CSE 847 Machine Learning
Project Progress Report

Submitted by:
Redwan Sony(sonymd@msu.edu)
MD Alamin (alaminmd@msu.edu)

Introduction:

This project is about predicting the number of COVID-19 cases based on the data from the previous two years. Our goal of this project is to conduct an extensive analysis on the time series data of the dataset curated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). To achieve this goal of predicting the spread of COVID-19 cases in the near future, we want to apply different machine learning prediction models. We intend to perform the comparative performance analysis of the various applied models.

Problem description:

The goal of this project is to predict the number of confirmed COVID-19 cases for the next week based on the time series analysis of the reported confirmed cases till the current day.

Description of the data:

Dataset: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series
The dataset contains 802 columns. However the number of columns will increase with the advance of time. The first 11 columns consist of different descriptors about the source of the data like the state, county, latitude, longitude and so on. The rest of the columns contain the cumulative number of confirmed cases till that day.

Work Done:

We have performed the data exploration part for our project. We analyzed the dataset containing the confirmed cases from January 1, 2020 to March 22, 2022. The dataset contains the cumulative number of cases till the mentioned day. So we first calculated the individual number of confirmed cases for a given day. For now, we are analyzing the cases in the United States of America. We also investigated statewise confirmed cases. The following section contains histograms denoting number of cases for the above mentioned time duration. Here, we only report the histograms for the top 10 states with respect to the number of cases and the lowest 10 states as well. We also report the histogram showing the total number of confirmed cases throughout the whole country. Some of the explored summary of the dataset is given in the following figures.

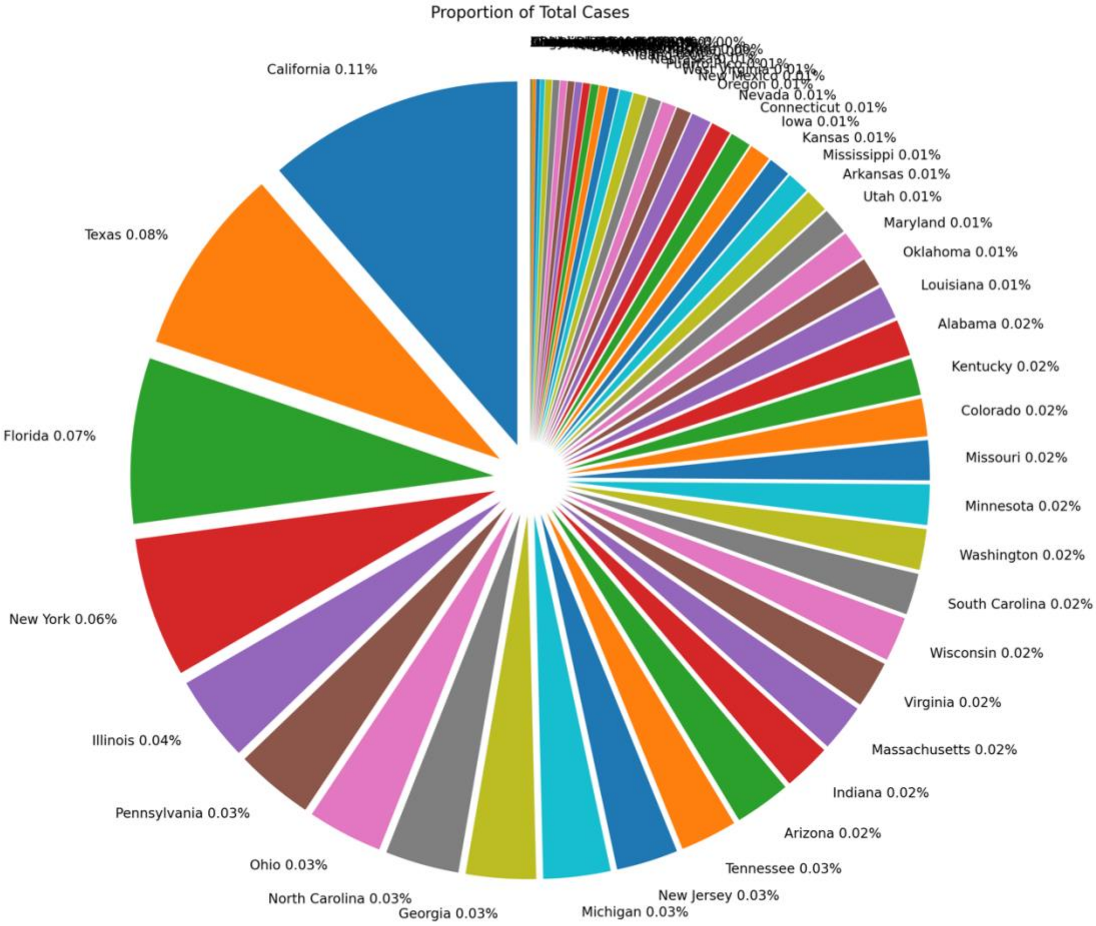

Figure 1: A pie chart showing the proportion of different states' total COVID-19 cases

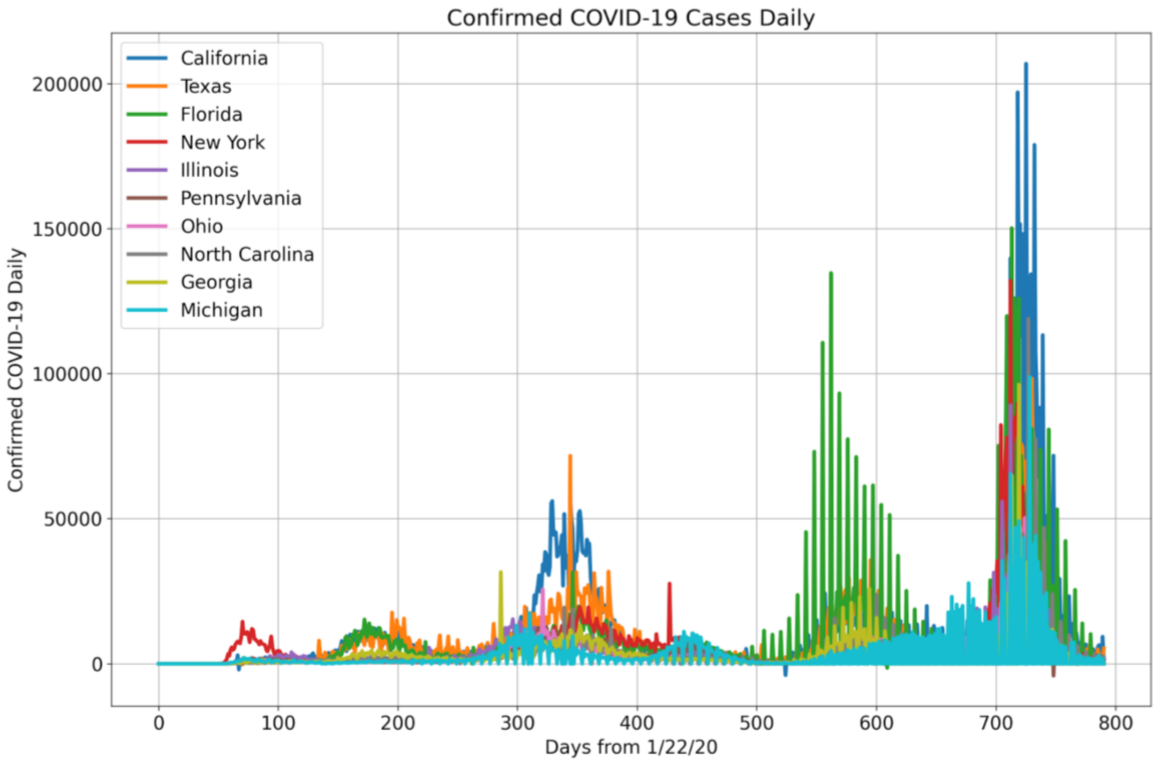

Figure 2: Top-10 States with average daily confirmed cases

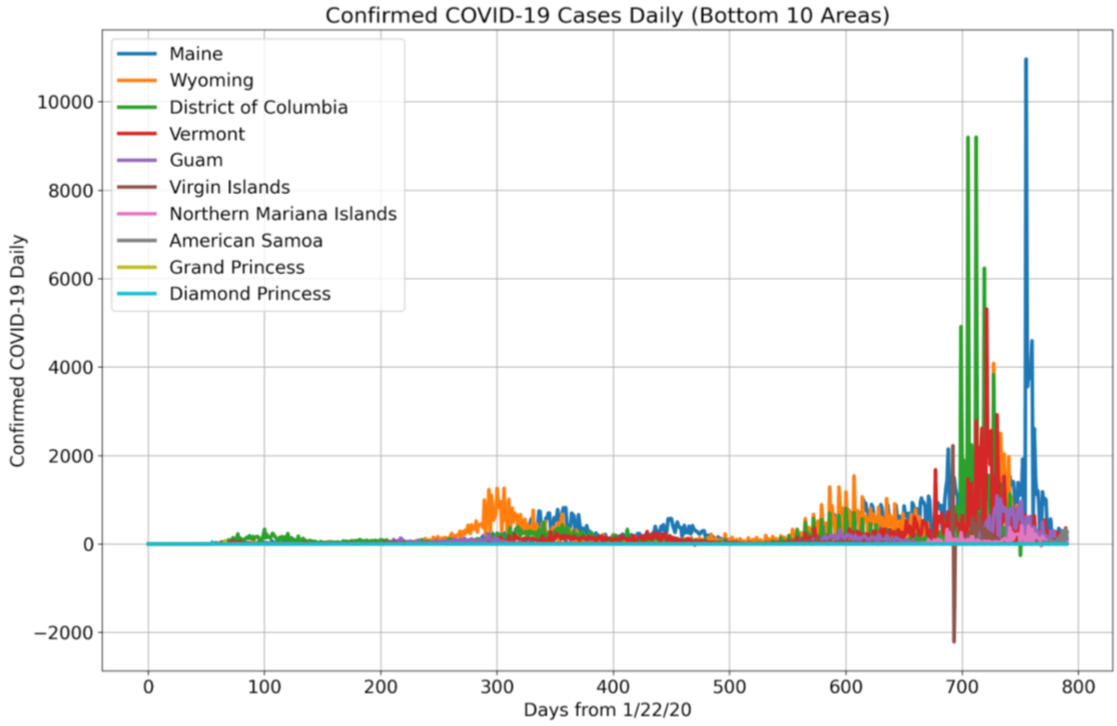

Figure 3: Bottom-10 States with average daily confirmed cases

Here one important finding is that there is some inconsistency in the data provided which requires further attention as we have some negative value in the daily cases.

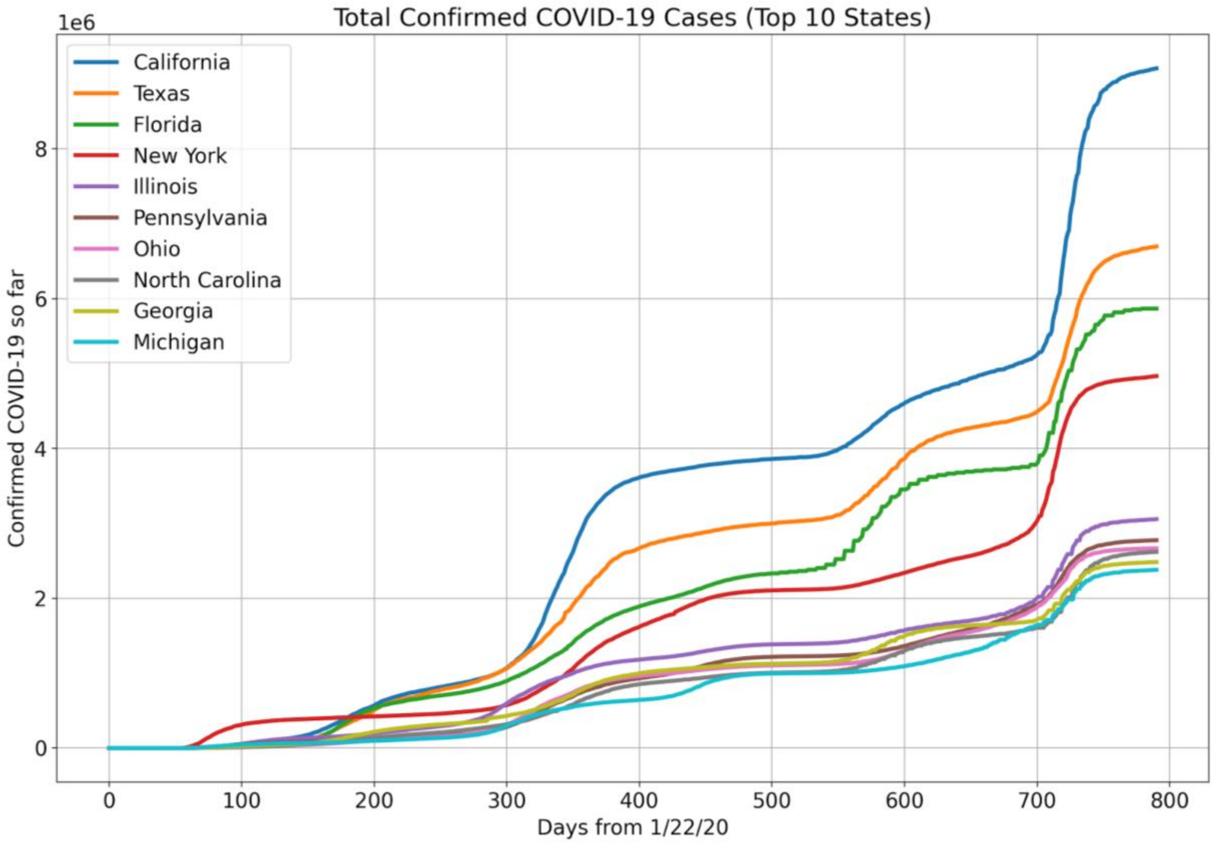

Figure 4: Top-10 States with total confirmed cases so far

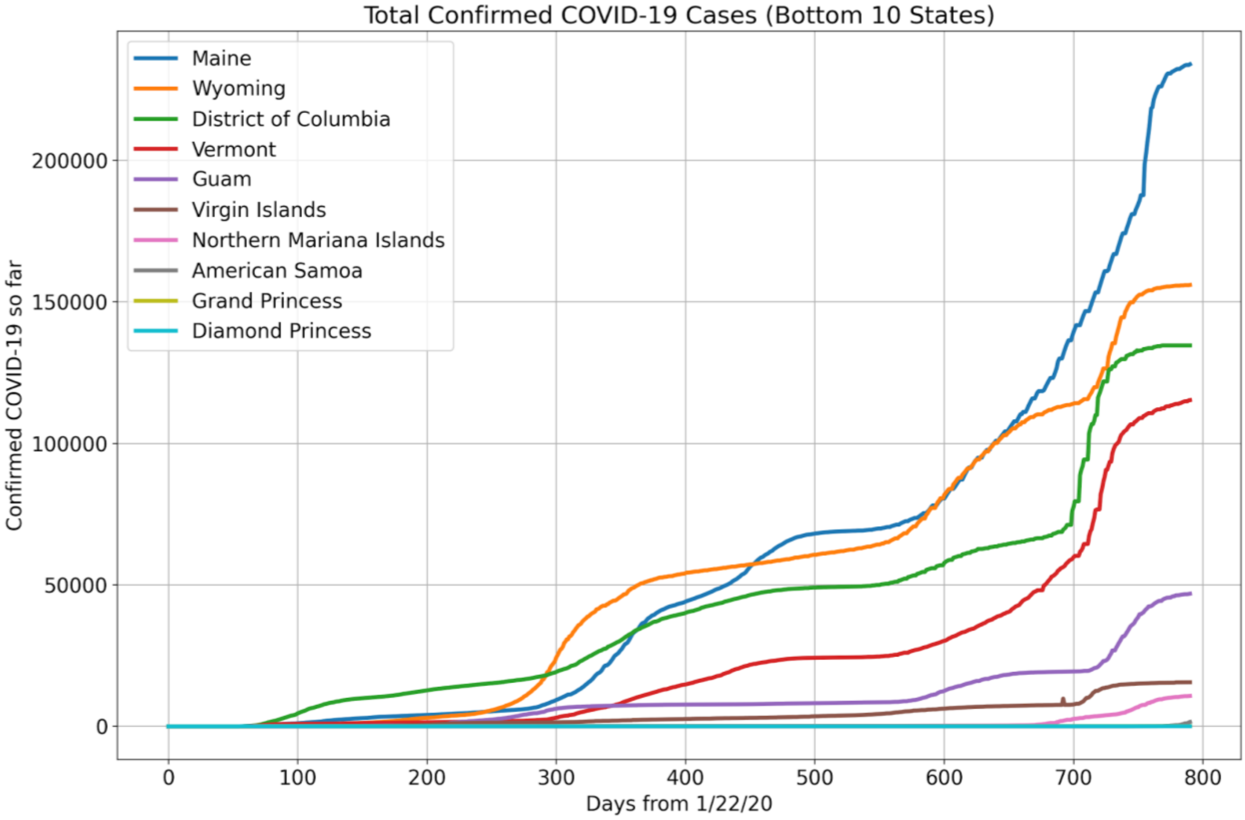

Figure 5: Bottom-10 States with total confirmed cases so far

Remaining Work:

- Creating the training and testing split. As this is a time series analysis, we can not perform random splitting. We will calculate the moving average for different time duration and split the data according to that.
- Selecting the prediction models based on our data exploration and applying them on the training data.
- Conducting the comparative performance analysis among the different prediction models.