can observe that if the predicted values and the target values are equal, then the points on the scatter plot will lie on the straight line. As we can see here, none of the model predicts the Compressive Strength correctly.

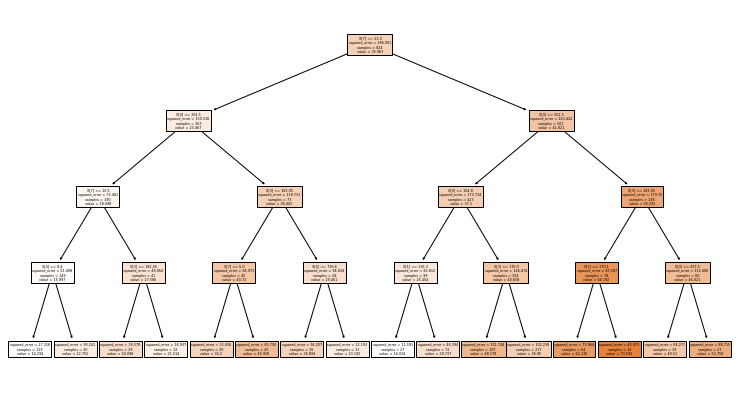
|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | MSE | RMSE | R^2 |
| Linear Regression | 91.92 | 9.58 | 0.68 |
| Lasso Regression | 93.31 | 9.58 | 0.68 |

From the above table we can say that there is not much difference between the performance with these two algorithms, in this case, we will look at other models.

Accuracy of the model: 68%

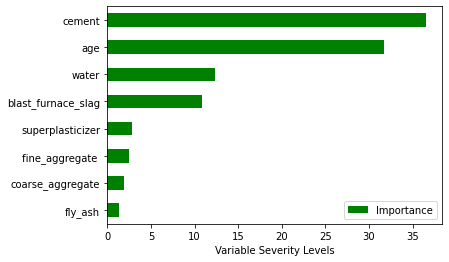
Decision Tree:

Decision tree is a Supervised Machine Learning Technique since, both values of independent and target variable are known. Decision tree is very flexible used for both the classification and regression techniques. It can be used when the dependent variable is not normally distributed. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data.



Fig(R): Decision Tree

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | MSE | RMSE | R^2 |
| Decision Tree | 77.55 | 8.80 | 0.75 |



Fig(N): Feature Importance of Decision Tree

Conclusion:

The Root Mean Squared Error (RMSE) has come down from 9.58 to 7.36, so the Decision Tree Regressor has improved the performance by a significant amount.

From the Fig(N): Feature Importance of Decision Tree, we can observe that top 4 contributors to strength are cement, age, water, superplasticizer, blast furnace slag.

Accuracy of Decision Tree: 77%

Support Vector Regressor:

Support Vector Regression is a supervised learning algorithm that is to predict discrete values. Support vector regressor uses the same principle as the as the SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points.

Regression model that try to minimize the error between the real and predicted values, the SVR tries to fit the best line within a threshold value. The threshold value is the distance between the hyperplane and the boundary line. The fit time complexity of SVR is more than quadratic with the number of samples which makes it hard to scale to datasets with more than a couple of 10000 samples.

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | MSE | RMSE | R^2 |
| Support Vector regression | 79.8 | 8.93 | 0.67 |

Conclusion:

From the above table we can say that, mean square error is 79.8 and Root mean square error is 8.93. As compared to previous model Decision Tree accuracy of the Support vector regression is to less.

Accuracy of Support Vector Regression: 67%

K Nearest Neighbour:

K Nearest Neighbour is one of the simplest machine learning algorithms based on supervised learning techniques. KNN algorithm assumes the similarity between the new case/ data and put the new case into the category that is most similar to the available categories. KNN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using KNN algorithm.

KNN is a non- parametric algorithm which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | MSE | RMSE | R^2 |
| K Nearest Neighbour | 86.61 | 9.31 | 0.64 |

Conclusion:

From the model K Nearest Neighbour mean square error is 86.61 and root mean square error is 9.31, so the K Nearest Neighbour model cannot gives the better accuracy as compare to the Decision tree. We use another model for the better accuracy.

Accuracy of K Nearest Neighbour: 64%

Random Forest:

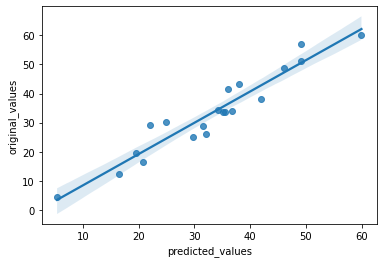
Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, **"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."** Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

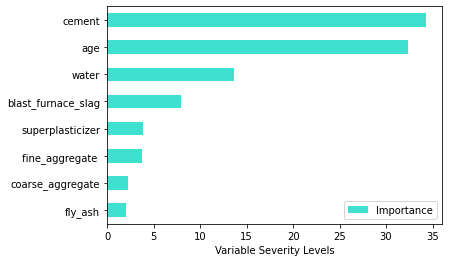
**The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.**

**Here are the Actual and Predicted values of Random Forest Regressor,**

|  |  |  |
| --- | --- | --- |
| **No**. | **Predicted values** | **Original values** |
| **456** | 37.8192 | 43.39 |
| **988** | 19.9001 | 16.5 |
| **809** | 19.6708 | 19.69 |
| **581** | 17.3325 | 12.37 |
| **549** | 34.76886833 | 41.68 |
| **312** | 46.2328 | 48.72 |
| **436** | 33.75658571 | 33.69 |
| **476** | 29.74332833 | 25.02 |
| **612** | 32.2665075 | 25.97 |
| **652** | 5.3437 | 4.57 |
| **941** | 30.86790833 | 28.94 |
| **67** | 25.2054 | 30.28 |
| **473** | 47.73625833 | 51.02 |
| **815** | 34.5781 | 33.8 |
| **437** | 40.3363 | 38.2 |
| **471** | 47.73625833 | 57.03 |
| **397** | 22.5109 | 29.16 |
| **621** | 34.837525 | 34.49 |
| **143** | 61.4591 | 60.2 |
| **197** | 36.7909 | 33.96 |



Fig(O): Reg Plot of Random Forest



Fig(P): Feature Importance of Random Forest

Conclusion:

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | MSE | RMSE | R^2 |
| Random Forest Regressor | 27.66 | 5.26 | 0.88 |

From the Fig(O), If the predicted values and the target values are equal, then the points on the reg plot will lie on the straight line, as we can see here more points are closer to the regression line that is the distance between the actual values and the predicted values is very less.

From the fig(O): Feature Importance of Random Forest, we can find the importance of the component for the better strength. From fig we observe that cement, age, water, blast furnace slag, superplasticizer this component provides more strength.

From the above table we observe that mean square error and the root mean square error is very less that is 27.66 and 5.26. The accuracy of the Random Forest Regressor model is better as compare to the previous model.

Accuracy of Random Forest Regressor: 88%

Ada Boost Regressor:

AdaBoost algorithm, short for Adaptive Boosting, is a Boosting Technique used as an Ensemble Method machine learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances. Boosting is used to reduce bias as well as variance for supervised learnings. It works on the principle of learners growing sequentially. Except for the first, each subsequent learner is grown from previously grown learners. In simple words, weak learners are converted into strong ones. The AdaBoost algorithm works on the same principle as boosting with a slight difference.

AdaBoost is an ensemble learning method (also known as “meta-learning”) which was initially created to increase the efficiency of binary classifiers. AdaBoost uses an iterative approach to learn from the mistakes of weak classifiers, and turn them into strongones.

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | MSE | RMSE | R^2 |
| Ada Boost Regressor | 67.97 | 8.24 | 0.72 |

Conclusion:

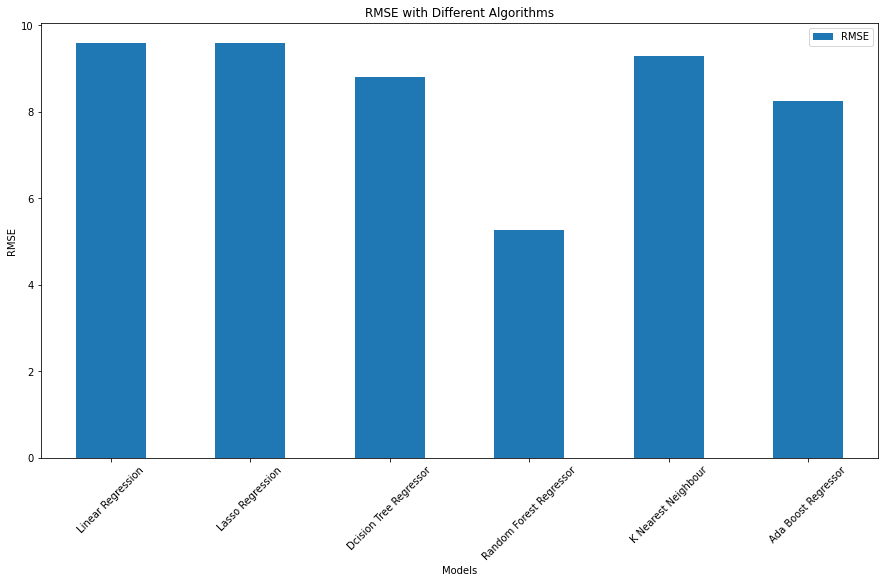
From the table Ada boost regressor model is best as compare to linear regression, lasso, decision tree and SVR.

Accuracy of Ada Boost Regressor: 72%

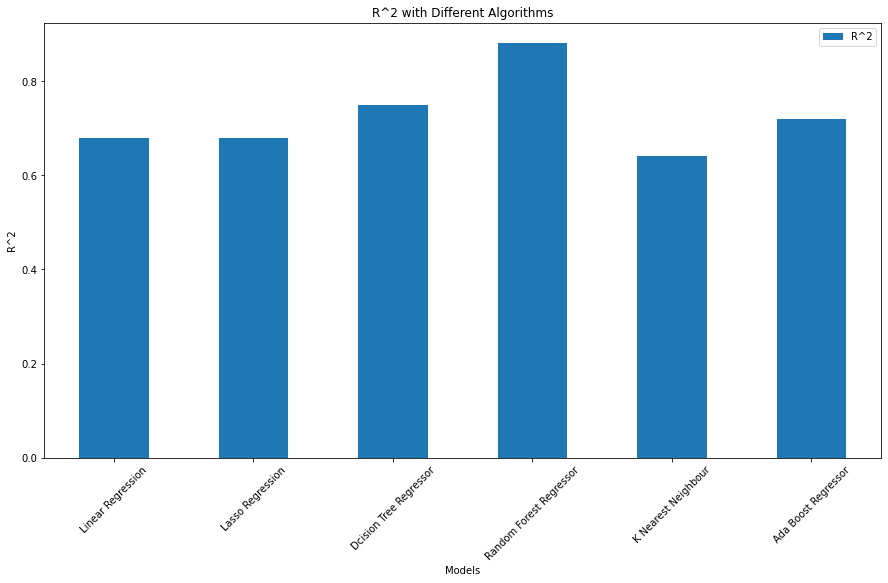
**OVERALL CONCLUSION**

Results:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Algorithm** | **R^2** | **RMSE** |
| **0** | Linear Regression | 0.68 | 9.58 |
| **1** | Lasso Regression | 0.68 | 9.58 |
| **2** | Decision Tree | 0.75 | 8.8 |
| **3** | Random Forest | 0.88 | 5.26 |
| **4** | K Nearest Neighbour | 0.64 | 9.3 |
| **5** | Ada Boost regressor | 0.72 | 8.24 |



Fig(P): RMSE vs Models



Fig(Q): R Square vs Models

From the table, we can observe that the model Random Forest gives 88% accuracy as compare to the other models.

From fig(P), root mean square error of random forest is 5.26 that is it is less than other models.

From the fig(Q), R square error of the random forest is 0.88 that is very high as compare to the other model.

We have analysed the Compressive Strength Data and used Machine Learning to Predict the Compressive Strength of Concrete. We have used Linear Regression and its variations, Decision Trees, SVR, Ada Boost and Random Forests to make predictions and compared their performance. Random Forest Regressor has the lowest RMSE and is a good choice for this problem.

Here the highest accuracy means it predicts the quality of concert by using training, which contains independent variables, and also it gives less error rate.

Reference:

1. Hongwei Song a, Ayaz Ahmad b c, Furqan Farooq b c, Krzysztof Adam Ostrowski c, Mariusz Maślak c, Slawomir Czarnecki d, Fahim Aslam e.

Predicting the compressive strength of concrete with fly ash admixture using machine learning algorithms

1. H N Muliauwan1, D Prayogo1, G Gaby1 and K Harsono1

Prediction of Concrete Compressive Strength Using Artificial Intelligence Methods

1. Alexey N. Beskopylny 1,\* , Sergey A. Stel’makh 2, Evgenii M. Shcherban’ 3, ,Levon R. Mailyan 4 Besarion Meskhi 5

Concrete Strength Prediction Using Machine Learning Methods

1. B. Ouyang, Y. Song, Y. Li, F. Wu, H. Yu, Y. Wang, G. Sant, and M. Bauchy

Predicting Concrete’s Strength by Machine Learning:

Balance between Accuracy and Complexity of Algorithms

1. Chongchong Qi,1,2 Binhan Huang,2 Mengting Wu,2 Kun Wang,1 Shan Yang,2 and Guichen Li3,\*

Concrete Strength Prediction Using Different Machine Learning Processes: Effect of Slag, Fly Ash and Superplasticizer

1. Sarmad Dashti Latif. Concrete compressive strength prediction modeling utilizing deep learning long short-term memory algorithm for a sustainable environment
2. MELTEM ÖZTURAN1, BIRGÜL KUTLU1, TURAN ÖZTURAN2

COMPARISON OF CONCRETE STRENGTH PREDICTION

TECHNIQUES WITH ARTIFICIAL NEURAL NETWORK

APPROACH

1. Vimal Rathakrishnan, Salmia Bt. Beddu & Ali Najah Ahmed

Predicting compressive strength of high-performance concrete with high volume ground granulated blast-furnace slag replacement using boosting machine learning algorithms

Websites:

* [www.analyticsvidhya.com](http://www.analyticsvidhya.com)
* [www.ineuron.ai](http://www.ineuron.ai)
* [www.kaggle.com](http://www.kaggle.com)