~~Comparative Evaluation of Artificial Neural Networks, Multiple Linear Regression, K-nearest Neighbour, Random Forests, Gradient Boosting, and ARIMA. A Case Study of Rainfall Forecasting in Selangor.~~

Forecasting rainfall in Selangor by using machine learning techniques

DATA SCIENCE PROJECT I

MUHAMMAD ZAMIL SYAFIQ BIN ZANZALI

DR TAY CHAI JIAN

Bachelor of Applied Science in Data Analytics with Honours

Centre for Mathematical Sciences

Universiti Malaysia Pahang Al-Sultan Abdullah

January 2025

اونيۏرسيتي مليسيا ڤهڠ السلطان عبد الله
UNIVERSITI MALAYSIA PAHANG
AL-SULTAN ABDULLAH

FACULTY SUPERVISOR'S DECLARATION

I hereby declare that I have checked this Data Science Project I Report. In my opinion, this project proposal is adequate in terms of its scope and quality for the award of Bachelor of Applied Science in Data Analytics with Honours.

_____

(Faculty Supervisor's Signature)

Full Name : DR TAY CHAI JIAN

Position :

Date :

اونيۏرسيتي مليسيا ڤهڠ السلطان عبد الله
UNIVERSITI MALAYSIA PAHANG
AL-SULTAN ABDULLAH

STUDENT'S DECLARATION

I hereby declare that the work in this Data Science Project I Report is based on my original work.

_____

(Student's Signature)

Full Name : MUHAMMAD ZAMIL SYAFIQ BIN ZANZALI

ID Number : SD22046

Date :

ACKNOWLEDGEMENT

Acknowledgements, acknowledgements.

ABSTRACT

Accurate rainfall forecasts are essential in planning and management of floods, water resources, and agriculture. This is particularly critical in a region like Selangor in Malaysia where rainfall patterns are highly variable. In this region extreme events such as too much or too little rainfall are often. This unpredictability poses a challenge to effective planning by the government. This study aims to identify machine learning algorithms that can be used for accurate rainfall forecasting, compare the performance of different models, and select the best model that can be deployed as part of an early warning system. Historical data collected between 2012 and 2020 will be used. ~~The target variable will be precipitation~~. The predictors will be average temperature, wind speed, and relative humidity. The machine learning algorithms that will be evaluated are: Artificial Neural Networks, Multiple Linear Regression, K-nearest Neighbour, Random Forests, Gradient Boosting, and ARIMA. Data will be pre-processed by mean imputation of missing values, removing observations outside expected range, and normalization. Root mean squared error, mean absolute error, and R squared will be used to evaluate model performance. This study will demonstrate the value of machine learning in rainfall forecasting and provide actionable insights that will be used in strategic planning and management.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

Figure number should follow the Chapters.

LIST OF SYMBOLS

LIST OF ABBREVIATIONS

ML	Machine Learning
NWP	Numerical Weather Prediction
MAE	Mean Absolute Error
SVM	Support Vector Machine
ANN	Artificial Neural Network
RMSE	Root Mean Square Error
RF	Random Forest
DT	Decision Tree
k-NN	K-Nearest Neighbor
IQR	Interquartile Range
R^2	Coefficient of Determination

CHAPTER 1

Introduction

1.1 Research Background

Rainfall patterns in Selangor region of Malaysia fluctuate widely partially driven by the tropical climate. In Selangor precipitation patterns are significantly influenced by tropical climate with the heaviest rainfall happening between October and December. November is the peak of this season where 324 mm of rainfall is experienced across 28 days. In October 222 mm of rainfall is experienced while in December 246 mm of rainfall is experienced. At the beginning of the year the amount of rainfall is relatively lower. January and February receive 148 mm and 102 mm respectively. However, April receives a rainfall of 241 mm which is comparable to precipitation received in peak season. During the summer months of June and July relatively lower rainfall amounts of 145 mm and 135 mm respectively are received (Nomadseason, 2025). These seasonal patterns have a major influence on local ecosystems as well as agriculture activities and water management. The Malaysian Meteorological Department (2025) analysed annual rainfall data from 1951 to 2023 and found there has been an upward trend in the amount of rainfall received in the country. This points to climate change that can lead to higher temperatures, rising sea levels, often occurrence of extreme events such as floods, disruption of habitats and agricultural activities, and economic losses. These fluctuations make it difficult to accurately forecast climate patterns. Climatic events such as frequent and heavy rainfall can lead to crop failure, floods, and water contamination. Similarly, seasons such as the monsoon have a significant influence on rainfall and its distribution. The Department of Irrigation and Drainage Malaysia (2025) describes monsoon rains as “typically of long duration with intermittent heavy bursts and the intensity can occasionally exceed several hundred mm in 24 hours”. This can lead to floods in urban areas and disruption of agricultural activities. Accurate forecasting will help the Selangor State Government in mitigating the effects of these events. Equipped with accurate forecasts the state government can put in place well planned emergency as well as disaster and preparedness strategies.

Machine learning models have become a critical tool in analysis of meteorological data. When comparing machine learning models with conventional Numerical Weather Prediction models, it has been observed machine learning models are superior at detecting intricate numerical and non-linear patterns in data (Bouallègue et al., 2024). This makes machine learning a suitable approach for predicting rainfall in a tropical region like Selangor. Large amounts of meteorological data can be analysed using machine learning techniques such as support vector machines (SVM), gradient boosting, and artificial neural networks (ANN) to provide accurate temporal estimations. These methods that will be discussed later, use historical data such as temperature, humidity, wind speed, and rainfall to provide accurate forecasts which were hitherto impossible using traditional techniques such as linear regression.

The problem is critical in places such as Selangor, where rainy conditions have not been accurately forecasting posing several difficulties. Hydrological functions enhanced by better rainfall predictions enable timely decisions in crop production, disaster management including floods and landslides, and water management. Due to improved accuracy levels of predictions, stakeholders will be in a position to save structures from destruction, people from hunger as well as resources from wastage.

Recent advances in machine learning have expanded possibilities for improving rainfall forecasting. Machine learning methods like support vector machines, gradient boosting, and artificial neural networks have shown great potential in capturing both temporal and spatial patterns of rainfall. These models are able to improve forecasts by continuously learning from new data. In Selangor, using machine learning techniques and local meteorological data presents an opportunity to develop a forecasting system that is highly accurate.

1.2 Problem Statement

Climate change has received significant global attention due to disastrous events it can cause. Rainfall is a major meteorological factor that is influenced by climate change. In Malaysia, rainfall patterns have changed causing floods and droughts. Selangor is one the states that has been affected by these changes in rainfall patterns. Disastrous floods happened consecutively in the years 2006 to 2008 and in the years 2010 and 2011. The years 1997, 1998, and 2008 had catastrophic dry periods (Talib et al., 2024). ~~Agricultural decisions and productivity are significantly influenced by environmental variables particularly the amount of water available and rainfall. In Selangor the influence of these variables is significant and a~~

threat to agricultural productivity. High and low rainfall affects crops. Although it is possible to mitigate low rainfall through irrigation, high rainfall usually damages crops and results in low agricultural productivity. Mitigation measures such as changing crop cycles and combining crop cycles have not been adequate. To adequately solve these problems technological solutions are required (Alam, 2021).

One of the technological solutions that can be used is availing accurate rainfall predictions. However, due to irregular occurrence of rainfall in Timur Region Selangor accurate prediction is difficult. This situation can harm farming, cause floods, and cause difficulties in water resources planning. Traditional models such as linear regression may not provide accurate precipitation forecast especially in the tropics because the atmospheric behaviour is not easy to predict. For example, Kassem et al. (2021) reported artificial neural networks were superior to linear regression in predicting monthly rainfall in Northern Cyprus. That study showed artificial neural networks were better at capturing relationships in coordinates, meteorological variables, and rainfall resulting in more accurate prediction compared to linear regression. Traditional models such as linear regression are weak at capturing complex relationships especially when they are non-linear. Compared to models such as support vector machines and artificial neural networks, linear regression models are poor at handling non-linear relationships. Conversely, support vector machines and artificial neural networks are difficult to interpret, computationally costly, and require large amounts of data (Goodfellow et al., 2023; Murphy, 2022). Modern meteorological research does not face the limitations of small datasets and limited computational power that were prevalent several decades ago. Meteorological instruments and IoT sensors have enabled accumulation of large datasets. This situation enables use of advanced machine learning models such as support vector machines and artificial neural networks in predicting rainfall. Specifically, in Selangor large volumes of meteorological data are available. Therefore, these advanced machine learning models can be used to accurate predict rainfall patterns. Insights obtained will be useful in agricultural, infrastructure, public health, and water management planning.

1.3 Research Questions

The specific research questions that will be investigated in this study are:

i. What are the machine learning models that can be used for rainfall prediction in Selangor?

ii. How does the performance of different machine learning models differ?

iii. What is the best model in forecasting rainfall pattern in Selangor?

1.4 Research Objectives

The broad objective of this study is to investigate the use of machine learning models in predicting rainfall in Selangor region of Malaysia. The specific objectives are:

i. To employ machine learning models that can be used for predicting rainfall in Selangor.

ii. To estimate and assess the performance of different machine learning models using performance metrics such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R-squared.

iii. To identify the best model for forecasting rainfall patterns in Selangor by comparing performance metrics and selecting the model with highest accuracy.

1.5 Research Scopes

This research deals with rainfall prediction for Selangor, Malaysia where the rainfall has irregular tropical pattern and significantly affects sectors such as water supply and flood control, agriculture. These problems will be addressed in this work by utilising and comparing a number of machine learning algorithms with support vector machines (SVM), gradient boosting, and artificial neural networks (ANN). These methods were chosen due to the possibility of the interpretation of which dependencies – both linear and nonlinear ones – are present in the data. In the present study, meteorological data from Sepang/KL International Airport is employed for data analysis where necessary climatic factors embracing average temperature, relative humidity, wind velocity, and precipitation for the years between 2012 and 2020 are utilised. This is to make certain that the data collected are accurate and reliable to

increase the efficiency of data analysis after it has been fed into the system therefore data cleaning, normalization of data, handling of missing values and feature engineering will be undertaken.To fully assess predictive performance, the model will be evaluated using measures like the Coefficient of Determination (R^2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). What is the software used?

1.6 Significance of Study

The research focus on using machine learning for rainfall forecasting in Selangor. Machine learning techniques utilize historical data to identify complex relationships, resulting in more precise and current forecasts. This study improves the scientific understanding of based on rainfall forecasting by evaluating how well different machine learning algorithms capture detailed tropical rainfall patterns. It represents a major breakthrough in environmental prediction and building resilience since it expands the use of machine learning for tropical weather forecasting and offers a structure that can be adjusted for different climates. The forecasting results could help the government in enhancing disaster readiness.

Start Chapter 2 at a new page.

CHAPTER 2

Literature Review

2.1 Introduction

In tropical regions such as Selangor in Malaysia where extreme events such as high or low rainfall happen; accurate rainfall forecasting is critical. When managers are provided with accurate predictions, they are better placed to put mitigation measures in place. These measures can help in management of disruptive events such as floods, agricultural crop failure, and disruptions in water supply. Machine learning has emerged as a powerful technique for analysing rainfall data, discovering patterns in meteorological data, and accurately predicting rainfall. This chapter presents an exhaustive review of existing literature on use of machine

learning for forecasting rainfall. Specifically, the strengths and limitations of each study are evaluated to identify research gaps that can be addressed in this study and future studies.

2.2 Challenges in Rainfall Forecasting

Numerous studies have well documented challenges faced when predicting rainfall. Kundu et al. (2023) have discussed some of these challenges. These authors note the primary challenge is the wide variability in rainfall patterns. Other challenges are scarcity of relevant meteorological variables such as soil, humidity, wind and temperature parameters which are essential. When these variables are not available the accuracy of prediction models is severely affected. Other human activities such as deforestation can also negatively affect the accuracy of rainfall prediction models. Even when advanced methodologies are used accurate prediction of rainfall is challenging as large volumes of data and collaboration are required.

The National Oceanic and Atmospheric Administration (2024) notes forecasting weather phenomena is a difficult skill that requires meticulous observation and analyzing large amounts of data. Weather phenomena can be characteristically thunderstorms covering large areas or a small area that can last for a few hours or several days. The phases involved in weather forecasting are observation, prediction, and dissemination of results.

Ray et al. (2021) discuss various challenges faced in predicting rainfall driven by landfalling tropical cyclones in India. Rainfall from these tropical cyclones especially when approaching landfall varies widely and is usually asymmetric. This pattern is often caused by wind, speed, land surface, and moisture parameters. That study found that increase or decrease in intensity of a tropical storm as it approached the coastline during landfall can change the characteristics of rainfall over land.

Selangor is a typical tropical environment characterized by widely fluctuating rainfall patterns. This variation makes accurate rainfall prediction challenging. These challenges arise because rainfall patterns are influenced by intricate relationships among atmospheric factors like variations temperature, humidity, and windspeed. Rainfall predictions are usually obtained from large scale computerized simulations of weather systems. Use of traditional prediction

methods like numerical weather prediction fails at capturing events that happen in isolated areas. Furthermore, this problem is severe in areas that have widely varying rainfall patterns such as Selangor. These models are further limited by their high cost and their lack of flexibility to adjust to changes in rainfall patterns in real time. Machine learning is a viable alternative for overcoming challenges faced by these traditional models. Particularly, machine learning models are suited to capturing complex and non-linear relationships that exist in meteorological data. With these capabilities machine learning models are an essential tool for discovering patterns that exist in historical meteorological data.

2.3 Overview of Machine Learning Techniques for Rainfall Prediction

Machine learning models are well suited to capture non-linear relationships that are a common feature in meteorological variables like temperature, windspeed, humidity, and precipitation. This makes machine learning models a robust technique for analysing meteorological data. This section presents an exhaustive review of literature that has examined use of different machine learning models for rainfall prediction.

Wani et al., (2024) compared use of "artificial neural network (ANN), random forest (RF), support vector regression (SVR), k-nearest neighbour (KNN), long-short term memory (LSTM), bi-directional LSTM, deep LSTM, gated recurrent unit (GRU), and simple recurrent neural network (RNN)" for rainfall forecasting. Time series techniques such autoregressive integrated moving average (ARIMA) and Box-Cox transformations were also investigated. Evaluation of prediction accuracy using metrics such as using root mean square error and mean absolute error revealed deep learning techniques achieved the highest accuracy in rainfall prediction. Machine learning techniques were second while time series models were third. When comparing individual deep learning methods, the order from the highest to lowest accuracy was "bi-directional LSTM, LSTM, RNN, deep LSTM, and GRU". When comparing individual machine learning models, the order ~~the order~~ from the highest to the lowest accuracy was "ANN, KNN, SVR, and RF". In subsequent sections literature on use of these different methods for rainfall prediction will be examined.

2.3.1 Support Vector Machines (SVM)

Support vector machines is one of the supervised techniques that has been successfully applied in predicting quantitative variables and classification of categorical variables. A key strength of support vector machines is their ability to find a separation even when the feature space is not linearly separable. Support vector machines achieve this by projecting to a high dimensional feature space. This projection happens in two stages. In the first stage a separation space is identified. The second stage involves modification of data to achieve separation using a hyperplane. These new features in the data are then used to make a decision on the category where a new record will be assigned.

Support vector machines are suited to capturing non-linear relationships and have been used for time series applications such as "prediction, pattern recognition, and multiple non-linear regression". Support vector machines avail options of four kernel functions which are linear, polynomial, radial basis function (RBF) and sigmoid. The latter three functions are non-linear and suitable for capturing non-linear relationships. This makes support vector machines appropriate for weather data that usually have complex and non-linear relationships (Ashok & Pekkat, 2024).

Several studies have used support vector machines to predict rainfall using historical meteorological data. Praveena et al. (2023) used support vector machine learning and logistic regression. In that study hyperparameter tuning achieved an impressive accuracy of 88%. This demonstrates the superiority of support vector machines in capturing complex relationships in high dimensional data. Thus, support vector machines are a viable technique for predicting rainfall using complex meteorological variables.

Hayaty et al. (2023) used support vector machines to predict rainfall in a city in Indonesia. The objective of that study was to investigate the performance of support vector machines. The predictors of rainfall were "temperature, humidity, and wind speed". The support vector machine model achieved an accuracy of 82% and ROC curve evaluation showed the model had a score of 0.74. From the results of that study, it is evident support vector machines are good at distinguishing positive and negative rainfall events.

Hapsari et al. (2020) used a support vector machined optimized using stochastic gradient descent to predict rainfall. Use of this optimization was novel as rainfall forecasting usually uses a linear threshold. The predictors were "atmospheric pressure, sea level pressure, wind

direction, wind speed, and relative humidity". A training and test subset ratio of 80% to 20% was used. Simulation results revealed support vector machines had better accuracy compared to traditional methods such as time series forecasting. This is an indication support vector machines are a promising tool for meteorological forecasting as it can accommodate more predictors to better capture the relationship to rainfall.

Due to bias predictions obtained from numerical weather prediction (NWP), Yin et al. (2022) compared support vector machines to other methods for rainfall forecasting in Japan. The other methods investigated were "quantile mapping (QM), cumulative distribution function (CDFt), and a combination of support vector machines and QM". Results revealed when support vector machines were used alone, there was a significant improvement in correlation. However, support vector machines faced the limitation of underestimating hourly and heavy rainfall occurrences. QM and CDFt successfully mitigated bias in SVM but had limitations in mitigating rainband location. A hybrid method incorporating support vector machines and QM demonstrated consistency in predicting extreme events although the model was observed to overestimate rainfall.

Al-Mahdawi et al. (2023) used support vector machines and monthly data from 1901 to 2022 to predict rainfall. Results showed the forecast accuracy of different months varied. For example, the root mean square, mean squared error, and mean absolute error of months such as June, July, and August were zero suggesting a very high accuracy. However, in January these metrics were observed as 5.51, 30.38, and 3.03 suggesting a relatively low accuracy. Despite the lower accuracy observed in some months, support vector machines are still a useful technique for forecasting rainfall.

Du et al. (2021) used a support vector machine optimized using particle swarm optimization to classify rain events. Use of this optimization is novel as a linear threshold is frequently used. An 80% to 20% ratio was used for splitting the training and testing subsets and a radial basis kernel was used. The variables used in the study were "atmospheric pressure, sea level pressure, wind direction, wind speed, relative humidity, and precipitation". Data were pre-processed by checking expected range of values and normalization. Results showed support vector machines with particle swarm optimization were a promising technique for forecasting precipitation accurately.

Velasco et al. (2022) used support vector regression to predict rainfall using data collected over a 17- month period. A radial basis function kernel was used. Other parameters used were “c = 100, g = 1, e = 0.1, and p = 0.001. The model achieved a mean square error of 3.46 demonstrating an acceptable accuracy between actual and predicted values. This suggests with proper data pre-processing and parameter tuning support vector regression is a viable technique for rainfall forecasting.

2.3.2 Gradient Boosting

Gradient boosting is an ensemble learner that provides high prediction accuracy using multiple weak learners. Usually, decision trees are selected as the weak learner. This learning proceeds through sequential fitting of residuals from the previous learner to a new learner and updating the ensemble. This process is iterated until a pre-specified criterion is met (Masui, 2024).

Numerous studies have used gradient boosting to predict rainfall. Anwar et al. (2020) used gradient boosting to predict rainfall using daily data collected over a seven-year period. The predictors were temperature, humidity, sun exposure, and wind parameters. The model showed relative humidity and minimum temperature were the most important predictors of rainfall. Model evaluation showed the model achieved root mean square error and mean absolute error values of 2.7 and 8.8 respectively. An often-observed limitation of gradient boosting which is overfitting was evident in that study. The best root mean square error was observed at five iterations. After that the test error started increasing which is an indication of overfitting.

Poola and Sekhar (2021) used XGBOOST and monthly data collected between 1987 and 2017 to predict rainfall. R statistical software was used to analyze data. Autocorrelation and partial autocorrelation functions were used for model assessment. The model achieved an accuracy of 95%.

Nuthalapati and Nuthalapati (2024) compared various models such as “k-nearest neighbors, support vector machine, gradient boosting, XGBOOST, logistic regression, and random forest” to asses their performance in predicting rainfall using daily data. Predictors were temperature, wind, and sun parameters. The accuracies of the models were: 76.87%, 77.55%, 70.07%, 80.95%, 80.95%, and 72.79%. These results demonstrate the superiority of XGBOOST over other machine learning algorithms.

Cui et al. (2021) used a hybrid SSA-LightGBM model consisting of singular spectrum analysis (SSA) and LightGBM. SSA was used to decompose the time series while a LightGBM was used to capture trend and variation. The hybrid SSA-LightGBM was superior to use of LightGBM or LSTM alone.

Sanches et al. (2024) investigated use of XGBOOST for predicting rainfall using daily data collected between 1989 and 2019 in Sao Paulo. Classification and regression were done. Results showed in classification an accuracy of 90% was achieved. In the regression task a mean absolute error of 3mm was observed.

Maaloul and Lejdel (2023) compared five algorithms which are "random forest, decision tree, naïve Bayes, gradient boosting, and artificial neural networks". Comparison of these models revealed gradient boosting achieved an accuracy of more than 98% in rainfall forecasting. This demonstrates the superiority of gradient boosting compared to other models.

Zhuang and DeGaetano (2024) Used LightGBM to classify rainfall events using daily data collected over 10 years in different parts of Australia. The target variable was a yes/no indicator of rainfall. The predictors were temperature, rainfall, sunshine, wind, humidity, and cloud parameters. The LightGBM parameters tuned were "number of estimators, learning rate, number of leaves, lambda, and alpha". A 10-fold cross-validation was incorporated into model training. One subset of 60% was used for model training while two subsets each consisting of 20% were used for model testing. Comparison to other models published in the literature revealed LightGBM had comparable accuracy with random forest and gradient boosting but had better accuracy when compared to k-nearest neighbor and a support vector machine using a linear kernel.

2.3.3 Random Forest (RF)

Random Forest (RF) is an ensemble learning method that creates multiple decision trees using subsets of data and features selected randomly. There are two major steps in the algorithm. In the first step every tree in the forest is trained using a bootstrap sample. This happens at every node where a random subset is used to identify the best split. This minimizes overfitting and correlation among trees. In the second step predictions are obtained by combining output from all the trees. In classification tasks majority voting is used while in regression averaging is used. This approach makes random forests achieve high accuracy,

robustness, suitable for complex data, and minimizes the weakness of individual decision trees (Talekar and Agrawal 2020).

Several studies have investigated use of random forests for rainfall prediction. Raniprima et al. (2024) compared random forests and decision trees for prediction a binary outcome rain/not rain. The predictors were temperature, humidity, and wind parameters. Results showed a random forest had an accuracy of 95.65% while a decision tree had an accuracy of 94.85%. This result suggests random forests were marginally better than decision trees.

Hsu et al. (2024) compared random forest and CatBoost for predicting rainfall using data collected from 1998 to 2018 in Taiwan. The target was a binary variable indicating rain or no rain. The predictor variables were "temperature, humidity, air pressure, wind direction, and airspeed". Data were pre-processed by removing null or missing values, transformation of rainfall into a categorical variable, and normalizing predictors. Results showed the precision, recall, F1, and support metrics of random forest and Catboost were 0.70, 0.70, 0.70, 11302, and 0.69, 0.69, 0.69, 11302 respectively. Thus, random forest was marginally better than CatBoost.

Raut et al. (2023) compared random forest regression, linear regression, support vector regression, and decision trees for predicting rainfall. The predictors were temperature and coastline characteristics. Results showed the random forest significantly outperformed all the other models.

Wolfensberger et al. (2021) compared random forest and non-polarimetric quantitative precipitation estimation (QPE) for predicting rainfall in Switzerland. The predictors were radar data. Evaluation results revealed use of random forest minimized bias and error in the predicted amount of rainfall. This was particularly observed in rainfall with "large and solid or mixed" characteristics. However, the random forest model had challenges with overestimating bias especially in low rainfall situations. Still the random forest provided faithful predictions that were an improvement over non-polarimetric QPE.

Primajaya and Sari (2021) used a random forest to predict rainfall in Indonesia. The predictors were temperature, mean sea/station pressure, WDSP, and MXSPD. A 10-fold cross-validation was used. Mean absolute error and root mean square error values of 0.35 and 0.46 were observed on the test data.

2.3.4 Decision Tree (DT)

A Decision Tree (DT) splits data into branches based on feature thresholds, aiming to reduce impurity (measured by metrics like Gini Index or entropy) at each node. Starting from a root node, the tree grows by dividing the dataset into smaller subsets until reaching leaf nodes that represent the predicted outcomes. Decision Trees are simple, interpretable, and effective for classification and regression tasks. However, they are prone to overfitting, especially with complex data, which can be mitigated through pruning or combining multiple trees in ensemble methods like Random Forest (Talekar and Agrawal 2020).

Multiple studies have used decision trees for predicting rainfall. Bhardwaj and Duhoon (2021) compared tree methods such as "Quinlan M5 algorithm, reduced error pruning tree, random forest, logit boosting, Ada boosting". The objective was to predict rainfall using daily data collected over 17 months in India. Evaluation of MAE, RAE, RRSE, RMSE, and MAPE performance metrics showed a random forest model outperformed all the other models under investigation.

Resti et al. (2023) used a decision tree to predict a binary target of occurrence of rainfall events. The predictors were "temperature, precipitation, sunshine, wind direction, wind speed, humidity, and cloud type". The model achieved an accuracy of 98.53% which indicates decision trees are a viable technique for rainfall prediction.

Sharma et al. (2021) used a decision tree to identify critical predictors of extreme precipitation events in Fiji Islands. The intensity of tropical cyclones was not an important predictor as weaker tropical cyclones can result in more rainfall compared to intense cyclones. The most important predictor of rainfall was tropical cyclone minimum distance from land followed by "TC cluster grouping, seasonality, and duration". These results suggest decision trees are useful for risk evaluation.

Nurkholis et al. (2022) used a C5.0 decision tree to predict three categories of rainfall which were low/medium/high. Predictors were date, temperature, humidity, sun, and speed parameters. The highest accuracy was obtained using a 5-fold CV in the test subset.

2.3.5 Artificial Neural Networks (ANN)

Artificial neural networks are well suited to capture non-linear relationships in data. The architecture of an artificial neural network consists of three layers. The first layer is the input layer, the second layer is a hidden layer, and the third layer is an output layer. The purpose of the input layer is to get data input and pass it to the hidden layer. The purpose of the hidden layer is to perform computations. The purpose of the output layer is to provide predictions. One or multiple neurons are used to provide final model output (Ashok & Pekatt, 2024).

Nayat et al. (2020) carried out a literature review on use of artificial neural networks in predicting rainfall. Literature spanning over 25 years was reviewed. Artificial neural networks were found superior compared to statistical and numerical methods such as multiple regression and ARIMA. Similarly, Nandakumar et al. (2020) carried out a literature survey on use of artificial neural networks for rainfall prediction. They concluded artificial neural networks provided more accurate forecasts compared to mathematical and numerical techniques.

Several studies have used ANNs to predict rainfall. Kala et al. (2021) used a Feed Forward Neural Architecture to predict rainfall. Predictors were “temperature, cloud cover, vapor pressure, and precipitation”. Data pre-processing was done using normalization and a split ratio for 60:40 for training and test subsets. Results showed the neural network had an accuracy of 93.55% and a root mean square error of 0.254. These results demonstrate the value of neural networks in rainfall prediction.

Mislan et al. (2020) used a backpropagation neural network architecture having two hidden layers to predict monthly rainfall collected between 1986 and 2008 in Indonesia. Data were preprocessed using sigmoid normalization and split into train and test subsets using a ratio of 60% to 40%. The model achieved a mean square error of 0.00096 indicating a high accuracy in predicting rainfall.

Aizansi et al. (2024) used a multilayer perceptron architecture to predict monthly rainfall using data collected between 1959 and 2017. A model using this architecture was compared to LSTM and climatology forecasts. Predictors were wind, temperature, pressure, humidity, and meridian parameters. Data was pre-processed by filling in missing values with the median and normalizing variables. Data were split into training, validation, and testing subsets. Models were evaluated using RMSE, MAE, MAPE, and R squared metrics. On the testing subset the performance metrics of the multilayer perceptron model were 72.41, 51.97,

61.64, and 0.432 respectively. The performance metrics of LSTM were 76.65, 54.53, 61.66, and 0.369 respectively. These results indicate the multilayer perceptron model had better performance compared to LSTM.

Lee et al. (2022) used an optimized artificial neural network to predict rainfall using monthly data collected between 1966 and 2017. A "three-layered feed-forward neural network" architecture. Predictors were 11 variables which were used in the simple model. Variable importance revealed there were five predictors that were most important and were used to build an optimal model. RMSE on train, validation, and test subsets were 25.84%, 32.72%, and 34.75% respectively. These results demonstrate artificial neural networks can be successfully used to predict rainfall.

2.3.6 Logistic Regression

Logistic regression is a statistical method used to model the probability of a binary outcome based on one or more predictor variables. It works by applying a logistic (sigmoid) function to a linear combination of input features, transforming the results into a range of probabilities between 0 and 1. A threshold of 0.5 is usually used to classify observations into one of the classes. Maximum likelihood estimation is used to estimate unbiased coefficients. Logistic regression is particularly suitable when there is a need for simplicity and interpretability (Anshul, 2024).

Numerous studies have used logistic regression to predict rainfall. Imon et al. (2022) used a logistic regression model to predict rainfall using daily data collected between 1989 and 2004. Predictors were evaporation, temperature, and humidity parameters. Data were preprocessed by checking outliers. The model had a classification accuracy of 95.25% indicating logistic regression is useful for predicting rainfall.

Ejike et al. (2021) used a logistic regression model to predict next-day rainfall events using one-year daily data collected in Australia. The predictors were "temperature, pressure, humidity, sunshine, evaporation, cloud cover, wind direction, and wind speed". A 70% subset was used for training and a 30% subset for testing. Significant predictors of rainfall were wind speed and pressure. Model evaluation showed an accuracy of 84% demonstrating the usefulness of logistic regression in rainfall forecasting.

Khan et al. (2024) compared “logistic regression, decision trees, multi-layer perceptron, and random forest”. Each model was selected due to its strengths. Logistic regression is simple and easy to interpret. Decision trees are well suited to capture non-linear relationships but have the limitation of overfitting. Multi-layer perceptrons are similarly suited to non-linear relationships but are computationally costly and required special hyper parameter tuning. Random forests overcome the limitation of overfitting but are difficult to interpret. Data were pre-processed by encoding categorical variables, converting date to an appropriate format, identifying missing values, and selecting relevant features. Model evaluation showed logistic regression had an accuracy of 82.80% while neural network model had an accuracy of 82.59%. These results demonstrate the usefulness of logistic regression as a simple, easy to interpret, and accurate technique for rainfall forecasting.

2.3.7 K-Nearest Neighbor (K-NN)

K-Nearest Neighbor is a non-parametric method used for classification and regression. A distance metric such as Euclidean, Manhattan, or Minkowski is used to capture the similarity between observations. In classification tasks majority voting is used to assign a class. In regression tasks an average is used for prediction (Yu & Haskins, 2021).

Multiple studies have used K-NN for rainfall prediction. Moorthy and Parameshwaran (2022) developed a hybrid model (WOAK-NN) consisting of a whale optimization algorithm (WOA) and K-NN to predict rainfall using daily data collected from 2013 to 2017. Twenty predictors were used. Model evaluation using MAE, F-measure, and accuracy revealed the hybrid WOAK-NN had better performance and was not computationally costly as it used lazy learning.

Huang et al. (2021) developed an improved K-NN which they referred to as WKNN. The objective of this innovation was to provide robustness which is usually affected by the choice of the k parameter. This improved model was compared to linear and radial support vector machines. Model evaluation revealed performance of WKNN was at par with linear and radial techniques.

Findawati et al. (2020) compared “Naïve Bayes, K-nearest neighbor, and C4.5” to forecast rainfall using data collected from 2015 to 2018. The predictors were temperature, humidity, wind, radiation, and rain parameters. Data were preprocessed by normalizing variables to bring them to a common range. Data were split into a train and test subset. The

various parameters of k used were 3, 5, and 7. Comparison of these values of k showed the highest accuracy was obtained with k = 7. Comparison of the models revealed K-NN had the highest accuracy demonstrating its value in rainfall forecasting.

Yu and Haskins (2021) compared "deep neural network, wide neural network, deep and wide neural network, reservoir computing, long short term memory, support vector machine, and K-nearest neighbor" for precipitation forecasting using data collected over 11 years. Predictors were precipitation, temperature, humidity, wind, pressure, and visibility parameters. Data were preprocessed by min-max and z-score normalization. R squared, correlation, MSE, and RMSE were used for model evaluation. Model evaluation revealed K-NN had the highest R squared, MSE, and RMSE.

Setya et al. (2023) compared linear regression and K-NN for predicting monthly rainfall in Indonesia using data collected between 2021 and 2023. Predictors were sunshine, temperature, wind, and humidity parameters. Data were split into train and test subsets. Models were compared using RMSE and MAE. Results showed K-NN had better performance compared to linear regression.

Dawoodi and Patil (2020) used K-NN to predict precipitation using daily data from North Marashtra using data collected from 2009 to 2018. K-NN achieved an accuracy of 96% indicating its potential usefulness.

2.3.8 ARIMA

Use of ARIMA for rainfall forecasting is well established. Vijayalakshmi et al. (2022) compared ARIMA and linear regression for predicting annual rainfall. ARIMA has three parameters. The parameter p captures the autoregressive process. The parameter q captures the moving average process. The parameter q captures the order of differencing required to achieve stationarity. Results showed ARIMA had higher accuracy in predicting seasonal and annual rainfall and is thus suitable in agricultural applications.

Bari et al. (2022) used a Box-Jenkins ARIMA approach to predict rainfall using data collected between 1980 and 2010 in Sylhet region of Bangladesh. The ARIMA model developed could be used in flood, tourism, crop cycle, and urban planning management.

Used an ARIMA model to predict annual rainfall using data collected between 2015 and 2020 in Assam-Meghalaya region. AIC was used to select the best model. Results showed predictions could be used to plan for earlier crop harvesting when accurate prediction of monsoon rains were available.

Kumar and Sharma (2024) used ARIMA to predict monthly monsoon rainfall during the months of June, July, August and September. The range of observed RMSE values was 13.88 to 51.15 mm while R squared ranged between 0.685 and 0.881. These results show the usefulness of ARIMA in rainfall forecasting.

These machine learning methods offer diverse approaches to rainfall forecasting, each with unique strengths and limitations. Selection of an appropriate algorithm requires considerations such as interpretability, simplicity, computational cost and specific objectives.

2.4 Summary

Authors	**Techniques**	**Data Frequency**	**Main Result**
Praveena et al. (2023)	Support vector machines, Logistic Regression	Daily	Both techniques achieve optimized results after hyperparameter tuning.
Hayaty et al. (2023)	Support vector machines	Daily	Support vector machine had an accuracy of 72%
Hapsari et al. (2020)	Support vector machines	Daily	Stochastic gradient optimization had better performance compared to time series
Yin et al. (2022)	QM, CDFt, support vector machines	Monthly	A hybrid SVM-QM model outperformed the other models

Al-Mahdawi et al. (2023)	Support vector machines	Monthly	Support vector machines had low MAE, RMSE, and MSE in some months but useful forecasts were obtained
Du et al. (2021)	Support vector machines	Daily	Swarm optimization was useful for improving accuracy
Velasco et al. (2022)	Support vector machines	Monthly	A radial basis kernel produced acceptable accuracy as measured by MSE
Nuthalapati. (2024)	Decision tree, K-Nearest Neighbor, Random Forest, Gradient Boosting, Logistic Regression	Daily	Gradient Boosting and Logistic Regression achieve the highest accuracy of 80.95%
Anwar et al. (2020)	XGBOOST	Daily	Best RMSE was obtained at five iterations
Poola and Sekhar (2021)	XGBOOST	Monthly	Model had a high accuracy of 95%

Nuthalapati and Nuthalapati (2024)	KNN, SVM, gradient boosting, XGBOOST, logistic regression, random forest	Daily	XGBOOST had superior performance compared to the other models
Cui et al. (2021)	SSA, LightGBM	Daily	A hybrid SSA-LightGBM was superior to either model
Sanches et a. (2024)	XGBOOST	Daily	An accuracy of 90% in classification and MAE of 3mm in regression were observed
Maaloul and Leidel (2023)	Random forest, decision tree, naïve bayes, gradient boosting, neural networks	Daily	Gradient boosting had the highest accuracy
Zhuang and DeGaetano (2024)	LightGBM	Daily	LightGBM had similar performance to random forest and gradient boosting but had higher accuracy than KNN and linear kernel SVM

Raniprima et al. (2024)	Random forest, decision tree	Daily	Random forest had a marginally higher accuracy than decision tree
Hsu et al. (2024)	Random forest, CatBoost	Daily	Random forest had better performance compared to CatBoost
Raut et al. (2023)	random forest regression, linear regression, support vector regression, and decision trees	Daily	Random forest had best performance compared to the other models
Sanaboina. (2024)	Artificial Neural Network	Daily	Yield accuracy of 88.65%
Primajaya and Sari (2021)	Random forest	Daily	MAE and RMSE values of 0.35 and 0.46 were observed
Bhardwaj and Duhoon (2021)	"Quinlan M5 algorithm, reduced error pruning tree, random forest, logit boosting, Ada boosting"	Monthly	Random forest had best performance

Resti et al. (2023)	Decision tree	Daily	An accuracy of 98.53% was observed
Sharma et al. (2021)	Decision tree	Daily	Decision trees are useful for risk evaluation
Nurkholis et al. (2022)	C5.0 decision tree	Daily	A high accuracy was observed
Kaya et al. (2023)	Feed forward neural network	Daily	An accuracy of 93.55% and RMSE of 0.254 were observed
Mislan et al. (2022)	Back propagation neural network	Monthly	MSE of 0.00096 was observed
Aizansi et al. (2024)	Multi-layer perceptron neural network, LSTM, climatology forecasts	Monthly	Multi-layer perceptron outperformed LSTM

Ejike et al. (2021)	Logistic regression	Daily	An accuracy of 84% was observed
Khan et al. (2024)	“Logistic regression, decision trees, multi-layer perceptron, and random forest”	Daily	Logistic regression had highest accuracy
Moorthy and Parmershawaran (2022)	WOAK, KNN	Daily	Hybrid model consisting of WOAK and KNN outperformed either model
Huang et al. (2020)	WKNN, support vector machine	Daily	WKNN was at par with support vector machine
Lee et al. (2022)	Artificial neural network	Monthly	RMSE value of 34.75% was observed on test subset
Findawati et al. (2021)	“Naïve Bayes, K-nearest neighbor, and C4.5”	Daily	KNN had highest accuracy

Yu and Haskins (2021)	"Deep neural network, wide neural network, deep and wide neural network, reservoir computing, long short term memory, support vector machine, and K-nearest neighbor"	Monthly	KNN had highest MSE and RMSE
Setya et al. (2023)	Linear regression, KNN	Monthly	KNN had better RMSE and MAE compared to linear regression
Dawoodi and Patil (2020)	KNN	Daily	An accuracy of 96% was observed
Wolfensberger et al. (2021)	Random forest, QPE	Daily	Random forest was better than QPE

~~*Table 2.1 Table of Summary*~~

CHAPTER 3

METHODOLOGY

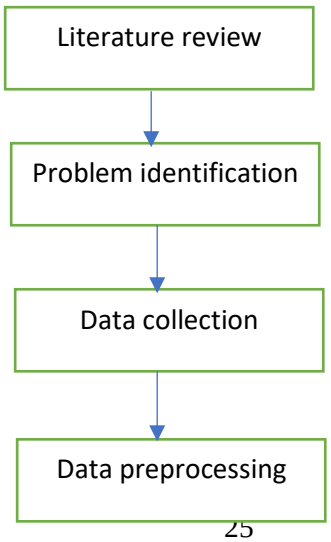
3.1 Introduction

This chapter presents the steps that will be followed in identifying the machine learning algorithm that provides the highest accuracy in predicting rainfall in Selangor. The steps involved are exhaustive review of available literature, identifying the problem to be investigated, collecting relevant data, pre-processing data to assure its suitability, model training, tuning model parameters, and evaluating models. This structured approach will ensure all critical steps are followed. It is expected this approach will help in meeting study objectives.

3.2 Research Design

This research design will act like a blueprint that will be followed in every stage of the study. The core objective is to compare machine learning algorithms and identify the algorithm that provides the highest prediction accuracy. A data driven approach is followed whereby historical weather data such as precipitation, temperature, humidity, and windspeed are the foundation of the study. A data science lifecycle that involves data gathering, pre-processing, parameter tuning, and model evaluation is followed.

3.3 Data Science Methodology



Please make the figure in one page.
Don't separate like this.

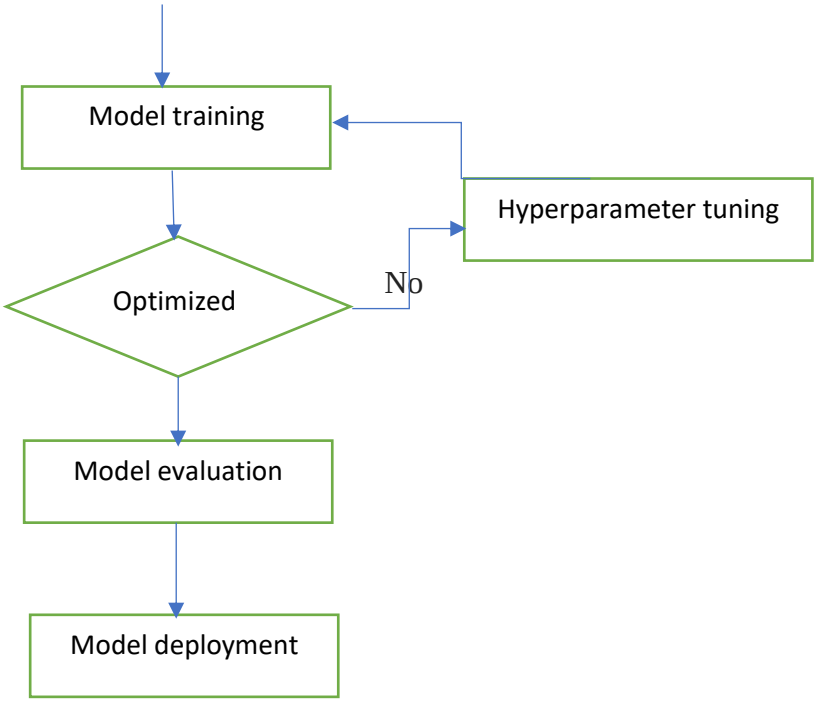


Figure 1 Data Science Methodology

Please correct the format of writing the figure caption.

3.3.1 Literature Review

The first step in carrying out a study is reviewing available literature. Extant literature on machine learning models used for predicting rainfall was reviewed. From reviewed literature it was evident machine learning is an established technique in rainfall forecasting. Reviewed literature revealed machine learning models are primarily used for forecasting the amount of rainfall or classifying rainfall to several categories such as rain/no rain or intensity of rainfall such as low/medium/high. To a lesser extent machine learning were also used to identify critical factors that affect rainfall. Commonly used machine learning methods were support vector machines, decision trees, K-nearest neigbour, logistic regression, gradient boosting, XGBOOST, linear regression, and artificial neural networks. With the exception of logistic regression all the other machine learning models can be used to predict a quantitative amount of rainfall. It was evident in almost all studies a train and test subset were used. This provides a subset for training the model and another subset not used for training that will be used to evaluate model performance. Reviewed literature showed data preprocessing steps such as checking missing values, imputing missing values, checking out of range values, and normalizing quantitative variables to a common range are critical to performance of a machine learning model. From the literature it was observed that some machine learning models have hyperparameters that need to be tuned to achieve high prediction accuracies. These principles that are well established in the literature will be incorporated in this study.

3.3.2 Problem Identification

Climate change has resulted in disruption of established weather patterns. This is a global phenomenon that can lead to extreme rainfall events such as too little or too much rainfall. These events have significant impact on public health, infrastructure, and agriculture. Although economic activities in Selangor are not primarily agricultural, extreme rainfall events need proper planning and mitigation. As a largely urbanized area, flooding from extreme rainfall events such as too much rainfall can cause major disruptions in infrastructure such as public transport, water supply, and waste management. Similarly, too little rainfall can disrupt water supply in urban areas. In rural areas of Selangor where crops such as palm and rubber are grown as well as livestock rearing, these extreme rainfall events can be debilitating. Too little or too much rainfall can cause crop failure. Literature reviewed showed mitigation measures such as changing types of crops or crop cycles were not adequate. These challenges make accurate rainfall prediction an essential strategy in planning and management within the Selangor state government. It is these challenges that were the main motivation of this study. This study aims to investigate if machine learning models can be used to produce accurate rain forecasts. These forecasts will be extremely useful to state government planners.

3.3.3 Data Collection

A dataset consisting of five variables which are date, average temperature, wind speed, relative humidity and precipitation will be used. Use of these variables is well established in the literature. The target variable will be precipitation and the main objective of this study is to evaluate performance of machine learning models in predicting this variable. The predictors will be the other variables except date. The date variable will be useful in building time series models such as ARIMA. The selected dataset consists of daily observations covering the period between 2012 and 2020.

3.3.4 Data Preprocessing

The selected dataset is expected to have some data quality issues. Exploratory data analysis will be used to identify missing values, values that are not within the expected range, and to understand the distribution of variables. Any missing values will be replaced with the mean value to avoid altering the distribution of variables. Any values that are not withing the

expected range will be dropped in the analysis. To ensure all variables have an equal contribution to the model, each variable will be normalized. This will ensure all variables have a common range. In addition, the original daily data were combined into weekly data to reduce noise and show bigger trends in climate behaviour. A ratio of 80% to 20% will be used to split the dataset into train and test subsets. The train subset will be used for model training while the test subset will be used for model evaluation. These principles are well established in reviewed literature.

3.3.5 Model Training

Are you using all these methods?

The models that will be investigated in this study are: artificial neural networks, support vector machines, decision trees, multiple linear regression, K-nearest neighbour, random forests, gradient boosting, and ARIMA. With the exception of linear regression all the other models have a set of parameters that will need to be tuned to achieve the highest prediction accuracy. These parameters are discussed for each model.

The artificial neural network has three architectural parameters that specify the general structure. They are layers, neurons in each layer, and activation functions. The layers and number of neurons will be used to achieve a balance between overfitting and long training time. Activation functions such as ReLu, Tanh, and Sigmoid will be used to capture non-linear patterns in the data. Various training parameters such as learning rate, batch size, epochs, and optimization methods such as SGD, RMSprop, and Adam will be examined to understand their influence on model accuracy. The dropout rate, L1, and L2 will be used to control overfitting.

The hyperparameters of a support vector machine that will need tuning are kernel, regularization, and epsilon. A non-linear relationship is expected in the data. Therefore, only radial basis and polynomial kernels will be examined. The regularization parameter will be tuned to control overfitting in the model. Epsilon will be tuned to control prediction accuracy.

The K-nearest neighbour hyperparameters that will be tuned are neighbours and distance metrics. The number of neighbours will be used to control overfitting. Various distance metrics such as Euclidean, Manhattan, and Minkowski will be examined.

The random forest hyperparameters that will be tuned are: maximum depth, samples per leaf/tree, maximum features/leaf nodes, and split criterion. Tuning will ensure the model adequately captures the relationships in the data while avoiding overfitting or underfitting.

Gradient boosting parameters such as trees, learning rate, depth, split, subsampling, and features will be tuned to minimize overfitting and maximize prediction accuracy.

An ARIMA model requires optimal identification of p, d, and q parameters. Visual inspection and stationarity tests will be used to identify an optimal differencing order. The autocorrelation and partial autocorrelation functions will be used to identify optimal p and q parameters.

The R statistical software will be used for exploratory data analysis and model training. This software was selected because it is freely available and provides extensive data visualization and algorithm capabilities.

3.3.6 Model Evaluation and Comparison

Three model evaluation metrics which are Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2) will be used to examine performance of models under investigation.

RMSE captures the square root of the average squared differences between predicted and actual observations. It shows the extent of large errors and is useful for identifying large deviations in rainfall predictions. RMSE is easy to interpret as it is expressed in units of the response variable but has the limitation of not adequately capturing the influence of outliers. The formula for RMSE is:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Where:
- y_i: Actual of observation i
- $\hat{y}_i$: Prediction of observation i
- n: Number of observations
- Σ: Summation from 1 to i

MAE captures the average difference in the absolute predicted and actual values. This provides a simple measure of prediction accuracy. MAE differs from RMSE as it considers all errors equal, making it robust against outliers. The formula for MAE is:

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

Where:
- y_i: Actual of observation i
- $\hat{y}_i$: Prediction of observation i
- n: Number of observations
- Σ: Summation from 1 to i

R^2 captures the extent to which the model explains the variation in the target variable. An R^2 value close to 1 shows the model is very good at capturing a high degree of the variation, while a value close to zero is indicative of poor predictive performance. The formula for R^2 is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where:
- y_i: Actual of observation i
- $\hat{y}_i$: Prediction of observation i
- n: Number of observations
- Σ: Summation from 1 to i

These metrics will be very helpful in understanding the models under investigation. The MAE and RMSE will provide a quantitative value that indicates the difference between actual and predicted rainfall values. This will be useful in identifying the model that provides the best accuracy. R^2 indicates the extent of model overfitting or underfitting. Therefore, comparison of these three metrics will provide a comprehensive performance evaluation.

Tables will be used to present the performance metrics of each model. This will facilitate easy comparison of the various models.

3.3.7 Deployment

The selected machine learning model will be deployed as a prototype application to demonstrate its practical use. This application could be integrated into an early warning system or a web-based platform to provide real-time rainfall forecasts for stakeholders such as farmers, city planners, and disaster management authorities. Deployment may involve creating a Python-based application with APIs to deliver actionable insights effectively.

```
# Clean and define both together
features = ['Temp_avg','Relative_Humidity', 'Wind_kmh']
data = data.dropna(subset=features + ['Rain_Today'])  # ensure no NaNs anywhere

X = data[features]
y = data['Rain_Today']

# Now split safely
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=89
)
```

```
model = LogisticRegression()
model.fit(X_train, y_train)
```

```
LogisticRegression()
```

```
y_pred = model.predict(X_test)

print("\nClassification Report:")
print(classification_report(y_test, y_pred))

print("\nAccuracy Score:", accuracy_score(y_test, y_pred))
```

Make sure you use the correct format for figure caption.
The numbering of the figures should follow the chapter.

Figure 2 Logistic Regression

Figure 2 shows the coding on how Logistic Regression works. The input features are Temp_avg, Relative_Humidity, Wind_kmh and the target feature is Rain_Today. The reason of Rain_Today be a target feature is because it is the output model to predict.

```
from sklearn.ensemble import RandomForestClassifier
```

```
X = data[['Temp_avg','Relative_Humidity', 'Wind_kmh']]
y = data['Rain_Today']
```

```
rf_model = RandomForestClassifier(random_state=89)
rf_model.fit(X_train, y_train)
```

```
RandomForestClassifier(random_state=89)
```

```
y_pred_rf = rf_model.predict(X_test)                    # predicted class (0 or 1)
y_pred_proba_rf = rf_model.predict_proba(X_test)[:, 1]  # predicted probability for class 1
```

Figure 3 Random Forest

Figure 3 shows the coding on how Random Forest works. The input features are Temp_avg, Relative_Humidity, Wind_kmh and the target feature is Rain_Today.

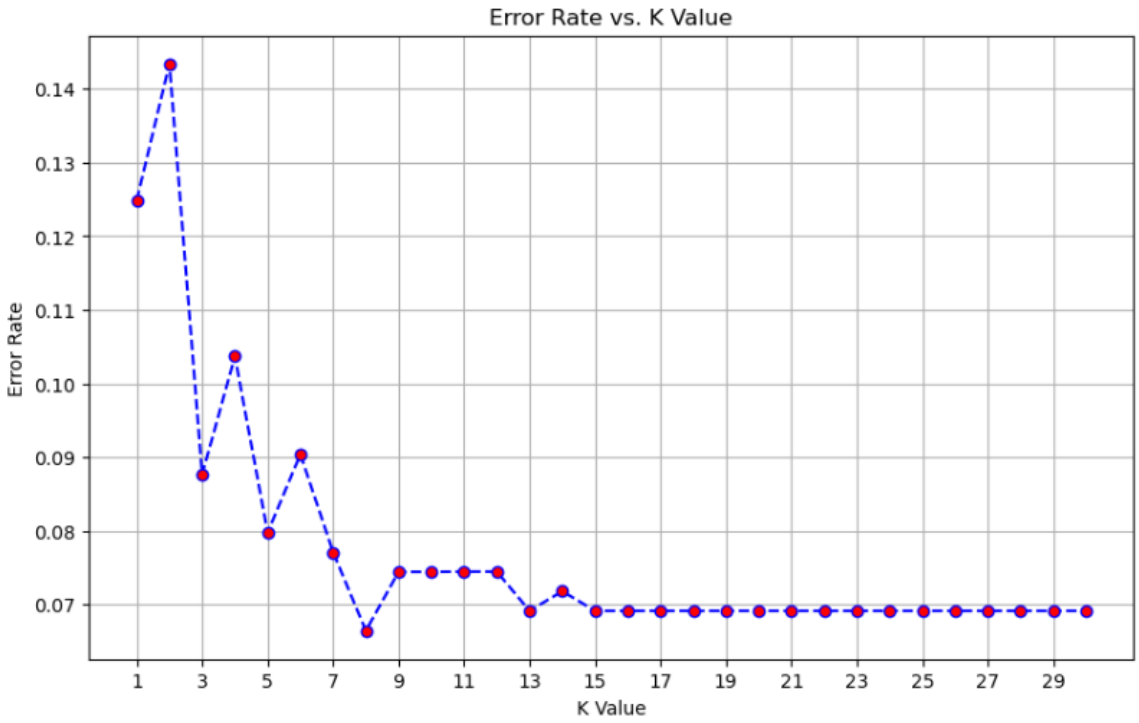
```
from sklearn.model_selection import cross_val_score
from sklearn.neighbors import KNeighborsClassifier
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import label_binarize
from sklearn.multiclass import OneVsRestClassifier
```

```
# Function to find the best k value
def find_best_k(X_train, y_train, max_k=30):
    error_rate = []  # To store error rates for each k

    # Loop through values of k from 1 to max_k
    for k in range(1, max_k + 1):
        knn = KNeighborsClassifier(n_neighbors=k)
        # Perform cross-validation and compute the mean error rate
        scores = cross_val_score(knn, X_train, y_train, cv=5, scoring='accuracy')
        error_rate.append(1 - scores.mean())  # Error = 1 - Accuracy

    return error_rate
```

```
# Find error rates for k values
max_k = 30  # Check k values from 1 to 30
error_rates = find_best_k(X_train, y_train, max_k=max_k)

# Plot the error rates
plt.figure(figsize=(10, 6))
plt.plot(range(1, max_k + 1), error_rates, color='blue', linestyle='dashed', marker='o', markerfacecolor='red')
plt.title('Error Rate vs. K Value')
plt.xlabel('K Value')
plt.ylabel('Error Rate')
plt.xticks(range(1, max_k + 1, 2))
plt.grid()
plt.show()
```

Figure 4 K-Nearest Neighbour

Figure 4 shows the coding of K-Nearest Neighbour (KNN). Import important tools and the input features and target feature will be the same. For KNN, we must find the optimal value of k hence why there is a coding on how to find the optimal value of k.




```
# Find the best k value
optimal_k = np.argmin(error_rates) + 1  # Add 1 because index starts from 0
print(f"The best value for k is: {optimal_k}")
```

```
The best value for k is: 8
```

Figure 5 Finding Optimal value of K

We choose the lowest value of error. Hence, we choose 8 because the best value of k is 8.

```
# Train the KNN model with the best k value
final_knn_model = KNeighborsClassifier(n_neighbors=optimal_k)
final_knn_model.fit(X_train, y_train)

# Make predictions
y_pred_final_knn = final_knn_model.predict(X_test)
y_pred_proba_final_knn = final_knn_model.predict_proba(X_test)[:, 1]
```

```
accuracy_knn = accuracy_score(y_test, y_pred_final_knn)
print(f"Accuracy of KNN with k={optimal_k}: {accuracy_knn:.2f}")
```

Figure 6 K-NN

Then, we use the best value of K to make predictions.

```python
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import classification_report, confusion_matrix
import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense

model = Sequential([
    Dense(16, input_dim=X_train.shape[1], activation='relu'),
    Dense(8, activation='relu'),
    Dense(1, activation='sigmoid')  # sigmoid for binary classification
])

model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])

history = model.fit(X_train, y_train, epochs=50, batch_size=16, validation_split=0.2, verbose=1)

loss, accuracy = model.evaluate(X_test, y_test)
print(f"\nTest Accuracy: {accuracy:.2f}")

y_pred = (model.predict(X_test) > 0.5).astype("int32")
print("\nClassification Report:")
print(classification_report(y_test, y_pred))
```

Figure 7 Artificial Neural Network

Figure 7 shows how Artificial Neural Network (ANN) works. Import important tools, build the model, and train it.

The results in Chapter are not enough.

CHAPTER 4

DATA ANALYSIS, RESULTS AND DISCUSSION

4.1 Introduction

This chapter will present the expected outcomes from the study. After carefully following the methodology developed earlier all study objectives will be achieved. The broad objective of the study is to investigate the potential of using machine learning in planning and management of extreme rainfall events in Selangor. Insights obtained from machine learning predictions will be used for agriculture, disaster, and water management planning.

4.2 Expected Outcomes

This study is expected to meet its objectives. The first objective is to employ machine learning for rainfall prediction. This objective has been addressed through a comprehensive review of existing literature, which demonstrates the effectiveness of machine learning algorithms such as artificial neural networks, support vector machines, random forests, linear regression, K-nearest neighbours, gradient boosting, and ARIMA in forecasting rainfall. The literature also emphasizes the importance of practices such as data quality checks, data normalization, and appropriate train/test splits for ensuring model accuracy. Additionally, widely used evaluation metrics including RMSE, MAE, and R-squared will be adopted in this project to assess model performance.

The second objective will be to train identified machine learning algorithms using the data specified in the methodology. This objective has not been achieved. The methodology specified earlier will be followed in training each of the selected models. It is expected careful tuning of parameters will train models that balance computational cost, accuracy, and overfitting.

The third objective will be to identify the machine algorithm that provides the highest accuracy in rainfall prediction. This objective has not yet been met and it will only be achieved after examining results from objective 2. After training the models on the train subset, the

performance of the models on the test subset will be examined using evaluation metrics and test subset. It is expected comparison of evaluation metrics will identify the algorithm with the highest accuracy.

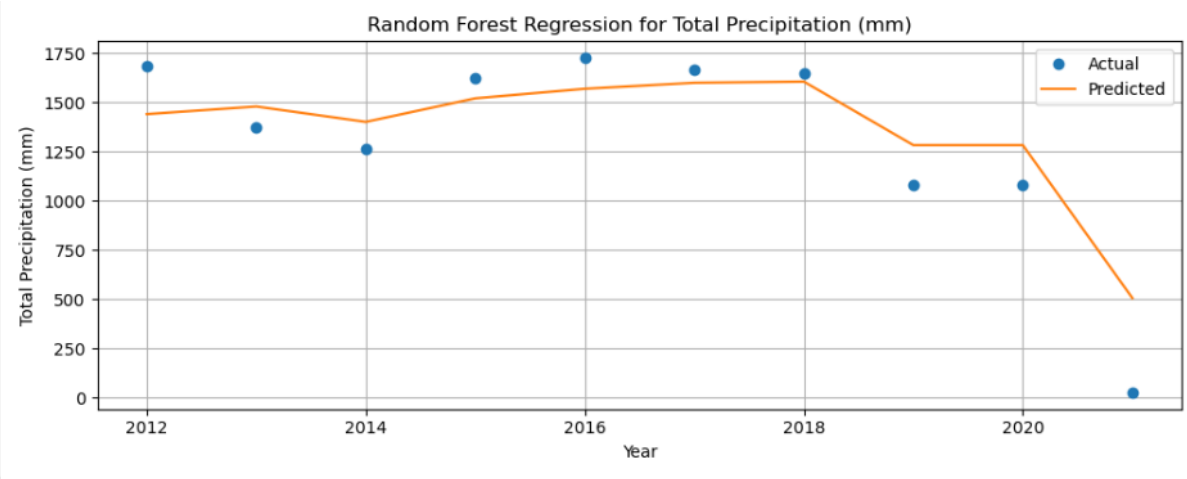


Figure 8 Random Forest Regression

The blue dots represent the actual total precipitation per year, while the orange line shows the predicted values from the model. The model uses yearly averages of temperature, humidity, and wind speed to estimate total precipitation.

Please correct the format.

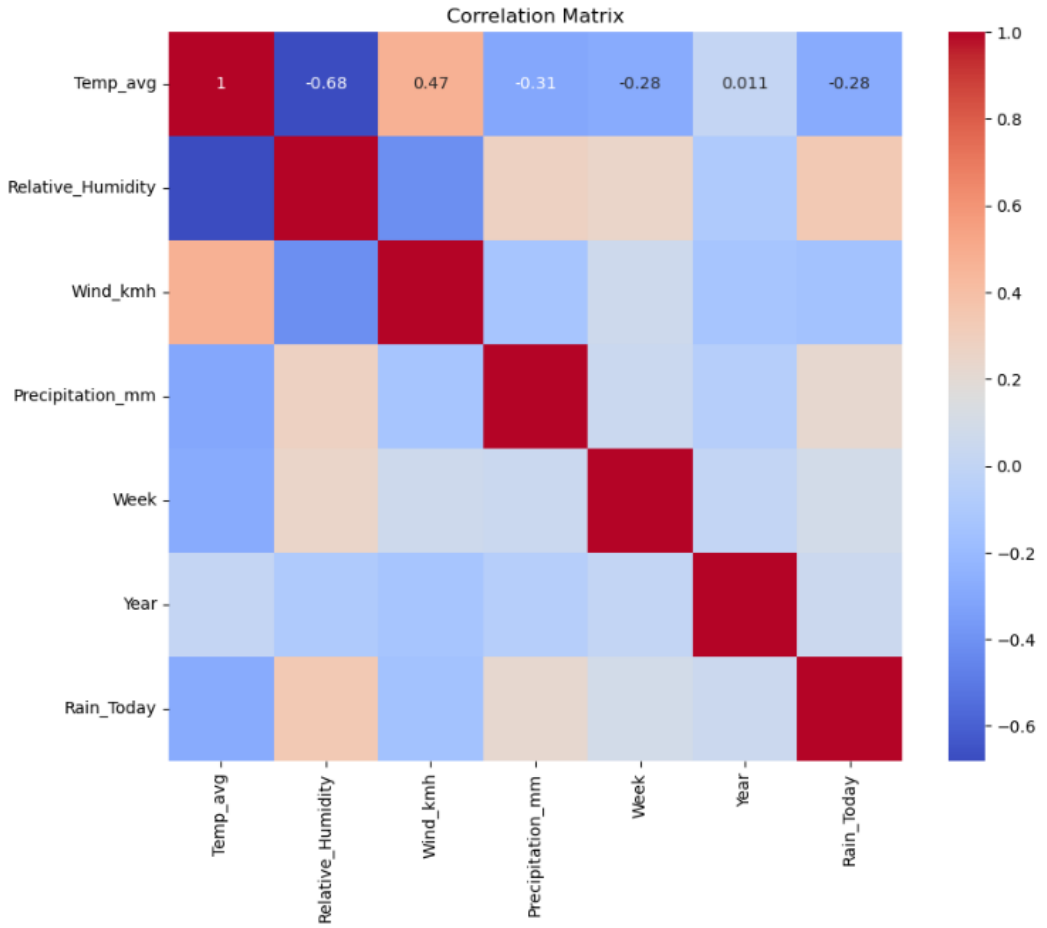


Figure 9 Correlation Matrix

The correlation matrix shows some important relationships between the weather variables. Temperature and humidity have a strong negative relationship, meaning when the temperature goes up, humidity usually goes down. Temperature also has a moderate positive link with wind speed, so higher temperatures often come with stronger winds. There is a weak negative connection between temperature and rainfall, suggesting that hotter days tend to have less rain. Rainfall and the "Rain Today" variable have a moderate positive link, which makes sense since more rain usually means it rained that day. The week and year don't strongly affect the other variables, but they may still help track changes over time. Overall, temperature, humidity, and wind are useful for predicting rainfall.

4.3 Conclusions

In conclusion, this research will build and evaluate machine learning models capable of accurately forecasting rainfall in Selangor. Using weather data and appropriate machine learning algorithms it is expected this study will identify a machine learning algorithm that can be incorporated into an early warning system. Such an early warning system will be critical to success of agriculture, infrastructure, and water management planning within Selangor. This study will demonstrate the value and limitations of using machine learning algorithms in rainfall prediction.

The findings are expected to provide actionable insights for various stakeholders, enabling better resource management, flood prevention, and agricultural planning. However, just like any other study this study will also have limitations. These limitations will only be fully clear after the project is completed. The findings of this study will then require interpretation in consideration of limitations.

REFERENCES

Abbot, J. (2024). Rainfall forecasting at long lead times for eastern Australia using artificial neural networks. *Neural Computing and Applications*, *36*(11), 5927–5953. https://doi.org/10.1007/s00521-023-09386-z

Aïzansi, A. N., Ogunjobi, K. O., & Ogou, F. K. (2024). Monthly rainfall prediction using artificial neural network (case study: Republic of Benin). *Environmental Data Science*, *3*. https://doi.org/10.1017/eds.2024.10

Al-Mahdawi, H. K., Alkattan, H., Subhi, A. A., Al-Hadrawi, H. F., Abotaleb, M., Ali, G. K., Mijwil, M. M., Towfeek, A. K., & Helal, A. H. (2023). Analysis and prediction of rainfall using support vector machine (SVM) in the city of Najaf. *Deleted Journal*, *2023*, 46–54. https://doi.org/10.58496/bjml/2023/009

Anshul. (2024, December 30). *Logistic Regression: A Comprehensive Tutorial*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/#:~:text=Logistic%20Regression%20is%20another%20statistical,pass%20this%20exam%20or%20not.

Anwar, M. T., Winarno, E., Hadikurniawati, W., & Novita, M. (2021). Rainfall prediction using Extreme Gradient Boosting. *Journal of Physics Conference Series*, *1869*(1), 012078. https://doi.org/10.1088/1742-6596/1869/1/012078

Ashok, S. P., & Pekkat, S. (2024). Performance assessment of rainfall forecasting models for urban Guwahati City using machine learning techniques and singular spectrum analysis. *Journal of Water and Climate Change*, *15*(4), 1565–1587. https://doi.org/10.2166/wcc.2024.465

Bari, S. H., Shourov, M. M. H., Rahman, M. T. U., & Ray, S. (2021). Forecasting monthly precipitation in Sylhet City using ARIMA model. *ResearchGate*. https://www.researchgate.net/publication/272744442_Forecasting_Monthly_Precipitation_in_Sylhet_City_Using_ARIMA_Model

Bhardwaj, R., & Duhoon, V. (2021). Study and analysis of time series of weather data of classification and clustering techniques. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2020, Volume 1* (pp. 257-270). Springer Singapore.

Bochenek, B., & Ustrnul, Z. (2022). Machine Learning in Weather Prediction and Climate Analyses—Applications and Perspectives. *Atmosphere*, *13*(2), 180. https://doi.org/10.3390/atmos13020180

Bouallègue, Z. B., Clare, M. C. A., Magnusson, L., Gascón, E., Maier-Gerber, M., Janoušek, M., Rodwell, M., Pinault, F., Dramsch, J. S., Lang, S. T. K., Raoult, B., Rabier, F., Chevallier, M., Sandu, I., Dueben, P., Chantry, M., & Pappenberger, F. (2024). The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context. *Bulletin of the American Meteorological Society*, *105*(6), E864–E883. https://doi.org/10.1175/bams-d-23-0162.1

Cui, Z., Qing, X., Chai, H., Yang, S., Zhu, Y., & Wang, F. (2021). Real-time rainfall-runoff prediction using light gradient boosting machine coupled with singular spectrum analysis. *Journal of Hydrology*, *603*, 127124. https://doi.org/10.1016/j.jhydrol.2021.127124

Du, J., Liu, Y., Yu, Y., & Yan, W. (2021). A prediction of precipitation data based on support vector machine and particle swarm optimization (PSO-SVM) algorithms. *Algorithms*, *10*(2), 57. https://doi.org/10.3390/a10020057

Ejike, O., Ndzi, D. L., & Al-Hassani, A. H. (2021, June). Logistic regression based next-day rain prediction model. In *2021 International Conference on Communication & Information Technology (ICICT)* (pp. 262-267). IEEE.

Ehteram, M., Ahmed, A. N., Khozani, Z. S., & El-Shafie, A. (2023). Convolutional Neural Network -Support Vector Machine Model-Gaussian Process Regression: A New Machine Model for Predicting Monthly and Daily Rainfall. *Water Resources Management*, *37*(9), 3631–3655. https://doi.org/10.1007/s11269-023-03519-8

Findawati, Y., Astutik, I. I., Fitroni, A. S., Indrawati, I., & Yuniasih, N. (2022, December). Comparative analysis of Naïve Bayes, K Nearest Neighbor and C. 45 method in weather forecast. In *Journal of Physics: Conference Series* (Vol. 1402, No. 6, p. 066046). IOP

Publishing.

Goodfellow, I., Bengio, Y., & Courville, A. (2021). Deep Learning. MIT Press.

Hapsari, D. P., Utoyo, M. I., & Purnami, S. W. (2020). *A prediction of rainfall data based on support vector machine with stochastic gradient descent*. https://www.semanticscholar.org/paper/A-Prediction-of-Rainfall-Data-Based-On-Support-With-Hapsari-Utoyo/e7589c2e3d814b077d617a03fd4026a493a6f8f3

Hayaty, N., Kurniawan, H., Rathomi, M. R., Chahyadi, F., & Bettiza, M. (2023). Rainfall Prediction with Support Vector Machines: A Case Study in Tanjungpinang City, Indonesia. *BIO Web of Conferences*, *70*, 01003. https://doi.org/10.1051/bioconf/20237001003

Hill, A. J., Schumacher, R. S., & Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado. (2021). Forecasting Excessive Rainfall with Random Forests and a Deterministic Convection-Allowing Model. In *Weather and Forecasting* (Vol. 36, pp. 1693–1711) [Journal-article]. https://doi.org/10.1175/WAF-D-21-0026.1

Huang, M., Lin, R., Huang, S., & Xing, T. (2022). A novel approach for precipitation forecast via improved K-nearest neighbor algorithm. *Advanced Engineering Informatics*, *33*, 89–95. https://doi.org/10.1016/j.aei.2017.05.003

Hsu, S., Sharma, A. K., Tanone, R., & Ye, Y. (2024). Predicting rainfall using Random Forest and CatBoost models. *Proceedings of the World Congress on Civil, Structural, and Environmental Engineering*. https://doi.org/10.11159/icgre24.146

Imon, A. H. M. R., Roy, M. C., & Bhattacharjee, S. K. (2022). Prediction of rainfall using logistic regression. *ResearchGate*. https://doi.org/10.1234/pjsor.v8i3.535

Kala, A., & Vaidyanathan, S. G. (2020, July). Prediction of rainfall using artificial neural network. In *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 339-342). IEEE.

Kassem, Y., Gökçekuş, H., Çamur, H., & Esenel, E. (2021). Application of artificial neural network, multiple linear regression, and response surface regression models in the estimation of monthly rainfall in Northern Cyprus. *Desalination and Water Treatment*, *215*, 328–346. https://doi.org/10.5004/dwt.2021.26525

Khan, M. U. S., Saifullah, K. M., Hussain, A., & Azamathulla, H. M. (2024). Comparative analysis of different rainfall prediction models: A case study of Aligarh City, India. *Results in Engineering*, *22*, 102093. https://doi.org/10.1016/j.rineng.2024.102093

Kundu, S., Biswas, S. K., Tripathi, D., Karmakar, R., Majumdar, S., & Mandal, S. (2023). A review on rainfall forecasting using ensemble learning techniques. *e-Prime - Advances in Electrical Engineering Electronics and Energy*, *6*, 100296. https://doi.org/10.1016/j.prime.2023.100296

Lee, J., Kim, C., Lee, J. E., Kim, N. W., & Kim, H. (2023). Application of artificial neural networks to rainfall forecasting in the Geum River Basin, Korea. *Water*, *10*(10), 1448. https://doi.org/10.3390/w10101448

Mantri, R., Raghavendra, K. R., Puri, H., Chaudhary, J., & Bingi, K. (2021). Weather Prediction and Classification Using Neural Networks and k-Nearest Neighbors. *School of Electrical Engineering, Vellore Institute of Technology, Vellore, India.* https://doi.org/10.1109/icscc51209.2021.9528115

Maaloul, K., & Lejdel, B. (2023). Big data analytics in weather forecasting using gradient boosting classifiers algorithm. In *Communications in computer and information science* (pp. 15–26). https://doi.org/10.1007/978-981-99-4484-2_2

Masui, T. (2024, February 18). All You Need to Know about Gradient Boosting Algorithm − Part 1. Regression. *Medium*. https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502

Malaysian Meteorological Department (2025). Climate Change. https://www.met.gov.my/en/pendidikan/perubahan-iklim-and-kesan-rumah-hijau/

Mislan, N., Haviluddin, N., Hardwinarto, S., Sumaryono, N., & Aipassa, M. (2022). Rainfall monthly prediction based on artificial neural network: A case study in Tenggarong Station, East Kalimantan - Indonesia. *Procedia Computer Science*, *59*, 142–151. https://doi.org/10.1016/j.procs.2015.07.528

Moorthy, R. S., & Parameshwaran, P. (2021). An optimal K-Nearest neighbor for weather prediction using whale optimization algorithm. *International Journal of Applied Metaheuristic Computing*, *13*(1), 1–19. https://doi.org/10.4018/ijamc.290538

Murphy, K. P. (2022). *Machine Learning: A Probabilistic Perspective*. MIT Press.

nomadseason. (2025, January 18). Selangor Climate https://nomadseason.com/climate/malaysia/selangor.html

Nandakumar, S. D., Valarmathi, R., Juliet, P. S., & Brindha, G. (2021). Artificial neural network for rainfall analysis using deep learning techniques. *Journal of Physics Conference Series*, *1964*(4), 042022. https://doi.org/10.1088/1742-6596/1964/4/042022

Nayak, D., Mahapatra, A., & Mishra, P. (2023). A Survey on Rainfall Prediction using Artificial Neural Network. *International Journal of Computer Applications*, *72*(16), 32–40. https://doi.org/10.5120/12580-9217

N, R., S, S., & S, K. (2022). Comparison of Decision Tree Based Rainfall Prediction Model with Data Driven Model Considering Climatic Variables. *Irrigation & Drainage Systems Engineering*, *05*(03). https://doi.org/10.4172/2168-9768.1000175

Analysis of Weather Data for Rainfall Prediction using C5.0 Decision Tree Algorithm. (n.d.). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/10180907

Nuthalapati, N. S. B., & Nuthalapati, N. A. (2024). Accurate weather forecasting with dominant gradient boosting using machine learning. *International Journal of Science and Research Archive*, *12*(2), 408–422. https://doi.org/10.30574/ijsra.2024.12.2.1246

Praveena, R., Babu, T. R. G., Birunda, M., Sudha, G., Sukumar, P., & Gnanasoundharam, J. (2023). Prediction of Rainfall Analysis Using Logistic Regression and Support Vector Machine. *Journal of Physics Conference Series*, *2466*(1), 012032. https://doi.org/10.1088/1742-6596/2466/1/012032

Poola, K., & Sekhar, P. H. (2021). Prediction of rainfall by using extreme gradient boost (XG boost) in Vishakapattanam area, Andhra Pradesh. *www.mathsjournal.com*. https://www.mathsjournal.com/archives/2021/vol6/issue3/PartB/6-3-20

Primajaya, A., & Sari, B. N. (2021). Random Forest Algorithm for prediction of precipitation. *Indonesian Journal of Artificial Intelligence and Data Mining*, *1*(1), 27. https://doi.org/10.24014/ijaidm.v1i1.4903

Rainfall Classification using Support Vector Machine. (2021, November 11). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9640773?casa_token=1jc-AwxnD1gAAAAA:_JzhzW9vmtvscvyGwyG7u7-jPfVR8lXNz27_ZcP3VQWUAjK3HXJSz0-1lvFi81_PsxiSQZvMpTE

Rainfall Prediction using Machine Learning & Deep Learning Techniques. (2020, July 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9155896?casa_token=LxLdxSohNKIAAAAA:K2lvNktG3-9fWX1utc5z0aImybAwWQ7xdL-wUtsI5XEqbIRSYKDlO5ok2NBLW__Ux68Mh7njgHQ

Raniprima, S., Cahyadi, N., & Monita, V. (2024). Rainfall prediction using random forest and decision tree algorithms. *Journal of Informatics and Communication Technology (JICT)*, *6*(1), 110–119. https://doi.org/10.52661/jict.v6i1.253

Raut, A., Theng, D., & Khandelwal, S. (2023, October). Random Forest Regressor Model for Rainfall Prediction. In *2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS)* (Vol. 1, pp. 1-6). IEEE.

Ray, K., Balachandran, S., & Dash, S. K. (2021). Challenges of forecasting rainfall associated with tropical cyclones in India. *Meteorology and Atmospheric Physics*, *134*(1). https://doi.org/10.1007/s00703-021-00842-w

Raniprima, S., 1, Cahyadi, N., Monita, V., & School of Electrical Engineering, Telkom University, Indonesia. (2024). Rainfall Prediction Using Random Forest and Decision Tree Algorithms. In *Journal of Informatics and Communications Technology* (Vol. 6, Issue 1, pp. 110–119) [Journal-article].

Bhardwaj, R., & Duhoon, V. (2021). Study and analysis of time series of weather data of classification and clustering techniques. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2020, Volume 1* (pp. 257-270). Springer Singapore.

Sharma, K. K., Verdon-Kidd, D. C., & Magee, A. D. (2021). A decision tree approach to identify predictors of extreme rainfall events – A case study for the Fiji Islands. *Weather and Climate Extremes*, *34*, 100405. https://doi.org/10.1016/j.wace.2021.100405

Sanaboina, C. S. (2024). A comparative study of different machine learning techniques for forecasting rainfall. *International Journal of Computing and Artificial Intelligence*, *5*(2), 211–219. https://doi.org/10.33545/27076571.2024.v5.i2c.117

Sanches, R. G., Miani, R. S., Santos, B. C. D., Moreira, R. M., Neves, G. Z. D. F., Bourscheidt, V., & Rios, P. A. T. Using Xgboost Models for Daily Rainfall Prediction. *Available at SSRN 4778138*.

Setya, B., Nurhidayatullah, R. A., Hewen, M. B., & Kusrini, K. (2023, October). Comparative Analysis Of Rainfall Value Prediction In Semarang Using Linear And K-Nearest Neighbor Algorithms. In *2023 5th International Conference on Cybernetics and Intelligent System (ICORIS)* (pp. 1-5). IEEE.

Talib, S. a. A., Idris, W. M. R., Neng, L. J., Lihan, T., & Rasid, M. Z. A. (2024). Irregularity and time series trend analysis of rainfall in Johor, Malaysia. *Heliyon*, *10*(9), e30324. https://doi.org/10.1016/j.heliyon.2024.e30324

Talekar, B. (2020). A Detailed Review on Decision Tree and Random Forest. *Bioscience Biotechnology Research Communications*, *13*(14), 245–248. https://doi.org/10.21786/bbrc/13.14/57

The Department of Irrigation and Drainage Malaysia. (2025). Rainfall Data. https://publicinfobanjir.water.gov.my/hujan/data-hujan/?lang=en

The National Oceanic and Atmospheric Administration. (2024). The Challenges and Complexities of Weather Forecasting. https://www.weather.gov/car/weatherforecasting

Velasco, L. C., Aca-Ac, J. M., Cajes, J. J., Lactuan, N. J., & Chit, S. C. (2022). Rainfall Forecasting using Support Vector Regression Machines. *International Journal of Advanced Computer Science and Applications*, *13*(3).

https://doi.org/10.14569/ijacsa.2022.0130329

Vijayalakshmi, C., Sangeeth, K., Josphineleela, R., Shalini, R., Sangeetha, K., & Jenifer, D. (2022, December). Rainfall Prediction using ARIMA and Linear Regression. In *2022 International Conference on Computer, Power and Communications (ICCPC)* (pp. 366-370). IEEE.

Wolfensberger, D., Gabella, M., Boscacci, M., Germann, U., & Berne, A. (2021). RainForest: a random forest algorithm for quantitative precipitation estimation over Switzerland. *Atmospheric Measurement Techniques*, *14*(4), 3169–3193. https://doi.org/10.5194/amt-14-3169-2021

Wani, O. A., Mahdi, S. S., Yeasin, M., Kumar, S. S., Gagnon, A. S., Danish, F., Al-Ansari, N., El-Hendawy, S., & Mattar, M. A. (2024). Predicting rainfall using machine learning, deep learning, and time series models across an altitudinal gradient in the North-Western Himalayas. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-77687-x

Weather Prediction and Classification Using Neural Networks and k-Nearest Neighbors. (2021, July 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9528115?casa_token=MnoohukKgpAAAAAA:aALPNXgQFoXmpjkSQkxLTadKyvYqQ2C2A_R5TjgPrn-3O935SeVpNb0e_Z8Hdho4Ai19eJ-Ugvw

Yin, G., Yoshikane, T., Yamamoto, K., Kubota, T., & Yoshimura, K. (2022). A support vector machine-based method for improving real-time hourly precipitation forecast in Japan. *Journal of Hydrology*, *612*, 128125. https://doi.org/10.1016/j.jhydrol.2022.128125

Yu, N., & Haskins, T. (2021, March 28). *KNN, An Underestimated Model for Regional Rainfall Forecasting*. arXiv.org. https://arxiv.org/abs/2103.15235?utm_source

Yudianto, M. R. A., Agustin, T., James, R. M., Rahma, F. I., Rahim, A., & Utami, E. (2021). Rainfall Forecasting to Recommend Crops Varieties Using Moving Average and Naive Bayes Methods. *International Journal of Modern Education and Computer Science*, *13*(3), 23–33. https://doi.org/10.5815/ijmecs.2021.03.03

Zhuang, H., Lehner, F., & DeGaetano, A. T. (2024). Improved Diagnosis of Precipitation Type with LightGBM Machine Learning. *Journal of Applied Meteorology and Climatology*, *63*(3), 437-453.