Programming Assignment 3

Members:

Erik Hale (netid: emh170004)

Shiva Kumar( netid: sak220007)

Part a)

Accuracy (Error rate) , F-Score, Precision and Recall on Test Set:

```
{'accuracy': 0.04832474226804124, 'precision': 0.9310344827586208, 'recall': 0.9,
 'f1_score': 0.9152542372881356}
```

Top Three words for Spam:

```
For Spam, the Top three Most Words are:
Word:  e
Class-conditional likelihoods 0.019965486051210533
Log-Likelihood -3.9137501935987267


Word:  t
Class-conditional likelihoods 0.014577763648539254
Log-Likelihood -4.228257949040139


Word:  a
Class-conditional likelihoods 0.01393300342658023
Log-Likelihood -4.273494905918018
```

Top Three words for No Spam:

```
For No-Spam, the Top three Most Words are:
Word:  e
Class-conditional likelihoods 0.05302573376325244
Log-Likelihood -2.936977940564855


Word:  t
Class-conditional likelihoods 0.03755363218906419
Log-Likelihood -3.281985176047472


Word:  a
Class-conditional likelihoods 0.0336552525510734
Log-Likelihood -3.391586141430148
```

Part b)

Test Sentence 1:

```
Posterior likelihood for Spam:  0.019798421397929738
Posterior likelihood for Not Spam:  0.05044974369435297
Log Posterior Likelihood for Spam  -130.5233661281371
Log Posterior Likelihood for Not Spam  -134.1042047089683
Sample Email: Congratulations! Your raffle ticket has won yourself a house
. Click on the link to avail prize
Is Spam
```

Test Sentence 2:

```
Posterior likelihood for Spam:  0.005122393516892269
Posterior likelihood for Not Spam:  0.013022268364794785
Log Posterior Likelihood for Spam  -113.15448152900261
Log Posterior Likelihood for Not Spam  -110.85224133145476
Sample Text: Hello. This email is to remind you that your project needs to
 be submitted this week
 Is Not Spam
```

Test Sentence 3:

```
Posterior likelihood for Spam:  0.01489812422461753
Posterior likelihood for Not Spam:  0.03637255350371147
Log Posterior Likelihood for Spam  -52.623353933645376
Log Posterior Likelihood for Not Spam  -53.84551001499095
Sample Email: Congrats! Click Link for a New Car !
Is Spam
```

Test Sentence 4:

```
Posterior likelihood for Spam:  0.016191508813701426
Posterior likelihood for Not Spam:  0.03480742475651997
Log Posterior Likelihood for Spam  -76.94694669628703
Log Posterior Likelihood for Not Spam  -63.26732359115838
Sample Text: Thanks! I forwarded the attached file to my boss.
 Is Not Spam
```

## Part c)

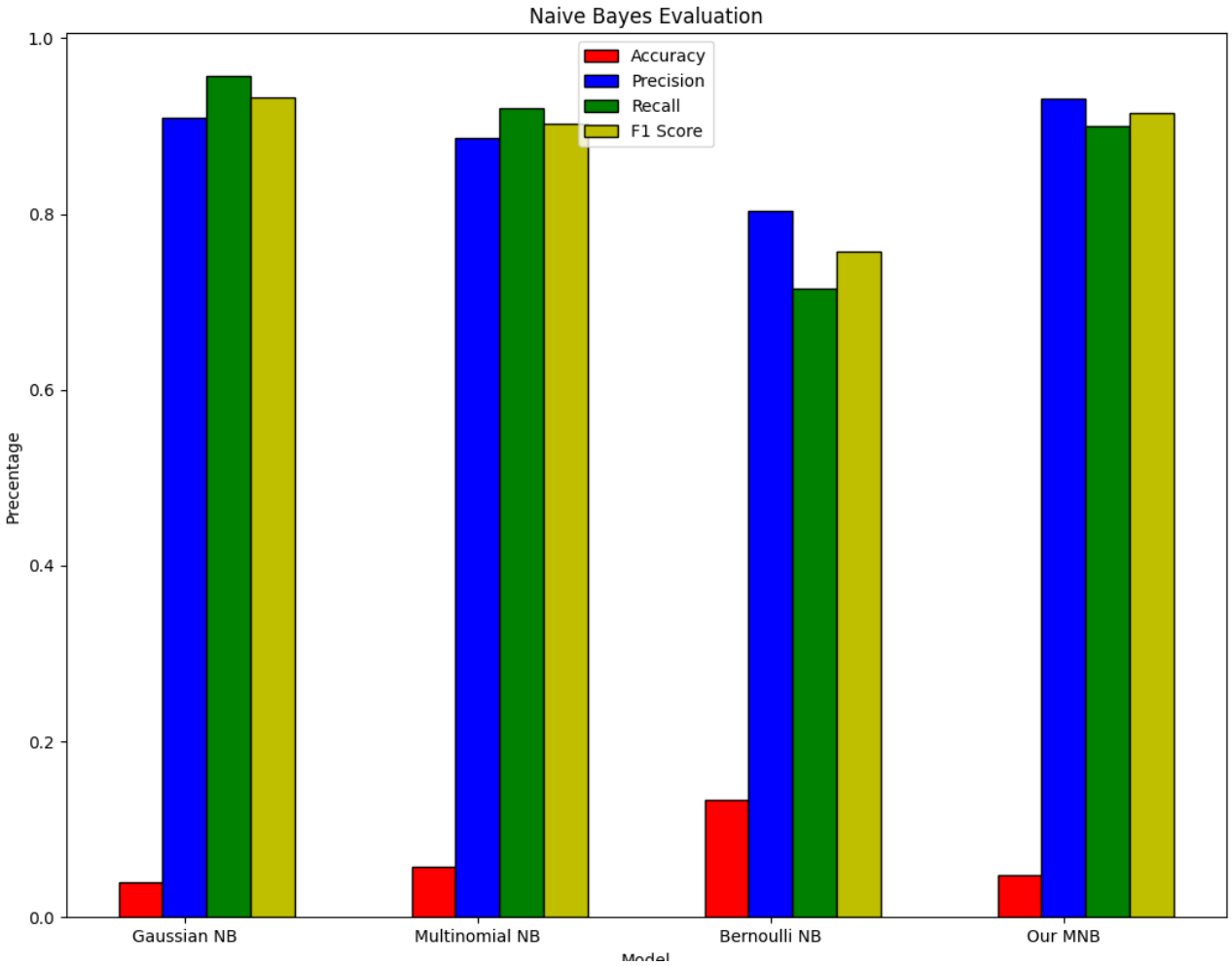
Based off the data collected, the models that had the highest F1 score also had high precision values and recall but low accuracy scores. The Gaussian Naïve Bayes model has the highest F1 score of ~93.3%, accuracy is ~4%, and a precision is ~90.9% while the lowest model has an F1 score of ~45.0%, accuracy is ~71%, and a precision of ~29.0%. It is important to note that although our model has the highest precision, even higher than that of the Gaussian model, it still has a lower F1 score because both the Precision and Recall are used explicitly in the equation F1 = 2 * ((P * R) / (P + R)), in which the Gaussian model has the better statistics.

The evaluation of each model is not a surprise when we consider how each method functions. Bernoulli NBs use a binomial distribution meaning that words are classified as true or false depending on if it is present in the email, Multinomial NBs do a better job of representing what data is in the emails by counting the number of times a word occurs / the frequency of the word, giving a more accurate prediction of a spam vs non-spam email. Gaussian NBs uses a normal distribution to predict if an email is spam, it does this well because unlike the other two models it takes in continuous data to formulate if an email is spam, and thus provides the best f1 score of the models.

For whether to maximize recall or precision, it really comes down to how strict you want the system to be. In the case that we would want to place more emphasis on getting our normal emails (and if we get one or two in our normal inbox then that is okay) then we would want to minimize the False Positives (meaning it was not spam and it was labeled as spam) meaning we should focus more on precision. If we wanted to make sure that every spam email goes in the spam folder (and at the expense of normal emails being spam), then we focus on minimizing the false negatives (say if an email was spam and was not labeled as spam), then we would want to maximize the recall. The F1 score takes both precision and recall into account which allows the best of both worlds as a higher F1 score usually means higher precision and recall scores.

| | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
|---|---|---|---|---|
| Gaussian | 3.99% | 90.93% | 95.77% | 93.29% |
| Multinomial | 5.73% | 88.65% | 92.29 | 90.29% |
| Bernoulli | 13.34% | 80.30% | 71.56% | 75.68% |
| Our MNB | 4.83% | 93.10% | 90% | 91.53% |

## Part d)




## Part e)

The Pickle file is included in the submission