Question 7

[CM7]Interpretability

1. Original feature space is used here
2. Explain the performance of NB compared to the decision tree approaches in Question 1.

Accuracy of different tree methods with the best hyper parameters tuned are compared with that of the Naïve Bayes here for each of the 3 labels.

Label	Accuracy Decision Tree	Accuracy Naïve Bayes	Accuracy Random Forest	Accuracy Gradient Tree Boosting
Confirmed	95.95%	92.39%	96.28%	96.03%
Deaths	92.22%	80.89%	92.39%	92.14%
Recovered	91.40%	71.29%	93.30%	93.13%

Naïve Bayes is a Probabilistic model and easier to implement

Although, it can be observed that it does a good job with “Confirmed” label classification whose accuracy values are comparable to those of tree based techniques, Compared to all the tree based classification techniques, it can be observed that Naïve Bayes tends to fall short in keeping up with the Decision Trees, Random Forests and Gradient Boosting Techniques for the covid dataset classification.

Naïve Bayes makes an assumption that each feature in independent and identically distributed and if this is not the case, the accuracy of classification is affected.
With the features in Covid dataset, this assumption cannot be assured.
Also the labels in the covid dataset are unbalanced, it could also be one of the reasons for this performance of Naïve Bayes.

3. Can you use the learned parameters of NB to make an interpretation about the data and compare this to the single Decision Tree model?

Label	Accuracy Decision Tree	Accuracy Naïve Bayes
Confirmed	95.95%	92.39%
Deaths	92.22%	80.89%
Recovered	91.40%	71.29%

From the above comparison, we can observe that Decision Tree has better performance compared to Naïve Bayes (Note: even with the best var smoothing value of 1e-1,it falls short) for our dataset for all the 3 labels with Confirmed label having highest accuracy of 95.95%, followed by Deaths and Recovered with 92.22% and 91.40% respectively.