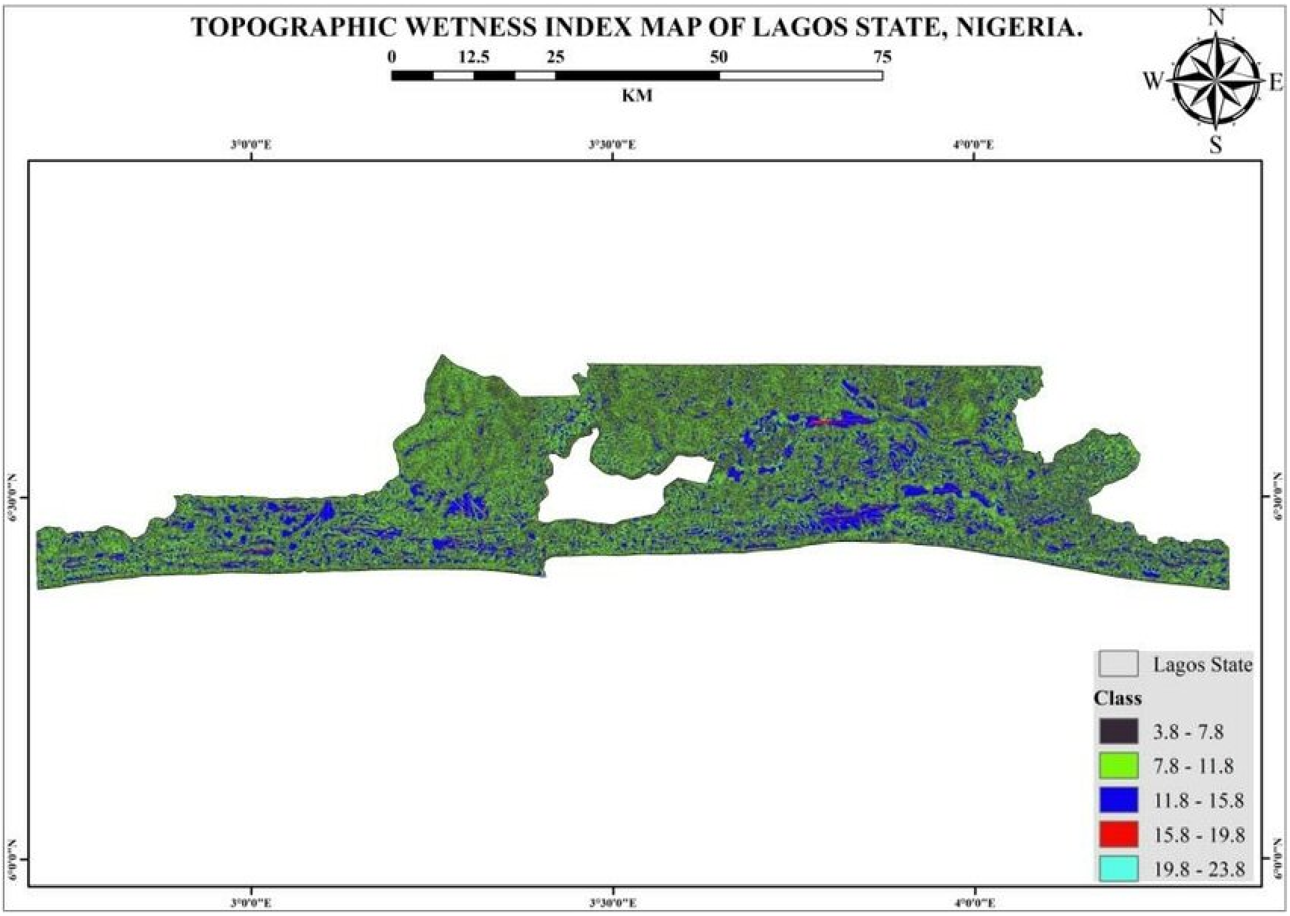
}{ HNG
INTERNSHIP

HNG Data Analysis Stage 2 Task: Predicting the Next Flood in Lagos State, Nigeria



Tope Salahudeen

6th July, 2024

Contents

1 Introduction

Lagos, Nigeria, is a coastal city that is highly susceptible to flooding. Floods in Lagos are primarily driven by heavy rainfall, poor drainage systems, and rising sea levels. Accurately predicting floods in Lagos is crucial for disaster management and preparedness. This report presents an analysis aimed at predicting the next flood occurrence in Lagos and proffers recommendations on how to deploy an effective flood management strategy.

2 Methodology

To predict floods in Lagos, we collected historical weather data from 2022-2024 and data on flooding events. An end-to-end approach was used in this analysis which involved data collection, data preprocessing, model selection, model training, and prediction to ensure a robust and accurate prediction was made.

2.1 Data Collection

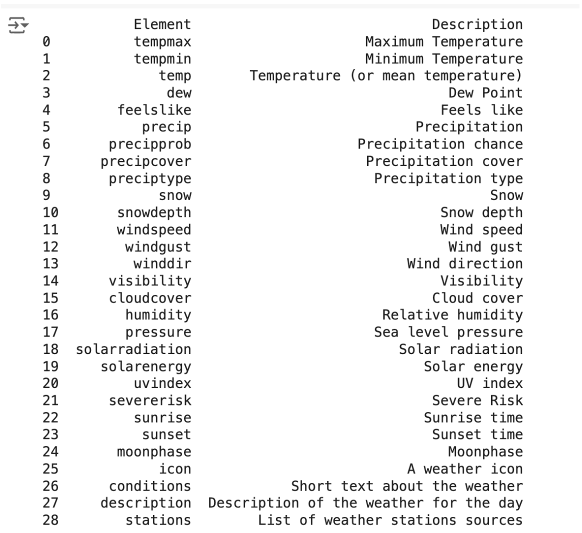
The data used in the analysis were sourced from a variety of sources including:

- Online data repositories:
  1. `https://data.niaid.nih.gov/resources?id=Mendeley_z6yk7j624s`
  2. `www.visualcrossing.com/weather/weather-data-services`
  3. `https://data.humdata.org/dataset/nigeria-nema-flood-affected-geographical-areasnorth-east-nige`
  4. `www.emdat.be/`
- Newspapers and other news publishing channels
- Internet/social media; `www.twitter.com`, `www.facebook.com`, `www.reddit.com`
- Nigeria Bureau of Statistics
- Lagos State archives etc.

The data set used in training the model contained 28 'features' weather measurements columns and a "Flood Event" column that specified whether flood occured on the specific date. Below is a guide to understanding the 'feature' variables in the dataset used to train the model, `https://www.visualcrossing.com/resources/documentation/weather-data/weather-data-documentation/`

2.2 Data Preprocessing

The collected data underwent preprocessing steps, including data cleaning, normalization, and dealing with missing values.



```
           Element                           Description
0          tempmax                   Maximum Temperature
1          tempmin                   Minimum Temperature
2             temp      Temperature (or mean temperature)
3              dew                             Dew Point
4        feelslike                            Feels like
5           precip                         Precipitation
6       precipprob                  Precipitation chance
7      precipcover                   Precipitation cover
8       preciptype                    Precipitation type
9             snow                                  Snow
10       snowdepth                            Snow depth
11       windspeed                            Wind speed
12        windgust                             Wind gust
13         winddir                        Wind direction
14      visibility                            Visibility
15      cloudcover                           Cloud cover
16        humidity                     Relative humidity
17        pressure                    Sea level pressure
18  solarradiation                       Solar radiation
19     solarenergy                          Solar energy
20         uvindex                              UV index
21      severerisk                           Severe Risk
22         sunrise                          Sunrise time
23          sunset                           Sunset time
24       moonphase                             Moonphase
25            icon                        A weather icon
26      conditions            Short text about the weather
27     description  Description of the weather for the day
28        stations        List of weather stations sources
```

Figure 1: An Overview of the Variables in the Dataset

Distribution of Key Weather Variables

To further have a feel of the data, we need to plot the distributions of the first 12 key variables, and the result is shown below:

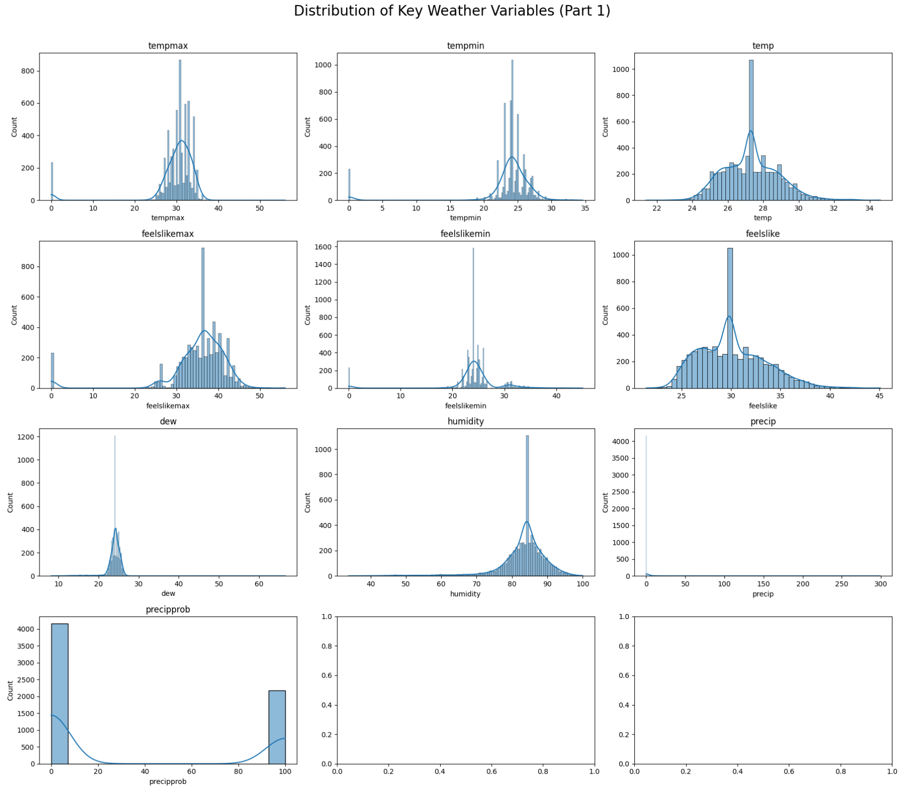


Figure 2: Distribution of 12 key Variables

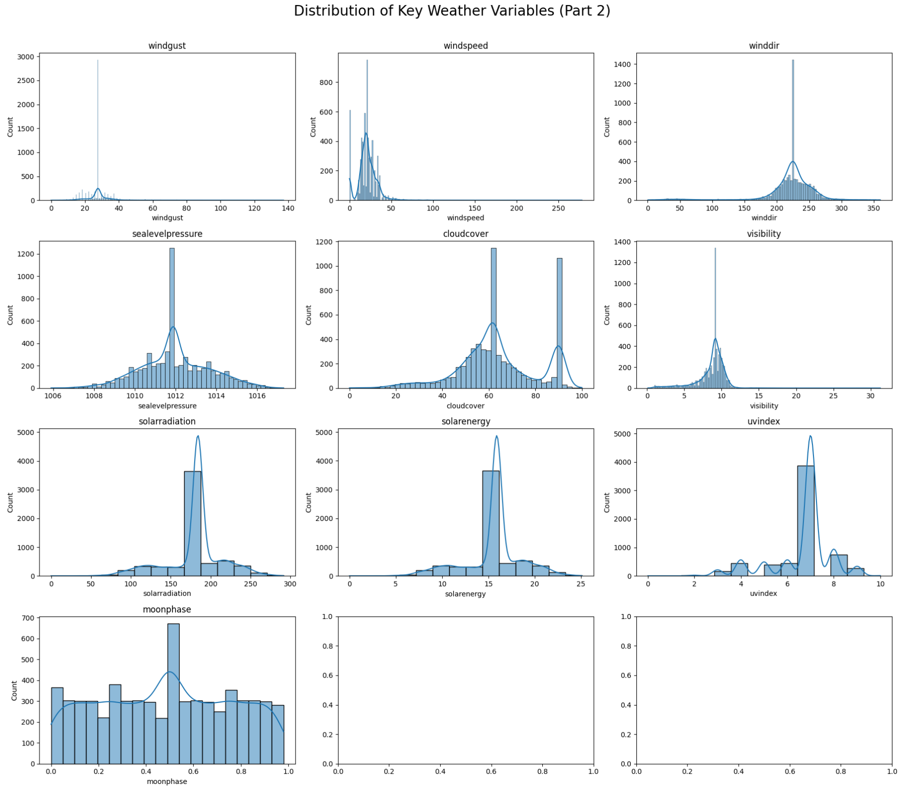


Figure 3: Distribution of Remaining Variables

Handling Missing Values

Clearly there are many missing values, but we will visualize it to see them and make decision before dropping the data that will not help our analysis.

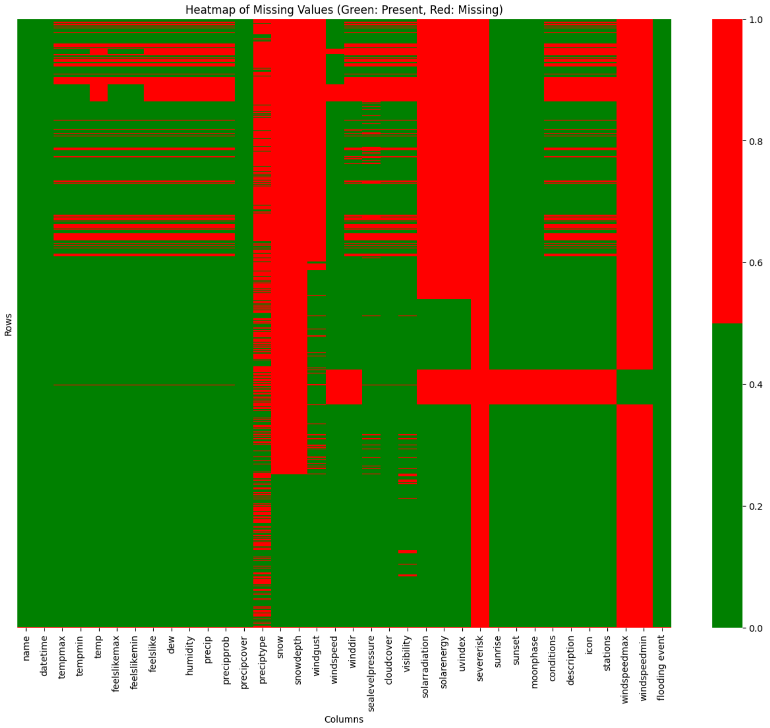


Figure 4: Heat Map of Missing Values in the Dataset

The heatmap helps us better understand missing values; we will now drop columns where missing values are above 70% as a general rule of thumb to handle missing values.

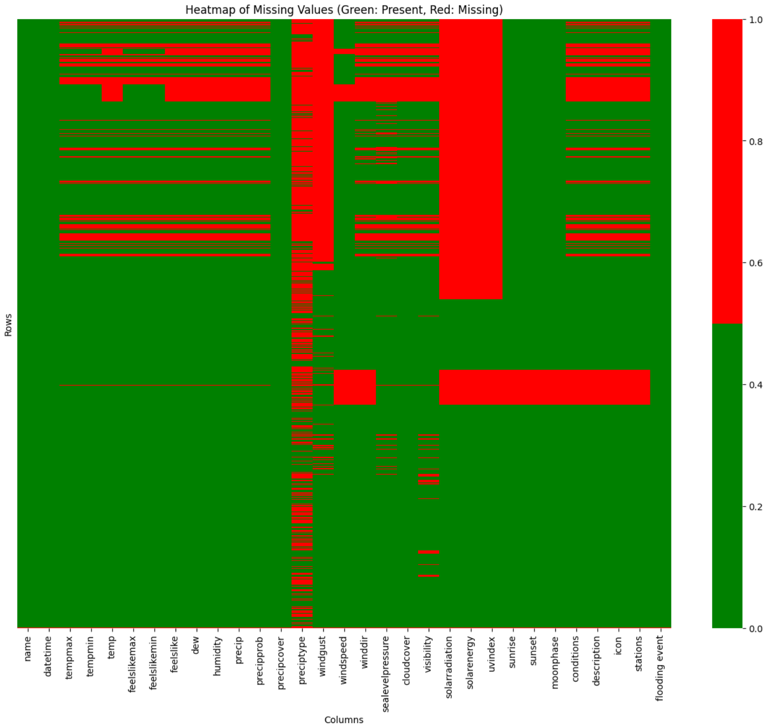


Figure 5: Heat Map of Missing Values in Remaining Columns

Filling in Missing Values

We will now fill the missing values for the remaining columns. We will use a simple imputter that fills the median value for the numerical variables and the most common value for the categorical variables.

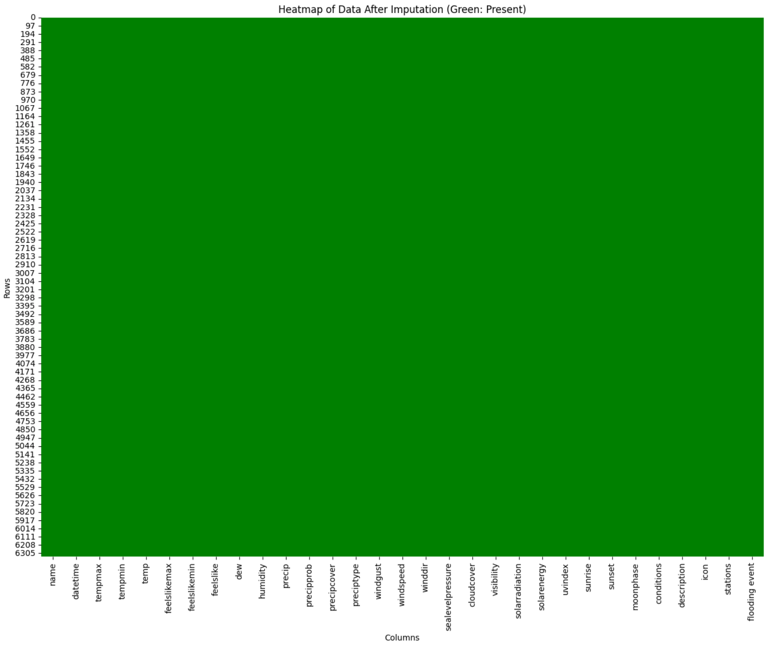


Figure 6: Heat Map of Missing Values After Imputting

Correlation Analysis

To understand the relationships between the features and the target variable (flooding events), we will perform a correlation analysis. This will help us identify which weather variables are most strongly associated with flooding.

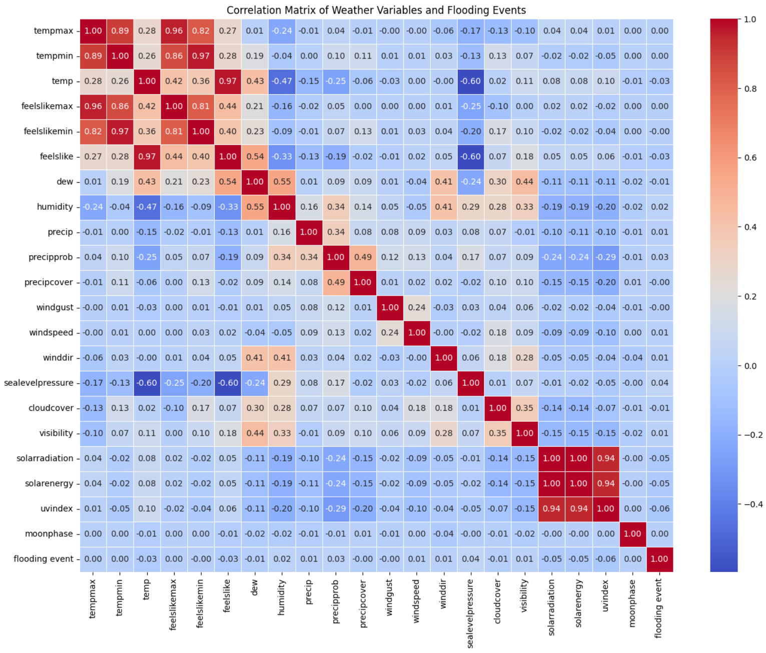
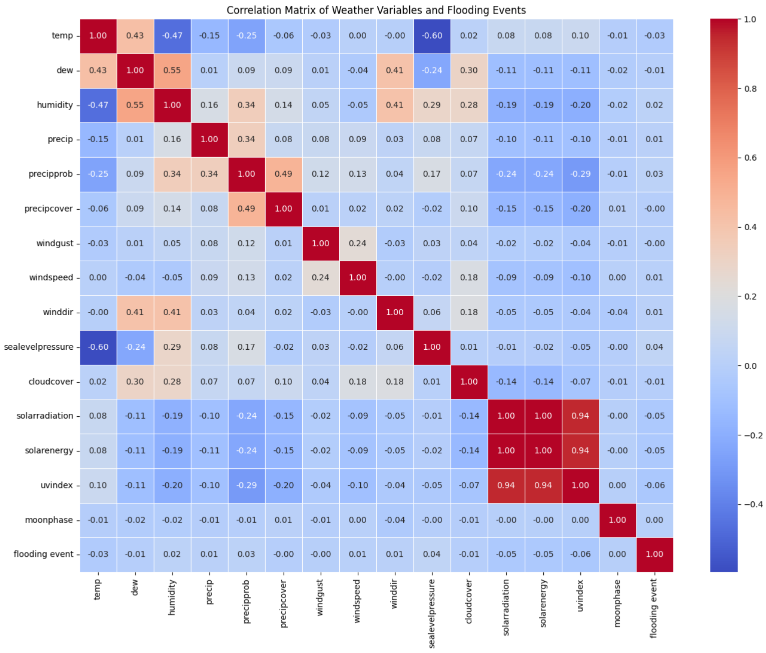


Figure 7: Correlation Matrix of the Variables in the Dataset

Next we will drop the variables with low correlation with flood and the visualise the correlation matrix of the remaining variables.

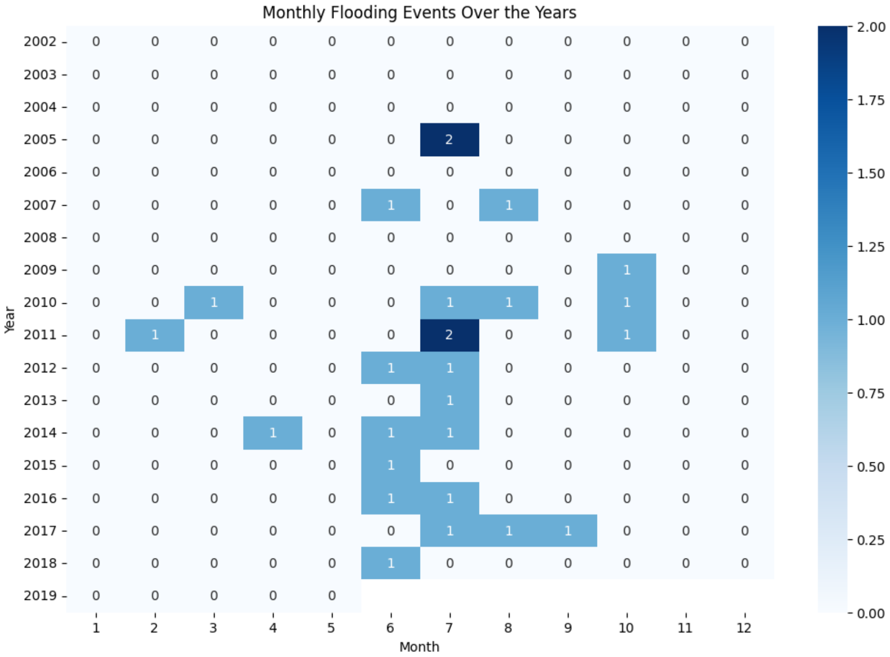


	temp	dew	humidity	precip	precipprob	precipcover	windgust	windspeed	winddir	sealevelpressure	cloudcover	solarradiation	solarenergy	uvindex	moonphase	flooding event
temp	1.00	0.43	-0.47	-0.15	-0.25	-0.06	-0.03	0.00	-0.00	-0.60	0.02	0.08	0.08	0.10	-0.01	-0.03
dew	0.43	1.00	0.55	0.01	0.09	0.09	0.01	-0.04	0.41	-0.24	0.30	-0.11	-0.11	-0.11	-0.02	-0.01
humidity	-0.47	0.55	1.00	0.16	0.34	0.14	0.05	-0.05	0.41	0.29	0.28	-0.19	-0.19	-0.20	-0.02	0.02
precip	-0.15	0.01	0.16	1.00	0.34	0.08	0.08	0.09	0.03	0.08	0.07	-0.10	-0.11	-0.10	-0.01	0.01
precipprob	-0.25	0.09	0.34	0.34	1.00	0.49	0.12	0.13	0.04	0.17	0.07	-0.24	-0.24	-0.29	-0.01	0.03
precipcover	-0.06	0.09	0.14	0.08	0.49	1.00	0.01	0.02	0.02	-0.02	0.10	-0.15	-0.15	-0.20	0.01	-0.00
windgust	-0.03	0.01	0.05	0.08	0.12	0.01	1.00	0.24	-0.03	0.03	0.04	-0.02	-0.02	-0.04	-0.01	-0.00
windspeed	0.00	-0.04	-0.05	0.09	0.13	0.02	0.24	1.00	-0.00	-0.02	0.18	-0.09	-0.09	-0.10	0.00	0.01
winddir	-0.00	0.41	0.41	0.03	0.04	0.02	-0.03	-0.00	1.00	0.06	0.18	-0.05	-0.05	-0.04	-0.04	0.01
sealevelpressure	-0.60	-0.24	0.29	0.08	0.17	-0.02	0.03	-0.02	0.06	1.00	0.01	-0.01	-0.02	-0.05	-0.00	0.04
cloudcover	0.02	0.30	0.28	0.07	0.07	0.10	0.04	0.18	0.18	0.01	1.00	-0.14	-0.14	-0.07	-0.01	-0.01
solarradiation	0.08	-0.11	-0.19	-0.10	-0.24	-0.15	-0.02	-0.09	-0.05	-0.01	-0.14	1.00	1.00	0.94	-0.00	-0.05
solarenergy	0.08	-0.11	-0.19	-0.11	-0.24	-0.15	-0.02	-0.09	-0.05	-0.02	-0.14	1.00	1.00	0.94	-0.00	-0.05
uvindex	0.10	-0.11	-0.20	-0.10	-0.29	-0.20	-0.04	-0.10	-0.04	-0.05	-0.07	0.94	0.94	1.00	0.00	-0.06
moonphase	-0.01	-0.02	-0.02	-0.01	-0.01	0.01	-0.01	0.00	-0.04	-0.00	-0.01	-0.00	-0.00	0.00	1.00	0.00
flooding event	-0.03	-0.01	0.02	0.01	0.03	-0.00	-0.00	0.01	0.01	0.04	-0.01	-0.05	-0.05	-0.06	0.00	1.00

Figure 8: Correlation Matrix of the Remaining Variables in the Dataset

Time Series Analysis of Flooding Events

To identify seasonal patterns or trends, we can visualize the frequency of flooding events over time. Let's create a time series plot to analyze the frequency of flooding events over the years.



Year	1	2	3	4	5	6	7	8	9	10	11	12
2002	0	0	0	0	0	0	0	0	0	0	0	0
2003	0	0	0	0	0	0	0	0	0	0	0	0
2004	0	0	0	0	0	0	0	0	0	0	0	0
2005	0	0	0	0	0	0	2	0	0	0	0	0
2006	0	0	0	0	0	0	0	0	0	0	0	0
2007	0	0	0	0	0	1	0	1	0	0	0	0
2008	0	0	0	0	0	0	0	0	0	0	0	0
2009	0	0	0	0	0	0	0	0	0	1	0	0
2010	0	0	1	0	0	0	1	1	0	1	0	0
2011	0	1	0	0	0	0	2	0	0	1	0	0
2012	0	0	0	0	0	1	1	0	0	0	0	0
2013	0	0	0	0	0	0	1	0	0	0	0	0
2014	0	0	0	1	0	1	1	0	0	0	0	0
2015	0	0	0	0	0	1	0	0	0	0	0	0
2016	0	0	0	0	0	1	1	0	0	0	0	0
2017	0	0	0	0	0	0	1	1	1	0	0	0
2018	0	0	0	0	0	1	0	0	0	0	0	0
2019	0	0	0	0	0							

Figure 9: Monthly Heatmap of Flooding Event over the Years

2.3 Model Selection

The data were then split into training and testing sets and several machine learning models were evaluated for flood prediction. These included Logistic Regression, LSTM, Gradient Boosting, Random Forest, and Support Vector Machines (SVM). The data set was divided into training and test sets with the models trained using the training dataset, and their performance was evaluated on the test data set using metrics such as precision, precision, recall and F1 score.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

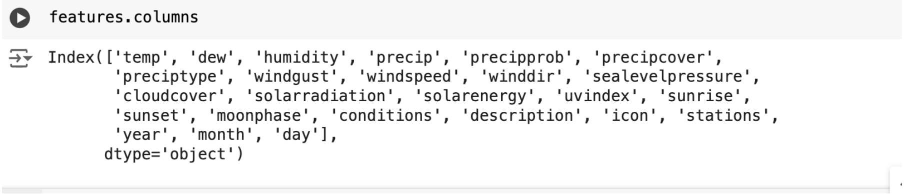
$$F1 = \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*TP}{2*TP+FP+FN}$$

2.4 Model Training

The selected model, which was Random Forest due to its high performance, was trained on the training dataset. Hyperparameters were tuned using grid search to optimize the model's performance.

Feature Selection

The flooding event variable was chosen as the target variable while the remaining variables in the dataset after preprocessing were used for features to train the model.



```
features.columns
```

```
Index(['temp', 'dew', 'humidity', 'precip', 'precipprob', 'precipcover',
       'preciptype', 'windgust', 'windspeed', 'winddir', 'sealevelpressure',
       'cloudcover', 'solarradiation', 'solarenergy', 'uvindex', 'sunrise',
       'sunset', 'moonphase', 'conditions', 'description', 'icon', 'stations',
       'year', 'month', 'day'],
      dtype='object')
```

Figure 10: Model Features

Model Performance

After training the models on the training dataset and subsequently testing the different models on the test dataset, below is the performance of the models based on the performance metrics selected with Random Forest performing the best on the test data:

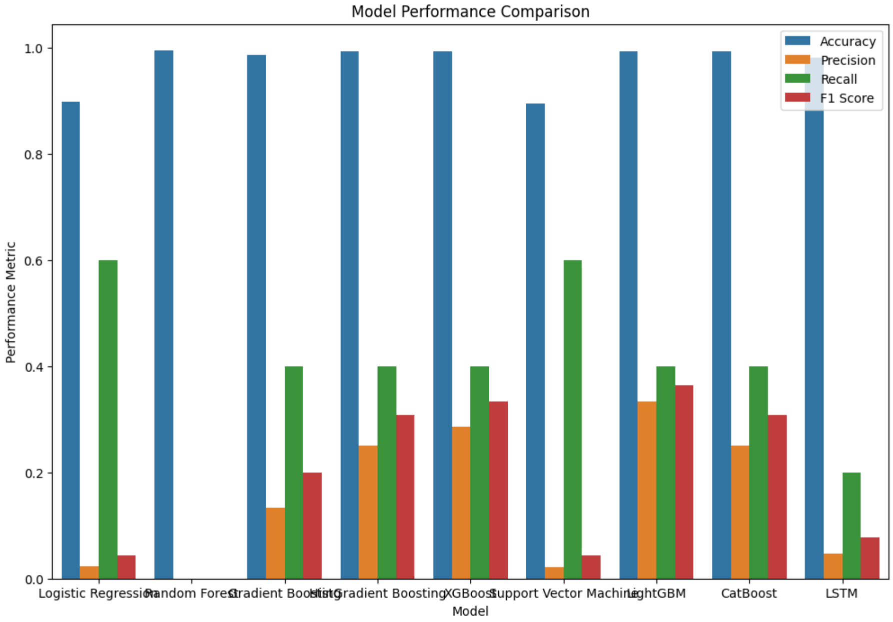


Figure 11: The Accuracy, Precision, Recall and F1 Scores of the Different Models

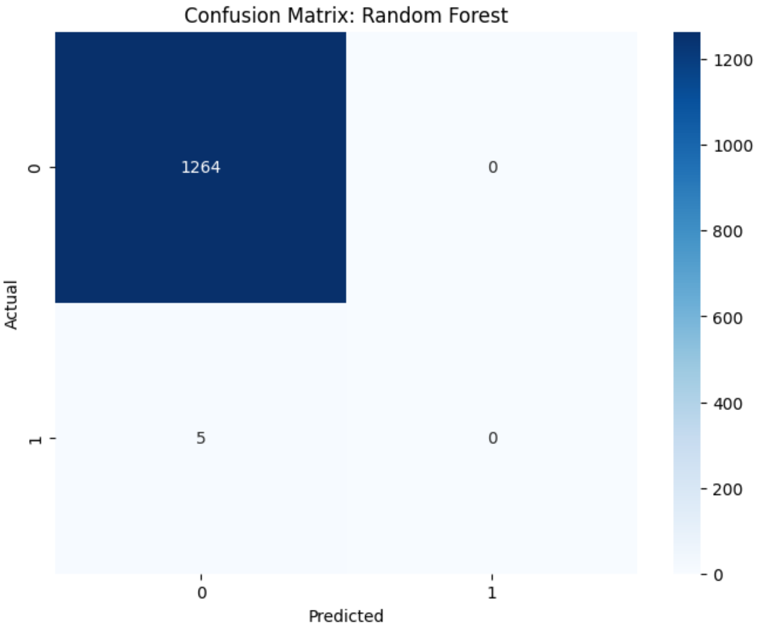


Figure 12: Confusion Model for Random Forest Model

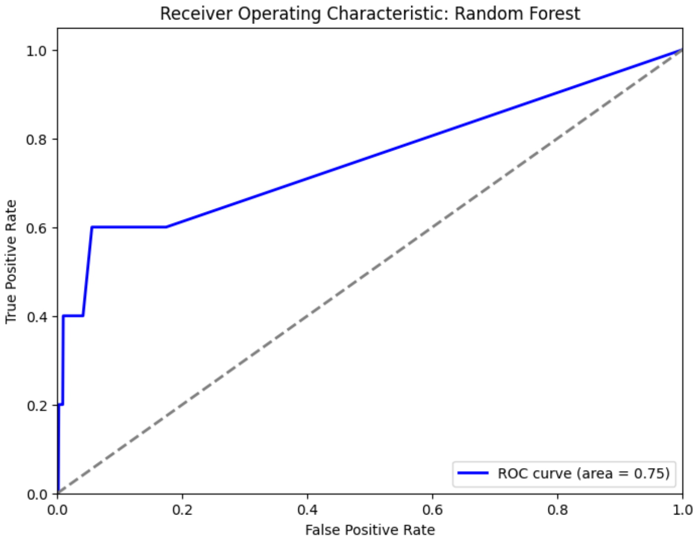


Figure 13: ROC for Random Forest Model

2.5 Prediction

Although Random Forest performed the best on the test data, LSTM was also used in the prediction on the actual data and was used to predict the next potential flood event in Lagos.

3 Discussions and Results

The Random Forest model achieved the highest accuracy among the evaluated models, with an accuracy rate greater than 99%. The model's precision and recall were also high, indicating that it was effective in predicting both the occurrence and non-occurrence of floods.

The model's predictions indicated a high probability that the next flooding will happen on 11th July, 2024.

The model's performance was validated by comparing its predictions with actual flood events that occurred during the testing period. The comparison showed a strong correlation between the predicted and actual events, reinforcing the model's reliability.

4 Recommendations

Based on the analysis and predictions of the model, several recommendations are put forward:

1. Improvement of the city's drainage systems to handle excessive water flow better.
2. Regular cleaning and maintenance of drainage channels to prevent clogs.
3. Implementation of early warning systems to alert residents of potential flood events.
4. Adoption of sustainable urban planning practices to mitigate flood risks.
5. Collaboration with meteorological agencies for accurate and timely weather forecasts and data collection.

5 Conclusion

This report presented an analysis aimed at predicting the next flood event in Lagos, Nigeria, using a machine learning and data analysis approach. The Random Forest model was identified as the most effective method for this prediction task, demonstrating high accuracy and reliability. The analysis highlighted the significance of timely prediction and the necessity for proactive measures to mitigate flood risks. Implementing the recommended actions can significantly reduce the adverse impacts of flooding in Lagos, thereby safeguarding lives and property.