**Capstone Project – Covid 19**

**Table of Contents**

**1. Problem Statement**

**2. Project Objective**

**3. Data Description**

**4. Data Pre-processing Steps and Inspiration**

**5. Choosing the Algorithm for the Project**

**6. Motivation and Reasons for Choosing the Algorithm**

**7. Assumptions**

**8. Model Evaluation and Techniques**

**9. Inferences from the Same**

**10. Future Possibilities of the Project**

**11. Conclusion**

**12. References**

**Problem Statement :**

Given data about COVID-19 patients, write code to visualize the impact and analyse the trend of rate of infection and recovery as well as make predictions about the number of cases expected a week in future based on the current trends.

**Project Objective :**

The project objective is to analyse the impact and trend of COVID-19 infection and recovery rates using visualization techniques, and to generate predictions for the number of COVID-19 cases expected a week into the future based on the current trends identified through the time series models

**Data Description :**

* This dataset contains a detailed account of the impact of COVID-19, including the geographical location and key metrics such as confirmed cases, deaths, and recoveries.
* The data also contains of latitude and longitude analysis, and the date column enables the know of trends over time.
* The WHO Region information adds allowing for regional comparisons.
* The dataset will be used to create interactive visualizations and time-series models to better understand and predict the progression of the pandemic.

**Data Pre-processing Steps and Inspiration :**

* Columns were given appropriate names and datatypes of columns were changed where ever needed .
* Depending on the specific analysis goals, filter the dataset to include only relevant columns for visualization and modelling. For instance, you might focus on "Date," "Confirmed," "Deaths," "Recovered," and "Country/Region."

**DATA INSIGHTS:**

* The number of confirmed COVID-19 cases has been increasing since February 2020, with a significant spike in April and May 2020.
* The number of recovered cases has also been increasing, but at a slower pace compared to the confirmed cases.
* The number of active cases has been fluctuating, with a slight decrease in June and July 2020.
* The number of deaths has been increasing since February 2020, with a significant spike in April and May 2020.
* There is a strong positive correlation between the number of confirmed cases and the number of active cases (0.91), indicating that as the number of confirmed cases increases, so does the number of active cases.
* The most affected country due to covid was USA as we can see it has highest deaths, active and confirmed cases.

**Choosing the Algorithm for the Project:**

For the initial analysis and forecasting of COVID-19 trends, Facebook Prophet is a strong candidate due to its simplicity, flexibility, and suitability for time series data with patterns and seasonality.

**Motivation and Reasons for Choosing the Algorithm :**

* The COVID-19 dataset is inherently time-series in nature, with observations recorded over time. Facebook Prophet is specifically designed for time series forecasting, making it well-suited for capturing temporal patterns, trends, and seasonality inherent in the spread of the virus.
* Facebook Prophet is designed to be user-friendly and requires minimal tuning of parameters.
* Prophet automatically detects changepoints, which are points in time where the time series experiences a sudden change in its trajectory.

**Assumptions :**

* It is assumed that the reported COVID-19 data, including confirmed cases, deaths, and recoveries, is accurate and reliable.
* The analysis assumes that factors influencing the spread of COVID-19, remain relatively stable over the period of analysis.
* The assumption is made that any observed seasonal patterns in the data are consistent

**Model Evaluation and Techniques :**

* For model evaluation technique used was finding RMSE values of both the model and whose RMSE value was less is considered the best model
* RMSE value of Fb prophet was 1984197.70
* As a result, Fb prophet model was chosen.

**Inferences from the Same :**

* The large RMSE indicates room for improvement in the model .
* The large RMSE suggests that the model's predictions deviate from the actual values by a substantial amount. The scale of COVID-19 cases (confirmed, deaths, or recovered) may contribute to the magnitude of errors.
* This could be due to the presence of an outlier or extreme value in the data, which the model might not have captured fully.
* Overall, the Fb prophet model seems to be a good fit for the data

**Future Possibilities of the Project :**

* Model can be improved by adjusting hyperparameter tuning .
* We could try other time series models or ensemble methods or deep learning models to see if we can achieve better accuracy.
* Adjusting seasonality-related parameters in the Prophet model might improve its ability to capture recurring patterns.

**Conclusion :**

* Data cleaning and preprocessing were crucial in improving the accuracy of the time series models.
* Various techniques were used to gather important information for future forecasting.
* Facebook's Prophet was chosen as the best performing model