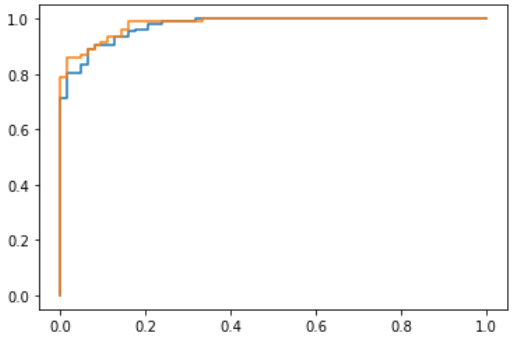
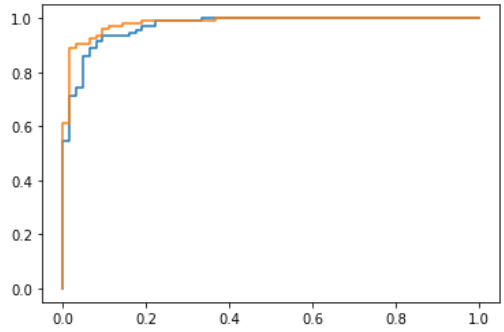
Vusal Ismayilov - 64190012 - Homework 5

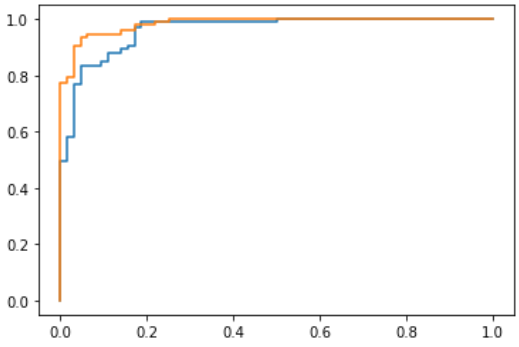
The renowned Breast cancer Wisconsin dataset is provided in this question, and we must design SVM and Naive Bayes classifiers for the given classification task. The development instructions are supplied, and they are improved in this solution for better data pretreatment and feature extraction in order to construct more optimum models.

The input and output values for SVM and Bayes functions are as follows:
SVM (c):
Input: 1 Output: 0.912
Input: 2 Output: 0.964
Input: 3 Output: 0.906

Bayes (var_smoothing):
Input: 1e-9 Output: 0.918
Input: 1e-8 Output: 0.976
Input: 1e-7 Output: 0.935







The Bayes model has the maximum classification accuracy, which is 97.6 percent.

Furthermore, because the classifier parameters are so important in predicting the evaluation measure, we tried to improve the code by using GridSearchCV, which is responsible for finding the optimal parameters for the task and optimizing the classifier parameters.

The regularization parameter c is used in SVM to control overfitting.

var smoothing is a portion of the biggest variance of all features is added to variances to ensure computation stability.

The most significant distinction between both the models you're developing in terms of "features" is that Naive Bayes sees them as standalone, however, SVM considers their connections to some extent, as long as you choose a non-linear kernel (Gaussian, RBF, poly, etc.). So, if you have relationships, as you very certainly do given your situation, an SVM will be better at capturing them, and thus at the classification task you need.

Rather than measuring absolute numbers, the AUC score assesses how well predictions are scored. AUC is insensitive to categorization thresholds. It assesses the accuracy of the model's predictions regardless of the categorization level used. The chance that a classifier would rate a randomly chosen positive example higher than a randomly chosen negative example is equal to the AUC of the classifier, i.e. P(score(x+)>score(x)).