Fischer Iris Dataset Analysis

Kirtimaan Gogna
2019AIM1014@iitrpr.ac.in
Indian Institute of Technology
Ropar

Abstract

The purpose of this report is to perform exploratory data analysis on the iris dataset. We will learn how to load and hadle data. It has only 4 attributes and 150 rows, meaning it is small and it is easy to fit in the memory. It gives an opportunity to apply various machine learning algorithms and compare the results.

1 Introduction

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. This is a very famous and widely used dataset by everyone trying to learn machine learning and statistics. 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 centimetres. The fifth column is the species of the flower observed. We are going to load the dataset directly from the scikit learn library.We are going to:

- Visualize the distribution of sepal length and sepal width.
- Scatter plot of Petal length and petal width.
- Make boxplots of for each of the four attributes, sepal length, sepal width, petal width and petal length.

After performing the exploratory data analysis and after plotting these graphs we are going to apply various supervised and unsupervised learning algorithms on it. These are

- Logistic Regression
- Gaussian Naive bayes
- K-means clustering for unsupervised learning method.

2 Exploratory Data Analysis

2.1 Scatter plots

Scatter plots give us a good idea of how much one variable is affected by another.As shown in Fig, there is a high correlation between the sepal length and the sepal width of the Setosa iris

flowers, while the correlation is somewhat less high for the Virginica and Versicolor flowers: the data points are more spread out over the graph and do not form a cluster like shown in the case of the Setosa flowers. The scatter plot that maps the petal length and the petal width displayed as Fig tells a similar story: The graph indicates a positive correlation between the petal length and the petal width for all different species that are included into the "iris" data set. More importantly, the scatter plots reveal a strong classification criteria. Setosa has the smallest petals versicolor has medium-sized petals and virginica has the largest petals.

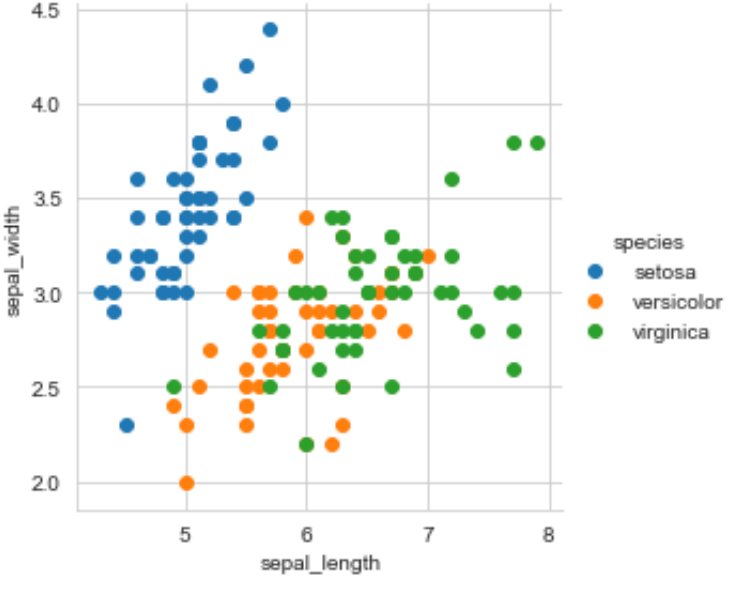

Figure 1: Sepal length and sepal width scatter plot

2.2 Histograms

What we see from those plots is that sepal length and sepal width do not vary much across species, however, petal length and petal width are quite different for different species. As shown in Figure 1 and Figure 2, petal length of versicolor and virginica are approximately normally distributed with different means and similar variability. Also, species setosa lies far away from these two species.

2.3 Boxplots

fter you check the distribution of the data by plotting the histogram, the second thing to do is to look for outliers. Identifying the outliers is important because it might happen that an association you find in your analysis can be explained by the presence of outliers. The best tool to identify the outliers is the box plot. Through box plots, we find the minimum, lower quartile (25th percentile), median (50th percentile), upper quartile (75th percentile), and a maximum of an continues variable. The function to build a boxplot is boxplot().

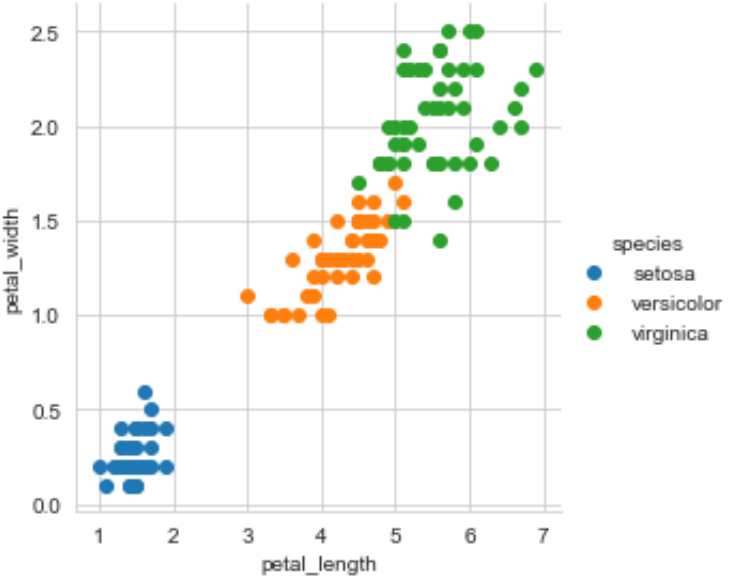

Figure 2: Petal length and Petal width scatter plot

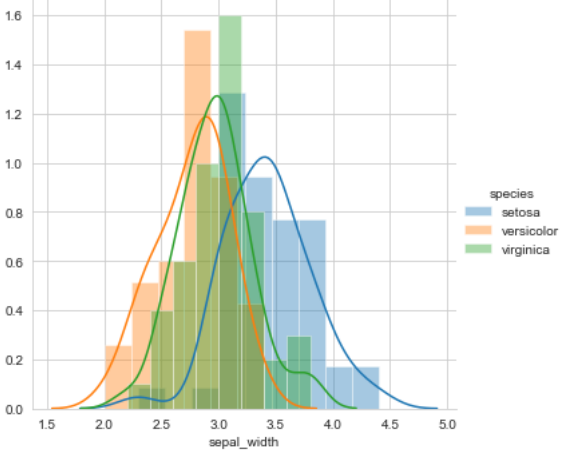

Figure 3: Histogram for sepal width

3 Supervised learning algorithms

3.1 Logistic Regression

Logistic regression is a model that uses a logistic function to model a dependent variable. Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent variable and one or more nominal, ordinal, interval or ratio-level independent variables. We are ap-

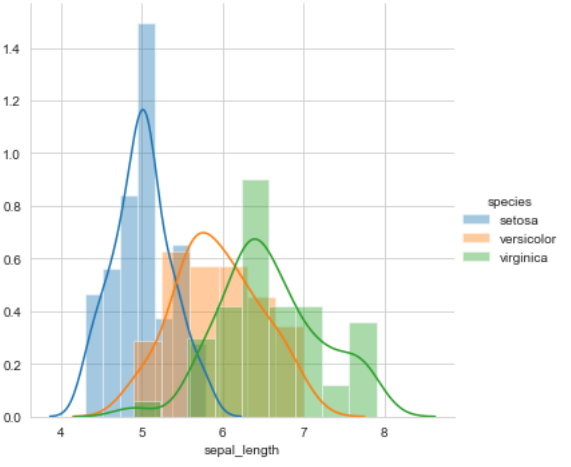

Figure 4: Histogram for sepal length

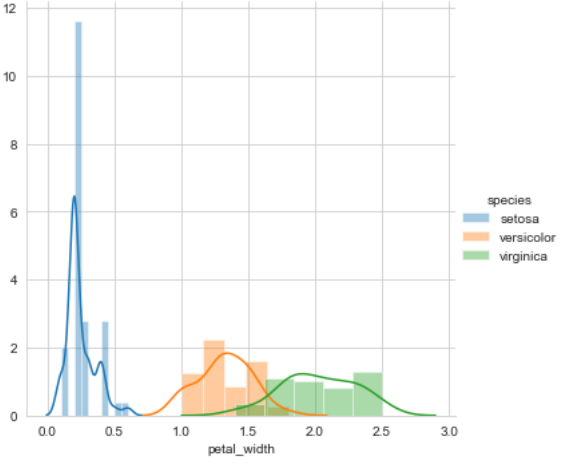

Figure 5: Histogram for petal width

plying cross validation score also on top of it.

Species	precision	Recall	f-1 score
0(setosa)	1.00	1.00	1.00
1(Versicolor)	1.00	1.00	1.00
2(virginia)	1.00	1.00	1.00
-	-	-	-
accuracy score	-	-	1.00
macro avg	1.00	1.00	1.00
Weighted avg	1.00	1.00	1.00

Classification report for Logistic Regression

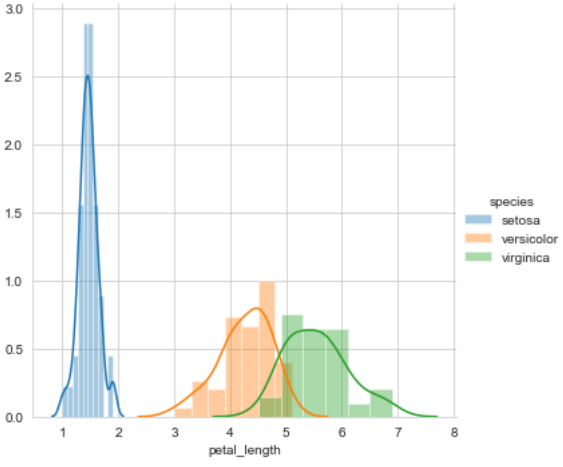

Figure 6: Histogram for petal length

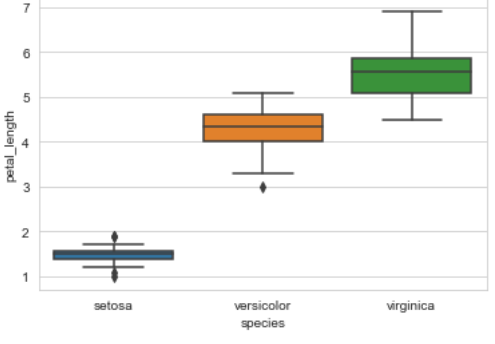

Figure 7: Box plot for petal width

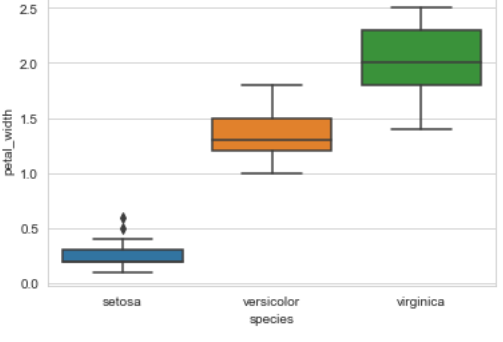

Figure 8: Box plot for petal width

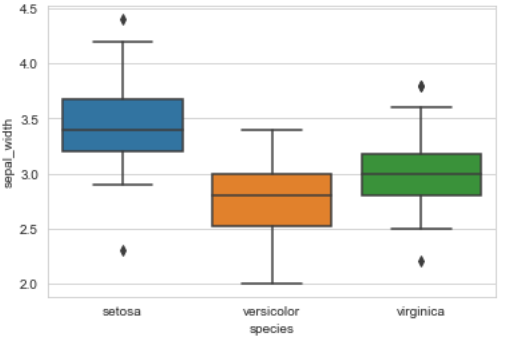

Figure 9: Box plot for sepal width

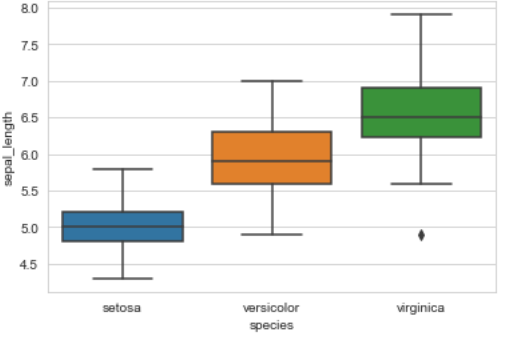

Figure 10: Box plot for sepal length

3.2 Gaussian Naive bayes

The Gaussian Naive Bayes is one classifier model. Beside the Gaussian Naive Bayes there are also existing the Multinomial naive Bayes and the Bernoulli naive Bayes. I picked the Gaussian Naive Bayes because it is the simplest and the most popular one.

Species	precision	Recall	f-1 score
0(setosa)	1.00	1.00	1.00
1(Versicolor)	1.00	1.00	1.00
2(virginia)	1.00	1.00	1.00
-	-	-	-
accuracy score	-	-	1.00
macro avg	1.00	1.00	1.00
Weighted avg	1.00	1.00	1.00

Classification report for Gaussian Naive bayes

4 Unsupervised learning

KMeans Clustering K-means clustering is one of the simplest unsupervised machine

learning algorithms.K-means converges in a finite number of iterations. Since the algorithm iterates a function whose domain is a finite set, the iteration must eventually converge The computational cost of the k-means algorithm is O(k*n*d), where n is the number of data points, k the number of clusters, and d the number of attributes.Compared to other clustering methods, the k-means clustering technique is fast and efficient in terms of its computational cost. It's difficult to predict the optimal number of clusters or the value of k. When not dividing the dataset into 2 parts that is training and the test set. We do not need to because we do not consider the labels assigned to the dataset. Therefore we will apply k means clustering on the entire dataset and check for the accuracy on it. In the next experiment I divided the dataset into 2 set, and check the accuracy scores.

Species	precision	Recall	f-1 score
0(setosa)	1.00	1.00	1.00
1(Versicolor)	0.77	0.96	0.860
2(virginia)	0.95	0.72	0.82
-	-	-	-
accuracy score	-	-	0.89
macro avg	0.91	0.89	0.89
Weighted avg	0.89	0.89	0.89

Classification report for KMeans custering

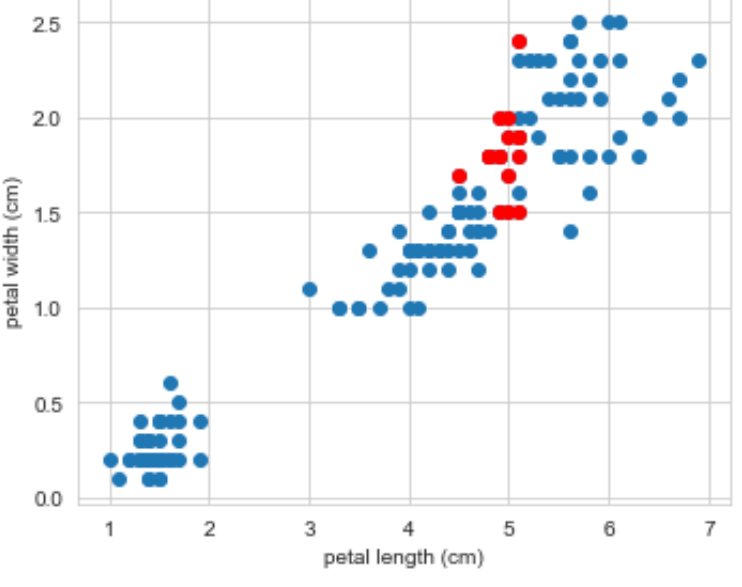

Figure 11: Misclassified points in Kmeans clustering

5 Conclusions and learnings

Model	1	2	3	4	5	mean
Logistic Regression	0.92	0.953	0.91	0.953	1	0.9560
Gaussian Naive bayes	0.96	1	0.831	1	0.913	0.94

KFold cross score

- Logistic Regression: We used it for finding various classes. We are getting an accuracy of 100 percent on the test set.
- Gaussian Naive bayes: We are getting an accuracy of 100 percent on the test set
- K-means clustering for unsupervised learning method.There are no labels, so the accuracy is a little bit less than other models/