TOPIC – COVID 19 Prediction

Business Problem:

Coronavirus disease 2019 is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The disease has since spread worldwide, leading to an ongoing pandemic. Symptoms of COVID-19 are variable, but often include fever, cough, headache, fatigue, breathing difficulties, and loss of smell and taste. Symptoms may begin one to fourteen days after exposure to the virus.

COVID-19 transmits when people breathe in air contaminated by droplets and small airborne particles containing the virus. The risk of breathing these in is highest when people are near, but they can be inhaled over longer distances, particularly indoors. Transmission can also occur if splashed or sprayed with contaminated fluids in the eyes, nose, or mouth, and, rarely, via contaminated surfaces. People remain contagious for up to 20 days and can spread the virus even if they do not develop symptoms. This project proposal mainly aims at exploring COVID-19 through data analysis and projections.

Datasets:

The datasets are fetched from several websites which are mentioned as follows:

1. John Hopins University - CSSEGISandData/COVID-19: Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE (github.com).
2. Apple mobility - COVID-19 - Mobility Trends Reports - Apple
3. World Health Organization - Coronavirus disease (COVID-19) (who.int)
4. CDC - Coronavirus Disease 2019 (COVID-19) | CDC

After combining all data from various data sources, the dataset will have the following columns:

1. State – State name.
2. Country – Country name.
3. Last Update – Last updated date of the data.
4. Latitude – Latitude of the state.
5. Longitude – Longitude of the state.
6. Confirmed cases – Number cases confirmed at present.
7. Deaths – Number of deaths happened so far.
8. Recovered – Number of cases recovered.
9. Active cases – Number of present active cases.

Methods:

This project concentrates more on visualization. So, I'm planning to use the seaborn library for all the visualization. Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures. Seaborn helps you explore and understand your data. Its plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

For the prediction, we can include Support Vector Machine, Polynomial Regression, and Bayesian Ridge regression. Support vector machines are a set of supervised learning methods used for classification, regression, and outliers detection. Polynomial regression, like linear regression, uses the relationship between the variables x and y to find the best way to draw a line through the data points.

Ethical consideration:

In this project, we are going to use the data which is available for public use from websites like apple mobility, John Hoppins University, World Health Organization, etc. So, there are no ethical issues in handling data. But the ethical issue might raise when we release the results of this project because it's experimental and we don't know exactly how the model will behave for future predictions. So, there are some potential threats that the result may mislead the people and government of each country in the world. Thus, to avoid this kind of issue, we need to test the model thoroughly with all possible parameters.

Challenges:

With the resurgence of machine learning and artificial intelligence, never has it been easier to implement predictive algorithms both new and old. With just a few lines of code, state-of-the-art models can be readily accessible at the fingertips of the budding data enthusiast, ready to conquer whatever insurmountable digital task may lay at hand. But a little bit of knowledge can be a dangerous thing. While much of machine learning can be attributed to statistics and programming what is equally important, but often skipped over in favor of instant gratification, is domain knowledge. There are a few Challenges in using the Bayesian Regression algorithm which is as follows:

1. The inference of the model can be time-consuming.
2. If there is a large amount of data available for our dataset, the Bayesian approach is not worth it and the regular frequentist approach does a more efficient job.

Reference:

1. COVID-19 - Wikipedia
2. Coronavirus (COVID-19) Visualization & Prediction | Kaggle