CPSC 483 Assignment-3 Report - Tommy Vu
Assignment #3

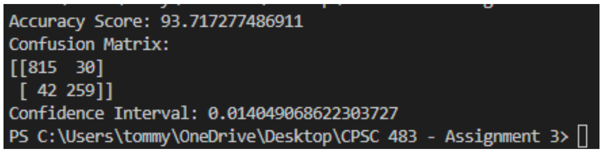
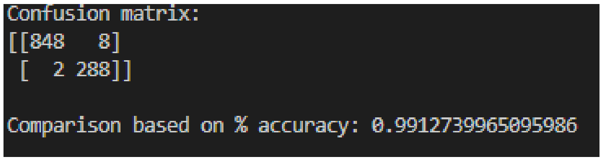
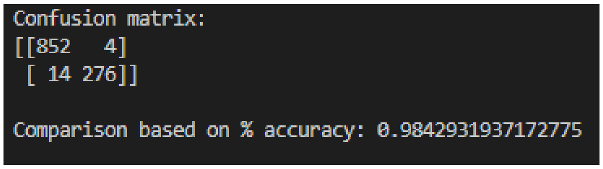
This assignment can be completed individually or by a team.
Total score: 100
Due date: 11/20

Objective: Text data classification

Write Python programs to classify the emails by Spam or non-spam category using the Naïve Bayes, KNN, and Support Vector Machine (SVM) and assess the classifier performance.

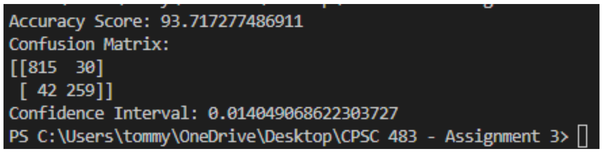
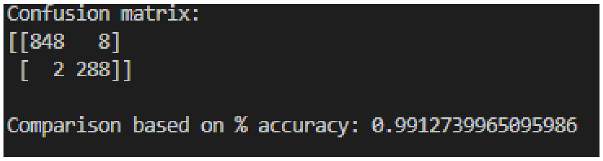
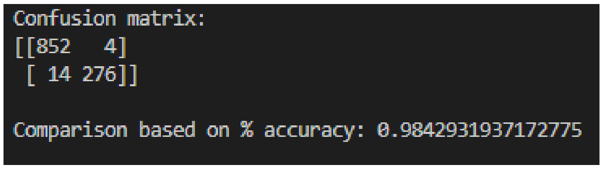
Only Python programs written using Python 3.0 or higher will be accepted. NO Jupyter
notebook or any Python variant will be accepted for efficient grading.

Write an analysis report including the following elements:
(1) Confusion matrix for each classifier

KNN	 Accuracy Score: 93.717277486911 Confusion Matrix: [[815 30] [42 259]] Confidence Interval: 0.014049068622303727 PS C:\Users\tommy\OneDrive\Desktop\CPSC 483 - Assignment 3>
Naïve-Bayes	 Confusion matrix: [[848 8] [2 288]] Comparison based on % accuracy: 0.9912739965095986
SVM	 Confusion matrix: [[852 4] [14 276]] Comparison based on % accuracy: 0.9842931937172775

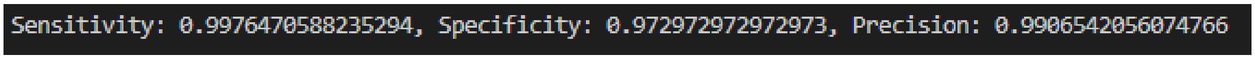
(2) Performance comparison based on the % accuracy

Comparing KNN, Naïve-Bayes, and SMV, the Naïve-Bayes and SVM are fairly accurate, whereas the KNN is the least accurate.

KNN has an accuracy score of ~93.72 which is fairly accurate.	 Accuracy Score: 93.717277486911 Confusion Matrix: [[815 30] [42 259]] Confidence Interval: 0.014049068622303727 PS C:\Users\tommy\OneDrive\Desktop\CPSC 483 - Assignment 3>
Naïve-Bayes has a % accuracy of 0.99127 which is very good.	 Confusion matrix: [[848 8] [2 288]] Comparison based on % accuracy: 0.9912739965095986
SVM has a % accuracy of 0.9842 which is also up there.	 Confusion matrix: [[852 4] [14 276]] Comparison based on % accuracy: 0.9842931937172775

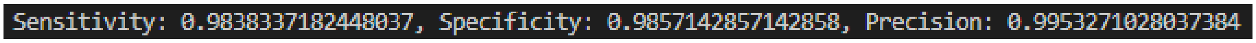
(3) Performance comparison based on sensitivity, specificity, and precision.

Naive-Bayes:



```
Sensitivity: 0.9976470588235294, Specificity: 0.972972972972973, Precision: 0.9906542056074766
```

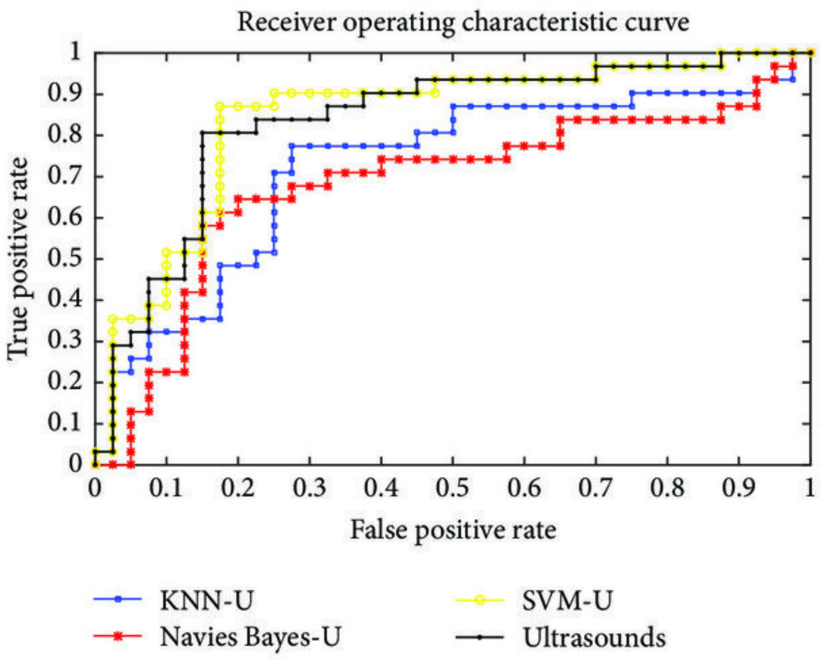
SVM:



```
Sensitivity: 0.9838337182448037, Specificity: 0.9857142857142858, Precision: 0.9953271028037384
```

Comparing the two based on sensitivity, specificity, and precision, it appears that Naive-Bayes beats SVM in sensitivity, whereas SVM beats Naive-Bayes in specificity and precision. Overall, both are fairly similar.

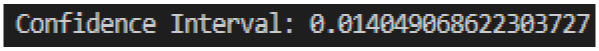
(4) Comparison of ROC curves



According to the graph, it can be seen that the most efficient ROC cruise is the SVM. It is mostly accurate throughout. However, as for KNN and Naives Bayes, they are fairly close. At the beginning,they are similar, then the Naive Bayes take over in accuracy, and towards the end, KNN is better. It can be seen that the KNN scaled better as the values get larger, whereas Naive Bayes is only very accurate for a moment. Overall, SVM is the most efficient.

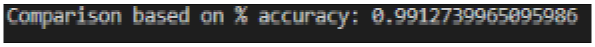
(5) Comparison of confidence interval (95%) for prediction accuracy

KNN:



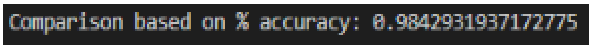
```
Confidence Interval: 0.014049068622303727
```

Naive-Bayes:



```
Comparison based on % accuracy: 0.9912739965095986
```

SVM:



```
Comparison based on % accuracy: 0.9842931937172775
```

Looking at the confidence intervals, it can be seen that the most accurate one would be the naive-bayes. SVM comes a close second and KNN comes in last.

(6) Your own conclusion based on the classifier performance and your recommendation. For example, which classifier would you recommend for Spam filtering and why?

Based on the results of this assignment, the Naïve Bayes is a good method for spam filter due to the fact that time costs little on training. Furthermore, testing an input message requires much time using Naïve Bayes, but the results are good enough for it to be used efficiently.