**STOCK MARKET PREDICTION**

System Design Specification Document

COLLEGE OF COMPUTING AND INFORMATION SCIENCES

|  |  |
| --- | --- |
| NAME | REGISTRATION NO. |
| KALUUMA MAHAD | 17/U/1405/EVE |
| KINTU ARNOLD | 17/U/5188/EVE |
| sLUBWAMA ISAAC | 17/U/5739/PS |
| OLARA SAMUEL OBADIAH | 17/U/9634/EVE |

*Supervisor*

**Mr. Grace Kamulegeya**

**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.

* **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.

* **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.

* **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.