STOCK MARKET PRICE PREDICTION

& FORECASTING

A Course Project report submitted

in partial fulfillment of requirement for the award of degree

BACHELOR OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

by

V.PAVAN KUMAR **2103A51381**

Under the guidance of

Mr. D. RAMESH

Assistant Professor, Department of CSE.

SR UNIVERSITY

Department of Computer Science and Artificial Intelligence

SR UNIVERSITY

Department of Computer Science and Artificial Intelligence

CERTIFICATE

This is to certify that project entitled **“STOCK MARKET PRICE PREDICTION** " is the bonafied work carried out by **V . PAVAN KUMAR** bearing Roll No **2103A51381** as a Course Project for the partial fulfillment to award the degree **BACHELOR OF TECHNOLOGY** in **ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING** during the academic year 2022-2023 under our guidance and Supervision.

Mr. D. RAMESH
Asst. Professor,
S R University,
Ananthasagar ,Warangal

Dr. M. Sheshikala
Assoc. Prof .& HOD (CSE)
S R University,
Ananthasagar ,Warangal

ACKNOWLEDGEMENT

We express our thanks to Course co-coordinator **Mr.** D.Ramesh**, Asst. Prof.** for guiding us from the beginning through the end of the Course Project. We express our gratitude to Head of the department CS&AI, **Dr. M.Sheshikala, Associate Professor** for encouragement, support and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dean, School of Computer Science and Artificial Intelligence, **Dr C. V. Guru Rao**, for his continuous support and guidance to complete this project in the institute.

Finally, we express our thanks to all the teaching and non-teaching staff of the department for their suggestions and timely support.

ABSTRACT

The Stock Market Price Prediction project aims to develop a predictive model that can forecast the future stock prices of a particular company or market index based on historical data. The project involves collecting and preprocessing data, selecting appropriate features, training a suitable machine learning algorithm, and evaluating the performance of the model. The main objective is to create an accurate and reliable predictive model that can assist investors in making informed decisions about their investment strategies.

A balanced dataset can influence the performance of a classification method. The project uses various machine learning techniques such as regression analysis, time series analysis, and artificial neural networks to predict future stock prices and analyze market trends. The final product can potentially benefit individuals, financial institutions, and companies in their investment decisions by providing insights into the expected stock prices and market behavior.

Table of Contents

INTRODUCTION

1.1 OVERVIEW

Stock market price prediction is a popular area of research in the field of finance and machine learning. The stock market is one of the most important indicators of the economy, and predicting the prices of stocks is crucial for investors and traders who seek to make informed decisions and maximize their profits . In this project, we aim to develop a machine learning model that can predict the future prices of stocks based on historical data and other financial indicators. The model will be trained on a large dataset of historical stock prices and other financial indicators, such as volume, volatility, and market trends.

The goal of this project is to build an accurate and reliable model that can predict the future prices of stocks with a high degree of precision. The model will be evaluated using various metrics, such as mean squared error (MSE) and root mean squared error (RMSE), and will be compared to other existing models and techniques in the literature. Overall, the project aims to provide valuable insights into the field of stock market price prediction and contribute to the development of more accurate and reliable models for predicting the future prices of stocks

1.2. PROBLEM STATEMENT

To develop a model which can help us to predict the price of the stock market values of companies with low error rate and a high precision of accuracy. The model will not tell the future, but it might forecast the general trend and the direction to expect the prices to move.

1.3. EXISTING SYSTEM

Firstly, we collect the data set from the online source: Kaggle. The data set represents the stock price . The dataset includes all the information about stock prices from 27 October,2014 to 24 October ,2022. The second step involves filtering and cleaning the data set. This involves removing all the incomplete data from the rows. It also involves filtering out unnecessary features present in the data collected.

1.4. PROPOSED SYSTEM

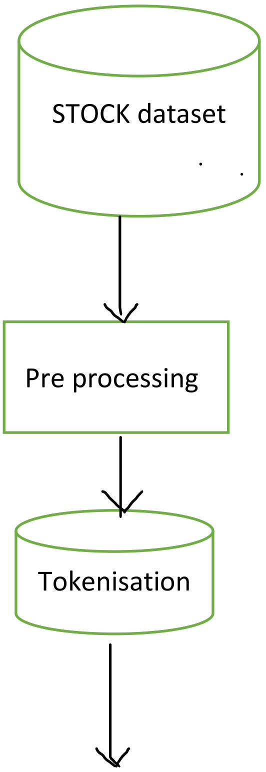
Training, followed by testing the dataset. We train our model, using the algorithm and the features taken into account to assist our model, to predict the future price of the stocks of company. Moving on to the testing part, we test the data to measure the accuracy of the algorithm that our model is using to predict the price of the stock.

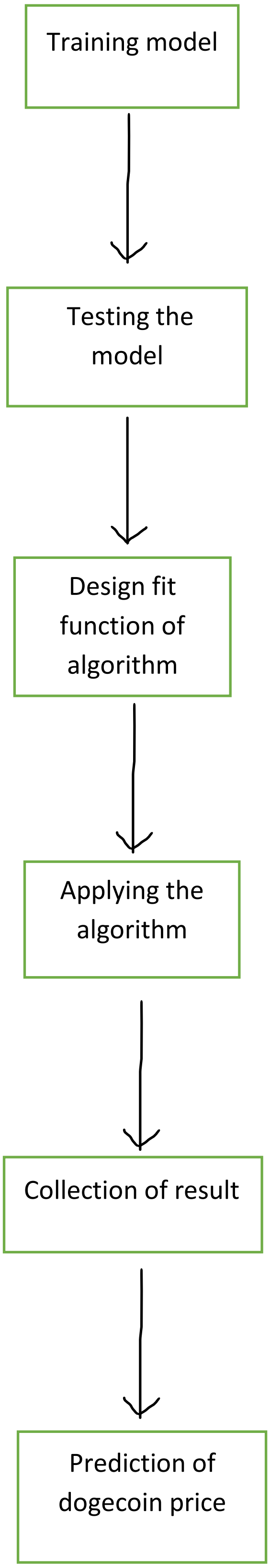
1.5. OBJECTIVES

The main objective of this research is to develop a model which can help us to predict the price of the stocks used , with low error rate and a high precision of accuracy. The model will not tell the future, but it might forecast the general trend and the direction to expect the prices to move

1.7. ARCHITECTURE

The architecture of the proposed system is as displayed in the figure below. The major components of the architecture are as follows: Dogecoin dataset, preprocessing, tokenization, training the model, test the model, design fitness function, application of algorithm, results collection and prediction of Dogecoin disease.




Training model
Testing the model
Design fit function of algorithm
Applying the algorithm
Collection of result
Prediction of dogecoin price

2.1.1 LITERATURE SURVEY

This would involve a comprehensive review of academic research, books, and other publications related to various aspects of the stock market. Some of the key topics that may be covered in such a survey include:

Efficient market hypothesis: The efficient market hypothesis (EMH) is a theory that suggests that financial markets are efficient and that all available information is already reflected in stock prices. A literature survey of the stock market would review the various studies that have been conducted to test this theory and evaluate its applicability to real-world markets.

Market microstructure: Market microstructure refers to the study of the process by which securities are traded in financial markets. This includes topics such as order flow, bid-ask spreads, and market liquidity. A literature survey of the stock market would examine the various studies that have been conducted to better understand market microstructure and its impact on stock prices.

Behavioral finance: Behavioral finance is an interdisciplinary field that combines psychology and finance to explain how investors make decisions. A literature survey of the stock market would review the various studies that have been conducted in this field to better understand how investor behavior affects stock prices.

Financial econometrics: Financial econometrics is the application of statistical methods to financial data in order to make inferences about the underlying economic processes. A literature survey of the stock market would review the various econometric techniques that have been developed and applied to stock market data to better understand stock price behavior.

3.DATA PRE-PROCESSING

3.1.1 DATASET DESCRIPTION

Sno	Attributes	Description
1.	OPEN	The opening price of the time period.
2.	HIGH	The highest price of the time period.
3.	LOW	The lowest price of the time period.
4.	CLOSE	The closing price of the time period.
5.	VOLUME	This is the volume in the transacted Company.
6.	ADJ CLOSE	The Adjacent closing price of the time period.

3.2 DATA CLEANING

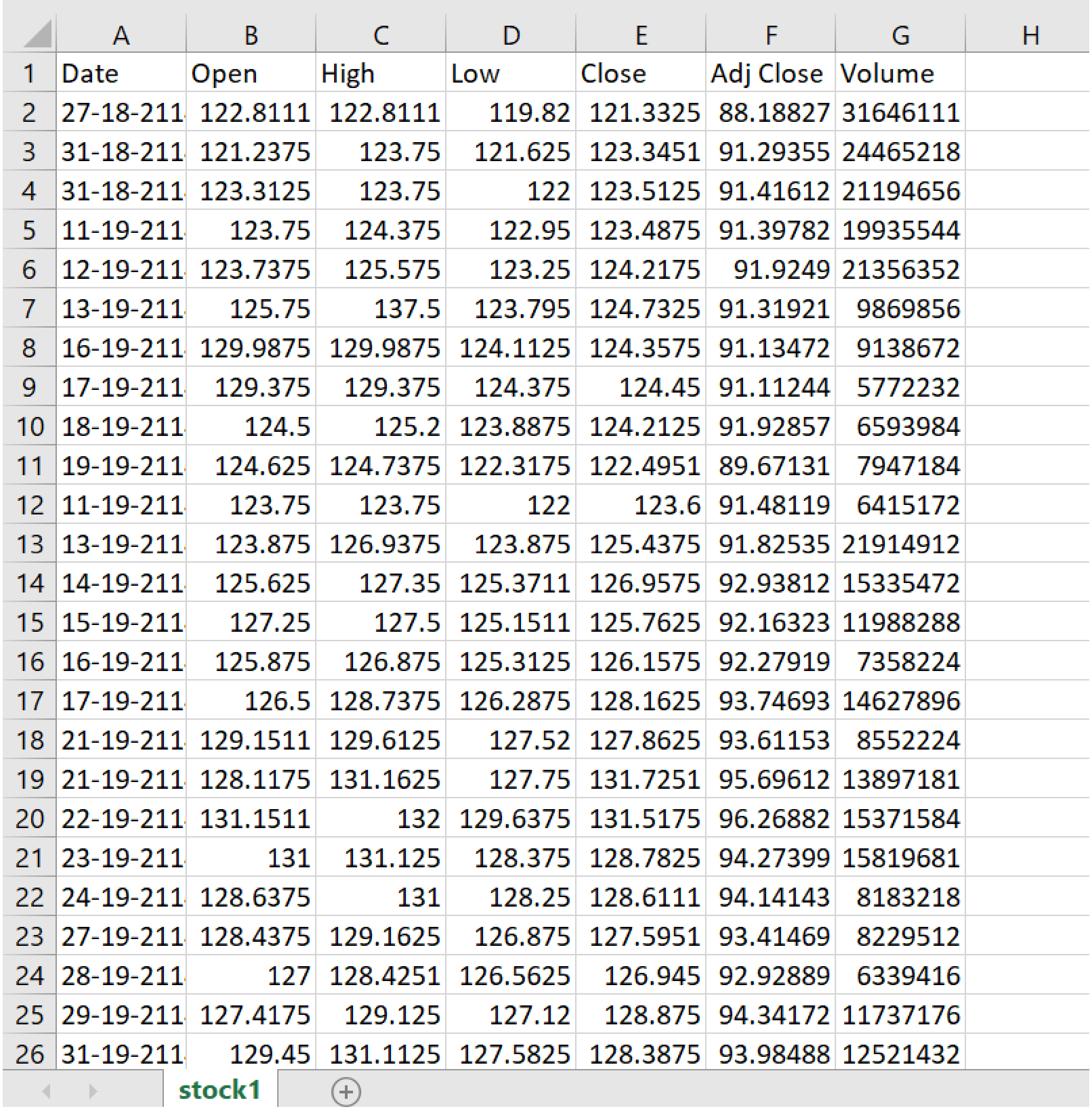
Data cleaning, also known as data cleansing, is the process of identifying and correcting or removing errors, inconsistencies, and inaccuracies from datasets. It helps ensure that the data is accurate, complete, and usable for analysis.Data auditing involves examining the dataset for any missing, duplicated, or inconsistent data. This can be done using descriptive statistics, visualization techniques, or automated tools. Data filtering involves removing any irrelevant or unnecessary data from the dataset. This can be done based on predefined criteria, such as data range, data type, or data quality. Data standardization ensuring that the data is consistent in format and structure. This can include converting data types, standardizing date formats, or converting units of measure. Data validation checking the data for accuracy and completeness. This can be done using data profiling techniques or by comparing the data with external sources. Data transformation involves modifying the data to make it suitable for analysis. This can include combining variables, creating new variables, or aggregating data.

Data imputation involves filling in missing values in the data. This can be done using various techniques, such as mean imputation, regression imputation, or hot-deck imputation. Data integration involves combining data from different sources to create a unified dataset. This can be done using data matching techniques or by merging datasets based on common variables. Overall , data cleaning is a critical step in data preparation that helps ensure that the data is accurate, complete, and ready for analysis.

3.4 DATA VISUALISATION

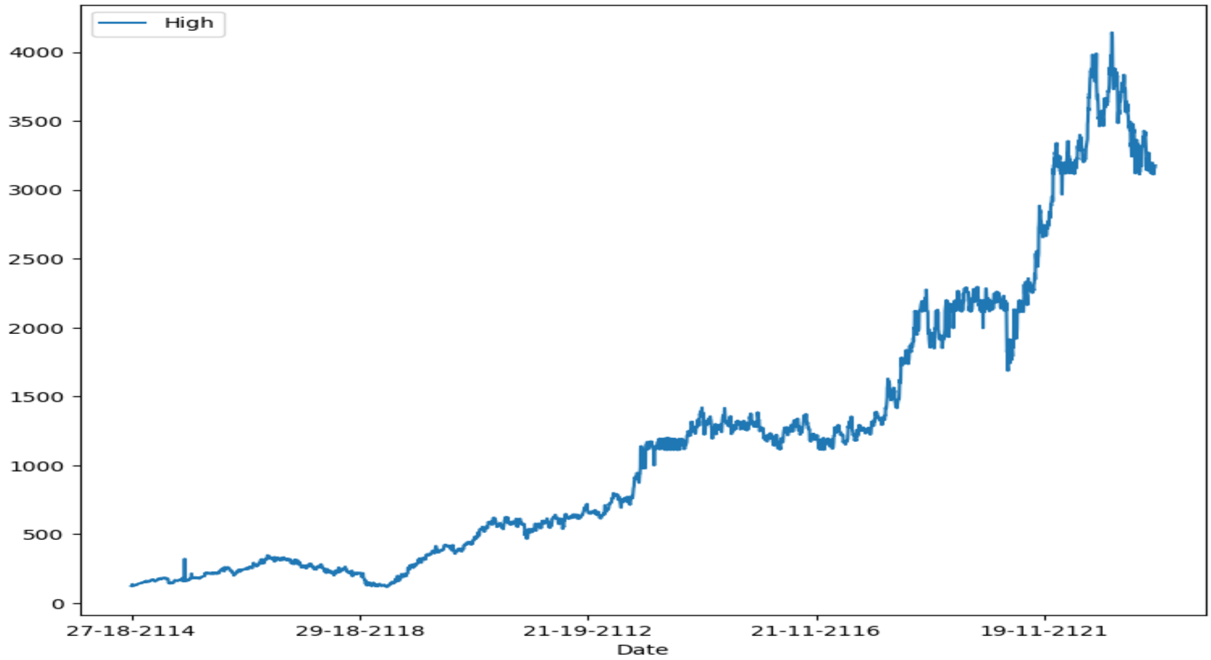
The historical Stock data set contains seven feature variables and two target variables output.

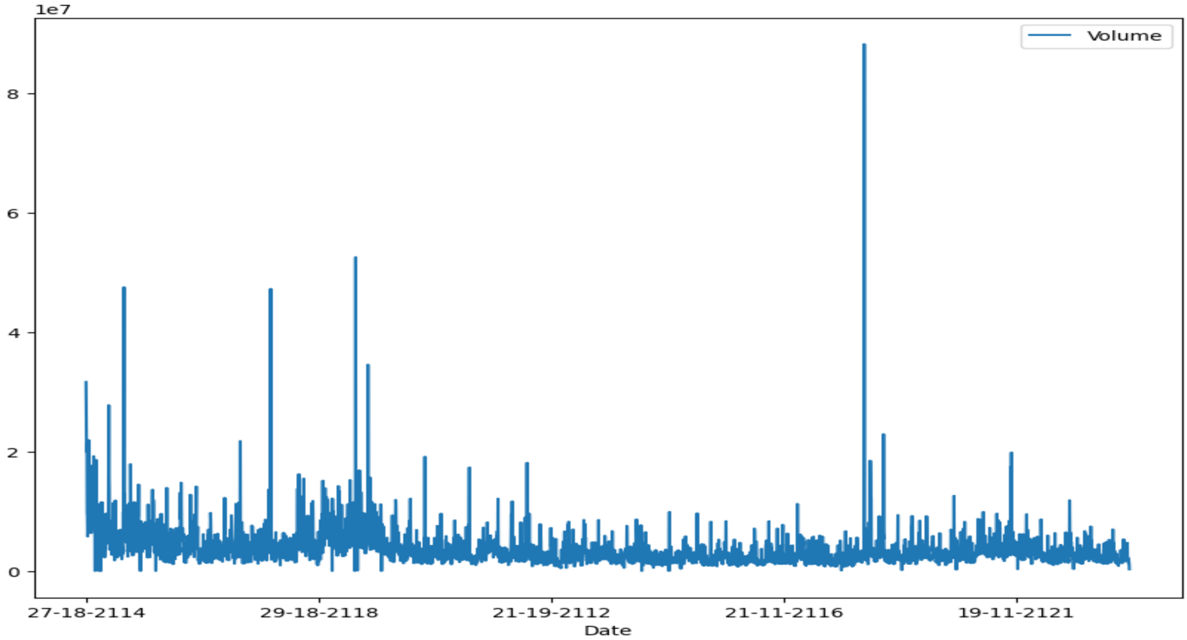
DATASET

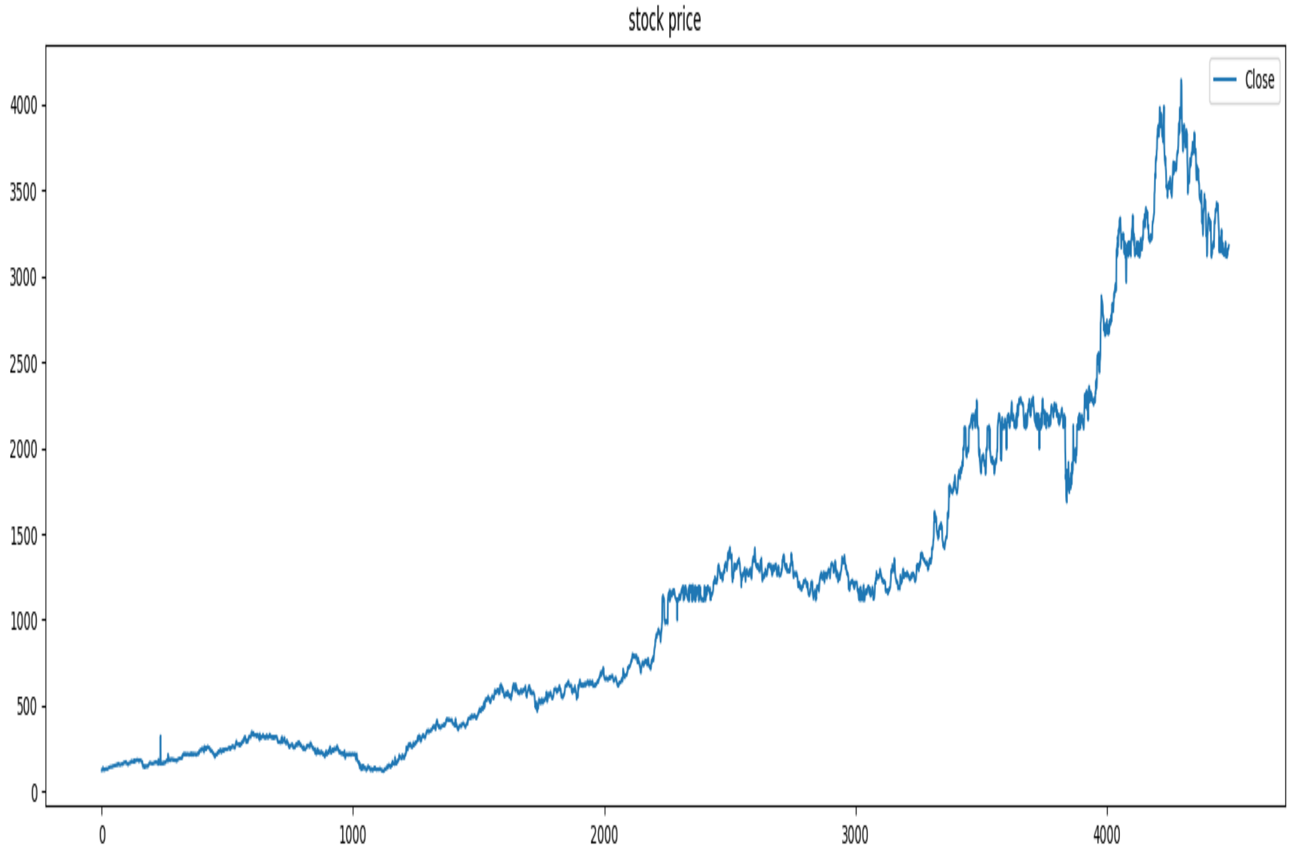


	A	B	C	D	E	F	G	H
1	Date	Open	High	Low	Close	Adj Close	Volume	
2	27-18-211	122.8111	122.8111	119.82	121.3325	88.18827	31646111	
3	31-18-211	121.2375	123.75	121.625	123.3451	91.29355	24465218	
4	31-18-211	123.3125	123.75	122	123.5125	91.41612	21194656	
5	11-19-211	123.75	124.375	122.95	123.4875	91.39782	19935544	
6	12-19-211	123.7375	125.575	123.25	124.2175	91.9249	21356352	
7	13-19-211	125.75	137.5	123.795	124.7325	91.31921	9869856	
8	16-19-211	129.9875	129.9875	124.1125	124.3575	91.13472	9138672	
9	17-19-211	129.375	129.375	124.375	124.45	91.11244	5772232	
10	18-19-211	124.5	125.2	123.8875	124.2125	91.92857	6593984	
11	19-19-211	124.625	124.7375	122.3175	122.4951	89.67131	7947184	
12	11-19-211	123.75	123.75	122	123.6	91.48119	6415172	
13	13-19-211	123.875	126.9375	123.875	125.4375	91.82535	21914912	
14	14-19-211	125.625	127.35	125.3711	126.9575	92.93812	15335472	
15	15-19-211	127.25	127.5	125.1511	125.7625	92.16323	11988288	
16	16-19-211	125.875	126.875	125.3125	126.1575	92.27919	7358224	
17	17-19-211	126.5	128.7375	126.2875	128.1625	93.74693	14627896	
18	21-19-211	129.1511	129.6125	127.52	127.8625	93.61153	8552224	
19	21-19-211	128.1175	131.1625	127.75	131.7251	95.69612	13897181	
20	22-19-211	131.1511	132	129.6375	131.5175	96.26882	15371584	
21	23-19-211	131	131.125	128.375	128.7825	94.27399	15819681	
22	24-19-211	128.6375	131	128.25	128.6111	94.14143	8183218	
23	27-19-211	128.4375	129.1625	126.875	127.5951	93.41469	8229512	
24	28-19-211	127	128.4251	126.5625	126.945	92.92889	6339416	
25	29-19-211	127.4175	129.125	127.12	128.875	94.34172	11737176	
26	31-19-211	129.45	131.1125	127.5825	128.3875	93.98488	12521432	

GRAPHS PLOTTED BETWWEN FEATURE AND TARGET VARIABLES:







4. METHODOLOGY

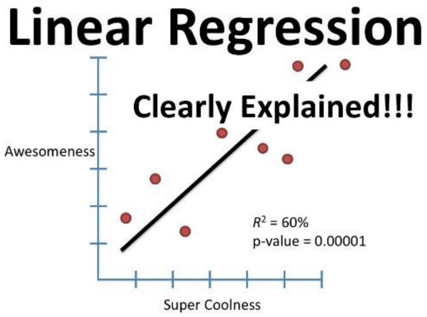
4.1 PROCEDURE TO SOLVE THE GIVEN PROBLEM

In this project Bitcoin price prediction and prediction we use three approaches:

- Linear regression
- Decision Tree
- K-Nearest Neighbour
- Support Vector Machine

Linear regression

Linear regression is a supervised machine learning method that is used by the Train Using AutoML tool and finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.

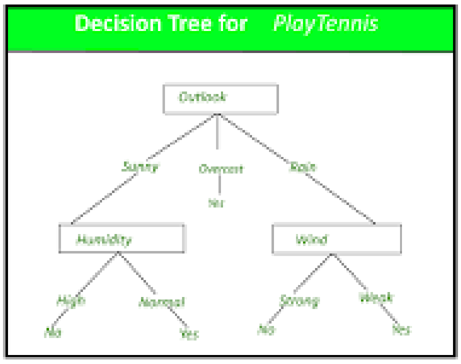


Advantages of linear regression algorithm:

- It handles overfitting pretty well using dimensionally reduction techniques, regularization, and cross-validation.
- Linear regression performs exceptionally well for linearly separable data.
- Easier to implement, interpret and efficient to train.
- One more advantage is the extrapolation beyond a specific data set.

DECISION TREE

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

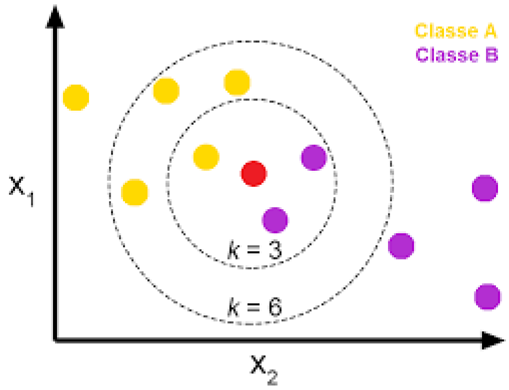


K-Nearest Neighbour

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

KNN Formula:

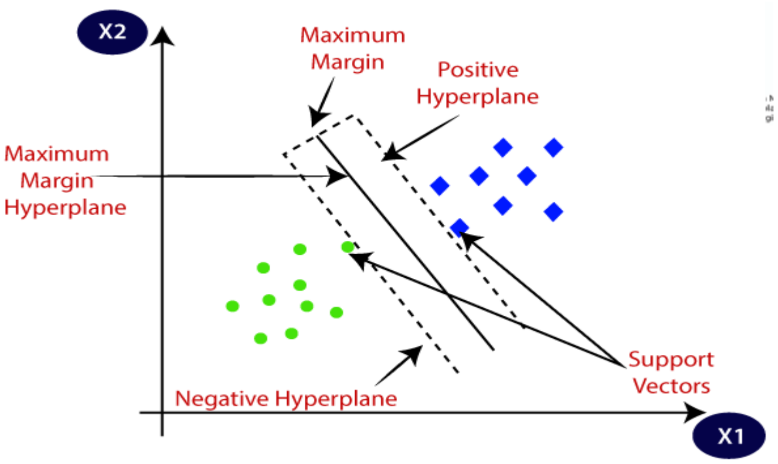
$$D(a,b) = \sqrt{\sum_{i=1}^{n} (b_i - a_i)^2}$$



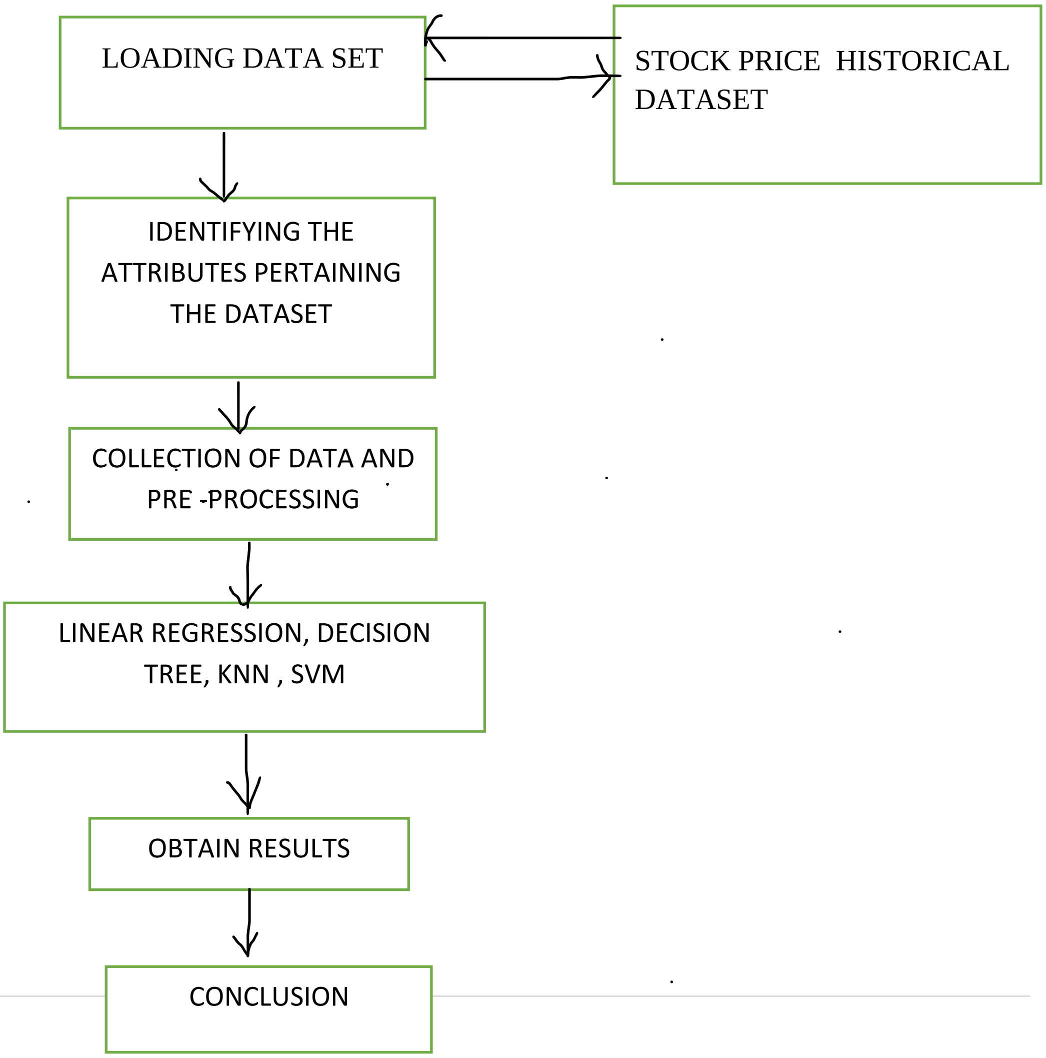
Support Vector Machine

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.



4.2 MODEL ARCHITECTURE


LOADING DATA SET
STOCK PRICE HISTORICAL DATASET
IDENTIFYING THE ATTRIBUTES PERTAINING THE DATASET
COLLECTION OF DATA AND PRE -PROCESSING
LINEAR REGRESSION, DECISION TREE, KNN , SVM
OBTAIN RESULTS
CONCLUSION

4.3 Requirement Specifications (S/W & H/W)

Hardware Requirements

✓	**System**	: 11th Gen Intel(R) Core(TM) i5-1155G7 @ 2.50GHz 2.50 GHz
✓	**RAM**	: 16 GB
✓	**Hard Disk**	: 1 TB
✓	**Input**	: Keyboard and Mouse
✓	**Output**	: PC

Software Requirements

✓	**OS**	: Windows 10, Mac, Linux
✓	**Platform**	: Google Colaboratory / Jupyter Notebook
✓	**Program Language**	: Python

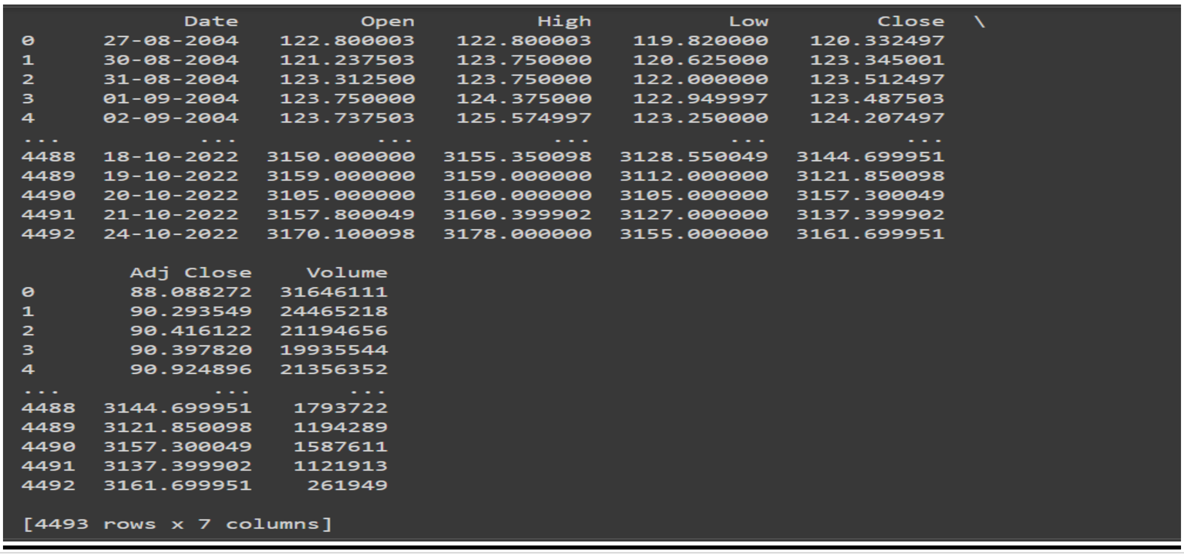
5. RESULTS

CODE

Dataset:

```
import pandas as pd
d=pd.read_csv('/content/stock1.csv')
print(d)
```

output:



```
            Date         Open         High          Low        Close  \
0     27-08-2004   122.800003   122.800003   119.820000   120.332497
1     30-08-2004   121.237503   123.750000   120.625000   123.345001
2     31-08-2004   123.312500   123.750000   122.000000   123.512497
3     01-09-2004   123.750000   124.375000   122.949997   123.487503
4     02-09-2004   123.737503   125.574997   123.250000   124.207497
...          ...          ...          ...          ...          ...
4488  18-10-2022  3150.000000  3155.350098  3128.550049  3144.699951
4489  19-10-2022  3159.000000  3159.000000  3112.000000  3121.850098
4490  20-10-2022  3105.000000  3160.000000  3105.000000  3157.300049
4491  21-10-2022  3157.800049  3160.399902  3127.000000  3137.399902
4492  24-10-2022  3170.100098  3178.000000  3155.000000  3161.699951

        Adj Close    Volume
0       88.088272  31646111
1       90.293549  24465218
2       90.416122  21194656
3       90.397820  19935544
4       90.924896  21356352
...           ...       ...
4488  3144.699951   1793722
4489  3121.850098   1194289
4490  3157.300049   1587611
4491  3137.399902   1121913
4492  3161.699951    261949

[4493 rows x 7 columns]
```

Linear regression:

```
From sklearn.model_selection import train_test_split
From sklearn.linear_model import LinearRegression
From sklearn.pipeline import make_pipeline
from sklearn.metrics import mean_squared_error as mse
from sklearn import metrics

X = d.drop('Volume', axis=1)
y = d['Volume']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)

regressor = LinearRegression()

regressor.fit(X_train, y_train)

y_pred = regressor.predict(X_test)
y_pred
```

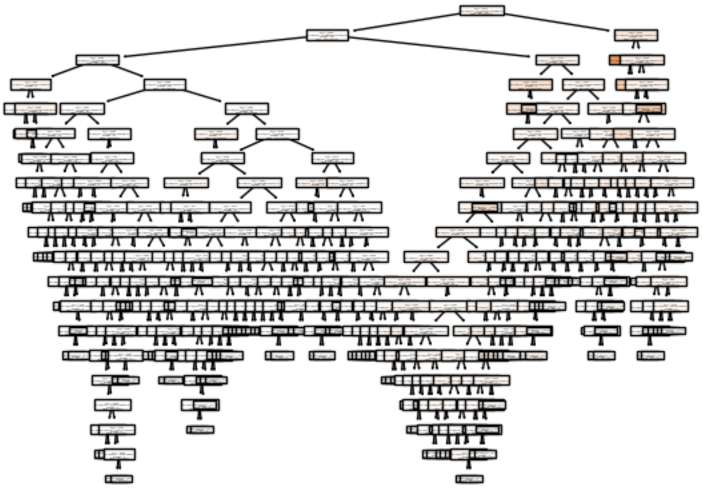
output:

```
array([2316539., 2411631., 3223314., ..., 2841821., 2975313., 3665111.])
```

Decision Tree:

```
from sklearn.tree import DecisionTreeRegressor
model=DecisionTreeRegressor()
model.fit(x_train,y_train)
y_pred=model.predict(x_test)
print(y_pred)
from sklearn.metrics import mean_squared_error
print(mean_squared_error(y_test,y_pred))
from sklearn import tree
tree.plot_tree(model,filled=True)
```

output:



K-Nearest Neighbour:

```
from sklearn.neighbors import KNeighborsRegressor
X = d.drop('Volume', axis=1)
y = d['Volume']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)

k = 3
clf = KNeighborsClassifier(n_neighbors=k)

clf.fit(X_train, y_train)

y_pred = clf.predict(X_test)
print("Mean Squared Error:", metrics.mean_squared_error(y_test, y_pred))
print("Accuracy:", metrics.accuracy_score(y_test, y_pred))
```

output:

```
Mean Squared Error: 15717510465481.559

Accuracy: 0.001483679525222552
```

Support Vector Machine:

```
from sklearn.svm import SVC

X = d.drop('Volume', axis=1)

y = d['Volume']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)

clf = SVC(kernel='linear')

clf.fit(X_train, y_train)

y_pred = clf.predict(X_test)

print("Accuracy:", metrics.accuracy_score(y_test, y_pred))
```

output:

```
[2784115 2456491 3183514 ... 1568224 2686842 2383132]

Accuracy: 0.002225519287833828
```

6. CONCLUSION AND FUTURE SCOPE

Even the most sophisticated algorithms and models can only provide probabilistic estimates of future stock prices, and these estimates may not always be accurate. Additionally, stock prices can be influenced by a wide range of factors, including macroeconomic trends, geopolitical events, and unexpected news, which can make them difficult to predict.

It's important to approach any stock market prediction project with caution and to always keep in mind the inherent uncertainty and volatility of the market. It's also important to remember that past performance is not necessarily indicative of future results.

In conclusion, while stock market prediction can be an interesting and potentially lucrative endeavor, it's important to approach it with a realistic understanding of the limitations and risks involved.The regression model, implemented here, is a basic model that takes into consideration only a few features that affect the stock price.

7.REFERENCES

[1]

[2] Agarwal, Basant, et al. "Prediction of dogecoin price using deep learning and social media trends." *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems* 8.29 (2021): e2-e2.

[3] Patra, Gyana Ranjan, and Mihir Narayan Mohanty. "Price Prediction of Cryptocurrency Using a Multi-Layer Gated Recurrent Unit Network with Multi Features." *Computational Economics* (2022): 1-20.

[4] Medzihorský, Juraj. "Dogecoin price prediction–can be a determinism supposed?." *Ekonomika a spoločnosť* 2.22 (2021): 67-81.

[5] Murphy, Marissa, et al. "Cryptocurrency Price Predictions Using High Performance Computing." (2021).