COVID-19 predictor

Abstract

Coronavirus disease (COVID-2019) is a dangerous growing quickly pandemic that is spreading quickly across the world. The Kingdom of Saudi Arabia (KSA) registered the first case of COVID-19 on 2 Mar 2020. Since that, the number of infections has been increasing daily. The World Health Organization (WHO) reported 248 million cases with more than 5 million deaths worldwide in. Nov 2021. The KSA has taken several measures to control the spread of COVID-19, including imposing a curfew on the cities of the Kingdom stopping Umrah and performing Hajj in reduced numbers from within the Kingdom, that had an effect in limiting the spread of the virus. Recently, many types of research and studies concerning the impact of (Covid-19) in all respects especially on health and economy. We propose to generate a more accurate diagnosis model of (COVID-19) based on patient symptoms by applying Machine Learning, for a supervised learning task to analyze the data. We aim to generate a classification model for (COVID-19) dataset to predict whether a patient will be infected or not. Also, what is the common reasons to get affected with Covid-19?

Design:

Data were collected from India on year 2020. Classifying accurately via machine learning models will help the person know if he infected or not, then he\she can isolate himself to prevent the disease from spreading in the community COVID is a contagious disease.

Data:

The dataset contains 5434 rows × 21 columns[Breathing Problem, Fever, Dry Cough, Sore throat, Running Nose, Asthma , Chronic Lung Disease, Headache, Heart Disease, Diabetes, Hyper Tension, Fatigue, Gastrointestinal, Abroad travel, Contact with COVID Patient, Attended Large Gathering, Visited Public Exposed Places, Family working in Public Exposed Places, Wearing Masks, Sanitization from Market, COVID-19][1] the most important elements by which we can determine whether a person will be infected or not is COVID-19 is the target and author is the features some of this features is symptoms like[Breathing Problem, Fever, Dry Cough, Sore throat, Running Nose] and some of it about diseases in humans such as[Asthma , Chronic Lung Disease, Headache, Heart Disease, Diabetes, Hyper Tension, Fatigue, Gastrointestinal] and the third group of features about the activities and actions of the person [Abroad travel, Contact with COVID Patient, Attended Large

Gathering, Visited Public Exposed Places, Family working in Public Exposed Places, Wearing Masks, Sanitization from Market]. All of the features and target contain two values Yes or No.

Algorithms:

Feature Engineering

- **data analysis** our data shape is (5434, 21) and the Data type is object and by looking at the column names in our dataset, we find the trailing whitespace problem so We can remove this by calling map on the columns list and stripping the whitespace with strip.
- **finding missing value** there is no missing value in our dataset**.**
- **feature transformation** we convert the features into integer by using LabelEncoder to converting string labels into numbers No = 0 and Yes = 1.
- **data visualization** we can see our data set imbalance because there is a big different between column [No, 1051] and column [Yes, 4383] so our dataset not very huge using over-sampling it more efficient than the under-sampling.
- **correlation between features** the most important symptoms and signs that refers the person has affected with Covid-19? as appear in the Correlation is Breathing Problem, Fever, Dry Cough, Sore throat. The Common reasons to get affected with Covid-19 is Contact with COVID Patient, Abroad travel and Attended Large Gathering.
- **feature selection** feature that we going to delete:
  Running Nose / Asthma /Chronic Lung Disease / Headache / Heart Disease / Diabetes / Fatigue / Gastrointestinal / Wearing Masks / Sanitization from Market

Models

Logistic regression, k-nearest neighbors, and Decision Tree classifiers were used before settling on Decision Tree as the model with strongest cross-validation performance.

Model Evaluation and SelectionThe entire training dataset was split into 80/20 train.

Decision Tree 10-fold CV scores: 96.7%

Tools:

- The main tools Jupyter Notebook for writ code
- Pandas (pandas.read_csv) to Read a comma-separated values (csv) file into DataFrame.
- import matplotlib.pyplot as plt
- import seaborn as sns
- sklearn.preprocessing to import LabelEncoder
- sklearn.model_selection to import train_test_split
- sklearn.metrics to import accuracy_score
- sklearn for: machine learning models, cross_val_score, metrics and preprocessing.
- imblearn.over_sampling to import SMOTE and RandomOverSampler.
- Collections to import Counter

Communication:

Presentation.

[1] https://covid19.who.int/table

[2] https://www.kaggle.com/hemanthhari/symptoms-and-covid-presence

Name: Wafa Hamdan Alshehri

Email: wafa.h457@gmail.com