12-15-2022

COVID-19 PREDICTION USING MACHINE LEARNING

Parashuram Singaraveni
Grand Valley State University

COVID-19 PREDICTION USING MACHINE LEARNING

Parashuram Singaraveni

A Project Submitted to

GRAND VALLEY STATE UNIVERSITY

In

Partial Fulfilment of the Requirements

For the Degree of

Master of Science in Applied Computer Science

School of Computing and Information Systems

December 2022

The signatures of the individuals below indicate that they have read and approved the project of Parashuram Singaraveni in partial fulfilment of the requirements for the degree of Master of Science in Applied Computer Science.

Dr. Robert Adams, Project Advisor	Date 12-08-2022
Dr. Robert Adams, Graduate Program Director	Date 12-08-2022
Dr. Paul Leidig, Unit head	Date 12-08-2022

ABSTRACT:

All around the globe, humankind faces a disastrous situation that witnessed COVID-19 outbreak. The COVID-19 pandemic caused severe loss of human life across the world. Most of the countries had been socially and economically weakened. The health sector faced lots of challenges in diagnosing the COVID patients, vaccinating the people, identifying the people who are infected by the virus. At the earlier stage, it has been difficult to identify the symptoms in infected person that is caused by the virus. Months later, symptoms were identified and, disease detecting machines were invented. But still, time taking for the results from the machines were able to slowly process the results. In my project, I came up with the application that predicts whether patients are infected by the virus or not.

INTRODUCTION:

COVID-19 which is commonly known as the Corona Virus, has been a hot topic ever since it was first reported in December 2019, leaving millions of people lives at risk. COVID-19 started spreading across the globe, which in turn killed more than six and a half million people. The virus can spread via close contact or through respiratory droplets. Due to the deadly nature of the virus, a pandemic has been announced by the World Health Organization. There were cases where people showed symptoms and some people are asymptomatic.

This project is focused on predicting whether a patient have COVID or not based on the symptoms the patient possess. With the help of the technological resources available we can gather information of the symptoms of the COVID patients and use the data obtained to train a machine learning algorithm in the form of a decision tree model in order to predict whether a person is affected with COVID-19 without taking any kind of test. In my project, I have collected the dataset of symptoms of the patients along with results and trained the dataset.

The project was built based on Decision Tree Classifier and coded in Python using Spyder IDE and Streamlit framework. The Decision Tree model was processed in Colab using the trained dataset and downloaded the resulting model file. The Streamlit framework was used to build a web application in which entries of the symptoms are entered manually to predict COVID, then using the classifier model the result is shown.

BACKGROUND RELATED WORK

Machine learning is a branch of artificial intelligence and computer science which uses data to learn and adapt that imitates human thinking without any explicit instructions. It evaluates the patterns in the collected data and helps in decision making. To understand ML by considering a real time example like getting shopping suggestions while using social media platforms, here ML plays a role to get automated suggestion based on what we were interested in. If we go little deeper, ML takes our previous shopping list and browsing history without our knowledge and that provides automated suggestions. Machine Learning is more like doing research on real time data that predicts the results, but there is no assurance of getting accurate results. Moreover, Machine Learning process was difficult in data acquisition, model selections, time and space consuming.

METHODS:

Obtaining the dataset was a crucial task in this process. The data was collected from the online website Kaggle and further evaluated by clearing unwanted data. Then the model was further trained to obtain the classifier. Many techniques were applied on the dataset to get the understanding of the variations in data. Graphical observations were performed on the dataset which showed how the data varied among themselves. By using the built-in library in Python, the required operational techniques were applied to the dataset file. Ultimately a Decision Tree Classifier was used by importing from sklearn.tree and the model saved.

After the model was created, the next part of the project was started that is building a web application using the Streamlit framework. This process starts with downloading the Anaconda application and importing the required libraries using PyPi commands for Streamlit. The configurations of the applications are made and the web application is built which is coded in Python. The web application is equipped with the fields of the symptoms where inputs have to be given and the output for the entries are displayed whether the person having COVID or not.

Google Colab:

Google Colab offers an open-source platform to write and execute the Python code. In my project Google Colab was used to create and verify the decision tree model file and to generate the saved file.

Streamlit:

Streamlit is an open-source application framework in which python programming is used, which is used to create the web application for data science and machine learning related projects. In this project, the Streamlit framework was used to build an interactive web application by importing the ML model saved from Google Colab.

TRAINING DATASET:

The dataset in this project was obtained from the Kaggle website, a data science company (kaggle.com). It consists of the data with symptoms and the results of the patients. The dataset was divided into training(80%) and testing data(20%) where the training dataset was used for training and to create the model. After creating the classifier, the test dataset was used to verify accuracy of the trained model.

PREPROCESSING:

The dataset file which is in CSV format have to pre-processed before the creation of the classifier. The file undergoes stages like identifying the missing values and clearing those values. Then further the variance in the dataset is identified using in-built functions.

CLASSIFIER:

In this project sklearn was used which is built-in library in Python that is used for machine learning. It provides various classifiers like Decision Tree Classifier, Random Forest Classifier etc. that helps in creating the model and various tools like StandardScaler, train_test_split, accuracy_score that helps in training and testing the dataset. Random Forest Classifier, SVM Classifier, Linear Regression and Decision Tree Classifier were used. Decision Tree Classifier's model have the highest accuracy score compared to other classifiers.

Decision Tree Classifier which is an algorithm in Machine Learning has been used for predicting the output. Based on the study of the dataset, the classifier behaves as human intelligence in predicting the output. It was the best algorithm in predicting the output based on the data provided. Finally, the classifier is saved into SAV file.

Decision Tree:

The Decision Tree is a supervised learning technique which is mostly used for classification problems. This algorithm behaves like human thinking in decision making. This model helps in predicting the class of the datasets in which algorithm performs on every data in the dataset and come up with result.

ORGANIZATION:

Google Colab is used to perform the machine learning algorithms to train the model. To build the web application, Anaconda application was used as a platform to code in python and to run the streamlit application. Anaconda offers Spyder application which has to be launched. Spyder IDE is mostly used in data science related works which is user-friendly and allows the user to get installed the required libraries, packages and modules with command-line codes (PyPi) in the machine learning terminal. The data fields that appear in the web page are created in the Spyder. Using the Streamlit terminal, run the python file and the functional web page is created.

DISCUSSION:

Heatmaps gives the graphical representation of the correlation among the dataset. It shows the clear vision on how data are related among themselves and how they differ from themselves. Using the colour pattern, it shows the visualization of the strength of the data. In our dataset, the data is distributed in disproportionate. The levels of the colour are formed from light blue to dark blue. As the level increases the chances of having COVID is increased. In the heatmap, there are many points where the level is low (light blue). Even at some points the level is too high that indicates that the respective fields are related with the result. In the heatmap, on the X and Y axes symptoms are mapped and their proportionality is displayed as colour palettes. The intervals on the graph describes the chances(seriousness) of having COVID. The intersection of respective symptoms and other symptoms shows their proportionality or co-relation among them.

Heatmap

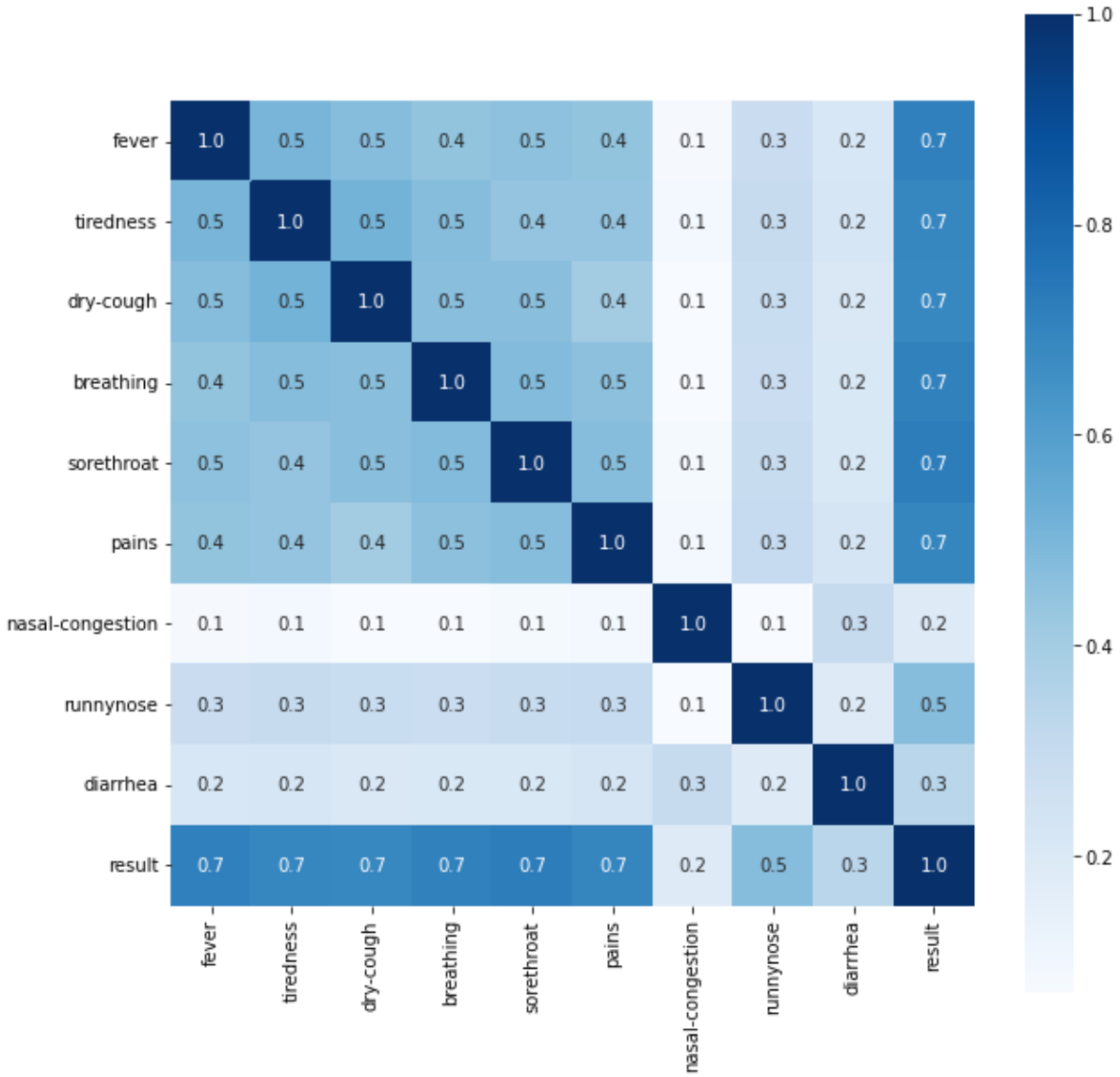

Fig-1

In Fig-1, symptoms like fever, tiredness, dry-cough, breathing, sore-throat, pains have close relation with result having 0.7. With respective to the fever, it is co-related with tiredness and dry-cough having 0.5 and it is less co-related with symptoms like runny nose, diarrhoea, nasal-congestion and pains having 0.3, 0.2, 0.1 and 0.4 respectively.

Graph:

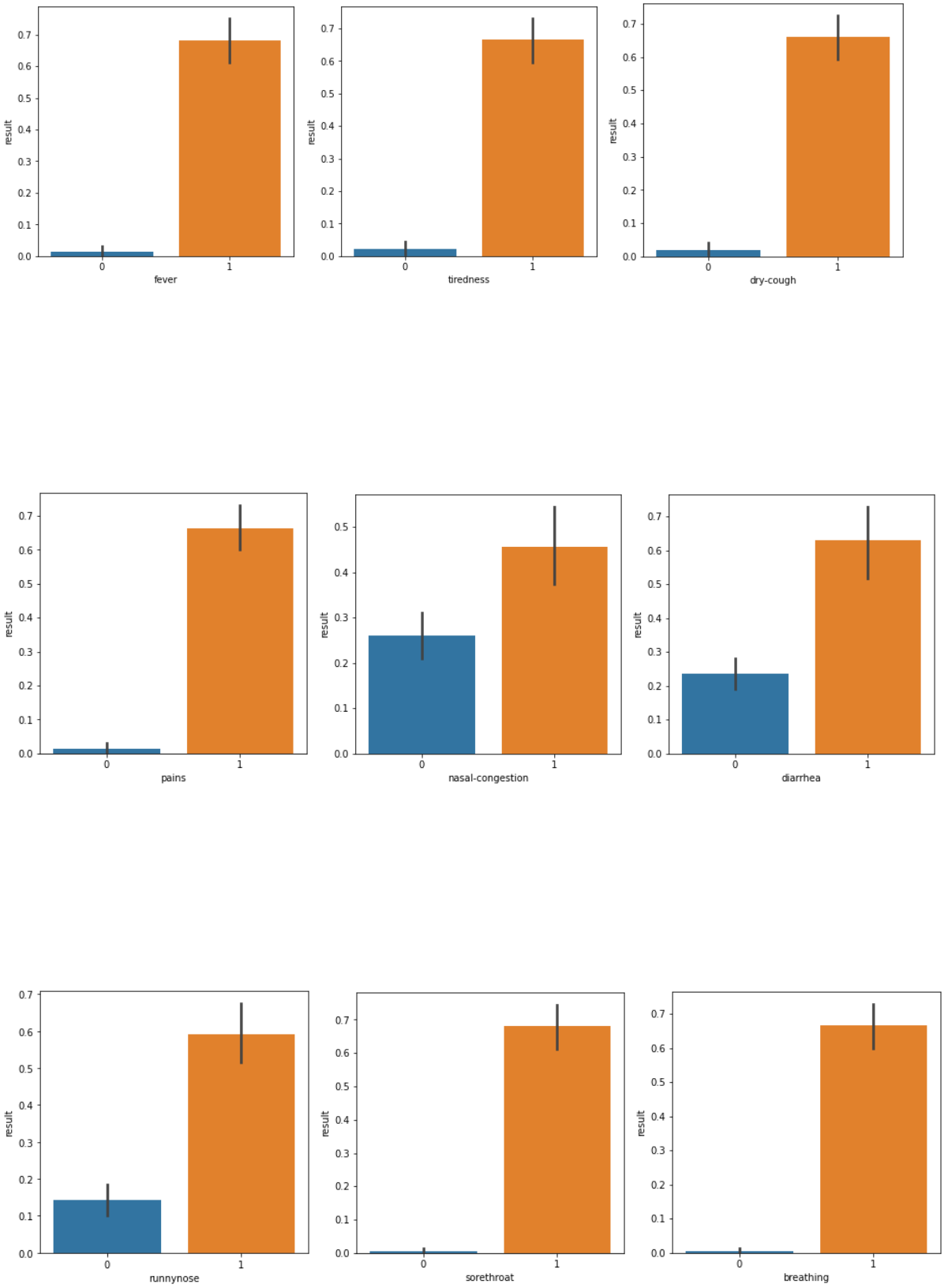
result
fever
tiredness
dry-cough
pains
nasal-congestion
diarrhea
runnynose
sorethroat
breathing

Fig-2

In the above plots, the symptoms are plotted with result data. Plots gives the proportionality of the symptoms along with the result, that is, on how they vary among themselves.

- 0 (blue bar) represents chance to be negative w.r.t symptom (no COVID)
- 1(orange bar) represents chance to be positive w.r.t symptom (having COVID)

The Decision Tree model have accuracy score of 85%. It predicted accurate results from the test data.

CONCLUSION:

The purpose of this project was to identify the presence of COVID in the patients by only looking at their symptoms using machine learning. The application is user-friendly and can be used by anyone. The project had resulted in the ability to predict the result accurately using the real time dataset. Results from the application allows the user to make more informed decision on whether to take a real COVID test.