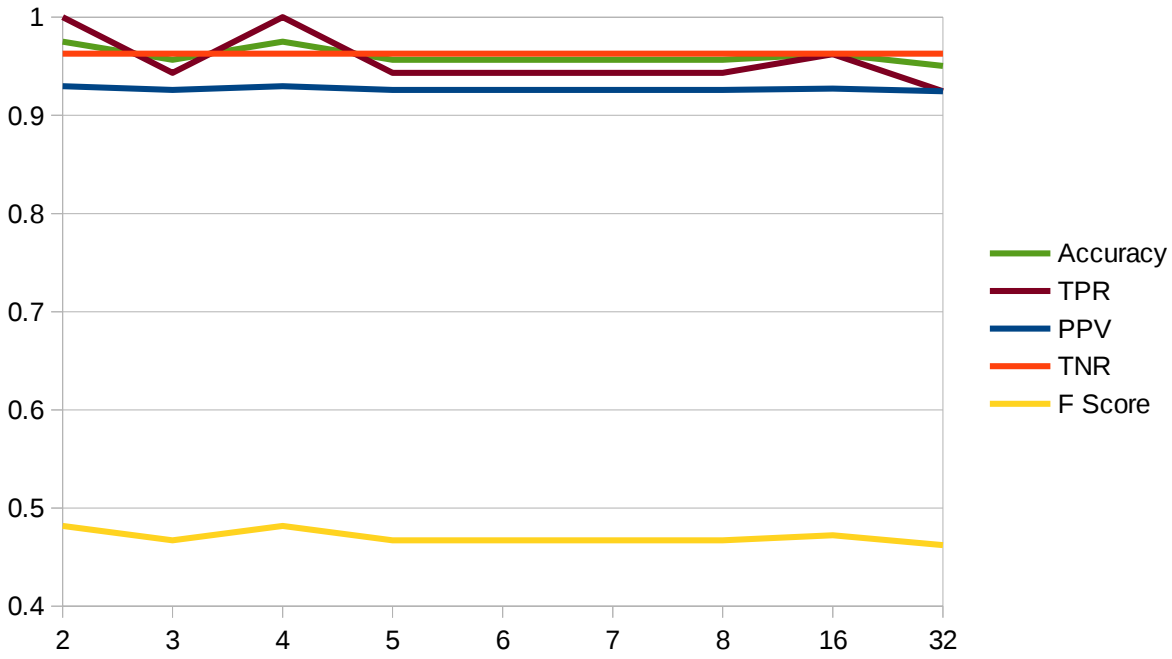
Jacob Vargo

CS 425 Intro to Machine Learning

Project 3

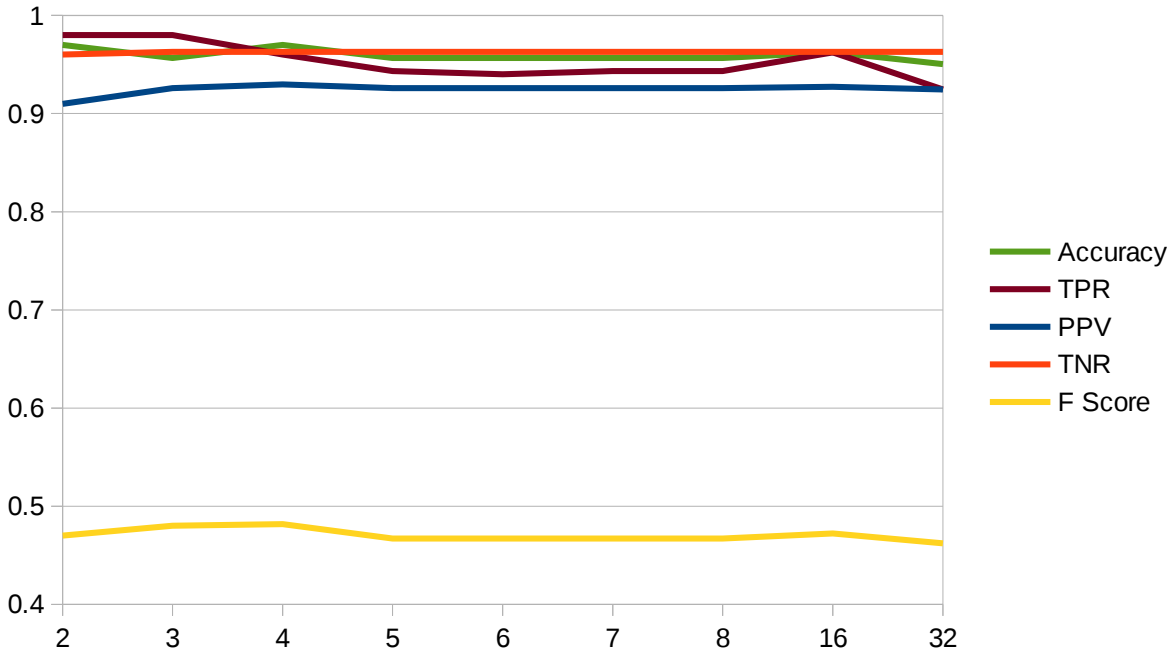
While preprocessing the raw data I found there were sixteen instances of missing data entries that were marked with a '?'. I decided the best way to handle the missing data entries was to replace the values with a '1' so as to be within range of the normal entries. I then split my data into 3 groups: test, validation, training. I put half of the data in the training set and gave both the validation and test sets a quarter of the data.

In my k-Nearest Neighbor algorithm, I used the training set to find the nearest neighbors and then get a prediction. In the case that the number of nearest neighbors that were malignant equaled those that were benign, I chose to assign the prediction to be malignant. I then used the validation set to find the best suited k for the algorithm.



Based on my data, I decided that 4 would be my best choice for k since is tied with k=2 with the highest rating in every performance meter and having a higher k allows the algorithm to be more robust. My performance ratings on the test set were: Accuracy = 0.99, TPR = 0.98, PPV = 0.98, TNR = 0.99, F Score = 0.49.

In my decision tree classifier, I used the training data to construct the decision tree. I then used the validation set to find the best maximum tree depth.



Based on my results, I decided that a maximum depth of 3 would be my best choice. My performance ratings on the test set were: Accuracy = 0.98, TPR = 0.97, PPV = 0.97, TNR = 0.99, F Score = 0.49.