TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

In one form or another, we have all heard the word stock. Stock is specifically associated with friends and businesses that have been commercialized and are settling into the marketization environment. The second word for stock is share, which is frequently used in daily speech [4]. People even refer to it as an investment strategy since they believe it to be a long-term investment that will guarantee and offer a comfortable retirement income.

A small portion of a corporation can be purchased by acquiring its stock [8]. People make the same investments to gain a long-term advantage that they believe has less worth right now but has the potential to increase over time. It is an investment that addresses long-term goals with reasonable aims over the long term [1][6]. It is not the same, but the value of the share you buy in now must provide you with the best income possible future.

The resources and the factors used to push the market off or on the set are unpredictable, just like the market itself [2]. It has never existed on the same scale, and its pattern is still unpredictable now. Although certain methods for approximating values and making predictions have been developed, all of the available resources cannot be relied upon and remain unpredictable in nature.

The greatest approach to find reliability is to be aware of the market environment and conduct research into it, which is why many agents have made a career out of it and are very successful [11][7]. They provide predictions and offer advice, but their fees are greater and the stock evaluation is never less than uniform.

Even within a single day, the market might experience numerous highs and lows depending on the resources, timing, and actions of external and internal agents [3][6]. A fascinating starting point is stock.

1.2 PROJECT PURPOSE

A software prediction technique called stock market prediction reveals the risk that is involved in stock market investment [12]. It makes predictions about stock prices and exchange rates while taking into account users' prior knowledge and statistical analysis.

Data is regarded as the digital fuel that provides opportunities for higher yearn and future terms [16]. The adage "knowledge is power" also applies to stocks [1]. The stock market is erratic and constantly shifting [9]. The same's ascent and decline are uneven and difficult to categories. Dependencies on similar arrangements involving adaptable resources and the people driving them.

The opening stock market for the following day is determined by investments made during a fiscal day. It is completely integrated with the level of finances and revenue creation and has its dependencies [5]. The stock is massive and quite active. The project's major goal is to predict turning curves, introduce predictability, go through processes, and use algorithms to arrive at a useful resource source.

There is a pattern to everything. The pattern is the method of derivation, and the stock is no exception. In daily life, stock moves in predictable patterns [6][10]. Increases in some resources can raise prices for some while lowering them for others, The source and the result are determined based on the polarity of the flow, which might be positive, neutral, or negative. The given polarity's correlation is identified, and a reliable source and source effectiveness are established.

This project aids in bridging the resources and empowering the populace to comprehend the generation and the vulnerabilities that must be seen and forecasted as well as to know and trade the most out of stock [2][11]. The improvement of the same is accomplished with the resource graph, which forces a user or client to analyses the same and take the wants and significant information before dealing and take into consideration those things for the yield that the person is prepared to invest on [1]. The forecast for stocks is made for the coming week using the data sources that are now available [12]. The report's main goal is to highlight how difficult it is to foresee things.

1.3 Literature Survey

The literature review is one of the crucial elements in keeping consistency. These are the essential actions that must be taken during the development process [9][15]. The software development requires legitimate materials that are readily available. This section aids in learning about the information that has been developed and identifying its application and execution in the present [1]. The economy and the quality of the product are crucial to development [13]. The support and resource flow must be tracked and calculated once the invention has completed the building phase [14]. This is also referred to as the research phase because all of the research is done during this time to maintain the flow.

1.3.1 MACHINE LEARNING

Machine learning is one of the best words used today. Machine learning has become a crucial component of modern technology, whether it is used at work or in other settings. Despite the fact that it is evolving and developing quickly, development and implementation of the same are still ongoing [2]. Automation, which was once just an idea, is now a reality thanks to machine learning, which itself brought about a random change in today's worlds.

Today, it is a phrase with potential. One of the actions that the entire company is interested in [15]. It is a guiding pillar for the future that will guide humanity toward a better evolutionary future in which customization and labor work can be cut in half and the safety of survival can be delayed in order to stand tall for the better exploitation of the human mind [12]. Considering that, it has proven dangerous for many more, regardless of their area of interest [3]. Since machines are thought to be the most effective and errors are kept to a minimum, the degree of work flow can be hazardous, and future development of the same may result in a thousand lying idle in homes, having a greater influence on unemployment and livelihood. This also poses a threat to society in other ways.

The main factor that powers machine learning in this graphic is statistics [13][8]. It deals with computing statistics from a broad perspective and processing the results to provide an output that is driven by data, making it more logical and resource-friendly [5]. In addition, it optimizes resources, and its dependability and efficiency are unbeatable by any standard.

Despite having an evolutionary origin, it has successfully adapted to computational and digitalization concepts [4]. It includes many computing topics, including Data Mining, Statistical Analysis, Resource Optimization, and Automation. The machine in this case is capable of processing the outcome just like a person would [10]. Both the initiator and the derivable of this process are possible [16]. Even unstructured or semi-structured data can be processed to produce an approximative answer, and statistical flow is primarily reasonable with data driven patterns. [6] The closest value to each equation's aligned field is discovered, and the distance between them is calculated.

The following are a list of the same's classifications:

1.3.1.1 SUPERVISED LEARNING

In supervised learning, the computer is supervised to generate the necessary input. [5] It is a mathematical model where the inputs and outputs are known beforehand and are given to the machine to obtain the anticipated output in order to calculate efficiency [2]. This is the machine's learning phase. Here, the same's feeding and derivation are measured.

The machines in this case filter the inputs and pick up knowledge from the functional unit. Compute it, store it in memory for later processing, and if a matching pattern is discovered, use it, learn from it, and plot a result using that information.

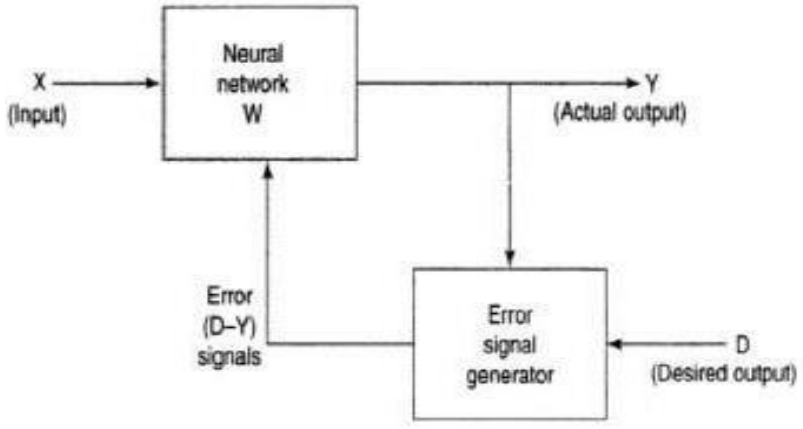

Figure1.1: Supervised learning

This procedure is dependent. The machine is entirely dependent on the user, who must supply the inputs, assess their effectiveness, and make necessary corrections as iterations proceed. The network is an ANN [11]. Vectors are taken into account during the training process.

In the illustration above, upper Both an input and an output vector are present [2][5]. The output vector originates from the input vector and provides an output flow. If an error signal is produced, iteration is required; however, if none is produced, the output field is derived, the output result is accurate, and no change is required for the same.

1.3.1.2 Unsupervised Learning

Unsupervised learning focuses on independent learning. It also goes by the name "self-learning algorithm." Just the input vector is known and provided in this case. Therefore, the variance of the outcome is related to the input variables. The input variables are aggregated and clustered in this case [7]. The key component of this technique is the cluster.

Test Data are transmitted, and with each iteration, the system learns from the data and progresses closer to the conclusion section [4]. The data collection lacks labelled items; thus, the machine has to classify and categories the remaining items. The main component of it is cluster and communalization.

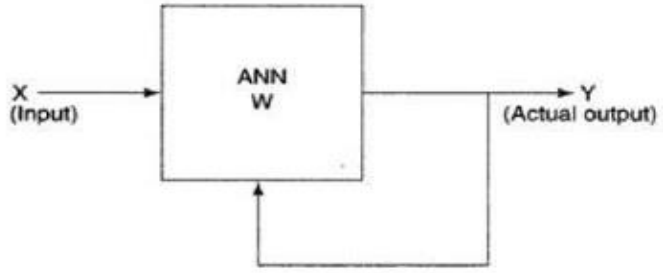

Figure1.2: Unsupervised Learning

As shown in the above graphic, in this ANN network the output required to be self-derived and matched with the cluster set to offer the result after the input is processed by the

function. The result goes through iteration if interpretation is missing. For efficient use in subsequent situations, all the data sets are generated and integrated into a cluster set.

Feedbacks in such cases only address similarities and are not reciprocated [13][10]. If similarities between the datasets are discovered, the previous functionalities are used to apply and derive the data. If not, it learns and recognizes for the other people.

1.3.1.3 Reinforcement Learning

A reinforced method is employed in this form of learning. Its topic is the growth of knowledge. This type of learning is neither supervised nor unsupervised. To inform the user of the output and its origin, they employ dynamic approaches.

These algorithm sets do not presuppose the environmental set. Even more advanced and complicated mechanisms, such as genetic algorithms, utilize them [2]. For the improvement of the establishment's efficiency, they are widely applied and in progress. These algorithms are employed in video games and vehicle resource automation.

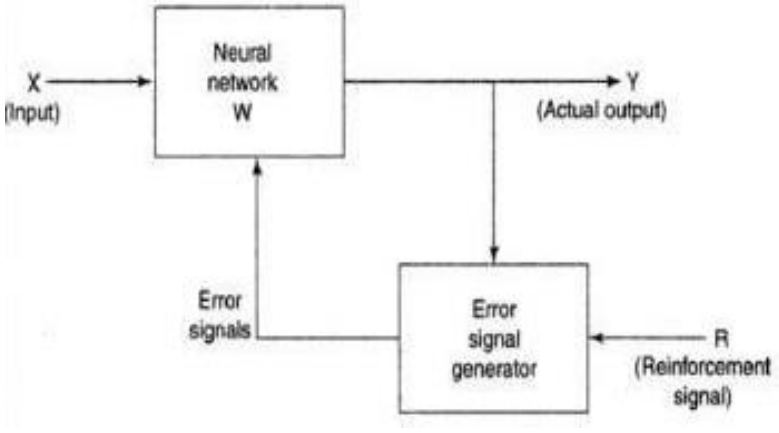

Figure1.3: Reinforcement Learning

The input vector is transmitted to an ANN model, where the functionalities of the same are kept, as shown in the picture. If the result is accurate, a reward is given to the user, moving them forward a level to complete additional tasks. If not, a signal called an error is sent out

for the same. The accuracy level is determined and communicated to the user informing them of the same.

To get the most out of it and finish the assignment to go up the success ladder, the user looks at the match and pass down percentages and tries additional iteration keys [2][6]. The machine operates in a similar manner. The machine repeats the same process and adds a reinforced signal to the mistake signal in order to learn from the error and improve the outcomes.

1.4 Problem Statement

Since then, stocks have been on an erratic curve. Its core had always been alive and indulgent. With relation to time, its popularity had increased. More people than ever before find the same things fascinating and interesting. The organization is in a similar situation. Instead of investing and getting a loan approval from the bank, the organization had built it as a superior source of revenue generating. From a business perspective, it is much more effective and less hectic.

CHAPTER 2

SYSTEM REQUIREMENT SPECIFICATIONS

2.1 REQUIREMENT SPECIFICATION:

A web-based interactive computational tool for starting with Jupiter Notebook papers is called IPython Notebook. The word "notebook" is a significant entity in and of itself to symbolize the integration with other entity sets. The primary document format from the same for execution, which complies with the short on the schema and the input and output methods, is JSON. It provides strong language set integration and a wide range of choice flexibility.

The same data integration is at its best. The integration of large data and its ability to roughly analyses chunks of values in time results in better performance and more powerful computational capabilities. The same can be used to do a variety of tasks on data, including cleansing, transforming, modelling, and visualizing.

2.1.1 HARDWARE REQUIREMENTS:

- Processor : Intel i5 or above
- RAM : Minimum 225MB or more.
- Hard Disk : Minimum 2 GB of space
- Input Device : Keyboard
- Output Device : Screens of Monitor or a Laptop

2.1.2 SOFTWARE REQUIREMENTS:

- Operating system : Windows
- IDE : Jupyter Notebook
- Data Set : .csv file
- Visualization : mat plot lib, pandas.
- Server : Web Server with HTTP process.

2.2 SYSTEM REQUIREMENTS:

The study of requirements is a crucial step in developing a product. It is crucial in figuring out whether an application is feasible. The software and hardware requirements needed to develop the product or application are specified by requirement analysis. Software requirements, hardware requirements, and functional requirements make up the majority of requirement analysis. [4,6] Requirement's analysis is a broad term used in the fields of systems engineering and software engineering to describe the processes involved in identifying the requirements that must be met for a new or modified product or project, taking into account the potentially conflicting requirements of the various stakeholders, and analyzing, documenting, validating, and managing software or system requirements.

A systems or software project's success or failure depends on the results of the requirement analysis. The requirements ought to be well-documented, usable, quantifiable, testable, traceable, tied to recognized business opportunities or needs, and sufficiently defined for system design.

The foundation for a contract between clients and vendors or contractors for how the software product should operate is established by the software requirements specification. Prior to the more detailed system design stages, software requirements specification conducts a thorough review of the requirements with the intention of minimizing subsequent redesign. [1,4] Additionally, it must offer a solid foundation for forecasting product prices, risks, and timelines. Software requirements specifications, when used properly, can aid in preventing software project failure.

2.2.1 FUNCTIONAL REQUIREMENTS:

In the technical perspective, functional requirements address the functionality of the software. It improves and describes the component flow and structural flow of the same.

The functional statement works with categorizing and learning from the same dataset's raw datasets. Then the datasets are grouped into clusters, and any degradation of the groupings is examined for efficiency. After the dataset has been cleaned, the machine learns the data and identifies the pattern set for it. It then goes through a number of iterations and produces results.

2.2.2 NON-FUNCTIONAL REQUIREMENTS:

The non-functional requirement addresses external factors of a non-functional nature. It serves as an analytical tool. Under the same conditions, the operations' performance is judged. [3] The extra effects and requirements assist stock to acquire the most recent updates and integrate in a single step where the technicians may work on and fix any bugs or draughts that may exist because stock is practical and always evolving.

Its effectiveness and hit gain ratio are adhered to non-functional constraints. The code's suitability for implementation, additional efficacy, and the search for a security console. Because of its portability and integration, the System is dependable and its performance is maintained.

CHAPTER 3

SYSTEM DESIGN AND MODELLING

3.1 Implementation

Here in this project, we have done two tasks one of this is finding the best algorithm which predicts best for the given stock dataset on the basis of this stock the future stocks will be predicted on the webpage we will create. So, basically this is research work which we have done on our project. Next part is Creating an interactive web page on which we have used the stock predicting algorithm which we have found as a best performing algorithm and this will be user friendly where the user just has to specify the stock tikker of the stock which we want to predict here we have set the prediction to next ten days.

3.2 System Architecture

The System Architecture is distributed into four phases. In first phase which is Data Splitting Phase the Dataset which is imported is preprocessed and cleaned that is all the null values if present is removed and sort the dataset. After this processes the dataset is further splitted into training and testing set. In Second Phase which is Data Training Phase the training dataset is passed to the algorithms in use i.e., Decision Tree, LSTM, etc. Next is Third Phase which is Testing and Evolution Phase where the testing data and trained model are used and based on which the final output is obtained which is the up and down movement of the graph which specifies the up and down values of the dataset of the stock. Also, the R2-Score, Mean Absolute Error (MAE), Root Mean Square Error (RMSE) are the Error Count which is identified in this phase itself and the error count and the best fit actual and predicted training and testing dataset are plotted on the graph and are compared to obtain the best performing algorithm out of the give five algorithms for the prediction of stock price is obtained for the best prediction of the stock price. Only the best performing algorithm is further used in phase four for best outcome. The fourth Phase or the Final Phase is Creating the User Interactive Webpage where the user just has to put the stock ticker from the yahoo finance website and the rest all work is done by the Webpage itself which is downloading the dataset of the stock

ticker provided by the user and giving the description, the previous five days price of the stock and graphical representation of the stocks historical price. After this the Webpage will import the dataset to the best performing algorithm and obtain the graph of training and testing dataset for the actual and predicted price of the stock and also show the Error Counts of the algorithm for the given dataset and also show the next ten days predicted price of the stock by the algorithm.

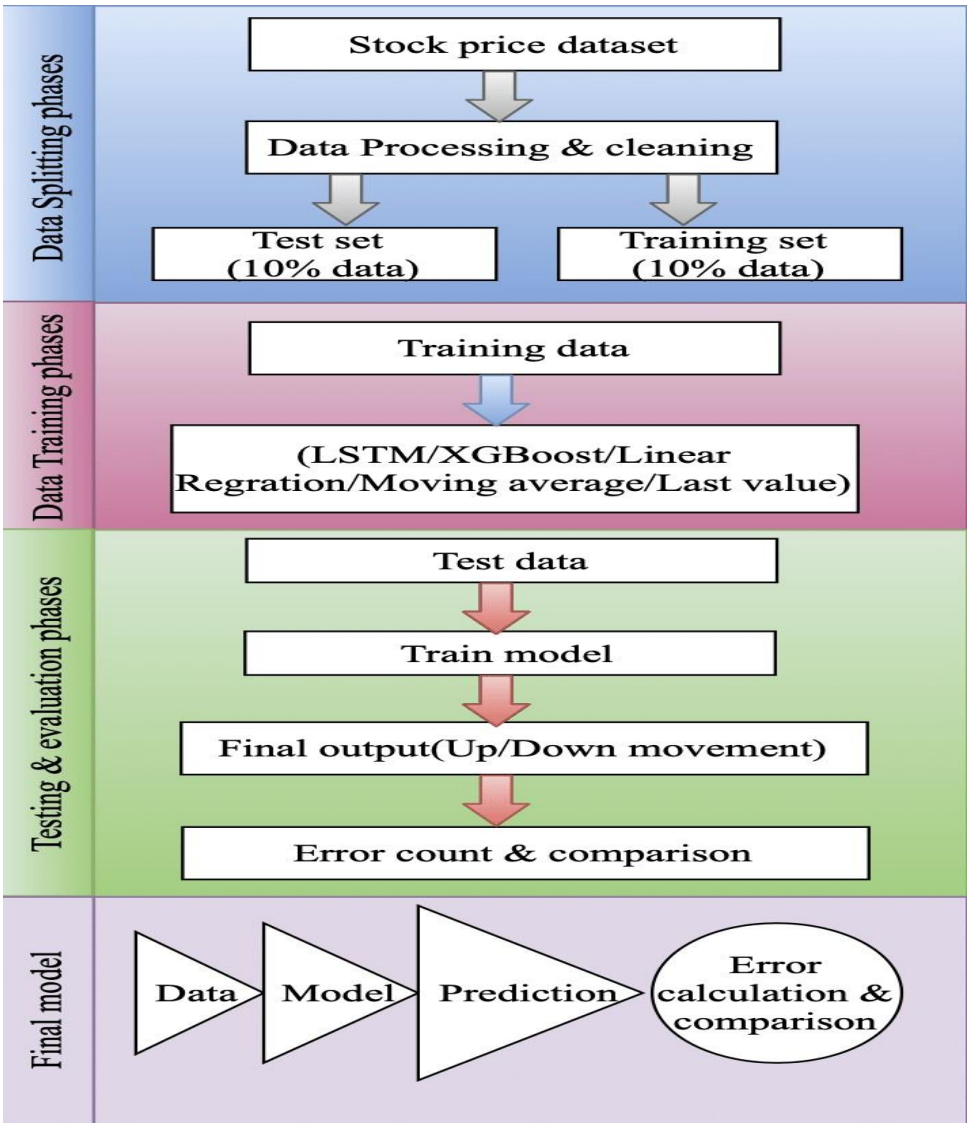

Figure3.1: System Architecture

3.3 Flow Chart

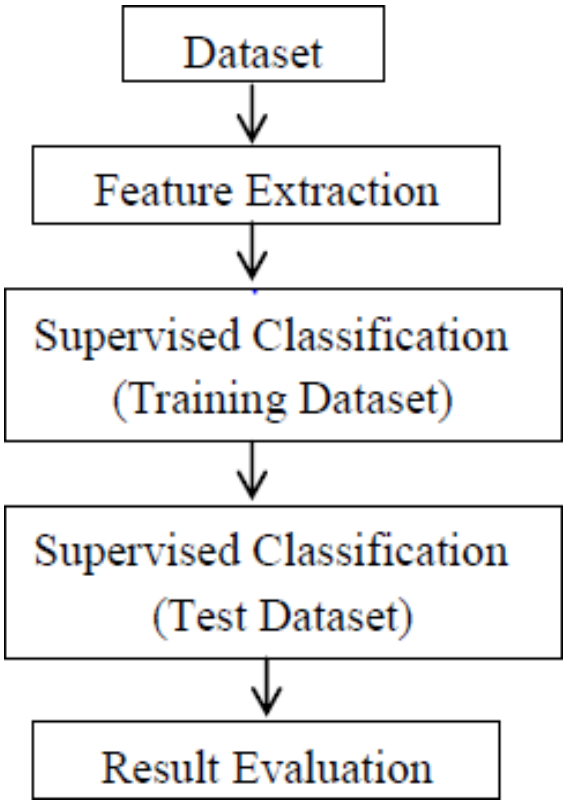

Figure3.2: System Flow

Here in the Figure, the system Flow is given that is first the dataset is given to the machine then based on that data we will do the Exploratory data analysis which will be further used to analyses which all the columns are require for our project base on this data we will distribute or split our dataset into the training and testing dataset based on which we will proceed with using these in our algorithms and evaluate them further.

3.4 Proposed System

The nature of stock is erratic and liberal. Impressive and reluctant in nature is the follow of the same. The finest hit goal for the same is locating predictability and obtaining the closest. Even now, it is still feasible to estimate anything precisely and accurately.

The pricing and rate of stock are both affected by a number of constraints. Before drawing any conclusions and starting to create the report, those restrictions had to be taken into account.

The dataset from which the proposed system will get input will be segmented feature-wise and classified below. The classification method utilized is supervised, and different machine-level algorithmic techniques are used to it.

In order to carry out the visualization and plotting tasks, test cases are constructed from the training dataset and implemented. The generated results are passed forward and graphically shown.

In this proposed system, we focus on predicting the stock values using five machine learning algorithms. In this proposed system, we have trained the machine from the various data points from the past to make a future prediction. We took the previous year's price data of the required stock to train the model. We majorly used six machine-learning libraries to solve the problem. The dataset we used is from the previous year's stock markets collected from the public database available online, 80 % of data is used to train the machine and the rest 20 % to test the data.

3.5 Modelling

The development of our prototype consists of the following modules:

3.5.1 Data Consideration

To make the project runs smoothly it's required that we make plan and design some accepts like Data Collection and Data Preprocessing which are defined below.

3.5.1.1 Data Collection

The Data collection Both for Research and Webpage differently for better understanding which is one of the important and basic things in our project. The right dataset must be provided to get robust results. Our data mainly consists of previous year stock prices. We will be taking and analyzing data from Nasdaq for research work and Yahoo finance for Webpage. The Dataset of the Stock is imported from any online available stock data provider sites by using which the research is done and the best performing algorithm is identified and based on this research the best performing algorithm is used in the webpage. For the

Webpage the dataset is imported from the yahoo finance website by using the ticker of the stock which is to be predicted using the algorithms.

```
In [5]: df.head()
```

Out[5]:

	Date	Open	High	Low	Close	WAP	No. of Shares	No. of Trades	Total Turnover	Deliverable Quantity	% Deli. Qty to Traded Qty	Spread H-L	Spread C-O
0	2022-05-26	421.4	422.55	404.95	420.70	414.20	1167082.0	24443.0	483402074.0	327447.0	28.06	17.60	-0.70
1	2022-05-25	428.8	428.95	413.60	417.05	419.93	509378.0	8254.0	213903135.0	106642.0	20.94	15.35	-11.75
2	2022-05-24	422.3	427.50	416.00	425.60	422.00	1123617.0	22290.0	474166530.0	372537.0	33.16	11.50	3.30
3	2022-05-23	421.1	431.10	419.70	421.60	424.70	693741.0	11490.0	294635127.0	124833.0	17.99	11.40	0.50
4	2022-05-20	410.0	421.35	409.20	417.95	417.54	1687548.0	32120.0	704612468.0	695716.0	41.23	12.15	7.95

Tabel3.1: The Dataset Used for Research

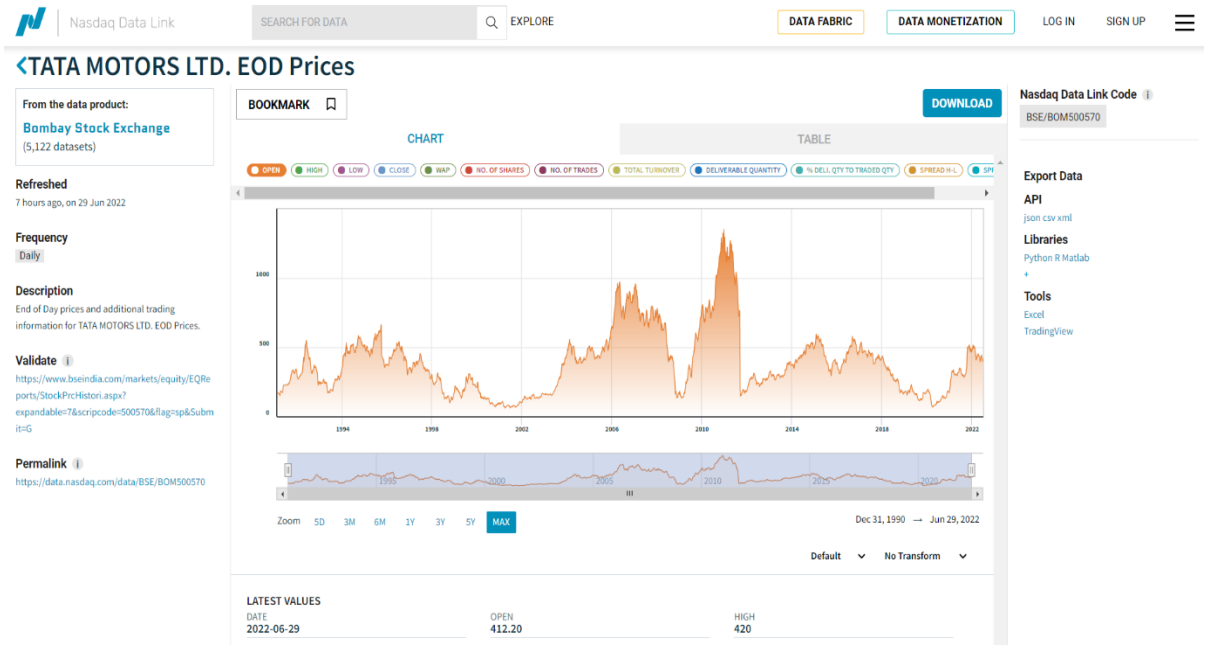

Figure3.3: Site Used for Research Work

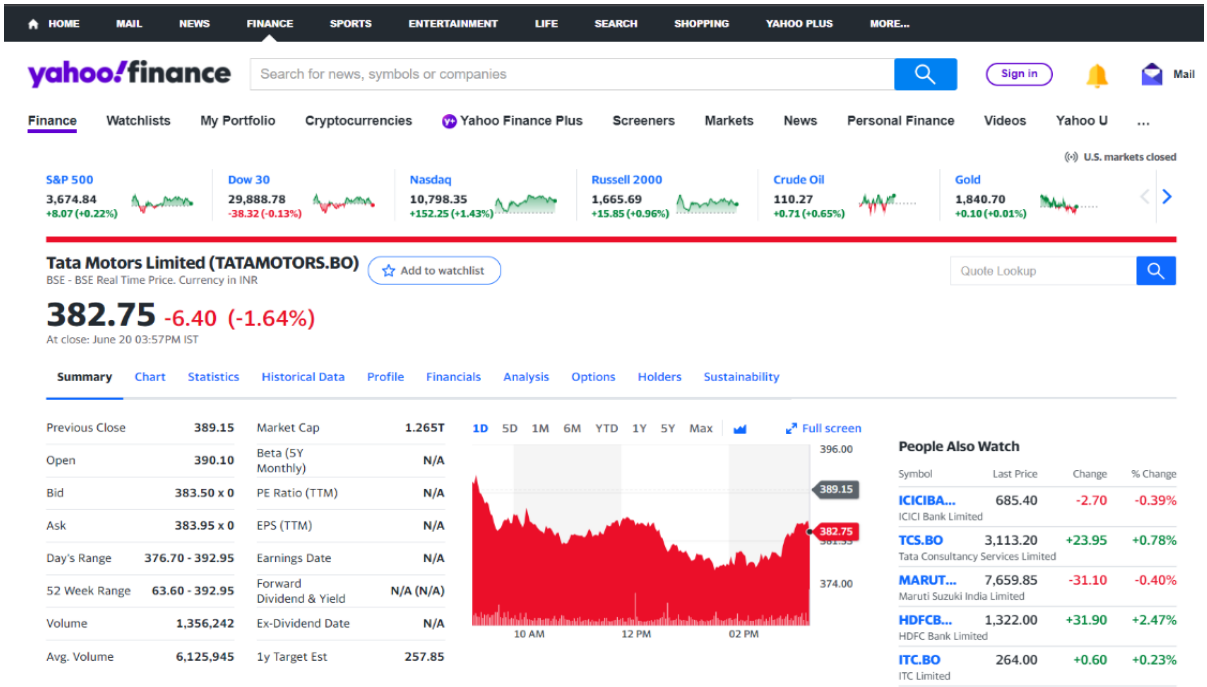

Figure3.4: Site Used for Webpage Prediction

3.5.1.2 Data Preprocessing

Human can understand any type of data but machine can't our model will also learn from scratch so it's better to make the data more machine readable. Raw data is usually inconsistent or incomplete. Data preprocessing involves checking missing values and remove the missing values or overcome the values and make it reliable for our module.

3.5.1.3 Future Selection

Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data by removing the columns that are not necessary in the project to be used in the algorithm and it is basically the columns which are not so important or which are not relevant to our project and we usually drop such columns from our dataset.

3.5.1.4 Splitting Training and Testing Datasets

Training model is a process in which a machine learning (ML) algorithm is fed with sufficient training data to learn from Similar to feeding somethings, machine/model should also learn by feeding and learning on data. The data set extracted from Nasdaq Website will be used to train the model. The training model uses a raw set of data as the undefined dataset which is collected from the previous fiscal year and from the same dataset a refine view is presented which is seen as the desired output. For the refining of the dataset various algorithms are implemented to show the desired output and Testing is referred to as the process where the performance of a fully trained model is evaluated on a testing set. Both the actual and predicted Stock price will get their train and test datasets based on which we will get our best performing algorithm which will be further used in our Webpage.

3.5.2 Algorithm Implementation

Below are the Machine Learning Algorithms implemented during the building of the project.

3.5.2.1 Linear Regression Algorithm

The linear regression algorithm is one of the well-known ones used in machine learning. It falls under statistical analysis as well as machine learning. It is used to examine the relationship between two variables, one of which has a known dependency with a known value and the other of which does not. The value of the unidentified reliance is compared to the values of the known dependencies, and the conclusion is drawn from this.

There are two sorts of dependencies on the variable chances. When both dependencies exhibit growth and are fully dependent and supportive of the changes flow, positive linear regression is the regression flow. [2] The regression flow known as negative regression occurs when one reliance prevents the expansion of another. This graph flow enters the picture if one dependency has a tendency to increase while the other one is shrinking.

The building block of linear regression, Single Linear Regression (SLR), is what they are. It is predicated that the two dependencies are linearly aligned and that altering the values of one will have an equivalent impact on the other.

A variation on the SLR algorithm, multi linear regression takes dependencies into account while considering several foundations. Even residual errors are handled by it.

In short, Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

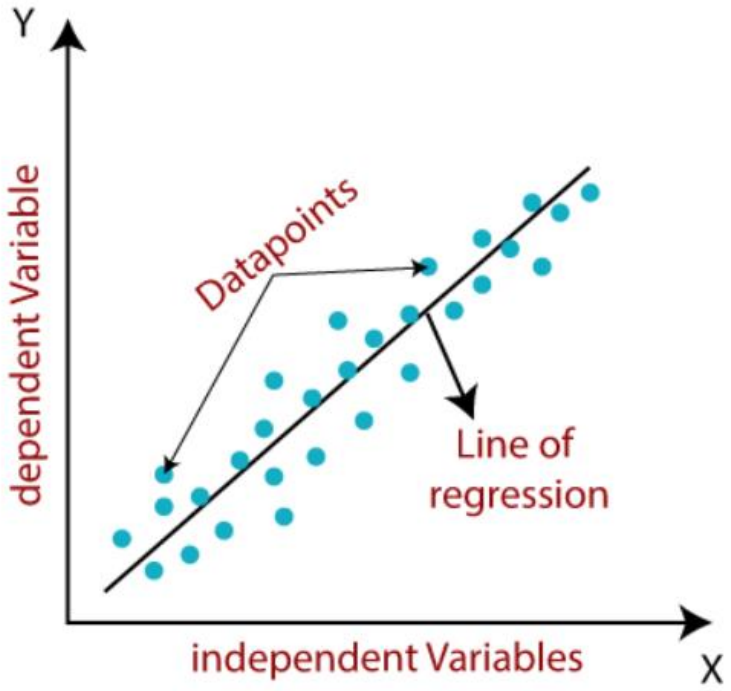

Figure3.5: Linear Regression Algorithm

3.5.2.2 KNN Algorithm

One of the machine learning algorithms that falls under both classification and regression. This module involves supervised learning. It's a crucial component of machine learning. It frequently appears in the data mining process.

Accordingly, this machine learning technique is utilized to resolve dataset regression and classification problems, in addition to its highly needed pattern recognition and intrusion detection capabilities. As implied by the name, this method deals with datasets that are close to one another.

Data modules, distance, and vector modules are calculated and shown on the dataset to show how similar they are. [1] Using the given dataset and the constant "k," which might be an integer number, the closest point is determined. The data are shown with the determined individual distance. The most acceptable terminology for this is Euclidean.

The distance data are arranged in ascending order and are aligned. The array is sorted using the closest distance index, "k," and the closest distance index is chosen. The dataset in question here deals with a large variety of values and their vicinity; it is scattered in nature and has many different categories. It is more practical because of the dispersion of the same. It addresses the proximity of the data.

Each division is broken down into little dataset pieces that are used to determine proximity and create results based on it. It's a simple algorithm, and its operation is clear to understand. Since it makes no assumptions about the dataset at the outset, these datasets are referred to as non-linear datasets.

It is a practical and adaptable technique that can be applied to both classification and regression of the data sets. The yield factor, which produces a favorable result set and is extremely accurate and effective, is the best.

In short, a k-nearest-neighbor is a data classification algorithm that attempts to determine what group a data point is in by looking at the data points around it. An algorithm, looking at one point on a grid, trying to determine if a point is in group A or B, looks at the states of the points that are near it.

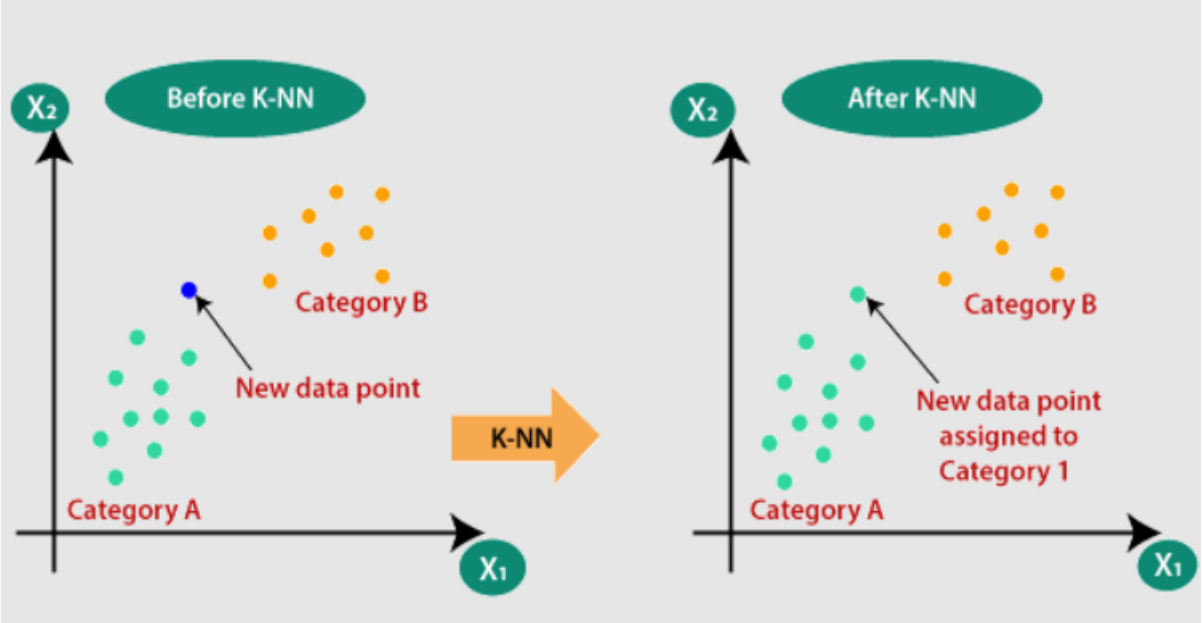

Figure3.6: KNN Algorithm.

3.5.2.3 Random Forest Algorithm

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.

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, there are two assumptions for a better Random Forest classifier they are as followed:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.

- The predictions from each tree must have very low correlations.

In short, Random Forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

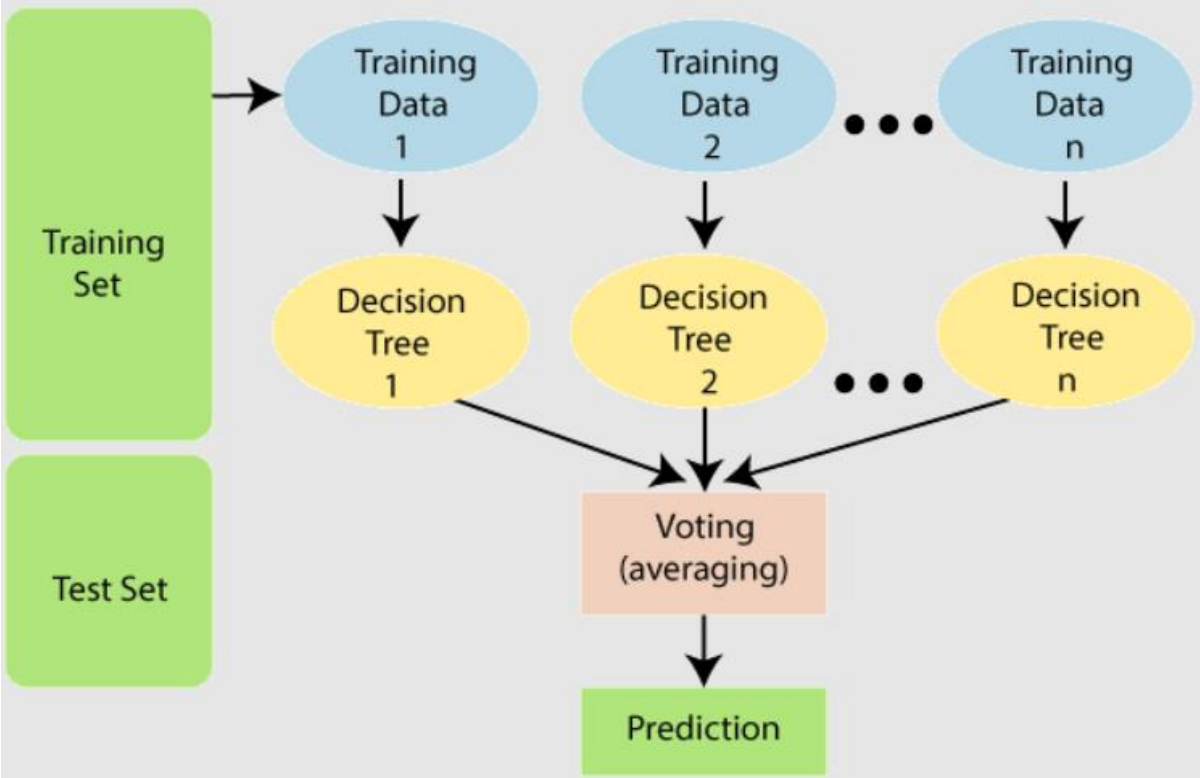

Figure3.7: Random Forest Algorithm.

3.5.2.4 SVM (Support Vector Machine) Algorithm

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional

space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. There are two different categories that are classified using a decision boundary or hyperplane.

SVM can be of two types:

- Linear SVM: Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
- Non-linear SVM: Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

In short, a support vector machine (SVM) is a type of deep learning algorithm that performs supervised learning for classification or regression of data groups. In AI and machine learning, supervised learning systems provide both input and desired output data, which are labeled for classification.

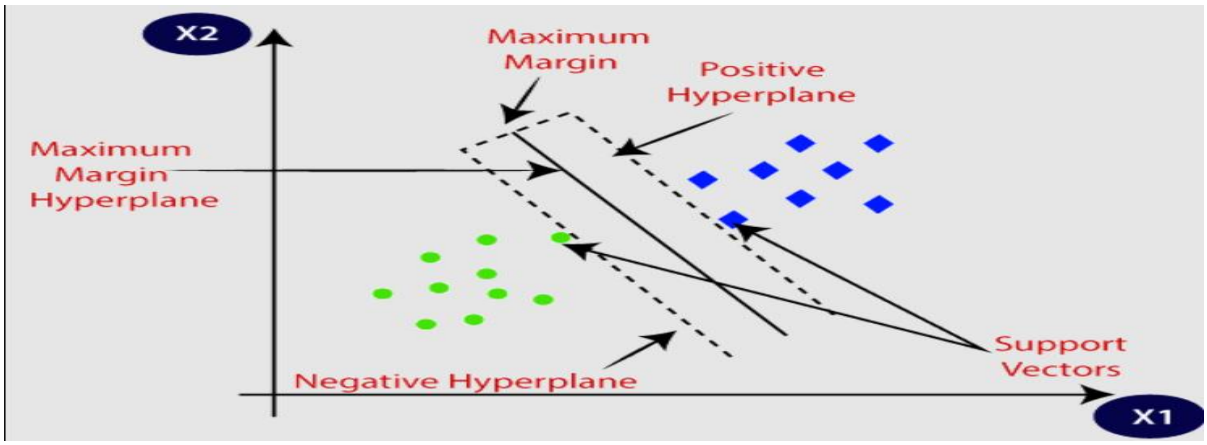

Figure3.8: SVM Algorithm.

3.5.2.5 LSTM (Long and Short Memory) Algorithm

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. It was proposed in 1997 by Sepp Hochreiter and Jurgen schmidhuber. Unlike standard feed-forward neural networks, LSTM has feedback connections. It can process not only single data points (such as images) but also entire sequences of data (such as speech or video).

Problems with sequence prediction have existed for a very long period. In the field of data science, they are regarded as among the most challenging issues to resolve. These cover a broad range of issues, from sales forecasting to spotting trends in stock market data, from deciphering movie storylines to identifying speech patterns, from language translations to anticipating your next word on your iPhone's keypad.

Long short-Term Memory networks, or LSTMs, have been determined to be the most efficient solution for almost all of these sequence prediction issues thanks to recent advancements in data science.

In many ways, LSTMs are superior to RNN and conventional feed-forward neural networks. They have the ability to selectively remember patterns for extended periods of time, which accounts for this. This article's goal is to introduce LSTM and show how to apply it to practical issues.

On the other hand, LSTMs only make minor adds and multiplications to the data. With LSTMs, the information travels via a system called cell states. LSTMs are able to selectively recall or forget things in this way. Three separate dependencies exist between the information at a specific cell state. They are used by businesses to transport goods around for various purposes. This technique is used by LSTMs to transfer information.

Information may be added to, modified, or removed as it passes through the various layers, just like a product may be moulded, painted, or packed while it is moving along a conveyor belt.

A general LSTM unit is composed of a cell, an input gate, an output gate, and a forget gate. The cell remembers values over arbitrary time intervals, and three gates regulate the flow of information into and out of the cell. LSTM is well-suited to classify, process, and predict the time series given of unknown duration.

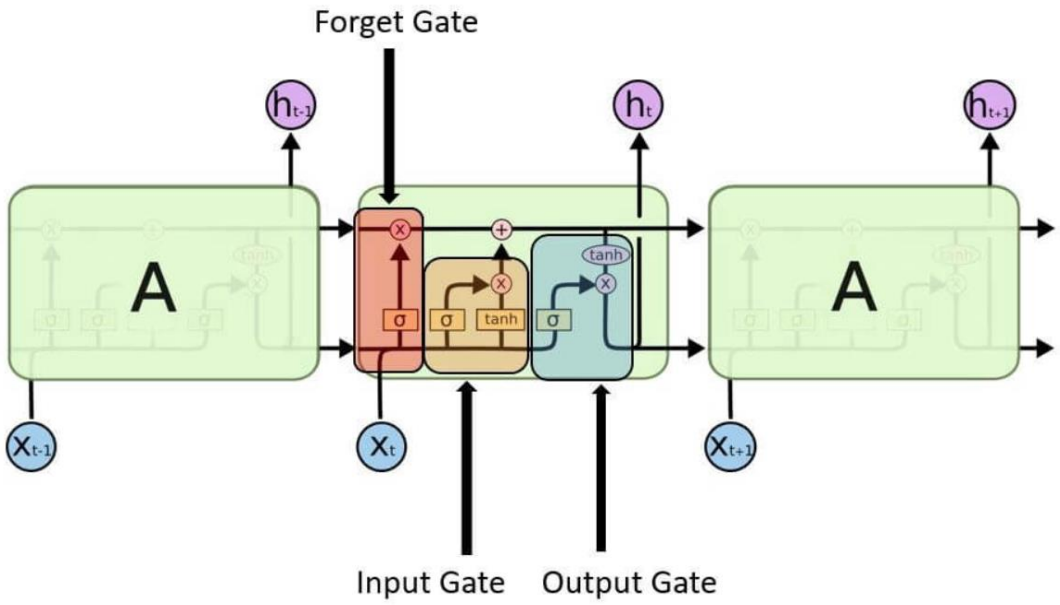

Figure3.9: LSTM Algorithm.

1. Input gate- It discover which value from input should be used to modify the memory. Sigmoid function decides which values to let through 0 or 1. And tanh function gives weightage to the values which are passed, deciding their level of importance ranging from -1 to 1.

$$i_t = \sigma(W_i.[h_t - 1, x_t] + bi)$$
$$C_t = \tanh(W_C.[h_t - 1, x_t] + b_C)$$

2. Forget gate- It discover the details to be discarded from the block. A sigmoid function decides it. It looks at the previous state (ht-1) and the content input (Xt) and outputs a number between 0(omit this) and 1(keep this) for each number in the cell state Ct-1.

$$f_t = \sigma(W_f.[h_t - 1, x_t] + b_f)$$

3. Output gate- The input and the memory of the block are used to decide the output. Sigmoid function decides which values to let through 0 or 1. And tanh function decides which values to let through 0, 1. And tanh function gives weightage to the values which are passed, deciding their level of importance ranging from -1 to 1 and multiplied with an output of sigmoid.

$$O_t = \sigma(W_o[h_t - 1, x_t] + b_o$$
$$h_t = o_t * \tanh(C_t)$$

In short, Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. It is more suitable for predicting big datasets values such as stock datasets.

3.5.3 Finding Algorithm with Least Error Values

Here the required stock dataset in csv formats is taken and all the preprocessing and cleaning is done. Then the given Dataset is splitted into training and testing dataset and this training and testing dataset is further used for different Algorithms.

Here five algorithms are used to predict the future stock price of the give stock dataset. The algorithms used here are Linear Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, Long-Short Term Memory (LSTM).

The R-Square or R2-Score is obtained for all the five algorithms here R2-Score shows how well the given algorithm is performing for the given dataset, here if the R2-Score is closed to zero then the given model is not performing well for the given dataset but if the R2-Score is close to one then the model is performing well for the given dataset.

Also, the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for Training and Testing dataset for both actual and predicted score is obtained to observe the performance of the given stock dataset.

By using matplotlib module's pyplot the graph for the training and testing for both actual and predicted Price is observed and based on Graph and R2-Score it is identified that Long Short

Term Memory (LSTM) performs well as compared to other four algorithms so we use LSTM with the user interactive webpage.

3.5.4 Creating User Interactive Webpage

After identifying the Algorithm that performs well for the give dataset of the stock the next step is to make a user-friendly interactive webpage where the user with a single click can obtain the description of the stock the user has selected and also predict using the LSTM algorithm. The Streamlit module is used here for creating an interactive webpage for the user. Streamlit provides a wide set of codes for ease implementation of the webpage for the Prediction of the stock price by using only the ticker of the stock from the yahoo finance. The webpage contains the link of the yahoo finance page where we can find the ticker of the stock and there is a input shell where we can insert the ticker of the stock and by using which the LSTM algorithm will be given the dataset of the stock for the given stock ticker. Based on the give dataset LSTM will do the training and testing of the dataset and will predict the next 10 days Price of the give stock

CHAPTER 4

SOFTWARE TESTING

Testing of any product comprise of giving the product an arrangement of test information and watching if the product carries on not surprisingly, if the product neglects to carry on obviously, then the conditions under which of disappointment happens are noted for investigating and amendment. At last, the framework in general is tried to guarantee that blunder in past countenances is revealed and the venture acts as determined. Testing is an important phase in the development life cycle of the product. This is the phase, where the remaining errors, if any, from all the phases are detected. Hence testing performs a very critical role for quality assurance and ensuring the reliability of the software. During the testing, the program to be tested was executed with a set of test cases and the output of the program for the test cases was evaluated to determine whether the program was performing as expected. Errors were found and corrected by using the below stated testing steps and correction was recorded for future references. Thus, a series of testing was performed on the system, before it was ready for implementation. It is the process used to help identify the correctness, completeness, security, and quality of developed computer software. Testing is a process of technical investigation, performed on behalf of stake holders, i.e., intended to reveal the quality- related information about the product with respect to context in which it is intended to operate. This includes, but is not limited to, the process of executing a program or application with the intent of finding errors.

The quality is not an absolute; it is value to some person. With that in mind, testing can never completely establish the correctness of arbitrary computer software; Testing furnishes a ‘criticism’ or comparison that compares the state and behavior of the product against specification. An important point is that software testing should be distinguished from the separate discipline of Software Quality Assurance (SQA), which encompasses all business process areas, not just testing. There are many approaches to software testing, but effective testing of complex products is essentially a process of investigation not merely a

matter of creating and following routine procedure. Although most of the intellectual processes of testing are nearly identical to that of review or inspection, the word testing is connoted to mean the dynamic analysis of the product-putting the product through its paces. Some of the common quality attributes include capability, reliability, efficiency, portability, maintainability, compatibility and usability. A good test is sometimes described as one, which reveals an error; however, more recent thinking suggest that a good test is one which reveals information of interest to someone who matters within the project community. Consequently, a progression of testing was performed on the framework, before it was prepared for usage. It is the procedure used to help recognize the accuracy, fulfilment, security, and nature of created PC programming. Testing is a procedure of specialized examination, performed for the benefit of partners, i.e., proposed to uncover the quality-related data about the item as for connection in which it is planned to work. This incorporates, however is not restricted to, the procedure of executing a project or application with the goal of discovering lapses.

There are numerous ways to deal with programming testing, yet viable testing of complex items is basically a procedure of examination not only a matter of making and taking after routine method. Albeit a large portion of the scholarly procedures of testing are almost indistinguishable to that of audit or investigation, the word testing is indicated to mean the dynamic examination of the item putting the item through its paces. A portion of the normal quality traits incorporate capacity, unwavering quality, productivity, versatility, viability, similarity and ease of use. A decent test is now and then depicted as one, which uncovers a slip; nonetheless, later thinking recommends that a decent test is one which uncovers data of enthusiasm to somebody who matters inside of the undertaking group.

4.1 BASICS OF SOFTWARE TESTING

4.1.1 Black Box Testing:

Black Box testing is done to find the following:

- Incorrect or missing functions
- Interface errors
- Errors on external access
- Performance error

4.1.2 White Box Testing:

This allows tests to:

- Check whether all independent paths within a module have been exercised at least once.
- Exercise all logical decisions on their false sides
- Execute all loops and their boundaries and within their boundaries
- Exercise the internal data structure to ensure their validity
- Ensure whether all possible validity checks and validity lookups have been provided to validate data entry.

4.2 TESTING TYPES

Following are the different types of testing:

- Unit testing
- Integration Testing
- System Testing
- Performance Testing
- Validation Testing
- Acceptance Testing

Let us consider each testing and discuss on it in detail. Firstly, we move to the first testing and give its detail description.

Unit Testing

Singular part is tried to guarantee that they work accurately. Every part is tried freely, without other framework segment. This framework was tried with the arrangement of legitimate test information for every module and the outcomes were checked with the normal yield. Unit testing centers around confirmation exertion on the littlest unit of the product outline module. This is otherwise called MODULE TESTING. This testing is done amid stages, every module is observed to work agreeable as respects to the normal yield from the module.

Integration Testing

Mix testing is another part of testing that is for the most part done keeping in mind the end goal to reveal mistakes related with stream of information crosswise over interfaces. The unit-tried modules are assembled together and tried in little section, which make it less demanding to seclude and revise mistakes. This Design and implementation of an IoT based forest environment monitoring system 2020-21 Page 64 Department of Computer Science and Engineering, TOCE approach is proceeded with unit I have coordinated all modules to frame the framework all in all.

System Testing

Framework testing is really a progression of various tests whose basic role is to completely practice the PC based framework. Framework testing guarantees that the whole incorporated programming framework meets prerequisites. It tests a design to guarantee known and unsurprising outcomes. A case of framework testing is the setup arranged framework mix testing. Framework testing depends on process depiction and streams, underscoring pre-driver process and incorporation focuses.

Performance Testing

The execution testing guarantee that the yield being delivered inside as far as possible and time taken for the framework aggregating, offering reaction to the clients and demand being send to the framework so as to recover the outcomes.

Validation Testing

The approval testing can be characterized from multiple points of view, however a straightforward definition is that. Approval succeeds when the product capacities in a way that can be sensibly expected by the end client.

Acceptance Testing

This is the last phase of testing procedure before the framework is acknowledged for operational utilize. The framework is tried inside the information provided from the framework procurer instead of recreated information.

Chapter 5

EXPERIMENTATION AND RESULTS

The required stock dataset in .csv formats is taken and all the preprocessing and cleaning is done. Then the given Dataset is splitted into training and testing dataset and this training and testing dataset is further used for different Algorithms. Here five algorithms are used to predict the future stock price of the give stock dataset. The algorithms used here are Linear Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, Long-Short Term Memory (LSTM).

The R-Square or R2-Score is obtained for all the five algorithms here R2-Score shows how well the given algorithm is performing for the given dataset, here if the R2-Score is closed to zero then the given model is not performing well for the given dataset but if the R2-Score is close to one then the model is performing well for the given dataset.

Also, the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for Training and Testing dataset for both actual and predicted score is obtained to observe the performance of the given stock dataset.

By using matplotlib module's pyplot the graph for the training and testing for both actual and predicted Price is observed and based on Graph and R2-Score it is identified that Long Short-Term Memory (LSTM) performs well as compared to other four algorithms so we use LSTM with the user interactive webpage.

Based on this research we have designed an interactive webpage by which we can use the yahoo finance stock ticker and find the prediction of next 10 days price of the stock.

Below Plots can be seen to know which algorithm is performing well for the given stock price dataset, all five Machine Learning Plot are presented below.

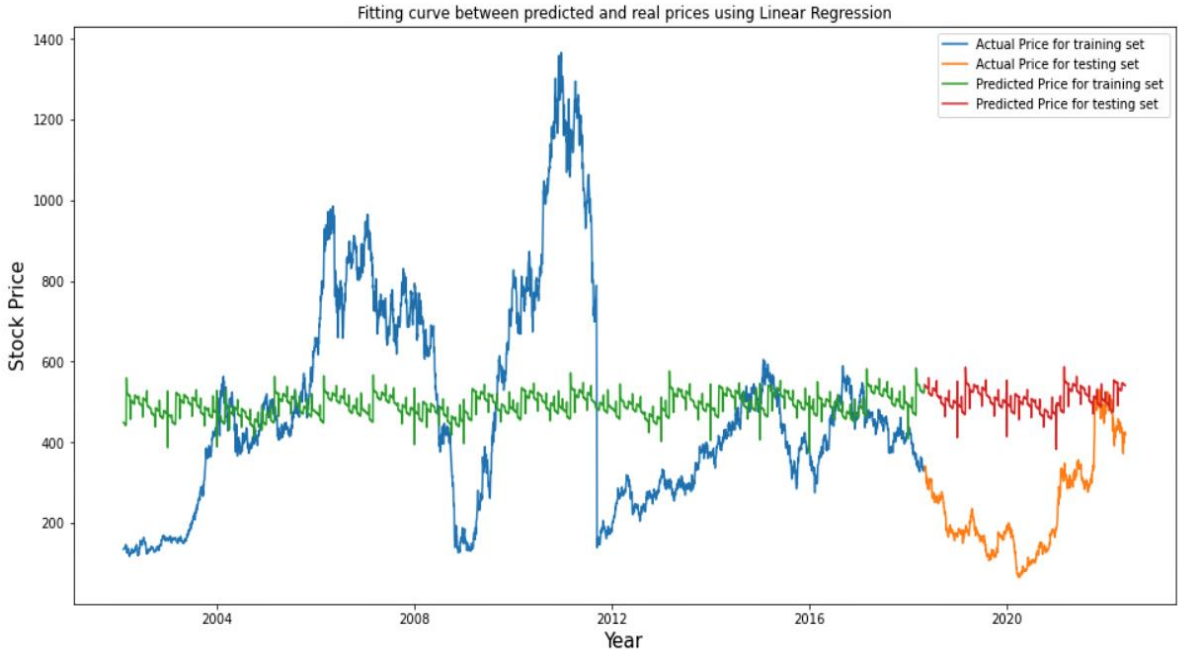

Figure5.1: Linear Regression Plot.

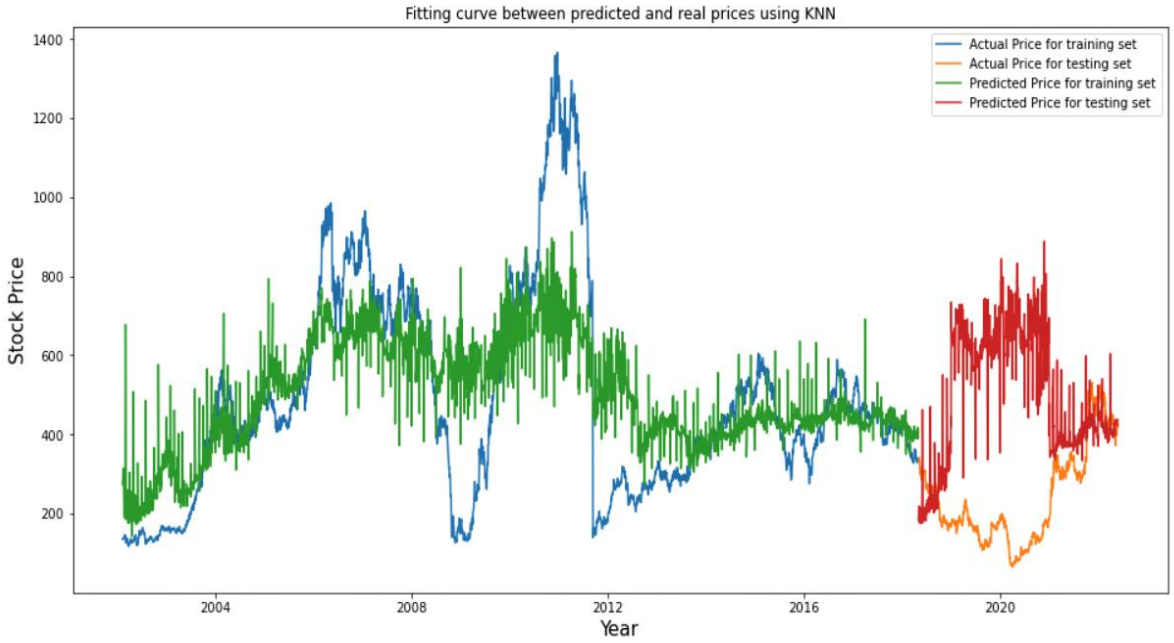

Figure5.2: KNN Algorithm Plot

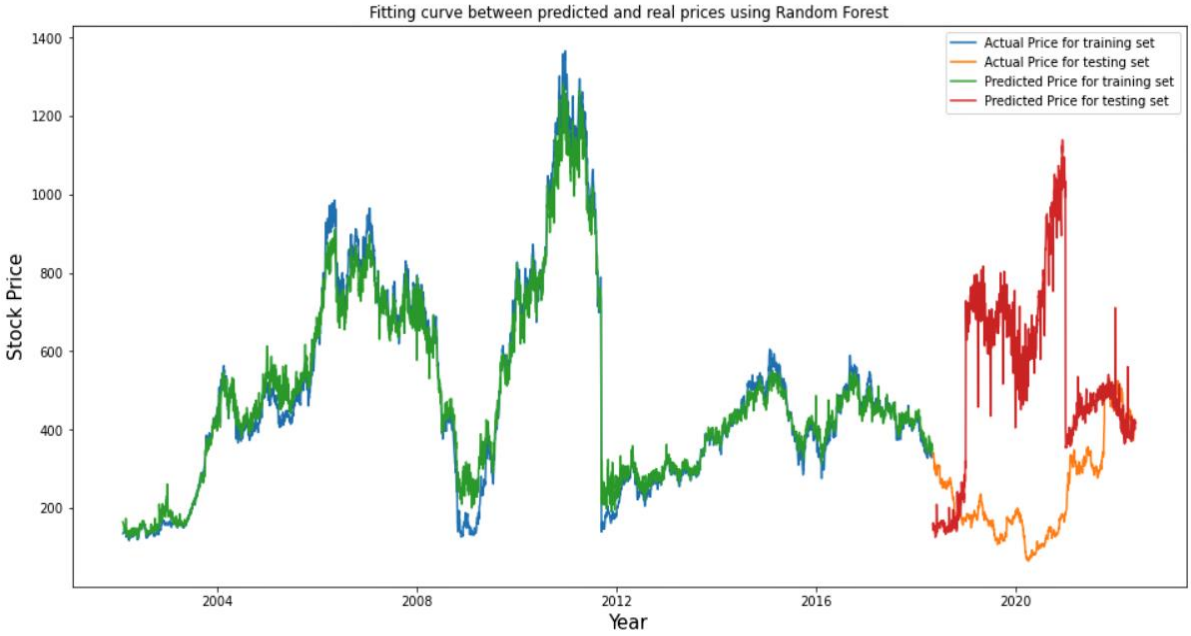

Figure5.3: Random Forest Algorithm Plot

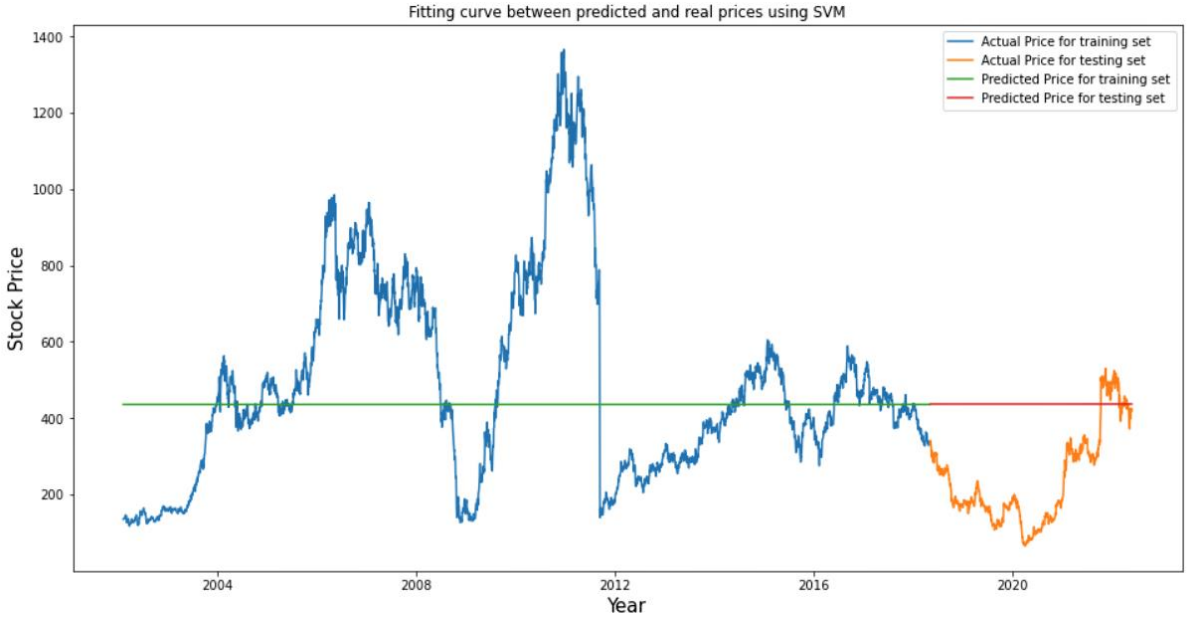

Figure5.4: SVM (Support Vector Machine) Algorithm Plot.

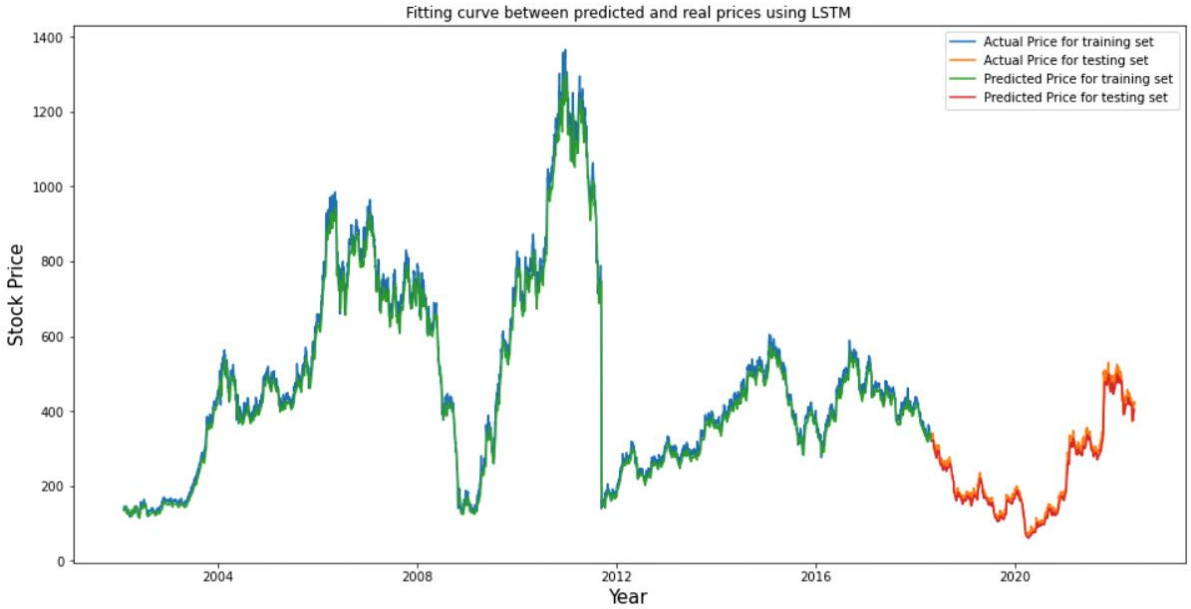

Figure5.5: LSTM (Long and Short Memory) Algorithm.

Here, in the above plots we can see that the LSTM Algorithm performs well as compared to all other four algorithm because of which we decided to implement LSTM in the Prediction Webpage.

Appendix A: Snapshots

	Method	R-Square Error(Training)	R-Square Error(Testing)	Mean Absolute Error(Training)	Mean Absolute Square Error(Testing)	Root Mean Square Error(Training)	Root Mean Square Error(Testing)
0	Linear Regression	0.009144	-4.809720	203.181620	267.204697	259.923719	290.307108
1	KNN	0.491967	-7.975387	132.594837	285.093765	186.117236	360.833332
2	SVM	-0.045406	-2.611942	195.339557	205.305532	266.982688	228.902456
3	Random Forest	0.975434	-11.775526	28.937902	335.242405	40.926747	430.496341
4	LSTM	0.992287	0.989749	14.024654	9.277777	22.772291	12.194355

Table5.1: The R-Square Error, MAE, RMS of all five ML algorithms in use.

	Days	Linear Regression	KNN	SVM	Random Forest	LSTM
0	Day1	541.232178	283.000000	435.999910	253.4635	428.036713
1	Day2	540.980760	251.294444	435.999915	213.7250	430.446472
2	Day3	540.729342	247.822222	435.999920	225.2810	435.348145
3	Day4	540.477924	228.183333	435.999926	227.8460	439.208435
4	Day5	543.560307	289.394444	436.000004	196.3880	442.597443
5	Day6	543.308889	254.055556	436.000010	168.5015	446.360535
6	Day7	543.057471	217.688889	436.000015	153.1005	450.290131
7	Day8	542.806053	187.694444	436.000020	169.2960	454.455231
8	Day9	542.554635	269.438889	436.000026	158.2890	458.830963
9	Day10	545.637018	401.833333	436.000104	222.8715	463.325012

Table5.2: Next ten days Stock Price prediction of all five ML algorithms in use.

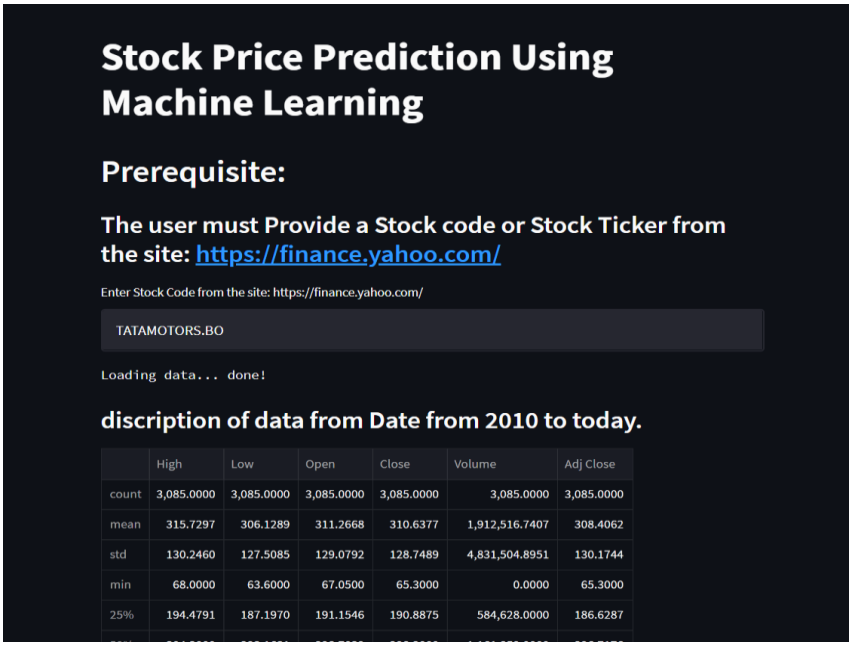

Figure5.6: The Start Page of the Webpage

Figure5.7: The Chart Containing the Actual and Predicted Price of the Training and Testing dataset in Webpage.

	Dates	LSTM Predicted Closing Value
0	Day1	388.3955993652344
1	Day2	383.78363037109375
2	Day3	378.366943359375
3	Day4	375.2313232421875
4	Day5	373.22552490234375
5	Day6	370.7102355957031
6	Day7	368.0760498046875
7	Day8	365.6556396484375
8	Day9	363.28314208984375
9	Day10	360.85858154296875

Figure5.8: Next ten days Stock Price prediction of LSTM in Webpage

CHAPTER 6

CONCLUSION AND FUTURE ENHANCEMENTS

6.1 CONCLUSION

To Conclude the Price of Stocks are tried to predict based on the previous few years dataset of the stock by using machine learning algorithms, from five algorithms the best algorithm is identified based on which the Webpage is created using the streamlit module of python. We believe that this stock price prediction using machine learning will be able to help many investors in minimizing their risk in investing and help them to predict the future days price of the stock.

6.2 LIMITATION AND FUTURE ENHANCEMENT

In this project the prediction is purely based on the previous year price datasets of the required stock so the investment purely based on this model is risky and stock price may depend on many other factors like the stock trend or the rumor about ups and downs of the stock. In such cases the stock price can change rapidly and prediction on such factors are yet to be discovered in future, if it is possible to link the current trend to this model the model will perform more accurately and preciously.

REFERENCES

[1] Advances in Distributed Computing and Artificial Intelligence Journal Regular Issue, Vol. 8 N. 4 (2019), 97-116 eISSN: 2255-2863

[2] Strader, Troy J.; Rozycki, John J.; ROOT, THOMAS H.; and Huang, Yu-Hsiang (John) (2020) "Machine Learning Stock Market Prediction Studies: Review and Research Directions," Journal of International Technology and Information Management: Vol. 28 : Iss. 4 , Article 3.

[3] International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181, Vol. 9 Issue 06, June-2020

[4] Hu, Z.; Zhao, Y.; Khushi, M. A Survey of Forex and Stock Price Prediction Using Deep Learning. Appl. Syst. Innov. 2021, 4, 9.

[5] Meng, T.L.; Khushi, M. Reinforcement Learning in Financial Markets. Data 2019, 4, 110.

[6] Shi, L.; Teng, Z.; Wang, L.; Zhang, Y.; Binder, A. DeepClue: Visual Interpretation of Text-Based Deep Stock Prediction. IEEE Trans. Knowl. Data Eng. 2019, 31, 1094–1108.

[7] Ballings M, Poel D V D, Hespeels N, Gryp R. Evaluating multiple classifiers for stock price direction prediction. Expert Systems with Applications, 2015, 42(2asdg0), pp. 7046–56.

[8] Milosevic N. Equity Forecast: Predicting Long Term Stock Price Movement Using Machine Learning. arXiv, 2016.

[9] Luca D P, Honchar O. Recurrent Neural Networks Approach to the Financial Forecast of Google Assets. International Journal of Mathematics and Computers in simulation, 2017, vol. 11, pp. 7–13.

[10] Roondiwala M, Patel H, Varma S. Predicting Stock Prices Using Lstm. International Journal of Science and Research (IJSR), 2017, vol. 6, pp. 1754–1756.

[11] Yang B, Gong Z J, Yang W. Stock Market Index Prediction Using Deep Neural Network Ensemble. 36th Chinese Control Conference (CCC), 2017, pp. 26–28.

[12] Rout, Ajit Kumar, P. K. Dash, Rajashree Dash, and Ranjeeta Bisoi. (2017) "Forecasting financial time series using a low complexity recurrent neural network and evolutionary learning approach." Journal of King Saud University-Computer and Information Sciences 29 (4) : 536-552

[13] Zhang J, Cui S, Xu Y, Li Q, Li T. A novel data-driven stock price trend prediction system. Expert Systems with Applications, 2018, 97(1), pp. 60–69.

[14] Hossain M A, Karim R, Thulasiram R K, Bruce N D B, Wang Y. Hybrid Deep Learning Model for Stock Price Prediction. IEEE Symposium Series on Computational Intelligence (SSCI), 2018, pp. 18–21.

[15] Powell N, Foo S Y, Weatherspoon M. Supervised and Unsupervised Methods for Stock Trend Forecasting. Paper presented at the 40th Southeastern Symposium on System Theory (SSST), 2008, pp. 203-205.

[16] Babu M S, Geethanjali N, Satyanarayana B. Clustering Approach to Stock Market Prediction. International Journal of Advanced Networking and Applications, 2012, vol. 3, pp. 1281-1291.