Chapter3:Methodology

3.1 Introduction

3.2 Data Collection

3.2.1 Traffic Data

3.2.2 Weather Data

3.2.3 Public Transport Ridership and Vehicle Registration Data

3.3 Data Preprocessing

3.4 Exploratory Data Analysis (EDA)

3.5 Model Evaluation:

3.6 Deployment and Monitoring:

3.1 Introduction

The research problem investigated in this study is predicting traffic congestion in Malaysia using machine learning algorithms. The goal is to develop models that can accurately forecast traffic jams, enabling better traffic management and planning. This research will follow the data science project life cycle.

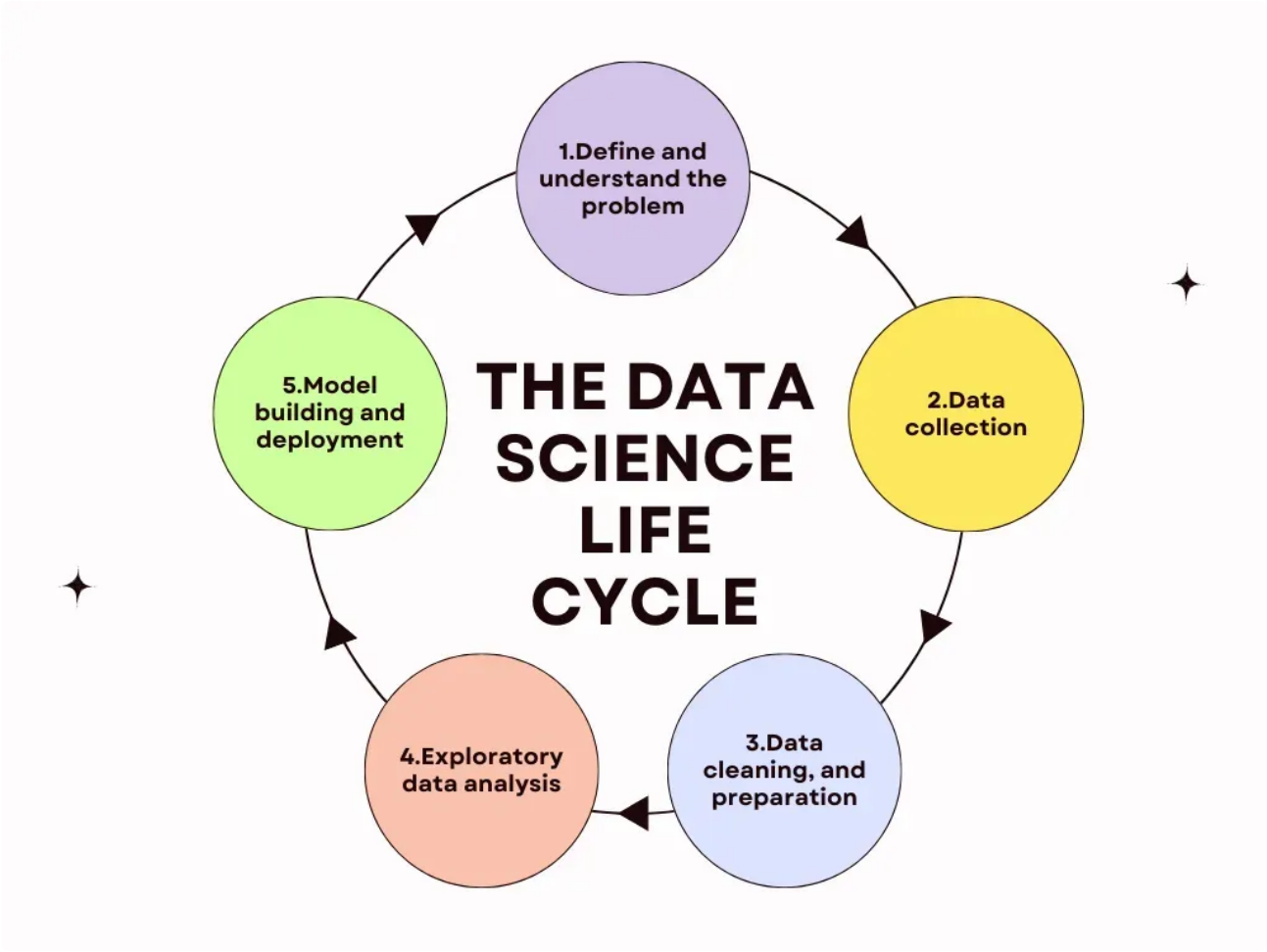

Figure1: Data Science Project Life Cycle

3.2 Data Collection

3.2.1 Traffic Data

Accessing Traffic Data from Xmap.ai: API Registration: Registered for an API key on the Xmap.ai platform API Requests: Utilizing the API key, perform HTTP GET requests to the Xmap.ai endpoints to get traffic data. The data get in JSON format and contains various traffic metrics. Data Storage: Stored the collected data in CSV , to process it in the analysis step.

3.2.2 Weather Data

Fetching Weather Data from OpenWeatherMap: API Registration: Signed up for an API key on the OpenWeatherMap website API Requests: Use the API key to make requests to the OpenWeatherMap historical weather data endpoint.

3.2.3 Public Transport Ridership and Vehicle Registration Data

Obtaining Public Transport Ridership and Vehicle Registration Data: Data Download: Navigate to the Ministry of Transport, Malaysia's open data portal and acquire the datasets. These are dataset files in CSV format. Data Import: Import the downloaded CSV files into an analysis environment using a variety of data manipulation libraries. Data Cleaning: The dataset will prospectively require cleaning to account for incomplete data and inconsistencies, all of which can influence the accuracy of any further data handling.

3.3 Data Preprocessing

The collected data will be thoroughly cleaned to ensure accuracy. This involves imputing any missing or null values, converting different time formats to a uniform standard, and incorporating new derived features. This step is crucial to prepare the data for effective analysis and modeling.

3.4 Exploratory Data Analysis (EDA)

Comprehensive visualization and analysis of the data will be conducted using graphical libraries. This step will help identify important patterns, trends, and insights within the data. Graphical representations will make it easier to understand the data’s behavior and identify any anomalies or significant trends.

3.5 Model Evaluation:

Machine learning models will be developed and trained to gain insights and learn complex patterns in the dataset. This involves selecting appropriate algorithms, tuning model parameters, and training the models on the preprocessed data to ensure they can accurately predict traffic congestion.

3.6 Deployment and Monitoring:

The final stage involves implementing the best-performing model in a production environment and monitoring its behavior. This ensures the model continues to perform well with new data and can provide real-time traffic predictions. Ongoing monitoring will help detect any issues and allow for timely updates to the model as needed.