APPLIED INFORMATION TECHNOLOGY AND COMPUTER SCIENCE
e-ISSN: 2773-5141

AITCS

Vol. 5 No. 2 (2024) 446-460
https://publisher.uthm.edu.my/periodicals/index.php/aitcs

Comparative Analysis of Machine Learning Algorithms on Temperature and Wind Speed Prediction

Pang Chin Hoo[1], Noor Zuraidin Mohd Safar[2*]

[1,,2] *Fakulti Sains Komputer dan Teknologi Maklumat,*
Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, 86400, MALAYSIA

*Corresponding Author: zuraidin@uthm.edu.my
DOI: https://doi.org/10.30880/aitcs.2024.05.02.024

Article Info

Received: 13 June 2024
Accepted: 28 September 2024
Available online: 15 December 2024

Keywords

Weather forecasting, Random Forest, Multiple Linear Regression, meteorological data, performance evaluation

Abstract

This research explores weather forecasting, emphasizing the effectiveness machine learning algorithms in addressing the challenges posed by traditional forecasting techniques. The objective is to evaluate and compare the performance of two machine learning algorithms, Random Forest and Multiple Linear Regression. This research focuses on predicting temperature and wind speed by utilizing two datasets, one from three meteorology stations in Batu Pahat, and another from Kuala Lumpur obtained from Oikolab, covering hourly data from 2018 to 2020. The data is pre-processed and split into 70% training and 30% testing subsets. The performance of the two selected algorithms were assessed using Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), and R-Squared (R^2). The results indicate that Random Forest outperformed Multiple Linear Regression, showing higher R-squared values and lower error rates, thus demonstrating its superior ability to handle complex meteorological data.

1. Introduction

Weather forecasting is the use of science technology to predicts the atmospheric conditions for a certain area and time frame [1]. Weather forecasting plays a pivotal role in various sectors, including agriculture, transportation, disaster management, and daily life planning. Weather forecasting is predicting how the air will change based on observations of the current environment made from the ground using tools like satellites, radiosondes, Doppler radar, boats, and airplanes. Traditional weather forecasting methods rely on numerical models, which can be highly computational and inaccurate in certain circumstances. There are two types of machine learning, supervised and unsupervised learning [2]. Supervised learning involves predetermined output attribute besides the use of input attributes, which is used to classify or predict outcome accurately. In contrast, unsupervised learning involves pattern recognition without the involvement of a target attribute, which is used to analyze and cluster unlabeled datasets. The intricacy of atmospheric science and the constraints of available data and technology present a number of difficulties and obstacles for machine learning-based weather forecasting. Therefore, the purpose of this project is to perform and compare machine learning algorithms that is related to weather forecasting.

The primary challenge in weather forecasting is achieving accurate predictions for meteorological parameters such as temperature, precipitation, humidity, and wind speed. Traditional forecasting methods struggle with the complex and nonlinear relationships within weather data, often leading to reduced prediction accuracy. Furthermore, the accuracy of short-term weather forecasts, which are crucial for daily planning and

various industries, heavily depends on the quantity and quality of the data used to train machine learning models. Inadequate data can negatively impact economic sectors like agriculture and transportation.

The primary objective of this research is to investigate the application and effectiveness of machine learning algorithms in improving the accuracy of weather forecasting. This study will utilize datasets from the meteorology office and Oikolab to provide a broad and thorough foundation for analysis. This research also aims to predict temperature and wind speed by using sophisticated modeling techniques. MATLAB will be employed as the analytical tool to implement and evaluate the chosen machine learning algorithms. Additionally, the effectiveness of the Random Forest and Multiple Linear Regression algorithms in precisely capturing the complexities and dynamics of weather patterns will be compared in this study.

The expected outcome of this research is to make a comprehensive study on machine learning algorithms for weather forecasting. This research also includes a comprehensive study on machine learning algorithms for weather forecasting. The outcomes of applying machine learning techniques to the input data will be evaluated in order to assess the models' effectiveness and precision in forecasting weather patterns.

2. Related Work

This section discussed information about weather forecasting, machine learning and its categories and the existing research on weather forecasting using machine learning algorithm.

2.1 Weather Forecasting

The process of predicting weather conditions for the future is known as weather forecasting [3]. It involves the use of scientific principles and technology to analyze and interpret atmospheric data, with the aim of providing accurate and timely information about upcoming weather conditions. The main objective of weather forecasting is to assist people in planning and making educated decisions based on expected weather patterns for themselves, their communities, and different industries including emergency management, transportation, and agriculture. A mix of computer simulations, mathematical models, and observational data are used in this forecasting procedure.

There are 3 main types of weather forecasting that exist in this world, they are short-range forecasting, medium-range forecasting and long-range forecasting. Short-range forecasts typically cover a period of up to 48 hours or 2 days. This type of weather forecast focuses on immediate weather conditions and are crucial for daily planning and activities. It relies heavily on real-time observational data from weather stations, satellites and radar. Next, medium-range forecasts extend from about 3 to 10 days. This type of weather forecast provides a broader view of expected weather patterns and are valuable for planning events, travel and agricultural activities. Besides, long-range forecasts often stretch beyond 10 days and can last for many months. They give a general view for trends and climatic patterns, but are less precise compared to short and medium-term predictions.

2.2 Machine Learning

Machine learning is an area of AI that focuses on building algorithms and computational models that allow computers to learn and make predictions or decisions without being explicitly programmed. The basic principle underlying machine learning is to give computers the ability to recognize patterns, learn from data, and improve their performance over time. In conventional programming, humans directly set rules and codes to guide computers on how to complete a task. Machine learning systems, on the other hand, learn from data and examples, allowing them to generalize and make predictions or conclusions on new and unseen data.

There are six main components in a generic model of machine learning which are shown in Figure 1, they are collection and preparation of data, feature selection, choice of algorithm, selection of models and parameters, training and performance evaluation [4]. The first process is collection and preparation of data, where the main task in machine learning is to collect and prepare data as input to the algorithms. These data need to be cleaned and pre-processed to a structured format in order to improve the quality of data so that the machine learning model can interpret it correctly. The second process is feature selection, where the unwanted features need to be removed in data preprocessing. The third process is choice of algorithm, where choosing the best algorithm for the problem to get the best possible results. The fourth process is selection of model and parameters, where most of the machine learning algorithms need user interaction to choose the best values for certain parameters. The fifth process is training, where the model needs to be trained based on the selected algorithm. The last process is performance evaluation, where the model must test to evaluate on performance parameters such as accuracy, precision and recall.

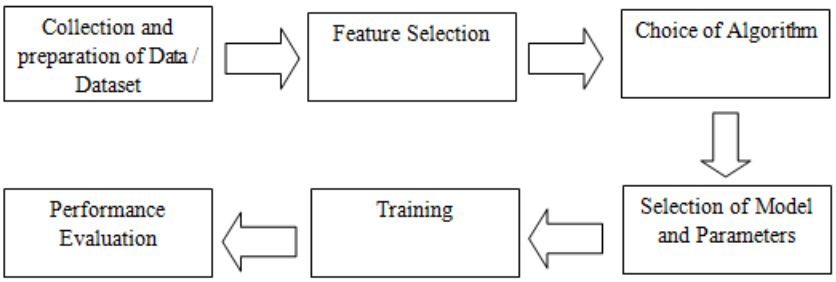

Fig.1 *Component of a generic machine learning*

There are two main categories of machine learning, supervised and unsupervised learning. Supervised learning is a type of machine learning that trains computers to anticipate outcomes and detect patterns using labeled information. There are two common tasks in supervised learning which are classification and regression, where classification separates the data and regression fits the data. Fig. 2 shows the workflow of supervised learning.

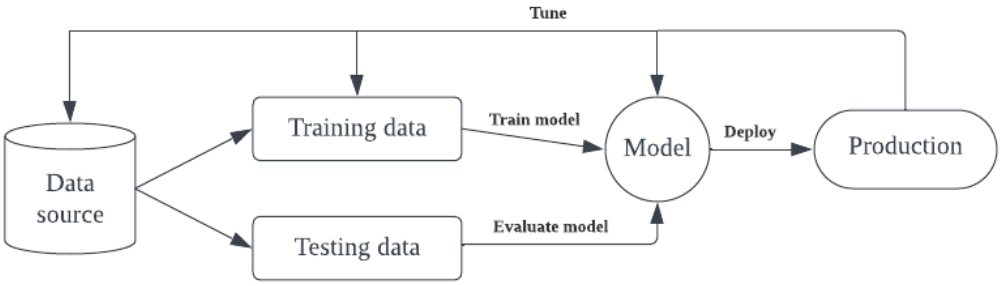

Fig. 2 *Supervised machine learning workflow*

Unlike supervised learning, unsupervised learning examines unlabeled datasets without requiring human intervention. The purpose of unsupervised learning is to discover the underlying structure of a dataset, categorize it based on similarities, and display it in a compressed format. There are some common tasks in unsupervised learning such as clustering, density estimation, dimensionality reduction and so on. Fig. 3 shows the workflow of unsupervised learning.

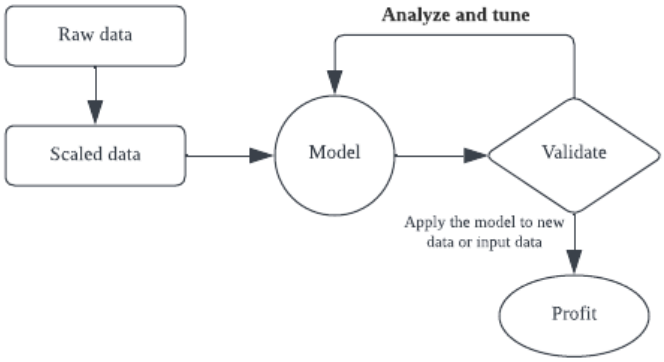

Fig. 3 *Unsupervised machine learning workflow*

2.3 Algorithm used in weather forecasting

There are many algorithms that can be used for weather forecasting and each of them has its own strength and weakness. The selected algorithms in this research are Random Forest and Multiple Linear Regression. These two algorithms are used to compare which algorithm has higher accuracy in predicting weather conditions. The description of these two algorithms will be explained in the next section.

Applied Information Technology and Computer Science Vol. 5 No. 2 (2024) p. 446-460

2.3.1 Random Forest

Random Forest, also known as Random Decision Forest (RFA), is a classification and regression algorithm that uses several decision trees to perform tasks such as classification and regression. It is classified as supervised learning and its main benefit is that it can be used for classification as well as regression. Random Forest is made up of a large number of decision trees (training samples) that are constructed in a randomized manner or as an ensemble. When a test dataset is applied to these decision trees, each decision tree makes a class prediction, and the class with the most votes is chosen as the model class [5]. Figure 4 shows the algorithm of random forest in machine learning.

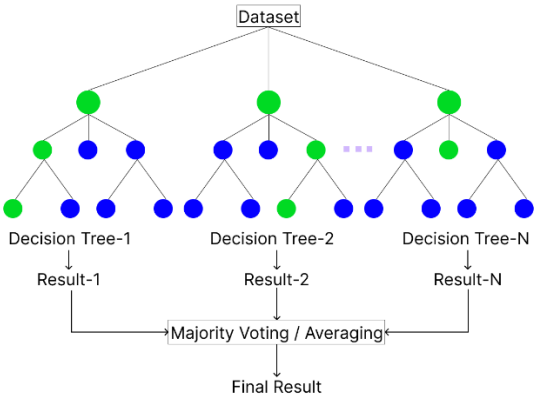

Fig. 4 *Random Forest Algorithm*

2.3.2 Multiple Linear Regression

Regression is an approach for supervised learning that is used to model and make predictions for continuous variables [6]. The purpose of univariate regression is to a linear equation that connects a dependent variable and a single independent variable by analyzing their relationship. Multiple linear regression is a statistical technique that model the relationship between a dependent variable and two or more independent variables. It expands the concept of simple linear regression to include more than one explanatory variable, which mean there are several independent variables (X_i) and one dependent variable (Y). This algorithm is commonly used to find the line or "best fit" curve for the dataset. This process shows how the features influence results [7]. Multiple linear regression can be expressed with the following equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \varepsilon \quad (1)$$

2.4 Comparative Study

This section shows the existing research study that are similar to this research paper. Many research has been done on weather prediction on different machine learning algorithms. These existing research studies have its own result but has the same research objectives. Table 1 shows the previous research studies about weather prediction using machine learning algorithm.

Table 1 *Previous research studies of weather prediction using machine learning*

Title of the research	Algorithm Used	Description	Result
Short Term Weather Forecasting Comparison Based on Machine Learning Algorithms [8]	Decision Tree, AdaBoost Regressor, Random Forest	This research seeks to determine which algorithms perform best in predicting weather conditions and analyze the potential benefits of integrating machine learning into meteorology. The parameters used are wind speed, relative humidity, dew point and air pressure.	Random Forest perform better result and accuracy compared to the other two algorithms

Table 1: (cont)

Analysis of Data Mining Techniques for Weather Prediction [9]	C.45 Decision Tree, Naïve Bayes	This research aims to extract meaningful information from the vast amount of weather-related data available, and use this information to make more accurate predictions about future weather conditions. The parameters used in this research study is temperature, humidity, wind speed, cloud cover and rainfall	The performance of C4.5 Decision Tree gives a better result compared to Naïve Bayes
Comparative Analysis of Machine Learning Algorithms for Weather Prediction using Error Detection [10]	Linear Regression, Random Forest, Polynomial Regression, Cart Regression	This study investigates the use of machine learning approaches in weather prediction, addressing the challenge of accurately forecasting weather patterns. The parameters used in this research are temperature, apparent temperature, humidity, wind speed, wind bearing and visibility	Random Forest algorithm gives the smallest value of error hence results in a better error detection compared to other 3 algorithms
Machine learning techniques to predict daily rainfall amount [11]	Multivariate Linear Regression, Random Forest, XGBoost Gradient Descent	The objective of this research project is to improve agricultural productivity and lessen the effects of irregular rainfall patterns in Bahir Dar City, Ethiopia by investigating the crucial relationship between atmospheric variables and daily rainfall intensity.	XGBoost obtained the lowest mean absolute error and root mean square error which are 3.58 and 7.85 respectively.
Machine Learning to Forecast Rainfall Intensity [12]	Multivariate Linear Regression, Random Forest, Neural Network	This research incorporates machine learning into a climate finance decision support system, concentrating on the effects of rainfall on Italian vineyards. It also uses machine learning to identify meteorological parameters that impact rainfall and predicts its intensity on a quarterly basis	Based on the performance findings, Random Forest performed better than both Linear Regression and Neural Network
A study on the evaluation of different regressors in Weather Prediction [13]	Linear Regression, Polynomial Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression	The study examines many machine learning regression models for the analysis of meteorological data sets in order to provide precise weather forecasts. The efficacy of each model is ascertained by comparing performance indicators such as MSE, RMSE, MAE, and R-squared.	The study found out that Random Forest and Decision Tree regressors generate less error rates and perform better when compared to other proposed machine learning models.

Table 1: (cont)

Prediction of Environmental Earth Surface Temperature using hybrid machine learning model [14]	Linear Regression, Random Forest Regressor	The study applied linear regression and random forest regression to predict temperatures using Berkeley Earth Climate Change temperature dataset, achieving accuracies of 99.52% and 99.58% respectively, with low mean square errors.	The overall results show that Random Forest performs better model fit compared to Linear Regression with 99.58%, where it scores 0.06% more than Linear Regression model

3. Methodology

3.1 Introduction

In this chapter, the research methods and procedures used for the machine learning in weather prediction will be discussed. Methodology is the systematic approach or set of principles guiding research or problem-solving in a particular field. It covers all of the methods, techniques, and procedures needed to provide dependable and consistent outcomes.

3.2 Research Framework

A research framework helps researchers organize their ideas, define the scope of the study, and direct the whole research process. For this reason, it is essential to a research report as it offers a systematic and structured approach to the investigation. This research will use Random Forest and Naïve Bayes to predict the temperature and windspeed in the proposed machine learning model. Figure 3 shows the research framework for chosen machine learning algorithm in weather prediction.

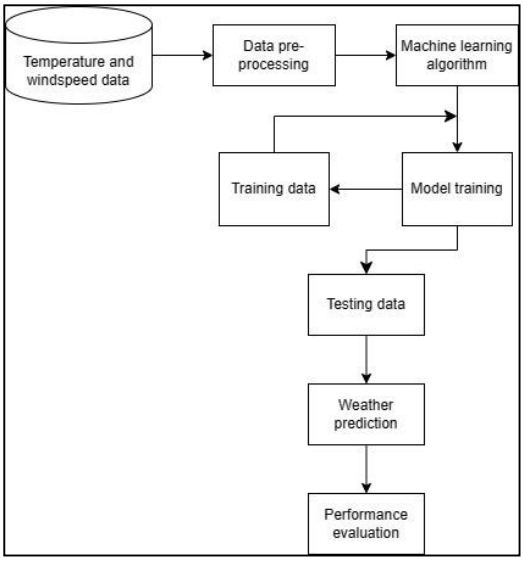

Fig. 5 *Research Framework for weather prediction*

3.3 Dataset Selection

Datasets are crucial to machine learning as without them, machine learning algorithms would have no reason to advance. In this study, 2 datasets will be used for performing the analysis. Dataset A is from 3 meteorology station in Batu Pahat. The selected dataset consists of 26304 hourly data values starting from year 2018 until year 2020 with parameters such as pressure, dry bulb temperature, dew point, relative humidity, mean surface wind and rainfall. Table 2 describes the information about the dataset A.

Table 2 *Dataset A Description*

Attributes	Type	Description
Year	Numerical	Year of the data that was recorded
Month	Numerical	Month of the data that was recorded
Day	Numerical	Day of the data that was recorded
Hour (MST)	Numerical	Hour of day in Mountain Standard Time
Pressure MSL (Hpa)	Numerical	Atmospheric pressure at mean sea level in hectopascals
Dry Bulb Temp. (^{0}C)	Numerical	Temperature measured by thermometer freely exposed to the air
Dew Point (^{0}C)	Numerical	Temperature when air becomes saturated with moisture and dew form
Relative Humidity (%)	Numerical	Percentage of moisture in air
Mean Surface Wind Direction (^{0})	Numerical	Average wind blowing direction
Mean Surface Wind Speed	Numerical	Average speed of wind in m/s
Rainfall Duration (min)	Numerical	Total rainfall duration in minutes
Rainfall Amount (mm)	Numerical	Total amount of rainfall measured in millimeters

For dataset B, the historical weather data is obtained from Oikolab, particularly in Kuala Lumpur. This dataset consists 26304 hourly weather data same to dataset A, starting from year 2018 until year 2020 with parameters such as relative humidity, mean sea level pressure, temperature, wind speed, precipitation and other parameters. Table 3 describes the information about the dataset B used in this study.

Table 3 *Dataset B Description*

Attributes	Type	Description
datetime (UTC)	Date	Date and time in Coordinated Universal Time (UTC)
temperature (^{0}C)	Numerical	Ambient temperature
dewpoint_temperature	Numerical	Temperature when air becomes saturated with moisture and dew form
relative_humidity (%)	Numerical	Percentage of moisture in air
wind_speed (m/s)	Numerical	Speed of wind in meter per second
wind_direction	Numerical	Wind blowing direction
total_cloud_cover	Numerical	Fraction of sky covered by clouds
total_precipitation (mm)	Numerical	Total amount of precipitation in millimeters
Mean_sea_level_pressure (Hpa)	Numerical	Atmospheric pressure at mean sea level in hectorpascals

Applied Information Technology and Computer Science Vol. 5 No. 2 (2024) p. 446-460

3.4 Features Selection

A critical stage in the creation of machine learning models is feature selection, particularly in the context of weather forecasting, where the choice of important meteorological features has a big influence on prediction accuracy. By utilizing initial findings from machine learning models, the features were chosen through a combination of feature relevance evaluations and correlation analyses. For Dataset A, the selected features for predicting dry bulb temperature, dewpoint, and mean surface wind speed include hour, pressure, relative humidity, and wind direction. On the other hand, Dataset B predicts temperature, dew point, and wind speed using pressure, relative humidity, and wind direction.

3.5 Data Pre-processing

Data pre-processing is one of the most crucial process in a training of machine learning in weather prediction. This process includes cleaning and converting raw data into a format appropriate for analysis and modelling. The goal of data pre-processing is to enhance the quality and accuracy of the data, making it more suitable for the specific requirements of the machine learning algorithm being used. The most obvious advantage of data pre-processing is that it enhances transparency in the machine learning process by analyzing every single stage and subsequently presenting a fairer and better model.

In this research study, the chosen dataset was examined for missing data and outliers using MATLAB. Generally, data pre-processing often entails processes such as data cleaning, data smoothing and data grouping. Data cleaning helps to find, clean and delete table rows with missing data. There are some functions that used throughout data cleaning process in MATLAB such as ismissing(), fillmissing() and rmmising(). Data smoothing is a method that helps reveal underlying patterns or trends in raw data by eliminating noise or unwanted fluctuations. Functions such as smoothdata(), movmean(), movmedian() and detrend() can be used during the data smoothing process.

3.6 Data Splitting

Data splitting is a process of splitting dataset into separate subsets, typically training set and testing set. This phase is crucial for the comparison and evaluation of the selected algorithms by identifying which models are the best for predicting weather. In this study, Dataset A and Dataset B will use the same ratio for training to testing which is 70:30, where 70% of the dataset will be used to train the models and the remaining 30% of the dataset will be used for evaluation purpose. Table 4 shows the dataset distribution according to 70:30 ratio for both datasets.

Table 4 *Distribution of data*

Dataset	Total Data	Training (70%)	Testing (30%)
A	26304	18143	7891
B	26304	18143	7891

3.7 Evaluation Metrics

In the world of machine learning, the performance of a machine learning model is evaluated using measurements known as evaluation metrics. These metrics offer numerical insights into the model performances in performing its predictions compared to the actual values.

3.7.1 Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is a commonly used metric in regression tasks to evaluate the performance of a machine learning model. This metric is the average of the absolute differences between the actual observation and the predicted outcome over the test sample, where each unique difference is given equal weight.

$$MAE = \frac{1}{N}\sum_{i=1}^{n}\left|x_i - \hat{x}\right| \tag{2}$$

3.7.2 Mean Absolute Error (MSE)

Mean Squared Error (MSE) is another commonly used metric in regression tasks to evaluate the performance of a machine learning model. The average squared difference between the data set's actual and predicted values is measured by MSE. It gives a measure of the average squared deviation of predictions from the actual values.

$$MSE = \frac{1}{N}\sum_{i=1}^{n}\left(x_i - \hat{x}\right)^2 \tag{3}$$

3.7.3 Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) used to measure the average magnitude of errors based on the quadratic scoring rule. RMSE is obtained by taking the square root of the MSE.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(x_i - \hat{x}\right)^2} \tag{4}$$

3.7.4 R-Squared (R^2)

R-squared is also known as coefficient of determination. It is a statistical measure used in machine learning to evaluate the goodness of fit of a model. It indicates the proportion of the variance for a dependent variable that is predictable from the independent variables. R-squared values lie between 0 and 1, where 0 indicates poor fitting line or the model does not fit with the dataset, while 1 indicates the best fitting.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(x_i - \hat{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \tag{4}$$

3.8 Hardware Requirement

Every experiment in this research study will be carried out using MATLAB tools in order to determine the achievement of the objectives. Therefore, a hardware is required for a successful completion of this entire research study. The hardware used in this research study is as shown in Table 3.

Table 3 *Hardware Requirement*

Hardware	Description
MSI GF-63 THIN	• Intel® Core™ i5-10500H @ 2.50Ghz • RAM: 16GB • Storage: 1TB SSD • Operating System: Windows 11 Pro 64-bit

3.9 Software Requirement

The data analysis tool that used in this research study is MATLAB. MATLAB, an acronym for MATrix LABoratory, is a high-performance programming language and environment that is mostly utilized for data analysis, visualization, and numerical computation. It was developed by MathWorks and it offers academics, engineers, and scientists a strong platform for working with data and resolving challenging mathematical issues. In this research study, MATLAB is mainly used to pre-process the data before the data is implemented into an experiment environment. Currently, there are 63 useful toolboxes which available in MATLAB. Next, MATLAB is also used for plotting chart to analyze and visualize the result. The system requirement of MATLAB for Windows operating system is shown in Table 4 as follows.

Table 4 *Software Requirement*

Configuration	Recommend Requirement
Operating System	• Windows 11 • Windows 10 (version 21H2 or higher) • Windows Server 2019 • Windows Server 2022

Applied Information Technology and Computer Science Vol. 5 No. 2 (2024) p. 446-460

Table 4: (cont)

Processor	Any Intel or AMD x86-64 processor with four or more cores and AVX2 instruction set support
RAM	16GB
Storage	3.8GB for just MATLAB 4-6GB for a typical installation

4. Result and Analysis

4.1 Introduction

This chapter will focus on the discussion of the obtained results and the analysis about the final result of the selected dataset for both algorithms, Random Forest and Multiple Linear Regression. The ultimate goal of this research is to compare the performance of Random Forest and Multiple Linear Regression for predicting temperature and wind speed and evaluate their performance based on accuracy and evaluation metrics. This chapter will present the dataset analysis findings in tables and graphs.

4.2 Evaluation

The performance of the 2 algorithms used in this study will be evaluated based on accuracy, MAE, MSE, RMSE and R-Squared.

4.2.1 Mean Absolute Error (MAE)

For Dataset A, Random Forest provides more accurate prediction compared to Multiple Linear Regression. Specifically, the MAE value for Random Forest shows a relatively lower error for the dry bulb temperature, dew point and mean surface wind speed, with values of 0.6800, 0.6669 and 0.3185 respectively. In contrast, Multiple Linear Regression gives higher MAE values of 1.0787 for dry bulb temperature, 1.1230 for dew point and 0.4465 for mean surface wind speed. Figure 6 illustrates the MAE values for Dataset A.

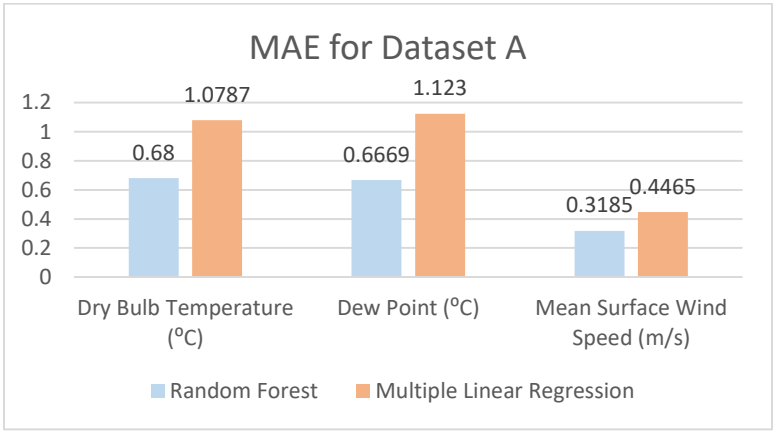

Fig. 6 *MAE value for Dataset A*

For Dataset B, Random Forest still perform better than Multiple Linear Regression. Specifically, the MAE value for Random Forest shows a relatively lower error for the temperature, dew point and wind speed, with values of 0.6736, 0.6505 and 0.4269 respectively. In contrast, Multiple Linear Regression gives higher MAE values of 0.7037 for temperature, 0.7417 for dew point and 0.4616 for wind speed. Figure 7 illustrates the MAE values for Dataset B.

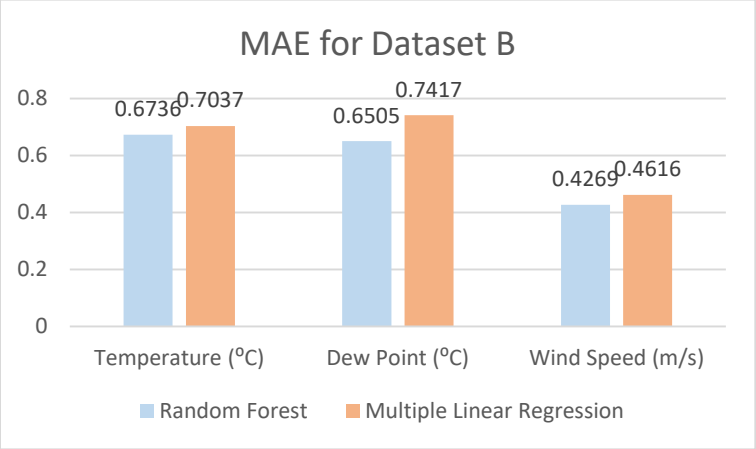

Fig. 7 *MAE value for Dataset B*

4.2.2 Mean Square Error (MSE)

For Dataset A, Random Forest provides more accurate prediction compared to Multiple Linear Regression. Specifically, the MSE value for Random Forest shows a relatively lower error for the dry bulb temperature, dew point and mean surface wind speed, with values of 0.7908, 0.7600 and 0.2102 respectively. In contrast, Multiple Linear Regression gives higher MSE values of 2.1205 for dry bulb temperature, 2.3592 for dew point and 0.3162 for mean surface wind speed. Figure 8 illustrates the MSE values for Dataset A.

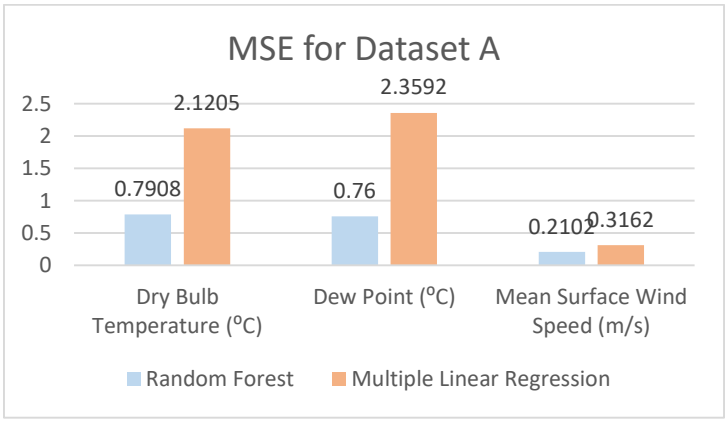

Fig. 8 *MSE value for Dataset A*

For Dataset B, Random Forest still perform better than Multiple Linear Regression. Specifically, the MSE value for Random Forest shows a relatively lower error for the temperature, dew point and wind speed, with values of 0.7542, 0.7049 and 0.3122 respectively. In contrast, Multiple Linear Regression gives higher MSE values of 0.8219 for temperature, 0.9123 for dew point and 0.3663 for wind speed. Figure 9 illustrates the MSE values for Dataset B.

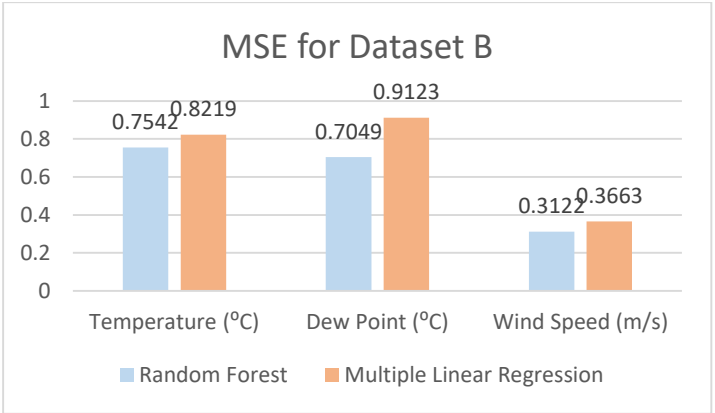

Fig. 9 *MSE value for Dataset B*

Applied Information Technology and Computer Science Vol. 5 No. 2 (2024) p. 446-460

4.2.3 Root Mean Square Error (RMSE)

For Dataset A, Random Forest provides more accurate prediction compared to Multiple Linear Regression. Specifically, the RMSE value for Random Forest shows a relatively lower error for the dry bulb temperature, dew point and mean surface wind speed, with values of 0.8893, 0.8717 and 0.4585 respectively. In contrast, Multiple Linear Regression gives higher RMSE values of 1.4562 for dry bulb temperature, 1.5360 for dew point and 0.5623 for mean surface wind speed. Figure 8 illustrates the RMSE values for Dataset A.

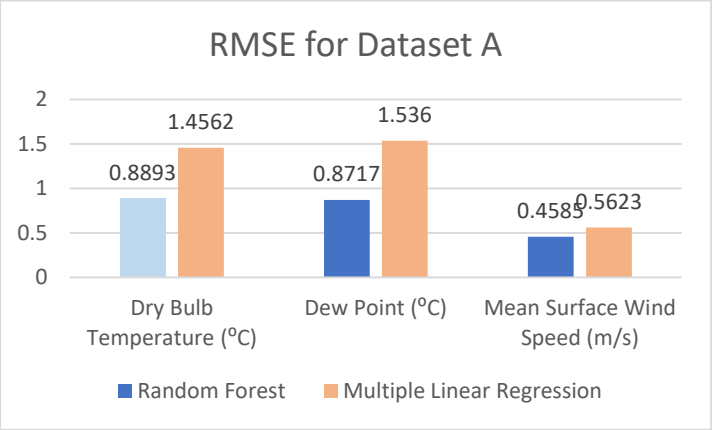

Fig. 10 *RMSE value for Dataset A*

For Dataset B, Random Forest still perform better than Multiple Linear Regression. Specifically, the RMSE value for Random Forest shows a relatively lower error for the temperature, dew point and wind speed, with values of 0.8685, 0.8396, and 0.5587 respectively. In contrast, Multiple Linear Regression gives higher MSE values of 0.9066 for temperature, 0.9551 for dew point and 0.6052 for wind speed. Figure 11 illustrates the RMSE values for Dataset B.

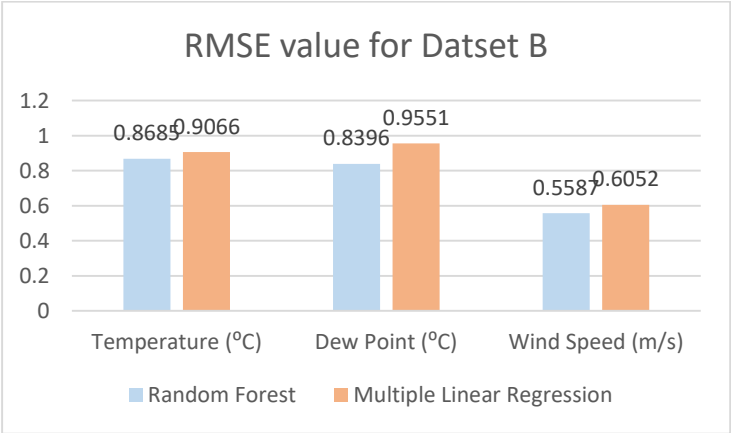

Fig. 11 *RMSE value for Dataset B*

4.2.4 R-Squared (R^2)

For Dataset A, Random Forest provides more accurate prediction compared to Multiple Linear Regression. Specifically, the R2 value for Random Forest showed a relatively lower error for the dry bulb temperature, dew point and mean surface wind speed, with values of 0.9258, 0.7456 and 0.4626 respectively. In contrast, Multiple Linear Regression gives higher R2 values of 0.8021 for dry bulb temperature, 0.2077 for dew point and 0.2136 for mean surface wind speed. Figure 12 illustrates the R^2 values for Dataset A.

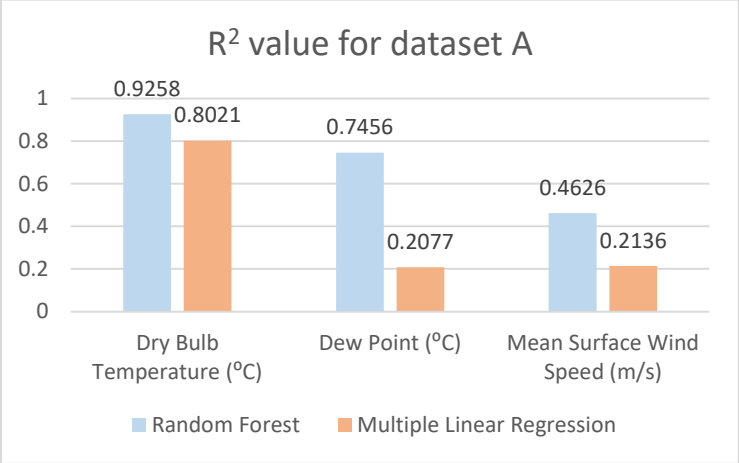

Fig. 12 *R^2 value for Dataset A*

For Dataset B, Random Forest still perform better than Multiple Linear Regression. Specifically, the R2 value for Random Forest shows a relatively lower error for the temperature, dew point and wind speed, with values of 0.9048, 0.3236 and 0.2954 respectively. In contrast, Multiple Linear Regression gives higher R2 values of 0.8946 for temperature, 1.095 for dew point and 0.1947 for wind speed. Figure 13 illustrates the R^2 values for Dataset B.

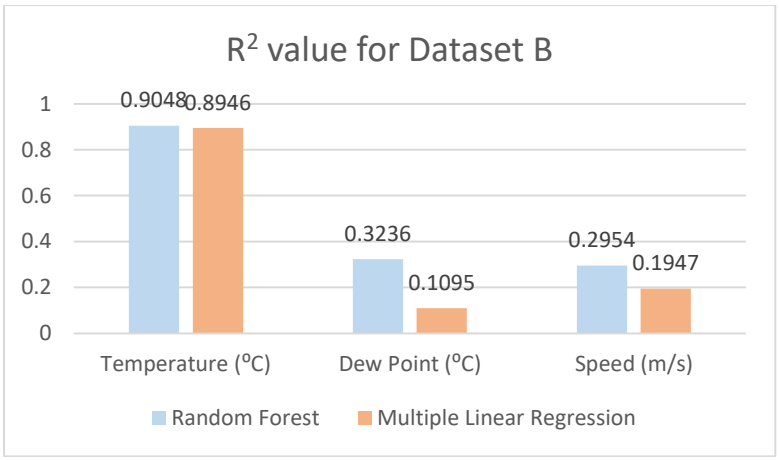

Fig. 13 *R^2 value for Dataset B*

4.3 Model Performance Summary

For Dataset A, Random Forest demonstrated a superior performance with lower error rates across all performance metrics compared to Multiple Linear Regression. For MAE, Random Forest recorded a lower value of 0.5552, indicating a minimal deviation from the true values when compared to Multiple Linear Regression, which is 0.8827. Similar trends were seen in the MSE and RMSE, with Random Forest reporting lower values of 0.5870 and 0.7662 respectively compared to 1.5986 and 1.2644 for Multiple Linear Regression. This indicates that a lesser variance and deviation in the predictions made by Random Forest. As for R2, Random Forest recorded a value of 0.7113, which was considerably higher than Multiple Linear Regression with a value of 0.4078. This shows that Random Forest explained a greater proportion of variance and provide a better model fit and predictability compared to Multiple Linear Regression.

In comparison to Dataset A, both models for Dataset B showed a little decline in performance metrics. However, Random Forest still perform a superior performance when compared to Multiple Linear Regression. For MAE, Random Forest recorded a lower value of 0.5837, indicating a minimal deviation from the true values when compared to Multiple Linear Regression, which is 0.6356. Similar trends were seen in the MSE and RMSE, with Random Forest reported lower values of 0.5904 and 0.7684 respectively compared to 0.7001 and 0.8368 for Multiple Linear Regression. This indicates that a lesser variance and deviation in the predictions made by Random Forest. As for R2, Random Forest recorded a value of 0.5080, which was slightly higher than Multiple Linear Regression with a value of 0.3996. This shows that Random Forest explained a greater proportion of variance and provide a better model fit and predictability compared to Multiple Linear Regression. Table 5 shows the average

performance of Random Forest model and Multiple Linear Regression model in terms of accuracy, MAE, MSE, RMSE and R2

Table 5 *Average performance of selected algorithms*

Dataset	Algorithm	MAE	MSE	RMSE	R^2
A	Random Forest	0.5552	0.5870	0.7662	0.7113
	Multiple Linear Regression	0.8827	1.5986	1.2644	0.4078
B	Random Forest	0.5837	0.5904	0.7684	0.5080
	Multiple Linear Regression	0.6356	0.7001	0.8368	0.3996

5. Conclusion

In conclusion, the primary goal of this research was to compare the performance of Random Forest and Multiple Linear Regression algorithms for predicting temperature and wind speed using the two selected datasets. The results show that Random Forest significantly outperforms Multiple Linear Regression for both datasets in various performance metrics such as MAE, MSE, RMSE, and R-squared, demonstrating its superior ability to handle complex meteorological data. This underscores the potential of Random Forest in enhancing weather prediction accuracy, which is critical for sectors like agriculture and transportation. Future research should explore additional machine learning models like Support Vector Machine (SVM), Artificial Neural Networks (ANN), XGBoost, and Long Short-Term Memory networks (LSTM), using larger and more diverse datasets as well as employing advanced validation techniques to further enhance weather forecasting reliability and accuracy.

Acknowledgement

The authors would like to thank the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia for its support.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design:*** *Pang Chin Hoo, Dr. Noor Zuraidin bin Mohd Safar;* ***data collection:*** *Pang Chin Hoo, Dr. Noor Zuraidin bin Mohd Safar;* ***analysis and interpretation of results:*** *Pang Chin Hoo, Dr. Noor Zuraidin bin Mohd Safar;* ***draft manuscript preparation:*** *Pang Chin Hoo, Dr. Noor Zuraidin bin Mohd Safar. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] P. Hewage, M. Trovati, E. Pereira, and A. Behera, "Deep learning-based effective fine-grained weather forecasting model," *Pattern Analysis and Applications*, vol. 24, no. 1, pp. 343–366, 2021, doi: 10.1007/s10044-020-00898-1.

[2] S. Badillo *et al.*, "An Introduction to Machine Learning," *Clin Pharmacol Ther*, vol. 107, no. 4, pp. 871–885, Apr. 2020, doi: 10.1002/cpt.1796.

[3] N. Singh, S. Chaturvedi, and S. Akhter, "Weather Forecasting Using Machine Learning Algorithm," in *2019 International Conference on Signal Processing and Communication (ICSC)*, IEEE, Mar. 2019, pp. 171–174. doi: 10.1109/ICSC45622.2019.8938211.

[4] J. Alzubi, A. Nayyar, and A. Kumar, "Machine Learning from Theory to Algorithms: An Overview," *J Phys Conf Ser*, vol. 1142, p. 012012, Nov. 2018, doi: 10.1088/1742-6596/1142/1/012012.

[5] S. Mishra, A. Shukla, S. Arora, H. Kathuria, and M. Singh, "Controlling Weather Dependent Tasks Using Random Forest Algorithm," in *2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, IEEE, Dec. 2020, pp. 1–8. doi: 10.1109/ICAECC50550.2020.9339508.

[6] S. Ray, "A Quick Review of Machine Learning Algorithms," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, IEEE, Feb. 2019, pp. 35–39. doi: 10.1109/COMITCon.2019.8862451.

[7] A. Zermane, M. Z. Mohd Tohir, H. Zermane, M. R. Baharudin, and H. Mohamed Yusoff, "Predicting fatal fall from heights accidents using random forest classification machine learning model," *Saf Sci*, vol. 159, p. 106023, Mar. 2023, doi: 10.1016/j.ssci.2022.106023.

[8] C. A. Anjum Era, M. Rahman, and S. T. Alvi, "Short Term Weather Forecasting Comparison Based on Machine Learning Algorithms," in *2023 Intelligent Methods, Systems, and Applications (IMSA)*, IEEE, Jul. 2023, pp. 369–374. doi: 10.1109/IMSA58542.2023.10217753.

[9] F. Sheikh, S. Karthick, D. Malathi, J. S. Sudarsan, and C. Arun, "Analysis of Data Mining Techniques for Weather Prediction," *Indian J Sci Technol*, vol. 9, no. 38, Oct. 2016, doi: 10.17485/ijst/2016/v9i38/101962.

[10] N. Kaur and N. Singh, "Comparative Analysis of Machine Learning Algorithms for Weather Prediction using Error Detection," in *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, IEEE, Jan. 2023, pp. 1094–1098. doi: 10.1109/ICSSIT55814.2023.10061132.

[11] C. M. Liyew and H. A. Melese, "Machine learning techniques to predict daily rainfall amount," *J Big Data*, vol. 8, no. 1, p. 153, Dec. 2021, doi: 10.1186/s40537-021-00545-4.

[12] M. E. Bruni, V. Lazzaroli, G. Perboli, and C. Vandoni, "Machine Learning to Forecast Rainfall Intensity," in *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, IEEE, Jun. 2023, pp. 1762–1767. doi: 10.1109/COMPSAC57700.2023.00272.

[13] A. P. Rodrigues, R. Fernandes, and P. Vijaya, "A study on the evaluation of different regressors in Weather Prediction," in *2022 International Conference on Artificial Intelligence and Data Engineering (AIDE)*, IEEE, Dec. 2022, pp. 13–18. doi: 10.1109/AIDE57180.2022.10060814.

[14] A. Dawood, A. Alsehibani, S. Dawood, and A. Rehman, "Prediction of Environmental Earth Surface Temperature using hybrid machine learning model," in *2023 Sixth International Conference of Women in Data Science at Prince Sultan University (WiDS PSU)*, IEEE, Mar. 2023, pp. 127–132. doi: 10.1109/WiDS-PSU57071.2023.00036.