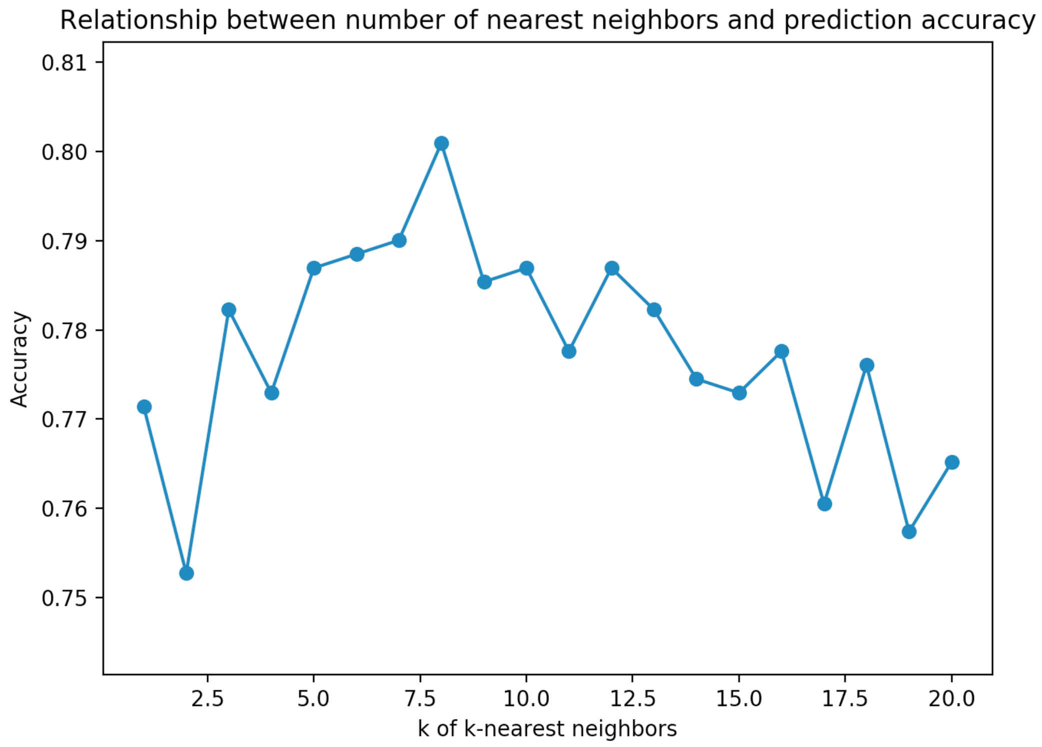
Elissa Ye

16791: Hw 2

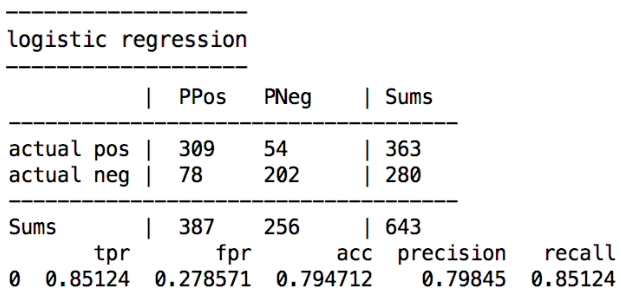
Part 4:
(a)



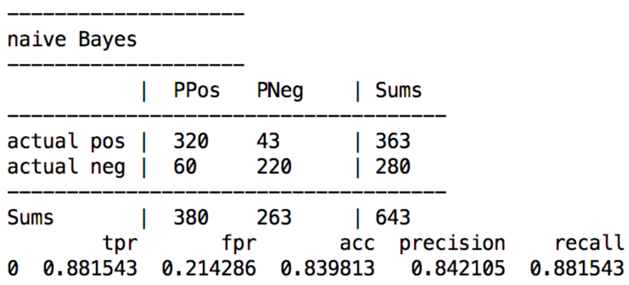
The accuracy of knn is maximized at 8-nearest neighbors, and thus knn can be best generalized for 8-nearest neighbors We can assume that from 1 to 7 neighbors the model maybe overfitting and beyond 8 neighbors the model may be underfitting.

(b)

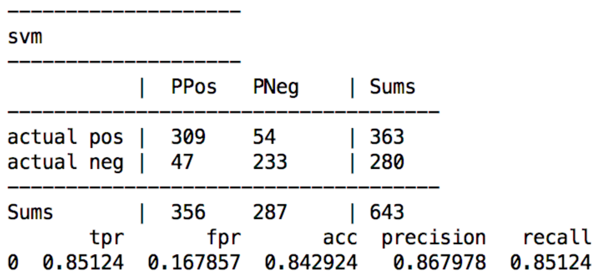
The metrics for logistic regression, naïve Bayes, SVM, and the default model are summarized below.



```
--------------------
logistic regression
--------------------
           |  PPos   PNeg    | Sums
--------------------------------------
actual pos |  309    54      | 363
actual neg |  78     202     | 280
--------------------------------------
Sums       |  387    256     | 643
       tpr       fpr       acc  precision   recall
0  0.85124  0.278571  0.794712    0.79845  0.85124
```

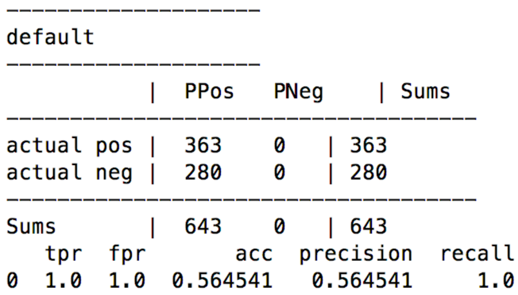


```
--------------------
naive Bayes
--------------------
           |  PPos   PNeg    | Sums
-------------------------------------
actual pos |  320    43      | 363
actual neg |  60     220     | 280
-------------------------------------
Sums       |  380    263     | 643
        tpr       fpr       acc  precision    recall
0  0.881543  0.214286  0.839813   0.842105  0.881543
```

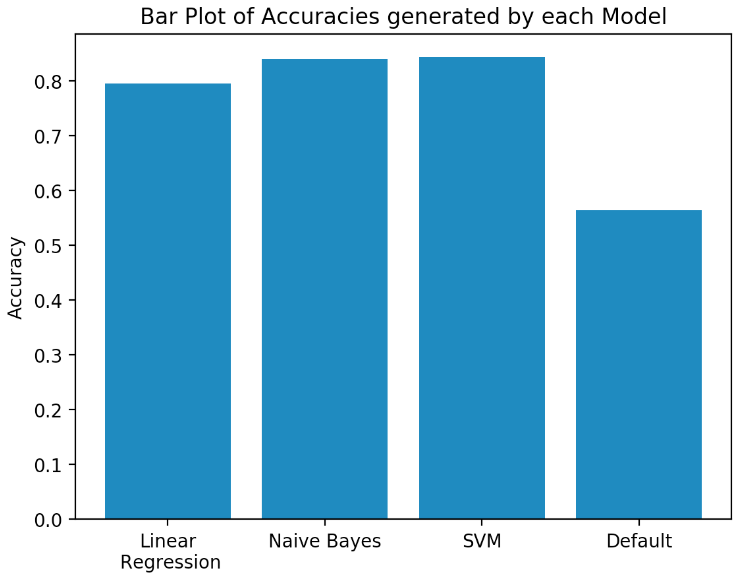


```
--------------------
svm
--------------------
           |  PPos   PNeg    | Sums
-------------------------------------
actual pos |  309    54      | 363
actual neg |  47     233     | 280
-------------------------------------
Sums       |  356    287     | 643
       tpr       fpr       acc  precision    recall
0  0.85124  0.167857  0.842924   0.867978  0.85124
```

Note: SVM is set to kernel ‘rbf’ and C=100 (Penalty parameter C of the error term). These parameters for SVM seem to optimize accuracy for this data in SVM.

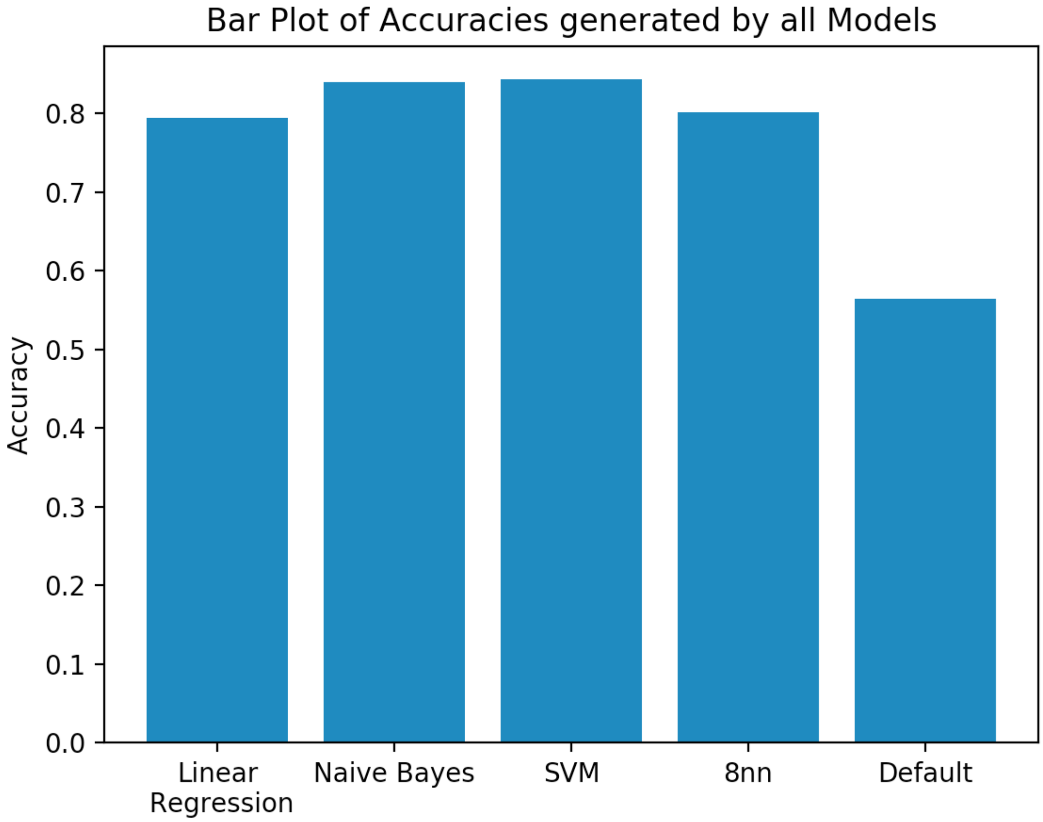


```
--------------------
default
--------------------
           |  PPos   PNeg    | Sums
-------------------------------------
actual pos |  363    0    | 363
actual neg |  280    0    | 280
-------------------------------------
Sums       |  643    0    | 643
   tpr  fpr       acc  precision  recall
0  1.0  1.0  0.564541   0.564541     1.0
```



Among these 4 models, the SVM model has the highest accuracy. The accuracy of the default model is around 0.56.

(c)



The SVM model seems to work best for this data and is closely followed by the Naïve Bayes model.