CS 498 AML: Homework 1

Rauhul Varma (rvarma2)

Problem 3.1

a. Building and running a Naïve-Bayes classifier on the Pima Indians Diabetes dataset yielded the following results over 10 folds:

```
training score (μ, σ): (0.7530081, 0.007002156)
testing score (μ, σ): (0.7535948, 0.03304803)
```

b. Building and running a Naïve-Bayes classifier which removes 0 entries on the Pima Indians Diabetes dataset yielded the following results over 10 folds:

```
training score (μ, σ): (0.7525203, 0.009030533)
testing score (μ, σ): (0.7346405, 0.0426036)
```

c. Building and running a Naïves-Bayes classifier using the klaR and caret packages on the Pima Indians Diabetes dataset yielded the following result confusion matrix and statistics:

```
          Reference
Prediction  0  1
         0 87 24
         1 13 29

               Accuracy : 0.7582
                 95% CI : (0.6824, 0.8237)
    No Information Rate : 0.6536
    P-Value [Acc > NIR] : 0.003479

                  Kappa : 0.4386
 Mcnemar's Test P-Value : 0.100178

            Sensitivity : 0.8700
            Specificity : 0.5472
         Pos Pred Value : 0.7838
         Neg Pred Value : 0.6905
             Prevalence : 0.6536
         Detection Rate : 0.5686
   Detection Prevalence : 0.7255
      Balanced Accuracy : 0.7086

       'Positive' Class : 0
```

d. Building and running an SVM classifier using SVMLight yielded the following accuracy:

```
testing score: 0.751634
```

Problem 3.3

a. Building and running a Naïve-Bayes classifier on the Processed Cleveland dataset yielded the following results over 10 folds:

```
testing score (μ, σ): (0.8226154, 0.06284531)
```

b. Building and running a Naïve-Bayes classifier on the Processed Cleveland dataset for predicting the values of the disease attributes (0-4) yielded the following results over 10 folds:

```
testing score (μ, σ): (0.6279, 0.07202452)
```

and the following confusion matrix and statistics:

```
            Reference
Prediction  0  1  2  3  4
         0 23  4  0  0  0
         1  1  2  2  3  0
         2  0  1  2  2  0
         3  0  1  1  0  1
         4  0  0  0  0  0

Overall Statistics

               Accuracy : 0.6279
                 95% CI : (0.4673, 0.7702)
    No Information Rate : 0.5581
    P-Value [Acc > NIR] : 0.2223

                  Kappa : 0.3728
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.9583  0.25000  0.40000  0.00000  0.00000
Specificity            0.7895  0.82857  0.92105  0.92105  1.00000
Pos Pred Value         0.8519  0.25000  0.40000  0.00000      NaN
Neg Pred Value         0.9375  0.82857  0.92105  0.87500  0.97674
Prevalence             0.5581  0.18605  0.11628  0.11628  0.02326
Detection Rate         0.5349  0.04651  0.04651  0.00000  0.00000
Detection Prevalence   0.6279  0.18605  0.11628  0.06977  0.00000
Balanced Accuracy      0.8739  0.53929  0.66053  0.46053  0.50000
```