IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 8, Issue.10, October 2020 E-ISSN: 2347-2693

Analysing COVID- 19 Cases by Eliminating False Negatives and False Positives through Machine Learning Approach

M. Vennela[1*], G. Lavanya Devi[2], P.R.S. Naidu[3]

[1,2,3]Dept. of CSSE, Andhra University College of Engineering, Andhra Pradesh, India

[*]*Corresponding Author: vennelaminnekanti31@gmail.com, Tel.: 8790687455*

DOI: https://doi.org/10.26438/ijcse/v8i10.7174 | Available online at: www.ijcseonline.org

Received: 09/Oct/2020, Accepted: 20/Oct/2020, Published: 31/Oct/2020

Abstract: Novel coronavirus (COVID-19 or 2019-nCoV) pandemic which doesn't have neither a clinically proven vaccine nor drugs. As the No. of Cases Increasing Day by Day, the public was panicking. In the process of increasing Number of Tests, Some Rapid tests are also Taking place. If you take these rapid tests into consideration, where we are getting false results like False Positives, True Positives which results in a panic among the people who have tested. Due to these false results, the public was in a panic situation. To avoid that panic among the public, we define machine learning approach to predict the COVID where it represents False Positive rate, and True Positive rate through ROC(Receiver Operating Characteristic) curve and, we also get a Confusion Matrix which visually represents True Negatives, False Positives, False Negatives and True positives and it also generates a new dataset from a given dataset, by eliminating them without sampling using clinical spectrum data of 'SARS-Cov-2 exam result'.

Keywords: Logistic Regression, ROC (Receiver Operating Characteristic) Curves, Confusion Matrix.

I. INTRODUCTION

COVID-19 or nCOVID-19 is a pandemic disease which is spreading very quickly in the world from one person to another person. It was first traced out in the city named Wuhan in China in December 2019. As it was found in the form pneumonia. Fever, Cough, Tiredness, breathing problems are COVID-19s most common symptoms. Till now it doesn't have any Vaccines or Drugs, still, people are recovering by increasing immunity power. As COVID-19 testing becomes vastly available around the globe, it is also crucial for the health care providers and public health officials to understand its limits and the impact of false results that can have efforts to retard the pandemic [1]. As the no. of cases are increasing in the world daily people are alarming with the outspread of disease. So, for tracing out disease quickly, the number of testing samples were increased to trace quickly to break the chain. In the process of increasing tests, some Rapid testings were also increased. As the rapid tests are doing there some erroneous data was reported like False positives, False negatives where it may panic the persons who get tested. These false reports make people more panic as there are only limited facilities for quarantining persons in many countries across the globe. These increasing Diagnostic Rapid tests can be erroneous in two ways. A person infected incorrectly flags false positives, which is tricky for him which includes unnecessary quarantine and contact tracing. False-negative results are more problematic, since the infected persons who might be asymptomatic may not be in isolation, then he is a person who can be the reason for transmitting the disease and infect others [2]. By using Machine Learning approach, we can predict COVID may give some accurate results as today many people accessing online data and predicting the outspread of coronavirus. We know how machine learning is growing in today's technical world because of its algorithms and its predictions. We can achieve it by using python.

Machine Learning: Machine Learning means the Ability of a machine to learn itself without being explicitly programmed. Machine learning is a branch of artificial intelligence [3].It is one of the rousing technologies that one would have ever come across. It gives the computer to make machines to learn and perform the tasks that are more similar to humans: The ability to learn. It is deliberately being used today in many more places or areas of work than one would expect. There many algorithms like logistic regression were there to predict the Data.

Python: Python is a Dynamic Type language. It is simple, robust and can be code within less no. of lines. Due to its libraries like NumPy, pandas, scilearn which were many data science libraries makes it easy to code and build Machine Learning models[3].

In Present research, we propose an approach to eliminate the errors such as False Positives, True Negatives, False Negatives, True Positives by implementing Logistic regression classifier coupled with ROC curves and Confusion Matrix.

Rest of the paper is organised as follows, Section II contain the models and methods of the proposed work, Section III contain the related work which is useful for Covid-19 cases, Section IV contain the existing system, Section V contain the Proposed system and results of the work which is obtained after filtering the dataset and the

last section VI contain the conclusion and future work of the project.

II. MODELS AND METHODS

Logistic Regression: It is used to predict the probability of certain diseases based on predicting attributes. It will give logarithmic values and also a probability.

Logistic Model: it is mainly used in the study of analysis and distribution, also it is used to calculate the risk factors of a certain disease, and also used to predict the probability of occurrence of a certain disease based on its risk factors and certain attributes. We can predict the development and transmission law of epidemiology through logistic regression analysis very roughly [4].

$$Qt = \frac{a}{1 + e^{b-c(t-to)}} [1]$$

ROC Curves: ROC curves Means Receiver operating Characteristic curves is a graphical plot which defines the symptomatic ability of a binary classifier system as its discriminated threshold value is diversified. It was developed for the operators of military radar receivers, that is why it is named like that, where the curve was plotted against True positive (Y-axis) and false positive (X-axis) values.

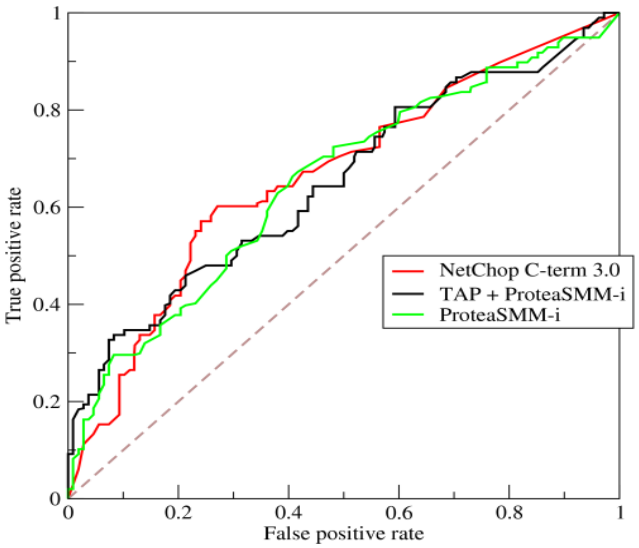

Fig.2.1 Example for ROC Curve

Confusion Matrix: It is a table which describes the performance of a classification model also known as a classifier on a set of test data for which the true values are known. We use confusion matrices to Define the True Positives, False Positives, True negatives, False Negatives and to Know whether the result is correct or not. It will compare Actual value and Predicted value to know it is right or not.

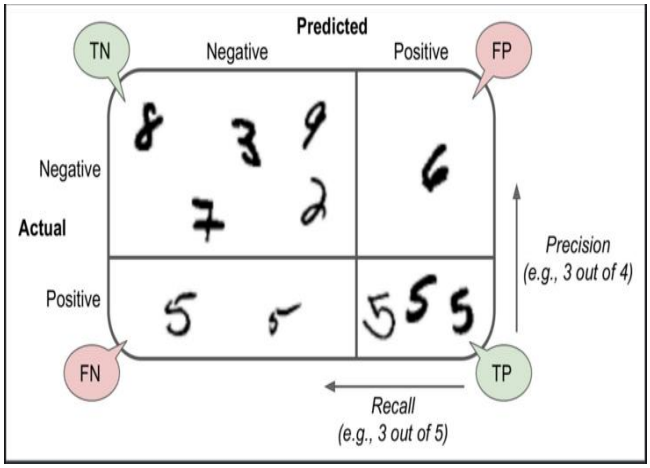

Fig.2.2 Example of the Confusion matrix

Dataset: Here we use a dataset named SARS-Cov-2 exam result which contains 5645 fields and 111 attributes which has some erroneous Data here we are removing that erroneous data and generating new data set.

Pre-processing: here we clean data by Check patient id and remove duplicate rows, we are placing empty fields with NAN, making positive as 1 and negative as 0.

III. RELATED WORK

On April 9, 2020, the article named False-negative COVID-19 test results may lead to a false sense of security published in Science Daily by mayo clinic by jay furst says that as the corona tests become more widely spread around the globe [1], there may be a risk of getting some erroneous values like false-negatives which means he has a disease but gets he has nothing in the test he may spread the disease in our project we are avoiding it by removing false negative values.

On June 5, 2020, a journal was published in The New England Journal of medicine by Steven Woloshin, M.D., Neeraj Patel, B.A., and Aaron S. Kesselheim, M.D., J.D., M.P.H. says that while many countries lifting lockdown slowly, and also corona teste were increasing rapidly they may inappropriate in two ways like False-positive which says a person is tested positive or labelled as positive even he doesn't have the disease if was taken into quarantine he may affect by the persons in quarantine, similarly, if the person tested negative[2] even he has affected which is false negative value then he shouldn't be required to be in quarantine as he Stested negative then he will be the person who spread the disease in his area.

The paper which was by Lin Jia1 Kewen Li , Yu Jiang Xin Guo1 Ting Zhao in this paper they are predicting the COVID19 Cases based on different kinds of mathematical models: logistic model, Bertalanffy model, Gompertz model in this they predict epidemic predictions of COVID data [4].

The paper which was by Sanjib halder, in this paper they are predicting the progress rate of the virus infection in males and females based on the overall population in the area by regression analysis techniques[5].

The paper which was by Saroj S. Date, in this paper they are predicting the increasing cases for a particular set of time in India by taking the number of cases registered till particular date by applying time series algorithm[6].

IV. EXISTING SYSTEM

In Existing System logistic regression used to classify the data. It will predict the data continuously. During the prediction of Data, there may be a chance to get the wrong predictions when we consider the rapid testing. Due to these wrong predictions, people may be alarmed considering the spreading of Coronavirus widely.

Disadvantages:

1. ***It may give wrong predictions.***
2. ***It will Give false predictions like false positives, false negatives also.***

V. PROPOSED SYSTEM

In the proposed system, we will rectify the errors of Existing System and we will Graphically Represent False Positives, True Negatives, False Negatives, True Positives by using ROC (Receiver Operating Characteristic) curves and we will eliminate these values and we will Generate a new Dataset from the Existing dataset and form confusion matrix. Here we also modify data by Checking patient id and remove duplicate rows, bringing Patient age quantile (0-19), replace 0 for negative and 1 for positive, modifying the name by checking whether Patient admitted to a regular ward, modifying the name if the patient is admitted in semi ICU, at hematocrit attribute making blanks as null after all modifications done on all attributes we are plotting ROC curve on True positive rate against false positive rate.

Results Obtained:

ROC Curve of this project is:

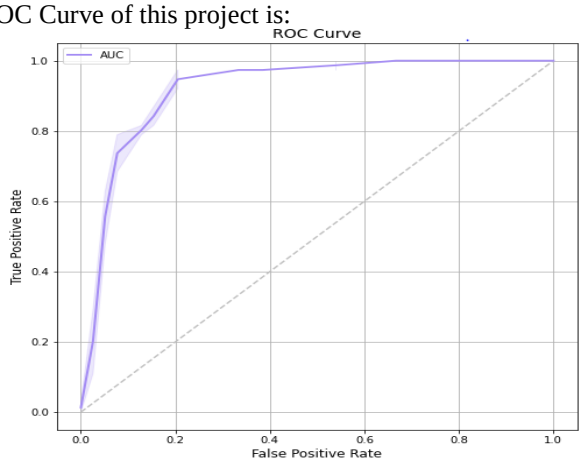

Fig 5.1 ROC curve for SARS- COVID19 Dataset

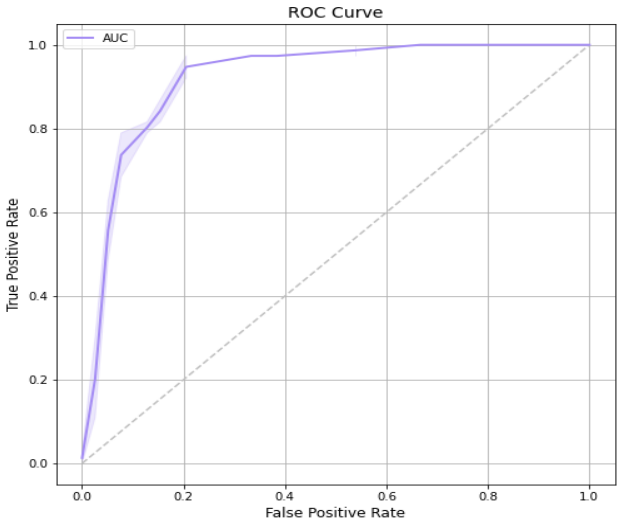

Fig 5.2 ROC curve for SARS-COVID19 Dataset

5.1 Confusion Matrix: It describes whether the data is correct or not based on matrix format. It forms between predicted values and actual values. Each row will represent the actual class, and the column will represent the predicted class.

5.2 The terminology of Confusion matrix:

True positive: This is also known as sensitivity, recall or probability of detection. It is the prediction that is correct.

True Negative: True Negative means how correctly it detected negative case or false values.

False Positive: even if there is a wrong prediction it shows that prediction is as correct. It called as type Error

False Negative: opposite to false-positive it shows even correct prediction as wrong. It is also called a type-II Error.

Accuracy: Overall, how often it predicts whether the classifier is correct .

Accuracy = TP+TN/(TP+TN+FN+FP) (1)

Precision: How often it predicts correctly.
Precision = TP/(TP+FP) (2)

F1 Score: Harmonic mean of precision and sensitivity
F1 Score = 2TP/(2TP+FP+FN) (3)

Recall: it is also called a true positive rate which defines how well it predicted
Recall= TP/(TP+FN) (4)

Confusion Matrix:

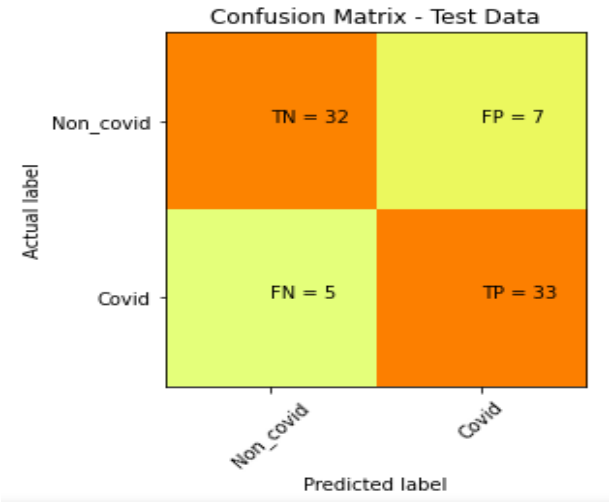

Fig 5.2: Confusion Matrix of Data

VI. CONCLUSION AND FUTURE WORK

In this paper, by using logistic classifiers we will pre-process the Data and we will eliminate the False Positives, False negatives, True positives, True Negatives and represent them using ROC(Receiver Operating Characteristic) curves and we form a confusion matrix based on the obtained values after making some cleaning and filtering to the dataset. After removing False Positives and False Negative values it generates a new data set from the Given Dataset.

In future work, we can extend this project by pre-processing the data and we can perform EDA (Exploratory Data Analysis) and we will predict whether he is positive or negative based on various factors.

1. ***We will pre-process the Data before proceeding for the next step.***
2. ***After pre-processing the Data, we will perform EDA (Exploratory Data Analysis) on the Data.***

REFERENCES

[1] Jay Furst-"False-negative COVID-19 test results may lead to a false sense of security". Source: mayo clinic

[2] Steven Woloshin, M.D., Neeraj Patel, B.A., and Aaron S. Kesselheim, M.D., J.D., M.P.H.-"False Negative Tests for SARS-CoV-2 Infection — Challenges and Implications, Journal published on June 5, **2020**, in The New England Journal of medicine".

[3] John Wiley& sons "Machine Learning: Hands-on for Developers and Technical Professionals"

[4] Lin Jia Kewen Li, Yu Jiang Xin Guo Ting zhao, "Prediction and analysis of Coronavirus Disease", Populations and Evolution, **2020.**

[5] Sanjib Halder "A Mathematical Model to Forecast & Compare Covid-19 Outbreak in Male & Female using Polynomial Regression Analysis"-IJCSE ,**vol.8,issued on 5,May 2020.**

[6] Saroj S. Date "Forecasting novel Covid-19 confirmed cases in India using Machine Learning Methods" –IJCSE, **vol.8, issued on 6,June 2020.**

AUTHORS PROFILE

Ms. Minnakanti Vennela currently pursuing Master Degree in Computer Networks in Andhra University College of Engineering(A), Andhra Pradesh, India. She completed M.Tech in Computer Science Engineering at Andhra University. Her Research Interest include Machine learning, object-oriented programming, computer graphics.

Dr.G.Lavanya Devi currently working as Assistant Professor in Andhra University College of Engineering(A), Andhra Pradesh, India. Her research area includes Fuzzy Theory, Data Mining, Bioinformatics, Object Oriented Analysis and Design, Database Systems, Cryptography and Network Security, Algorithm Analysis, Automata Theory, Computer Communication and Networks.

P.R.S.Naidu currently working as Assistant Professor in MVGR College of Engineering(A), Andhra Pradesh, India. His research area includes Machine Learning, Cryptography and Network Security, Linux Network Administration

