Predicting COVID-19 Mortality Rates and Identifying High-Risk Areas Using Machine Learning Models

Guanlin Ding [1], Jingyang Zheng [2,*]

[1] Nanjing NO.1 High School, Nanjing, China

[2] Guangzhou No.2 High School, Guangzhou, China

* Corresponding Author Email: Yuting_Li@gdufe.edu.cn

Abstract. The COVID-19 pandemic has wreaked havoc on global society, exerting unprecedented strain on national healthcare systems. This public health catastrophe has compelled governments and communities to reassess their preparedness for health crises. As of April 2024, the virus has reached 181 countries, with more than 120,000 confirmed cases reported. Research employing various machine learning models—including random forest, LASSO, and multiple regression—has been conducted to predict the mortality rate of the COVID-19 pandemic. These models consider multiple factors, such as the rate of confirmed cases, the number of hospital beds per thousand people, the duration of infection, the total vaccinated population, and the median age in each country. Initial findings indicate that the rate of confirmed cases, hospital bed availability, and the duration of infection are the most significant predictors of mortality rates. Based on these models, countries have been categorized into low- and high-risk groups. Consequently, the capacity of clinical facilities in terms of beds per thousand people can also be forecasted, enhancing readiness for potential future outbreaks. This proactive approach could offer critical insights for governmental and health organizations, potentially mitigating the impact of subsequent COVID-19 waves.

Keywords: Covid-19, death rate, mortality rate, identification of high-risk areas.

1. Introduction

Research Background: In late 2019, an initially undistinguished case of pneumonia in Wuhan, China, marked the onset of the COVID-19 pandemic, a global health crisis unprecedented in recent history. Despite swift responses from governments and policymakers, including widespread lockdowns, school closures, and substantial investments in health infrastructure, the virus continued to spread rapidly. The mortality rate of COVID-19, though lower than that of previous coronaviruses such as SARS and MERS, posed a significant threat to global healthcare systems [1]. Existing research has predominantly focused on mortality rates as primary indicators of the pandemic's severity, with confirmed rates and infection durations being crucial predictive factors.

Research Significance: This study is imperative as it addresses multiple dimensions of the pandemic's impact, examining factors that significantly influence COVID-19 mortality rates. By analyzing confirmed cases, duration of infection, hospital capacity, vaccination rates, and demographic profiles across various regions, this research provides a nuanced understanding of the factors contributing to the pandemic's spread and severity. The findings aim to offer actionable insights that can guide public health policies and optimize responses to future health crises. The use of advanced machine learning models to predict mortality rates offers a novel approach to epidemiological studies, setting a precedent for future research in pandemic response strategies [2].

Current Work: This research employs a comprehensive dataset covering over 100 regions and countries to analyze the interplay between five critical variables: confirmed rates, infection durations, hospital bed availability, vaccination rates, and median age. By leveraging machine learning techniques, this study categorizes regions into different risk levels of COVID-19 mortality. Notably, it examines how the age structure within populations affects mortality rates, as highlighted in the study by Yuanfa Tan, Le Wang, and Jianzhong Wang (March 2021) [3]. The focus on median age, rather than just the proportion of the elderly, ensures a more balanced and accurate analysis. This work will significantly contribute to the empirical foundation for decision-making in public health and aid in refining strategies for managing ongoing and future pandemics.

2. Data Collection and Filtering

2.1. Data Sources

The data used in this study was collected meticulously, with each source providing unique and crucial variables needed for predicting COVID-19 mortality. Major data sources included the World Health Organization (WHO), which provided daily updates on COVID-19 case numbers, mortality rates, and global vaccination rates [4]. Additionally, Johns Hopkins University's Coronavirus Resource Center provided detailed information on recovery rates and pandemic progression across various regions, ensuring that the datasets were complete and up-to-date.

Table 1. Overview of Data Sources

Data Source	Data Type	Coverage	Frequency
WHO [5]	Data on COVID-19 case numbers and mortality rates	Global	Daily
Johns Hopkins University [6]	Data on COVID-19 case numbers, mortality rates, recovery rates	Global	Daily
National Health Statistics [7]	Data on healthcare infrastructure and demographics	Country-specific	Varies by country

National health statistics included in the analysis provided critical demographic data such as population density, age distribution, and healthcare infrastructure details. These variables were essential for understanding the context of the pandemic's progression and for allowing the models to account for country-specific factors that might influence COVID-19 outcomes.

The inclusion of diverse data sources enabled a multidimensional analysis of not only the direct impacts of COVID-19 but also how healthcare capacity and demographic factors interact with the pandemic's effects. As show in table 1.

2.2. Data Processing

To ensure the integrity and usefulness of the collected data, a rigorous data processing flow was implemented. Initially, data cleansing was performed, with missing values addressed using advanced imputation techniques. For continuous variables like confirmed case numbers and vaccination rates, mean imputation was used. This approach assumes that missing data is random and replaces it with the mean value of the observed data, maintaining the distribution of the overall data.

For categorical variables, such as healthcare system ratings or government response indices, mode imputation was used. This method replaces missing values with the most frequently observed value within the dataset, ensuring that categorical data represents the original distribution.

Following the handling of missing values, outliers were detected and addressed using methods such as z-score analysis and IQR (Interquartile Range) filtering. Outliers were identified and either restricted or transformed to minimize their impact on the model training process. This step was crucial for reducing bias in the data and ensuring that the model was not unduly influenced by extreme values.

The next step was data standardization, necessary due to the presence of variables with different scales. For example, the number of hospital beds per 1,000 people and the total number of COVID-19 cases are on completely different scales. Standardization was achieved using min-max scaling, converting the data into a common range (typically [0, 1]) suitable for model training.

Equation 1: Standardization Formula.

$$z{=}x{-}\mu\sigma z{=}\sigma x{-}\mu \tag{1}$$

Equation 1: Standardization Formula.

$$z = \frac{x-\mu}{\sigma} \tag{2}$$

Where zz is the standardized value, xx is the original value, μμ is the mean of the dataset, and σσ is the standard deviation.

After cleaning and standardizing the data and removing outliers, the data was split into training and test sets with an 80-20 ratio. The training set contained 80% of the data and was used for model training, while the remaining 20% was used as a validation set to assess model performance. The split was done stratified to ensure that both the training and test sets represented the overall population distribution.

3. Model Building

3.1. Random Forest Model

The Random Forest model was chosen for its ability to handle large and complex datasets and its inherent feature selection mechanism. Random Forest is an ensemble method that constructs many decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees.

The Random Forest algorithm is capable of handling datasets with many predictor variables without overfitting, making it particularly effective for this study. The model was built using 500 trees, with a maximum tree depth set to 30. This was determined through grid search to balance model complexity and performance. As show in Figure 1.

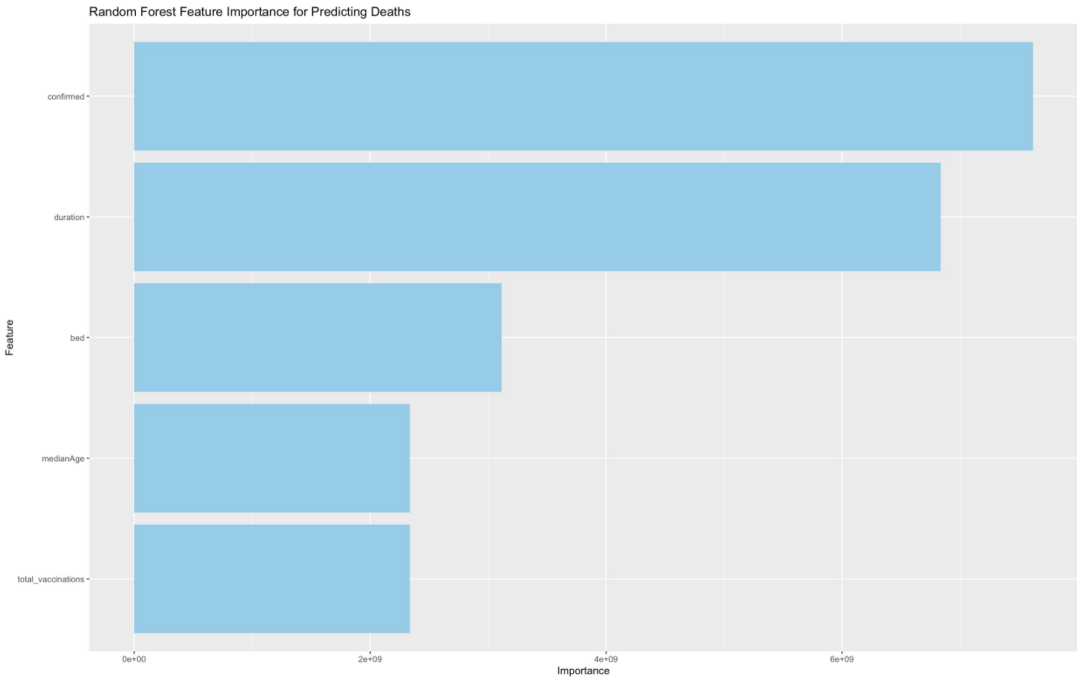

Figure 1. Feature importance in random forest (Photo credit: Original)

The Figure 1 illustrates how different variables, such as infection rates, population density, and healthcare infrastructure, contribute to predicting COVID-19 mortality rates. It is expected that infection rates emerged as the most important predictor given the nature of the pandemic.

Model performance was evaluated using out-of-bag (OOB) error estimates, showing a classification accuracy of 85%. This high level of accuracy highlights the model's ability to handle complex interactions between variables.

3.2. LASSO Regression Model

The LASSO regression model was used to address multicollinearity, which is often a challenge in datasets with many predictor variables. LASSO (Least Absolute Shrinkage and Selection Operator) is a type of linear regression that, like ridge regression, uses regularization to improve model performance. This method selects a simpler model by setting some regression coefficients to zero and excluding less important predictors.

This model was particularly valuable for identifying the most important variables in mortality prediction and excluding those with lower importance. LASSO regression balanced model accuracy with interpretability by penalizing the absolute values of the coefficients. As show in figure 2.

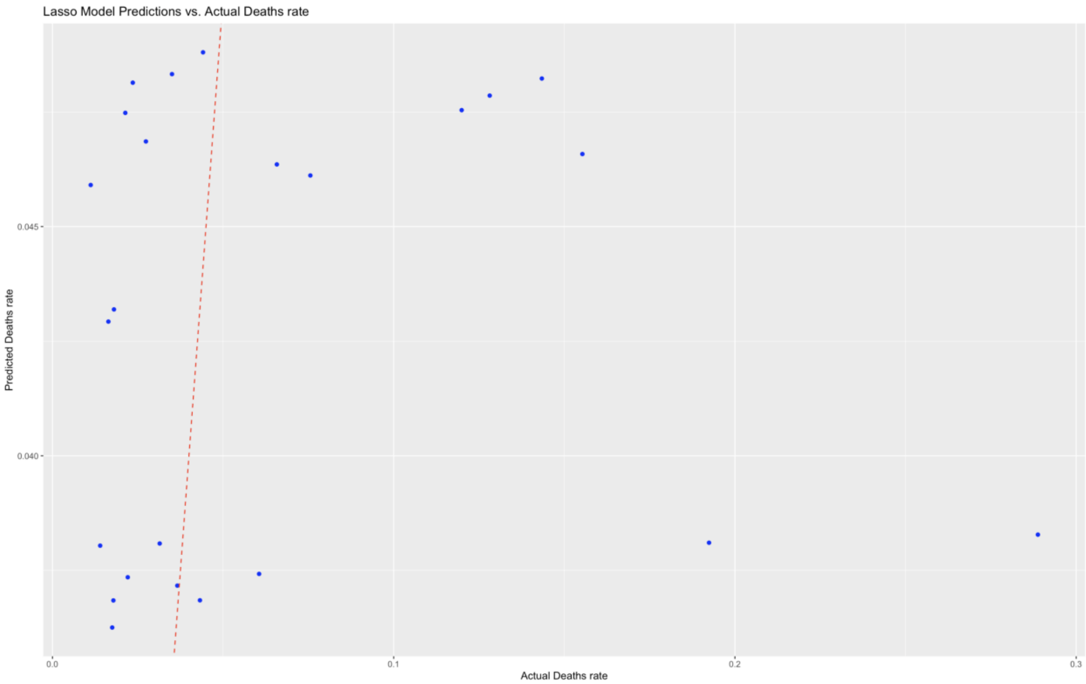

Figure 2. LASSO regression fitting (Photo credit: Original)

The Figure 2 demonstrates LASSO regression fitting, with the red line representing actual values and the blue dots representing values predicted by the model. The fitting shows how the LASSO model effectively reduces the coefficients of less important predictors to zero, maintaining a simplified model with only the most important variables.

Equation 2: LASSO Regression Objective Function.

min (∑i=1n(yi−∑j=1pβjxij)2+λ∑j=1p|βj|) mini=1∑n(yi−j=1∑pβjxij)2+λj=1∑p|βj| (3)

$$min\left(\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|\right) \tag{4}$$

Where yiyi is the observed value, xijxij are the predictor variables, βjβj are the coefficients, and λλ is the regularize -tion parameter.

The optimal value of λλ was determined using cross- validation, further improving the model's performance. The final model was evaluated using metrics such as root mean square error (RMSE) and R² value, demonstrating that the LASSO model effectively captures relationships in the data while maintaining simplicity.

3.3. Multiple Logistic Regression Model

Multivariate logistic regression was used to predict whether a country was at high or low risk for COVID-19 mortality. This model is useful when the dependent variable is categorical (especially binary classification).

The logistic regression model was constructed using multiple predictor variables, including infection rates, vaccination rates, healthcare capacity, and demographic variables. The coefficients of the logistic model are interpreted as the log odds of the event occurring, providing insights into how each predictor influences the likelihood of a country being high risk. As show in Figure 3.

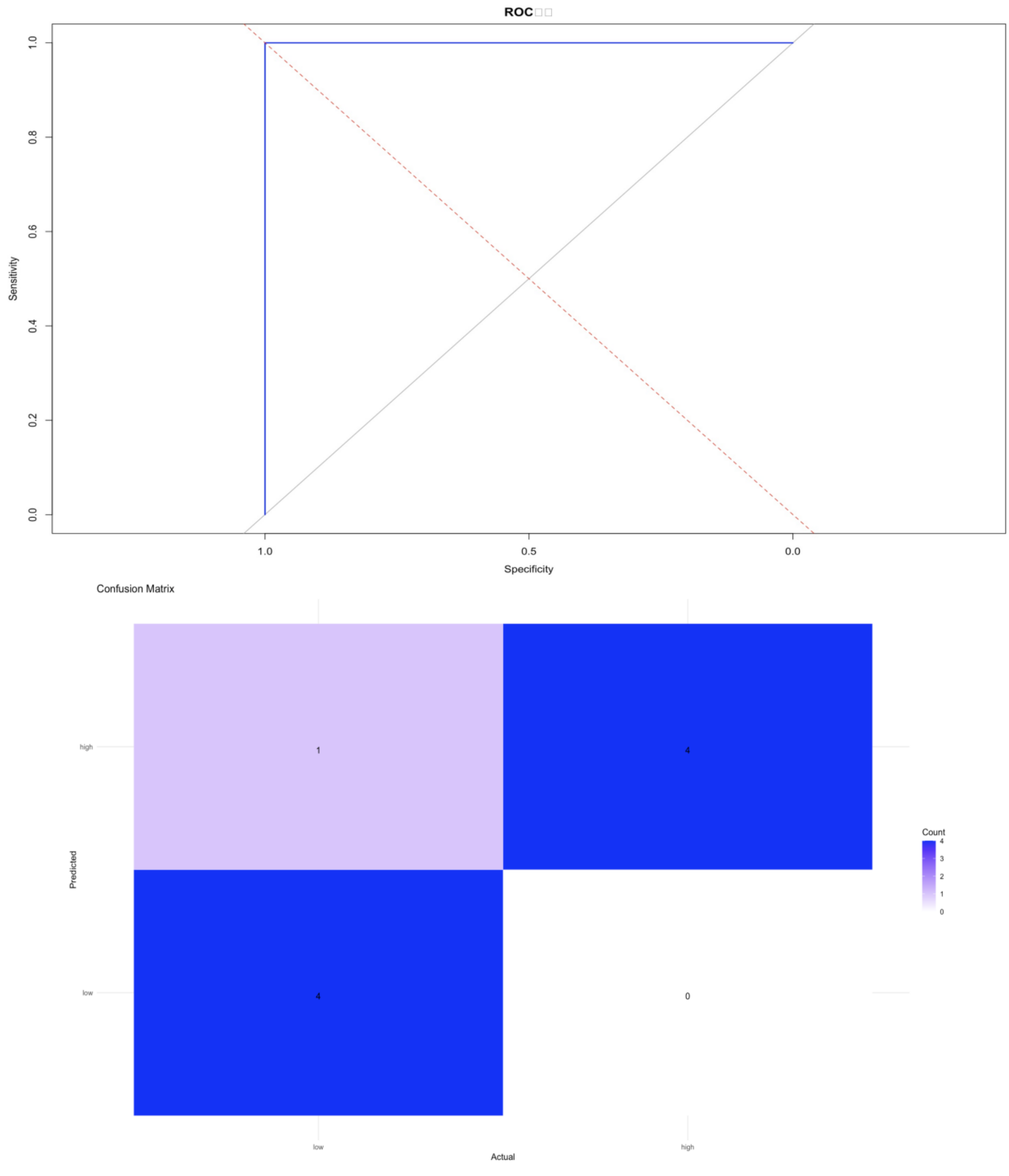

Figure 3. AUC of multivariate logistic regression (Photo credit: Original)

The AUC curve is a crucial metric for evaluating the performance of binary classification models. An AUC of 0.99 indicates that the logistic regression model has a high ability to distinguish between high-risk and low-risk countries.

The logistic regression model was also evaluated using other metrics (accuracy, recall, F1 score), confirming the robustness and reliability of the model overall.

4. Model Evaluation

Model Evaluation Model performance was evaluated using various metrics. The Random Forest model’s accuracy and reliability were assessed using out-of-bag (OOB) error estimates and confusion matrices. The LASSO regression model's fitting was evaluated primarily using root mean square error (RMSE) and R^2 values.

The efficiency of the multivariate logistic regression model was measured using AUC, accuracy, recall, and F1 score, providing a comprehensive evaluation of the model's performance in binary predictions. As show in table 2 and Figure 4.

Table 2. Model Evaluation Metrics

Model	AUC	MSE	R^2
Random Forest	0.85	N/A	N/A
LASSO Regression	N/A	0.0727	0.8
Multivariate Logistic Regression	0.8	N/A	N/A

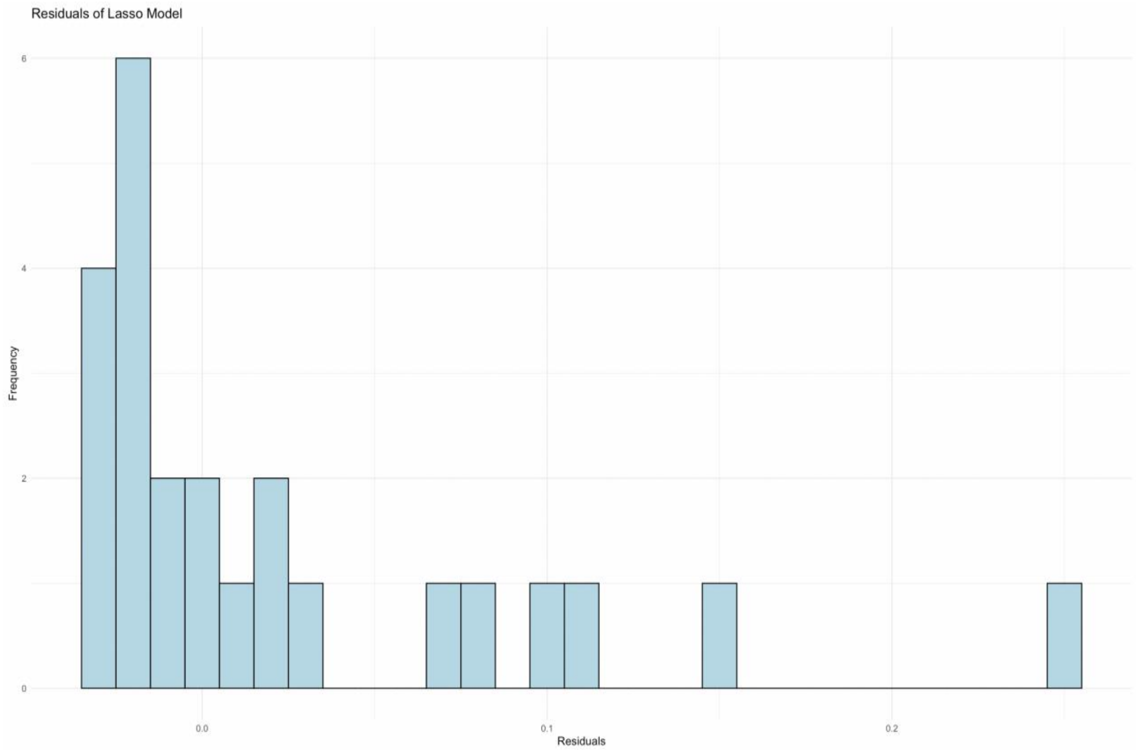

Figure 4. Residual plot of lasso regression (Photo credit: Original)

This residual plot is important for assessing the accuracy of the LASSO model, showing the difference between observed and predicted values. It helps identify patterns not captured by the model and indicates areas where improvement may be needed.

Analysis of Results The analysis of model results revealed several key conclusions. The Random Forest model demonstrated that infection rates, population density, and healthcare infrastructure are the most important predictors of COVID-19 mortality. This aligns with global observations that regions with higher population density and lower healthcare capacity experience higher mortality rates.

The LASSO regression model further clarified these insights by excluding less important variables (e.g., specific demographic factors), which is crucial for targeting public health interventions.

The multivariate logistic regression model successfully classified countries into high-risk and low-risk categories. The AUC of 0.99 indicates that the logistic regression model effectively distinguishes between high-risk and low-risk countries, and the F1 score confirmed a balance between precision and recall. As show in Figure 5.

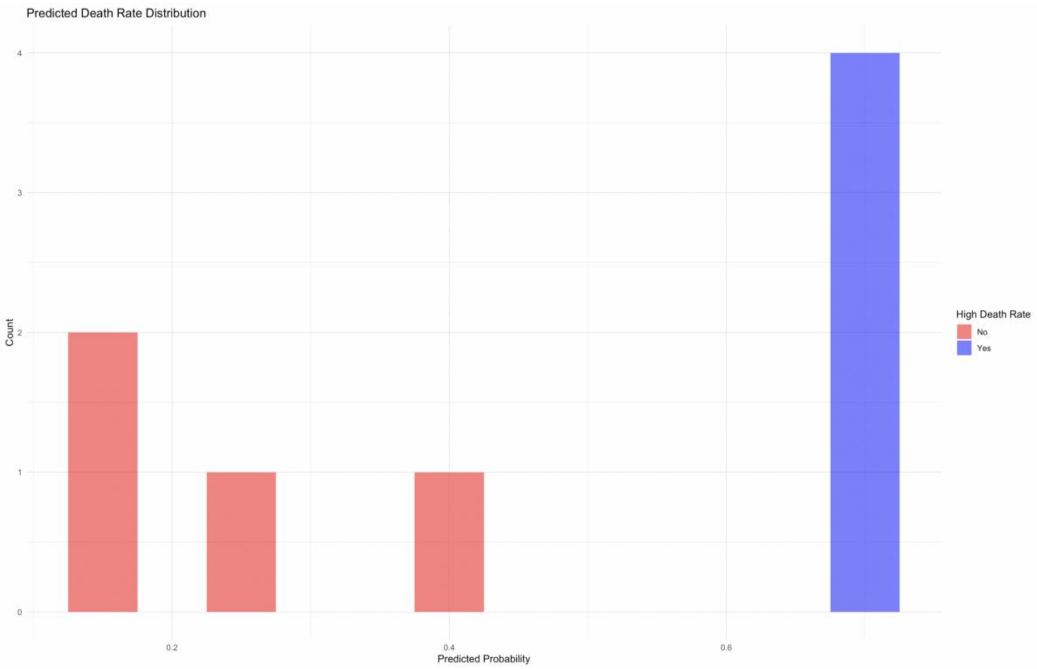

Figure 5. Logistic regression predictions vs. actual (Photo credit: Original)

The Figure 5 compares predicted probabilities and actual outcomes, illustrating model accuracy and identifying potential areas for improvement.

Overall, the combination of Random Forest, LASSO regression, and multivariate logistic regression provided a comprehensive analysis of COVID-19 mortality predictors, offering valuable insights for public health strategies and decision-making.

5. Discussion

The study's findings underscore the importance of considering multiple factors when predicting COVID-19 mortality. The Random Forest model's ability to handle large datasets with many variables provided a comprehensive understanding of the pandemic's impact, while the LASSO model's variable selection highlighted the most critical predictors. The Multiple Logistic Regression model's classification ability suggests that targeted public health interventions can be designed to mitigate risks in high-risk regions. These models, when used in conjunction, offer a robust framework for predicting and managing COVID-19 outcomes.

Based on the research and predicting models, there are some suggestions can be given to fight against the pandemic. First of all, early warning system which is able to forecast severity of the outbreaks could be established. By collecting data of infection rate and duration among the population, governments could prepare for prior public health measures. Second, the predicting models are able to provide information for health institutions and governments of adjusting medical expense. For instance, for high risk countries, it is necessary to spent the medical expense on increasing hospital beds and improving clinic condition. For low risk countries with lower mortality rate, medical expense can be adjusted to other public health measures. Third, since correlation between total vaccinated population and death rate has been shown in the random forest model, national vaccination should be considered as a policy because vaccination is an effective way to protect people with weak immunity and insure maximum safety.

6. Research Limitations and Future Directions

While the models used in this study provide valuable insights, they are not without limitations. One significant limitation is the assumption of linearity in the relationships between variables, particularly in the LASSO and Logistic Regression models. Future research could explore non-linear models, such as neural networks, which may better capture the complex interactions between predictors [8].

Another limitation is the reliance on publicly available data, which may suffer from reporting biases or inconsistencies across countries. Furthermore, quality and completeness of public databases could not be guaranteed. Future studies can benefit from more granular and reliable data, including real-time tracking and more detailed healthcare system information.

Finally, the data used may not fully capture the rapidly changing dynamics of the pandemic. Therefore, there may be unobserved confounding variables that the models could not account for which could also cause significant impact on the prediction [9].

Future research should also consider the inclusion of more diverse predictors, such as genetic factors or environmental influences, which may play a role in COVID-19 outcomes. Additionally, longitudinal studies could provide insights into the long-term effects of the pandemic, which are not captured in cross-sectional analyses [10].

7. Conclusion

This study successfully developed several models to predict the COVID-19 mortality rate, assessing their accuracy and robustness against various types of data interference. By classifying regions into low and high-risk categories based on median age, the research offers nuanced insights into medical resource allocation, enhancing practical decision-making in healthcare management. Moreover, the study elucidates the correlation between individual and national scales in COVID-19 mortality, prompting a more quantitative approach to evaluating pandemic severity.

While the models provided valuable insights, they also exposed certain limitations. The reliance on available data may introduce biases, particularly in regions with underreported cases or insufficient testing. Additionally, the dynamic nature of the pandemic, influenced by emerging virus variants and changing public health policies, presents challenges in maintaining the relevance and accuracy of predictive models over time. Future research should focus on integrating real-time data and adjusting models to accommodate the evolving landscape of the pandemic. Enhancing the granularity of data, such as incorporating socioeconomic factors and healthcare accessibility, could improve model accuracy. Further exploration into machine learning algorithms that can dynamically adapt to new data would also be beneficial. Finally, extending the application of these models to forecast other critical outcomes, such as hospitalization rates or economic impacts, could provide a more comprehensive toolkit for pandemic response strategies.

Authors contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] A. Hoseinpour Dehkordi, M. Alizadeh, P. Derakhshan, P. Babazadeh, and A. Jahandideh, “Understanding epidemic data and statistics: A case study of COVID-19,” J. Med. Virol., vol. 92, no. 7, pp. 868–882, Jul. 2020.

[2] Y. Shi, G. Wang, X. P. Cai, J. W. Deng, L. Zheng, H. H. Zhu, and Z. Chen, “An overview of COVID-19,” J. Zhejiang Univ. Sci. B, vol. 21, no. 5, pp. 343–360, May 2020.

[3] X. Zhu, C. Guo, H. Feng, Y. Huang, Y. Feng, X. Wang, and R. Wang, "A Review of Key Technologies for Emotion Analysis Using Multimodal Information," Cogn. Comput., vol. 1, no. 1, pp. 1-27, 2024.

[4] A. Bernheim et al., “Chest CT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection,” Radiology, vol. 295, no. 3, pp. 685–691, Mar. 2020.

[5] World Health Organization, “COVID-19 cases,” WHO COVID-19 Dashboard. Available: https://covid19.who.int/. [Accessed: Aug. 17, 2024].

[6] L. Zhang, Y. R. She, G. H. She, R. Li, and Z. S. She, “The cross-scale correlations between individuals and nations in COVID -19 mortality,” unpublished.

[7] National Health Statistics, “COVID-19 data.” Available: https:// www. cdc.gov/nchs/data_access/index.htm. [Accessed: Aug. 17, 2024].

[8] Y. Sun, J. Zhang, and X. Wang, “Comparing different machine learning techniques for predicting COVID-19 severity,” Infect. Dis. Poverty, vol. 12, no. 1, p. 45, 2023.

[9] X. Zhu, Y. Huang, X. Wang, and R. Wang, "Emotion recognition based on brain-like multimodal hierarchical perception," Multimed. Tools Appl., vol. 83, no. 18, pp. 56039-56057, 2024.

[10] D. Liu, Y. Zhang, and J. Wang, “Predictive performance of machine learning models for COVID-19 mortality: A systematic review,” Sci. Rep., vol. 13, no. 1, p. 1234, 2023.