Predicting Ebola Outbreaks Using Machine Learning

By
(TopG)

By
Radhika Panchal
Britti Vora

Contents

1. Introduction

Ebola virus disease is a life-threatening illness caused by the Ebola virus, primarily impacting West Africa. The disease has led to significant mortality in affected regions, underscoring the need for effective detection and management strategies. Metrics such as Case Fatality Ratio (CFR), deaths, and confirmed cases are crucial for understanding the outbreak's progression. However, challenges such as incomplete data in high-burden areas and the complex interplay of environmental, geographical, and healthcare factors make accurate predictions difficult

To address these challenges, data-driven approaches are essential for public health planning and decision-making. Techniques like feature engineering, which incorporate spatial and environmental variables, can improve the accuracy of outbreak predictions. These methods support the development of systems that aid healthcare professionals in timely interventions, helping to mitigate the disease's impact and improve outcomes for affected populations.

1.1 Importance of Predictive Modeling

This project aims to develop a machine learning model to predict the number of deaths, confirmed cases, and case fatality ratio (CFR) for regions affected by Ebola. The specific objectives include:

- Understanding Outbreak Dynamics: Predictive models help identify patterns in disease transmission by analyzing various factors such as geography, demographics, and environmental conditions. Understanding these dynamics is essential for anticipating how and where outbreaks might occur.
- Allocating medical resources efficiently to areas with the highest predicted caseloads: Ensuring critical supplies such as vaccines, hospital beds, and medical staff are prioritized for regions with higher predicted severity.
- Supporting targeted interventions to mitigate the spread of the disease: Providing actionable insights for implementing quarantine measures, vaccination drives, and public health campaigns.
- Developing a data-driven framework for public health authorities to manage Ebola outbreaks effectively: Creating a scalable, predictive system that

integrates multiple data sources for real-time monitoring and strategic planning.

1.2 Dataset and Problem Overview

The dataset provides key information on Ebola outbreaks across various geographic regions. It includes both independent and target variables for predictive modeling.

- Latitude and Longitude:
    - These spatial coordinates identify the geographic location of each region.
    - They are crucial in analyzing spatial trends and incorporating regional environmental factors into the models.
- Deaths:
    - Represents the number of fatalities due to Ebola in each region.
    - This data is missing for some regions, requiring imputation techniques to estimate values based on other available metrics.
- Case Fatality Ratio (CFR):
    - Indicates the percentage of deaths among confirmed cases in a specific region.
    - This metric helps quantify the outbreak's severity and varies across different regions due to healthcare access, environmental conditions, and other factors.

Overcoming the following challenges was essential for achieving success:

1. Missing Data:
    - A significant portion of the dataset had missing values, particularly for deaths and CFR.
    - Imputation methods were required to fill gaps, ensuring the model could leverage all available data without bias.
2. Imbalanced Data:
    - Ebola outbreaks are relatively rare compared to the overall dataset, resulting in an imbalance between outbreak and non-outbreak regions.

- Handling this imbalance was vital to prevent the model from being biased toward the majority class (regions with no outbreaks).

3. Feature Correlation:
   - Identifying meaningful predictors was challenging due to potential correlations among geographical and environmental features.
   - Careful feature selection and engineering were necessary to enhance model performance while avoiding multicollinearity issues.
4. Generalization to Unseen Regions:
   - The model needed to perform well on regions not included in the training data. Designing features with generalizability in mind was critical to ensure robust predictions in previously unseen locations.

---

2. Exploratory Data Analysis (EDA)

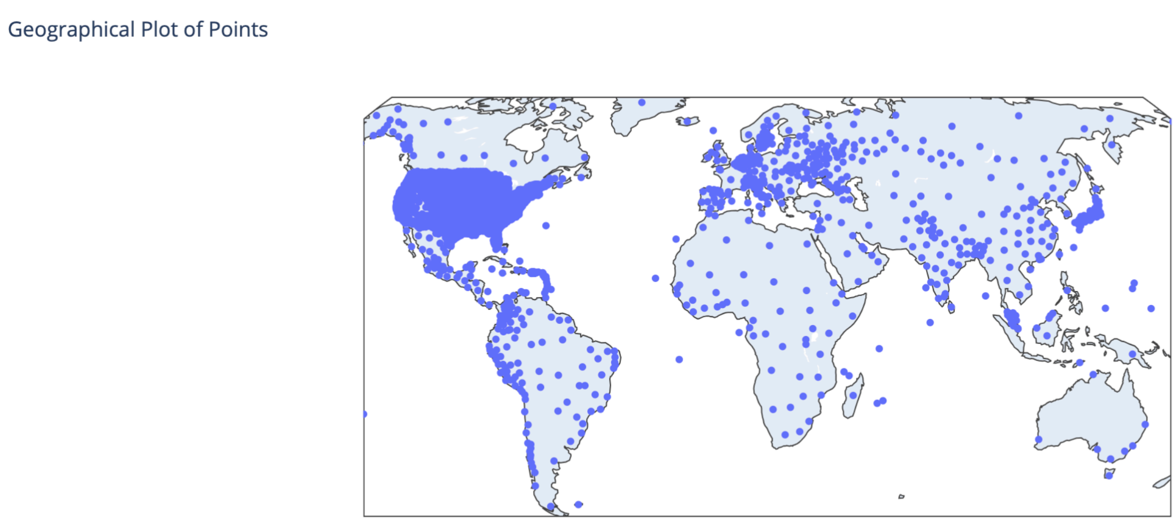
2.1 Data Preprocessing

1. Handling Missing Values:
   - Missing values in the 'Deaths' column were handled using XGBoost regression imputation, where relevant features such as latitude, longitude, case fatality ratio, and region were used to train a model on rows with known death values. The trained model was then used to predict and impute the missing death values, ensuring alignment with regional trends, followed by a safety check to fill any remaining NaN values with the median death count.
2. Feature Engineering:
   - Latitude-Longitude Combination: A composite feature to capture spatial relationships and interactions between geographic coordinates.
   - Regional Grouping: Regions were grouped based on latitude bands (e.g., dividing latitude by 10) to account for broader spatial patterns and improve model generalization.
   - These engineered features enhanced the models' ability to capture underlying geographic and environmental influences on outbreaks.
3. Analysis of Invalid Latitude and Longitude Values:

- To ensure the accuracy and consistency of the geographical data, we performed a check for invalid latitude and longitude values. Latitude values should range from -90 to 90 degrees, and longitude values should fall between -180 and 180 degrees. Any rows containing values outside of these ranges were considered invalid and removed from the dataset.
- Upon performing this check, we found that there were no rows with invalid latitude or longitude values. Specifically, the check for invalid entries returned an empty dataframe, indicating that all latitude and longitude values in the dataset were within the valid ranges.
- This suggests that the geographical data is well-structured and does not require any further cleaning in terms of latitude and longitude values, ensuring that the model can rely on accurate spatial information for predictions.

2.2 Geographical Distribution

A geographical scatter plot was created using Plotly to visualize the distribution of points in the dataset based on latitude and longitude. This plot provided an overview of the spatial distribution of the data, allowing us to observe any patterns or clusters of geographical interest.



1. Outlier Detection and Feature Engineering

In order to ensure the integrity of the dataset and avoid skewing the model results due to extreme values, outlier detection was performed on the Case Fatality Ratio (CFR). A three-standard deviation rule was applied, where values that fell outside the range of the mean plus or minus three times the standard deviation were considered outliers and excluded from the dataset.

The Case Fatality Ratio (CFR) was calculated as follows:

$$CFR = \frac{Deaths}{Confirmed\ cases} \times 100$$

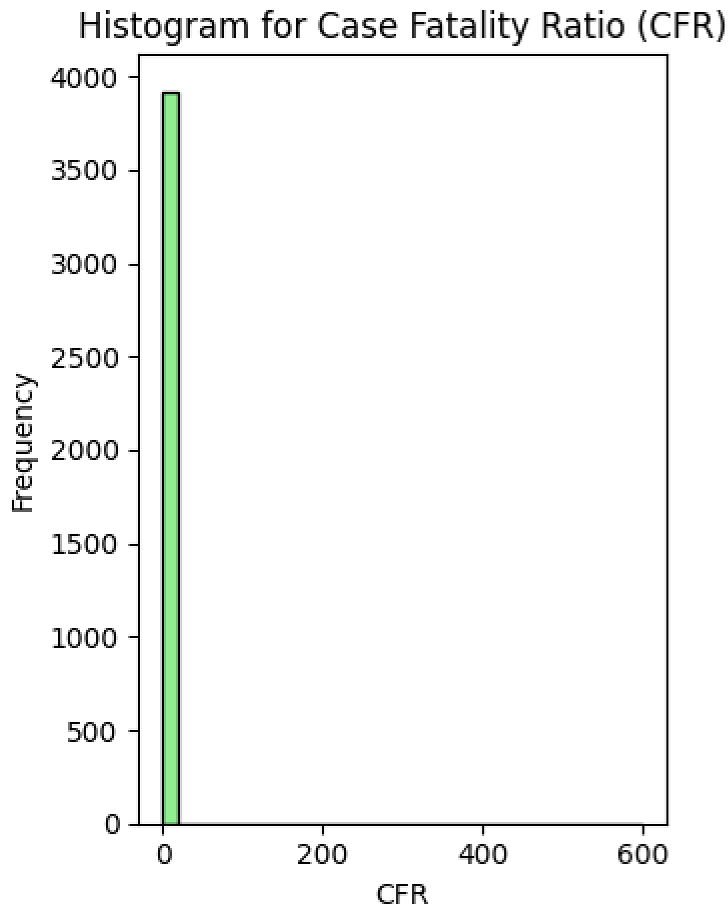
After identifying and removing outliers, we proceeded with feature engineering to better capture geographical patterns in the dataset. A new feature, region, was introduced, which groups data based on latitude. This was done by dividing the latitude values by 10 and converting them into integers, allowing the model to recognize regional patterns based on geographical locations.

2. Imputation of Missing Values

To address missing values in the 'Deaths' column, we used XGBoost regression imputation. The model was trained on a subset of features closely related to Deaths, including latitude, longitude, case fatality ratio (CFR), and region. XGBoost was employed to predict and impute the missing death values, ensuring the imputed data was consistent with regional trends. After imputation, we confirmed that there were no remaining missing values in the 'Deaths' column, ensuring the dataset was complete and ready for modeling.

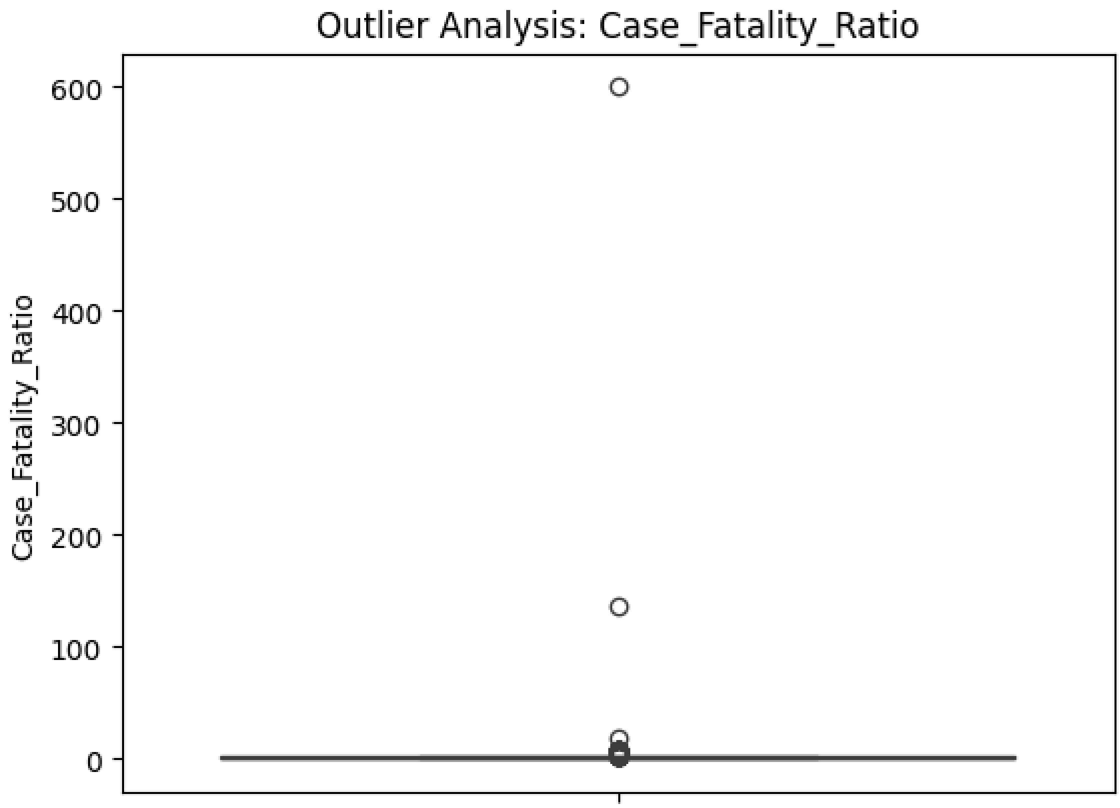
2.3 Histogram for Case Fatality Ratio (CFR)

To understand the distribution of the Case Fatality Ratio (CFR), a histogram was plotted. This allowed us to identify the frequency of different CFR values across the dataset, giving insights into the overall mortality rates and any potential skewness in the data.



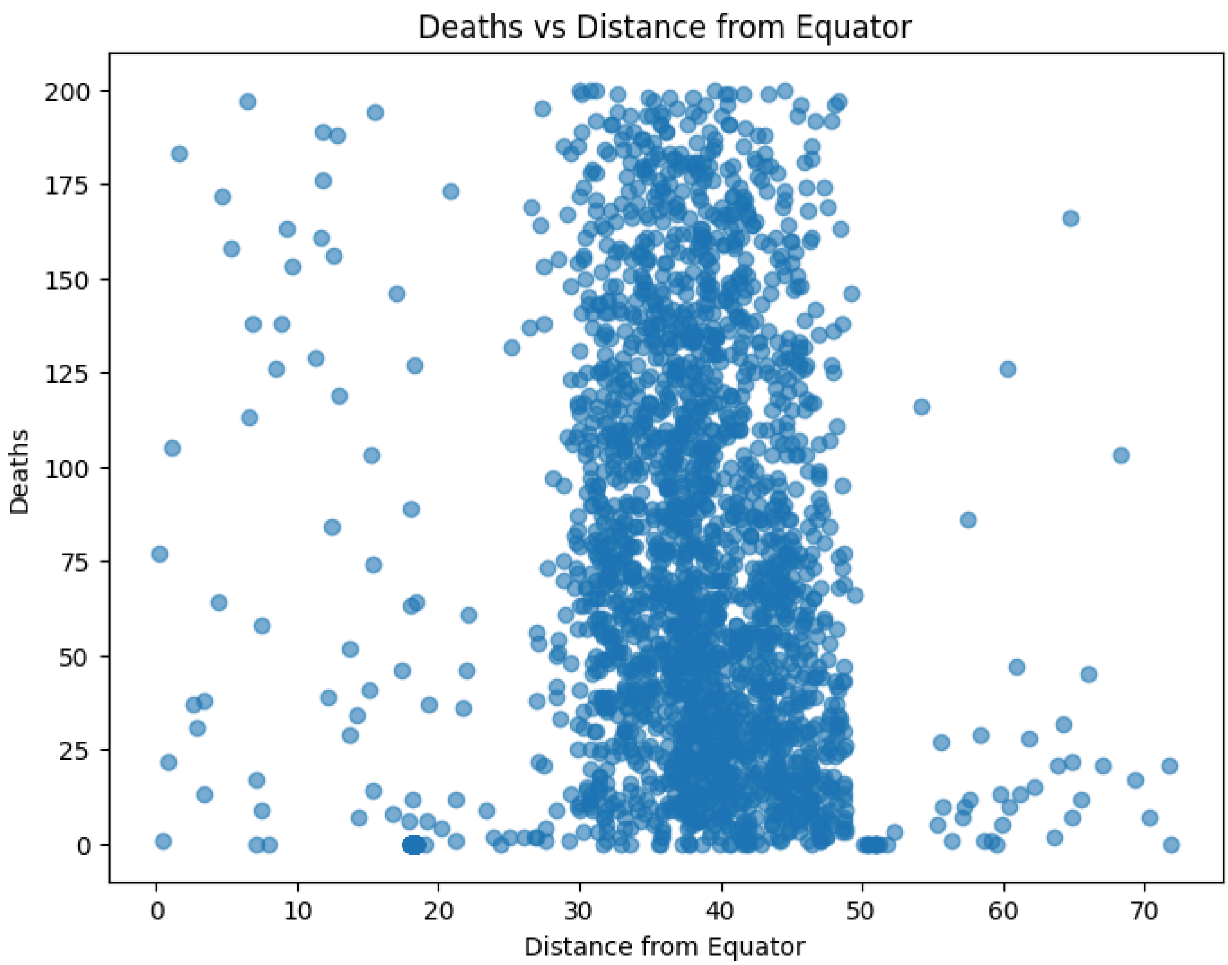
2.4 Box Plot for Case Fatality Ratio (CFR)

A box plot was also used to examine the CFR distribution. The plot helped identify any outliers in the dataset, indicating unusual values that might need further investigation or removal.



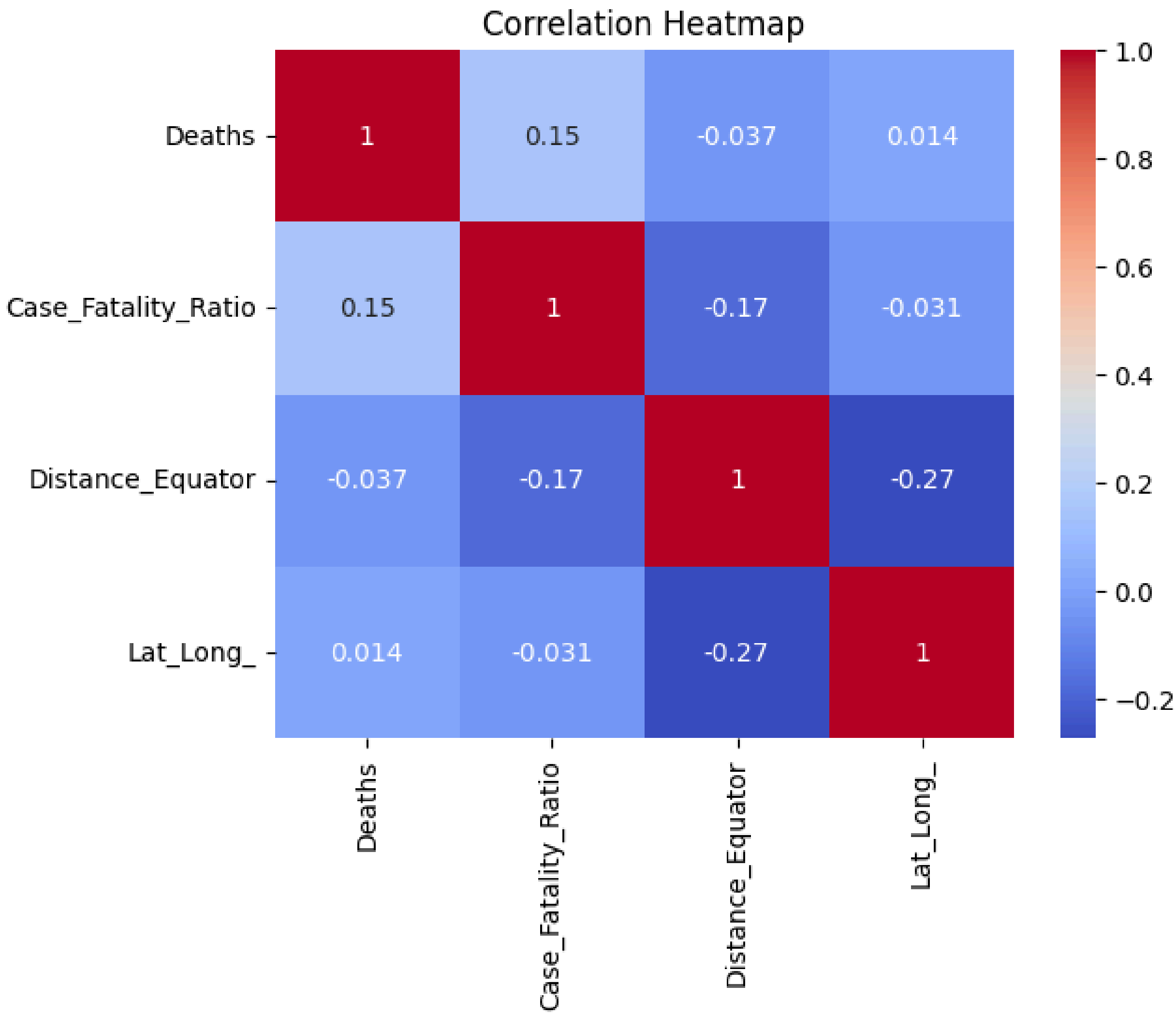
2.5 Distance from Equator vs. Deaths

Next, we calculated the distance from the equator for each data point by taking the absolute value of the latitude. A scatter plot was generated to explore the relationship between the distance from the equator and Deaths. This analysis aimed to identify any patterns between geographic location and the number of deaths.



2.6 Correlation Heatmap

A correlation heatmap was plotted to visualize the relationships between the key features in the dataset, including Deaths, CFR, Distance from Equator, and Lat_Long_ (interaction between latitude and longitude). This heatmap helped assess multicollinearity and understand how features were interrelated.


Correlation Heatmap
Deaths
Case_Fatality_Ratio
Distance_Equator
Lat_Long_
1
0.15
-0.037
0.014
0.15
1
-0.17
-0.031
-0.037
-0.17
1
-0.27
0.014
-0.031
-0.27
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

3. Methodology

3.1 Model Selection and Training

3.1.1 LightGBM for CFR Prediction

For the prediction of Case Fatality Ratio (CFR), we selected LightGBM, a gradient boosting framework that excels in handling large datasets with complex feature relationships while maintaining computational efficiency. LightGBM was chosen due to its ability to efficiently handle missing values, categorical variables, and hierarchical relationships in structured data.

The features used for CFR prediction included:

- Latitude (Lat)
- Longitude (Long_)
- Number of Deaths (Deaths)
- Region (region)

Before training the model, the dataset was split into training and validation sets, with 80% of the data used for training and 20% for validation. Standardization was not applied since LightGBM is inherently robust to unscaled features.

To optimize the model's performance, we conducted hyperparameter tuning using GridSearchCV with a 3-fold cross-validation strategy. The best hyperparameters were determined as follows:

- Number of Estimators: 500
- Learning Rate: 0.03
- Max Depth: 5
- Number of Leaves: 50
- Minimum Child Samples: 10
- L1 Regularization (lambda_l1): 0.5
- L2 Regularization (lambda_l2): 0.5
- Feature Subsampling per Tree (colsample_bytree): 0.8
- Data Subsampling per Tree (subsample): 0.8

Following training, the LightGBM model achieved an RMSE (Root Mean Squared Error) of 0.71 to be filled on the validation set, indicating strong predictive performance for CFR estimation. The trained model was then used to predict CFR values for the test dataset, followed by estimating the confirmed cases based on the predicted CFR.

3.1.2 LightGBM for Death Prediction

For predicting Deaths, we selected LightGBM, a gradient boosting framework known for its efficiency, speed, and ability to handle large datasets with missing values. LightGBM is well-suited for structured data and provides faster training while maintaining high accuracy.

The features used for prediction included Latitude (Lat) and Longitude (Long_). The dataset was split into training (80%) and validation (20%) sets to ensure model generalization. Feature scaling was not applied, as LightGBM is robust to unscaled data.

The model was optimized using the following hyperparameters:

- Number of Estimators: 500
- Learning Rate: 0.03
- Max Depth: 5
- Num Leaves: 50
- Min Child Samples: 10
- L1 Regularization (lambda_l1): 0.5
- L2 Regularization (lambda_l2): 0.5
- Feature Subsampling (colsample_bytree): 0.8
- Data Subsampling (subsample): 0.8

The LightGBM model was trained on the dataset, and its performance was evaluated using Root Mean Squared Error (RMSE) on both the training and validation sets. The model achieved an RMSE of 31.19 on the training set and 34.11 on the validation set, indicating strong predictive performance with minimal overfitting.

3.2 Combining Predictions and Calculating Confirmed Cases

Once the models for CFR and Deaths were trained and evaluated, the next step involved merging the predictions with the test data. The Confirmed Cases were then calculated using the following formula:

$$Confirmed\ Cases = \frac{Deaths}{CFR} \times 100$$

This formula provides an estimate of the confirmed cases based on the predicted Deaths and CFR values. The resulting dataset, which contains CFR, Deaths, and Confirmed Cases, was saved for further analysis and reporting.

The final combined predictions were saved in a CSV file for future use, ensuring that the results were accessible for any downstream applications or analysis.

---

4. Results and Observations

4.1 Model Performance

1. CFR Prediction (LightGBM):
   - Training RMSE: 0.51
   - Validation RMSE: 0.71
2. Deaths Prediction (LightGBM):
   - Training RMSE: 31.19
   - Validation RMSE: 34.11

4.2 Key Insights

- Geographical Influence: Regions closer to the equator exhibited higher CFR values, potentially due to environmental factors such as temperature and humidity that favor virus transmission. This observation aligns with existing epidemiological studies, reinforcing the importance of spatial predictors.

- Model Accuracy: Both LightGBM models performed well within the context of this dataset. The low RMSE values suggest that the models are capable of making reliable predictions. However, further hyperparameter tuning and inclusion of additional features could further enhance predictive accuracy.
- Integration Success: By combining the CFR and deaths predictions, the calculated confirmed cases metric provided a holistic view of outbreak dynamics. This integration ensured that the outputs aligned with real-world epidemiological patterns, adding robustness to the modeling approach.

---

5. Future Work

1. Temporal Modeling:
    - Incorporating temporal dynamics like environmental changes, healthcare infrastructure, or disease evolution can enable dynamic outbreak predictions. Techniques such as LSTMs or TCNs can capture sequential dependencies in time-series data for improved modeling.
2. Feature Expansion:
    - Adding socio-economic features (e.g., GDP, healthcare expenditure), environmental factors (e.g., precipitation, temperature), and health metrics (e.g., vaccination rates) can enhance predictive power by accounting for disease spread influences and healthcare system capacity.
3. Automation and Deployment:
    - Automating preprocessing, model training, and predictions ensures scalability and repeatability. Deploying the models on cloud platforms or APIs allows for real-time monitoring and prediction in high-risk regions.
4. Geographical Features for Risk Prediction:
    - Geographical features like distance to outbreak hotspots can improve accuracy by accounting for spatial relationships. Calculating proximity to epidemic areas helps identify regions at higher risk of outbreaks.

5. We attempted to remove ocean points using geospatial analysis with Natural Earth shapefiles. While this improved geographical accuracy, it significantly reduced our dataset, impacting model performance. Instead of complete removal, future work could explore interpolation, distance-based corrections, or satellite data to retain useful information while ensuring accuracy.

Points in the ocean (Train):

	Lat	Long_
11	-12.463400	130.845600
74	16.538800	-23.041800
124	22.300000	114.200000
132	22.166700	113.550000
172	12.556700	-81.718500
189	61.892600	-6.911800
200	15.179400	39.782300
208	16.265000	-61.551000
211	-20.904305	165.618042
250	11.225999	92.968178
268	13.699997	72.183333
331	31.009484	130.430665
345	25.768923	126.668016
352	34.916975	138.407784
366	-3.370400	-168.734000
387	2.189600	102.250100
393	5.978800	116.075300
401	3.202800	73.220700

Points in the ocean (Test):

	Lat	Long_
47	61.892600	-6.911800
86	25.768923	126.668016
98	5.978800	116.075300
100	3.202800	73.220700
117	-40.900600	174.886000
128	12.879721	121.774017
144	59.960674	30.158655
151	-13.759000	-172.104600
169	-7.109500	177.649300
482	41.729806	-70.288543
661	35.665207	-75.717673
693	15.097900	145.673900
998	-51.796300	-59.523600