UTM
UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

PRESERVING CULTURAL HERITAGE SITES THROUGH RANDOM FOREST AND XGBOOST ALGORITHM FOR MICROCLIMATE MONITORING AND PREDICTION

Iman Aidi Elham bin Hairul Nizam
School of Computing, Faculty Engineering
81310, Johor Bahru, Malaysia
imanaidielham@gmail.com

Assoc. Prof. Dr. Mohd Shahizan Othman
School of Computing, Faculty Engineering
81310, Johor Bahru, Malaysia
shahizan@utm.my

Abstract—This study aims to develop an accurate prediction model for microclimate data in the heritage-rich cities of Johor Bahru and Melaka, Malaysia. The research focuses on key microclimate variables including temperature, rainfall, humidity, and wind speed. Historical data will be obtained from the Copernicus Climate Data Store (CDS) to train and validate the prediction models. The study will compare the performance of two machine learning algorithms, Random Forest and XGBoost, to determine the most effective method for microclimate prediction in these specific urban environments. A user-friendly dashboard will be developed using HTML to visualize both historical data and predictions, making the information accessible and interpretable. The accuracy and reliability of the prediction models will be evaluated using standard statistical measures including Mean Absolute Error, Root Mean Square Error, and R-squared score. The models will be trained using cross-validation techniques for hyperparameter tuning, and their performance will be assessed on a held-out test set representing the year 2023. This research aims to provide a valuable tool for local meteorologists, urban planners, and researchers interested in the microclimatic conditions of Johor Bahru and Melaka. The resulting predictive model and dashboard could serve as a foundation for future studies on the impact of microclimate on urban planning and heritage conservation in these historically significant Malaysian cities.

Keywords — Rainfall, Temperature, Humidity, Wind Speed, Random Forest, XGBoost, Root Mean Square Error, Mean Absolute Error, R-squared score

I. INTRODUCTION

Cultural heritage sites are the foundation of our global historical values, connecting us to ancestral traditions and shaping our cultural identity. These sites face various risks of damage and deterioration due to microclimate conditions, including temperature, humidity, airborne pollutants, and air speed (Fabbri & Bonora, 2021). Particularly in developing nations, these impacts pose significant challenges to heritage preservation (Pioppi et al., 2020). Safeguarding these sites is crucial not only for preserving cultural identity but also for promoting cultural and tourism-driven economic development (Alcaraz Tarragüel et al., 2012).

In recent years, the administration of cultural heritage sites has gained worldwide focus through improved detection, monitoring, and assessment methods. Initiatives are underway to enhance and preserve these resources by adopting suitable adaptation measures and sustainable management approaches (Guzman et al., 2020). To address these challenges, this thesis focuses on the application of advanced machine learning algorithms, specifically Random Forest and XGBoost, for microclimate monitoring and prediction at cultural heritage sites. These algorithms were chosen for their ability to handle complex, non-linear relationships in environmental data and their robustness in dealing with the high variability often present in microclimate measurements.

By leveraging these techniques, this research aims to contribute to the preservation of cultural heritage sites under changing environmental conditions, supporting sustainable and efficient conservation efforts. The developed models and dashboard have the potential to provide heritage site managers and conservators with accurate, timely predictions of microclimate conditions, enabling proactive conservation measures and more efficient resource allocation.

The research objectives are:

(a) To investigate and identify the most suitable machine learning algorithms for analyzing microclimate data, recognizing patterns, trends, and predictions in the heritage site's area.
(b) To evaluate the accuracy of the developed machine learning models.
(c) To develop and design a dashboard that displays microclimate data trends and predictions.

This approach of combining advanced machine learning techniques with practical, user-friendly tools represents a significant step forward in the field of heritage preservation, offering a data-driven solution to the complex challenges of microclimate management in cultural sites.

II. LITERATURE REVIEW

This section presents a comprehensive review of literature relevant to the application of machine learning algorithms in microclimate monitoring and prediction for cultural heritage preservation. It begins with an overview of the study areas and data sources. Subsequently, it delves into a systematic review of existing research related to this topic. The section then explores the data analysis techniques, focusing on two key algorithms: Random Forest and XGBoost. Finally, it discusses the methods for measuring the accuracy of prediction results for microclimate parameters in the context of cultural heritage sites.

A. *Study Area*

This research focuses on two significant cultural heritage sites in Malaysia. The first is A Famosa in Melaka, a 16th-century Portuguese fortress located in the historic city. It is one of the oldest surviving European architectural remains in Southeast Asia, situated at approximately 2.1936° N, 102.2501° E. The second site is the Sultan Ibrahim Building in Johor Bahru, an iconic administrative building built in the early 20th century. It stands as a prime example of colonial architecture, located at approximately 1.4616° N, 103.7622° E. Both sites are subject to the tropical climate of Malaysia, characterized by high temperatures, humidity, and significant rainfall throughout the year.

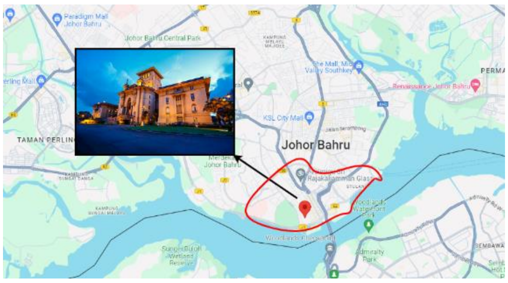


Figure 1: Sultan Ibrahim Building, Johor Bahru

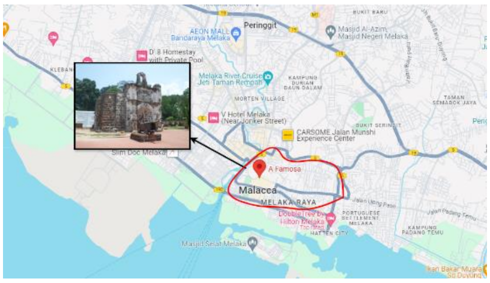

Peringgit
A Famosa
Malacca
MELAKA RAYA
Melaka River Cruise
CARSOME Jalan Munshi Experience Center
PORTUGUESE SETTLEMENT MELAKA

Figure 2: A Famosa, Melaka

B. *Data Source*

The dataset for this study was obtained from the Copernicus Climate Data Store (CDS), which provides comprehensive climate and environmental data. The data includes monthly records of rainfall, humidity, temperature, and wind speed from 1940 to 2023 for the locations of the selected heritage sites. This extensive historical data allows for a robust analysis of long-term microclimate trends and patterns.

C. *Microclimate Parameters*

The study focuses on four key microclimate parameters: rainfall, humidity, temperature, and wind speed. Rainfall, measured in millimeters (mm), is crucial for understanding moisture-related risks to heritage structures. Humidity, expressed as a percentage, significantly impacts the preservation of materials in heritage sites. Temperature, recorded in degrees Celsius (°C), can cause thermal stress on heritage structures through its fluctuations. Wind speed, measured in meters per second (m/s), influences erosion and pollutant deposition on heritage sites. These parameters collectively provide a comprehensive picture of the microclimate conditions affecting the cultural heritage sites.

D. *Random Forest Algorithm*

Random Forest, introduced by Breiman in 2001, is an ensemble learning method that has gained popularity in various fields, including environmental modeling and heritage preservation. It operates by constructing multiple decision trees and merging them to get a more accurate and stable prediction. Random Forest is particularly suited to microclimate prediction due to its ability to handle high-dimensional data without overfitting, its robustness to noise and outliers in the dataset, its capability to rank the importance of input variables, and its effective management of both categorical and continuous variables. In the context of microclimate monitoring for heritage sites, Random Forest can effectively model complex interactions between various environmental parameters, providing insights into the most influential factors affecting the microclimate around cultural heritage structures.

E. XGBoost Algorithm

XGBoost (Extreme Gradient Boosting) is an advanced implementation of gradient boosting machines. It has gained prominence in recent years due to its high performance and accuracy in handling large-scale, multi-dimensional datasets. XGBoost offers several advantages for microclimate prediction, including efficient handling of missing data through built-in mechanisms, regularization to prevent overfitting, parallel and distributed computing capabilities for faster processing, and built-in cross-validation at each iteration. In the realm of cultural heritage preservation, XGBoost's ability to process large volumes of climate data efficiently makes it particularly suitable for real-time microclimate monitoring and prediction. Its robust performance in the presence of complex, non-linear relationships between variables can provide accurate forecasts of microclimate conditions, enabling proactive conservation measures.

F. Performance Evaluation Metrics

Both algorithms will use Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and R-squared for accuracy measure. These are the most common accuracy measure that has been used for regression. The formula for MAE, RMSE and R-squared are as follow:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Where,
$\hat{y}$ $-$ *predicted value of y*
$\bar{y}$ $-$ *mean value of y*

Figure 3: Formula of MAE, RMSE, R-squared

MAE measures the average magnitude of errors in a set of predictions, without considering their direction. It's calculated as the average of the absolute differences between predicted and actual values. In this study, MAE represents the average deviation of predicted temperature, humidity, rainfall and wind speed from their actual values. A lower MAE indicates better model performance.

RMSE is the square root of the average of squared differences between predicted and actual values. It gives more weight to large errors due to the squaring operation. In microclimate modeling, RMSE provides a measure of the typical magnitude of prediction errors, expressed in the same units as the climate variable being predicted. Like MAE, a lower RMSE indicates better model performance.

R-squared represents the proportion of variance in the dependent variable (e.g., temperature, humidity, rainfall, windspeed) that is predictable from the independent variables. It ranges from 0 to 1, with 1 indicating perfect prediction. In microclimate modeling, a higher R-squared suggests that the model explains a larger portion of the variability in the climate data at heritage sites.

III. RESEARCH METHODOLOGY

This section, describe the methodology for the analysing of the research framework which describes the methods used from data collection until evaluating the result. This section aims to achieve a fuller understanding of the research before implementing it. This methodology explains step by step for accomplishing the research objectives.

A. Data Collection

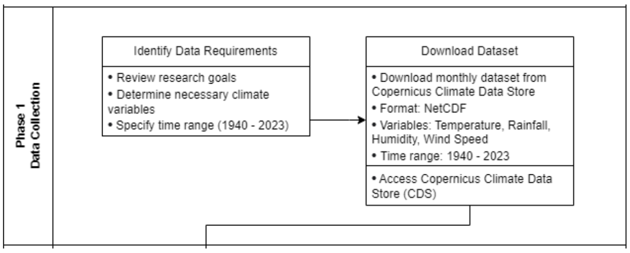

Phase 1
Data Collection
Identify Data Requirements
• Review research goals
• Determine necessary climate variables
• Specify time range (1940 - 2023)
Download Dataset
• Download monthly dataset from Copernicus Climate Data Store
• Format: NetCDF
• Variables: Temperature, Rainfall, Humidity, Wind Speed
• Time range: 1940 - 2023
• Access Copernicus Climate Data Store (CDS)

Figure 4: Activities in Data Collection Phase

This initial phase focuses on identifying and acquiring the necessary climate data. It begins with a review of research goals and determining the required climate variables. The time range is specified as 1940-2023. The data is then downloaded from the Copernicus Climate Data Store in NetCDF format, including variables such as temperature, rainfall, humidity, and wind speed. This phase ensures that the appropriate data is collected to support the research objectives.

B. *Data Processing*

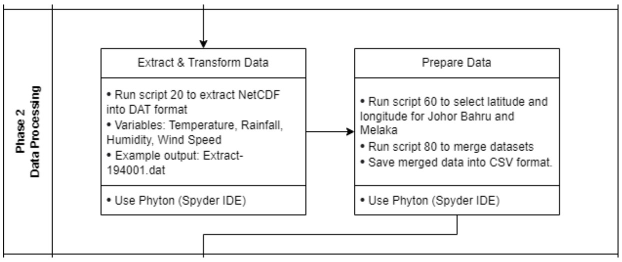

Phase 2
Data Processing
Extract & Transform Data
• Run script 20 to extract NetCDF into DAT format
• Variables: Temperature, Rainfall, Humidity, Wind Speed
• Example output: Extract-194001.dat
• Use Phyton (Spyder IDE)
Prepare Data
• Run script 60 to select latitude and longitude for Johor Bahru and Melaka
• Run script 80 to merge datasets
• Save merged data into CSV format.
• Use Phyton (Spyder IDE)

Figure 5: Activities in Data Processing Phase

The second phase involves extracting and transforming the raw data into a usable format. Using Python in the Spyder IDE, a script (script 20) is run to convert the NetCDF files into DAT format, maintaining the key climate variables. Another script (script 60) is then used to select specific latitude and longitude data for Johor Bahru and Melaka. The datasets are merged and saved in CSV format, preparing the data for analysis in the subsequent phases.

C. *Data Analysis and Modelling*

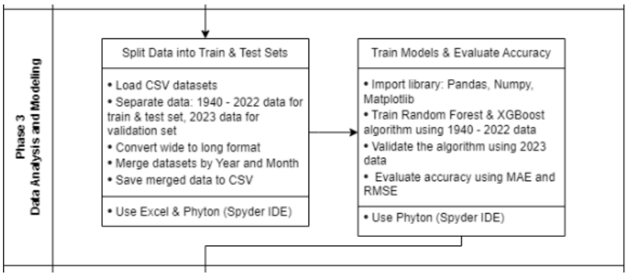

Phase 3
Data Analysis and Modeling
Split Data into Train & Test Sets
• Load CSV datasets
• Separate data: 1940 - 2022 data for train & test set, 2023 data for validation set
• Convert wide to long format
• Merge datasets by Year and Month
• Save merged data to CSV
• Use Excel & Phyton (Spyder IDE)
Train Models & Evaluate Accuracy
• Import library: Pandas, Numpy, Matplotlib
• Train Random Forest & XGBoost algorithm using 1940 - 2022 data
• Validate the algorithm using 2023 data
• Evaluate accuracy using MAE and RMSE
• Use Phyton (Spyder IDE)

Figure 6: Activities in Data Analysis and Modelling Phase

This critical phase focuses on preparing the data for modeling and conducting the analysis. The CSV datasets are loaded and split into training (1940-2022) and testing (2023) sets. The data is converted from wide to long format and merged by year and month. Using Python libraries such as Pandas, Numpy, and Matplotlib, Random Forest and XGBoost algorithms are trained on the historical data. The models are then validated using the 2023 data, with accuracy evaluated using MAE, RMSE and R-squared metrics.

D. *Data Analysis and Modelling*

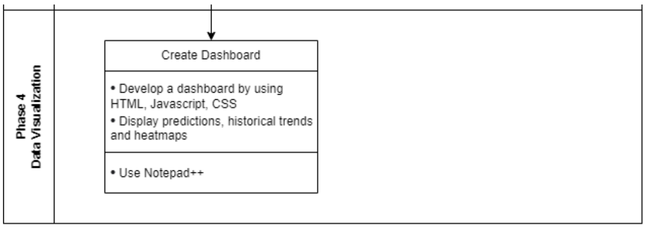

Phase 4
Data Visualization
Create Dashboard
• Develop a dashboard by using HTML, Javascript, CSS
• Display predictions, historical trends and heatmaps
• Use Notepad++

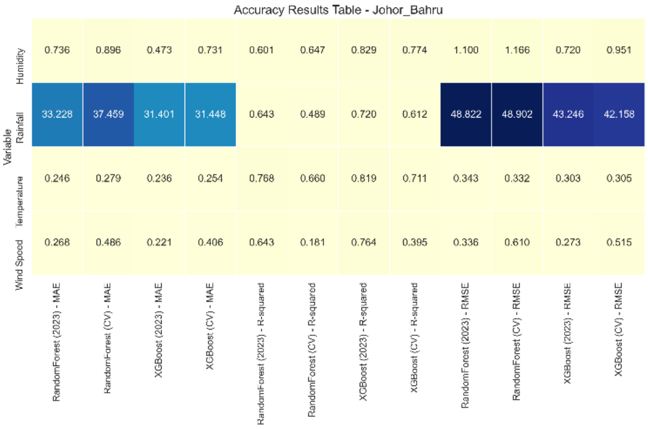
Figure 7: Activities in Data Visualization Phase

The final phase involves creating a dashboard to visualize the results. Using HTML, JavaScript, and CSS in Notepad++, a dashboard is developed to display predictions, historical trends, and heatmaps. This visualization tool allows for an intuitive presentation of the analysis results, making it easier to interpret the climate predictions and trends for the cultural heritage sites in question.

IV. RESULT AND ANALYSIS

A. *Accuracy Results*

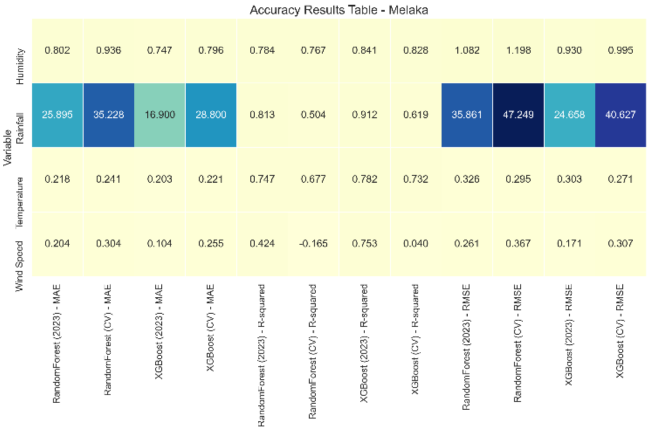
In investigating suitable machine learning algorithms for microclimate data analysis, both Random Forest and XGBoost demonstrate effectiveness, but their performance varies across different climate variables and locations. For temperature prediction, both algorithms show high accuracy with R-squared values ranging from 0.660 to 0.819, indicating strong predictive power. XGBoost generally outperforms Random Forest in temperature prediction, especially for Johor Bahru (R-squared of 0.819 vs 0.768).


Accuracy Results Table - Johor_Bahru

Variable	RandomForest (2023) - MAE	RandomForest (CV) - MAE	XGBoost (2023) - MAE	XGBoost (CV) - MAE	RandomForest (2023) - R-squared	RandomForest (CV) - R-squared	XGBoost (2023) - R-squared	XGBoost (CV) - R-squared	RandomForest (2023) - RMSE	RandomForest (CV) - RMSE	XGBoost (2023) - RMSE	XGBoost (CV) - RMSE
Humidity	0.736	0.896	0.473	0.731	0.601	0.647	0.829	0.774	1.100	1.166	0.720	0.951
Rainfall	33.228	37.459	31.401	31.448	0.643	0.489	0.720	0.612	48.822	48.902	43.246	42.158
Temperature	0.246	0.279	0.236	0.254	0.768	0.660	0.819	0.711	0.343	0.332	0.303	0.305
Wind Speed	0.268	0.486	0.221	0.406	0.643	0.181	0.764	0.395	0.336	0.610	0.273	0.515

Figure 8: Accuracy Results Table for Johor Bahru

The accuracy results for Johor Bahru show varying performance across different climate variables and models. For temperature prediction, both Random Forest and XGBoost models perform well, with R-squared values ranging from 0.711 to 0.819, indicating good predictive power. The MAE for temperature is low (around 0.25°C), suggesting accurate predictions. Humidity predictions show moderate accuracy, with R-squared values between 0.601 and 0.829. The models struggle more with rainfall predictions, as evidenced by the higher MAE and RMSE values. This is typical for rainfall due to its highly variable nature. Wind speed predictions show mixed results, with R-squared values ranging from 0.181 to 0.764, indicating that some models (particularly XGBoost) perform better than others for this variable. Overall, the Random Forest model using cross-validation (CV) tends to perform slightly better across most variables, especially for temperature and humidity predictions.



Accuracy Results Table - Melaka

Variable	RandomForest (2023) - MAE	RandomForest (CV) - MAE	XGBoost (2023) - MAE	XGBoost (CV) - MAE	RandomForest (2023) - R-squared	RandomForest (CV) - R-squared	XGBoost (2023) - R-squared	XGBoost (CV) - R-squared	RandomForest (2023) - RMSE	RandomForest (CV) - RMSE	XGBoost (2023) - RMSE	XGBoost (CV) - RMSE
Humidity	0.802	0.936	0.747	0.796	0.784	0.767	0.841	0.828	1.082	1.198	0.930	0.995
Rainfall	25.895	35.228	16.900	28.800	0.813	0.504	0.912	0.619	35.861	47.249	24.658	40.627
Temperature	0.218	0.241	0.203	0.221	0.747	0.677	0.782	0.732	0.326	0.295	0.303	0.271
Wind Speed	0.204	0.304	0.104	0.255	0.424	-0.165	0.753	0.040	0.261	0.367	0.171	0.307

Figure 9: Accuracy Results Table for Melaka

For Melaka, the accuracy results show some similarities and differences compared to Johor Bahru. Temperature predictions are quite accurate, with R-squared values ranging from 0.677 to 0.782 and low MAE values (around 0.22°C). Humidity predictions show good accuracy, with R-squared values between 0.767 and 0.841, which is slightly better than Johor Bahru. Rainfall predictions in Melaka show a notable improvement compared to Johor Bahru, with higher R-squared values (up to 0.912 for XGBoost) and generally lower MAE and RMSE values. This suggests that the models are better at capturing rainfall patterns in Melaka. However, wind speed predictions in Melaka show more variability and lower accuracy compared to Johor Bahru, with R-squared values ranging from -0.165 to 0.753. The negative R-squared for the Random Forest (CV) model indicates poor performance for wind speed prediction. In Melaka, the XGBoost model using cross-validation tends to perform better across most variables, particularly for rainfall and humidity predictions.

The cross-validation results (CV) generally align with the 2023 test set results, suggesting that the models are robust and not overfitting. However, there are some discrepancies, particularly in wind speed predictions, which may indicate changing patterns or the need for more data to improve model stability.

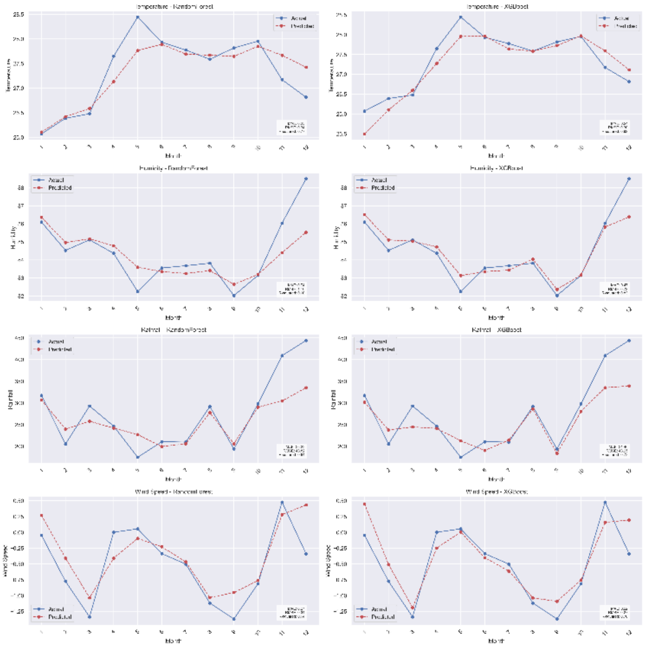


Figure 10: Actual vs Predicted Line Graph for Johor Bahru in 2023

For Johor Bahru, both Random Forest and XGBoost models demonstrate strong performance in predicting temperature trends, with XGBoost showing a slight edge, particularly in capturing peak temperatures. However, both models struggle to accurately predict the sharp temperature drops in the final months of the year.

In terms of humidity predictions, XGBoost outperforms Random Forest, especially in capturing the significant humidity increase towards the year's end. Random Forest tends to underestimate humidity fluctuations, particularly in the latter half of the year.

Rainfall predictions prove challenging for both models, especially during extreme events in the last quarter of the year. While XGBoost shows marginally better performance in capturing overall rainfall patterns, both models significantly underestimate the high rainfall in the final three months.

For wind speed, both models capture the general trend but struggle with extreme values. XGBoost demonstrates slightly higher accuracy, particularly in the middle months, but both models have difficulty predicting sharp wind speed changes, especially in the last quarter of the year.

Overall, for Johor Bahru, XGBoost consistently outperforms Random Forest across all variables, albeit by a small margin. It shows particular strength in capturing subtle data variations.

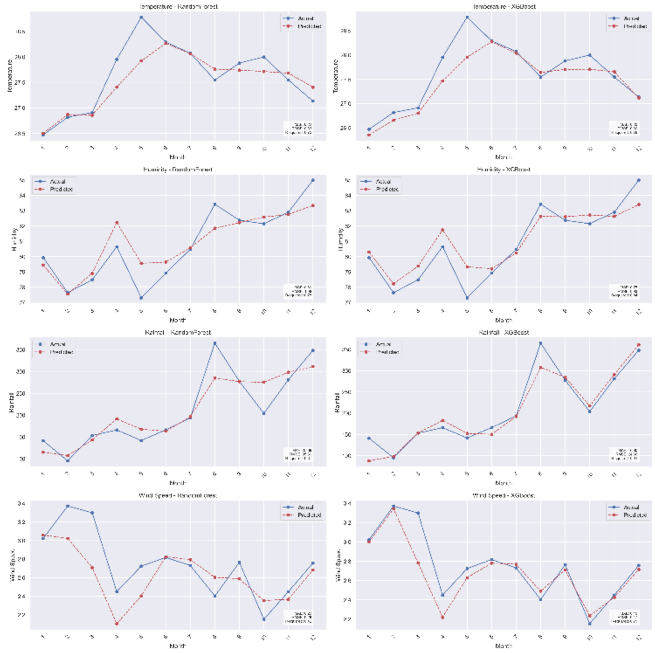


Figure 11: Actual vs Predicted Line Graph for Melaka in 2023

In Melaka, both Random Forest and XGBoost models perform similarly well in temperature prediction, accurately capturing the overall trend including the mid-year peak. XGBoost shows a slight advantage in accuracy, particularly in the latter half of the year.

Humidity predictions in Melaka prove challenging for both models, with difficulties in capturing sharp fluctuations. While XGBoost appears marginally better at tracking the overall trend, both models miss extreme low and high humidity points, particularly in the middle and end of the year.

Rainfall predictions in Melaka are complicated by extreme events, especially in months 8 and 12. XGBoost demonstrates a slight edge in capturing the overall rainfall pattern, but both models struggle with accurately predicting high rainfall months.

Wind speed predictions in Melaka are particularly challenging, with both models struggling to accurately capture peaks and troughs. XGBoost shows marginally better performance, especially in the latter half of the year, but both models miss significant wind speed drops in the middle and end of the year.

For Melaka, the performance difference between Random Forest and XGBoost is less pronounced compared to Johor Bahru. However, XGBoost still maintains a slight overall advantage across the variables.

In conclusion, XGBoost generally outperforms Random Forest for both locations, though the margin is often small. Both models face significant challenges in predicting extreme weather events and sharp fluctuations across all variables. Temperature predictions are generally the most accurate, while rainfall and wind speed predictions present the greatest difficulties. These results highlight the complex nature of weather prediction and the ongoing challenges in accurately forecasting diverse climate variables.

B. Correlation Matrix

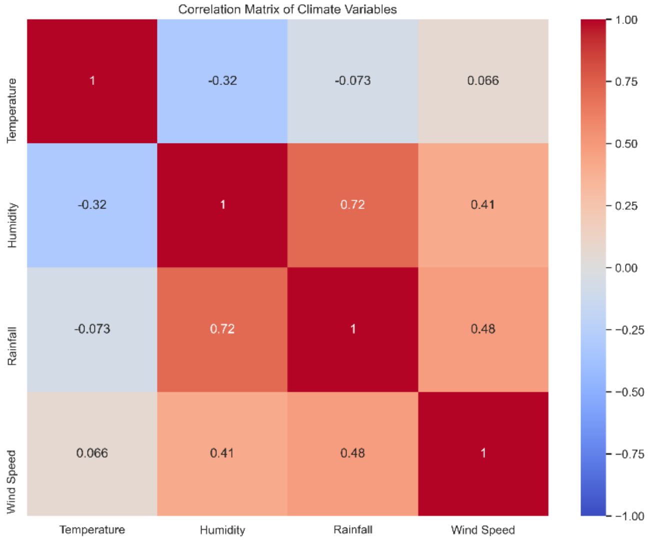


Figure 12: Correlation Matrix of Climate Variables in Johor Bahru

Firstly, there's a moderate negative correlation (-0.32) between temperature and humidity. This suggests that as temperature increases, humidity tends to decrease slightly, which is a common relationship in many climates. Next, there's a strong positive correlation (0.72) between rainfall and humidity. This indicates that higher humidity is often associated with increased rainfall, which is expected as moisture in the air contributes to precipitation. Wind speed has weak positive correlations with humidity (0.41) and rainfall (0.48), suggesting that windier conditions are slightly associated with higher humidity and more rainfall. There's a very weak negative correlation (-0.073) between temperature and rainfall, indicating that there's almost no linear relationship between these variables in Johor Bahru.

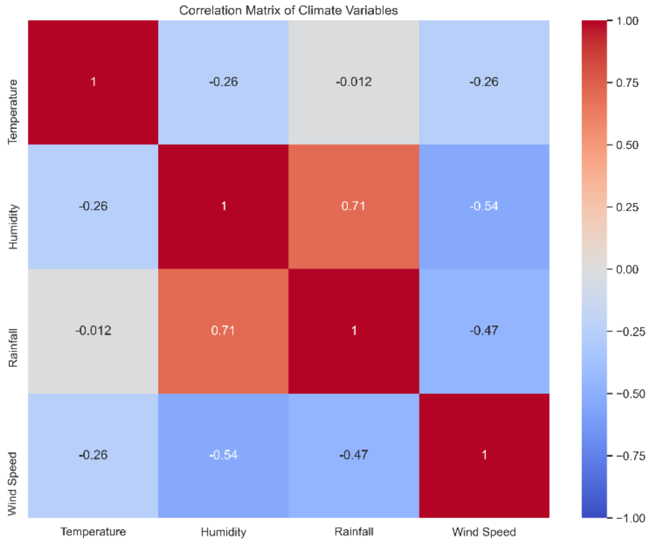


Figure 13: Correlation Matrix of Climate Variables in Melaka

Similar to Johor Bahru, there's a negative correlation (-0.26) between temperature and humidity, though it's slightly weaker. There's a strong positive correlation (0.71) between rainfall and humidity, very similar to Johor Bahru, reinforcing the relationship between moisture in the air and precipitation. Interestingly, wind speed in Melaka shows negative correlations with all other variables, particularly strong with humidity (-0.54) and rainfall (-0.47). This suggests that windier conditions in Melaka are associated with lower humidity, less rainfall, and slightly lower temperatures. There's practically no correlation (-0.012) between temperature and rainfall in Melaka, indicating that these variables don't have a linear relationship.

The relationship between humidity and rainfall is consistently strong and positive in both locations, highlighting the importance of humidity in precipitation processes. The negative correlation between temperature and humidity is present in both locations, though slightly stronger in Johor Bahru. The most notable difference is in the wind speed relationships. In Johor Bahru, wind speed is positively correlated with humidity and rainfall, while in Melaka, these correlations are negative. This could indicate different local climate dynamics or geographical influences affecting wind patterns and their relationship to other weather variables. Both locations show very weak or no correlation between temperature and rainfall, suggesting that other factors may be more important in determining rainfall patterns in these areas.

C. *Dashboard Development*

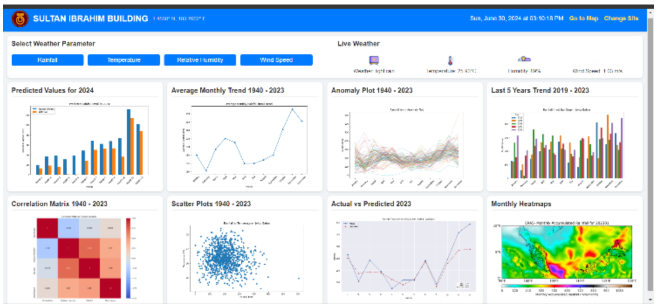


Figure 14: Microclimate Monitoring and Prediction Dashboard Johor Bahru

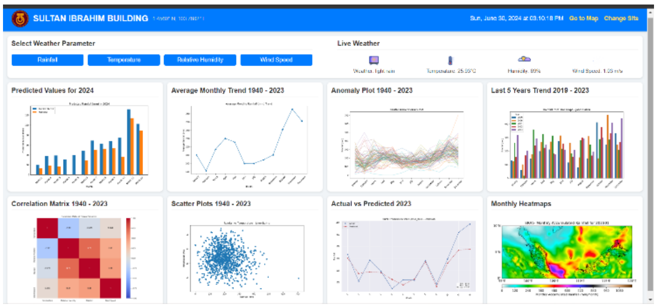


Figure 15: Microclimate Monitoring and Prediction Dashboard Melaka

The dashboard was developed using HTML, JavaScript, and CSS to provide an intuitive visualization of the microclimate data and predictions for the cultural heritage sites. It features multiple interactive components, including selectable weather parameters, live weather display, and various charts showing predicted values, historical trends, and correlations between climate variables. The dashboard presents data through bar charts, line graphs, scatter plots, and heatmaps, offering a comprehensive view of both historical patterns and future predictions. Key elements include a actual vs predicted value for 2023, prediction for 2024, average monthly trends from 1940-2023, anomaly plots, and a correlation matrix to illustrate relationships between different weather parameters. The interface allows users to easily switch between different heritage sites and weather variables, making it a versatile tool for researchers and site managers to analyze microclimate conditions and trends relevant to cultural heritage preservation.

V. CONCLUSION

ACKNOWLEDGEMENT

In preparing this thesis, I was in contact with many people, seniors, supervisors, and friends. They have contributed to my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main thesis supervisor, Assoc. Prof. Dr. Mohd Shahizan Othman, for the encouragement, guidance, critics and enormous patience throughout his sharing knowledge of completion of this thesis. Without his continued support and interest, this thesis would not have been the same as presented here.

REFERENCE

[1] Alang Othman, M., Ghani, A. A., & Alang Othman, M. S. (2020). Distribution of rainfall events in northern region of Peninsular Malaysia. Paper presented at the IOP Conference Series: Earth and Environmental Science.

[2] Ali, P. J. M., Faraj, R. H., Koya, E., Ali, P. J. M., & Faraj, R. H. (2014). Data normalization and standardization: a technical report. Mach Learn Tech Rep, 1(1), 1-6.

[3] Armain, M. Z. S., Hassan, Z., & Harun, S. (2021). Climate change impact under CanESM2 on future rainfall in the state of Kelantan using Artificial Neural Network. Paper presented at the IOP Conference Series: Earth and Environmental Science.

[4] Dubey, A. D. (2015, 21-22 Dec. 2015). K-Means based radial basis function neural networks for rainfall prediction. Paper presented at the 2015 International Conference on Trends in Automation, Communications and Computing Technology (I-TACT-15).

[5] El Abbassi, M., Overbeck, J., Braun, O., Calame, M., van der Zant, H. S. J., & Perrin, M. L. (2021). Benchmark and application of unsupervised classification approaches for univariate data. Communications Physics, 4(1). doi:10.1038/s42005-021-00549-9

[6] Engel, T., Charão, A., Kirsch Pinheiro, M., & Steffenel, L. A. (2015). Performance improvement of data mining in Weka through multi-core and GPU acceleration: opportunities and pitfalls. Journal of Ambient Intelligence and Humanized Computing, 6, 377-390. doi:10.1007/s12652-015-0292-9

[7] Hidayat, R., Yanto, I. T. R., Ramli, A. A., Fudzee, M. F. M., & Ahmar, A. S. (2021). Generalized normalized euclidean distance based fuzzy soft set similarity for data classification. Computer Systems Science and Engineering, 38(1), 119-130. doi:10.32604/CSSE.2021.015628

[8] Hu, W., & he Pan, Q. (2015). Data clustering and analyzing techniques using hierarchical clustering method. Multimedia Tools and Applications, 74(19), 8495-8504. doi:10.1007/s11042-013-1611-9

[9] Jenitha, G., & Vennila, V. (2014, 8-8 July 2014). Comparing the partitional and density-based clustering algorithms by using WEKA tool. Paper presented at the Second International Conference on Current Trends In Engineering and Technology - ICCTET 2014.

[10] Krishnaveni, N., & Padma, A. (2020). Weather forecast prediction and analysis using sprint algorithm. Journal of Ambient Intelligence and Humanized Computing. doi:10.1007/s12652-020-01928-w

[11] Kwon, O. H., & Park, S. H. (2016, 2016//). Identification of Influential Weather Factors on Traffic Safety Using K-means Clustering and Random Forest. Paper presented at the Advanced Multimedia and Ubiquitous Engineering, Singapore.

[12] Litoriya, R. (2012). Comparison of the various clustering algorithms of weka tools. 2, 73-80.

[13] Mahmud, M., & Kumar, T. (2008). Forecasting severe rainfall in the equatorial Southeast Asia. Geofizika, 25(2), 109-128. Retrieved from <Go to ISI>://WOS:000263081500002

[14] Mehta, K., Modi, S., Tomy, C., & Singh, A. (2019, 29-30 March 2019). Analysis of Stocks by the Use of Clustering and Classification Algorithms. Paper presented at the 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN).

[15] Mercioni, M. A., & Ş, H. (2018, 8-9 Nov. 2018). Evaluating hierarchical and non-hierarchical grouping for develop a smart system. Paper presented at the 2018 International Symposium on Electronics and Telecommunications (ISETC).

[16] Muhammad, N. S., Abdullah, J., & Julien, P. Y. (2020). Characteristics of Rainfall in Peninsular Malaysia. Paper presented at the Journal of Physics: Conference Series.

[17] Naik, A., & Samant, L. (2016). Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime. Procedia Computer Science, 85, 662-668. doi:https://doi.org/10.1016/j.procs.2016.05.251

[18] Naseem, R., Deris, M. M., Maqbool, O., & Shahzad, S. (2019). Euclidean space based hierarchical clusterers combinations: an application to software clustering. Cluster Computing, 22(3), 7287-7311. doi:10.1007/s10586-017-1408-0

[19] Nidheesh, N., Nazeer, K. A. A., & Ameer, P. M. (2020). A Hierarchical Clustering algorithm based on Silhouette Index for cancer subtype discovery from genomic data. Neural Computing and Applications, 32(15), 11459-11476. doi:10.1007/s00521-019-04636-5

[20] Olatayo, T., & Taiwo, A. (2019). Statistical Modelling and Prediction of Rainfall Time Series Data Statistical Modelling and Prediction of Rainfall Time Series Data.

[21] Parmar, A., Chauhan, D., & Bansal, K. L. (2017). PERFORMANCE EVALUATION OF WEKA CLUSTERING ALGORITHMS ON LARGE DATASETS. International Journal of Advanced Research, 5, 2209-2216. doi:10.21474/IJAR01/4661

[22] Peng, Y., Zhang, Y., Kou, G., Li, J., & Shi, Y. (2012). Multicriteria Decision Making Approach for Cluster Validation. Procedia Computer Science, 9, 1283-1291. doi:https://doi.org/10.1016/j.procs.2012.04.140

[23] Preetha, V. (2021, 8-10 April 2021). Data Analysis on Student's Performance based on Health status using Genetic Algorithm and Clustering algorithms. Paper presented at the 2021 5th International Conference on Computing Methodologies and Communication (ICCMC).

[24] Saad, S. M., & Ismail, N. (2020). Multifractal properties of temporal rainfall series in peninsular Malaysia. Paper presented at the IOP Conference Series: Earth and Environmental Science.

[25] Shah, C., & Jivani, A. (2013, 28-30 Nov. 2013). Comparison of data mining clustering algorithms. Paper presented at the 2013 Nirma University International Conference on Engineering (NUiCONE).

[26] Suhaila, J., & Jemain, A. A. (2009). Investigating the impacts of adjoining wet days on the distribution of daily rainfall amounts in Peninsular Malaysia. Journal of Hydrology, 368(1), 17-25. doi:https://doi.org/10.1016/j.jhydrol.2009.01.022

[27] Teodoro, P. E., de Oliveira-Júnior, J. F., da Cunha, E. R., Correa, C. C. G., Torres, F. E., Bacani, V. M., . . . Ribeiro, L. P. (2016). Cluster analysis applied to the spatial and temporal variability of monthly rainfall in Mato Grosso do Sul State, Brazil. Meteorology and Atmospheric Physics, 128(2), 197-209. doi:10.1007/s00703-015-0408-y

[28] Tiwari, M. (2012). Performance analysis of Data Mining algorithms in Weka. IOSR Journal of Computer Engineering, 6, 32-41. doi:10.9790/0661-0633241

[29] Wu, J. (2012a). Cluster Analysis and K-means Clustering: An Introduction. In J. Wu (Ed.), Advances in K-means Clustering: A Data Mining Thinking (pp. 1-16). Berlin, Heidelberg: Springer Berlin Heidelberg.

[30] Wu, J. (2012b). The Uniform Effect of K-means Clustering. In J. Wu (Ed.), Advances in K-means Clustering: A Data Mining Thinking (pp. 17-35). Berlin, Heidelberg: Springer Berlin Heidelberg.

[31] Xin, F., & Abraham, Z. (2012, 2012//). Extreme Value Prediction for Zero-Inflated Data. Paper presented at the Advances in Knowledge Discovery and Data Mining, Berlin, Heidelberg.

[32] Ye, L., Jabbar, S. F., Abdul Zahra, M. M., & Tan, M. L. (2021). Bayesian Regularized Neural Network Model Development for Predicting Daily Rainfall from Sea Level Pressure Data: Investigation on Solving Complex Hydrology Problem. Complexity, 2021, 6631564. doi:10.1155/2021/6631564

[33] Zhou, K., & Yang, S. (2020). Effect of cluster size distribution on clustering: a comparative study of k-means and fuzzy c-means clustering. Pattern Analysis and Applications, 23(1), 455-466. doi:10.1007/s10044-019-00783-6