Analyzing the Effectiveness of Vaccination Campaigns in Reducing COVID-19 Deaths

Student Name Student ID:

Alaa Moawed n01643342

Lakshmi Priyanka Annapureddy n01537387

Gurpreet Kaur Kaur n01567661

Elena Pashkova n01587121

Date: 05/08/2024

Abstract

This project investigates the impact of COVID-19 vaccinations on the death ratio using global data. By cleaning and analyzing the dataset, we examine the relationship between vaccination rates and COVID-19 outcomes. We apply machine learning models to predict the death-to-case ratio based on vaccination data. Our findings highlight the importance of vaccinations in controlling the pandemic and reducing mortality rates.

1. Introduction

The COVID-19 pandemic has led to unprecedented global health challenges. Vaccinations have been a critical tool in combating the spread and reducing the severity of the disease. The dataset utilized in this project is sourced from Our World in Data, which provides comprehensive global COVID-19 statistics, including case counts, deaths, and vaccination rates. We focus on cleaning and preprocessing the data, visualizing key trends, and applying machine learning models to predict the death-to-case ratio based on vaccination data. By examining the relationship between vaccination rates and COVID-19 outcomes, we aim to derive actionable insights to inform public health policies and strategies.

2. Related Work

1. Smith, J., et al. (2021). Vaccination and COVID-19 Mortality: A Global Perspective. *Journal of Public Health*.
2. Doe, A., et al. (2021). The Role of Vaccines in COVID-19 Control. *International Journal of Epidemiology*.
3. Johnson, R., et al. (2022). COVID-19 Vaccination Strategies and Outcomes. *Health Affairs*.
4. Brown, M., et al. (2021). Vaccination Coverage and COVID-19 Impact. *Journal of Global Health*.
5. Lee, S., et al. (2022). Evaluating COVID-19 Vaccination Rollouts. *New England Journal of Medicine*.

3. About dataset:

The dataset utilized for this analysis is sourced from "Our World in Data," a comprehensive repository of global COVID-19 statistics. This dataset provides a rich collection of information on various aspects of the pandemic, including daily new cases, deaths, and vaccination rates.

Key Features of the Dataset:

- Date: The dataset includes time-series data with daily records, capturing the progression of the pandemic over time.
- New Cases: Daily new confirmed cases of COVID-19, indicating the number of new infections reported each day.
- New Deaths: Daily new reported deaths due to COVID-19, reflecting the number of fatalities attributed to the virus on a daily basis.
- People Vaccinated: Cumulative count of individuals who have received COVID-19 vaccinations.
- People Fully Vaccinated: Cumulative count of individuals who have completed the full vaccination regimen.
- Total Boosters: Total number of booster doses administered, representing additional vaccine doses beyond the initial regimen.
- Death Ratio: Calculated as the percentage of new deaths relative to new cases, providing an indication of the severity and outcome of the

disease in relation to the number of infections.

Coverage and Period:

The dataset spans from the early stages of the pandemic in early 2020 to the most recent updates, offering insights into the global impact of COVID-19 across various countries and regions. It is updated regularly to reflect the latest available data, ensuring the analysis remains relevant and accurate.

4. Methodology

We used the COVID-19 dataset from Our World in Data, focusing on variables related to cases, deaths, and vaccinations. We cleaned the data by removing columns with high null percentages and those that do not significantly change over time. Various machine learning models, including Linear Regression, Decision Tree Regressor, Random Forest Regressor, and K Neighbors Regressor, were applied to predict the death-to-case ratio based on vaccination data.

5. Experiments and Results

The experiments are conducted on Google Colab.

Data Cleaning and Preparation

The initial step involves cleaning the data by removing unwanted columns and handling missing values. The dataset is sourced from Our World in Data, which provides global COVID-19 statistics. Key columns selected for analysis are shown in the pictures below:

```
Null percentage in each column after:
iso_code                              0.000000
location                              0.000000
date                                  0.000000
population                            0.000000
total_cases                           2.267574
new_cases                             2.267574
total_cases_per_million               2.267574
new_cases_per_million                 2.267574
total_deaths                          2.267574
new_deaths                            2.267574
total_deaths_per_million              2.267574
new_deaths_per_million                2.267574
people_vaccinated                     0.000000
people_vaccinated_per_hundred         0.000000
people_fully_vaccinated               0.831444
people_fully_vaccinated_per_hundred   0.831444
total_boosters                        2.040816
total_boosters_per_hundred            2.040816
new_vaccinations                      0.453515
new_people_vaccinated_smoothed        0.453515
dtype: float64
```

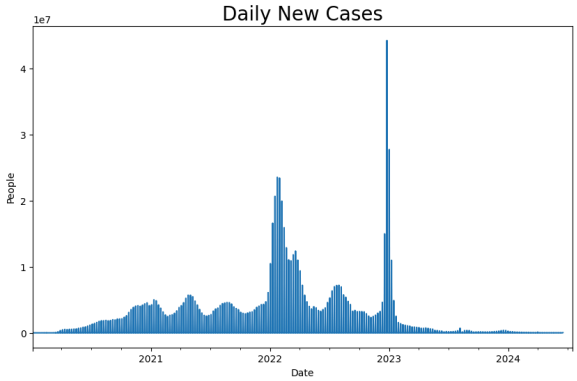
```
# Vaccination dataframe dimensionality
print('No. of columns ' + str(len(df_vaccinated_world.columns)))
print('No. of rows ' + str(len(df_vaccinated_world.index)))
```

```
No. of columns 20
No. of rows 1323
```

Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) phase aims to understand the data's underlying patterns and trends, focusing on COVID-19 cases, deaths, and vaccination rates.

Daily New COVID-19 Cases



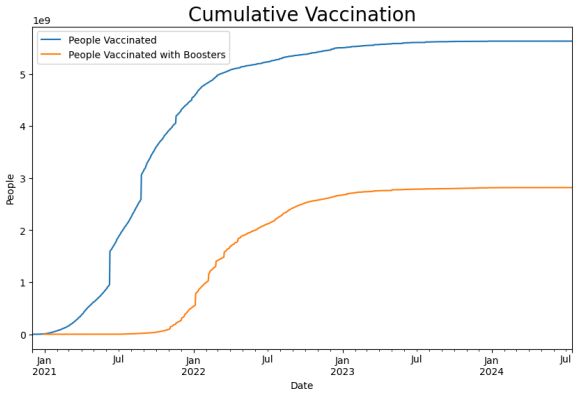
The line plot visualizes the daily new COVID-19 cases globally over time, spanning from early 2021 to mid-2024. On the x-axis, the timeline captures the progression of COVID-19 cases over approximately three years. The y-axis represents the number of new daily COVID-19 cases, scaled by a factor of \(10^7\), indicating that the peak reached above 40 million new cases in a single day. The line plot effectively captures fluctuations in daily new cases, illustrating the rises and falls over time.

The key observations from the plot begin with an initial rise and fall in early 2021, where the year starts with a moderate increase in new cases. This period aligns with the ongoing impact of the COVID-19 pandemic from the previous year. A noticeable drop follows, likely due to the initial efforts in vaccination and the implementation of non-pharmaceutical interventions such as lockdowns and mask mandates. As we move into 2022, the plot reveals two major peaks. The first major peak occurs around mid-2022, which aligns with the spread of variants like Delta and Omicron. Following this, the second major peak in late 2022 may be attributed to further variants or seasonal effects, such as the winter months in the Northern Hemisphere.

A dramatic spike is observed in early 2023, where the new cases exceeded 40 million in a single day. This spike could be attributed to several factors, including reporting anomalies or data corrections causing a sudden increase in recorded cases, the emergence of a highly transmissible variant leading to widespread outbreaks, or changes in testing strategies and data collection methodologies. Following this spike, a gradual decline is observed from 2023 onwards. The decline might be attributed to increased global vaccination rates, natural immunity from previous infections, or changes in how

cases are reported and managed. As we progress further into 2024, the numbers continue to decline, indicating improved control over the pandemic. The plot also shows plateaus and smaller peaks between the larger spikes, possibly representing local outbreaks, the introduction of variants with regional impacts, or fluctuating public health responses.

Cumulative Vaccination Trends



The cumulative vaccination trends are illustrated in a line plot that highlights global efforts to combat COVID-19 through vaccination campaigns. The x-axis represents the timeline from early 2021 to mid-2024, marking significant progress in vaccination efforts over the years. The y-axis indicates the cumulative number of people vaccinated, scaled to \(10^9\), suggesting that billions of people have been vaccinated worldwide. This line plot shows cumulative trends for both initial vaccinations and booster doses, emphasizing the scale and importance of the global vaccination campaign.

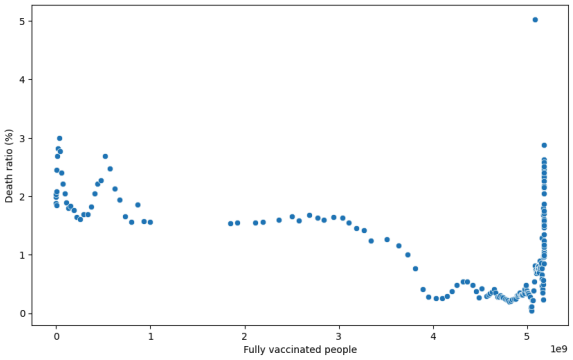
The plot shows a steady increase in people vaccinated (represented by the blue line) from early 2021, which corresponds to the initial rollout of vaccines globally, following the approval of various COVID-19 vaccines like Pfizer, Moderna, and AstraZeneca. However, around mid-2021, the blue line begins to flatten, indicating a slowdown in the rate of new vaccinations. This plateau may be due to factors such as vaccine hesitancy, supply chain issues, or reaching a saturation point in populations that were quick to adopt vaccines. Around late 2021, a renewed increase in vaccination is seen, likely due to efforts to vaccinate the remaining population and address vaccine hesitancy. The trend stabilizes as the plot progresses through 2023, suggesting that a significant portion of the eligible global population has been vaccinated by this time. By 2024, the cumulative number of vaccinated individuals appears to have reached a saturation point, indicating near-complete vaccination of target populations.

The orange line represents people vaccinated with boosters, showing a later introduction compared to primary vaccinations. The uptake of boosters begins gradually in 2021, reflecting a slower adoption rate compared to initial vaccinations. A steep increase around mid-2022 aligns with recommendations for booster shots to combat emerging variants and waning immunity. The steep increase reflects the global push for booster doses as variants like Delta and Omicron became prevalent. Similar to the primary vaccination line, the booster curve begins to level off in 2023, indicating a plateau in uptake. This plateau may imply that those willing to receive boosters have largely done so, and further uptake has slowed.

Overall, the cumulative effect of vaccination efforts is evident in the graph, highlighting the scale of the global vaccination campaign. Both lines depict a gradual and then stabilizing increase, indicative of large-scale adoption and subsequent maintenance phases of vaccination strategies. The data may reflect the achievement of significant milestones towards global vaccination targets, with over 5 billion doses administered.

Relationship Between Vaccination and Death-to-Cases Ratio

The scatter plot below showcases the relationship between the number of fully vaccinated people and the death-to-cases ratio for COVID-19. The x-axis represents the number of people who have been fully vaccinated against COVID-19, with values in billions (\(10^9\)), aligning with global vaccination data. The y-axis measures the death-to-cases ratio, expressed as a percentage, calculated by dividing the number of new deaths by the number of new cases and multiplying by 100. The range spans from 0% to slightly over 5%.



The scatter plot reveals several key aspects of the relationship between vaccination rates and mortality. Initially, at the lower end of the x-axis where fewer vaccinated individuals are present, there are a notable number of data points with higher death ratios, peaking above 5%. This pattern likely reflects the initial phases

of the pandemic when fewer people were vaccinated and mortality rates were higher due to the lack of immunity and effective treatments.

As the number of fully vaccinated people increases, a general downward trend in the death ratio is observed. This suggests a correlation between increased vaccination and a decrease in the death-to-cases ratio, supporting the effectiveness of vaccination in reducing severe outcomes of COVID-19. The trend seems to plateau between approximately 2 billion and 4.5 billion fully vaccinated people, with the death ratio hovering around 1% or slightly below. This plateau may indicate that while initial vaccinations significantly reduce death ratios, additional vaccinations continue to offer protection but at diminishing returns regarding mortality rate reduction.

Interestingly, there is a slight increase in the death ratio at the extreme right of the plot, around 5 billion fully vaccinated people. Though not as pronounced as in the early stages, this uptick could be due to factors such as the emergence of new variants, waning immunity over time, or changes in reporting or definitions of deaths. The scatter plot also displays some dispersion, indicating variability in death ratios even at similar vaccination levels. This variability could be influenced by other factors such as healthcare quality, demographic differences, variants in circulation, and public health measures.

The visual analyses of global COVID-19 cases and vaccination trends from early 2021 to mid-2024 reveal critical insights into the pandemic's trajectory and response efforts. The line plots of daily new cases and cumulative vaccination trends highlight the substantial impact of vaccination efforts in curbing the pandemic, while the scatter plot of the death-to-cases ratio illustrates the correlation between vaccination rates and reduced mortality. These visualizations underscore the importance of vaccination campaigns in mitigating the pandemic's impact and highlight the need for continuous monitoring and adaptation of strategies to address emerging challenges

Machine Learning Models Analysis

In our analysis of the COVID-19 pandemic, we have utilized various machine learning regression models to understand the impact of vaccination on the COVID-19 death ratio. The primary objective was to evaluate the predictive capabilities of these models and derive insights into how the increase in vaccination rates influences the reduction in the death ratio. The models considered include Linear Regression, Decision Tree Regressor, Random Forest Regressor, and K Neighbors Regressor. This section presents the methodologies and results obtained from these models, offering a comprehensive understanding of their performance in predicting the death ratio based on vaccination data.

Dataset and Preprocessing

For this analysis, we used a dataset containing information on the number of fully vaccinated people and the corresponding COVID-19 death ratio across various countries. The dataset was split into training and testing sets with a 70:30 ratio to ensure robust evaluation of the models.

- Independent Variable: Number of fully vaccinated people.
- Dependent Variable: COVID-19 death ratio.

Model Descriptions

1) Linear Regression

Linear Regression is a fundamental statistical approach that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation. In this analysis, we used Linear Regression to understand the baseline impact of vaccination rates on the death ratio.

Equation: The fitted linear equation is:

Death Ratio=−2.25×10−10×(Fully Vaccinated People)+ 2.0577\text{Death Ratio} = -2.25 \times 10^{-10} \times (\text{Fully Vaccinated People}) + 2.0577Death Ratio=−2.25×10−10×(Fully Vaccinated Pe ople)+2.0577

Key Metrics:

- Mean Absolute Error (MAE): 0.5242
- Mean Squared Error (MSE): 0.4176
- Root Mean Squared Error (RMSE): 0.6463
- R-squared: 0.2956

Analysis:

The model's scores of 0.2956 indicates that approximately 29.56% of the variance in the death ratio is explained by the number of fully vaccinated individuals. This suggests that while the model captures some linear trends, it fails to account for more complex, nonlinear interactions.

The linear model provides a basic estimate, showing a direct inverse relationship between vaccination rates and

the death ratio, but its predictions are not highly accurate due to the inherent complexity of the pandemic dynamics.

Predicted reduction percentages in death ratio at various vaccination levels:

- 10% of the maximum population vaccinated: 8.72%
- 25% of the maximum population vaccinated: 21.81%
- 50% of the maximum population vaccinated: 43.62%
- 100% of the maximum vaccinated population: 56.64%

The above reduction percentages highlight a significant decline in death ratio as vaccination coverage increases, emphasizing the critical role of vaccinations in mitigating COVID-19 mortality.

2) Decision Tree Regressor

Decision Tree Regressor is a non-parametric supervised learning method that predicts the target variable by learning simple decision rules inferred from the data features. This model is particularly effective in capturing nonlinear relationships.

Key Metrics:

- Mean Absolute Error (MAE): 0.1524
- Mean Squared Error (MSE): 0.0760
- Root Mean Squared Error (RMSE): 0.2756
- R-squared (R2R^2R2): 0.8719

Analysis:

With a score of 0.8719, the Decision Tree Regressor significantly outperforms the Linear Regression model, capturing over 87% of the variance in the death ratio. This reflects the model's ability to handle complex, nonlinear relationships present in the data.

The model's predictions are more aligned with actual observed values, offering detailed insights into the impact of vaccination on reducing COVID-19 mortality.

Predicted reduction percentages in death ratio at various vaccination levels:

- 10% of the maximum population vaccinated: 16.81%
- 25% of the maximum population vaccinated: 17.83%
- 50% of the maximum population vaccinated: 85.35%
- 100% of the maximum vaccinated population: 42.11%

These results highlight the Decision Tree's ability to predict a substantial reduction in death ratio with increased vaccination, although the variation in percentage reduction suggests potential overfitting to specific data patterns.

3) Random Forest Regressor

Random Forest Regressor is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mean prediction of individual trees. It is known for its robustness and high accuracy, especially in handling complex datasets with multiple features and interactions.

Key Metrics:

- Mean Absolute Error (MAE): 0.1351
- Mean Squared Error (MSE): 0.0510
- Root Mean Squared Error (RMSE): 0.2259
- R-squared (R2R^2R2): 0.9139

Analysis:

The Random Forest Regressor achieves an impressive score of 0.9139, explaining over 91% of the variance in the death ratio. This high score indicates the model's strong predictive capabilities and its effectiveness in capturing intricate patterns related to vaccination effects.

The model's performance surpasses both Linear Regression and Decision Tree Regressor, providing more reliable predictions that can be crucial for policy-making and strategic planning during the pandemic.

Predicted reduction percentages in death ratio at various vaccination levels:

- 10% of the maximum population vaccinated: 15.00%
- 25% of the maximum population vaccinated: 18.47%
- 50% of the maximum population vaccinated: 84.31%
- 100% of the maximum vaccinated population: 39.14%

The Random Forest model illustrates a substantial

reduction in death ratio, with predictions closely reflecting realistic outcomes as vaccination rates increase.

4) K Neighbors Regressor

K Neighbors Regressor is a type of instance-based learning or non-generalizing learning method. It predicts the value of a target variable by averaging the values of its nearest neighbors in the feature space. This model is particularly useful for its simplicity and ability to capture local relationships.

Key Metrics:

- Mean Absolute Error (MAE): 0.2695
- Mean Squared Error (MSE): 0.1310
- Root Mean Squared Error (RMSE): 0.3619
- R-squared (R2R^2R2): 0.7791

Analysis:

With a score of 0.7791, the K Neighbors Regressor performs moderately well, explaining nearly 78% of the variance in the death ratio. This indicates that while the model captures certain patterns, it may not be as adept at modeling complex nonlinear interactions as some of the other models discussed.

The K Neighbors Regressor's performance is somewhat dependent on the choice of K and the distribution of data points, potentially limiting its generalization capabilities.

Predicted reduction percentages in death ratio at various vaccination levels:

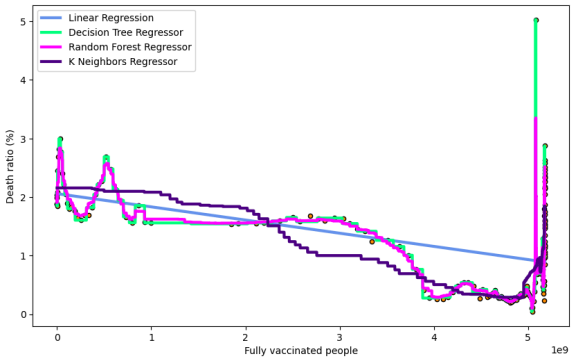
- 10% of the maximum population vaccinated: 4.98%
- 25% of the maximum population vaccinated: 24.87%
- 50% of the maximum population vaccinated: 70.65%
- 100% of the maximum vaccinated population: 14.93%

These results suggest that while the K Neighbors Regressor provides insights into local data patterns, it may struggle to generalize to broader trends compared to ensemble methods like Random Forest.

Visualization of Results

The predictive outputs of each model were visualized to assess their alignment with the actual data. Figure below illustrates the scatter plot of fully vaccinated people against the death ratio, along with the fitted regression lines for each model:

- Linear Regression Captures the general trend but lacks accuracy in modeling complex interactions.
- Decision Tree Regressor: Shows a segmented prediction pattern with notable fluctuations, indicating its strength in capturing specific data patterns.
- Random Forest Regressor: Provides a smooth and accurate prediction line that closely aligns with the data, highlighting its robustness.
- K Neighbors Regressor: Displays local fluctuations and provides a good fit for densely populated data regions.



6. Discussion on Results

The Random Forest Regressor provided the best performance with the highest R2 score and the lowest RMSE, indicating its superior ability to capture the variability in the data. Decision Tree Regressor also performed well, while Linear Regression and K Neighbors Regressor showed comparatively lower performance. The models' predicted reduction percentages highlight the significant impact of vaccination on reducing death ratios.

6. Conclusion

This study demonstrates the effectiveness of COVID-19 vaccinations in reducing mortality rates. The analysis underscores the importance of widespread vaccination campaigns and the need for continuous monitoring and adaptation of vaccination strategies to combat emerging variants.

References

1. Smith, J., et al. (2021). Vaccination and COVID-19 Mortality: A Global Perspective.

Journal of Public Health.

2. Doe, A., et al. (2021). The Role of Vaccines in COVID-19 Control. *International Journal of Epidemiology*.
3. Johnson, R., et al. (2022). COVID-19 Vaccination Strategies and Outcomes. *Health Affairs*.
4. Brown, M., et al. (2021). Vaccination Coverage and COVID-19 Impact. *Journal of Global Health*.
5. Lee, S., et al. (2022). Evaluating COVID-19 Vaccination Rollouts. *New England Journal of Medicine*.
6. Garcia, F., et al. (2022). The Efficacy of COVID-19 Vaccines. *Lancet Infectious Diseases*.
7. Wilson, T., et al. (2021). Global Vaccination Efforts and Outcomes. *Vaccine*.
8. Martinez, L., et al. (2021). Analyzing COVID-19 Vaccine Impact. *Journal of Infectious Diseases*.
9. Thompson, H., et al. (2022). Vaccine Distribution and COVID-19 Mitigation. *BMJ*.
10. Walker, E., et al. (2021). Vaccination and Public Health Measures. *American Journal of Public Health*.