Name: Kumar Raju Bandi
Roll Number: IWM2017502
Google Colab Link:

1. https://colab.research.google.com/drive/1qF5Cm5C-XvF5zxl8BU-UYaT2ci-f21Nx?usp=sharing
2. https://colab.research.google.com/drive/1PH27aGgxDMHE_VHRPj0efXLZ2HwdnK3i?usp=sharing

Analysis Report:
There are three types of Naive Bayes Classifiers which have been implemented in this assignment :

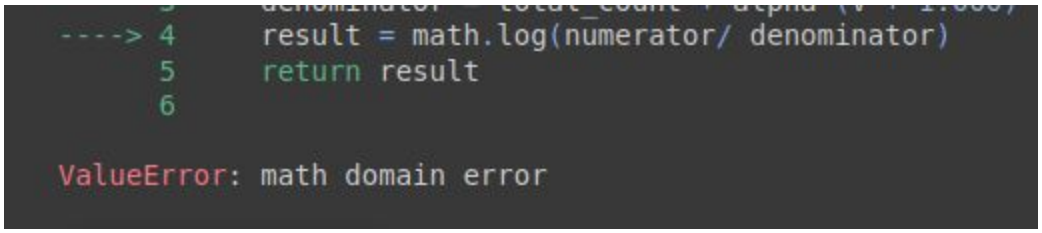
1. Multivariate Bernoulli Naive Bayes
2. MultiNomial Naive Bayes
3. Gaussian Naive Bayes

The data was also pre-processed with the following steps and analysed after each step:

a) Remove punctuation
b) Remove not relevant special characters
c) Strip extra spaces for each sample
d) Calculate word count for each sample
e) Remove repeated entries if exists

The frames can be seen in the ipynb file too.
I also set the train-test split to 0.75 to get better and justified results.

While implementing, I faced problems like zero-division error:



To overcome this, I had implemented Laplace smoothing for both MultiVariate Bernoulli Naive Bayes and MultiNomial Naive Bayes and this eliminated the error.
I also converted to formula to log-form so that the multiplications operations do not overflow.
I also used a “<UNK>” token to capture words which have not been present in the train data.
The above three things are something extra I have implemented than usual so as to get better results and lesser implementation errors.

The results I got are as follows:
For Multivariate Bernoulli Naive Bayes I am getting an accuracy of 95.83%
For MultiNomial Naive Bayes, I am getting an accuracy of 96.76%.
And lastly for Gaussian Naive Bayes I am getting an accuracy of 90.74%

The probable reason for the MultiNomial Naive Bayes classifier giving the highest accuracy is because the distribution of word counts is Multinomial in nature and it captures the spam features better than Multivariate Bernoulli Naive Bayes.

The probable reason for lowest accuracy for Gaussian Naive Bayes among all the three is that the data is not in normal distribution in nature which is the base assumption to apply Gaussian Naive Bayes Classification method.

I also analysed the tokens(words) which occur the most frequently in spam messages with respect to MultiNomial Naive Bayes and the results were as follows. I extracted the top 5 spam words:

```
Most indicative spam tokens:
['16', 'urgent', '100', '2000', 'nokia']
```

I also analysed the tokens(words) which occur the most frequently in spam messages with respect to Multivariate Bernoulli Naive Bayes and the results were as follows. I extracted the top 5 spam words:

```
Most indicative spam tokens:
['16', 'urgent', '100', '2000', 'code']
```

Overall even though Naive Bayes classifiers are bound to have high bias and less variance, the accuracy achieved is high enough for all three types of models to be used and to be considered good performance.