COVID-19 PREDICTION BASED ON SYMPTOMS

KESAV NAGENDRA
Electronics and communication
Vellore institute of Technology
kesavnagendra3@gmail.com

GAGAN KARTHIK K
Electronics and communication
Vellore institute of Technology
gagankomma@gmail.com

KATTA JAYA KRISHNA
Electronics and communication
Vellore institute of Technology
jayk7179@gmail.com

U SAI PRANEETH
Electronics and communication
Vellore institute of Technology
upraneeth895@gmail.com

RAVI YASWANTH SAI
Electronics and communication
Vellore institute of Technology
yaswantsai.ravi34@gmail.com

SATYA SAI SUJAN N
Electronics and communication
Vellore institute of Technology
satyasaisujan777@gmail.com

Abstract- Coronavirus disease is an infectious disease caused by the SARS CoV-2 virus. Most people infected with the virus will experience mild to moderate respiratory illness and recover without requiring special treatment. However, some will become seriously ill and require medical attention. COVID-19 testing began to rise all over the world during this pandemic and there is reduce of proper medical resources all over the world. In this project we make a machine learning model which will predict the possibility of having COVID-19. We are trying to use the prediction model and we use the symptoms which will cause covid-19 to predict weather you have covid or not using the training data set.

Keywords- COVID-19, Algorithm, Predictive Analysis, Gradient, Efficiency.

I. INTRODUCTION

The covid-19 outbreak has brought the health crisis all over the world. There are many people who fear COVID for small symptoms and others who neglect the to test them due to negligence. The limited facilities also hinder in testing a huge amount of people at a time. Through this project we create a simple prediction system for covid which will reduce the pressure on medical facilities and to reduce the anxiety of the patient . For this project we have collected the data of the covid patients and their symptoms from Kaggle which include data of 98000+ patients. We are going to build various machine learning models. The data is prepared for training of different machine learning models and the efficiencies for each Machine Learning is calculated.

II. TOOLS

A. *SOFTWARE:*

- PYTHON
- JUPYTER NOTEBOOK
- TENSORFLOW & SKICIT LIBRARY
- IBM WATSON STUDIO
- IBM MACHINE LEARNING LIBRARY

B. *OTHERS:*

- COVID-19 PATIENT DATA
- HEALTH RECORDS

III. METHODOLOGY

We have collected the data of the covid patients and their symptoms from Kaggle which include data of around 1 lac patients. We are going to build various machine learning models and prepare the data for training of different machine learning models and we will check the efficiencies for each Machine Learning model and select the best model.

We will be taking Data from Outside which includes all the parameters required for classification or determining the output. In this project we will be collecting the patient data which includes various parameters i.e. Symptoms; and by using this methodology, we will be able to determine whether the person is affected by Covid-19. Our main objective of this project is Predicting so we will be using Predictive Analysis.

A. *MOTIVATION OF THE PROJECT:*

The novel coronavirus (COVID-19) outbreak has produced devastating effects on the global economy and the health of entire communities. The huge population in India is the limitation to the number of tests that can be done with the facilities we posses.

There are many people who fear small symptoms as covid and others who neglect their's as normal. This unnecessary fear or unaware negligence is leading to mental tension or further spread of covid.

Our Main motivation is to build a classifier prediction model to predict the status of coronavirus CovID-19 which will reduce the medical pressure and replaces the lack of medical resources.

Another motivation is to make a project where people can find the possibility of covid by giving their current symptoms which can help in reducing the unnecessary fear and further spreading.

B. *ADVANTAGES OF EXISTING SOLUTIONS:*

- Prediction models for covid-19 are very useful for academic literature to support medical decision making at a time when they are urgently needed.
- collaborative efforts and data sharing to also allow an investigation of the stability and heterogeneity in their performance across populations and settings.
- These solutions allows us to estimate the pandemic's behavior within an acceptable degree of uncertainty.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

- With this information, we can calculate the demand for acute medical services; determine the timeframes for partially or completely lifting containment measures.
- Estimating health care demand allows planning for required health technologies. (PPE, ventilators, etc.)

C. LIMITATIONS FOR EXISTING SOLUTIONS:

- There is an inherent uncertainty to predictive modelling that can be introduced at any step of the model-building process.
- Units of analysis, such as population or patient subgroups, can have a broad degree of heterogeneity, which requires different sets of assumptions to be applied for each group.
- unreliable predictions could cause more harm than benefit in guiding clinical decision.

D. PREDICTIVE ANALYSIS:

Predictive analytics is the form in which analytics is used in making predictions about future outcomes via analyzing and checking previous data.

Predictive analytics includes a combination of scientific methods and scientific techniques as discussed below:

Data Mining: In order to manage large amounts of data sets either structured or unstructured to recognize hidden patterns and relationships among variables provided, data mining is aimed to.

Statistical Modelling: statistical data models can be developed depending on the context of what needs to be anticipated using the same collected data as for data mining. Once the model is built, the new data is fed to models to predict future outcomes.

Machine learning: ML can deploy iterative methods and techniques to identify patterns from large data sets and build models. For example, recommendation engines are widely used for online shopping recommendations as predictions are made from using customers' prior purchasing and browsing behavior.

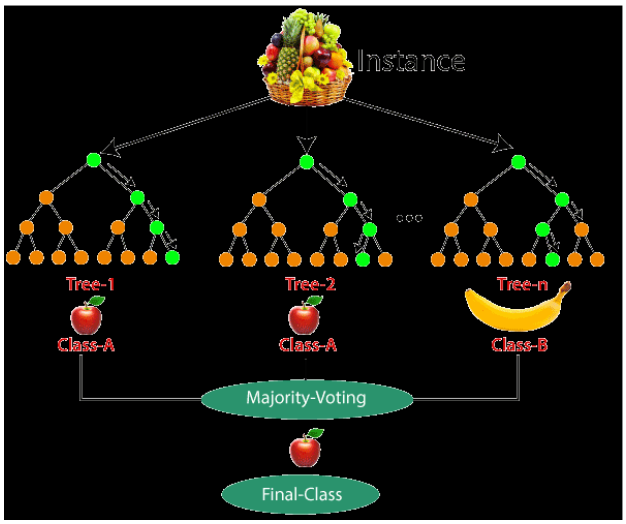
We will be using Machine learning as our main scientific method for predicting the output.

IV. ALGORITHM

Mainly our project is focussed on prediction of Covid-19 based on input so we will be creating different classification model since the output is binary (i.e Yes/No).

So for our project we have used 4 different classification algorithms to check and predict the best model for our project.

A. RANDOM FOREST CLASSIFIER



Random forest is a Machine Learning Algorithm that comes under supervised category and it is widely used in taking decisions or the problems which have specific output.It is commonly used as classification or regression model. This algorithm builds n-number of trees and gets output in each tree. In case of classification algorithm it takes the majority output and in the case of regression it takes the average of every output.

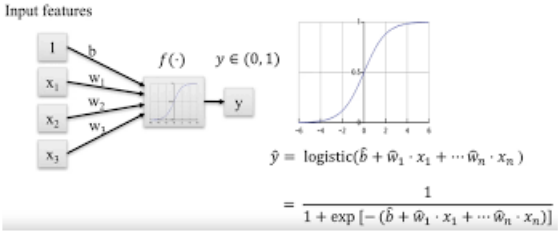
a. PROS:

- It is used as both classification algorithm and regression algorithm.
- It is simple algorithm which uses averaging and majority voting rather than complex equation.
- It is very stable algorithm.
- It does well in when the data contains null/complex values.

b. CONS:

- Training time is more compared to other models due to its complexity. Whenever it has to make a prediction each decision tree has to generate output for the given input data.
- It is very complex since it has to do same algorithm for n times.

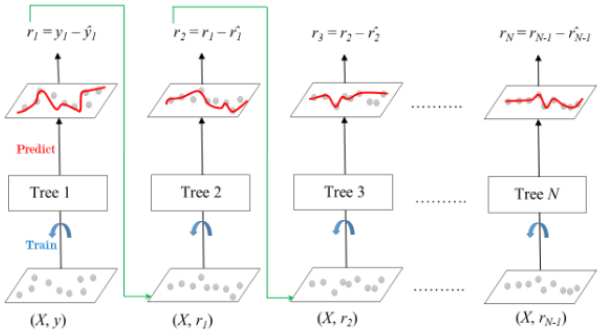
B. Logistic Regression:



IJSER

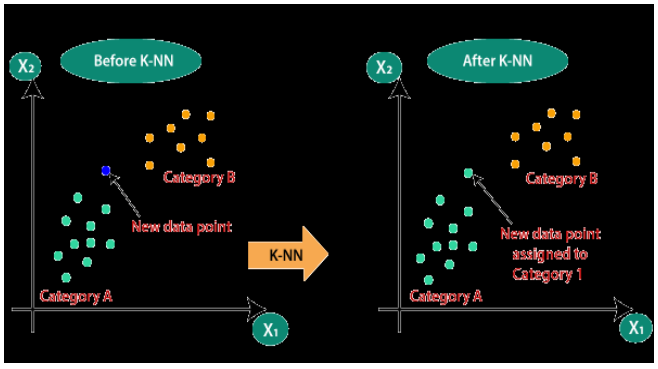
Logistic regression predicts the output of a dependent variable. The variable will be substituted in sigmoid function or in any function such that the output will be 1 or 0 (i.e Yes or No) if the value of sigmoid function is greater than threshold value then the value will be 1 or else it will be zero. Therefore the outcome will be a categorical and discrete value. It can be either Yes or No, 0 or 1, true or False, etc.

C. Gradient booster classification:



Gradient Boosting is very famous boosting algorithm. In this gradient boosting algorithm, each predictor corrects the value of its predecessor's error. In contrast the weights of the training instances are not changed, instead of that each predictor is trained using the residual errors of predecessor as labels. The equation can be written as
y(pred) = y1 + (eta * r1) + (eta * r2) + ... + (eta * rN)

D. K-Nearest Neighbours(KNN):

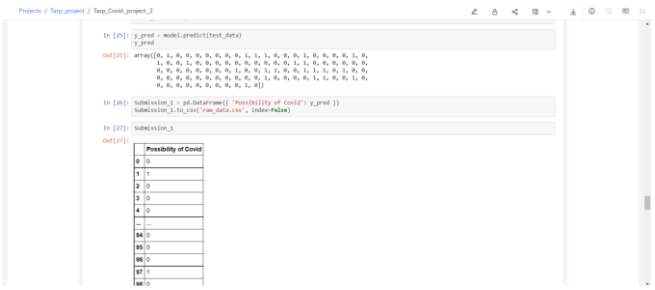


K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

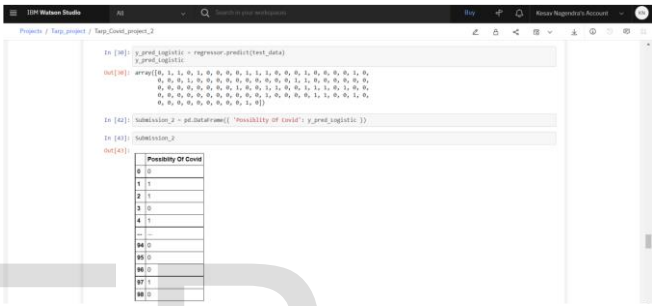
K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

V. RESULTS

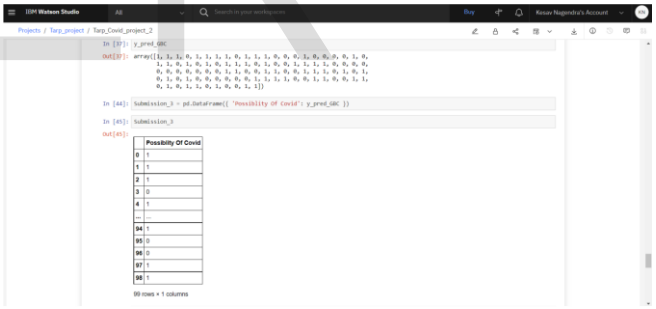
Random Forest:



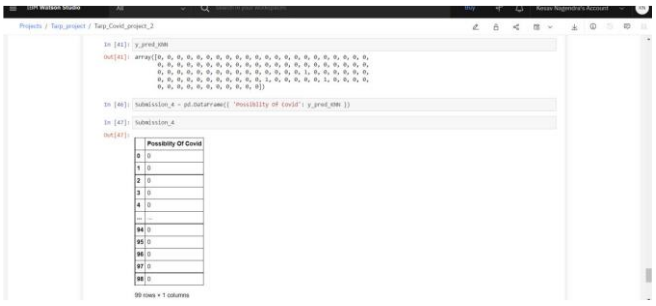
Logistic Regression:



Gradient Booster Classification:



K-nearest Neighbours(KNN):



IJSER

MODEL	EFFICIENCY	SQUARE MEAN ERROR
RANDOM FOREST	82.82	0.17171717
LOGISTIC REGRESSION	78.78	0.232323232
GRADIENT BOOSTING	64.64	0.454545454
K-NEAREST NEIGHBOURS(KNN)	87.87	0.1313131313

VI. CONCLUSION

We have collected the health records of patients through internet and filtered the data and downloaded some data through Kaggle and prepared our own data set.

We have trained our data set with 4 different classification algorithms and we have found out that KNN is indeed the best Algorithm for prediction of Covid among our results.

KNN algorithm is simple and it will be very useful in predicting the simple tasks such as weather prediction and health related predictions.

Rather than KNN all other algorithms are very complex and it will be taking lot of time and effort and we can not set the weights as weights will be changing for every small change in data so it will be very difficult to control the prediction outputs.

IJSER

VII. ACKNOWLDGEMENT

We would like to express our special thanks to PROF. CHRISTINA JOSEPHINE MALATHI A for providing us with the opportunity to present our project and report and as well as for guiding us on our every stride. We would also like to thank VIT, Vellore for providing us with this platform.

VIII. REFERENCES

[1] Zoabi, Yazeed, Shira Deri-Rozov, and Noam Shomron. "Machine learning-based prediction of COVID-19 diagnosis based on symptoms." npj digital medicine 4.1 (2021): 1-5.

[2] Kushwaha, Shashi, et al. "Significant applications of machine learning for COVID-19 pandemic." Journal of Industrial Integration and Management 5.04 (2020): 453-479.

[3] Muhammad, L. J., et al. "Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset." SN computerscience2.1(2021):1-13.

[4] Struyf, Thomas, et al. "Signs and symptoms to determine if a patient presenting in primary care or hospital outpatient settings has COVID-19." *Cochrane Database of Systematic Reviews* 2 (2021).

[5] L. W. Mary and S. A. A. Raj, "Machine Learning Algorithms for Predicting SARS-CoV-2 (COVID-19) – A Comparative Analysis," *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*, 2021.

[6] V. Bhadana, A. S. Jalal and P. Pathak, "A Comparative Study of Machine Learning Models for COVID-19 prediction in India," 2020 IEEE 4th Conference on Information & Communication Technology (CICT), 2020.

[7] Ibrahim, Ibrahim, and Adnan Abdulazeez. "The role of machine learning algorithms for diagnosing diseases." *Journal of Applied Science and Technology Trends* 2.01 (2021): 10-19.

[8] Callejon-Leblic, María A., et al. "Loss of smell and taste can accurately predict COVID-19 infection: a machine-learning approach." *Journal of Clinical Medicine* 10.4 (2021): 570.

[9] P. Podder and M. R. H. Mondal, "Machine Learning to Predict COVID-19 and ICU Requirement," 2020 11th International Conference on Electrical and Computer Engineering (ICECE), 2020.