COVID 19 DETECTION USING MACHINE LEARNING

Submitted by : PRASTUT MAHTHA

Registration No. 25BAI11044

***Abstract*—Technological advancement has a profound effect on all spheres of life, whether in the medical field or in any other field. Artificial intelligence has shown promising results in health care by making its decisions by analyzing and processing data. To prevent the spread and development of a life-threatening disease, the most important step is its early diagnosis. COVID-19 is a highly contagious disease, and has become a global epidemic that needs to be addressed as soon as possible. Due to its rapid speed of spreading comes the need for a system which can be used to detect the virus. With the increase in use of technology, lots of data about COVID-19 is readily available at our fingertips, which can be used to obtain important information about the virus. In this project, we compared the accuracies of different machine learning algorithms in predicting COVID-19 and used the most accurate one in the final model testing.**

I. Introduction

In December 2019, the novel coronavirus appeared in the city of Wuhan in China [1] and was reported to the World Health Organization (WHO) on 31 December 2019. The virus posed a global threat and was named COVID-19 by the WHO on the 11th. February 2020. W.H.O declared the outbreak a public health emergency [2] and stated the following; "the virus is spread through the respiratory tract when a healthy person comes in contact with an infected person". An infected person shows symptoms within 2-14 days. According to W.H.O the symptoms and signs of moderate to severe conditions are dry cough, fatigue and fever while in severe cases dyspnea, fever and fatigue may occur. People with other illnesses such as asthma, diabetes, and heart disease are at greater risk of contracting the virus and may become seriously ill. A system which can be used to detect the virus has become necessary due to the rapid spread of the virus, killing hundreds of thousands of people. Machine learning classification algorithms, data sets and machine learning software are essential tools for designing the COVID-19 predictive model.

This project aims to compare different machine learning algorithms like K-nearest neighbors, Random forest and Naive Bayes with respect to their accuracies and then use the best one among them to develop a system which predicts whether a person has COVID or not using the data provided to the model.

II. Related Work

A strong and accurate diagnosis of COVID-19 can save millions of lives and deliver a large amount of information on which machine learning (ML) models can be trained. ML may provide helpful inputs in this regard, especially for making diagnostics based on clinical literature, radiography images, etc. Studies in [3] are shown to effectively differentiate COVID-19 patients in 85% of cases using a support vector (SVM) algorithm. The study analyzed the results of COVID-19 tests done in Hospital Israelita Albert Einstein (HIAE) in São Paulo, Brazil. It was one of the main Covid testing centers in the country during the initial weeks of the outbreak. This study was conducted by a task force whose purpose is to respond to the COVID-19 emergency. It tested the performance for positivity of COVID of some machine learning algorithms (neural networks, gradient boosted trees, random forests, logistic regression and SVM). The study in [4]used several dividers, including logistic regression, multilayer perceptron (MLP) and XGBoost in the same database from the Brazil hospital [9]. It classified COVID-19 patients with an accuracy above 91%. The work in [5] developed and evaluated an ML algorithm for Covid-19 diagnosis. The algorithm was designed based on demographics and lab features. They collected data from the UCLA Health System in Los Angeles, California. It included all emergency rooms and inpatient cases receiving SARS-CoV-2 PCR testing with a set of 1,455 ancillary laboratory features from 1 March 2020 till 24 May 2020. They tested with some ML models and used a combination of these for the final classification. The developed algorithm had a sensitivity of 0.093 and a specificity of 0.64. The function in [7] predicts COVID-19 with 91% accuracy and 89%, respectively. In

addition, a prediction of ICU / semi-ICU needs was made in 98% of cases [6]. Since very little work is done on diagnostics and prediction using text, we have used machine learning models to classify clinical reports into COVID positive or COVID negative.

III. METHODOLOGY

The proposed methodology consists of 4 steps. In step 1 data collection is being performed and step 2 gives an overview of preprocessing, step 3 exploratory analysis is being performed to understand the data set and the last step, step 4 includes the hyperparameter tuning by grid search CV

A. Data Collection

As the WHO has declared the Coronavirus pandemic as a health emergency, researchers and hospitals have provided open access to data related to the epidemic. We procured a data set from kaggle.com and it has 5434 × 21 rows of columns. This dataset contains 20 variables that could be determinants in the prediction of COVID-19, as well as one class attribute that defines if COVID-19 is found.

B. Data Preprocessing

The process of converting raw data into a comprehensible format is known as data preprocessing. Real-world data may have noise, missing values, or be in an incompatible format that prevents it from being directly used in machine learning models. Preprocessing of data is an essential step in which we clean the data and make it compatible i.e. suitable to be used in a machine learning model. This also enhances the accuracy and efficiency of the model. The main steps in data preprocessing are as follows:

1) Removing features: From the figure 2 we can conclude that wearing masks and sanitization from the market are two features that have only one value that is 'no' as they don't affect our predictions we can simply just drop those columns off our dataset.

2) Encoding Categorical Data: Labeling Coding is a popular form of code flexible code management for categories. In this process, each label is given a whole number based on alphabetical order. All attributes of our dataset are of 'yes' or 'no' type so we have used label encoding to convert it to 0 and 1 for the model to understand the dataset better. Table 4 shows the dataset after applying label encoding.

3) Splitting the Dataset: The next stage in machine learning data preprocessing is to split the dataset. A machine learning model's dataset should be split into two parts: training and testing.
We divided the data into an 80:20 split. This means that we use 80% of the data to train the model while keeping the remaining 20% for testing. We take all the 20 independent attributes into x and the dependent column 'COVID-19' into y as we aim to predict if the patient is COVID positive or not.

	missing_values	percent_missing %
Breathing Problem	0	0.0
Fever	0	0.0
Dry Cough	0	0.0
Sore throat	0	0.0
Running Nose	0	0.0
Asthma	0	0.0
Chronic Lung Disease	0	0.0
Headache	0	0.0
Heart Disease	0	0.0
Diabetes	0	0.0
Hyper Tension	0	0.0
Fatigue	0	0.0
Gastrointestinal	0	0.0
Abroad travel	0	0.0
Contact with COVID Patient	0	0.0
Attended Large Gathering	0	0.0
Visited Public Exposed Places	0	0.0
Family working in Public Exposed Places	0	0.0
Wearing Masks	0	0.0
Sanitization from Market	0	0.0
COVID-19	0	0.0

Fig. 1. No. of missing values and missing percentage of all the attributes

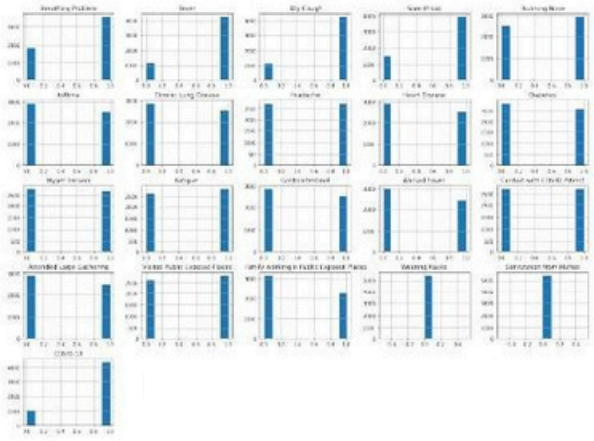

Fig. 2. Histogram of all the attributes

C. Exploratory Data Analysis

Exploratory data analysis is used to evaluate various datasets with the aim of summarizing them based on their key characteristics. This summary can be visualized using statistical graphics and other data visualization techniques. EDA helps data scientists in different ways:-

- Increasing the understanding of data
- Detecting a variety of data patterns
- Improved comprehension of the problem statement

D. Hyperparameter tuning by grid search CV

Its main goal is to discover the optimal parameters where the model's efficiency is the best or highest and the error rate is the minimum. We have used the gridsearchcv tool to produce the best combination of parameters, based on accuracy score as the scoring metric when all the different parameters are fed into the parameter grid.

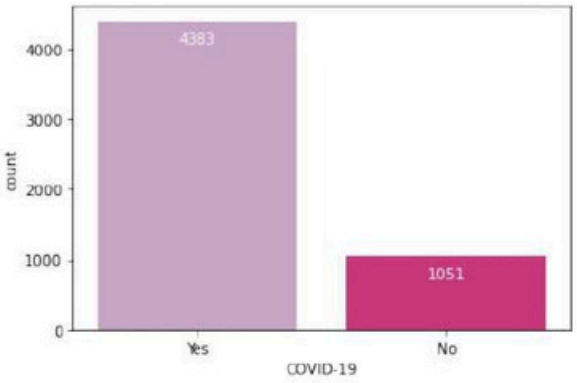

Fig. 3. Histogram showing the number of patients with covid positive and negative

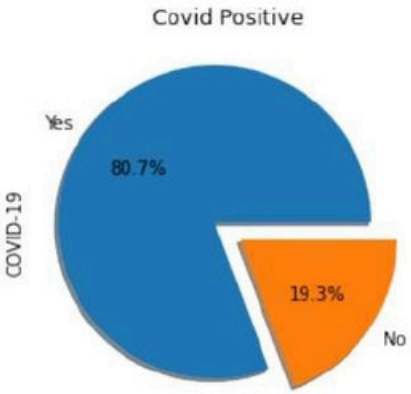

Fig. 4. Piechart showing the number of patients with covid positive and negative

IV. Implementation

With the growth of computer technology, predictive modeling is changing. We are now able to make predictable modeling more efficient, and less expensive than before. In our project, we use various classification algorithms to predict and use a grid search CV to find the most advanced solution for each algorithm. Some of the categorization algorithms that have been employed include:

A. Logistic Regression

Logistic regression is a data categorization technique that uses machine learning. This algorithm, models the odds of the potential outcomes of a single experiment using a logistic function. The easiest way to understand the influence of numerous independent factors on a single outcome variable is to use logistic regression, which was designed for this purpose. In general, the algorithm calculates the probability of belonging to a particular class. We have two classes here, y=0,1.

B. K Nearest Neighbors

The oldest supervised machine learning algorithm for classification is KNN, which classifies a given instance according to the majority of categories among its k-nearest neighbours in the dataset. The distance between the item to be categorized and every other item in the data set is calculated by the algorithm.

C. Random Forest

This classifier is a meta-estimator that adapts to decision trees on the dataset's different sub-samples and utilizes the average to increase the model's predicted accuracy and control over-fitting. In most circumstances, this random forest classifier seems to be more accurate than decision trees, and it also minimizes overfitting. At the Random Forest level, average over all the trees is the final feature importance. The feature's importance sum value on each tree is numerically calculated and divided by the total number of trees:

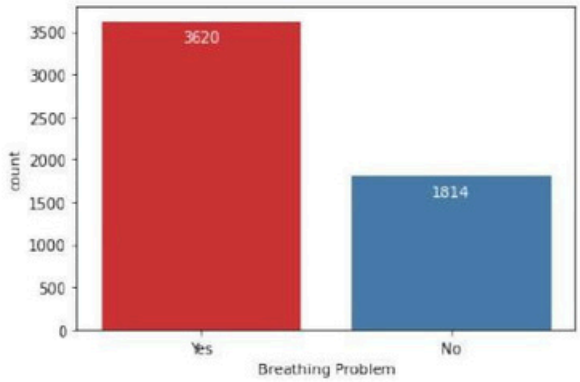

Fig. 5. Histogram showing the number of patients having breathing problems

V. Result and Discussion

To evaluate the effectiveness of the Machine Learning algorithms applied in this experiment, we decided to adopt the Accuracy, Mean squared error, Precision, Recall and F-Measure which are widely used in domains such as information retrieval, machine learning and other domains that involve binary classification.

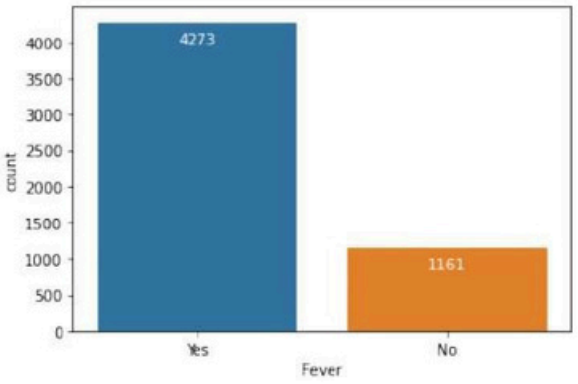

Fig. 6. Histogram showing the number of patients having fever

Figure 7 shows that all of the algorithms performed well during the training process, owing to the fact that the hyperparameters had been fine-tuned. However, there were minor discrepancies in accuracy, as indicated by the blue bar. The Random Forest Tree method has the best accuracy of all the algorithms, with a score of 98.39 percent. Additionally, R2 scores, mean squared errors and ROC scores are plotted in the bar chart using red, green and purple respectively. With 98.37 percent accuracy, the KNN is the next most acceptable algorithm to use.

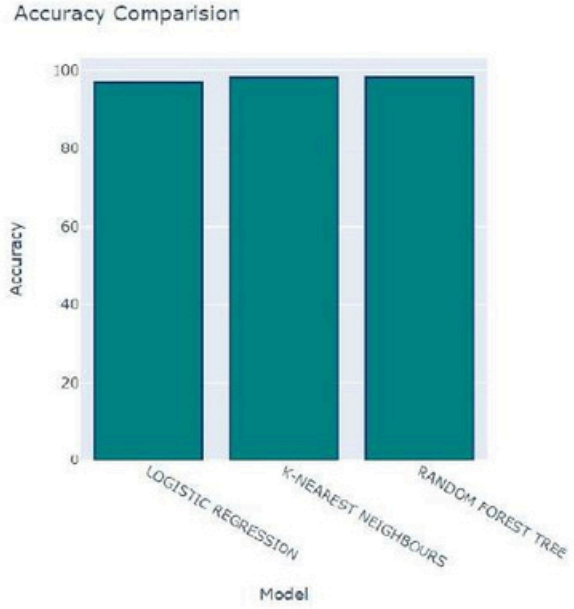

Fig. 7. Analysis of algorithms by their accuracy

The Mean Squared Error gives the difference between the expected value of the class and the actual class. This allows us to see how well the built model estimates the labels of the samples in comparison with the original values in the dataset. A low MSE model can be considered to be more efficient

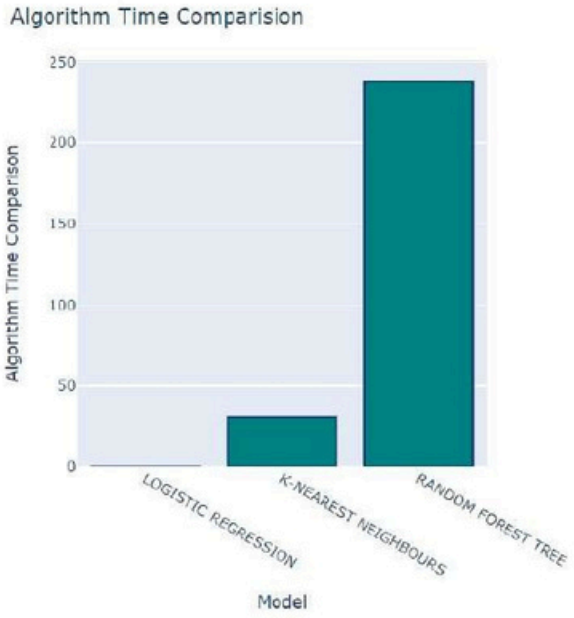

Fig. 8. Analysis of algorithms by time taken for fitting the model

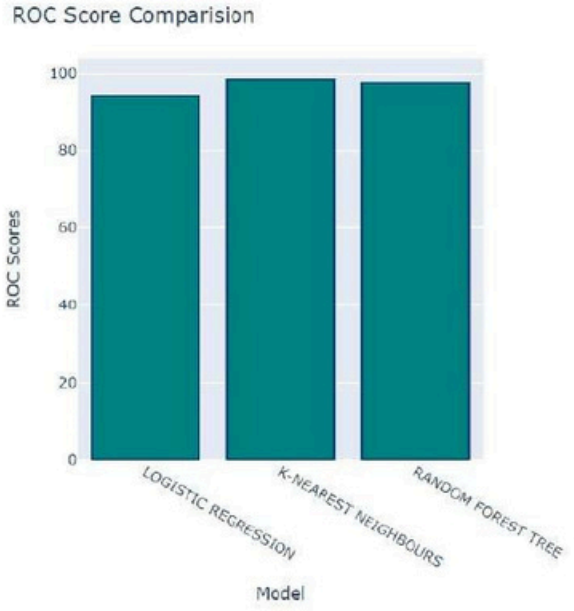

Fig. 9. Analysis of algorithms with ROC score

compared to a model having high MSE. The lowest MSE was attained by the Random Forest algorithm, with 2.207% followed by KNN having 2.57%.

R2 scores measure how well the model fits over the data. So here we see that Random Forest have the highest score of 85.51% followed by KNN with 83.1%. Finally, as shown in Figure 10, the time it took to develop the model was taken into account, as this is an important factor in determining the

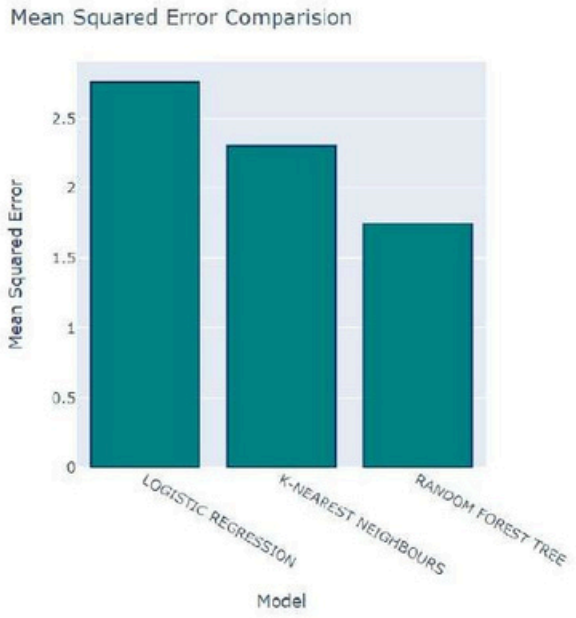

Fig. 10. Analysis of algorithms with MSE

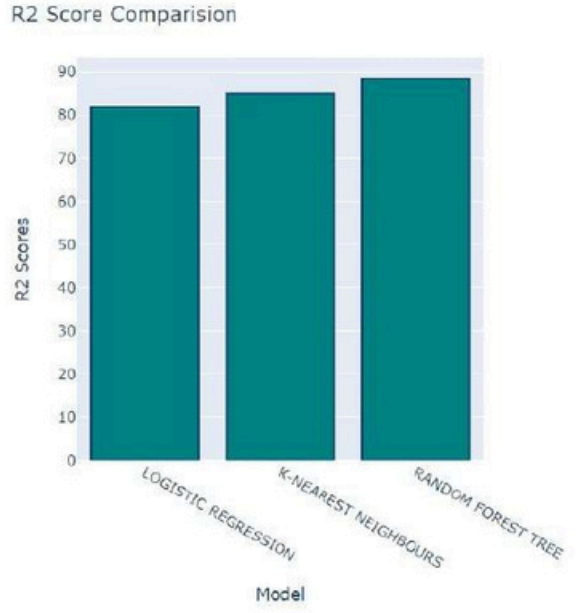

Fig. 11. Analysis of algorithms with R2 score

best algorithm for our prediction model. Based on the results, Logistic Regression was the fastest algorithm with 0.05s to train a classifier model.

VI. CONCLUSION

The goal of this work was to use the three supervised machine learning techniques to create a COVID-19 presence predicting model. The model's performance was evaluated

	Accuracy	MSE	R2 score	ROC score	Running time
KNN	98.37%	2.57	83.1	98.58	24.252
Logistic Regression	97.03%	3.036	80.086	93.23	0.038
Random Forest	98.39%	2.207	85.51	97.41	213.331

TABLE I
COMPARISON OF METRICS FOR KNN, LOGISTIC REGRESSION AND RANDOM FOREST

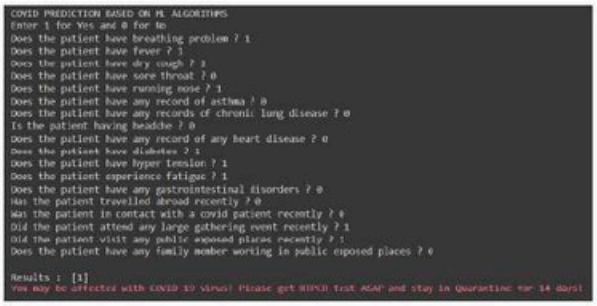

Fig. 12. Prediction model takes input from the user and gives a result - COVID Negative

in a comparative analysis. The results show that the KNN classifier with number of neighbors to be considered equal to 2 is the best machine learning algorithm, having an accuracy of 98.37%, and 0.026 mean absolute error considering the runtime for training. In comparison to other methods, the model takes average time but gives good accuracy.

This research can be used as a supporting tool for decision-making by doctors, with the established model assisting in recognising COVID-19 presence in a person based on their symptoms. Individuals who are suffering COVID-19-related symptoms can also use it to assess if they would be tested positive or negative for COVID-19. The model that has been developed here can be employed to deploy an app with the following features:

Fig. 13. Prediction model takes input from the user and gives a result - COVID Positive

- Individuals can quickly determine whether they are at risk of transmitting COVID-19 based on their symptoms.
- Medical practitioners can employ this test as a primary health assessment for COVID detection.
- Assisting businesses in limiting physical interaction with clients who may be infected with COVID-19;

Extra information or diagnoses from hospital records, persons who contracted the virus, COVID-19 survivors, patients under assessment, or management can all be included for future research. A software which can predict the severity of COVID-19 can indeed be deployed to provide further information about the steps that must be taken and the interventions that should be considered.

REFERENCES

[1] *Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ (2020) A new coronavirus associated with human respiratory disease in china. Nature 44(59):265–269* https://www.nature.com/articles/s41586-020-2008-3 [1]

[2] *Medscape Medical News, The WHO declares public health emergency for novel coronavirus (2020)* https://www.medscape.com/viewarticle/924596 [2]

[3] *A. F. M. Batista, J. L. Miraglia, T. H. R. Donato and A. D. P. C. Filho, "COVID-19 diagnosis prediction in emergency care patients: a machine learning approach", medRxiv, 2020* https://www.medrxiv.org/content/10.1101/2020.04.04.20052092v2.full.pdf [3]

[4] *M. R. H. Mondal, S. Bharati, P. Podder and P. Podder, "Data analytics for novel coronavirus disease", Informatics in Medicine Unlocked Elsevier, vol. 20, pp. 100374, 2020* https://www.researchgate.net/publication/342195015_Data_analytics_for_novel_coronavirus_disease [4]

[5] *D. Goodman-Meza, A. Rudas, J. N. Chiang, P. C. Adamson, J. Ebinger et al., "A machine learning algorithm to increase COVID-19 inpatient diagnostic capacity", PLOS ONE, vol. 15, no. 9, pp. e0239474, 2020.* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7508387/ [5]

[6] *P. Schwab, A. D. Schütte, B. Dietz and S. Bauer, "Clinical predictive models for COVID-19: systematic study", J Med Internet Res, vol. 22, no. 10, pp. e21439, 2020.* https://www.jmir.org/2020/10/e21439/ [6]

[7] *Y. Sun, V. Koh, K. Marimuthu, O. T. Ng, B. Young, S. Vasoo, M. Chan et al., "Epidemiological and clinical predictors of COVID-19", Clin Infect Dis, vol. 71, no. 15, pp. 786-792, Jul 2020.* https://academic.oup.com/cid/article/71/15/786/5811426 [7]

[8] *Z. Meng, M. Wang, H. Song, S. Guo, Y. Zhou, W. Li et al., "Development and utilization of an intelligent application for aiding COVID-19 diagnosis", medRxiv, 2020.* https://www.medrxiv.org/content/10.1101/2020.03.18.20035816v1 [8]