Binary Classification of Long COVID Using Health-Related Survey Data

Tom Lupicki
tlupick1@jh.edu

Kenan Rustamov
krustam1@jh.edu

Akshat Chauhan
achauh14@jh.edu

1 Problem Statement

According to survey data released by the National Center for Health Statistics, as of September 2024, 61.6% of United States adults report having ever had COVID, and 17.9% report having ever experienced Long COVID, and 5.3% report that they are currently experiencing Long COVID (National Center for Health Statistics, 2022-2024). Worldwide, the total number of people that have had long COVID is 400 million, and the New York Times reports that the economic cost of the condition is $1 trillion annually (Belluck, 2024).

Long COVID may cause extended symptoms like brain fog and problems with taste or smell, and potentially cause or worsen other illnesses like heart disease, stroke, and diabetes (Staff).

However, there is still no standardized diagnostic test to identify Long COVID, as it is determined by physicians based on prior health records and reported symptoms. The goal of this study is to develop a machine learning-based binary classification model that can predict the presence of Long COVID from health-related survey data.

2 Importance

Long COVID presents a significant public health concern due to its high prevalence and long lasting negative effects the condition can have on people's lives. The current reliance on qualitative analysis by medical professionals for the diagnosis of Long COVID creates challenges for timely diagnosis and puts pressure on an already resource-limited setting.

A machine learning model for predicting Long COVID could be used as a pre-screening mechanism and has the potential to deliver fast, cheap identification of patients that are high risk for developing Long COVID. High-risk individuals and doctors can work together at an earlier stage to alleviate or avoid future symptoms.

3 Data

For this study we utilize the CDC's Behavioral Risk Factor Surveillance System (BRFSS) telephone survey data, available for download from the CDC website (CDC, 2024). This is a large-scale survey conducted across the United States every year. In the 2023 survey, Kentucky and Pennsylvania were excluded due to insufficient data collection, leading to approximately 433,000 rows of data in the total survey. This survey is routinely used for machine learning applications on health related problems.

The data collected includes responses for questions regarding a large variety of health factors, including "health status and healthy days, exercise, hypertension, cholesterol, chronic health conditions, demographics, [etc.]"[1]

The survey is split into Core, Optional, and State based questions. Due to the nature of the Optional and State based questions, they contain a lot of missing data per row. In this paper, we will focus mainly on the Core survey responses to select for nationwide data and not any specific state or subset of states, this leaves us with 74 core survey responses that can be used as features for our model.

There are approximately 200,000 responses on incidence of Long COVID in the 2023 survey. The data is of mixed type with both categorical and numerical features. An individual can respond with "Don't Know", "Refused", and there are partially complete surveys leading to more missing data. Some survey data questions are dependent on answering previous questions or answering previous questions with a specific response. This also leads to missing data.

Our binary target variable from the dataset is a yes or no answer to the survey question "Do you currently have symptoms lasting 3 months or longer

[1] https://www.cdc.gov/brfss/annual_data/2023/pdf/Overview_2023-508.pdf

that you did not have prior to having COVID-19?" This question is asked only to survey respondents who indicate that they have had COVID in the past.

There is a strong class imbalance for the binary Long COVID incidence question with approximately 85% of NO responses and 15% of YES responses without accounting for missing data. This is considered in the rest of the paper.

4 Methodology

4.1 Missing Data Challenge

As mentioned before individuals can respond to the questions with “Don’t know/Refused” according in the survey. We decided to use only the core questions along with some health indicating variables (age, race, BMI) from the dataset as they contained most of that data and had the least missing values enabling us to retain most of the dataset.

Given our data and domain knowledge modeling the mechanism of missingness in our dataset is challenging and likely intractable. It is also reasonable to believe that our data is MNAR because for any variable, for example income level, it is reasonable to believe that the mechanism of missingness depends at least to some extent on that variable. Additionally, because of the nature of our data originating in a survey, we notice that sequentially the number of missing answers increases, likely due to respondents ending the survey early. We also having missing data in the form of "Don’t know" and "Refused" answers. While these two answers do differ in their meaning, for the purposes of our project we treat them the same for numerical variables due to it being intractable to impute these with different methods that reflect the semantics of the answer. For categorical variables, we retain "Don’t know" and "Refused" as valid categories, under the motivation that they may hold information that would be lost when imputing or dropping.

We experiment with the following varying assumptions about the status of our dataset with regard to missing data, and preprocess the dataset in three ways:

1. **Assume MCAR:** Remove rows with NaN and numerical "Don’t know" or "Refused," retaining these as categories for categorical data. Models trained this way are Logistic Regression, Decision Tree, Random Forest,AdaBoost, HistGradientBoosting, LightGBM and XGBoost.

2. **Assume Missingness Contains Information:** Retain NaN rows for models that handle them and treat categorical "Don’t know"/"Refused" as categories without imputation.

3. **Assume MAR:** Use IterativeImputer (HistGradientBoostingRegressor) to impute missing values while keeping categorical "Don’t know"/"Refused" as categories. Fit the imputer on the training set and transform the test set.

4.2 Preprocessing

4.2.1 Feature Selection

Feature	Question
DIABAGE4	How old were you when you were first told you had diabetes? (Missing for non-diabetics)
NUMHHOL4	Not including cell phones or numbers used for computers, fax machines, or security systems, do you have more than one landline telephone number in your household?
NUMPHON4	How many of these landline telephone numbers are residential numbers?
CPDEMO1C	How many cell phones do you have for your personal use?
FLSHTMY3	During what month and year did you receive your most recent flu vaccine that was sprayed in your nose or flu shot injected into your arm?
HIVTSTD3	Not including blood donations, in what month and year was your last HIV test?
COVIDPO1	Have you ever tested positive for COVID-19?

Table 1: Dropped features and their corresponding questions.

For training, we limit our target variable coded as ’COVIDSM1’ to a binary 1 for positive samples and 0 for negative samples. We also include imputed age, race, and calculated BMI. We then proceed to do light feature selection on survey questions that are dates of events or questions that relate to the survey but not necessarily health data. These include age when diagnosed with diabetes for respondents with diabetes, number of landline phones in the household, number of residential phone numbers, number of cell phones, date of most recent flu vaccine, and date of the respondent’s last HIV test.

4.2.2 Standardization and Transformation

For categorical variables, we used Pandas to label columns, retaining "Don't know" and "Refused" as valid responses. To handle dependent follow-up questions, a "Missing" category (-1) was encoded to capture potential signals in the data patterns. For numerical data, we transformed the raw data to standardized units. Coded values such as '777' which indicated missing or refused responses that were converted to NaN.

Numerical preprocessing involved transforming raw data to standardized units, such as converting height and weight to metric measurements. Coded values like 777 or 999, indicating missing or refused responses, were converted to NaN.

4.2.3 Handling Missing Data

For specific features, logic based imputation is applied, i.e. missing values are filled based on logical assumptions derived from the context of the data. For example, if a respondent indicates that they never exercise, the missing value for "minutes spent exercising" is imputed as 0 instead of being left as NaN. This ensures that such straightforward cases are handled realistically.

For more complex missing values, Multivariate Imputation by Chained (MICE) as implemented in scikit-learn was used [2]. MICE treats each feature with missing value as a regression problem. Missing values are imputed iteratively based on the correlation between feature in question and the other features in the dataset. This imputation continues until convergence. Here, we used a Histogram-based Gradient Boosting Regression Tree[3] as the estimator for imputation because it natively handles categorical data, making it well-suited for the mixed data types present in the dataset. For numerical variables, "Don't know" and "Refused" responses are treated as NaN and imputed using the above methods. However, for categorical variables, "Don't know" and "Refused" are retained as valid categories. This decision is based on the rationale that these responses might hold meaningful information that would be lost through imputation or dropping.

To avoid data leakage, the imputation process is conducted only on the training set, and the trained imputer is subsequently used to transform the test set. This ensures consistency and prevents information from the test set influencing the training process. By combining logic-based rules with statistical imputation techniques ,we address missingness while preserving the integrity and predictive power of the dataset.

4.2.4 Dimension Reduction

We used Factor Analysis of Mixed Data (FAMD) as implemented in the Prince library to attempt to curb the large dimensionality of our data (Halford). While PCA is effective for continuous data and MCA is effective for categorical data, neither of them are effective for a mixed data type dataset (Abdi and Williams, 2010; Abdi and Valentin, 2006). FAMD attempts to solve this issue by combining the methods from PCA and MCA. It does this by normalizing numerical and categorical data, one hot encoding categorical data, then reducing the dimensionality to k dimensions that explain most of the data.

4.3 Handling Data Imbalance

The dataset had a strong class imbalance, with significantly higher number of negative cases (85%) compared to positive cases (15%). To address, this imbalance, we initially considered SMOTE-NC, a variant of Synthetic Minority Over-sampling Technique (SMOTE) that can handle mixed datasets (Bowyer et al., 2011). However, SMOTE-NC uses a nearest neighbors method to construct synthetic samples, and this may lead to potentially logically-inconsistent synthetic samples due to the existing dependence between some of the variables in the complete dataset.

Instead, we assigned higher weights to the minority class, ensuring that it is prioritized during model training using the native class imbalance weighting found in many of our models. We hope to improve our model's performance on the imbalanced classes using this technique.

4.4 Models

We decided the following machine learning models owing to their popularity and proven effectiveness in Medical Classification tasks (Ahsan et al., 2022). Our models chosen are described below. We also include Histogram-based Gradient Boosting Classification Tree (HistGradientBoosting) and LightGBM in our selection of models

[2] https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html

[3] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingRegressor.html

Many of the models also natively handle NaN values such as Descision Trees, Random Forest, HistGradientBoosting, XGBoost, and LightGBM. All of the models used support weighting functionality meant to handle class imbalances.

Logistic Regression[4] is a linear classification model that predicts the probability of binary outcomes. It uses a logistic function to map predictions between 0 and 1, making it efficient for linearly separable data.

Decision Tree [5] partitions data into subsets by evaluating feature thresholds, creating a flowchart-like structure. They are largely used for categorical data due to the nature of categorical data being easily split into two branches.

Random Forest [6] is an ensemble learning method that combines multiple decision trees to improve accuracy and robustness. Each tree is trained on a random subset of features and data, preventing overfitting to the data.

AdaBoost [7] AdaBoost builds a strong classifier by combining several weak classifiers, typically shallow decision trees. It assigns higher weights to misclassified samples in each iteration, ensuring the model focuses on harder-to-predict instances.

HistGradientBoostingClassifier [8] is a fast and memory-efficient boosting algorithm optimized for tabular data. It creates histograms from numerical features for faster processing. missing values.

XGBoost (Chen and Guestrin, 2016) and LightGBM[9] are two different implementations of gradient boosting decision trees.

5 Evaluation Metrics

To evaluate the performance of the machine learning models, we employ the following metrics:

[4] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[5] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[6] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[7] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html

[8] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingClassifier.html

[9] https://github.com/microsoft/LightGBM

5.1 Accuracy

Accuracy measures the proportion of correctly predicted instances out of the total instances. It is defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where:

- TP = True Positives (correctly predicted positive instances)
- TN = True Negatives (correctly predicted negative instances)
- FP = False Positives (incorrectly predicted positive instances)
- FN = False Negatives (incorrectly predicted negative instances)

5.2 Precision

Precision, also known as Positive Predictive Value, measures the proportion of correctly predicted positive instances out of all predicted positive instances. It is given by:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

5.3 Recall

Recall, also referred to as Sensitivity or True Positive Rate, calculates the proportion of correctly predicted positive instances out of all actual positive instances. It is defined as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

5.4 F1-Score

The F1-Score is the harmonic mean of Precision and Recall, providing a balance between these two metrics. It is given by:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

6 Results and Discussion

6.1 Experiment 1: Dropping All Missing Values

Here, we drop all the missing values in the dataset and evaluate the performance of each model with and without FAMD.

Table 2: Performance metrics of models without FAMD (Drop All Missing Values).

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.6896	0.2382	0.5949	0.3401
Decision Tree	0.7745	0.1888	0.2054	0.1967
Random Forest	0.8661	0.6222	0.0117	0.0230
AdaBoost	0.6878	0.2372	0.5963	0.3394
HistGradientBoosting	0.6833	0.2363	0.6069	0.3401
XGBoost	0.7075	0.2392	0.5390	0.3314
LightGBM	0.6860	0.2373	0.6032	0.3406

Table 3: Performance metrics of models with FAMD (Drop All Missing Values).

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.6834	0.2320	0.5862	0.3324
Decision Tree	0.7801	0.1830	0.1834	0.1832
Random Forest	0.8658	0.7410	0.0028	0.0056
AdaBoost	0.6567	0.2192	0.6060	0.3220
HistGradientBoosting	0.6826	0.2307	0.5824	0.3304
XGBoost	0.7335	0.2330	0.4279	0.3017
LightGBM	0.6883	0.2321	0.5709	0.3301

In Table 2 and Table 3 we see the inclusion of FAMD in Experiment 1 did not yield any significant improvement in performance compared to without FAMD as measured by the F-1 Score and Recall. Our best performing model, based on F-1 Score and Recall in this experiment was HistGradientBoosting without FAMD although Logistic Regression, AdaBoost, HistGradientBoosting, LightGBM, and XGBoost are all relatively similar. Descision Tree and Random forest exhibit high Accuracy but very low F-1 Scores and Recall.

6.2 Experiment 2: Keeping All Missing Values

Here, we test the performance of the models that can natively handle missing values without FAMD. FAMD cannot be used in this case as the implementation of FAMD that we use does not handle missing values. If we were to use an implementation that handles missing values, then it would do so through imputation. We already to FAMD on imputed data in our next experiment.

Table 4: Performance metrics models natively handling missing values

Model	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.7729	0.1953	0.2117	0.2032
Random Forest	0.8634	0.5301	0.0106	0.0209
HistGradientBoosting	0.6803	0.2387	0.6111	0.3433
XGBoost	0.7032	0.2408	0.5436	0.3338
LightGBM	0.6830	0.2399	0.6077	0.3440

In Table 4 we see that HistGradientBoosting and LightGBM performed best, balancing Recall and F1 score. As before, Logistic Regression, AdaBoost, HistGradientBoosting, LightGBM,and XGBoost are all relatively similar, and Descision Tree and Random forest exhibit high Accuracy but very low F-1 Scores and Recall.

6.3 Experiment 3: Imputing the Missing Values

Here, we test the performance of the models on the dataset where the missing values have been imputed with IterativeImputer with and without FAMD.

Table 5: Performance metrics of models without FAMD (Impute Missing Values).

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.7107	0.2295	0.4883	0.3123
Decision Tree	0.7772	0.1782	0.1820	0.1801
Random Forest	0.8648	0.3977	0.0110	0.0213
AdaBoost	0.6741	0.2202	0.5600	0.3161
HistGradientBoosting	0.6696	0.2190	0.5677	0.3160
XGBoost	0.6942	0.2214	0.5063	0.3081
LightGBM	0.6719	0.2199	0.5655	0.3167

Table 6: Performance metrics of models with FAMD (Impute Missing Values).

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.6834	0.2320	0.5862	0.3324
Decision Tree	0.7801	0.1830	0.1834	0.1832
Random Forest	0.8658	0.7410	0.0028	0.0056
AdaBoost	0.6567	0.2192	0.6060	0.3220
HistGradientBoosting	0.6826	0.2307	0.5824	0.3304
XGBoost	0.7391	0.2360	0.4199	0.3022
LightGBM	0.6883	0.2321	0.5709	0.3301

In Table 5 and Table 6 we see the inclusion of FAMD in Experiment 3 did not yielded slight performance to LightGBM, AdaBoost, Logistic Regression, and XGBoost although not enough to be a significant change to without FAMD as measured by the F-1 Score and Recall. LightGBM displayed the best performance with FAMD and AdaBoost with FAMD. As before, Logistic Regression, AdaBoost, HistGradientBoosting, and XGBoost are all relatively similar, and Descision Tree and Random forest exhibit high Accuracy but very low F-1 Scores and Recall.

7 Conclusion

Our models demonstrate slight predictive capability for predicting Long COVID with an accuracy of approximately .7 and recall of approximately .6. Our findings could serve as an initial step in identifying individuals at high risk for developing Long COVID. However, our precision is considerably

low across most of our models, leading to positive Long COVID classifications for individuals who do not have Long COVID. The high class imbalance, missing data, and inherent lack of predictive strength of the data most likely limited our model's ability to predict Long COVID. The dataset lacks large amount of detailed healthcare data (vitals, blood pressure, medication history) that a doctor would have access to when seeing patients that may have classify as having Long COVID or be at risk of developing Long COVID post-COVID infection.

8 Possible Extensions

Our work provides an important insight into the effectiveness on using machine learning techniques to predict positive cases of Long COVID given information that covers an individual's demographics, lifestyle, and healthcare history. Possible extensions to our paper include the following.

Next year's BRFSS may include the same question to Long COVID as our current BRFSS allowing us to merge both years data for prediction. This could help alleviate issues with data imbalance by having enough examples in the imbalanced class for prediction.

In the future, a different dataset, with more medical information about patients could be used as a more predictive dataset for Long COVID. This is similar to the approach by (Pfaff et al., 2022). This health information dataset would have the benefit of not being self-reported dataset, avoiding many of the pitfall of self-reported survey data mentioned previously. Expert analysis could be used to reduce our feature set to the most predictive feature.

Another potential area for extensions based on our work is research into why people may be inclined to refuse to answer specific survey questions in order to gain insight into the mechanism of missingness for one possible dimension of our missing data. having this greater insight would allow us and future researchers to develop methods to more accurately impute missing data values in similar surveys that are meant to comprehensively survey national populations, especially when questions about factors like income are included that have far greater rates of refusal to answer than other questions.

References

Hervé Abdi and Dominique Valentin. 2006. Multiple correspondence analysis.

Hervé Abdi and Lynne J. Williams. 2010. Principal component analysis. *WIREs Computational Statistics*, 2(4):433–459.

Md Manjurul Ahsan, Shahana Akter Luna, and Zahed Siddique. 2022. Machine-learning-based disease diagnosis: A comprehensive review. *Healthcare*, 10(3).

Pam Belluck. 2024. About 400 million people worldwide have had long covid, researchers say. *The New York Times*.

Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. 2011. SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813.

CDC. 2024. 2023 brfss survey data and documentation.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754.

Max Halford. Prince.

National Center for Health Statistics. 2022-2024. Household pulse survey. *U.S. Census Bureau*.

Emily R Pfaff, Andrew T Girvin, Tellen D Bennett, Abhishek Bhatia, Ian M Brooks, Rachel R Deer, Jonathan P Dekermanjian, Sarah Elizabeth Jolley, Michael G Kahn, Kristin Kostka, et al. 2022. Identifying who has long covid in the usa: a machine learning approach using n3c data. *The Lancet Digital Health*, 4(7):e532–e541.

Mayo Clinic Staff. Long covid: Lasting effects of covid-19.