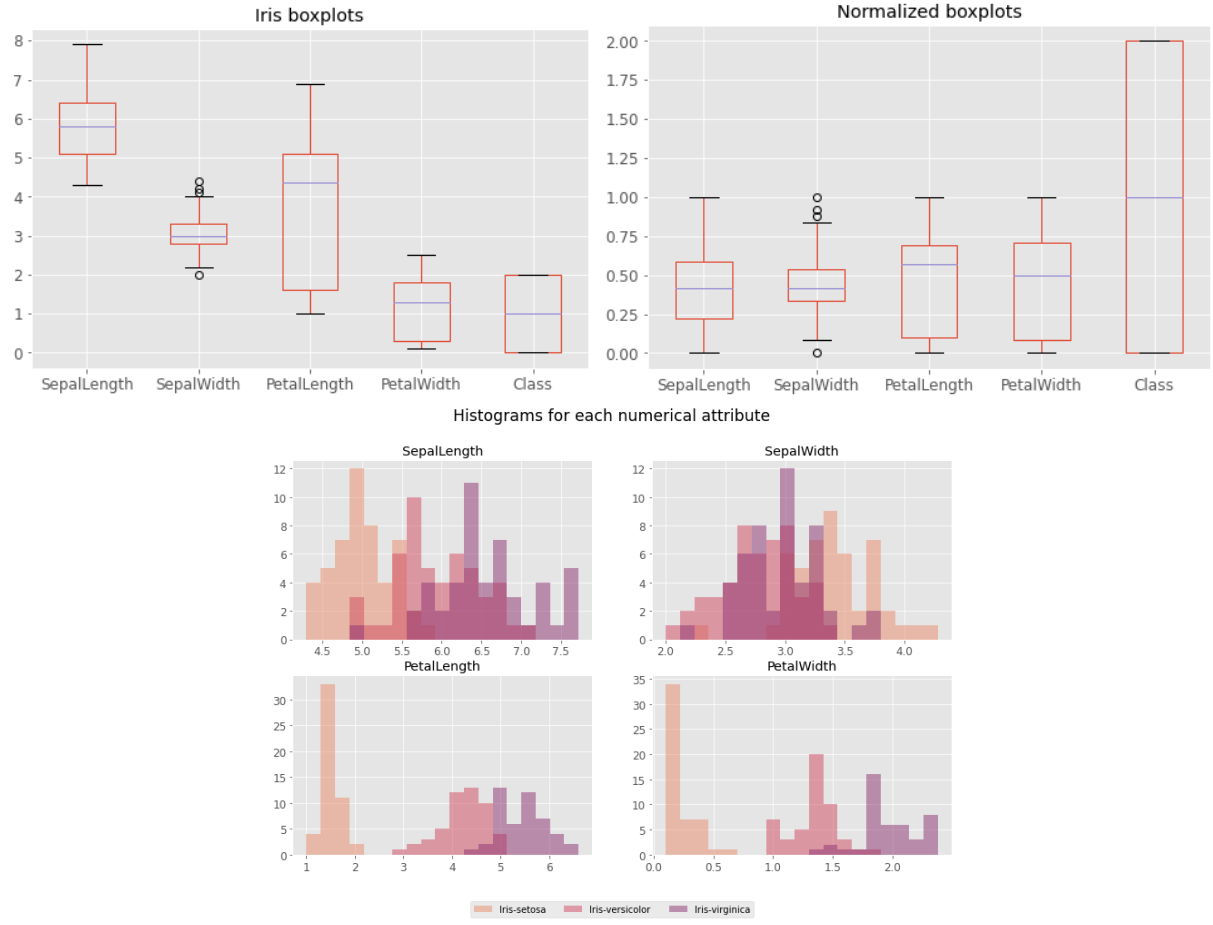
Mia Rodgers: https://github.com/miamrodgers/4310-ML/blob/main/ME3_NativeBayes/NaiveBayes.ipynb
Alex Larsen: https://github.com/alarsen123/ML-HW/blob/main/ME3_NativeBayes/ME3_NativeBayes/NaiveBayes.ipynb
ME3

For this assignment, we worked with the iris data set from the sklearn datasets. First, we normalized the data set using a min-max normalization. I dropped the 'Class' and 'Name' columns since we do not want to normalize categorical data.

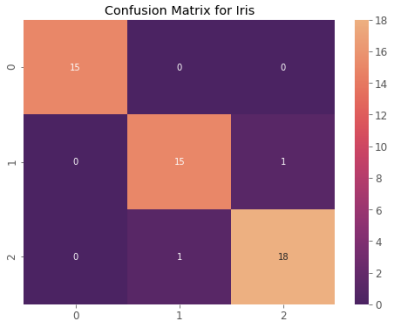
Next, we looked at the data. The original and normalized data sets have the same info() and count from describe, but are different for the other summary statistics. As expected the min and max changed to 0 and 1 respectively for all of the normalized columns and all other attributes also reflect a successful normalization. When we looked at the boxplots, we saw that they were different. But of course the histograms were the same for both.

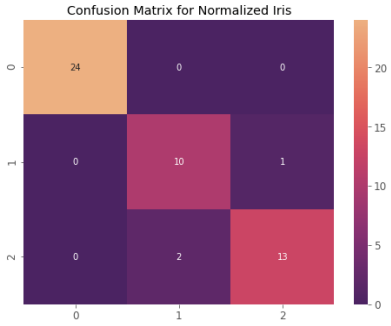


The boxplots reflect the normalization because the minimum and maximum values are all the same in the second plot, whereas they are all different in the first one.

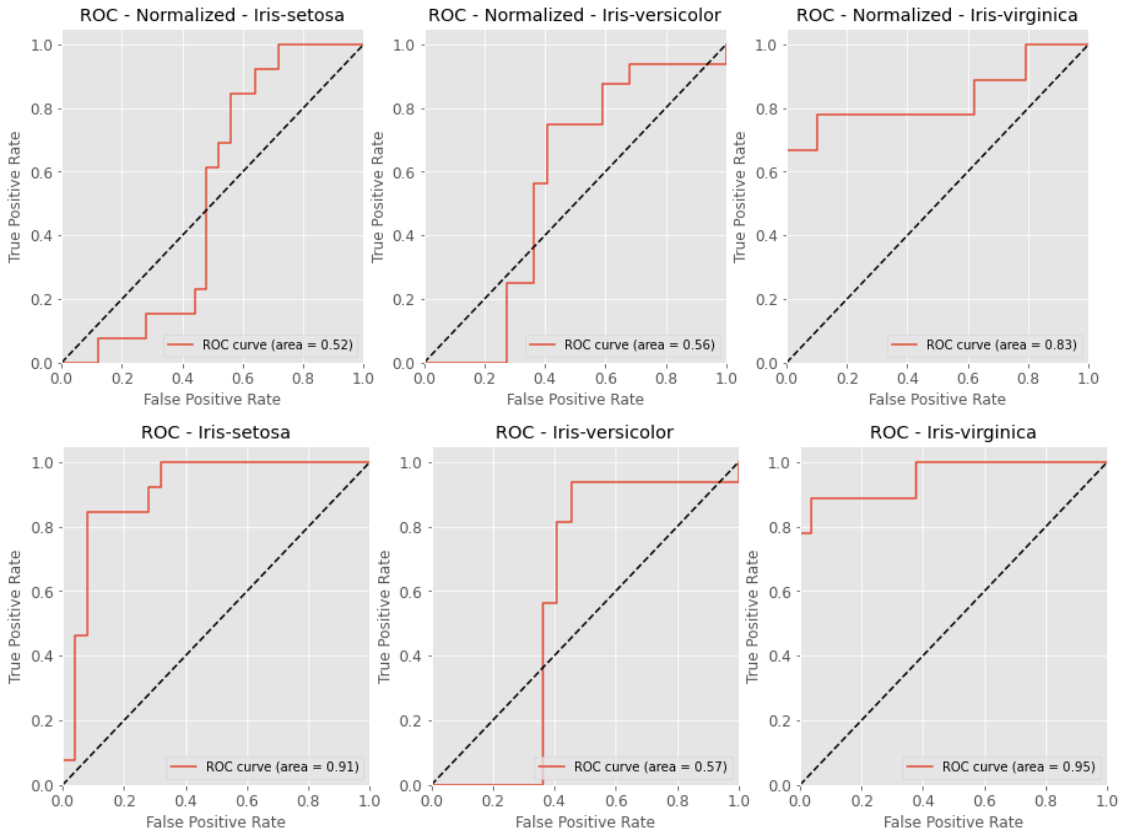
Then, we created a Naïve Bayes classifier for each of the data sets with k-fold cross-validation using three splits. We then looked at the performance metrics and ROC curves for each of

them. For the non-normalized data set, the accuracies ranged from 0.90 to 0.98 and for the normalized set, it ranged from 0.92 to 0.98. The weighted averages for the precision, recall, and f1-score ranged from 0.94 to 0.96 for both data sets.





The confusion matrices are pretty similar, but there are some noticeable differences. Overall, there is not much of a difference between the normalized and non-normalized data sets when it comes to modeling.



From the ROC curves, it seems that the non-normalized data is more accurate for the setosa iris, but about the same for the other two.