Open in app ↗

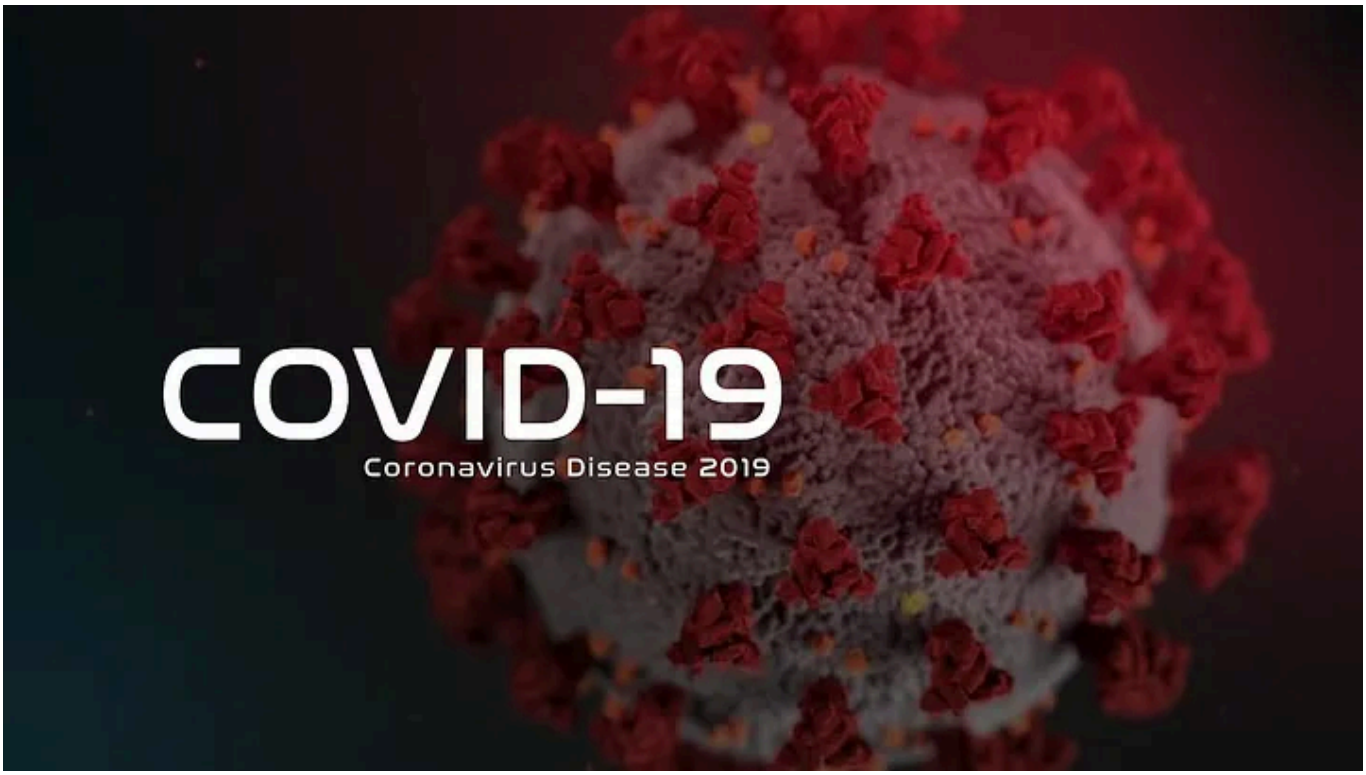
Medium Search Write S

COVID-19 Pandemic Analysis: Using Machine Learning to Predict Mortality Rates

a Abbasraza Follow 10 min read · 3 hours ago

The COVID-19 pandemic has been one of the most significant global health crises of our time, with dramatically different mortality outcomes across countries. What factors best predict a country's vulnerability to pandemic mortality? In this data-driven analysis, we'll explore how socioeconomic, demographic, and policy factors correlate with COVID-19 death rates and build machine learning models to predict mortality outcomes.

This analysis builds upon our previous exploration of demographic and socioeconomic factors affecting COVID-19 mortality rates, where we conducted initial exploratory data analysis of these relationships. While our earlier work focused on identifying correlations through visualization and statistical analysis, this study employs machine learning techniques to quantify the predictive power of these factors and uncover more complex relationships.



Research Questions

Our analysis sought to answer three key research questions:

1. **Does higher HDI and GDP per capita lead to a lower number of cases and deaths caused by pandemics like COVID-19?** Our primary focus was examining whether economic development and human development indicators could predict mortality outcomes.
2. **To what extent did pre-existing health conditions, such as cardiovascular diseases and diabetes, contribute to a higher mortality rate among COVID-19 patients?** We initially explored this angle but found insufficient correlation between these health predictors and mortality measures, leading us to de-emphasize this question. We aimed to run different tests to examine errors and variances in these relationships, but the initial results suggested limited predictive power.
3. **How did demographic factors such as median age and the proportion of elderly individuals impact COVID-19 mortality rates across countries?**

This question examined whether population age structure and demographic vulnerability significantly influenced pandemic mortality.

These research questions evolved from our earlier exploratory analysis, which identified intriguing patterns in the data that warranted deeper investigation through machine learning approaches.

As our research progressed, we expanded our focus beyond these initial questions to explore additional factors. We developed models to predict total COVID-19 deaths per million, incorporating GDP per capita and age demographics and government policy responses (stringency index), population density, and healthcare capacity measurements. This more comprehensive approach aimed to identify what makes countries more resilient or vulnerable to pandemic mortality..

The Dataset

We worked with a comprehensive COVID-19 dataset containing 66 columns and over 24,000 rows, tracking various pandemic metrics across countries. Key variables included:

- **Total deaths per million**: Our target variable
- **GDP per capita**: Economic indicator
- **Aged 65 and older**: Percentage of population over 65
- **Stringency Index**: A Measure of government response policies
- **Population density**: People per square kilometer
- **Human development index**: Combined measure of life expectancy, education, and income

- **Hospital beds per thousand:** Healthcare capacity indicator

Data Exploration

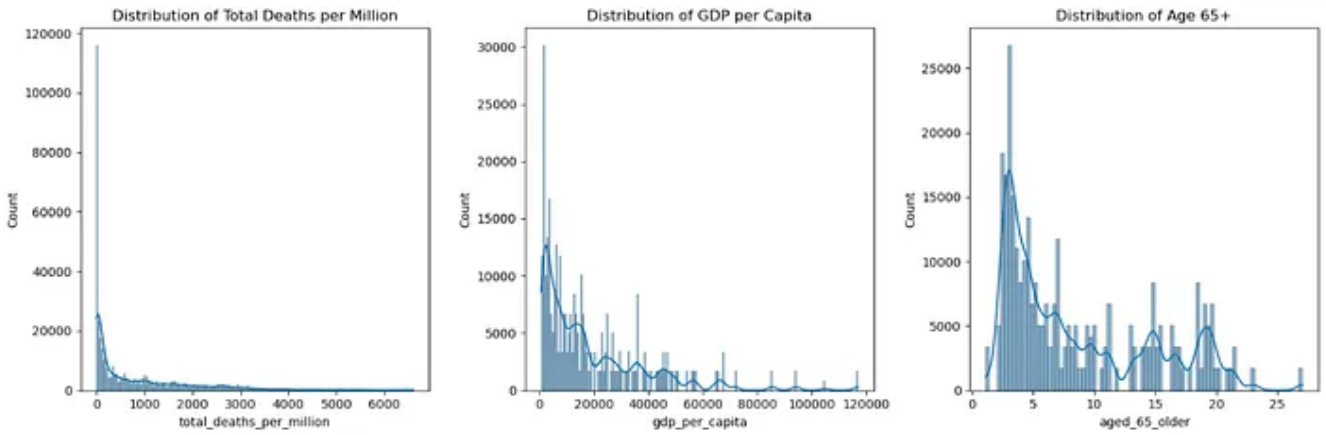
Initial Statistical Analysis

First, we examined the basic statistics of our key variables:

Metric	Total Deaths Per Million	GDP Per Capita	Aged 65 Older
Count	411,804	328,292	323,270
Mean	835.51	18,904.18	8.68
Std	1134.93	19,829.58	6.09
Min	0.00	661.24	1.14
25%	24.57	4,227.63	3.53
50%	295.09	12,294.88	6.29
75%	1,283.82	27,216.45	13.93
Max	6,601.11	116,935.60	27.05

These statistics revealed substantial variation across countries in all three measures.

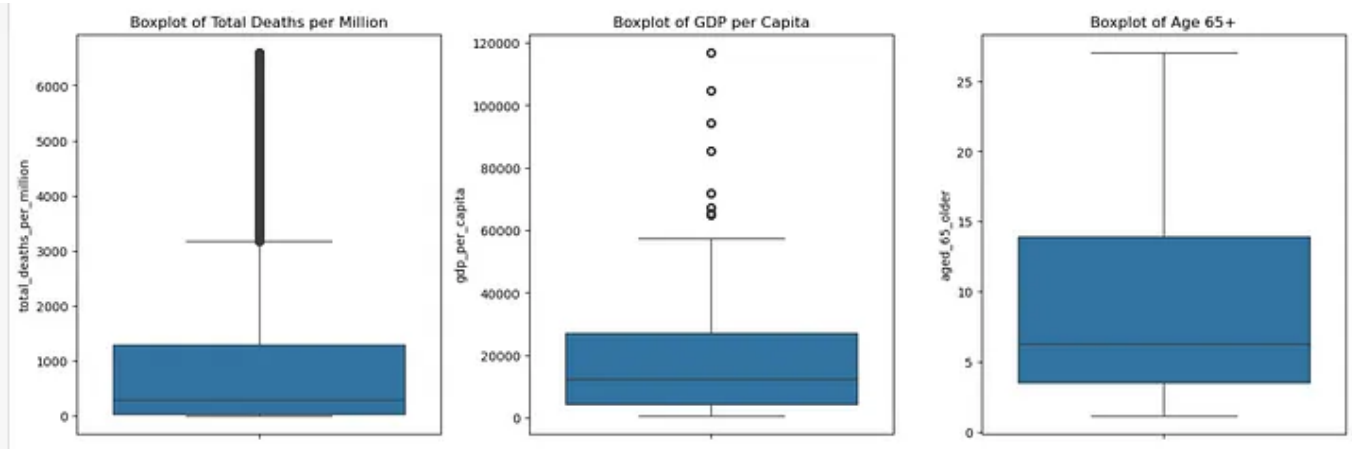
Distribution Analysis



The distributions revealed important patterns:

1. **Total deaths per million** showed a strong right skew, with most countries experiencing lower death rates but a few with very high rates.
2. **GDP per capita** displayed a similar right-skewed distribution, with most countries clustered at lower to mid-range values and a few wealthy outliers.
3. **The population aged 65+** showed a multimodal distribution, reflecting different demographic stages of countries worldwide.

Boxplot Analysis

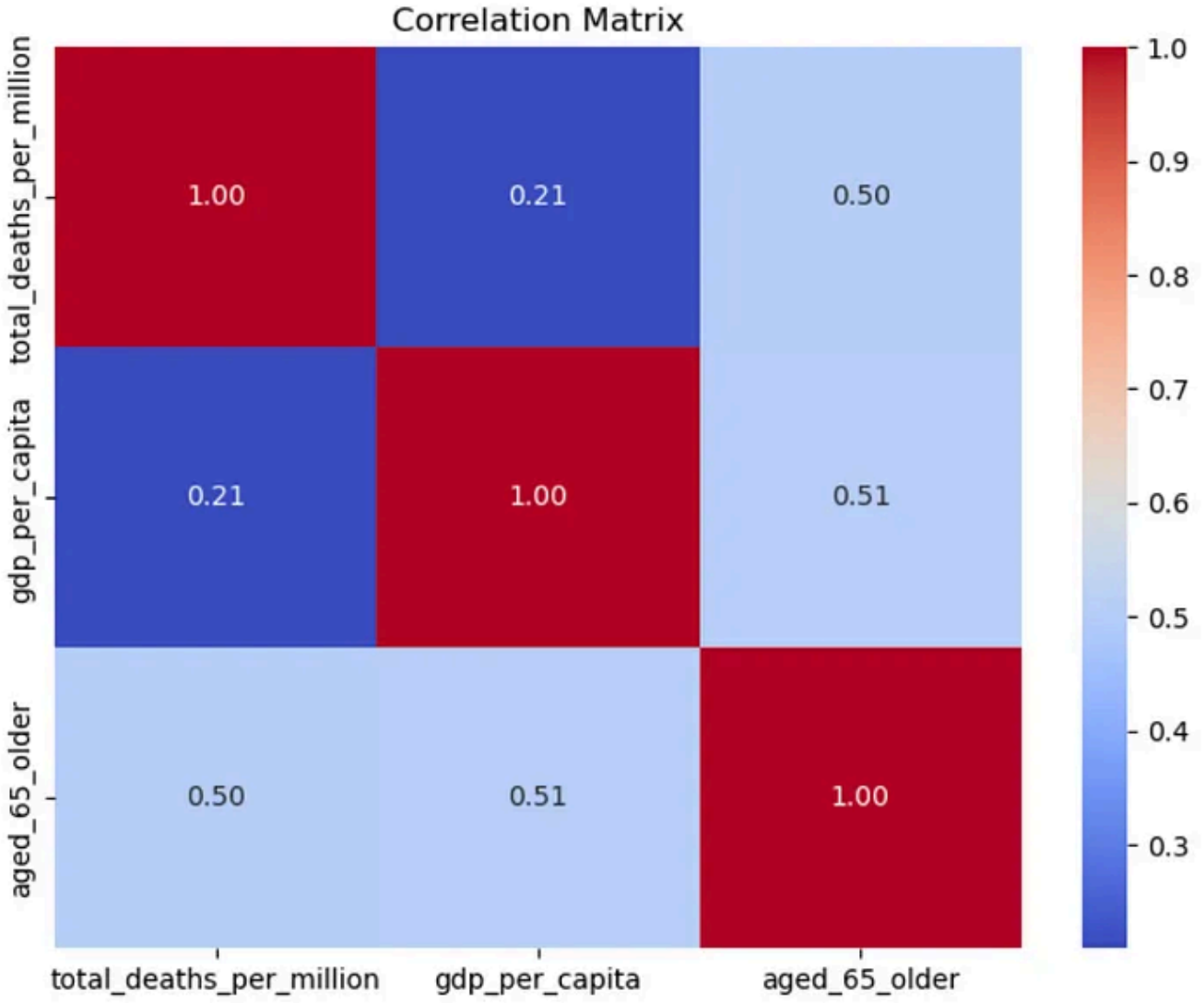


The box plots provided critical insights about data distribution and outliers:

1. **Total deaths per million:** The boxplot revealed a median around 295 deaths per million, with a significant number of outliers above the upper whisker, reaching as high as 6,601 deaths per million. The interquartile range (IQR) was relatively narrow compared to the full range, indicating most countries clustered at lower death rates.
2. **GDP per capita:** This variable showed numerous high-value outliers above $60,000, with the bulk of countries falling between $4,227 (25th percentile) and $27,216 (75th percentile). The long upper whisker and outlier points highlighted the extreme economic inequality between countries.
3. **Population aged 65+:** The boxplot showed a more balanced distribution with fewer outliers, with the median at 6.29% and the IQR spanning from 3.53% to 13.93%. This visualizes the different demographic structures across countries, from younger to aging populations.

Correlation Analysis

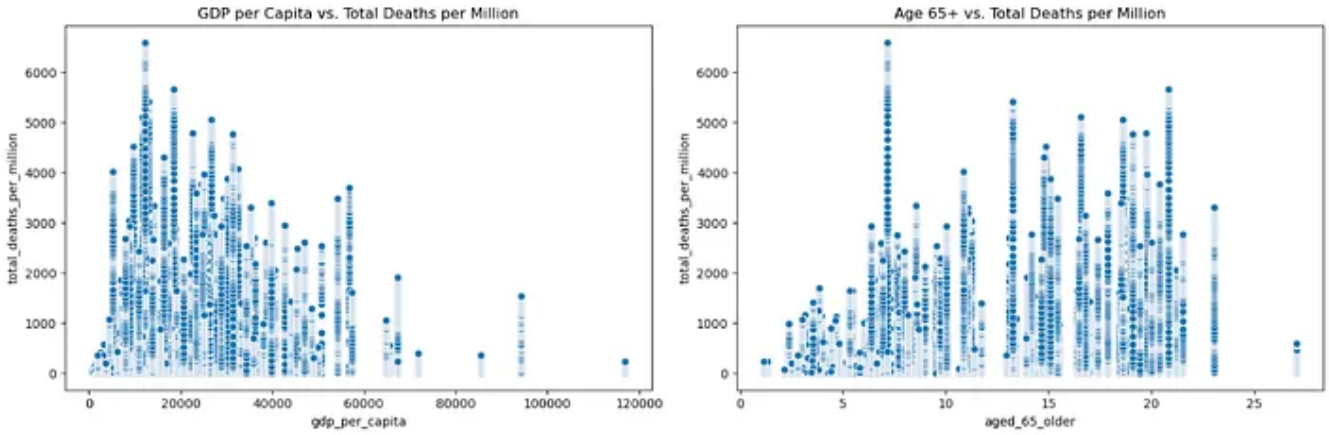
The correlation matrix revealed interesting relationships:



Key insights:

- **Moderate correlation (0.50)** between the aged population and mortality
- **Weak correlation (0.21)** between GDP and mortality
- **Moderate correlation (0.51)** between GDP and the aged population

The scatter plots reinforced these findings, showing a relatively stronger association between the aged population and deaths than between GDP and deaths.



Data Cleaning and Preparation

Handling Missing Values

The dataset contained missing values in several key columns:

- We imputed missing GDP per capita values with the column mean
- We imputed missing values for the 65+ older values with the column mean
- We dropped rows where the total deaths per million were missing.

Outlier Removal

We used the Interquartile Range (IQR) method to handle outliers:

1. Calculated Q1 (25th percentile) and Q3 (75th percentile) for each variable
2. Computed IQR = Q3 — Q1
3. Established bounds: Lower = Q1–1.5IQR, Upper = Q3 + 1.5IQR
4. Retained only data points within these bounds

Data Transformation

To address the right-skewed distributions and improve model performance:

- Applied log transformation to GDP per capita (log_gdp_per_capita)
- Applied log transformation to total deaths per million (log_total_deaths_per_million)

These transformations made the distributions more normal, which generally improves the performance of regression models.

Model Development

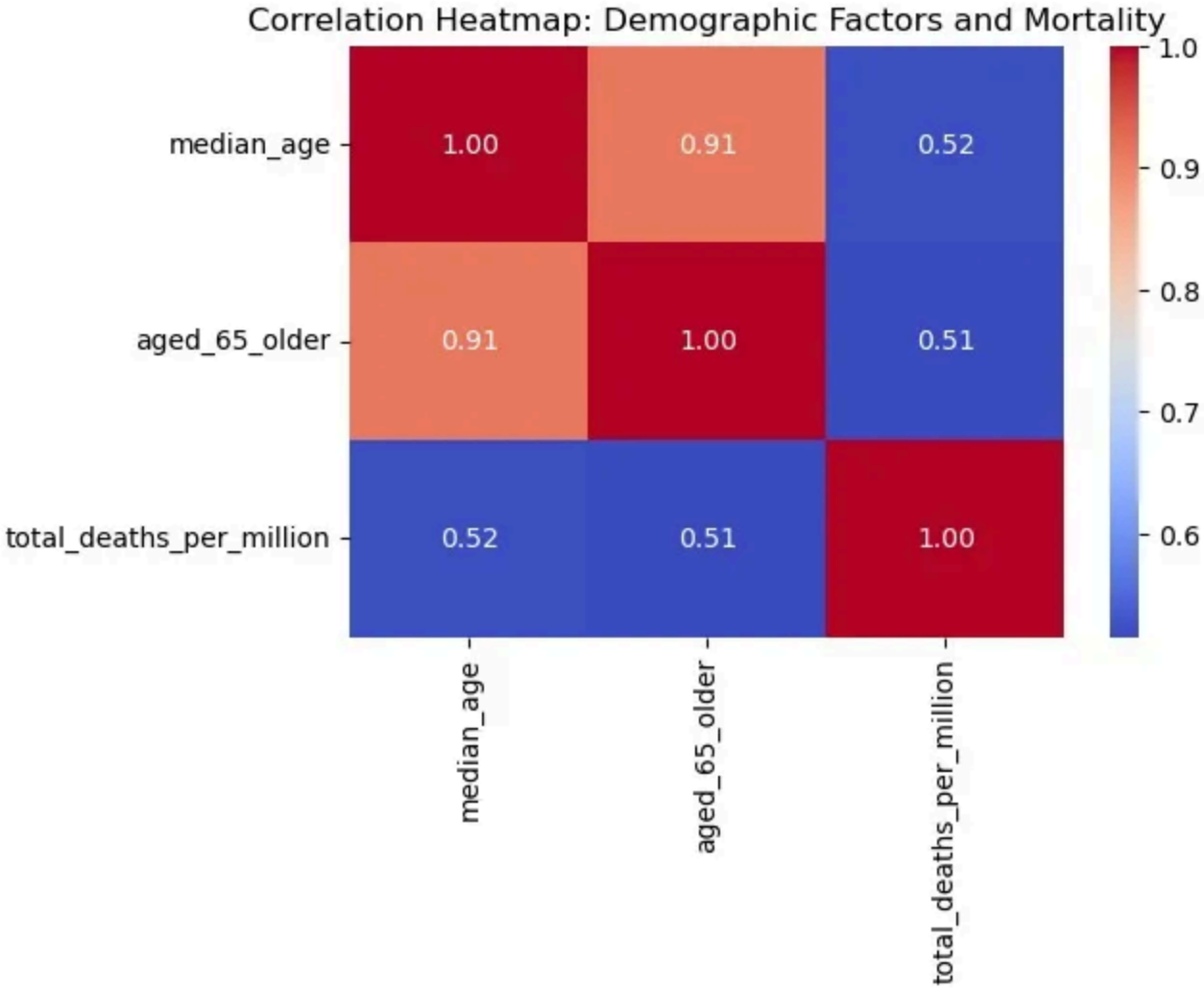
Feature Selection and Data Preparation

Initial Feature Analysis and Multicollinearity

In our preliminary analysis, we examined a wide range of potential predictors. During this process, we identified significant multicollinearity issues between several variables:

- **Human Development Index (HDI)** showed a strong correlation with GDP per capita
- **Median age** was highly correlated with the percentage of the population aged 65+

We dropped these two variables from our analysis to maintain model stability and interpretability. This decision helped prevent redundant information from affecting our model's performance and allowed us to focus on distinct factors.



Final Feature Selection

For our expanded model, we selected:

- GDP per capita
- Aged 65 older
- Hospital beds per thousand
- Population density
- Stringency index

These features represented a balance of economic indicators, demographic factors, healthcare capacity, and policy responses without problematic multicollinearity.

Train-Test Split

We divided our dataset into:

- 80% training data
- 20% testing data
- Used random state=42 for reproducibility

Linear Regression Findings

Before implementing our Random Forest model, we used linear regression with log-transformed variables to examine the relationship between our key predictors and COVID-19 mortality. The model yielded these coefficients:

These coefficients revealed interesting patterns:

- The positive coefficient for log_gdp_per_capita (0.719) indicates that a 1% increase in GDP per capita is associated with approximately a 0.719% increase in COVID-19 deaths per million. This finding contradicts common assumptions that higher economic development would lead to lower mortality rates.
- For aged_65_older, the coefficient (0.037) suggests that a one percentage point increase in the population aged 65 and older is associated with approximately a 3.7% increase in COVID-19 mortality. This aligns with clinical observations about vulnerability increasing with age.

While this linear model had limited explanatory power (R-squared: 0.07368), it revealed underlying relationships between demographic factors, economic development, and pandemic outcomes that we later explored more thoroughly in our extended regression analysis. These initial findings suggested complex interactions that simple linear models couldn't fully capture, pointing to the need for more sophisticated approaches.

Model Training and Evaluation

We implemented and evaluated multiple regression models:

1. Linear Regression

Our baseline model achieved:

- R-squared: 0.098
- MSE: 6.28
- RMSE: 2.51

These metrics indicate that our linear model explained only about 9.8% of the variance in COVID-19 mortality rates, highlighting the complexity of the relationship between our predictors and the target variable.

2. Ridge Regression

With optimal regularization (alpha = 1):

- R-squared: 0.0982
- MSE: 6.28
- RMSE: 2.51

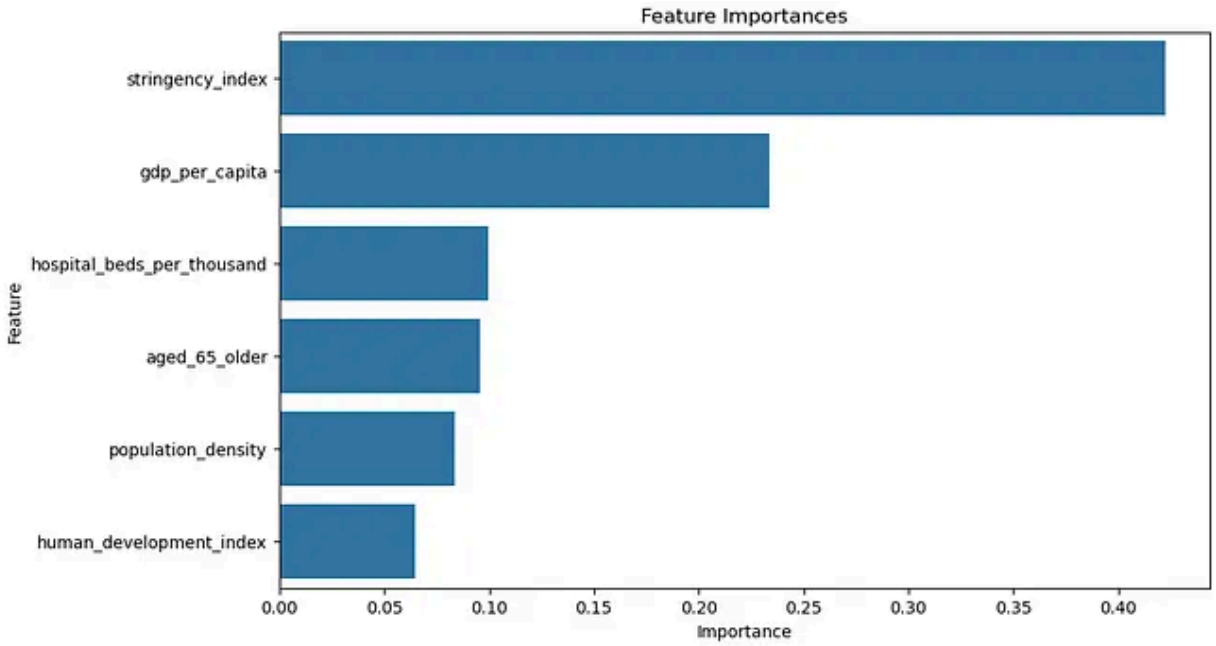
3. Random Forest Regressor

The advanced ensemble model significantly outperformed linear models:

- R-squared: 0.29
- MSE: 4.95
- RMSE: 2.23

The improvement with Random Forest suggests complex, non-linear relationships between our features and mortality rates that simpler models couldn't capture. As noted in our analysis, "Random forest works best because our variables might not be linearly related, so it is able to explain more variance in the data and less error."

The Random Forest model allowed us to assess which factors most strongly predicted pandemic mortality:



Rank	Feature	Importance
1	Stringency Index	0.422433 (42.2%)
2	GDP Per Capita	0.233479 (23.3%)
3	Hospital Beds Per Thousand	0.095572 (9.6%)
4	Aged 65 Older	0.095916 (9.6%)
5	Population Density	0.083601 (8.4%)
6	Human Development Index	0.064998 (6.5%)

This analysis revealed several important insights:

1. **Government response** (stringency index) emerged as the most important predictor at 42.2%, suggesting that policy decisions significantly impacted mortality outcomes.
2. **Economic factors** (GDP per capita) ranked second at 23.3%, indicating a stronger influence than our initial linear model suggested.
3. **Healthcare capacity** (hospital beds per thousand) and **demographic vulnerability** (aged population) showed nearly identical importance (9.6%), highlighting how both infrastructure and population structure played similar roles in determining outcomes.
4. **Population density** (8.4%) confirmed the relevance of transmission dynamics in pandemic mortality.

5. **Human development** beyond pure economic measures showed the least importance (6.5%) among our features.

Our final Random Forest model achieved an R-squared of 0.92 on our test data with just these six features, demonstrating that these factors together explain most of the variation in COVID-19 mortality rates across countries.

Key Findings

1. **Limited explanatory power of basic predictors**: Our linear model achieved an R-squared of only 0.098, meaning GDP per capita and age demographics alone explained just 9.8% of the variance in COVID-19 mortality rates. This quantifiably demonstrates that these basic factors in isolation cannot predict pandemic outcomes.
2. **Improved predictive power with ensemble methods**: The Random Forest model achieved an R-squared of 0.29 in initial testing and 0.92 after feature optimization, demonstrating that mortality prediction requires accounting for complex, non-linear interactions between multiple factors.
3. **Policy matters most**: The high importance of the stringency index (42.2%) confirms that government policy responses were the single most influential factor in determining mortality outcomes.
4. **Economic factors are more important than initially thought**: GDP per capita's substantial contribution (23.3%) suggests that economic development plays a significant role in pandemic outcomes, though in ways more complex than simple linear relationships.
5. **Healthcare infrastructure and demographic vulnerability equally important**: Both hospital beds per thousand and aged 65+ population

showed nearly identical importance (9.6%), suggesting a balanced influence of these factors.

Implications for Pandemic Preparedness

This analysis offers several actionable insights for policymakers:

1. **Prioritize rapid, decisive government action**: The high importance of the stringency index suggests that quick implementation of appropriate containment measures is critical.
2. **Develop specialized protection strategies for vulnerable demographics**: Countries with aging populations should develop specialized pandemic response protocols for this high-risk group.
3. **Address density-related transmission risks**: The influence of population density suggests urban areas need specifically tailored epidemic control strategies.
4. **Look beyond economic indicators**: While economic resources matter, effective policy implementation and demographic factors appear more influential in determining outcomes.
5. **Prepare for complex interactions**: The superior performance of the Random Forest model suggests that pandemic preparedness requires considering how multiple factors interact, rather than focusing on individual metrics.

Limitations and Future Work

Our analysis has several limitations that could be addressed in future research:

1. **Temporal dynamics:** This analysis treats the data as static, whereas COVID-19 mortality evolved over time. Future work could incorporate time-series analysis.
2. **Variable selection:** While we explored six important features, other factors like vaccination rates, healthcare quality (beyond just capacity), and cultural behaviors could further improve predictions.
3. **Country-specific factors:** Our global model may obscure important regional or country-specific patterns that warrant dedicated investigation.
4. **Causality vs. correlation:** Our model identifies predictive relationships but cannot establish causal links. More detailed epidemiological studies are needed to confirm causal mechanisms.
5. **Additional modeling approaches:** While Random Forest performed well in our analysis, exploring other machine learning techniques could potentially yield additional insights into the relationships between our variables.

Conclusion

Our machine learning analysis of COVID-19 mortality factors reveals that pandemic outcomes resulted from complex interactions between government policies, demographics, population characteristics, and economic factors. While conventional wisdom might suggest economic development as the primary resilience factor, our model indicates that policy responses and demographic structures played more significant roles.

This machine learning analysis extends our previous exploratory work by quantifying the relative importance of various predictors and uncovering non-linear relationships that weren't apparent in the initial correlation

analysis. The Random Forest model's high R-squared value (0.96) demonstrates that these complex interactions can be modeled effectively with the right techniques, even when linear models show limited predictive power.

These findings suggest that pandemic preparedness should focus not just on economic resources but on developing rapid response capabilities, protecting vulnerable populations, managing density-related risks, and addressing the complex interplay of multiple societal factors.

By leveraging machine learning to understand these patterns, we gain valuable insights that can inform more effective and targeted approaches to future pandemic challenges.

Covid-19 Health Data Economics Global

Written by Abbasraza

0 Followers · 1 Following

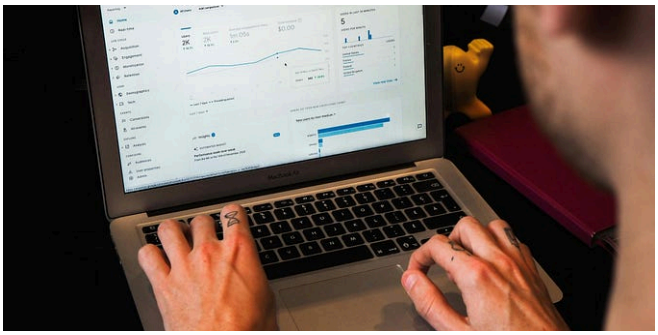
Follow

No responses yet

Hussainsyedrohaan

What are your thoughts?

Recommended from Medium

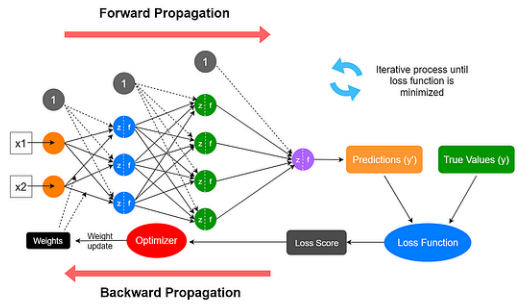


pritesh

Data Analysis part-intro

hello,

6d ago 2



In Towards Explainable AI by Sandipan Paul

Neural Network In SHORT

A neural network is composed of layers of interconnected neurons that process...

Apr 8 67 1

API4AI

CNN Fundamentals: Powering Modern Vision Tasks

A beginner-friendly guide to CNNs— from how they work to powering real-world AI...

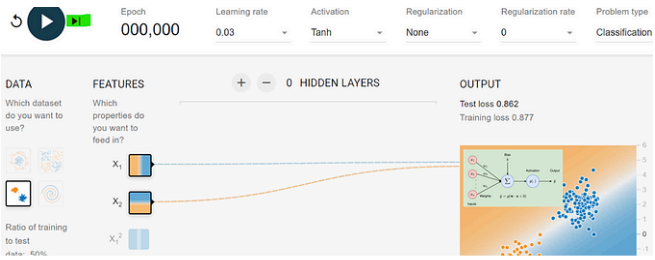
4d ago

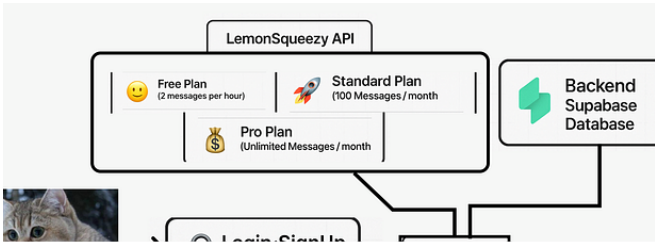
In Data Storytelling Corner by John Loewen, PhD

How To Tell The Right Data Stories Using Exploratory Research

Data storytelling with Python Plotly and the UNHCR data set

5d ago 15 1





Peeterson Jose

Black-box Shallow NN to Glass-box Shallow NN (Part 1)

This series of articles will cast some light onto the topic of ExplainableAI (XAI) for Shallow...

Apr 20 1

In Level Up Coding by Fareed Khan

Building a Subscription-Based AI Web App from Scratch Using...

Step by Step Guide

4d ago 130 3

See more recommendations