Iris Dataset

Gibril

Introduction

Fisher’s Iris dataset is famous for testing different classification algorithms. It contains measurements (in cm) of sepal length and width and petal length and width, respectively, for 50 flowers of the 3 Iris species (Iris setosa, versicolor, and virginica).

The aim of this project is to use various supervised learning techniques to classify the species of a given Iris flower.

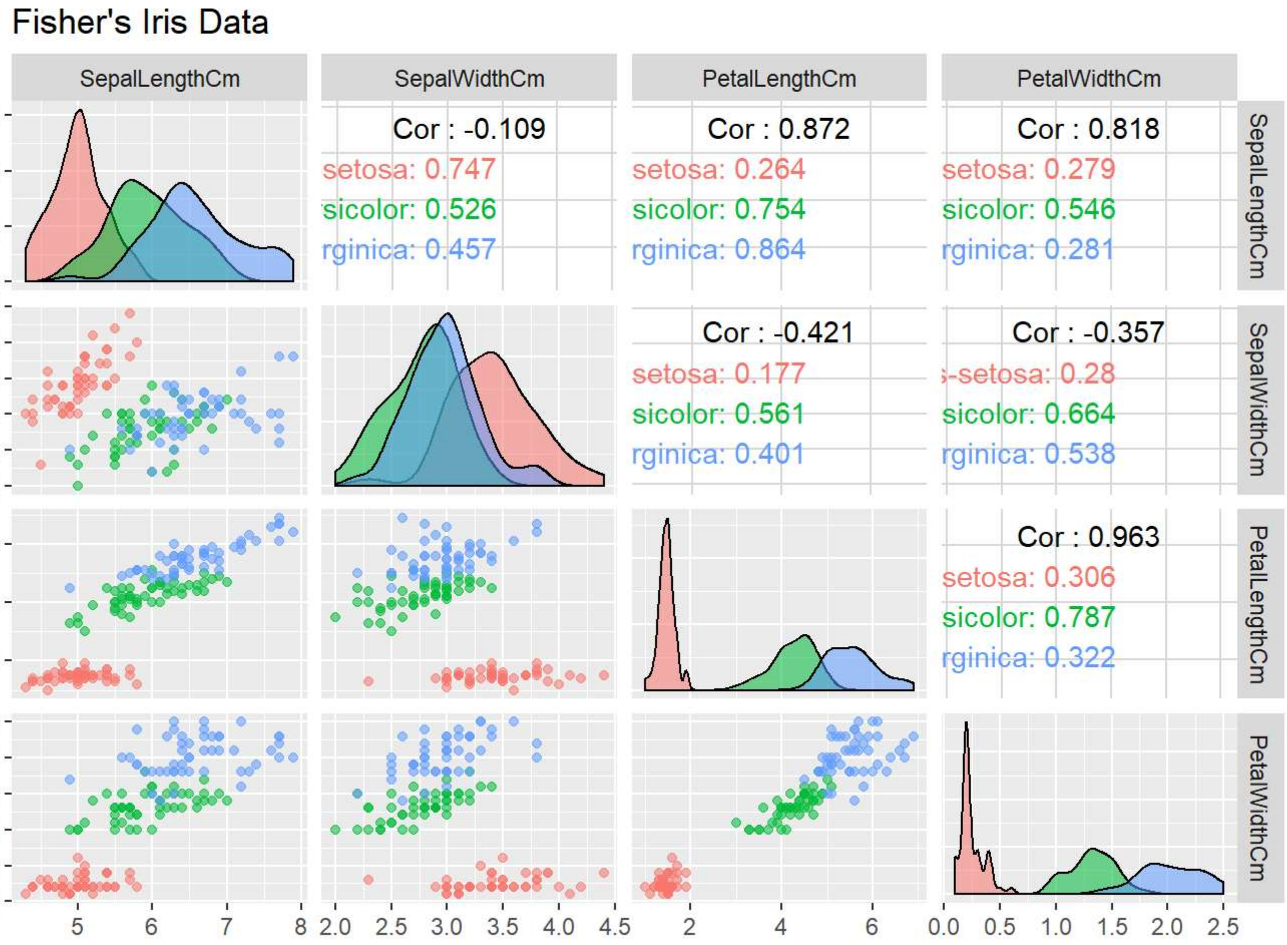
Dataset

The dataset is freely available on the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/Iris). It includes 50 samples of each of the three species and some measured properties about each flower. As will be seen the graphs below, one flower species can be linearly separated from the other two whereas the other two are not linearly separable from each other.

The variables in the datasets are:

- Id
- SepalLengthCm
- SepalWidthCm
- PetalLengthCm
- PetalWidthCm
- Species
    - Iris-setosa
    - Iris-versicolor
    - Iris-virginica

The graph below show the distributions and pairwise relationships of the continuous variables in the dataset.



And finally, a correlation matrix of the variables in the dataset (excl. Id). The new variable `SpecNum` is the coded version of the `Species` variable.

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	SpecNum
SepalLengthCm	1.000	-0.109	0.872	0.818	0.783
SepalWidthCm	-0.109	1.000	-0.421	-0.357	-0.419
PetalLengthCm	0.872	-0.421	1.000	0.963	0.949
PetalWidthCm	0.818	-0.357	0.963	1.000	0.956
SpecNum	0.783	-0.419	0.949	0.956	1.000

The table shows that `Species` is highly correlated to `SepalLengthCm`, `PetalLengthCm` and `PetalWidthCm`. From the above scatterplot matrix, we see that these variables can be used to classify `Species` with less error and so will be prioritised in variable selection.

Training Method(s)

Results

Summary