Unlocking Insights from the Shadows: A Machine Learning Approach to Analyze COVID-19 Patient Data in Mexico

Fangxu Gu*, Chi-Feng Ho*, Chi-Hong Ho*, Lavender Yu*, Adrian Vasquez*
*INDENG 142, University of California, Berkeley
Dec. 15th, 2023

***Abstract*—In the past three years, everyone suffered from the coronavirus disease, COVID-19. Most people were infected with the disease several times, and a partial amount of people passed away. Healthcare workers already tried their best to support the patients during this period, but if they can understand the insights of the Covid-19 dataset, more patients can be identified in the early stages and provide better treatments. This project accessed the Mexican Government's Centers for Disease Control and Prevention (CDC)'s COVID-19 dataset in 2020. Authors utilize machine learning techniques to discover the insight of the dataset and analyze the patient cases in Mexico. The authors will try different classification models to increase the predictability accuracy of the patient's status(alive/dead). Through the result comparison, the CART model with cross-validation has the best performance with an accuracy of 85.3% and a False Negative Rate of 10.3%, compared to other classifiers.**

***Index Terms*—COVID-19, machine learning, classification, CART model, accuracy.**

I. Introduction

A. Background

The COVID-19 pandemic had a significant global impact over the past three years. Many people unfortunately died due to the pandemic's spread. Therefore, identifying the elements that lead to COVID-19 patient deaths is critical. Our objective in this study is to identify the most important COVID-19 mortality indicators and examine the influence of these factors on death rates. When it comes to tracking COVID-19 mortality data, it is assumed that all patients' deaths were directly attributed to COVID-19 complications.". However, the reality is that these patients may have underlying health issues. Given the interconnection of the immune system, potential health issues, and people's ability to resist the virus, we are curious about whether underlying health conditions could be a significant factor affecting the mortality rate. We intend to give a complete analysis of COVID-19 mortality predictions by employing various statistical and machine-learning methodologies. We may learn more about the risk factors linked to COVID-19 mortality by identifying these predictors and further developing more effective patient care and public health strategies.

B. Objective

The research objective is to identify the main influential factor responsible for COVID-19 mortality. Many factors cause patients with COVID-19 to die; regardless the Mexican Government CDC labels those patients' deaths as due to COVID-19. By examining the impact of the other factors (patients' medical history, external factors) on death rates, we aim to provide a comprehensive analysis that transcends traditional mortality data tracking.

C. Significance of the Project

Pandemics are a tragic experience for any society to undergo, and any improvement to our understanding can be incredibly beneficial. Through our research, we may learn more about the risk factors linked with COVID-19 mortality by diving into the dataset and further developing more effective patient care and public health strategies. While healthcare workers already strike their best to keep us safe from COVID-19, a better prediction model can empower doctors/physicians/public health officers to develop even more effective patient care and public health strategies. This research extends beyond the immediate pandemic challenges, informing proactive measures for future health crises.

II. Data Collection and Data Cleaning

A. Sources

The dataset utilized in this study was sourced from the official website of the Mexican Government's CDC, available at [1].

B. Raw Data & Information

The dataset comprises approximately four million anonymized entries encompassing patient-related information, including pre-existing conditions. The raw dataset contains 40 distinct features, ranging from patients' birthplaces to specific diseases prevalent in the Latino population. All information is presented in the Spanish language.

C. Data Cleaning

The group lacked proficiency in the Spanish language; therefore, the authors translated the Spanish text into English for research purposes. This study is specifically centered around COVID-19 patients and pertinent information. Initially, the authors filtered the data to extract valuable information and removed irrelevant variables, such as Country of nationality, Record ID, Language, and record day, from the dataset. To exclusively focus on COVID-19 patients, we excluded individuals who did not test positive for the COVID-19 infection.

The research involved several steps to prepare the data for analysis. Initially, any missing data needed to be removed. Several data cells in the original dataset are blank, thus we substituted the word "nan" in those cells for research purposes.

Binary data, denoted by Yes/No signs, was converted into categorical variables for improved visualization. The first death data field simply included two distinct values: 9999-999, or their date of death. To facilitate comprehension, the date of death was transformed into a binary classification: "alive" or "deceased." Age bins were created using the frequency distribution of patient ages to provide an equitable distribution of patients within each age group. A histogram was made to display the distribution of the data across the adjusted variables to give a visual representation of the data characteristics. These methods for data preparation ensured that the information was well-organized, clean, and suitable for further investigation.

Ultimately, the dataset was refined to include 21 unique features, encompassing 10,000 distinct patient records.

1) Categorical Data:

- Sex: Female/Male.
- Age: Age of the patient.
- Classification: COVID test findings. Values 1-3 indicate varying degrees of COVID diagnosis, while 4 or higher signifies the absence of COVID or inconclusive test results.
- Patient Type: Type of care received by the patient ('home' or 'hospital').

2) Medical Conditions:

- Pneumonia: Presence or absence of air sac inflammation.
- Pregnancy: Pregnancy status.
- Diabetes: Presence or absence of diabetes.
- copd: Presence or absence of Chronic Obstructive Pulmonary Disease (COPD).
- Asthma: Presence or absence of asthma.
- Inmsupr: Immunocompromised status.
- Hypertension: Presence or absence of hypertension.
- Cardiovascular: Presence or absence of heart or blood vessel-related disease.
- Renal Chronic: Presence or absence of chronic renal disease.
- Other disease (OTHER_COM): Presence or absence of other diseases.
- Obesity: Presence or absence of obesity.
- Smoking: Tobacco use status.

3) Logistical Information:

- Origin: Indicates whether the patient received treatment in medical units of the first, second, or third level.
- Medical Unit: Type of institution within the National Health System providing care.
- Intubed: Indicates whether the patient was connected to a ventilator.
- ICU: Indicates whether the patient was admitted to an Intensive Care Unit.
- Status: Indicates whether the patient died in 2020.

III. Multiple Correspondence Analysis

We utilized Multiple Correspondence Analysis (MCA) to study the relationship between different categorical variables [2]. Since the data sources from the previous section discussed are categorical data, we can only apply MCA to study the relationship between categories of different variables. The idea of MCA in our study is to uncover the patterns and associations to provide insights into the structure of the data.

The main idea behind MCA is to reduce its dimension, just like Principal Component Analysis (PCA) but with discrete categorical variables. We apply the Python package [3] and the SPSS analysis tool to trace different dimensions that are used to explain the observed relationship between different categorical variables. Each dimension represents a pattern or structure, and a corresponding eigenvalue represents the contribution of that dimension to the overall variance. Larger eigenvalues indicate that the dimension contains more information.

The investigation included the inertia distribution as a critical component in Fig. 1. Out of the entire dataset inertia, the first two dimensions explained 20.51% of the variance, well above the standard value of 6.91%. This large explanatory power suggested that a considerable amount of the variability in the dataset was captured by these dimensions. Furthermore, it was suggested that the first ten dimensions be taken into account since taken as a whole, they contained 52.1% of the inertia, which was more than the reference value of 33.4%. This supported the idea that these dimensions held important information about the connections between different factors and people.

Based on positive and negative coordinates, the several planes (1:2, 3:4, 5:6, 7:8, and 9:10) offered insights into the opposition between groups. Different factor frequencies were seen in each group, highlighting the importance of variables such as patient type, ICU status, pneumonia, intubation, age, hypertension, pregnancy, sex, and diabetes in characterizing these groups.

We mostly concentrate on analyzing the graph of the component 1 and 2 since it has an interpretability of around 21% of the total dataset. Based on the unique factor frequencies found in the groups across positive and negative coordinates, the x-axis labeled "Health Severity" and the y-axis labeled "Criticality" are justified shown in Fig. 2. Based on the measured factor frequencies, the x-axis, designated as "Health Severity," depicts the continuum of health problems ranging from less severe (negative coordinates) to more severe (positive coordinates). The y-axis, designated "Criticality," concurrently indicates the degree of crucial health circumstances; lower negative coordinates denote fewer critical conditions, while higher positive coordinates indicate a more critical health state.

Through the study, we learned that high frequency for factors of ICU_N, PNEUMONIA_Y, INTUBATED_N, AGE_61+, HYPERTENSION_Y, and DIABETES_Y, indicated a positive coordinate on the axis. Pneumonia, intubation, and the high incidence of diabetes and hypertension would all point toward the high severity of health issues and a higher chance of COVID-19 death. On the other hand, Group 3 has a negative coordinate on the axis. Those people in the cluster have high frequencies of Alive, AGE_31–40, home PATIENT_TYPE, Female, and PREGNANCY_N. This reveals a less serious state of illness and decreases the risk of

intubation and pneumonia. Groups 2 and 4 exhibit intermediate patterns in factor frequencies, according to differing levels of health severity and criticality. These groups also share positive and negative coordinates.

Moreover, the MCA also provides a good framework to analyze how each variable correlates to each other. In Fig. 3, we labeled different input variables in a map diagram to assess the correspondence level of the variables. Connecting each factor point with the origin of the graph, the angle formed between two separate rays will describe the correlation between two variables. Indeed, if the angle is pretty small (variables A & B), we would conclude that there is a high chance that if a patient is satisfied with A, they are also satisfied with B. As the angle between two variables increases, their relevance to each other decreases.

It is clear that the dimensions of x and y are the critical components in determining the similarity of different variables. The map can help us illustrate the connection between some facts to the potential risks, such as suffering from chronic renal disease and COPD. We would statistically guess we should provide a COPD check for the patient who has chronic renal disease because it could be a factor that affects their likelihood of COVID-19 death.

IV. Methodology & Modeling

We used five different methods to model the COVID-19 dataset and aimed to maximize the model's accuracy. After the data cleaning, we split the dataset (a total of 10,000 data points) for a 70%-30% training and testing set, in which 7000 data points were used for model training and the rest of the points were used to examine the performance of the training model.

A. Logistic Regression

Most of the data we kept for analysis were binary variables, therefore, we ran the datasets through a logistic model [4] to assess the accuracy of predicting COVID deaths with different variables. We mainly focused on the data-cleaned table with 21 categorical variables and studied how patients' body conditions would become an associated problem leading to COVID-19 deaths. Our model exhibits strong performance, boasting a high accuracy of 0.903 and an impressive Area Under the ROC Curve (AUC) of 0.94, as illustrated in Figure 4. This underscores its excellent discriminative ability. The high accuracy may be due to the high correlation with the target variable we selected. Further, because there are no serious outliers in the dataset, it is easier for the model to learn patterns in the data. In the current model, we observe a False Negative Rate (FNR) of 0.372, indicating a potential risk of falsely predicting patients who would not die as experiencing mortality. This issue could lead to significant consequences. To mitigate the FNR, we would apply other more robust ML algorithms (RF, CV) to prevent model overfitting and improve the generalization ability of data.

B. LASSO

Aside from reducing the overfitting error, we tend to reduce some unrelated coefficients to zero for implementing feature selection. Therefore, we apply the Lasso Classification in our study to approach the COVID classification and implement robustness. Because Lasso penalizes coefficients, forcing the model to fit new data better to prevent overfitting, we set the regularization strength to 0.01 to regulate each feature in the model. Indeed, the larger alpha values result in a stronger regularization, and to avoid regularizing the feature, we stopped at 0.01. Compared with traditional regression models, Lasso classification can achieve model sparsity by making some features' weights zero. In addition, in Lasso, we use one-hot encoding to reduce collinearity and improve the generalization performance of the model. Our accuracy of the model is 89.95%. Nevertheless, the LASSO model exhibits a concerning False Negative Rate (FNR) of 0.51, implying that half of the individuals experiencing mortality are erroneously predicted to be alive. This misclassification could potentially result in delays in receiving timely treatment. Consequently, the LASSO model may not be the most optimal choice in this particular scenario.

C. CART - Decision Tree and with CV

We are leveraging the CART Classification Model [5] in this project due to the fact that it is highly interpretable and provides intuitive explanations of the decision-making process. We configured key parameters, setting the minimum leaf node samples to 6, maximum leaf nodes to 6, and assigning class weights alive equal to 1 and death equal to 4. By setting the minimum leaf node to require at least 6 samples, we aim to avoid creating excessively small leaf nodes, contributing to a more generalized and less overfit model. Similarly, capping the maximum leaf nodes at 6 serves as a means to control tree complexity and mitigate overfitting risks. Additionally, the class_weight parameter addresses class imbalance by assigning a higher weight to the 'died' class compared to the 'alive' class, ensuring that the model pays increased attention to the minority class, which, in this dataset, is the 'died' category. After fitting the model to the training set, predictions were made on the test set, and the model's accuracy was calculated. The printed accuracy for the model is 84.55% and FNR is 15.1%.

Besides, the Decision Tree Classifier may lead to the problem that the minority classes, in our case, the 'died' class, perform poorly in the prediction model because the Decision Tree (DT) tends to predict the dominant class. The utilization of different technical parameters can yield distinct results. Therefore, we employed cross-validation [6] in the decision tree model to identify the most optimal parameters. We used Gini impurity and entropy to measure the quality of each parameter. Based on the baseline model, we observed the ratio between alive and dead is 4:1, so we specified the weight for the 'dead' class 4 times for the 'alive' class. The advantage is that it allows for increased focus on the minority class of the dataset, thereby reducing the FNR. It is beneficial to

comprehend the settings of the classification tree, as illustrated in the figure. In this configuration, the accuracy improves to 85.4%, with a corresponding reduction in the FNR to 0.103.

D. Random Forest and with Cross Validation

We opted for the Random Forest classifier (RF) [7] to address potential model overfitting and effectively handle nonlinear relationships within the dataset. The development of the RF classifier in Fig. 5, employed the bagging and sampling method to construct multiple random trees and assess their performance. The final accuracy of the RF model is determined by averaging the classification performance across these multiple random trees. In the initial execution of the RF classifier, cross-validation was not utilized, resulting in a test accuracy of 0.801 and the lowest FNR recorded at 0.0816. Subsequently, implementing cross-validation in the RF classifier significantly improved the accuracy to 0.867, albeit with a slightly higher FNR of 0.115. Additionally, the exploration of larger datasets, multiple iterations of cross-validation, and fine-tuning of hyperparameters have the potential to further enhance both the accuracy and stability of the RF classifier's performance.

E. Support Vector Machine (SVM)

SVM is effective in solving regression and classification issues [8]. For the COVID classification assignment in our project, we utilized 21 criteria. Our goal is to employ SVM to identify the hyperplane that best divides the classes. Probability = true indicates that we ask for every output to be the class probability, and in the end, we obtain 89.95% accuracy and 0.51 FNR in SVM learning. Our accuracy is higher than Baseline and BERT. Apart from its increased accuracy, the model's performance in accurately predicting negative instances is inferior to that of properly recognizing positive instances, as evidenced by the fact that other measures, such as TPR in SVM, are lower than FNR. Stated differently, this model fails to account for a comparatively high proportion of positive cases, which might have consequences based on the details of COVID-19 death prediction.

V. Discussion & Conclusion

Our primary focus on learning the critical factors that influence COVID-19 mortality in our project has indeed led us to some constructive findings. We aimed to unravel the inherent complexity of mortality data tracking and acknowledge the complex factors that contribute to patients' deaths under COVID-19 according to the data from the Mexican Government.

The machine learning models exhibiting noteworthy results across various criteria are presented in Fig. 6. Logistic regression demonstrated its effectiveness in handling binary data, achieving an accuracy of 90.3% with a 37.2% FNR. LASSO showcased its utility in minimizing overfitting and selecting relevant features, achieving an accuracy rate of 89.95% with an FNR of 51.1%. CART decision trees provided insights into nonlinear connections and interpretability, achieving an accuracy of 84.6% with an FNR of 53.3%. The cross-validated CART model improved accuracy to 85.3% while reducing the FNR to 10.3%. Random Forest showcased its robustness in handling complex data structures, with accuracy rates of 80.1% and 86.7% and FNRs of 8.16% and 11.5%, both without and with cross-validation. SVM outperformed CART and baseline models with an accuracy of 89.95% but exhibited a higher FNR of 51.1%.

However, the parameter used to measure the best performance is not always "accuracy". For example, in healthcare studies, we do not want to wrongly classify patients, whether a patient with negative results or a healthy man. Specifically, providing treatment to already healthy people does not cause too much harm, and in contrast, misidentifying a positive patient as negative would fail to provide timely and effective assistance to patients who may indeed die. Thus, we introduced the FNR, Recall Rate, and Precision in Fig. 7 to compare the effectiveness of the ML classification models. The technical indicator for evaluating the model should focus on lowering the False Negatives, which means having a higher True Positive Rate (Recall Rate). A detailed comparison is provided in the appendix section, revealing that the most favorable performance is achieved by CART with CV attaining an accuracy of 85.3% and an approximate 90% recall rate. Although Random Forest without CV achieves a higher recall rate at around 91.2%, the significantly higher False Positive Rate is a concern. Classifying too many patients as likely to die, when they will not, could lead to the unnecessary allocation of medical resources and the denial of treatment to patients who genuinely require it. Consequently, we assert that CART with CV exhibits superior performance in this context.

In our study, it is important to acknowledge certain limitations. We exclusively considered data from the year 2020, neglecting numerous factors influencing medical technology development, such as new treatments and vaccines. Furthermore, our current dataset focuses exclusively on data from the Mexican government, primarily representing the Mexican population. It is essential to recognize that various subpopulations, such as those of Asian or African descent, may respond differently to COVID-19.

To enrich our understanding and enhance the inclusivity of our analysis, we plan to incorporate data from diverse sources in future endeavors. This may involve leveraging information from global health organizations or regional health databases, allowing us to attain a more comprehensive perspective on the nuanced impacts of COVID-19 across different populations. We may keep expanding the dataset size to enhance the generalization capabilities to reduce overfitting. Also, it's better to study more quantitative datasets for further classification than using binary values, because, for instance, the difference between individual disease measurements may affect COVID-19 mortality from different levels. Yet, based on more solid modeling and in-depth studies, we could specify more specific factors that provide more nuanced insights into the mortality of the epidemic disease.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to Professor Grigas for his invaluable guidance and unwavering support throughout the development of this project. We are also deeply appreciative of GSI Qiran for her assistance and valuable insights during the OH. We are truly thankful for the opportunity to benefit from their mentorship.

REFERENCES

1 Secretaría de Salud (2023). Datos abiertos: Bases históricas. Dirección General de Epidemiología. Retrieved from https://www.gob.mx/salud/documentos/datos-abiertos-bases-historicas-direccion-general-de-epidemiologia

2 Le Roux, B., Rouanet, H., & Sage Publications. (2010). *Multiple Correspondence Analysis*. Sage Publications.

3 Halford, M. (2021, March 18). *prince*. GitHub. https://github.com/MaxHalford/prince

4 Grigas, P. (2023). *Lectures 5 and 6 - Logistic Regression - Part 1*. Course Notes. Pages 19-23, Sep. 6th, 2023.

5 Grigas, P. (2023). *Lecture 9 - CART - Part 1*. Course Notes. Pages 34-59, Sep. 20th, 2023.

6 Grigas, P. (2023). *Lectures 10 and 11 - CART and Cross-Validation*. Course Notes. Pages 23-45, Sep. 25th, 2023.

7 Grigas, P. (2023). *Lecture 12 - Random Forests*. Course Notes. Pages 29-34, Oct. 2nd, 2023.

8 James, G., Witten, D., Hastie, T., Tibshirani, R. (2023). *An Introduction to Statistical Learning with Applications in Python*. Publisher.

9 Grigas, Paul. (2023). *Lectures 5 and 6 - Logistic Regression - Part 1*. Course Notes. Pages 99-110, Sep. 6th, 2023.

10 Grigas, Paul. (2023). *Lectures 7 and 8 - Logistic Regression - LDA*. Course Notes. Pages 24-37, Sep. 13th, 2023.

APPENDIX

The following are the figures, diagrams, and charts used in this project:

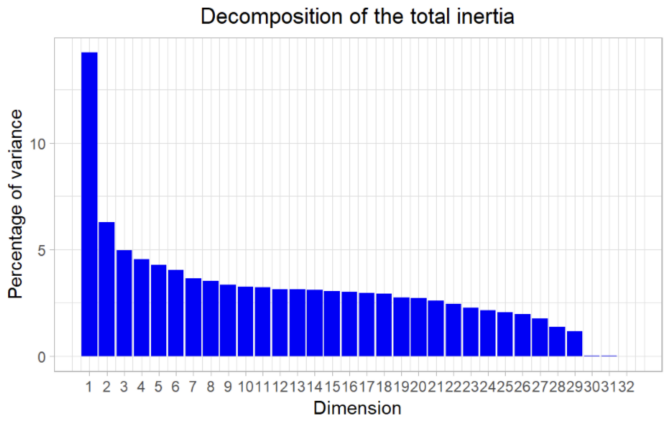

Fig. 1: Total Inertia Decomposition Chart

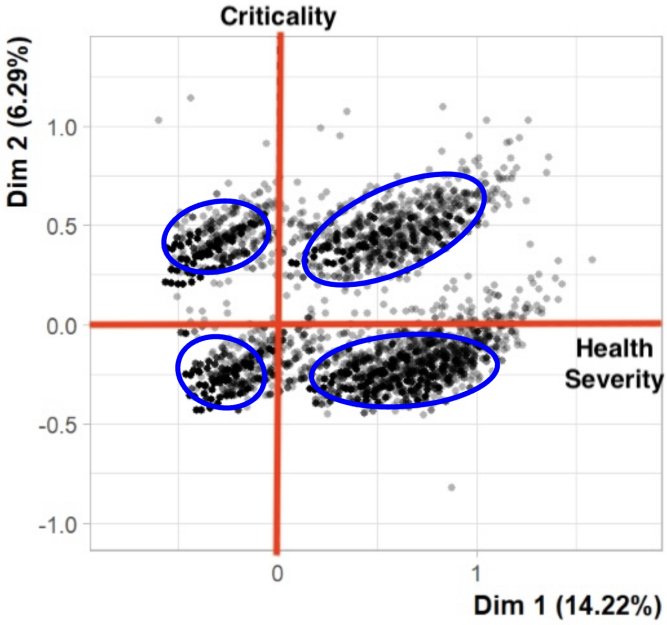

Fig. 2: Multiple Correspondence Analysis Biplot

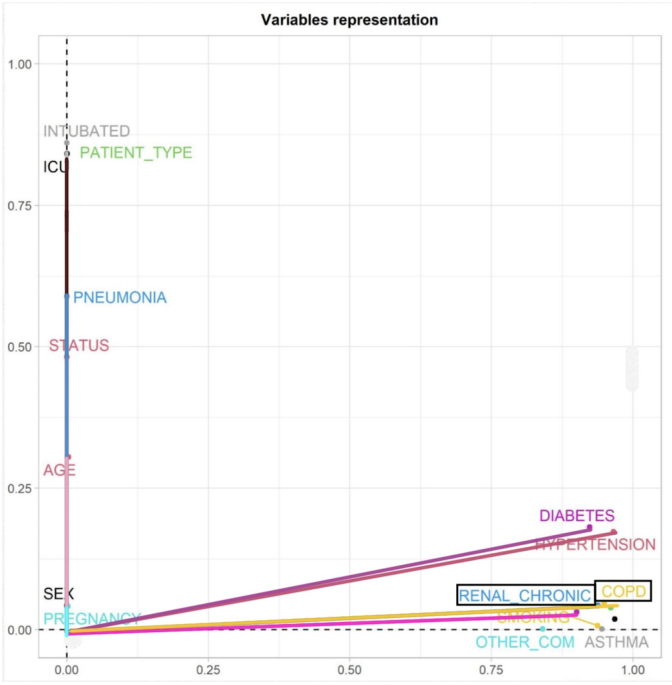

Fig. 3: MCA Factor Map

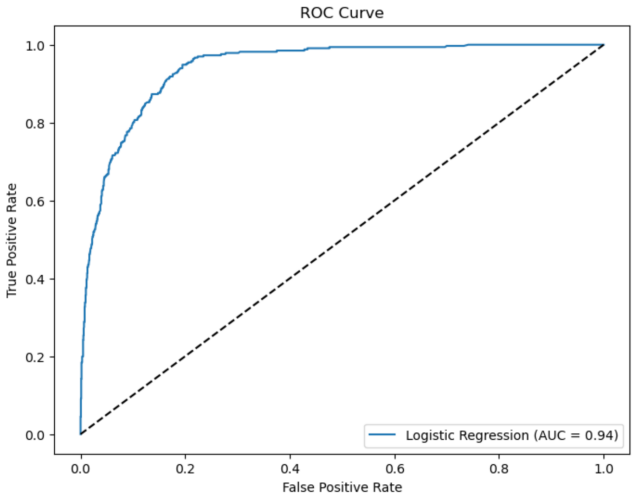

Fig. 4: ROC Curve

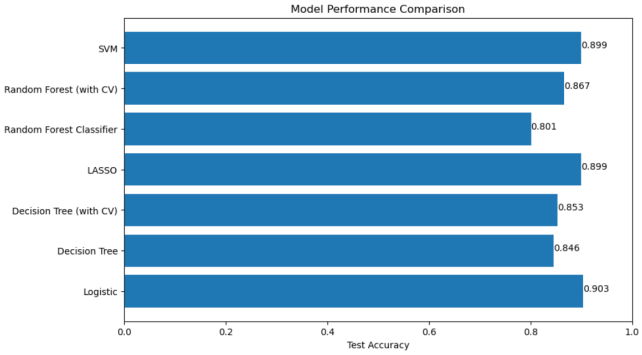

Fig. 7: Accuracy Comparison Bar Chart

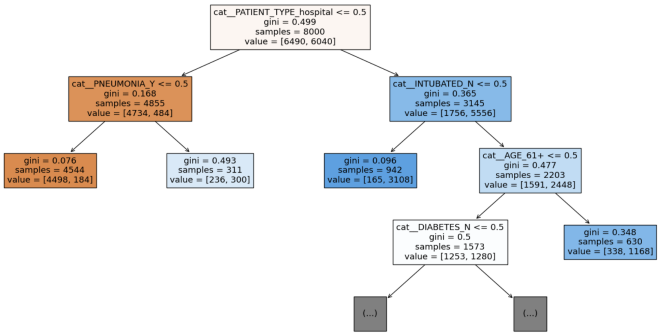

Fig. 5: Decision Tree Diagram (max_depth=3)

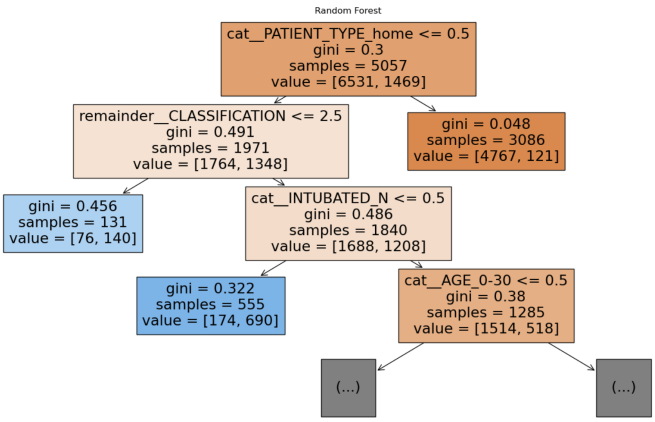

Fig. 6: Random Forest Tree Diagram (max_depth=3)

Model	Accuracy	TPR	FPR	FNR	Precision
Logistic Regression	0.903	0.628399	0.0425404	0.371601	0.74552
Decision Tree	0.8455	0.848943	0.155183	0.151057	0.52037
Cross-validated Decision Tree	0.853	0.897281	0.155782	0.102719	0.533214
Lasso	0.8995	0.489426	0.0191732	0.510574	0.835052
Random Forest	0.801	0.918429	0.222289	0.081571	0.45037
Cross-validated Random Forest	0.8665	0.885196	0.137208	0.114804	0.561303
SVM	0.8995	0.489426	0.0191732	0.510574	0.835052

Fig. 8: Test Result Table