Lachlan Sinclair

2/2/2020

Week 4 Assignment 1

Reflections

a. Looking at the 8 histograms for Iris-virginica and Iris-versicolor there are some noticeable difference in the dispersion of the measurements. However, there are significant overlaps in the ranges of the data for some of these histograms. The sepal width of both species is a good example of this, they both have a large concentration of instance of lengths around 3. This overlapping of data occurs throughout the observable factors which would cause problems for a predictive model built exclusively around these factors. That is why I suspect a model built without feature engineering would cause a significant percentage of samples to be assigned to the wrong species.

b. Looking at the scatter plots produced, Iris-setosa is well separated from the other two species in all 6 graphs. Iris-versicolor and Iris-virginica on the other hand are extremely close to one another in the graphs. They are somewhat separate however if you were to look at the graph without color, they appear in one large section, causing there to be only two distinct groups in the scatterplots. This would cause issues for the simple unsupervised learning technique as there would be no way for it to accurately put the line in that should separate the species Iris-versicolor and Iris-virginica. Therefor it would be quite difficult to use a simple unsupervised learning technique to accurately predict the species.