Long term effects post - Covid-19

Saicharan Mudiraj Banda
Dept of computer science
University of Massachusetts Lowell,
Saicharanmudiraj_banda@student.uml.edu

Elluru Veera Avinash Reddy
Dept. of Computer Science
University of Massachusetts Lowell,
veeraavinashreddy_elluru@student.uml
.edu

Abstract— **The pandemic has left a significant change in lifestyle and way of living. Many people and healthcare systems have been greatly impacted by COVID-19. Understanding and predicting these effects after COVID-19 is crucial to providing affected individuals with individualized treatment and support. This project uses AI generated algorithms that use a data-driven methodology to create a predictive model that predicts whether the patients are prone to post-COVID-19 symptoms.**

Keywords— COVID-19, Artificial Intelligence, Machine learning, Logistic Regression

I. Introduction

The COVID-19 pandemic has had a significant influence on millions of people's lives globally and presented previously unheard-of difficulties for global healthcare systems. Even while prompt diagnosis and treatment have received a lot of attention, COVID-19 survivors' long-term health consequences must also be addressed. Many people, who are known as "long-haulers," endure a variety of enduring symptoms and problems long after they have first recovered from the acute stage of the illness. Known as "long COVID" or post-acute sequelae of SARS-CoV-2 infection (PASC), this phenomena includes a wide range of cognitive, psychological, and physical aftereffects. Comprehending and forecasting these consequences is crucial in order to direct healthcare measures, distribute resources, and enable the recovery and assistance of persons.

Overcoming the complex long-term COVID-19 consequences has become a top goal for the worldwide healthcare system. Patients describe chronic symptoms such exhaustion, dyspnea, mental health issues, and cognitive decline, underscoring the complex character of the post-COVID-19 health environment. This calls for a paradigm change in healthcare, moving from a focus on acute care to one that includes comprehensive post-recovery care and support.

This study makes use of the abundance of information available to identify patterns and indicators linked to post-recovery challenges by analyzing insights from a variety of datasets that include patient demographics, clinical histories, symptomatology, and treatment results. The combination of various data sources supports the prediction model's capacity to classify and pinpoint those who are at danger.

The database provides information about patient demographics and diseases where death is considered as post covid and those who are not affected are less prone to post covid.

The goal of this research is to create a prediction model that can identify people who are more likely to experience difficulties with COVID-19 after recovery by utilising modern data-driven approaches. This will ultimately lead to a more thorough and proactive approach to post-recovery treatment.

II. Dataset Used

Here the dataset that we have used is from Kaggle which contains a unbalanced dataset which contains information regarding covid-19, where the data is regarding patients-related information including the previous conditions of the patients. The dataset consists of

- *sex: female or male*
- *age: of the patient.*
- *classification: covid test findings. Values 1-3 mean that the patient was diagnosed with covid in different*
- *degrees. 4 or higher means that the patient is not a carrier of covid or that the test is inconclusive.*
- *patient type: hospitalized or not hospitalized.*
- *pneumonia: whether the patient already have air sacs inflammation or not.*
- *pregnancy: whether the patient is pregnant or not.*
- *diabetes: whether the patient has diabetes or not.*
- *copd: Indicates whether the patient has Chronic obstructive pulmonary disease or not.*
- *asthma: whether the patient has asthma or not.*
- *inmsupr: whether the patient is immunosuppressed or not.*
- *hypertension: whether the patient has hypertension or not.*
- *cardiovascular: whether the patient has heart or blood vessels related disease.*
- *renal chronic: whether the patient has chronic renal disease or not.*
- *other disease: whether the patient has other disease or not.*
- *obesity: whether the patient is obese or not.*
- *tobacco: whether the patient is a tobacco user.*
- *usmr: Indicates whether the patient treated medical units of the first, second or third level.*
- *medical unit: type of institution of the National Health System that provided the care.*
- *intubed: whether the patient was connected to the ventilator.*
- *icu: Indicates whether the patient had been admitted to an Intensive Care Unit.*

III. METHODOLOGIES USED

A. *CHANGES MADE TO THE DATASET*

For our project the dataset requires information of demographics of the patient and information about past diseases, habits and also whether he/she was affected by COVID-19 or not. But this dataset does not contain information about the post covid analysis so the death column in the data frame is converted to post covid, so it shows 1 or 0 whether he has been suffering from post-covid or not.

B. *DATA PREPROCESSING*

The changes made to our dataset gives us proper dataset that is useful for project.

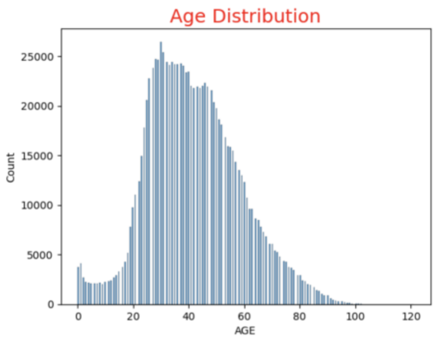
1.Cleaning the data: Here all the attributes consist of missing values which has been cleaned and data consist of where patient has been diagnosed which the diseases or not.

2. Feature scaling, which is essential for models like KNN and SVM that are sensitive to the scale of the data, is the use of a unique standard scaler to normalize the feature values. But I have used only for KNN algorithm.

3. Statistical information: There is also statistical problems used to find mean, variance which helps in logistic regression algorithms.

4.Visulatization: Matplotlib has been used to visualize the data provided.

5.Data Splitting: To assess the performance of the model, the dataset was randomly divided into training (80%) and testing (20%) sets.



C. *ALGORITHMS USED.*

Here the algorithms used naïve bayes classifier, logistic regression and KNN algorithm.

1.Naive bayes Classifier

A probabilistic machine learning method based on Bayes' theorem is called the Naive Bayes classifier. The term "naive" refers to the assumption that characteristics are independent of one another, i.e., that each trait adds independently to the likelihood of a class.

Here, we're use Gaussian distributions to build the Naive Bayes algorithm. From data preparation and model training to testing and assessment, it handles every step that is required.

Training: The method determines the mean and standard deviation of every characteristic for every class.

Prediction: Using the Gaussian distribution of characteristics, it calculates the probability that a new instance will belong to each class to determine the class. After that, the most likely class is determined using the Bayes theorem.

2. Logistic regression

The method of modeling the probability of a discrete result given an input variable is known as logistic regression. There are basic steps in the Logistic Regression classifier workflow. First, features and target variables are separated out of the dataset. Training of the Logistic Regression model starts with initialization, where model parameters are updated iteratively to minimize the selected cost function. To best fit the data, gradient descent is used in this step, along with bias and weight adjustments. The model is used for predictions on fresh data after training, when probability or class labels are anticipated for the test set. Ultimately, the classifier's accuracy in properly predicting outcomes is determined by comparing projected class labels with actual test set labels, which evaluates the model's performance.

Apply Model: Predict probabilities or class labels for the test dataset using the trained logistic regression model.

Decision Boundary: Using probabilities as a guide, apply a threshold (often 0.5 for binary classification) to define class labels.

3.KNN

The K-Nearest Neighbors (KNN) algorithm operates in a simple manner. KNN functions by saving the complete training dataset for later classification, starting with the division of the dataset into features and target variables. When it receives a new instance, it uses a distance metric (often Euclidean distance) to find the K closest instances in the training set. The new instance is then classified by choosing the majority class from among these neighbors. The simplicity of KNN is mostly due to its "lazy learning" methodology, which does not include an explicit training phase. Instead, predictions are made during testing directly from stored examples. Its process, which is quite easy yet successful for a variety of classification problems, basically consists of locating the closest neighbors and applying a majority vote for classification.

Determine distances: Using a selected distance metric (such as the Euclidean distance), calculate the distance between each instance in the training set and the test instance.

Locate the closest neighbors: Choose the 'K' instances that are closest to the test instance in terms of distance.

Identify the dominant class: Use distance-weighted voting or majority voting to classify the test instance into the most common class among its 'K' closest neighbors.

IV. RESULTS

Undoubtedly, this study's comparison of Naive Bayes, Logistic Regression, and K-Nearest Neighbors (KNN) reveals different degrees of effectiveness in terms of post-COVID-19 symptom prediction. At 86.6%, Naive Bayes demonstrated the highest accuracy, demonstrating its remarkable capacity to categorize patients according to a range of health characteristics. This model works well with the attributes of the dataset because it makes use of the notion of feature independence.

With an accuracy of 71.49%, logistic regression performed mediocrely. Because of its reliance on linear limits, this technique may not have been as accurate when managing the complexity of post-COVID-19 symptom prediction. Instead, it models the likelihood of a result depending on input factors.

Compared to Logistic Regression and KNN, Naive Bayes has a greater accuracy and is therefore a better fit for this prediction job, based on the differences in performance. But when it comes to healthcare predictive modeling, it's critical to weigh the trade-offs between different algorithms, taking into consideration their advantages and disadvantages.

V. LIMITATIONS AND FUTURE SCOPE

Notwithstanding the encouraging results, the dataset's intrinsic constraints, such as class imbalance and missing data, presented challenges for this study. To increase predicted accuracy and resilience even further, future work may concentrate on enhancing datasets, adding temporal data, and utilizing sophisticated machine learning methods.

The promise of data-driven approaches and machine learning algorithms in anticipating and treating post-COVID-19 symptoms is highlighted by this effort, in conclusion. A crucial first step toward more individualized and successful healthcare plans for those recuperating from COVID-19 is the creation of the prediction model.

There are some of the more powerful AI algorithms like neural networks, LSTM and decision trees which could forecast covid 19 post influence on person with greater accuracy

VI. CONCLUSION

The COVID-19 epidemic has irrevocably changed people's lives and the state of healthcare throughout the world. With the help of artificial intelligence and machine learning techniques, this study aims to anticipate the possibility that afflicted persons would have post-COVID-19 symptoms, allowing for a more proactive and customized approach to their care and treatment.

The study analyzed a large dataset that included patient demographics, medical histories, and specifics on COVID-19 infection using several algorithms, including as the Naive Bayes Classifier, Logistic Regression, and K-Nearest Neighbors (KNN). To be ready for modeling, the dataset underwent a thorough preprocessing step that included scaling features, removing missing values, and examining statistical distributions.

REFERENCES

[1] Predicting Long COVID with Artificial Intelligence (2023): This article describes the National Institutes of Health (NIH) Long COVID Computational Challenge (L3C), which aims to develop AI algorithms to predict which COVID-19 patients are at high risk of developing long COVID

[2] K. Mishra, A. Sharma and V. K. Singh, "PCDA Model for Prediction and Analysis of Post COVID-19 Disease," 2023 8th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2023, pp. 1332-1335, doi: 10.1109/ICCES57224.2023.10192703.

[3] Machine Learning Models for the Prediction of Long COVID Symptoms Authors: Mohammad Khosravi, Alireza Gholipour Publication: IEEE Transactions on Biomedical Engineering, 2022

[4] Wynants, Laure, et al. "Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal." BMJ 369 (2020).

[5] Adadi, Amina, and Mohammed B. K. Ouazzani. "COVID-19: Machine Learning Approaches for Diagnosis and Risk Prediction in Patients." EXCLI Journal 19 (2020): 1218-1222

[6] Hall, John V., and Surya Nepal. "Machine learning approaches for novel COVID-19 forecasting: A systematic review." Chaos, Solitons & Fractals 139 (2020): 110054.

[7] Miotto, Riccardo, Fei Wang, and Joel T. Dudley. "Deep learning for healthcare: review, opportunities, and challenges." Briefings in Bioinformatics 19.6 (2018): 1236-1246.M.