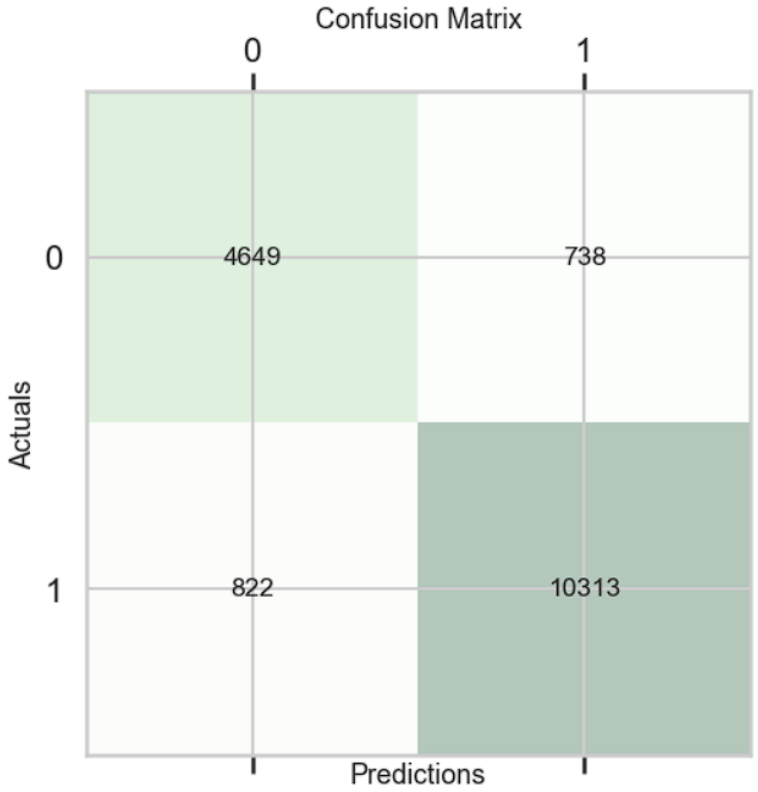
Pamela Gayle B. Domingo
BS in Computer Science IV

HW4

Link to GitHub: https://github.com/PamelaDomingo/homework-4/blob/main/hw4_domingo.ipynb

Guide questions:

1. What is the effect of removing stop words in terms of precision, recall, and accuracy? Show a plot or a table of these results.




```
Accuracy Score =  0.9055804382036073
Precision Score =  0.9332187132386209
Recall Score =  0.9261787157611137
```

Out of the 16522 emails of the test set, the model scored with 90.56% accuracy, 93.32% precision, and 92.62% recall. Additionally, the confusion matrix shows that the results of the prediction using the model there were 10313 True Positives, 7383 False Positives, 4649 True Negatives, and 822 False Negatives. By removing the stop words, the model was able to predict spam and ham emails using only the filtered unique words in the vocabulary.

2. Experiment on the number of words used for training. Filter the dictionary to include only words occurring more than k times (1000 words, then k > 100, and k = 50 times). For example, the word “offer” appears 150 times, that means that it will be included in the dictionary.

3. Discuss the results of the different parameters used for Lambda smoothing. Test it on 5 varying values of the λ (e.g. λ = 2.0, 1.0, 0.5, 0.1, 0.005), Evaluate performance metrics for each.

4. What are your recommendations to further improve the model?

   It will be ideal to use datasets that are in csv file format to make it easier to extract and clean the dataset. Furthermore, it is recommended that the model use a dictionary of real-life existing words to countercheck the vocabulary to ensure that the words being analyzed are actual words that humans can understand. Lastly, the model already produced high scores for accuracy, precision, and recall but it still needs faster algorithms to classify thousands of emails as spam or ham.