IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

USING MACHINE LEARNING CLASSIFICATION ALGORITHMS TO PREDICT COVID-19 DISEASE

[1]Dr.T.Ramaswamy, [2]Dr.S.P.V.Subba Rao, [3]Y.Pranati, [4]C.V.Sindhu Lahari, [5]S.Sanjana

[1]Associate Professor,Dept.of ECE,
[2]Professor & HoD,Dept.of ECE
[3]B.Tech(ECE),
[4]B.Tech(ECE),
[5]B.Tech(ECE)
[1]Sreenidhi Institute of Science and Technology,Telangana,India

Abstract: The project may be a demonstration of using Machine Learning Classification Algorithms to predict if an individual has the Covid-19 disease from the symptoms given. Machine Learning could be a novel technique which aids in disease prediction likewise as in diagnostics. Classification algorithms are utilized in machine learning. They use input training data to predict whether following data will fall under one among the established categories. Different classification algorithms are used here to work out their operation and therefore the accuracy that they present in accurately predicting the prognosis. In this project we are applying LR(Logistic Regression) to coach a model to predict the prognosis..

Index Terms **- Classification Algorithms, Logistic Regression, Machine Learning, Covid-19**

I. INTRODUCTION

Machine learning (ML) could be a kind of computing (AI) that permits software to enhance its accuracy at predicting outcomes without being explicitly programmed to try to do so. Machine learning differs from AI in this it's one amongst – but not the sole – steps within the process of developing a narrow AI. Machine learning algorithms can learn from their mistakes and with time, things will recover. To train, three common tactics are employed to coach machine learning algorithms. There are three sorts of machine learning: supervised learning, unsupervised learning, and reinforcement learning. One of the foremost basic forms of machine learning is supervised learning. The machine learning algorithm has been honed over time on labelled data during this case. Irrespective that exact data labelling is critical for this method to figure, when employed in the right settings, supervised learning are often quite effective. In supervised learning, the ML algorithm is given a little training dataset to figure with. This training dataset may be a subset of the larger dataset, and it serves to supply the algorithm with a rudimentary understanding of the difficulty, the answer, and also the data sets to be controlled. In terms of characteristics, the training dataset is sort of like the ultimate dataset. The ultimate dataset contains the labelled parameters that the algorithm requires to resolve the matter. The program then establishes a cause-and-effect relationship between the variables within the dataset by

II. LITERATURE SURVEY

The Classification algorithm could be a Supervised Learning technique that uses training data to see the category of fresh data. Software learning from a dataset or observations and so classifying fresh observations into one among several classes or groupings is defined because the process of classification. Binary Classifier: Used when there are only two possible outputs to a classification problem. Multi-class Classifier: Employed when a classification problem involves quite two outcomes.

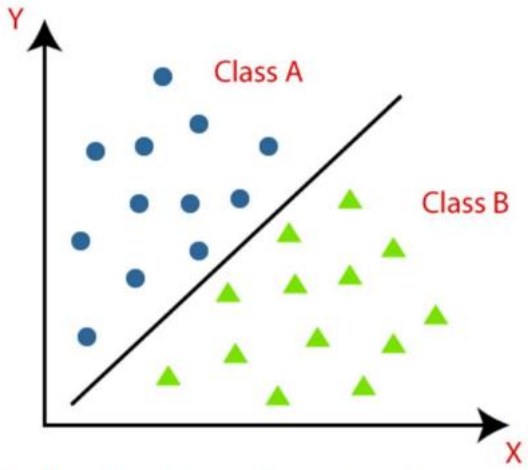

Fig 1 Working of Classification Algorithms

III. WORKING

The Logistic regression may be a statistical procedure for predicting binary classes. the end result or target variable is dichotomous in nature. The term dichotomous refers to the actual fact that there are only two potential classes. The target variable being categorical may be a special case of rectilinear regression. Logistic Regression's Properties: •In logistic regression, the variable quantity reflects the binomial distribution, and estimation is finished using maximum likelihood. • There's no R-Square, hence model fitness is set using Concordance and KS-Statistics.

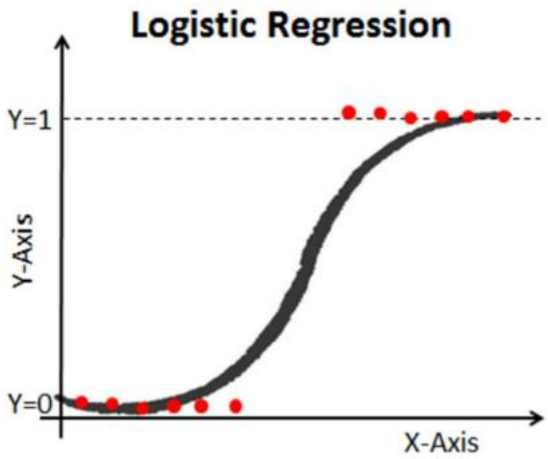

Fig 2 Logistic Regression

IV. TRAINING AND TEST ENVIRONMENT

We are training the model to predict the Covid-19 disease from the given symptoms.The aim of this project is to produce a method of early detection of Covid-19 disease from their symptoms. While 1,813,188 COVID-19 deaths were documented in 2020, World Health Organization approximates that a minimum of 3,000,000 people died as a results of the virus.If we are able to detect signs early enough, healthcare workers will have enough time to properly treat patients and save lives. We are then obtaining a concise summary of the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5861480 entries, 0 to 5861479
Data columns (total 10 columns):
 #   Column               Dtype
---  ------               -----
 0   test_date            object
 1   cough                int64
 2   fever                int64
 3   sore_throat          int64
 4   shortness_of_breath  int64
 5   head_ache            int64
 6   corona_result        object
 7   age_60_and_above     object
 8   gender               object
 9   test_indication      object
dtypes: int64(5), object(5)
memory usage: 447.2+ MB
```

Fig 3 Concise Summary of the Dataset

Our data set contains 9 features with below possible outputs.

```
Feature: cough with [0 1] Levels
Feature: fever with [0 1] Levels
Feature: sore_throat with [0 1] Levels
Feature: shortness_of_breath with [0 1] Levels
Feature: head_ache with [0 1] Levels
Feature: corona_result with ['Negative' 'Positive'] Levels
Feature: age_60_and_above with ['Yes' 'No'] Levels
Feature: gender with ['female' 'male'] Levels
Feature: test_indication with ['Other' 'Contact with confirmed' 'Abroad'] Levels
```

Fig 4 Features and possible outputs in the dataset.

In order to predict any of the 11 vector borne diseases in our data set accurately, we must make sure that our data set is balanced i.e. it's almost equal amount of knowledge on each disease. Now the following thing we are checking is that if the predictors are correlated or not. If the predictors are perfectly uncorrelated, then the effect on the response because of a predictor doesn't depend upon the opposite predictors within the model. So, we are calculating the correlation of the predictors in our model i.e., the symptoms using which we estimate the prognosis and plotting a heatmap to pictorially determine the correlation of predictors with each other.

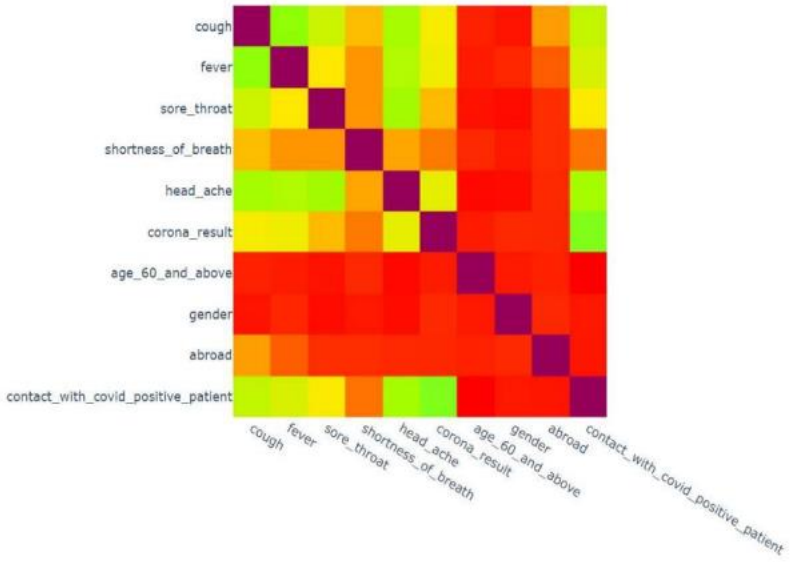

Fig 5 Heatmap for the featurews in the dataset.

our model are uncorrelated. In our dataset, the information is arranged in rows and columns. The rows comprise data of the patients demonstrating which symptoms they possess and which they don't. The columns describe the symptoms. In this covid-19 dataset, the matter to unravel is given the symptoms of patients, it's to be determined whether the prognosis is covid positive or negative. this is often a classification problem. The original dataset has 10 features(symptoms). The objective of feature selection is to pick fewer features out of those available within the original dataset and still achieve high accuracy. Chi squared test is employed to see the most effective features to unravel the matter.

	Important Variables
0	cough
1	fever
2	sore_throat
3	shortness_of_breath
4	head_ache
5	age_60_and_above
6	contact_with_covid_positive_patient
7	gender
8	abroad

Fig 6 Importance Features determined after Chi Square Test.

We can see that a number of our columns don't seem to be informative during this dataset. as an example, test date won't really help us to differentiate the disease.

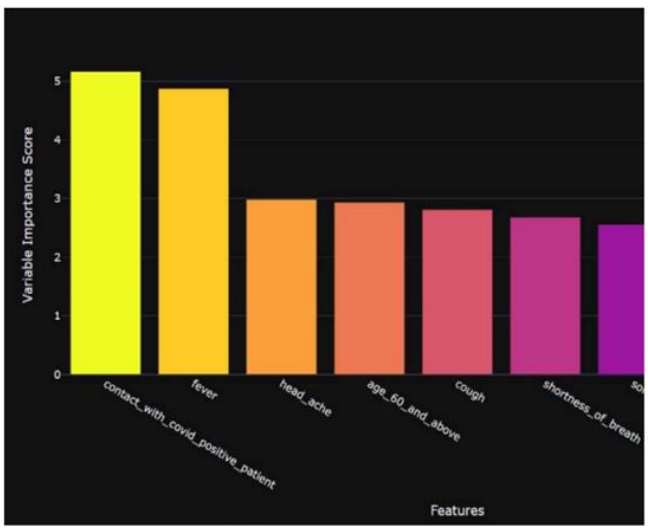

Fig 7 Variable Importance Score.

Removing uninformative features reduces the quantity of resources. With a lesser number of features : The models are easier to grasp and Model training is quicker, and therefore the model's space requirements are smaller A p-values, also referred to as a probability value, may be a number that indicates how likely it's that your data occured by random chance. The p-values for the various features(columns) in our dataset is shown below.It represents a feature's importance score.

```
cough                                1.479529e-30
fever                                5.462412e-30
sore_throat                          3.417364e-28
shortness_of_breath                  6.245624e-26
head_ache                            2.056296e-25
Age_60_and_above                     8.582283e-02
contact_with_covid_positive_patient  1.257875e-01
gender                               1.738650e-01
abroad                               2.200208e-01
Length: 64, dtype: float64
```

Fig 8 P-values of each features.

V. RESULT

The performance of a machine learning algorithm is then evaluated. It involves taking a dataset and separating it into two subgroups. The training dataset is that the first subset, which is employed to suit the model. The second subset isn't wont to train the model; instead, the dataset's input element is given to the model, which then makes predictions and compares them to the expected values. The test dataset is that the name given to the second dataset. Train Dataset: this is often the info set that's wont to fit the machine learning model. Test Dataset: This dataset is employed to assess how well a machine learning model fits. The goal is to estimate the machine learning model's performance on new data that wasn't accustomed train the model. A classification report is employed to assess the accuracy of predictions from a classification algorithm. The accuracy is estimated on the idea of what percentage of the predictions are correct and the way many are incorrect. True Positives, False Positives, True Negatives, and False Negatives are accustomed predict the metrics of a classification report. There are four techniques to work out if the forecasts are correct or incorrect: True Negative (TN): when a case was negative and was predicted to be negative. True Positive (TP): when a case was positive and was predicted to be positive. False Negative (FN): when a case was positive but predicted to be negative. False Positive (FP): when a case was negative but was predicted to be positive. F1 score - Percentage of correct positive predictions. F1 Score = 2*(Recall * Precision) / (Recall + Precision) When there are one or more independent variables, Multinomial Logistic Regression could be a classification technique which extends the logistic regression algorithm to tackle multiclass possible outcome problems. This model is employed to estimate the possibilities of a categorically variable quantity with two or more alternative outcome classes. When the dependent categorical variable has two outcome classes, the logistic regression model is utilized.

LOGISTIC REGRESSION :
Under Logistic Regression, we follow the equation –
Linear Regression Equation:

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots\ldots + \beta nXn$$

Fig 9 Equation of Logistic Regression.

Creating Where, y depends variable and x1, x2 ... and xn are explanatory variables. Where β indicates slope. As we are running the algorithm multiple times, the classifier adjusts the values of the slopes to raised suit the info. As such, these adjustments induce changes within the accuracy. For the dataset, we are training the model and generating a classification report for it. The maximum accuracy that we obtain for the training data is 97 percent.

Classification Report for Train Data

	precision	recall	f1-score	support
0	1.00	0.96	0.98	188938
1	0.94	1.00	0.97	113501
accuracy			0.98	302439
macro avg	0.97	0.98	0.98	302439
weighted avg	0.98	0.98	0.98	302439

Recall on Train Data: 0.9984
Specificity on Train Data: 0.964
Accuracy on Train Data: 0.9769
Precision on Train Data: 0.9434
F1 Score on Train Data: 0.9701

Fig 10 Classification Report of the Training data

For the column contact_with_covid_positive_patient 0 represents no contact and 1 represents there was a previous contact. For the column age_60_and_above 0 represents patient is below 60 years and 1 represents patient aged above 60. For the column gender 0 represents male and 1 represents female. For the column abroad, 0 represents patient is not returned from abroad and 1 represents that patient is returned from abroad. For the column Actual Prognosis, 0 represents that the patient is tested negative and 1 represents positive. For the column Predicted Prognosis, 0 represents that patient is predicted to be COVID negative and 1 represents that patient is predicted to be positive. In the following graph, every column (i.e., symptom) is compared to the actual prognosis which illustrates how the 2 categories (0 and 1) are affected by COVID.

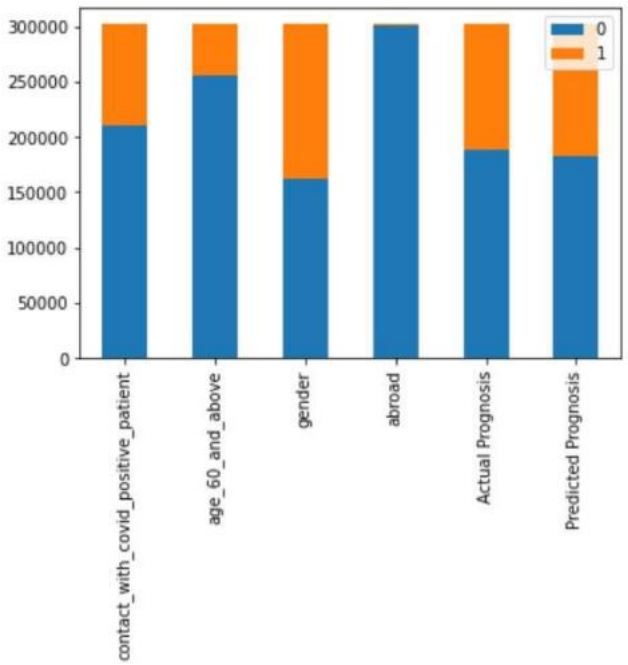

Fig 11 Graph representing inputs for different features in Training data.

For the column Actual Prognosis, 0 represents that the patient is tested negative and 1 represents positive. For the column Predicted Prognosis, 0 represents that patient is predicted to be COVID negative and 1 represents that patient is predicted to be positive. In the above graph, those 2 columns are compared to show how the actual result varies from predicted one.

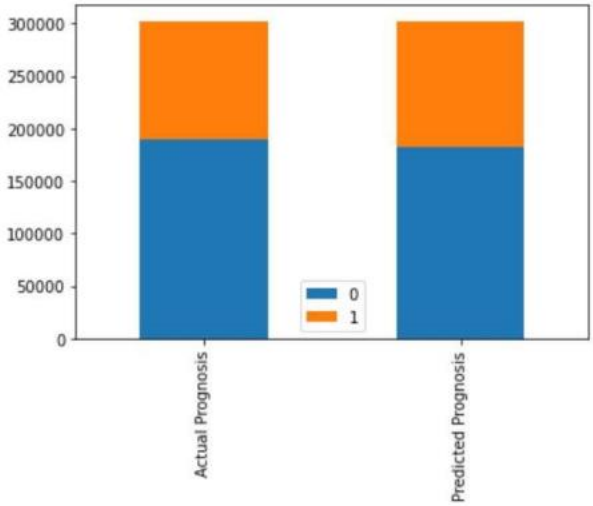

Fig 12 Graph representing ratio of values for Actual and Predicted Prognosis in Training data.

Classification Report for Test Data

	precision	recall	f1-score	support
0	1.00	0.96	0.98	81097
1	0.94	1.00	0.97	48520
accuracy			0.98	129617
macro avg	0.97	0.98	0.98	129617
weighted avg	0.98	0.98	0.98	129617

```
Recall on Test Data:  0.9987
Specificity on Test Data:  0.9644
Accuracy on Test Data:  0.9772
Precision on Test Data:  0.9437
F1 Score Test Data:  0.9704
```

Fig 13 Classification report for the Testing data.

For the column contact_with_covid_positive_patient 0 represents no contact and 1 represents there was a previous contact. For the column age_60_and_above 0 represents patient is below 60 years and 1 represents patient aged above 60. For the column gender 0 represents male and 1 represents female. For the column abroad, 0 represents patient is not returned from abroad and 1 represents that patient is returned from abroad. For the column Actual Prognisis, 0 represents that the patient is tested negative and 1 represents positive. For the column Predicted Prognosis, 0 represents that patient is predicted to be COVID negative and 1 represents that patient is predicted to be positive. In the following graph, every column(i.e symptom) is compared to the actual prognosis which illustrates how the 2 categories(0 and 1) are affected by COVID.

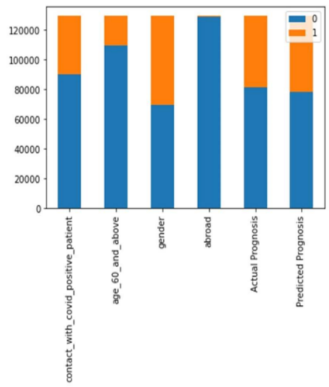

Fig 14 Graph representing inputs for different features in Test data.

For the column Actual Prognosis, 0 represents that the patient is tested negative and 1 represents positive. For the column Predicted Prognosis, 0 represents that patient is predicted to be COVID negative and 1 represents that patient is predicted to be positive. In the above graph, those 2 columns are compared to show how the actual result varies from predicted one.

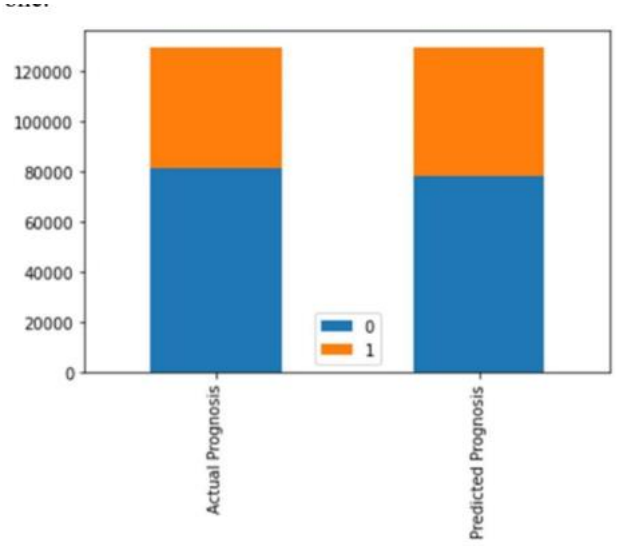

Fig 15 Graph representing ratio of values for Actual and Predicted Prognosis in Test data.

VI. Future Scope

When it involves healthcare, machine learning techniques is employed in a range of how to boost disease prediction, diagnosis, and treatment while also enhancing overall healthcare operations. Implementing machine learning effectively allows healthcare workers to form better decisions, discover trends and breakthroughs, and increase the efficiency of research and clinical trials. Using historical and real-time data, machine learning allows to make models that swiftly assess data and provides outcomes. Healthcare providers can use machine learning to form better decisions about patient diagnoses and treatment alternatives, leading to an overall improvement in healthcare services

VII. Conclusion

In this paper we've generated models using Logistic Regression to predict whether an individual has Covid-19 disease given the symptoms. The scope of Classification Algorithms is proscribed to labelled data but it's very powerful.

References

[1] Chan EYS, Cheng D, Martin J (2021) Impact of COVID-19 on excess mortality, life expectancy, and years of life lost in the United States. PLoS ONE 16(9): e0256835.https://doi.org/10.1371/journal. pone.0256835

[2] Barbosa TP, Costa FBP, Ramos ACV, Berra TZ, Arroyo LH, Alves YM, et al. Morbimortalidade por COVID-19 associada a condições crônicas, serviços de saúde e iniquidades: evidências de sindemia. Rev Panam Salud Publica. 2022;46:e6. https://doi.org/10.26633/RPSP.2022.6

[3] F. Sebastiani. Machine Learning in Automated Text Categorization ACM computing surveys (CSUR), 2002vol. 34, issue 1, pp. 1–47

[4] Peng, Joanne & Lee, Kuk & Ingersoll, Gary. (2002). An Introduction to Logistic Regression Analysis and Reporting. Journal of Educational Research - J EDUC RES. 96. 3-14. 10.1080/00220670209598786.