MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR- 474005

DEPARTMENT OF IT

MINOR PROJECT REPORT
IT III Year V Semester

PROJECT TITLE- ML Algorithm Identifier

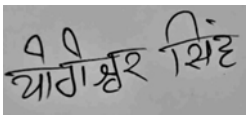
SUBMITTED TO- Prof. Vikas Sejwar
SUBMITTED BY- Khushi Bhadoria (0901IT191031) &
Sakshi Talreja (0901IT191052)

CERTIFICATE

This is to certify that Khushi Bhadoria (0901IT191031) & Sakshi Talreja (0901IT191052) minor project, " ML Algorithm Identifier " is a genuine record of a project completed under our supervision and guidance in partial fulfilment of the requirements for the award of a Bachelor of Technology in Information Technology in the Department of Information Technology, Madhav Institute of Technology and Science, Gwalior.

(Prof.Vikas Sejwar)

Mentor



(Prof. YOGESHWAR SINGH)

Mentor

CONTENT

Category- Machine Learning(ML) in python

Objective- To find out the best ML algorithm for particular dataset.

Problem Identification- Finding the most appropriate algorithm for a particular problem is also one of the challenges faced by developers. This project deals with such types of problems and also with the methods through which the performance of ML models can be improved.

Introduction- Artificial Intelligence is the most popular component of this digital world, and automation has become the basic need of the era, which seems endless. Machine Learning is one of the areas under artificial intelligence, whereas neural network is part of supervised machine learning algorithms. This project aim is to analyze some machine learning algorithms and provide the best algorithm suitable for particular data based on some popular hyperparameters. It also ascertains that ensemble learning increases accuracy and provides better results. Ensemble Learning is the intelligent technique in Machine Learning that improves the overall accuracy by combining the results from two or more machine learning models. This project also focuses on the size of the data. Results also demonstrate that the accuracy of the model doesn't always increase on increasing the size of data. Here data is everything, which means data provided by the user plays an essential role in decision making for the most appropriate machine learning algorithm to meet the specific requirement.

Advantages-

This Project will help the developers to find out the best ML algorithm for the particular dataset whether it is regression-based or classification-based.

This project also deals with the methods through which the accuracy of the ML model can be enhanced.

Software Tools-

- Google Colab

Colab notebooks are stored in Google Drive, or can be loaded from GitHub. Colab notebooks can be shared just as you would with Google Docs or Sheets.

- RAM used

740 MB

Material and Methods-

First data pre-processing is done in such a way that works for every classification and regression-based dataset.

The company on which ML algorithms are applied to stock data is Wipro. Here Linear Regression, KNN, Random Forest, and SVR are applied to each dataset. K-folds are also applied to each algorithm to check whether the accuracy of the model is increasing or decreasing.

In this project, the value of K (neighbors) in KNN is 7. The number of K-folds used is 5 for each algorithm for every dataset. In Random Forest, 100 estimators and 10 maximum depth is taken. RBF kernel is taken into consideration in SVR.

Sample Dataset

Data is obtained from yahoo finance for Wipro stock price prediction. SMA open, SMA close, RS, RSI are calculated manually on an excel sheet from the given data. Dataset is divided into 70-30 ratio for training and testing purposes respectively.

Result Obtained

Bar graphs are plotted through mat plot lib to compare the actual and predicted values that are obtained from the results of a particular ML algorithm. Here, all the parameters have been used to predict the value of RSI.

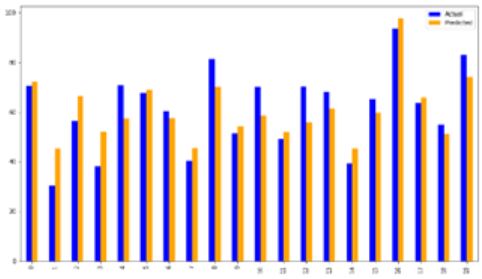


Fig 1a. Linear Regression representation

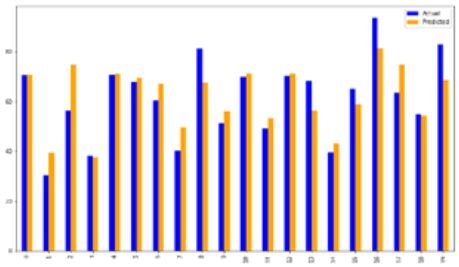


Fig 1b. KNN representation

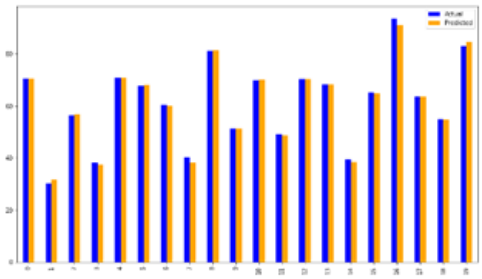


Fig 1c. Random Forest representation

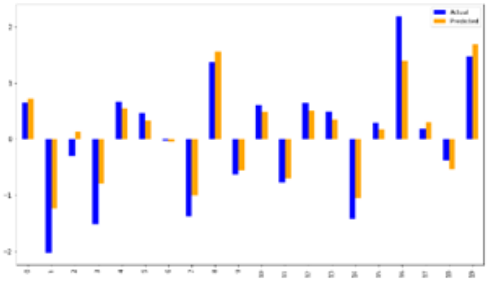


Fig 1d. SVR representation

Fig1. Representation of comparison between actual and predicted values of algorithms on Wipro dataset.

These representations show the comparison between actual and predicted values of each algorithm. The X-axis shows the number of rows from datasets that were taken for the representation of the graph. Y-axis shows the height of bars in a graph.

Here Fig 2, proved that increasing the number of records of the dataset not always increases the accuracy of the model, but sometimes it decreases. Sample data of 250 records are taken and calculated accuracy by calculating the r2 scores of the series of first 10, 20, 30, 40…. records in the dataset.

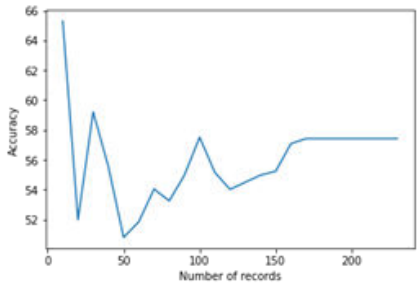


Fig 2a. Linear Regression

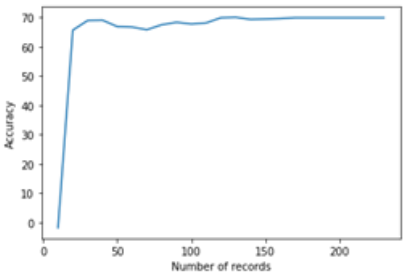
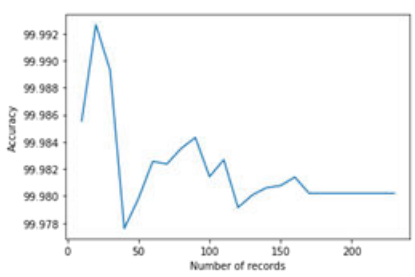


Fig 2b. KNN



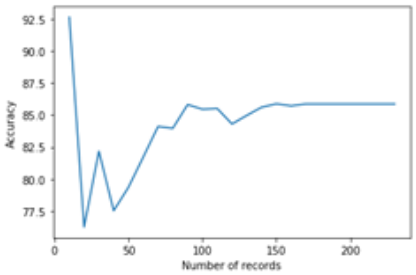


Fig 2c. Random Forest **Fig 2d.** SVR

Fig 2. Representation of Accuracy Graphs of algorithms according to sets of Wipro dataset.

Fig 3 shows a graphical representation of comparison between different ML algorithms based on training and testing accuracy, training and testing RMSE, and K-Fold accuracy. The X-axis shows the names of algorithms that were applied to the dataset. Y-axis shows a scale from 0 to 100 to represent the accuracy and RMSE of the models. Here blue and orange line represents accuracy of models (that are shown on the x-axis) on training and testing datasets respectively. The green and red lines represent the RMSE of model on training and testing datasets respectively. The purple line shows accuracy of model after applying 5 k-folds to the dataset.

The best fitted algorithm is chosen based on this representation. The best fit model for any dataset is the algorithm that has the least root mean squared error (RMSE) and greatest accuracy.

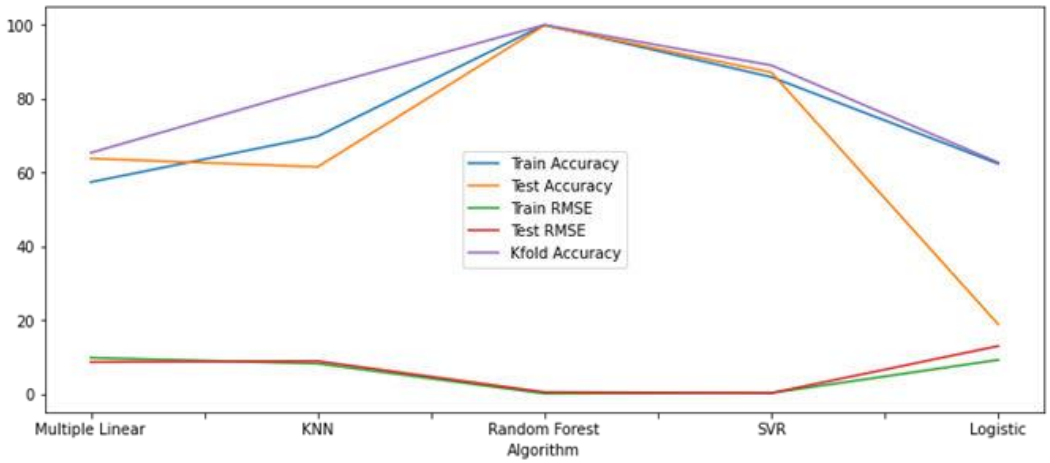


Fig 3. Representation of comparison between ML algorithms on Wipro dataset

Table1. Table of comparison between ML algorithms on Wipro dataset

Algorithm	Train Accuracy	Test Accuracy	Train RMSE	Test RMSE	K-Fold Accuracy
Linear Regression	91.917956	92.035244	4.697926	4.341280	92.490195
KNN	80.330996	71.126960	7.328865	8.265666	88.475987
Random Forest	99.982572	99.935008	0.218156	0.392158	99.982044
SVR	96.156071	93.270596	0.199908	0.246216	96.250439

This table shows all the values on which basis the best algorithm is decided. According to this, random forest is obtained as the best algorithm for the dataset.

Conclusion-

In this project, the research is conducted on the accuracy of Machine Learning models. Analysis of our results suggests that Random Forest is the best-fitted supervised machine learning algorithm for predictions in the stock market among all the algorithms that were applied to the dataset. Analysis of algorithms is done by comparing the accuracy and RMSE of the models. We also discussed the methods through which the accuracy of the ML model can be enhanced. Applying K-folds to the dataset also improved the performance of the model. A complete machine learning model depends on the quality and quantity of data. Sometimes more data can decrease the accuracy of the model. It also depends upon the model parameters and the tuning of hyperparameters. The number of records in the dataset is also an important factor that affects the performance of the model. This project proved this through graphical representations between accuracy and the number of records. Finally, we have summarized the solution to real-world problems faced by developers.