David Strickland

CS5402 Intro to Data Mining

Homework 3

Python file located at the following github repository: https://github.com/ds5b4/CS5402-IntroToDataMining-Fall2018/tree/master/HW3

Question 1:

Accuracy Rating: 0.75

Precision: 0.9

Recall: 0.8181818181818182

F1: 0.8571428571428572

Question 2:

['Win' 'Win' 'Win' 'Lose' 'Win' 'Lose' 'Win' 'Win' 'Win' 'Lose' 'Win' 'Lose']

Question 3:

In this instance Naïve Bayes performs worse than ID3 as all 4 metrics accuracy, precision, recall, and F1 are lower. The values for ID3 are 0.91666, 0.91666, 1.0, and 0.956521 for accuracy, precision, recall, and F1 respectively. This is most likely due to two factors. Naïve Bayes operates on the assumption that all the features are completely independent. In this case the media feature may be related the Is\_Home\_or\_Away feature, making this assumption incorrect and causing unintended weight to be placed on these two features. Additionally, Naïve Bayes works best with large train datasets, and while the train dataset is larger than the test dataset in this example, it is much smaller than many other datasets, not even containing 50 data points. Without sufficient training, the closer to random, the results will be, as seen by the .75 accuracy rating.

\*Comparison results are the same results from HW2 and as such I have no results for C4.5 to measure on other than the general knowledge that C4.5 is generally better than ID3