CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter will elaborate more on the methodology used in this research which includes the research design framework, steps and techniques used for the research. The overall research workflow consists of four phases which will be discussed further in this chapter. Each step of the framework will be elaborated in detail on how it will be implemented to this research including the techniques that will be used.

3.2 Research Workflow

The research methodology framework has four main phases which are Phase 1 Literature Review, Phase 2 Problem Definition, Phase 3 Experiment & Evaluation and Phase 4 Result Documentation. The framework starts with Phase 1 which is the literature review of the study, it related to the process of gathering information related to microclimate monitoring and prediction and machine learning techniques as shown in Figure 3.1. The framework then continues with the second phase which is the problem definition phase of the thesis, the problem on microclimate monitoring & prediction. The next phase is the experiment and evaluation phase which is how to conduct the research and how to evaluate data of the results acquired. The last phase is the result documentation phase where end results will be reported.

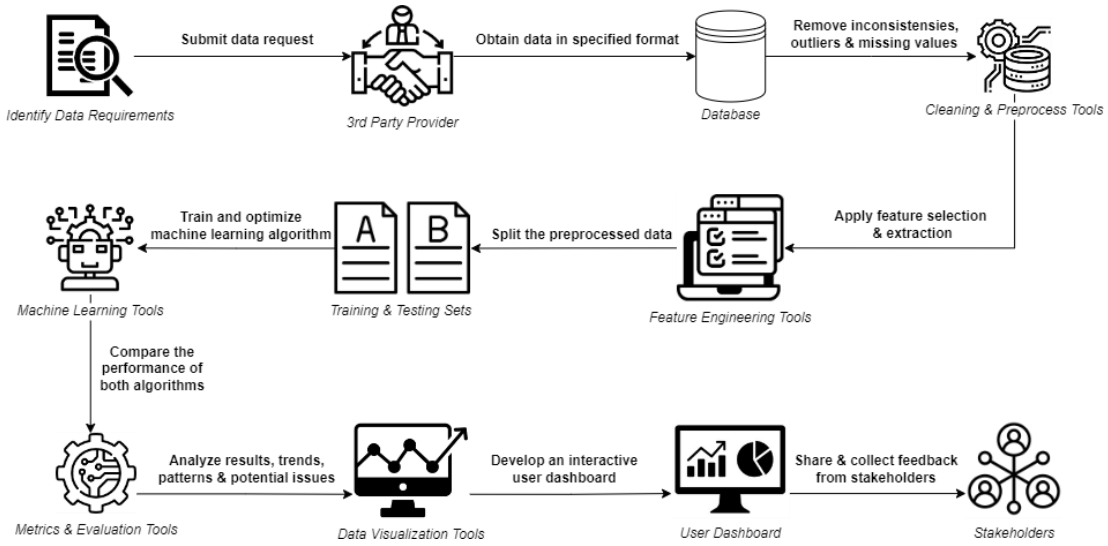


Figure 1: Research Workflow

3.2.1 Phase 1: Literature Review

In the first phase, a comprehensive literature review is conducted to gather relevant information on preserving cultural heritage sites through microclimate monitoring and prediction using Random Forest and XGBoost algorithms. Various scholarly sources, including journals, articles, and theses, are explored to understand the current state of research in this field. The literature review delves into topics such as data collection methods, pre-processing techniques, and the application of machine learning models. By examining existing studies, this phase helps identify gaps, challenges, and potential solutions for effectively monitoring and predicting microclimate conditions at heritage sites. The insights gained from the literature review form the foundation for the subsequent phases and guide the research towards developing a robust methodology.

3.2.2 Phase 2: Data Collection and Preprocessing

In the second phase, the focus shifts to obtaining microclimate data from the Malaysian Meteorological Department for a specific heritage site in Johor Bahru. This data, encompassing temperature, relative humidity, and wind, rainfall and solar

radiation measurements, serves as the basis for subsequent analysis and modelling. To ensure data quality, a rigorous pre-processing stage is undertaken, which involves cleaning the raw data and addressing any missing values or outliers. Techniques such as interpolation, statistical analysis, and data imputation are employed to enhance the integrity and accuracy of the collected data. Additionally, feature engineering techniques are applied to extract meaningful features from the raw data, enabling the capturing of temporal dependencies and relationships between variables. This phase prepares the dataset for further analysis and model development in the subsequent phases.

3.2.3 Phase 3: Machine Learning Model Development

The third phase revolves around the development of machine learning models for microclimate monitoring and prediction. The pre-processed data is divided into training and testing sets, with the training set used to train and optimize the Random Forest and XGBoost algorithms. Various parameters and hyperparameters are fine-tuned using techniques like grid search and cross-validation to achieve optimal model performance. The trained models are then evaluated using appropriate assessment metrics, such as mean absolute error and mean squared error, to assess their predictive capabilities. This evaluation process helps determine the effectiveness and performance of the Random Forest and XGBoost algorithms in accurately predicting microclimate patterns for the designated heritage site. The models' performance and generalizability are crucial factors in ensuring the reliability and usefulness of the developed models for microclimate monitoring and prediction.

3.2.4 Phase 4: Dashboard Development

In the fourth phase, a user-friendly dashboard is designed and developed to visualize and present the microclimate data in a comprehensible manner. The dashboard provides real-time insights into the microclimate conditions of the heritage site, displaying key metrics such as temperature, humidity, and wind speed. The trained machine learning models are integrated into the dashboard to provide

recommendations for preventive maintenance actions based on the analysed data. Additionally, visualization tools and interactive features are incorporated to facilitate a better understanding of trends and patterns in the microclimate data. The dashboard serves as a valuable tool for local authorities and stakeholders involved in the preservation of cultural heritage sites, enabling them to make informed decisions and take proactive measures to mitigate potential issues that may impact the site's condition.

3.3 Justification of Tools, Techniques and Data

The chosen tools for this research include data collection from the Malaysian Meteorological Department (MET Malaysia), data pre-processing techniques, machine learning algorithms (Random Forest and XGBoost), and dashboard development. These tools have been selected based on their suitability for addressing the research objectives and providing valuable insights for preventive maintenance strategies at the designated heritage site in Johor Bahru.

Microclimate data from MET Malaysia is essential for understanding the environmental conditions at the heritage site. Temperature, humidity, and wind speed are crucial parameters that can affect the preservation of heritage structures. By obtaining this data, we can assess the impact of these environmental factors on the site's structural integrity and identify potential maintenance issues.

The research utilizes machine learning techniques to analyse the collected microclimate data and develop predictive models for preventive maintenance. Random Forest and XGBoost algorithms are chosen due to their proven effectiveness in handling complex relationships between variables and handling both regression and classification tasks. These algorithms can capture temporal dependencies in the data and provide accurate predictions for future microclimate conditions. More comprehensive justification of research is discussed below:

1. Preservation of Cultural Heritage: The preservation of heritage sites is of paramount importance to maintain cultural identity and historical significance. By utilizing advanced data analysis techniques, this research aims to provide insights into the impact of microclimate conditions on the designated heritage site in Johor Bahru. The findings can help authorities develop targeted preventive maintenance strategies to ensure the site's long-term preservation.

2. Data-Driven Decision Making: By collecting and analysing microclimate data, this research enables data-driven decision making for preventive maintenance. Traditional approaches may not consider the dynamic nature of microclimate conditions and their impact on heritage sites. The use of machine learning algorithms allows for a more comprehensive understanding of the relationships between environmental factors and potential maintenance issues, leading to more informed decision making.

3. Efficiency and Cost-Effectiveness: Implementing preventive maintenance strategies based on predictive models can result in cost savings and increased efficiency. By identifying trends, patterns, and potential issues, maintenance activities can be prioritized, scheduled, and targeted accordingly. This approach minimizes reactive maintenance efforts, reduces costs associated with emergency repairs, and optimizes resource allocation.

4. Real-Time Monitoring and Visualization: The development of a user-friendly dashboard integrating real-time microclimate data and machine learning models provides a powerful tool for monitoring and maintenance planning. The visualization tools within the dashboard help users understand trends and patterns in the data, facilitating proactive decision making. This allows local authorities to respond promptly to changing microclimate conditions and potential threats to the heritage site.

5. Stakeholder Engagement and Collaboration: The research encourages collaboration between local authorities, heritage site management teams, and relevant stakeholders. By involving these parties in the evaluation and testing

phases, their feedback can be gathered to refine the system and ensure its usability and effectiveness. Engaging stakeholders throughout the research process increases their ownership and facilitates the adoption of preventive maintenance strategies.

3.4 Chapter Summary

This chapter summarized the four phases of the research study. The literature review phase involved a comprehensive review of relevant literature, providing a solid knowledge base for the subsequent phases. The data collection and pre-processing phase focused on obtaining and cleaning microclimate data, while the machine learning model development phase involved training and evaluating Random Forest and XGBoost algorithms. The final phase focused on developing a user-friendly dashboard that visualizes the microclimate data and provides maintenance recommendations.

CHAPTER 4

RESEARCH DESIGN AND IMPLEMENTATION

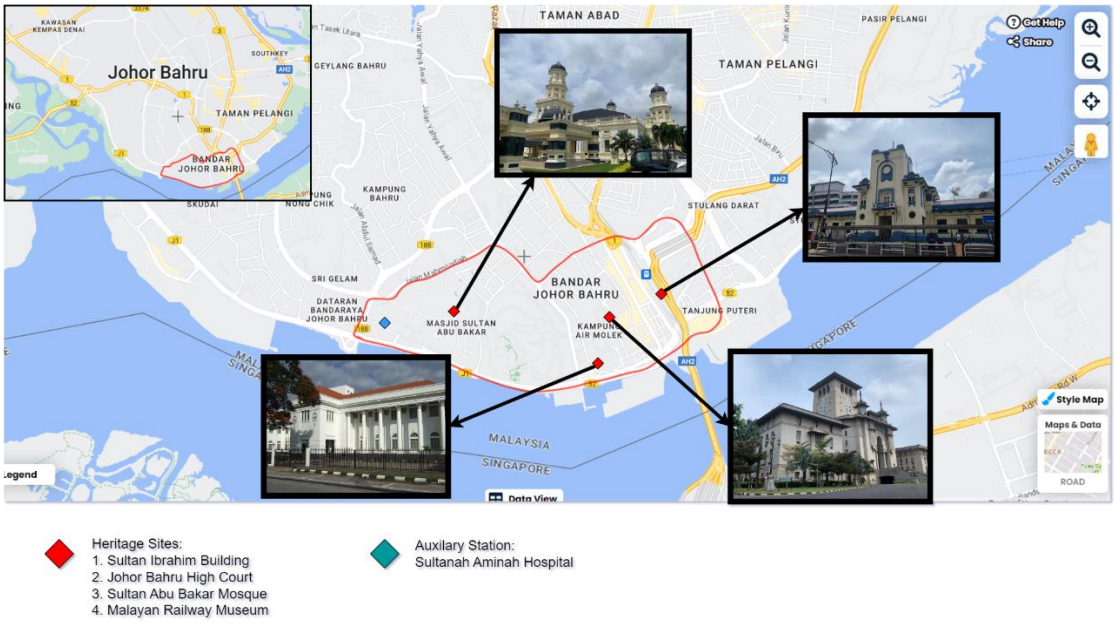
4.1 Introduction

This chapter discusses in depth the research design and implementation of the research methodology described in the previous chapter. The proposed solution will be broken down into several steps, including the data collection, pre-processing, feature extraction, training and testing and machine learning models development.

4.2 Proposed Solution

The dataset used in this study is collected and extracted from Malaysian Meteorological Department. The dataset that is used in this study mainly from microclimate data such as temperature, humidity, wind speed, rainfall, and solar radiation. The dataset must go through a few experimental steps in order to obtain the sentiment score of the social media data. First, data must be extracted from social media. Following that, the data is subjected to pre-processing and feature extraction. The data then moves on to the next step, which is the feature vector and sentiment classifier implementation step. The final step in this research is to calculate the sentiment score for each tweet. Python code will be used to run each process in this experiment.

4.3 Research Area Zone Mapping



In the context of this research, a research area zone mapping was established to focus on the preservation of cultural heritage sites in Johor Bahru, Malaysia. The selected heritage sites for this study are:

1. Sultan Ibrahim Building: This building holds historical and architectural significance as the state secretariat building of Johor. It represents the rich cultural heritage of the region.

2. Johor Bahru High Court: The Johor Bahru High Court is a notable judicial institution that plays a crucial role in the administration of justice. It possesses historical and legal importance.

3. Sultan Abu Bakar Mosque: Known for its exquisite Moorish-inspired architecture, the Sultan Abu Bakar Mosque is a revered religious site. It serves as a symbol of faith and cultural identity.

4. Malayan Railway Museum: The Malayan Railway Museum, located in Johor Bahru, showcases the historical development and significance of the railway system in Malaysia. It offers insights into the country's transportation heritage.

To ensure comprehensive monitoring and analysis of the microclimate conditions in the research area, the chosen auxiliary station is the Sultanah Aminah Hospital. This auxiliary station, situated close to the selected heritage sites, serves as an essential data collection point for microclimate variables. By selecting an auxiliary station in proximity to all the chosen heritage sites, the research study aims to capture accurate and representative microclimate data specific to the designated research area zone.

By focusing on the Sultanah Aminah Hospital auxiliary station, which covers all the selected heritage sites, the research can effectively monitor and analyse the microclimate conditions in the vicinity of these sites. This approach ensures that the data collected and analysed is directly relevant to the preservation efforts and maintenance strategies of the cultural heritage sites in Johor Bahru.

4.4 Experiment Design

4.4.1 Microclimate Data Collection Process

In this research study, data collection is a crucial step in developing the prototype dashboard. To ensure the accuracy of the results, data was collected from the Malaysian Meteorological Department (MET Malaysia) website and stored in our database. A duration of 30 years was chosen to gather a substantial amount of data, enabling more reliable and robust predictions. The data was specifically obtained from the auxiliary station of Hospital Johor Bahru, which is in close proximity to the research area.

Five categories of microclimate data were collected for this research, namely temperature, relative humidity, wind, rainfall, and solar radiation. A comprehensive set of attributes was acquired for temperature data, consisting of eight different types. Similarly, for relative humidity data, five distinct attributes were collected. In the case of rainfall, four attributes were available for analysis. However, due to limited availability of wind data at the

Hospital Johor Bahru auxiliary station, only one attribute, namely the hourly surface wind, could be obtained. Consequently, to enhance the predictive capabilities of the research study, the inclusion of solar radiation and rainfall data was deemed necessary. All of the collected data is in the CSV file format and accessible from the MET Malaysia website.

The meticulous collection and inclusion of these various microclimate data categories and attributes aim to ensure a comprehensive analysis and prediction process within the research study. By incorporating multiple data sources, a more holistic understanding of the microclimate conditions at the designated heritage site can be achieved, ultimately facilitating the development of an effective and informative dashboard prototype.

4.4.2 Pre-processing and Feature Extraction of Microclimate Data

After the data are collected, they must undergo the pre-process, and feature extract. The data collected are cleaned by using Phyton to remove any outliers to ease the pre-processing process. These two processes are required in the development of machine learning models to obtain a clean dataset, which will make algorithms more convenient and accurate.

```
import pandas as pd
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import f_regression
```

Figure 2: Import necessary libraries

```
data = pd.read_csv('microclimate_data.csv')
```

Figure 3: Read the microclimate data from a CSV file

```
imputer = SimpleImputer(strategy='mean')
data[['temperature', 'relative_humidity', 'wind', 'solar_radiation', 'rainfall']] =
imputer.fit_transform
(data[['temperature', 'relative_humidity', 'wind', 'solar_radiation', 'rainfall']])

scaler = StandardScaler()
data[['temperature', 'relative_humidity', 'wind', 'solar_radiation', 'rainfall']] =
scaler.fit_transform
(data[['temperature', 'relative_humidity', 'wind', 'solar_radiation', 'rainfall']])
```

Figure 4: (Pre-processing) Handle missing values using mean imputation and standardize the data

```
X = data[['temperature', 'relative_humidity', 'wind', 'solar_radiation', 'rainfall']]
y = data['target_variable']  # Target variable

kbest_selector = SelectKBest(score_func=f_regression, k=5)
X_new = kbest_selector.fit_transform(X, y)
selected_features = X.columns[kbest_selector.get_support()]
```

Figure 5: (Feature Engineering) Extract relevant features using SelectKBest with f_regression as the scoring function

4.4.3 Splitting of Data into Training and Testing Sets

In the below code snippet, the train_test_split function from scikit-learn is used to split the pre-processed data (X_new) and the corresponding target variable (y) into training and testing sets. The test_size parameter is set to 0.2, indicating that 20% of the data will be used for testing, while the remaining 80% will be used for training. The random_state parameter is set to 42 to ensure reproducibility of the split. Then, it will proceed with training and evaluating machine learning models using the X_train, y_train for training, and X_test, y_test for testing.

```
from sklearn.model_selection import train_test_split

# Split the preprocessed data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_new, y, test_size=0.2, random_state=42)
```

Figure 6: Split the pre-processed data into training and testing sets with an 80-20 ratio

4.4.4 Training and Evaluation of Machine Learning Models

The next step after splitting the data into training and testing sets is to train and evaluate the machine learning models. The figure shows the code for training and evaluating the Random Forest and XGBoost algorithms: In this code snippet, the Random Forest and XGBoost models are initialized and trained using the training data (X_train and y_train). Then, predictions are made on the testing data (X_test) using the trained models. The mean absolute error (MAE) and mean squared error (MSE) are calculated to evaluate the performance of both models.

```
from sklearn.ensemble import RandomForestRegressor
from xgboost import XGBRegressor
from sklearn.metrics import mean_absolute_error, mean_squared_error

# Initialize and train the Random Forest model
rf_model = RandomForestRegressor()
rf_model.fit(X_train, y_train)

# Make predictions on the testing set using the trained Random Forest model
rf_predictions = rf_model.predict(X_test)

# Calculate evaluation metrics for the Random Forest model
rf_mae = mean_absolute_error(y_test, rf_predictions)
rf_mse = mean_squared_error(y_test, rf_predictions)

# Initialize and train the XGBoost model
xgb_model = XGBRegressor()
xgb_model.fit(X_train, y_train)

# Make predictions on the testing set using the trained XGBoost model
xgb_predictions = xgb_model.predict(X_test)

# Calculate evaluation metrics for the XGBoost model
xgb_mae = mean_absolute_error(y_test, xgb_predictions)
xgb_mse = mean_squared_error(y_test, xgb_predictions)

# Print the evaluation results
print("Random Forest - Mean Absolute Error:", rf_mae)
print("Random Forest - Mean Squared Error:", rf_mse)
print("XGBoost - Mean Absolute Error:", xgb_mae)
print("XGBoost - Mean Squared Error:", xgb_mse)
```

Figure 7: Train and evaluate the Random Forest and XGBoost algorithms

4.5 Parameter and Testing Methods

In order to evaluate the effectiveness of the proposed solution, several parameters are measured during the testing phase. These parameters provide valuable information about the performance and efficacy of the developed models. The parameters to be measured include:

4.5.1 Parameters to be Measured

1. Prediction Accuracy: The accuracy of the predictive models in capturing and predicting the microclimate conditions is measured. This indicates how well the models are able to estimate the actual values of temperature, humidity, wind speed, solar radiation, and rainfall.

2. Mean Absolute Error (MAE): MAE is measured to determine the average absolute difference between the predicted and actual values. It provides insights into the average prediction error across all the microclimate parameters.

3. Mean Squared Error (MSE): MSE is calculated to assess the average squared difference between the predicted and actual values. It measures the overall variance between the predicted and actual values.

4.5.2 Testing Procedure

1. Splitting Data: The pre-processed microclimate data is divided into training and testing sets, typically using an 80:20 split. The training set is used to train the machine learning models, while the testing set is used for evaluating the performance.

2. Model Evaluation: The trained Random Forest and XGBoost models are applied to the testing set to make predictions for the microclimate parameters. The predicted values are then compared with the actual values from the testing set.

3. Calculation of Evaluation Metrics: The evaluation metrics, such as MAE and MSE, are calculated based on the predicted and actual values. These metrics provide quantitative measures of the model's performance.

4. Analysis and Interpretation: The evaluation results are analysed to gain insights into the accuracy and effectiveness of the trained models. The performance of the models is assessed based on the evaluation metrics and compared to determine the superior algorithm for microclimate prediction.

4.6 Chapter Summary

In this chapter, the research experimental design and implementation of the proposed solution for preserving cultural heritage sites through microclimate monitoring and prediction are summarized. The steps for data collection, pre-processing, model development using

Random Forest and XGBoost, evaluation and testing are outlined. The experimental setup and parameter details are provided, along with the evaluation metrics used to assess the performance of the models. The next chapter will present the results and analysis of the experiments, highlighting the insights gained and the recommendations for preserving cultural heritage sites.