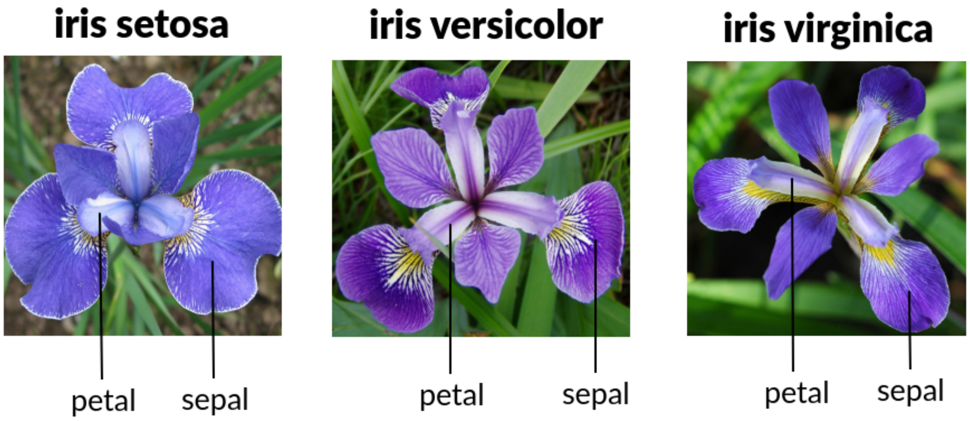
# CPSC 393 Assignment 1: Report
# Using Support Vector Machines (SVMs) on the Iris Dataset

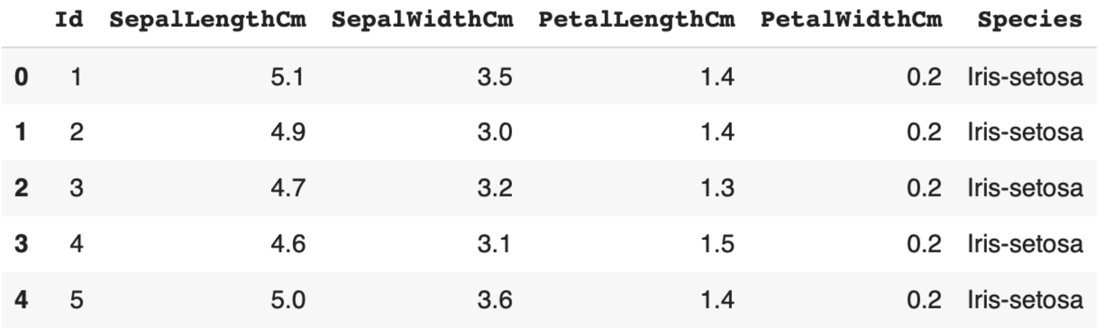
## INTRODUCTION

The dataset used for this Assignment relates to Iris-Setosa classification, a species of flowering plants. The features in this dataset include flower ID, Sepal Length (Cm), Sepal Width (Cm), Petal Length (Cm), Petal Width (Cm), and Species (target). Using this information, we will classify whether the species is Iris-Setosa or Not-Iris-Setosa using a Support Vector Machine (SVM) hyperplane to separate the data points into two different classes based on their features. Below is a visual representation of the sepal and the petal of this species of flower.



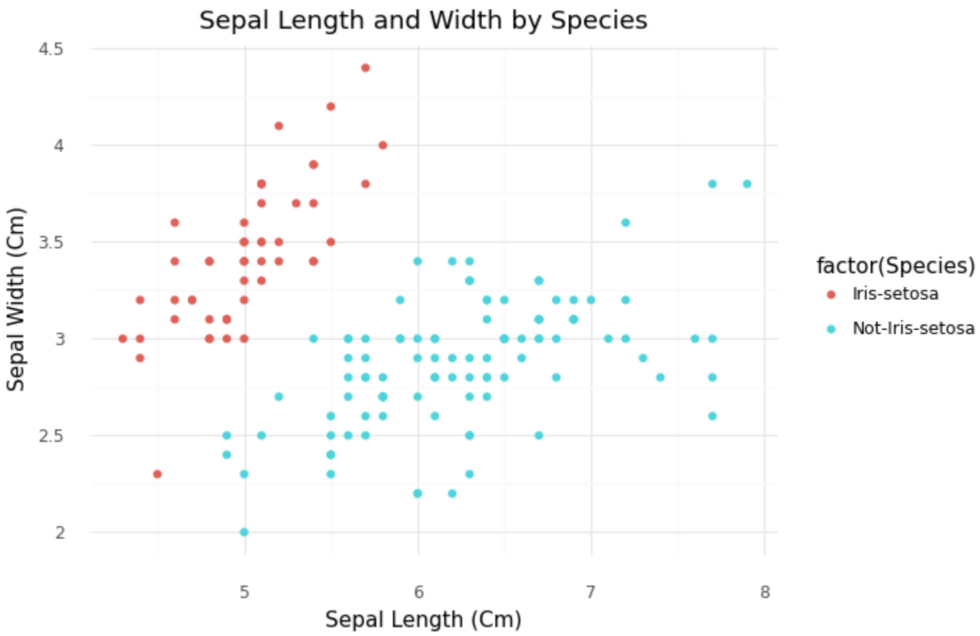
## ANALYSIS

After loading the Iris dataset from a CSV file, exploring the dataset was helpful. Using the shape function, the size of the dataset was 150 rows and 6 columns, and the head function showed the first 5 lines of the CSV file. This dataset has the features: ID, Sepal Length (Cm), Sepal Width (Cm), Petal Length (Cm), Petal Width (Cm), and Species.

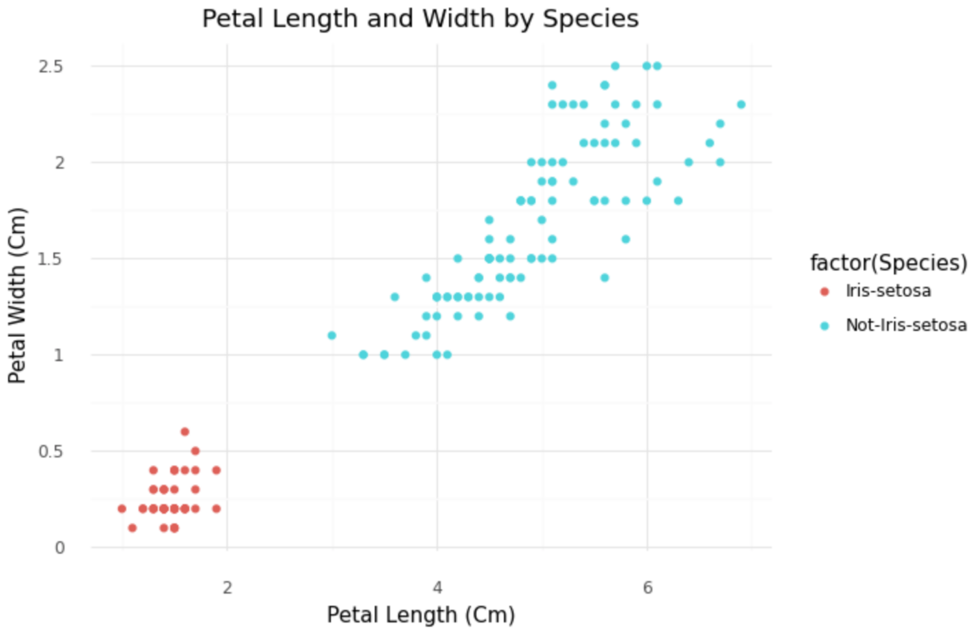


| | Id | SepalLengthCm | SepalWidthCm | PetalLengthCm | PetalWidthCm | Species |
|---|---|---|---|---|---|---|
| **0** | 1 | 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa |
| **1** | 2 | 4.9 | 3.0 | 1.4 | 0.2 | Iris-setosa |
| **2** | 3 | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa |
| **3** | 4 | 4.6 | 3.1 | 1.5 | 0.2 | Iris-setosa |
| **4** | 5 | 5.0 | 3.6 | 1.4 | 0.2 | Iris-setosa |

Then, using a scatterplot to display the Sepal Length and the Sepal Width by species showed that the Iris-Setosa group had a larger Sepal width and a shorter Sepal length, while the Not-Iris-Setosa group had a smaller Sepal width and longer Sepal length. This graph shows two groups by species, with a little overlap but otherwise pretty defined.



Using another scatterplot to display the Petal Length and the Petal Width by species showed that the Iris-Setosa group had a smaller Petal width and a shorter Petal length, while the Not-Iris-Setosa group had a larger Petal width and longer Petal length. Again, this graph shows two groups even more polarized than the Sepal dimensions.



The next step was to transfer the data into data frame format and to change the Species into a dummy variable (0 = Not-Iris-Setosa, 1 = Iris-Setosa) for simplicity in further analysis.

## METHODS

The Iris dataframe was then split into a training set and a test set to understand model performance using train_test_split( ), which split the data into 70% training and 30% test. To classify this data based on the Sepal length and width and the Petal length and width, Support Vector Machines (SVMs) will be used - one of the most popular and widely used supervised machine learning algorithms. The SVM classifier separates data points using a hyperplane with the largest amount of margin, then finds an optimal hyperplane (essentially a divide) to help classify new data points into different classes. With the Iris dataset, SVMs will be used to classify species of flowers using the features listed prior. The SVM algorithm is implemented using a kernel, which transforms an input data space into the required form. The kernel trick used takes a low-dimensional input space and transforms it into a higher dimension, and for this assignment a linear kernel was used. By importing the SVM module and creating a support vector classifier (SVC) object with a linear kernel, the model was then fit on the training set, and the prediction was performed only on the test set.

## RESULTS

To evaluate how accurately the classifier and model can predict and group the correct species of flower, the accuracy, precision, and recall values were computed. Accuracy is calculated by comparing actual test set values and predicted values and resulted in 1.0, or 100% accuracy. Precision is calculated by dividing the true positives by anything that was predicted as a positive and resulted in 1.0, or 100% precision. Lastly, Recall is calculated by dividing the true positives by anything that should have been predicted as a positive and resulted in 1.0, or 100% recall. All of these evaluation values for the model are very high and the model performed extremely well. We can conclude that SVM classifiers offer good accuracy and fast prediction for classifying flower species. This dataset shows a clear margin of separation and clearly defined groups.

## REMARKS

Due to the simplicity and quick implementation of Support Vector Machines, I think this is an excellent choice for analysis in determining whether a flower is of the species Iris-Setosa or Not-Iris-Setosa. It ran very quickly and was effectively able to classify this dataset with 100% accuracy, precision, and recall. It is helpful to implement an SVMs classifier and create a hyperplane to accurately classify different groups of flower based on the Sepal/Petal length and width.

## REFERENCES

https://www.datacamp.com/tutorial/svm-classification-scikit-learn-python
https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
https://medium.com/geekculture/svm-classification-with-sklearn-svm-svc-how-to-plot-a-decision-boundary-with-margins-in-2d-space-7232cb3962c0
https://canvas.chapman.edu/courses/44489/files/3815553?module_item_id=1426616