Machine Learning Assignment 1
Garvin P. Bulkin (927842)

Question 1:
From observing the results across the performance of both methods, we can see that for 2 out of 3 of the datasets tested, discretizing yields better results. For example, for the wine dataset (refer to appendix A), for each of the metrics, gaussian produced the lower results. It was 23% less accurate, and the F1 score is 50% higher when using non-gaussian methods. This means that discretizing these specific dataset leads to higher accuracy, and even simultaneously high precision and recall.

The phenomenon can be explained by the distribution of the datasets. Gaussian Naïve Bayes tend to perform well on data that conforms to Gaussian distribution, but real-life data does not always follow Gaussian distribution. The sample sizes for the wine and WDBC dataset are also relatively small, which makes it less likely to conform to normal distribution, adding to the reasons why the gaussian model did poorly.

It is also important to note that I did not use any outlier removal methods with the dataset. Since the 2 out of 3 datasets used are relatively small, the mean and standard deviation can be affected greatly by outliers. Further explaining the relatively poor performance of Gaussian Naïve Bayes, especially in the wine dataset (178 instances).

However, we can also see that for the adult dataset, the Gaussian Naïve Bayes performed better than the Naïve Bayes with the discretized attributes. This diminishing performance in the Naïve Bayes model can be attributed to the size of the dataset. The adult data has the largest number of instances (~30,000), making the distribution close to gaussian, hence why gaussian has performed better for this set.

To clearly see the visualization of the performance, I have created a confusion matrix for the accuracy of each class with both models. (Appendix A)

Question 2:
For this question, I have chosen to implement a baseline model which predicts the class label at random. You can find results and comparison between Naïve Bayes (non-gaussian) and the baseline model.

You can clearly see a trend between the accuracy of the baseline model and the number of class labels for each dataset. This is because the accuracy will be very close to the actual probability of getting an answer correct (1/# of possible labels). We can compare them side by side using the following table:

Dataset	# of Labels	Probability (1/labels)	Actual Accuracy
Wine	3	33.33%	38%
Mushroom	2	50%	50%
Car	5	20%	26%
Adult	2	50%	49%

We can also see that the F1 score is also following a similar pattern. If we repeat the baseline test with different seeds and average it, as we do more tests, it would converge to the probability value.

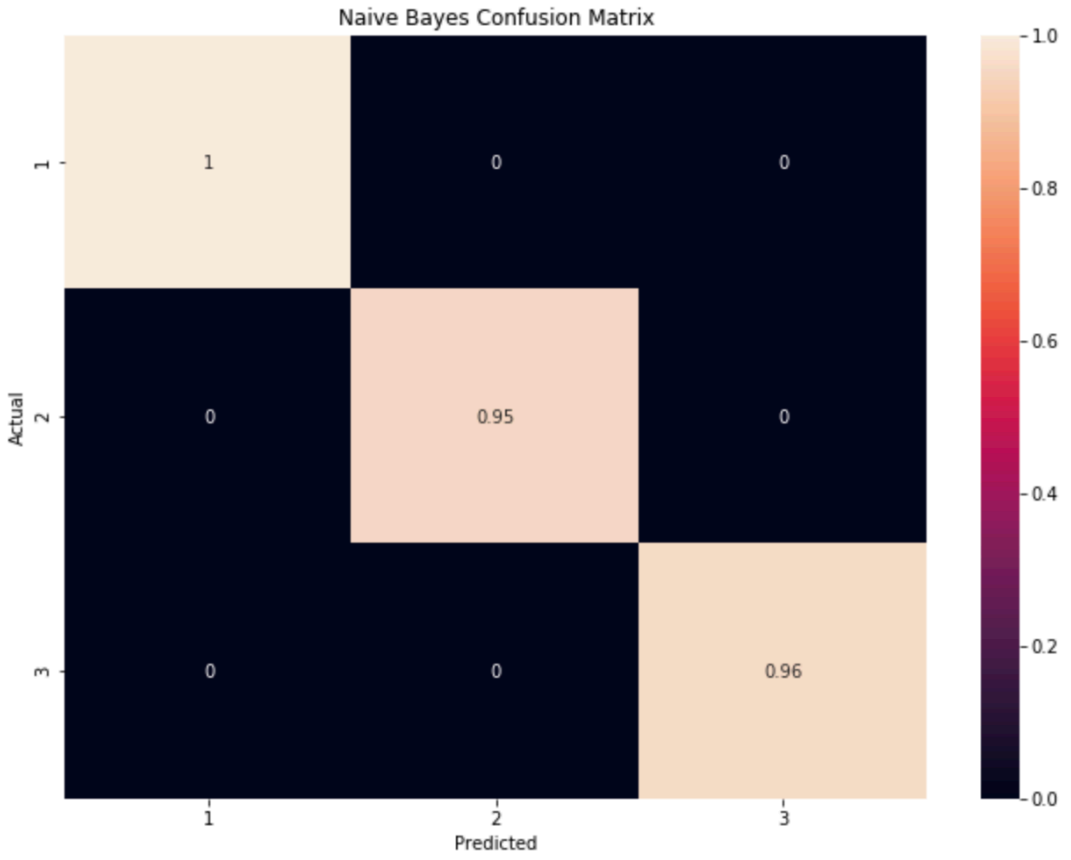
Relative to the baseline, my Naïve Bayes Model performed much better and is not affected greatly by the number of possible labels. Looking at the results in appendix B, it does not show any significant diminishing accuracy as the number of possible class labels increased. For example, the accuracy for mushroom and wine of my model has negligible difference (~2%), in contrast, the random baseline dropped 12% in accuracy. Even for datasets with 5 possible class label (car.data), the Naïve Bayes model is able to still produce 87% accuracy. This tells us that my Naïve Bayes model performs well under the stress of having an increasing number of possible class labels.

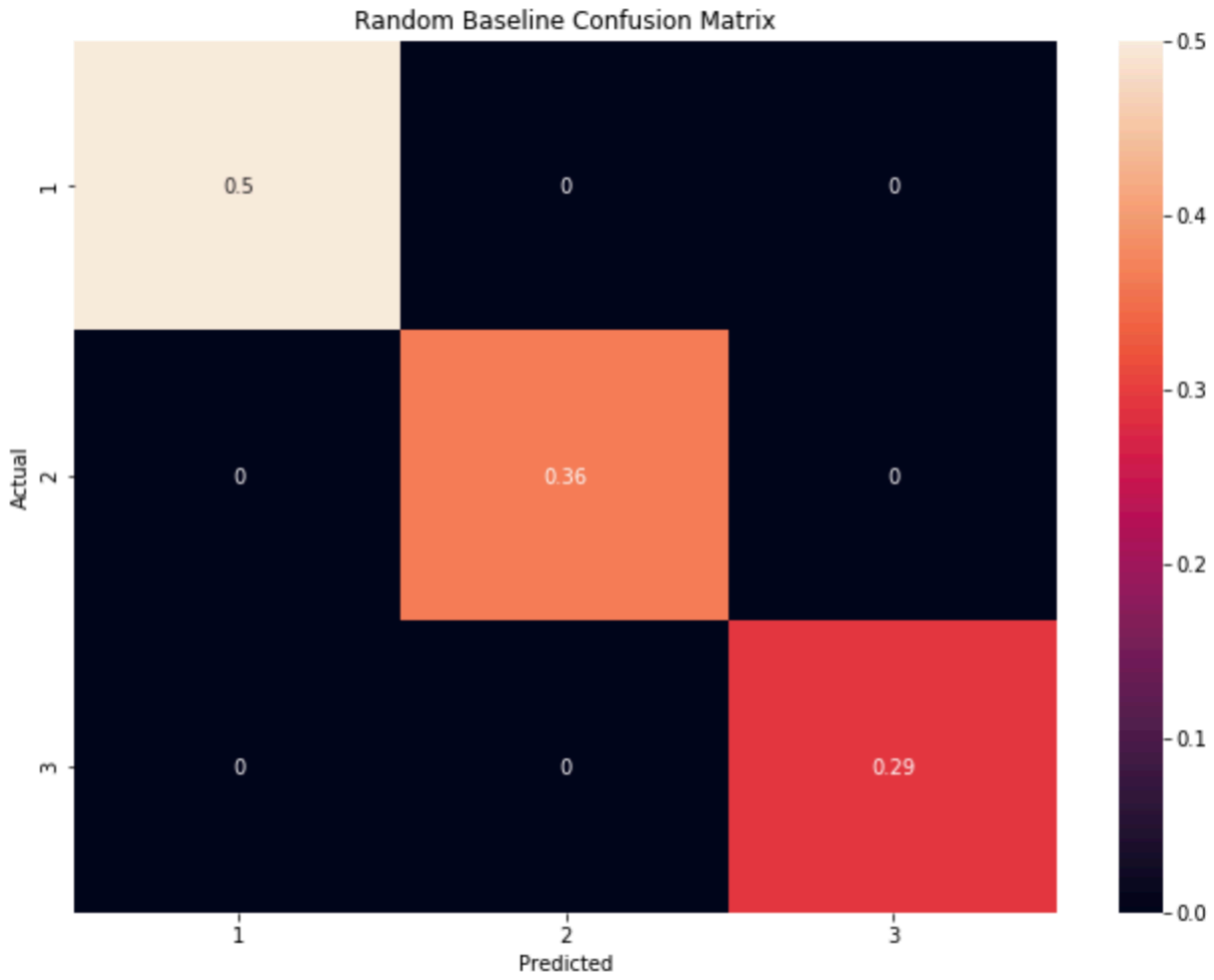
To summarize, the performance of the random baseline decreases with number of classes because it is more likely to provide an incorrect label. This type of baseline is useful for determining how your model performs as you increase the number of possible classes in the dataset.

Appendix

Appendix A: Naïve Bayes vs Gaussian Naïve Bayes evaluation results
Wine Dataset :

```
Gaussian Naive Bayes:
Accuracy: 72%
Error-rate: 28%
Precision: 73%
Recall: 56%
F1 Score: 0.63

Naive Bayes:
Accuracy: 98%
Error-rate: 2%
Precision: 99%
Recall: 96%
F1 Score: 0.97
```





WDBC dataset:

```
Gaussian Naive Bayes:
Accuracy: 71%
Error-rate: 29%
Precision: 77%
Recall: 80%
F1 Score: 0.78

Naive Bayes:
Accuracy: 95%
Error-rate: 5%
Precision: 95%
Recall: 97%
F1 Score: 0.96
```

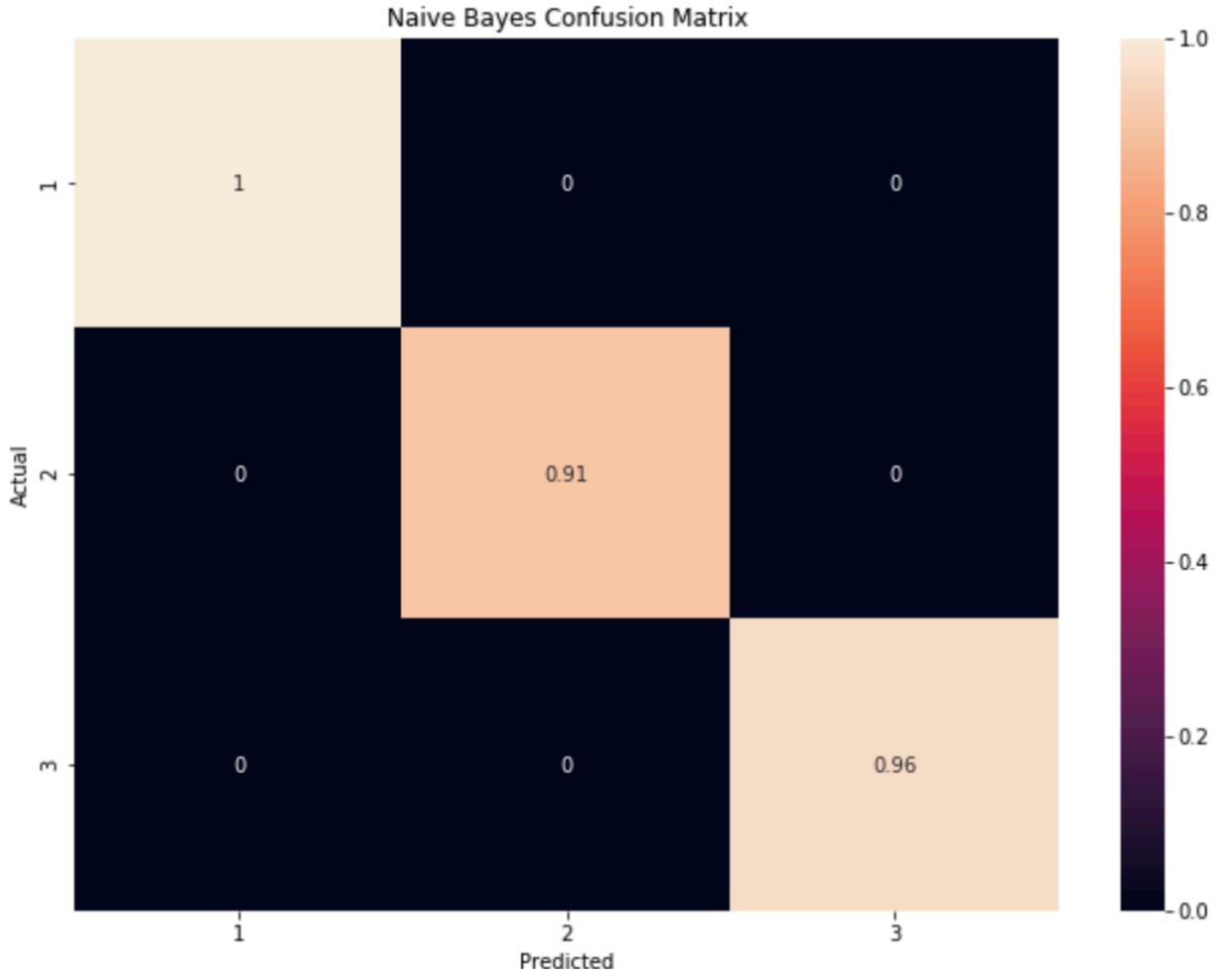
Adult dataset:

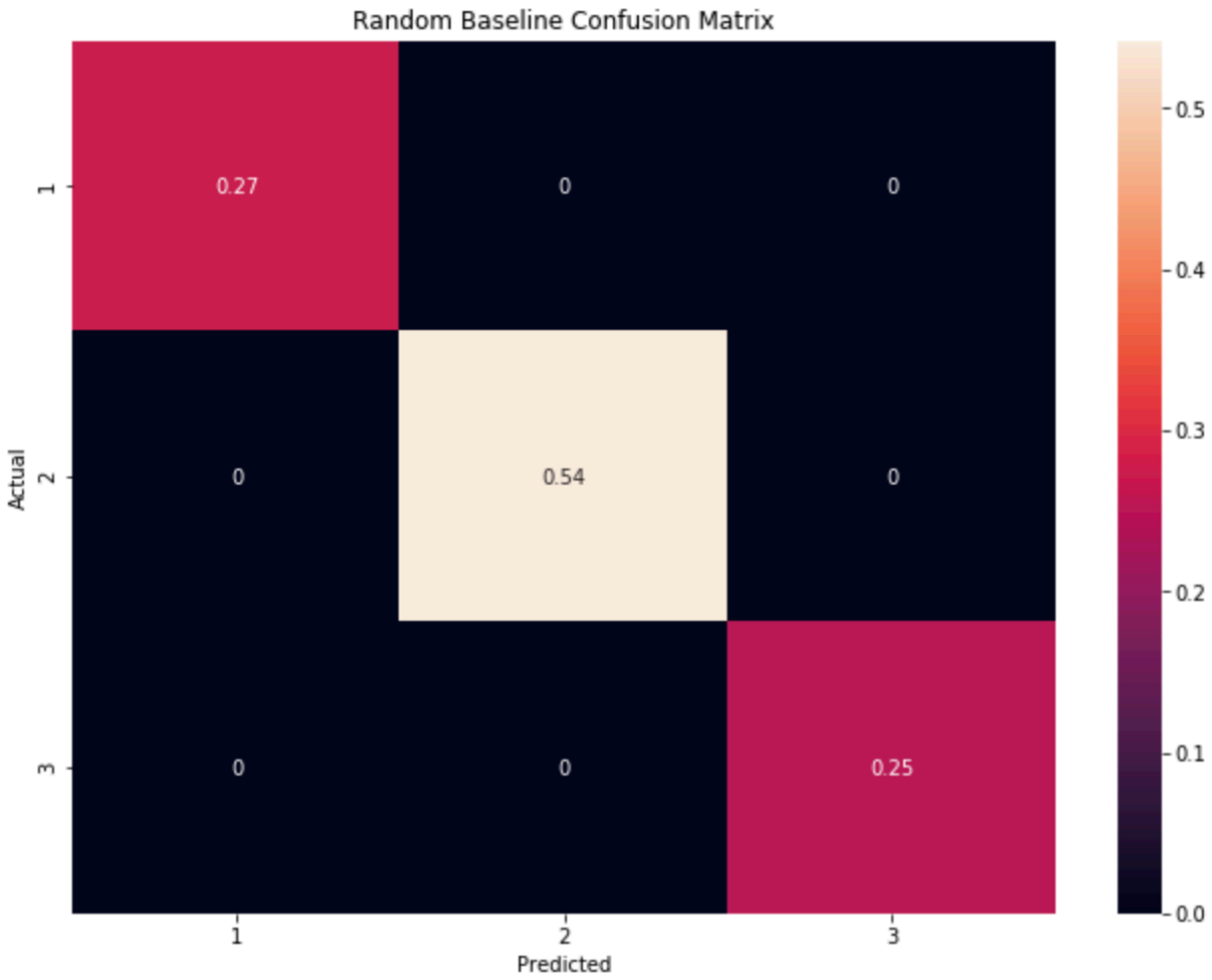
```
Gaussian Naive Bayes:
Accuracy: 84%
Error-rate: 16%
Precision: 93%
Recall: 85%
F1 Score: 0.89

Naive Bayes:
Accuracy: 78%
Error-rate: 22%
Precision: 93%
Recall: 76%
F1 Score: 0.84
```

Appendix B: Random Baseline

Wine dataset:

```
Naive Bayes:
Accuracy: 95%
Error-rate: 5%
Precision: 94%
Recall: 90%
F1 Score: 0.92

Random Baseline:
Accuracy: 29%
Error-rate: 71%
Precision: 30%
Recall: 17%
F1 Score: 0.22
```





Mushroom Dataset:

```
Naive Bayes:
Accuracy: 98%
Error-rate: 2%
Precision: 98%
Recall: 100%
F1 Score: 0.99

Random Baseline:
Accuracy: 49%
Error-rate: 51%
Precision: 61%
Recall: 47%
F1 Score: 0.53
```

Car Dataset:

```
Naive Bayes:
Accuracy: 87%
Error-rate: 13%
Precision: 80%
Recall: 46%
F1 Score: 0.58

Random Baseline:
Accuracy: 26%
Error-rate: 74%
Precision: 26%
Recall: 9%
F1 Score: 0.13
```

Adult Dataset:

```
Naive Bayes:
Accuracy: 78%
Error-rate: 22%
Precision: 93%
Recall: 76%
F1 Score: 0.84

Random Baseline:
Accuracy: 49%
Error-rate: 51%
Precision: 74%
Recall: 50%
F1 Score: 0.59
```