A Machine Learning Approach to Predicting COVID-19 Mortality Rates Based on Key Features

Aishah Mustapha

Date: November 2024

Problem Statement

Existing models for predicting COVID-19 mortality are limited by oversimplified assumptions, resulting in reduced accuracy. This project aims to develop a robust machine learning model to improve mortality predictions and support data-driven public health decisions.

Solution Approach

Data Preparation

- Cleaned and standardized global COVID-19 datasets.
- Integrated population data, addressing missing values and outliers.
- Selected key features relevant to predicting COVID-19 mortality.
- Conducted Exploratory Data Analysis (EDA) to uncover trends and relationships.

Model Selection

- Chose XGBoost Regressor for its high accuracy and scalability.

Evaluation Metrics

- Validated model performance using Mean Squared Error (MSE), R^2, and Mean Absolute Error (MAE).

Exploratory Data Analysis (EDA)

- **Key Trends**
  - **Active Cases**: Strongest predictor of mortality trends.
  - **High Death Rates**: Observed in the US, San Marino, and the UK, reflecting overwhelmed healthcare systems.
- **Geographical Insights**
  - **US**: High death toll due to large population and strained resources.
  - **San Marino**: Exceptionally high mortality despite its small size.

- **Implications**
  - Overburdened healthcare systems correlate with higher mortality rates.
  - Tracking **active cases** and **new deaths** is essential for predicting future trends.

Key Features and Their Impact

Selected Features

- **Confirmed Cases**
- **Active Cases**
- **New Deaths**
- **Recovered Cases**

Feature Importance

- Active Cases: 75.03%
- Confirmed Cases: 16.85%
- New Deaths: 2.20%

Insights

- **Active Cases** are the most critical factor in predicting mortality trends.
- Monitoring trends in **Confirmed Cases** and **New Deaths** provides actionable insights.

Results

Model Performance

- **Mean Squared Error (MSE):** 1,114,649.87
- **R-Squared (R^2):** 0.97
- **Mean Absolute Error (MAE):** 162.05

Interpretation

- The model explains **97% of the variance** in COVID-19 deaths.
- Accurate predictions with an average deviation of **162 deaths** from actual values.

What's Next?

1. Model Improvement

- **Hyperparameter tuning** for better accuracy.
- Add **time-series features** to capture trends.

2. Broader Validation

- Test on **additional regions** for generalization.

3. Deployment

- Create a **real-time dashboard** for decision-makers.

4. Future Research

- Explore the impact of **vaccination** and **healthcare capacity**.

Conclusion & Recommendations

Conclusion

- **XGBoost Regressor** effectively predicted COVID-19 mortality rates with high accuracy.
- **Key Features**: Active cases, confirmed cases, and new deaths were critical in mortality predictions.
- **Performance**: Achieved high R^2 (0.97) and low MSE (1,114,649.87).

Recommendations

- **Public Health**: Use model insights for resource allocation and policy formulation.
- **Further Research**: Explore additional variables like vaccination data and healthcare infrastructure.
- **Model Expansion**: Include more regions and real-time data for better generalization.