Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 7803 - 7809
Received 25 April 2021; Accepted 08 May 2021.

Machine Learning-Based Probabilistic Evaluation of Covid-19

Dr.M.Purushotham Reddy[1], P.Sirisha Reddy[2], B.Spandana Singh[3], Shalom Raj[4]

[1]Department of Information Technology, Institute of Aeronautical Engineering, JNTU, Hyderabad, India.
[2]Department of Information Technology, Institute of Aeronautical Engineering, JNTU, Hyderabad, India.
[3]Department of Information Technology, Institute of Aeronautical Engineering, JNTU, Hyderabad, India.
[4]Department of Information Technology, Institute of Aeronautical Engineering, JNTU, Hyderabad, India.

ABSTRACT

The COVID 19 pandemic has affected the world very badly as the number of cases is not declining any day. The growth and the mutation of the virus in a different form have created a lot of concern among the researchers working on it. This project enables the prioritization of tests by taking the symptoms from a person and feeding the input data to it which in turn gives an output of estimation of how much the person is infected with COVID 19.

Keywords
Covid-19,Coronavirusdisease,MachineLearning, Estimation, Health informatics, Naïve Bayes, Logisticregression,Classifiers, Tests.

I. Introduction

Coronavirus is indeed a viral infectious disease with Wuhanas the initial epicenter. As of November 8, 2020, the totalnumber of confirmed cases worldwide was 50,395,239, with1,258,235 deaths and 35,629,514 recoveries. Coronavirus 2,alsoknownasSARS coronavirus 2, is a virus that wasresponsible for the COVID-19 pandemic. When an infectedindividualisclosetoanother, the virus spreads quickly.Whenanindividualisinfectedtothepointthattheycan'tsee a simple vision, the concentrations of more than a fewrespiratorsdrop.Suchsalivawhichgetsinhaledintoabalanced person's lungs through the nasal passage may alsoformsuchdroplets,transmittingthe disease directly fromthesource.

Thosethatcomeintodirectcontactwithcontaminatedproductsand then expose their eyes, nose, or mouth aremorelikely to contract the infection quickly. The virus'slongevity was found to be greater on the surface of steel orplastic materials, lasting up to 72 hours, even though thevirus'slifetimeisshortoncopperandcardboard.Thedisease's basic symptoms included mild signs of a commoncold, such as pain, a blocked nose, a scratchy throat, andotherdiscomforts.

Coronavirus infection is a major concern for senior citizenswith high blood pressure, heart arrest, and diabetic patients.Tostopinhalingthedroplets,theWHOsuggestsmaintaining a social distance of around 3 feet and washingtheir hands for 20 seconds with alcohol-based hand rubs andwater to help kill the virus early. This research will benefitgreatlyfromaggregateddataonthesymptoms of otherinfectedpatientstoimprovethediagnosisofcoronavirus-affectedpeople.

COVID-19isdiagnosedmainlybased on the symptomsmentionedabove.Furthermore,weuseseveralmachinelearningalgorithmstodeterminetheprobabilityofcontracting the disease. For such analysis, we developed amodelthatusesarandomforestclassification model toassesstheriskofcontractingCOVID-19infectionbasedonaperson'ssymptoms.

II. Motivation

Coronaviruswasfirstidentifiedduringthefirstfewmonths of this year. Even though months have passed, novaccine or cure for this deadly virus has been created. If wecan even recognize the disease ahead of time during thesedifficulttimes,itmightbepossibletocontactmedicalauthoritiesintimetodiagnoseandreceivetreatment.

III. Methodology

Following the loading of the COVID-19 symptom dataset and the division of the attributes into independent and relative variables, the dataset is split into two parts: one for the train data and the other for the test data that we can determine. The first is the majority; we put it to the test using various machine learning algorithms to get the best possible result. The Random Forest Classification algorithm, we discovered, offers the highest degree of precision and has high efficiency as compared to other algorithms. The objective element is the target value,

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 7803 - 7809
Received 25 April 2021; Accepted 08 May 2021.

which receives some of the other symptoms' attributes as well as the individual's requisite information and displays the desired result. The perplexity is depicted in Figure 1.

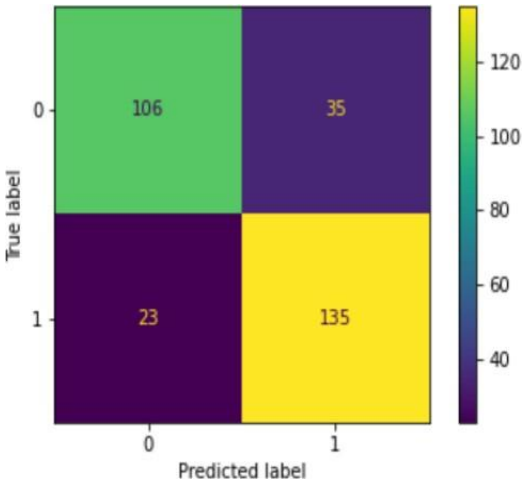


Fig.[1]Randomforestclassifiersconfusionmatrix

Following model training, the most recent dataset is often used to forecast the presence of an outbreak over a specific individual's lifetime. The Covid consequences are the final result of the analysis that we must decide. We propose the design that is being used to measure the sufferer's final coronavirus outcome using a variety of algorithms. The test dataset is tested against the trained dataset in the final stage, and the degree of precision is determined. The entire process is shown in Figure 2 below

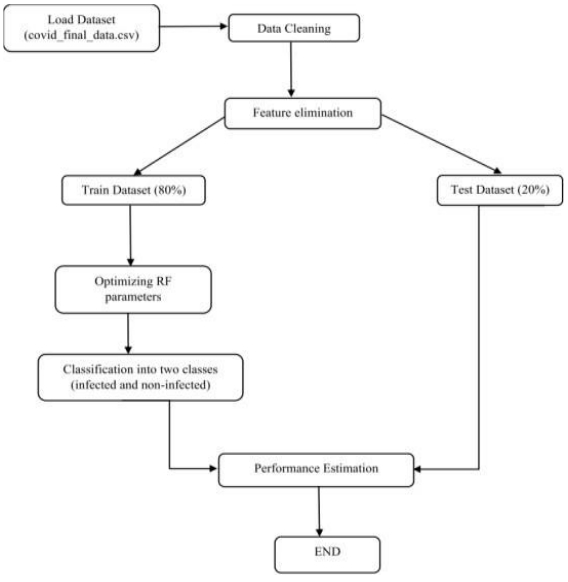
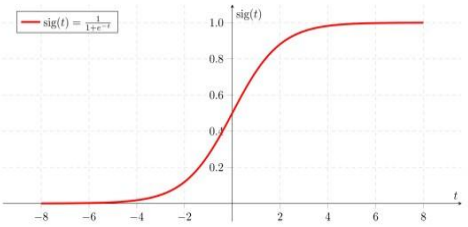


Fig.[2]Thealgorithmisrepresentedasablock diagram

A.LogisticRegression

This method of regression is a famous method afterLinearRegression.Thoughbothofthemareregressionalgorithms the underlying difference is that the first is usedfor classification tasks, while the second is used to predict orforecast data given to the model. One such algorithm alsopredicts the value using a linear equation, but the outputrange is between negative and positive infinity. To have aclassification output we will be having binary data for which1 means Yes and 0 means No which sums it up to have therange 0 to 1. For squeezing the output data to be in thisrange,thesigmoidfunctionisused in which the belowfigure3showstherepresentationoftheranges

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 7803 - 7809
Received 25 April 2021; Accepted 08 May 2021.



Below is the graphical representation of Logistic Regression input and output scenario.

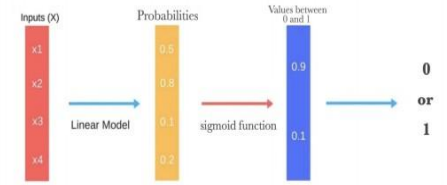


Fig. [3]Logistic Regression Input and Output scenario.

B. Decision Tree Classifier

It is a visual representation of how to distinguish examples. Since the data is continuously divided by a parameter, it's called Supervised Machine Learning.

- A Decision Tree's Nodes are used to evaluate or verify the outcome of a particular property.
- Edges/Branch: Based on the results of a test, connect to the next node or leaf.
- Terminal nodes that project the result are known as leaf nodes.

Thefollowingisagraphicalrepresentationoffigure4:-

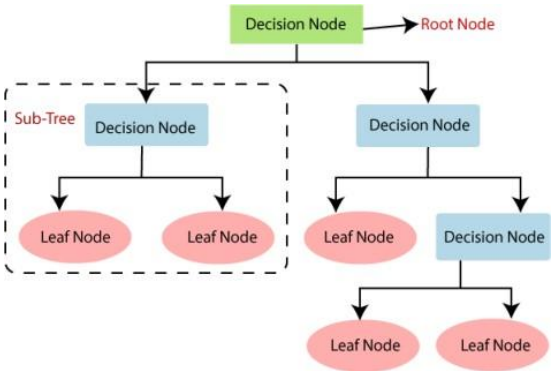


Fig. [4] shows a graphical representation of the decision tree classifiers for a given dataset.

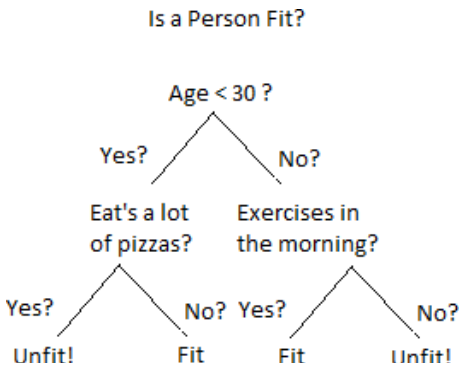
There are two kinds of Decision Trees:

- Classification Trees
- Trees of Regression

Trees of classification (Yes/No types):

A classification tree is shown in the image below, with the result being a variable such as "suitable" or "unsuitable. "The decision variable may be categorical or discrete in this case.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 7803 - 7809
Received 25 April 2021; Accepted 08 May 2021.



To make such a tree, a method known as binary recursive separation is used. It's an analysis procedure that is defined as ways into partitions, which are then subdivided more on each branch.

Trees of Regression (Continuous data types):
Regression trees are decision trees that target a variable that can take repeated values. The price of a house or the duration of a patient's stay in a hospital are two examples that best characterize regression trees.

C. Making a Decision Tree:

A list of Instruction examples is separated into the smallest subsections throughout this method of practice, and a correlated decision tree is constructed incrementally. The test dataset will be overlaid by a decision tree, which will return it after the course. The basic idea is to use a decision tree to divide the state space across clusters, or heavy areas, and then empty the scattered provinces.
In Decision Tree Classification, a specific example is categorized by putting it through a series of tests to decide the class mark. A decision tree is a hierarchical order in which these tests are organized. The Divide-and-Conquer method is used in these Decision Trees.
Decision Tree Classifier
We begin first at subtree and split the content further into aspects that yield the largest content gain using the algorithm. When we get closer to a final decision, we'll be able to reduce the confusion. We can repeat this partitioning process adaptively at many of the leaf nodes till the leaves are perfect, ensuring all of the observations from each root node relate to a certain group. To measure efficiency, this type of practice can set a limit on the size of the tree. We're undermining the authenticity a little here because the final leaves might still be impure. by a method of reducing ambiguity in the lead-up to a final decision.

D. Random Forest Classifier

A supervised learning algorithm is random forests. This can be used for classification as well as regression. They are the most adaptable and practical algorithm. The forest would be more resilient if you plant more trees. Random forests, on the other hand, build decision trees on randomly selected data samples to get a prediction from each tree and then vote on the best solution. These Random forests often serve as a good indicator of how important a function is. Random forests can be used for a variety of things, including recommendation engines, picturerecognition, and feature selection. Sorting is often used to detect fraudulent behavior, forecast diseases, and recognize faithful loan applicants. This is the foundation of the Boruta algorithm, which selects the most appropriate features in a dataset like the one shown in Figure 5.
Itoperatesinfourstages:

1) Takerandomsamplesfromacollectionofresults.
2) Outofeachsample,buildadecisiontreeandgetpredictionresultsfromeachdecisiontree.
3) Castavoteforeachexpectedoutcome.
4) Asthefinalprediction,choosethepredictionoutcomethatreceivedthemostvotes.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 7803 - 7809
Received 25 April 2021; Accepted 08 May 2021.

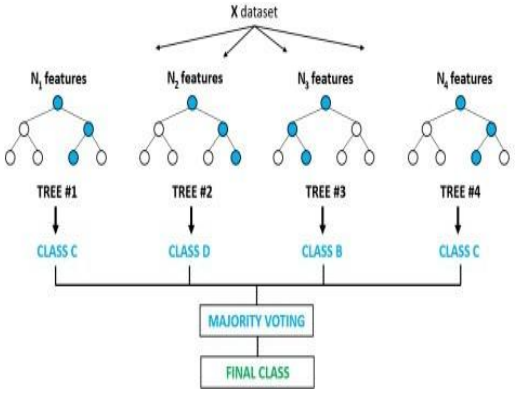


Fig. [5] shows the final result of the random forest classifier

E. Naïve Bayes

The Naive classifiers, which are a form of possibility classifier, are created using the Predictive law. It is referred to as "Naive" because it needs fixed distinctions between input parameters. You can use Bayes optimization to evaluate your performance. As the name implies, this pattern implies that all of the data items are "Naïve," that is, unrelated.

Naïve Bayes algorithm that you can use to analyze your results. As the name implies, this algorithm assumes that all of the variables in the dataset are "Naive," meaning they are unrelated. It is one of the most widely used classification algorithms for determining the dataset's base accuracy.

Naive Bayes precedence could be that:

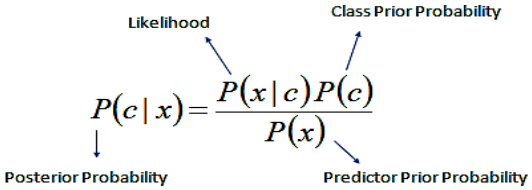
- It is a more straightforward method of determining the class of the sample data set. It performs admirably in multi-class forecasting.
- When the assumption of autonomy is met, a Bayes Decision outperforms other frameworks such as regression models, and it needs less training data.

The major drawback of the Naive Bayes is that:

- Furthermore, simple Bayes is regarded as a poor estimator since the chance outcomes are not considered legitimate.

The presumption of significant risk factors is another flaw in Bayes. In the actual world, gettinga statistical technique that is entirely distinct is virtually impossible.

The dataset for Naive Bayes is shown in figure 6 below.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Fig. [6] shows the probability derivation of the Naive Bayes classifiers.

F. Support Vector Machine

The Algorithm is a supervised neural network capable of identification, replication, and pattern discovery. From the standpoint of the algorithm, it finds a higher dimensional space in the N-dimensional field that accurately identifies knowledge sources.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 7803 - 7809
Received 25 April 2021; Accepted 08 May 2021.

Those certain Support Vector Machines, which are focusing on mathematical learning systems, are amongst the most reliable estimation methods. The Classifier can execute semi inference as well as rigid grouping, fully filtering their samples into greater features extracted. A clear picture of the SVM is shown in the figure below 7.

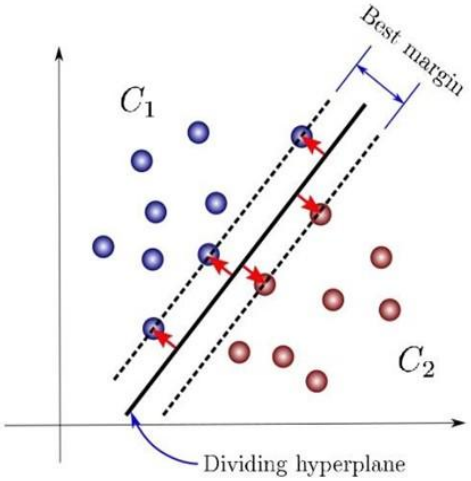


Fig. [7] visual representation of the SVM graph

To differentiate the two classes of knowledge points, a variety of alternate hyperplanes can be used. The objective is to find a plane with the largest margin, or the most evenly distributed data points from both classes. It helps by widening the margin gap, allowing more data points to be classified as untrustworthy.
Hyperplanes are inspection procedures that assist in the grouping of pieces of information. Points on either side of the plane are frequently classified as belonging to various groups. The selection of features determines eachhyperplane's scale. If there are only two input factors, the centroid is just a line; if there are three input components, the pixel can become a double plane. Whenever the set of instances exceeds three, it's difficult to imagine everything.

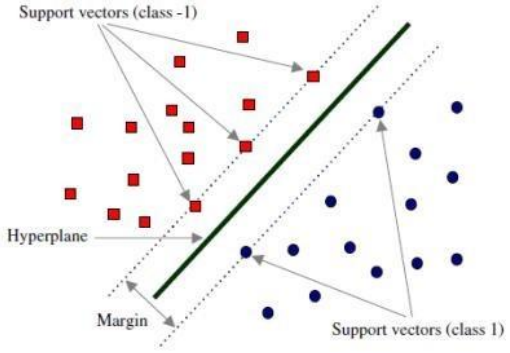


Fig. [7.1] shows the possible hyperplanes are division curves in the decision boundaries.

Using the sigmoid equation, During logistic regression, we can obtain the product of the response variable and compact the absolute value of [0,1]. If the collapsed coefficient is lesser than a predefined threshold (0.5), it is marked 1, otherwise, it is marked 0. The performance of the linear model is obtained in this Support vector machine, and if it is significantly higher than 1, it is calculated for one class, and if it is -1, it is determined again with class. After the edge values in the Support vector machine are modified to 1 and-1, we get this stabilization range with values([-1,1]), something that operates like a boundary.

IV. References

[1] "PatientMedicalDataforNovel Coronavirus COVID-19" from theWolfram Data Repository, Wolfram Study, 2020.https://doi.org/10.24097/wolfram.11224.data

[2] What are the COVID-19 symptoms?https://www.who.int/health-topics/coronavirus
[3] MoritzKraemer,Earlydescriptionsofthe2019nCoVoutbreakepidemiologicaldetails
[4] E. Dong, H. Du, and L. A Gardner Lancet Infectious Diseases havecreatedaninteractiveweb-basedtooltomonitorCOVID-19inreal-time.https://doi.org/10.1016/S1473-3099(20)
[5] Y.Zoabi, S. Deri-Rozov, and N. Shomron. COVID-19 diagnosispredictionbasedonsymptomsusingmachinelearning.(2021).
[6] M.RoserandH.Ritchie,"2019istheyearofthecoronavirus(COVID-19",https://www.mayoclinic.org/diseasesconditions/coronavirus/diagnosis-treatment/drc20479976
[7] Menni,C.etal.Real-timetrackingofself-reported symptoms topredictpotentialCOVID-19.Nat.Med.26,1037–1040(2020).
[8] Coronavirus: rolling out population monitoring for COVID-19 in theNHS,Whittington,A.M.etal.OpinionoftheBritishMedicalJournal(2020).
https://blogs.bmj.com/bmj/2020/02/17/coronavirus-rolling-out-community-testing-for-covid-19-in-the-nhs/
[9] COVID-19EpidemicAnalysis Using Machine Learning and DeepLearning Algorithms, Penn, N. S., Sonbhadra, S. K., and Agarwal,S.medRxiv,(2020). https://doi.org/10.1101/2020.04.08.20057679
[10] GenderDifferencesinCOVID-19Patients:AFocusonSeverityandMortality,Jin,J.-M.etal.PublicHealth8(Front)(2020).
[11] Coronavirus Disease 2019 (COVID-19) - How It Spreads, Centers forDisease Control and Prevention, 4th March 2020.ttps://www.cdc.gov/coronavirus/2019-ncov/prepare/transmission.html
[12] CoronavirusDisease,M.RoserandH.Ritchie(COVID-19).OxfordMartin,2020,OurWorldinData.