

Iris Analysis Report

Samrudha Medhekar | Data Analytics in R | 31/07/2018

# Executive

An extensive analysis is performed on various features and species of Iris flowers and the correlation between them. Predictive techniques like the Decision Tree Algorithm is implemented to predict the species of flowers based on their features.

* There are 150 observations with 4 features each (sepal length, sepal width, petal length, petal width).
* There are no null values, so we don't have to worry about that.
* There are 50 observations of each species (setosa, versicolor, virginica).

## INtroduction

The data analysis on the Iris dataset was performed by Samrudha Medhekar. We explore the data to:

Understand the data

Summarize the data

Clean and Prune the data

Understand relationships between atrributes

Think about and source other data which maybe useful in answering the question

Get a preliminary feel for the types of models we think would best fit the data

## Background:-

The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician 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 as 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".

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.

## Purposes:-

The dataset describes particular biological characteristics of various types of Iris flowers, specifically, the length and width of both pedals and the sepals, which are part of the flower’s reproductive system. New types of sorting models and taxonomy algorithms often use the Iris flower data set as an input, to examine how various technologies sort and handle data sets. Programmers might, for example, download the Iris flower data set for the purposes of testing a decision tree, or a piece of machine learning software. For this reason, the Iris flower data set is built into some coding libraries, in order to make this process easier.

## Limitations:-

Only 3 species of Iris flowers are used for analysis, over plenty of other species, which effectively lessens the ability of the machine to predict species features and variations accurately.

## Methods:-

The following methods were used to perform the analysis :-

1. Summary and Str function :- Summary() and str() are used for performing exploratory analysis of the dataset to observe the structure and features of the dataset.
2. Boxplot () – Boxplot function is used to visualize the results of the summary function in a graphical form. Boxplots are plotted for the datset based on the species.
3. Pie() – Pie function is used for plotting the pie chart of the species distribution in the dataset.
4. Subset() – Subset method is used to classify the Iris dataset based on the species of the flowers.
5. Cor() – Cor () method is used for finding the correlation between different features of the dataset.
6. Decision tree prediction :- A decision tree is modelled to predict the species of the flowers based on the features in the dataset.

## Sample:-

The analysis was carried out by Samrudha Medhekar under the guidance of Neeraj Sir for Suven Consultants.

## Instrumentation:-

The following tools and libraries were used to perform the analysis:-

1. R Studio :- For performing the analysis in R language.
2. MS Excel :- For fetching the dataset.
3. Library(party) :- For plotting the decision tree.

## Results:-

The results have shown that classification of the flower species depend on their petal length and petal widths.

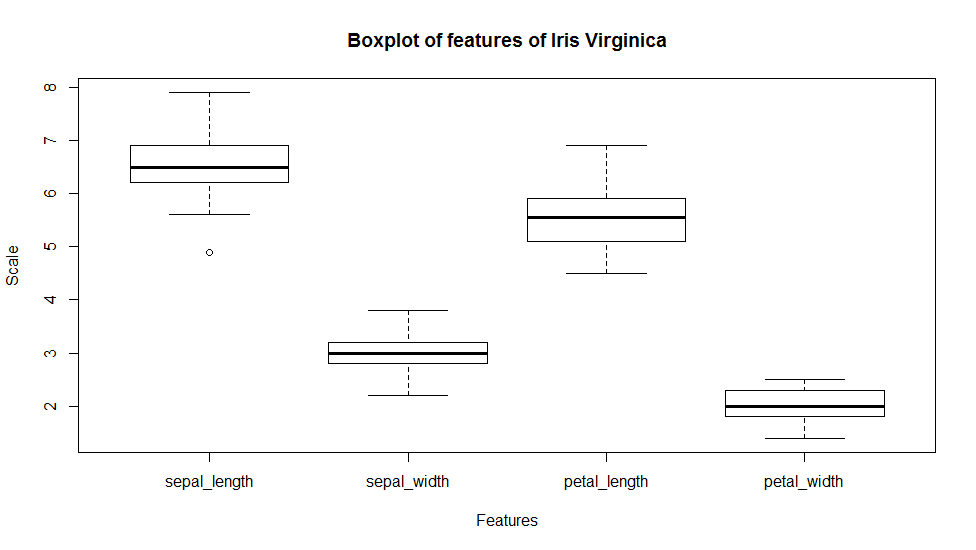
1.The flowers having petal lengths greater than 1.9 and petal width greater than 1.7 belong to the species of Iris Virginica.

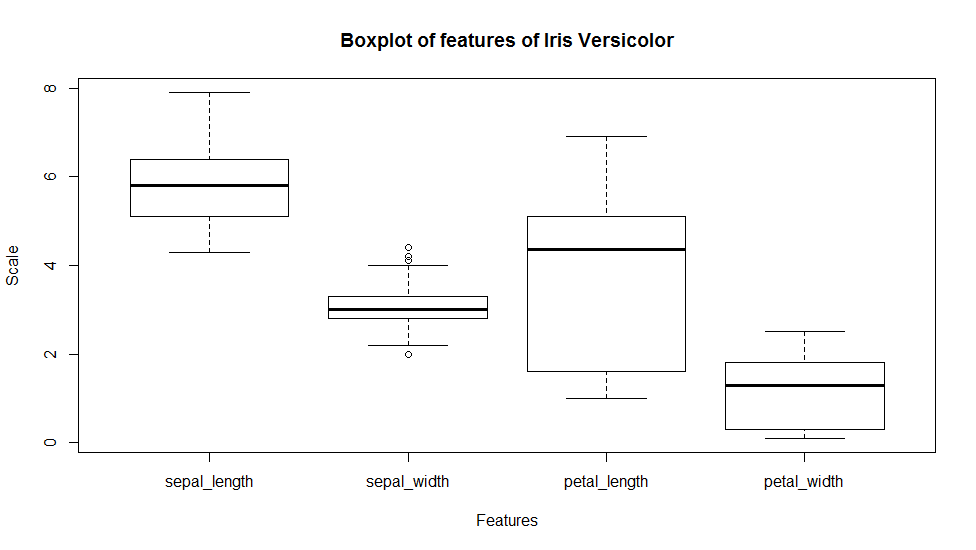
2.The flowers having petal widths less than 1.7 and petal lengths less than 4.8 belong to the species of Iris Versicolor.

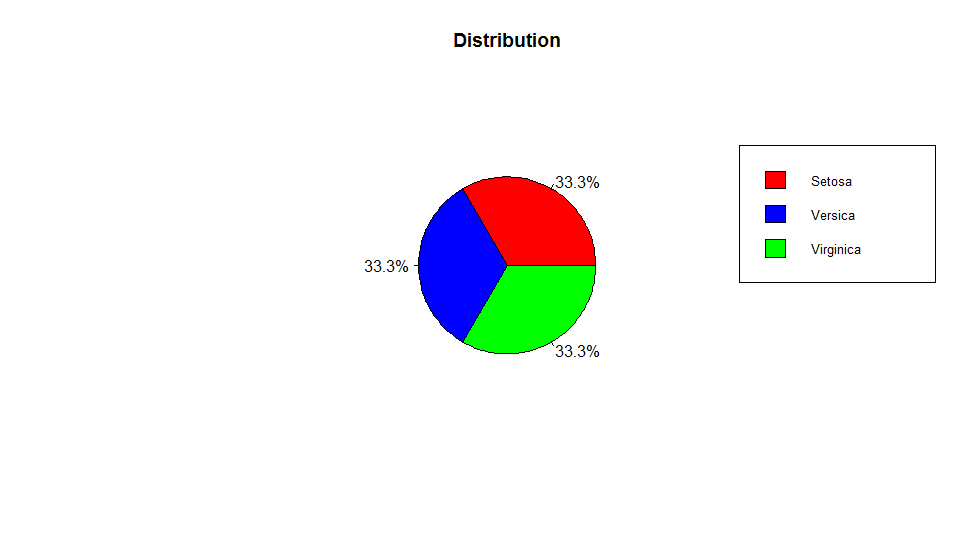
3. The flowers having petal lengths less than 1.9 belong to the species of Iris Setosa.

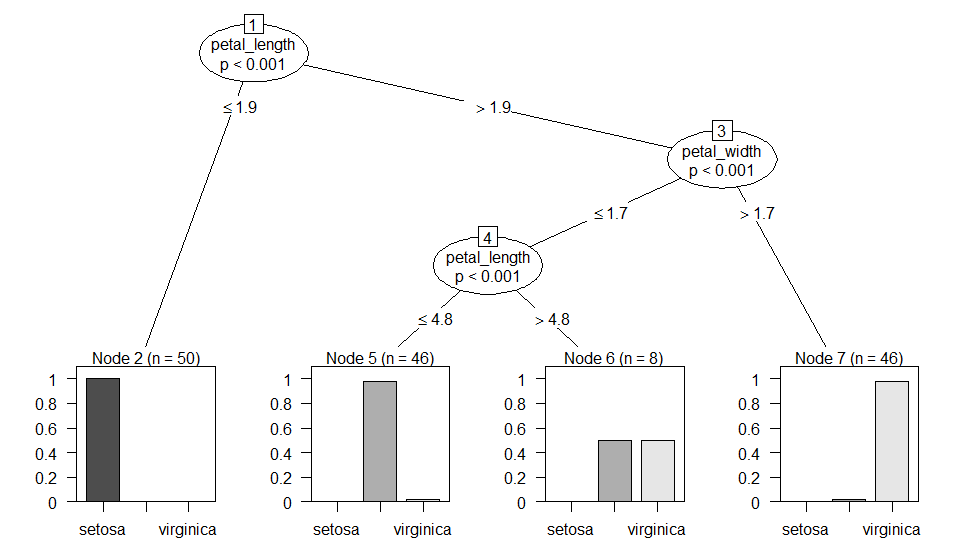
4. The flowers having petal widths less than 1.7 and petal lengths greater than 4.8 are equally distributed between the species of Iris Virginica and Iris Versicolor.











## Recommendations:-

The analysis of the data gives satisfactory results in species prediction. It is recommended to make use of more options in charts and plots to effectively visualize the data. Use of predictive algorithms is also recommended. The analysis can be extended to different species of flowers and their features.

## SUMMARY:-

The overall findings of the analysis indicates a high correlation between sepal length and petal length (87.17%) as well as between petal length and petal width (96.27%). There are no missing values which makes the data set easy to analyze. The three species are equally distributed, ie., each have 50 observations each.

## REFERENCES:-

<https://www.kaggle.com/antoniolopez/iris-data-visualization-with-r>

<https://rpubs.com/rpadebet/269829>

<http://msudatascience.com/blog/2016/8/27/quick-analysis-in-r-with-the-iris-dataset>

<https://rpubs.com/moeransm/intro-iris>

<http://www.lac.inpe.br/~rafael.santos/Docs/R/CAP394/IntroEDA-Iris.html>