Assignment-2 Report (Data Analytics with Fisher Iris Dataset)

Samreet Singh Dhami
2018CSB1120
IIT Ropar
Rupnagar,Punjab

IRIS DATA CHARACTERSTICS

Number of Instances: 150 (50 in each of three classes)

Number of Attributes: 4 numeric, predictive attributes and the class

Attribute Information: sepal length, sepal width, petal length, petal width

Class: Iris-Setosa, Iris-Versicolour, Iris-Virginica

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 for 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 width of the sepal and petal respectively, in centimetres. The fifth column is the species of the flower observed (class label).

EXPLORATION OF DATA

Visualization techniques are often specialized to the type of data being analyzed. Indeed, new visualization techniques and approaches, as well as specialized variations of existing approaches, are being continuously created, typically in response to new kinds of data and visualization tasks.

1. Visualizing the pdf and cdf of sepal width by grouping the sepal length int 10 bins

2. Scatter Plot of Petal Length and Petal Width

3. Box Plots of each of the four attributes: Sepal width, Sepal length, Petal length, Petal width

OBSERAVTIONS

1. The lower and upper ends of the box indicate the 25th and 75th percentiles, respectively, while the line inside the box indicates the value of the 50th percentile. The top and bottom lines of the tails indicate the 10th and 90th percentiles. Outliers are shown by points.

```
:Summary Statistics:

============== ==== ==== ======= ===== ====================
                Min  Max   Mean    SD   Class Correlation
============== ==== ==== ======= ===== ====================
sepal length:   4.3  7.9   5.84   0.83    0.7826
sepal width:    2.0  4.4   3.05   0.43   -0.4194
petal length:   1.0  6.9   3.76   1.76    0.9490  (high!)
petal width:    0.1  2.5   1.20   0.76    0.9565  (high!)
============== ==== ==== ======= ===== ====================
```

We can observe from box plots of training set that

PETAL LENGTH: very compact distribution of setosa with many outliers. Moreover, outliers are present in other species. Clearly, this is not a good characterstics for separating the flowers.
SEPAL LENGTH: compact distribution for setosa and outliers can be observed in virginica
SEPAL WIDTH: outliers present in setosa and virginica
PETAL WIDTH: very compact for setosa with outliers

2. From the scatter plot, it is observed that there is a clear demarcation of points of setosa from the other species

3. By using sepal width (as seen in the distribution curve of sepal width), we can't do anything because its all messed up and we can't separate the flowers

4. By looking at the classification report, Gaussian naive bayes and logistic regression are good models for training the dataset of iris ,but Kmeans Algorithm has an unpredictive nature. It depends highly on how the training and testing split has occured. It,sometimes,have high accuracy upto 0.9333 but sometimes goes down upto 0.05

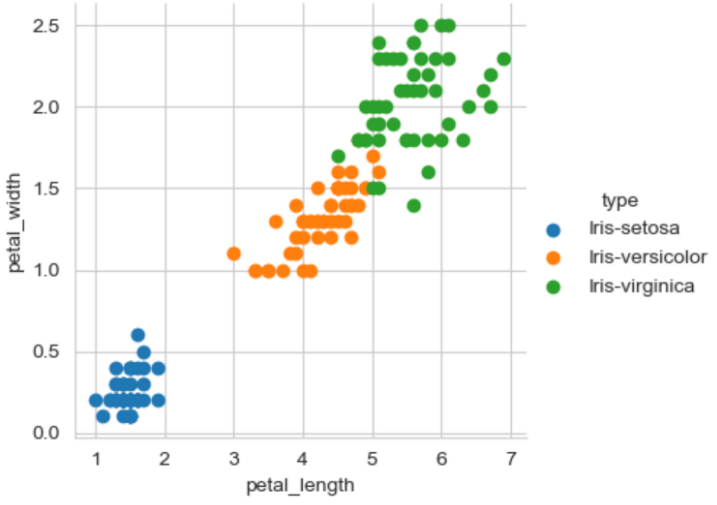

Figure 1: Scatter Plot

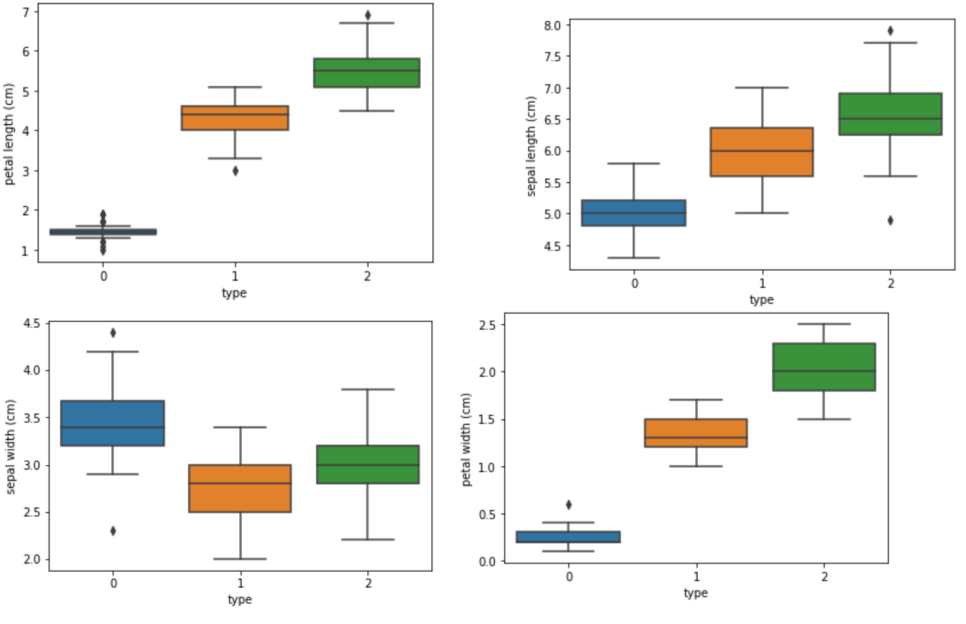

Figure 2: Box Plots

GaussianNB(priors=None, var_smoothing=1e-09)

	precision	recall	f1-score	support
0	1.00	1.00	1.00	9
1	1.00	0.83	0.91	12
2	0.82	1.00	0.90	9
accuracy			0.93	30
macro avg	0.94	0.94	0.94	30
weighted avg	0.95	0.93	0.93	30

Figure 3: Naive-bayes report

	precision	recall	f1-score	support
0	1.00	1.00	1.00	9
1	1.00	0.92	0.96	12
2	0.90	1.00	0.95	9
accuracy			0.97	30
macro avg	0.97	0.97	0.97	30
weighted avg	0.97	0.97	0.97	30

Figure 4: Logistic Regression Report