|  |  |  |
| --- | --- | --- |
| NAME | REGISTRATION NUMBER | SIGNATURE |
| KALUUMA MAHAD | 17/U/1405/EVE |  |
| KINTU ARNOLD | 17/U/1405/EVE |  |
| OLARA SAMUEL OBADIAH | 17/U/1405/EVE |  |
| LUBWAMA ISAAC | 17/U/1405/PS |  |

**Problem statement**

The problem with regards to the investors is that they do not know which are the most appropriate company stocks to invest their money. Stock market analysis and prediction help to solve this problem by using technical analysis, which uses statistical figures of a particular dataset to identify the trends in the stock market of a particular company and therefore predicting the company’s future profitability hence helping the investors in picking the right companies to invest in.

**Background to the dataset**

The dataset was obtained from a Yahoo! Finance, which is a traditional data source that contains structured data which can be used in deriving a data analytics pipeline to analyze and provide insight to the dataset.

The dataset to be used for the stock prediction and analysis is historical data, of a daily stock activity for a period of 10 years of Apple Inc. (AAPL) from 13th July 2009 to 13th July 2019 downloaded from [finance.yahoo.com](file:///C:\Users\LUBWAMA%20ISAAC\Downloads\finance.yahoo.com). This sums up to a total of 2519 records of the particular variables for each day excluding the weekends where stocks are not calculated.

Here is a brief sample of the dataset:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Date | Open | High | Low | Close | Adj Close | Volume |
| 13/07/2009 | 19.934286 | 20.334286 | 19.647142 | 20.334286 | 17.813419 | 120875300 |
| 14/07/2009 | 20.290001 | 20.454287 | 20.165714 | 20.324286 | 17.804667 | 86811900 |
| 15/07/2009 | 20.719999 | 21 | 20.617144 | 20.982857 | 18.381586 | 121396800 |
| 16/07/2009 | 20.822857 | 21.145714 | 20.795713 | 21.074286 | 18.461687 | 98392700 |
| 17/07/2009 | 21.297142 | 21.717142 | 21.232857 | 21.678572 | 18.991062 | 150538500 |
| 20/07/2009 | 21.895714 | 22.148571 | 21.555714 | 21.844286 | 19.136227 | 183881600 |

The dataset has seven **variables** that include Date, Open, High, Low, Close, Adj Close and Volume.

The columns Open and Close represent the starting and final prices at which the stock is traded on a particular day.

High and Low represent the maximum and minimum prices of a share on that day.

Close is the closing price of a stock on the day while adj stock is the closing price that has been evaluated to include any distributions and corporate actions e.g. dividends, stock splits, and new stock offerings.

The Volume is the number of shares that changed hands during a given day.

All the numeric variables above are calculated in the currency of US Dollars.

**Data Pipeline**

The data pipeline details a step by step route taken to build the system that will solve the stock prediction problem from the point of loading the data to the deployment of the system. Below is a diagrammatic illustration of the pipeline to be used.



Here is a breakdown of all the steps included in the data pipeline;

1. Loading of data:

Data that is in form of a dataset is to be read in form of a CSV file from its source as stated above observed to know the format of the major variables to be used e.g. text, numbers, audio or video.

1. Data wrangling

Data is passively observed to know whether it can be analyzed and put into use for the required objective. Various tasks might be carried out in the data to ensure its justification. Since the major variables to be used in the dataset are already in the numeric format then label encoding will not be necessary.

The missing value in the data, for example, will have to be looked for to ensure that the data is clean and can be used. Outlier removal might also be necessary for the dataset since the records are many and there might be a possibility of an outlier in any variable, a value that is way higher than all the other values in the column.

Feature scaling will be used to scale all features to have zero mean and unit variance. The data is split into training and testing datasets. The testing dataset will be used in the evaluation step so that performance metrics are completely independent of training and represent an unbiased estimate of actual performance.

We shall use the first 1024 samples as a training dataset for learning and the remaining 504 as a testing dataset for evaluation

1. Visualization

Data is represented as visual objects contained in graphics and also graphically in terms of tables, pie charts, bar graphs and many other kinds of graphs that help us analyze and derive a model that will help us.

1. Transformation of the machine learning algorithm

The prediction algorithms are used on both train and test data

1. Evaluation of the algorithm

The algorithms to be used are moving average, linear regression and support vector machine (**SVM**)

1. Deployment

After using the regression models to build the system then it should be made accessible to the investors who are supposed to use it. The model is then to be deployed as a web-based system that will help the investors log in, insert data from any company, run the data pipeline on the data and then be able to know the most appropriate stocks in which to invest in.