**Covid-19 Detection by Symptoms Using Machine Learning**

Md. Kamrul Hasan

Department of Computer Science & Engineering

East West University

Dhaka, Bangladesh

[2017-1-60-150@std.ewubd.edu](mailto:2017-1-60-150@std.ewubd.edu)

Nasim Bahadur

Department of Computer Science & Engineering

East West University

Dhaka, Bangladesh

[2018-1-60-073@std.ewubd.edu](mailto:2018-1-60-073@std.ewubd.edu)

***Abstract:* The covid-19 pandemic has created havoc around the world. Millions of people**

**have been infected and millions are at risk. It is very tough for health organization and government to test each and every citizen for covid test. However, there are many symptoms of this virus. We can use machine learning to detect covid-19 virus. In this paper, we have implemented four different machine learning algorithms to detect covid-19 accurately. We have chosen the best algorithm, based on the accuracy.**

**First, we have collected a dataset of the symptoms of covid-19. Then we have processed the data and implemented the algorithms. We have implemented Decision tree, Naïve Bayes, Neural network and Random forest. By performing these machine learning algorithms in our dataset, we have seen the accuracy of neural network is the highest with 98% accuracy. However, there are many aspects considered.**

***Keywords: Covid-19, machine learning algorithm, naïve Bayes, decision tree, random forest, accuracy.***

1. **Introduction**

Covid-19 has created a global pandemic around the world. It has stopped the whole world. Almost 3.4 million people have died from corona virus or covid-19[1]. Daily new cases increasing. Till date almost 16 crore people got effected from it [2]. So, it is really tough for any government or country to take test for covid-19 for all the population one country has.

Sometimes one country might capable to take test for covid-19 but it is too much time consuming. So, we have to rely on technology for it. As covid virus has some certain symptoms that is sign of corona virus [3]. In this project we have tried to implemented machine learning to determine the covid-19 corona virus via symptoms. There are many machine learning

algorithms. We have tried four algorithms. Which is Neural Network, Naïve Bayes, Decision tree, Random forest.

**a. Objective**

Our objective is to reduce the testing time. So, that prevention speed will be faster. Also, the accuracy of the test is also an objective so the chances of error is lesser. Here, fast testing is the main priority. A patient or person can do his/her test automatically using device via the symptoms. So, that person don’t have to go to hospital for test because it is also increasing the chances of exposing the person to covid-19.

**b. Motivation**

As millions of people affected by this virus and medical facility dealing with effected patients. It is getting tougher to get manual test. Which also increases the chances of getting affected. So, the motivation behind this project is to test at home by own,

-To reduce testing time

-To determine the affected rate and give medical support quickly.

-To minimize the testing total time, spend on collecting data one by one.

-To give faster and accurate results.

1. **Data Preprocessing**

We trained and tested our method by Symptoms and COVID Presence dataset. This dataset is publicly available [3]. There is no way to unacknowledged that a good quality dataset provides a good prediction and generalize the ability of our model. However, the dataset we choose for our experiment may not be used directly in our model because there is a lot of noise such as absence of data and unsuitable format of the dataset. Hence, preprocessing of data is essential to process the original data. Data processing and data cleaning both we have done in our project. First of all, in our dataset, it is checked that is there any column or row where exits null. If it is found, we will remove the whole row. Making suitable data for ML classification, we have converted data of each column and (Yes, No) converted into (1, 0). After that we find out the correlation between each column with the decision factor column/class. We got two columns (Sanitization from Market and Wearing Masks) which have no correlation with decision factor class and that is why, we have dropped these two columns from our existing dataset. And we rename our decision factor class, COVID-19 as COVID.

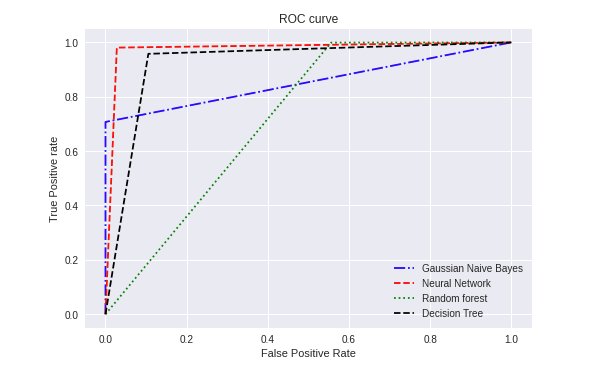
1. **Methodology**

Besides, we separate our main dataset with several proportions such as training dataset and testing dataset. We allocate 80% data of dataset for training and 20% data of main dataset for testing our ML classification model. During this project we have taken four popular classification model like Decision Tree, Naïve Bayes, Random Forest and Neural Network. After trained our model with training dataset then we provide test data to our classification models for getting our prediction value. Then we compare returned prediction value with our actual output. And we measure and print the accuracy score, f1 score, precision score and recall score of our ML classification models. For dataset classification here, we use some free python library like pandas, numpy, sklearn, scipy, matplotlib, graphviz, pydotplus and so on. Sklearn library contains features like classification, regression, clustering and various algorithm. We have used sklearn classification library.

1. **Experimental Result**

A compatible tool for performance assessment of classification problem is confusion matrix which consists of 4 parts:

1. True -Positive (TP): The amount of positive samples predicted to be positive.
2. False- Positive (FP): The amount of negative samples predicted to be positive.
3. True -Negative (TN): The amount of negative samples predicted to be negative.
4. False- Negative (FN): The amount of positive samples predicted to be negative.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method** | **Different Metrics** | | | |
| **Accuracy** | **Precision** | **Recall** | **F1-score** |
| **Neural Network** | 98% | 99% | 98% | 98% |
| **Decision Tree** | 96% | 97% | 98% | 97% |
| **Random Forest** | 88% | 87% | 100% | 93% |
| **Naïve Bayes** | 75% | 100% | 69% | 82% |

1. **Challenges**

**Dataset:** Getting the perfect dataset is a challenge.

**Cleaning Process:** The dataset processing is a tough job. As many datasets don’t have the expected value we want. Some algorithms require numeric values. So, it is necessary to clean datasets.

**Deciding Factors:** As there are many machine learning algorithms, it is tough job to choose one of them. It is tough to say which one will give accurate result before implementation.

1. **Limitations**

There are certain limitations in our project the main limitations we have is the dataset. As we do not have a big dataset. We cannot perform a big operation such as the spread of covid in a city. Another limitation is, there are several algorithms in machine learning where we can process our data. But we have implemented some of that. Where there is more opportunity to choose more algorithm, which can be more efficient.

1. **Future work:**

To enhance the accuracy of our prediction, we have to analyze more with COVID Symptoms dataset and others ML classification model. In future work, we will try various model which is designed for classification and besides, we will try to train and test our models for other COVID Symptoms dataset.

1. **Conclusion**

We have implemented our datasets in four different machine learning algorithms. We have seen tremendous result in some algorithm where the accuracy is close to perfection. The four algorithms we used are Gaussian Naïve Byes, Random Forest, Neural Network, Decision tree. From the implementation we have seen that Neural network has the best precision rate with 99.0%. Decision tree has the second best with a precision rate of 97.0%. Naïve Byes precision was 100% but as its accuracy level is low with only 75%, so we cannot rely on it. The accuracy percentage of Neural Network is highest amongst all. With an accuracy level of 98%.

The close behind was decision tree with 96%. So, we can say that Neural Network is the suitable amongst all. But there is a problem choosing this algorithm which is, it is best when it comes to big dataset. It works fine with the big dataset. Other than that, the accuracy level is best in Neural Network. Decision tree is better when the dataset is small compare to the dataset of Neural Network. But the accuracy of Decision tree is less than Neural Network.

So, if we want to consider the best algorithm for covid-19 detection we can say that Neural Network can detect more precisely and accurate result. Which can prevent the spread of covid-19 as people will test according to their symptoms and it will give the results. Neural network can perform better result for big datasets so it is much more suitable for a large number of populations which is a very good thing. So, we have in this paper we have not only implemented machine learning for covid detection but also discovered which machine learning algorithm is best for this type of dataset. Although there are many algorithms to implemented which might be come with more accuracy and suitability.

**Reference**

[1] Worldometer, ‘COVID Reports’. https://www.worldometers.info/coronavirus/ (accessed May 24, 2021).

[2] CDC, ‘Corona Symptoms’. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html (accessed May 24, 2021).

[3] Kaggle, ‘Dataset’. https://www.kaggle.com/hemanthhari/symptoms-and-covid-presence?fbclid=IwAR1VW3jnpkNaMeiIeqnaKqwgY6L8iPcBljB3VcP2dh5TNJsqh2HcMs-Qjqg (accessed May 23, 2021).