Exercise 9 WriteUp

In this exercise, we implemented two different classification algorithms: decision tree and Naive Bayes. Upon training each model with a 80/20 training testing split, we then tested the models, produced confusion matrices of those results and compared the models effectiveness. There is no such thing as a perfect model, and even given the same data set, you will get different results. Here are the accuracies confusion matrices generated for each model:

##### Decision Tree (Acc: .79):

|  |  |
| --- | --- |
| True Negative: 934 | False Positive: 677 |
| False Negative: 664 | True Positive: 4237 |

##### Naive Bayes (Acc: .79):

|  |  |
| --- | --- |
| True Negative: 499 | False Positive: 1112 |
| False Negative: 244 | True Positive: 4657 |

What is interesting about these results is that, for this specific train/test split, we actually get the same accuracy. Accuracy is a composite of multiple things, however, so by looking at the confusion matrices we can, ironically enough, clarify the differences between the two models. Our decision tree was much better at avoiding false positives while the naive Bayes model had significantly more false positives, but less false negatives. For this data set when we are attempting to predict income, it is not necessarily clear whether we are more concerned about protecting against false negatives or false positives.

If, however, this data set instead was attempting to predict whether or not a patient had heart disease, we would certainly want to favor the Bayes model. This is because when dealing with a sensitive prediction like a diagnosis, we would much rather make the mistake of telling someone they do have heart disease when they don’t then telling someone who is sick they are healthy. As a result, even with the same accuracy measure, one model can be clearly preferred to another.