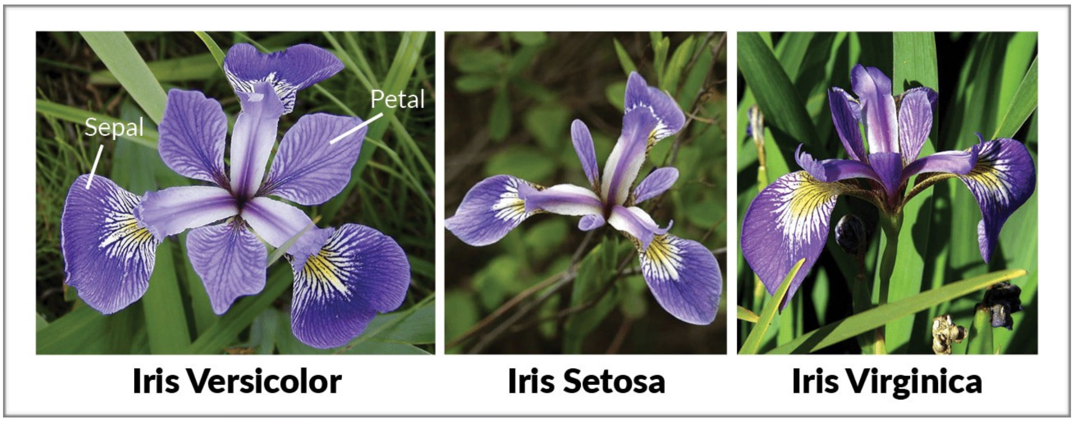
Classification of Iris Flower

1. Introduction

The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician, eugenicist, and biologist Ronald Fisher in his 1936 paper The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis. It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species. Two of the three species were collected in the Gaspé Peninsula "all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus".



It would be interesting and useful to classify the different types of Iris flowers having the sepal and petal sizes.

2. Data acquisition and cleaning

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. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other. The Iris dataset can be downloaded in the link below:

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

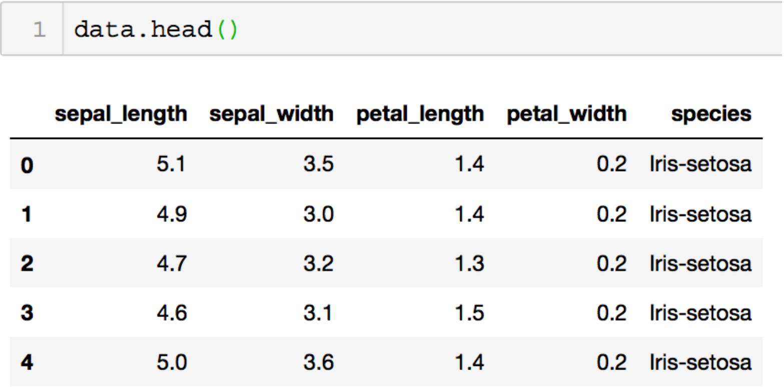
In this study I try to clustering Iris Dataset used SVM and KNN.

Attribute Information:

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

class: -- Iris Setosa -- Iris Versicolour -- Iris Virginica

The structure of the dataset is like this:



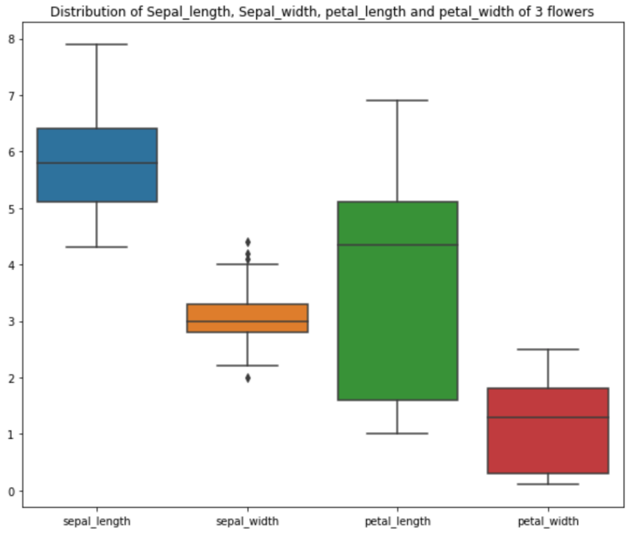
```
data.head()
```

	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

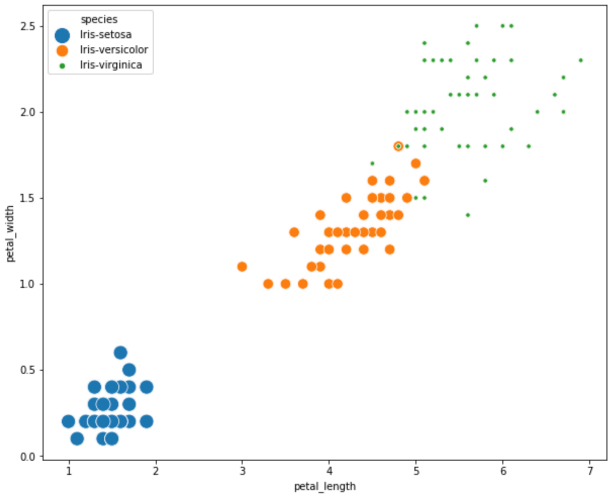
The features are **sepal_length, sepal_width, petal length** and **petal** width. The target is the **species** which its attributes are Iris Setosa, Iris Versicolour and Iris Virginica.

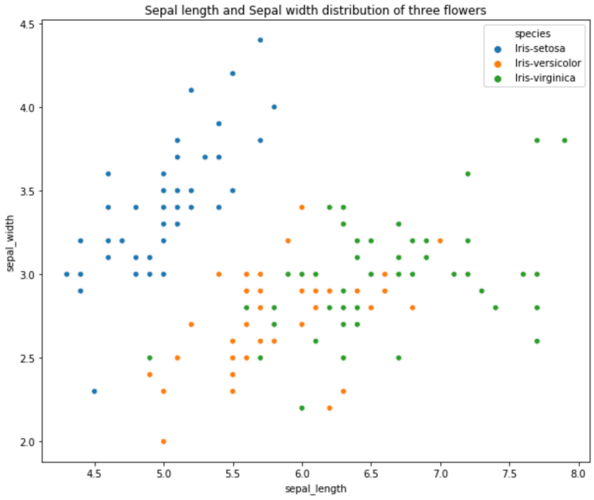
3. Exploratory Data Analysis

In order to get more familiar with data, I plot the 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:

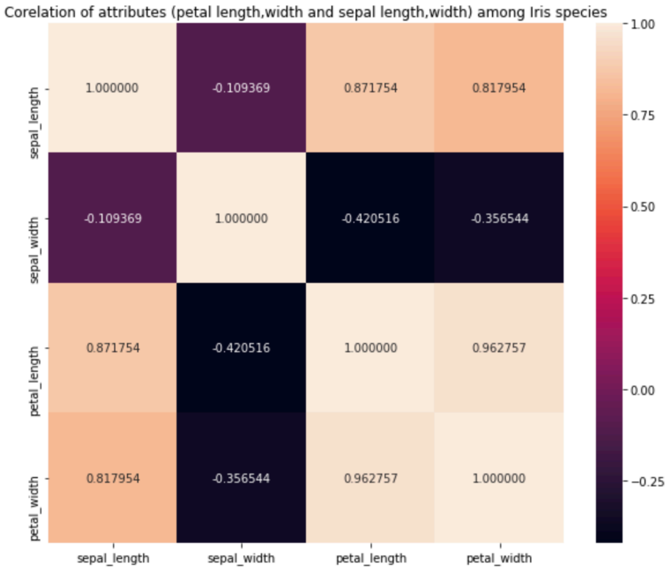


Distribution of Sepal_length, Sepal_width, petal_length and petal_width of 3 flowers





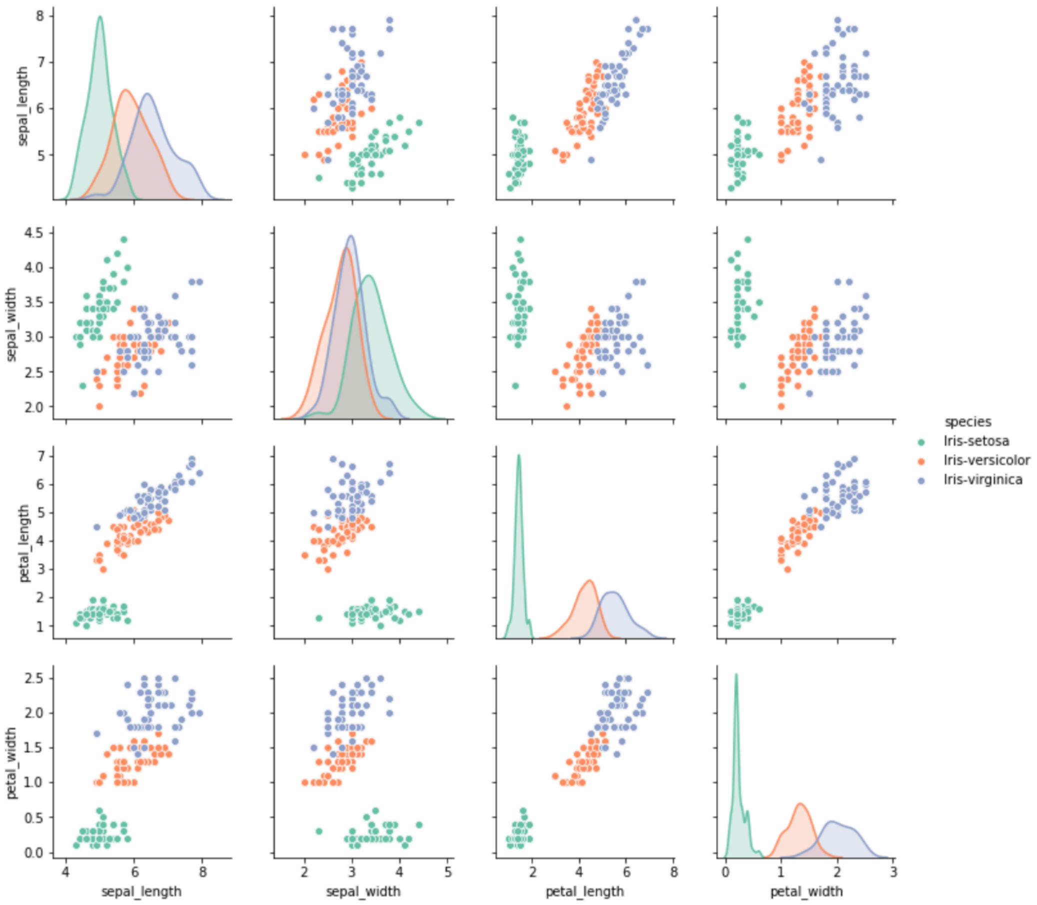
How does one variable compares to others? Are these correlated? The Correlation of all the attributes are plotted as:



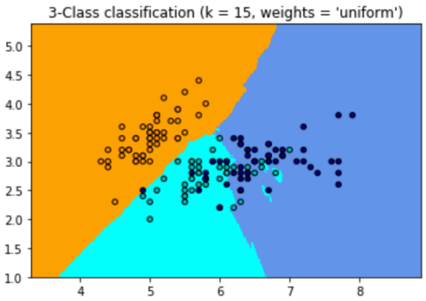
4. Predictive Modelling

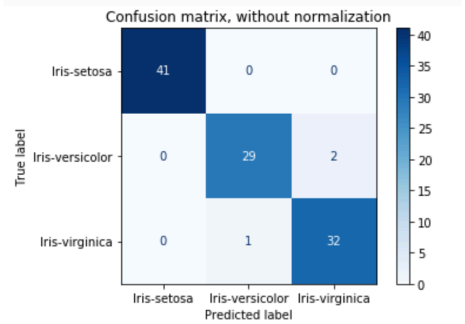
Even if we already know the classes for the 150 instances of irises, it could be interesting to create a model that predicts the species from the petal and sepal width and length. Two models that are easy to create and understand are K-nearest neighbor (KNN) and support vector machine (SVM).

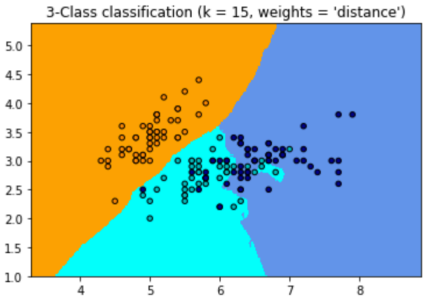
First of all let’s take a look at the Scatterplot matrices which are very good visualization tools and may help identify correlations or lack of it:

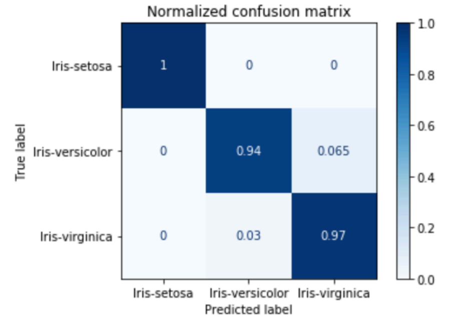


Applying the KNN classification, the results and the confusion matrix are:


3-Class classification (k = 15, weights = 'uniform')


Confusion matrix, without normalization


3-Class classification (k = 15, weights = 'distance')


Normalized confusion matrix

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
```

However using SVM algorithm also, the same accuracy has been achieved.

5. 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.