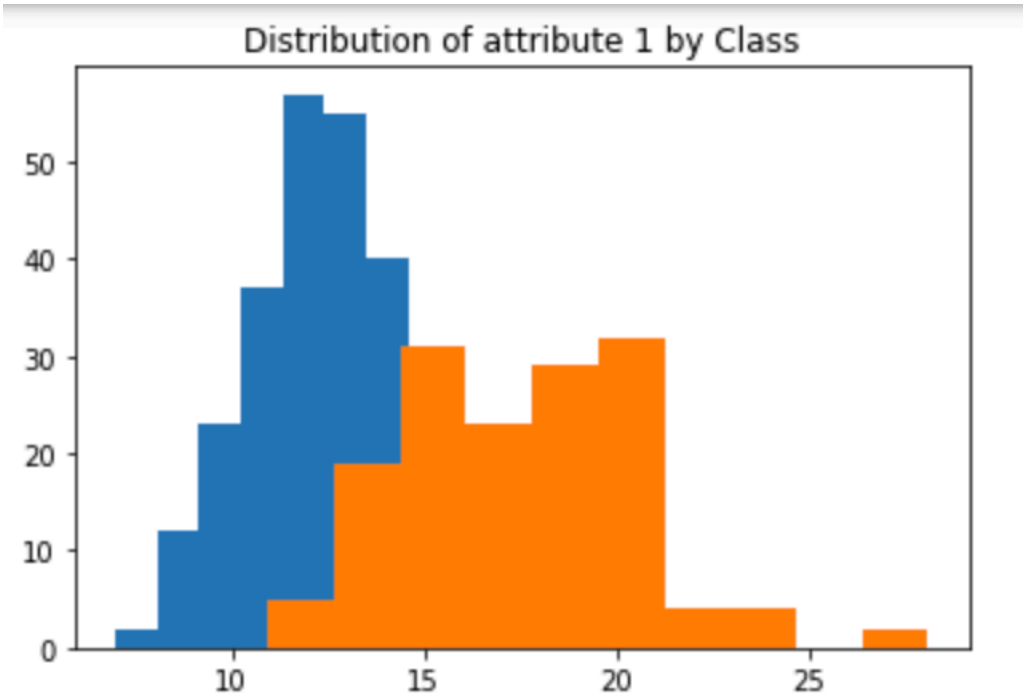
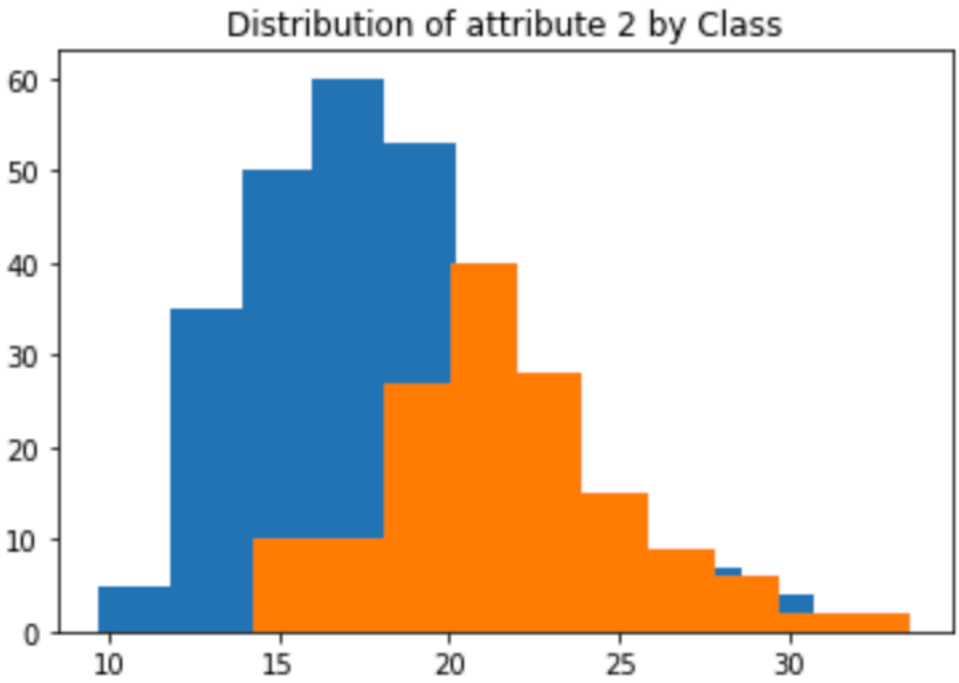
Student name: Chan Jie Ho
Student number: 961948

Q1

Try discretising the numeric attributes in these datasets and treating them as discrete variables in the na¨ıve Bayes classifier. You can use a discretisation method of your choice and group the numeric values into any number of levels (but around 3 to 5 levels would probably be a good starting point). Does discretizing the variables improve classification performance, compared to the Gaussian na¨ıve Bayes approach? Why or why not?

```
Average error rate reduction = -15.47863247863246%
```





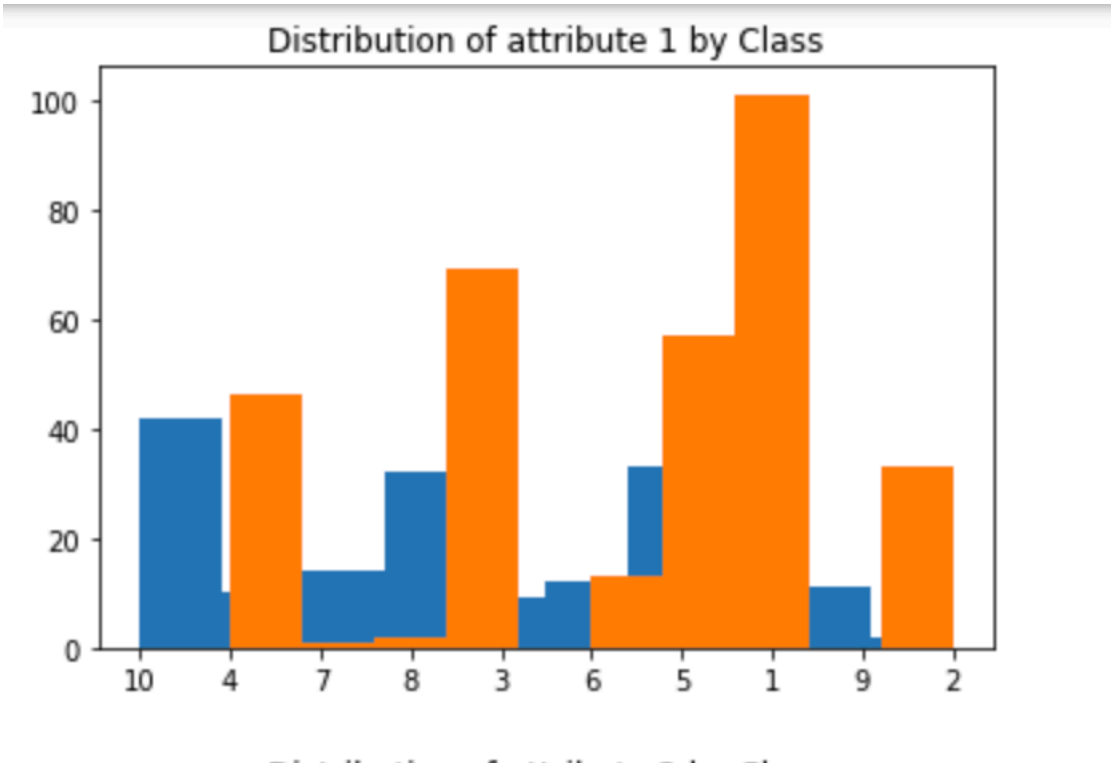
As shown in the negative average error reduction rate, the discretization method does slightly worse as compared to the Gaussian Naïve Bayes model, however this is a very small reduction. One reason that the Gaussian model performed better is likely because the underlying data is already normally distributed, or has a Gaussian distribution. The Gaussian model takes these distributions into consideration and is also less susceptible to outliers as compared to the discretization method. This is further emphasised due to the fact that I used the equal width binning method. Had I used a k-means clustering method to perform the binning, the discretization method may have very likely performed even better than the Gaussian method.

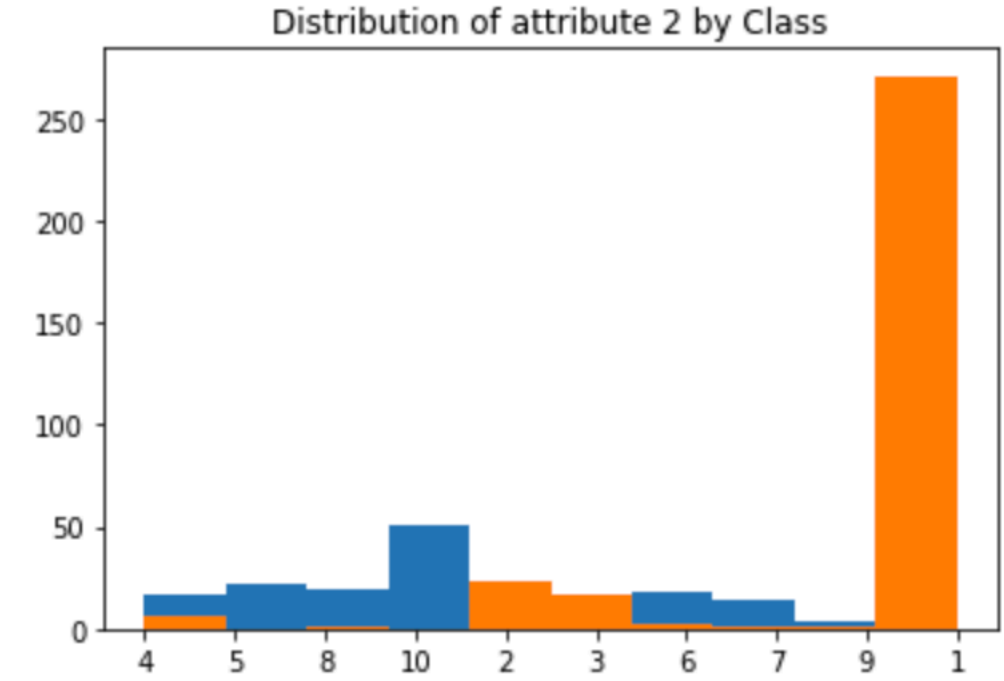
Q2

Implement a baseline model (e.g., random or 0R) and compare the performance of the na¨ıve Bayes classifier to this baseline on multiple datasets. Discuss why the baseline performance varies across datasets, and to what extent the na¨ıve Bayes classifier improves on the baseline performance.

```
Average error rate reduction = -13.647953216374281%

defaultdict(float, {'B': 249.0, 'M': 149.0})
```





The baseline model that I had chosen was the 0R model and this generally performed a lot worse on the datasets, however, some datasets did not see a significant reduction in the error rate and may even see a positive reduction in the error rate. One pattern I noticed was that it tended to perform worse on the nominal datasets. This is mainly due to the fact that the distribution of the classes is not equal. Going further by plotting out the dataset, we can see that many of the dataset, such as the above, are not actually normally distributed and one can see a much higher frequency in some attribute values (not sorted because they are

nominal and sorted would be inappropriate) and this would imply that the class and the attributes have a significant relationship that is being overlooked when we do not consider the attributes.