**Ebola Data Analysis and Prediction Report**

**Executive Summary**

This report analyzes Ebola outbreak data from 2014-2016, focusing on data preprocessing, exploratory analysis, and predictive modeling. Using a Random Forest Regressor, the analysis developed a highly accurate model to predict cumulative deaths based on country, cumulative cases, new cases, and new deaths. The model achieved an R-squared score of over 99.9% on both training and test data, indicating exceptional predictive performance.

**Introduction**

The Ebola virus disease (EVD) outbreak of 2014-2016 was one of the most severe public health emergencies in recent history. This analysis examines data collected during this period to understand patterns in cases and mortality and develop a predictive model for death counts.

**Data Overview**

The dataset contains records of Ebola cases and deaths across multiple countries from August 2014 to March 2016, with key features including:

* Country name
* Date of reporting
* Cumulative cases
* Cumulative deaths
* New cases
* New deaths
* Case fatality rate (calculated)

**Data Preprocessing**

The following data preparation steps were performed:

1. **Missing Value Analysis**: Initial exploration identified 8 missing values in the 'cumulative no. cases' column.
2. **Handling Missing Values**: Records with missing values were removed to ensure data integrity.
3. **Data Aggregation**: Data was grouped by country and date to eliminate duplicate entries, reducing the dataset from 2485 to 2371 records.
4. **Feature Engineering**:
   * Case Fatality Rate (CFR) was calculated as: (cumulative deaths/cumulative cases) \* 100
   * New cases and new deaths were derived by comparing consecutive records for each country
5. **Data Type Conversion**: Float columns were converted to integers where appropriate, improving processing efficiency.

**Exploratory Data Analysis**

Key statistical findings include:

* **Case Distribution**: Considerable variation in case numbers across countries, with maximum cumulative cases reaching 14,122
* **Mortality Rates**: The overall case fatality rate averaged around 29.85%
* **Data Skewness**: Both case and death distributions show positive skew, with mean values significantly higher than median values, indicating the presence of outliers

**Predictive Modeling**

**Model Development**

* **Target Variable**: Cumulative number of deaths
* **Features Used**: Country, cumulative number of cases, new cases, new deaths
* **Algorithm**: Random Forest Regressor with 300 estimators
* **Feature Encoding**: One-hot encoding was applied to the country feature
* **Training Strategy**: 80% training, 20% testing split

**Model Performance**

The model demonstrated exceptional predictive capability:

* **Training R-squared**: 0.9999 (99.99%)
* **Test R-squared**: 0.9994 (99.94%)
* **Mean Squared Error (Test)**: 1920.999
* **Mean Absolute Error (Test)**: 7.683

**Predictive System**

A functional prediction system was developed that allows users to input:

* Country
* Cumulative cases
* New cases
* New deaths

The system then predicts the expected cumulative deaths based on these inputs.

**Conclusions**

The analysis demonstrates that:

1. Ebola mortality can be predicted with very high accuracy using just a few key variables.
2. The relationship between cases and deaths appears to be highly consistent within countries once patterns are established.
3. The Random Forest model effectively captures the complex relationships between variables across different countries and time periods.

**Technical Implementation**

The model and related components were saved for future use:

* Full Random Forest model (EbolaModelReg.pkl)
* Column transformer for preprocessing (ct.pkl)
* Country encoder for transforming country names (country\_encoder.pkl)

These components enable seamless integration of the predictive system into operational applications for epidemic monitoring and response planning.