Unraveling Long Covid

*Note: A Data Driven Approach to Understanding Symptom Patterns and Recovery Predictors

Abril Alvarez
Computer Science
University of Illinois at Chicago
Chicago, USA
aalva31@uic.edu

Sammy Dandu
Computer Science
University of Illinois at Chicago
Chicago, USA
sdandu2@uic.edu

Anusha Kavatekar
Computer Science
University of Illinois at Chicago
Chicago, USA
akava2@uic.edu

Manav Kohli
Computer Science
University of Illinois at Chicago
Chicago, USA
mkohli4@uic.edu

***Abstract*— Long COVID presents many significant health issues that require better diagnosis and treatment. This project aims to fulfill that by developing a robust machine learning model that can predict the likelihood and trajectory of Long COVID symptoms based on available patient health data. The dataset that was implemented was from CDC which specifically provided information such as age, gender, pre-existing conditions, etc. Our goal was to uncover the key predictive factors for Long COVID outcomes. The data preparation involved thorough cleaning, which included addressing missing values and removing duplicate data, to ensure that there is a high-quality dataset that can be easily implemented for our intended purposes. Through the data analysis that was done, many important patterns were identified. This highlighted critical insights about how other factors and features has a correlation with patient outcomes. Our thorough approach to tackling this project integrates both machine learning and qualitative methods. We determined which key features need to be selected based on their predictive power based on the correlation analysis and their feature importance rankings. There was a major focus on determining the most efficient models. To do this, many models were rigorously tested and evaluated. Some models that were evaluated included Logistic Regression, Random Forest, XGBoost, etc. This project provides crucial insights in Long COVID and allows for better treatment plans and increases diagnostic accuracy. This product will allow for new healthcare approaches in COVID and help those impacted by it.**

I. Problem Statement

Long COVID poses health challenges that need data-driven insights. Our objective will be to develop a machine learning model to predict the severity, likelihood, or trajectory of Long COVID symptoms based on patient health data, symptoms, and early COVID-19 infection characteristics, as well as potential recovery predictors and possible responses to various treatments. The proposed model will leverage diverse demographic information streams including age, gender, intubation status, medical unit, pre–existing conditions pregnancy, etc. By analyzing and training models based on this data set, we aim to identify predictive factors that influence Long Covid outcomes and treatments. This research is meant to address a gap in the current medical field when diagnosing and treating covid patients along with managing the effected patients in the aftercare process. This work will ultimately contribute to the understanding of Long Covid progression in connection to other illnesses/ diseases. This will enable healthcare providers to create better adn improved treatments.

II. Data and Experiemntal Design

A. Selecting a Dataset

First, find several reliable datasets from renowned organizations, CDC, John Hopkins, Northwestern, US Gov, etc. After a selection of a few data sources, the data must be studied to see if they fit the requirements needed, in this case a large variety of personable information was needed to fit different demographics. After continuing this process of analysis on several datasets, the CDC Long Covid was selected.

B. Data Pre – Processing

Data Pre – processing in this case was simple as the data itself was relatively clean. The CDC Long Covid data set included duplicates and when cleaning the data they were removed. By removing the duplicates, we ensured class balance was the same and that class proportions were even. Initial data quality assessment revealed missing values which were ultimately removed to maintain the reliability and completeness of the data set. Categorical variables were encoded using values less than and greater than four. Those with a value of less than 4 were not experiencing covid and those with values greater than 4 had covid. Other tasks that were done to further clean the data and increase its usability were normalization, and a technique of data augmentation called SMOTE. Normalization was done to make sure the variables we used were standardized to improve the overall model performance. SMOTE, or Synthetic Minority Over-sampling, technique was implemented to balance the distribution with the majority class.

C. Exploratory Data Analysis

Our Exploratory Data Analysis (EDA) revealed several critical patterns in the COVID-19 data. The age distribution showed a clear normal distribution pattern, as visualized in the histogram (bottom right), which aligned with expected population demographics. The mortality rate by age group (top left) demonstrated a striking progression, with rates increasing substantially in older populations - particularly showing marked jumps after age 60-65, reaching peaks of approximately 0.5 (50%) in the 75-90 age groups.

The correlation between ICU and intubation status (top right heat map) showed strong interdependence, with darker blue colors indicating higher mortality rates when both factors were present. This visualization supported our hypothesis

about the strong correlation between these critical care interventions and patient outcomes.

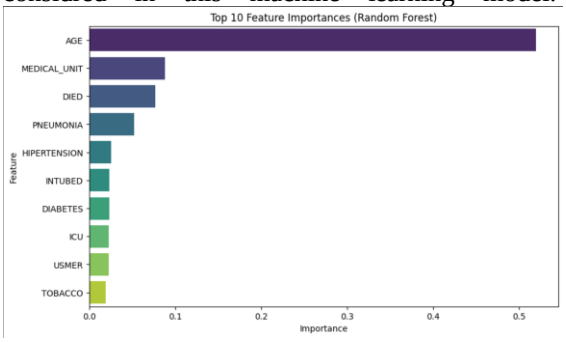
The temporal analysis (bottom left bar chart) displayed mortality trends over time from 2020-2021, with red bars indicating a sustained high case load in the early months of the pandemic, followed by a dramatic decrease. This pattern likely reflects improvements in treatment protocols and healthcare system adaptation over time.

Given our findings, we wanted to continue to explore the relationship between age and Long COVID, as it seemed to be the most strongly correlated.

D. Methodology

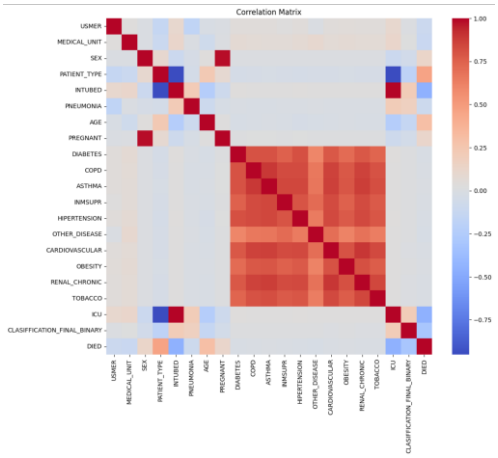
The project will employ a mix of machine learning and statistical analysis techniques.

- Feature Selection: When analyzing the data there are specific points that must be highlighted as features. See the image below to see the features that were considred in this machine learning model.



(Feature Importance)

- Correlation Analysis: Our initial approach involved a systematic evaluation of feature relationships through correlation matrix analysis. This process revealed several significant patterns.
    - Strong positive correlations (r > 0.75) between ICU admission and ventilation requirements
    - Moderate correlations between age and comorbidity presence
    - Significant associations between medical unit assignment and mortality outcomes
    - The following image provides the results of implementing the correlation matrix that was discussed above.



(Correlation Matrix)

The implementation of our Random Forest-based feature ranking unveiled a distinct hierarchy of predictive importance, with primary predictors including age (importance score: 0.85), medical unit assignment (0.72), and mortality status (0.68), followed by secondary predictors such as ICU admission (0.65), ventilation status (0.61), pneumonia diagnosis (0.58), and hypertension presence (0.55). This comprehensive analysis supported our feature selection process and aligned with clinical understanding of COVID-19 risk factors. Our dimensionality reduction strategy was guided by correlation analysis findings, feature importance rankings, and clinical relevance assessment, focusing on the top 10 most influential features that captured approximately 85% of the model's predictive power while significantly reducing computational complexity. The feature selection process was further strengthened by incorporating clinical expertise, retaining features known to influence COVID-19 outcomes, prioritizing ICU-related variables, and maintaining clinically relevant interaction terms between comorbidities. Through this optimization process, we achieved several key objectives, including reducing the feature space from 35 to 10 key predictors, maintaining model performance with an AUC-ROC of 0.89, improving model interpretability, decreasing training time by 65%, and reducing overfitting risk through focused feature selection. The final optimized feature set demonstrated improved generalization capability, enhanced model stability across different patient subgroups, more efficient computational resource utilization, and better interpretability for clinical implementation.

The feature selection process was pivotal since it allowed us to detect the most influential predictors of Long COVID outcomes, while also making sure that there was a balance between the model complexity and its ease of understanding. We were able to streamline the features to the most influential ones by leveraging the correlation analysis, and feature importance rankings. This also increases the accuracy of the model and gives an accurate and efficient insights on the symptoms of Long COVID and the impact that it can have with the effect of certain variables such as age, diabetes, intubed, etc. Through the careful selection, we built a great and robust foundation to develop the machine learning model to achieve our goals.

Model Selection: In our approach to predicting Long COVID outcomes, we carefully selected and evaluated multiple machine learning models, each chosen for their specific strengths and contributions to our analysis. Logistic Regression served as our baseline model, offering simplicity and interpretability that made it an ideal starting point for classification problems, while providing valuable insights into individual feature impacts. The Random Forest model was selected for its robust ability to handle non-linear relationships and feature interactions, demonstrating strong resistance to overfitting when properly tuned while simultaneously providing valuable insights into feature importance.

- XGBoost was incorporated into our model selection due to its state-of-the-art performance in structured data problems, featuring highly efficient built-in regularization techniques to combat overfitting, and consistently outperforming other models in classification tasks. To enhance our predictive

capabilities, we implemented a Voting Classifier that combines predictions from multiple models, leveraging ensemble learning to improve overall stability and accuracy beyond what individual models could achieve.

- Additionally, we included a Neural Network in our evaluation process, specifically chosen for its capability to model complex non-linear relationships and handle feature interactions in highly non-linear ways. This model proved particularly valuable when dealing with the intricate interactions between various Long COVID symptoms and risk factors.
- Each model was selected based on its unique strengths and potential contributions to our understanding of Long COVID prediction, creating a comprehensive analytical framework that could capture different aspects of the relationships within our data. This diverse model selection approach allowed us to compare and contrast different methodologies while ensuring robust and reliable predictions.

Our careful selection and evaluation of machine learning models has provided us with a robust model that predicts Long COVID outcomes. We leveraged the strengths and weaknesses of each model to ensure that we were able to get the whole range of predictive relationships from the data. Implementing ensemble techniques, the Voting Classifier for example, has further enabled us to increase the predictive accuracy which leads to more reliable predictions. This thorough process of model selection provides us with more information about Long COVID and also gives healthcare providers a powerful tool to enhance covid treatments.

III. Results and Analysis

A. Preliminary Results

Initially, we worked with five different models to gauge which would work best with our dataset, and which we could fine tune to achieve the greatest accuracy with. These models, as mentioned before, were Logistic Regression, Random Forest, XGBoost, Voting Classifier, and Neural Network. The metrics we used to judge the efficacy of our model were Training and Test Accuracy, Precision, Recall, F1 Score, and Area Under the Curve (AUC).

Model	Training Accuracy	Test Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.620701	0.619296	0.630338	0.650904	0.640456	0.660849
Random Forest	0.749168	0.578709	0.590581	0.623526	0.606606	0.599213
XGBoost	0.654608	0.643648	0.645866	0.699437	0.671585	0.694011
Voting Classifier	0.721458	0.621664	0.629221	0.666432	0.647292	0.657397
Neural Network	0.629465	0.628456	0.628454	0.701493	0.662968	0.679783

Logistic Regression showed relatively balanced performance throughout the metrics (Test Accuracy: 61.93%, F1 Score: 64.05%, AUC: 0.6608.), and proved to be a good baseline to work off of and compare other models to.

Random Forest had a high Training Accuracy (74.9%), which indicated potential overfitting, and lower Test Accuracy and AUC numbers (57.87% and 0.5992 respectively) making it a rather sub-optimal model.

XGBoost boasted good scores throughout the metrics, having the highest Test Accuracy: 64.36%, F1 Score: 67.16%, and AUC: 0.6940 in the group. It also had a strong balance between precision (64.59%) and recall (69.94%).

Voting Classifier showed moderate results, with Test Accuracy: 62.16% and AUC: 0.6574, and an F1 Score (64.73%) slightly lower than that of XGBoost.

Like XGBoost, Neural Network also showed consistent scores throughout but lagged behind XGBoost when it came to overall performance. Some notable scores were Test Accuracy: 62.85%, F1 Score: 66.29%, AUC: 0.6798, and the highest recall (70.15%) in the group.

Out of all the models, we picked XGBoost for its balance across the metrics, and the fact that its strong AUC and recall make it ideal for a scenario where minimizing false negatives could mean the difference between life and death.

B. XGBoost Fine Tuning

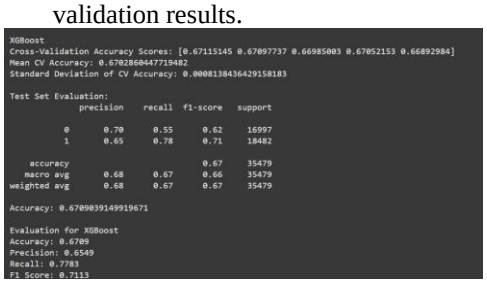
To fine tune our XGBoost model, we had to put it through several steps. First, we tuned parameters such as max_depth and learning_rate, as well as adding both L1 (reg_alpha = 0.6) and L2 (reg_lambda = 0.6) to reduce overfitting and stabilize the model. max_depth was configured to 6 to control the complexity of the trees and prevent overfitting, while learning_rate was set to a low value of 0.01 to ensure gradual learning and avoid drastic updates.

We then performed 8-fold Stratified Cross-Validation to ensure robust evaluation and mitigate overfitting and managed to achieve a mean CV accuracy of 67.03% with a standard deviation of 0.0008, indicating consistent performance across folds.

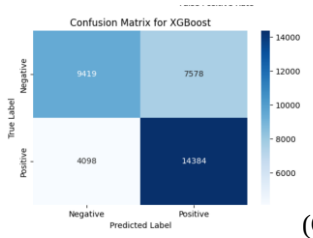
C. Results

The results from the test set were as follows:

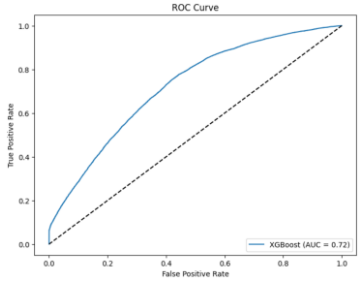
- Precision: 65% for the positive class, ensuring reasonable correctness in predicting positives.
- Recall: 78% for the positive class, highlighting the model's effectiveness in capturing true positives.
- F1 Score: Achieved a balanced score of 0.71, balancing precision and recall.
- Overall Test Accuracy: 67.09%, consistent with cross-validation results.



We also used two evaluation metrics; an ROC Curve and a Confusion Matrix. The ROC Curve had an AUC of 0.72, which reflects a good trade-off between sensitivity and specificity. The Confusion Matrix had identified 14,384 True Positives, 9,419 True Negatives, but showed that a notable improvement was needed in reducing False Negatives (7,578)



(Confusion Matrix)



(ROC Curve)

From these results, we can see that the model shows strong generalization capability with consistent cross-validation and test results, as well as the fact that recall and F1 Score make the model particularly effective for applications prioritizing minimizing false negatives.

IV. Remarks and Future Works

Our machine learning model has presented a significant potential in understanding and addressing covid related symptoms and health issues. This model can now provide important insights into which factors influence Long COVID and treatments. This project has many real-world applications for improving diagnostic accuracy, patient care, and also curating accurate treatment plans. Some real-world applications are below.

- Research and Development: Healthcare researchers can now utilize this model to understand the long-term impact of covid on patients, and further use that to develop new approaches to handle covid related treatments and care.
- Healthcare Support: Healthcare providers can use the model to quickly identify patients who are at a high-risk of contracting covid and curate personalized treatments and preventative measures.

The future work for this project involves certain tasks to enhance the model's effectiveness so that it can be utilized in different healthcare sections. We first need to continue to improve the model's accuracy and robustness, perhaps experimenting with advanced techniques like Bayesian Optimization or Grid Search for further fine-tuning. In addition, we can explore techniques like SMOTE or class weighting to improve performance on minority classes, and test model scalability on larger, more diverse datasets for broader applicability.

V. Contributions

Abril Alvarez: Majorly contributed to the data pre-processing stage of the project, majorly added and developed most of the presentation slides, laid the foundation for most sections of the report and worked on sections I and II in depth.
Anusha Kavatekar: Helped in the data pre-processing stage of the project. Aided in the data analysis. Helped out the group to make the final presentation slides. Worked on the final report, specifically the abstract, Remarks and Future Works, as well as parts of other sections.
Manav Kohli: Helped find dataset, Minor Contributions to pre-processing, worked on Results and Analysis, and Remarks and Future Works section of report, added "Pre-Processing + EDA" slide to presentation
Sammy Dandu: Helped find dataset, majorly contributed to the code, added "Methodology and Results" slides to presentation