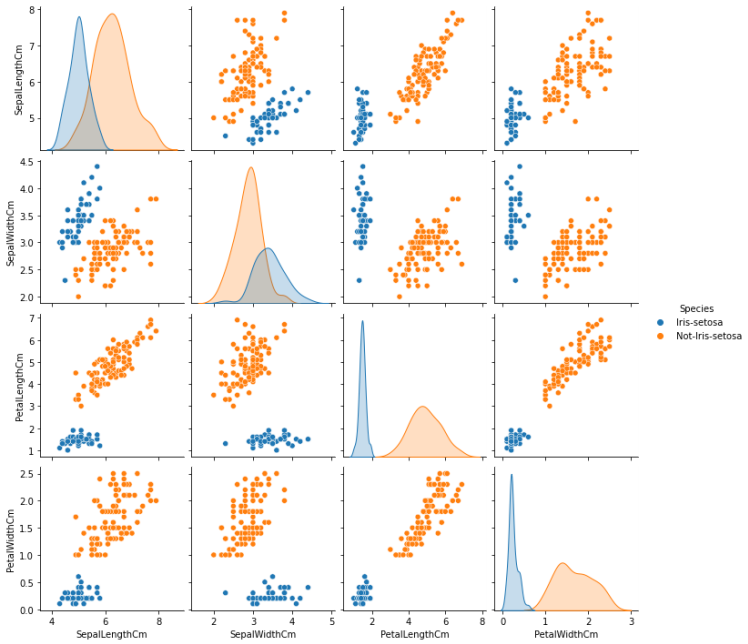
Assignment 1
Nora Mirabal
09/08/2022
CPSC 393-01

Introduction

The dataset provided for assignment 1 provides data on the classification of species as Iris-Setosa or not Iris-Setosa. The dataset has 6 variables, Id, SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm, and Species. The purpose of this assignment is to classify the species using the variables.

Analysis

The first step I took in analyzing the data set was opening it in excel to get an overhead view of the information the dataset provided. Once I was familiar with the variable names and had a better idea of the data I was going to work with I opened the file as a csv in python as a dataframe. I uploaded the csv using file.upload() and io, which allows me to choose the file from my computer's file explorer. Using the .shape method I found that the original dataset has 6 columns and 150 rows. Using the .head() method I was able to view the first 5 rows to assess that the csv file was properly formatted like the excel file. I also checked the data frame for null values using the .isnull().sum() functions together and found that there were no null values for the variables. After analyzing the shape of the data frame I used seaborn and matplotlib to view the relationship between all of the variables. In this step I dropped the ID variable as it is only used to catalog the number of plants collected and doesn't give us a deeper understanding of the relationship between the parts of the plant and its classification.



The graph separates the Iris-Setosa and Not-Iris-setosa species by color to help visualize the difference between the relationship of the variables between species. We can see from these graphs that the species have almost 0 overlap between the relationships of variables together. For example Iris-Setosa have a

petal width of less than 1 cm and a petal length of around less than 2 cm while not Iris-Setosa have much larger petal lengths and widths. From the graph we can also see that there is some imbalance between the two species as the not-Iris-Setosa group is visibly larger.

Methods

The purpose of this assignment was to use Support Vector Machines. SVMs are supervised learning methods which in this case are being used for classification. I set the X variables as SepalLengthCm, SepalWidthCm, PetalLengthCm, and PetalWidthCm and I set the y variable as Species. I disregarded the ID variable once again because it doesn't add any value to our classification. After setting my variables I split the data set into an 80% training set and 20% test set. I created a SVM classifier with a linear kernel and used this model to fit the X train and the y train. I chose a linear kernel after reviewing the relationship between the variables because it seemed that a single line could separate the two species groups accurately. Once I trained the model I predicted the responses for the test dataset and stored that in a y pred variable.

Results

Using the y test set and the y pred I was able to analyze the results of the model. The accuracy, precision, and recall all came out to be 1.0. This means that the model was 100% accurate when classifying the species of the plant and 100% accurate in labeling positive tuples. It's always important to take into account any possibility of data leakage or overfitting of the model. However in this case we don't suspect either of those things because the size of the data set is quite small and the difference between the data collected for Not-Iris-Setosa and Iris-Setosa, it makes sense that the model would have an accuracy score of 1.

Remarks

The data set was much smaller than many I've used for classification before so it's interesting how SVM differs in its classification methods from other supervised learning methods. It was interesting as well to see an accuracy score of 1 but after a deeper look at all of the variables it was clear the model was able to distinguish the two species easily because their variable values differed so greatly. Overall the dataset was easy to work with because it didn't need any cleaning.