A MINI - PROJECT REPORT ON

Predicting Stock Prices Using Decision Tree Model

Submitted to the partial fulfillment of the requirement
for the 18CSE355T Data Mining and Analytics course and
for the award of the degree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENE AND ENGINEERING

Submitted by
Mr. Piyush Devansh
[Registration No. RA1811003030352]
Mr. Siddharth Shukla
[Registration No. RA1811003030345]
GROUP ID: 2020DDDD

Under the guidance of
Mrs. MEGHA AGARWAL
(Assistant Professor, Department of Computer Science and Engineering)

DEPARTMENT OF COMPUTER SCIENE AND ENGINEERING
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
(Deemed to be University under Section 3 of the UGC Act, 1956)

NOV 2020

BONAFIDE CERTIFICATE

Certified that this project report titled "**Predicting Stock Prices Using Decision Tree Model**" is the bonafide work of Mr. Piyush Devansh [Registration No. RA1811003030352] and Mr. Siddharth Shukla [Registration No. RA1811003030345] , who carried out the project work under my supervision and submitted to the partial fulfillment of the requirement for the 18CSE355T Data Mining and Analytics course and for the award of the degree of Bachelor of Technology in Computer Science and Engineering of SRM Institute of Science and Technology.

Mrs. MEGHA AGARWAL

Supervisor

ABSTARCT

Our aim is to analyze previous stock data of a certain companies registered in NSE (National Stock Exchange) with help of certain parameters that affect stock value keeping the changing foreign exchange rates in mind. The decision taken will be based on decision tree classifier which is one of the data mining techniques. To build the proposed model, the CRISP-DM methodology is used over real historical data of companies listed in NSE.This will also help us to determine the values that particular stock will have in near future to help investors in growth of their portfolio and monitor the investment. This study tries to help the investors in the stock market to decide the better timing for buying or selling stocks based on the knowledge extracted from the historical prices of such stocks.

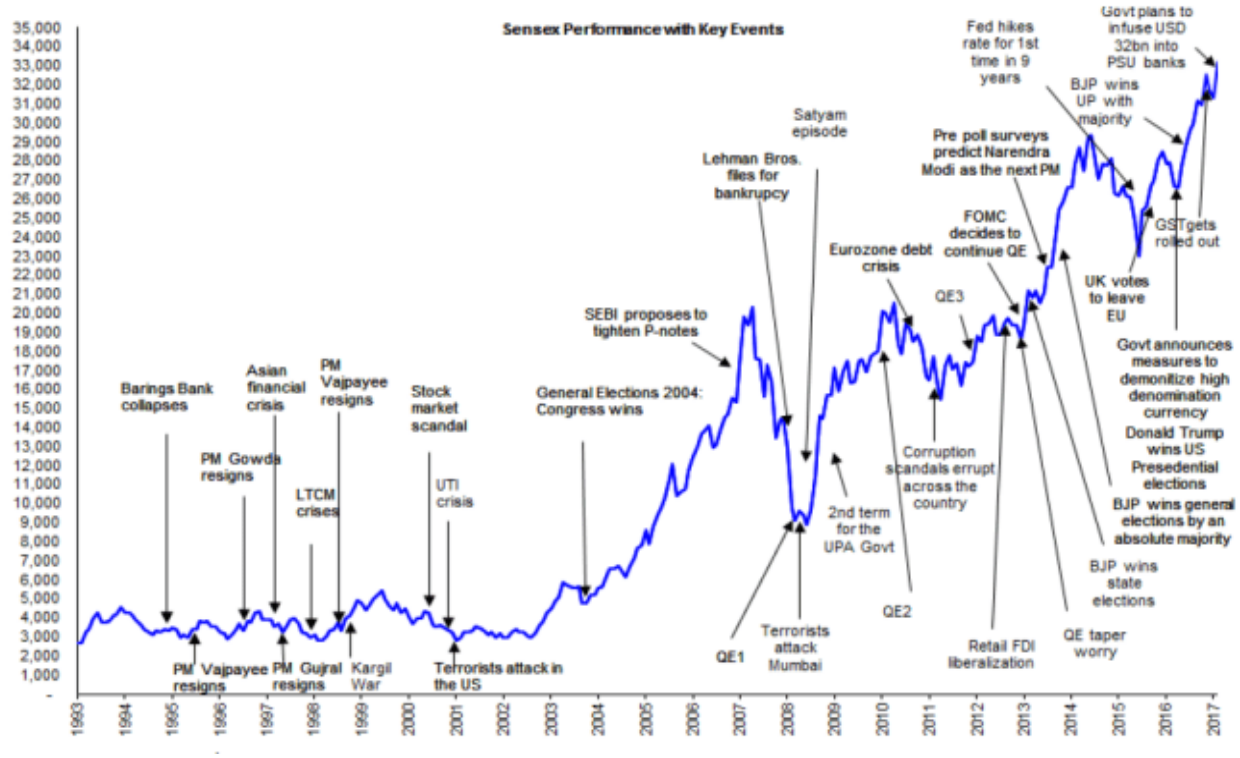


TABLE OF CONTENTS

INTRODUCTION

Stock market prediction is the act of trying to determine the future value of a company stock or other financial instrument traded on an exchange. The successful prediction of a stock's future price could yield significant profit. The efficient-market hypothesis suggests that stock prices reflect all currently available information and any price changes that are not based on newly revealed information thus are inherently unpredictable. Others disagree and those with this viewpoint possess myriad methods and technologies which purportedly allow them to gain future price information.

Technical analysis as illustrated in and refers to the various methods that aim to predict future price movements using past stock prices and volume information. It is based on the assumption that history repeats itself and that future market directions can be determined by examining historical price data. Most of the techniques used in technical analysis are highly subjective in nature and have been shown not to be statistically valid. Recently, data mining techniques and artificial intelligence techniques like decision trees, rough set approach, and artificial neural networks have been applied to this area . Data mining refers to extracting or mining knowledge from large data stores or sets. Some of its functionalities are the discovery of concept or class descriptions, associations and correlations, classification, prediction, clustering, trend analysis, outlier and deviation analysis, and similarity analysis. Data classification can be done in many different methods; one of those methods is the classification by using Decision Tree. It is a graphical representation of all possible outcomes and the paths by which they may be reached. Decision trees and artificial neural networks can be trained by using an appropriate learning algorithm.

TECHNOLOGY USED

MACHINE LEARNING:

Machine learning is a study of computer science that provides computers the ability to learn without being explicitly programmed. Machine learning is used to study algorithms that learn from and make predictions on data. Machine learning is related to computational statistics, which also focuses on prediction making. Within the field of data analytics, machine learning is a method used to devise complex models and algorithms that end themselves to prediction; in commercial use, this is known as predictive analytics. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves. The learning process begins with observations or data, examples, direct experience, or instruction, in order to find patterns in data and make better decisions in the future based on the examples provided. The objective is to allow the computers to earn automatically without human assistance and adjust actions accordingly.

WEKA TOOL: Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. Weka supports several standard data mining tasks, more specifically, data pre-processing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as one flat file or relation, where each data point is described by a fixed number of attributes normally, numeric or nominal attributes.

METHODOLOGY

Data analytics :

Data analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. Data mining is a data analysis technique which focuses on modelling and knowledge discovery typically for predicting the future. Data analytics is performed on the Given database so that the data can be cleaned and data that is not required can be deleted. It refers to qualitative and quantitative techniques and processes used to enhance productivity and business gain. Data is categorized and extracted to identify and analyse similar behavioural data and patterns, and techniques vary according to organizational requirements. Data analysis is linked to data visualization. It is used to make relationships between different columns of the database so that it can be used for prediction and visualization. Data visualization is the way of presenting data in a pictorial or graphical format. It enables decision makers to see analytics presented visually, to understand difficult concepts or to identify new patterns. Visualization helps to make charts and graphs for more detail, thereby changing what data you see and how it's processed. This helps us to understand which features of data have strong relations between them.

Classification:

It is a Data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts. Classification is the problem of identifying to which of a set of categories (subpopulations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known. Example: Before starting any Project, we need to check it's feasibility. In this case, a classifier i s required to predict class labels such as ' Safe' and ' Risky' for adopting the Project and to further approve it.

Understanding the collected data:

The National Stock Exchange (NSE) contains the historical prices of the companies listed in the exchange from the year 1992. As the amount of such data is very large and complicated, the decision was taken to choose three companies listed in the exchange. These companies are “Punjab National Bank”, its’ code in the stock market “PNB” and it belongs to the banking sector, “National Thermal Power Corporation ”, its’ code is “NTPC” and it belongs to the power sector, and “Bharat Heavy Electrical Limited”, its’ code is “BHEL” and it belongs to the industrial sector. The period that was selected is from April 2018 to May 2019, which presented the current and actual status of the market at that period of time before COVID-19. The data collected contains 6 attributes. Table 1 shows the 6 attributes selected with their descriptions and their possible values. The class attribute is the investor action whether to buy or sell that stock and it is named, “Action”. The data of this attribute was taken also from NSE database, which is the net position of one of the biggest brokers dealing with the above mentioned stocks every day. The net position could be either buying or selling that stock for that day.

Attribute	**Description**	**Possible Values**
Previous	Previous day close price of the stock	Positive, Negative, Equal
Open	Current day open price of the stock	Positive, Negative, Equal
Min	Current day minimum price of the stock	Positive, Negative, Equal
Max	Current day maximum price of the stock	Positive, Negative, Equal
Last	Current day close price of the stock	Positive, Negative, Equal
Action	The action taken by the investor on this stock	Buy, Sell

Preparing the data:

When the data was collected, all the values of the attributes selected were continuous values. Data transformation was applied by generalizing data to a higher-level concept so as all the values became discrete. The criterion that was made to transform the numeric values of each attribute to discrete values depended on the previous day closing price of the stock. If the values of the attributes open, min, max, last were greater than the value of attribute previous for the same trading day, the numeric values of the attributes were replaced by the value Positive. If the values of the attributes mentioned above were less than the value of the attribute previous, the numeric values of the attributes were replaced by Negative. If the values of those attributes were equal to the value of the attribute previous, the values were replaced by the value Equal. Table show the same sample after selecting the 6 attributes and after transforming them to discrete values.

Previous	Open	Maximum	Minimum	Last	Action
45.82	45.99	46	45.41	45.67	Sell
45.67	45.68	45.68	45.2	45.3	Buy
45.3	44.8	45.3	44.41	44.9	Buy
44.9	44.8	44.9	44.3	44.87	Sell
44.87	44.87	45.55	44.85	45.3	Buy
45.3	45.25	46	45.25	45.82	Buy
45.82	45.99	46.4	45.99	46.3	Buy
46.3	46.3	46.7	46	46.02	Buy
46.02	46.09	46.25	45.55	45.63	Sell

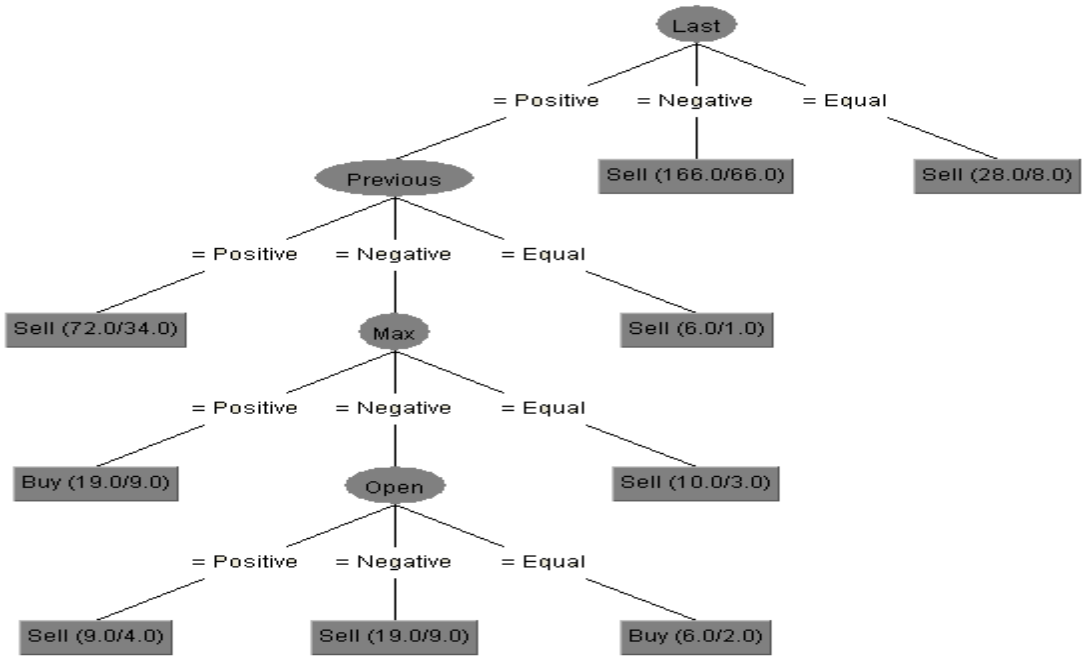
Previous	Open	Maximum	Minimum	Last	Action
Positive	Positive	Positive	Negative	Negative	Sell
Negative	Positive	Positive	Negative	Negative	Buy
Negative	Negative	Equal	Negative	Negative	Buy
Negative	Negative	Equal	Negative	Negative	Sell
Negative	Equal	Positive	Negative	Positive	Buy
Positive	Negative	Positive	Negative	Positive	Buy
Positive	Positive	Positive	Positive	Positive	Buy
Positive	Equal	Positive	Negative	Negative	Buy
Negative	Positive	Positive	Negative	Negative	Sell

Building the model :

The next step was to build the classification model using the decision tree technique. The decision tree technique was selected because the construction of decision tree classifiers does not require any domain knowledge, thus it is appropriate for exploratory knowledge discovery. Also, it can handle high dimensional data. Another benefit is that the steps of decision tree induction are simple and fast. Generally, decision tree accuracy is considered good. The decision tree method depends on using the information gain metric that determines the most useful attribute. When the decision tree model was applied on the data of the three companies using the WEKA software version 3.5, the root attribute for both PNB and NTPC company was the Open, while the attribute Last was the root for the decision tree of the BHEL company. As the process of building the tree goes on, all the remaining attributes were used to continue with this process. C4.5 algorithm was used in building the decision trees and the pruning technique was used in the C4.5 algorithm in order to reduce the size of the produced decision trees.

Result Analysis:

Bharat Heavy Electrical Limited



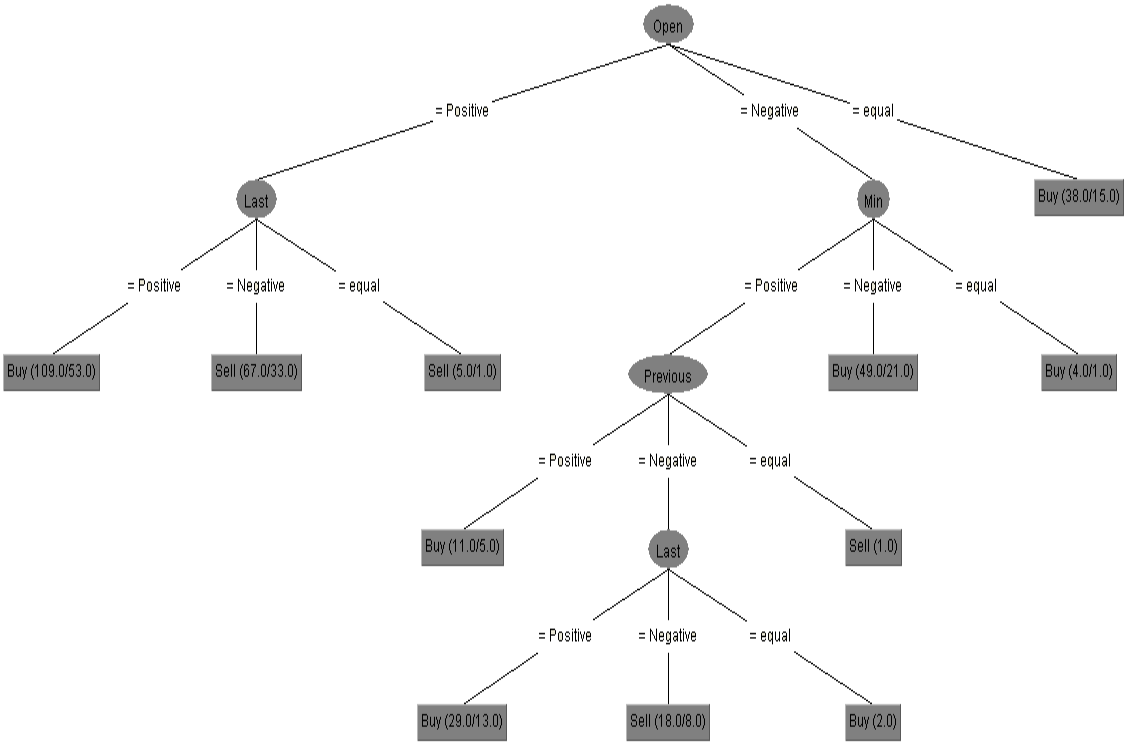
Holdout Percentage 66%

```
=== Summary ===

Correctly Classified Instances          91               53.2165 %
Incorrectly Classified Instances        80               46.7835 %
Kappa statistic                          0.8611
Mean absolute error                      0.0604
Root mean squared error                  0.4714
Relative absolute error                 46.9697 %
Root relative squared error             29.6116 %
Total Number of Instances              171
```

C.V 10

```
=== Summary ===

Correctly Classified Instances         265               52.7893 %
Incorrectly Classified Instances       237               47.2107 %
Kappa statistic                          0.3474
Mean absolute error                      0.0212
Root mean squared error                  0.4344
Relative absolute error                 46.9656 %
Root relative squared error             29.1047 %
Total Number of Instances              502
```

Punjab National Bank



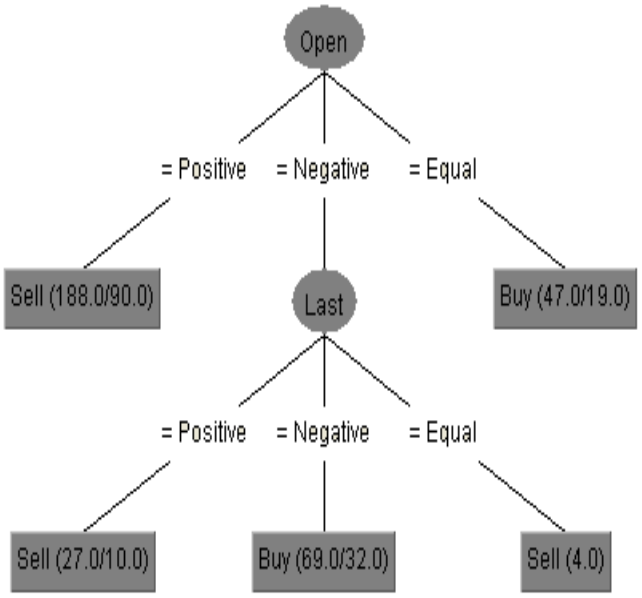
Holdout Percentage 66%

```
=== Summary ===

Correctly Classified Instances          83               48.6667 %
Incorrectly Classified Instances        87               52.3333 %
Kappa statistic                          0.3520
Mean absolute error                      0.4444
Root mean squared error                  0.4714
Relative absolute error                 36.9697 %
Root relative squared error             49.6116 %
Total Number of Instances              170
```

C.V 10

```
=== Summary ===

Correctly Classified Instances         237               47.4865 %
Incorrectly Classified Instances       262               53.5135 %
Kappa statistic                          0.8611
Mean absolute error                      0.0604
Root mean squared error                  0.4714
Relative absolute error                 46.9697 %
Root relative squared error             59.6116 %
Total Number of Instances              499
```

National Thermal Power Corporation



Holdout Percentage 66%

```
=== Summary ===

Correctly Classified Instances          94               54.9712 %
Incorrectly Classified Instances        80               45.0288 %
Kappa statistic                          0.8643
Mean absolute error                      0.0073
Root mean squared error                  0.6040
Relative absolute error                 13.8898 %
Root relative squared error             39.2678 %
Total Number of Instances              171
```

C.V 10

```
=== Summary ===

Correctly Classified Instances         264               52.5909 %
Incorrectly Classified Instances       238               47.4091 %
Kappa statistic                          0.5833
Mean absolute error                      0.0219
Root mean squared error                  0.1047
Relative absolute error                 41.6667 %
Root relative squared error             62.5497 %
Total Number of Instances              502
```

Open Price Prediction

Determining the next day opening price for the selected stock listed on NSE. Let's take the previsions history stocks for Punjab National Bank from April 2018 to May 2019.We obtain the data from the website of quandl by using the quandl library and process into comma separated vales(.csv) and followed by the process of data cleaning and transformation of data frame. The detailed view of the codes processed are given on the next page followed by the heatmap generated using seaborn library to see the corelation between the different attributes of the data frame. After that we use train_test_split to train the specific data based on history is loaded in DecisionTreeClassifier to predict the next day open price value for the selected PNB share of NSE.

- **Decision Trees (DTs)**

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. For instance, in the example below, decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model.

As with other classifiers, takes as input two arrays: an array X, sparse or dense, of size [n_samples, n_features] holding the training samples, and an array Y of integer values, size [n_samples], holding the class labels for the training samples.

- **train_test_split**

sklearn.model_selection.**train_test_split**(**arrays*, ***options*)

Split arrays or matrices into random train and test subsets. Quick utility that wraps input validation and next(Shuffle Split().split(X, y)) and application to input data into a single call for splitting (and optionally subsampling) data in a one-liner.

Code Window :

Using Python Spider (version 3.7) data derived from quandl library to fetch Stock data from Quandl website using. Further code proceed as follows-

```
import pandas as pd
import matplotlib.pyplot as plt
data = pd.read_csv("PNB.csv")
data.info()
data.head()
data.isnull().sum()
```

Importing Library and Loading Dataset

```
import seaborn as sns
plt.figure(1 , figsize = (17 , 8))
cor = sns.heatmap(data.corr(), annot = True)
```

Correalations between Data

```
x = data.loc[:,'High':'Turnover (Lacs)']
y = data.loc[:,'Open']
x.head()
y.head()
```

Selecting Attributes from Dataframe

```
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeRegressor
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size = 0.1,random_state = 0)
```

Training and Testing

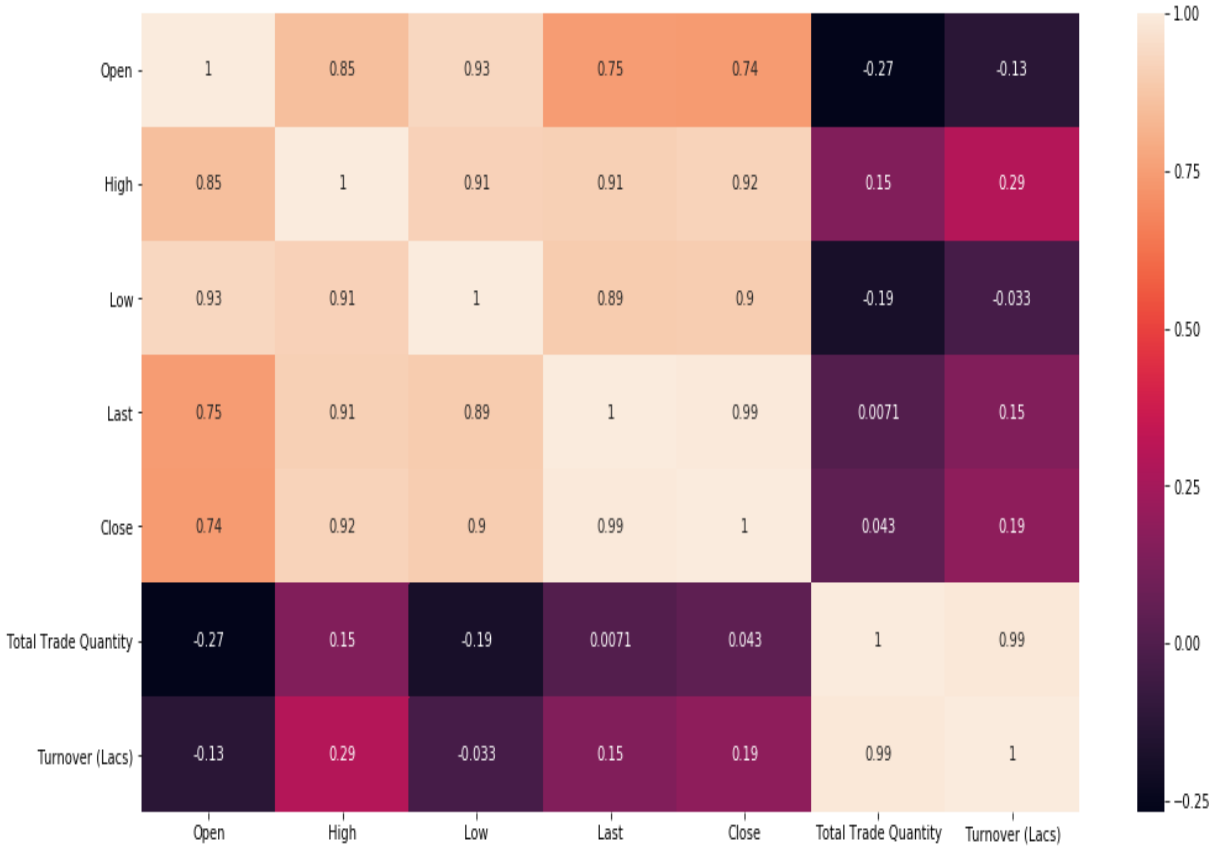
```
Classifier = DecisionTreeRegressor()
Classifier.fit(x_train,y_train)
```

DecisionTree Model

```
test = [[46.50 ,43.10 ,44.40, 44.45, 13889470.0 ,6219.22]]
prediction = Classifier.predict(test)
print(prediction)
```

Prediction on random Data

Heatmap Visualization : Correlation Between Data



Output:

```
runfile('C:/Users/HP/.spyder-py3/untitled0.py', wdir='C:/Users/HP/.spyder-py3')
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20 entries, 0 to 19
Data columns (total 8 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Date                 20 non-null     object
 1   Open                 20 non-null     float64
 2   High                 20 non-null     float64
 3   Low                  20 non-null     float64
 4   Last                 20 non-null     float64
 5   Close                20 non-null     float64
 6   Total Trade Quantity  20 non-null     float64
 7   Turnover (Lacs)      20 non-null     float64
dtypes: float64(7), object(1)
memory usage: 1.4+ KB
[45.95]
```

RESULTS AND DISCUSSION :

In order to evaluate the model, the WEKA software was used to calculate the accuracy of the classification model. Two evaluation methods were used, the K-Fold Cross Validation (K-CV) where K=10 folds and the percentage split method where 66% of the data was used for training and the remainder for testing. Evaluation methods used was C4.5 decision tree classification methods. Analysis result shows the accuracy of all the classifiers generated using both classification methods and evaluation methods. The resultant classification accuracy from the decision tree model is not very high for the training data used and it varies from one company to another. The reason for such a low accuracy is that the company's performance in the stock market is affected by internal financial factors such as; news about the company, financial reports, and the overall performance of the market. Also, external factors can affect the performance of the company in the market such as; political events and political decisions. Thus, it can be difficult to have a model that gives a high accuracy classification for all the companies at the same time as the performance of these companies differs.

In order to determine the approx. open price value of the share price we use python programming language to train(sklearn.model_selection & train_test_split) and build a decision tree classifier from (DecisionTreeRegressor) with gives the result. The actual value on that selected day was + 46.20 and our predicated value is 45.95 so we are very close at real value.

CONCLUSIONS AND FURTHER SCOPE

Use of decision tree classifier on the historical prices of the stocks to create decision rules that give buy or sell recommendations in the stock market. Such proposed model can be a helpful tool for the investors to take the right decision regarding their stocks based on the analysis of the historical prices of stocks in order to extract any predictive information from that historical data. The results for the proposed model were not perfect because many factors including but not limited to political events, general economic conditions, and investors' expectations influence stock market.

After developing our model, and to show its performance we would implement a risk strategy to check the profits we would gain based on our predictions and a few enhancements can be done and studied for our prediction model. One direction is to add extra technical indicators used in stock market. Another direction would be trying different time-frames for grouping our data. Finally, we could try to enhance the prediction of the exact price.

References

https://www.kaggle.com/datasets

https://www.datacamp.com/projects

https://www.cs.waikato.ac.nz/ml/weka/

https://youtu.be/l7R9NHqvl0Y (Data Mining Decision Tree J 48 session 3.4)