**Case Study using IRIS data set**

**Data set description**: The Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in 1936. The data was collected to quantify the morphological variation of *Iris* flowers of three related species namely, setosa, verginica and versicolor. The data consists of 50 samples from each of the three species of IRIS. 4 features were measured from each sample: the length and the width of the sepals and petals of the flower in centimeters. Fisher developed a machine learning model to classify the species from each other using the combination of 4 features. This is one of the benchmark data set popularly being used machine learning practitioner.



|  |
| --- |
|  |
| **Details of the IRIS data set** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 150 | **Area:** | Life |
| **Attribute Characteristics:** | Real | **Number of Attributes:** | 4 | **Date Donated** | 1988-07-01 |
| **Associated Tasks:** | Classification | **Missing Values?** | No | **Number of Web Hits:** | 315055 |

**Attribute Information:**

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

Load the “Iris data.xls” and perform the following operations:

1. Describe the data with its statistical properties.
2. Find out the distribution of the given data set.
3. Express the relation of each feature with the Class and comment on the significance of each feature in the data set. Also, suggest which combination of the feature would represent the class effectively.
4. Suppose, 3 samples are picked randomly from the total sample space of size 150, then what is the probability that at least one sample is picked from each class.
5. Establish a feature to feature relation and comment on the results based on your observation.
6. Perform an appropriate statistical hypothesis testing to determine correlation between the features.
7. Transform the data set using min-max scaling. Repeat all the operations mentioned from question 1 to question 6 and compare the results.