2021 5th International Conference on Computer, Communication and Signal Processing (ICCCSP) | 978-1-6654-3277-1/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCSP52374.2021.9465516

Extracting Insights and Prognosis of Corona Disease

Dr. V. Chandra Shekhar Rao
Associate Professor
Department of CSE
KITS,Warangal
Telangana ,India
vcsrao.cse@kitsw.ac.in

Swapna Gampa
Student
Department of CSE
KITS,Warangal
Telangana, India
gampaswapna8@gmail.com

VijaySai Rama
Student
Department of CSE
KITS,Warangal
Telangana ,India
vijaysai219@gmail.com

Hemanth Anumala
Student
Department of CSE
KITS,Warangal
Telangana ,India
Hemanthanumala7@gmail.com

Akshay Gadepally
Student
Department of CSE
KITS,Warangal
Telangana ,India
gadepallyakshay1999@gmail.com

ABSTRACT - COVID-19, Corona Virus Disease-2019, belongs to the genus of Coronaviridae. A pandemic with no immunogen or has neither clinically well-tried immunogen nor medicine making unpredictable havocs within the human lives in each country throughout the globe. In this paper, we put forward a web app that enables users to predict covid19 results in real time using an online intelligent device. This app is jam-packed with information that allows the user to describe their COVID symptoms. It then processes user-specific details to see whether a person is affected with covid19. This entire process is done using an intelligent data mining technique to develop data mining prediction models which are further used for the guessing of COVID-19, with an epidemiological data of COVID-19 patients of South Korea. An analysis of datasets is done to understand how a person is affected. These Prediction models are built using machine learning algorithms like decision tree (DT), logistic regression (LR), random forest (RF), and K-nearest neighbor (KNN) and their performances are computed and evaluated. These algorithms were directly applied to the data with python as programming language to develop the different models. The output of this study have proven that the model developed with a Random forest DM algorithm stands to be the best model to predict the infected patients more effectively among the models developed with different algorithms, with an overall accuracy of 98.83%

Keywords: COVID-19, Decision Tree Classifier, Logistic Regression, Machine Learning, Data mining

I. INTRODUCTION

Corona is indeed a huge virus family that has been linked to illnesses ranging from the flu virus to more serious conditions including Extreme Acute Respiratory Syndrome (SARS). Since December 2019, outbreaks of the COVID-19 virus have been causing global health issues. The World Health Organization reported it a pandemic in March 2020. COVID-19 had affected 186 countries and territories across the world as of March 21, 2020, with over 280,000 reported cases and 11,842 victims. The seriousness of the cases puts a strain on medical facilities and creates an intensive care resource shortage.

As COVID-19 spreads rapidly across the world, the ability to classify cases at risk of death has become an urgent yet difficult requirement. Prediction models will assist in health resource management and preventive planning. Data mining algorithms and techniques are an sophisticated artificial intelligence (AI) technique for developing predictive models, finding novel, useful, and true hidden patterns in datasets, and performing data analysis. Data mining has been commonly used in the healthcare industry for a number of purposes, including disease prevention, contamination regeneration are evaluated. They have the ability to derive valuable information from raw data without explicitly doing so. The main aim of the prediction tasks is to construct models that can estimate the mapping of inputs to outputs depending on a specimen of data called as training-data.

The models were created using dataset available on internet extracted from Korea Centers for Disease Control and Prevention (KCDC), and dataset instances of contaminated 2019-nCoV pandemic symptoms were considered. For non-visible inputs,

978-1-6654-3277-1/21/$31.00 ©2021 IEEE

trained models can be used to predict outputs. For exploratory analysis, these approaches are more versatile and accurate than traditional statistical analysis. The information collected may be applied to a number of areas, including the healthcare industry. In the proposed model, to build the models in the proposed model, DM algorithms such as DT, Logistic regression, Random Forest, and K-Nearest neighbor were implemented directly to the dataset using the Python. A UI is created to get the entries to include basic information, symptoms. By optimizing, this classifier aims to reveal patients with covid19, which reduces the clinical burden and could reduce mortality.

II. EXISTING SYSTEM

Nowadays, all other people are affected by the COVID-19 virus. While we believe that all physicians and medical experts have done good research and have a great deal of experience, treating COVID-19 has become a challenging task. The WHO and doctors preserve data from infected patients and work together to find out more about this new virus.

Assuming, they have collected symptoms of the patients who were affected with COVID previously; it would take a lot of time for them to compare those symptoms with the symptoms a person is currently suffering with.
As a result, analysis of the patient's COVID-19 outcome is challenging.

The disadvantages of the existing system:

- The current COVID-19 prediction system is the manual method.
- Several papers have been published on COVID-19 forecasting using a variety of machine learning algorithms and data mining techniques, but no systems have been implemented and tested to date..
- The existing system, i.e. the manual system where people go exclusively to places where COVID-19 testing is conducted, for example, hospitals, PHC, and campaigns.
- A few samples are taken from people using nose and tongue Q-tips and sent to a lab for analysis and prediction.
- Once the result is obtained, individuals are reported as COVID-19 and indicate precautions to avoid COVID-19.
- The entire process takes approximately 1.5 days and an individual is not informed of the infection, which can result in the spread of the disease during this period.

III. PROPOSED SYSTEM

The system we are proposing is designed to deal with the disadvantages of the current system. Machine learning techniques are used to create a predictive data mining model that can accurately predict whether an individual has COVID-19. It would assist medical institutions in making decisions.

Here, we create a web app that allows anyone who wants to check whether they have corona to enter symptoms from the comfort of their own home. This web app would analyze these symptoms with that of the symptoms in the data set with the help of different algorithms (Random Forest, Decision tree, Logistic regression, KNN) and give the result directly to the person within seconds without the dangers of spreading the virus making it safer and efficient.

IV. METHODOLOGY

The methodology here is to implement a COVID-19 prediction system that utilizes a data extraction technique that includes logistic regression (LR), random forest (RF), decision tree (DT), and k nearest neighbor (KNN). This procedure can be expressed as the “knowledge Discovery Process”, this process includes the following steps:

A. Dataset Collection

Korean center for disease control provide many datasets and made them available on the Kaggle website. There are a large number of datasets in website regarding coronavirus, out of which we chose the dataset which is more appropriate for our use i.e. prediction of covid-19. We used a South Korean epidemiological dataset of COVID infected persons.

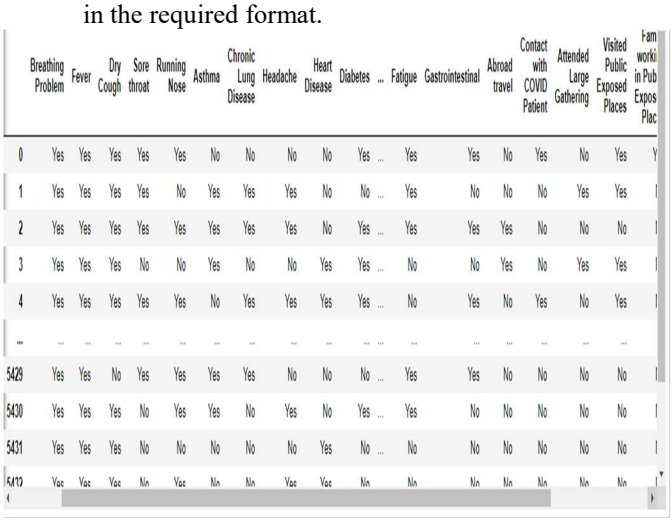
B. Dataset Description

The dataset has 5434 instances with 21 attributes that includes patient ID, global number, breathing problem, fever, dry cough, sore throat, running nose, chronic lung disease, asthma, headache, heart disease, diabetes, fatigue, abroad travel, contact with covid patients, attended the large gathering, visited public places, family working in public places, wearing masks, sanitization from market COVID.

C. Cleaning and preprocessing dataset

Pre-Processing of the data is the first step in analyzing the dataset. It is basically cleaning the data and obtaining into the format where we can give valid insights and improve the data quality. In most of the cases, Dataset is obtained from the websites in two ways. One to directly obtain from the required website in csv/excel/text format. The other way is it is obtained by web scraping from different websites. Web scraping means that retrieving and extracting

data from the websites where it is not directly shown in the required format.



	Breathing Problem	Fever	Dry Cough	Sore throat	Running Nose	Asthma	Chronic Lung Disease	Headache	Heart Disease	Diabetes	...	Fatigue	Gastrointestinal	Abroad travel	Contact with COVID Patient	Attended Large Gathering	Visited Public Exposed Places	Fam worki in Pub Expos Plac
0	Yes	Yes	Yes	Yes	Yes	No	No	No	No	Yes	...	Yes	Yes	No	Yes	No	Yes	Y
1	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	No	...	Yes	No	No	No	Yes	Yes	I
2	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	...	Yes	Yes	Yes	No	No	No	I
3	Yes	Yes	Yes	No	No	Yes	No	No	Yes	Yes	...	No	No	Yes	No	Yes	Yes	I
4	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	...	No	Yes	No	Yes	No	Yes	I
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5429	Yes	Yes	No	Yes	Yes	Yes	Yes	No	No	No	...	Yes	Yes	No	No	No	No	I
5430	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes	...	Yes	No	No	No	No	No	I
5431	Yes	Yes	Yes	No	No	No	No	No	Yes	No	...	No	No	No	No	No	No	I
5432	Yes	Yes	Yes	No	Yes	No	No	Yes	Yes	No	...	No	No	No	No	No	No	I

Fig.1. Example of the dataset used after cleaning.

A normal webpage may contain different kind of data. When we inspect a webpage, all we see is the HTML tags that are embedded. Content of the webpage is extracted and retrieved into the required format; it is downloaded into spreadsheet or to a database.

We often need large training dataset for the model in Machine Learning. We collect the data by scraping sources on the web. Web scraping plays a critical component in the project. There are automated tools/ libraries such as “Beautiful Soup”, “Selenium”, “Scrapy” and the others. Beautiful Soup can be used to scrape the website with just few lines of the code. It extracts the text between the HTML tags. It calls the Constructor by the desired parsing method. Whereas Selenium was developed to facilitate automated testing. But it is found as off label use as a web scraper. As it is used for the automating testing and the web scraper at the same time, it is considered to be Volatile. Unlike, Beautiful Soup and Selenium, Scrapy is known to be fast and efficient as it is written with popular networking framework in python and that gives scrapy some asynchronous capabilities.

In this case, the dataset obtained from the Kaggle site, where only require relevant attributes of the symptoms are extracted. The dataset has 5434 instance and attributes like breathing problem, fever, dry cough, sore throat, running nose, chronic lung disease, diabetes, fatigue, abroad travel, contact with covid-19 patients, attended the large gathering, visited public places, family working in public places, wearing masks, sanitization from the market are only taken into account and the attributes like patient ID and Global number are excluded from this dataset. Then the union of all these is done with all the common symptoms changed into column data making it easy for data application. There are certain steps before we say that the dataset is Pre-Processed and ready for analysis.

They are:

a) Understand the dataset:

Check the dataset format, its attributes, and unique values in a particular to get a basic idea of what the dataset is about. Understanding the dataset is the crucial and important step to proceed further. To get the top 10 rows of the dataset we can use df.head() method, incase if we want to the bottom 10 rows, we can use df.tail(). By utilizing these methods, we can see the records in the dataframe and values can be identified.

b) Check missing values:

We know that the data obtained is mostly raw and inconsistent. One must purify the data. And the very first step to do so, after understanding the dataset is to check for missing values in the dataframe. There is a method called isnull() which can be applied to the whole dataframe in one go or we can apply it column wise to know the null values/missing records of that particular field belonging to that particular column. It is easier to apply for the dataframe instead of checking each column separately. The output of this will show if there are no. of values present among Total values.

c) Drop missing values:

Once Missing Values are identified in the data frame, Missing values should be removed only if the record is unaffected. If it removes the entire row, there is no point in dropping the missing values. Instead, we should go for replacing the missing value. Replacing with appropriate values is suitable only if we do not want to lose the information. For example, if a dataset with numeric row has 5 attributes, out of which 3 are null, they can be either replaced with 0 or a relatively large number as an indication that it is not null. So, that the computation won’t go wrong.

Those are the basic thumb rules for preprocessing the data. In this project, the dataset obtained is neither having two columns which are neither “Yes” nor “No”. They are “Age” column and “Body Temperature” column. These columns have different set of values for each record, which are not constant. These different values need to be grouped accordingly so that they fall into the certain range. Accessing the column with ranges is far easier than accessing it with unique value per record. It does not serve the purpose of computation. We can either drop the columns or get it into the required format. If first option is chosen, there is a loss of data, we know that every output is dependent on different attributes; it may lead to false prediction of the output. We should only drop the column if it is redundant. In this case, age group of different

persons is grouped into three different categories: “17-35”, “36-64” and “65+”. Also, the body temperature is classified into three categories: persons having body temperature of range “96-100.4”, “100.41-102.2”, “102.21+”. And these values are replaced if a particular value falls under mentioned ranges.

D. Label Encoding

Label Encoder is a part of Scikit Learn Preprocessing Package, where it converts the unique values into numerics starting from 0.. If there are 2 different values “Yes” and “No” namely, then they are transformed into “1” and “0” each of them indicating their previous values. Machine always understands the inputs in the form of numeric’s, it is better if we move into that format before we proceed with the further steps of Data Training. Label Encoding can be done in multiple ways. One way is to apply fit transform method on the whole dataframe. The other way is to apply for column by column. The significance by utilizing the second way of label encoding is that, if the dataset has only certain column that needs to undergo label encoding, it is helpful. After encoding the dataframe, we save into a new dataframe instead of overriding it.

	Breathing Problem	Fever	Dry Cough	Sore throat	Running Nose	Asthma	Chronic Lung Disease	Headache	Heart Disease	Diabetes	...	Fatigue	Gastrointestinal	Abroad travel	Contact with COVID Patient	Attended Large Gathering	E
0	1	1	1	1	1	0	0	0	0	1	...	1	1	0	1	0	
1	1	1	1	1	0	1	1	1	0	0	...	1	0	0	0	1	
2	1	1	1	1	1	1	1	1	0	1	...	1	1	1	0	0	
3	1	1	1	0	0	1	0	0	1	1	...	0	0	1	0	1	
4	1	1	1	1	1	0	1	1	1	1	...	0	1	0	1	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
5429	1	1	0	1	1	1	1	0	0	0	...	1	1	0	0	0	
5430	1	1	1	0	1	1	0	1	0	1	...	1	0	0	0	0	
5431	1	1	1	0	0	0	0	0	1	0	...	0	0	0	0	0	
5432	1	1	1	0	1	0	0	1	1	0	...	0	0	0	0	0	
5433	1	1	1	0	1	1	0	1	0	1	...	1	0	0	0	0	

5434 rows × 21 columns

Fig.2. Example of data after label encoding.

E. Feature Selection

The data which we got from the Kaggle website has few unnecessary attributes which have been removed and only the data which can be used for prediction have been chosen.

F. Split the given data set accordingly

The data was divided into 70 percent learning (also known as training data) for predictive model development and 30 percent testing (test data) to verify the efficacy of the developed models. The training dataset was trained using four algorithms: LR, Decision Trees, KNN, and Random Forests. The Python programming language was used to build the models.

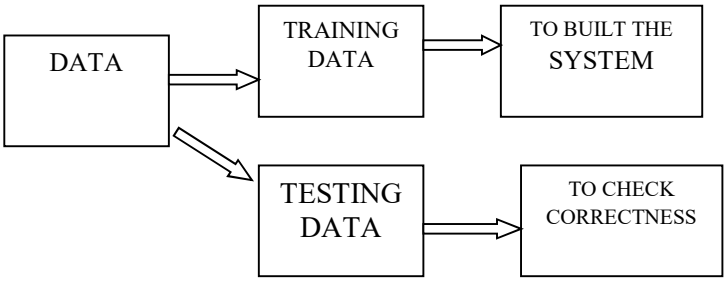


Fig.3. Example of splitting of the dataset.

G. finding the accuracy of the model:

As a final step, the performance evaluation i.e. finding the accuracy of the model is done which helps us to determine which algorithm is best suited to predict COVID-19. In this proposed system we use a confusion matrix as a performance evaluator

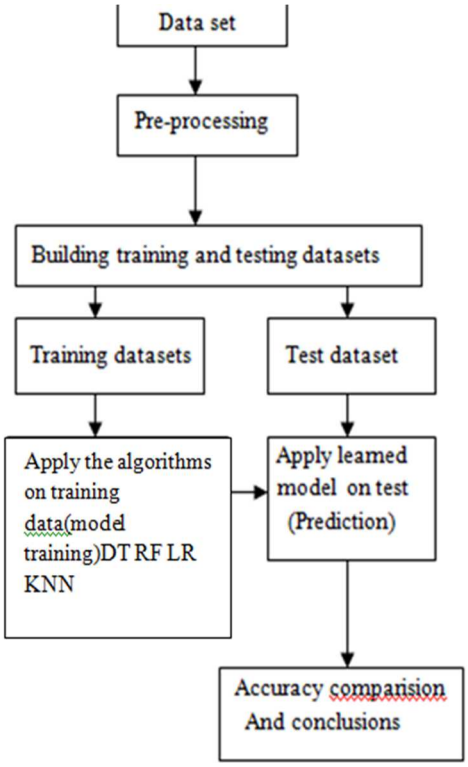


Fig.4. Example Flowchart for prediction

V. DATA MINING TECHNIQUES/ ALGORITHMS

In ML, classification is performed in 2 levels that involve learning as well as prediction. The ML model is constructed depending on provided training-data in the learning phase. This model is utilized for predicting the answer for provided data in the prediction stage.

A. Logistic Regression (LR)

One of the most popular and well-known algorithms for solving a classification problem is logistic regression. LR is used for describing data and examining the dependence of one dependent variable and one or more independent variables (categorical or continuous).

binary logistic regression method is utilized in condition if dependent variable has 2 values, such as 0 and 1, yes and no, or true and false. Multinomial LR, on the other hand, is used when the dependent variable has more than two values.
It's named after 'Logistic Regression,' since the technology behind it is very familiar to Linear Regression.

Problem Formulation

When we apply the LR of a dependent variable named y1 to a set of non dependent variables named x1 = (x1... xn).
We find logistic regression function (x) for every instance u = 1... G so that the estimated replies/responses p(xi) are nearer to original responses yi as possible.

If the underlying mathematical dependence remains unchanged, you can use the logistic regression function (x) to predict outputs for new and unknown inputs once you've learned it.

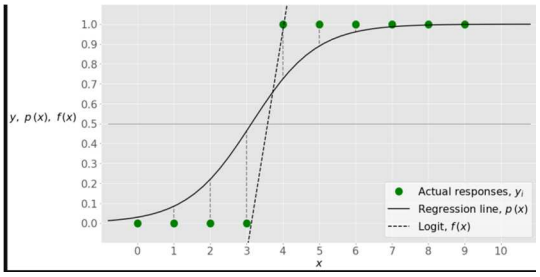


Fig.5. Example of a graph for single or mono logistic regression.

The findings classified as zeros are shown in white circles, while those classified as ones are shown in green circles.

The logistic regression formula is as follows:
We can model a non-linear relationship linearly using a logarithmic transformation on the outcome variable. This is the Logistic Regression equation.

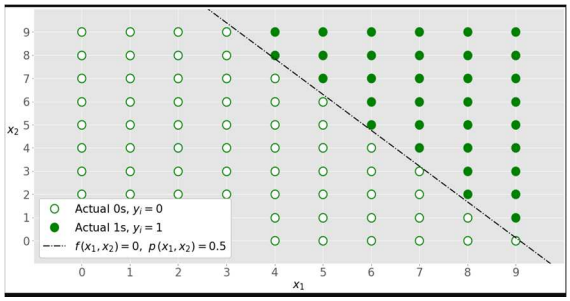


Fig.6. Example of a graph for multi logistic regression.

$$Log \qquad (q/1-q) \qquad (1)$$

Assume that the variables which depend on others are binary i.e. 1 & 0, with 0 denoting a false (negative) return/attribute and 1 denoting a true (positive) return. Consequently, the discrete values average would be the percentage of true attributes. If q represents a percentage of considerations with a result 1, then (1-q) denotes likelihood of 0 outcome.

B. Decision trees (DT)

The DT algo is supervised training algorithm in which input is constantly divided based on feature. Two entities, decision nodes and leaves, can be used to illustrate the tree. Both classification and regression problems are solved using decision trees.

Designing the outline to design a way to proceed or conducting appropriate model evaluation are examples of what it entails. It is being utilized to break down tough issues or segments. Each DT branch would result in a unique outcome. While determining a target class for an object, we begin at the top of trees. The parameters of main instance and record's instances are compared. It makes its decision after running a series of tests.

The aim of using a DT is to create a training model which could be utilized to determine the target / attr ibute value by automatically learning through previ ous information (training data).

We utilize the CART algo that abbreviates for Classification and Regression Tree algorithm, to design a desired tree. A decision tree asks questions & then divides tree into sub trees based on the responses (Yes/No). A decision tree's basic form is depicted in the diagram below:

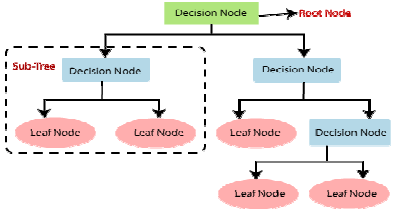


Fig.7. Example of decision tree.

Measures for feature extraction (i.e. Attribute selection) when constructing a Decision tree, the question of how to choose the optimal parameters for the root node and sub-nodes arises. As a result, there is a technique known as Attribute Selection Measure or ASM in short, that can be used to solve such problems. We could quickly pick the best

attribute for the nodes of the graph using this measure.

Attribute Selection Measures

When building a Decision tree, the question of choosing the appropriate value/instance for main node & child node arises. As a result, there is a method called ASM that can be used to solve such problems. one will quickly pick the optimal parameter for tree's nodes using this measurement. Information Gain and the Gini Index are two widely used ASM techniques.

C. *Random Forest (RF)*

RF creates a "forest" from a set of DT's that are functional prototypes through utilizing "bagging" technique. The underlying concept of the bagging approach is that integrating various learning models boosts result. A random forest's hyperparameters are almost identical to those of a DT or a bagging classification algorithm. Fortunately, you can use the classifier-class of random forest instead of combining a decision tree and a bagging classifier. To cope for regression problems in an RF, we could utilize the algorithm's regressor. As the number of trees grows, the model becomes more random. Rather than seeking for the most suitable feature, it searches for the best attribute among a smaller data set while dividing a node.

Also, as a result, there seems to be a great deal of variance, that contributes towards a more accurate model. The hyperparameters in the random forest are used to either improve the model's predictive ability or make it more efficient. The random forest has benefit of calculating relative significance of each feature on the prediction.

.

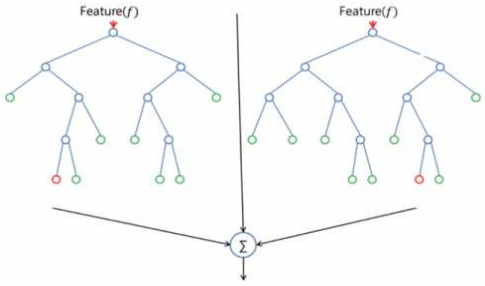


Fig.8. Example of random forest.

D. *K Nearest Neighbor*

The K-NN algo is depended on the method of supervised learning and falls under category of fundamental ML algo's. The K-Nearest neighbor supposes that the fresh instances and already present instances are identical and puts the new instance in group which closely resembles to the already available group.

A small k value in this means that noise will have a greater impact on the result, while a large value indicates that it will be computationally expensive, which contradicts KNN's basic theory (that points that are near might have similar densities or classes). It's simple to choose k by setting

$$k = n\ (1/2).$$

Usually, data scientists would choose:

a. If the number of classes is two, an odd number.
b. Setting

$$k = \text{sqrt}(t) \qquad (2)$$

Is another easy way to pick k.

Here, t is the sum of data points available throughout the training set.

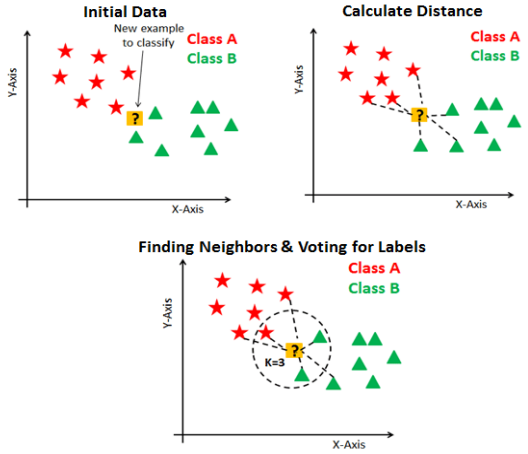


Fig.9. Example of K nearest neighbors.

VI. TECHNOLOGIES USED

A. *Jupyter notebook:*

Jupyter allows you to integrate computer program, , numeric performance, graphical representations, and digital content in a legal book.

B. *Visual Studio (IDE)*

A software application that provides comprehensive (IDE) is a feature-rich framework developed by Microsoft that can be used for a variety of tasks.

C. *Python*

Python code is comprehendible by humans, which makes it easier to build models for machine learning. It is used to write machine learning code.

VII. EXPERIMENTAL RESULTS

Data mining and machine learning predictive tasks are carried out using the Python programming language.. Python is used to implement data mining algorithms in conjunction of aid of repositories with a particular intent such as NumPy, pandas, and others.

In logistic regression, despite the fact that the model's anticipated precision of 97%, the

quantity of misclassified tests seemed, by all accounts, to be 56. Added, feature importance is not given equally example Visiting public exposed spots are given more significance than numerous other significant indications which assists us with foreseeing Covid. A portion of the symptom's importance values fell into the negatives (ignored). Nonetheless, some are given more significance than others.

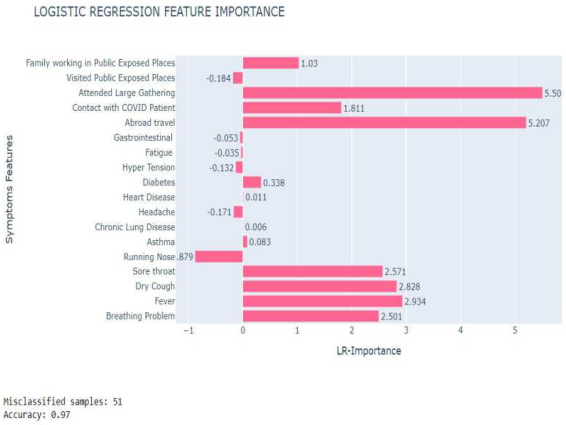


Fig.10. Feature importance of logistic regression

In K-nearest neighbor, despite the fact that the model's anticipated precision of 98%, the quantity of misclassified tests gave off an impression of being 35. Added, a portion of the features are not considered and others are neglected completely.

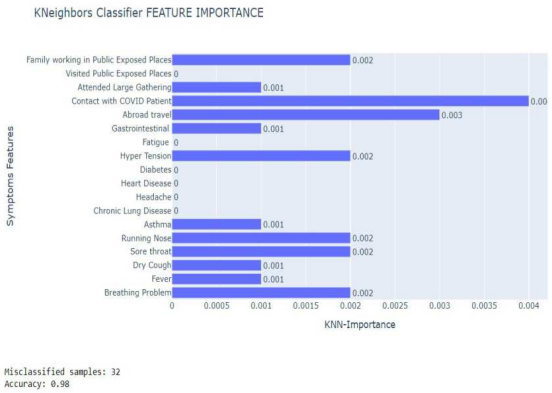


Fig.11. Feature importance of k nearest neighbor.

In the decision tree despite the fact that the model's anticipated exactness of 95%, the quantity of misclassified tests gave off an impression of being 85(which is extremely high when contrasted with different classifiers). Added, just 7 highlights out of 18 are considered i.e. prediction is managed without thinking about the entirety of the symptoms. So the decision tree expectation is viewed as wrong.

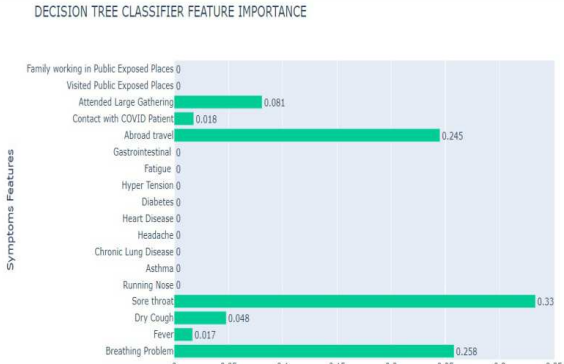


Fig.12. Feature importance of Decision tree

The random forest model was anticipated with a precision of 98%, yet additionally, the quantity of misclassified tests seemed, by all accounts, to be 27 (low) contrasted with different models. Added, every one of the symptoms is considered for the expectation of infection. So, we choose random forest as the best classifier.
This model has shown that every one of the attributes is important for predicting the COVID-19. From the model, we conclude prediction of COVID-19 is effectively done utilizing a random forest classifier.

Fig.13. Feature importance of Random forest

VIII. PERFORMANCE EVALUATION OF DEVELOPED MODEL

To evaluate the accuracy of Data mining models, validation techniques are used. Using data mining or machine learning algorithms, the techniques assess the model's consistency and productivity. Accuracy is one of the major performance assessment techniques/strategies for the data mining model. The amount of instances of a dataset predicted right for the model created by the DM algo is said to be accuracy.
Expressed this way:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

True positives (TP): Predicted positive and are actually positive.

False positives (FP): Predicted positive and are actually negative.

True negatives (TN): Predicted negative and are actually negative.

False negatives (FN): Predicted negative and are actually positive.

The model created using Random Forest proved to be high productive, with the greatest prediction performance of 98.22 percent.

IX. RESULT AND DISCUSSIONS

A. Results based on accuracy:

A direct application of Decision tree, LR, RF, K-NN data mining techniques is done on dataset utilizing the programming language called Python. In any case, built model generated using the RF algorithm was found to be the most reliable, with 98.22 percent accuracy, giving it the appearance of being the most advanced among the others.

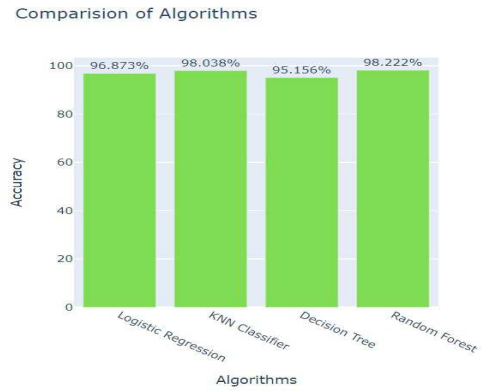


Fig.14. Graph comparing accuracy of algorithms

B. GUI:

The graphical user interface is created using a visual studio through which the user inputs values which are taken as a test set and used for the prediction of COVID-19.

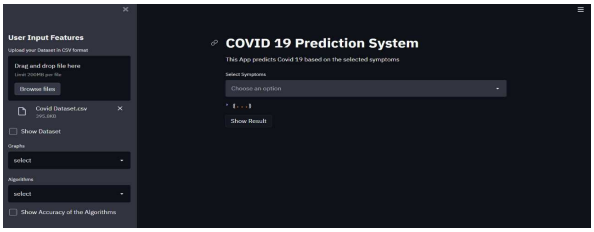


Fig.15. Graphical user interface

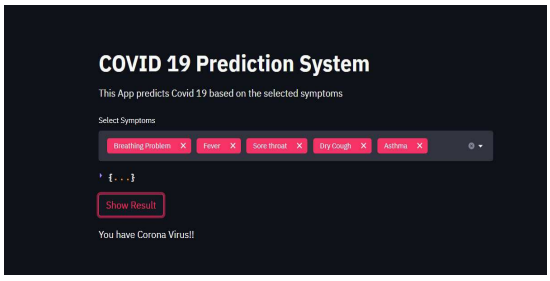


Fig.16. web app predicting corona virus based on symptoms provided by user

IX. CONCLUSION

In order to predict COVID-19, the paper compared various machine learning algorithms. Using a COVID-19 epidemiological dataset from South Korea, we used data mining techniques to develop various models for the prediction of corona virus affected patients. The dataset was subjected to DT, LR, RF, and also K-NN algorithms utilizing pythonprogramming language.

The model trained using Random Forest is proven to be more effective having an accuracy of 98 .22 percent, succeeded by K-NN of 98.0 percent, LR of This is succeeded by K-NN of 98.03 percent, LR of 96.83 percent, and Decision tree 95.186 percent. Experts in medical services will benefit more from the existing models in combating the pandemic.

References

1. Obenshain, M.K,“ Application of Data Mining Techniques to Healthcare Data”, Infection Control and Hospital Epidemiology, 25(8), 690–695, 2004.
2. International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 04 Issue: 03 | Mar -2017 www.irjet.net p-ISSN: 2395-0072
3. Corona virus disease 2019 (COVID-19), https://www.maoclinic.org/diseasesconditions/ corona virus/diagnosis-treatment/drc20479976, (Last Accessed 12.05.2020
4. Advice on the use of point-of-care immunodiagnostic tests for COVID-19, https://www.who.int/newsroom/ commentaries/detail/advice-on-the-use-ofpointof-care-immunodiagnostic-tests-forcovid-19, (Last Accessed 12.05.2020)
5. Viral Testing Data in the U.S., https://www.cdc.gov/coronavirus/2019-ncov/casesupdates/testing-in-us.html, (Last Accessed 12.05.2020)
6. What are the symptoms of COVID19?,https://www.who.int/emergencies/diseases/ novelcoronavirus-2019/question and-answershub/q-adetail/qacoronaviruses#:~:text=symptoms (Last Accessed: 11.05.2020)
7. Chapman, P., Clinton, J., Kerber, R. Khabeza, T., Reinartz, T., Shearer, C., Wirth, R.: “CRISP-DM 1.0: Step by step data mining guide”, SPSS, 1-78, 2000
8. Dataset which is extracted from kaggle website is used for developing this model

https://www.kaggle.cohemanthhari/symptoms-and-covid-presence
9. Bibhuprasad Sahu, Sujata Dash, Sachi Nandan Mohanty, Saroj Kumar Rout. "Ensemble Comparative Study for Diagnosis of Breast Cancer Datasets", International Journal of Engineering & Technology, 2018
10. Nazmus Sakib Patwary, Protap Kumar Saha, Ifthakhar Ahmed. "Nurse care activity recognition challenge using a supervised Exclude quotes Off Exclude bibliography Off Exclude matches Off methodology", Proceedings of the 2019
11. International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers - UbiComp/ISWC '19.