Data Analytics with the Fisher Iris Dataset

Ankit Bhadu
2018CSB1073
Indian Institute of Technology
Ropar

Abstract

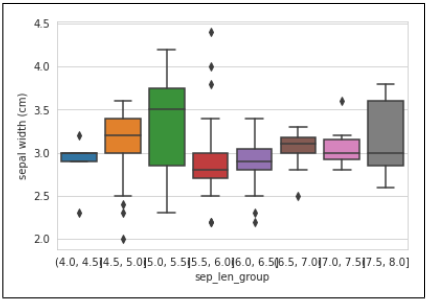
This document is based on analysis and application of various Classification techniques on the Fisher Iris Dataset. The Iris flower data set or Fisher's Iris data set is a multivariate data set which is a very famous and widely used Dataset for machine learning and statistics.The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). The primary objective is to predict the species using classification techniques

1 Introduction

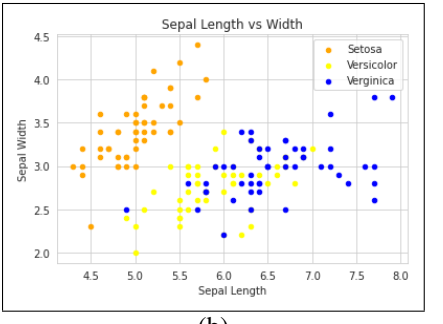
We will be analysing two types of classifications algorithms and a clustering algorithm:
1) Logistic Regression
2) Naïve Bayes
3) K Means Clustering

1.1 Exploratory Data Analysis

1) **Cleaning and Pruning :-**
First step was to check if any data point is empty or out of the ordinary. It was found that all data points were complete. Since the data set is already so small no pruning was required. Standardization of data reduced accuracy and hence was not done.
2) **Visualization:-**
(i) **Sepal Width and Sepal Length**
Here firstly we divide sepal length into 8 bins and see distribution of sepal width over the bins. It tells us how sepal width and sepal length vary with each other. Another way to visualize is to make a scatter plot between the two wherein we see how it is easy to classify on the basis of sepal length as compared to sepal width as it is all mixed up.Althpugh both sepal width and sepal length are not good at distinguishing species.Now we look at petal length and width.
(ii) **Petal Width and Petal Length**
It is easy to see that petal length and petal width are much more effective features to classify species as there is minimal overlap here between different species.
(iii) **Visualization through box plots**
We can re affirm our previous observation with box plots shown in Figure 3. There is minimal overlap between petal length and petal width of different species. However, sepal width and sepal length have a lot of overlapping inter species data points.
(iv) **Pair Plot**
Pair plots are a quick way to draw observations. Here we can see that by using petal length we can easily distinguish species 0 from the other two simply by setting a threshold.

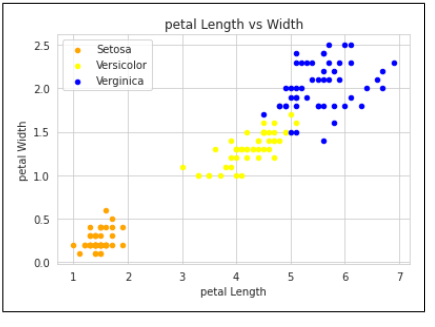


(a)

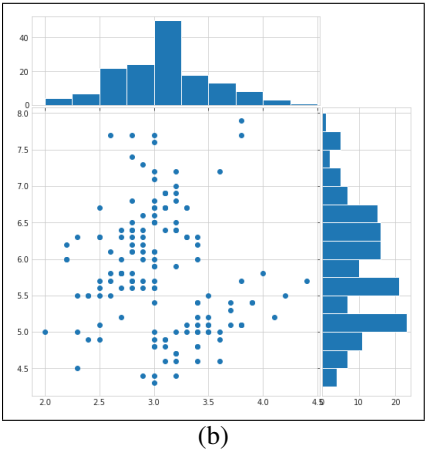


(b)

Figure 1: visualizing sepal width vs sepal length



(a)



(b)

Figure 2: visualizing petal width vs petal length and distribution of sepal width and length

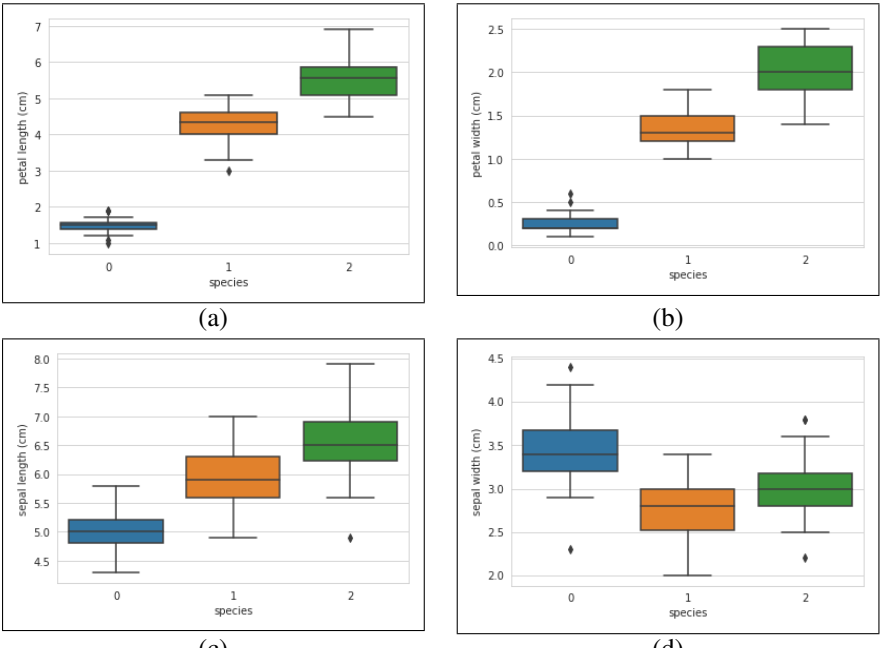


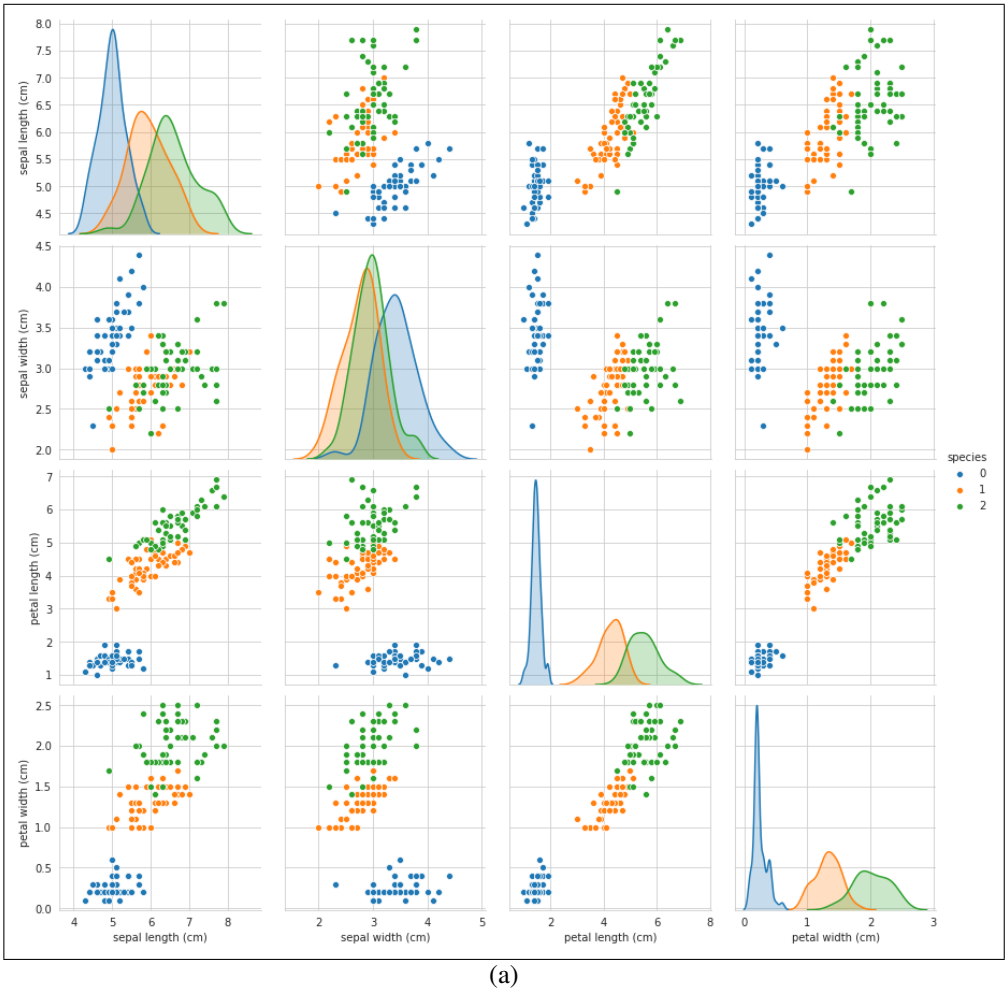
Figure 3: visualization through box plots for each feature

1.2 Logistic Regression

Logistic regression is not a regression algorithm but a probabilistic classification model. Classification in Machine Learning is a technique of learning, where an instance is mapped to one of many labels. The machine learns patterns from data in such a way that the learned representation successfully maps the original dimension to the suggested label/class without any intervention from a human expert.Logistic regression has a sigmoidal curve. Logistic regression is originally a binary classifier but we can adopt the one vs all strategy for multi-class classification. Here we consider one class as +ve and all other as -ve and do so for each class. Then we predict the label by choosing the label with highest probability as predicted by different iterations.

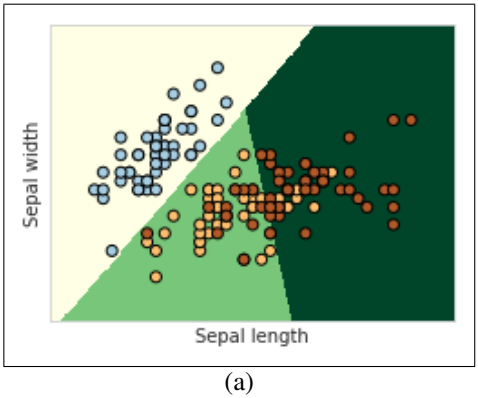
We have used the one vs all strategy and also 5-fold cross validation.Mean accuracy over the 5 iterations with different test and training sets was observed to be **96 percent**.

Here it is easy to observe why the misclassification of some data points happens. Since the decision boundary is linear as can be observed the attached figures, but the data points in actual cannot be separated by straight lines therefore the misclassification happens.As can be observed from the confusion matrix a total of 6 out of 150 were misclassified as they lie around the boundary separating species 1 and 2. However no data point of species 0 is misclassified.



(a)

Figure 4: Pair Plot



(a)

```
[[50  0  0]
 [ 0 47  3]
 [ 0  3 47]]
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	50
1	0.94	0.94	0.94	50
2	0.94	0.94	0.94	50
accuracy			0.96	150
macro avg	0.96	0.96	0.96	150
weighted avg	0.96	0.96	0.96	150

(b)

Figure 5: Logistic Regression Decision boundary graph and results

```
[[50  0  0]
 [ 0 47  3]
 [ 0  4 46]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        50
           1       0.92      0.94      0.93        50
           2       0.94      0.92      0.93        50

    accuracy                           0.95       150
   macro avg       0.95      0.95      0.95       150
weighted avg       0.95      0.95      0.95       150
```

Figure 6: Classification report for Naive Bayes Classifier

1.3 Naive Bayes Classifier

Unlike some classifiers, multi-class labeling is trivial with Naive Bayes. For each test example i, and each class k we find:

$$\arg\max_{\mathbf{k}} P(class_k|data_i) \tag{1}$$

In other words, we compute the probability of each class label in the usual way, then pick the class with the largest probability.
Here too we have used 5-fold cross validation using function cross val predict. Cross val predict makes iterations with different test and train sets each time and in this way finally it has predictions for the whole data as each data point was considered as a test point in some iteration.
As can be observed from confusion matrix a total of 7 out of 150 were misclassified and again they didn't belong to species 0. Hence accuracy was found to be **95 percent**.

1.4 K-means Algorithm

K-means clustering is an unsupervised learning algorithm which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest centroid. The algorithm aims to minimize the squared Euclidean distances between the observation and the centroid of cluster to which it belongs.
Although an unsupervised machine learning technique, the clusters can be used as features in a supervised machine learning model. However we won't be doing that but simply using clusters as a mode of classification. Since we can dictate the amount of clusters, it can be easily used in classification where we divide data into clusters which can be equal to or more than the number of classes.
The strategy followed to get the actual labels from the cluster labels was as follows:-
We find how many of the a particular cluster's elements belong to a particular species. Then the species which has the maximum count is assigned to the clusters. Therefore more than one cluster can have the same species label.

```
KFold(n_splits=5, random_state=None, shuffle=False)
Number of clusters is 30
Inertia : 0.0
Homogeneity :       1.0
[[30]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        30

    accuracy                           1.00        30
   macro avg       1.00      1.00      1.00        30
weighted avg       1.00      1.00      1.00        30

Number of clusters is 30
Inertia : 0.0
Homogeneity :       1.0000000000000007
[[20  0]
 [ 0 10]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        20
           1       1.00      1.00      1.00        10

    accuracy                           1.00        30
   macro avg       1.00      1.00      1.00        30
weighted avg       1.00      1.00      1.00        30

Number of clusters is 30
Inertia : 0.0
Homogeneity :       1.0
[[30]]
              precision    recall  f1-score   support
```

(a)

```
[[30]]
              precision    recall  f1-score   support

           1       1.00      1.00      1.00        30

    accuracy                           1.00        30
   macro avg       1.00      1.00      1.00        30
weighted avg       1.00      1.00      1.00        30

Number of clusters is 30
Inertia : 0.0
Homogeneity :       1.0000000000000004
[[10  0]
 [ 0 20]]
              precision    recall  f1-score   support

           1       1.00      1.00      1.00        10
           2       1.00      1.00      1.00        20

    accuracy                           1.00        30
   macro avg       1.00      1.00      1.00        30
weighted avg       1.00      1.00      1.00        30

Number of clusters is 30
Inertia : 0.0
Homogeneity :       1.0
[[30]]
              precision    recall  f1-score   support

           2       1.00      1.00      1.00        30

    accuracy                           1.00        30
   macro avg       1.00      1.00      1.00        30
weighted avg       1.00      1.00      1.00        30
```

(b)

Figure 7: Metrics for various combinations of test and training sets along with confusion matrix and classification report

Firstly size of clusters was kept equal to no of target classes and then gradually increased to see whether accuracy increases. Indeed the accuracy increased as we increased the no of clusters although the Inertia decreases. Homogeneity also increases as no of clusters are increased.
Then we perform classification on test set using 5 fold cross validation. Accuracy is found to be 100 percent when clusters are 30 and 99.4 percent when no of clusters are kept at 3.
Since the accuracy remains 100 percent across the 5 iterations of 5 fold cross validation one cannot say model is overfitting. More no of clusters increases accuracy because a particular species itself can more than one type of characteristic combinations of petal and sepal length and widths.

1.5 Principal Component Analysis

Principal Component Analysis(PCA) helps to reduce dimensionality and thus enables better visualization of data. Each of the principal components is chosen in such a way so that it would describe most of the still available variance and all these principal components are orthogonal to each other. Here we had earlier observed how species 1 and 2 were hard to distinguish by looking at individual scatter plots relating to the features one by one. Now we can look at all features at once using PCA. Attached figure based on PCA proves how to the misclassified case must belong to species 1 and 2 as they are mixed up and it is hard to draw a line separating them.

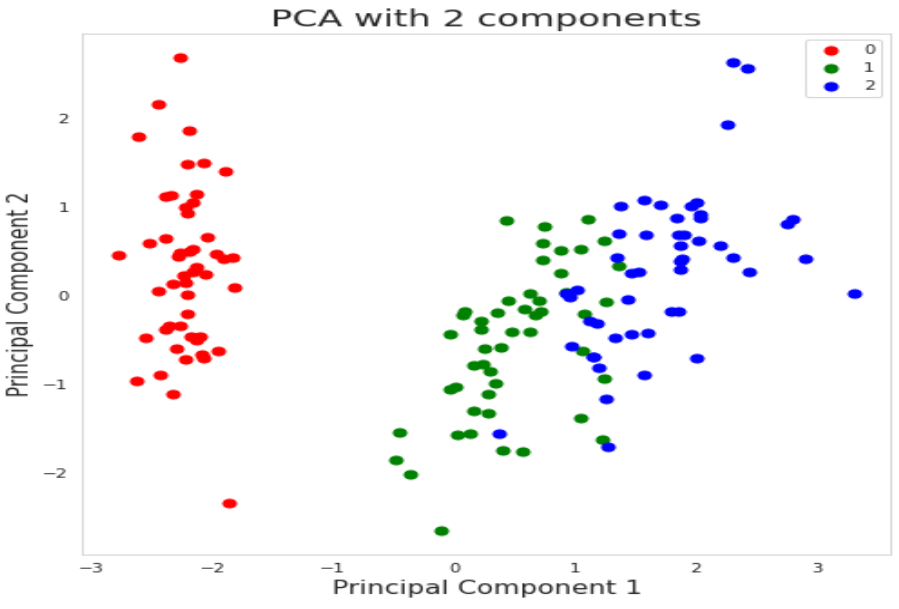


Figure 8: Visualization using PCA

1.6 Conclusion

To conclude the above findings it can be said that looking at the accuracy we can say k-means algorithm outperformed others even though it is basically a clustering algorithm and not a classifier. However, we need to test it on a bigger dataset to say this with confidence as the currently used dataset is quite small. One of the species was easily distinguishable due its starkingly different features as compared to the other two which had some overlaps hence decreasing classifier's accuracy. K Fold cross validation was highly beneficial as otherwise we could have had all test data from one species which is easily identifiable and would have had 100 percent accuracy in each model which would have been misleading.