Data Analysis on Iris Dataset

BY JAINAB KHATRI

(200180107002)

PYTHON FOR DATA SCIENCE(3150713)

Data Analysis

Data Analysis is the process of examining the data to obtain the useful statistical data which helps in solving the problem.

We use python for data analysis because it open source, free, flexible, easy to learn and python have huge set of libraries which are used for data analysis and its visualization.

Steps for Data analysis are:

1. Define the goal
2. Collect the required data
3. Cleaning the data
4. Analyse the data
5. Visualize and present the data

Steps for data Analysis

1. Define the goal:

   Defining the goal means we should probably know the problem statement i.e. why are we doing this data analysis.

2. Collecting the data:

   Gather the required data from the valid source.

3. Clean the Data:

   Cleaning the data means doing Exploratory data analysis which is also known as data exploration in which we convert the data in complete useable form.

   Here we remove the unwanted data , duplicates ,basic errors and white spaces from the dataset.

4. Analyse the data:

   Here we actually analyse the given dataset i.e. extracting actual information that we want from data.

5. Visualize and present the data:

   Visualization means presentation of your information in the graphical format so that the people can understand it better. And it helps in comparing the two data easily.

What is Iris dataset and what are the attributes of this dataset?

The iris dataset is a multivariate set of data which comprises of 150 samples of 3 different species (50 each) and 6 columns.

There are 3 kind of flowers and that are Iris setosa , iris virginica , iris versicolor .

It is also knows as fisher’s iris dataset.

The attributes are:

1. Sepal length
2. Sepal width
3. Petal length
4. Petal width

iris setosa

iris versicolor

iris virginica

Analysing the iris dataset

The dataset has been downloaded from https://www.kaggle.com/ .

To analyse the data we have import the libraries of python and convert the dataset in form of data frames by reading the csv file using read_csv() method.

Data visualizing is done using line graph, bar graph and histography.

Python file is attached with this file for code.

Thank you