<https://machinelearningmastery.com/machine-learning-in-python-step-by-step/>

**Your First Machine Learning Project in Python Step-By-Step**

# Background

The Iris Dataset was collected by Edgar Anderson in 1935 and published by Ronald Fisher in 1936 (Wikipedia, 2019). It is a set of measurements of three species of Iris flowers. The measurements taken are:

* Petal length
* Petal width
* Sepal length
* Sepal width

Fifty of these measurements were recorded for each of the three species of Iris flower giving a dataset of 150 rows and 5 columns. The measurements of width and length are decimal numbers to one decimal place precision. The fifth column holds a text value denoting which species of Iris flower the measurements are associated with.

However, recent studies have shown that there is more than one Iris data set in the public domain. Indeed, Bezdek et al (Bezdek, 1999) found that there is at least one other version which has errors in the dataset.

Notwithstanding the above, the Iris dataset (in its various versions) has been used both as a pedagogical tool and as a test dataset for Machine Learning (ML) algorithms (Joseph, 2018). Machine Learning is defined as “**teaching computers how to learn without the need to be programmed for specific tasks. In fact, the key idea behind ML is that it is possible to create algorithms that learn from and make predictions on data”** (Maglaras, 2018)

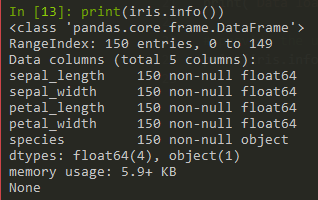
**This document will explore the Iris dataset with a view to describing it such that the reader can know what it contains and tells.**

# ****The process****

**As with any dataset, the first step in the process is to visually look through the dataset to:**

1. **Understand what the data contained shows**
2. **Get a sense of the data**
3. **Be aware of any errors or anomalies**
4. **Know the key attributes of the data**

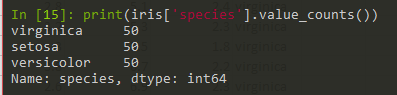
**Having browsed visually over the data, all columns seem to be homogenous. However, we can get Python to perform a quick overview of the data using *info()*. From this we have:**



**This shows that:**

* **There are 150 rows in the data**
* **There are 5 columns in the data**
* **There are no *NULL* values in any of the data**
* **The 4 length and width columns are decimal data (float64)**
* **Each column has a specific label**

**We can further see from counting the rows by ‘species’ type that:**

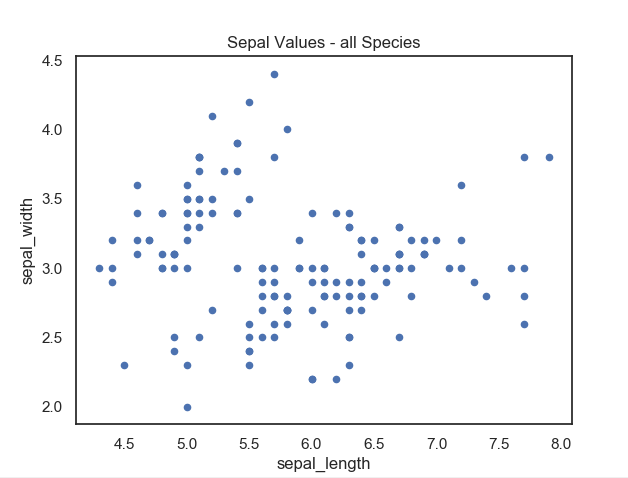


This shows that the data is spread evenly amongst three values in ‘species’. Namely; ‘virginica’, ‘setosa’ and ‘versicolor’ with 50 rows of data each.

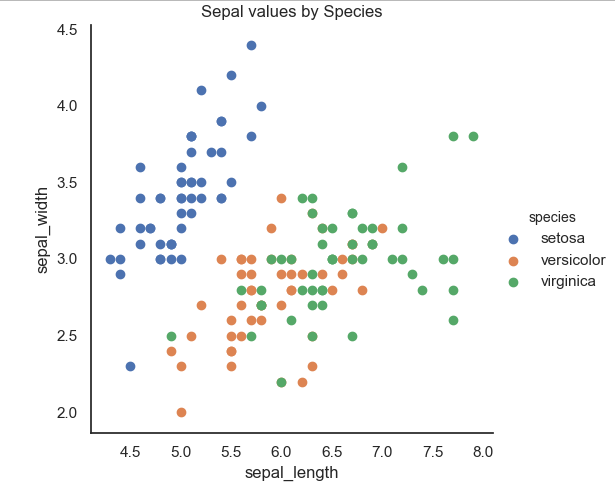
In summary, in the Iris dataset there are four numerical variables and one categorical variable with 50 rows of data in each category and no null values.

## High level analysis

Using python, we can start to examine the Iris dataset at a high level. Plotting all values in Sepal length against Sepal width we get:

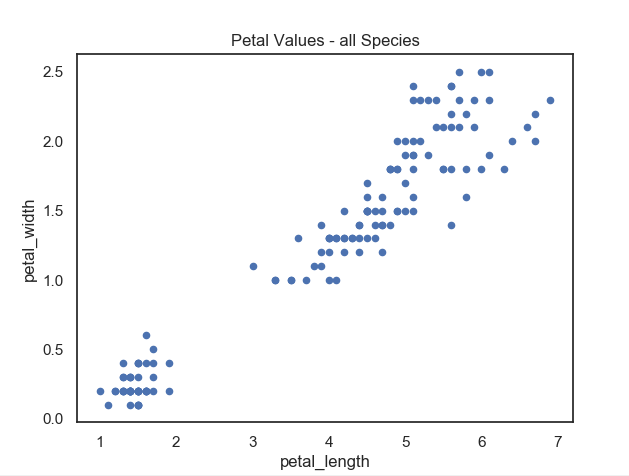


At first glance, we can see that the data largely aggregates into two groups. The first predominantly in the upper left quadrant and the second predominantly in the lower left and right quadrants. Further use of python to identify the values in the scatter plot by species shows us:



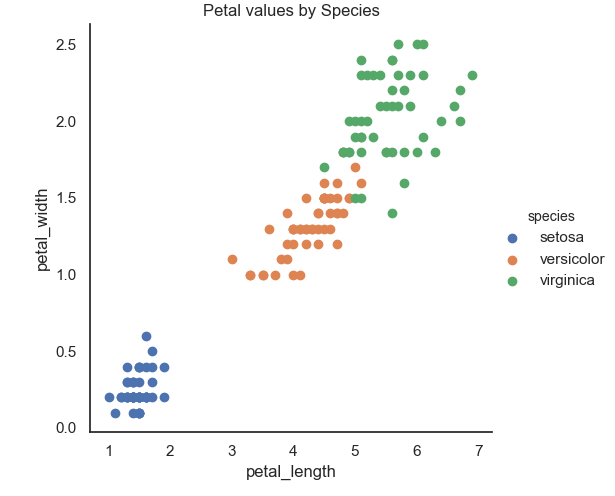
This shows us that our first observation above was correct in that the values which aggregated into the upper left quadrant belong to one species (*Setosa*). Furthermore, in the more scattered second group in the lower left and right quadrants, we can now see that within that are to fairly distinct groups with those in the lower left quadrant belonging to *Versicolor* and the remainder belonging to *Virginica*.

Similarly, plotting all values in Petal length against Petal width we get:



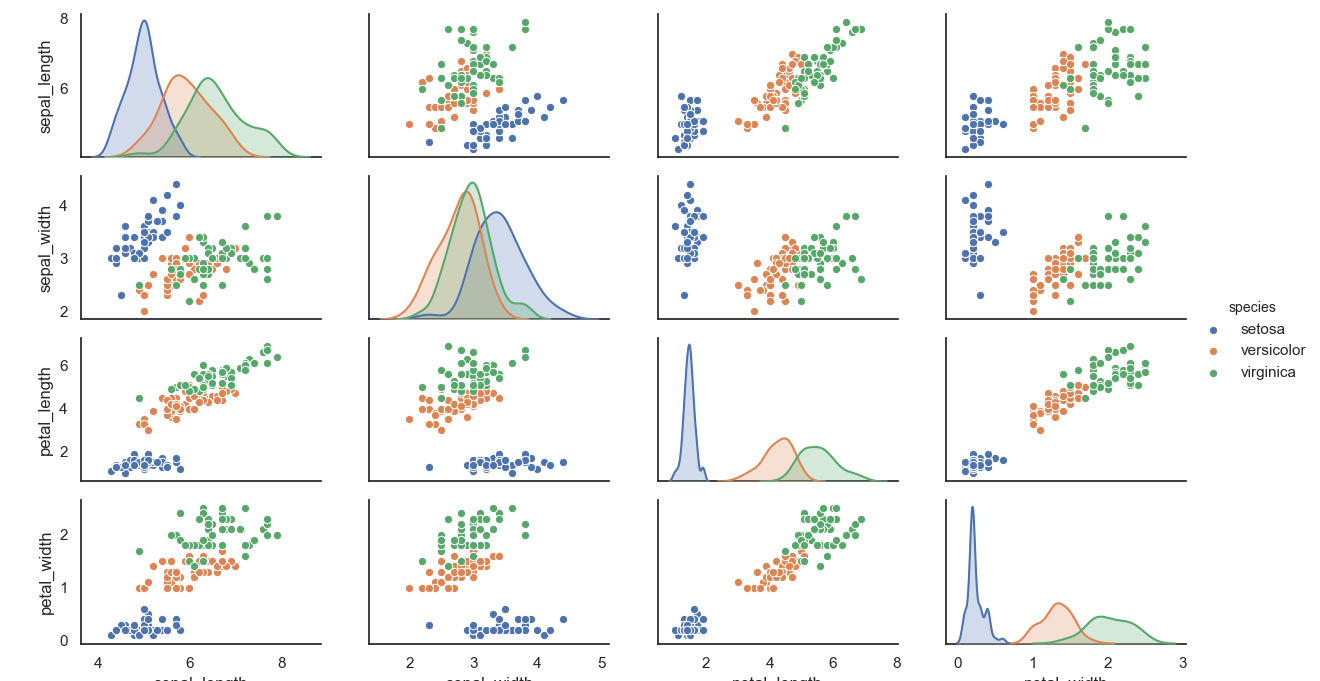
However, we can see on this occasion, the data is divided clearly into two groups. The first in the lower left quadrant (small petal width and length) and the second group largely within the upper right quadrant (large petal width and length).

As with the Sepal values we can identify the values by species giving us:



This shows us more clearly that species *Setosa* again is a distinct and different set of values to the other two species. However, in the Petal values we can see a clearer distinction between the remaining two species (*Versicolor* and *Virginica*).

However, what we can’t see is the relationship between the Sepal and Petal values by species.



Sepal length

Sepal width

Petal length

Petal width

From the above grid of plots it is clear that from this dataset we can identify the species of Iris based on the interrelationship between the Sepal length, Septal width, Petal length and Petal width. This is more clearly shown in the RadViz plot below: