Flood Prediction in Lagos, Nigeria

Abstract

This report explores the potential for predicting flood occurrences in Lagos, Nigeria, using machine learning models. We analyzed historical weather data (2004-2024) and flood occurrences to identify patterns and relationships. The data revealed an imbalanced distribution, with very few flood events compared to non-flood periods. To address this, we employed undersampling techniques to create a balanced dataset for training.

Four machine learning models were evaluated for flood prediction: Logistic Regression, Random Forest, XGBoost, and Stacking Classifier. The Random Forest and XGBoost models achieved the highest accuracy (100% each). We utilized the Random Forest model to predict future flood events and identified a potential flood on July 11th, 2024.

While the model offers valuable insights for proactive planning and disaster risk management, limitations exist. The model's accuracy depends on the quality of historical data and may not capture all flood-influencing factors like sudden storm intensity changes. Additionally, flood predictions are probabilistic, and unforeseen circumstances can influence the actual outcome.

Continuous monitoring, data collection, and model refinement are crucial for improving prediction accuracy and ensuring the safety of Lagos residents.

1. Introduction

Flooding is a major natural hazard that disrupts lives, damages property, and poses a threat to public safety. In Lagos, Nigeria, a rapidly growing megacity, the risk of flooding is particularly concerning due to several factors, including its low-lying coastal location, increasing urbanization, and potential changes in weather patterns.

This report aims to analyze historical weather data and flood occurrences in Lagos to gain insights into flood risk and explore the possibility of predicting future flood events.

2. Data Acquisition and Preprocessing

The weather data for Lagos from 2004 to 2024 was extracted from Weather Data Services . Additionally, historical flood occurrence data for the same period was gathered from various sources, including social media (Twitter, YouTube), research papers, news articles (CNN, etc.), and government websites. These datasets were merged to create a comprehensive dataset for analysis and prediction.

The combined dataset included the following attributes:

- **Datetime:** Essential for tracking historical flood events and setting up a timeline.
- **Rain Occurrence**: A key factor in flooding.
- **Precipitation Probability:** The likelihood of precipitation can help assess flood risk.
- **Wind Direction:** May influence where rain concentrates and contribute to flooding.
- **Sea Level Pressure:** Abnormally high or low pressure can impact coastal flooding.
- **Flood Occurrence:** Crucial for identifying patterns of past flood events.
- **Temperature (max & min):** While not directly causing floods, extreme temperatures can affect evaporation rates.
- **Humidity:** High humidity can contribute to heavier rainfall.
- **Wind Speed & Wind Gust:** Strong winds can worsen storm surges and coastal flooding.

2.1 Data Cleaning and Preprocessing

During data cleaning, we addressed missing values using different strategies depending on the variable. For example, missing rain occurrence values were replaced with zero, while missing sea level pressure and wind gust values were imputed using the median to avoid data loss. We then performed descriptive analysis to understand the distribution and central tendency of the data.

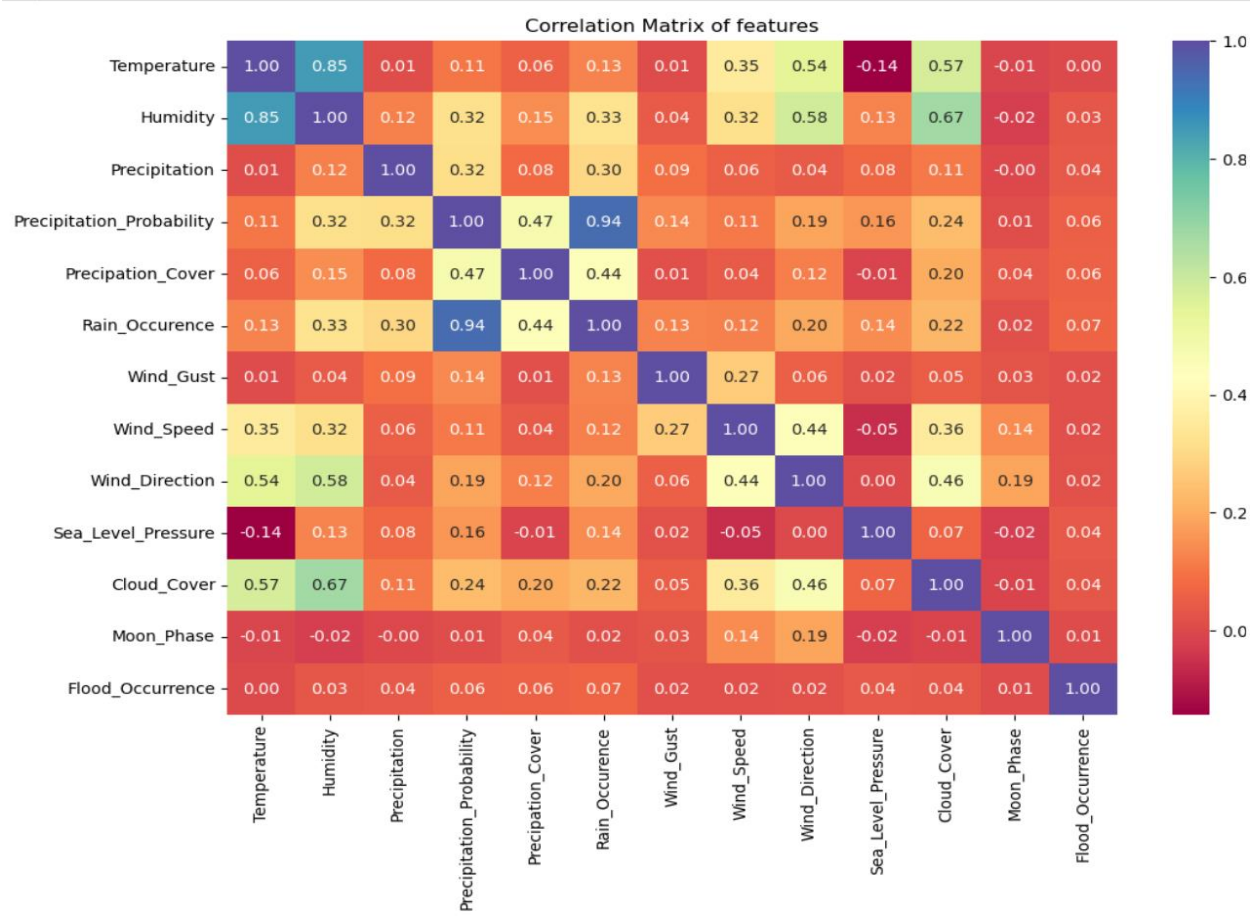
```
1 data.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Temperature	7357.0	26.158407	5.928894	0.0	26.00	27.30	28.50	34.60
Humidity	7357.0	78.954832	18.663443	0.0	79.60	83.40	86.60	100.00
Precipitation	7357.0	4.944502	18.254626	0.0	0.00	0.00	1.70	299.00
Precipitation_Probability	7357.0	42.231888	49.396236	0.0	0.00	0.00	100.00	100.00
Precipation_Cover	7357.0	5.354873	13.244749	0.0	0.00	0.00	4.17	100.00
Rain_Occurence	7357.0	0.453989	0.497912	0.0	0.00	0.00	1.00	1.00
Wind_Gust	7357.0	29.001658	9.295310	0.0	25.60	28.10	31.70	137.20
Wind_Speed	7357.0	21.233818	12.964308	0.0	15.10	20.50	25.90	277.90
Wind_Direction	7357.0	197.798967	79.277413	0.0	198.90	221.60	241.00	360.00
Sea_Level_Pressure	7357.0	1012.032486	1.747282	1005.9	1010.80	1011.90	1013.20	1017.30
Cloud_Cover	7357.0	57.556395	20.712817	0.0	49.30	57.90	68.50	100.00
Moon_Phase	7357.0	0.459155	0.300545	0.0	0.19	0.46	0.72	0.98
Flood_Occurrence	7357.0	0.006117	0.077975	0.0	0.00	0.00	0.00	1.00

Temperature: The temperature data shows a moderate range, with most values clustering around the mid-20s (°C). The presence of a minimum value of 0.0°C may indicate some data anomalies or outliers that could be worth investigating.

Precipitation: The precipitation data is highly variable, with many days having no rainfall at all, as evidenced by the median and 25th percentile values being 0.0 mm. However, there are also extreme rainfall events, as indicated by the maximum value of 299.0 mm. This variability is crucial for understanding flood risks and patterns.

We sought for linear relationships between the variables and the target yet the correlation matrix showed that none of the variables have a strong linear relationship with the target variable Flood_Occurrence.



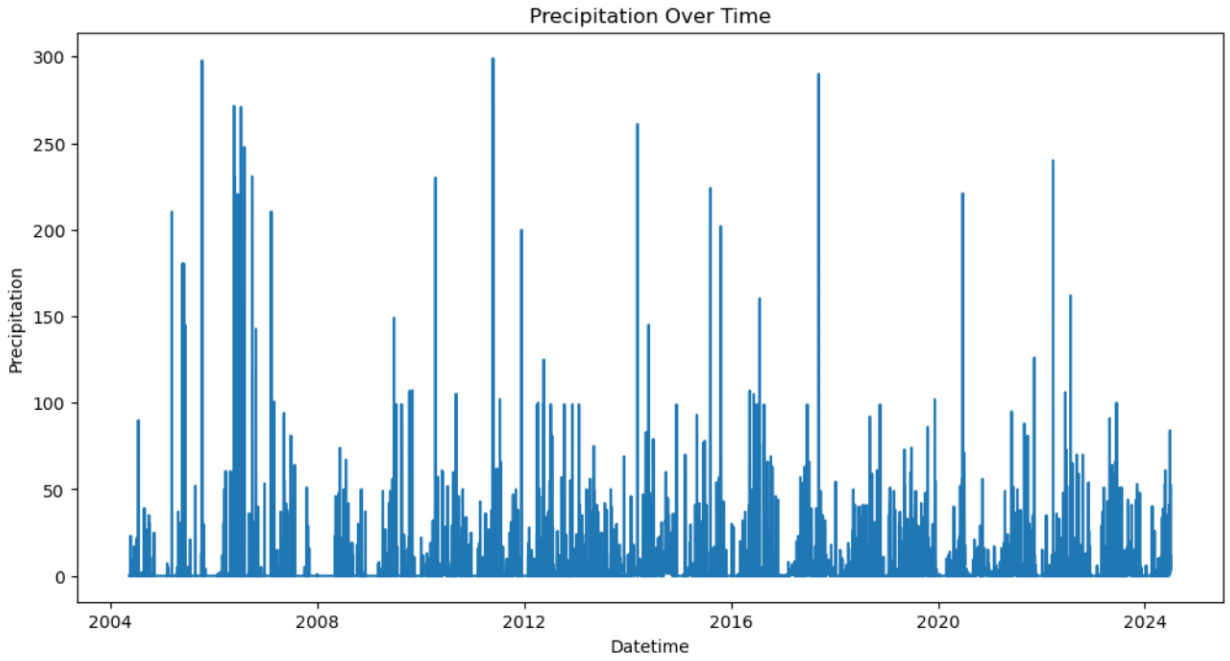
This suggests that predicting floods based on these individual features alone might be challenging, and a more complex model or additional features might be required to improve prediction accuracy. The low correlations indicate that the model may need to rely on a combination of these features and possibly other data to make accurate flood predictions.

2.2 Data Exploration and Visualization

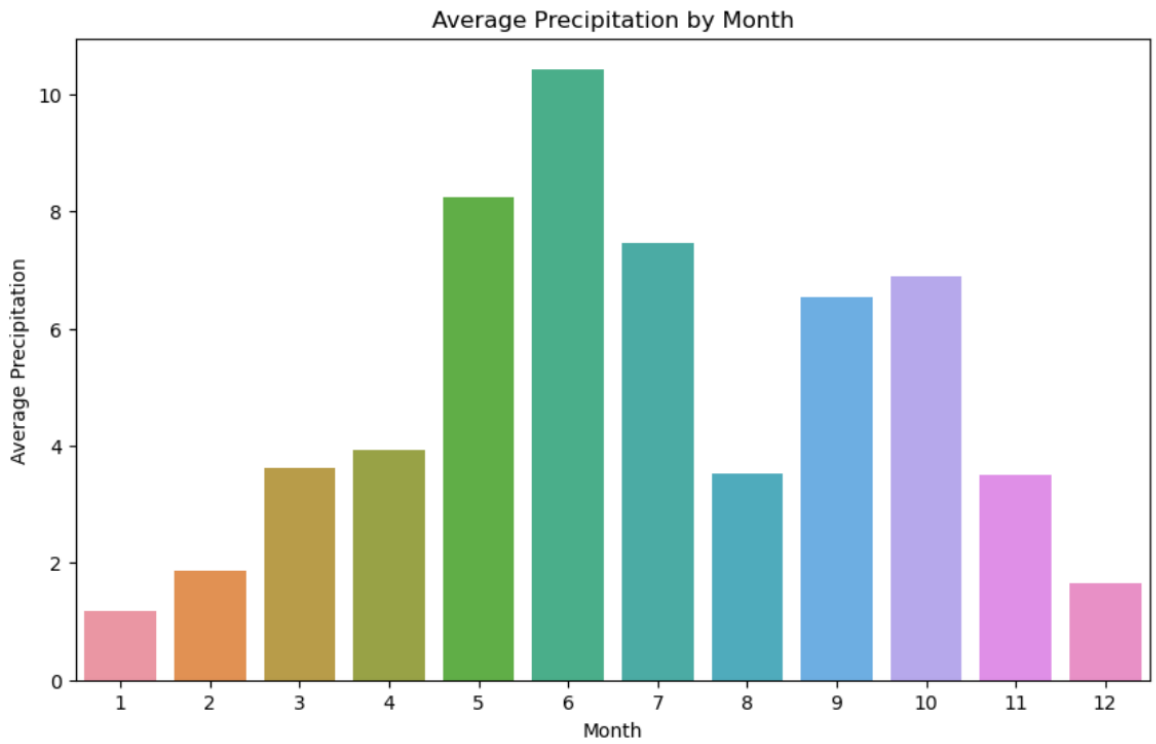
We also performed some feature selection and feature engineering for further analysis and to aid visualization.

We extracted, Day, Month and year from the Datetime in order to visualize year to year and month to month patterns.

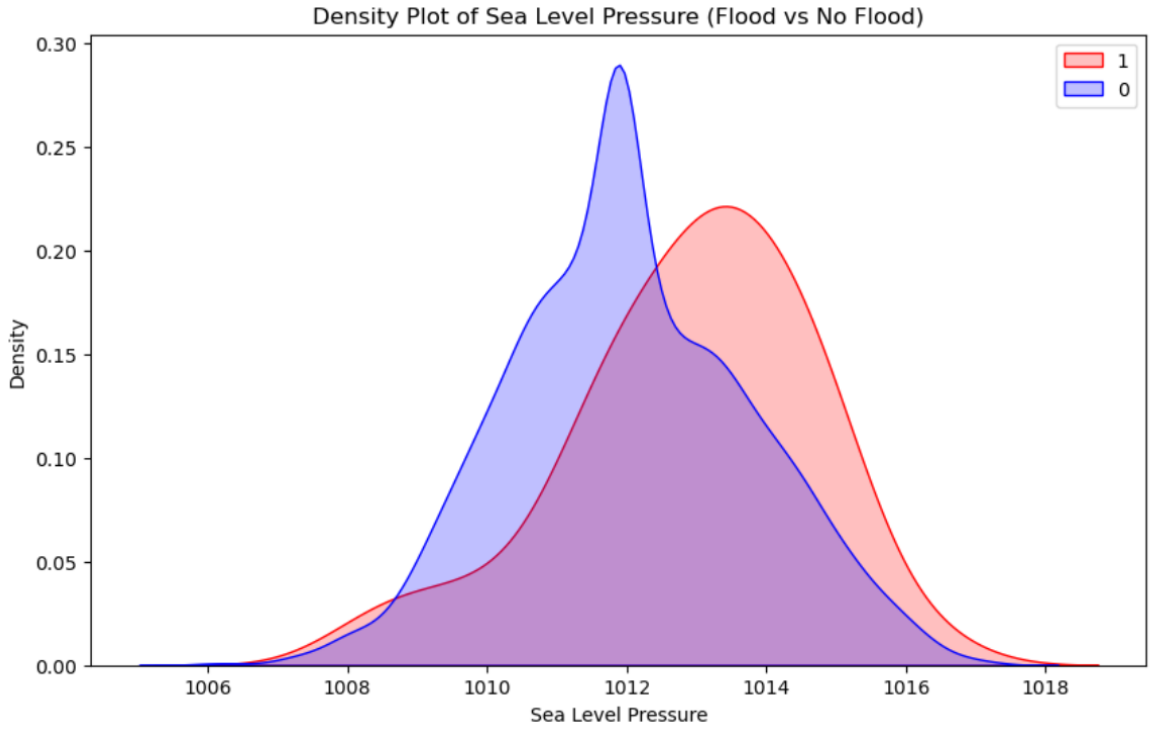
We employed various data visualization techniques to explore the relationships between weather variables and flood occurrences. Here are some key findings:



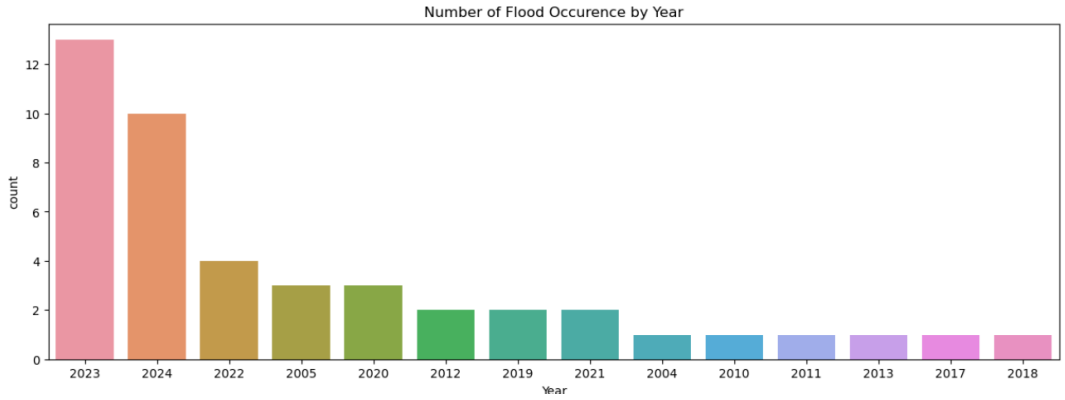
Precipitation Over Time: This revealed significant year-to-year variability in precipitation levels, with peaks indicating periods of heavy rainfall and potential extreme weather events. While the data suggests some cyclical patterns, more recent years seem to have sustained higher precipitation levels, although not as extreme as some earlier events.



Average Precipitation by Month: It was identified that June is the month with the highest average precipitation, followed by May, and July, suggesting a rainy season spanning these months. January, December, and February displayed the lowest average precipitation, indicating drier periods.



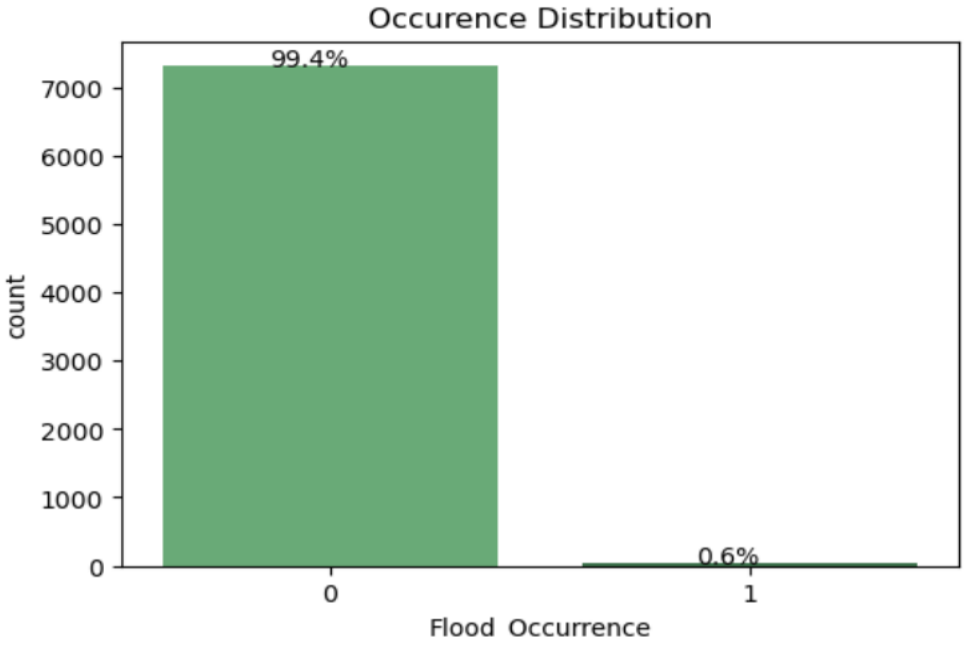
Sea Level Pressure Distribution: The density plot showed that both flood and non-flood days have peak densities around 1012 hPa for sea level pressure. However, the distribution for flood days had a broader range, with higher densities between 1010 hPa and 1016 hPa. This suggests that sea level pressure alone might not be a strong predictor but could be a contributing factor.



Flood Occurrences by Year: The chart revealed a recent increase in flood occurrences, with 2023 having the highest number. This trend suggests a potential change in environmental conditions or urban development patterns that warrant further investigation and safeguarding implementation.

NB: More visualization works can be found in my notebook.

2.3 Imbalanced Dataset



The bar chart depicting the flood occurrence distribution reveals a critical aspect of the data: imbalance. Only a tiny fraction (0.6%) represents flood occurrences, while the vast majority (99.4%) corresponds to periods with no flooding. This imbalance poses a challenge for machine learning models, as they can become biased towards the majority class during training. Consequently, models might prioritize predicting "no flood" events, leading to poor performance in identifying the crucial minority class (flood occurrences).

To address this imbalance, we employed the Undersampling technique. This approach involves reducing the majority class to the minority class (flood occurrences) until the dataset has an equal number of samples for both classes (flood and no flood). Balancing the dataset ensures the model is exposed to a fair representative sample of flood events during training. This helps the model learn the characteristics of both flood and non-flood events more effectively, ultimately improving its ability to predict flood occurrences with better accuracy and reliability.

3. Machine Learning Model Selection

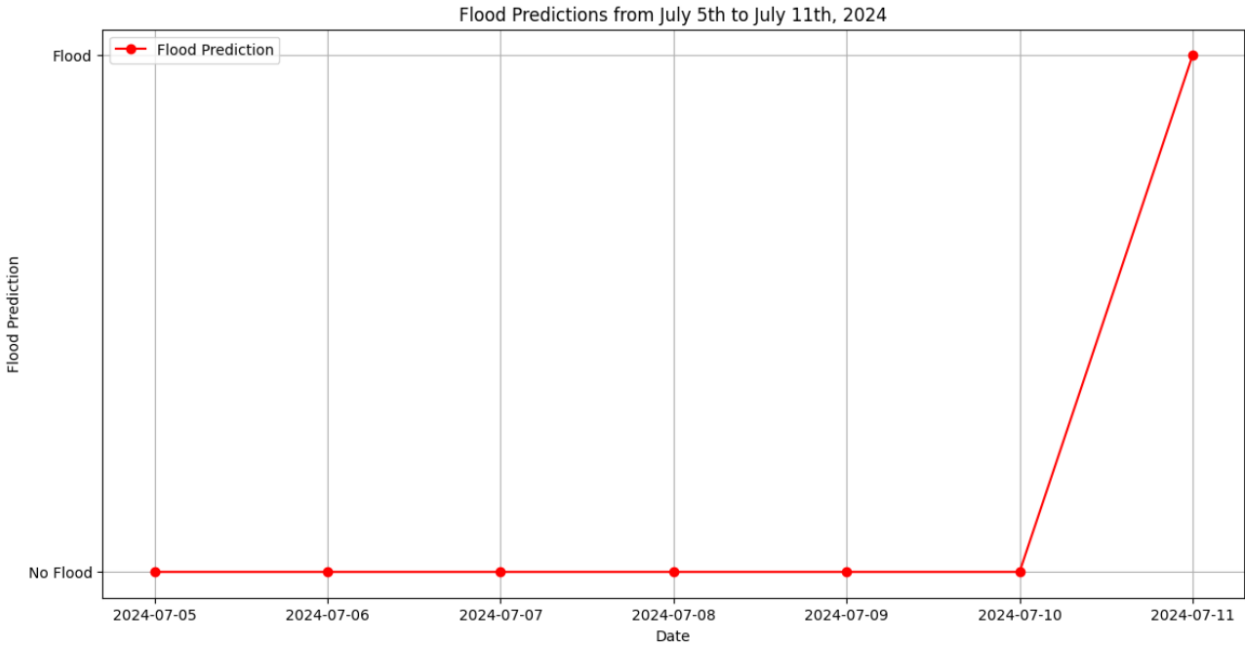
We employed four machine learning models for flood prediction: Logistic Regression, Random Forest, XGBoost, and Stacking Classifier. The rationale behind selecting these models is as follows:

- **Logistic Regression:** This is a well-established linear model that excels at classification tasks. It provides interpretable results, allowing us to understand the relationships between the weather variables and flood occurrences. While interpretability is valuable, logistic regression might struggle with complex, non-linear relationships that might exist in the data.
- **Random Forest:** This ensemble method combines multiple decision trees, leading to robustness and flexibility in handling various data types. Random forests can effectively capture non-linear relationships between features and the target variable (flood occurrence).
- **XGBoost:** This is a powerful gradient boosting algorithm known for its ability to handle complex data and achieve high predictive accuracy. XGBoost can potentially outperform other models in capturing intricate relationships within the weather data that influence flood risk.
- **Stacking Classifier**: This meta-learning approach leverages multiple models (Logistic Regression and Random Forest, in our case) to create an ensemble model. The stacking classifier aims to combine the strengths of individual models, potentially leading to superior performance compared to any single model used in isolation.

4. Model Performance and Future Prediction

The Random Forest and the XGB Models performed best scoring **100%** each**.** Stacking Model scored accuracy of **84%** and **72%** for Logistic Regression model. emerged as the best performing model, followed by XGBoost. This suggests that the ensemble approach was most effective in capturing the complex relationships between weather variables and flood occurrences in the Lagos dataset. Logistic regression, while interpretable, might not have been as adept at handling the non-linearities within the data.

To predict the next likely date of a flood, we utilized the trained Random Forest classifier model. We would first obtain future weather forecasts from Weather Data Services and use these predicted weather conditions as input to the model. The Random Forest would then analyze these features and provide a prediction for the likelihood of a flood occurring on that specific date. It's important to remember that model predictions are probabilistic, and a high predicted probability of flood occurrence should be interpreted with caution and potentially followed by further investigation from relevant authorities.



The flood prediction chart from July 5th to July 11th, 2024, indicates that a flood is predicted to occur on July 11th, 2024. The Random Forest Model has identified this specific date as having conditions conducive to flooding, while the other dates within this range are predicted to have no flood occurrences.

5. Limitations and Ongoing Development

Flood prediction remains a complex task, and our model is subject to several limitations:

- The model's accuracy relies on the quality and completeness of historical data. Incomplete or inaccurate data can lead to biases in the predictions.
- While the model incorporates various weather variables, it might not capture all factors influencing floods, such as sudden changes in storm intensity, tidal patterns, or drainage system capacity.
- Uncertainty in Predictions: Flood predictions are inherently probabilistic. The model assigns a likelihood of a flood occurring, but the actual outcome may vary depending on unforeseen circumstances.
-

6. Conclusion

Addressing the data imbalance and employing a combination of machine learning models, we were able to develop a system for predicting flood occurrences in Lagos. This flood prediction report provides valuable information for proactive planning and disaster risk management in Lagos. By taking necessary precautions and staying informed, authorities and residents can enhance their preparedness for potential flooding events.

The Random Forest and the XGBoost Model achieved the most accurate predictions. However, it's crucial to acknowledge that flood prediction remains a complex task. Ongoing monitoring, data collection, and model refinement are essential for improving prediction accuracy and ensuring the safety of Lagos residents.

Additionally, the model predicted a flood event for July 11th, 2024. It is crucial to remember that this prediction is based on the model's analysis of historical data and current weather forecasts. Continuous monitoring and refinement of the model are essential to improve the accuracy of future predictions.

APPENDIX

I. My notebook
II. The merged dataset
III. The Forecasting data