Question 6

[CM6]Naive Bayes Classifier

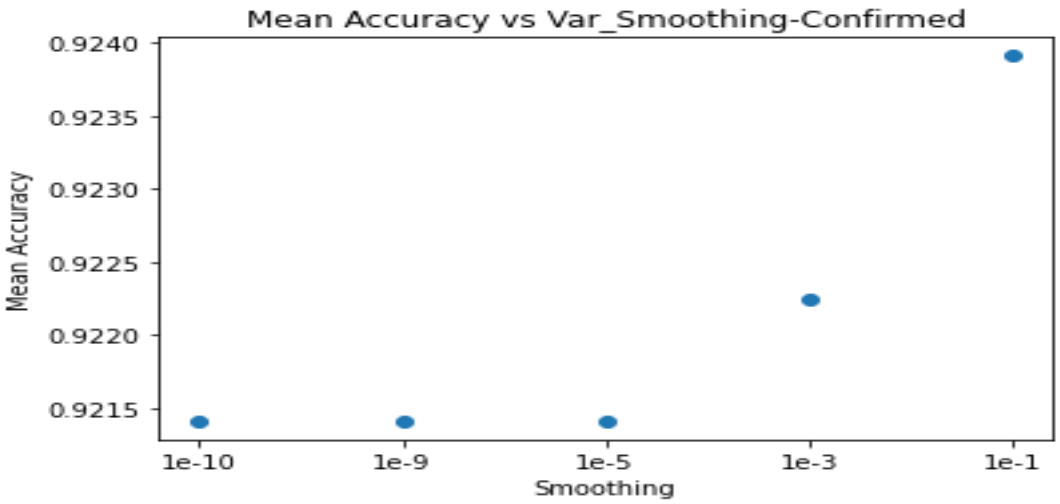
1. Original Feature Space is used for this classification
2. Hyper parameter tuning is performed using 10-fold cross validation on each label

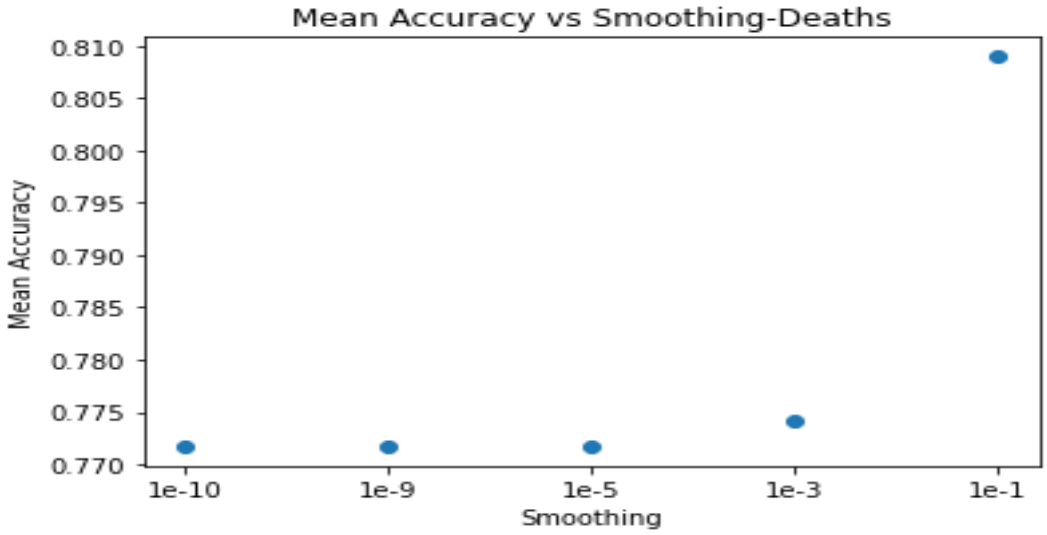
```
DTbase = GaussianNB()
param_grid = {
    'var_smoothing': [1e-10, 1e-9, 1e-5, 1e-3, 1e-1]
}

DT_fit = GridSearchCV(estimator=DTbase, param_grid=param_grid, cv = 10, refit='accuracy_score')
DT_result = DT_fit.fit(Original_data_copy.iloc[:, 3:14], y["Confirmed"])

results_df = pd.DataFrame(DT_result.cv_results_)
results_df
```

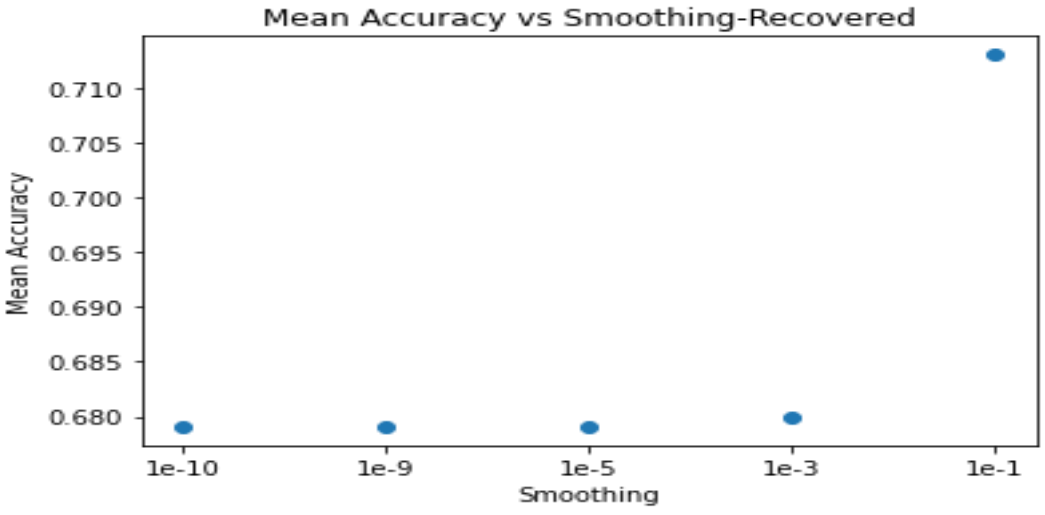
Results of Hyperparameter tuning:

	Confirmed	Deaths	Recovered
param_var_smoothing	mean_test_score	mean_test_score	mean_test_score
1e-10	0.921412	0.771742	0.679077
1e-09	0.921412	0.771742	0.679077
1e-05	0.921412	0.771742	0.679077
0.001	0.922245	0.774222	0.679910
0.1	0.923912	0.808953	0.712989

Plot of accuracy vs var_smoothing values for each of the labels are as below :







Label	Best Var_smoothing	Accuracy (With Smoothing)
Confirmed	1e-1	92.39%
Deaths	1e-1	80.89%
Recovered	1e-1	71.29%

Can you explain the impact of the smoothing parameter?

- var_smoothing is a stability calculation to widen (or smooth) the curve and therefore account for more samples that are further away from the distribution mean
- We can observe from the plot that as the var smoothing increases, the curve is smoothened and the samples which are far away from the distribution mean are also considered value .Therefore, the accuracy of prediction also increases for all the 3 labels.
- If we have no occurrences of a class label and a certain attribute value together, then the frequency-based probability estimate of Naïve Bayes will be zero. We can get over this Zero Probability Phenomena by using smoothing.