Tech Saksham

Capstone Project Report

STOCK MARKET FORECAST

Government College of Engineering

Tirunelveli-627007

NM ID	NAME
4D38130E6798A4DA0D948CCCD530E	NANDHINI.N

Trainer Name:

P.RAJA

Master Trainer:

Dr.K.SUMANGALA

ACKNOWLEDGEMENT

I would like to express our sincere gratitude to all those who have contributed to the completion of this project.

First and foremost, I extend my heartfelt thanks to our project supervisor, P.Raja for their invaluable guidance, support, and encouragement throughout the duration of this project. Their expertise and constructive feedback have been instrumental in shaping my work and pushing me towards excellence.

I am also deeply thankful to Dr.K.Sumangala for their assistance and insights, which have enriched my understanding of the subject matter and enhanced the quality of my project.

Furthermore, I would like to acknowledge the contributions of my classmates and friends who provided me with encouragement, advice, and assistance whenever needed.

.....

ABSTRACT

The stock market is a very important activity in the finance business. Its demand is consistently growing. Stock market prediction is the process of determining the future value of company stock or other financial instruments traded on a financial exchange. For some decades Artificial Neural Network (ANN), which is one intelligent data mining technique has been used for Stock Price Prediction. It has been trusted as the most accurate consideration. This paper surveys different machine learning models for stock price prediction. We have trained the available stock data of American Airlines for this project. The programming language that we have used in this paper is Python. The Machine Learning (ML) models used in this project are Decision Tree (DT), Support Vector Regression (SVR), Random Forest (RF), and ANN. The data here is split into 70% for training and 30% for testing. Thedataset contains stock data for the last 5 years. From the simulation results, it is shown that Random Forest performs better as compared to others. Thus, it can be used in the real-time implementation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

CHAPTER 1

INTRODUCTION

1. Problem Statement:

The problem statement highlights the challenges faced by stockbrokers in making trading decisions based on traditional methods such as experience, price trends, and fundamental analysis. These methods are subjective and short-sighted, potentially leading to significant losses for investors. As a result, there's a need for a more reliable tool that can provide guidance on proper trading methods and their consequences.

2. Problem Definition:

Machine learning methods offer a solution by leveraging technical and fundamental analysis to predict future stock market movements. By analyzing historical stock prices and other relevant data, machine learning algorithms can generate insights and predictions to assist traders in making informed decisions.

3. Expected Outcomes:

The expected outcomes of addressing this problem statement include:

Improved decision-making: Implementing machine learning models can lead to more informed and data-driven trading decisions, reducing the reliance on subjective methods and potentially minimizing losses for investors.

Enhanced prediction accuracy: By leveraging historical data and advanced algorithms, machine learning techniques can improve the accuracy of stock market predictions, allowing traders to anticipate market movements more effectively.

Increased investor confidence: A reliable tool for analyzing stock prices and predicting market trends can instill confidence in investors, encouraging greater participation in trading activities.

Reduced risk: By providing insights into potential market fluctuations and identifying trends, machine learning models can help mitigate risks associated with trading, leading to more stable and profitable investment strategies.

Development of robust trading tools: Addressing this problem statement can pave the way for the development of sophisticated trading tools and platforms that incorporate machine learning capabilities, catering to the evolving needs of traders and investors.

1.4. Organization of the Report

The remaining report is organized as follows:

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

CHAPTER 2

LITERATURE SURVEY

CHAPTER 2

LITERATURE SURVEY

1. Paper-1

Stock market data of American airlines from 2-08-2013 to 2-07-2018 has been used a dataset in this project.

1. Brief Introduction of Paper:

Since the introduction of the Stock Market so many predictors are constantly trying to predict stock values using different Machine Learning algorithms such as Support Vector Regressor (SVR), Linear Regression (LR), Support Vector Machine (SVM), Neural Networks Genetic Algorithms, and many more [5] on stocks of various companies. There is a diversity in many papers based on different parameters. Many different ML algorithms are used by different authors based on different parameters. Some authors believe that Neural Networks have given better performance as compared to other approaches [5]. Like, in paper [12] Hiransha M and GopalKrishnan E. A has trained four models Multi-Layer Perceptron (MLP), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM) and it was observed that CNN has performed better than the other three networks. On the other hand, many authors believe that Support Vector Regression which is known to solve regression and prediction problems gives better performance as seen in paper [13] by Haiqin Yang, Laiwan Chan, and Irwin King. In paper [5] Paul d. Yoo has trained 3 models Support Vector Machine, CaseBased Reasoning classifier (CBR), and Neural Networks (NN) from which Neural has given the most appropriate prediction. Sumeet et al [18] has done an approach where they have combined two distinct fields for stock exchange analysis. It merges price prediction based on real time data as well as historical data with news analysis. In this paper LSTM(Long Short-Term Memory) is used for prediction. The datasets are collected from large sets of business news in which relevant and live data information is present. Then the results of both analyses are

combined to form a response which helps visualize recommendation for future increases. So, in many papers, it has been seen that neural networks give the expected prediction value.

2.1.2. Techniques used in Paper:

In this project, prediction is carried out by using these ML algorithms. These are Decision Tree, Support Vector Regression, Random Forest, and Artificial Neural Network.

Decision Tree Methodology:

Decision Tree Methodology It is a supervised ML, which is used for both regressions as well as classification. That is how it is also called CART Classification and Regression Trees. In this algorithm, two nodes are present namely Decision Node which is for making the decisions and can be divided into multiple branches and Leaf Node which gives the output of decisions and this node can't be further divided into many nodes. The following is the formula for Leaf Node: $Information\ Gain = Class\ Entropy - Entropy\ Attribute$ (1) Branches-Here decision rules are set by which nodes can be divided further. For Prediction, it starts from the root node, compares values of the real attribute with the root attribute, and based on that comparison it follows the branch and jumps to the next node. This process continues until it reaches the leaf node of the tree.

Entropy-It is a metric that helps in measuring error in a given attribute. The formula to find entropy is: -

Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no) (2) Here, (S) implies the Total number of samples. P (yes) refers to the Probability of S and P (no) means the Probability of no.

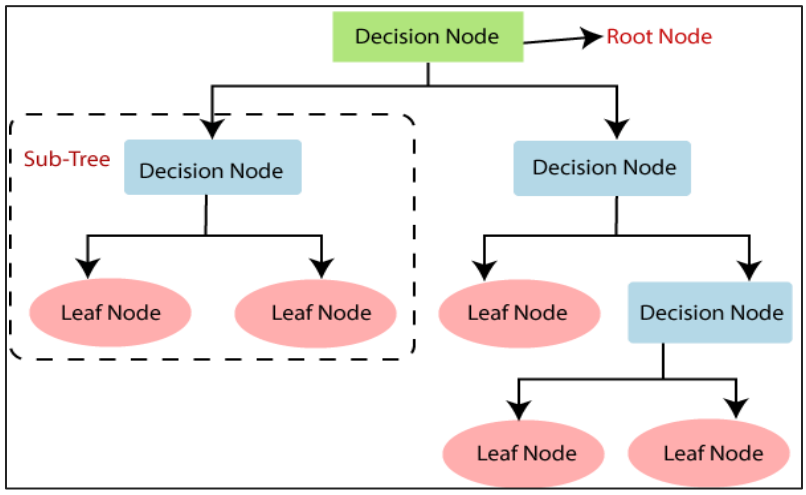


Figure 1: Decision Tree Classifier Process

Support Vector Regression Methodology:

It is a Supervised Machine learning algorithm used for regression analysis. It finds the function that helps us approximate mapping based on the training sample from an input domain to real numbers. The Terminologies contained in this are Hyperplane -this is the line that is used to predict the continuous output. Kernel helps to find hyperplanes in higher dimensional space without increasing the computational cost of it and the decision boundary is a simplification line that differentiates positive examples and negative examples.

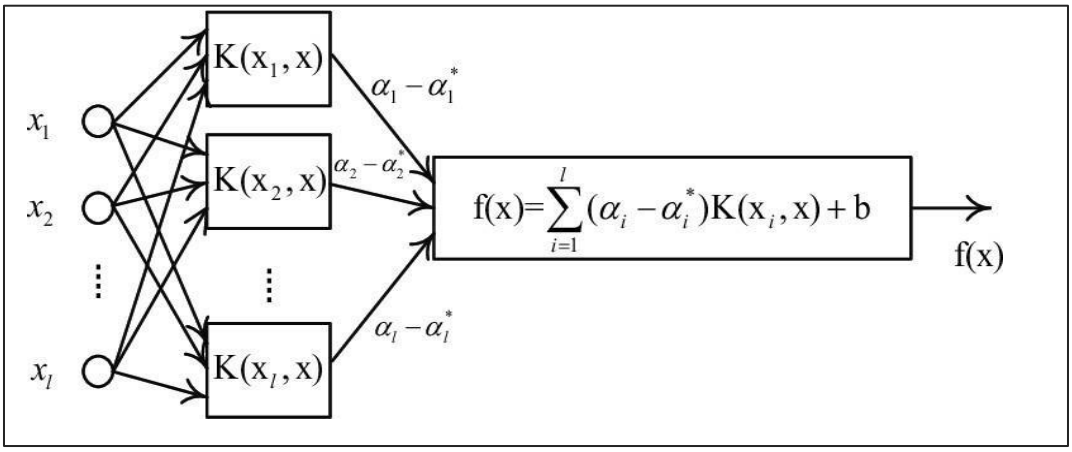


Figure 2. Support Vector Regression

Random Forest Methodology:

Random forest is a supervised Machine Learning algorithm that is used for Regression analysis. This overcame the problem of overfitting as seen in the decision Tree [12]. It is an ensemble learning method. The steps for prediction are first a random k data point is picked from the training set then accordingly the decision tree is built. Then choose the number of trees we want to build and again follow the previous steps. From every new data point, make N tree Trees predict the value of Y for data points and assign new data points across all of y predicted Y values.

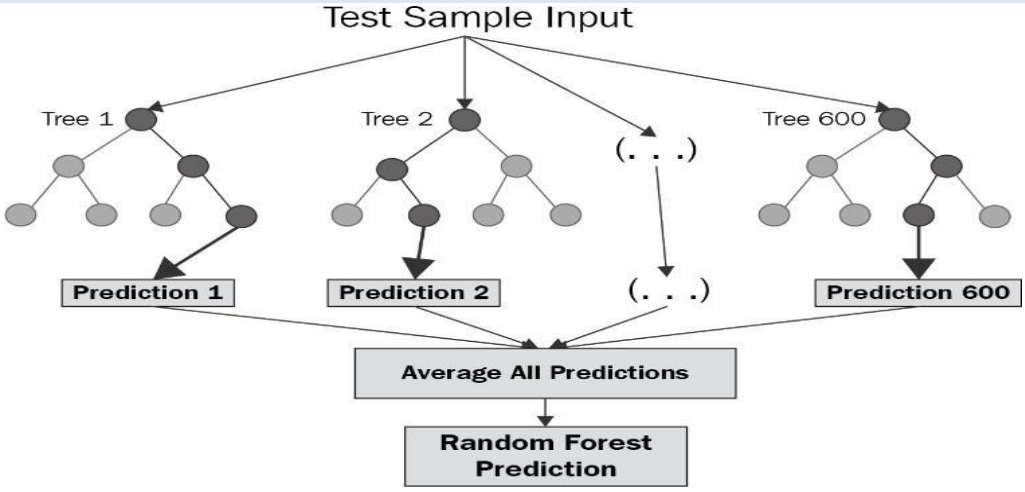


Figure 3. Random Forest Procedure

Artificial Neural Network Methodology:

An artificial Neural network is an interconnection of nodes that is like the biological neuron in our body but not similar. For the last few decades, ANN has been used for Stock Price Prediction [12]. It contains three layers, first is the Input Layer – this layer takes different inputs variable from the user then, the hidden layer-This layer is present between 166 the input layer which identifies all hidden features and patterns and the last layer is the Output layer- This layer provides the final output. ANN takes different inputs and multiplies them with the specified weights for each with an activation function for the activation of neurons. The formula of the transfer function is: $\sum Wi * Xi + b\ n\ i{=}1$

Here, b is the threshold value. Xi is input and value and Wi is the weight.

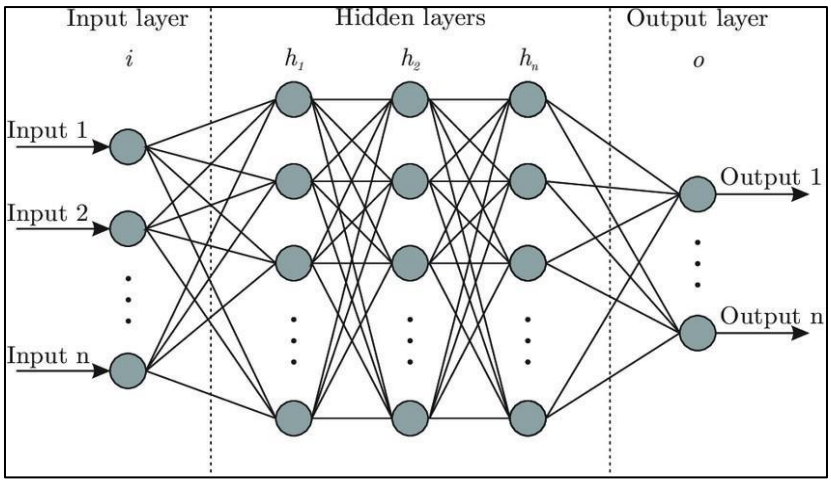


Figure 4. Artificial Neural Network Procedure

CHAPTER 3

PROPOSED METHODOLOGY

CHAPTER 3

PROPOSED METHODOLOGY

3.1 System Design

Stock market data of American airlines from 2-08-2013 to 2-07-2018 has been used as a dataset in this project. This dataset has 1258 rows and 7 columns. Each row represents the information for a single day. For columns, the following are the feature description.

1. Data Preprocessing: It includes searching for essential missing or null values and replacing them with mean values Searched for categorical value and if there is any unnecessary data then those values are dropped.

2.Data Splitting: The processed data has been divided into 70% training data and 30% testing data using the train test split method. Here 881 data is taken as training data and the rest 377 is kept for testing. The training data values are taken from the date 2013-02-08 to 2016-08-09 and the testing data are from 2016-08-10 to 2018-02-06.

Table 1: Dataset Feature Description Table

Sl. No	Feature	Description
1.	Date	It shows the date in the format: yy-mm-dd.
2.	Open	It shows the price of the stock at market opening.
3.	High	It shows the highest price reached on that day.
4.	Low	It shows the lowest price reached on that day.
5.	Close	It shows the lowest price reached on that day.
6.	Volume	It shows the number of shares traded on that day.
7.	Name	This is the name of the stock's ticker.

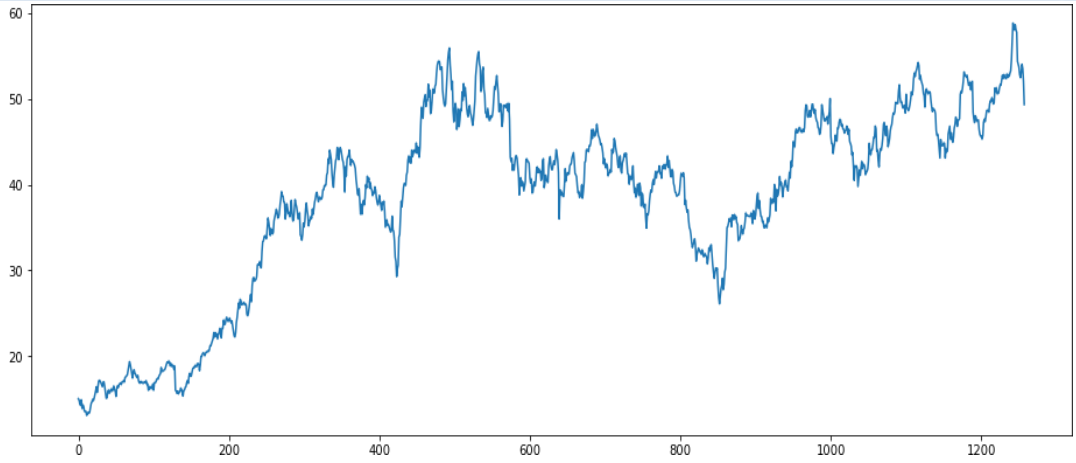


Figure 5: Opening Price Graph

3.Data Scaling: Standardization and Normalization are done on the data using Minmax Scaler and Standard Scaler to limit the ranges of variables to make them comparable on common grounds using ML methods.

4.Feature Selection: The selection of features is a very important task to predict future values. If we consider the worst features then the prediction can go wrong. In this paper, the attribute or feature used for feature extraction is the opening price or the 'open' column of American Airlines stocks. A data structure has been created with 7 timesteps and 1 output.

5.Prediction: We have adapted Machine Learning Approaches to find the prediction. In this case, training the model is very necessary. Random Forest, Decision Tree, and Support Vector Regression models have been used to do the prediction work.

6.Error Calculation: There are 4 types of error calculations present for evaluation. In this paper, we have used the MAPE method to find the error. Performance evaluation is done using MAPE values of all the models.

3.2 Modules Used

Data Collection and Preprocessing:

Pandas: For data manipulation, loading datasets, handling missing values, and preprocessing tasks.

NumPy: For numerical operations and working with arrays, essential for preprocessing data.

Feature Selection and Engineering:

Scikit-learn: Provides various feature selection techniques and tools for dimensionality reduction.

Feature-engine: A feature engineering library for handling categorical variables, missing data, and feature scaling.

Model Selection and Training:

Scikit-learn: Offers a wide range of machine learning models like linear regression, decision trees, random forests, gradient boosting, and support vector machines.

TensorFlow or PyTorch: For implementing deep learning models such as recurrent neural networks (RNNs) or long short-term memory networks (LSTMs) for time series analysis.

Model Evaluation:

Scikit-learn: Provides functions for cross-validation, hyperparameter tuning, and evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R-squared.

Deployment:

Flask or Django: Web frameworks for building RESTful APIs or web applications to deploy the trained models.

FastAPI: A modern, fast (high-performance) web framework for building APIs with Python based on standard Python type hints.

Monitoring and Maintenance:

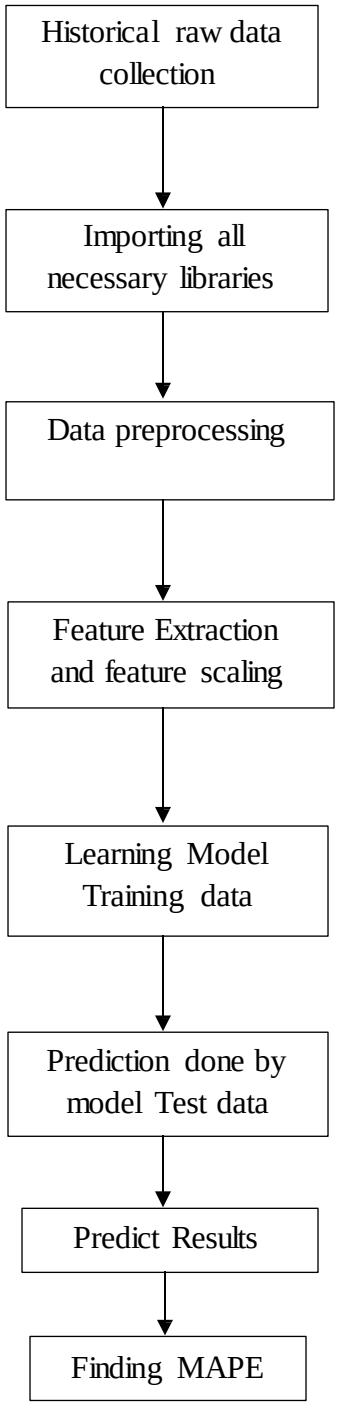
Custom scripts or monitoring tools to track model performance over time and trigger retraining or updates when necessary.

Docker: For containerizing the application for easier deployment and scalability.

Kubernetes: For container orchestration and management in a production environment.

3.3 Data Flow Diagram

A Data Flow Diagram (DFD) is a graphical representation of the "flow" of data through an information system, modeling its process aspects. A DFD is often used as a preliminary step to create an overview of the system, which can later be elaborated. DFDs can also be used for the visualization of data processing (structured design).



As shown in the figure above all the historical data were collected first and followed by the importation of all necessary libraries such as NumPy, Pandas, matplotlib, Seaborn, mean squared error, etc. In the next step, various data processing methods have been performed such as drop, isnull, etc. Then feature extraction and feature scaling techniques have been implemented using Min Max Scaler and sc fit transform. In the next step we have trained the data and learned the model required. In the next step various machine learning model which we have learned have been applied such as Decision Tree, Support Vector, Artificial Neural Networks, and Random Forest. Then we have got the prediction results. Out of all the 4 algorithms, Random Forest has the lowest MAPE value i.e.- 0.36.

3.4 Advantages

Objective Decision Making: Machine learning algorithms can provide unbiased and objective analysis of stock prices, removing personal biases from decision-making processes.

Comprehensive Analysis: Machine learning models can analyze vast amounts of data including historical prices, company financials, market sentiment, and more to provide a comprehensive view of the market.

Long-term Perspective: By considering a wide range of factors, machine learning models can help investors take a more long-term perspective, rather than being influenced by short-term trends or personal experiences.

Risk Management: Machine learning models can assess the risk associated with different investment options, helping investors make informed decisions and mitigate potential losses.

Increased Participation: Providing a tool that guides investors with proper trading methods and consequences can increase investor confidence and encourage greater participation in trading.

Adaptability: Machine learning models can adapt to changing market conditions and continuously improve their predictions over time, providing valuable insights for investors.

Efficiency: By automating the analysis process, machine learning can save time and resources for both investors and stockbrokers, allowing them to focus on other aspects of their work.

3.4 Requirement Specification

1. Hardware Requirements:

High-performance CPU: A multicore processor with sufficient processing power to handle the computational load of data preprocessing, model training, and prediction.

Sufficient RAM: Adequate memory to store and manipulate large datasets efficiently during data processing and model training.

Storage: Sufficient storage capacity to store historical stock market data, pre-processed datasets, and trained machine learning models.

Graphics Processing Unit (GPU) (Optional): GPU acceleration can significantly speed up certain machine learning tasks such as model training, especially for deep learning algorithms.

Network Connectivity: Stable internet connection to access real-time data sources, APIs, and cloud-based services.

2. Software Requirements:

Operating System: The system should be compatible with popular operating systems like Windows, Linux, or macOS.

Programming Languages: Proficiency in programming languages such as Python, R, or Java for implementing machine learning algorithms, data preprocessing, and software development.

Machine Learning Libraries: Familiarity with machine learning libraries such as TensorFlow, Scikit-learn, PyTorch, or Keras for building and training prediction models.

Data Processing Tools: Utilize tools like Pandas, NumPy, or Apache Spark for data cleaning, manipulation, and preprocessing tasks.

Visualization Tools: Integration with visualization libraries like Matplotlib, Seaborn, or Plotly for generating interactive visualizations of historical stock data and prediction results.

Development Environment: Use integrated development environments (IDEs) such as Jupyter Notebook, PyCharm, or Visual Studio Code for coding, debugging, and testing machine learning models.

Database Management System (DBMS): Integration with DBMS like MySQL, PostgreSQL, or MongoDB for storing and retrieving historical stock market data efficiently.

Web Development Frameworks (Optional): If developing a web-based application, familiarity with frameworks like Django, Flask, or React.js for building user interfaces and backend services.

Deployment Platforms: Understanding of cloud platforms such as AWS, Google Cloud, or Microsoft Azure for deploying and hosting the stock market prediction tool.

Version Control: Proficiency in version control systems like Git for collaborative development, code management, and version tracking

CHAPTER 4

IMPLEMENTATION and RESULT

CHAPTER 4

IMPLEMENTATION and RESULT

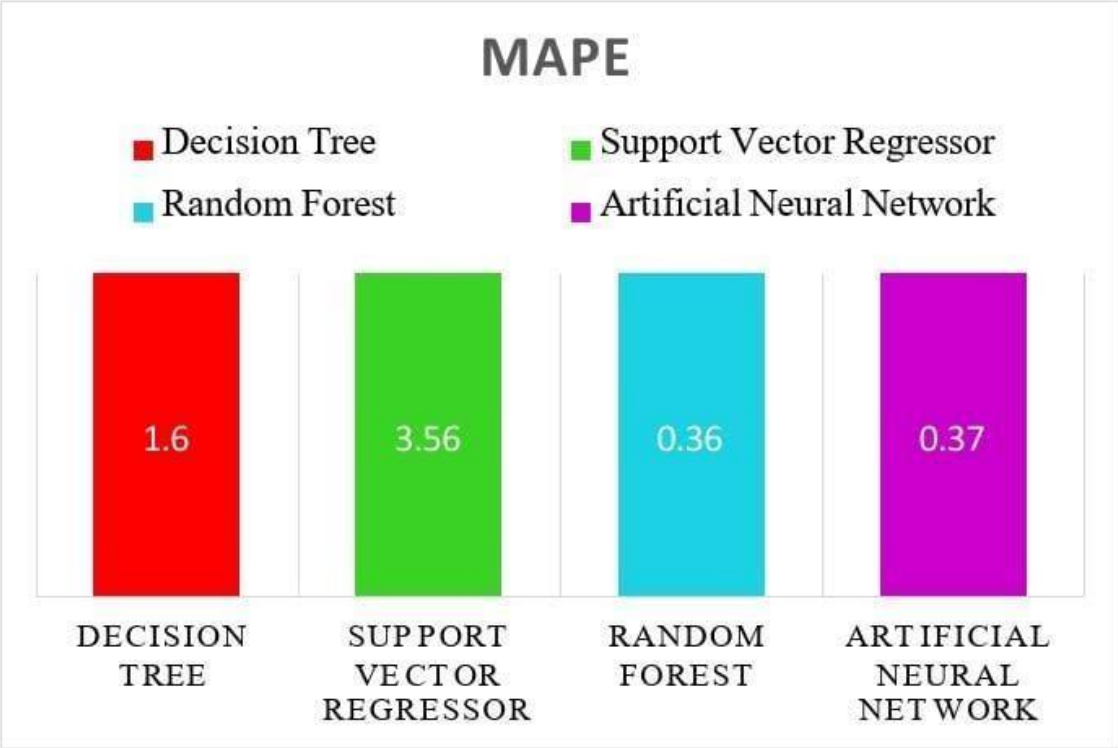
Results and Discussion

The main objective of this project is to examine several different prediction techniques to predict future stock prices based on past returns. And here it is visible that Random Forest is the best algorithm for this research giving a MAPE value of 0.36. This algorithm shall be used to predict opening prices shortly. The following is the table to show the MAPE values using Machine Learning Algorithms.

Table 2

MAPE

S.No	Model	MAPE
01	Decision Tree	1.60
02	Support Vector Regression	3.56
03	Random Forest	0.36
04	Artificial Neural Network	0.37


MAPE
Decision Tree
Support Vector Regressor
Random Forest
Artificial Neural Network
1.6
3.56
0.36
0.37
DECISION TREE
SUPPORT VECTOR REGRESSOR
RANDOM FOREST
ARTIFICIAL NEURAL NETWORK

CHAPTER 5

CONCLUSION

CHAPTER 5

CONCLUSION

1. ADVANTAGES:

1. Objective Decision Making
2. Comprehensive analysis
3. Long term perspective
4. Risk Management
5. Increased Participation
6. Adaptability
7. Efficiency

2. SCOPE:

The project was majorly aimed at creating an efficient tool that will help stockbrokers and investors properly invest in the stock market. Five years American Airlines stocks have been preprocessed and four machine learning algorithms have been used – Random Forest, Support Vector Regressor, Decision Tree, and Artificial Neural Network on this project. Based on calculations, estimations, and observations, we conclude that Random Forest hasthe lowest Mean Absolute Percentage Error (MAPE) value of 0.36 followed by Artificial Neural Networks with the value of 0.37, then Decision Tree showing MAPE value of 1.6 and the highest in SVR showing a value of 3.5. Artificial Neural Network has been used inthis project, giving a MAPE value of 0.37 which is the second least MAPE value provided.So, in the future, it is intended to work on advanced ANN evolutionary techniques like Genetic Algorithm to decrease the MAPE values for better implementations.

GITHUB LINK

DEMO VIDEO LINK:

REFERENCES

1. Bhattacharjee, Indronil, and Pryonti Bhattacharja. "Stock Price Prediction: A Comparative Study between Traditional Statistical Approach and Machine Learning Approach." 2019 4th International Conference on Electrical Information and Communication Technology (EICT). IEEE, 2019.
2. Mehta, Yash, Atharva Malhar, and Radha Shankarmani. "Stock Price Prediction using MachineLearning and Sentiment Analysis." 2021 2nd International Conference for Emerging Technology (INCET). IEEE, 2021.
3. Sharma, Ashish, Dinesh Bhuriya, and Upendra Singh. "Survey of stock market prediction using machine learning approach." 2017 international conference of electronics, communication and aerospace technology (ICECA). Vol. 2. IEEE, 2017.
4. Hegazy, Osman, Omar S. Soliman, and Mustafa Abdul Salam. "A machine learning model for stock market prediction." arXiv preprint arXiv:1402.7351 (2014).
5. Yoo, Paul D., Maria H. Kim, and Tony Jan. "Machine learning techniques and use of event information for stock market prediction: A survey and evaluation." International Conference onComputational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA- IAWTIC'06). Vol. 2. IEEE, 2005.
6. S. Chakravarty, B. K. Paikaray, R. Mishra and S. Dash, "Hyperspectral Image Classification using Spectral Angle Mapper," 2021 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE), 2021, pp. 87-90, doi: 10.1109/WIECON- ECE54711.2021.9829585.
7. Wanjawa, Barack Wamkaya, and Lawrence Muchemi. "ANN model to predict stock prices at stock exchange markets." arXiv preprint arXiv:1502.06434 (2014).
8. Reddy, V. Kranthi Sai. "Stock market prediction using machine learning." International Research Journal of Engineering and Technology (IRJET) 5.10 (2018): 1033-1035.
9. Ravikumar, Srinath, and Prasad Saraf. "Prediction of Stock Prices using Machine Learning (Regression, Classification) Algorithms." 2020 International Conference for Emerging Technology (INCET). IEEE, 2020.
10. Pathak, Ashish, and Nisha P. Shetty. "Indian stock market prediction using machine learning and sentiment analysis." Computational Intelligence in Data Mining. Springer, Singapore, 2019.595-603.
11. Deepak, Raut Sushrut, Shinde Isha Uday, and D. Malathi. "Machine learning approach in stock market prediction." International Journal of Pure and Applied Mathematics 115.8 (2017): 71- 77.
12. Hiransha, M., et al. "NSE stock market prediction using deep-learning models." Procedia computer science 132 (2018): 1351-1362.
13. Yang, Haiqin, Laiwan Chan, and Irwin King. "Support vector machine regression for volatile stock market prediction." International Conference on Intelligent Data Engineering and Automated Learning. Springer, Berlin, Heidelberg, 2002.
14. Kohli, Pahul Preet Singh, et al. "Stock prediction using machine learning algorithms." Applications of Artificial Intelligence Techniques in Engineering. Springer, Singapore, 2019. 405-414.
15. Moedjahedy, Jimmy H., et al. "Stock Price Forecasting on Telecommunication Sector Companies in Indonesia Stock Exchange Using Machine Learning Algorithms." 2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS). IEEE, 2020.

16. Mohanty, Sachi Nandan, et al., eds. Recommender System with Machine Learning and Artificial Intelligence: Practical Tools and Applications in Medical, Agricultural, and Other Industries. John Wiley & Sons, 2020.
17. Jain, Sarika, et al. "Human Disease Diagnosis Using Machine Learning." Intelligent Data Communication Technologies and Internet of Things. Springer, Singapore, 2021. 689-696.
18. Sarode, Sumeet, et al. "Stock price prediction using machine learning techniques." 2019 International Conference on Intelligent Sustainable Systems (ICISS). IEEE, 2019.