VI SEM U.G. Project Report 2021

On

"Data Visualization and Prediction of Slump by Machine Learning Algorithms"

Supervisor

Dr. Suresh Kumar

(Assistant Professor)

Submitted By

Shivam Kumar Singh

DEPARTMENT OF CIVIL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY (BANARAS HINDU UNIVERSITY)

VARANASI-221005, INDIA

20 April 2021

Roll. No.:18065067

Department of Civil Engineering

Indian Institute of Technology, (BHU) Varanasi

VARANASI - 221005, INDIA

Dr. Suresh Kumar

Assistant Professor

Department of Civil Engineering

IIT(BHU) Varanasi

20 April 2021

This is to certify that the present work entitled “Data Visualization and Prediction of Slump by Machine Learning Algorithms” has been carried out by Mr. Shivam Kumar Singh under my direct supervision and guidance during his academic semester VI. He has conducted his studies very sincerely, meticulously and methodically and the results of the work are embodied in this report. I wish him success in all his future endeavours.

Dr. Suresh Kumar

Supervisor

E-mail: skumar.civ@iitbhu.ac.in

Acknowledgement

During the period of my project in this University, several respectful and affectionate persons helped directly and indirectly to my project. Without their support it would be impossible for me to accomplish my work, that's why I wish to dedicate this section to recognize their support.

First and foremost, I would like to thank my guide Prof. Suresh Kumar for guiding me thoughtfully and efficiently through this project, giving me an opportunity to work at my own pace along my own lines, while providing me with very useful directions whenever necessary.

I offer my sincere thanks to all other persons who knowingly or unknowingly helped me complete this project. I perceive as this project as a big milestone in my career development. I will strive to use gained skills and knowledge in the best possible way, and I will continue to work on their improvement, in order to attain desired career objectives. Hope to continue cooperation with all in the future.

Shivam Kumar Singh

Junior Undergraduate

Dept. of Civil Engineering & Technology

Indian Institute of Technology (BHU) Varanasi-221005

Abstract

Concrete workability, quantified by concrete slump, is an important property of a concrete mixture. Concrete slump is generally known to affect the consistency, flowability, pumpability, compactibility, and harshness of a concrete mix. Hence, an accurate prediction of this property is a practical need of construction engineers. This research proposes a machine learning model for predicting concrete slump based on the various Machine Learning algorithms and techniques have been applied to various datasets to automate the analysis of large and complex data. Many researchers, in recent times, have been using several machine learning techniques to help the core industry . This project presents various models based on such algorithms and techniques and analyzes their performance.

Introduction

Concrete workability is defined as the effort required to manipulate a freshly mixed quantity of concrete with minimum loss of homogeneity . This property of concrete is generally known to affect the consistency, flowability, pumpability, compactibility, and harshness of a concrete mix. Thus, concrete workability is a very crucial factor that must be considered in order to produce high quality concrete .

The slump test is the most common method for assessing the flow properties of fresh concrete; the slump provides a measure of workability . Using this test, the slump can be derived by measuring the drop from the top of the slumped fresh concrete. In the task of concrete mixture design, the prediction of concrete flowability is critical for on-site construction. As the complexity of concrete construction escalates, there is an increasing pressure on material engineers to achieve high workability as well as to maintain the necessary mechanical properties to meet design specifications.

Concrete has been increasingly utilized in high-rise building and infrastructure development projects and special ingredients are often employed to make the material satisfy a specific set of performance requirements . Superplasticizers are often included to enhance the concrete workability . This situation makes the concrete mixes to be highly complex materials and modeling their properties becomes a very challenging task. There are complex and nonlinear relationships between the characteristics and the components that constitute the concrete mixes .

Objective

The aim of the project is to explore the problems in concrete compressive strength and make a suitable prototype to solve the problem.

The objective is to implement various machine learning algorithms such as logistic regression, k-nearest neighbors, support vector machine, random forest and decision tree in order to predict heart disease. Other features include performing data visualization of dataset.

Methods

In this study, the concrete compressive strength and slump values were predicted using the ML regression models, namely, the regression tree, RF, support vector machines, artificial neural network, partial least square, bagging, and FL. Datasets were randomly split into 70% for the training set and 30% for the independent test set. The training data were used to train the ML model. The independent test data were applied for the evaluation of the model's performance. The 10-fold cross-validation procedure helped in the estimation of the ML model skills.

The ML preprocessing steps were applied to the raw datasets before they could be utilized for the regression method training. The datasets were not normally distributed according to the Shapiro–Wilk normality test results. Many normalization methods have been previously developed to normalize the dataset . In this study, four different normalization methods (i.e., min-max, decimal, sigmoid, and z-score) were applied to derive the most successful normalization method for the raw dataset. Then, the K-nearest neighbor (KNN) regression method was applied to the normalized datasets. The prediction results were compared to determine the most suitable normalization method. Later, the raw datasets were normalized with the determined normalization technique.

The ML regression models were trained to predict the f_c and S values. The correlation coefficient (R), root mean squared error (RMSE), and mean absolute error (MAE) metrics were employed to compare the models' prediction performance. According to these statistical results, the most successful regression method was determined to predict the f_c and S values. Afterward, the feature selection method was used to obtain the subset with fewer features, and the prediction accuracy was examined. All regression methods and computations were performed using the *R* programming language. The prediction process is illustrated in Figure 1 in the form of a flow diagram.

Machine Learning Methods

The ML regression method estimates the output value using the input samples of the dataset. Such a procedure is also termed as the training set. The purpose of the regression method is to minimize the error between the predicted and actual outputs . Herein, seven different regression methods (i.e., DT, RF, support vector machine, partial least squares, artificial neural networks, bootstrap aggregation (bagging), and FL) were used to predict the

concrete compressive strength and slump values. Additionally, the K-nearest neighbor method was applied to determine the suitable normalization method for the dataset. These methods are briefly described below.

➢ **Regression**: Regression is a statistical concept which is used to determine the weight of relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables). Two basic types of regression are linear regression and polynomial regression. Also, there are several non-linear regression methods that are used for more complicated data analysis such as logistic regression.8

➢ **Decision Tree**: A Decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences including chance event outcomes and utility. It is one of the ways to display an algorithm. Decision trees are commonly used in operations research, specifically in decision analysis to help and identify a strategy that will most likely reach the goal. It is also a popular tool in machine learning. A Decision tree can easily be transformed to a set of rules by mapping from the root node to the leaf nodes one by one. Finally by following these rules, appropriate conclusions can be reached. It's easy to use and implement. Not needed too much data preparing. It looks like it's similar to human thinking form. Categorical data can be used.

➢ **Random Forest**: Random Forest is an ensemble learning method (also thought of as a form of nearest neighbor predictor) for classification and regression techniques. It constructs a number of Decision trees at training time and outputs the class that is the mode of the classes output by individual trees. It also tries to minimize the problems of high variance and high bias by averaging to find a natural balance between the two extremes. Both R and Python have robust packages to implement this algorithm. Random Forest is a simple, flexible machine learning algorithm which is widely used for both classification and regression. Basically it consists of many decision trees. The final results depend upon the results of these decision trees.

➢ **KNN (K Nearest Neighbors**): KNN is a machine learning algorithm which is widely used for classification. Main aims of this algorithm are finding k nearest data to the point which is going to be classified. By looking its neighbors, algorithm decides to put data into which class. KNN algorithms use a data and classify new data points based on a similarity measures

Material and Methodology

The following machine learning algorithms have been used:

➢ Logistic Regression

➢ K-nearest neighbors

➢ Random Forest

➢ Decision Tree

For prediction of the slump test, data is collect from https://archive.ics.uci.edu/ml/datasets/Concrete+Slump+Test .

Data Set Information:

This database contains 103 attributes, but all published experiments refer to using a subset of 14 of them.. The "target" field refers to the SLUMP (cm). It is continous valued features.

Attribute's Information:

➢ **Cement**:

➢ **Slag**:

➢ **Fly ash**:

➢ **Water**:

➢ **SP**:

➢ **Coarse Aggr.**:

➢ **Fine Aggr.**:

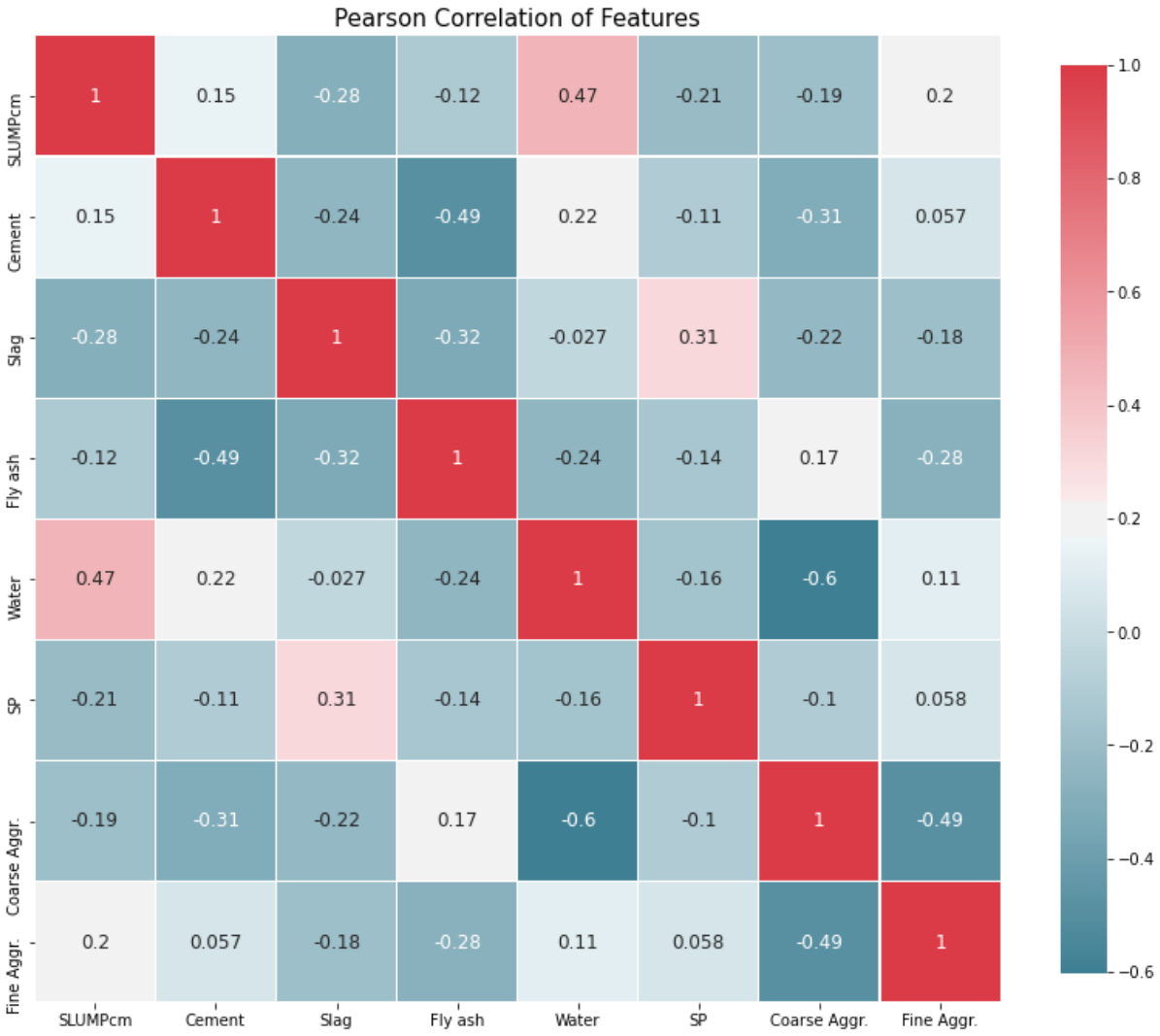
➢ **SLUMP(cm)**:

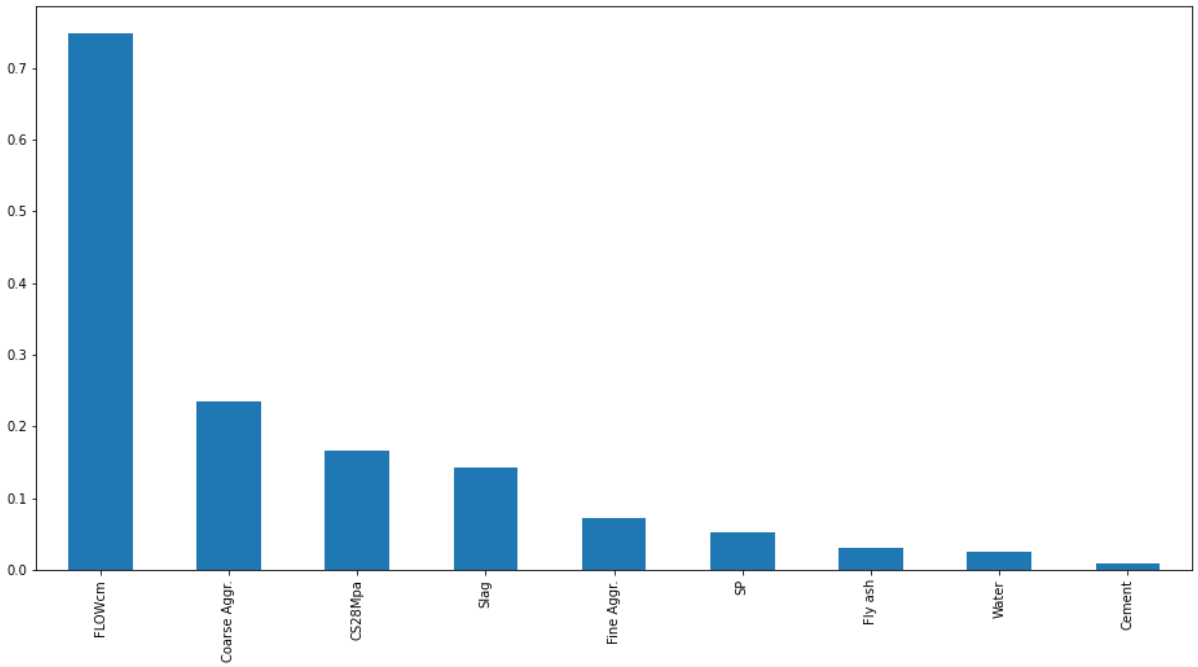
➢ **Compressive Strength (28-day)(Mpa)**:

Work Plan

The data has been analyzed by performing visualization plots and charts. Then, the machine learning algorithms like logistic regression, k-nearest neighbors,, random forest and decision tree have been implemented on the extracted data and prediction will be done by each algorithms. Then score for each algorithm has been calculated and the algorithms have been rated by their scores. On the basis of scores of each algorithm, models have been compared that how efficient and precise is the algorithm in order to make prediction.

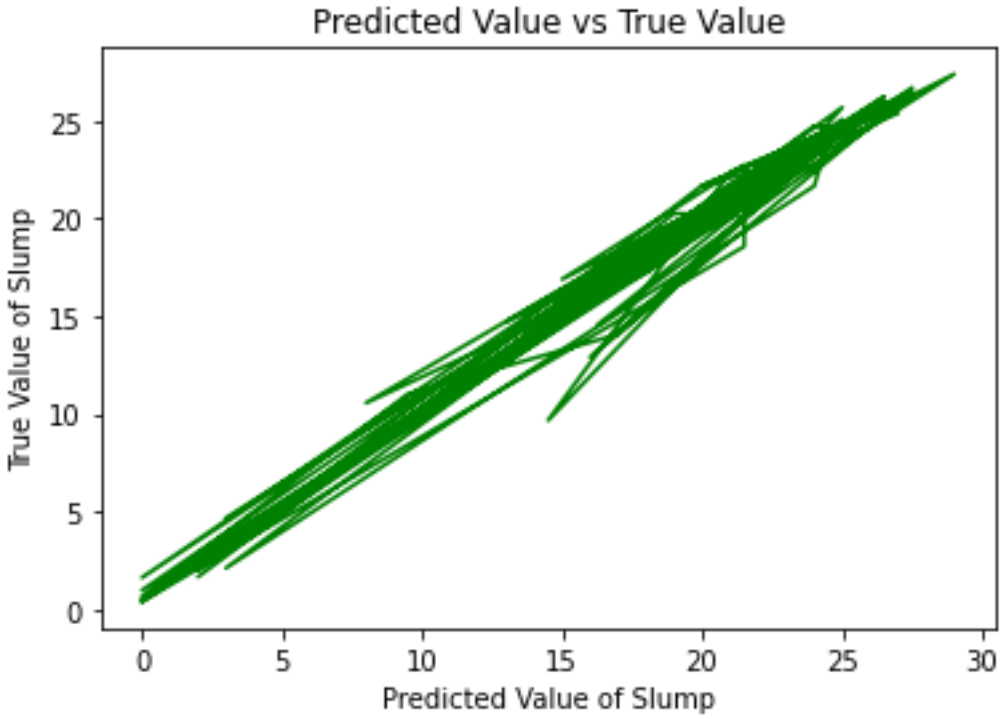
Data Visualization





Importance of feature for prediction of Slump

Cement	0.009402
Water	0.025598
Fly ash	0.030626
SP	0.051964
Fine Aggr.	0.072745
Slag	0.142414
CS28Mpa	0.167032
Coarse Aggr.	0.234495
FLOWcm	0.748318



Result

The following scores were resulted after implementing the following machine learning algorithms :-

Algorithm	**% Score**
Linear Regression	87.00699168258601
Decision Tree	95.25661779751831
K Nearest Neighbors	35.32459179508549
Random Forest	95.07270139432367

The complete project with Jupyter Notebook is uploaded on GitHub repository which link is given below: https://github.com/shivamkrsing?tab=repositories

Conclusion

After implementing several machine learning algorithms, the highest score is found by Linear Regression and Decision Tree with 0.952. The worst score is found by K Nearest Neighbors with 0.350