Deliverable 1

1) Team

a) Members

Aleena Jimmy George (801254680)
Anupama Ramesh (801261344)
Gayathri Jayagopal (801259246)
Santhan Pal Bandari (801259376)
Sneha Chilumula (801161898)

b) Communication plan

Git Repo : https://github.com/j-gayathri/BigDataProject

Meeting notes : GroupMeetingNotes

Meeting platform : Zoom

2) Selection of domain and data

Link : COVID-19 Data Lake

The above is a dataset with up-to-date curated dataset on the spread of COVID-19

3) Business Problem or Opportunity, Domain Knowledge:

COVID-19 has been a concern for over a year now. The data collected regarding the COVID helps in a lot of research. The dataset includes a lot of information such as the number of covid positive cases, the count of positive cases identified every day, the death toll, and a lot more. Firstly, the information is required to learn about the growth of the cases. Lockdowns were imposed on several parts of the world and a lot of restrictions were put on social meetings or day-to-day activities. The effectiveness of the restrictions imposed can be seen from the past data.

Furthermore, the rate of increase in positive cases, especially in the past, alerts us about the seriousness of the situation and the measures to be taken to curb the further spread, if any.

Finally, the data helps us in projecting the future positive count and be prepared in terms of medical services to be provided.

4) Research Objectives and Question(s):

Objectives that the research caters to:

- Generate better forecasts of hotspots and trends
- Dashboards could be built to track infections and collaborate to efficiently deploy vital resources like hospital beds and ventilators

Questions that answered:

- What are the county-level predictions for number of deaths
- What county-level predictions are allocated to hospitals within counties proportional the their total number of employees
- What is the threshold number of cumulative predicted deaths for a hospital?

5) Data Understanding

a) Exploratory Data Analysis

b) Dashboard

6) Data Preparation

7) Analytics, Machine Learning

8) Evaluation and Optimization

9) Results

10) Future Work, Comments

What was unique about the data?

Typically, while predicting COVID-19 death toll by location the previous couple of years death toll data is used.

But our data set is unique as it also takes into consideration elements such as number of people fully vaccinated, GDP per capita, extreme poverty, cardiovascular death rate, diabetes prevalence, number of smokers, hospital beds per thousand, etc. This helps

when predicting the death rate due to COVID-19 for people with certain health conditions or live in a particular region.

What were the problems you faced? How did you solve them?

While cleaning the data set, we identified there are “null” values present in quite a few rows in multiple columns that where needed for our use. To fix this we replaced the null value with the mean of that particular column.

Future Work:

With this project it is seen that it’s possible to predict the COVID-19 death toll among people with specific conditions (in their environment or health), to a certain degree of accuracy. In our future work we would like to collect data for a specific health condition (like Sickle cell disease, Chronic liver disease, etc.) and predict the people affected by COVID-19 as well as the death toll.