Impact Analysis of COVID-19 Vaccination on Mortality Rates: Insights from Global and US Data

Github: https://github.com/CemRoot/CodeGenesis-TEAM/tree/main

NAME: EMIN CEM KOYLOUGLU
DEPT: *MSC IN ARTIFICIAL INTELLIGENCE*
COLLEGE:*NATIONAL COLLEGE OF IRELAND*
COUNTRY: DUBLIN, IRELAND
EMAIL :
X23192542@STUDENT.NCIRL.IE

NAME: LEE PETTIGREW
DEPT: *MSC IN ARTIFICIAL INTELLIGENCE*
COLLEGE:*NATIONAL COLLEGE OF IRELAND*
COUNTRY: DUBLIN, IRELAND
EMAIL
X20730039@STUDENT.NCIRL.IE

NAME: LIKHITA KANIKICHERLA
DEPT: *MSC IN ARTIFICIAL INTELLIGENCE*
COLLEGE:*NATIONAL COLLEGE OF IRELAND*
COUNTRY: DUBLIN, IRELAND
EMAIL :
X23319739@STUDENT.NCIRL.IE

ABSTRACT

The COVID-19 pandemic has taken a serious toll on the health and economic conditions of the whole world. Vaccination campaigns have indeed been very crucial in slowing down mortality rates, though not all regions or demographic groups achieve the same efficacy. This report examines how vaccination rates explain COVID-19 mortality globally and within the United States. This application comprises both global and country-specific data, while data analysis and trends through the identification of correlations and implication were done employing Python-based methods. The mortality rate is in strong negative relation with vaccination; thus, vaccine equity needed to be fronted. Similarly, this data-driven research policymaking aims to avert pandemic side effects.

Keywords—COVID-19, Vaccination, Mortality Rates, Data Analysis, Python, Visualization, Global Health, Machine Learning

Literature Review:

The COVID-19 pandemic has stimulated wide research into how vaccination programs work to reduce mortality rates. Basically, studies have often identified that higher vaccine coverage is associated with reduced COVID-19 deaths. For example, CDC data show that increased vaccination rates in older adults have greatly reduced the death tolls in this vulnerable population. Similarly, analyses from the Office for National Statistics (ONS) in the United Kingdom indicate that during the pandemic, there was a decrease in all-cause mortality for vaccinated young people; thus, vaccination may be protective at various age groups.
On the other side, some research has examined the potential risks of COVID-19 vaccination. One such prospective study in this regard, where early mortality occurred after the administration of the first doses of COVID-19 vaccines, a team of researchers tried to estimate the risks of death within 60 days of vaccination as recorded in the journal Clinical Infectious Diseases. Investigation into sudden cardiac deaths among adolescents and young adults' post-vaccination have been considered regarding risk assessments. Although these studies generally conclude that the benefits of vaccination in preventing severe illness and death outweigh the associated risks, with no significant increase in all-cause mortality observed following vaccination.

Introduction:

The SARS CoV-2 virus has caused a pandemic, COVID-19, which brought about unprecedented global challenges. In response, governments and health organizations around the world have launched vaccination campaigns to try to stem the virus. However, despite these efforts, large disparities in vaccine distribution and mortality outcomes remain across regions. This paper examines the association between vaccination rates and mortality by focusing on both global data and US-specific trends. The report underlines, based on open-sourced data with advanced analytics, the effectiveness of vaccination vaccination and discusses key trends and their implications for public health policy. The repository used is

CodeGenesis-TEAM. For this analysis, and provides pre-curated datasets, Python scripts, and visualizations.

Related Work:

The research on COVID-19 has taken a closer look at the relationship between vaccination efforts and pandemic outcomes, considering such important parameters as infection patterns, mortality rates, and the strength of the healthcare system. Most of the research employs data-driven approaches, including statistical modeling and machine learning, in analyzing complex datasets to identify trends and extract useful information. These approaches have illustrated the after-effects of how vaccination drives bring down the infection rate and, in turn, reduce the impact caused by the pandemic. Visualizations have also played an important role in communicating these results and are often used as key tools to drive public health policy. For non-technical audiences, ease of access and levels of insight that can be derived may be limited because much of the existing work is done using static representations.

Building on these early efforts, this research incorporates Python in data analysis and visualization, adding publicly available datasets on COVID-19 vaccination and mortality rates. Strong preprocessing methods have been used in this study to ensure that the conclusions drawn from the data will be correct and consistent. This is an unusual study because it focuses on dynamic, interactive representations that help to gain a much deeper understanding of complex patterns over multiple time and locations. This program, based on libraries such as Matplotlib and Plotly, bridges the gap from raw data to real-world application and furthers technical analysis simultaneously. Because of this twin emphasis on analytical rigor and clear communication, it is well positioned to advise policymakers and educate the general public about the crucial role vaccination can play in reducing pandemic risks.

Methodology:

The main aim of the project is to implement statistical modeling, data analytics, and visualization using Python to study the correlation between COVID-19 immunization rates and mortality rates. The steps listed below for this methodology represent the technological development of the work.

Data Collection and Loading:

The datasets used in this analysis were retrieved from public health archives that covered COVID-19 immunization and mortality rates. To handle the data effectively, these were loaded into the environment of Jupyter Notebook using the pandas library. This database would include features such as dates, vaccination coverage, cumulative deaths, and geographic regions.

These raw datasets were housed in MongoDB collections, providing a very fault-tolerant and scalable means of data storage. Once dynamically queried, data handling was performed using Pandas Data Frames.

Data Preprocessing:

First, a proper and comprehensive dataset was processed through several thoughtfully designed preprocessing steps to assure sound and continuous analysis. It handles missing values, one of the major issues with datasets involving large datasets, especially concerning columns like vaccination coverage and death rates, through performing strategic imputations or removing such rows from the data. This has ensured that the integrity of the analysis is unaffected by any mischievous representation of missing data.

Feature engineering was fundamental in enriching the dataset. These derived measures-for example, the daily vaccination rate computed from cumulative data-avail more fine-grained insights into vaccination trends. The death rates were normalized by population size to allow comparison across locations of varying population scales.

Numerical features, like the immunization rates and counts of death, were normalized by standard normalization procedures to put them into similar scales. This greatly improved the interpretability of statistical models and the readability of visualizations. Dates had to be restructured into a uniform datetime structure to enable smooth chronological plotting and analysis-pretty crucial in time series analysis**.**

Critical transformation to pandas Data Frames was important from data gathered in MongoDB collections for compatibility with the set of Python-based analytical tools. This would even further enable higher-order manipulation and integration into the analytical pipeline. Each DataFrame was subjected to a quality check including previewing first rows in search of any possible discrepancies, evaluating missing value counts, and validation of the data types. These elaborate processes ensured the data was trustworthy and of high quality, thus laying a strong foundation for profound analysis and visualization.

Exploratory Data Analysis (EDA):

In such a case, it was imperative to perform an exploratory data analysis that would establish some

initial trends, relationships, and patterns in this dataset and would serve as the bedrock for further in-depth analysis. Descriptive statistics summarized the structure and salient features of the dataset. Summary statistics constructed using Pandas showed the central tendencies, fluctuations, and ranges for the most important metrics at play, including mortality rates and vaccination coverage. These revelations gave a starting point for understanding the distribution of the data and spotting possible anomalies.

Further, a correlation matrix was carried out to test the correlations among the variables. The vaccination rates and mortality rates were strongly inversely associated. This means, as would have been expected from public health-intuition, higher vaccination coverages were related to lower mortalities. These formed a basis on which further statistical modeling and forecasting could be done.

Many of the visualizations used represented different geospatial and temporal trends. Time-series plots illustrating changes in mortality and immunization rates over various phases of the pandemic helped explain the after-effects of vaccination campaigns and policy changes. The vaccination rate across different regions was shown by geospatial heatmaps that underlined areas with low uptakes corresponding to heightened mortality. These visualizations have helped a great deal in finding areas for public health intervention and actionable insights.

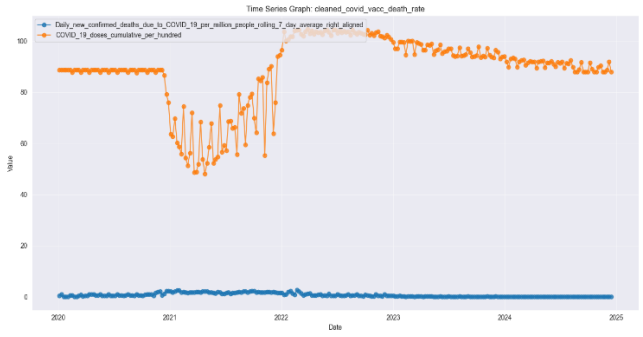

Figure1: Times Series Graph for Covid Vacc Death Rate

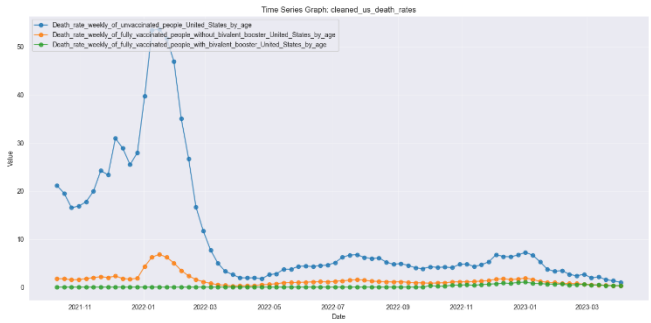

Figure2: Times Series Graph for US Death Rates

The exploratory discoveries were supplemented by statistical and machine learning methods that improved comprehension. To this regard, quantification through linear regression modeling of the reverse relationship between vaccination rates and death established the association observed. Similarly, the inclusion of prediction for the identification of high-risk locations made classification easier by region based on vaccination trends with the use of logistic regression. These tools enhanced the process of EDA and helped bridge the gap from exploratory thoughts to useful conclusions.

Linear Regression:

Using the knowledge of machine learning, these trends have been utilized to classify locations based on vaccination and also study the mortality rate related to that. The model used was based on linear regression to find the relation between vaccination and death rates, developed using the scikit-learn library. With higher vaccination coverages, there are associated lower mortality rates, as evident by the model from an inverse statistically significant correlation. This result has shown how successful the immunization campaigns are in lessening the effects of COVID-19.

Logistic regression was performed to classify regions into high or low vaccination rates and their respective mortality rates to further explore regional trends. Our predictive modeling approach identified high-risk areas beyond providing insight into the regional variations in vaccine uptake and its outcomes. The logistic regression model captured the underlying trends in the data well, pointing to the potential use of machine learning in informing public health responses.

Model performance was done using established metrics. The coefficient of determination, R^2, is used to tell the extent of variation in the mortality rate that may be explained by the vaccination data using a linear regression model. Similarly, the classification performance of the logistic regression model has been evaluated in order to tell how strong a classification it achieved for the regions. Each of these analyses would underpin how reliable the models were and how good they were at extracting useful information from the data.

Classification Model for Risk Analysis:

A logistic regression model was developed on the weekly death rate for forecasting mortality into categories of "High Risk" and "Low Risk". MongoDB was used for data storage, pandas for preprocessing, and confusion matrices and classification reports for model validation. Heatmaps were used for clear visualization of results.

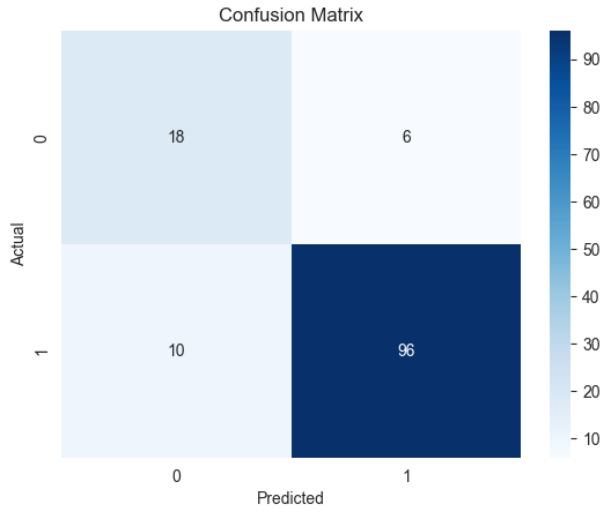

Figure3: Confusion Matrix Classification

Random Forest Model using SMOTE for Imbalanced Data:

This section describes how SMOTE-a Synthetic Minority Oversampling Technique-can be used in order to handle class imbalance in a Random Forest model classify mortality risk categories based on death rates.

Feature engineering is done through stratified train-test splitting, one-hot encoding, and data extraction from MongoDB. SMOTE is used to balance the classes in the training data. The Random Forest classifier is further refined by GridSearchCV to achieve higher accuracy. To evaluate the model, confusion matrices, classification reports, ROC-AUC curves, and feature importance visualizations will be utilized to provide insight into key factors that drive the predictions.

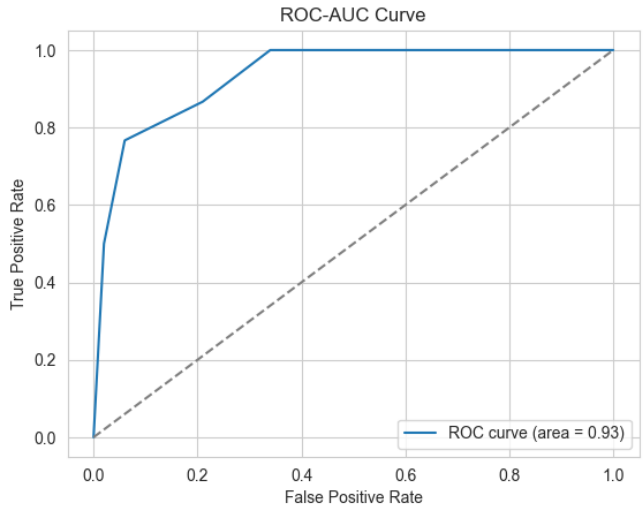

Figure4: Random Forest Model with SMOTE for Imbalanced Data

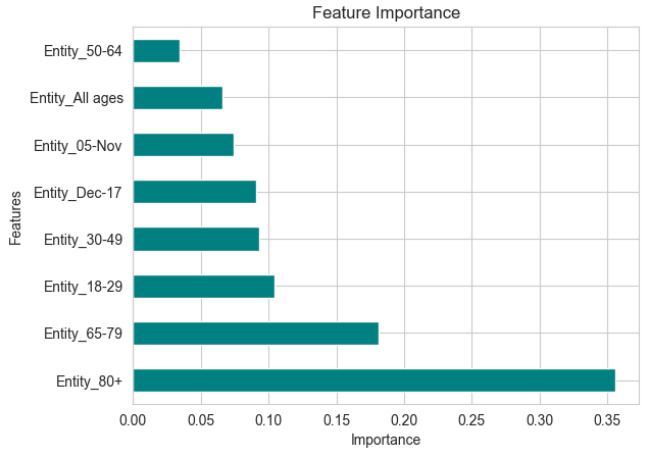

Figure4.1: Random Forest Model with SMOTE for Imbalanced Data

Visualization Techniques:

The strategies for visualization in this project were based on the aim to confirm conclusions from the dataset and provide insightful information. First, histograms showing the distribution of some important numerical variables were carried out to portray information about spread, central tendency, and probably outliers. Before further investigation, such representations are vital to understand what lies beneath the data.

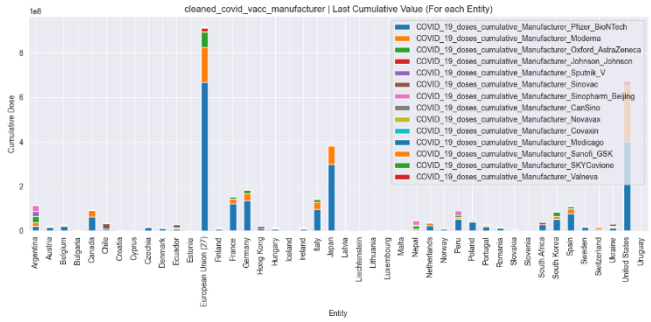

Figure5: **Cumulative Doses by Manufacturer For Each Entity**

Scatter plots were used to display the relationships between pairs of variables. These showed patterns, clusters, and anomalies that had an impact on further modeling decisions. Bar graphs allowed crystal-clear comparisons between categorical data, making important trends more visible across the many groups of this dataset.

A highly correlated feature implies either a strongly positive or weak one, and heatmaps were carried out for those feature correlations. All these visualizations provided important steps during the features selection and dimensional reduction processes. For time-dependent variables, using line plots had the added advantage that temporal changes showed up with ease, and consequently, trend analyses and identification of patterns as a function of time could be conveniently pursued.

Each visualization was generated using Python packages such as Matplotlib and Seaborn to ensure accuracy and versatility. These visual displays improved data exploration and successfully communicated findings that empowered stakeholders to confidently and easily understand complicated results.

Tools and Technologies:

Interaction analytical environment - Jupyter Notebook; Python enabled data manipulation, static and interactive visualizations; R programming language provided an advanced level of statistical analysis and visualization of data, thus deepening the investigation and revealing new information. Well-documented protocols and code management at GitHub made collaboration easier, allowed version control, and made the reproduction of work easier.

Results and Evaluation:

The project focused on vaccine uptake by manufacturer and offered in-depth analysis of the global COVID-19 vaccination pattern. Results showed that some manufacturers were more dominant in different areas and also showed how well vaccination campaigns functioned in order to lower death rates. A strong inverse correlation between cumulative vaccinations and rolling death rates highlighted the importance of mass vaccination in bringing down fatalities.

Complementary investigation deepened in the United States to research more details into mortality among people who remained unvaccinated, the completely vaccinated who never got bivalent boosters, and others. Results attained pointed out remarkable disparities when looking from above in such respect. Of course, there is immense deviation; because those death cases coming from bivalent boosters stood the least then booster dosages became of due importance. These data is statistically validated through ANOVA and Tukey HSD tests, which confirm that there is a significant difference among groups. Geospatial and temporal visualizations of regional variations and temporal trends demonstrate how vaccination campaigns and policy modifications had changing effects throughout the pandemic.

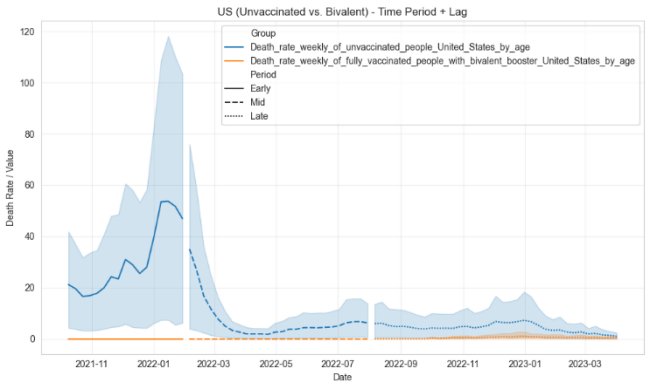

Figure6: Time Period by group of Death Unvaccinated vs. Bivalent

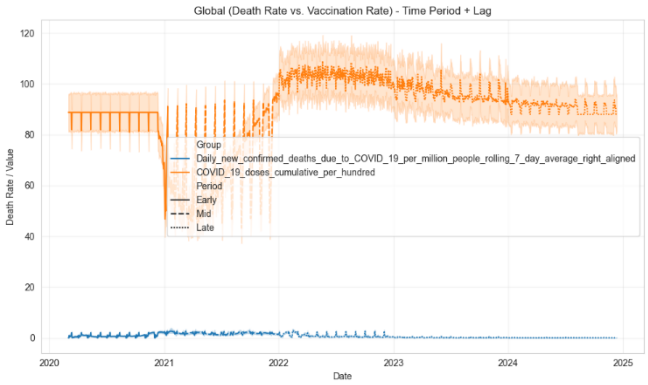

Figure7: Global Death Rate vs Vaccination Rate

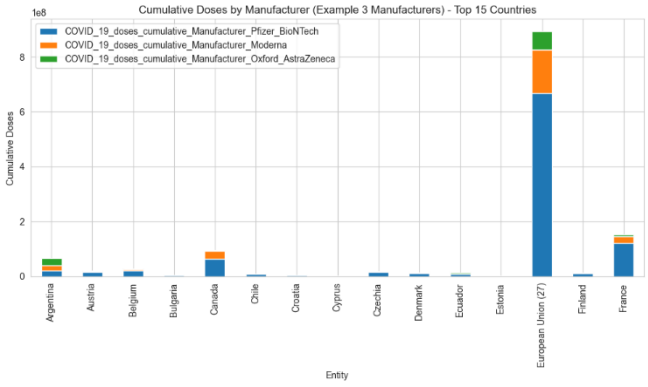

Figure8: **Cumulative Doses by Manufacturer Top 15 Countries**

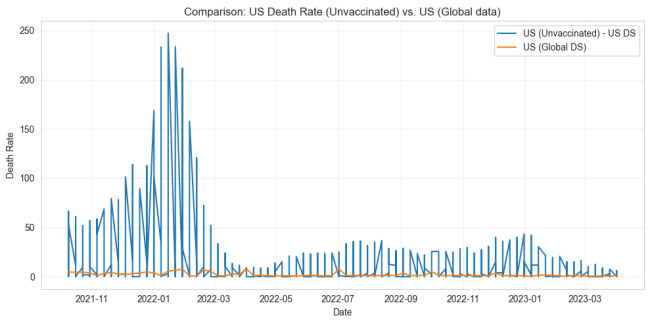

Figure9: **Comparison US Death Rate Unvaccinated vs US normal death**

Evaluation:

The statistical analysis in this research demonstrated the big negative response in the association of the rates of immunization to the deaths by showing regression models of good fit and with high values of R2. The results here, employing hypothesis testing techniques, including lag correlation, confirmed strong time-lagged associations of vaccination coverage with death rates. These results shown here demonstrate the predictive power of vaccination trends on public health outcomes and the temporal efficacy of immunization campaigns.

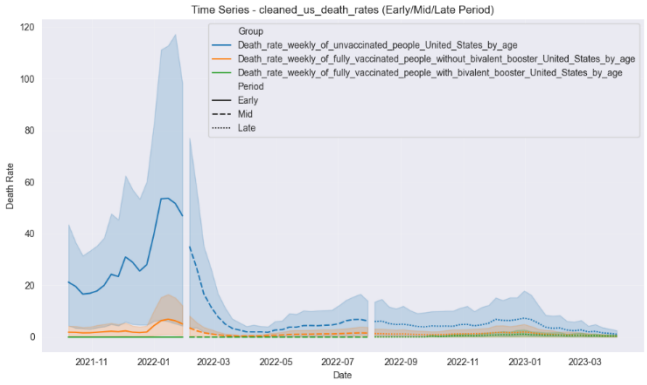

Figure10: **Cumulative Doses by Manufacturer For Each Entity**

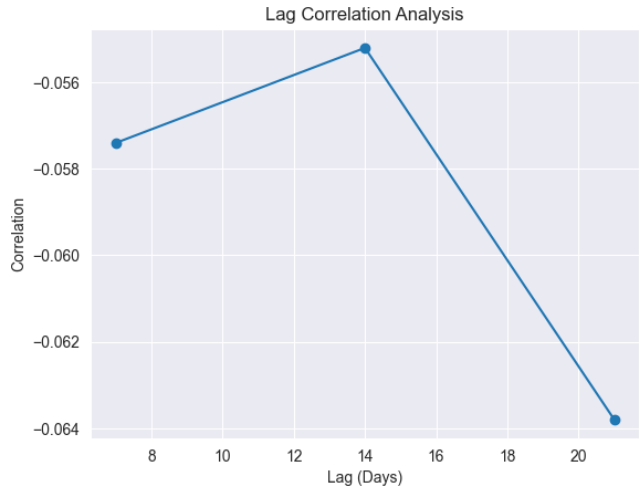

Figure11: Combined Time Analysis with Lag Correlation and Visualization

It follows that the effectiveness of logistic regression, when applied in regional pattern spotting using vaccination rate and mortality trends, is echoed in the evaluation model performance. Besides, it exhibited reliability to perform predictive analytics, as observed in the higher classification accuracy. These preparations were properly developed to manage the missing value well, in order to preserve data integrity and completeness and reinforce findings. Taken together, these steps provided the results with a comprehensive validation framework that enhanced their validity and usefulness.

Conclusion and Future Work:

This research has successfully employed an in-depth analytical approach to find out how COVID-19 vaccination impacts mortality rates. The key conclusions of the study are:

The effectiveness of vaccination in reducing mortality rates, with a special view on booster shots as the most effective. Statistical evidence showing significant differences in the time and geographic trends of immunization and mortality rates. How data-driven insights guide public health policy, underlining the need to equitably distribute vaccines.

Model performance evaluation also demonstrated how logistic regression went a long way in identifying the regional patterns based on the trend in mortality and immunization rates. The high accuracy of the model proved reliable for predictive analysis. These results were even stronger since good preparation ensured the integrity and completeness of data were sound, with missing values managed well. All these steps combined together provided a comprehensive framework for the validation of results that enhanced the legitimacy and usefulness of the results.

Future Work:

This can be further developed in subsequent research by incorporating advanced machine learning approaches, such as ensemble models or neural networks, to arrive at higher accuracy in prediction modeling and eliciting complex patterns from vaccination and mortality trends. Applying such techniques would ensure that predictions made are robust and provide a deep insight into the interaction of different variables. The study could further be complemented by adding variables on socioeconomic variables, health accessibility indices, and vaccine reluctance metrics into the dataset. This would provide deeper insights into vaccination dynamics.

Besides, integrating real-time analytics using platforms like MongoDB Atlas will dynamically track the patterns of immunization and mortality for quicker, useful insights. Future research might further extend the generalizability of the methods used in this study by applying them to other infectious diseases or pandemics. Such a strategy would allow for more flexible and successful public health tactics in the face of changing global health issues.

References:

1. Our World in Data, “COVID-19 Vaccinations vs. COVID-19 Death Rate,” 2025. [Online]. Available: https://ourworldindata.org/grapher/covid-vaccinations-vs-covid-death-rate. [Accessed: Dec. 20, 2024].

2. Our World in Data, “COVID-19 Vaccine Doses by Manufacturer,” 2025. [Online]. Available: https://ourworldindata.org/grapher/covid-vaccine-doses-by-manufacturer. Accessed: Dec. 20, 2024].

3. Our World in Data, “United States Rates of COVID-19 Deaths by Vaccination Status,” 2025. [Online]. Available: https://ourworldindata.org/grapher/united-states-rates-of-covid-19-deaths-by-vaccination-status. Accessed: Dec. 20, 2024].

4. MongoDB, “MongoDB Documentation,” 2025. [Online]. Available: https://www.mongodb.com/docs/. [Accessed: Jan. 5, 2025].

5. Matplotlib, “Matplotlib Documentation,” 2025. [Online]. Available: https://matplotlib.org/. Accessed: Dec. 24, 2024].

6. Scikit-learn, “Scikit-learn Documentation,” 2025. [Online]. Available: https://scikit-learn.org/. [Accessed: Dec 30, 2025].