EXECUTIVE

The Iris Flower Dataset is a popular multivariate dataset that was introduced by R.A. Fisher as an example for discriminant analysis. The Iris

dataset has been analysed via two distinct methods. First, plotted the dataset onto scatterplots, boxplots, histogram and decision tree to determine patterns in the data in relation to the Iris classifications. Second, series of methods on the dataset to extract relevant statistical information from the dataset.With these two methods, I can make concrete predictors about the dataset.

INTRODUCTION

The Iris Flower Dataset is a popular multivariate dataset that was introduced by R.A. Fisher as an example for discriminant analysis. The data reports on four characteristics of the three species of the Iris Flower, sepal length, sepal width, petal length, and petal width. The goal of a discriminant analysis is to produce a simple function that, given the four measurements, will classify a flower correctly. This is the beginning of creating “predictors” in order to try to make a more educated guess on a record in a dataset. This article will attempt to analyse this dataset to try to draw some conclusions from the model

purpose of this analysis is to classify the types of the flowers based on their sepal length sepal width petal length and petal width

also we can study that the lengths and widths are affecting the type of the flower or not

the minimum length or width as well as maximum length and the width of every type of the flower

we are making the decision tree to exactly classify the types of the flowers according to strong fields like sepal length petal width sepal width petal width

Methods

Methods involved in this analysis is plotting of histogram ,boxplot, functions like correlation, decision tree algorithm based methods and KNN algorithm

RESULTS

We are drawing the histogram of Petal length ,so we can find out which length of the flower is minimum medium and maximum

Minimum and maximum petal length are having low frequency and medium petal length are having more frequency

We have plotted the BOXPLOT of subset of dataset

We have plotted the HISTOGRAM of subset of dataset

We have found out correlation between sepal length and Sepal width also the correlation between Petal length and Petal width

correlation between sepal length and sepal width is -0.1093692

correlation between petal length and petal width is 0.9627571

We have found out the correlation matrix for SETOSA respective of every column ie sepal length, sepal width, petal length, petal width with same four

we have also classified the data using decision tree .

as we have classified the 80 percent of data as training as 20 percent of data as testing data

We have used KNN to seperate the dataset into training and testing set .We are training our data using KNN and then checking is it working properly

or not by testing data. And KNN on dataset is working properly

We have used rpart and rplot for drawing the decision tree

petal length less than 2.5 are setosa and others are further classified. From this remaining data petal length less than 5 are versicolor and remaining are virginica

SUMMARY

In evaluating the Iris dataset, I found it important to be able to reference graphs of the data with the statistical analysis run from the application. In

the future, it would be a nice feature to be able to add visualization with the analysis, instead of static graphs generated in Excel. Along with the framework mentioned earlier in the article, having the data stored as XML and using VTK to render the data would be a nice step in creating interactive visualizations of the data. This would be the next step on the journey to create an interface for interactive visualizations