Classifiers

Report 3

Discriminant Analysis

Author:
Piotr Pawełko
Piotr Wojsa

Laboratory date:
06.05.2022

1. Visualization of datasets

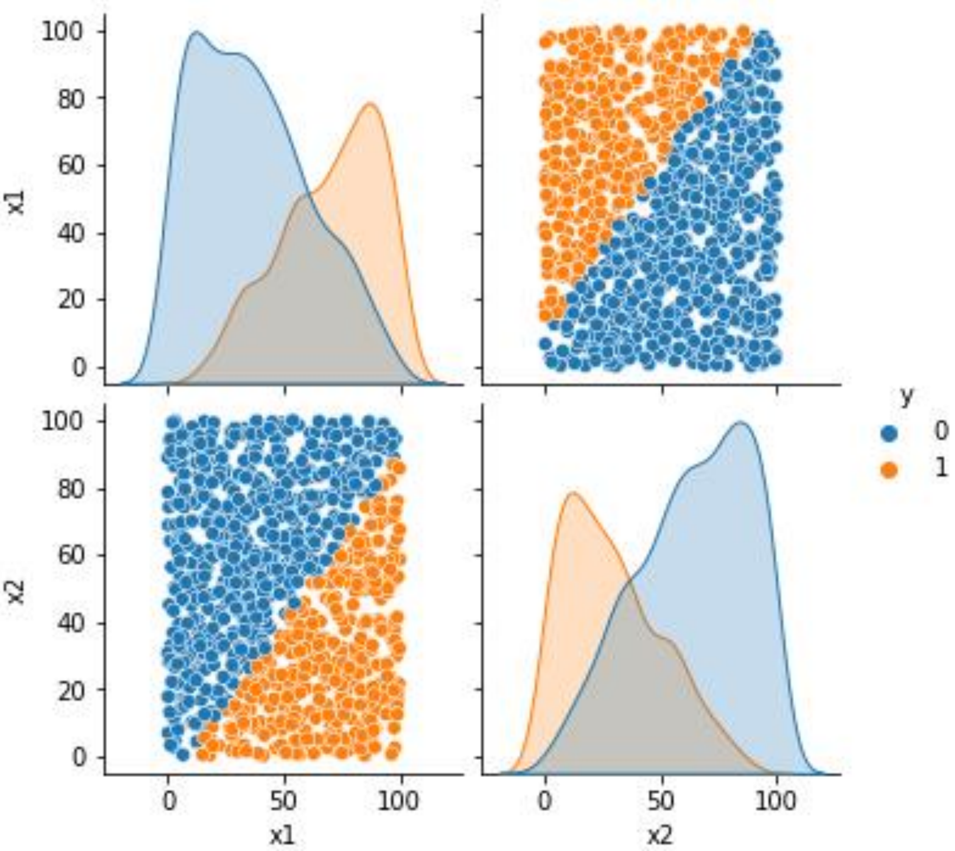

Figure 1 Pairplot of dataset4a

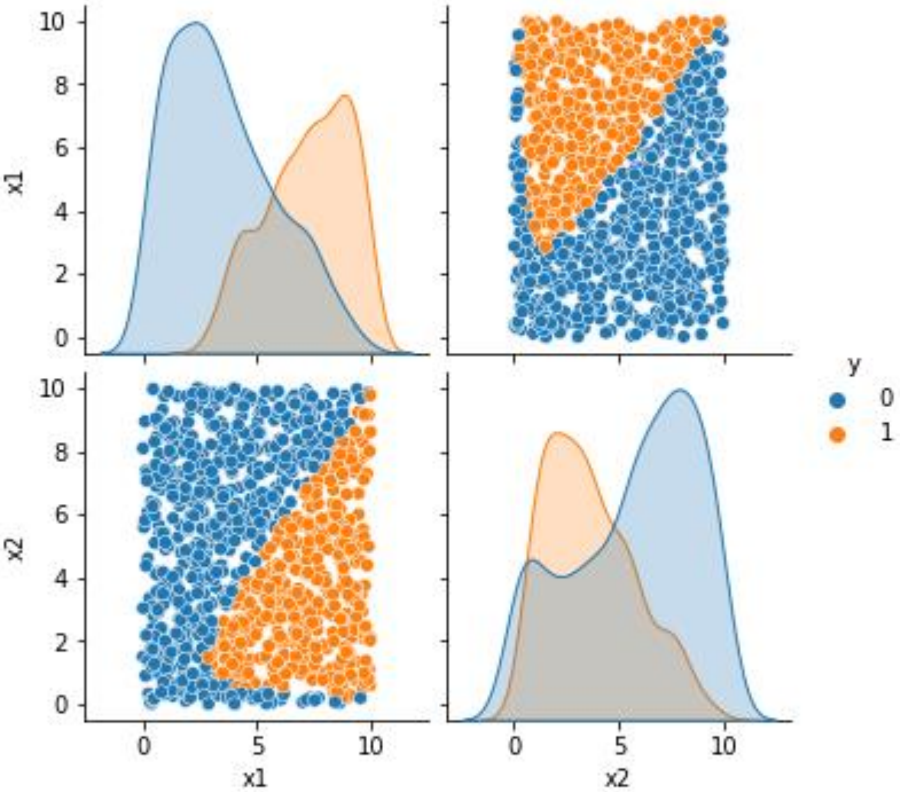

Figure 2 Pairplot of dataset4b

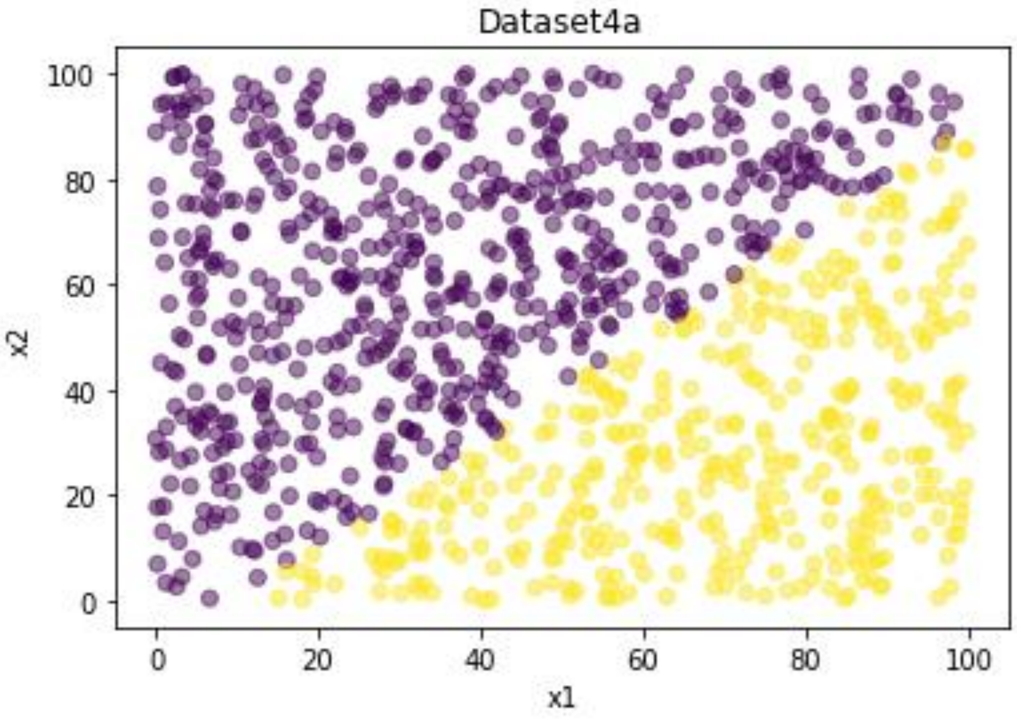

Figure 3 Scatterplot of data4a

0 – purple color

1 – yellow color

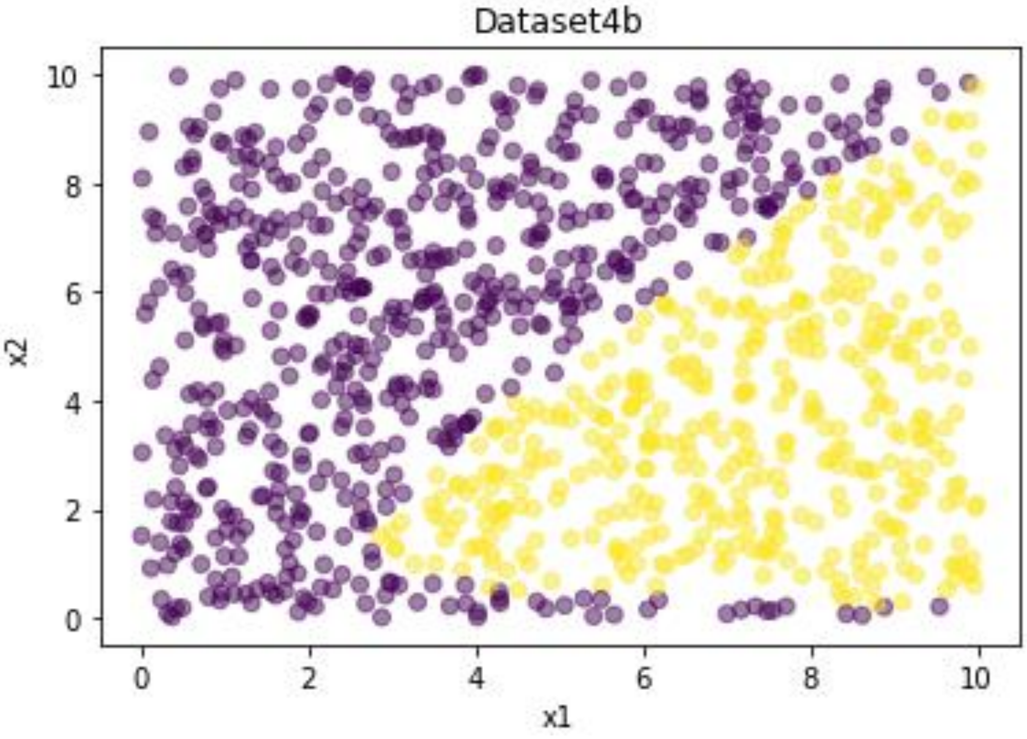

Figure 4 Scatterplot of data4b

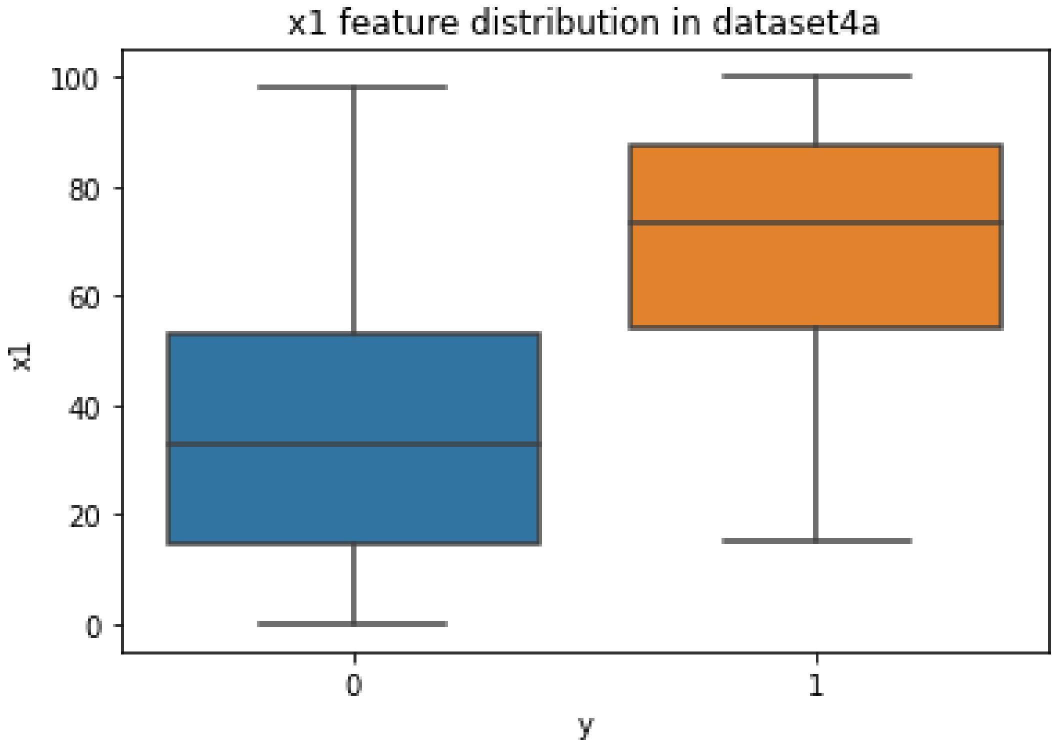

Figure 5 Boxplot in dataset4a for x1 feature

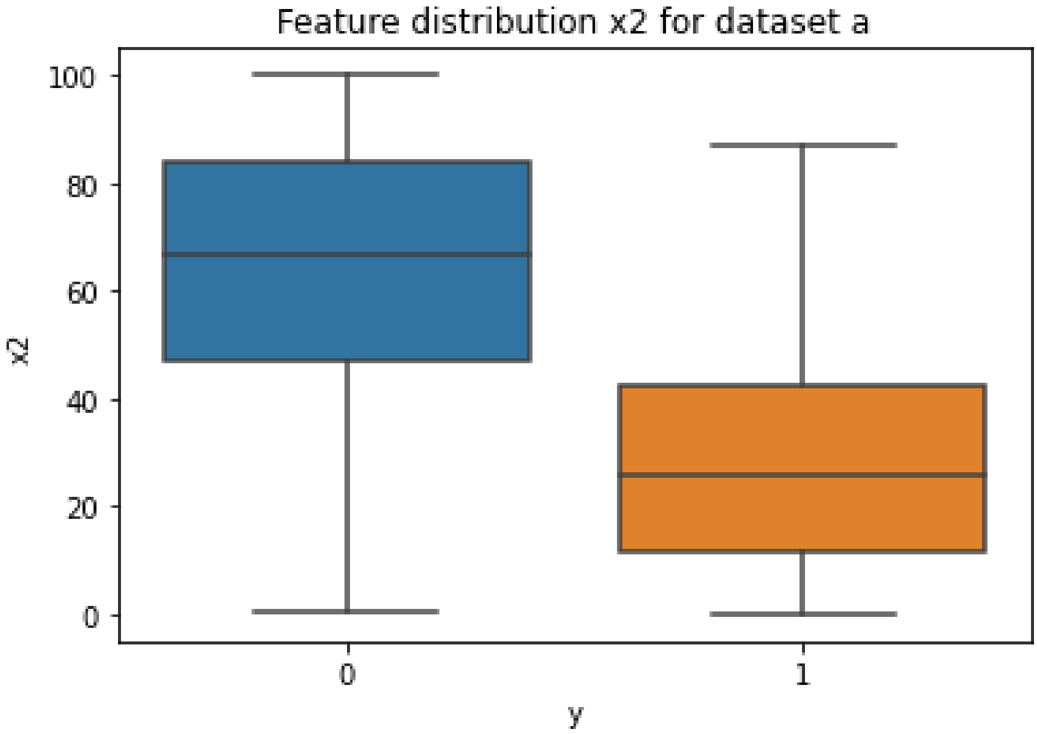

Figure 6 Boxplot in dataset4a for x2 feature

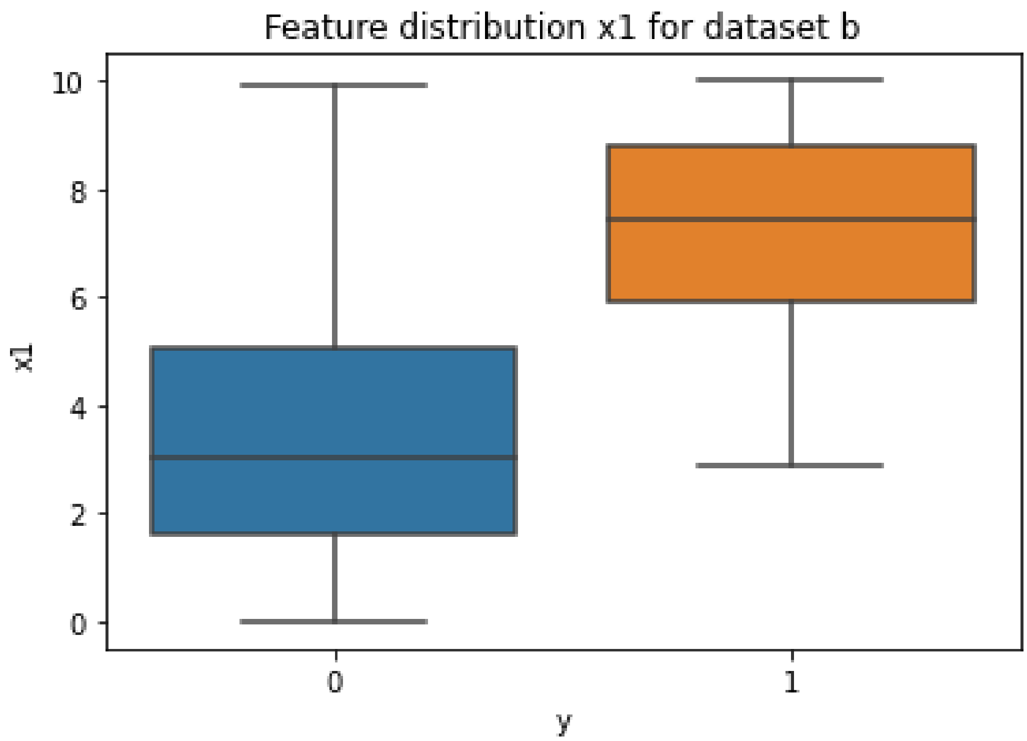

Figure 7 Boxplot in dataset4b for x1 feature

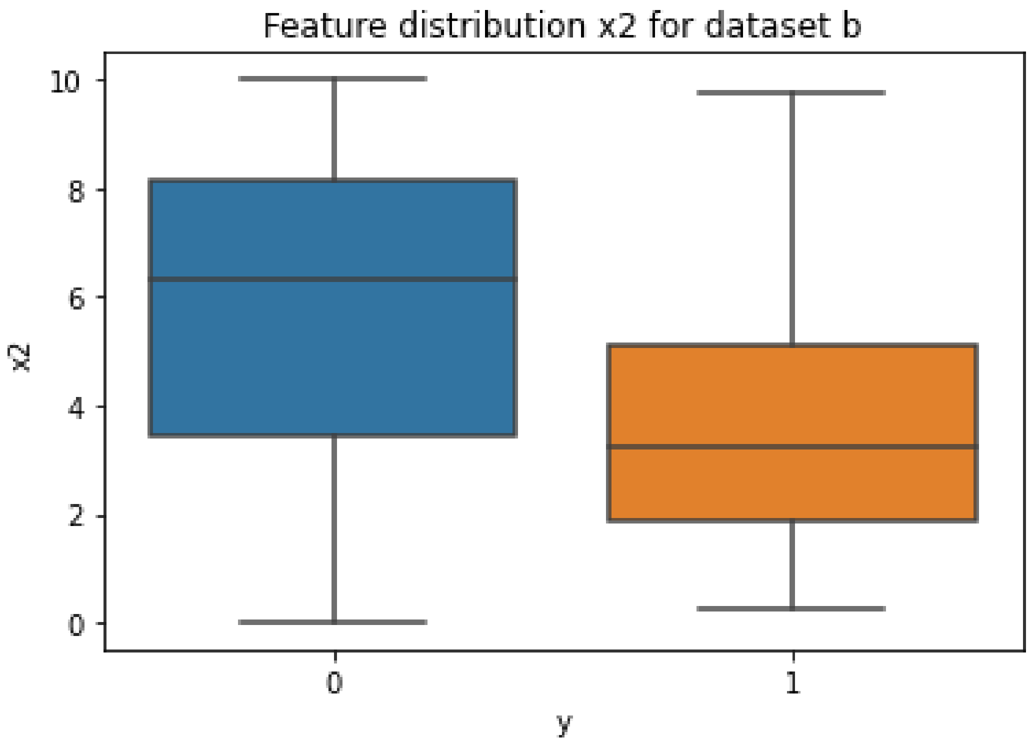

Figure 8 Boxplot in dataset4b for x2 feature

Dataset4a is linearly separable unlike the dataset4b. On scatterplot there is clearly visible linear line between outcome of 2 classes, dataset4b classes are more mixed specially on the bottom of plot. It is also visible on a boxplots. For both features in dataset4a boxes are not overlapping each other, unlike in dataset4b where feature x2 is overlapping in both classes.

2. Classification

2.1 LDA

Confusion matrix for dataset4a:

	Positive	Negative
Positive	122	0
Negative	0	178

Accuracy = 100%

Confusion matrix for dataset4a:

	Positive	Negative
Positive	102	17
Negative	22	159

Accuracy = 87%

2.2 QDA

Confusion matrix for dataset4a:

	Positive	Negative
Positive	118	0
Negative	4	178

Accuracy = 98.67%

Confusion matrix for dataset4b:

	Positive	Negative
Positive	101	10
Negative	23	166

Accuracy = 89%

2.3 SVM – gaussian kernel

Confusion matrix for dataset4a:

	Positive	Negative
Positive	122	0
Negative	0	178

Accuracy = 100%

Confusion matrix for dataset4b:

	Positive	Negative
Positive	117	5
Negative	7	171

Accuracy = 96%

2.4 SVM – linear kernel

Confusion matrix for dataset4a:

	Positive	Negative
Positive	122	0
Negative	0	178

Accuracy = 100%

Confusion matrix for dataset4b:

	Positive	**Negative**
Positive	112	18
Negative	12	158

Accuracy = 90%

As it was expected the best result with the biggest accuracy we got for the dataset4a, because data there is linearly separable, what makes classification more accurate. Overall best classification for linearly and non-linearly separable dataset was accomplished by SVM method with gaussian radial basis function(rbf) kernel, the accuracy for dataset4a was 100%, and 96% for dataset4b. LDA method did good in linearly separable dataset where it got 100% accuracy, but it was the worst for the second dataset, where it got the accuracy of 87%. The QDA method was the only one of all methods that did not score 100% accuracy in linearly separable dataset, but the result was also on high level of 97.67%, for "b" dataset it got 89% with the biggest number of false positives from all methods. Using linear kelner in SVM method did not improved scores for the non-linearly separable dataset, the number of false positives and false negatives increased, and the overall score has decreased to 90%.

3. Code for training(just for SVM, because for other methods it was very similar)

```
## SVM data b kernel - linear
from sklearn import svm

b_SVM_linear = svm.SVC(probability=True, kernel='linear')
b_SVM_linear.fit(b_train, b_train)
predict_b_SVM_linear = SVM_b.predict(b_test)
proba_b_SVM_linear = b_SVM_linear.predict_proba(b_test)

TN_b, FP_b, FN_b, TP_b = confusion_matrix(b_test,predict_b_SVM_linear).ravel()

Accuracy_sb = (TN_b+TP_b)/(TN_b+FP_b+FN_b+TP_b)*100
```