# CHAPTER 3
# RESEARCH METHODOLOGY

## 3.1. Data Collection

This study will be using the dataset of ADRO from 01/01/2010 until 31/12/2020 [15]. The dataset contains information such as Date, Open, High, Low, Close, Adjusted Close and Volume or OHLC data. The three coal stocks  are chosen as there are no other sectoral dataset that can be used to be downloaded and analyzed. The data, which starts from the First of January 2010 and ends on 31st of December 2020, is considered relevant enough to be investigated because its duration will be adequate to be a good sample despite some minor crash that happened in Jakarta Stock Exchange (JKSE). Also, ADRO is chosen because there has not been much study on Indonesia’s Stock Market and this paper will be a reference in the future for those having interest in developing a more sophisticated machine learning model to predict Stock Price or Stock Market Indices.

## 3.2. Analysis Objective

The models created from the OHLC Data from the datasets (ADRO) will be used to create machine learning models and those models will be evaluated with some performance metric.

## 3.3. Data Preparation

The dataset consists of a few columns namely Date,Open,High,Low,Close,Adjusted Close and Volume, to simplify the columns for analysis, they are expanded to be Open,High,Close,Volume,Adjusted Close,Year,Month,Day. Other than simplifying the columns, the data is also cleaned from null values.

These are the explanation for the column :

1. Date : Date of the transaction of the stock
2. Open : The opening price of stock in the particular day
3. High : The highest price of stock in the particular day
4. Low : The lowest price of stock in the particular day
5. Close : The close price of stock in the particular day

6. Adjusted Close : The close price after dividend payment in the particular day
7. Volume : Volume transaction in the particular day.

**Table 3.1 Dataset from YahooFinance**

| | Date | Open | High | Low | Close | Adjusted Close | Volume |
|---|---|---|---|---|---|---|---|
| 0 | 2010-01-01 | 1730 | 2075 | 1730 | 1890 | 955.029 | 2332183503 |
| 1 | 2010-02-01 | 1850 | 1960 | 1760 | 1830 | 924.711 | 2010715500 |
| 2 | 2010-03-01 | 1840 | 1980 | 1820 | 1960 | 990.40 | 2018508000 |
| 3 | 2010-04-01 | 2025 | 2250 | 1970 | 2200 | 1111.674 | 2097459500 |
| 4 | 2010-05-01 | 2125 | 2175 | 1700 | 2000 | 1010.61 | 1990892503 |

This are the visualization of a data set taken from YahooFinance and it is imported to Google Collab.After that data will be cleaned and normalized with Min-Max normalization whereby it helps the user to read the data easily. This is the formula of Min-Max normalization:

$$X_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

## 3.4. Machine Learning Models

### *3.4.1. ANN Flowchart*

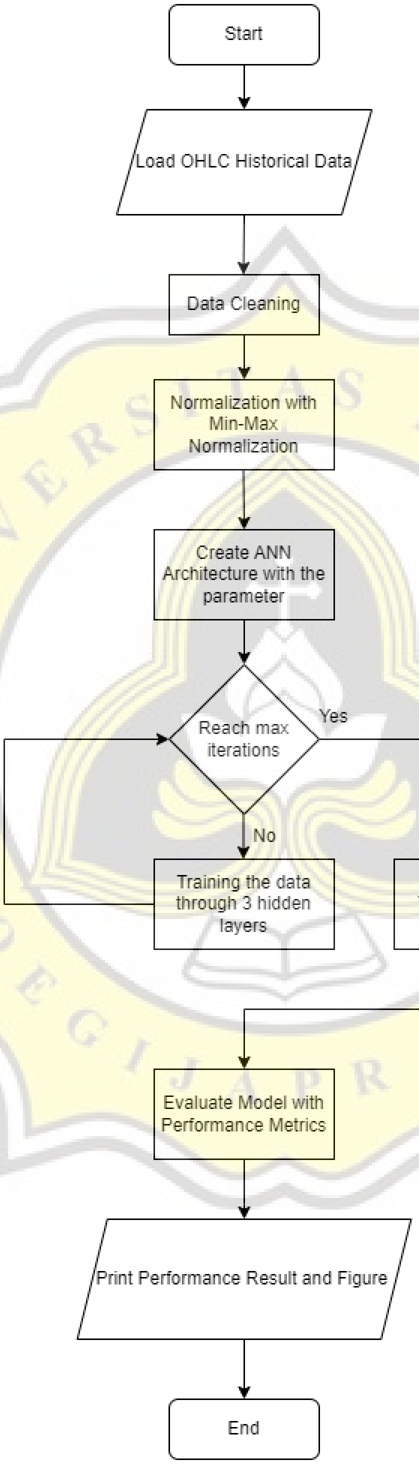
Start
Load OHLC Historical Data
Data Cleaning
Normalization with Min-Max Normalization
Create ANN Architecture with the parameter
Reach max iterations
Yes
No
Training the data through 3 hidden layers
Test the data set
Evaluate Model with Performance Metrics
Print Performance Result and Figure
End

**Figure 3.1 Research scheme to use ANN machine learning models on OHLC Historical data.**

This is the diagram of how ANN model works.First it imports OHLC Historical data taken from Yahoo Finance.The next step in preparation process is to import libraries and check the dataset.And then the dataset will be go through 3 hidden layers and 1 output layers.Data set will be divided into validation test and training test and then these data will be evaluated with performance metrics.The next part is to analyze the result from the evaluation.

### *3.4.2. Random Forest Flowchart*

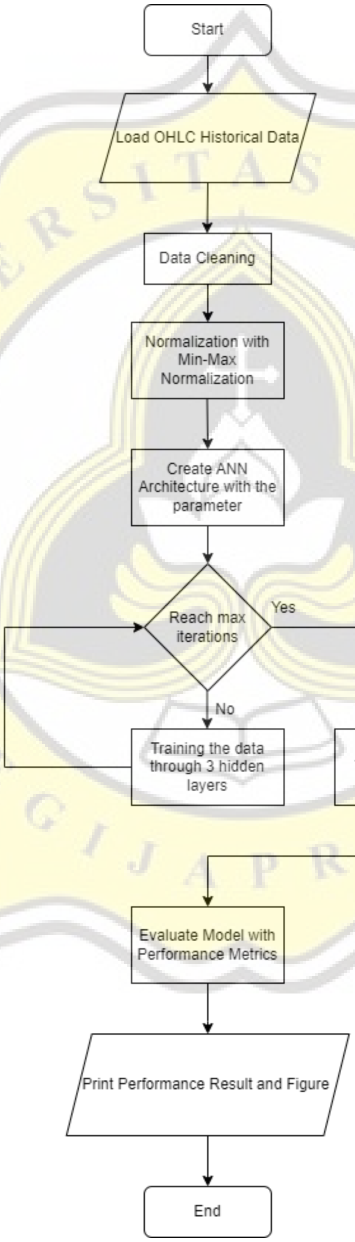
Start
Load OHLC Historical Data
Data Cleaning
Normalization with Min-Max Normalization
Create ANN Architecture with the parameter
Reach max iterations
Yes
No
Training the data through 3 hidden layers
Test the data set
Evaluate Model with Performance Metrics
Print Performance Result and Figure
End

**Figure 3.2 Research scheme to use Random Forest machine learning models on OHLC Historical data.**

This is the diagram of how Random Forest model works.First it imports OHLC Historical data taken from Yahoo Finance.The next step in preparation process is to import libraries and check the dataset.And then the dataset will be split into 80% for training and 20% for validation test with 1500 number of forest and the number of columns to each decision tree in form of max_depth(10) and random state with value of 1.The result will be averaging for regression to get the final decision and will be evaluated will performance metrics.The next part is to analyze the result from the evaluation.

### *3.4.3. ARIMA Model Flowchart*

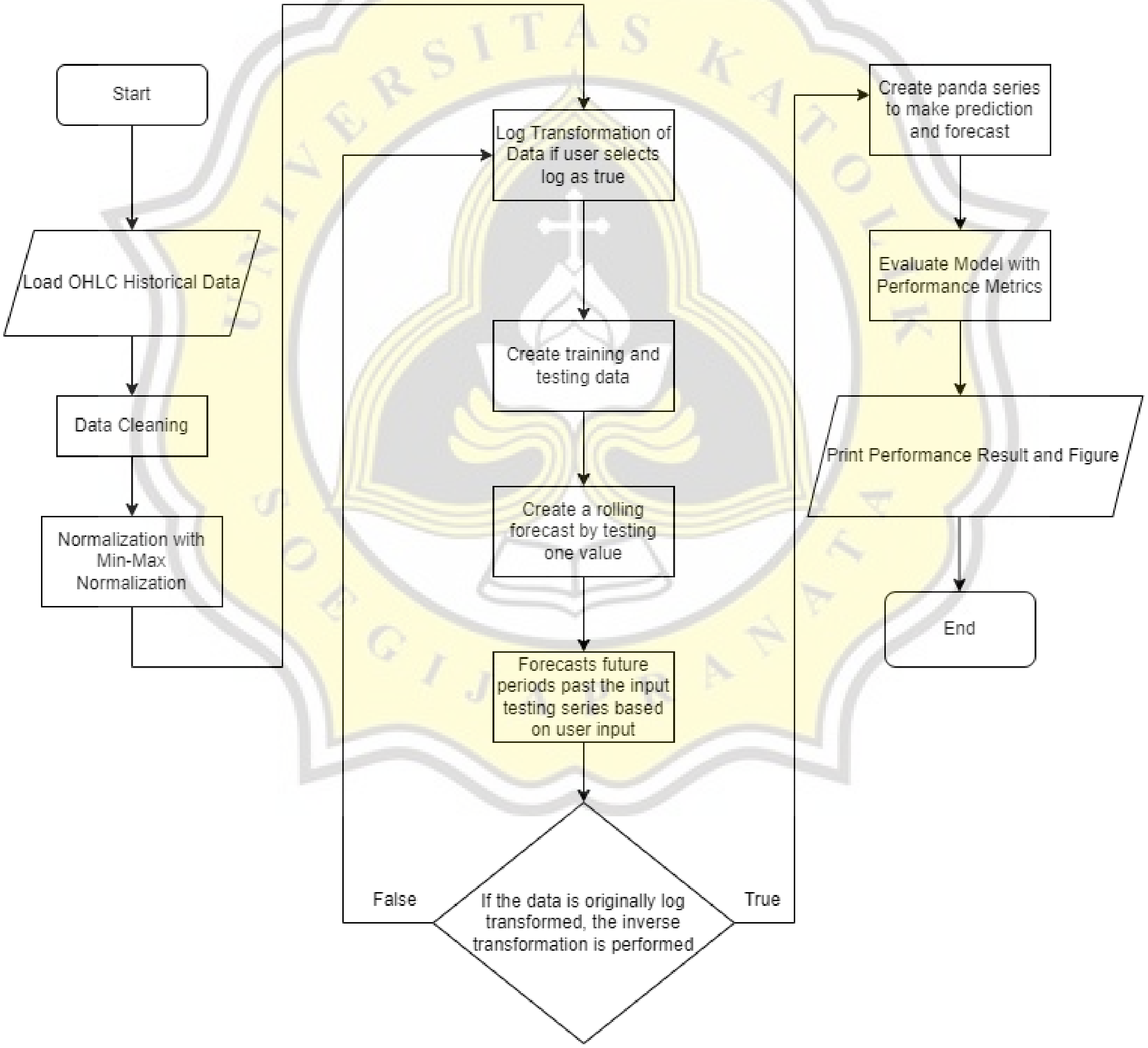
Start
Load OHLC Historical Data
Data Cleaning
Normalization with Min-Max Normalization
Log Transformation of Data if user selects log as true
Create training and testing data
Create a rolling forecast by testing one value
Forecasts future periods past the input testing series based on user input
If the data is originally log transformed, the inverse transformation is performed
False
True
Create panda series to make prediction and forecast
Evaluate Model with Performance Metrics
Print Performance Result and Figure
End

**Figure 3.3 Research scheme to use ARIMA machine learning models on OHLC Historical data.**

This is the diagram of how ARIMA model works.First it imports OHLC Historical data taken from Yahoo Finance.The next step in preparation process is to import libraries and check the dataset.And then the dataset will be split into training set and validation test.The forecast will rely upon yesterday closing price to forecast the next data.The result will be evaluated with performance metrics.The next part is to analyze the result from the evaluation.

### *3.4.4. Artificial Neural Network (ANN)*

Ayala et al. [5] stated that Artificial Neural Network are computational model that classify information in nodes and each connection between each nodes of different layer have different weight.The adjustment of weight in each connections is the fundamental for this model to minimizing a loss function that comes from data prediction error from the calculation.The output layer depends on the expected output whereas the total neurons and dimensionality in the input layer will be the deciding factor.

The architecture uses 4 hidden layers that consists of 3 hidden layers and 1 output layer.Those 3 hidden layers use relu function for activation and linear function for the output layer.The first hidden layer will execute 128 units and processed into 64 and then becomes 32.The final output will be 1 unit.The author uses 200 batch size and 200 epochs with 0,3 validation split.

### *3.4.5. Random Forest*

Random Forest is a methodology that works to classify and do regression task.The classification comes from prediction from a set of trees with majority vote in the case of classification.The final prediction that comes from the average of the tree's predictions will represent the regression task.

The author uses preprocessing.LabelEncoder() to transform the data and put them into categories.RandomForestRegressor are the algorithm being used with hyperparameter such as number of forest in form of n_estimator (1000), the number of columns to each decision tree in form of max_depth (10) and random state with value of 0.

### *3.4.6. ARIMA Model*

ARIMA model is a modification model from autoregressive moving average (ARMA).There are three parameters in ARIMA which are :

1. p as the lag order which represents the number of lag observations in the model.
2. D as the degree of differencing which represents the differences of the number of times the raw observation.
3. Q as the order of the moving average which represents the size of the moving average window.

Arima Model equation are

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} \varepsilon_t + \emptyset_1 \epsilon_{t-1} + \emptyset_2 \epsilon_{t-2} + .. + \emptyset_q \epsilon_{t-q} \#(1)$$

The regression model is lagged values of y, until p-th time in the past,as predictors.C is a constant, φs are parameters, p is the number of lagged observations in the model and ε is white noise at time T.The hyperparameters in the ARIMA model are p = 2 , d =1 , and q =1 with data split = 0,7.Future period = 12 which determine the future value in terms of seasonality.

## 3.5. Evaluation Methods

To measure the prediction performance of machine learning models, the author will use several performance metrics, specifically Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Mean Squared Error (MSE).

1. Mean Absolute Error (MAE) is performance metrics that measure absolutes differences between the actual observation and predicted value.This is the formula of MAE :

   MAE = $\frac{1}{n} \sum_{t=1}^{n} |y_t - \bar{y}_t| \#(2)$

   Because all the single differences are considered average equally,it is considered to have linear score.The effect of this is even for the larger error will contribute linearly to the total error.This becomes the drawback for this method.

2. Root Mean Squared Error (RMSE) is a quadratic scoring rules.It means that the total average of magnitude of error will be calculated.Because of the nature,the outlier will

bring more impact to the average and affect the result.It is defined by the following formula :

RMSE = $\sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y}_t)^2}$ #(3)

3. Mean Absolute Percentage Error (MAPE) represent how far in terms of average of the prediction from the real values.It is different from MAE because MAE represent the average magnitude of error in a model.The issue of MAPE is if the denominator is zero,it will be undefined for the result.It is defined the formula as :

   MAPE = $\frac{1}{n}\sum_{t=1}^{n}\frac{y_t - \overline{y_t}}{y_t}$ #(4)

4. Mean Squared Error (MSE) works by squaring the difference of the model prediction with the actual data and average it out throughout the whole dataset.This function works by ensuring the trained model will not produce outlier predictions with critical error.It is defined the formula as :

   MSE = $\frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y}_t)^2$ #(5)