FLOOD SUSCEPTIBILITY FORECAST USING MACHINE LEARNING ALGORITHMS

by Azubuike, Chibuike Henry
June, 2024

INTRODUCTION

The risk of floods has naturally increased due to climate change variability, rapid urbanization, and rapidly expanding spatial development. This hazard to human lives and the global economy has become quite severe. Over the years, flooding has become a recurring issue and a major concern in Nigeria. In 2020, there was a massive flooding event in the country, affecting 320 local government areas in 35 states (out of 36) including the FCT, this event also displaced over 129,000 persons, killed many others and destroyed many properties and farmlands (Vanguard News, 2020). In 2022, a similar flooding event took place, but this time it was worse and lasted for months, affecting over 4 million persons across the country and destroying over a million properties. The impact was so much more than the previous year because little or no measures were put in place to contain this (as is usually the case in some under developed/developing countries). The 2022 floods in various parts of Nigeria led to an estimated economic loss of $9.12 billion (The Nation News, 2023).

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

A major problem with these natural disasters is that in most cases they are unavoidable, but early detection or identification of these flooding events through Machine Learning

models can be used to create pre-alarming systems/signals on flooding events that could occur in future, and hence we can put up effective flood control measures that will help mitigate the severity of what would have been. This can be done by training machine learning algorithms with previous flooding data to create models that could be used to predict future flooding events.

OBJECTIVES, SCOPE AND DATA COLLECTION

The study focused on forecasting flood susceptibility, as well as enhancing the management of potential pluvial flooding risk through the application of machine learning techniques. The dataset for this study was gathered from the Ibadan Metropolis in Oyo state, South West Nigeria, by the Copernicus Climate Data Store and the United States Geological Survey (USGS) using the ArcGIS software. A total of 144,401 records and 8 conditioning variables out of 53 were gathered. 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, using two Machine Learning algorithms: LogisticRegression and RandomForest.

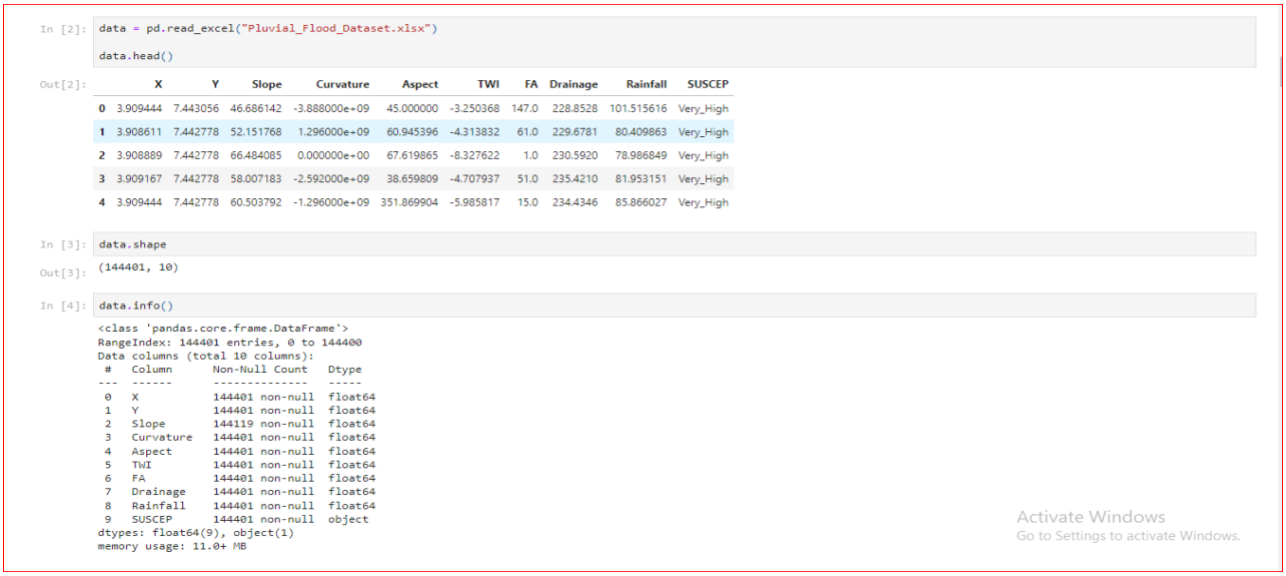
DATA PREPARATION

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, I loaded and inspected the data set; 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.



```
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

EXPLORATORY DATA ANALYSIS (EDA)

EDA 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. 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%.

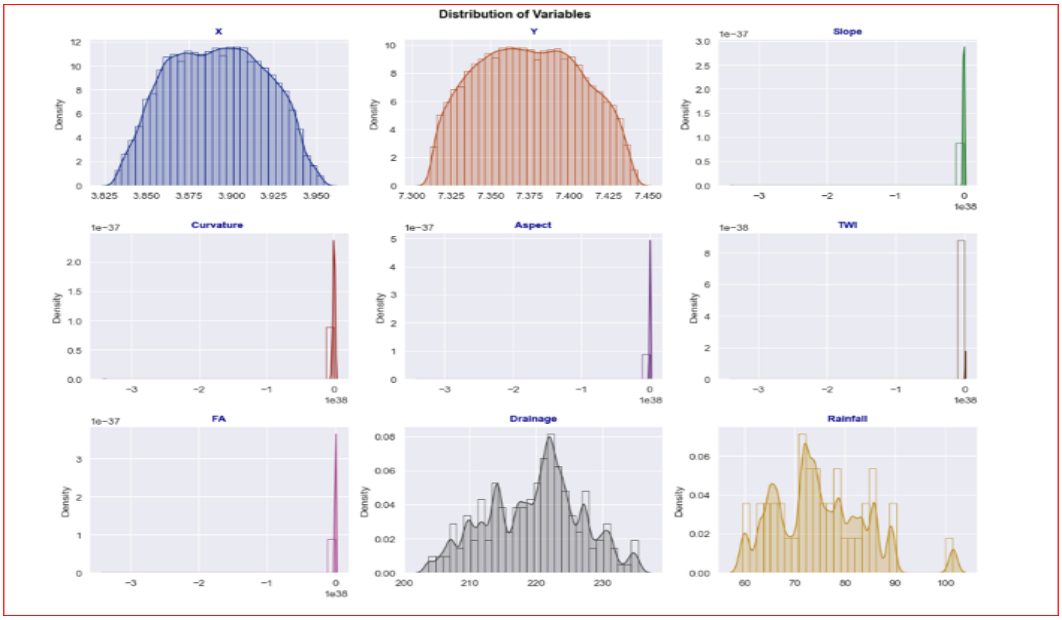


Figure 4: Numerical data distribution

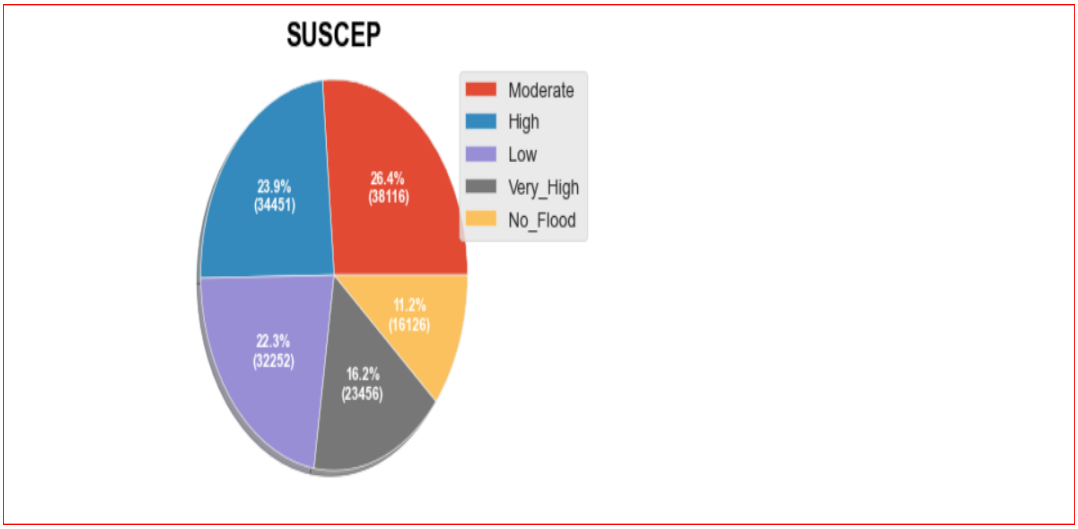


Figure 5: Categorical data distribution

2. <u>For the bivariate analysis</u>, a correlation matrix was plotted to show relationship between the variables. The plot revealed the following:
   - A moderate negative correlation between TWI and Slope
   - A moderate negative correlation between TWI and Curvature
   - A moderate negative correlation between FA and Curvature
   - A high positive correlation between FA and TWI

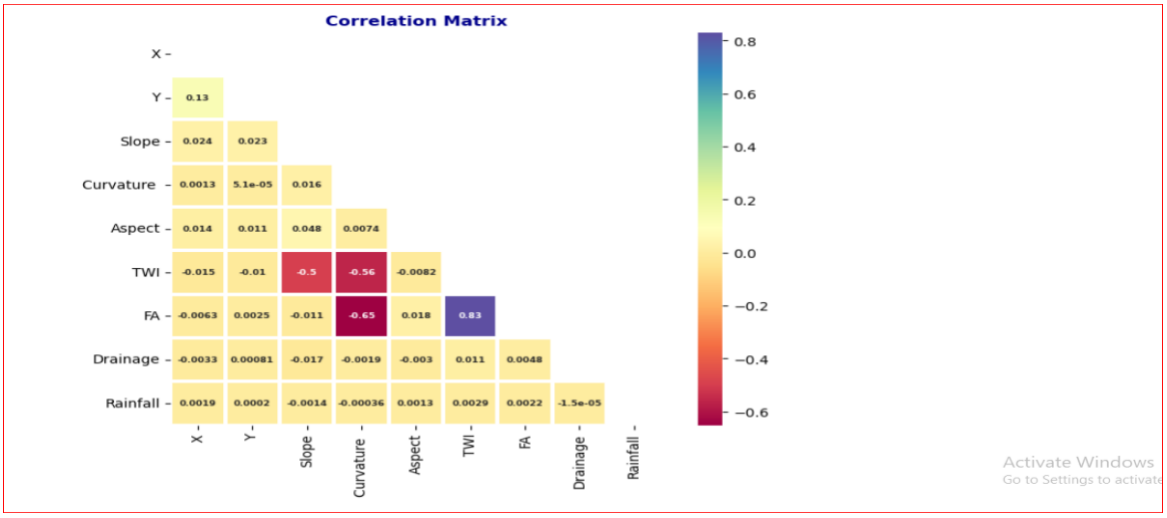


Figure 6: Correlation matrix

FEATURES AND PREDICTOR

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.

PREPROCESSING

This involved the following steps:

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

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

RESULTS

- The Logistic Regression model had a training accuracy of 98.7%, and a test accuracy of 98.6%.
- 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.

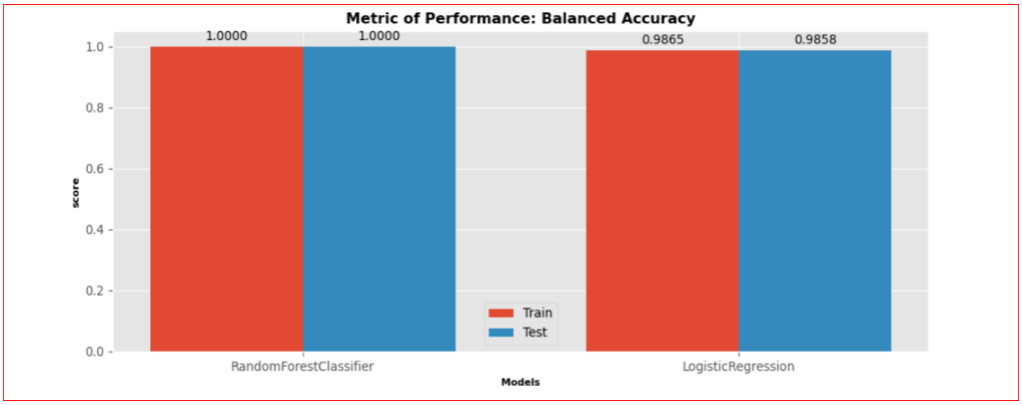


Figure 7: Training and test prediction accuracy for both models

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.

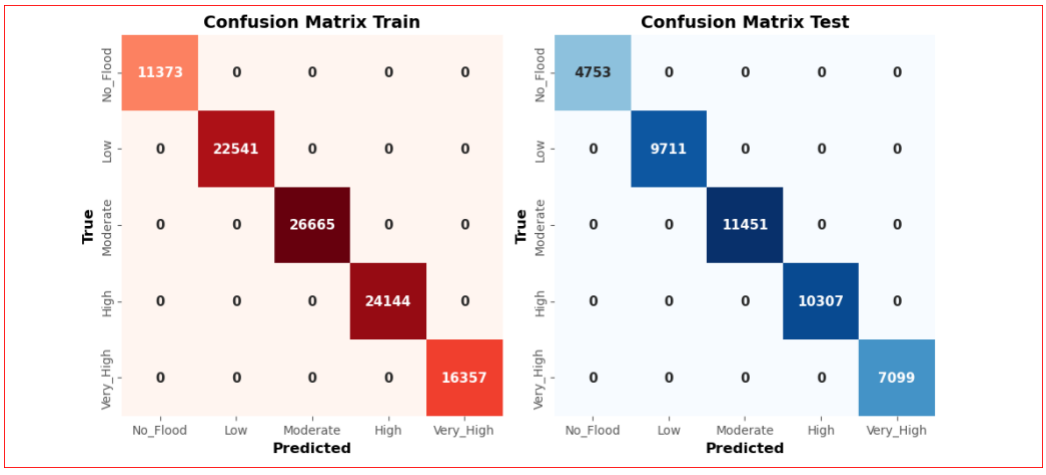


Figure 8: Evaluation metric for optimal model

CONCLUSION

This study focused on forecasting flood susceptibility in regions within Ibadan Metropolis in Oyo state, South West Nigeria, using LogisticRegression and RandomForestClassifier 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.

REFERENCES

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-says-nbs/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: 8th 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. Email: Henryc.azubuike@gmail.com.