Applying Data Science Methods to Analyze Covid-19 Data and Impact of Vaccination

Project Report for IITB DS203 Programming for Data Science

Alok Panigrahi
Dept. of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, Maharashtra
200100019@iitb.ac.in

Kushal Pawaskar
Dept. of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, Maharashtra
20d070059@iitb.ac.in

Yuvraj Singh Tanwar
Dept. of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, Maharashtra
20d100030@iitb.ac.in

***Abstract*—People all over the world have been suffering due to the pandemic-level outbreak caused by COVID-19. It has led to great changes in many sectors, including the lifestyle of people. It has also caused economic downfalls. But the world looks forward to get back to normalcy with vaccination numbers reaching the desired value. This study is meant for analyzing the dataset of various COVID-19 related factors, and to draw conclusions on the effect that vaccination has on these factors. The correlation between various attributes of the dataset has been described. Some Machine Learning frameworks have been utilized in predicting the chances of getting infected after complete vaccination. Based on this study, we conclude that, after vaccination the number of cases and number of deaths were found to decrease in several countries. Thus the mortality rate after vaccination is lower compared to that before vaccination, in general.**

I. Introduction

Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus.

Most people infected with the virus experience mild to moderate respiratory illness and recover without requiring special treatment. However, some become seriously ill and require medical attention. Older people and those with underlying medical conditions like cardiovascular disease, diabetes, chronic respiratory disease, or cancer are more likely to develop serious illness. Anyone can get sick with COVID-19 and become seriously ill or die at any age.

The best way to prevent and slow down transmission is to be well informed about the disease and how the virus spreads. Protection from infection can be obtained by staying at least 1 metre apart from others, wearing a properly fitted mask, and washing hands or using an alcohol-based rub frequently.

The virus can spread from an infected person's mouth or nose in small liquid particles when they cough, sneeze, speak, sing or breathe. These particles range from larger respiratory droplets to smaller aerosols. It is important to practice respiratory etiquette, for example by coughing into a flexed elbow, and to stay home and self-isolate until recovery if one feels unwell. [1]

This study aims towards analyzing the various attributes related to COVID-19, for different countries and make some predictions based on them.

II. Prior Work

- Data analysis of Covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods [2].
  In this study, data of COVID-19 between 20/01/2020 and 18/09/2020 for USA, Germany and the globe was obtained from the World Health Organization. Dataset consists of weekly confirmed cases and weekly cumulative cases. The distribution of data was examined and its parameters were obtained according to statistical distributions. Time series prediction model using Machine Learning was proposed to obtain the curve of disease and forecast the epidemic tendency. Linear regression, multilayer perceptron, random forest and support vector machines were used. The performances of the methods were compared according to the RMSE, APE, MAPE metrics and SVM was found to achieve the best trend. The predictions of the study were that the global pandemic would peak at the end of January 2021 and approximately 80 million people would be cumulatively infected.

III. Data and Methodology

The dataset D1 [3] was used in the study. D1 consists of data of various COVID-19 related attributes of 237 locations, on a datewise basis, with date ranging from around February-March,2020 to November 2021. The 237 locations consisted of 220+ countries, 6 continents and some groups based on economic factors. The attributes comprised of:

- location related features
- date
- cases related features
- deaths related features
- reproduction rate
- hospital related features
- tests related features

- positive rate
- vaccination related features
- stringency index
- population related data
- age related features
- GDP,HDI data
- cardiovascular death rate, diabetes prevalence
- smoking related data
- excess mortality related features

But this data consisted of a lot of missing values. Hence the attributes and records corresponding to highly unavailable data were eliminated from the dataset. The modified data consisted of 33 countries with 40 attributes. This dataset was loaded as a dataframe for further analysis. Then this data was further cleaned to take out the leftover missing values which were possible to be replaced by some value based on available data. Also the starting date and ending date for the data of different countries varied. Hence the dataframe was modified to include data only from May 1,2020 to October 1,2021 for each of the 33 countries.

IV. Experiments and Results

A. Exploratory Data Analysis

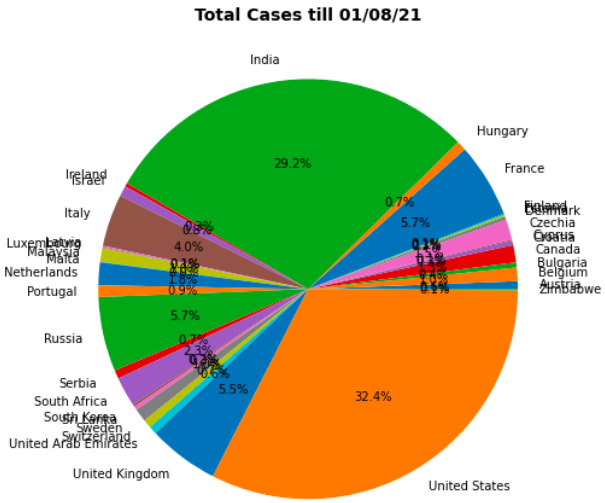
In statistics, exploratory data analysis (EDA) is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task [4]

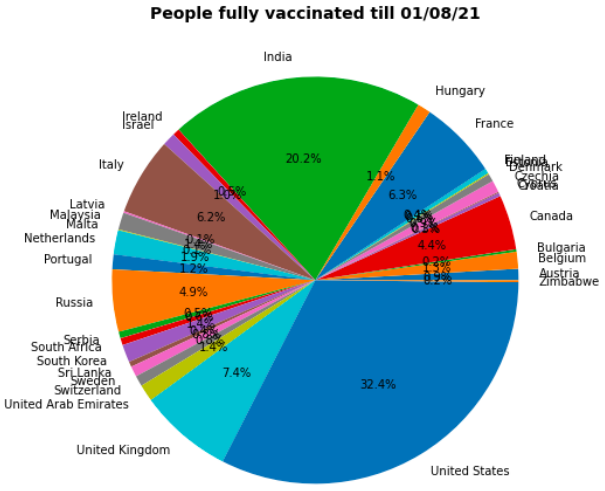
For EDA, we find out:

- Number of records and variables
- Data types of variables
- Number of unique values of each attribute
- Number of missing values of each attribute
  We already dealt with a lot of missing values earlier in data cleaning, by using the technique of replacing the missing values with the average of surrounding available values (in case of numerical values). This was done wherever it was relevant.
- Mode of the discrete variables
- Entropy of discrete variables
  The entropy of the data is low so the information gain is high
- Mean, variance, skew, minimum, maximum, median, 25th percentile, 75th percentile, inter-quartile range of continuous variables

We plot:

- Histogram for discrete variables
  It showed that for almost all the variables, the smaller values are more frequent.
- Histogram for continuous variables
  It showed that there were all three types of variables: the ones with smaller values more frequent, the ones with middle values more frequent, and the ones with larger values more frequent.
- Box-and-Whiskers plots for continuous variables
  These were plotted to get a graphical idea of distribution of values.
- Pie charts
  These were plotted to understand the proportion of contribution of each country to the total value of the particular attribute. For the variables which varied with date, the plots were made on a 3-monthly basis. For the variables where the data was independent of date, only a single plot was required.
  Some pie charts are:



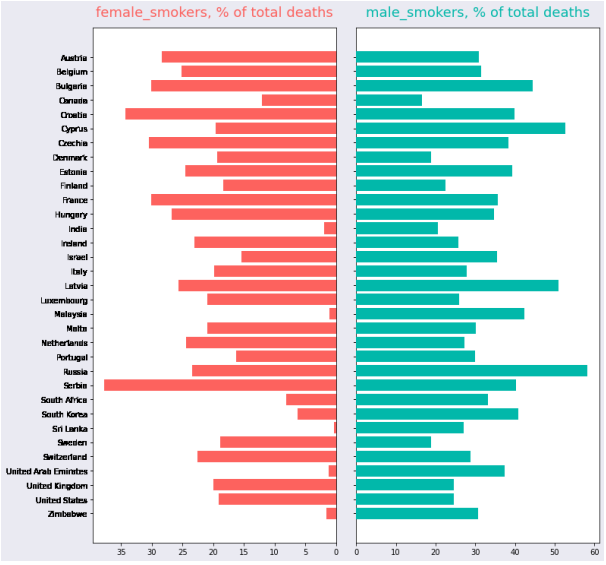


B. Descriptive Analysis

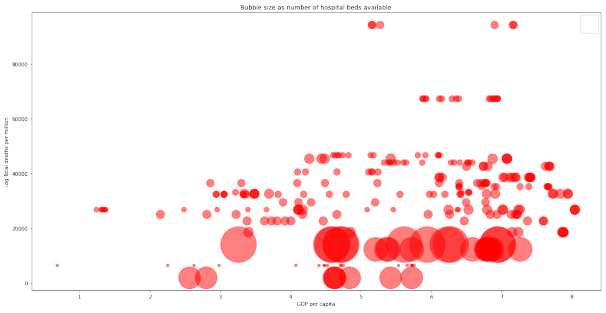
Descriptive Analysis is the type of analysis of data that helps describe, show or summarize data points in a constructive way such that patterns might emerge that fulfill every condition of the data. [5]

In Descriptive Analysis, we tried to find the correlation between pairs of variables.

- Line plot between total cases and GDP per capita
  For most countries with very high GDP, total cases remain low but some countries with high GDP also have high number of total cases.
- Line plot between total cases and median age
  For median age around 28-29, the number of total cases is high, possibly because of higher working force in this age group. Also for higher median ages the total cases are high, possibly because of weaker immunity.
- Line plot between hospital beds and total cases
  Areas with lesser number of hospital beds per thousand might be considered as those having poor medical awareness. Hence, here the total cases are high.
- Double bar graph for percentage of total deaths of male and female smokers

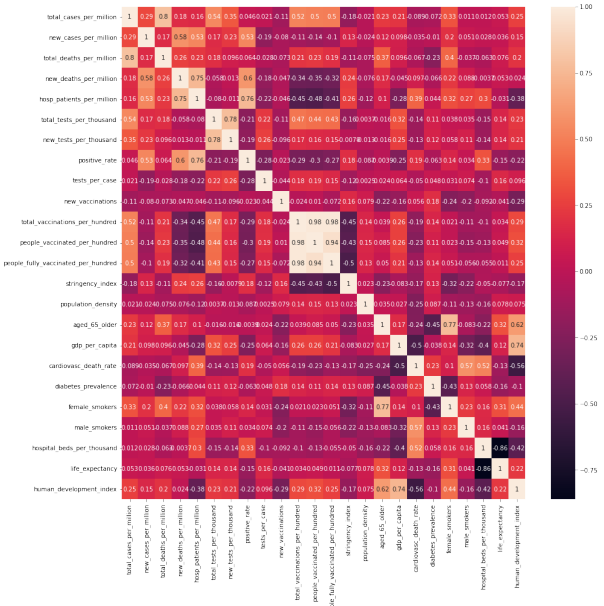


- Bubble plot with GDP per capita, total deaths per million, and hospital beds per thousand denoting the bubble size

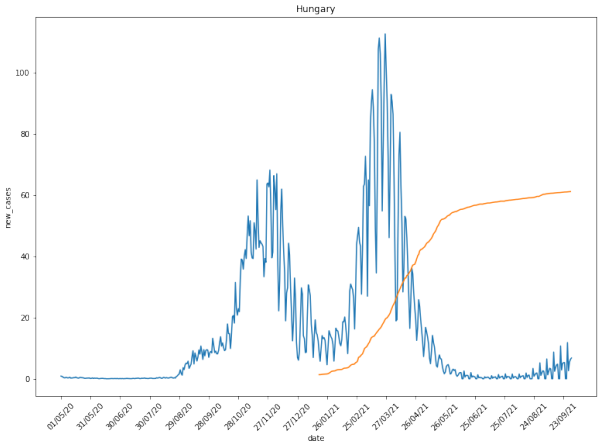


AS the size of bubbles increases, the total deaths decreases. For higher GDP there are more bubbles, i.e., more number of beds.

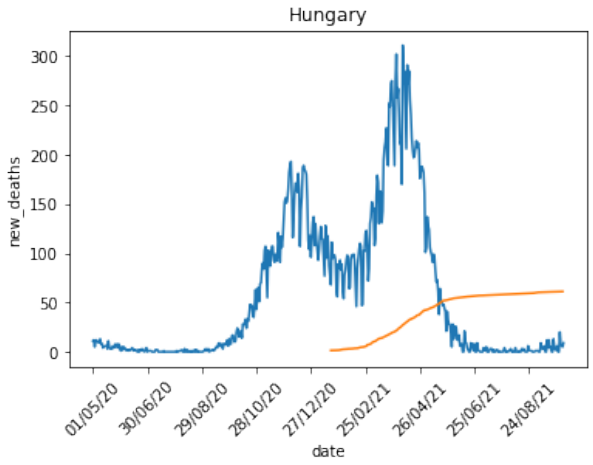
- Correlation heatmap for pairwise comparison of various attributes with new cases
  Only GDP is negatively correlated with new cases. Other attributes used are positively correlated.
- log plot of total deaths vs diabetes prevalence
- log plot of total cases vs total deaths
- log plot of total cases vs cardiovascular death rate
- Heatmap between pairs of variables to show correlation
  The features people_vaccinated_per_hundred, hospital_beds_per_thousand, human_development_index, stringency_index, aged_65_older, male_smokers, female_smokers have a high correlation with new_cases_per_million. Hence these features will be used in ML model for predictive analysis.



- Line plots of new cases and people vaccinated per hundred vs date for each country
- Line plots of new deaths and people vaccinated per hundred vs date for each country
  It is seen that as vaccinations increase, the number of cases and deaths decrease in general. Hence mortality rate is decreasing in general on vaccination.
  Example (Hungary)



(Orange curve is for people vaccinated per hundred)



C. Predictive Analysis

Predictive analysis encompasses a variety of statistical techniques from data mining, predictive modelling, and machine learning that analyze current and historical facts to make predictions about future or otherwise unknown events. [6]

In our case we tried to predict the new covid cases even after people getting vaccinated in a given area or country, if we know the certain factors such as GDP, HDI (Human Development Index), number of smokers, hospital beds per thousand, along with number of people vaccinated. We dropped the other factors since they were highly uncorrelated or redundant for our prediction, since we preferred our data to be normalised, to be fair across different area and its populations.

For predicting this we used following two models:

1)**Decision Tree**: A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Hyperparameter Tuning: Hyperparameter tuning is performed on the dataset with 5-fold cross-validation strategy. Best hyperparameters came out to be max features equal to 2, min sample split 2, and n estimators as 40.

Results: Choosing the best parameters, we get the R2 score as 63.84%.

2)**Random Forest Regressor**: The tree growing in Random Forests happens in parallel which is a key difference between AdaBoost and Random Forests. Random Forests achieve a reduction in overfitting by combining many weak learners that underfit because they only utilize a subset of all training samples.

Hyperparameter Tuning: Hyperparameter tuning is performed on the dataset with 5-fold cross-validation strategy. Best hyperparameters came out to be maximum depth equal to 8, min sample leaf 3, and minimum sample split 3.

Results: Choosing the best parameters, we get the R2 score as 78.09%

Thus, we dropped the Decision Tree because of poor performance.

And we finally used this model on our test data set and our accuracy came out to be around 76%.

V. Future Work and Conclusions

This paper gives a brief insight of different factors that affected the covid cases. This paper also includes a simple model predicting chances of covid in a given area with number of people vaccinated and other factors like GDP and HDI of the area or country. With this study we conclude that the mortality rate after vaccination is lower than before it. For a future analysis on a similar subject, it would be beneficial to include more features like the fearless attitude of people after getting vaccinated.

VI. Acknowledgements

We would like to express our very great appreciation to Prof.Amit Sethi, Prof. Manjesh Hanawal, Prof. Sunita Sarawagi, and Prof. S. Sudarshan for this opportunity to explore the application of data science.

References

[1] https://www.who.int/health-topics/coronavirus#tab=tab_1
[2] https://pubmed.ncbi.nlm.nih.gov/33281306/
[3] https://ourworldindata.org/covid-vaccinations
[4] https://en.wikipedia.org/wiki/Exploratory_data_analysis
[5] https://www.analyticssteps.com/blogs/overview-descriptive-analysis
[6] https://en.wikipedia.org/wiki/Predictive_analytics