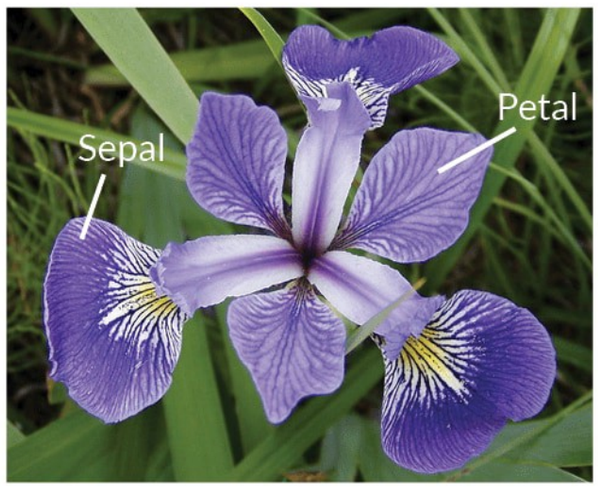
Classification of Iris flowers



Iris Versicolor

Iris Setosa

Iris Virginica

Introduction

. In this study we will do some exploratory data analysis on the famous Iris dataset.

. The Iris Dataset contains four features (length and width of sepals and petals) of 50 samples of three species of Iris (Iris setosa, Iris virginica and Iris versicolor).

. These measures were used to create a linear discriminant model to classify the species. The dataset is often used in data mining, classification and clustering examples and to test algorithms.

. Here, I applied two Machine Learning classification algorithms K-nearest neighbor (KNN) and support vector machine (SVM)

Dataset Structure

. The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.

The Iris dataset can be downloaded in the link below:

https://archive.ics.uci.edu/ml/machine-learning-databases/iris/

Attribute Information:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm

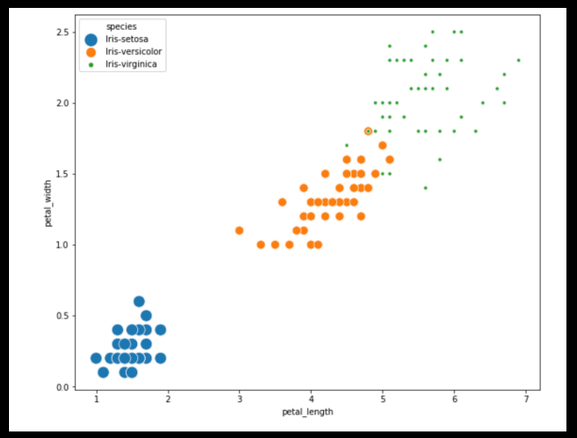
The features are **sepal_length, sepal_width, petal length** and **petal** width. The target is the **species** which its attributes are Iris Setosa, Iris Versicolour and Iris Virginica.

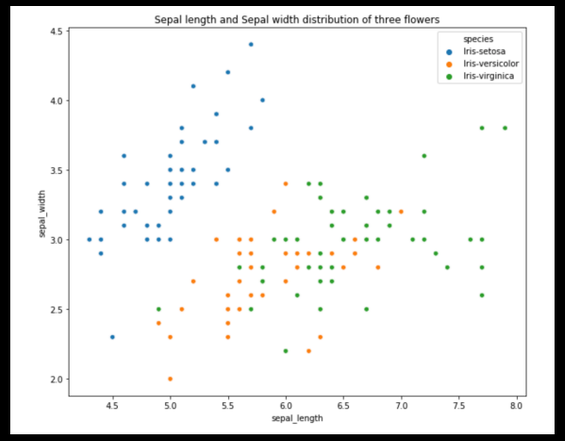
```
1 data.head()
```

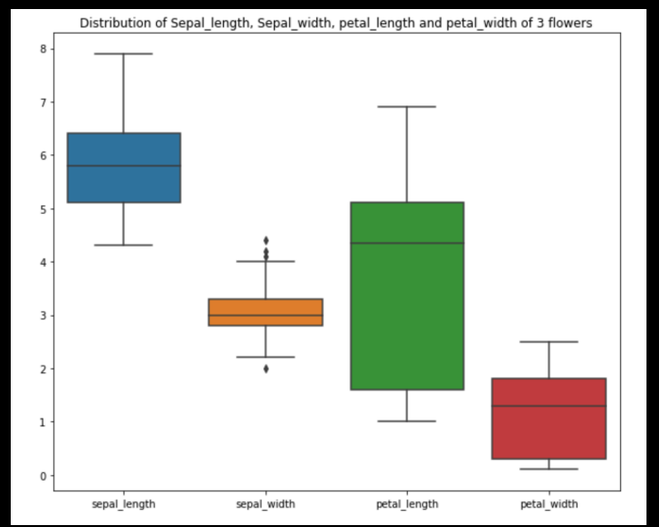
	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

Exploratory Data Analysis

Distribution of Sepal Length and width, Petal length and width of 3 kinds of the flowers as well as their scattered plots are shown at below

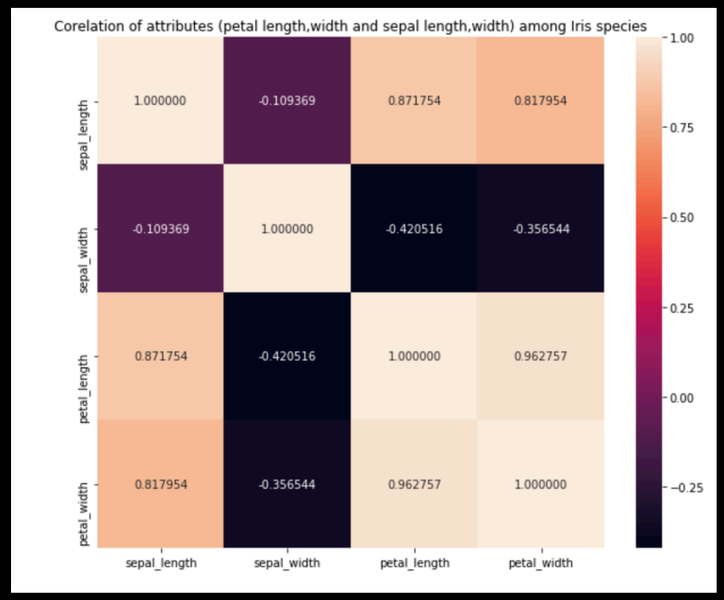






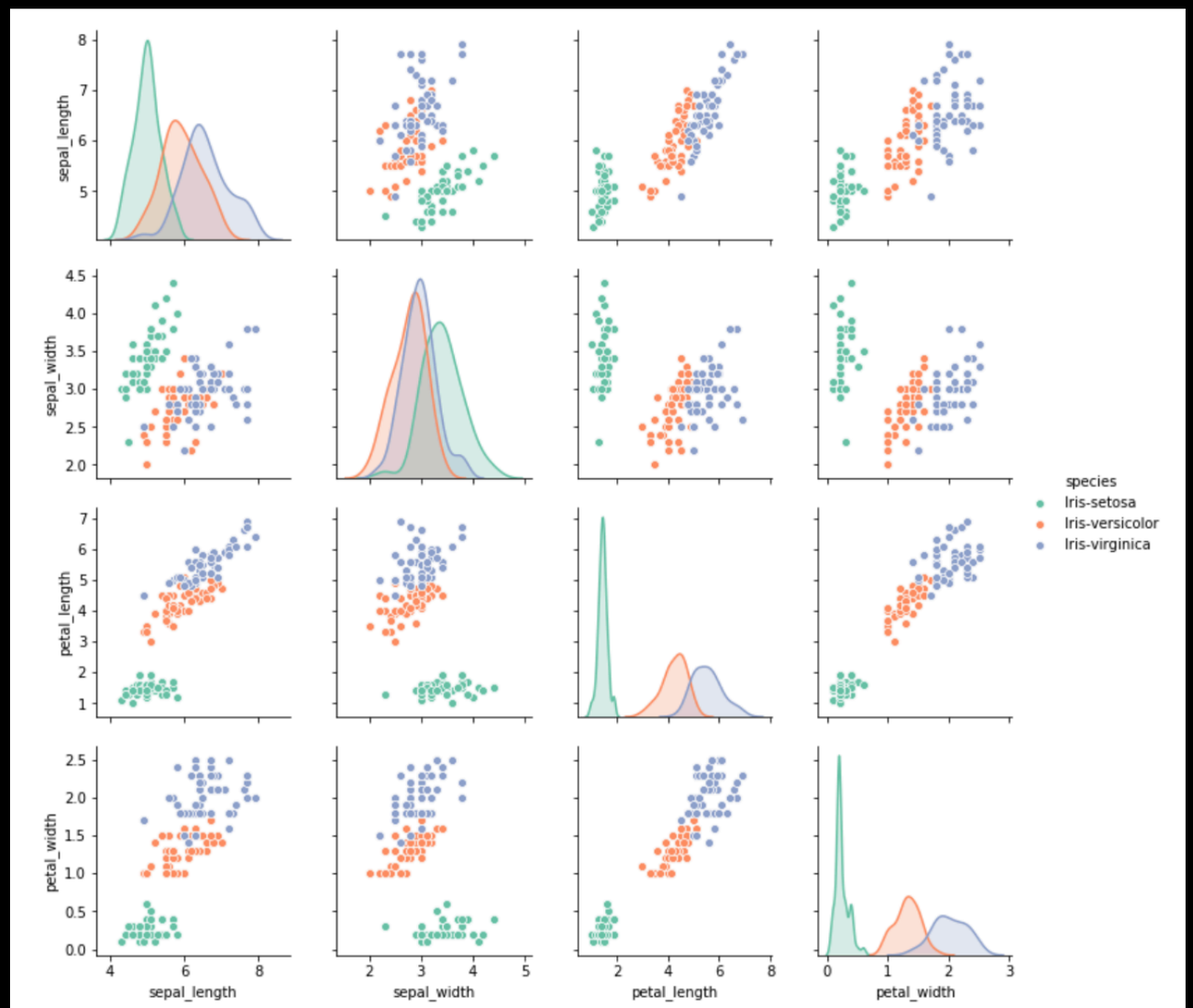
How does one variable compares to others?

Are these correlated? The Correlation of all the attributes are plotted as:



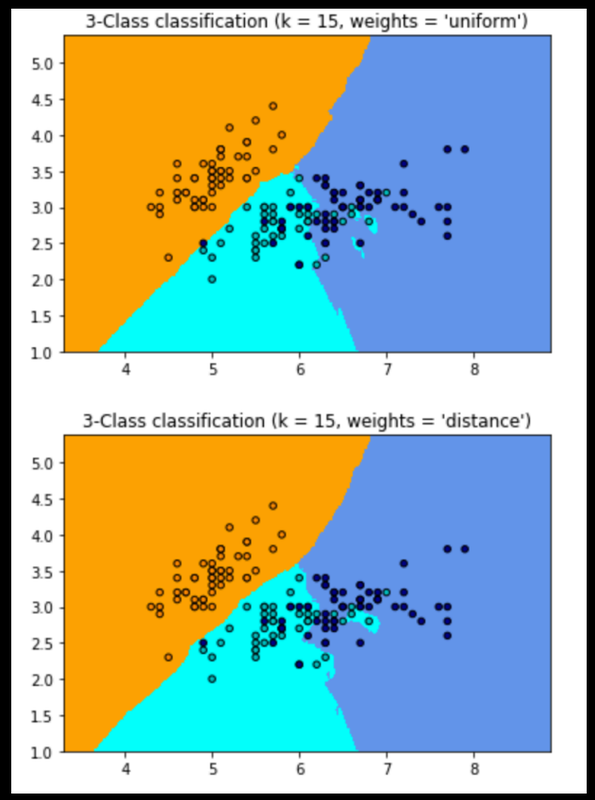
Predictive Modelling

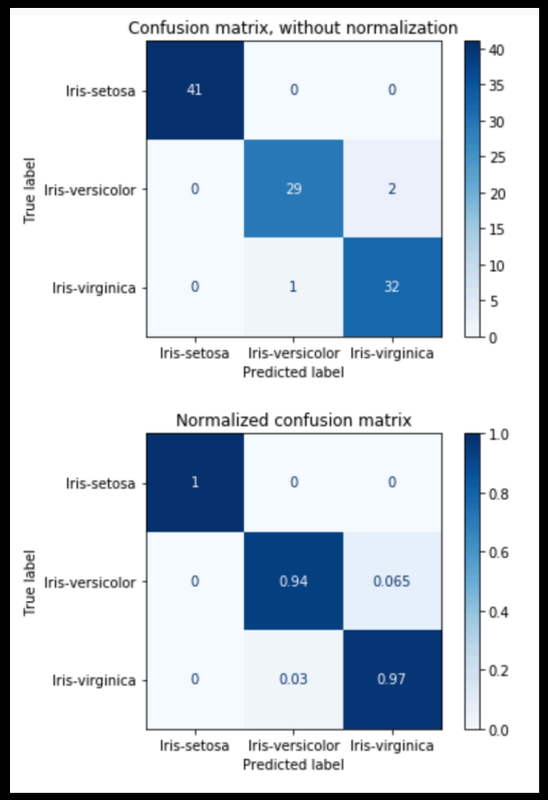
First of all let's take a look at the Scatterplot matrices which are very good visualization tools and may help identify correlations or lack of it:



Predictive Modelling

Applying the KNN classification, the results and the confusion matrix are:





In this case the model accuracy is: 0.977:

	precision	recall	f1-score	support
class 0	1.00	1.00	1.00	9
class 1	1.00	0.95	0.97	19
class 2	0.94	1.00	0.97	17
accuracy			0.98	45
macro avg	0.98	0.98	0.98	45
weighted avg	0.98	0.98	0.98	45

Conclusions and suggestions for future works

In this study, I classified Iris Dataset used SVM and KNN based on the petal and sepal sizes.
Use different classification algorithms to give alternative classes for the flowers,
and tag (e.g. by a new attribute) which instances were assigned different classes according
to the diffferent classifiers.

There are other classification and clustering algorithms such as decision tree,
Naive Bayes Random Forest and logistic regression which could be applied for this problem.