RAIS RESEARCH ASSOCIATION for INTERDISCIPLINARY STUDIES
November 21-22, 2024

DOI:10.5281/zenodo.14514169

Predicting COVID-19 Mortality Rates: An Analysis of Case Incidence, Mask Usage, and Machine Learning Approaches in U.S. Counties

Jacob Pratt[1], Serkan Varol[2], Serkan Catma[3]

[1]*University of Tennessee Chattanooga, Engineering Management and Technology Department, Chattanooga, Tennessee, U.S.A., nxm681@mocs.utc.edu*
[2]*University of Tennessee Chattanooga, Engineering Management and Technology Department, Chattanooga, Tennessee, U.S.A., serkan-varol@utc.edu*
[3]*University of South Carolina Beaufort, Business Administration, Bluffton, South Carolina, U.S.A., catma@uscb.edu*

ABSTRACT: The COVID-19 pandemic has necessitated the use of multidisciplinary approach to assess public health interventions. Data science has been widely utilized to promote interdisciplinary collaboration especially during the post-COVID era. This study uses a comprehensive dataset, including mask usage and epidemiological metrics from U.S. counties, to explore the correlation between public compliance with mask-wearing guidelines and COVID-19 mortality rates. After employing machine learning approaches such as linear regression, decision tree regression, and random forest regression, our analysis identified the random forest model as the most accurate model in predicting mortality rates due to its efficacy with the lowest error metrics. The models' performances were rigorously evaluated through error metric comparisons, highlighting the random forest model's robustness in handling complex interactions between variables. These findings provide actionable insights for public health strategists and policy makers, suggesting that enhanced mask compliance could significantly mitigate mortality rates during the ongoing pandemic and future health crises.

KEYWORDS: machine learning applications, predictive modeling for public health, COVID-19 analysis, pandemic, model comparison.

1. Introduction

During the COVID-19 pandemic, a crisis that has stretched global healthcare and public health systems to their limits, the scientific community has deployed a wide array of analytical tools and models to predict the course of the disease and evaluate the effectiveness of various containment measures. This study delves into the critical aspect of pandemic control: understanding the relationship between public mask-wearing practices and COVID-19 mortality rates, with a focus on predicting death rates (death per 100,000 individuals) using comprehensive datasets. Utilizing data on mask-wearing frequency from a survey conducted by Dynata, a survey company, this study seeks to provide insights into how behavioral practices such as mask usage can influence the pandemic's death toll at the county level across the United States.

A cornerstone of this research is the comparative analysis of three distinct predictive models: linear regression, random forest regression, and decision tree regression, each offering different strengths and perspectives in data analysis and prediction. By employing these models through the machine learning suite in Alteryx, a data analytics and visualization platform, the study not only aims to predict COVID-19 mortality rates with high accuracy, but also rigorously evaluates the performance of these models against each other. This comparative analysis is essential for identifying the most effective tool for epidemiological prediction based on available data, which can significantly influence public health decision-making and policy formulation.

The findings reveal that the random forest regression model provided the most accurate and robust predictions among the evaluated models, demonstrating lower error metrics and a high degree of reliability in capturing the complex interplay of factors affecting COVID-19 mortality rates. The study significantly highlights the influence of mask-wearing behaviors on

mortality outcomes, providing actionable insights that can inform public health strategies and policy decisions.

The significance of this study extends beyond its immediate findings. By providing a methodical comparison of predictive modeling techniques in the context of a public health crisis, it contributes to the broader field of epidemiological modeling, offering insights that could be pivotal for managing current and future pandemics. Through this analysis, the research underscores the profound impact of public compliance with health guidelines, such as mask-wearing, on controlling the spread of the virus. As the world still grapples with the ongoing challenges of COVID-19, the findings of this study aim to guide policymakers, public health officials, and the public in making informed decisions to mitigate the pandemic's impact effectively.

2. Literature Review

Research in the domain of COVID-19 predictive modeling has extensively utilized various regression and machine learning techniques to forecast the trajectory of the pandemic and assess intervention strategies. Khan et al. (2022) explored multiple regression models, including linear and support vector regression, to predict COVID-19 cases, demonstrating the utility of these models in handling complex epidemiological data, albeit with varying degrees of accuracy across different models. Similarly, Rohini et al. (2021) evaluated different machine learning algorithms, highlighting the superior predictive capability of K-nearest neighbors for COVID-19 severity predictions based on diverse metrics, though noting that some models were less effective with large datasets.

Mary and Raj (2021) assessed the effectiveness of various machine learning algorithms, such as Support Vector Machine (SVM) and K-Nearest Neighbor (KNN), for predicting COVID-19 cases. They found that SVM, with an accuracy of 85.2%, outperformed others in terms of precision and recall, confirming its robustness in predicting disease spread.

The comparative analysis of predictive models has been a focal point of recent studies. Majhi et al. (2020) utilized regression-based, decision tree-based, and random forest models, concluding that random forest algorithms provided the most accurate predictions due to their ability to manage nonlinear relationships and complex interactions within datasets. This finding aligns with the work by Shaikh et al. (2021), who emphasized the efficacy of polynomial regression over linear models, especially with higher degree polynomials providing more accurate predictions in complex scenarios such as multiple data sets across different regions.

Despite the advancements, several studies have identified key challenges and limitations within existing models. For instance, Mandayam et al. (2020) compared linear regression and support vector regression, uncovering significant discrepancies in predictive accuracy, with linear regression outperforming in simpler, linear datasets. This inconsistency in model performance across different types of data underlines the critical need for robust model evaluation and selection strategies.

Incorporating socio-economic factors into predictive models has emerged as a crucial aspect of recent research. Almalki et al. (2022) focused on the impact of food access and health issues on COVID-19 infections and deaths, utilizing regression analysis to establish a clear link between these socio-economic factors and pandemic outcomes. This approach is vital for understanding the broader implications of the pandemic and for designing targeted public health interventions.

Further enhancing the predictive accuracy of models through advanced techniques such as hyperparameter tuning and feature selection has shown promising results. Kaliappan et al. (2021) emphasized the role of machine learning techniques like XGBOOST and Random Forest in predicting the COVID-19 reproduction rate, noting that hyperparameter tuning significantly improved the model's performance.

Although not related to COVID-19, the study by Acharya, Armaan, and Antony (2019) compared various regression models in a different context—predicting graduate admissions. They highlighted that Linear Regression performed best in their analysis, which aligns with the findings in public health models where socio-economic and behavioral predictors are integral. This parallel between admissions and public health underscores the versatility and effectiveness of regression analysis across different domains of prediction.

While considerable progress has been made in predictive modeling of COVID-19, there remains a gap in integrating behavioral data, such as mask usage, with epidemiological models at a granular level, such as county-specific analysis in the United States. Additionally, the dynamic nature of the pandemic, with evolving virus strains and varying public health responses, calls for adaptive models that can incorporate real-time data and reflect current realities more accurately. Building on these insights, this study introduces an integration of mask-wearing data with traditional epidemiological data to predict COVID-19 mortality rates using advanced machine learning techniques. This approach not only enhances the granularity and relevance of predictive modeling in public health but also provides actionable insights for policymakers to tailor interventions more effectively.

3. Methodology

3.1. Data

The analysis draws on two datasets provided by the New York Times GitHub repository, focusing on the effects and responses of COVID-19 within the United States. The first dataset, us-counties-2021, captures daily counts and seven-day rolling averages of new COVID-19 cases and deaths per 100,000 residents at the county level for the year 2021. This dataset is vital for understanding the temporal trends of the pandemic, incorporating data adjustments for reporting anomalies sourced from state and national health agencies, as explained by the New York Times: "*The data in these files...is derived from the difference in cumulative cases from one day to another"* (Nytimes, 2021). The second dataset, mask-use-by-county, derived from a survey conducted by Dynata, where they ask participants how often they wear a mask when they are around other people, provides estimates of mask usage across U.S. counties. This dataset records the frequency of mask usage and reflects public health behavior during the pandemic, specifically between July 2 and July 14.

The integration of these datasets involved several steps to align and merge the data for comprehensive analysis. Given that the mask usage data represents a specific snapshot in time (July 2 to July 14), the seven-day rolling averages of deaths per 100,000 residents (deaths_avg_per_100k) from the us-counties-2021 dataset were averaged over these dates to match the period of the mask usage survey.

Using a GEO-ID to FIPS master key facilitated the matching of each county's geographic identifier from the us-counties-2021 dataset with the corresponding FIPS code from the mask usage dataset, ensuring that the health data corresponded accurately with the behavioral data for the same geographical regions.

The datasets were then merged based on county and state identifiers. The combined dataset included the geographic identifier (GEOID), county, state, daily average cases per 100,000 residents (**cases_avg_per_100k**), and daily average deaths per 100,000 residents (deaths_avg_per_100k), alongside the proportions of respondents who never, rarely, sometimes, frequently, or always wear masks. Additionally, any incomplete or anomalous records that could potentially skew the analysis were carefully removed to maintain the integrity of the data.

The descriptive table for this dataset is summarized below in Table-1, and provides a statistical summary of key variables, including the rolling average of deaths per 100,000 residents—the primary dependent variable of this study—as well as the frequency categories of mask usage, which serve as predictors. This arrangement not only aids in understanding the

distribution of the variables but also sets the stage for examining the potential correlations between mask-wearing habits and COVID-19 death rates at the county level. The table details max, min, and average values for each field, offering insights into the range and central tendencies of the data.

Table 1. Variable Descriptive Table

Attributes	***Description***	***Value***	***Field Summary***
Deaths_avg_per_100k (Dependent Variable)	Rolling average of deaths per 100 thousand residents in county	Continuous	Max: 4.7 Min: 0.006 Avg: 0.246
Cases_avg_per_100k	Rolling average of cases per 100 thousand residents in county	Continuous	Max: 121.947 Min: 0 Avg: 6.181
Never	Fraction of county population that never wear a mask	Continuous-Bounded (0-1)	Max: 0.432 Min: 0 Avg: 0.077
Rarely	Fraction of county population that rarely wear a mask	Continuous-Bounded (0-1)	Max: 0.365 Min: 0 Avg: 0.076
Sometimes	Fraction of county residents that sometimes wear a mask	Continuous-Bounded (0-1)	Max: 0.422 Min: 0.001 Avg: 0.117
Frequently	Fraction of county population that frequently wear a mask	Continuous-Bounded (0-1)	Max: 0.483 Min: 0.029 Avg: 0.203
Always	Fraction of county population that Always wear a mask	Continuous-Bounded (0-1)	Max: 0.88 Min: 0.146 Avg: 0.529

3.2. Random Forest Regression Model

The Random Forest Regression model was selected as one of the models in this research to predict average deaths per 100,000 people, utilizing predictors such as average cases per 100,000 and the frequency of mask usage (never, rarely, sometimes, frequently, always). This ensemble learning approach aggregates the predictions from multiple decision trees, enhancing the model's ability to capture complex interactions between variables Dong et al. (2019). The Random Forest model's effectiveness in synthesizing diverse perspectives into a coherent prediction aligns with the wisdom of the crowd theory, demonstrating the model's capacity for high accuracy and robustness in predictive tasks (Peterek et al., 2013).

In implementing the Random Forest Regression model within Alteryx for this study, the model was trained using an 80/20 data split—80% for training and 20% for validation to ensure a comprehensive evaluation of its performance on unseen data. Specific parameters were carefully chosen to tailor the model to the data and predictive objectives. The model was set up with 100 estimators, creating a diverse ensemble of trees to improve the prediction accuracy. Each tree was allowed to grow to a maximum depth of 10 levels, a constraint that helps prevent overfitting by limiting the complexity of the models. The mean squared error criterion was

applied to guide the decision-making process within each tree, focusing on minimizing variance for more accurate predictions. Furthermore, a minimum sample split of 2 was specified, ensuring that at least two samples are required to split a node, which aids in maintaining the generalizability of the model. This combination of parameters was chosen to optimize the balance between model complexity and predictive performance.

3.3. Decision Tree Regression Model

The Decision Tree Regression model was selected for its robust predictive capabilities, particularly for forecasting average deaths per 100,000 people using numerical predictors. This included variables like average cases per 100,000 and coded frequencies of mask usage, transformed into numerical values to indicate the frequency categories (never, rarely, sometimes, frequently, always). Decision Trees categorize or approximate outcomes by sequentially testing input variables, a method that is especially effective when dealing with numerical data. This approach, based on constructing a model through a series of binary decisions, leverages the most informative splits within the dataset to optimize prediction accuracy, a principle foundational to greedy learning algorithms (Blockeel et al. 2023).

Implemented within Alteryx, the Decision Tree Regression model was precisely configured to harness the full potential of its predictive accuracy. The model training utilized an 80/20 data split—80% for training and 20% for validation. This setup aimed to ensure a comprehensive evaluation of the model's performance on data. The selection of the mean squared error criterion was crucial for directing the decision-making process within the tree, aiming to minimize variance at each decision node and thereby refine the precision of predictions. To curb the complexity and prevent overfitting, a maximum tree depth of 10 was established. The consistency and reproducibility of the model's outcomes were guaranteed by initializing the construction process with a random seed of 2. Moreover, the criterion for further splitting a node—the requirement of a minimum of 2 samples—was set to foster a detailed yet generalizable learning from the data. The strategic application of these parameters, derived from the principles of recursive partitioning and heuristic optimization, was instrumental in developing a model adept at navigating the intricacies of numerical predictors within a regression context (Bertsimas et al. 2017).

3.4. Linear Regression Model

The Linear Regression model employed in this research was designed to predict average deaths per 100,000 people based on six numerical predictors: average cases per 100,000, and the frequency of mask-wearing categorized into five levels and quantified numerically. Although the relationship between these independent variables and the dependent variable was not perfectly linear, linear regression was chosen to facilitate comparison with the other models used in this study. This method assumes a linear relationship between the dependent and independent variables. Initial analysis typically involves scatter plots to visually assess the linearity of these relationships, ensuring that linear regression is suitable for modeling as non-linear relationships would require alternative statistical methods (Schneider et al. 2010). This essential step verifies that the model is appropriate for the data, even with the acknowledged imperfections in linearity.

This model was trained using an 80/20 split for the training and validation data, consistent with the methodology applied to the other models in this study. In this configuration, the model constructs the predicted outcome as a weighted sum of the predictors, with each weight quantifying the expected change in the mortality rate for a unit change in the respective predictor. This structure allows for direct assessment of each predictor's influence on the estimated mortality rates, providing critical insights into the factors significantly impacting the dependent variable. Schneider et al. (2010) emphasize the importance of understanding these relationships within the model's context and the characteristics of the data, aiding both in prediction accuracy and in the understanding of public health dynamics.

4. Results

4.1. Error Comparisons

In this comprehensive study, three distinct regression models—Linear Regression, Decision Tree Regression, and Random Forest Regression—were utilized to explore their effectiveness in forecasting COVID-19 mortality rates based on variables such as mask usage frequency and case incidence. The primary goal was to assess which models provide the most accurate and reliable forecasts that could significantly inform public health decisions. Among these, the Random Forest Regression model stood out for its superior performance, achieving the lowest Mean Squared Error (MSE) of 0.08483, Root Mean Squared Error (RMSE) of 0.29125, and Mean Absolute Error (MAE) of 0.15368, indicating high accuracy and robustness in its predictive capability. Despite this, its Model Absolute Percentage Error (MAPE) of 138.99 was higher than that of the Linear Regression model, suggesting some limitations in terms of percentage error relative to actual mortality rates.

In comparison, the Decision Tree Regression model displayed significant variability, with an MSE of 0.20111, an RMSE of 0.44845, and an MAE of 0.20794, suggesting a potential overfitting issue that could affect its reliability in diverse datasets. The Linear Regression model, achieving a lower MSE of 0.09181 and an RMSE of 0.30299, demonstrated good generalization capabilities. However, its MAPE of 117.854, although lower than that of the Random Forest model, suggests better performance in percentage terms but indicates a challenge in capturing more complex relationships within the data due to higher MSE and RMSE values. The performance metrics for each model are summarized in Figure 1, setting the stage for a detailed examination of their predictive behaviors.

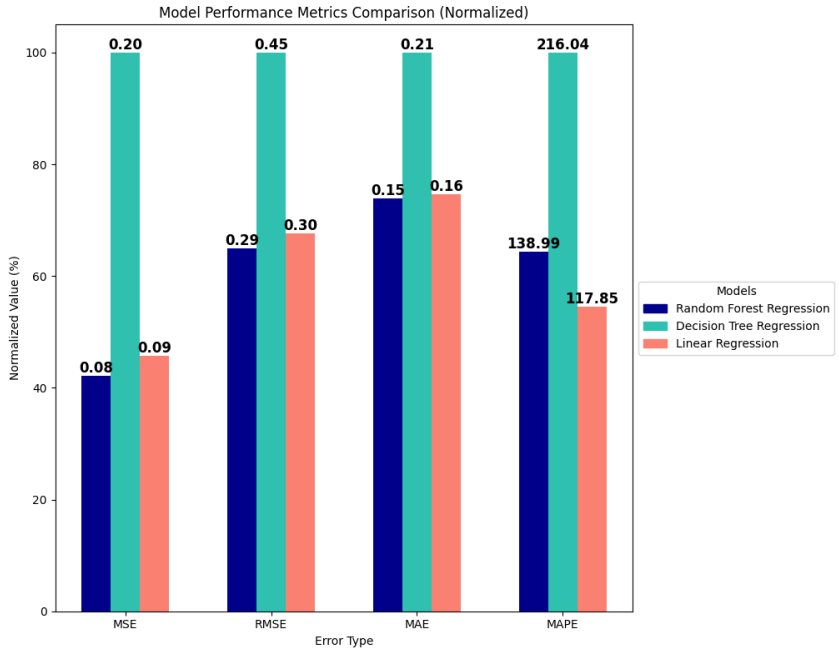

Figure 1. Error Comparison Of Models (MSE, RMSE, MAE, MAPE)

4.2. Predicted vs Actual Scatterplots

In this section, the predictive performance of three regression models—Linear Regression, Decision Tree Regression, and Random Forest Regression—is visually examined. The scatterplots (Figures 2, 3, and 4) provide a graphical representation of predicted versus actual COVID-19 mortality rates, which is crucial for assessing the practical utility of each model in public health scenarios.

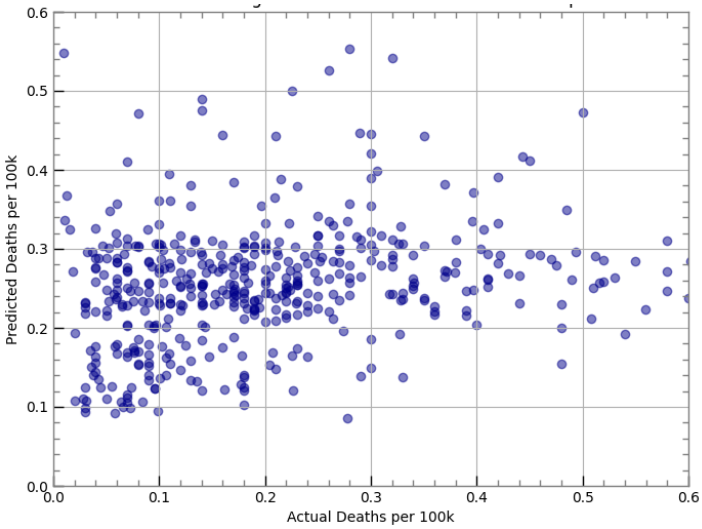

Figure 2. Predicted vs Actual Deaths Sample Scatterplot (Random Forest Regression Model)

The scatterplot for the Random Forest Regression model illustrates a well-balanced distribution of predictions. Notably, data points cluster more closely around the actual mortality rates, which implies a strong model capability in capturing the variations within the data without significant overfitting. This makes it particularly reliable for predictive tasks in public health. It's important to note that for clearer visualization, we restricted the axis to 0.6 to exclude large outliers that obscure the majority of the data. This adjustment ensures the focus remains on the most relevant results, reflecting the model's effectiveness in integrating the complexities of epidemiological data. These characteristics underscore the Random Forest model's robust predictions that can support nuanced public health strategies.

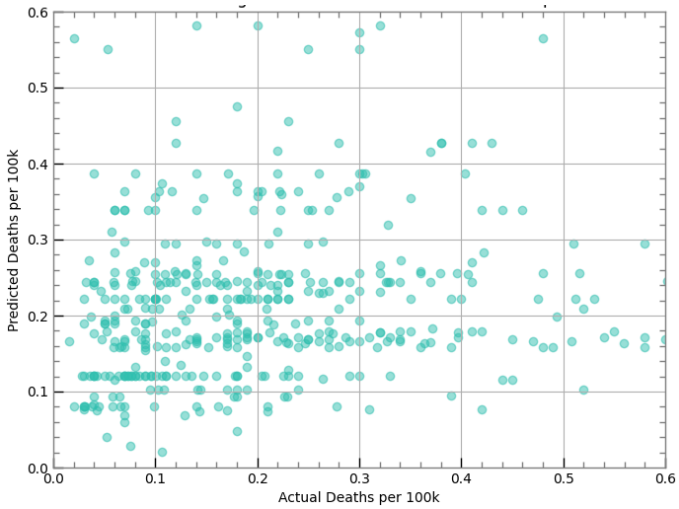

Figure 3. Predicted vs Actual Deaths Sample Scatterplot (Decision Tree Regression Model)

Conversely, the scatterplot for the Decision Tree Regression model, as shown in Figure 3, reveals a broader spread of predictions. This wide variation underscores the model's sensitivity to specific data characteristics, which results in highly variable predictions across different data points. Such variability may reduce the model's effectiveness in scenarios that demand consistent and reliable forecasts, particularly when managing diverse epidemiological conditions. The axis here is also limited to 0.6 to display the central mass of data and mitigate the impact of outliers on the visualization more effectively.

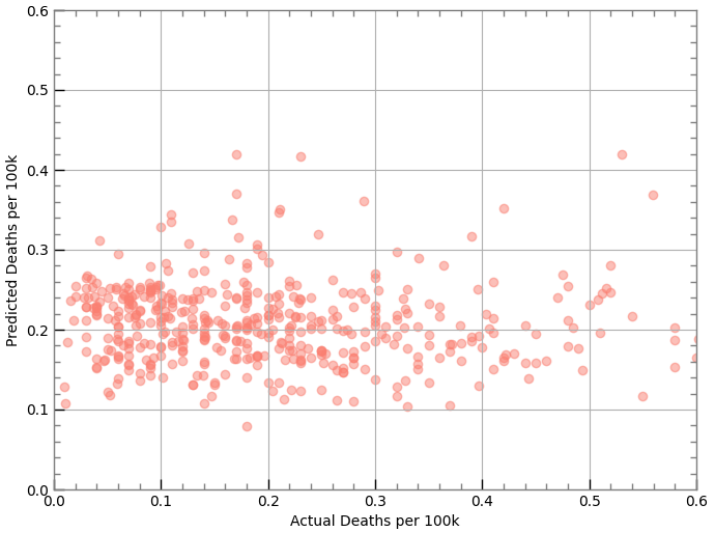

Figure 4. Predicted vs Actual Deaths Sample Scatterplot (Linear Regression Model)

Lastly, the scatterplot for the Linear Regression model (Figure 4) displays tighter clustering of predictions around the mean mortality rates. This pattern indicates that while the model is stable across different datasets, it may underestimate actual mortality rates, particularly in scenarios of higher severity. This consistency near the average, however, suggests that the Linear Regression model, despite its stability, might not fully capture the peaks in mortality trends which are crucial for planning emergency responses and allocating healthcare resources during critical times. The decision to limit the axis to 0.6 was similarly motivated by the desire to focus on the most interpretable part of the data, avoiding distraction from extreme outliers.

These visualizations not only illustrate the strengths and weaknesses of each model but also highlight the importance of choosing the right model based on the specific needs and characteristics of the data at hand. By understanding these dynamics, we can better tailor our models to deliver reliable and actionable insights for public health decision-making.

4.3. Gini Impurity and Ordinary Least Squares

The next results we delved into were derived from evaluating each feature using the criteria provided by Assisted Modeling in Alteryx, which utilizes both Gini Impurity and Ordinary Least Squares (OLS) measures as seen in Table 2 below. This process involves selecting the higher value of the two for each feature, either Gini or OLS, and then assessing whether this value falls within an acceptable range indicative of its predictive strength. This methodological approach is essential for confirming the reliability and effectiveness of the predictors used in our regression models.

Table 2. Gini Impurity and Ordinary Least Squares of the Variables

Feature	***Gini Impurity***	***Ordinary Least Squares***
Cases_avg_per_100k	16.23	9.60
NEVER	17.65	15.85
RARELY	16.20	10.88
SOMETIMES	16.59	14.38
FREQUENTLY	15.33	5.36
ALWAYS	17.99	17.33

For instance, the feature Cases_avg_per_100k demonstrated a Gini Impurity as its higher value at 16.23. This places it solidly within the "good predictor" range of 2 to 50. This indicates that the feature substantially enhances the model's ability to forecast COVID-19 mortality rates,

doing so in a manner that avoids the pitfalls of overfitting, thereby maintaining the model's general applicability across different datasets. Such balance is crucial for preserving the model's utility in varied epidemiological scenarios.

The NEVER feature similarly showed a high Gini Impurity of 17.65 as its higher measure, placing it within the same "good predictor" range. This demonstrates its strong, yet balanced, influence on model outcomes, ensuring that the feature contributes positively to the predictive accuracy without becoming excessively dominant, which could potentially limit the model's broader applicability.

Moreover, the feature ALWAYS had the highest Gini measure among all the features at 17.99. While it is the highest, it still falls under the "good predictor" category. This underscores its significant contribution to the model, enhancing its ability to accurately capture the complexities of the data. Importantly, since it does not escalate into the "highly associated" or "too highly associated" categories, it remains a crucial predictor, avoiding overdominance that could skew predictions and detract from the model's effectiveness across different public health contexts.

4.4. Residual Comparison

Lastly, the residuals were found by calculating the differences between the predicted deaths from each of the regression models and the actual observed deaths. Residuals are crucial for diagnosing the fit of a regression model, as they reflect the amount by which the prediction deviates from the actual data points. A residual plot, such as the one presented in figure 5, visually represents these differences.

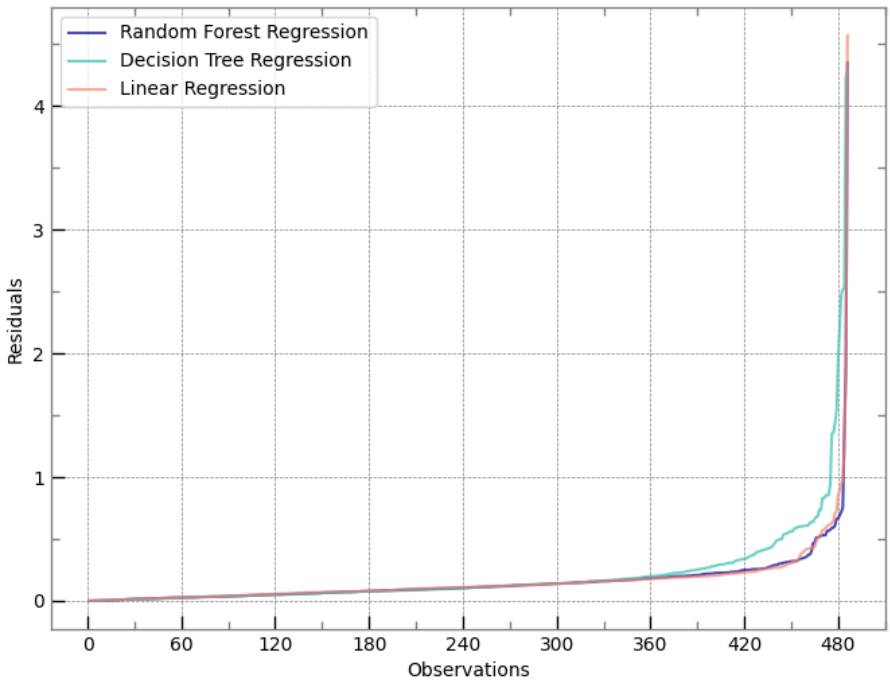

Figure 5. Residual vs Observation Plot for all the Models

To create the plot, the residuals were sorted from smallest to largest to understand the distribution of errors across the dataset. This is a common practice to identify patterns in the residuals that might indicate problems with the model, such as non-linearity or heteroscedasticity. The sorted residuals allow us to see if there is an increase in error for certain types of predictions, which can be indicative of model performance issues. The x-axis of the plot shows the observations in their sorted order, while the y-axis displays the absolute magnitude of the residuals. The plot features three lines, each representing one of the regression models: Linear Regression, Decision Tree Regression, and Random Forest Regression.

In the residual plot, the majority of observations show low residuals for all three models, indicating a generally close prediction to actual mortality rates. However, towards the higher end of observations, we observe an uptick in the residuals across the board. This increase suggests difficulties in accurately predicting higher mortality rates, which could be attributed to outliers or complex aspects not well-captured by the models. Notably, the residuals of the Decision Tree Regression model are significantly higher compared to the Random Forest and Linear Regression models in this range. This stark increase for the Decision Tree model points to potential overfitting or an insufficient model complexity to handle these extreme values. On the other hand, the Random Forest Regression model demonstrates a slightly better fit overall, with lower residuals that corroborate its superior metrics such as MSE and MAE reported earlier. The residual plot is a vital tool for diagnosing model performance, clearly indicating where improvements are needed and assisting in choosing the most effective model for predicting COVID-19 mortality rates, with the Random Forest model showing the most promise in this respect.

5. Discussion and Conclusion

This study utilizes machine learning models to predict COVID-19 mortality rates by incorporating case incidence data and mask usage frequencies across U.S. counties. While the Random Forest Regression model demonstrated adeptness in handling nonlinear relationships and complex interactions, its mean squared error (MSE) of 0.08483 only marginally improves upon the Linear Regression model's MSE of 0.09181. This modest difference underscores the effectiveness of Linear Regression in scenarios where relationships might approximate linearity more closely than anticipated.

The analysis of important variables revealed that mask-wearing frequencies, particularly the 'Always' category, significantly influenced model predictions, as indicated by its high Gini impurity score. However, the scatterplots provided a visual affirmation of model performance, showing that the Random Forest and Linear Regression models generally produced predictions that closely aligned with actual mortality rates. Notably, these models managed to cluster around the actual values with minimal spread, suggesting effective capture of the central tendency but at times failing to predict more extreme mortality rates accurately.

Residual plots from the study highlight areas where predictions diverge from actual outcomes. Especially in the Decision Tree model, large residuals were observed for outliers with high mortality rates, indicating potential overfitting or the model's sensitivity to extreme values. This effect was less pronounced in the Random Forest and Linear Regression models, yet the presence of outliers similarly influenced their performance. Error metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) also varied across models, reinforcing the necessity to balance model complexity with predictive accuracy.

Enhancements in handling outliers could significantly improve the accuracy of predictions. Employing robust regression techniques or selectively filtering outliers may help stabilize predictions across varied datasets. For the Linear Regression model, refining data preprocessing to enhance linear relationships or using variable selection techniques might optimize its performance. Increasing the granularity of mask usage data with a larger sample size per county could also provide more precise insights into behavioral impacts on mortality rates. Furthermore, tuning the complexity settings of the Decision Tree model might mitigate overfitting, possibly through parameter adjustments such as limiting maximum tree depth or increasing the minimum samples per leaf.

The findings contribute to the growing body of literature that evaluates the efficacy of machine learning models in epidemiological predictions. Similar to the work of Majhi et al. (2020), this research confirms the robustness of Random Forest models in handling complex data structures. Almalki et al. (2022) also emphasized the integration of socio-economic factors into predictive models, a concept partially explored here through behavioral data. The study’s

results resonate with those of Mary and Raj (2021), who found machine learning algorithms effectively predicting disease spread, advocating for their broader application in public health.

The comparative analysis of machine learning models in this study not only advances the understanding of predictive modeling in public health but also underscores the practical utility of integrating behavioral data into epidemiological studies. As global health challenges continue to evolve, the insights provided by this research are essential for enhancing public health response strategies, ultimately contributing to safer, more informed communities. Future improvements in model application and data integration can further augment the reliability and applicability of predictive models in public health decision-making.

References

Acharya, Mohan S., Asfia Armaan, and Aneeta S. Antony. 2019. "A comparison of regression models for prediction of graduate admissions." *International Conference on Computational Intelligence in Data Science (ICCIDS)*. https://doi.org/10.1109/iccids.2019.8862140.

Almalki, Abrar, Balakrishna Gokaraju, Yaa Acquaah, and Anish Turlapaty. 2022. "Regression analysis for COVID-19 infections and deaths based on food access and health issues." *Healthcare 10*(2): 324. https://doi.org/10.3390/healthcare10020324.

Bertsimas, Dimitris, Jack Dunn, and Aris Paschalidis. 2017. "Regression and classification using optimal decision trees." *IEEE MIT Undergraduate Research Technology Conference (URTC)*. https://doi.org/10.1109/urtc.2017.8284195.

Blockeel, Hendrik, Laurens Devos, Benoît Frénay, Géraldin Nanfack, and Siegfried Nijssen. 2023. "Decision trees: From efficient prediction to responsible AI." *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1124553.

Dong, Penghao, Huachen Peng, Xianqiang Cheng, Yan Xing, Xin Zhou, and Dongliang Huang. 2019. "A random forest regression model for predicting residual stresses and cutting forces introduced by turning in718 alloy." *IEEE International Conference on Computation, Communication and Engineering (ICCCE)*. https://doi.org/10.1109/iccce48422.2019.9010767.

Kaliappan, Jayakumar, Kathiravan Srinivasan, Saeed Mian Qaisar, Karpagam Sundararajan, and Chuan-Yu Chang. 2021. "Performance evaluation of regression models for the prediction of the COVID-19 reproduction rate." *Frontiers in Public Health*, *9*. https://doi.org/10.3389/fpubh.2021.729795.

Khan, Mohammad Ayoub, Rijwan Khan, Fahad Algarni, Indrajeet Kumar, Akshika Choudhary, and Aditi Srivastava. 2022. "Performance evaluation of regression models for COVID-19: A statistical and Predictive Perspective." *Ain Shams Engineering Journal 13*(2): 101574. https://doi.org/10.1016/j.asej.2021.08.016.

Majhi, R., Thangeda, R., Sugasi, R. P., & Kumar, N. 2020. "Analysis and prediction of covid-19 trajectory: A machine learning approach." *Journal of Public Affairs 21*(4). https://doi.org/10.1002/pa.2537.

Mandayam, Ashish U., A. C. Rakshith, S. Siddesha, and S. K. Niranjan. 2020. "Prediction of covid-19 pandemic based on regression." *Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*. https://doi.org/10.1109/icrcicn50933.2020.9296175.

Mary, L. William, and S. Albert Antony Raj. 2021. "Machine learning algorithms for predicting SARS-COV-2 (COVID-19) – a comparative analysis." *2nd International Conference on Smart Electronics and Communication (ICOSEC)*. https://doi.org/10.1109/icosec51865.2021.9591801.

Nytimes. 2021. *Covid-19-data/rolling-averages/US-counties-2021.csv at master · Nytimes/covid-19-DATA*. GitHub. https://github.com/nytimes/covid-19-data/blob/master/rolling-averages/us-counties-2021.csv.

Peterek, Tomáš, Pavel Dohnálek, Petr Gajdoš, and Maroš Šmondrk. 2013. "Performance evaluation of random forest regression model in tracking parkinson's disease progress." *13th International Conference on Hybrid Intelligent Systems (HIS 2013)*. https://doi.org/10.1109/his.2013.6920459.

Rohini, M., K. R. Naveena, G. Jothipriya, S. Kameshwaran, and M. Jagadeeswari. 2021. "A comparative approach to predict corona virus using machine learning." *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*. https://doi.org/10.1109/icais50930.2021.9395827.

Schneider, Alexandra, Gerhard Hommel, and Maria Blettner. 2010. "Linear regression analysis." *Deutsches Ärzteblatt International*. https://doi.org/10.3238/arztebl.2010.0776.

Shaikh, Saud, Jaini Gala, Aishita Jain, Sunny Advani, Sagar Jaidhara, and Mani Roja Edinburgh. 2021. "Analysis and prediction of COVID-19 using regression models and time series forecasting." *11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. https://doi.org/10.1109/confluence51648.2021.9377137.