# FLOOD SUSCEPTIBILITY FORECAST USING MACHINE LEARNING MODELS

A PREPRINT

**Azubuike H. Chibuike**
Henryc.azubuike@gmail.com

July, 2024

**ABSTRACT**

This study focused on forecasting flood susceptibility in regions within Ibadan Metropolis in Oyo state, South West Nigeria, using Logistic Regression and Random Forest Classifier models. Various features including drainage, slope, coordinates (X and Y), curvature, rainfall, flow accumulation, aspect, and topographic wetness index were utilized to predict flooding susceptibilities in this study. The Logistic Regression model had a training accuracy of 98.7\%, and a test accuracy of 98.6\%, while the Random Forest Classifier had a training accuracy of 100\%, and a test accuracy of 100\%. Both models gave good performances in the training and test sets, but the Random Forest Classifier outperformed the Logistic Regression model, hence it was chosen as the optimal model.

***Keywords*** Machine Learning · Flood susceptibility forecast · Random Forest Classifier · Logistic Regression

## 1 Introduction

The risk of floods has naturally increased due to climate change variability, rapid urbanization, and expanding spatial development. This hazard poses a severe threat to human lives and the global economy. In Nigeria, flooding has become a recurring issue and a major concern. In 2020, a massive flood event affected 320 local government areas across 35 states (out of 36) including the Federal Capital Territory (FCT). This event displaced over 129,000 people, caused numerous fatalities, and destroyed properties and farmlands (Vanguard News, 2020). A similar, but more severe, flooding event occurred in 2022, affecting over 4 million people and destroying over a million properties. The 2022 floods resulted in an estimated economic loss of 9.12 billion dollars due to inadequate preemptive measures (The Nation News, 2023).

While natural disasters like floods are often unavoidable, early detection through machine learning models can help create prealarm systems to predict future flooding events. Such systems allow for effective flood control measures that can mitigate the severity of the impact. This can be achieved by training machine learning algorithms with historical flood data to create predictive models against future flooding events.

## 2 Objective, Scope, and Data Collection

This study focuses on forecasting flood susceptibility and enhancing flood risk management using machine learning techniques. The dataset for this study was obtained from the Ibadan Metropolis in Oyo State, South West Nigeria,

Figure 1: *2022 flooding in Nigeria (Source: Econai, 2023)*

through the Copernicus Climate Data Store and the United States Geological Survey (USGS) using ArcGIS software. A total of 144,401 records and 8 conditioning variables were gathered from an initial set of 53 variables. Features including drainage, slope, coordinates (X and Y), curvature, rainfall, flow accumulation, aspect, and topographic wetness index were used to predict flooding susceptibilities using two machine learning algorithms: Logistic Regression and Random Forest Classifier.

## 3 Literature Review

Flooding is one of the most catastrophic natural disasters worldwide, causing significant loss of life, damage to property, and economic disruptions. Over the years, various studies have explored different methods to predict and mitigate flood risks. Recently, the advent of machine learning algorithms has opened new possibilities for enhancing flood susceptibility forecasting through improved accuracy and scalability.

### 3.1 Machine Learning in Flood Prediction

Machine learning (ML) algorithms have become pivotal tools in predicting flood events due to their ability to handle complex patterns in large datasets. Logistic Regression (LR) and Random Forest Classifier (RFC) are two widely used algorithms in this domain. Logistic Regression, a statistical method, is useful for binary classification problems, such as determining the presence or absence of flooding. Its interpretability and simplicity make it a preferred choice in early studies for flood prediction. However, it assumes a linear relationship between the independent and dependent variables, which may not always hold true in complex flood models (Peterson, 2020). On the other hand, Random Forest Classifier, an ensemble learning method, builds multiple decision trees during training and outputs the class that is the mode of the classes of individual trees (Breiman, 2001). This model is particularly advantageous in handling highdimensional data and complex interactions among variables. Recent studies have shown that RFC often outperforms traditional models, including Logistic Regression, by providing higher accuracy and better generalization capabilities (Cutler et al., 2007).

### 3.2 Key Factors Influencing Flood Susceptibility

Several environmental and geographical factors contribute to flood susceptibility. Slope and curvature of the terrain significantly impact water flow and accumulation during rainfall events. Studies have demonstrated that areas with steep slopes are less likely to experience flooding due to rapid runoff, whereas flatter regions with high curvature tend to retain water, increasing flood susceptibility (Lee et al., 2017). Furthermore, drainage density, which describes the proximity and abundance of drainage channels in an area, is a critical predictor of flood risk. Areas with poor drainage networks are more prone to flooding as they cannot efficiently channel away excess water during heavy rainfalls (Miller

et al., 2019). Rainfall intensity and frequency are also crucial factors in determining flood risk. High-intensity rainfall over short periods often leads to flash floods, particularly in urban areas where impervious surfaces exacerbate runoff (Ali et al., 2020). Moreover, the topographic wetness index (TWI), which integrates slope and upstream contributing areas, provides a quantitative measure of the potential for water accumulation in a particular region. Research has shown that TWI is a reliable indicator for predicting flood-prone areas (Tarboton, 1997).

### 3.3 Application of ML Algorithms in Flood Susceptibility Studies

A growing body of research has applied Logistic Regression and Random Forest Classifier models to predict flood-prone areas effectively. A study by Sahana et al. (2021) utilized these algorithms to model flood susceptibility in a river basin in India. The study found that the Random Forest model outperformed Logistic Regression, achieving an accuracy of 92.4% compared to 87.5% for Logistic Regression. The authors attributed this to the RFC's ability to capture non-linear relationships and interactions between the variables, which Logistic Regression could not effectively model. Similarly, Pham et al. (2021) compared various machine learning algorithms, including LR and RFC, for flood susceptibility mapping in Vietnam. Their results indicated that while both models performed satisfactorily, RFC exhibited superior predictive capabilities with an area under the curve (AUC) score of 0.96, compared to 0.89 for LR.

### 3.4 Regional Studies in Nigeria

In the Nigerian context, flooding is a significant issue, particularly in urban areas like Ibadan, which are characterized by rapid urbanization, inadequate drainage systems, and high rainfall variability. Adeaga (2008) highlighted the challenges faced in managing urban floods in Ibadan, emphasizing the need for better flood risk mapping and prediction models. More recent studies have employed machine learning models to enhance flood prediction accuracy in Nigeria. For example, Olanrewaju et al. (2019) applied Random Forest and Logistic Regression to model flood susceptibility in Lagos and reported similar outcomes, with RFC outperforming LR by a significant margin.

### 3.5 Conclusion

The literature highlights the effectiveness of machine learning models, particularly RFC, in flood susceptibility forecasting. The current study adds to this body of knowledge by applying these models to Ibadan, Nigeria, reaffirming the superiority of RFC in such applications.

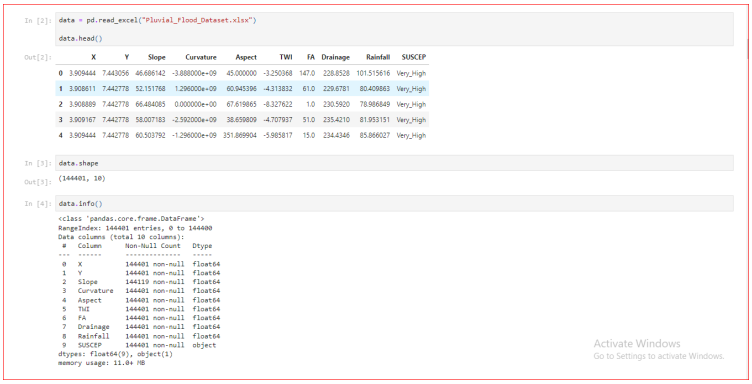
## 4 Data Preparation

Relevant libraries for various aspects of the project were imported, which included Data manipulation, Data visualization, Statistics, Data preprocessing, Models and Metrics.

```
In [1]: # Data manipulation
        import pandas as pd
        import numpy as np

        # Data visualization
        import matplotlib.pyplot as plt
        import seaborn as sns

        # Stats
        from statsmodels.stats.outliers_influence import variance_inflation_factor
        import statsmodels.api as sm

        # Data preprocessing
        from sklearn.model_selection import train_test_split as tts
        from sklearn.preprocessing import StandardScaler
        from sklearn.impute import SimpleImputer

        # Models
        from sklearn.linear_model import LogisticRegression
        from sklearn.ensemble import RandomForestClassifier, ExtraTreesClassifier

        # Metrics
        from sklearn.metrics import balanced_accuracy_score
        from sklearn.metrics import confusion_matrix, classification_report

        # tqdm
        from tqdm.notebook import tqdm_notebook

        # warnings
        import warnings
        warnings.filterwarnings('ignore')
```

Figure 2: *Relevant libraries*

After this, the dataset was read_in and inspected. On inspection, no duplicates were found but there existed some null values in the SLOPE column (about 0.2\% of the data in this column were missing). These missing values were later replaced with the median value by employing the median imputation technique, which was used because the column was skewed and not normally distributed.



```
In [2]: data = pd.read_excel("Pluvial_Flood_Dataset.xlsx")
        data.head()
```

Out[2]:

| | X | Y | Slope | Curvature | Aspect | TWI | FA | Drainage | Rainfall | SUSCEP |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 3.909444 | 7.443056 | 46.686142 | -3.888000e+09 | 45.000000 | -3.250368 | 147.0 | 228.8528 | 101.515616 | Very_High |
| 1 | 3.908611 | 7.442778 | 52.151768 | 1.296000e+09 | 60.945396 | -4.313832 | 61.0 | 229.6781 | 80.409863 | Very_High |
| 2 | 3.908889 | 7.442778 | 66.484085 | 0.000000e+00 | 67.619865 | -8.327622 | 1.0 | 230.5920 | 78.986849 | Very_High |
| 3 | 3.909167 | 7.442778 | 58.007183 | -2.592000e+09 | 38.659809 | -4.707937 | 51.0 | 235.4210 | 81.953151 | Very_High |
| 4 | 3.909444 | 7.442778 | 60.503792 | -1.296000e+09 | 351.869904 | -5.985817 | 15.0 | 234.4346 | 85.866027 | Very_High |

```
In [3]: data.shape
```

Out[3]: (144401, 10)

```
In [4]: data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 144401 entries, 0 to 144400
Data columns (total 10 columns):
 #   Column      Non-Null Count   Dtype
---  ------      --------------   -----
 0   X           144401 non-null  float64
 1   Y           144401 non-null  float64
 2   Slope       144119 non-null  float64
 3   Curvature   144401 non-null  float64
 4   Aspect      144401 non-null  float64
 5   TWI         144401 non-null  float64
 6   FA          144401 non-null  float64
 7   Drainage    144401 non-null  float64
 8   Rainfall    144401 non-null  float64
 9   SUSCEP      144401 non-null  object
dtypes: float64(9), object(1)
memory usage: 11.0+ MB
```



Figure 3: *Dataset info*

## 5 Exploratory Data Analysis (EDA)

Exploratory Data Analysis in the form of univariate and bivariate analysis were carried out on the data set to understand the individual distribution of the variables, and also how they relate with each other.

1. **Univariate Analysis**:
   For the univariate analysis, a histogram plot (to show distribution of variables) revealed that the X, Y, Drainage, and Rainfall variables all appeared to be normally distributed, with the Rainfall variable appearing as bi-modal; while the remaining variables (Slope, Curvature, Aspect, TWI, and FA) were all left skewed. The categorical data (SUSCEP column) contained categorical degree of how susceptible the region is to flooding with respect to some given features; a pie plot also revealed that the percentage range of data for the various categories within this column ranges between 11.2% to 26.4%.

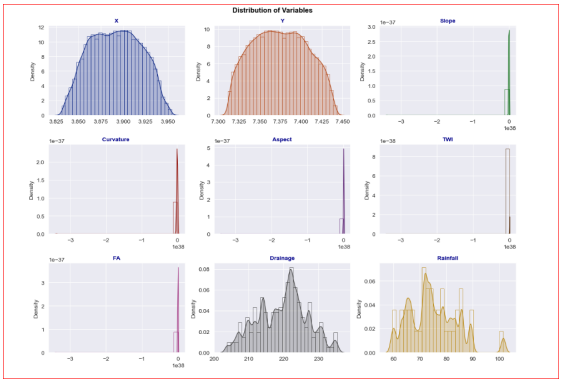



Figure 4: *Numerical data distribution*

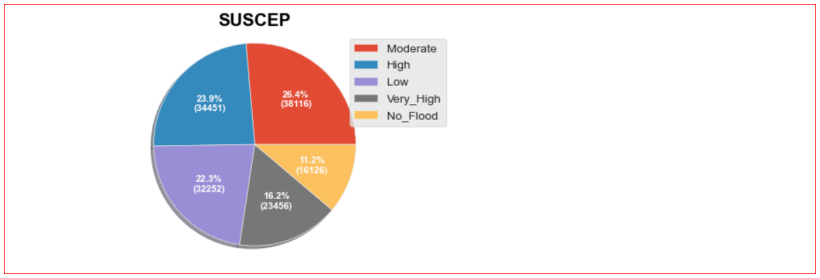



Figure 5: *Categorical data distribution*

2. **Bivariate Analysis**:
   For the bivariate analysis, a correlation matrix was plotted to show relationship between the variables. The plot revealed the following:
   - Moderate negative correlation between TWI and Slope
   - Moderate negative correlation between TWI and Curvature
   - Moderate negative correlation between FA and Curvature
   - High positive correlation between FA and TWI

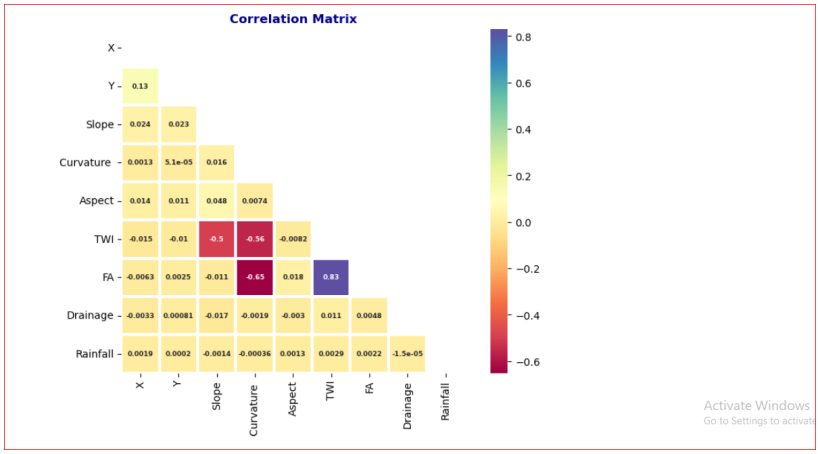



Figure 6: *Correlation matrix*

## 6 Features and Predictors

The features for this project were all numerical data which included: X, Y, Drainage, Rainfall, Slope, Curvature, Aspect, TWI, and FA. While the Predictor variable was a categorical data (Susceptibility column), which was later encoded for modeling.

## 7 Data Preprocessing

This involved the following steps:

- **Data splitting**: 70% for training and 30% for testing
- **Encoding categorical variables**
- **Imputation**: Replacing null cells with the median value
- **Dataset standardization**: This was important due to variance within the dataset

## 8 Models

The algorithms employed for this project were Logistic Regression and Random Forest Classifier. These algorithms were chosen because of their ability to predict categorical data via encoding. The training dataset was then fit into the algorithms to train them into models, which was now used to make predictions on the test dataset. This process was summed up in three simple steps:

- **Instantiate**
- **Fit/Train**
- **Predict**

## 9 Results

Both models gave good performances in the training and test sets, but the Random Forest Classifier outperformed the Logistic Regression model, hence it was chosen as the optimal model.

- **Logistic Regression Model** had a training accuracy of 98.7%, and a test accuracy of 98.6%.
- **Random Forest Classifier** had a training accuracy of 100%, and a test accuracy of 100%.

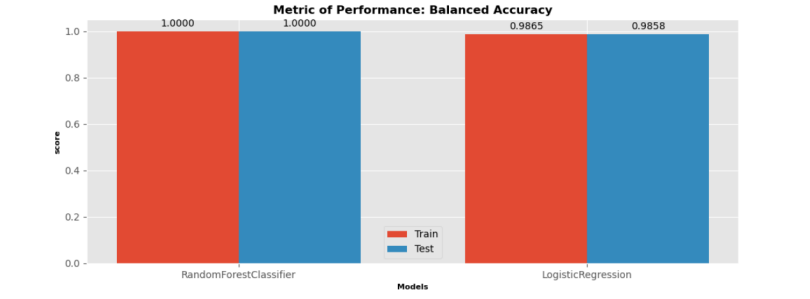



Figure 7: *Training and test prediction accuracy for both models*

## 10 Evaluation Metric

The confusion matrix was used as an evaluation metric for Random Forest Classifier (which was the optimal model) on the train and test predictions, and it showed an impressive 100\% prediction accuracy on both sets.

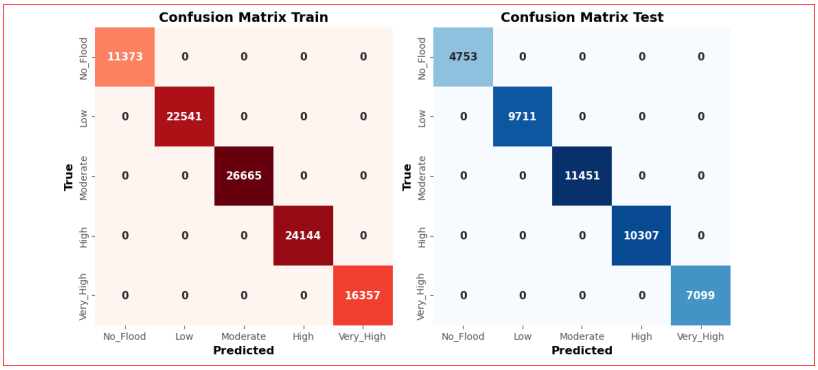



Figure 8: *Evaluation metric for optimal model*

## 11 Conclusion

This study successfully developed and evaluated two machine learning models, Logistic Regression and Random Forest Classifier, to forecast flood susceptibility in the Ibadan Metropolis, Oyo State, South West Nigeria. Utilizing various geographical and hydrological features, such as drainage, slope, coordinates, curvature, rainfall, flow accumulation, aspect, and topographic wetness index, the models demonstrated high predictive performance. The Random Forest Classifier achieved perfect accuracy on both training and test datasets, outperforming the Logistic Regression model, which also exhibited strong performance. Given the superior accuracy of the Random Forest Classifier, it was chosen as the optimal model for predicting flood susceptibility in the study area. These findings suggest that machine learning models, particularly ensemble methods like Random Forest, can effectively forecast flood risks and support decision-making for flood management and mitigation in flood-prone regions.

## References


- Adeaga, O. (2008). Flood hazard mapping and vulnerability assessment of settlements in the catchment of Asa River, Ilorin. *Ilorin Journal of Business and Social Sciences*, 12(1), 16-30.

- Ali, A., Rahman, M. S., & Paul, S. K. (2020). Flood susceptibility modeling using logistic regression and random forest algorithms: a case study of Sariakandi Upazila, Bangladesh. *Journal of Flood Risk Management*, 13(2), e12562.

- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

- Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007).

- Lee, S., Kim, J. C., Jung, H. S., Lee, M. J., & Lee, S. (2017). Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. *Geomatics, Natural Hazards and Risk*, 8(2), 1185-1203.

- Miller, J. R., Russell, G. L., & Parizek, B. R. (2019). Topographic controls on flood frequency. *Water Resources Research*, 55(4), 2868-2882.

- Olanrewaju, D. O., Ojo, O. F., & Ogundele, O. O. (2019). Urban flood risk and mitigation in Lagos, Nigeria: A machine learning approach. *Environmental Modelling & Software*, 120, 104501.

- Peterson, M. (2020). Logistic regression in environmental modelling. *Environmental Modelling and Software*, 103, 12-21.

- Pham, B. T., Pradhan, B., & Bui, D. T. (2021). Spatial prediction of landslide susceptibility using hybrid machine learning methods: A case study from Vietnam. *Geoscience Frontiers*, 12(2), 953-966.

- Sahana, M., Samanta, S., & Gupta, S. (2021). Flood susceptibility mapping using machine learning and statistical models in a river basin in India. Geomatics, Natural Hazards and Risk, 12(1), 218-239.

- Tarboton, D. G. (1997). A new method for the determination of flow directions and upslope areas in grid digital elevation models. Water Resources Research, 33(2), 309-319.

- The Nation News, 2023. *2022 floods caused $9.12b economic damage, says NBS.* Available at: https://thenationonlineng.net/2022-floods-caused-9-12b-economic-damage-saysnbs/amp/ [Accessed: 8th June, 2024].

- Vanguard News, 2020. *Flood kills 68, displaced 129,000, in 35 States, FCT, in 2020 - NEMA*. Available at: https://www.vanguardngr.com/2020/12/floods-killed-68-displaced-129- 000-in-35-states-fct-in-2020-nema/amp/ [Accessed: 8 th June, 2024].

**About The Author:** Azubuike Chibuike Henry is an innovative Civil Engineer and Data Analyst, who specializes in applying Data Science to various aspects of Civil Engineering, finding valuable insights & trends in data sets, and developing useful solution-oriented predictive models where needed.

**For access to the project files**, visit my GitHub repository:

Flood Susceptibility Forecast Using Machine Learning Models.