Comprehensive Analysis of Ebola Outbreak Prediction Model

Deepesh Yadav Atria University Bengaluru

January 26, 2025

1 Introduction and Project Overview

This document presents a comprehensive analysis of an Ebola outbreak prediction model developed using machine learning techniques and epidemiological principles. The model aims to predict outbreak patterns, death rates, and case fatality ratios across various geographical locations, providing valuable insights for public health interventions.

2 Data Overview & Characteristics

2.1 Dataset Structure

The analysis utilizes two primary datasets:

- **Training Dataset:** Comprises 3,925 records with:
  - Geographic coordinates spanning latitude ($-71.95°$ to $71.71°$)
  - Longitude range ($-178.12°$ to $178.07°$)
  - Death records: 2,458 complete entries (1,558 missing)
  - Case Fatality Ratio (CFR): 3,972 records
- **Test Dataset:** Contains 4,016 locations requiring predictions

2.2 Statistical Distribution

Key statistical metrics observed in the training data:

Metric	Mean	Std Dev	Min	Max
Deaths	70.75	55.36	0	200
CFR (%)	13.90	123.02	0	7570.77

Table 1: Training Data Statistics

3 Methodological Framework

3.1 Feature Engineering Architecture

3.1.1 Geographic Distance Implementation

The model employs the Haversine formula for accurate distance calculation:

$$d = 2R \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos\phi_1 \cos\phi_2 \sin^2\left(\frac{\Delta\lambda}{2}\right)}\right) \tag{1}$$

Where:

- d represents the great-circle distance
- R denotes Earth's radius (6371 km)
- ϕ_1, ϕ_2 are latitudes of points 1 and 2
- $\Delta\lambda$ represents the longitude difference

3.1.2 Outbreak Centers

Three primary outbreak centers were identified:

Location	Latitude	Longitude
Liberia	6.4281°	-9.4295°
Sierra Leone	8.4606°	-13.2317°
Guinea	9.9456°	-9.6966°

Table 2: Primary Outbreak Centers

3.2 Regional Analysis Framework

Implementation of K-means clustering with k=8 revealed significant regional variations:

- **Region 5:** Highest mortality (52.95 ± 60.84 deaths)
- **Region 3:** Lowest mortality (5.42 ± 22.76 deaths)
- **Regional CFR:** Variations indicate healthcare capacity differences

4 Model Architecture & Implementation

4.1 RandomForest Configuration

Optimized hyperparameters through extensive grid search:

- **max_depth:** 6 (Prevents overfitting while maintaining complexity)
- **min_samples_leaf:** 2 (Ensures prediction stability)
- **min_samples_split:** 6 (Optimizes node division)
- **n_estimators:** 400 (Balances complexity and performance)

4.2 Performance Metrics

Model evaluation revealed:

$$R^2 = 0.3906 \text{ (39.06\% variance explained)} \tag{2}$$

Distance effects analysis:

$$\text{Distance Effects} = \begin{cases} -86.08 & \text{Center 0 (strongest effect)} \\ -83.89 & \text{Center 1} \\ -83.64 & \text{Center 2} \end{cases} \tag{3}$$

5 Prediction Analysis & Results

5.1 Test Data Predictions

Comprehensive prediction statistics:

Metric	Mean	SD	Range
Deaths	84.35	23.35	50-151.14
CFR (%)	25.01	0.03	25-25.9
Confirmed Cases	337.23	93.37	193.05-604.27

Table 3: Test Data Prediction Statistics

6 Domain Knowledge Integration

6.1 Epidemiological Principles

6.1.1 Distance Decay Implementation

Correlation analysis revealed consistent patterns:

$$\begin{aligned} \text{corr}(\text{dist}_0, \text{deaths}) &= -0.071 \\ \text{corr}(\text{dist}_1, \text{deaths}) &= -0.068 \\ \text{corr}(\text{dist}_2, \text{deaths}) &= -0.073 \end{aligned} \tag{4}$$

6.1.2 Healthcare Capacity Integration

The model incorporates healthcare system capabilities through:

- Regional clustering patterns
- CFR range constraints (25-90%)
- Geographic accessibility factors

7 Model Limitations & Future Enhancements

7.1 Current Limitations

- **Performance Constraints:**

  - Moderate R^2 score (0.3906)
  - Geographic bias in training data
  - Limited temporal feature integration

- **Data Quality Issues:**
  - Significant missing death records (1,558)
  - Potential reporting biases
  - Geographic coverage gaps

7.2 Enhancement Strategies

- **Technical Improvements:**
  - Advanced feature engineering
  - Algorithm optimization
  - Temporal pattern integration

- **Data Enhancement:**
  - Additional data source integration
  - Healthcare infrastructure data
  - Improved geographic coverage

8 Application to New Locations

8.1 Implementation Framework

The model application process involves:

- **Geographic Processing:**
  - Coordinate standardization
  - Distance calculation
  - Regional cluster assignment

- **Prediction Pipeline:**

  - Death rate prediction
  - CFR calculation
  - Confirmed cases estimation

8.2 Validation Methodology

Quality assurance procedures include:

- Geographic validation
- Prediction range verification
- Epidemiological consistency checks
- Regional pattern validation

9 Conclusion

This comprehensive model demonstrates significant potential for predicting Ebola outbreak patterns while acknowledging current limitations. The integration of epidemiological principles with machine learning techniques provides a robust framework for public health applications. Future enhancements focusing on data quality and model sophistication will further improve prediction accuracy and reliability.