**HDSC November ’21 Capstone Project Presentation: Death Rate Prediction**

*By Team Scikit-learn*

Population data is essential for planning purposes. The knowledge of which is an integral perquisite for government/organisational planning. Population data are mainly predicted based on population growth rate, immigration rate and life expectancy. With the advent of COVID-19 pandemic, it is now paramount to put into consideration additional factors such as rate of infection, and COVID survival rate to generate a more accurate prediction for adequate planning.

This project is to perform analysis on COVID-19 data, to predict with high degree of accuracy and speed the death rate due to COVID-19 in countries around the world. Taking into consideration factors like population, vaccine administration, infection rate and severity of infected individuals.

**Aim and Objective**

The aim of this project is to come up with a ML application which can predict death rate due to Covid-19 in countries around the globe with high accuracy. The objective is to deploy the developed model to be able to make live predictions.

**Process Flow**

Data Splitting

Exploratory Data Analysis

Data Cleaning

Feature Engineering

Data Sourcing

Model Training and Evaluation

**Data Sourcing:** This process taken to extract and integrate data from trusted data bank. The data used for this project is sourced from Kaggle at <https://www.kaggle.com/josephassaker/covid19-global-dataset?select=worldometer_coronavirus_daily_data.csv>  
**Data Pre-processing:** There the raw data is transformed to an understandable format. The data raw is cleaned, analysed, integrated, and transformed to high-quality data for machine learning Algorithm. Exploratory analysis done is to derive meaningful insight from the dataset.

Data Cleaning: This process involved taking the best decision on how best to deal with incomplete, inaccurate, and irrelevant records from the source data. Missing values from the columns of interest ‘daily\_new\_cases’ and ‘active\_cases’ was about 1% of the dataset, the missing values werefilled with zeros (0) as an illustration.



Exploratory Data Analysis: Here the cleaned dataset was analysed to derive patterns using data visualization methods. To begin with correlation of all columns was studies, in other to have an overview understanding of relationships between records.

Interesting enough there was high correlations among key records, which led to further analysis.

Graphical user interface

Description automatically generated

Of which is a visualization of the margin between rate of infection and rate of death amongst the highest hit countries.

Chart

Description automatically generated

Chart, line chart

Description automatically generated

From the above plots, we can observe that, the rate of infect grows exponentially at the onset, as countries take precautions – lockdown, social distancing and the likes, infection rate also declined in same pattern. Also observed is on average low death rate compared to infection rate, except for few cases where the death rate shot up.

Furthermore, to observe the effect of population on Covid-19 activities,population of countries at both ends of the curve was tabulated withtotal confirmed cases.

Table

Description automatically generated

From the above table it is observed that top populated countries like China had confirmed cases reaching over90000, while least populated countries like Saint Barthelemy had confirmed cases slightly above 1000. But surprisingly, Barthelemy had the highest case population ratio.

But more interesting, was a discovery of multiple peaks in the charts of some countries. This peaks most likely depict second waves probably due to a new strain.

Graphical user interface, chart, application

Description automatically generated

China according to the news last year controlled the outbreak after the first hit which probably explains the presence of just a peak.

Further analysis indicates an interesting fact, the impact of the pandemic on countries may depend largely on measures put in place by individual countries to litigate further spread.A check on the degree of correlation of population with total confirm supports this view.

Graphical user interface, application

Description automatically generated

Here we had a moderate correlation, this is likely because country like China with high population, was able to manage the Pandemic averagely well.This explains whycountrylike China with high population had low total-confirm case ratio compared to American and Saint Barthelemy. Visualization was also done based on continents, it is interesting to note Both North and South America appear to be heavily hit by this pandemic.

Chart

Description automatically generated

Europe and Asia are also recording quite high numbers, followed by Africa with slightly less aggressive numbers. Australia/Oceania seems to be the continent fairing the best with this pandemic, with record low numbers.

Feature Engineering: This involved the use of domain knowledge to extract and transform features from the dataset, to improve performance of machine learning algorithm. With the use of time series several analysis was carried out including ADF test and then the data was split into train set and test set and finally fit into the models.

Chart, scatter chart

Description automatically generatedThe above plot is prediction of daily new death cases for the next few months.Several models were used to check for the best result using selected countries.

Text

Description automatically generated

**Conclusion**

It should be noted that the values predicted with these models should only be relied upon as far as the conditions during the model building remains. With significant events such as the development of a new vaccine, increase awareness and prescription drug could reduce death cases.