Covid-19 Zone Predictor

- Samruddha Patil

26/05/2020

1. Introduction

1.1 Background

The first case of the COVID-19 pandemic in India was reported on 30 January 2020, originating from China. As of 26 May 2020, the Ministry of Health and Family Welfare have confirmed a total of 145,380 cases, 60,491 recoveries (including 1 migration) and 4,167 deaths in the country. India currently has the fourth largest number of confirmed cases in Asia with number of cases breaching the 100,000 mark on 19 May 2020.
The Government Divided the entire nation into three zones – Green Zone, Red Zone, Orange Zone, relaxation will be allowed accordingly.
Red zone (Hotspots) – districts with high doubling rate and high number of active case**.**
Orange zone (Non-hotspots) – districts with fewer cases
Green zone – districts without confirmed cases or without new cases in last 21 days

1.2 Problem

Data that might contribute to determining zone of location might include his latitude and longitude that describe its position. This project aims to predict whether the location is in red zone,orange zone or green zone based on this data.

1.3Interest

Indian people will be interested in this project as everyone should know in which zone they are or in which zone there relatives are. Volunteers working for reduction of this pandemic should know in which area they should put more efforts and which zone should protect from deadly virus.

2. Data Acquisition and Cleaning

2.1 Data Sources

First different location were selected from the table in the website https://www.latlong.net/category/districts-102-16.html This website has 7 different pages with 100 different locations with their coordinates.
For getting the information about zones of these locations https://www.indiatoday.in/india/story/red-orange-green-zones-full-current-update-list-districts-states-india-coronavirus-1673358-2020-05-01 this website was useful.

2.2 Data Cleaning

Data scraping from 7 webpages of the website resulted in seven dataframes of location with their geographical coordinates. These 7 dataframes merged into 1 dataframe. The column of zones was added into this dataframe. Then manually zones of these locations were added to dataframe.
One more problem was there that names of the locations have country India in it. It is useless to have India in Indian locations. So, last 7 letters of names including 5 letters of India and a space and comma were removed. Thus we got the final dataframe containing name, place name, latitude and longitude.

2.3 Feature Selection

After data cleaning dataframe was consist of 684 locations with name latitude and longitude. This dataframe was split into X as a train data and y as a test data. This coordinates were plotted into a map using folium library.

3. Predictive Modelling

Algorithms required for this modelling was classification algorithms like K-nearest Neighbours, Decision Tree, Support Vector Machine and Logistic Regression. We don't know which algorithm will give best result so it is better to evaluate the algorithms using indices like jaccard index, F-1 score and log loss. As log loss can't be used for evaluation as it can only be used in the case of logistic regression.

After evaluation of various algorithms we get

Algorithm	Jaccard Index	F-1 score	Other Info
K-nearest Neighbours	0.761829	0.773723	K = 1
Decision Tree	0.794003	0.802920	Depth = 12
Support Vector Machine	0.630811	0.678832	
Logistic Regression	0.420313	0.547445	

4. Conclusion

We can clearly say that Decision Tree Algorithm with max depth 12 is the best algorithm for Covid-19 Zone Predictor.