# An Analysis of COVID-19 Dataset Using Machine Learning for Predicting High-Risk Patients

ZIWEI GUO
*Department of Data Analytics*
*Dickinson College*
Carlisle, US
gzw19914760905@outlook.com

***Abstract*—The COVID-19 pandemic has created a profound healthcare crisis, resulting in a critical shortage of medical resources worldwide. Analyzing large-scale COVID-19 datasets can facilitate the identification of high-risk patients and the anticipation of their medical resource requirements. This paper presents a comprehensive analysis of a COVID-19 dataset from Mexico, aiming to develop robust machine learning models that can accurately predict high-risk patients and their healthcare needs. We employ multiple machine learning models, including logistic regression, decision trees, bagging, random forests, gradient boosting, and support vector machines, to assess their effectiveness in risk prediction. Unlike previous studies, which often focus on a single model, this study provides a comparative analysis of multiple models to determine the most effective approach for predicting high-risk patients. The results of this study can support healthcare providers in prioritizing critical care patients, optimizing resource allocation, and ultimately enhancing patient outcomes during a public health emergency.**

***Index Terms*—COVID-19, machine learning, logistic regression, decision trees, bagging, random forests, gradient boosting, support vector machines, healthcare analytics**

## I. Introduction

The COVID-19 pandemic has placed an immense strain on healthcare systems globally, necessitating efficient strategies for resource allocation. Accurately predicting high-risk patients is essential for healthcare providers to allocate limited medical resources effectively and save lives. This study aims to leverage a COVID-19 dataset from Mexico to develop machine learning models capable of predicting high-risk patients based on their medical history and symptoms. We have employed logistic regression alongside other advanced machine learning models to explore the data and identify key predictors of mortality. By understanding these predictors, healthcare providers can intervene early, deliver targeted care to those most vulnerable, and thereby potentially reduce mortality rates while alleviating the burden on healthcare facilities.

## II. Dataset Description

The COVID-19 dataset includes 1,048,576 unique patient records from Mexico, encompassing 21 mostly binary independent variables. It captures patient demographic information, such as sex and age, along with comorbidities, including pneumonia, pregnancy, diabetes, chronic obstructive pulmonary disease (COPD), asthma, hypertension, cardiovascular disease, chronic renal disease, obesity, and tobacco use. Additionally, the dataset includes details of medical services provided to patients, such as patient type (hospitalized or non-hospitalized), intensive care unit (ICU) admission, intubation, and mechanical ventilation. The response variable, DATE_DIED, indicates whether a patient died or survived, making the dataset well-suited for developing predictive models of patient outcomes.

## III. Methodology

### A. Preprocessing

To preprocess the dataset, we eliminated irrelevant columns, such as the classification column that indicated whether the patient was a COVID-19 carrier, since all individuals in the dataset had tested positive. Observations with missing or unknown data were categorized under an ”unknown” class, which constituted approximately 3% of the dataset. Values other than ”1” and ”2” were treated as missing data, including outliers like 98 and 99. All character-type values were converted to numeric representations to facilitate model development. Subsequently, the dataset was randomly divided into a training set (80%) and a testing set (20%) to evaluate model performance. This preprocessing step was crucial for ensuring the dataset was standardized, reducing biases, and enhancing the reliability of machine learning algorithms used in the analysis.

### B. Exploratory Analysis

Exploratory analysis was performed to understand the mortality patterns without building a predictive model. Key findings include differences in mortality rates based on demographic characteristics, comorbidities, and treatment types, highlighting the importance of these factors in risk assessment.

The total death rate is 7.22 The death rate for men is 5.15 The death rate for women is 9.3 The death rate for patients who returned home is 0.76 The death rate for hospitalized patients is 34.92

The death rate for patients with pneumonia is 38.1 The death rate for patients without pneumonia is 2.4

The death rate for patients with diabetes is 21.58 The death rate for patients without diabetes is 5.25

The death rate for patients with COPD is 27.43 The death rate for patients without COPD is 6.86

The death rate for patients with asthma is 4.82 The death rate for patients without asthma is 7.25

The death rate for patients with immunosuppression is 18.48 The death rate for patients without immunosuppression is 7.01

The death rate for patients with hypertension is 19.3 The death rate for patients without hypertension is 4.95

The death rate for patients with cardiovascular disease is 21.07 The death rate for patients without cardiovascular disease is 6.89

The death rate for patients with chronic kidney disease is 28.95 The death rate for patients without chronic kidney disease is 6.77

The death rate for patients with other diseases is 15.35 The death rate for patients without other diseases is 6.93

The death rate for patients with obesity is 10.5 The death rate for patients without obesity is 6.56

The death rate for patients who use tobacco is 7.99 The death rate for patients who do not use tobacco is 7.09

The death rate for patients who received mechanical ventilation is 11.03 The death rate for patients who did not receive mechanical ventilation is 5

The death rate for patients who received intubation is 78.23 The death rate for patients who did not receive intubation is 26.17

The death rate for patients admitted to the ICU is 48.58 The death rate for patients not admitted to the ICU is 33.96

The death rate for pregnant female patients is 1.43 The death rate for non-pregnant female patients is 5.23

## IV. Logistic Regression Model

The first predictive model developed was a logistic regression model, which was fitted using the binomial family, with the dependent variable (DATE_DIED) regressed on all other independent variables. The logistic regression model assumes a linear relationship between the logit of the dependent variable and the predictor variables.

The logistic model exhibited the following performance metrics:

- Residual deviance: 19705 on 79965 degrees of freedom, significantly lower than the null deviance of 41333 on 79999 degrees of freedom, suggesting that the model captures significant variability in the data.
- Akaike Information Criterion (AIC): 19775, indicating that the model is well-suited for the data.

Key predictors of mortality included variables such as USMER, SEX, PATIENT_TYPE, INTUBATION, PNEUMONIA, AGE, DIABETES, HYPERTENSION, CARDIOVASCULAR, OBESITY, RENAL_CHRONIC, and TOBACCO. For instance, INTUBATION was strongly associated with an increased risk of mortality, while pregnancy was associated with a decreased likelihood of death. The Receiver Operating Characteristic (ROC) curve for the model had an Area Under the Curve (AUC) of 0.959, indicating excellent discrimination between positive and negative classes.

### A. Receiver Operating Characteristic (ROC) Curve

The Receiver Operating Characteristic (ROC) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings.

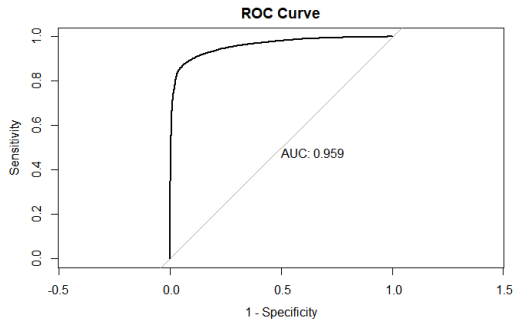


Fig. 1. ROC Curve demonstrating the performance of the logistic regression model. The area under the curve (AUC) is 0.959, indicating excellent discrimination between positive and negative classes.

## V. Imbalanced Data Handling

The dataset was highly imbalanced, with a small proportion of patients experiencing mortality. To address this imbalance, we applied an oversampling technique to balance the dataset. However, the accuracy of the logistic regression model on the oversampled data was only 7.35%, indicating that the model's classification threshold (set at 0.6) may not have been optimal. Moreover, the oversampling technique may have introduced bias, resulting in overfitting and poor generalizability. To improve model performance, alternative resampling methods, such as SMOTE (Synthetic Minority Over-sampling Technique) or cost-sensitive learning approaches, could be explored. SMOTE creates synthetic samples of the minority class, which helps to prevent overfitting that often results from simple duplication of minority instances. Cost-sensitive learning adjusts the cost function to penalize misclassification of the minority class more heavily, making it a promising approach for handling imbalanced data effectively.

## VI. Variable Selection Using Resubset Method

To optimize model performance and interpretability, we used the resubset method for variable selection. The model with 11 predictor variables (USMER, SEX, INTUBED, PNEUMONIA, AGE, PREGNANT, HYPERTENSION, IMMUNOSUPPRESSION, RENAL_CHRONIC, TOBACCO, ICU) exhibited the best balance of performance and complexity. By selecting the most relevant variables, we reduced model complexity, enhanced interpretability, and minimized the risk of overfitting. This approach also allowed us to focus on the key factors significantly impacting patient outcomes, providing clearer insights for healthcare professionals.

## VII. Machine Learning Models

### A. Decision Tree Model

The confusion matrix for the decision tree model is as follows:

| | | Predicted | |
|---|---|---|---|
| | | 1 | 2 |
| Actual | 1 | 504 | 146 |
| | 2 | 966 | 18384 |

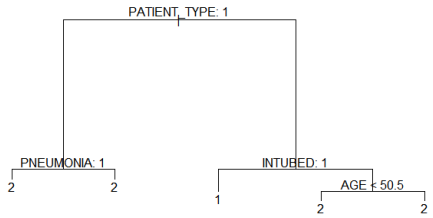


Fig. 2. Decision tree model visual representation.

A decision tree model was developed using four key variables (PATIENT_TYPE, PNEUMONIA, INTUBED, AGE), resulting in five terminal nodes and a residual mean deviance of 0.258. Although the decision tree model achieved a misclassification error rate of 5.4%, it lacked predictive accuracy compared to logistic regression. Nevertheless, decision trees offer a highly interpretable structure, which is particularly beneficial for clinical decision-making. Their simplicity allows healthcare professionals to easily understand the decision-making criteria, making the model useful for assessing patient risk and determining treatment priorities.

### B. Bagging and Random Forest Models

To enhance predictive performance, bagging was employed, resulting in an accuracy of 93.58% and a Kappa statistic of 0.5009. However, sensitivity remained relatively low (50.34%), whereas specificity was high (97.01%). The random forest model, which leverages multiple decision trees, achieved an accuracy of 94.73% and a balanced accuracy of 71.32%, with a sensitivity of 43.88% and specificity of 98.76%. Random forests generally improved accuracy but struggled with correctly identifying positive cases. Ensemble methods like bagging and random forests are effective at reducing variance and increasing model robustness, but challenges with sensitivity underscore the difficulties in classifying high-risk patients.

### C. Gradient Boosting Machine Model

Gradient boosting was employed with 500 iterations, resulting in an accuracy of 92.65% on the test set. The optimal threshold for binary classification was determined to be 1, based on model performance metrics. Gradient boosting combines multiple weak learners to form a strong predictive model, balancing accuracy and complexity. The effectiveness of the gradient boosting model is highly dependent on hyperparameter tuning, such as the number of iterations and learning rate, which significantly impact model performance.

### D. Support Vector Machine Model

The support vector machine (SVM) model demonstrated the highest overall accuracy of 94.44%, with a sensitivity of 99.21% but lower specificity (34.29%). Given the importance of accurately identifying positive cases (i.e., patients who died), the SVM model may be the optimal choice for predicting high-risk patients. SVMs are well-suited for high-dimensional datasets and can model non-linear relationships effectively. However, the low specificity suggests that the model overestimated high-risk patients, highlighting the need for further optimization or the use of a weighted SVM to improve the classification of negative cases.

### E. Lasso and Ridge Regression Models

To further refine model performance, we applied Lasso and Ridge regression techniques. The Lasso model failed to predict positive cases effectively, resulting in low sensitivity. In contrast, the Ridge model achieved a balanced accuracy of 78.83% and a sensitivity of 96.19%, outperforming Lasso in terms of prediction capability. Regularization techniques such as Lasso and Ridge are instrumental in preventing overfitting by penalizing large coefficients, particularly useful in datasets with multicollinearity. The Ridge model's improved performance suggests that regularization effectively reduced variance without sacrificing predictive power.

### F. Bagging, Random Forest, and Boosting Models

After developing a decision tree model, bagging was applied to enhance its performance. The bagging model achieved an accuracy of 93.58%, with high specificity but moderate sensitivity. The random forest model further improved accuracy to 94.73%. Boosting, using the Gradient Boosting Machine (GBM), achieved an accuracy of 92.65%, demonstrating robust performance but limited sensitivity. These ensemble methods, particularly random forests and boosting, enhanced predictive accuracy by combining multiple base models, reducing overfitting, and capturing complex patterns in the data.

### G. Support Vector Machine Analysis

The support vector machine (SVM) model was ultimately employed, achieving an overall accuracy of 94.44%. The SVM model's high sensitivity (99.21%) made it particularly effective in identifying true positive cases, although it struggled with specificity (34.29%). The SVM's ability to handle non-linear and high-dimensional data made it advantageous for this application. However, the trade-off between sensitivity and specificity suggests the need for further hyperparameter tuning or alternative approaches, such as using a weighted SVM to improve classification of negative cases.

## VIII. Conclusion

This study analyzed a COVID-19 dataset from Mexico using multiple machine learning models to predict high-risk patients. The models included logistic regression, decision trees, bagging, random forests, gradient boosting, and support vector machines (SVM). Logistic regression provided a reasonable baseline, while decision trees offered interpretability. Bagging and random forests improved accuracy but struggled with sensitivity. Gradient boosting and SVM achieved high accuracy, with SVM demonstrating the highest sensitivity in identifying high-risk patients. Overall, the SVM model performed best for scenarios requiring precise identification of high-risk patients, while random forests provided balanced accuracy for general use.

These models can assist healthcare professionals in prioritizing high-risk patients and optimizing resource allocation. While SVM showed the best classification accuracy, further model tuning and validation are needed to improve performance. This study provides a foundation for using machine learning in pandemic response efforts, and future research could explore deep learning models to capture more complex relationships, as well as hybrid models for enhanced performance. Cross-validation and external validation with datasets from other regions are recommended to improve generalizability and robustness.

## Acknowledgment

The authors would like to thank the professor Xiexin Liu and Mexican Government Data Portal for providing the dataset used in this study.

## References

[1] Mexican Government Data Portal, "Información referente a casos COVID-19 en México," Accessed: [Online]. Available: https://datos.gob.mx/busca/dataset/informacion-referente-a-casos-covid-19-en-mexico.