**STOCK MARKET PREDICTION**

**College of Computing and Information Sciences**

|  |  |  |
| --- | --- | --- |
| **NAME** | **REGISTRATION NUMBER** | **SIGNATURE** |
| KALUUMA MAHAD | 17/U/1405/EVE |  |
| KINTU ARNOLD | 17/U/5188/EVE |  |
| OLARA SAMUEL OBADIA | 17/U/9634/EVE |  |
| LUBWAMA ISAAC | 17/U/5739/PS |  |

**SUPERVISOR:****Mr. Grace Kamulegeya**

**Signature: …………………………………**

**Table of Contents**

[**1. CONCEPT PAPER 3**](#_Toc16471454)

[**1.1 Problem Statement 3**](#_Toc16471455)

[**1.2 Background of the Dataset 3**](#_Toc16471456)

[**1.3 Data Pipeline 4**](#_Toc16471457)

[**2. SOFTWARE REQUIREMENTS DOCUMENT 8**](#_Toc16471458)

[**2.1 Use Case Diagram 9**](#_Toc16471459)

[**3. SOFTWARE DESIGN DOCUMENT 12**](#_Toc16471460)

[**3.1 Data pipeline diagram 12**](#_Toc16471461)

**Table of Figures**

[Figure 1:Data Pipeline flowchart 5](#_Toc16473266)

[Figure 2: Use Case Diagram 10](#_Toc16473267)

[Figure 3:Data Pipeline Diagram 12](#_Toc16473268)

[Figure 4:Close Price trend over the years 16](#_Toc16473269)

[Figure 5:Close prices trend over the last month 16](#_Toc16473270)

[Figure 6:: line graph showing High, Low, Close and open Price 17](#_Toc16473271)

[Figure 7:Volume Trend over the last month 18](#_Toc16473272)

[Figure 8:Line graph showing Volume trend over the last month 19](#_Toc16473273)

[Figure 9:Normal Distribution of close prices over the years 19](#_Toc16473274)

# **CONCEPT PAPER**

## **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 of 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.



**Figure 1:Data Pipeline flowchart**

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. **Training Models**

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.

## **SOFTWARE REQUIREMENTS DOCUMENT**

***The users of the data pipeline***

For our pipeline, there will be two users (actors in our case diagram) that is the system administrator and the investor who is the client in this case.

The system administrator uses the pipeline to:

1. **Load data**

The data cannot be used without being loaded so it is important to load the data by reading it in form of a csv file from its source.

1. **Wrangle data**

Wrangling data consists of four main categories that all have their different justifications which include;

Feature scaling which helps to scale the spaces in the values of the variables and put them in the appropriate range to be worked on.

Outlier removal helps to remove distant values from the other values in the dataset and therefore match the values also in the same range.

Data cleaning which will include the splitting of data into two sectors that are train and test so that the model is trained and tested appropriately using the algorithms like linear regression

Linear encoding in this case will not be necessary because the values in the data set are already in numeric representation as they are supposed to be.

1. **Visualize the data**

The administrator uses the pipeline to represent the data in visual and graphical formats that can help the investors analyze the data themselves by using the various tables and graphs.

1. **Work algorithms and evaluate them**

Predictive algorithms are used on both the train and test data that were split in order to derive a model that will be trained and tested using the values in the dataset.

Both the system administrator and the investor use the pipeline for deployment. Deployment is important to the administrator because he runs the whole system and also mobilizes all the algorithms to be used in the system. He is also responsible for the storage of the data of the system.

The investors use only the deployment part of the pipeline because they don’t need to know what happens in the background and therefore, they are focused on the foreground hence they use it to see the stocks that were run and predicted in order to find out the most valuable stocks to purchase.

## **Use Case Diagram**



**Figure 2: Use Case Diagram**

**Explanation**

For our use case diagram the wrangle data use case includes data sampling, data cleaning, visualization of data and removal of outliers. Data sampling is a statistical analysis technique which uses data points from our data sets for manipulation and analysis to identify patterns in a larger data set, for data cleaning there is removal of inaccurate records from our data set to ensure accuracy and for data visualization there is encoding of data as visual objects contained in graphics. Outlier removal deals with removal of extreme values that deviate from other data observations. This is all accomplished within the data wrangle use case

The deployment use case is interacted with all the actors that is system administrator and investor. The system administrator manages and controls the operations of the deployment use case, the investor on the other hand interacts with it as a web interface for registration on to the system that is account creation which he uses to monitor the current and predicted stock prices thus choosing a more viable stock for investment.

Apply linear regression use case, here we use linear regression to model the relationship between a scalar response and an explanatory variable (one or more). Here we use the approach to fit a predictive model to an observed data set of values of the response and explanatory variables, after developing the model if additional values of the explanatory variables are collected without a response, the model can be used to make a prediction of the response. The algorithm is evaluated for accuracy and efficiency using the model evaluation metrics for regression.

### **SOFTWARE DESIGN DOCUMENT**

## **Data pipeline diagram**



**Figure 3:Data Pipeline Diagram**

Components of the Data pipeline include:

1. **Data source**

The Apple.inc dataset was downloaded as a CSV file from the yahoo finance website.

1. **Loading data**

After downloading the CSV file, we import the pandas as the library and load the CSV file into the pandas data frame. We use pandas because it offers data structures and operations for manipulating numerical tables and time series and handle large datasets.

1. **Preprocessing**

Data pre-processing involves transforming raw datainto an understandable format. It comprises of the following tasks.

1. **Feature Scaling**

Feature scaling helps to scale the spaces in the values of the variables and put them in the appropriate range to be worked on.

In feature scaling, we use the MinMaxScaler scaler algorithm which essentially shrinks the range such that the range is now between 0 and 1

A minmaxscaler is a method imported from the sklearn library under the preprocessing class.

Minmax scaler works better with small deviations and very sensitive to outliers in the data.

1. **Feature Extraction**

In feature extraction, we reduce the amount of data that must be processed by transforming the original data to a dataset with a reduced number of variables. We do this by dropping some of the unnecessary columns.

We use the pandas library for feature extraction because it offers data structures and operations for manipulating large datasets.

1. **Outliers removal**

**Outliers** refer to data points that are distant from other similar points. A histogram is used to visualize outliers in the data frame.

The concept of quartiles is implemented to remove outliers from the data frame. Using the NumPy package, we obtain low quantile and high quantile. Anything below the low quantile is an outlier and anything above the high quantile is an outlier. After we use the lambda method to filter into a new data frame.

1. **Structured Data.**

Structured data is highly-organized and formatted data that are easily represented for visualization.

1. **Visualization.**

* Data visualization uses [statistical graphics](https://en.wikipedia.org/wiki/Statistical_graphics), [plots](https://en.wikipedia.org/wiki/Plot_(graphics)), [information graphics](https://en.wikipedia.org/wiki/Infographic) and other tools to create a visual representation of the data. We use the following visualization components to represent our data.
* **Matplotlib**: it is generated using the Matplotlib Python Library. We use the library to plot graphs such as histograms, line plots, and bar charts
* **Line graphs** are easy to read and useful for making comparisons between different data sets. **Histograms** represent the frequency of the data occurring in the dataset and categories which are difficult to interpret in a tabular form which helps to visualize the distribution of the data. The matplot library allows access to several of Matplotlib's methods with less code. **Bar charts** are used torepresent data that shows changes over time, which helps people visualize trends in time series

1. **Data sampling**

Sampling involves splitting the dataset into two datasets, the training and testing datasets. The training dataset is used to train the prediction model the testing dataset is used to test the prediction model.

The training dataset takes 80 percent and the test dataset takes 20 percent of the original dataset. We use the scikit library which provides the Model selection tool. Inside the model selection library, we import the train\_test\_split class to split the dataset into various proportions.

We use the scikit-learn library because it is Simple and reusable in various contexts

1. **Training**

The training dataset is used as a learning set used to train the various machine learning algorithms.

1. **Evaluation**

* **Linear regression**

Linear Regression is a machine learning algorithm based on supervised learning.it performs a regression task by modeling a target prediction value (close prices) based on independent variables. We import the linear regression class from the scikit-learn library

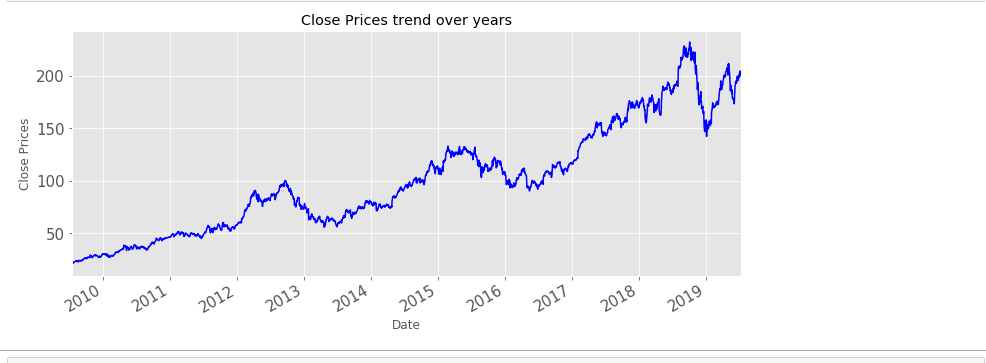
* **Simple Moving averages**

Simple Moving Average (SMA) takes the average of the target variable (close prices) over some set number of periods (years). Implementation of SMA uses the rolling mean function imported from the pandas library.

* **Support Vector regression Machine (SVM)**

It is a supervised machine learning classification algorithm. It's implemented using the scikit-learn python library.

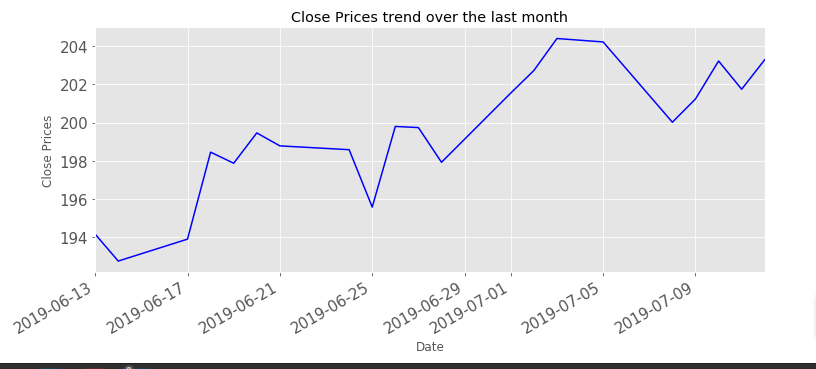
# **IMPLEMENTATION REPORT**



**Figure 4:Close Price trend over the years**

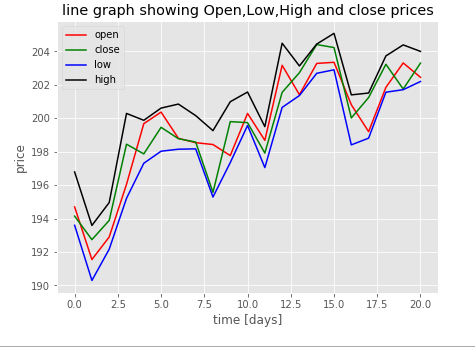
The line chart in figure 4 above, represents the historical close price trend that connects a series of data points with a continuous line over the years from 2009 to 2019. It gives a clear visualization of where the price of a stock has travelled over the given years. The closing price determines how well or poorly a stock performed, and this is very important not only to investors but also to financial institutions and other stakeholders.

After monitoring the closing prices, Investors and other stakeholders can base their decisions on this regarding their investment portfolios



**Figure 5:Close prices trend over the last month**

The line graph in *figure 5* above shows the trend of historical close prices over the last month from 2019-06-13 to 2019-07-12. This gives a clear view to investors and other stakeholders to determine performance over a specific time like a day as it is shown in the graph above.



**Figure 6:: line graph showing High, Low, Close and open Price**

**The** graph in figure 6 above represents four features which include:

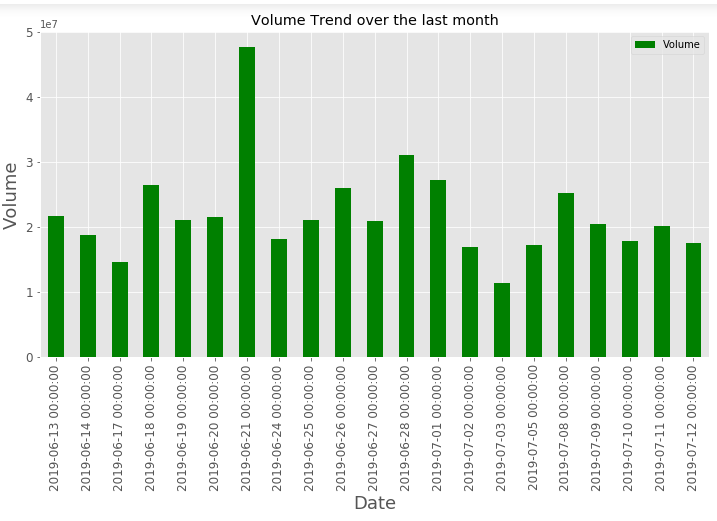
**Open and close prices:** Open price is represented with a red line while Close price is represented with a green line.

The opening price is the starting price where the stock trades at the opening bell of the stock market. The opening price is important for investors particularly for those interested in measuring short-term results for that day's trading activity.

The closing price is the final price at which a stock is traded on a given trading day. The closing price represents the most up-to-date [valuation](https://www.investopedia.com/terms/v/valuation.asp) of stock until trading commences again on the next trading day. The closing price of one day can be compared by investors to the previous closing price to measure market sentiment for a given stock over a trading day.

**Low and High prices**: Low price is represented with a blue line while the high price is represented with a black line. The low price represents the lowest price a particular stock has reached in the most recent period. The high price represents the highest price a particular stock has reached in the most recent period. A high stock price indicates that the business is doing well while a low price indicates that a business is doing poorly.

The low and high prices are important factors for investors and other stakeholders in determining a stock's current value and predicting future price movement.

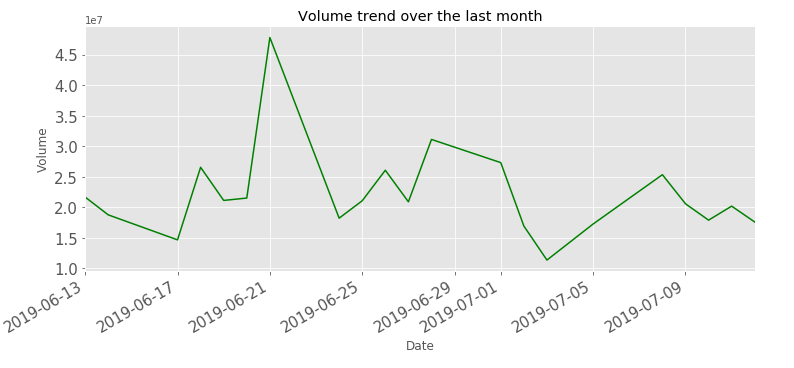


**Figure 7:Volume Trend over the last month**

The bar graph in figure 7 above shows the volume trend over the last month. Volume is a measure of how much of a given stock has been bought or sold over a particular span. Using volume to analyse stocks can boost profits and also reduce risk in business.

Decreasing volume and increasing prices show a lack of interest from investors and other stakeholders which is a warning of a potential change in the price direction.

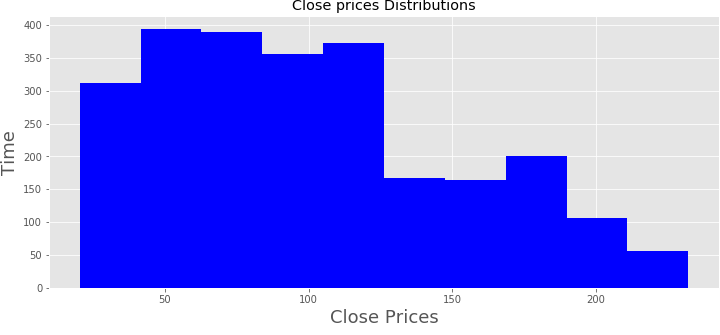
Large Volume and low prices indicate a high demand for stock or shares and the strength of any given price movement is measured primarily by the volume.



**Figure 8:Line graph showing Volume trend over the last month**

*The line graph in* figure 8 *above shows the trend of volume over the last month from 2019-06-13 to 2019-07-12.* Volume is used by investors and other stakeholders to confirm trends and chart patterns.

The trend in volume over time can be related to price trends to determine if a stock is gaining or losing momentum since the price is preceded by volume.



**Figure 9:Normal Distribution of close prices over the years**

The histogram in figure 9 above, provides a visual representation of close price distribution for the various trading days. It enables investors and other stakeholders to get information about a sample distribution without detailed statistical graphing or analysis.