Predicting COVID-19 Infection

1 Introduction

Since 2019, we have been experiencing the Covid-19 pandemic. The cause of this pandemic has had a significant impact on the lives of citizens as well as the global economy. One of the most important tasks in this pandemic has been determining whether a person is infected with Severe Acute Respiratory Syndrome coronavirus 2. (SARS-CoV-2).

To accomplish this, we are using a dataset that records all possible symptoms, physiological, and geographical parameters to predict infections.

The primary goal is to identify the best model for predicting COVID-19 infection in a person.

2 Dataset

The dataset used in the project is a covid-19 symptom dataset.(HUNGUND, 2019) It is a dataset used to detect the infection of Covid-19 in an individual. The dataset was created based on the WHO guidelines and Ministry of Health and Family Welfare, India. (WHO, 2019)(MoHFW, 2019)

2.1 Dataset Description

The shape of the dataset is 316800x27.

2.2 Samples

The dataset consists of 316800 records.

2.3 Measurements

Measurements: The dataset consists of 27 features, out of which 23 features describe the symptoms, age, gender, contact and country. The remaining four features describe the severity of infection.

3 Task

There are two distinct tasks found for this dataset. Task-1 involves making predictions to determine whether or not the individual is infected with the Covid-19 virus.

In Task-2, predictions are made to see what level of infection in a person or whether they have Covid-19.

3.1 Task-1

3.1.1 Data Cleaning

Data cleaning for Task-1 involves mapping target values, which is the process of reducing a four-class problem to a two-class problem. For this, the severity_severe, severity_mild, and severity_moderate, each of which has 3, 2, and 1 values, should be mapped to just one value, which is 1. Following this cleaning, the dataset's target column would only contain the numbers 0 and 1.

3.1.2 Model

The model we have used for Task-1 is decision tree classifier with three hyperparameters which are max_depth, min_samples_split and random_state that provides 75.17 percent accuracy for the task. The values of the above mentioned hyperparameters are obtained through grid search cv on the dataset.

3.2 Task-2

3.2.1 Data Cleaning

For task-2, initial dataset is considered where the target column has four values, 3, 2, 1, and 0, which, respectively, indicate severity_severe, severity_mild, severity_moderate, and None.

For task-2 make_blobs function from sklearn package is used on the dataset. This function gives blobs of points by using Gaussian distribution. There are different parameters that can be adjusted, including the number of samples and blobs that are generated. The make_blobs function is used in task-2 with the following three parameters: n_samples, n_features, centres, and random_state.

To fit the model, the output from make_blobs is considered.

3.2.2 Model

The model we have used for Task-2 is decision tree classifier with three hyperparameters which are max_depth, min_samples_split and random_state that provides 93.8 percent accuracy for the task. The values of the above mentioned hyperparameters are obtained through grid search cv on the dataset.

4 Conclusion and Future Scope

For both the tasks, Decision tree classifier is providing highest accuracy.
The work on this dataset can be extended to a web application which can be developed using the model that aids in predicting whether or not a person has Covid-19 infection by gathering the data needed for the prediction.

References

BILAL HUNGUND. 2019. Covid-19 symptom checker dataset. Kaggle, Dataset.

MoHFW. 2019. Ministry of health and family welfare india. MoHFW, MoHFW-Guidelines.

WHO. 2019. World health organisation guidelines on the symptoms of covid-19. WHO, WHO guidelines.