Alex Ho
ID: 1001294712
CSE 4334-001

Assignment 2 Report

## Problem 1:

### Part 1:

The hardest part about part 1 was understanding how to implement the Naïve Bayes Classifier for a 2D Gaussian dataset, but once I figured it out, the implementation was easy. I first got the priors for class 0 and class 1. I then found the mean and std for all X and Y's of class 0 and class 1 in the training data. I then used normpdf() on the test data with the mean and standard deviation of class 0 and 1. After I multiplied the X and Y values together, I found the posterior by getting the higher probability between class 0 and class 1. Finally, I set the predictor labels to their respective classes and set the error value to the number of false labels divided by the size of the test data.

**Part 2:**

**Accuracy**: 92.10%

**Positive = Class 0**

| | Actual Positive | Actual Negative |
|---|---|---|
| Predicted Positive | 462 | 41 |
| Predicted Negative | 38 | 459 |

Precision: 91.85%

Recall: 92.40%

**Positive = Class 1**

| | Actual Positive | Actual Negative |
|---|---|---|
| Predicted Positive | 459 | 38 |
| Predicted Negative | 41 | 462 |

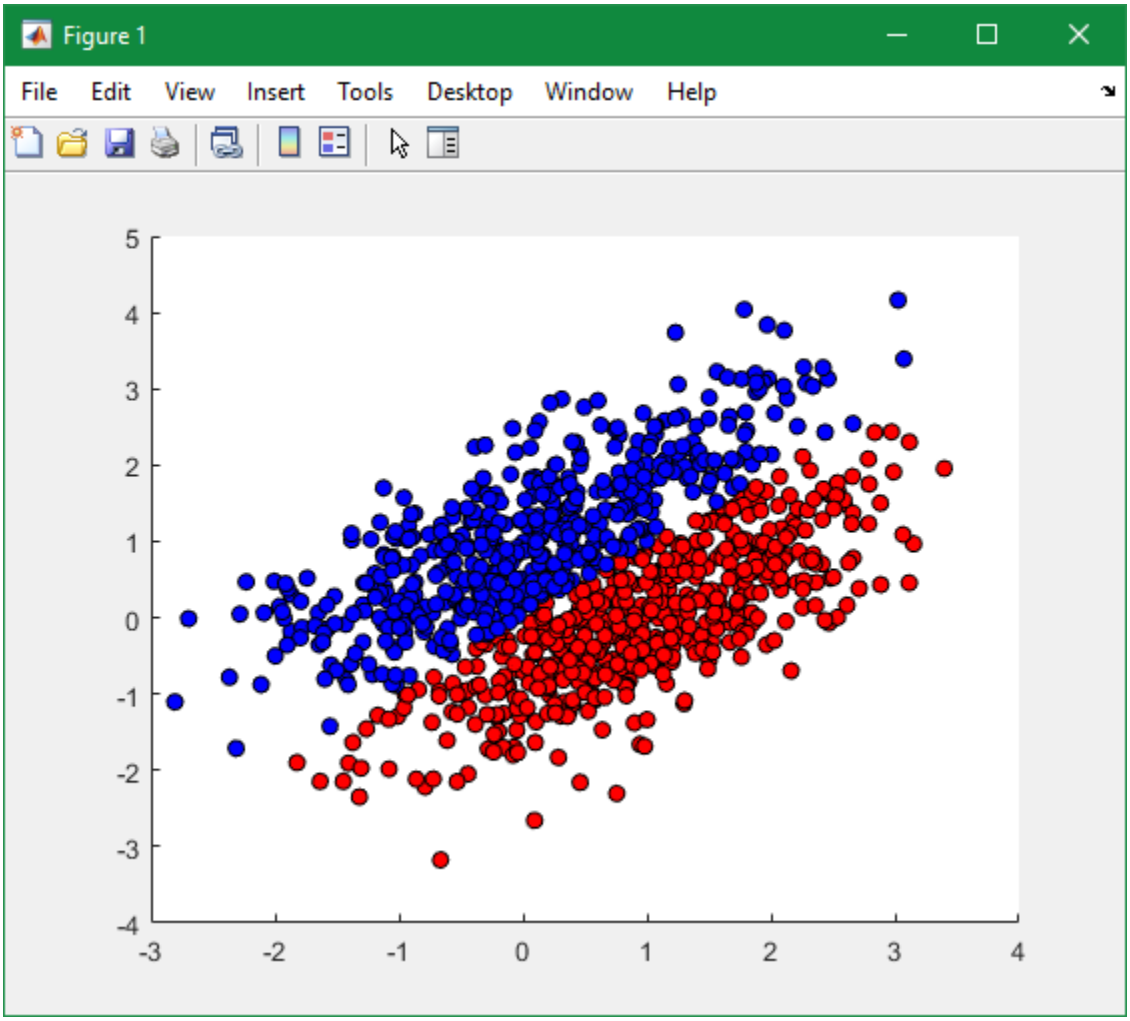
Precision: 92.35%

Recall: 91.80%

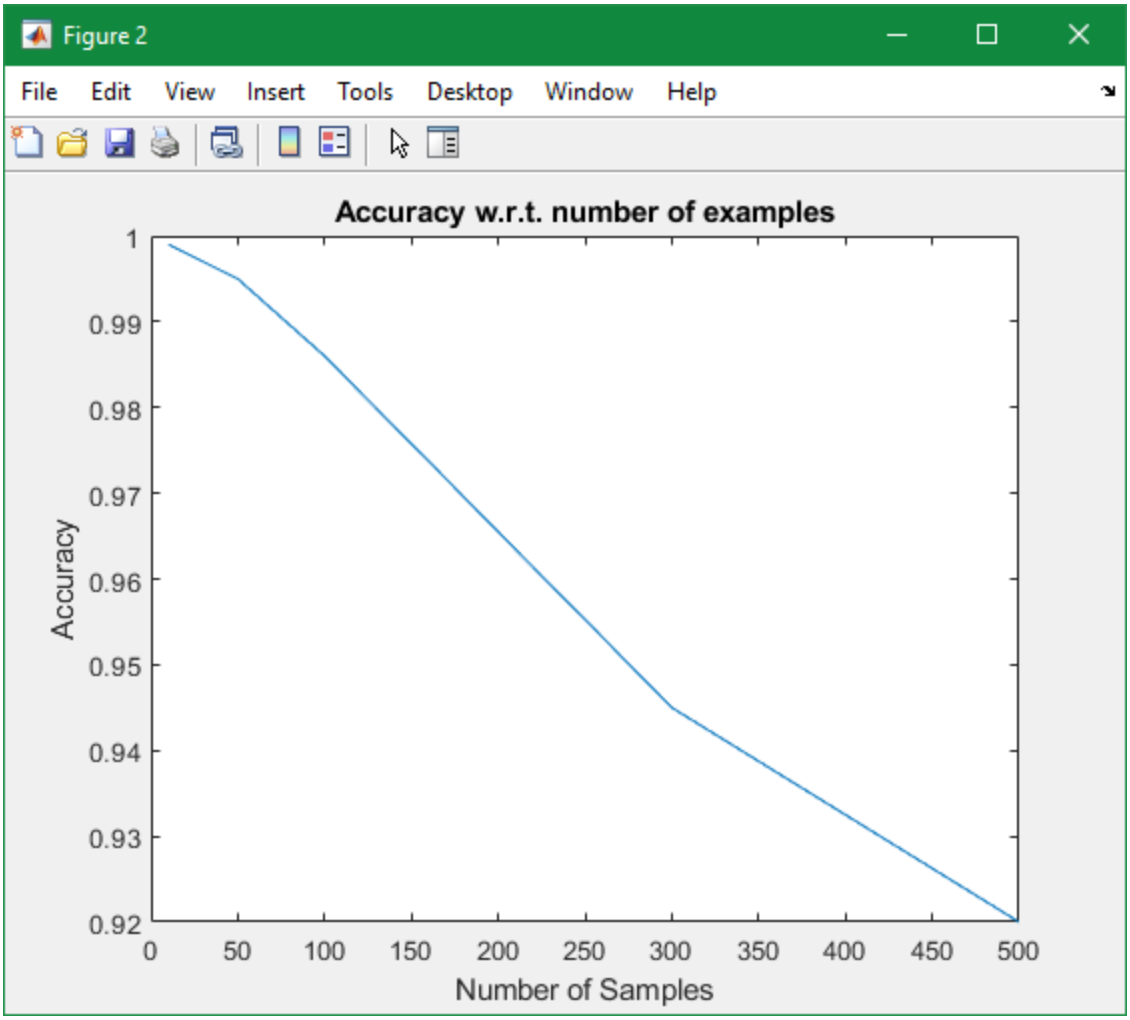
Class 0: Blue Dots

Class 1: Red Dots

**Part 1 Plotted:**



**Part 3:**




In part 3, there was a decrease in accuracy as the number of test samples increased. This is because with more samples, there is a higher chance to have points that are incorrectly classified at boundaries.

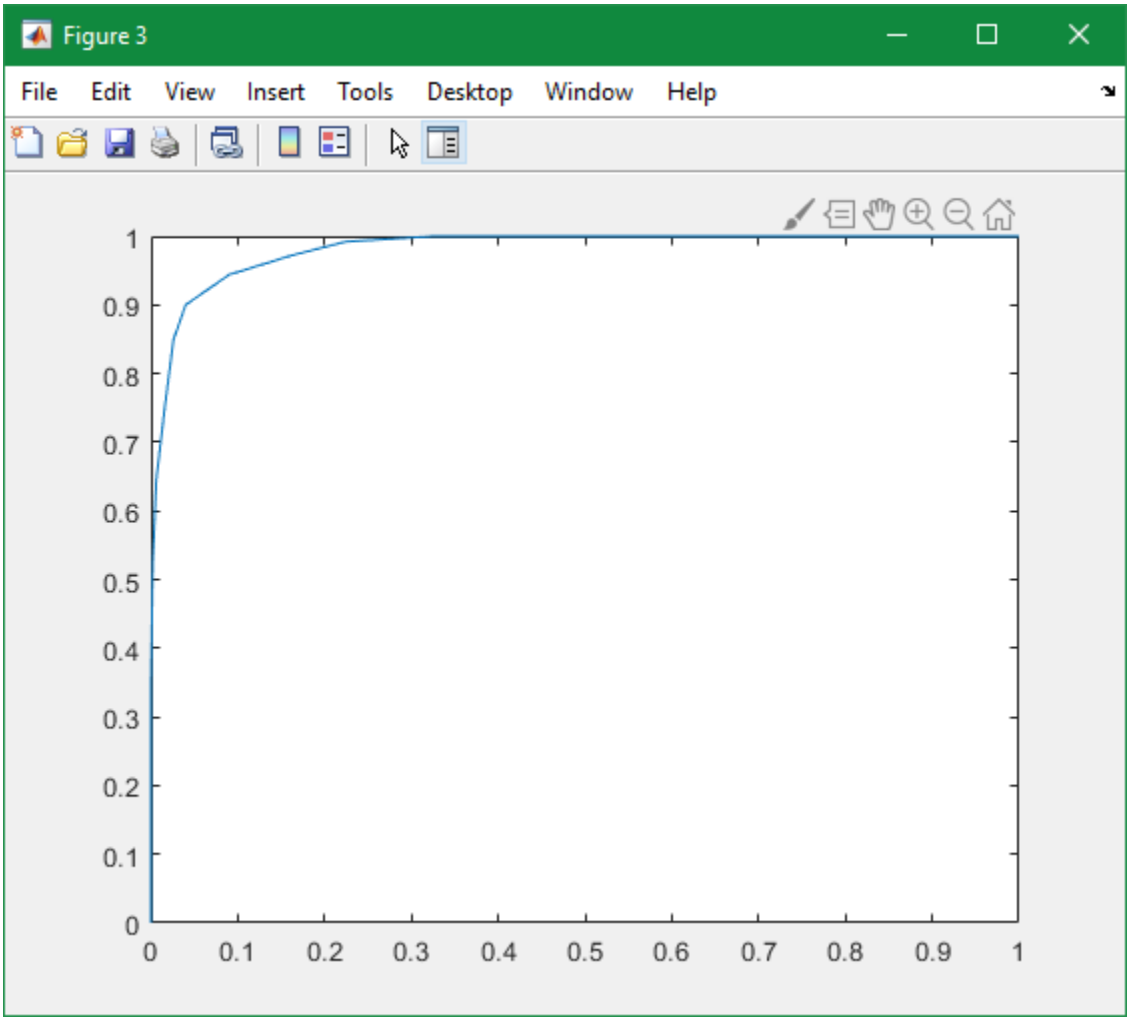
**Part 4:**

When the training data was changed to 700 of class 0 and 300 of class 1, the accuracy lowered significantly (500/500 split was 91.40%, 700/300 split was 78.90% on same testing data). This is because with insufficient examples, the model will overfit as it cannot predict the lacking region as accurately. There is also a higher chance for class 0 training data to be occur more often in the area that class 1 should be.

**Part 5:**

ROC of P1-2:

AUC: 0.6998



ROC of P1-4:

AUC 0.5388