CHAPTER 3

METHODOLOGY

3.0 Data Science Life Cycle

This methodology adopts a structured data science project life cycle mechanism that facilitates proper data collection, analysis, and model creation. The life cycle consists of seven key phases: problem identification, data collection, data preprocessing, analysis, data modelling, and assessing the model's performance and implementation. The above-mentioned phases are important for completing the project and obtaining the correct and precise forecasting results.

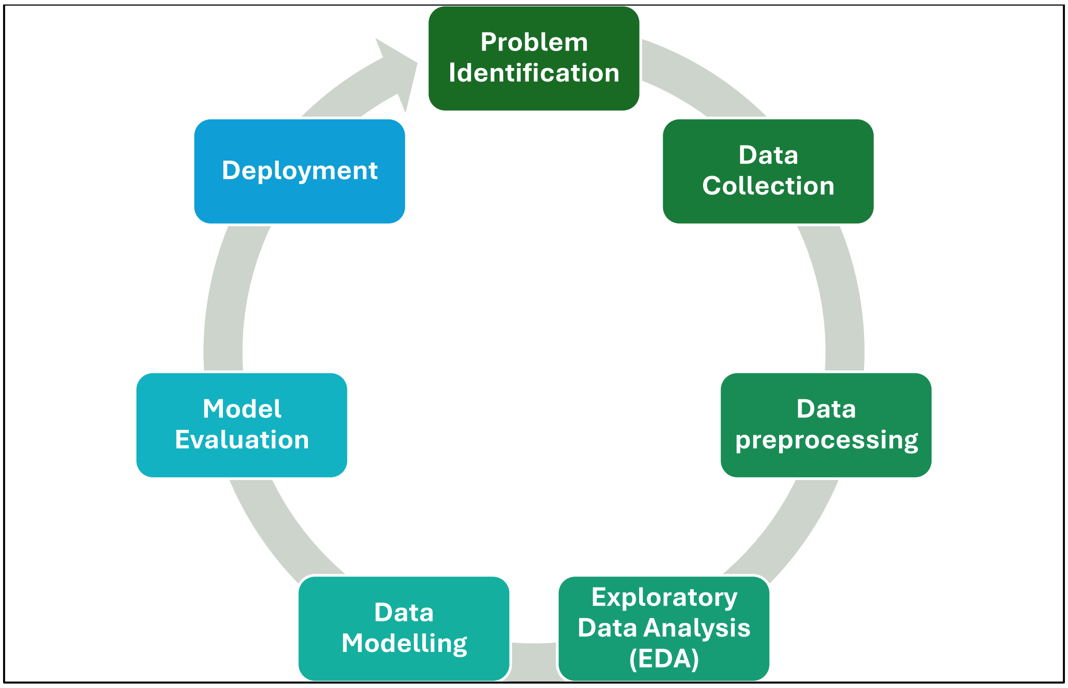

Figure 3.1 Data Science Life Cycle

a) Problem Identification

The first phase is crucial because it involves identifying the problem, guiding the entire study, and defining the research objectives. The main aim of this project is to predict energy consumption in Malaysia by employing regression models. In the problem definition process, one outlines issues that need to be addressed, exposed, and explained, as well as the areas of focus and the resulting findings. This phase creates certainty for all stakeholders because they will be fully aware of what is expected from them and what is expected in terms of the project outcome.

b) Data Collection

Data collection is crucial to gathering relevant data from various sources. For this study, data on historical energy consumption, economic indicators (GDP), demographic factors (population growth, urbanization), and climatic variables (temperature, humidity) are collected. Sources include national databases such as the Energy Commission of Malaysia and the Department of Statistics Malaysia, open-source websites such as Statista, and international organizations like the World Bank.

c) Data Preprocessing

Data cleaning is the process of selecting, integrating, validating, and transforming the collected data into a standard form for analysis. It deals with missing values, outliers' exclusion, data normalization, and transforming new features if needed. Other steps include feature scaling and normalizing: categorical features are converted into numerical format, and data are transformed and made uniform. This step is vital to get high-quality, reliable data in the modelling phase.

d) Exploratory Data Analysis (EDA)

One step of data analysis is called Exploratory Data Analysis (EDA), which determines relationships in the data sets. In this stage, the data is represented using graphical techniques like charts and graphs, computation of other central tendencies and dispersion measures, and searching for trends, seasonality, and other irregularities. Regarding the whole process, EDA assists in forming hypotheses about the data and the findings that feed into the next phase, modelling. It also allows us to define which fields are significant and possible transformations, if any, that are required.

e) Data Modelling

The model development phase entails using suitable regression methods to predict energy consumption. This study uses linear regression analysis, polynomial regression analysis, multiple linear regression analysis, and regularization analysis with items such as Ridge and Lasso. The models are estimated using historical data and checked for cross-validation to improve their performance.

f) Model Evaluation

Model evaluation is a critical phase that determines the developed models' effectiveness and uses suitable measures. Some of the measures of regression models include Mean Absolute Error (MAE), Mean Squared Error (MSE), and the Root Mean Squared Error (RMSE). The models are assessed from training and validation data sets for their ability to perform on unseen data. It also increases the chances of selecting the most promising model for deployment from various applicable models.

g) Deployment

The last step focuses on deploying the model into an application that will utilize it to make predictions at time intervals. The deployment also includes establishing systems to monitor the model after deployment and updating it with the new data when needed. This makes it possible to constantly update it so that it will be able to adapt well to conditions that might change with time.

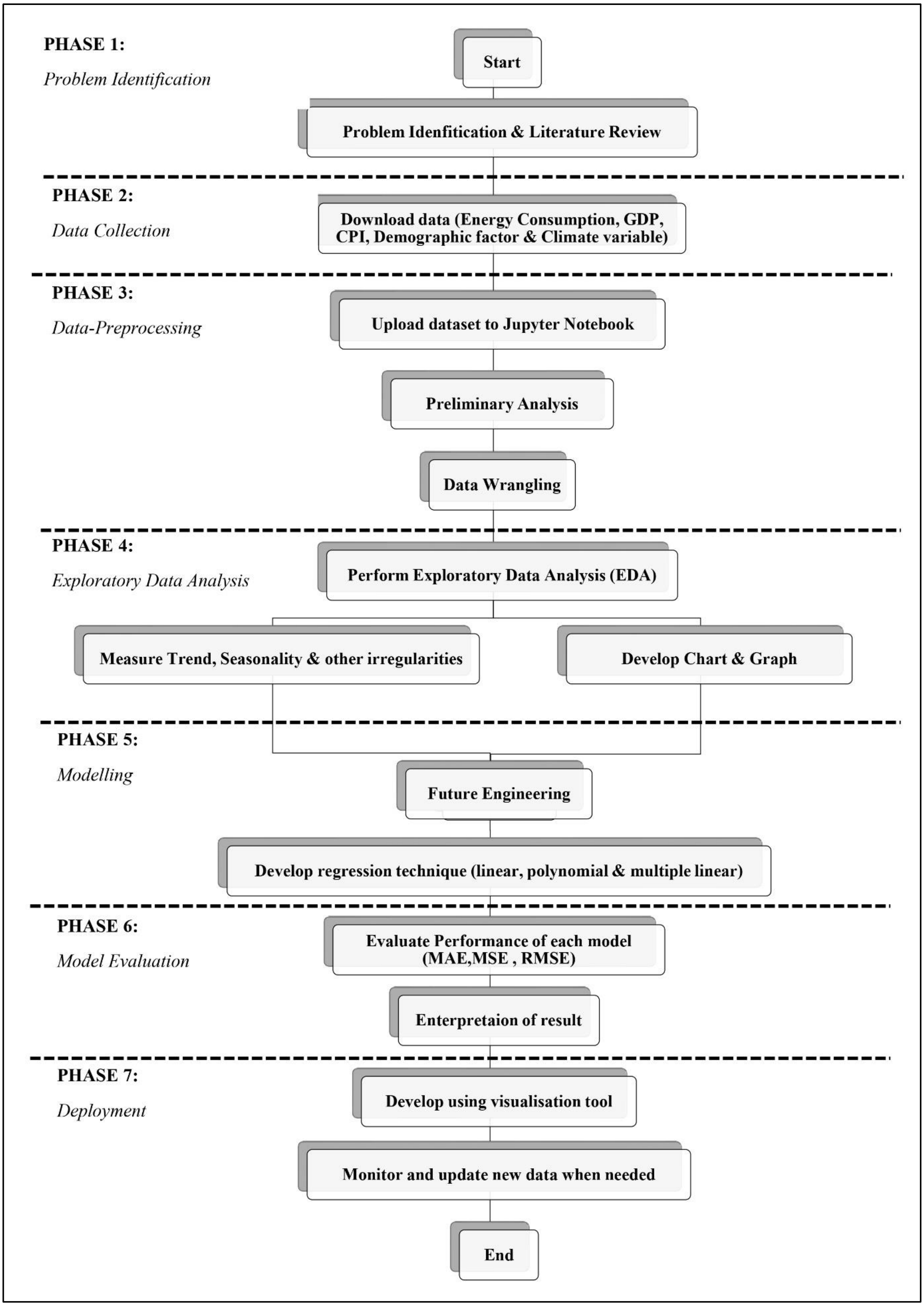

Figure 3.2 Research Framework of Energy Consumption in Malaysia

3.1 Problem Identification

The primary aim of this project is to leverage advanced regression techniques to enhance the accuracy and reliability of forecasting future energy consumption in Malaysia. However, several challenges need to be addressed to generate high-quality predictions:

a) **Data Collection and Quality**
   - **Navigating the Complexities of Diverse Data Formats**: Ensuring data consistency and integration from various sources, such as the Energy Commission of Malaysia, the Department of Statistics Malaysia, Tenaga Nasional Berhad, and global organizations such as the International Energy Agency (IEA) and the World Bank can be challenging. Different data formats and units of measurement need to be standardized.
   - **Ensuring Data Quality**: Data quality is crucial for creating accurate forecasting. Issues such as missing values, outliers, and inconsistent data entries must be identified and addressed to maintain the integrity of the dataset.
   - **Identifying Meaningful Patterns, Trends, and Correlations**: Extracting relevant insights from historical energy consumption data requires good techniques to identify underlying trends, seasonal variations, and correlations with economic, demographic, and climatic factors.

b) **Model Selection and Development**
   - **Selecting the Most Appropriate Regression Techniques Methods**: Choosing the suitable regression models (linear, polynomial, multiple linear regression) that align with the characteristics of energy consumption data is crucial. The models must be capable of capturing temporal dependencies and non-linear relationships present in the data.
   - **Capturing Complex Relationships**: Energy consumption is influenced by multiple factors, including economic growth, population dynamics, and climate conditions. The models need to capture these complex interactions effectively to provide accurate forecasts.

c) **Model Evaluation and Continuous Improvement**

- **Evaluating Model Performance**: Rigorous evaluation of the regression models using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) is essential to determine their accuracy and reliability.
- **Continuous Refinement and Updates**: Factors that influence energy consumption are dynamic and can update from time to time. The models need continuous refinement and updates with new data to maintain their relevance and accuracy. This involves setting up monitoring systems and retraining the models periodically to adapt to changing conditions.

3.2 Data Sources and Collection Methods

The dataset employed in this study was retrieved from different credible national and international databases to capture all the variables impacting energy usage in Malaysia. The primary data sources include:

1. **Energy Commission of Malaysia (Suruhanjaya Tenaga)**: Provides detailed reports and statistics on energy consumption by sector, including residential, industrial, and commercial energy use.
2. **Department of Statistics Malaysia (DOSM)**: This agency offers economic and demographic data, such as GDP, population growth, and urbanization rates.
3. **Tenaga Nasional Berhad**: Provides details about the electricity consumed in Malaysia.
4. **Climate Change Knowledge Portal**: Provide details about climate variables such as temperature and humidity
5. **World Bank and International Energy Agency (IEA)**: Provide global economic indicators and comprehensive energy statistics.

In data collection, these sources are accessed through online platforms containing the dataset or downloaded from PDF and organized in a standard format for analysis. In this way, all the variables are guaranteed to be listed and synchronized for further analysis at this step. The datasets and the description given are stated in Table 3.1 below.

Table 3.1 Dataset and the description given

No	Datasets	Description	Source
1	Amount of electricity consumed in Malaysia.xlsx	Amount of electricity consumed in Malaysia from 2014 to 2023 (in billion kilowatt hours)	• Department of Statistics Malaysia (DOSM) • Tenaga Nasional Berhad
2	Energy demand by sector.xsl	Final energy demand by sector (Industrial, Transport, Agriculture, non – Energy, Residential and Commercial)	Energy Commission of Malaysia (Suruhanjaya Tenaga)
3	Final electricity consumption by sector.xsl	Final electricity consumption by sector (Industrial, Transport, Agriculture, Residential and Commercial)	Energy Commission of Malaysia (Suruhanjaya Tenaga)
4	Household electricity consumption per capita in Malaysia.xlsx	Household electricity consumption per capita in Malaysia from 2000 to 2016 (in kilowatt-hours)	International Energy Agency (IEA)
5	Energy Indicator – Energy intensity per unit GDP.xls	Primary Energy Intensity (toe/GDP at 2015 Prices (RM million))	Energy Commission of Malaysia (Suruhanjaya Tenaga)
6	Urbanization in Malaysia based on year.xlsx	Percentage of Urbanization in Malaysia 2023	World Bank
7	Primary energy supply in Malaysia.xls	Primary Energy Supply (ktoe) include crude oil, petroleum products, natural gas, coal and coke, hydropower, biodiesel, solar, biomass and biogas.	Energy Commission of Malaysia (Suruhanjaya Tenaga)

8	Economic Indicator – GDP.xls	The dataset includes GDP at 2015 Prices (RM Million) and GDP at Current Prices (RM Million)	Department of Statistics Malaysia (DOSM)
9	GDP annual nomial supply.csv	The dataset includes series, date, sector and value	Department of Statistics Malaysia (DOSM)
10	GDP quarter nomial supply.csv	The dataset includes series, date, sector and quarter value.	Department of Statistics Malaysia (DOSM)
11	Mean Temperature Rainfall and Mean Humidity in Malaysia.csv	The dataset includes mean temperature and humidity according to the state of Malaysia.	Department of Statistics Malaysia (DOSM)
12	Average temperature in Malaysia.xlsx	List of average every state in Malaysia.	Climate Change Knowledge Portal
13	Economic Indicator – Population.xls	Population based on year.	Department of Statistics Malaysia (DOSM)
14	Population in Malaysia.csv	The population dataset includes date, sex, age and value	Department of Statistics Malaysia (DOSM)
15	World population Growth (annual %).csv	Percentage of annual growth by country.	World Bank

3.3 Data Pre-processing

Data pre-processing is vital in ensuring the data is clean, consistent, and ready to use before the analysis. This process involves data cleaning, transformation, and future engineering. Below are the detailed steps for preliminary analysis to prepare the data for modelling and analysis.

3.3.1 Data Collection and Integration

The data is collected from various resources, including the Energy Commission of Malaysia, Department of Statistics Malaysia, Tenaga Nasional Berhad, and international sources such as the World Bank and International Energy Agency (IEA). The datasets that are collected are historical data on energy consumption in Malaysia, sector-specific consumption, economic indicators (GDP), consumer price index (CPI), demographic data (population growth), and climatic variables (temperature and humidity). For data integration, the datasets are merged from different sources into a unified format to ensure they have a standard reference for proper merging. It also handles the discrepancies in data formats or units of measurement.

3.3.2 Data Cleaning

Data cleaning is a crucial part of the pre-processing process to ensure the dataset is accurate and reliable for the analysis. The first task we should do is handle the missing value within the dataset. This process will start with identifying any missing value and using appropriate techniques such as mean, median, or mode replacement, which are commonly used. For time series data, the methods of forward-fill and backward-fill can be applied to propagate the last known value to fill in the missing data. If certain data have a significant amount of missing data and are not critical, these data may be removed, or it can maintain the integrity of the dataset.

The next step is removing the duplicate data, as it is essential to check and eliminate any duplicate data to avoid redundancy and ensure that the data is unique. Duplicate data can distort analysis results and lead to inaccurate analysis and results, which may affect decision-making in the future.

Furthermore, outlier detection and treatment are other essential aspects of data cleaning. Outliers can be identified using statistical methods such as the Z-score or the Interquartile Range (IQR) and visualizations such as box plots. The outliers are treated to prevent them from skewing the analysis, and this can be done either by removing them or capping their values to a more reasonable range. By addressing this step, the dataset is refined and well-prepared for the analysis, ensuring the results are accurate and reliable.

3.3.3 Data Transformation

In data transformation, converting data into a suitable format for analysis is a part of the process. Several key steps are included in the process. First, normalization and scaling of numerical features are performed to ensure they have similar ranges. Min-Max scaling is A common technique that scales the data into a fixed range, usually 0 and 1. Standardization transforms the data to have a mean zero and a standard deviation to one, and robust scaling uses robust statistics for outliers such as the median.

Second, convert them into numerical formats that machine learning algorithms can utilize is essential for encoding categorical variables. This can be achieved by one-hot encoding, which will create a binary column for each category, or label encoding, which will assign a unique integer to each category. These methods aim to ensure that categorical data is represented in a way that maintains the integrity of the dataset.

Lastly, date-time feature extraction is performed to use the features from date-time columns such as year, month, and day. This step also involves creating cyclic features using (sine and cosine transformation to represent periodic data like months or days that efficiently capture the cyclical nature of time-related data. By transforming the data, all the datasets have become more structured and suitable for analysis and modelling, facilitating more accuracy and reliability.

3.3.4 Future Engineering

Future engineering involves creating new variables that capture additional information or enhance the model’s prediction. This includes generating lagged variables to capture temporal dependencies in time series data, creating interactions between features to capture the combined effects, developing polynomial features to non-linear relationships, calculating rolling to smooth out the short-term fluctuations, and creating dummy variables to capture seasonal effects such as months, quarters, or seasons. These steps ensure consistent data across different periods and sources and prepare more accurate and reliable modelling outcomes.