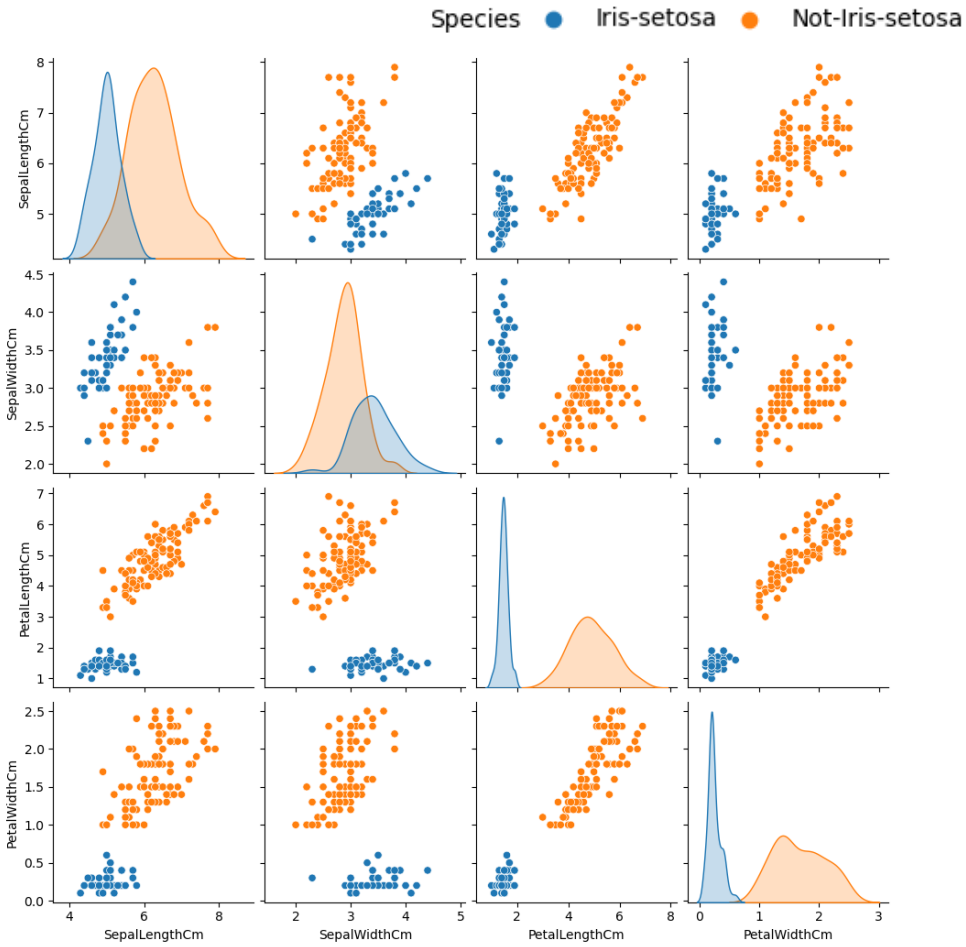
CPSC 393 Assignment #1 Report Nikki Poentis

Introduction

Support Vector Machines (SVMs) work to classify clusters by maximizing the margins that separate them. At the basic level, SVM models separate the groups of data points by building a linear hyperplane in the middle of the 2-dimensional support vectors–where is gets its name–based on the points that are the closest together. If the data points are able to be separated into groups easily with a single line, the SVM classifier should work well. If the data points are not easily separable, the points can be moved to a higher dimension so that it may become linearly separable.

Analysis

The dataset uses four quantitative variables–the width and length in centimeters of the flowers' petal and sepal–to classify each into the species categories of 'Iris-setosa' and 'Not-Iris-setosa'. The pairplot below shows the relationship between the variables and shows that there is a clear separation between the the Iris-setosa species and other flowers. Since it looks like the data points are separated nicely, the SVM model should have no issues classifying the data points into the species categories.

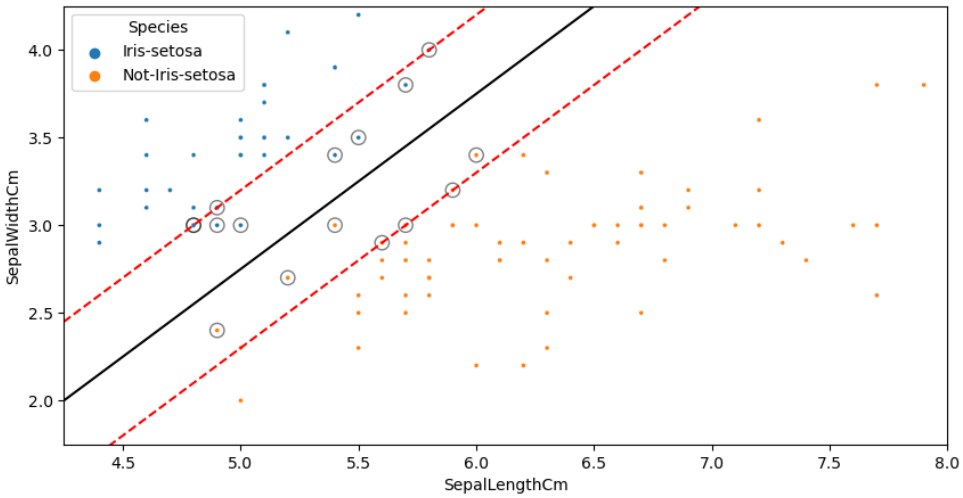


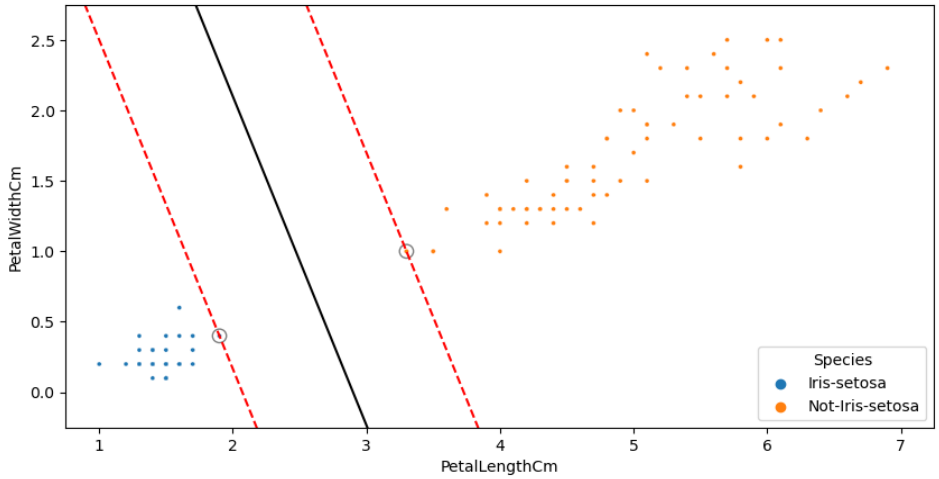
Results

After splitting the data into the testing and training sets, I then trained the model and used the testing data for all 4 flower measurements to produce predictions. Based on these predictions, the model resulted in an accuracy, precision, and recall score of 1. Since a 1 is the highest possible score (1.0 = 100%), this means that based on the actual classifications in the testing set, the model correctly classified the species for all of the flowers.

Discussion/Conclusion

In most cases, having 100% accuracy is not seen as a great thing but instead something to raise flags. While it could mean that the model is overfit, it is also possible that the variables are all perfectly linearly separated and the model actually works that well. I wanted to make sure that this was the case by seeing the support vectors and margins generated by the SVM classifier myself; however, it's highly impractical to visualize a 4-dimensions. Instead I broke up the data into two sets and built new SVM classifier for each. Since it was previously established that the difference between the two species is visually obvious, no matter which variables are plotted against each other, it technically doesn't matter how to split up the variables. I ended up just splitting it into the measures for sepals and petals.





Similar to the main SVM model generated for this assignment, the two classifiers give a 100% accuracy score. Both the plots, especially the one for the petal measures, clearly show the linear separation between ‘Iris-setosa’ and ‘Not-Iris-setosa’. Based on this observation, the accuracy level of the SVM model with all 4 variables make sense.