*In-house concrete compressive strength Prediction*

**Problem Statement**:-

**Concrete** is a composite material composed of fine and coarse aggregate bonded together with a water and cement paste that hardens over time, making the structure to resist the load applied on it. The engineering properties of concrete depends on various parameters such as the properties of the mixture, admixtures, water-cement ratio, aggregate (fine & coarse), age etc. Thus, the compressive strength of concrete too depends on the various parameters but the complex nature of this parameter is still pushing towards a more comprehensive and accurate method. So, the final model for predicting the compressive strength of the concrete is a nonlinear function of age and ingredients.

The target variable is the compressive strength of the concrete. Predicting the compressive strength of the concrete with a given amount of the ingredients is necessary so that we will be able to know what is the strength of concrete. Eventually, making the structure more strengthened and increasing the load carrying capacity of the structure.

**Data analysis:-**

The dataset consist of the following 9 components, namely:

1. Cement (kg in a m^3 mixture)
2. Blast Furnace slag (kg in a m^3 mixture)
3. Fly ash (kg in a m^3 mixture)
4. Water (kg in a m^3 mixture)
5. Super plasticizer (kg in a m^3 mixture)
6. Coarse Aggregate (kg in a m^3 mixture)
7. Fine Aggregate (kg in a m^3 mixture)
8. Age (day)
9. Concrete Compressive Strength (MPa, megapascals)
10. Cement (kg in a m^3 mixture): It shows the density of the cement, since the unit of the column is kg/m^3. The unit of density is also kilogram per cubic meter.
11. Blast Furnace slag (kg in a m^3 mixture): It shows the density of the Slag formed in the blast furnace. This is basically the residue which are left in the furnace after the heating. It remains in the fluid form.
12. Fly ash (kg in a m^3 mixture): It shows the density of the dust or ash left after heating the furnace.
13. Water (kg in a m^3 mixture): It shows the density of the water required for the burning process.
14. Super plasticizer (kg in a m^3 mixture): It also known as high range water reducers, are additives used in making high strength concrete. The column shows its density.
15. Coarse Aggregate (kg in a m^3 mixture): The column shows the density of the coarse.
16. Fine Aggregate (kg in a m^3 mixture): This shows the density of the fine aggregate used in the blast furnace.
17. Age (day): It gives the idea of how old the supplements are.
18. Concrete Compressive Strength (MPa, megapascals): Compressive strength is the load per unit area. It also gives us the yielding strength and life of the concrete used in the building.

**EDA Concluding Remarks:-**

Exploratory data analysis (**EDA**) is used by data scientists to analyse and investigate data sets and summarize their main characteristics, often employing data visualization methods. It can also help to determine if the statistical techniques you are considering for data analysis are appropriate.

**Scatterplot:** A scatterplot is a type of plot that displays values pertaining to typically two variables against each other, usually it is a dependent variable to be plotted against independent variable.

The scatterplot is plotted between each variable of the datasets against the target variable, i.e., the compressive strength of the concrete.

This gives the visualization of the datasets vs. the compressive strength of the concrete (CMS).

1. Age vs CMS: It shows that at the starting the strength was weak and it solidifies with the time and again it loses its strength.

2. Fine Aggregate vs CMS: It shows with the density of 750 to 900 of fine aggregate, maximum compressive strength lies.

3. Coarse Aggregate vs CMS: It shows with the density of 950 to 1100 of coarse aggregate, maximum compressive strength lies.

4. Superplasticizers vs CMS: It shows with the density of 5 to 12 of superplasticizers, maximum compressive strength lies.

5. Water vs CMS: It shows with the density of 160 to 210 of water, maximum compressive strength lies.

6. Fly ash vs CMS: It shows with the density of 100 to 175 of fly ash, maximum compressive strength lies.

7. Blast Furnace Slag vs CMS: It shows with the density of 0 to 200 of blast furnace slag, maximum compressive strength lies.

8. Cement vs CMS: It shows with the density of 100 to 400 of cement, maximum compressive strength lies.

**Correlation:** Correlation is a statistical matrix for measuring to what extent different variables are interdependent. Measures to what extent different variables are interdependent. Correlation does not imply causation.

Some columns are making positive correlation while some columns are making negative correlation. They are as mentioned below.

Columns making positive correlation are:-

1. Age
2. Superplasticizers
3. Blast Furnace Slag
4. Cement

Columns making negative correlation are:-

1. Fine aggregate
2. Coarse aggregate
3. Water
4. Flyash

The positive correlated columns has a great impact on the target column, while negative correlated columns has less or zero impact on the target column.

**Description of Datasets:** In description of the datasets, we determine the count, mean, standard deviation, minimum, 25%, 50%, 75% and maximum of each column in the datasets.

**Pre-processing Pipeline:-**

To check whether the datasets are normalised or not we use ‘SUBPLOTS’ and ‘DISPLOTS’. As the data are not normalised, so we can go for Data Cleaning.

**Data Cleaning:** The process of converting or mapping data from the initial “raw” form into another format, in order to prepare the data for further analysis. It is also known as ‘Data Pre-processing’ or ‘Data Wrangling’.

To do the data cleaning it is necessary to represent the data in the graphical form i.e., in the distribution form, as it will be easier in removing the skewness of the data.

Histogram is used to distribute the data, as the data is not normally distributed so we can say that there is skewness in the data.

**Skewness:** It is the measure of how much the probability distribution of a random variable deviates from the normal distribution.

Here we observe that

The columns having high skewness are:-

1. Age- 3.269

2. Superplasticizers- 0.907

3. Blast Furnace Slag- 0.801

These columns are having good correlation with the target column, so these columns cannot be dropped. So now we need to go for transforming the columns having skewness.

**Data Transformation:-** We need to transform the data columns with skewness. This is done by:-

Outliers Check: We need to plot the plot box visualization so as to check the outliers in each columns. Then we used **Z-Score** method to remove these outliers.

The percentage of data lost during removing the outliers is 4.76%. As it is less than 5% we can go for outliers removal. Now we can check the percentage of data lost in the removal of outliers by **IQR Method**, it is 8.64%.

Hence, we can go for Z-Score method as it results in less data loss.

At last, after getting the percentage of data loss, we will go for transforming the data by **Power Transform** method.

**Building Machine Learning Models:-**

The **Model Development:** A model can be thought of as a mathematical equation used to predict a value given one or more other values.

Relating one or more independent variables to dependent variables.

Usually the more relevant data you have the more accurate your model will be.

1. The data is converted into a single scale by **Standard Scaler** method.

2. In **Linear Regression Method**, different random state are chosen, to find at which random state the training r2 score and testing r2 score are the most similar.

Here, at random state=96, the model performs the best, with a r2 score of 81.2

**Regularization**:- The regularization term, or penalty, imposes a cost on the optimization function for overfitting the function or to find an optimal solution. In machine learning, regularization is any modification one makes to a learning algorithm that is intended to reduce its generalization error but not its training error.

This is a form of regression, that constraints (regularizes) or shrinks the coefficient estimates towards zero. In other words, this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.

**1. Lasso:-** Lasso is another variation, in which the above function is minimized. It’s clear that this variation differs from ridge regression only in penalizing the high coefficients.

At different CV value, the cross validation score and R2 score is found out.

Hence, at CV=3, it gives the best value.

Cross validation score= 72.366

R2 score= 81.187

Now, the errors we get are:-

Mean Absolute error= 5.4699

Mean Squared error= 50.089

Root Mean Square error= 7.0774

**2. Ridge regression:-** Ridge Regression method is modified by adding the shrinkage quantity. Now, the coefficients are estimated by minimizing this function. Here, λ is the tuning parameter that decides how much we want to penalize the flexibility of our model. The increase in flexibility of a model is represented by increase in its coefficients, and if we want to minimize the above function, then these coefficients need to be small. This is how the Ridge regression technique prevents coefficients from rising too high.

Hence, the values are found:-

Cross validation score= 71.6781

R2 score= 81.2729

Now, the errors we get are:-

Mean Absolute error= 5.60977

Mean Squared error= 52.1911

Root Mean Square error= 7.2243

**Ensemble Techniques:-** Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produces more accurate solutions than a single model would. This has been the case in a number of machine learning competitions, where the winning solutions used ensemble methods.

Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model.

**Decision Tree:-** A Decision Tree determines the predictive value based on series of questions and conditions. For instance, this simple Decision Tree determining on whether an individual should play outside or not. The tree takes several weather factors into account, and given each factor either makes a decision or asks another question.

Hence, the R2 score = 83.8322

Cross Val score = 32.057

Now, the errors we get are:-

Mean Absolute error= 3.8303

Mean Squared error= 45.0389

Root Mean Square error= 6.7125

**Random Forest Regression:-** Random Forest Models can be thought of as Bagging, with a slight tweak. Random Forest models decide where to split based on a random selection of features. Rather than splitting at similar features at each node throughout, Random Forest models implement a level of differentiation because each tree will split based on different features. This level of differentiation provides a greater ensemble to aggregate over, ergo producing a more accurate predictor.

Hence, the values are found:-

Cross validation score= 36.11565

R2 score= 92.3561

**Gradient Boosting Regression:-** Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. Gradient boosting is a type of machine learning boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error.

Hence, the values are found:-

Cross validation score= 60.2611

R2 score= 92.3561

**Support Vector Regression:-**

Hence, the values are found:-

Cross validation score= 68.951

R2 score= 80.962

**Concluding Remarks:-**

The best model is the **LASSO** since the difference between percentage score of cross validation and score of R2 score is optimum.

Hence, **CV=3**, and the value are as follows:-

**Cross validation score= 72.366**

**R2 score= 81.187**

Now, the **Errors** we get are:-

**Mean Absolute error= 5.4699**

**Mean Squared error= 50.089**

**Root Mean Square error= 7.0774**

Therefore, the above built model will be able to predict the concrete compressive strength with the required amount of the ingredients, as mentioned in the datasets.