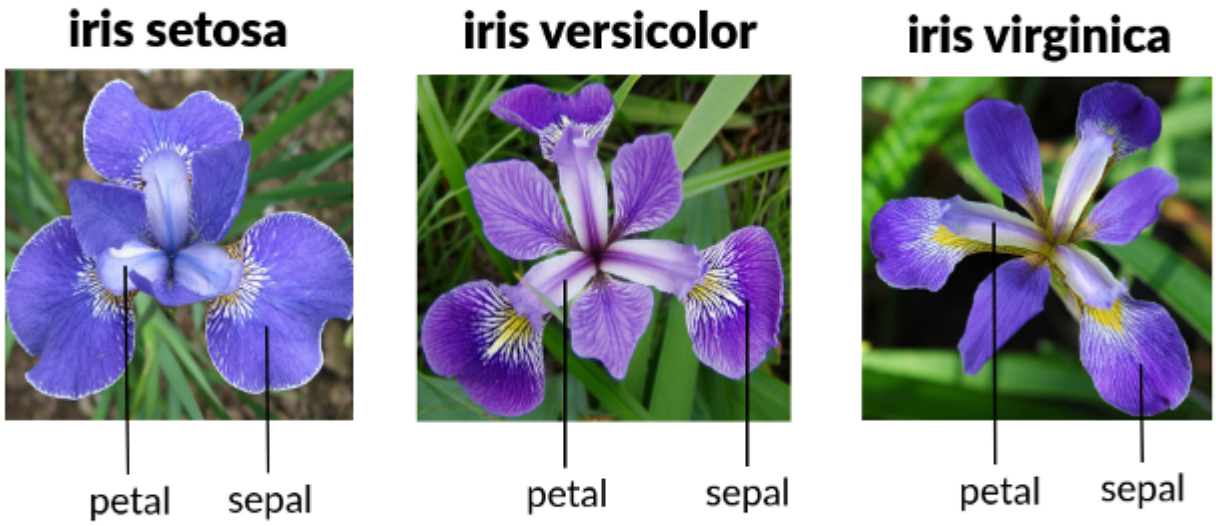
Iris classification with SVM on python

Studying and implementing a Support Vector Machine for classify the type of iris



We are going to create a model for classifying the type of iris based on the variables of the dataset.

In first place, we're going to identifying the variables

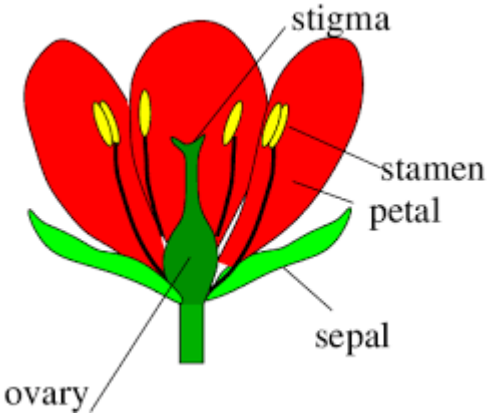
Sepal

The sepal is the part that forms the calyx of a flower; typically function as protection for the flower in bud, and often as support for the petals when in bloom.

We have two variables

- The Sepal length on centimetres

- The Sepal Width on centimetres



Petal

Petals are modified leaves that surround the reproductive parts of flowers.

We have two variables

- The Petal length on centimetres
- The Petal Width on centimetres

The Types of Flowers

Iris is a genus of 260–300 species of flowering plants with showy flowers. It takes its name from the Greek word for a rainbow, Iris.

In the dataset we have three types of iris:

- the Iris Setosa
- Iris Versicolour
- Iris Virginica

Let's code

Import packages

For this template we're use the next packages

Read the data

We are going the read the dataset directly from the UCI MACHINE LEARNING REPOSITORY, (https://archive.ics.uci.edu/ml/datasets/iris) but this dataset doesn't have any name, then at first place we will define our column names and then read the dataset

And then we have

	sepal_length_in_cm	sepal_width_in_cm	petal_length_in_cm	petal_width_in_cm	class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

For this example we just have a categorical columns then by pandas we encode the column

And then we have

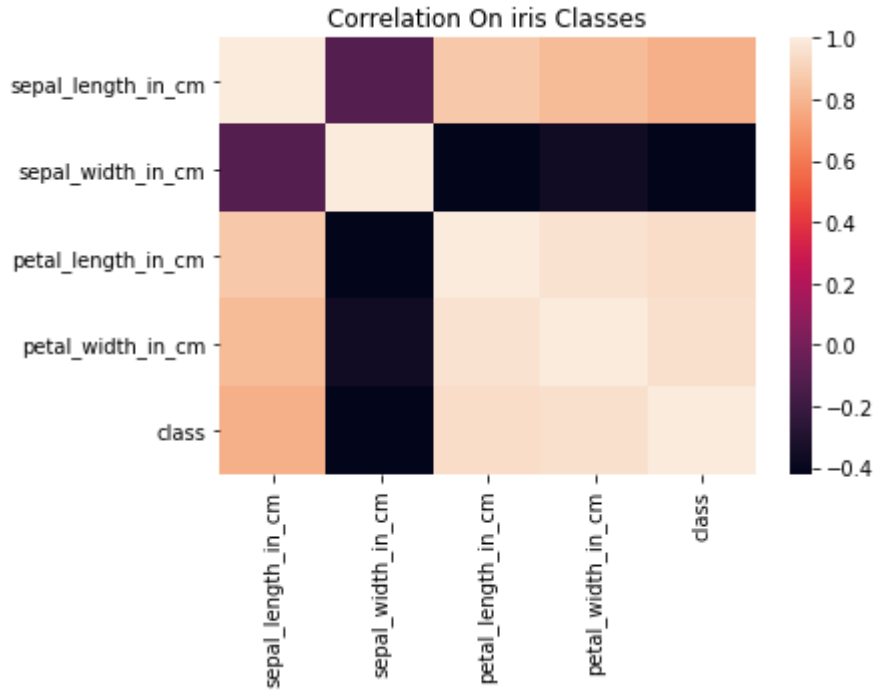
	sepal_length_in_cm	sepal_width_in_cm	petal_length_in_cm	petal_width_in_cm	class
0	5.1	3.5	1.4	0.2	1
1	4.9	3.0	1.4	0.2	1
2	4.7	3.2	1.3	0.2	1
3	4.6	3.1	1.5	0.2	1
4	5.0	3.6	1.4	0.2	1

Now we're going to analyze our data

In first place, all the dataset is organized equally, there is not any type of flower with more data, there are 50 rows for each flower so trying to count any quantity will be un–useful.

So let's look the correlation between the columns, in this way will see how important is a column for choosing which type of flower.

And the result is



For our project we must see the last column of the heat map, as we can see, the shape of the petals are the most correlation columns with the type of flowers, with lower correlation there is the sepal length which also haves a directly correlation and in last place there we have the negative correlation of the sepal width column, but this doesn't mean that is less important, is important but is inverse relation with the type of flower.

So all the columns are important for the model, in the case we want to quit some columns, the candidates will be the sepal columns with the sepal width in first place.

Splitting dataset

Creating the SVM model

For this classification problem, we'll use the SVM classifier, this by a personal choice, with a small dataset and the good parameters we will have an accurate model.

And finally for check the accuracy of the model, we'll use the confusion matrix and the cross validation

The results are:

```
[[13  0  0]
 [ 0 15  1]
 [ 0  0  9]]
Accuracy: 98.18 %
Standard Deviation: 3.64 %
```

We have a 98% of accuracy which is a very good model, and with the confusion matrix we can see that we have just only one misclassified data.

Conclusion

The iris classification problem is a good project for predicting the class and evaluating the columns to check the importance on the predictions