Post-COVID-19 Analysis on Healthcare

Shubh Sanjeevkumar Mehta Sarthak Choudhary
Niveditha Bommanahally Parmeshwarappa

project-nibomm-shumehta-sarchou

Abstract

This study presents a comprehensive analysis of the COVID-19 pandemic's impact on global healthcare systems, with a focus on guiding an equitable and resilient recovery. It leverages expansive healthcare datasets and advanced machine learning techniques, including Gradient Boosting, XGBoost, LightGBM, and other sophisticated methods, to construct models that provide insightful guidance for policymakers. The research aims to illuminate post-pandemic trends in medical resource allocation, patient care strategies, and public health practices, enabling better preparedness for future systemic challenges.Central to the project is an exploration of emerging healthcare patterns and the utilization of cutting-edge analytical approaches to understand the pandemic's influence on patient outcomes, medical system management, and public health frameworks. This research promises to deliver actionable insights for strengthening medicine's crucial capabilities and offers a comprehensive account of COVID-19's extensive impact. The findings, which will be detailed later, are anticipated to be instrumental in shaping public health strategies and medical responses to cater to high-risk groups effectively and manage geographic disparities in pandemic impacts.The study also evaluates the performance of various machine learning models, considering their effectiveness in forecasting COVID-19 outcomes and their potential applicability in real-world scenarios. The outcomes of this research are expected to provide indispensable knowledge for leaders and stakeholders in reconstructing a society profoundly affected by the pandemic, guiding them towards more informed, data-driven decision-making processes.

Keywords

COVID-19 Pandemic, Healthcare Systems, Machine Learning, Data Analysis, Public Health Policy, Resource Allocation, Patient Care Strategies, Gradient Boosting, Geographic Disparities, Predictive Modeling.

1 Introduction

The unprecedented COVID-19 pandemic has fundamentally transformed global healthcare systems, necessitating evidence-based analysis to guide an equitable, resilient recovery. This ambitious research initiative aims to illuminate post-pandemic trends in medical resource allocation, patient care strategies, and public health practices. Leveraging expansive healthcare datasets in coordination with sophisticated machine learning techniques, our interdisciplinary team will construct insightful models to advance policymaking and better prepare for future systemic shocks.

At the project's core is a multifaceted interrogation of emerging healthcare patterns using Gradient Boosting, XGBoost, LightGBM, and other cutting-edge analytical approaches. By combining pandemic healthcare data with these ingenious algorithms, we can derive actionable insights to reinforce medicine's most vital capabilities. More broadly, this research will produce a definitive account of COVID-19's vast influence on patient outcomes, medical systems management, and public health frameworks—indispensable knowledge for leaders working to reconstruct a society battered by calamity.

Previous work

In this work by Mohammad Pourhomayoun, a machine learning model was created to estimate COVID-19 mortality risk and help prioritize patient treatment. Using various methods, it predicted mortality with 89.98 percent accuracy.[4].In this work by Aktar et al, a meta-analysis and machine learning analysis identified chronic conditions like COPD, cardiovascular disease, and diabetes as increasing COVID-19 mortality risk. Age and gender were the most reliable mortality indicators according to this author.

Here Sakifa Aktar's study of COVID-19 patients in Brazil found higher death rates among the elderly and those with cardiovascular, lung, and neurological conditions.[6].The study used machine learning to link COVID-19 severity with comorbidities such as type 2 diabetes, cerebrovascular and cardiovascular diseases, COPD, cancer, and hypertension. Asthma was identified as a significant factor. Age and gender predicted mortality. Combinations of pneumonia, diabetes, and hypertension, along with ARDS and hypertension, were associated with COVID-19 death. These findings help identify high-risk patient groups for resource allocation and care planning.[5]

In the study by G. J. B. Sousa, one of the largest COVID-19 mortality analyses using UK patient data found greater risks for men, older people, those with diabetes, asthma, and other illnesses. Risk was also higher for Black and South Asian ethnic groups, this author found.[2]. The study analyzed 6,036 COVID-19 patients using k-mode clustering, revealing links between conditions like COPD and obesity with higher in-hospital mortality. Multimorbidity, especially with mental/neurological and cardiovascular issues, increased mortality risk. Notably, asthma correlated with lower mortality. Rigorous methodology, including sensitivity analyses and ICD-10 codes, enhanced insights into COVID-19 mortality risks.[3].

In this work by Elizabeth J. Williamson, used machine learning to categorize comorbidities. This author found that vascular disorders, neurological disease, heart failure, and related conditions were key comorbidities. Patients with more comorbidities had worse outcomes.[1].

In summary, older age, male gender, chronic conditions like respiratory disease, cardiovascular disease, and diabetes are associated with higher COVID-19 mortality risk across multiple works. Ethnic minorities also show elevated risk according to these authors.

2 Methods

2.1 Data Preprocessing and Quality Control

The dataset used in this study is focused on conditions contributing to COVID-19 deaths in the United States. It covers the period from January 1, 2020, to September 23, 2023, and comprises more than 600,000 records. The dataset encompasses a variety of countries, age demographics, and pre-existing health conditions. The specific columns included in the dataset are Data As Of, Start Date, End Date, Group, Year, Month, State, Condition Group, Condition, ICD10_codes, Age Group, COVID-19 Deaths, Number of Mentions, and Flag columns.

The project initiated with a primary emphasis on data preprocessing and quality control, recognizing the crucial importance of accurate healthcare data. Automated analysis using Python scripts was employed to identify missing values and inconsistencies in the data. A detailed examination of the dataset led to the implementation of different strategies for handling these issues.

In cases where columns had significant missing data that could potentially skew results, those records were removed to maintain analytical integrity. For columns with minimal missing data, a more nuanced approach was adopted. This involved median imputation for continuous variables and modes or KNN for categorical data, ensuring that the categorical variance was preserved.

Following the imputation process, rigorous analysis ensued, incorporating visual plots and statistical tests to ensure that no biases or anomalies were introduced. This step was crucial for establishing a high quality of data, which is paramount for the integrity of subsequent analyses.

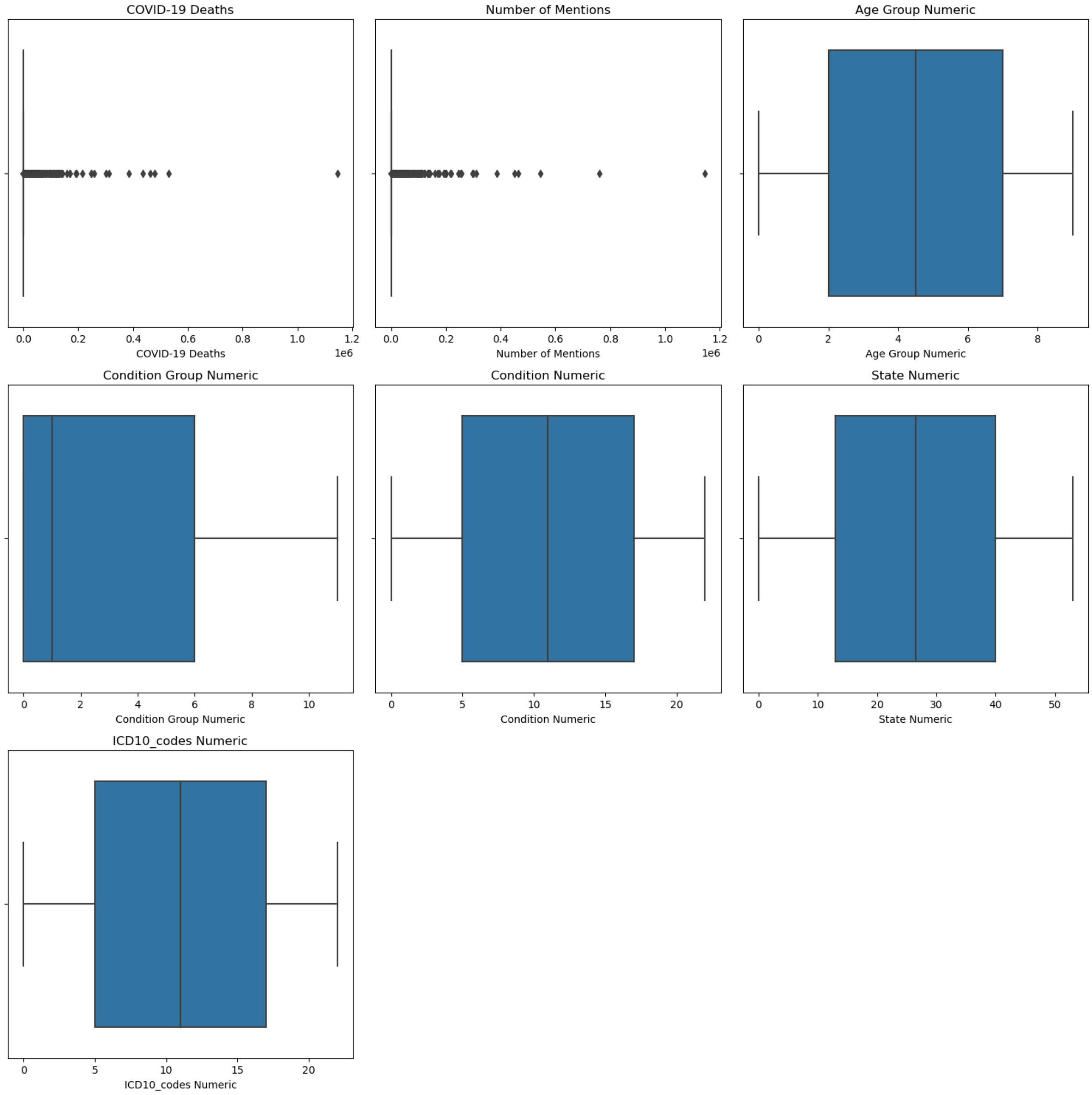

Figure 1: Box Plot analysis for variability of values in dataset

2.2 Feature Engineering ,Analysis and Selection

The next phase was feature engineering and selection, tailored to address the core research objectives related to the pandemic's impact. The focus was on filtering the data to precise segments based on date ranges, geographical regions, and demographics, ensuring relevance to the study. Python tools, such as Pandas, were used for transformations like one-hot encoding and labeling, which converted categorical data into numerical formats suitable for machine learning algorithms. This process retained the intrinsic properties of the data. Additionally, normalization and standardization techniques were applied to align divergent scales across features. New features were derived through aggregation, ratios, and interactions, considering the complex nature of healthcare data. Only variables showing high correlation, determined through techniques like PCA, were selected. This helped in balancing model efficiency with interpretive quality. We also conducted a comprehensive analysis of COVID-19 fatalities, focusing primarily on age demographics, condition and region to understand the pandemic's impact. Our methodology involved categorizing the data into various 'Age Groups','Condition' and summing the COVID-19 deaths within each group. This approach allowed us to compare the impact across different age groups, revealing which demographics were more affected, possibly due to higher vulnerability or comorbidities. This information is crucial for informing public health policies, including targeted vaccination campaigns and resource allocation.

We further refined our analysis by segmenting the data based on region, health condition, and age group. This multi-dimensional grouping of the dataset by 'State', 'Condition Group', and 'Age Group' provided a nuanced view of the pandemic's impact, considering regional differences, the influence of health conditions, and age factors. These insights are invaluable for understanding regional disparities and the interplay between health conditions and age in COVID-19 severity.

Additionally, we performed a time series analysis of COVID-19 deaths by age group, grouping data by 'Start Year' and 'Age Group' to observe trends over time. We visualized these trends using line plots, which helped in correlating the pandemic's progression with external events like pandemic waves or public health measures. This analysis was crucial in understanding how the impact of COVID-19 evolved over time across different age demographics.It combined age and health conditions. By grouping the data by 'Start Year', 'Age Group', and 'Condition', and employing the seaborn library for detailed visualization, we gained insights into how specific health conditions, along with age, influenced the death rates over time. This analysis was instrumental in identifying high-risk groups and conditions that led to more severe outcomes from COVID-19.

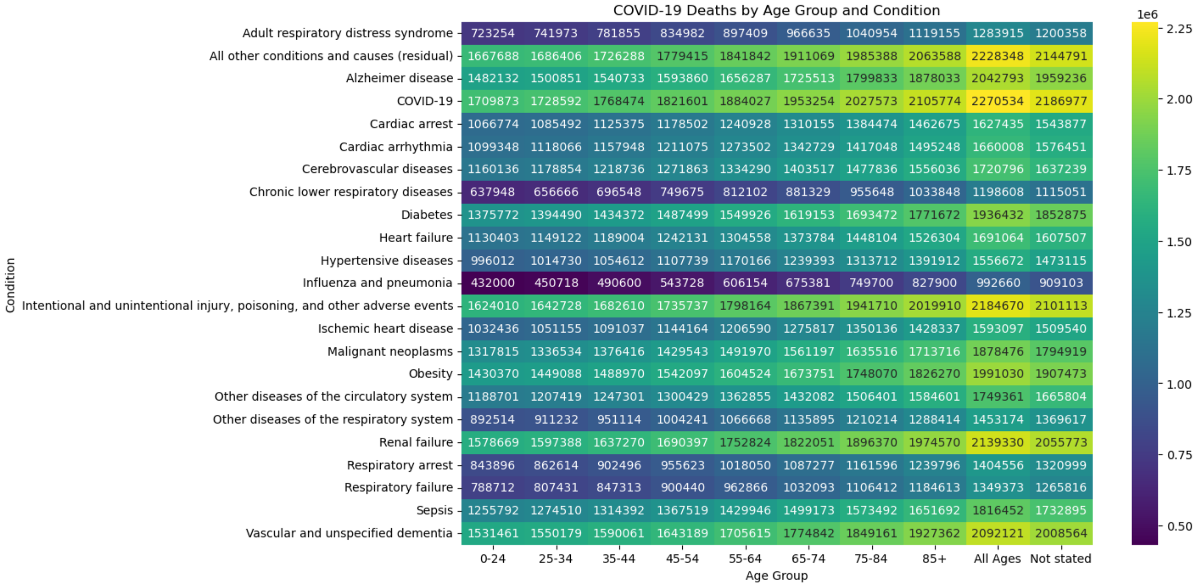

Figure 2: Heat map analysis COVID-19 Deaths by Age Group and Condition

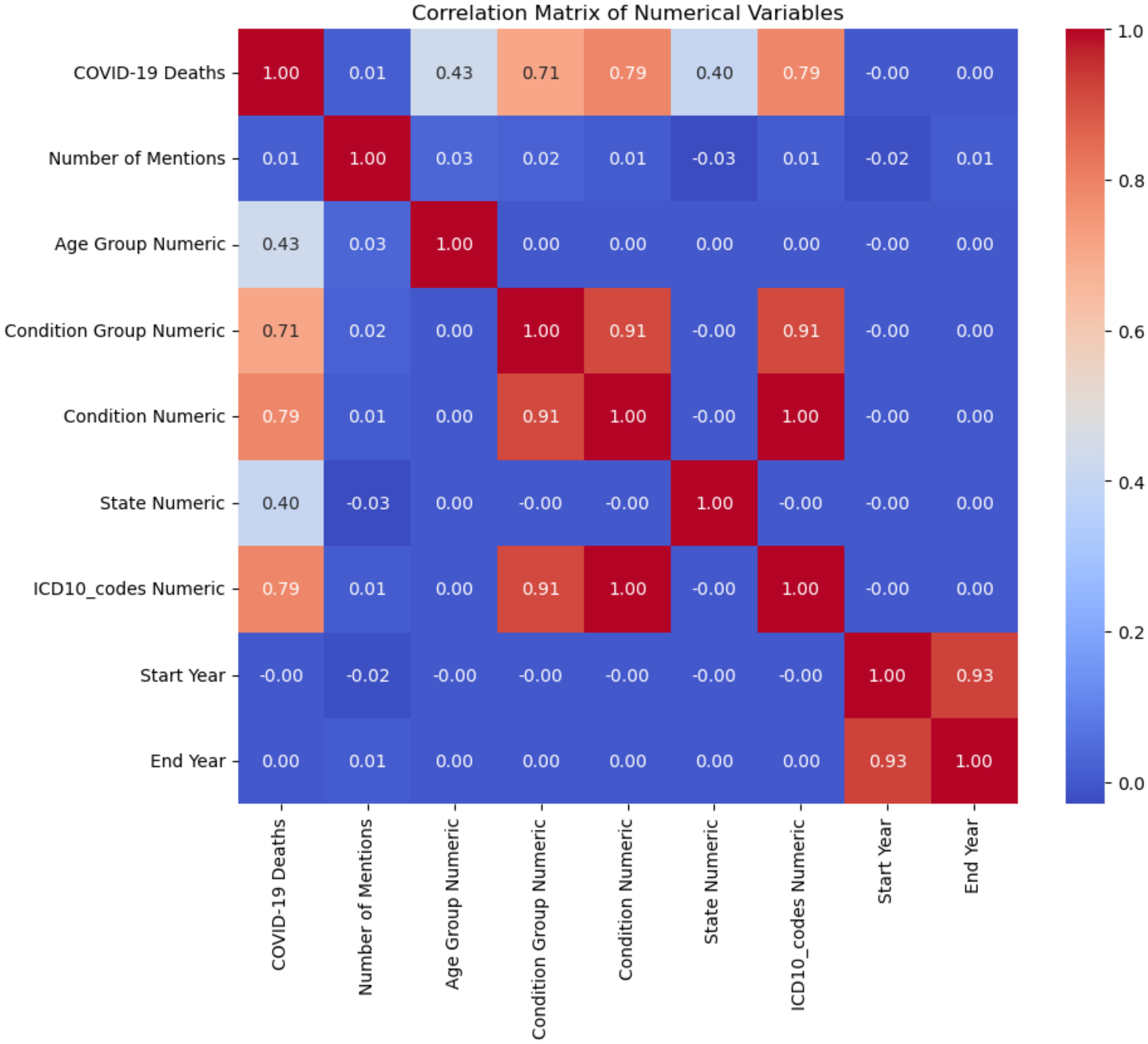

Figure 3: Correleation matrix analysis

2.3 Machine Learning Model Development

2.3.1 Multi-Layer Perceptron (MLP) Development

The choice of Multi-Layer Perceptrons (MLPs) was predicated on their exceptional capability to discern complex and non-linear patterns, which are prevalent in expansive healthcare datasets, particularly in the challenging context of a pandemic. The flexibility inherent in the MLP's architecture is particularly advantageous given the heterogeneous nature of COVID-19 data. The MLP model was intricately architected, featuring an input layer that corresponds to the number of features within our dataset. It included several hidden layers, meticulously engineered to capture and interpret the complex interrelationships inherent in the data. The output layer was specifically designed to align with our predictive objectives, whether for binary or multi-class classification, or for regression tasks. We incorporated advanced activation functions such as ReLU in the hidden layers to introduce necessary non-linearity, and tailored the output layer with either softmax or sigmoid functions, contingent upon the prediction requirements. The training regimen of the MLP involved a strategic implementation of backpropagation with gradient descent, meticulously optimizing a loss function selected to suit the unique characteristics of our healthcare prediction task.

$$\boldsymbol{Y} = \boldsymbol{f}^{(3)}\left(\boldsymbol{W}^{(3)} \cdot \boldsymbol{f}^{(2)}\left(\boldsymbol{W}^{(2)} \cdot \boldsymbol{f}^{(1)}\left(\boldsymbol{W}^{(1)} \cdot \boldsymbol{X} + \boldsymbol{b}^{(1)}\right) + \boldsymbol{b}^{(2)}\right) + \boldsymbol{b}^{(3)}\right)$$

In this equation:

- $\boldsymbol{Y}$ represents the output of the MLP.
- $\boldsymbol{f}^{(i)}$ denotes the activation function at the $\boldsymbol{i}$-th layer. Common choices for activation functions include ReLU (Rectified Linear Unit), sigmoid, and tanh.
- $\boldsymbol{W}^{(i)}$ and $\boldsymbol{b}^{(i)}$ are the weights and biases, respectively, at the $\boldsymbol{i}$-th layer of the network. These are the parameters that the network learns during the training process.
- $\boldsymbol{X}$ is the input vector to the network.

2.3.2 Gradient Boosting and XGBoost Implementation

Our selection of Gradient Boosting and XGBoost techniques was driven by their demonstrated robustness in diverse data scenarios, an essential attribute for managing the complexity and variability inherent in healthcare data. XGBoost, in particular, was chosen for its proficiency in efficiency and scalability, which is paramount when dealing with voluminous healthcare datasets. These models operate through the construction of a series of decision trees, each intelligently designed to rectify the inaccuracies of its predecessor, cumulatively resulting in a robust and comprehensive predictive model. A critical aspect of our approach was the fine-tuning of key hyperparameters, including the learning rate, number of trees, and depth of the trees. We also employed regularization techniques to curtail overfitting, thus ensuring the model's generalizability. The empirical evaluation of these models was rigorously conducted using cross-validation techniques, affirming their performance and reliability in predicting COVID-19 outcomes.

$$\begin{aligned}
\boldsymbol{F_0(x)} &= \arg\min_{\gamma} \sum_{i=1}^{n} \boldsymbol{L(y_i, \gamma)} \\
&\text{For } \boldsymbol{m} = 1 \text{ to } \boldsymbol{M}: \\
\boldsymbol{r_{im}} &= -\left[\frac{\partial \boldsymbol{L(y_i, F(x_i))}}{\partial \boldsymbol{F(x_i)}}\right]_{\boldsymbol{F(x)=F_{m-1}(x)}} \\
\boldsymbol{F_m(x)} &= \boldsymbol{F_{m-1}(x)} + \boldsymbol{\nu} \cdot \boldsymbol{h_m(x)}
\end{aligned}$$

In this equation:

$\boldsymbol{F_0(x)}$: The initial model, usually a constant prediction.
γ : The value that minimizes the loss function in the initial model.
$\boldsymbol{L(y_i, \gamma)}$: The loss function, measuring the difference between the predicted and actual value $\boldsymbol{y_i}$.
$\boldsymbol{n}$: The number of data points.
$\boldsymbol{y_i}$: The actual value for the $\boldsymbol{i}$-th data point.
$\boldsymbol{m}$: Represents the iteration step in the boosting process.
$\boldsymbol{M}$: The total number of boosting iterations.
$\boldsymbol{r_{im}}$: The negative gradient of the loss function at the $\boldsymbol{i}$-th data point for the $\boldsymbol{m}$-th model.
$\boldsymbol{F_m(x)}$: The prediction model at the $\boldsymbol{m}$-th iteration.
$\boldsymbol{\nu}$: The learning rate, scaling the contribution of each tree.
$\boldsymbol{h_m(x)}$: The base learner added at the $\boldsymbol{m}$-th step.

2.3.3 LightGBM

The incorporation of LightGBM was guided by its specific design attributes that cater to efficiency and speed, particularly when handling extensive healthcare datasets. Its innate ability to directly manage categorical features, without necessitating extensive preprocessing, renders it an ideal choice for healthcare data, which often comprises a diverse array of categorical and numerical variables. The development of LightGBM models in our project included the application of Gradient-based One-Side Sampling (GOSS), a technique that significantly enhances the balance of data distribution and expedites the training process — a critical factor in the analysis of time-sensitive healthcare data. Furthermore, LightGBM's capacity to provide detailed feature importance metrics was instrumental in decoding the influential factors within our COVID-19 dataset. We conducted a comprehensive comparative analysis of LightGBM against other models to evaluate its effectiveness in managing the specific intricacies of the healthcare data under study.

$$\boldsymbol{L(\phi)} = \sum_{i=1}^{n} \boldsymbol{L(y_i, \hat{y}_i)} + \sum_{k=1}^{K} \boldsymbol{\Omega(f_k)}$$

$$\boldsymbol{\Omega(f_k)} = \gamma \boldsymbol{T} + \frac{1}{2}\lambda \|\boldsymbol{w}\|^2$$

In this equation:

$\boldsymbol{L(\phi)}$: The overall objective function of XGBoost.

$\boldsymbol{n}$: The number of data points.

$\boldsymbol{y_i}$: The actual value for the $\boldsymbol{i}$-th data point.

$\boldsymbol{\hat{y}_i}$: The predicted value for the $\boldsymbol{i}$-th data point.

$\boldsymbol{K}$: The number of trees in the model.

$\boldsymbol{\Omega(f_k)}$: The regularization term for the $\boldsymbol{k}$-th tree.

γ : A parameter penalizing the complexity of the model.

$\boldsymbol{T}$: The number of leaves in the tree.

λ : The L2 regularization term on the weights.

$\|\boldsymbol{w}\|^2$: The square of the L2 norm of the leaf weights in the tree.

3 Results

In our analysis, we explored the links between health conditions and COVID-19 fatalities, geographic variations in the pandemic's impact, and its temporal dynamics. We discovered a significant prevalence of respiratory issues, cardiovascular diseases,alzheimer and diabetes among COVID-19 fatalities, highlighting the vulnerability of those with these conditions. Geographically, states with denser populations and urban centers, especially those with early outbreaks, saw higher death rates, indicating the influence of public health responses and healthcare infrastructure. Our time series analysis also showed fluctuating COVID-19 mortality rates, correlating with pandemic waves and the efficacy of public health measures, including vaccinations.

Evaluating our machine learning models, we focused on R-squared (R2) and Mean Squared Error (MSE) scores. XGBoost and LightGBM emerged as the standout models, with high R2 scores and low MSEs, indicating superior predictive accuracy and efficiency. XGBoost was slightly more accurate, while LightGBM was more efficient with large datasets. This comparative analysis highlights the need to choose models based on specific project requirements, balancing accuracy and dataset size.

In our comparative analysis, XGBoost and LightGBM stood out for their superior performance, as evidenced by their high R2 scores and low MSE values. This indicates their excellence in both accuracy and precision. The choice between XGBoost and LightGBM would be contingent upon specific project requirements, with XGBoost offering marginally higher accuracy and LightGBM providing enhanced efficiency in handling extensive datasets.

Model	**MLP**	**Gradient Boost**	**XG Boost**	**LightGBM**
R2 score	0.9785	0.8664	0.9331	0.9643
MSE score	548.6645	3402.1798	1703.5465	908.1637

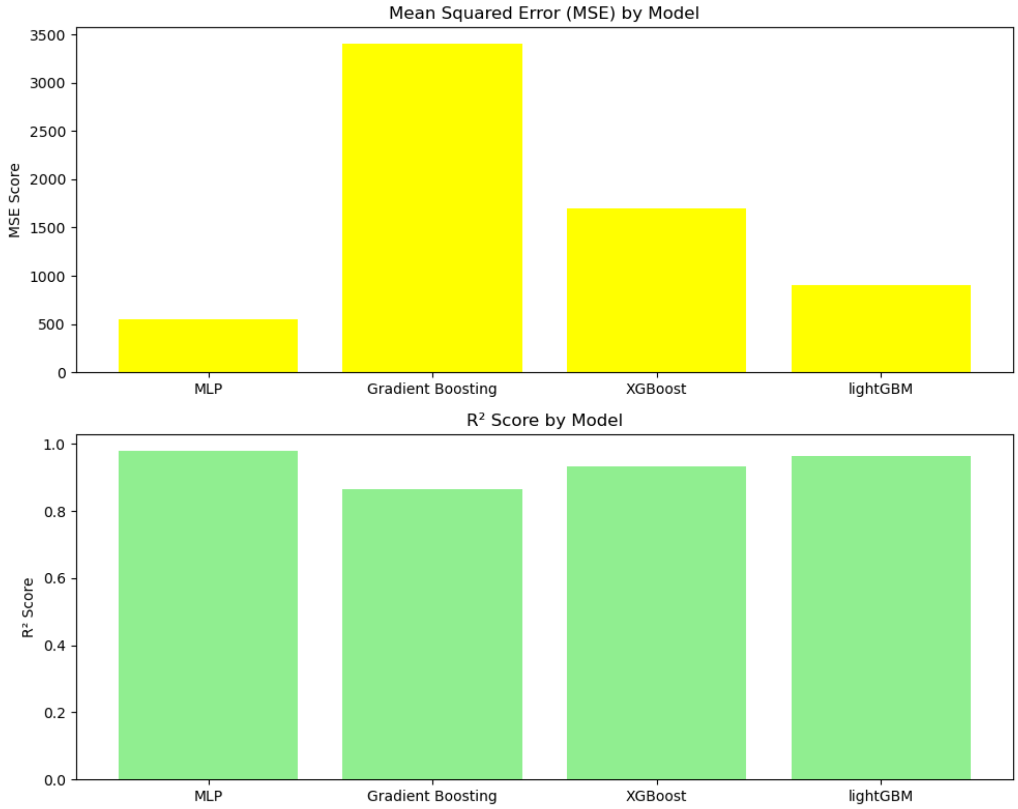

Figure 4: Result of various algorithm based on R2 Score and MSE Score

4 Discussion

Our analysis revealed significant correlations between specific health conditions and COVID-19 fatalities, underscoring the increased vulnerability of individuals with comorbidities like respiratory issues, cardiovascular diseases, and diabetes. This highlights the need for targeted public health strategies, such as prioritized vaccination and tailored treatment protocols for these high-risk groups. Geographic analysis showed notable disparities in COVID-19 impacts across states, with denser populations and urban centers experiencing higher death rates. This suggests the complexity of pandemic management, influenced by healthcare infrastructure, public health policies, and community behavior, and indicates the necessity for region-specific approaches in healthcare resource allocation and public health campaigns. Additionally, our time series analysis demonstrated how mortality rates varied with pandemic waves and public health measures, emphasizing the importance of timely and effective interventions in controlling the spread of the virus.

The performance evaluation of our machine learning models – MLP, Gradient Boost, XG Boost, and LightGBM – using R2 and MSE scores, revealed their effectiveness in forecasting

COVID-19 outcomes. XGBoost and LightGBM were particularly notable for their accuracy and efficiency, underscoring their potential in real-world applications. These models can be instrumental in predicting future trends of the pandemic, aiding in resource allocation, and informing policy decisions. The choice between these models should be based on the specific requirements of the task, with XGBoost being preferable for higher accuracy needs and LightGBM for its efficiency with large datasets.

In conclusion, the findings of our study provide crucial insights for healthcare policymakers, public health officials, and healthcare providers. They offer a data-driven basis for developing targeted and region-specific strategies to combat the COVID-19 pandemic. Additionally, the efficacy of the machine learning models in analyzing complex healthcare data suggests their broader application in pandemic modeling and healthcare resource management, potentially shaping future pandemic response strategies.

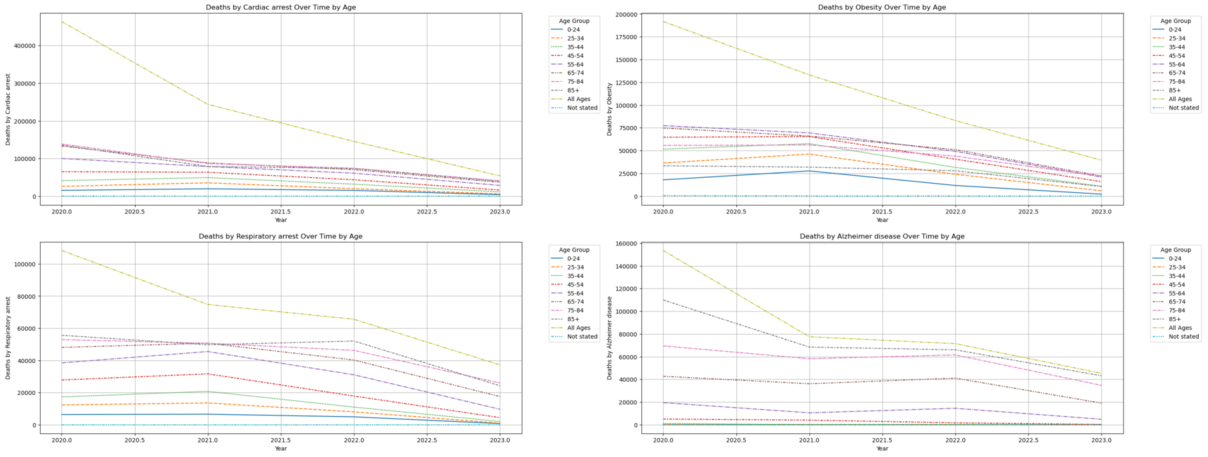

Figure 5: Result of most number of death by different reasons

5 Author Contribution

The project was successfully executed by a dedicated team of three individuals, each contributing their expertise to different facets of the analysis.

Shubh Mehta played a crucial role in the Exploratory Data Analysis (EDA) phase. His responsibilities included dataset imputation, data cleaning, integer encoding of multiple columns, and segmentation analysis based on age group, state, and conditions affecting COVID-19. Additionally, he diligently worked on various data manipulation techniques to balance and normalize the dataset. His efforts focused on reducing unwanted noisy data and identifying trends among different columns.

Sarthak Choudhary spearheaded the correlation analysis, uncovering multiple row relations highly correlated to COVID-19 deaths. He took charge of training the dataset and employed diverse algorithmic modeling techniques, ranging from neural networks like MLP to ensemble learning models like Gradient Boost. Sarthak optimized the training dataset, algorithms, and rigorously tested multiple algorithms to ensure their accuracy.

Niveditha Bommanahally Parmeshwarappa led the effort in discovering and acquiring relevant data. She proactively identified trends and statistical insights within the dataset, conducting a comprehensive time series analysis and contributing significantly to the overall data analysis process.

The collaborative efforts of Shubh Mehta, Sarthak Choudhary, and Niveditha Bommanahally Parmeshwarappa resulted in a well-rounded and thorough exploration of the data, providing valuable insights into the factors influencing COVID-19 outcomes.

References

[1] Krishnan Bhaskaran Seb Bacon Chris Bates Caroline E. Morton Helen J. Curtis Amir Mehrkar David Evans Peter Inglesby Jonathan Cockburn Helen I. McDonald Brian MacKenna Laurie Tomlinson Ian J. Douglas Christopher T. Rentsch Rohini Mathur Angel Y. S. Wong Richard Grieve David Harrison Harriet Forbes Anna Schultze Richard Croker John Parry Frank Hester Sam Harper Rafael Perera Stephen J. W. Evans Liam Smeeth Ben Goldacre Elizabeth J. Williamson, Alex J. Walker. *Factors associated with COVID-19-related death using OpenSAFELY.* PhD thesis, Nature, 2020.

[2] V. R. F. Cestari R. S. Florêncio T. M. M. Moreira G. J. B. Sousa, T. S. Garces and M. L. D. Pereira. Mortality and survival of covid-19. In *Proceedings of the Sample Conference*, volume 148. Cambridge University Press, 2021.

[3] Damien Bennett Lynsey Patterson Rachel Spiers David Gibson Hugo Van Woerden Anthony J. Bjourson Magda Bucholc, Declan Bradley. Identifying pre-existing conditions and multimorbidity patterns associated with in-hospital mortality in patients with covid-19, 2022.

[4] Mahdi Shakibi Mohammad Pourhomayoun. Predicting mortality risk in patients with covid-19 using machine learning to help medical decision-making. *Smart Health*, 20(100178), 2021.

[5] Md. Martuza Ahamad A. H. M. Kamal Jahidur Rahman Khan 4 Md. Protikuzzaman 1 Nasif Hossain 5ORCID A. K. M. Azad Julian M. W. Quinn Mathew A. Summers Teng Liaw Valsamma Eapen andMohammad Ali Moni Sakifa Aktar, Ashis Talukder. Machine learning approaches to identify patient comorbidities and symptoms that increased risk of mortality in covid-19, 2021.

[6] Md. Martuza Ahamad A. H. M. Kamal Jahidur Rahman Khan Md. Protikuzzaman Nasif Hossain Julian M.W. Quinn Mathew A. Summers Teng Liaw Valsamma Eapen Mohammad Ali Moni Sakifa Aktar, Ashis Talukder. *Machine Learning and Meta-Analysis Approach to Identify Patient Comorbidities and Symptoms that Increased Risk of Mortality in COVID-19.* Cornell University, 2020.