Table of Content

1.0 Introduction

A natural disaster known as flooding occurs due to climatic or climatological elements including temperature, precipitation, evaporation, wind, and earth's inherent qualities. Floods and flash floods are common in Malaysia, particularly during the monsoon season on the east coast. The nation experiences more frequent floods either naturally or because of changes in the monsoon brought on by a growth in urban slums. Consequently, it is important to prevent floods for several reasons, including the potential for severe and broad effects on economies, ecosystems, and people. It is related to SDG 15, "Life on Land," and is relevant to the impacts of floods due to its focus on protecting, restoring and promoting sustainable use of terrestrial ecosystems.

Hence, our company, MonSoonProTech, specializes in using modern technology to offer solutions for increased flood mitigation and prediction. The firm provides precise and fast flood forecasts, allowing communities to take preventative action to protect people and property. It does this by combining data analytics and machine learning monitoring. Therefore, five employees from our organization were sent to oversee the project.

No	**Name**	**Role**	**Responsibilities**
1	Abraham Lim Bing Sern	CEO	The overall leadership and the company's strategy.
2	Chong Wei Han	Co-founder	Ensuring the solutions' superior technology.
3	Lim Ka Quan	CTO	Innovation and strategies for technology.
4	Nurul Syahirah binti Abdul Karim	Consultant	Managing client consulting engagements.
5	Pravinkumar a/l Palanisamy	Data Analyst	Providing data-driven insights for decision making

1.1 Project Description

According to the World Health Organisation (WHO), the most common kind of natural disaster is floods, which happen when a water overflow submerges a normally dry area. Heavy rains, quick snowmelt, storm surges from tropical cyclones, or tsunamis in coastal locations are frequently the cause of floods. Normally, there are three main categories of floods which are flash floods, river floods and coastal floods. A flash flood is defined as a flood brought on by an abundance of rain or severe precipitation during a brief period of time usually less than six hours (US Department of Commerce, NOAA, National Weather Service, 2019). When a river's capacity is exceeded by persistent rain or snowmelt, river floods result. Coastal flood was created when Storm surges linked to tropical storms and tsunamis.

Our country Malaysia will face multiple waves of floods at the end of every year. The floods as a serious threat to the residents, houses and other structures. According to the Star (2023), the report said that flooding is worsening in Kelantan, Terengganu and Pahang with the number of evacuees increasing in all three states. Therefore, our goal is to predict whether a flood is likely to occur, which helps to lessen the potential damage caused by floods. By doing so, this prediction will provide residents and property owners with valuable information to take preventive measures and protect their homes.

1.2 Problem Statement

End-of-year floods have emerged as a recurrent challenge, inflicting substantial harm upon human communities, ecosystems, and wildlife. The unpredictable nature of these occurrences compounds the difficulty in effectively preventing and mitigating floods. Recognizing this critical issue, our initiative endeavours to leverage advanced machine learning algorithms to forecast the likelihood of flooding. Through this innovative approach, we seek to significantly diminish flood risks, empowering local communities with timely insights to proactively implement preventative measures and thereby minimize the potential damage inflicted by these inundations.

1.3 Objectives

The objectives of the study are:

1. To reduce the damage of the flood risk to human communities, ecosystems, and wildlife.
2. To develop a model to predict whether a flood is likely to occur.
3. To evaluate the model's performance based on a decision tree, support vector machines (SVM), random forest, gradient boosting, and XGBoost.

1.4 Scope

The project's goal is to address the rising frequency of floods in Sungai Pahang, Malaysia by developing advanced technologies for flood prediction and mitigation, with a particular emphasis on Kelantan, Terengganu, and our focus is Sungai Pahang. Five people working together will perform different tasks using data analytics and machine learning algorithms to lessen the damage that floods due to ecosystems, wildlife, and human settlements. Using techniques like decision trees and SVM along with data collecting from meteorological and environmental elements, the project will develop a machine learning model that can predict different types of floods. In line with SDG 15, which calls for the protection of terrestrial ecosystems, the programme involves communities and stakeholders and aims to improve prediction accuracy through ongoing research. The initiative aims to reduce the negative effects of floods by using creative, data-driven solutions, and it will do so by providing extensive documentation and regular updates to the public.

1.5 Limitation of the Project

The project to develop advanced flood prediction and mitigation technologies in Malaysia faces several limitations that need to be carefully considered. First and foremost, there are significant data challenges, as gathering accurate and comprehensive data for flood prediction can be a complex task. Incomplete or inconsistent data can hinder the effectiveness of the prediction model. Additionally, Malaysia's weather patterns are influenced by various factors, including monsoons and tropical cyclones, leading to unpredictable and rapidly changing weather conditions that make flood prediction more challenging. Limited resources, both in terms of technology and funding, could also potentially restrict the project's capabilities. The complexity of developing machine learning models for flood prediction, potential issues with community engagement, ethical and privacy concerns surrounding data collection, and the need to ensure compliance with regulations further add to the project's limitations. Careful consideration of environmental impact, ongoing model validation, and effective public communication strategies will be essential to address these challenges successfully.

The project is still essential in tackling Malaysia's annual floods, which are a major threat to human settlements, ecosystems, and animals. This is true even with these challenges. By using cutting-edge machine learning techniques to predict flooding, these dangers could be greatly decreased. The project's goal is to reduce the possible harm that floods may do by providing local communities with timely information and preventative actions. The study aims to mitigate flood damage to populations, ecosystems, and animals by creating a prediction model and assessing its efficacy through a range of machine learning methodologies. In line with Sustainable Development Goal 15, the initiative aims to safeguard terrestrial ecosystems with a particular emphasis on Kelantan, Terengganu, and Pahang.

2.0 Project Pipeline

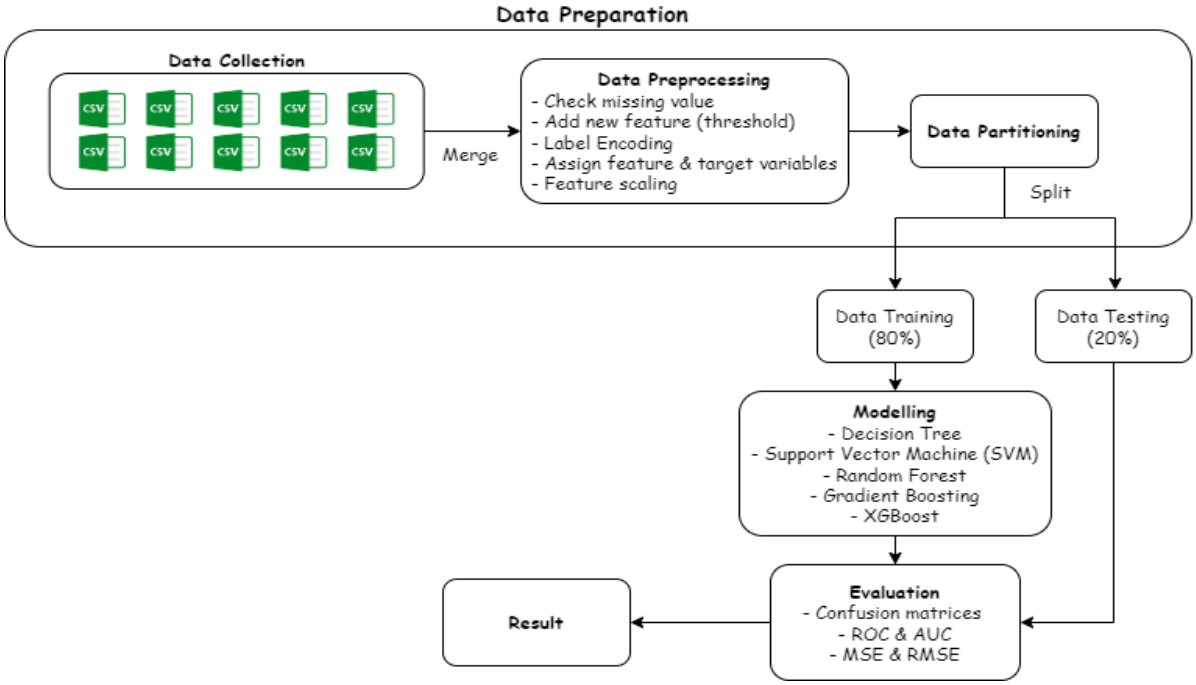

Figure 2.1 Project pipeline

Figure 2.1 shows the pipeline of the project about water level prediction. A few steps are involved in this pipeline, including data preparation, modelling, evaluation, and results. The first step in this study is data collection. There are 8 datasets collected which are streamflow, rainfall, weather, and water level in year 2021 and 2022 respectively. The study area for the data collection is located at Sungai Pahang, Pahang. For more easier to data preprocessing, the dataset is merged into one and several preprocessing techniques will apply such as check missing value, add new features (threshold), label encoding, assign feature and target variable, and feature scaling.

After done the data preprocessing, data partitioning will apply to split the dataset into two subset which is train set (80%) and test set (20%). The subsequent phase involves model development, specifically decision tree, support vector machine (SVM), random forest, gradient boosting and xgboost will be the model training and compared each other in performance evaluation such as confusion matrices, ROC and AUC, and MSE and RSME. Once the evaluation is done, analysis will be made and shown in result part.

3.0 Data Preparation

3.1 Data Description

The data represents daily measurements of various environmental factors, including rainfall, water level, streamflow, and weather conditions. However, the dataset has missing values, which should be considered when analyzing or using the data. Here is a description of each column in the dataset:

Table 3.1 Data description

Variables	Description	Data Type
Day	Day of the month	Integer
Rainfall	Amount of rainfall (mm)	Float
WaterLevel	The level of the water (m)	Float
Streamflow	The level of the stream flow ($m^3 sec^{-1}$)	Float
Weather	Temperature of weather (°C)	Float

3.2 Data Preprocessing

There are eight datasets to merge: rainfall, water level, streamflow, and weather datasets in 2021 and 2022, as shown in Figure 2.1. First, divide the data into one column using the 'Day' column as the identifying variable; the melt function is applied to every dataset. This procedure converts the datasets into a long format, stacking the temperature values into a new column while maintaining the 'Day' column. To add parameters to a database, append instances or observations, or eliminate repetitions and other erroneous data, combine the two datasets (2021 and 2022) for each variable into one. Then, merge these four datasets that have been incorporated into one.

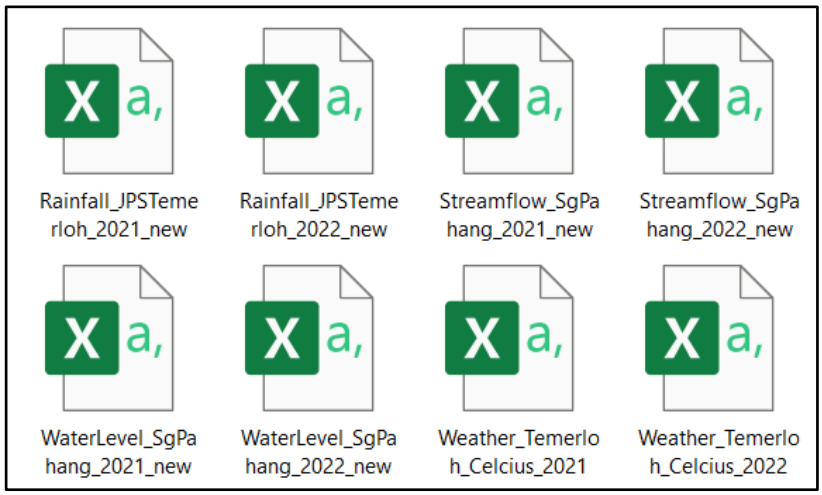

Figure 3.1 Datasets

The first step in data pre-processing is to check the null value in the dataset; null values exist in each attribute. In the initial data analysis, it was observed that some columns, particularly "Day", exhibited missing values due to variations in the number of days across different months. To address this, the rows with missing data in the "Weather" column were removed, eliminating any missing values in the dataset. Thus, the "WaterLevel" column's missing values for the relevant range (days 40 to 60) were filled in using a linear interpolation approach, and the interpolated values were then substituted. Subsequently, it was discovered that question marks indicated unknown data, so they were replaced with NaN first, and the missing rainfall data was filled in with zeros. Moreover, recheck to ensure that there is no more missing data.

Since this research is about classification, the thresholds will be added to classify the water level as either "normal", "alert", "warning" or "danger". This means that there will be a new column added to the dataset called "Threshold". Additionally, the feature and target variables are essential to the training data that the model utilizes to acquire patterns and provide predictions. So, the 'Rainfall', 'StreamFlow' and 'Weather' are assigned as feature variables (X), while the target variable for regression (y_regression) is 'WaterLevel'. Meanwhile, the target variable for classification (y_classification) represents the 'Threshold' labels, including 'normal', 'alert', 'warning' and 'danger' based on the water level.

Furthermore, the hold-out method is applied in the data training since it is the simplest way to partition data. The dataset was split into 20% for data testing and 80% for data training. The next step was using Standard Scaler for feature scaling. It aims to standardize the feature variables by eliminating the mean and scaling to unit variance. It is also known as Z-score normalization. The last step of the data preprocessing was instantiating the label encoder. Label encoder converts categorical labels such as 'normal', 'alert', 'warning', and 'danger' into numerical labels.

3.3 Data Analysis

3.3.1 Heat map

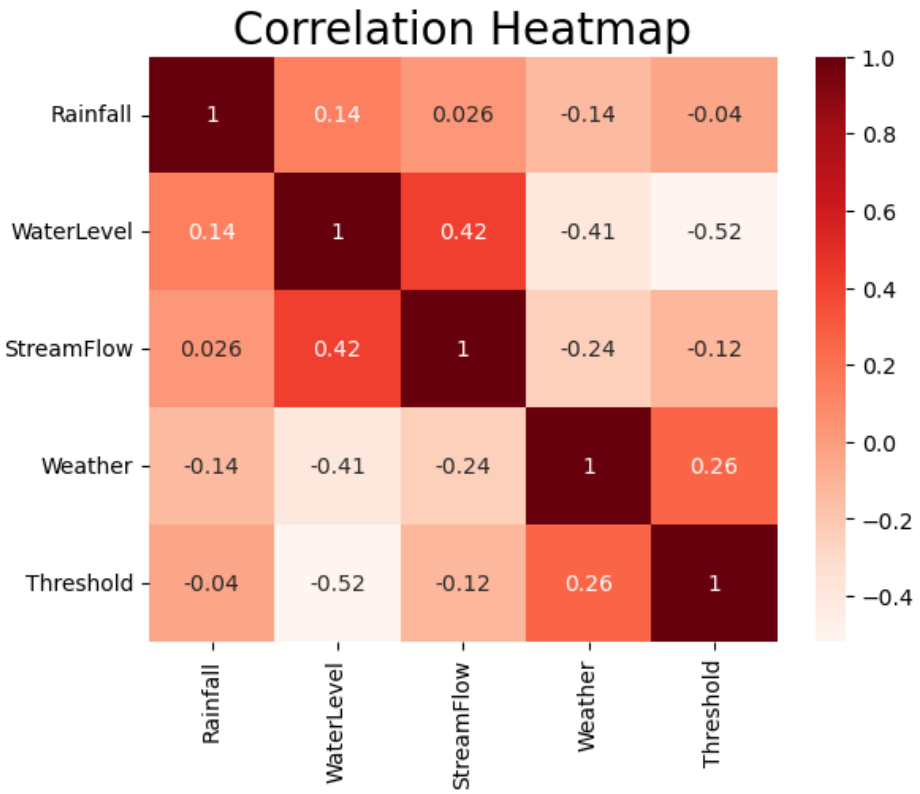

Figure 3.2 Heatmap

Based on the correlation heap map in Figure 3.2 above, there is a slightly strong correlation of about 0.42 between water level and stream flow. A weak correlation appears between water level and rainfall, with a value of 0.14. At the same time, the weather and water level show a slightly strong negative correlation of -0.41.

3.3.2 Line Graph (water level in year 2021)

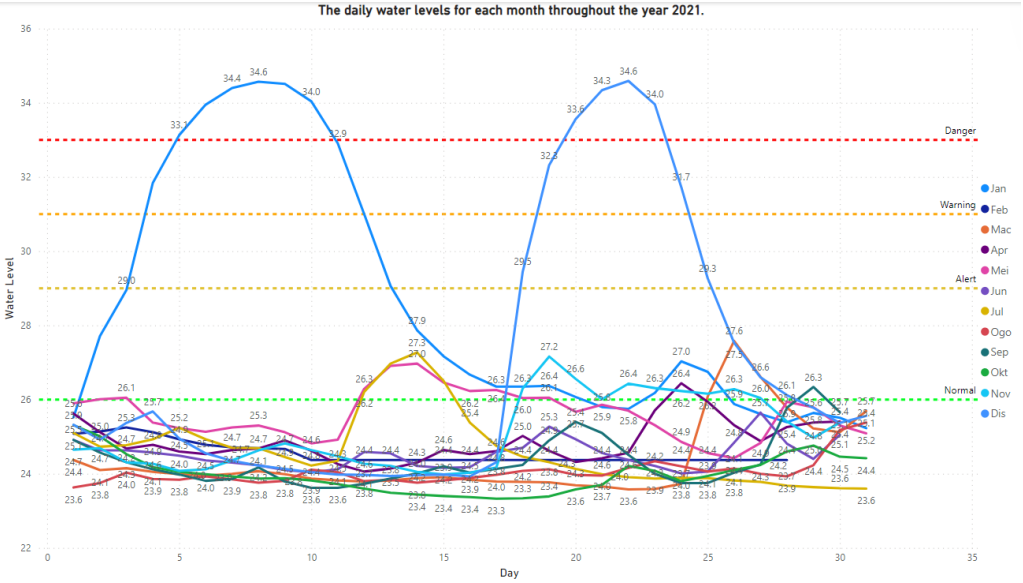

Figure 3.3 Daily water levels for each month throughout the year 2021

Figure 3.3 depicts the daily water levels for each month in 2021. Each month represents different colours with the labels on the right-hand side. There are four different horizontal colours in the graph labelled "Danger", "Warning", "Alert", and "Normal". The highest recorded level, reaching up to 34.6m in January and December respectively that peaks twice in the graph.

Moreover, most of the monthly water levels fluctuate between 23rd and 28th. This graph is valuable for tracking water levels and detecting recurring patterns or trends. For instance, if the water levels continually surpass the "Danger" threshold, it might suggest the need to take action to prevent floods.

3.3.3 Line graph (water level in year 2022)

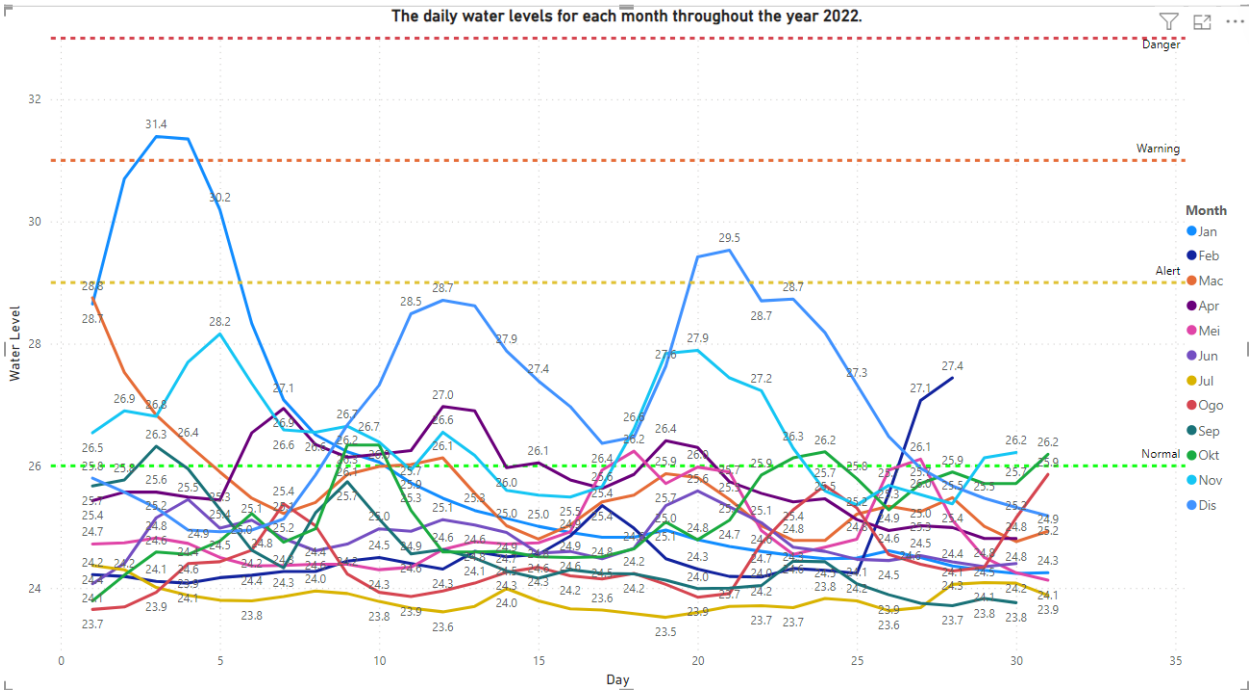

Figure 3.4 Daily water levels for each month throughout the year 2022

Figure 3.4 depicts the daily water levels for each month in 2022. Each month represents different colours with the labels on the right-hand side. There are four different horizontal colours in the graph labelled "Danger", "Warning", "Alert", and "Normal". The highest recorded level, reaching up to 31.4m, is represented by a blue line (Jan) that peaks once in the graph between day 3 and 4. Based on the graph reveals two distinct peaks, occurring on December 11th and December 21st, with values of 28.7m and 29.5m, respectively. Consequently, it can be inferred that there were nearly two instances of flood alerts in December, along with a single wave of flood warnings in early January.

4.0 Developing of Machine Learning Algorithm

4.1 Decision Tree

Decision Tree is a popular non-parametric supervised learning technique for applications involving classification and regression (Saini, 2021). An example of a hierarchical model used in decision support is a decision tree, which shows choices, possible outcomes, chance occurrences, resource costs, and utility. Its internal nodes, leaf nodes, branches, and root nodes make up its hierarchical tree structure, as shown in Figure 4.1.

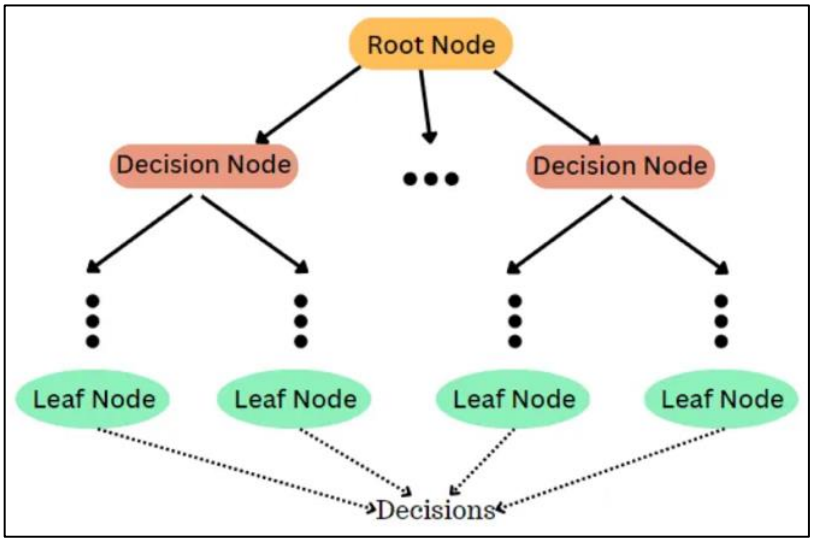

Figure 4.1 Decision Tree (Source: Nidhi, 2023)

4.2 Support Vector Machine (SVM)

According to Sunil (2019), a supervised learning machine learning approach called Support Vector Machine (SVM) may be used for regression or classification problems. However, most of its applications are in classification tasks, including text classification. Each data point is plotted as a point in n-dimensional space where n is represented as the number of features when using the SVM method. A specific coordinate represents the value of each feature. Figure 4.2 illustrates the best hyper-plane that effectively separates the two classes and is then found to do classification.

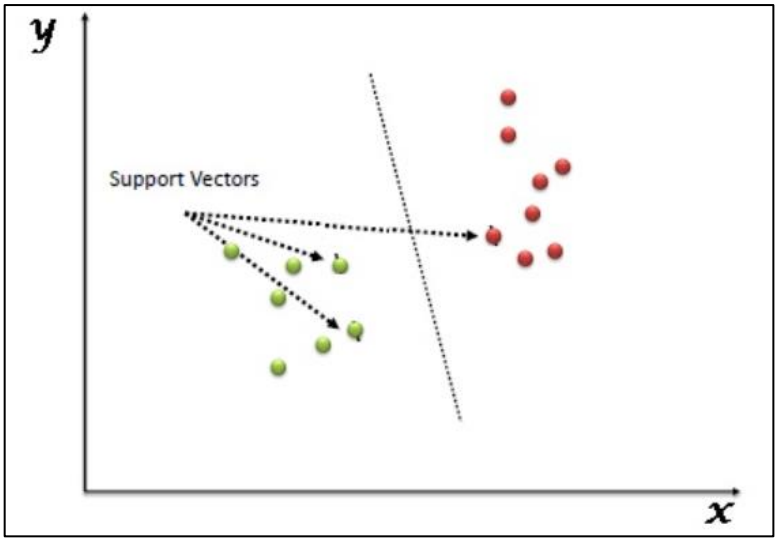

Figure 4.2 SVM (Source: Sunil, 2019)

4.3 Random Forest

Random forest is a versatile and user-friendly machine learning technique that consistently yields excellent results, even without hyper-parameter adjustments (Donges, 2021). Decision trees serve as the random forest's basis estimators. A random forest randomly chooses a collection of characteristics to determine the optimal split at each decision tree node. First, the raw dataset is divided into random subgroups using bootstrapping. Only a random subset of factors is considered at each decision tree node to determine the optimal split. Every subset is fitted using a decision tree model. The average of all the decision trees' predictions is used to determine the final forecast. Figure 4.3 shows the structure of the random forest.

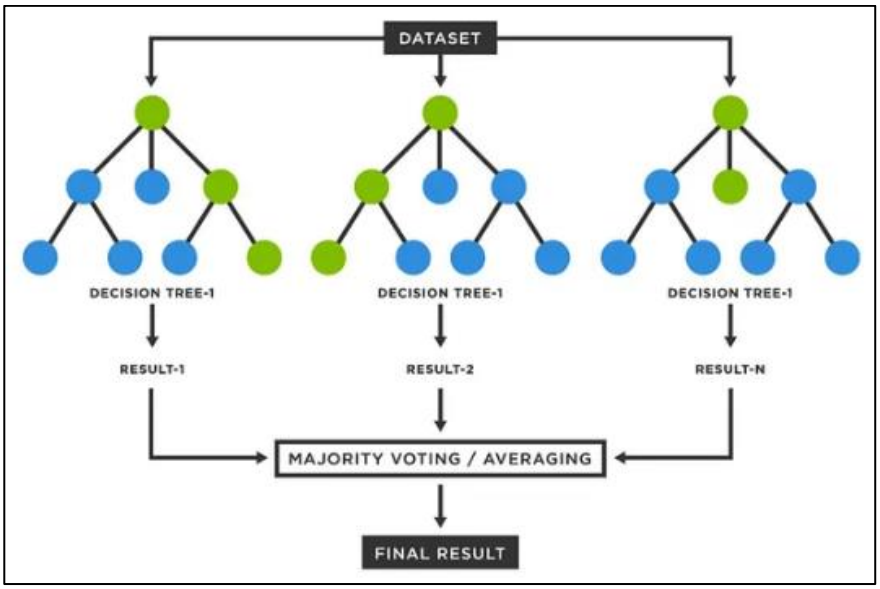

Figure 4.3 Random Forest (Source: Kumar, 2021)

4.4 Gradient Boosting

Gradient boosting is a machine learning method that improves the performance of a model by iteratively adjusting the weights of several weak learners (Saini, 2021b). This process reduces prediction errors and increases the accuracy of the model as it progresses. Figure 4.4 shows the algorithm for Gradient Boosting.

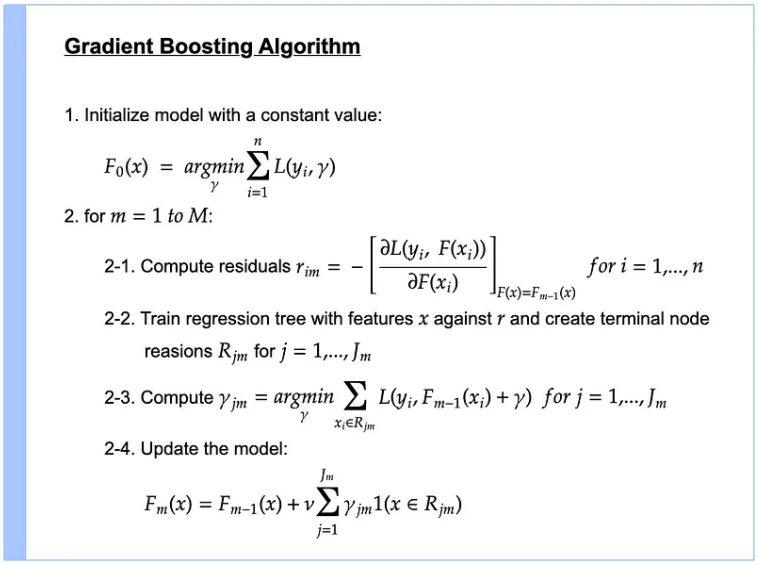

Figure 4.4 Gradient Boosting Algorithm (Source: Masui, 2022)

4.5 XGBoost

XGBoost is a machine learning package that implements a scalable and distributed gradient-boosted decision tree (GBDT) algorithm. The acronym XGBoost stands for Extreme Gradient Boosting. It offers parallel tree boosting and is the dominant machine learning package for regression, classification, and ranking issues. In addition, XGBoost has inherent functionality for parallel processing, enabling efficient training of models on extensive datasets. Figure 5 explains the XGBoost objective function analysis.

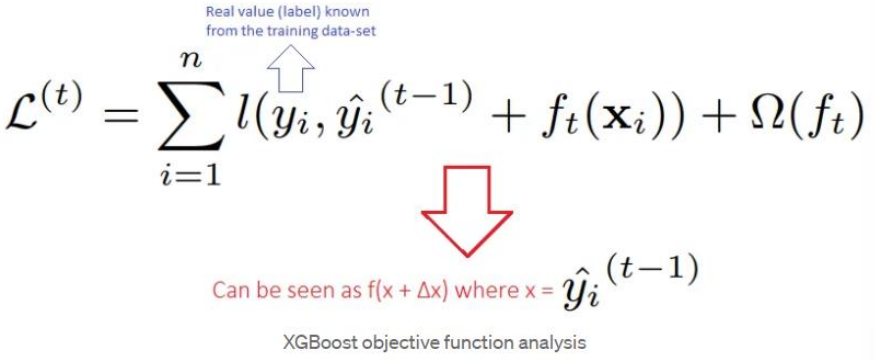

Figure 4.5 XGBoost objective function analysis (Source: Leventis, 2022)

5.0 Result and Conclusion

5.1 Interpretation of the Results

These models are trained to make predictions. The prediction summary is represented in matrix form by the confusion matrix. It assists in identifying the classes that the model is confusing with other classes by displaying the proportion of correct and incorrect predictions for each class (Tiwari, 2022). Most classification models are evaluated based on accuracy since it reduces model error and allows more datasets to be accurately predicted and classified. The fraction of true positive and true negative samples divided by the total number of samples is the approach used to determine the model's accuracy. The formula is constructed as below:

$$Accuracy = \frac{TP+TN}{Total\ samples(N)} \quad (5.1)$$

Besides that, precision and recall also show the performance of the model. Precision has the ability of a classification model to identify only the relevant data points, and it has the fraction of true positive within the true positive and false positive, which the formula is:

$$precision = \frac{TP}{(TP+FP)} \quad (5.2)$$

Recall (Sensitivity) has the ability of a model to find all the relevant cases within a dataset, and it has the fraction of true positive within the true positive and false negative, which the formula is:

$$recall = \frac{TP}{(TP+FN)} \quad (5.3)$$

Figure 5.1 shows the random forest classifier has the highest accuracy in predicting the water threshold value, which is 82.88%, and then SVM, gradient boosting, and XGBoost with 80.14%. The decision tree classifier has the lowest accuracy with only 34%.

In the precision and recall result from Figure 5.2, we also found that the random forest classifier gives high precision and recall, which means that the RF model effectively identifies and classifies instances with a low rate of false positives and false negatives. Following by xgboost, gradient boosting and support vector machine give the same high recall, but the precision is intermediate.

The accuracy and classification report of all the models is shown below:

Score	Model
0.828767	Random Forest
0.801370	Support Vector Machine
0.801370	Gradient Boosting
0.801370	XGBoost
0.753425	Decision Tree

Figure 5.1 Accuracy report

Model	precision	recall	f1-score	support
Random Forest	0.805479	0.828767	0.793750	146.0
XGBoost	0.769780	0.801370	0.780900	146.0
Gradient Boosting	0.767914	0.801370	0.776067	146.0
Decision Tree	0.759315	0.753425	0.754077	146.0
SVM	0.662671	0.801370	0.722114	146.0

Figure 5.2 Classification report

Figure 5.3 shows the performance of Receiver Operating Characteristics (ROC). Meanwhile, Figure 5.4 illustrates that XGBoost has the highest Area Under the Curve (AUC) value at 0.77, indicating the most outstanding performance. At the same time, the lowest AUC is the decision tree, with only 0.59. The AUC values quantify the ability of each model to differentiate between classes. A model's performance improves as the AUC value increases. The ROC curve shows that XGBoost outperforms Random Forest and Gradient Boosting as the most effective model for this task. SVM and Decision Tree exhibit comparatively lower accuracy in differentiating true positives from false positives.

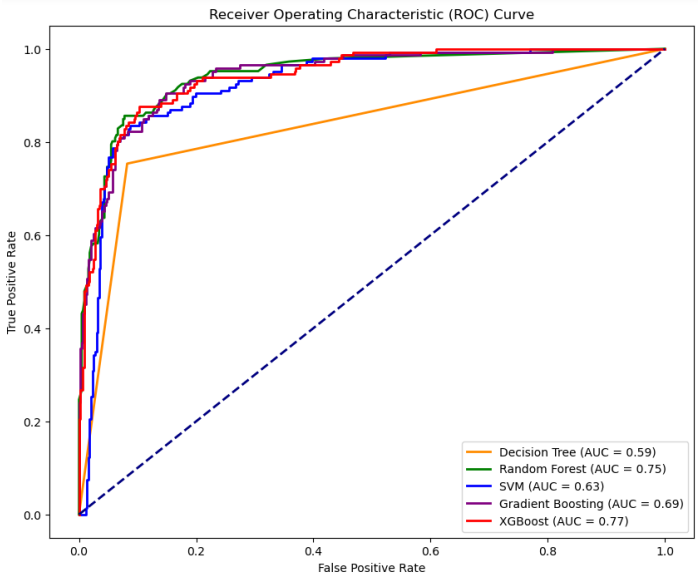

Figure 5.3 Receiver Operating Characteristics (ROC)

Model	AUC
XGBoost	0.767977
Random Forest	0.752685
Gradient Boosting	0.688898
SVM	0.626958
Decision Tree	0.587532

Figure 5.4 AUC report

Figure 5.5 shows the performance of Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). Among the models provided, the Random Forest stands out as the top performer due to its much lower MSE and RMSE values, with 0.58 and 0.76, respectively. With the greatest values of MSE and RMSE, the Decision Tree is the least effective of those provided with the MSE at 0.93 and RMSE at 0.97. The error rates indicate each model's prediction accuracy regarding the objective variable. The model's performance increases as the error rate drops.

Model	MSE	RMSE
Random Forest	0.582192	0.763015
Gradient Boosting	0.650685	0.806650
SVM	0.671233	0.819288
XGBoost	0.726027	0.852072
Decision Tree	0.931507	0.965146

Figure 5.5 MSE and RMSE report

5.2 Conclusion and Recommendations

By comparing the model, we can conclude that the random forest model in machine learning is the most suitable for predicting whether a flood is likely to occur. It performed the best on the evaluation metric with the highest accuracy of the model, 82.87% on the test data compared to the XGBoost, Gradient Boosting, Decision Tree and SVM.

Random Forest performed well overall, with high recall (82.88%), accuracy (80.55%), and an F1-score of 79.38. The models for XGBoost, Gradient Boosting and Decision Trees also demonstrated excellent accuracy, recall, and F1-score values. In contrast, the SVM achieved a lower accuracy (66.27%) and an F1 score of 72.21% as a trade-off for its notable recall (80.14%).

XGBoost was ranked first with an AUC of 76.78%, closely followed by Random Forest at 75.27%, according to the models' Area Under the Curve (AUC) values, which measure their ability to discriminate between classes. The AUC ratings of the Decision Tree, SVM, and gradient boosting models decreased. Lower values are preferred for Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), and Random Forest performed exceptionally well in this regard, with the lowest MSE (0.582) and RMSE (0.763). The MSE and RMSE values of the Gradient Boosting, SVM, XGBoost, and Decision Tree models showed an increasing order trend.

In summary, the Random Forest model always prevailed at this task when taking accuracy, precision, recall, F1-score, AUC, MSE, and RMSE. It also has strong performance in all metrics for XGBoost, Gradient Boosting, SVM and Decision Tree models, showing competitive but relatively significantly lower overall performance.

Therefore, we can improve our model by adding more relevant features to the dataset to enhance the flood prediction. It may include temporal features like time of year; for instance, flooding patterns may vary based on this factor or seasonality and historical trends. For example, in the case of flood occurrence, precipitation levels are the primary feature to be considered alongside others, such as soil moisture and river water levels, which depend heavily on climatic conditions because most disturbances happen owing to variations they can cause, significantly changing from.

Hence, it could remarkably enhance the model's predictive adequacy for flood prediction. It also enables the model to predict using more factors, improving accuracy and reliability. According to Hakim et al. (2023), through historical data analysis, predictive models can help authorities be alert about the probability of a flood, and people living near affected areas will have time to evacuate or take required measures.