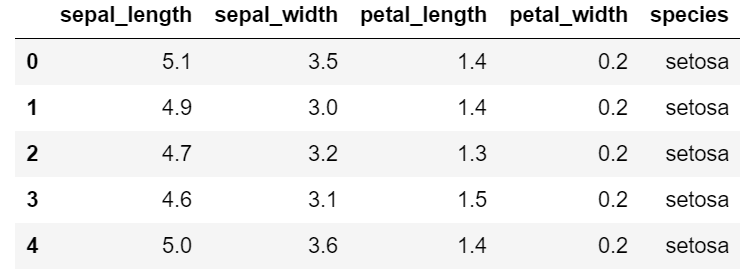
Cristhian Salvania Tolentino

Professor Celepcikay

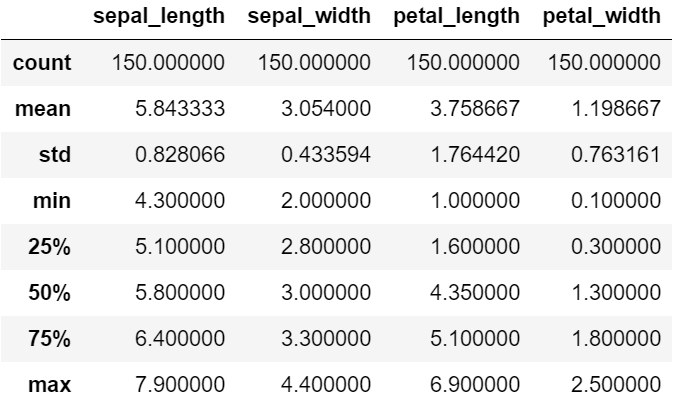
CS4319-21823

May 5, 2019

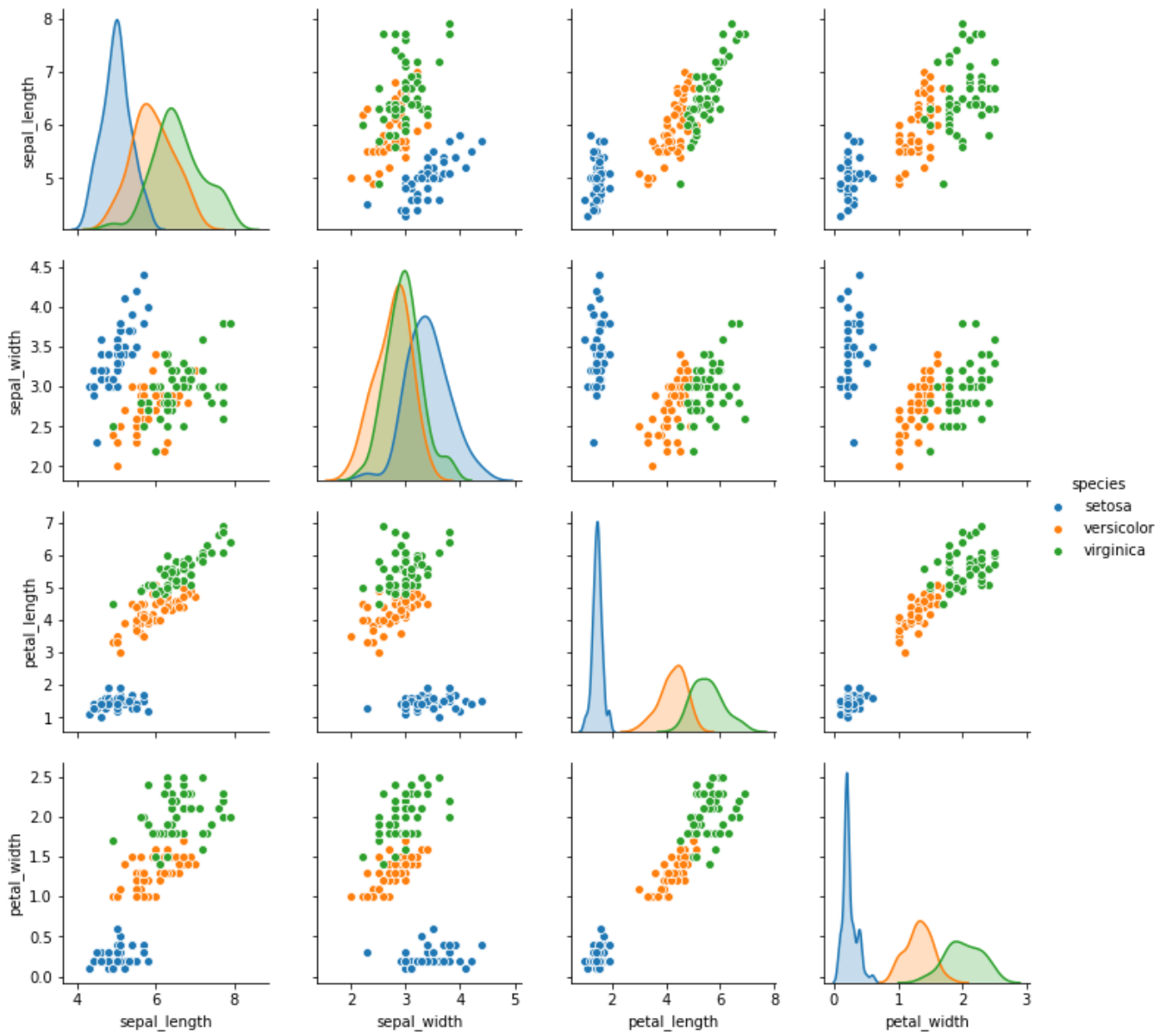
Final Project Part 2: Iris Dataset Decision Tree



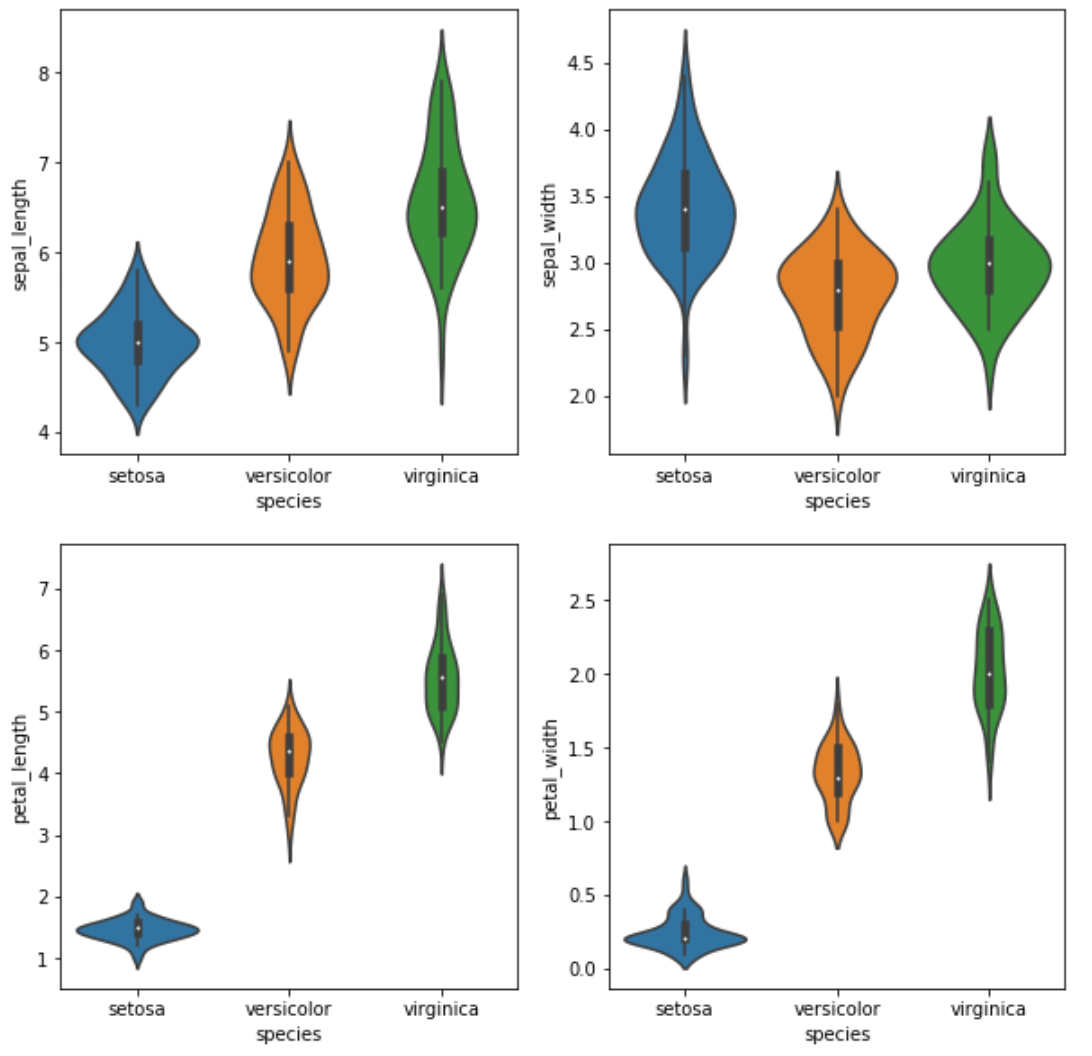
Each row represents an entry for a flower: four measurements and one class, which tells us the species of the flower.



From the summary statistics of the dataset, sepal lengths range from 4.3 cm to 7.9 cm. Sepal widths range from 2.0 cm to 4.4 cm. Petal lengths range from 1.0 cm to 6.9 cm. Petal widths range from 0.1 cm to 2.5 cm. 50% of flowers have sepal lengths between 5.1 and 6.4, sepal widths between 2.8 and 3.3, petal lengths between 1.6 and 5.1, and petal widths between 0.3 and 1.8.



From the scatterplot, it seems that the Setosa flowers are the smallest flowers and Virginica flowers are the largest flowers.

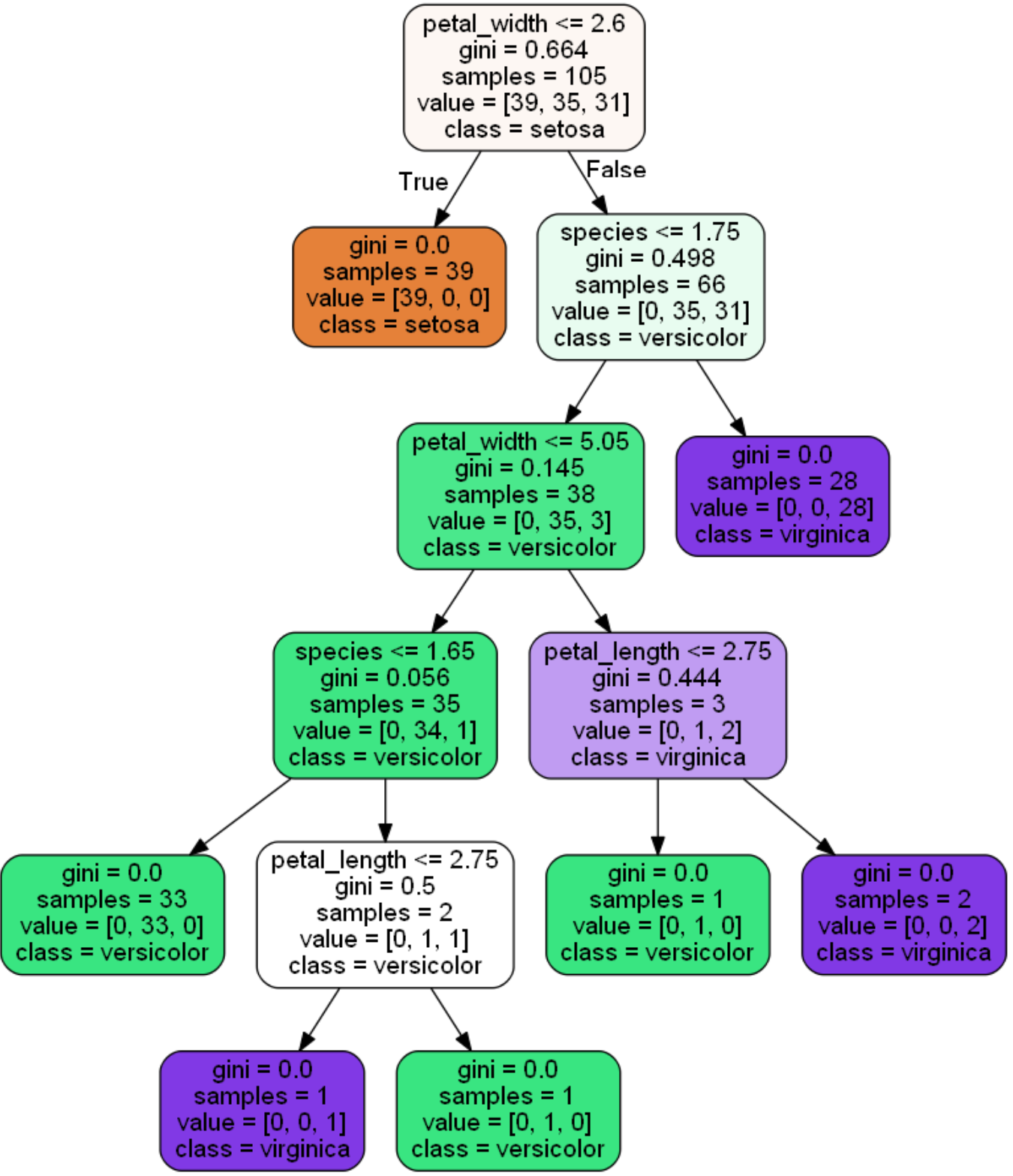


These violin plots show the relationships between species to measurements. The boxplot elements show the median weight for setosa flowers is lower than for other species of flowers. The shape of the distribution (skinny on each end and wide in the middle) show that the measurements of virginica are highly concentrated around the median.

**Decision Tree Model**

After data analysis, the model will have 4 independent variables: sepal length, sepal width, petal length, and petal width; while the dependent variable is species.

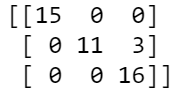
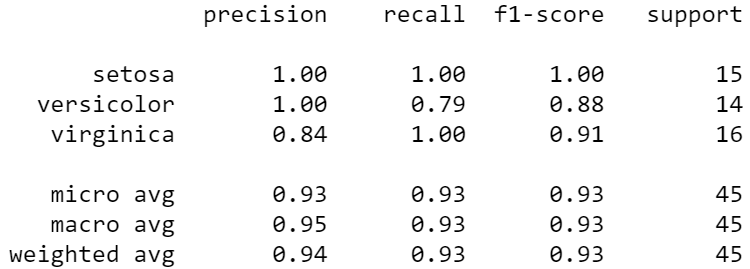
I have generated one decision tree using species as my training and test sets.



The Gini index is the measure of the impurity, the samples are the number of flowers left to classify, and the value is how many flowers are in class Setosa, how many flowers in class Versicolor, and how many flowers in class Virginica.

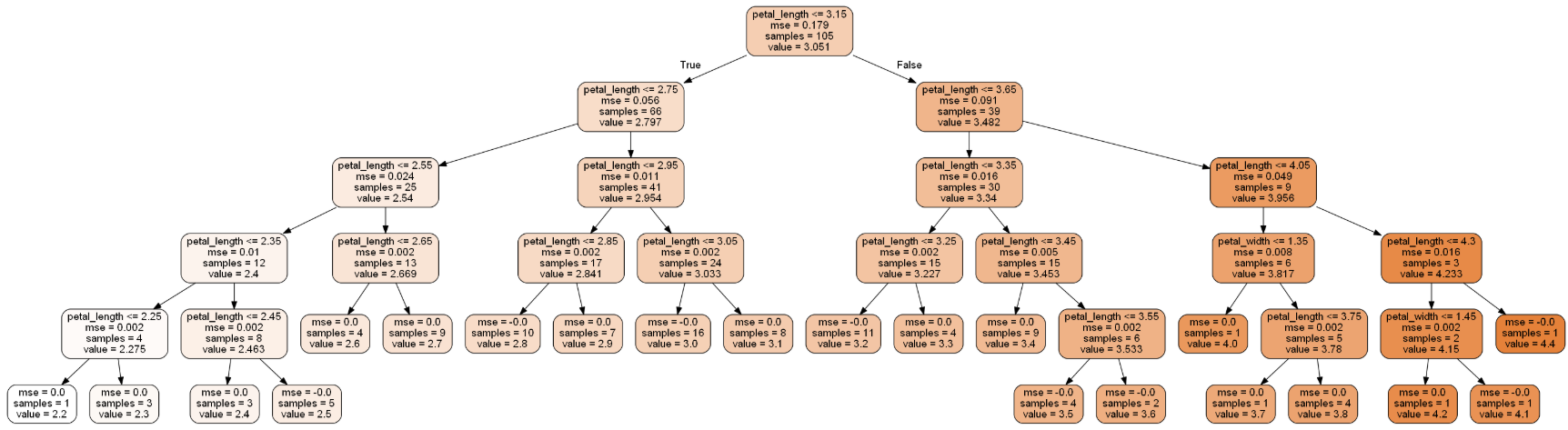
I see that I have 105 flowers left to classify: 39 are setosa, 35 are versicolor, and 31 are virginica. The root starts at how wide the petals are. 39 out of 105 flowers with petal widths less than or equal to 2.6 are classified as setosa. This leaves 66 flowers left: 35 versicolor and 31 virginica.

28 of 66 flowers are classified as virginica because they petal widths greater than 5.05. 1 flower with petal length of less than 2.75 is classified as versicolor and 2 flowers with petal length greater than 2.75 are classified as virginica. 1 flower with petal length less than 2.75 was classified as virginica and 1 flower with petal length greater than 2.75 is classified as versicolor.

Each row corresponds to a correct species, while columsn represent the predicted species. There are 15 flowers predicted as iris setosa flowers, 11 predicted as iris versicolor flowers, and 19 predicted as iris virginica flowers. In reality, there are 15 iris setosa flowers, 14 iris versicolor flowers, and 16 iris virginica flowers.

I tried to refine my results using a continuous attribute such as sepal width. Here is the decision tree with sepal width:



Sepal width is a continuous attribute with floating point values, so I had to use the DecisionTreeRegressor() method. Even so, just by looking at this decision tree, it was hard to determine which node belonged to which class. I do not think a decision tree is fit for continuous attributes. Therefore, I did not continue to make decision trees for the rest of the measurements.

**Conclusion**

In this second part of this project, this dataset is a lot smaller than the Lending Club so I was able to quickly observe on my own the Iris Virginica flowers are generally larger than Setosa and Versicolor. After training, testing, and building a decision tree, I found that my observations are correct.