Report: Predicting Deaths and Confirmed Cases from Ebola Virus

Vaibhav Sonkar (24m0301@iitb.ac.in); Sarita Yadav; Ishrat Jan

Abstract:

Machine learning offers significant potential in combating infectious diseases. By analysing vast datasets, these models can identify disease hotspots, forecast outbreaks, and uncover hidden patterns in transmission. This knowledge empowers researchers to understand the influence of factors like geography, demographics, and the environment, enabling early detection and targeted interventions.

Ebola virus poses significant public health risks. Early prediction of expected cases and deaths in various regions is critical to effective resource allocation, time intervention, containment efforts, guide healthcare service delivery, and accelerate the development of effective vaccines and treatments.

Challenge:

Given the case fatality ratio (CFR) for a few regions, along with the number of deaths reported in some areas (death data are missing for some locations), and the latitude and longitude for tracking the geographical location of the regions, participants need to build a machine learning model that can forecast the number of deaths, confirmed cases and case fatality ratio for the unreported regions.

Resources and Datasets:

Training data contained the following fields:

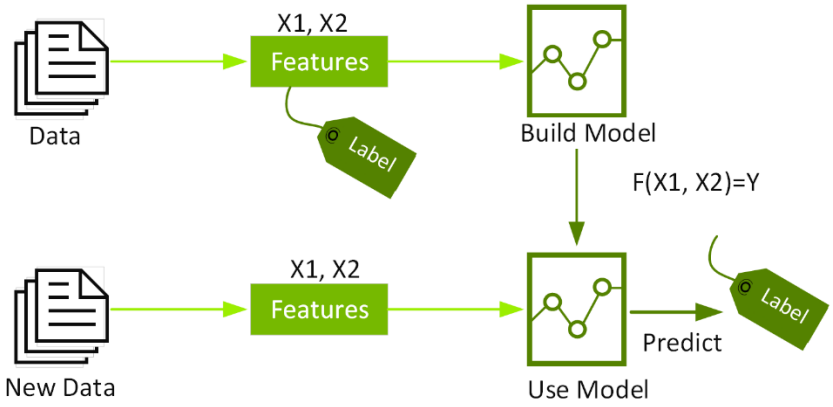
1. Latitude and longitude for geographic locations.
2. Deaths (values for some geographic locations).
3. Case fatality ratio for all geographic locations.

Some other fields that were added to the training and test data were:

1. **Population density**: It was added as a high population density will directly correlate to a higher number of Ebola cases. Ebola virus is directly transmitted and close proximity between people facilitates the transmission of the virus.
    a. The dataset used for population density was "NASA Gridded Population of the World, 2020" with 1km resolution.
2. **Health Security Score**: It analyses the preparedness of a country's health system to manage such disease spread.
    a. Data used was Global Health Security Index 2021 is an indicator of the overall health security conditions of a country.
3. **Temperature**: It was used as an indicator of the environment aspect in the spread of Ebola Virus.
    a. The data used was World Temperature Data above 2m ground level, developed by SOLARGIS and provided by the Global Solar Atlas (GSA)

Overview:

The approach of supervised Machine learning is applied for the training.



The machine learning model applied for training the model was XGBoost.

XGBoost: XGBoost is an implementation of Gradient Boosting and is a type of ensemble learning method. Ensemble learning combines multiple weak models to form a stronger model.

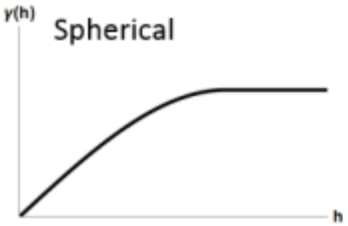
1. XGBoost uses decision trees as its base learners combining them sequentially to improve the model's performance. Each new tree is trained to correct the errors made by the previous tree and this process is called boosting.
2. It has built-in parallel processing to train models on large datasets quickly.

Spherical Kriging for Unknown death Estimation in the training data:

Kriging is an advanced geostatistical procedure that generates an estimated surface from a scattered set of points with z-values.

Kriging the distance or direction between sample points reflects a spatial correlation that can be used to explain variation in the surface. It fits a mathematical function to a specified number of points, or all points within a specified radius, to determine the output value for each location.

Kriging is a multistep process; it includes exploratory statistical analysis of the data, variogram modeling, creating the surface, and exploring a variance surface. Kriging is most appropriate when you know there is a spatially correlated distance or directional bias in the data, as deaths from Ebola should be spatially correlated.



$$\gamma(h) = \begin{cases} c_0 + c\left(\frac{3h}{2a} - \frac{1}{2}\left(\frac{h^3}{a^3}\right)\right) & 0 < h \le a \\ c_0 + c & h > a \\ 0 & h = 0 \end{cases}$$

Data Processing and Cleaning:

The training data was passed through a series of process:

1. The rows with Nan latitude, longitudes were dropped.
2. Unknown deaths were found using spherical kriging.
3. Outliers from the CFR were normalised.
4. Confirmed cases were calculated from CFR.
   CFR = (Deaths/Confirmed Cases) * 100
5. Additional fields like population density, temperature and heath security score were added.
6. Additional fields were filled, outliers were removed and infinite and Null values were removed.

Methodology:

Library Imports

The notebook imports necessary libraries for data processing, machine learning, spatial analysis, and visualization.

Key Libraries & Their Purpose:

1. numpy, pandas → For data manipulation and analysis.
2. matplotlib.pyplot, seaborn → For data visualization.
3. xgboost → For implementing XGBRegressor, a machine learning model.
4. sklearn.model_selection → For splitting the dataset into training and testing sets.
5. sklearn.metrics → For evaluating model performance using metrics like mean_squared_error and r2_score.
6. pykrige.ok → Implements Ordinary Kriging, a geostatistical interpolation technique.
7. rasterio → For handling raster (grid-based) spatial data.
8. folium, HeatMap → For creating interactive maps.
9. reverse_geocode → To convert coordinates into location names.

10. fuzzywuzzy.process → For matching and cleaning location names.

---

2. Loading Datasets

- Reads the dataset(s) using pandas.read_csv().
- Displays the first few rows using .head() to inspect the structure.
- Checks for missing values using .info() and .isnull().sum().

---

3. Data Cleaning and Processing

- Drops missing or unnecessary columns.
- Converts data types where needed.
- Handles outliers or invalid values.

3.1 Preparing Training Data

- Extracts features (X) and target variables (y).
- Splits the dataset into **training and testing sets** using train_test_split().

3.2 Preparing Test Data

- Prepares the test dataset separately, ensuring consistency in feature engineering.

---

4. Adding Population Density, Temperature, and Health Security Data

- Merges external datasets containing **population density, temperature, and health security index**.

- Uses pandas.merge() to combine datasets on a common column (e.g., country or region).
- Fills missing values using .fillna().
- The reason behind adding these features to the data are :

Population Density:

Reason: Population density reflects the number of individuals living in a specific area and can be a critical factor in understanding the spread of diseases, healthcare accessibility, and overall mortality. In densely populated areas, there might be higher transmission rates of contagious diseases, or the healthcare system might become overwhelmed, leading to higher mortality rates.

Impact: Areas with higher population density often experience greater strain on healthcare resources, which can negatively affect outcomes. Thus, including this feature allows the model to account for these variations in the spread of illness or the effectiveness of healthcare infrastructure.

Temperature:

Reason: Temperature can influence the severity of certain diseases (e.g., cold weather can exacerbate respiratory illnesses, while hot weather can affect heart-related conditions) and the general vulnerability of populations to environmental stress. It can also impact the effectiveness of preventive measures or the types of illnesses that are more common in certain climates.

Impact: Temperature influences behavior, exposure to environmental risks, and the spread of diseases. Including it helps to capture seasonal effects or geographic patterns that may contribute to increased mortality, such as heat waves or cold snaps.

Health Security Data:

Reason: Health security data, which can include indicators of healthcare system preparedness, access to medical care, or availability of resources (like hospitals and healthcare workers), is a critical factor in determining survival rates. A stronger healthcare system can lead to lower death rates, especially in emergencies or during disease outbreaks.

Impact: Including health security data ensures the model can assess the quality and accessibility of healthcare services in different regions, helping predict how effectively populations can respond to health crises and thus reducing mortality rates.

5. Estimating Missing Death Values using Spherical Kriging

- Uses **Ordinary Kriging** for spatial interpolation.
- Predicts missing values based on known data points.
- Generates a **spatial prediction grid** for missing deaths.

---

6. Adding Confirmed Cases Column

- Adds **confirmed Ebola case numbers** to the dataset.
- Possibly estimates missing values using Kriging or a statistical method.
- Uses .groupby() or .agg() to organize data by region.

---

7. Training the Model Using Gradient XGBoost for Death Prediction

- Implements **XGBoost Regression** to predict death counts.

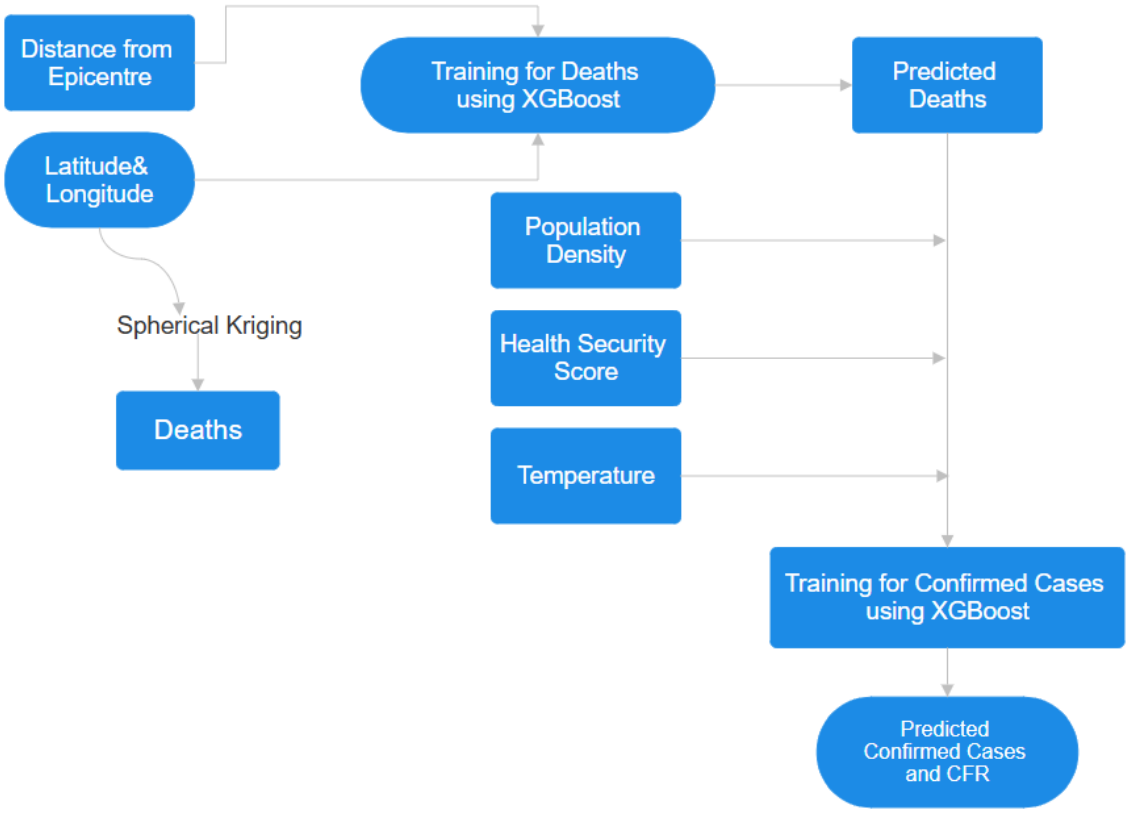
- Evaluates model performance using:
  - **Mean Squared Error (MSE)** to measure accuracy.
  - **R^2 Score** to check how well the model explains variance.

---

8. Training the Model Using XGBoost for Confirmed Cases Prediction

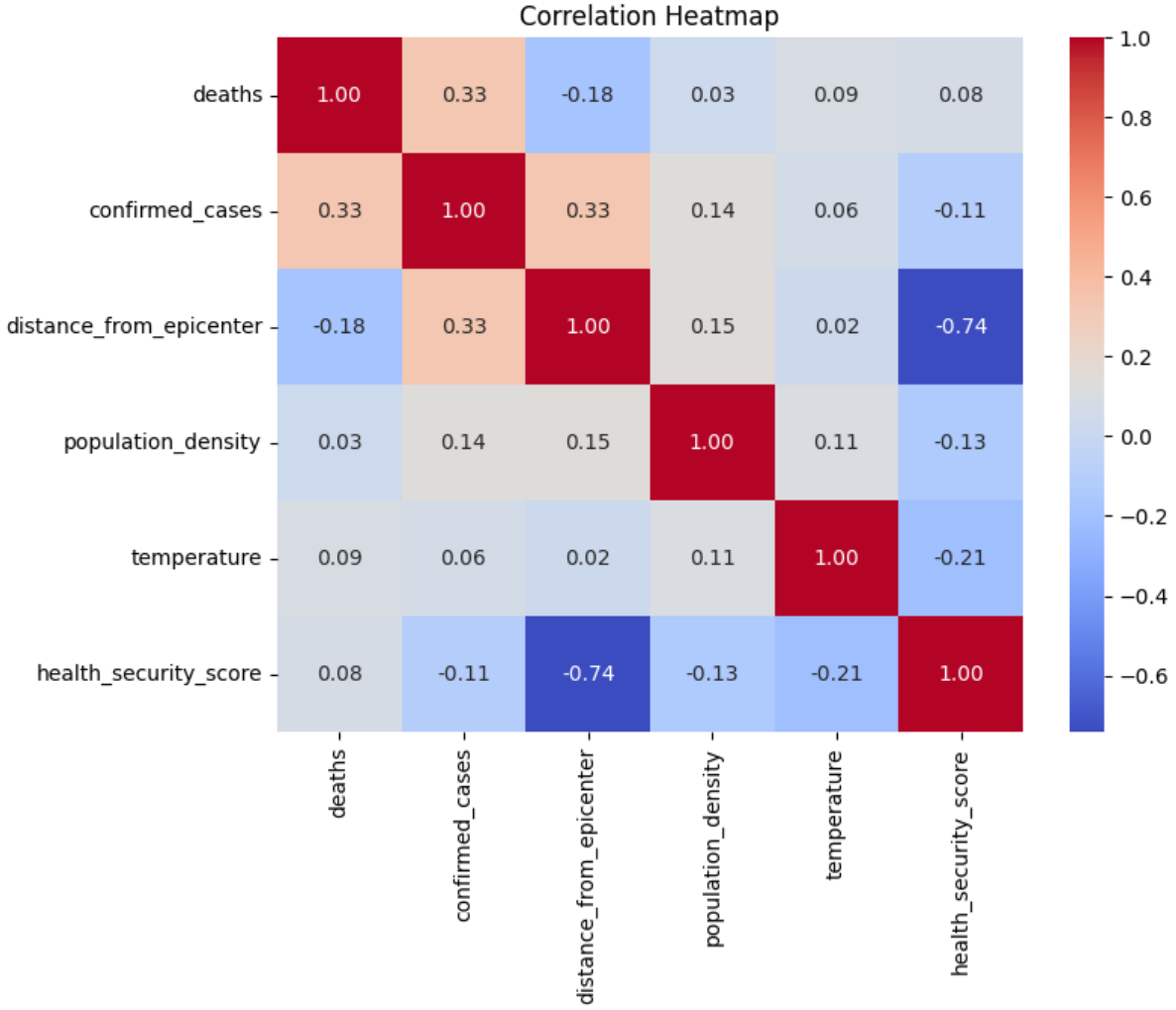
- Trains another **XGBoost model**, this time for predicting confirmed cases.
- Evaluates predictions using the same performance metrics.

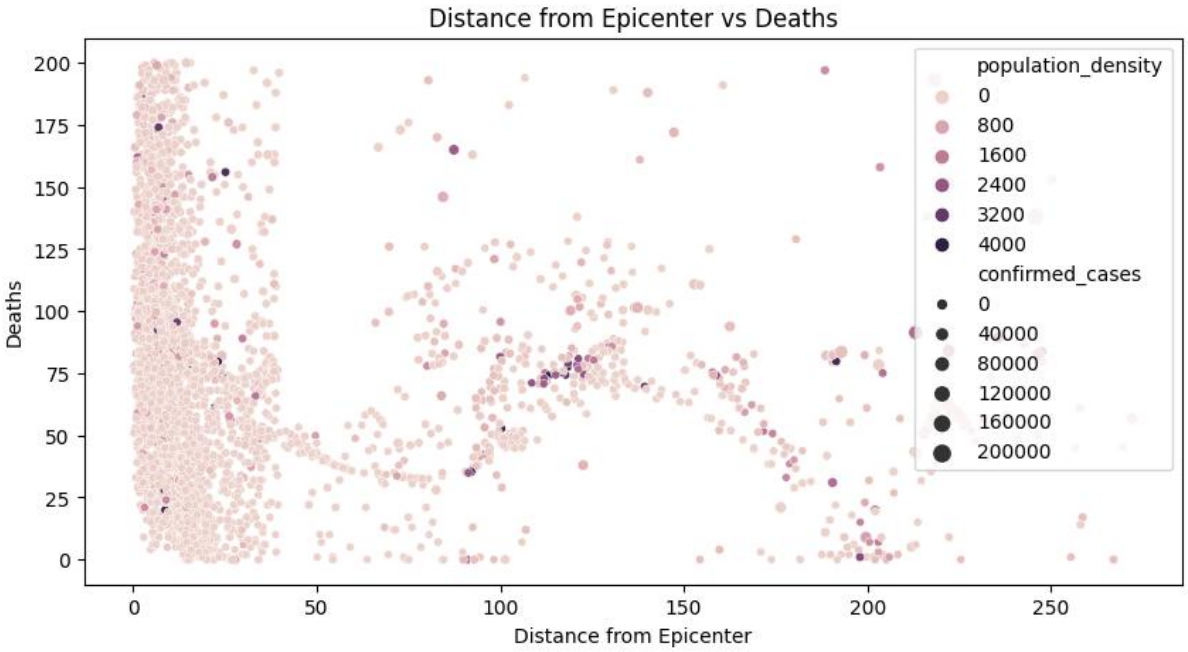
Flowchart:

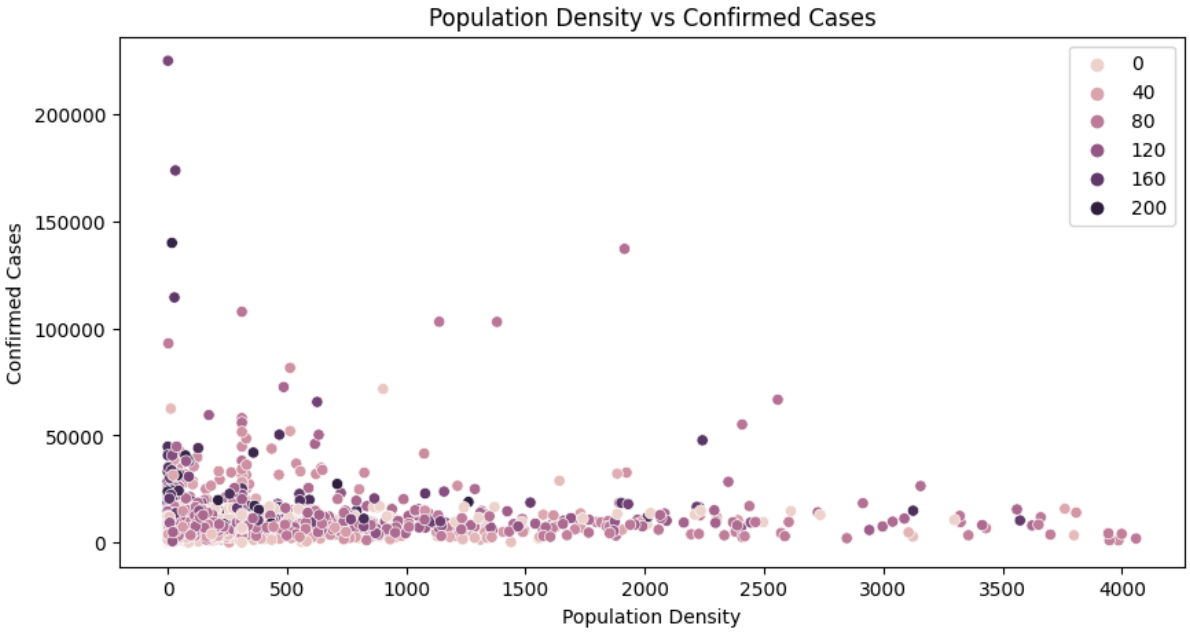


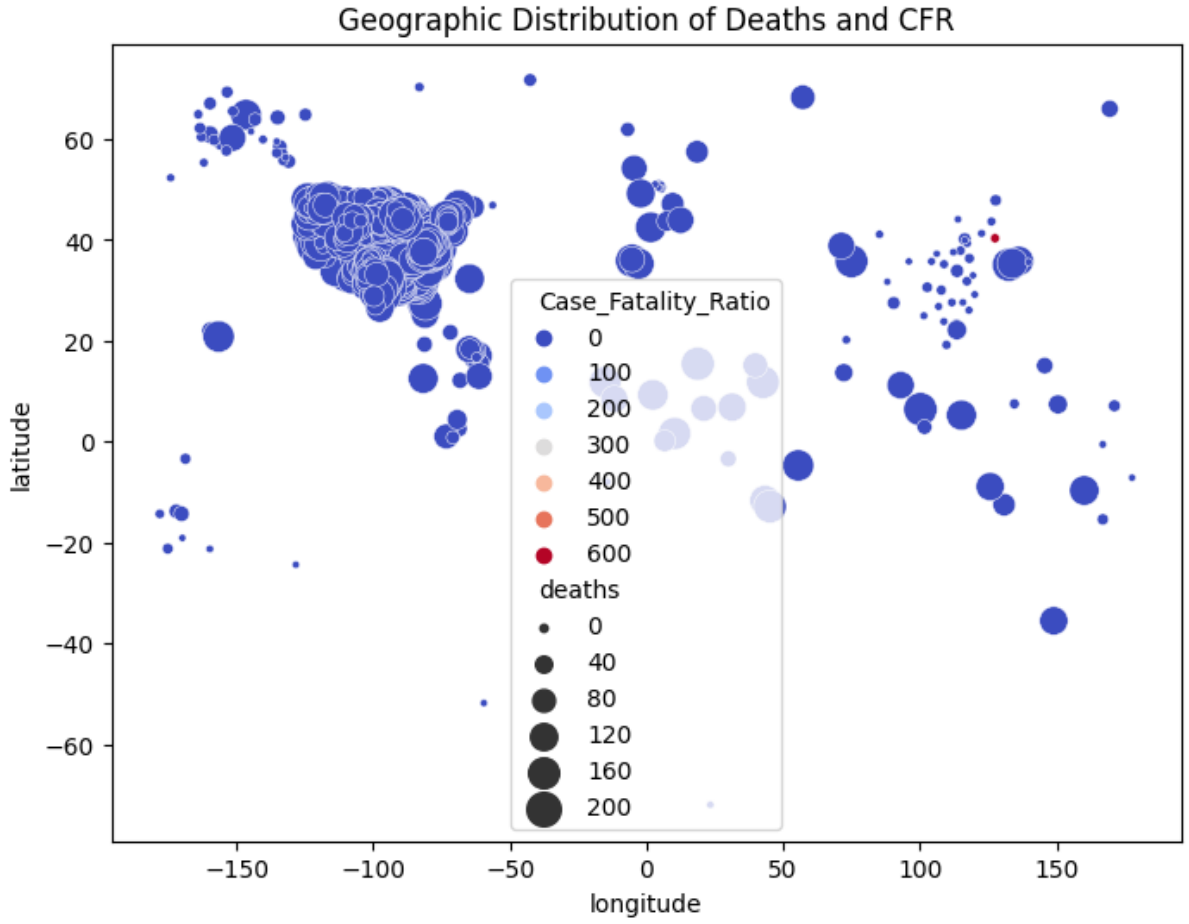
Visualizations:

Correlation Matrix:


Correlation Heatmap
deaths
confirmed_cases
distance_from_epicenter
population_density
temperature
health_security_score
1.00 0.33 -0.18 0.03 0.09 0.08
0.33 1.00 0.33 0.14 0.06 -0.11
-0.18 0.33 1.00 0.15 0.02 -0.74
0.03 0.14 0.15 1.00 0.11 -0.13
0.09 0.06 0.02 0.11 1.00 -0.21
0.08 -0.11 -0.74 -0.13 -0.21 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


Distance from Epicenter vs Deaths
Deaths
Distance from Epicenter
0
25
50
75
100
125
150
175
200
0
50
100
150
200
250
population_density
0
800
1600
2400
3200
4000
confirmed_cases
0
40000
80000
120000
160000
200000

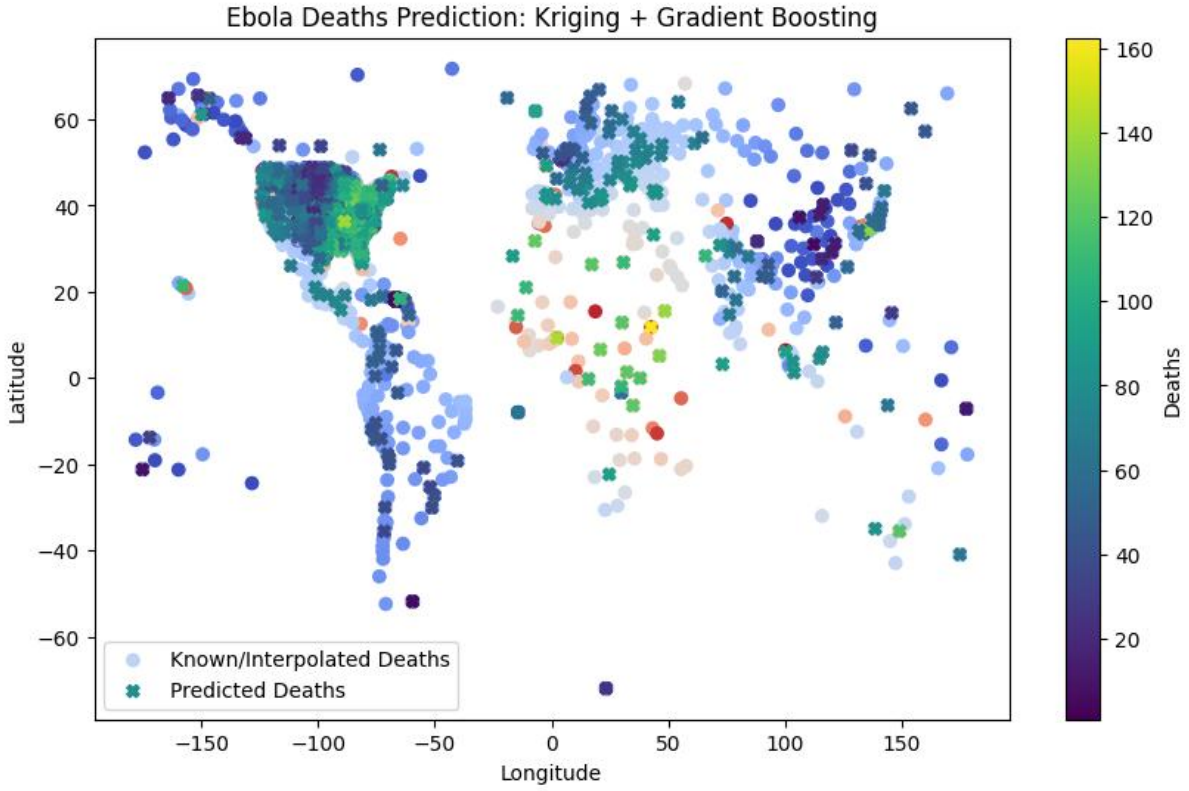

Population Density vs Confirmed Cases
Confirmed Cases
Population Density
0
40
80
120
160
200
0
50000
100000
150000
200000
0
500
1000
1500
2000
2500
3000
3500
4000

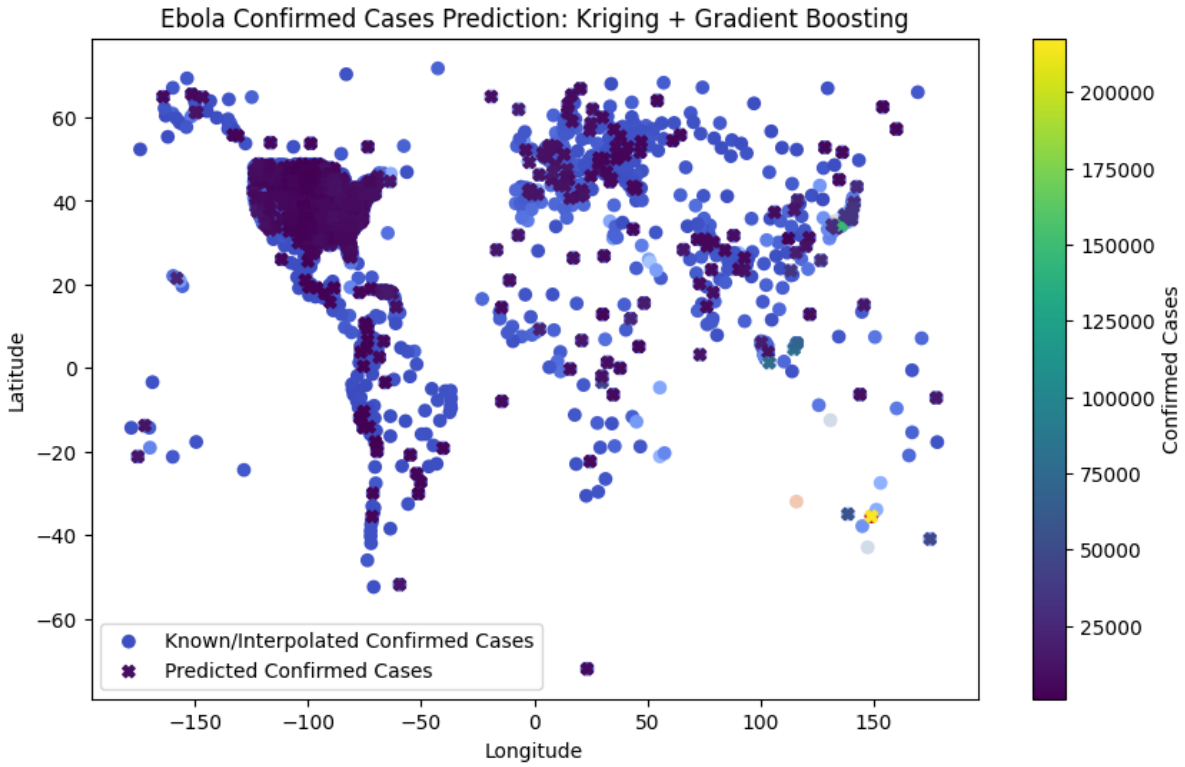

Geographic Distribution of Deaths and CFR
latitude
longitude
Case_Fatality_Ratio
0
100
200
300
400
500
600
deaths
0
40
80
120
160
200
60
40
20
0
−20
−40
−60
−150
−100
−50
0
50
100
150

Results:

The training Model has the following RMSE accuracies:

1. For Deaths: 33.4825
2. For Confirmed Cases: 5240.5937





References:

1. https://ghsindex.org/report-model/

2. https://data.subak.org/dataset/world-air-temperature-2m-above-ground-level-temp-gis-data-global-solar-atlas/resource/cb2fa04d-7d3c-4dbd-91d9-f7a1938be770

3. https://www.earthdata.nasa.gov/data/projects/gpw