## Stage IV (Basic Machine Learning)

Team Task

**Develop Linear and Non-Linear (polynomial) regression models for predicting cases and deaths in US**

For this task the superdataset generated in Stage1 of project is used.

* Performed data cleaning & adding additional column for daily cases and deaths to the data frame.
* Grouping the data by dates so we get daily cases and deaths details in a date wise fashion.
* Creating first occurrence column in the data frame for cases & deaths.

Below are the screenshots of the models created for both Linear and Non-Linear Regression :

**Linear Model New Cases:**

**Chart, line chart

Description automatically generated**

* From the plot of linear regression for cases we can see that the model fit the data in a good manner.
* We can observe the trend line is increasing as Number of cases per day is increasing.
* We can infer from the graph, as number of days are between 700 to 747 there is peak / surge in new cases of Covid\_19.
* We can infer That the highest peak of the cases is at the end of the graph for linear model.

**Linear Model New Deaths:**

**Chart, scatter chart

Description automatically generated**

* From the plot of linear regression for deaths, we can see that the model fit the data in a nice manner.
* We can observe the trend line is increasing as the number of deaths per day is increasing.
* We can infer from the above diagram that the trend is positive from the start, and it is increasing.
* We can infer as number of death cases in linear model is increasing or at the peak it is not affecting the trend line.

**Non-Linear New Cases:**

Chart

Description automatically generated

* For nonlinear cases, it can be seen that the regression lines are at the peak in the end.
* For nonlinear regression model of degree = 2, it fits the model in a linear manner.
* For nonlinear regression model of degree = 3, it fits the model in a curved manner.
* For nonlinear regression model of degree = 4, it fits the model in a multi curved manner.
* The model with degree = 4 fits the data in the best way.

**Non-Linear New Deaths:**

* For nonlinear deaths, it can be seen that the regression lines are at the peak in the end.
* For nonlinear regression model of degree = 2, it fits the model in a nonlinear curvature manner.
* For nonlinear regression model of degree = 3, it fits the model in a curved manner.
* For nonlinear regression model of degree = 4, it fits the model in a multi curved manner.

Chart, scatter chart

Description automatically generated

* The model with degree =2 has a declining curve at the end of the data points. While degree 3 and 4 have an increasing curve at the end of the data points.
* The model with degree = 4 fits the data in the best way.

### **Plot trend line along for the data along with the forecast of 1 week ahead:**

**Linear Model trendlines:**

1. **Cases Data.**

* For new cases, we have plotted the trendline using plotly graphs, by considering the number of days since first occurrence of cases on X axis and the number of daily cases reported on Y axis.
* The trendline shows that the data have an increasing trend. It starts with a few cases reported in the initial days which goes on increasing and reaches the peak at around 747 days.
* In-between there are a few small peaks, but the trend is overall increasing and positive.

Chart, line chart

Description automatically generated

**Linear model trendline for new cases data**

1. **Deaths Data**

* For new deaths, we have plotted the trendline using plotly graphs, by considering the number of days since first occurrence of cases on X axis and the number of daily deaths reported on Y axis.
* The trendline shows that the data have an increasing trend. It starts with a few deaths reported in the initial days which goes on increasing and reaches the peak at around 380 days and again at around 620 days.
* But the trend is overall increasing and positive.

Chart, scatter chart

Description automatically generated

**Linear model trendline for new deaths data**

**Trendline for predicted data of a week:**

1. **Cases Data:**

Chart, line chart

Description automatically generated

**Predicted trendline for the upcoming week’s cases data**

* For future prediction of upcoming week’s cases, we have generated data of next seven days using the model.predict().
* Then we have concatenated it with the original data for cases and days.
* After creating a new updated data frame, we used plotly to plot the trendline same as above.
* We can see that the trendline is increasing since the first day of cases reported.
* And it can be predicted from the trendline that in the upcoming week the cases are going to increase.
* Also, we can see a slight difference of trendline between the original new cases trendline and the upcoming week’s cases trendline.

1. **Deaths Data:**

Chart, scatter chart

Description automatically generated

**Predicted trendline for the upcoming week’s deaths data**

* For future prediction of upcoming week’s deaths, we have generated data of next seven days using the model.predict().
* Then we have concatenated it with the original data for deaths and days.
* After creating a new updated data frame, we used plotly to plot the trendline same as above.
* We can see that the trendline is increasing since the first day of deaths reported.
* And it can be predicted from the trendline that in the upcoming week the deaths are going to increase.
* Also, we can see a slight difference of trendline between the original new deaths trendline and the upcoming week’s deaths trendline.

**Inferences:**

1. **Predicted trendline for the upcoming week’s cases data**

* From the graphs of upcoming week trend for cases, where we try to compare with the original dataset, the new datapoints generated lie in range beyond 747 where 747 is the last data point in the original dataset.
* After modelling and predicting the new values we see the there was a rise in datapoints from 747 to 754. This range of days on the X-axis prove that the values have been predicted for a week data based on the model trained on the original dataset.
* The number of cases on 754th day is 212.6448K.

1. **Predicted trendline for the upcoming week’s deaths data**

* From the graphs of upcoming week trend for deaths where we try to compare with the original dataset, the new datapoints generated lie in range beyond 747 where 747 is the last data point in the original dataset.
* After modelling and predicting the new values we see the there was a rise in datapoints from 747 to 754. This range of days on the X-axis prove that the values have been predicted for a week data based on the model trained on the original dataset. The number of deaths on 754th day is 1406.975.