CHAPTER 3

RESEARCH METHODOLOGY

3.1. Literature of Research

A great deal of research has been conducted, some of which is obviously connected to the topic at hand. Before deciding to apply the model, the researcher decides to investigate some scholarly literature on forecasting stock prices using machine learning and deep learning. The researcher estimates the model to be utilized and the evaluation to be applied in this stage.

3.2. Data Preprocessing and Analysis

3.2.1. Data Collection

The material was collected from https://finance.yahoo.com (Yahoo Finance). The dataset utilized relies on three different companies:

1) BAC (Bank of America Corporation)
2) HDB (HDFC Bank Limited)
3) RY (Royal Bank of Canada)

3.2.2. Data Selection

Data selection is a phase that involves picking and removing usable and unnecessary data from the source. The data variable provided by the firm is as follows:

1) Date : The day on which trading on a newly issued stock begins.
2) Open : The first cost was paid on a day.
3) High : The biggest cost was paid on a day.
4) Low : The smallest cost was paid on a day.
5) Close : A closing cost was paid on a day.
6) Adj Close : A closing costs after dividends and stock splits have been deducted.
7) Volume : The total amount of deals that day.
8) HL_PCT : The percentages of the biggest price and smallest price for each day.
9) PCT_change : The percentages of first price and closing price for each day.

10) delta_open_close_day_before_% : The disparity between the closing price and the open next day's first price.

11) Open:30 days rolling : The mean of the preceding 30 days open price.

12) High:30 days rolling : A mean income of the preceding 30 days' high price.

13) Low:30 days rolling : A mean income of the preceding 30 days' low price.

14) Close:30 days rolling : A mean income of the preceding 30 days' close price.

15) Adj Close:30 days rolling : A mean income of the preceding 30 days' adj close price.

16) Volume:30 days rolling : A mean income of the preceding 30 days' volume price.

17) Label : Forecast out results

Only Adj Close, Volume, HL_PCT, and PCT_change were utilized from all variables in the study given.

3.2.3. Data Visualization

A) Heatmap Visualization

Figure 3.1 shows two-dimensional visual information that uses colors to represent the individual values in a matrix. The provided heatmap data will be further processed, and data visualization in the form of a heatmap and visualization of data distribution will be created to produce a clear representation. Visualization is used to create a clear image of the data so that additional analysis and action may be taken. A heatmap allows us to examine the connection between data and how significant it is. Heatmap visualization can be utilized with the seaborn library.

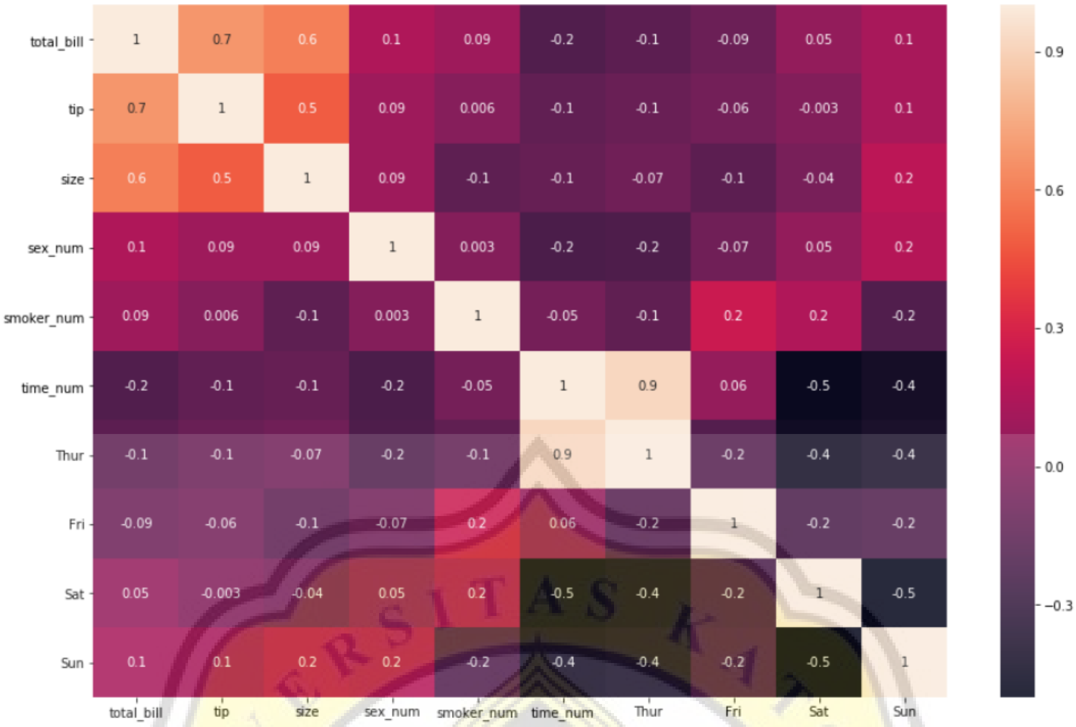

Figure 3.1 Heatmap

A) *Box Plot Visualization*

Figure 3.2 shows a summary of a set of data values. It can also be used to compare the distribution of data across data sets by generating boxplots for each one. The x-axis shows the data to be shown, while the y-axis shows the frequency distribution.

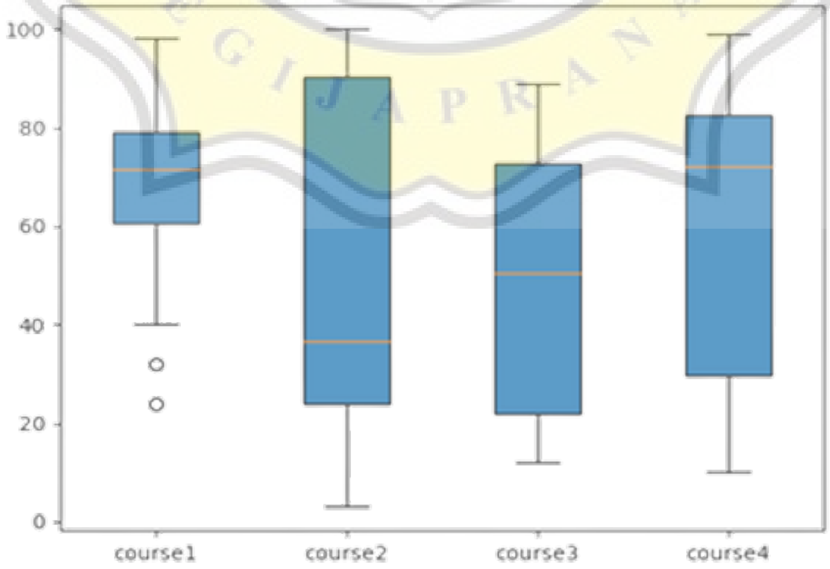

Figure 3.2 Box Chart

B) *Bar Chart Visualization*

Figure 3.3 shows data using rectangular bars whose length and height are proportional to the values they represent. Bar graphs can be drawn horizontally or vertically. A bar chart is used to show comparisons between several categories. The one axis of the graph illustrates the specific categories being compared, while the other shows the measured values linked with those categories. To produce bar charts, use the matplotlib application.

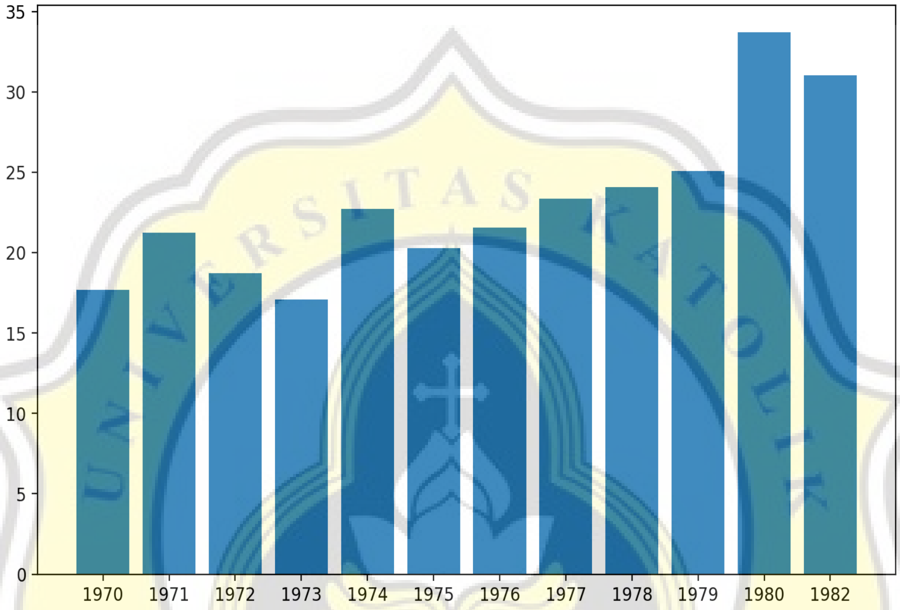

Figure 3.3 Bar Plot

3.2.4. Split Data

Before the data is divided into several training data and test data, the researcher determines how many adj close rows will be shifted to predict future stock prices, data that is not included in this number will be used as training data. After that, the researcher conducted 5 experiments. For the first experiment, researchers used 80% of the data that was not included in the shift rate. For the second experiment, researchers used 60% of the data that was not included in the shift rate. For the third experiment, researchers used 40% of the data that was not included in the shift rate. For the fourth experiment, researchers used 20% of the data that was not included in the shift rate. For the last experiment, researchers used 50% of the data that was not included in the shift rate.

3.2.5. Feature Scaling

Many machine learning or deep learning algorithms that measure convenience using Euclidean distance will fail to give satisfactory recognition for smaller features. Scaling crucial data in deep learning or machine learning models can be more effective. There are several ways for doing feature scaling, normalization, and standardization.

A) Standardization

Standardization ensures that all characteristics are planned in reference to an average value with a one-standard-deviation standard deviation. By decreasing the mean of each observation divided by the standard deviation it will reach standardization (1).

$$X_{new} = \frac{X - X_{mean}}{\sigma} \tag{1}$$

B) Normalization

Each value in a feature is reduced by the feature's minimum value, then divided by the range of values or the maximum value, and the decreased minimum value of the feature produces a new normalized value between 0 and 1 or -1 and 1 (2).

$$X_{new} = \frac{X_{old} - X_{min}}{X_{max} - X_{min}} \tag{2}$$

3.2.6. Models Evaluation

MAE, RMSE, and MAPE will be used to assess system performance. (1) MAE is defined as a mean absolute difference between measured and predicted values. (2) The RMSE is then computed by squaring the error (prediction) divided by the amount of data (= average), and finally rooted. (3) While MAPE is the absolute percentage of average error.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - x| \tag{1}$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}} \tag{2}$$

$$MAPE = \sum_{t=1}^{n}\left|\frac{y_i - \hat{y}_i}{\hat{y}_i}\right| x\ 100\% \tag{3}$$

3.3. Algorithms

The researcher employs a regression approach with five models such as LSTM, CNN, linear regression, lasso regression, and LSSVM. These model's performance will be assessed using numerous computations explained in the preceding section. Linear regression in quantitative research is to forecast the connection between variables X and Y. The reason the authors use linear regression is that despite its limitations, such as the fact that real data rarely shows a clear relationship between the dependent and independent variables, this linear regression can predict future values, making it useful for forecasting stock prices, sales, etc.

Lasso regression is an extension of linear regression in that the regularization parameters are multiplied by the sum of the absolute values of the weights and applied to the linear regression loss function (ordinary least squares). The gain using Lasso regression versus linear least squares regression is found in the bias-variance trade-off, which states that as alpha grows, the flexibility of the lasso regression's fit diminishes, resulting in a drop in variance but an increase in bias. The reasons why researchers utilize lasso regression to assist prevent overfitting since it has the capacity to set the coefficients for characteristics that are regarded unappealing to 0, hence lowering the model's complexity.

LSSVM is a variant of the standard SVM that employs equality constraints rather than inequality constraints and a squared loss function rather than the -insensitive loss function. LSTM is an RNN modification. The LSTM has three gates: A forget gate defines what information from the previous cell is to be disregarded, an input gate determines which data is aligned with the in-use cell, and an output gate determines which information should be transferred to the next hidden layer. The LSTM algorithm has the benefit of accepting input of varying durations. This capability is extremely handy for creating forecasting models with LSTM. The architecture of LSTM can be seen in Figure 3.1.

CNN is a subset of deep-neural-network. CNN is a technique for predicting future occurrences that uses a 1D array as input. CNN also has the benefit of being stronger at pattern recognition, more accurate at feature extraction, and faster at training. So that is the reason this research uses CNN. The architecture of CNN can be seen in Figure 3.2.

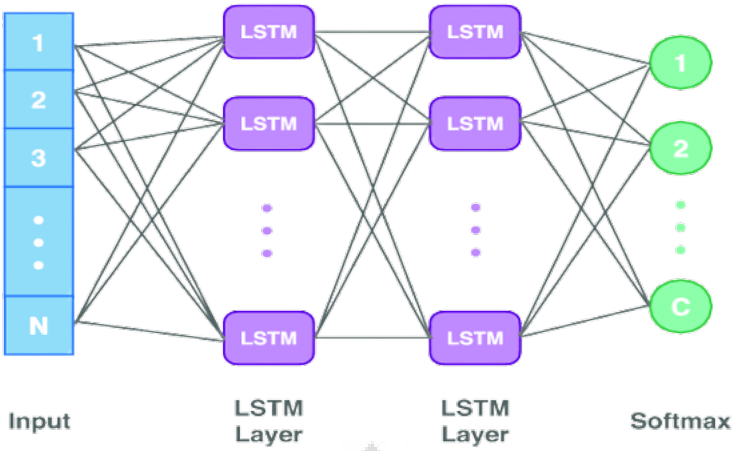

Figure 3.1 LSTM Architecture

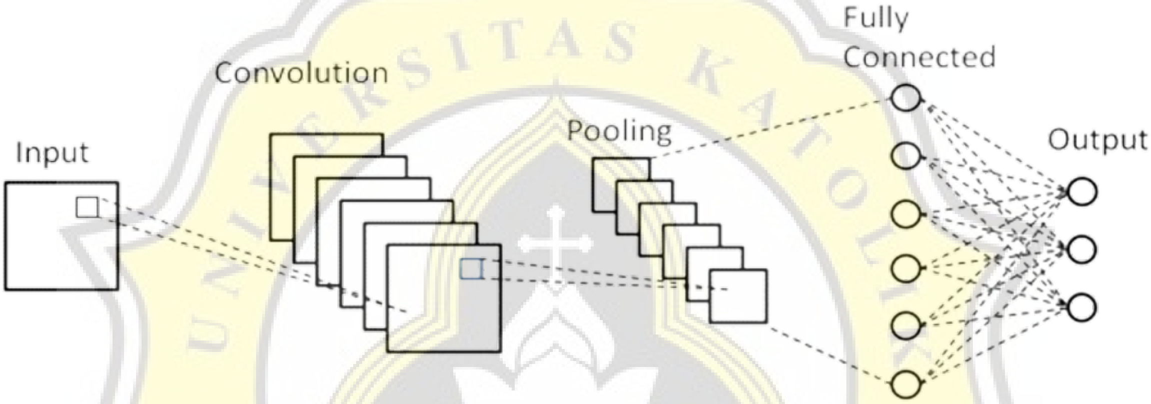

Figure 3.2 CNN Architecture