CHAPTER 3

METHODOLOGY

This chapter explains methodological approaches that will be used to analyze Malaysian GDP and unemployment rates. The organization of the chapter for this study will be in a way that presents the Data Science Project Life Cycle, data sources and data collection methods, data pre-processing, and the analytical tools to be employed in the current study.

3.1 Data Science Project Life Cycle

The Data Science Project Life Cycle encompasses the following stages: Some of the tasks that have been highlighted include feature extraction which is Data Collection, the process of initializing the data and making it ready for analysis, this is Data Pre-processing, the initial examination of the data with the aim of summarizing it, this is Exploratory Data Analysis (EDA), creating the model is Model Building, testing the model is Evaluation and using the model for prediction is Deployment. Each of the phases plays a crucial role in ensuring the compilation of the right data and the achievement of result accuracy.

1. **Data Collection:** Gathering of data to be analysed is the initial step that is taken in the process. This entails gathering data from website data. gov. my.
2. **Data Pre-processing:** This is the process of cleansing and preparing the collected data for analysis and to make it in a usable form. It involves attributes with missing values, scaling data, and generating new attributes.
3. **Exploratory Data Analysis (EDA):** EDA can be defined as the process of describing the data in simple words with the help of graphs. The Company learns from it to understand patterns of data components, as well as outliers, and even come up with hypotheses.
4. **Model Building:** In the following stage, statistical tools and machine learning techniques will apply on the data to forecast.

5. **Evaluation:** To be sure that the results are accurate and relevant, the models will be evaluated based on other performance measures also.
6. **Deployment:** The final phase refers to the use of developed and implemented models in the live setting in which the models would be beneficial for insights.

3.2 Data Sources and Collection Methods

This study will utilise a comprehensive set of datasets from Data.gov.my and other relevant sources. The primary datasets include:

1. **Monthly Unemployment Data** (lfs_month.csv)
    - **Source**: Data.gov.my
    - **Description**: Provides monthly data on overall unemployment rates in Malaysia.
    - **Columns:** date, lf, lf_employed, lf_unemployed, lf_outside, p_rate, ep_ratio, u_rate.
2. **Seasonally Adjusted Monthly Unemployment Data** (lfs_month_sa.csv)
    - **Source**: Data.gov.my
    - **Description**: Provides seasonally adjusted unemployment data to remove seasonal effects.
    - **Columns:** date, lf, lf_employed, lf_unemployed, p_rate, u_rate.
3. **Monthly Youth Unemployment Data** (lfs_month_youth.csv)
    - **Source**: Data.gov.my
    - **Description**: Focuses on unemployment rates among the youth.
    - **Columns:** date, unemployed_15_24, u_rate_15_24, unemployed_15_30, u_rate_15_30.
4. **Monthly Unemployment Duration Data** (lfs_month_duration.csv)
    - **Source**: Data.gov.my
    - **Description**: Contains data on the duration of unemployment periods.
    - **Columns:** date, unemployed, unemployed_active, unemployed_inactive, unemployed_active_3mo, unemployed_active_6mo, unemployed_active_12mo, unemployed_active_long.
5. **Monthly Unemployment Status Data** (lfs_month_status.csv)

- **Source**: Data.gov.my
- **Description**: Provides detailed status information of the unemployed.
- **Columns:** date, Value type, employed, employed_employer, employed_employee, employed_own_account, employed_unpaid_family.

6. **Annual Real GDP Data** (gdp_gni_annual_real.csv)
    - **Source**: Data.gov.my and World Bank
    - **Description**: Contains annual real GDP data.
    - **Columns:** date, series, gdp, gni, gdp_capita, gni_capita.
7. **Annual Nominal GDP Data** (gdp_gni_annual_nominal.csv)
    - **Source**: Data.gov.my and World Bank
    - **Description**: Contains annual nominal GDP data.
    - **Columns:** date, series, gdp, gni, gdp_capita,gni_capita.
8. **GDP Lookup Data** (gdp_lookup.csv)
    - **Source**: Data.gov.my
    - **Description**: Provides additional context or conversion between nominal and real GDP.
    - **Columns:** method, code, variable_en, variable_bm.
9. **GDP Nominal Supply Data** (gdp_annual_nominal_supply.csv)
    - **Source**: Data.gov.my
    - **Description**: Details GDP from the supply side.
    - **Columns:** series, sector, value.
10. **Malaysia Economic Indicator Data** (malaysia_economic_indicator.csv)
    - **Source**: Data.gov.my
    - **Description**: Includes various economic indicators for Malaysia.
    - Columns: leading, coincident, lagging, leading_diffusion, coincident_diffusion.

3.3 Data Pre-processing

Data pre-processing is an important phase of the data science project life cycle which involves cleaning of the collected raw data into a convenient form for analysis. The pre-processing steps undertaken in this study will include:

1. **Data Cleaning**
    - **Handling Missing Values**: Datasets with missing values will be identified and handled by either using suitable methods of removing the records if they are deemed unnecessary for the analysis.
    - **Outlier Detection and Treatment**: Outliers will be identified and resolved to prevent them from skewing the analysis results.
    - **Duplicate Removal**: Duplicate records will be identified and removed to ensure each data entry is unique.

2. **Data Transformation**
    - **Date Conversion:** Date columns will be converted to datetime format to facilitate time series analysis.
    - **Resampling and Interpolation:** Annual GDP data will be resampled to a monthly frequency to align with the monthly unemployment data. Missing values will be interpolated.

3. **Feature Engineering**
    - **New Features**: New features will be created to further enhance the analysis. For example, seasonally adjusted unemployment rates will be used to remove seasonal effects.
    - **Normalisation and Scaling**: The data will be normalised and scaled to ensure that all features contribute equally to the analysis.
4. **Data Merging**
    - **Combining Datasets**: The datasets will be merged based on the common date index to form a comprehensive dataset for analysis.

With these pre-processing steps, the data will be ready for the next stages of analysis, ensuring that it is clean, suitable, and consistent for the descriptive and time series analysis.

3.4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis or EDA is the process by which the data is explained, the important characteristics, which are; abnormalities and relationship of the data are defined. EDA will entail the use of mean, median, mode, standard deviation, density plots, histograms, box and whisker plots and time plots. It is beneficial at this stage to create hypotheses and decide which values need to be additionally analyzed.

1. **Descriptive Statistics:** These variables are; mean, median, standard deviation, variance will be computed in order to give summary statistics.
2. **Data Visualization:** Comparisons of the relative frequencies of the variables too, will be done by histograms and box-plots of the unemployment rates or GDP.
3. **Time Series Plots:** Such line graphs as time series will be developed to show unemployment rates and GDP at different time periods.

3.5 Time Series Analysis

Time series analysis is a statistical method of analysing data that is measured at different points in time. This research will utilise time series decomposition, ARIMA, and Prophet models to aggregate and forecast unemployment rates relative to GDP as well as other variables.

1. **Time Series Decomposition**

The historical unemployment rate time series will be separated into trend cycle, seasonal and irregular movements.

2. **ARIMA Model**
    - **Model Selection:** The future unemployment rates are going to be predicted with the help of the ARIMA that is an abbreviation for AutoRegressive Integrated Moving Average. Since it only deals with one time variable this paper deemed appropriate is the ARIMA model, The model is flexible enough to include the

other components of the time series data such as the AR, the differencing and the MA.

- **Parameter Estimation:** Choice of specific values of p, d and q in context of the ARIMA model will be made by plotting the ACF/PACF diagrams or by the Grid search or any frequent by applying the method included in the pm darima library.
- **Model Fitting:** In the next step, the current data pertaining to the unemployment rate will be discussed and will be analysed and/or forecasted using the ARIMA model for better understanding of the pattern of the data set.
- **Model Evaluation:** The effectiveness of the forecast acquired from the developed ARIMA model shall be assessed making use of Mean Absolute Error (MAE), Mean Squared Error (MSE). However, the diagnostic checks relating to residuals will also be done to check for White noise.
- **Forecasting:** The future unemployment rates will be forecasted through the application of the ARIMA model. When explaining the model to the audience, the forecast will be plotted along with records to show the output of the model.

3. **Prophet Model**
    - **Model Introduction:** Prophet is an open source forecasting tool created by Facebook for time series data with strong seasonal patterns and gaps. Prophet will be used as the second model to forecast the data in order to compare the results with the ARIMA model.
    - **Data Preparation:** The data will be transformed in a way that Prophet requires the time series to have columns named ‘ds’ (date) and ‘y’ (value).
    - **Model Fitting:** The Prophet model will be estimated on the unemployment rate data.
    - **Forecasting:** Forecasts will be made using the Prophet model. The next data frames will be generated for the forecast horizon and the predictions will be displayed.
    - **Model Evaluation:** The result of the Prophet model will be evaluated by the same criteria as in the case of the ARIMA model (MAE, MSE, RMSE). Moreover, the decomposition of the model into the constituent parts, which

include the trend, seasonal, and holiday effects, will be done to establish the percentage of impact of each part towards the forecast.

Therefore, using these time series analysis techniques, the study seeks to forecast future unemployment rates given past GDP and other economic variables. The findings from both ARIMA and Prophet models will be compared to identify the best method of forecasting in this case.

Conclusion

In this chapter, the method that will be applied to analyze the trends of GDP and the unemployment rate in Malaysia will be outlined. The Data Science Project Life Cycle will be employed in the research process, data acquisition, and data cleaning, EDA, Modelling, and Evaluation, and deployment. For the purpose of fulfilling the research objectives, this research will utilize multiple datasets and will utilize multiple analytical tools including time series decomposition, ARIMA models, and Prophet models in order to analyze the relationship between GDP and unemployment and present the results that can be useful for policymakers and economists.