PROJECT REPORT

"SymptoSense: COVID-19 Symptom Analyser"

SUBMITTED TO : R FOSSEE Semester Long Internship 2024

SUBMITTED BY: Rajeev Kumar

Institute Of Engineering & Technology Lucknow
(226021)

Quick View in Data Set:

- Country :: country name
- Age:: age in years
- Gender:(Male, Female,TransGender)
- Symptoms::(Fever,Tiredness,Dry-Cough,Difficulty-in-Breathing,Sore-Throat","Pains,Nasal-Congestion,Runny-Nose,Diarrhea)
- Experiencing_Symptoms::(Fever,Tiredness,Dry-Cough,Difficulty-in-Breathing","Pains,Nasal-Congestion,Runny-Nose)
- Severity::(Mild,Severe,Moderate)
- Contact::(Yes,No)

ABSTRACT

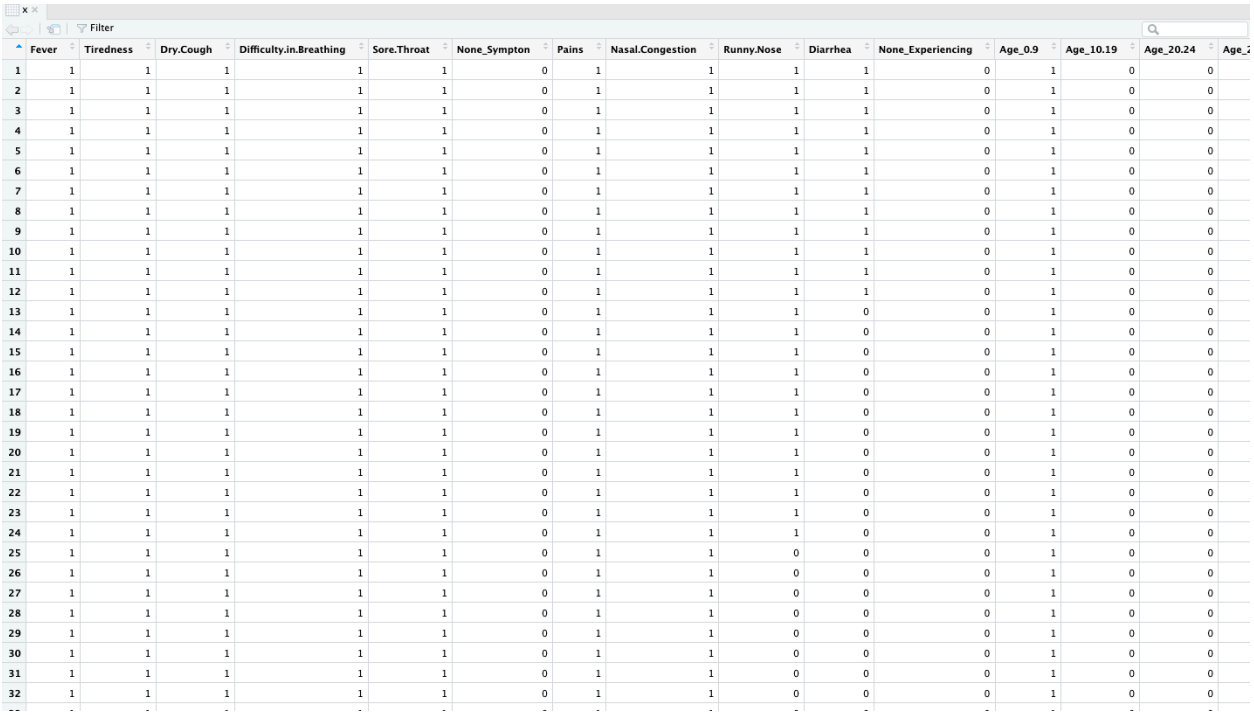
Technological advancement has a profound effect on all spheres of life, whether in the medical field or in any other field. Artificial intelligence has shown promising results in health care by making its decisions by analyzing and processing data. To prevent the spread and development of a life-threatening disease, the most important step is its early diagnosis. COVID-19 is a highly contagious disease, and has become a global epidemic that needs to be addressed as soon as possible. Due to its rapid speed of spreading comes the need for a system which can be used to detect the virus. With the increase in use of technology, lots of data about COVID-19 is readily available at our fingertips, which can be used to obtain important information about the virus. In this project, we compared the accuracies of different machine learning algorithms in predicting COVID-19 and used the most accurate one in the final model testing.

INTRODUCTION

In December 2019, the novel coronavirus appeared in the city of Wuhan in China [1] and was reported to the World Health Organization (WHO) on 31 December 2019. The virus posed a global threat and was named COVID-19 by the WHO on the 11th. February 2020. W.H.O declared the outbreak a public health emergency [2] and stated the following; “the virus is spread through the respiratory tract when a healthy person comes in contact with an infected person”. An infected person shows symptoms within 2-14 days. According to W.H.O the symptoms and signs of moderate to severe conditions are dry cough, fatigue and fever while in severe cases dyspnea, fever and fatigue may occur. People with other illnesses such as asthma, diabetes, and heart disease are at greater risk of contracting the virus and may become seriously ill. A system which can be used to detect the virus has become necessary due to the rapid spread of the virus, killing hundreds of thousands of people. Machine learning classification algorithms, data sets and machine learning software are essential tools for designing the COVID-19 predictive model. This project aims to compare different machine learning algorithms like K-nearest neighbors, Random forest and Naive Bayes with respect to their accuracies and then use the best one among them to develop a system which predicts whether a person has COVID or not using the data provided to the model.

Data Collection

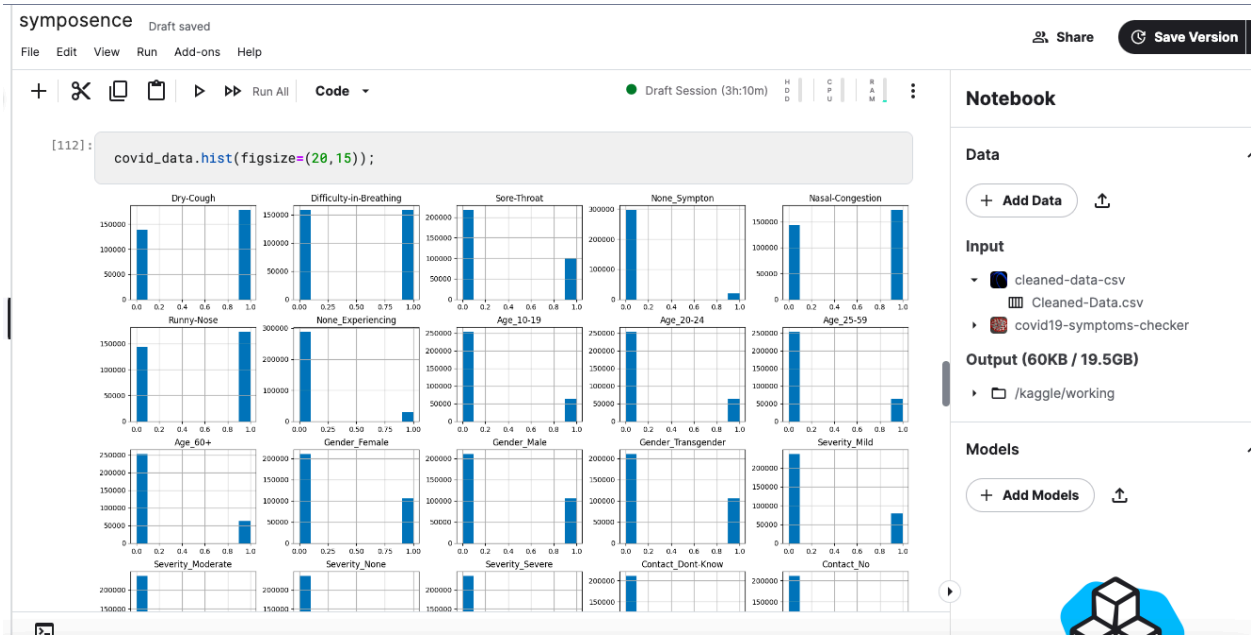
As the UNO & Indiangov.in has declared the Covid pandemic as a health emergency, researchers and hospitals have provided open access to data related to the epidemic. We procured a data set from kaggle.com and it has rows of columns 316800 x 27. This dataset contains 27 variables that could be determinants in the prediction of COVID-19, as well as one class attribute that defines if COVID-19 is found.



	Fever	Tiredness	Dry.Cough	Difficulty.in.Breathing	Sore.Throat	None_Sympton	Pains	Nasal.Congestion	Runny.Nose	Diarrhea	None_Experiencing	Age_0.9	Age_10.19	Age_20.24	Age_
1	1	1	1	1	1	0	1	1	1	1	0	1	0	0	
2	1	1	1	1	1	0	1	1	1	1	0	1	0	0	
3	1	1	1	1	1	0	1	1	1	1	0	1	0	0	
4	1	1	1	1	1	0	1	1	1	1	0	1	0	0	
5	1	1	1	1	1	0	1	1	1	1	0	1	0	0	
6	1	1	1	1	1	0	1	1	1	1	0	1	0	0	
7	1	1	1	1	1	0	1	1	1	1	0	1	0	0	
8	1	1	1	1	1	0	1	1	1	1	0	1	0	0	
9	1	1	1	1	1	0	1	1	1	1	0	1	0	0	
10	1	1	1	1	1	0	1	1	1	1	0	1	0	0	
11	1	1	1	1	1	0	1	1	1	1	0	1	0	0	
12	1	1	1	1	1	0	1	1	1	1	0	1	0	0	
13	1	1	1	1	1	0	1	1	1	0	0	1	0	0	
14	1	1	1	1	1	0	1	1	1	0	0	1	0	0	
15	1	1	1	1	1	0	1	1	1	0	0	1	0	0	
16	1	1	1	1	1	0	1	1	1	0	0	1	0	0	
17	1	1	1	1	1	0	1	1	1	0	0	1	0	0	
18	1	1	1	1	1	0	1	1	1	0	0	1	0	0	
19	1	1	1	1	1	0	1	1	1	0	0	1	0	0	
20	1	1	1	1	1	0	1	1	1	0	0	1	0	0	
21	1	1	1	1	1	0	1	1	1	0	0	1	0	0	
22	1	1	1	1	1	0	1	1	1	0	0	1	0	0	
23	1	1	1	1	1	0	1	1	1	0	0	1	0	0	
24	1	1	1	1	1	0	1	1	1	0	0	1	0	0	
25	1	1	1	1	1	0	1	1	0	0	0	1	0	0	
26	1	1	1	1	1	0	1	1	0	0	0	1	0	0	
27	1	1	1	1	1	0	1	1	0	0	0	1	0	0	
28	1	1	1	1	1	0	1	1	0	0	0	1	0	0	
29	1	1	1	1	1	0	1	1	0	0	0	1	0	0	
30	1	1	1	1	1	0	1	1	0	0	0	1	0	0	
31	1	1	1	1	1	0	1	1	0	0	0	1	0	0	
32	1	1	1	1	1	0	1	1	0	0	0	1	0	0	

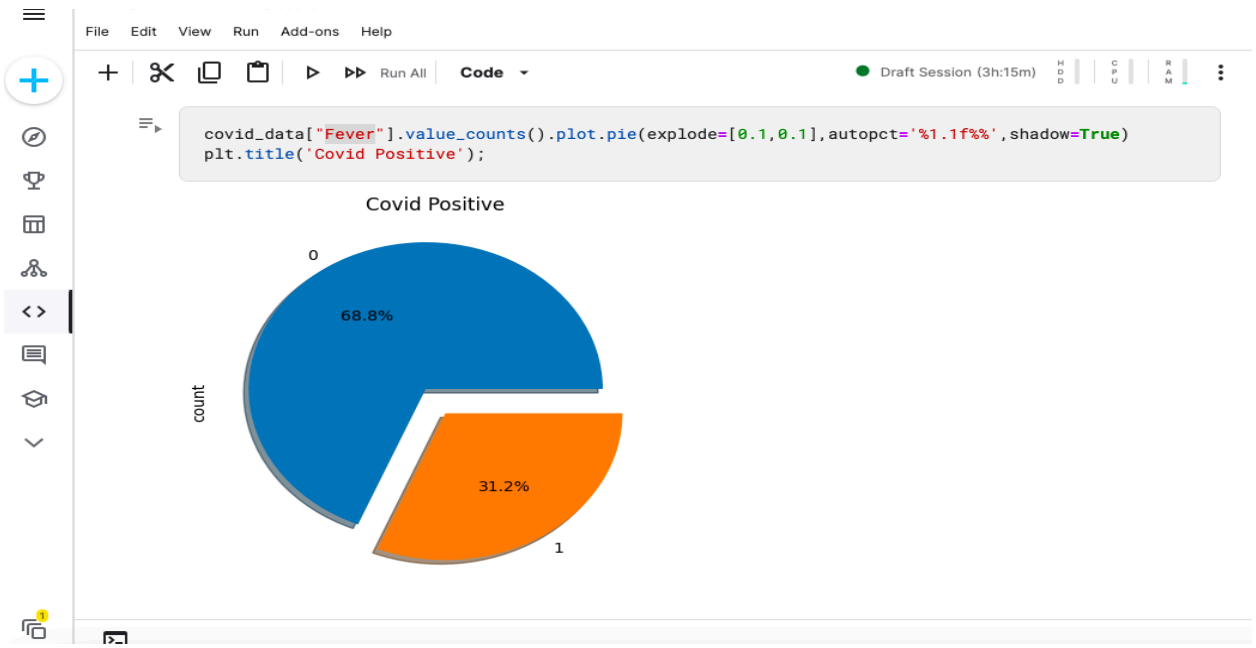
Splitting the Dataset

The next stage in machine learning data preprocessing is to split the dataset. A machine learning model's dataset should be split into two parts: training and testing. We divided the data into an 80:20 split. This means that we use 80% of the data to train the model while keeping the remaining 20% for testing. We take all the 20 independent Fig. 1. No. of missing values and missing percentage of all the attributes attributes into x and the dependent column 'COVID-19' into y as we aim to predict if the patient is COVID positive or not.



Hyperparameter tuning by grid search CV

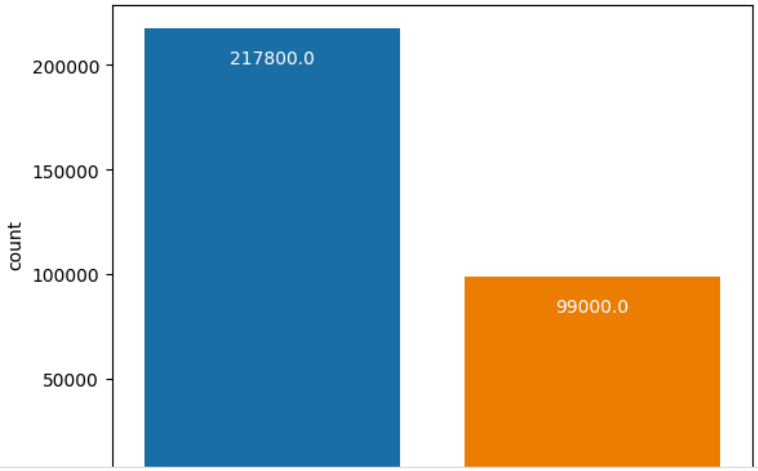
Its main goal is to discover the optimal parameters where the model's efficiency is the best or highest and the error rate is the minimum. We have used the gridsearchcv tool to produce the best combination of parameters, based on accuracy score as the scoring metric when all the different parameters are fed into the parameter grid.



RESULT AND CONCLUSION

To evaluate the effectiveness of the Machine Learning algorithms applied in this experiment, we decided to adopt the Accuracy, Mean squared error, Precision, Recall and F-Measure which are widely used in domains such as information retrieval, machine learning and other domains that involve binary classification.

```
[45]: ax = sns.countplot(x='Fever',data=covid_data)
      for p in ax.patches:
           ax.annotate(f'\n{p.get_height()}', (p.get_x()+0.4, p.get_height()+100), ha='center', va='top', c
      plt.show()
```



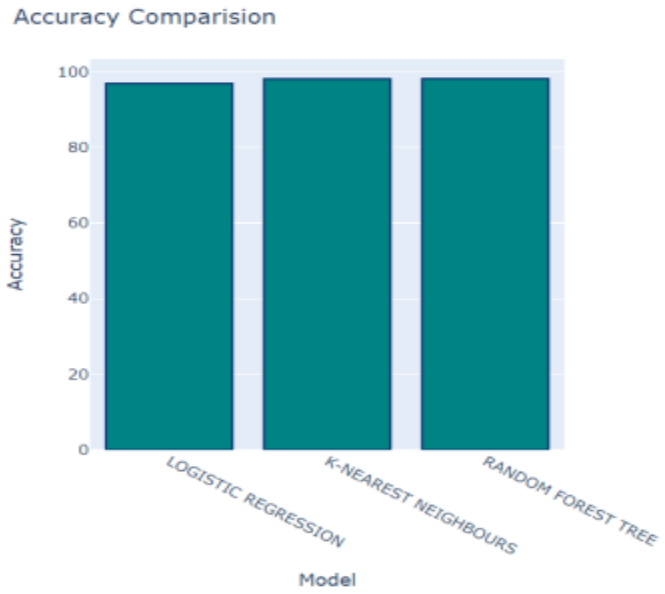


Fig. 7. Analysis of algorithms by their accuracy

LOGISTIC REGRESSION:

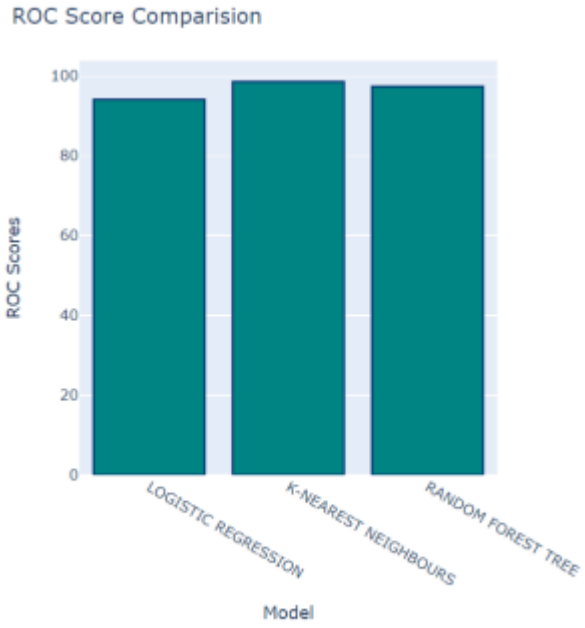
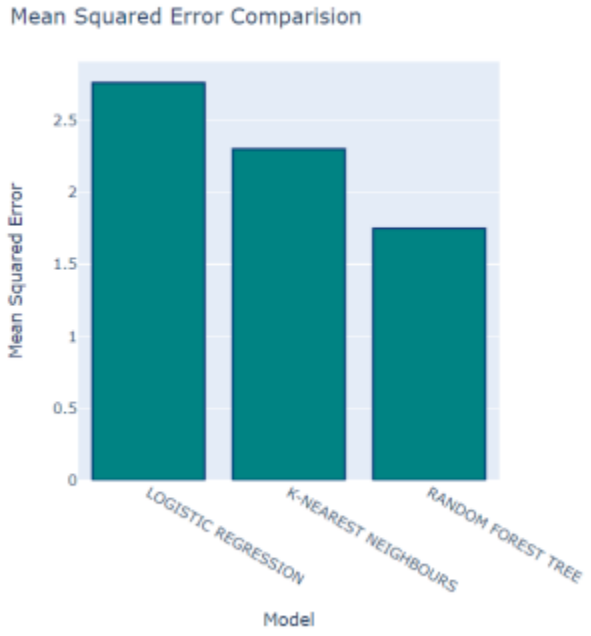


Fig. 9. Analysis of algorithms with ROC score

ROC:



	Accuracy	MSE	R2 score	ROC score	Running time
KNN	98.37%	2.57	83.1	98.58	24.252
Logistic Regression	97.03%	3.036	80.086	93.23	0.038
Random Forest	98.39%	2.207	85.51	97.41	213.331

TABLE I
COMPARISON OF METRICS FOR KNN, LOGISTIC REGRESSION AND RANDOM FOREST

```
COVID PREDICTION BASED ON ML ALGORITHMS
Enter 1 for Yes and 0 for No
Does the patient have breathing problem ? 1
Does the patient have fever ? 1
Does the patient have dry cough ? 1
Does the patient have sore throat ? 0
Does the patient have running nose ? 1
Does the patient have any record of asthma ? 0
Does the patient have any records of chronic lung disease ? 0
Is the patient having headche ? 0
Does the patient have any record of any heart disease ? 0
Does the patient have diabetes ? 1
Does the patient have hyper tension ? 1
Does the patient experience fatigue ? 1
Does the patient have any gastrointestinal disorders ? 0
Has the patient travelled abroad recently ? 0
Was the patient in contact with a covid patient recently ? 0
Did the patient attend any large gathering event recently ? 1
Did the patient visit any public exposed places recently ? 1
Does the patient have any family member working in public exposed places ? 0

Results :  [1]
You may be affected with COVID-19 virus! Please get RTPCR test ASAP and stay in Quarantine for 14 days!
```

The goal of this work was to use the three supervised machine learning techniques to create a COVID-19 presence predicting model. The model's performance was evaluated Accuracy MSE R2 score ROC score Running time KNN 98.37% 2.57 83.1 98.58 24.252 Logistic Regression 97.03% 3.036 80.086 93.23 0.038 Random Forest 98.39% 2.207 85.51 97.41 213.331 TABLE I COMPARISON OF METRICS FOR KNN, LOGISTIC REGRESSION AND RANDOM FOREST Fig. 12. Prediction model takes input from the user and gives a result - COVID Negative in a comparative analysis. The results show that the KNN classifier with number of neighbors to be considered equal to 2 is the best machine learning algorithm, having an accuracy of 98.37%, and 0.026 mean absolute error considering the runtime for training. In comparison to other methods, the model takes average time but gives good accuracy. This research can be used as a supporting tool for decisionmaking by doctors, with the established model assisting in recognising COVID-19 presence in a person based on their symptoms. Individuals who are suffering COVID-19-related symptoms can also use it to assess if they would be tested positive or negative for COVID-19. The model that has been developed here can be employed to deploy an app with the following features: Fig. 13. Prediction model takes input from the user and gives a result - COVID Positive • Individuals can quickly determine whether they are at risk of transmitting COVID-19 based on their symptoms. • Medical practitioners can employ this test as a primary health assessment for COVID detection. • Assisting businesses in limiting physical interaction with clients who may be infected with COVID-19; Extra information or diagnoses from hospital records, persons who contracted the virus, COVID-19 survivors, patients under assessment, or management can all be included for future research. A software which can predict the severity of COVID-19 can indeed be deployed to provide further information about the steps that must be taken and the interventions that should be considered.

REFERENCES

- Channappanavar R, Perlman S. Pathogenic human coronavirus infections: causes and consequences of cytokine storm and immunopathology. Seminars in immunopathology. 2017 Jul;39(5):529-39.
- International Committee on Taxonomy of Viruses (ICTV). Virus Taxonomy: The Classification and Nomenclature of Viruses The 9th Report of the ICTV (2011) [14 January 2020]. Available from: https://talk.ictvonline.org/ictv-reports/ictv_9th_report/.
- Yin Y, Wunderink RG. MERS, SARS and other coronaviruses as causes of pneumonia. Respirology (Carlton, Vic). 2018 Feb;23(2):130-7.
- World Health Organization (WHO). WHO Statement regarding cluster of pneumonia cases in Wuhan, China 2020 [14 Januray 2020]. Available from: https://www.who.int/china/news/detail/09-01-2020-who-statement-regarding-cluster-of-pneumonia-cases-in-wuhan-china
- World Health Organization (WHO). Novel Coronavirus – China 2020 [14 January 2020] Available from:

  https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/