AML ASSIGNMENT 1 – WRITEUP

Deepak Mewada

20CS91P02

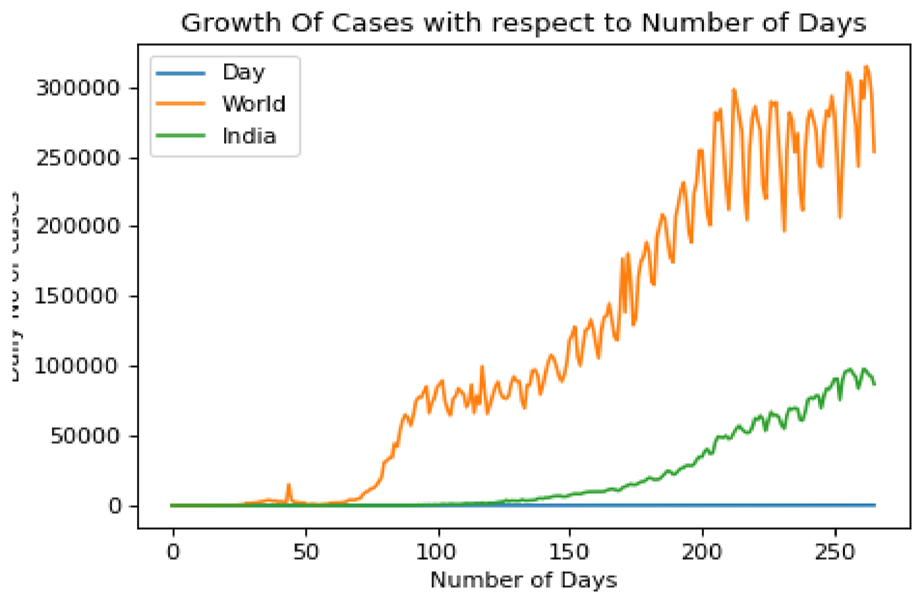
Details of Methods Used:

The objective of this assignment is to predict the number of new COVID-19 cases in India and the World using Gaussian Process Regression. So followed the steps given below to predict the new cases-

1. **Loading the Data & Examining its size/shape/ features**: First the .csv file is loaded in the programme environment then the shape and size of data is checked. Which is found to be as follows

```
(266, 4)
[[1 '31-12-2019' 27 0]
[2 '01-01-2020' 0 0]
  [3 '02-01-2020' 0 0]
...
[264 '19-09-2020' 309844 93337]
[265 '20-09-2020' 294862 92605]
[266 '21-09-2020' 253567 86961]]
```

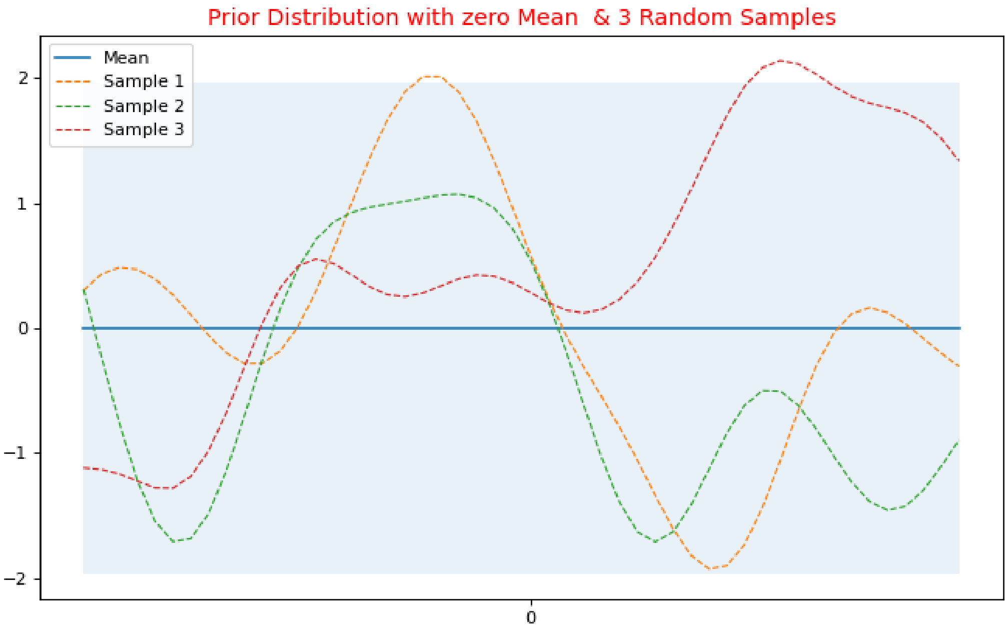
2. **Data Visualisation:** Data is visualised to see get the distribution of data. To assess which prior and what kernel parameters will suit our data.



3. **Data Pre-processing:** The data in .csv file is pre-processed to make is suitable for Gaussian Process Regression. The '*No. of Days*' are considered as Feature vector X and the '*No. of Cases/day*' as Target vector for Gaussian Process Regression. Then the data is divided in training set (data till September) and testing set (data from September onwards) according to number of days.

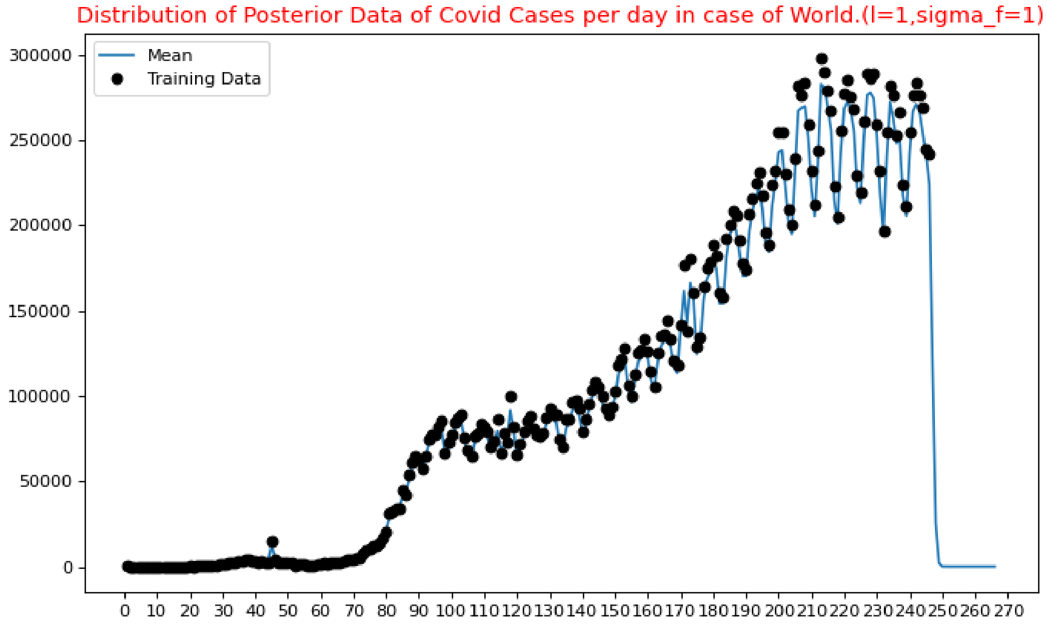
4. **Defining Kernel:** Kernel is a covariance function describes the covariance of the Gaussian process random variables. Here I used the squared exponential kernel, also known as Gaussian kernel or RBF kernel. The length parameter '**l**' controls the smoothness of the function and '**σf**' the vertical variation.

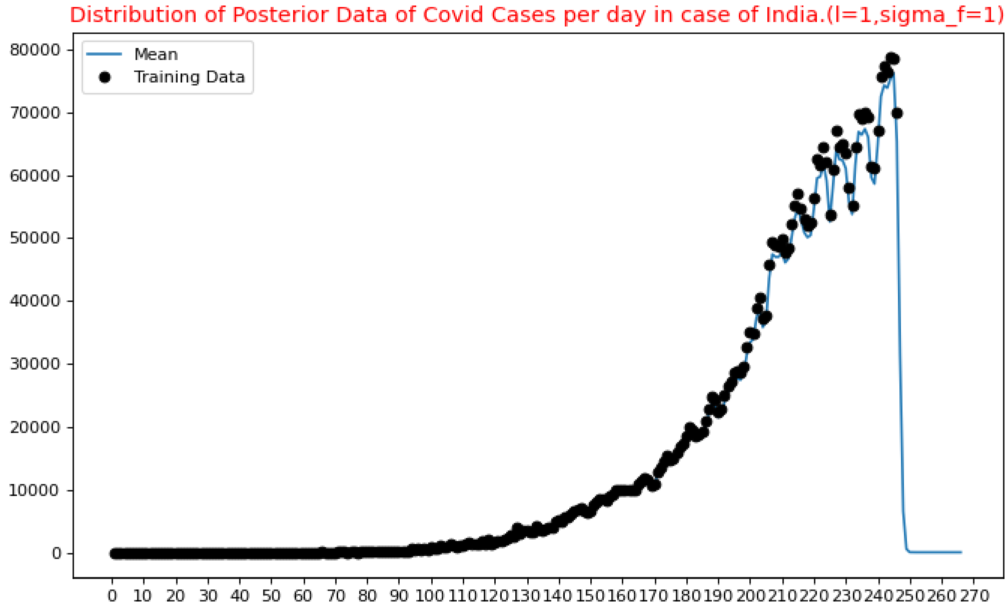
5. **Defining Gaussian Prior:** A Gaussian prior over function is defined. Here is the illustration of few samples drawn from distribution.



6. **Defining the Posterior Predictive function:** To compute the sufficient statistics i.e. mean and variance from x_train, y_train and new input x_test this posterior predictive function is defined. It returns posterior mean vector and covariance matrices.

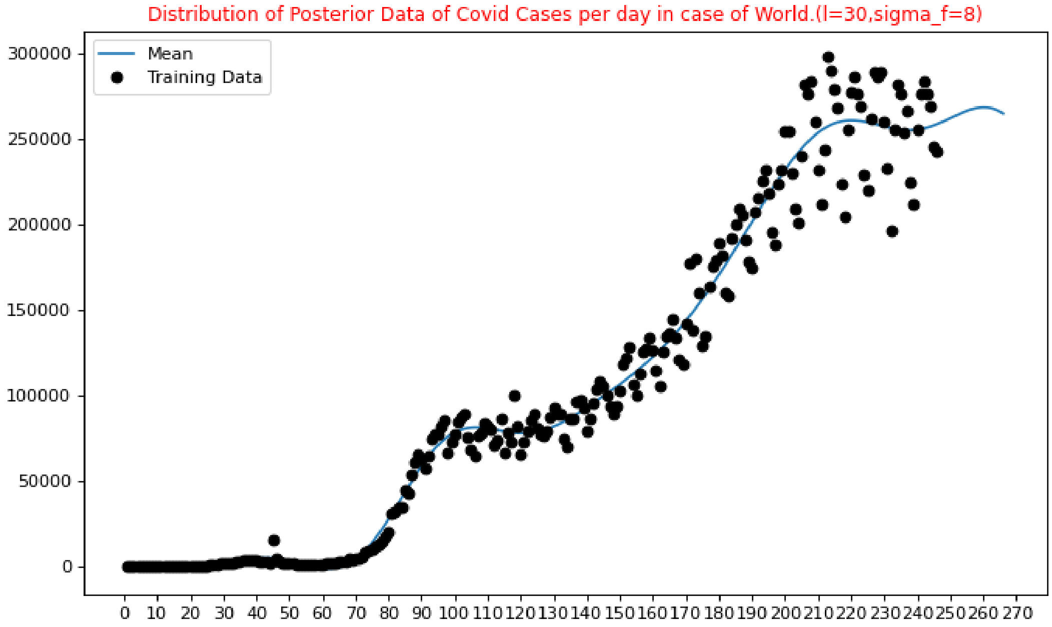
7. **Calling Predictive function (PPF) and plotting the data with Mean vector & Covariance matrices with default l=1, sigma_f=1:** Here the Posterior Predictive function (PPF) is called and the mean vector & covariance matrices for Posterior Predictive Distribution (PPD) is obtained. After that the this PPD is plotted along with the prediction for our new data x_test. Here we are considering mean line as the prediction for future data value.

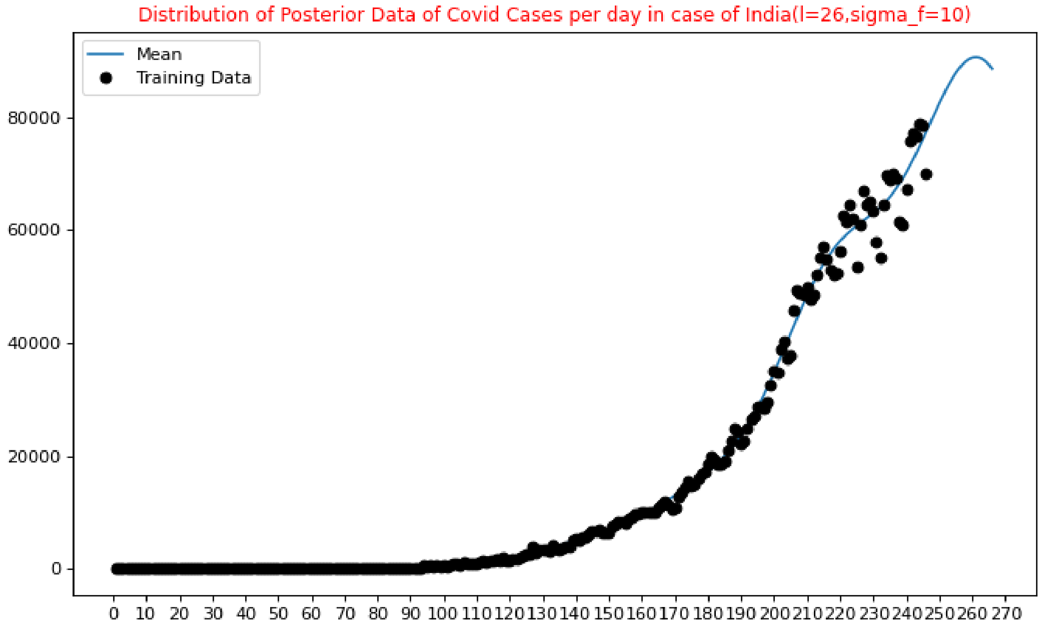




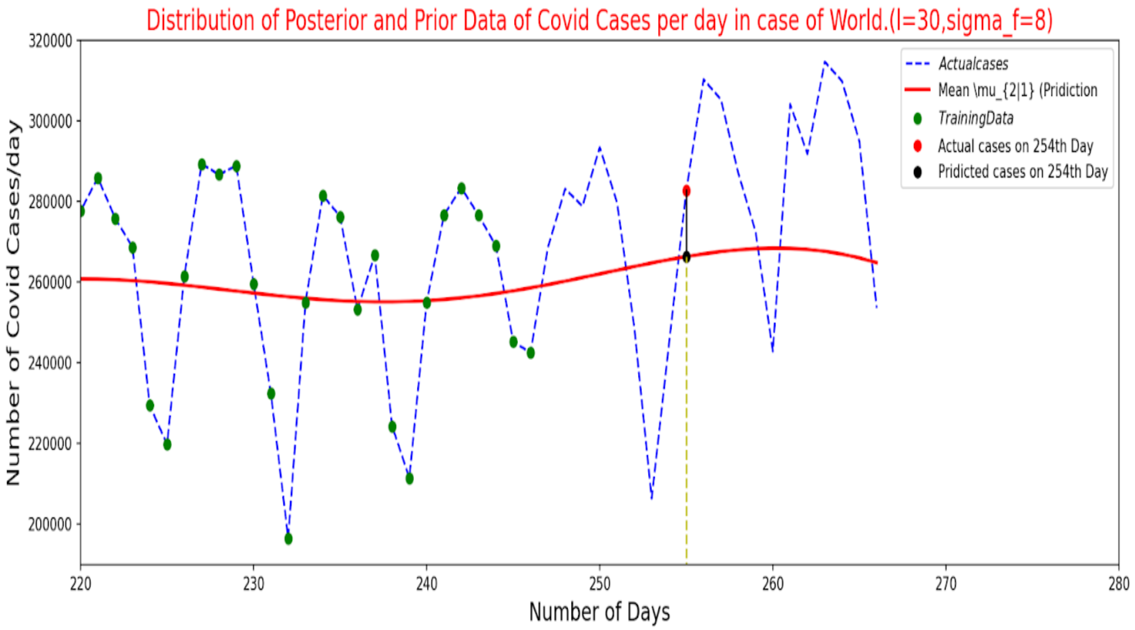
8. **Performance Measurement:** Defining function for Performance Measurement of our model with Root Mean Squared Error.

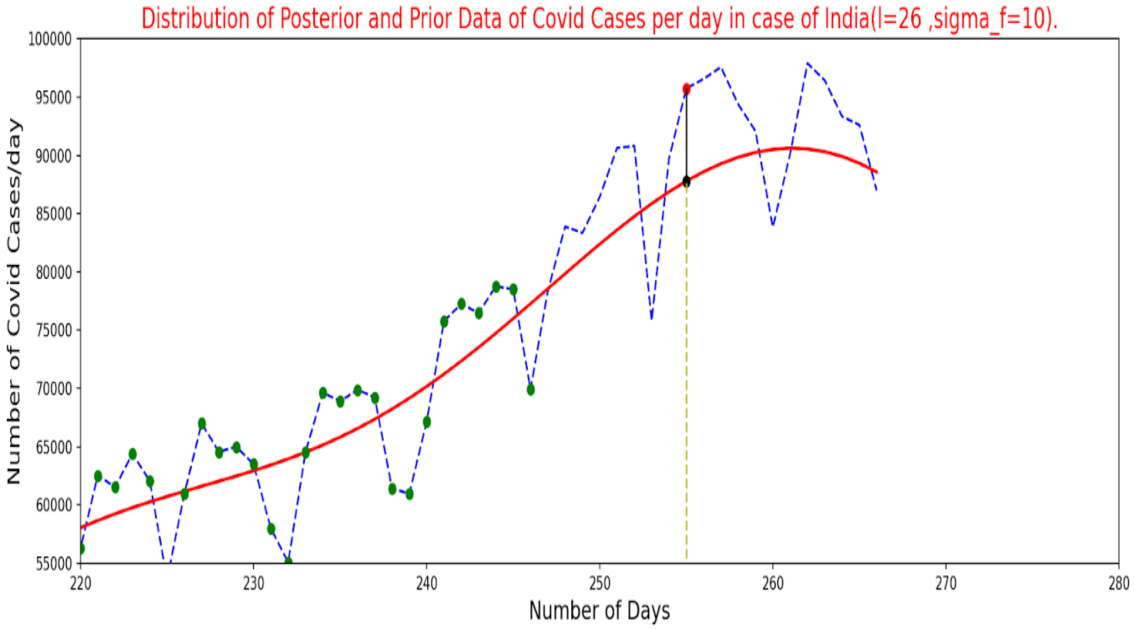
9. **BONUS-Hyperparameter Optimisation:** Performed hyperparameter optimisation to get best suitable values of '**l**' and '**sigma_f**' according to our data. The output obtained after hyperparameters optimisation is given below:(Hyper parameters are written above in title of figure)





10. **Executing the code and Finding Result:** After running code the last plot will give approximate number of cases in near future on 254th day.
Note: Here in the figure given below the yellow line points to prediction of Covid cases for 254th day. The black dot shows predicted cases on 254th day & the red dot shows the actual cases on 254th day. Black line joining red dot and black dot shows the error in our predicted data.


Distribution of Posterior and Prior Data of Covid Cases per day in case of World.(l=30,sigma_f=8)


Distribution of Posterior and Prior Data of Covid Cases per day in case of India(l=26 ,sigma_f=10).

Conclusion:

For India**, l = 26.0** and **sigma_f = 10.0**, we get optimal performance.
For World**, l = 30.0** and **sigma_f = 8.0**, we get optimal performance

Code:

Language:	Python
Platform:	Google Collab
File Name:	AML_Assignment1_20CS91P02.py