# Project Iris

## Project Outline:

1. Data Description
2. Descriptive Statistics
3. Further Data Analysis
4. References

## Data Description

Iris flower data set is a well-known multivariate data set created by British statistician and biologist Ronald Fisher back in 1936 to demonstrate the work of the discriminant analysis method developed by him Alsow data called Anderson's Iris data set because Edgar Anderson as a botanist collected the data to quantify the morphologic variation of Iris flowers.

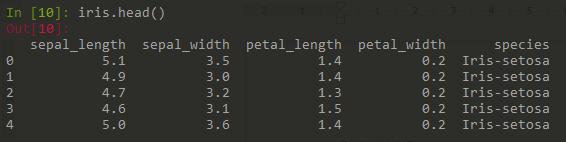
Fisher’s method became worldwide known data set example for data mining and data exploration for statisticians, mathematicians, programmers etc. Iris data set contains three different flower classes with numeric values for sepal and petal width and length.

The data set collected from three iris species (iris versicolor, iris setosa and iris virginica) and contained 150 records 50 each where flowers sepals and petals width and length measured in centimetres and all the species data carefully collected end measured for each flower.

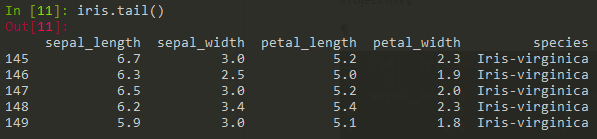
Traditionally this data set used for prediction, classification to identify the flowers as a certain type of irises. By visually observing this data you could see how data grouping in cluster with these three different types of irises. This data set has become a classic, and is often used in the literature to illustrate the work of various statistical algorithms.

## Descriptive Statistics

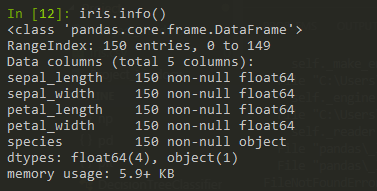
To view data set – top 5 lines:



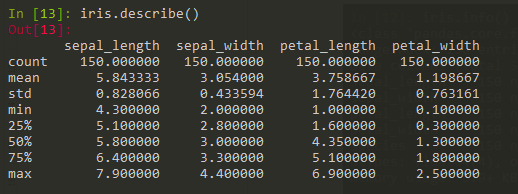
To view data set last 5 lines:



To gain basic info on data:

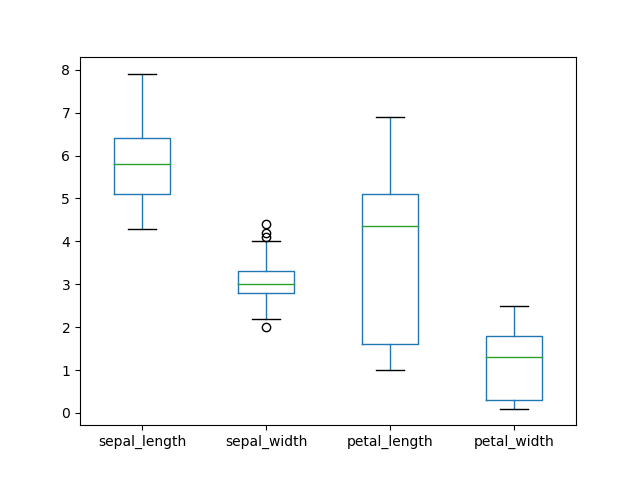


Basic stats of the numeric variables: shows number of observations, averages, standard deviation, min, max and 25th, 50th and 75th percentalies.

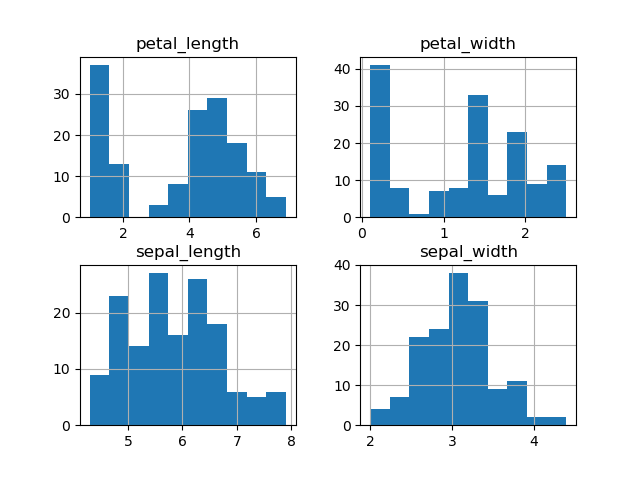


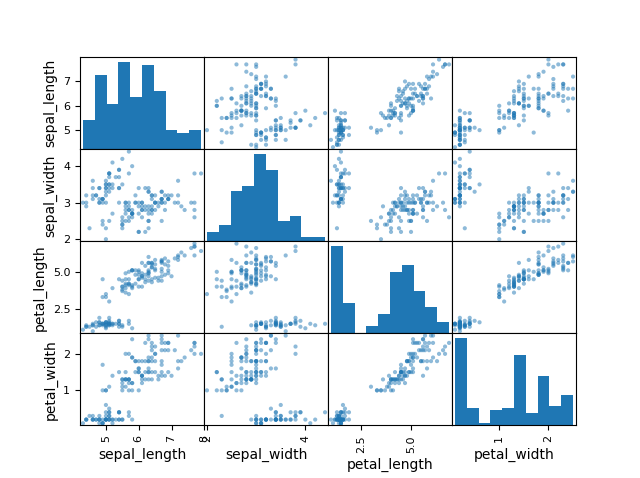
## Further Data Analysis

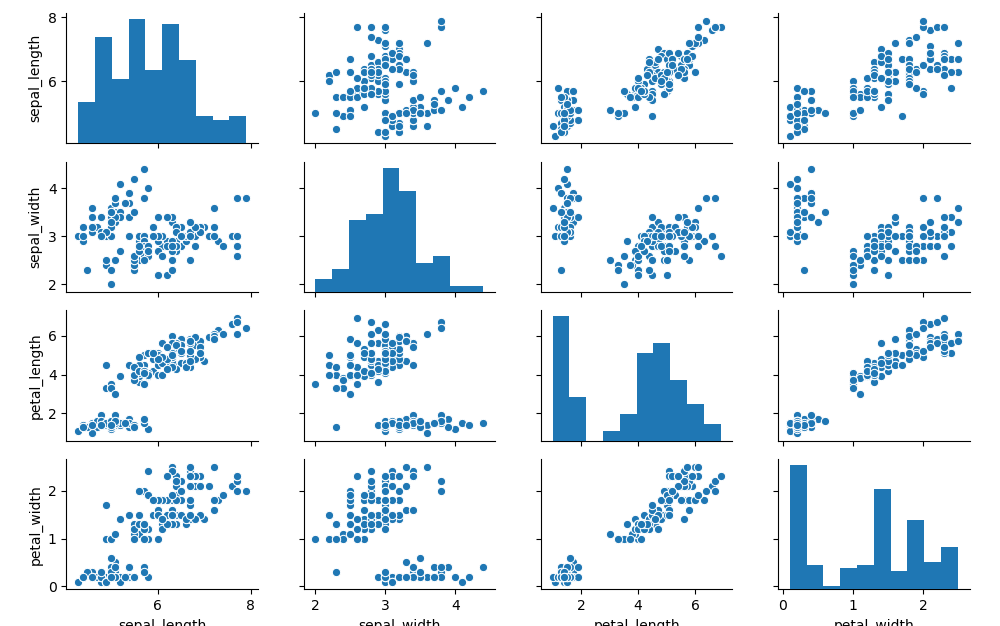
Box lot visualises statistic data average, median, min, max, percentiles. It allows one to quickly determine if the averages are representative. It also plots so-called outliers (circles on the plot).



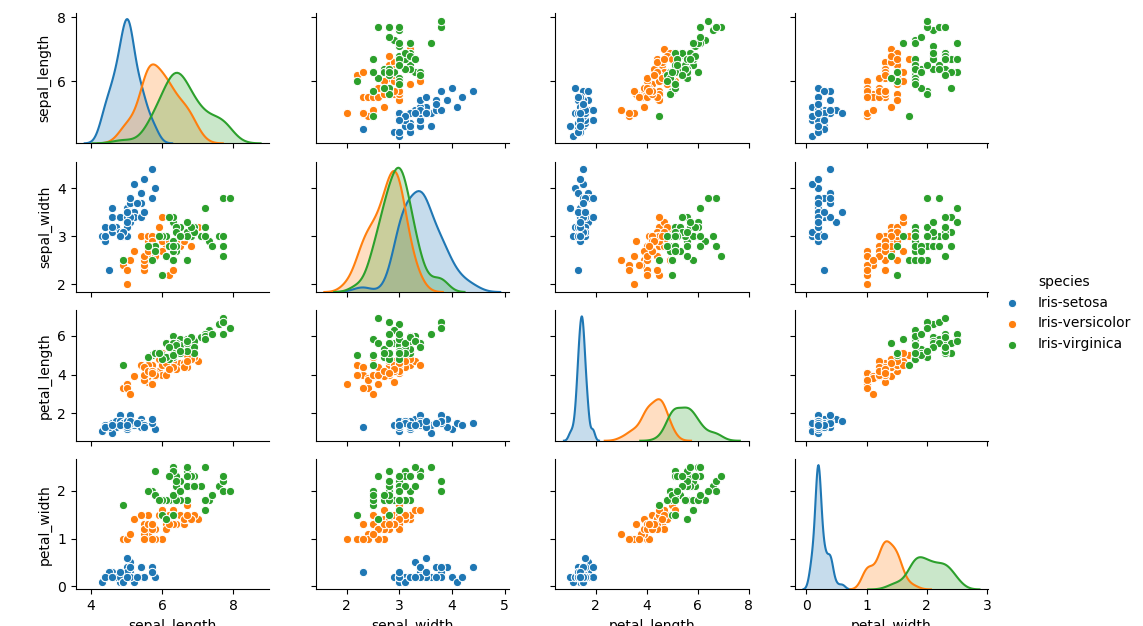
Histogram is another way to gain insights into data distribution.

This is important as some statistical analysis requires normal distribution of data. The only variable that approximates normal distribution is sepal\_width. Petal\_length and petal\_width variables have a very strange distribution. It looks like two different data sets.

Adding scatter plots helps to see further that parts of data on all variables stands separetely. The graph below 

Similar graph is available in seaborn module. It’s a matter of personal preference but it looks a bit better. searborn.scatter() allows to display the following:

Seaborn.scatter( by speacies) allows to separate data on a scatter plot with different colour for species.. Right away it is obvious that data for different species clusters together. Iris setosa stands apart on all dimensions.



Iris data set is very useful in testing classification algorisms in machine learning.

The following algorisms were used for this project, Python code adapted from https://medium.com/codebagng/basic-analysis-of-the-iris-data-set-using-python-2995618a6342:

LogisticRegression  
LinearDiscriminantAnalysis  
KNeighborsClassifier  
DecisionTreeClassifier  
GaussianNB  
SVC

The data is split in training and validation parts.

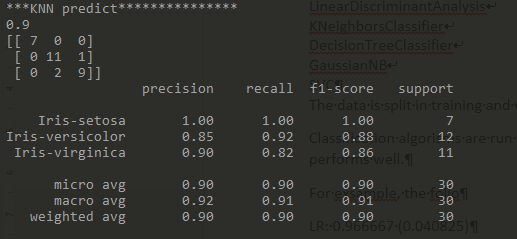
Classification algorisms are run on training part of data. The matrix allows to see which algorism performs well.

For example, the following is the output from all training algorisms:

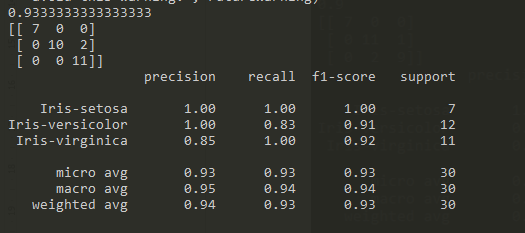
LR: 0.966667 (0.040825)  
LDA: 0.975000 (0.038188)  
KNN: 0.983333 (0.033333)  
CART: 0.966667 (0.040825)  
NB: 0.975000 (0.053359)  
SVM: 0.991667 (0.025000)

KNN and SVM seem to be performing very well. These can be used to make predictions:

KNN predicts with 90% accuracy:



While SVM has 93% accuracy:



## References

https://medium.com/codebagng/basic-analysis-of-the-iris-data-set-using-python-2995618a6342