Group members

Imane Higazy

Philip Morehead

Emenike Chimdi Raphael

Task 1. setting up

All done

Task 2. Business understanding

Background:

The presented project is about the impact of COVID-19 in the USA. This is presented through studying the healthcare system in USA via datasets about the number of patients, number of deaths, number of cured, and number of ICU patients. In addition, data about the medical tests each of the patients have undergone is as well listed. Thus; this project does not primarily have any business goals, although, it has an impact on the USA economy.

Goals:

(1) Providing insights about infection severity: Studying data about COVID 19 in all of the USA states will provide insights about severity of infection in each state and the overall infection rate in the USA.

(2) Providing insights about efficiency of the healthcare system, inferred from data about early detection of infection and data about the number of patients cured.

(3) Provide data about the overall death toll, resulting from the COVID infection

(4) Providing information about the efficiency of the medical tests performed, on the detection and follow up of infection progress.

(5) Provide guidelines to economists about the financial needs of developing the healthcare system.

(6) Provide guidelines to economists about the expected production rates, as per the COVID situation; considering the average age of patients.

Success Criteria:

(1) Data available is covering all of the USA states

(2) Similar projects have been carried in various countries using different datasets, which would provide guidelines for our project

(3) As COVID infection is still prevailing, updated data can be also searched for, to enhance more efficient prediction

- Assessing your situation

Inventory of resources

(1) Datasets of related projects would be referred to

(2) Datasets of the actual project provided would be the mail guideline

- Requirements, assumptions, and constraints (N/A)
- Risks and contingencies (N/A)
- Terminology (N/A)
- Costs and benefits (N/A)

- Defining your data-mining goals

Data-mining goals

(1) Find interesting patterns in the data set which might help in finding correlations that would provide more useful data about the infection that is still spreading in its various forms

(2) Data visualization would even provide more vivid picture of patterns that might have appeared unrelated.

Data-mining success criteria

(1) Testing various models against data provided would give broader view on the expected outcome

(2) Finding the adequate model for the data sets provided would be sort of prototyping for other future related projects

Task 3. Data understanding

<u>Gathering Data</u>

<u>Outline data requirements:</u>

- Number of deaths due to Covid-19
  - We need the daily death toll per US state over the period of one year.
  - We need the cumulative death toll in each state over discreet periods of time to be analyzed
- Number of hospitalizations due to Covid-19
  - We need the daily hospitalization rate of Covid-19 patients in each state over the period of one year.
  - We need the cumulative hospitalization rate of Covid-19 patients in each state over discreet periods of time to be analyzed.
  - We also need the number of people being sent daily to the intensive care unit of hospitals each day for each state
  - We also need the cumulative number of people being sent to the intensive care unit over discreet periods of time.
- Number Covid-19 test results
  - We need the daily number of positive Covid-19 test results of each state over the period of one year
  - We also need the daily number of negative Covid-19 test results from each state over the period of one year
  - We also need the cumulative numbers of Covid-19 test results over discreet periods of time.

<u>Verify data availability:</u>

All in all, the data present in our dataset fulfills the requirements for the goals of our project. We found it necessary to narrow the timeframe of our project to one year given that much of the earlier datapoints are lacking values. To correct this, we decided on a cutoff of one year because most of the dataset seems complete and intact.

Define selection criteria:

For our data gathering process we chose a public domain dataset from "https://www.kaggle.com/ramjasmaurya/covid-in-usa-states". It shows the condition of US states during the Covid-19 pandemic. Specifically, it contains the data we require regarding hospitalization rates, death rates, and Covid-19 testing. We were successfully able to load the data from the CSV file to the Jupyter notebook platform to begin cleaning up and analyzing the data. The main issue we faced with this dataset was the missing datapoints in specific parts of the file. We corrected this by adjusting our timeframe to one year from the latest datapoint.

Describing data:

The data we are using is contained in a CSV file obtained from a public domain Kaggle dataset. In it, we have 39 daily metrics from each US state and territory ranging from deaths, hospitalizations, recoveries, ICU numbers, Covid-19 test results, and number of people on ventilators. Each of the US states has more than a year of daily data, so the dataset has well over 2,500 rows. This seems to us to be sufficient for our stated goals, since we will only be analyzing one year's worth of datapoints and all our required fields are represented, along with many others which are not necessarily useful.

Exploring data:

The data shows that the mean daily Covid-19 death rate in the US is 3682 and the mean number of people hospitalized is 9262. The mean number of people in the ICU because of Covid-19 is 360 and the mean daily Covid tests being taken is 2186936. There are several data quality issues that need to be considered, such as the need to round numbers, remove negative numbers and null values. We hypothesize that states with higher number of daily Covid tests tend to have greater populations and therefore higher number of infections and hospitalizations. Furthermore, higher hospitalizations correlate to higher incidence of patients having to go to ICU and consequently dying. We hope to be able to predict the recovery rate for patients using the dataset.

Verifying data quality:

We think that the dataset we have chosen will be able to fulfill our project goals. There are, however several quality issues present that need to be further addressed and fixed if possible. Firstly, the numbers need to be all rounded to the nearest integer. Secondly, we need finish narrowing our timeframe to one year to remove a large number of bad datapoints and null values. Thirdly we need to decide what to do with other null values present throughout the dataset. As far as the columns go, we need to decide which ones are useful to our project and be able to define some more ambiguous columns. Ambiguous columns which we cannot define, we should discard.

Task 4 (Planning the project)

In Our project, we are going to work on covid -19 dataset from the united states. The dataset shows the situation of all the states in the US during the pandemic. The dataset has 41 column and about 20,780 rows. The dataset is in a csv format and it is one the project topics recommended from Kaggle.

The following tasks will be handled during the project

1. The first task to collect the data and clean it up. Cleaning means handing the missing/null values. This step is important because it ensure that the data used in the machine learning model is of good quality. Instead of dropping rows or columns with null values, we decided to replace the missing values (Nan) in each column with the mean of the column.
2. Exploratory data analysis is one of the preliminary steps in any data science project. In this task, The main purpose of EDA is to look at data before making any assumptions, identify obvious errors, as well as better understand patterns within the data, find interesting relationships among the attributes via visualizations.
3. Feature engineering will also be used to creat new columns or drop some columns that are not of importance the project goals.
4. Building machine learning regression models to predict death, hospitalizations and recovery
5. Evaluation of the machine learning models.

   So, the results we get from carry out the above task will be of great importance to the ministry of health/health board and the hospital management to effective control/reduce the spread of the virus. For example, US states with a high number of death as seen from the visualizations will mean that strict public health measure such as wearing of face masks, avoiding large gatherings, hand hygiene and sanitation needs to be enforced. Also in resource allocation, states with very high hospitalization, patients in intensive care and in need of ventilators will can easily be seen in the visualizations.

   Machine learning model will help to predict deaths and hospitalization to enable proper planning to reduce the incidence, spread, death and hospitaliations.

Method and tools for the tasks

Python pandas library and functions

Python seaborn library and functions

Python matplotlib.pyplot library and function

LinearRegression model from sklearn.linear_model

Lasso regression model from sklearn.linear_model

Ridge model from sklearn.linear_model

Metrics function from sklearn

Python numpy library and functions

Jupyter notebook