1 LDA

1.1 Data Pre-Processing & Visualisation

WHAT DOES data.csv HAVE ?

1. It has three columns x, y and label (2000 X 3)
2. Both x and y are float values while label is 0 or 1
3. No missing data

OUTLIERS

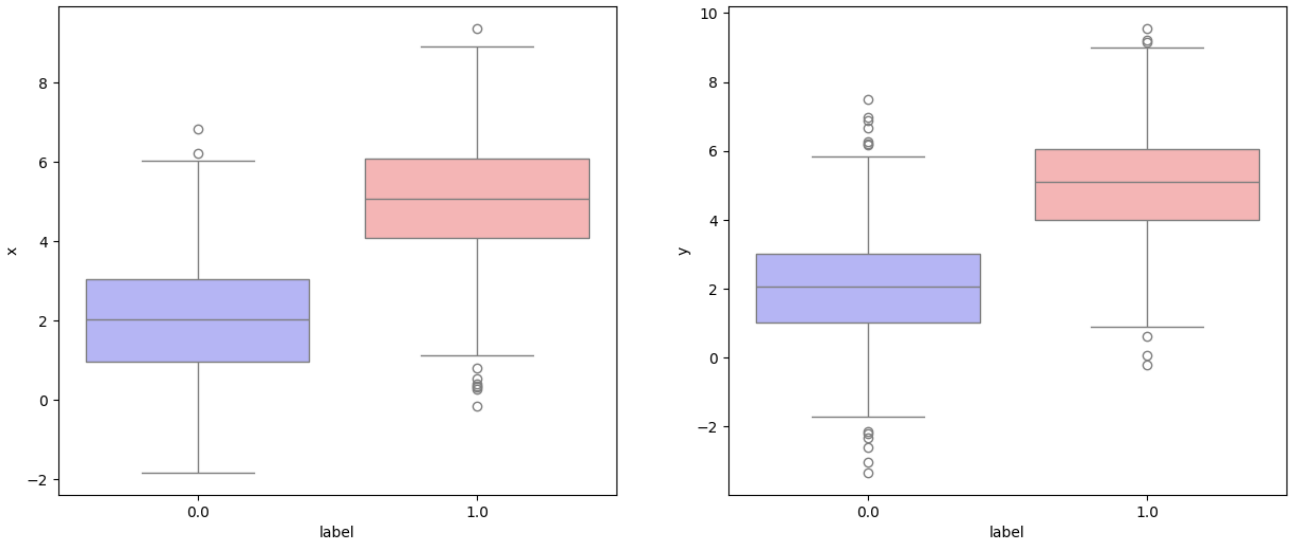


Figure 1: outliers

VISUALIZATION OF DATA

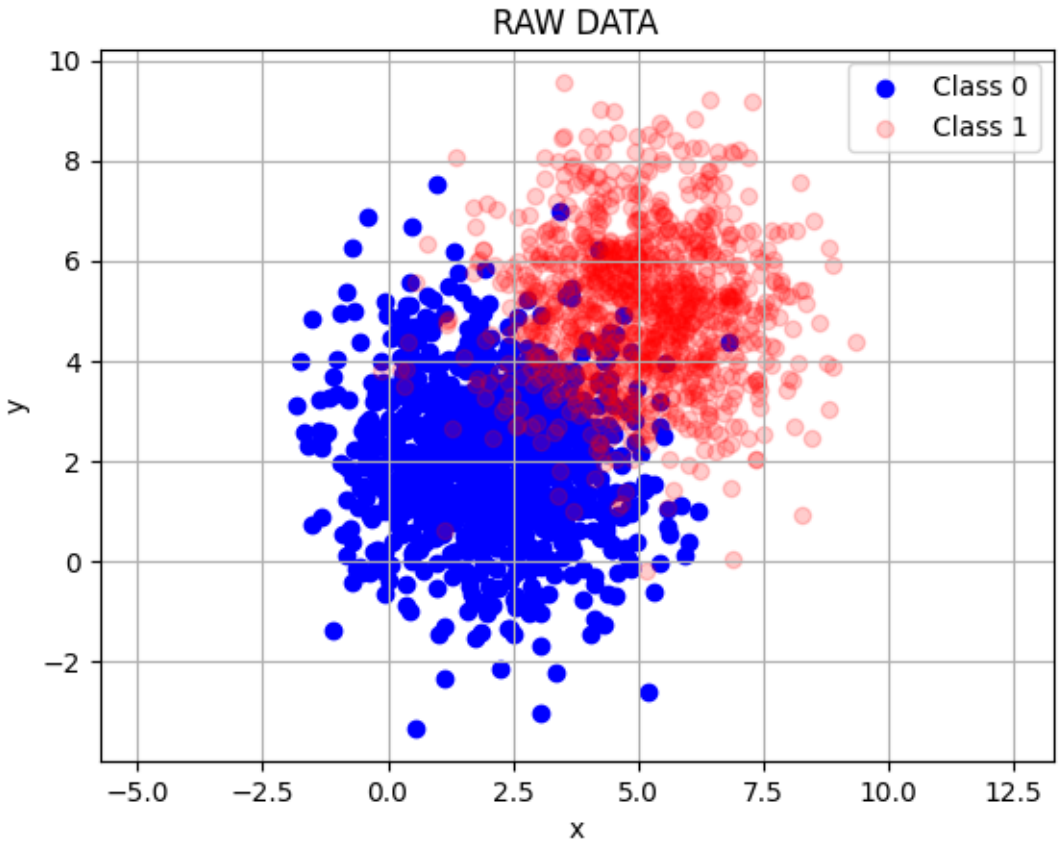


Figure 2: plotting data

1.2 Linear Discriminant Analysis

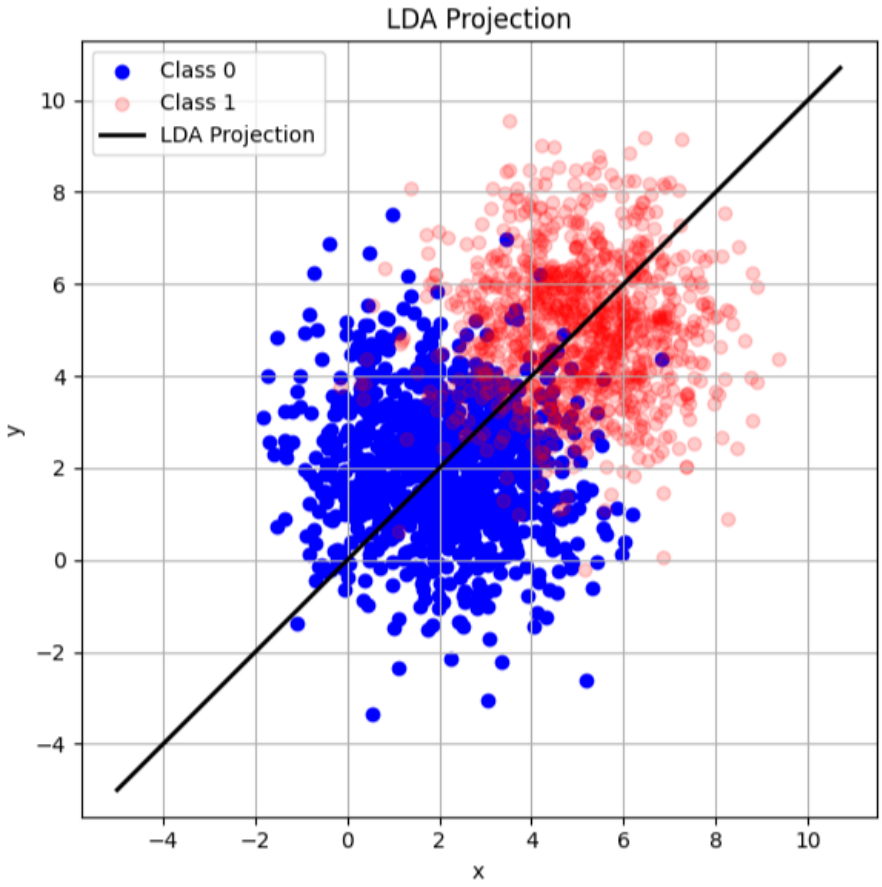


Figure 3: vector on raw data

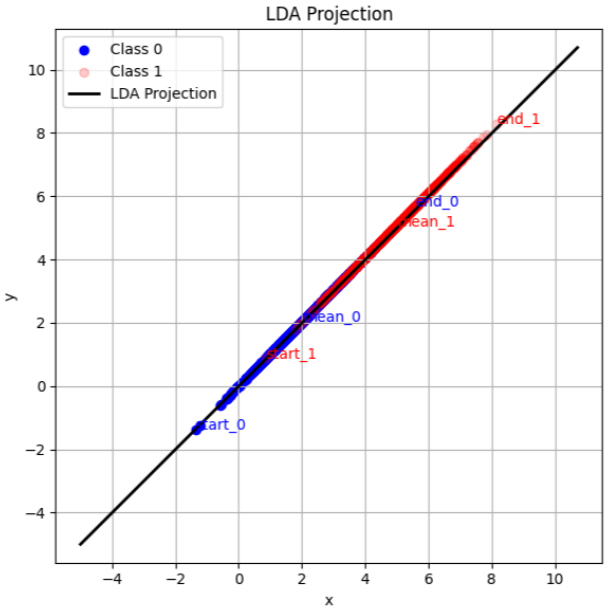


Figure 4: projection

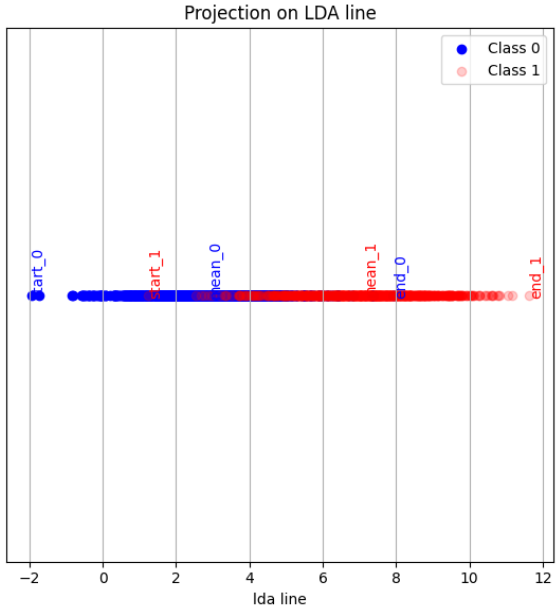


Figure 5: reorientation

1.3 performance of 1-NN neighbor classifier on original data vs projected data

BEFORE LDA :

1. CLASSIFICATION REPORT

label	precision	recall	f1-score	support
0	0.89	0.88	0.89	199
1	0.89	0.89	0.89	201

2. CONFUSION MATRIX

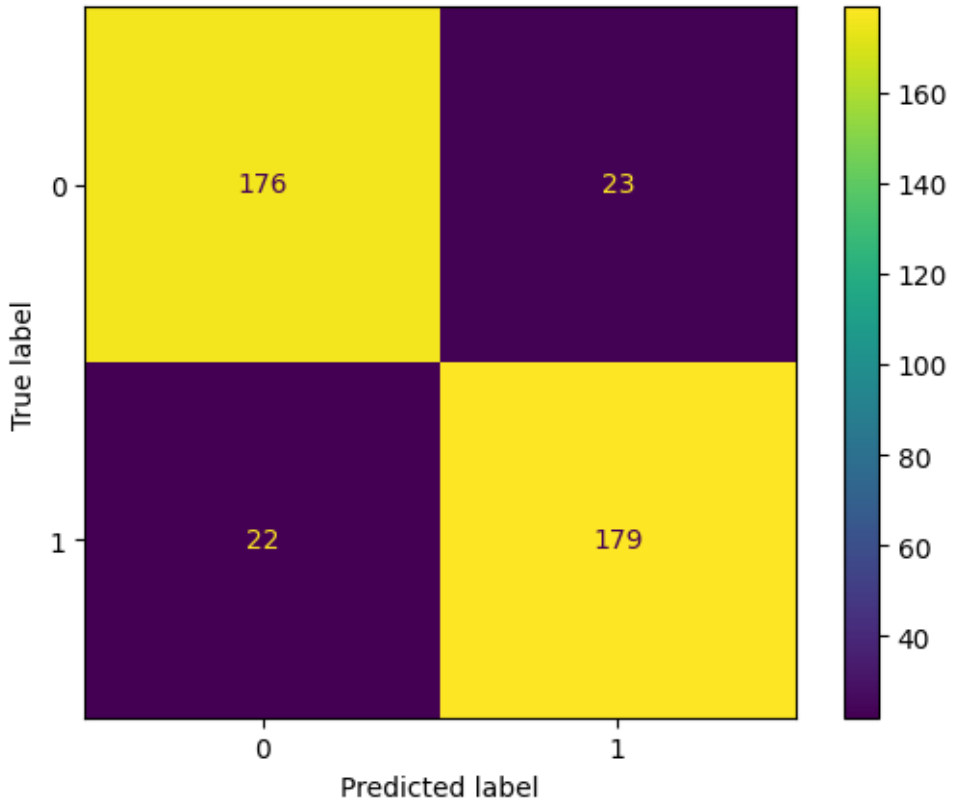


Figure 6: confusion matrix

AFTER LDA :

1. CLASSIFICATION REPORT

label	precision	recall	f1-score	support
0	0.88	0.88	0.88	199
1	0.88	0.88	0.88	201

2. CONFUSION MATRIX

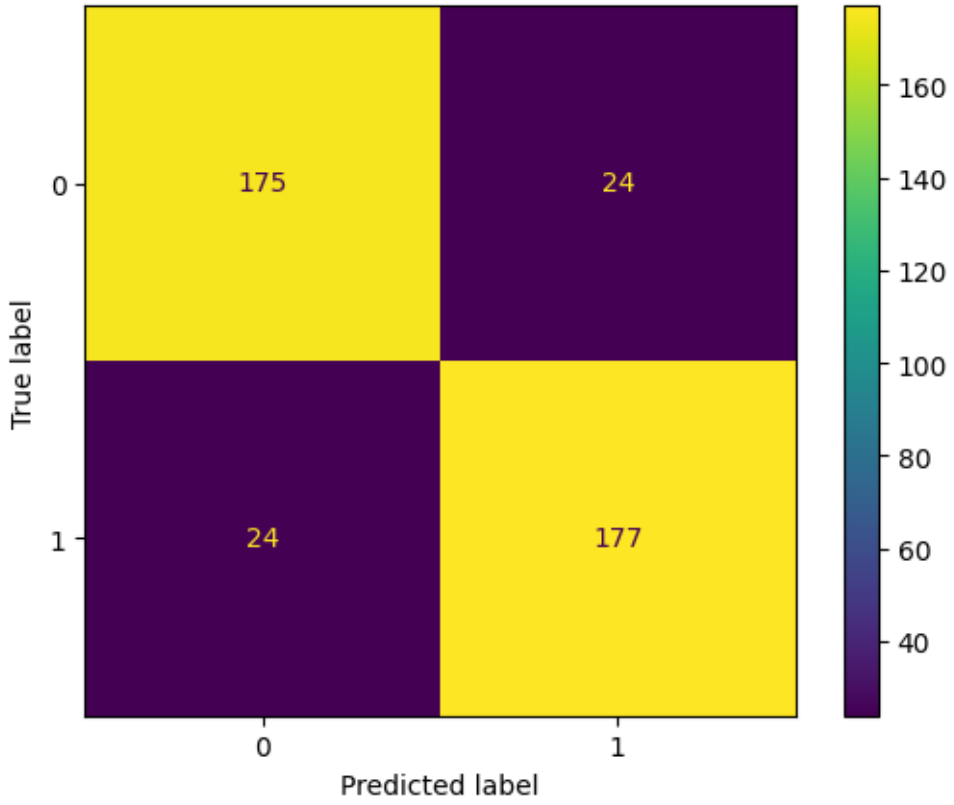


Figure 7: confusion matrix

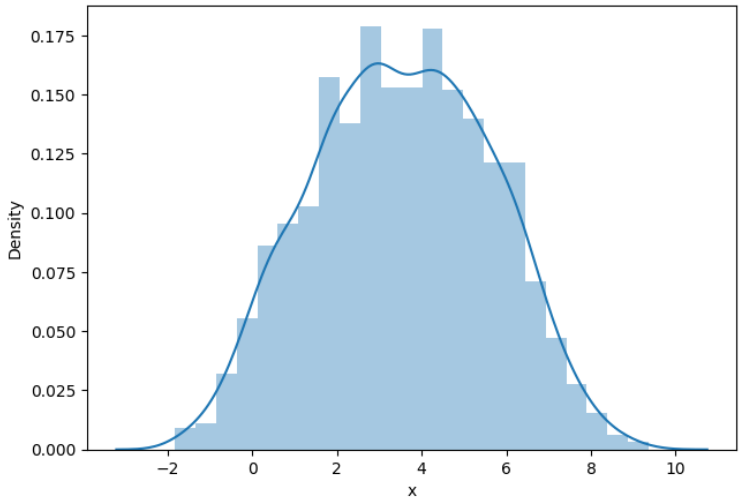
*** The accuracy is almost same for original and projected data**

Reasons for misclassifications :

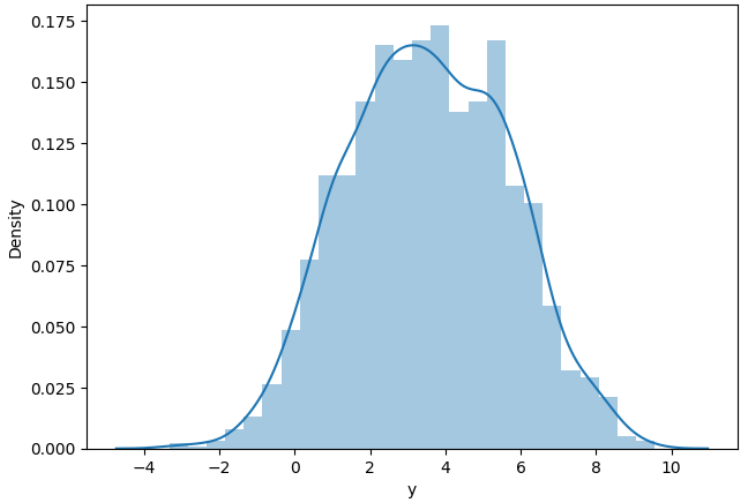
1. As 1NN is very sensitive to outliers, which here they are present so it is possible for wrong prediction
2. Projecting data on 1D always invloves in loss of information
3. We are forcefully assuming that the decision boundary is LINEAR
4. Overlapping between boundary, ie mixing of classes

1.4 REMOVING OUTLIERS

X PLOT



Y PLOT



VISUALIZATION OF DATA

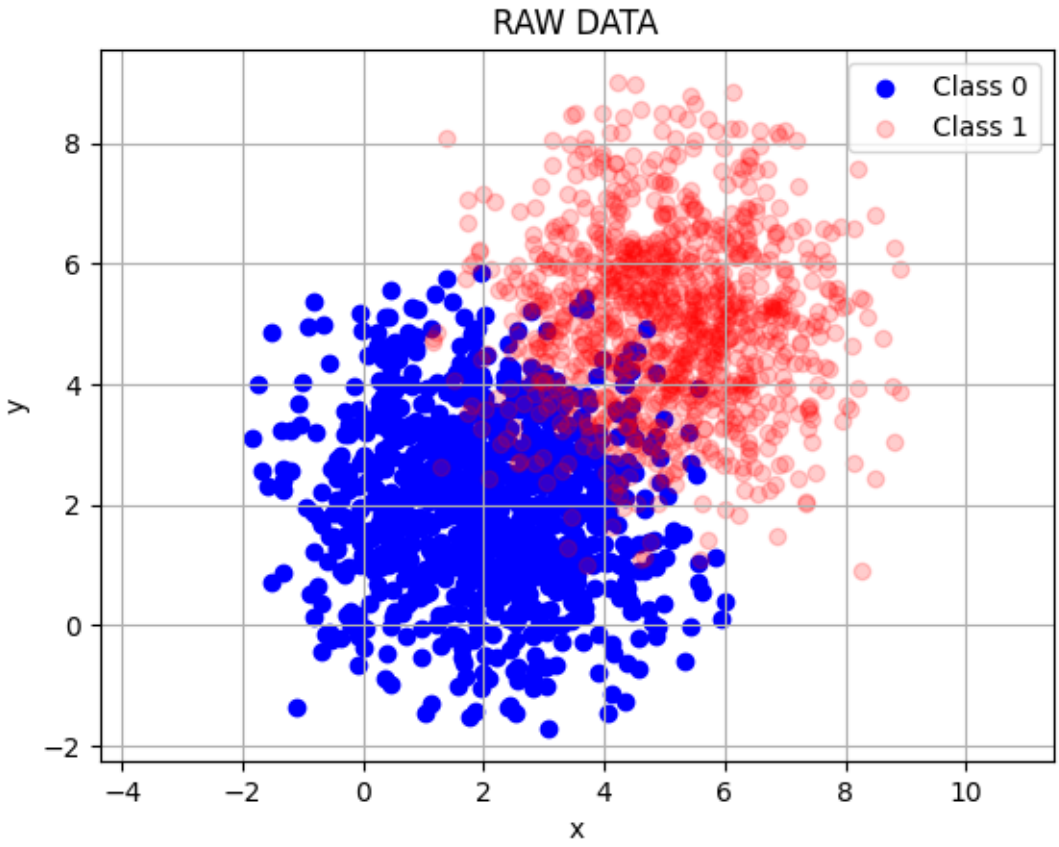


Figure 8: data after removing outliers

BEFORE LDA after removing outliers:

1. CLASSIFICATION REPORT

label	precision	recall	f1-score	support
0	0.89	0.87	0.88	184
1	0.89	0.91	0.90	211

2. CONFUSION MATRIX

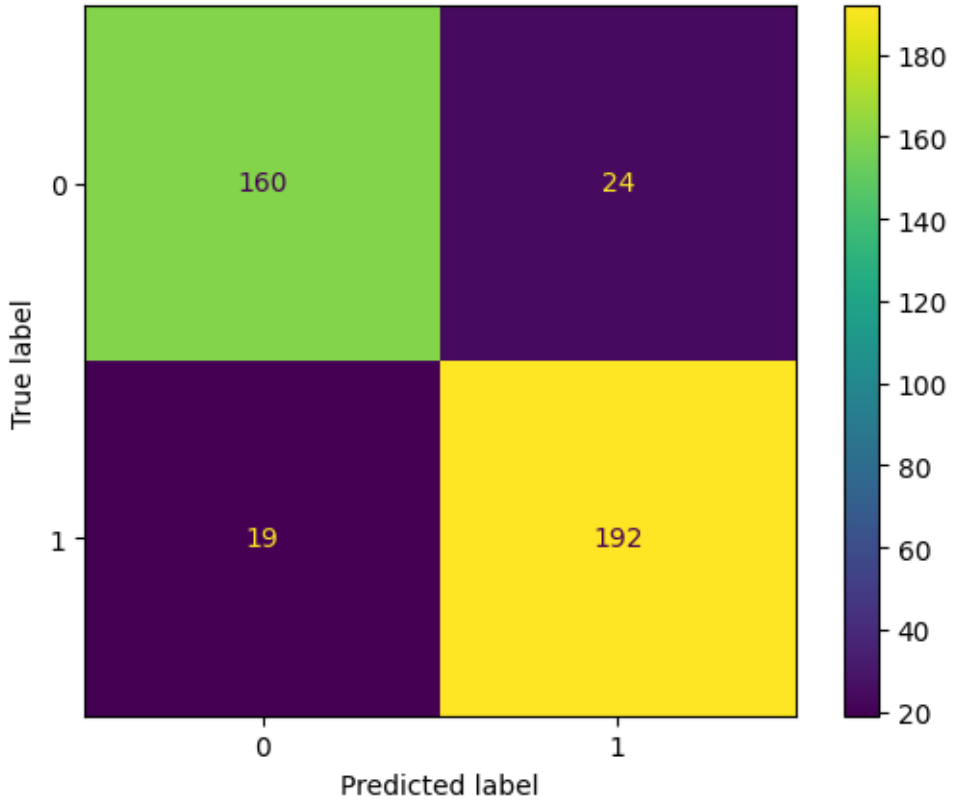


Figure 9: confusion matrix

AFTER LDA after removing outliers:

1. CLASSIFICATION REPORT

label	precision	recall	f1-score	support
0	0.93	0.89	0.91	184
1	0.91	0.94	0.92	211

2. CONFUSION MATRIX

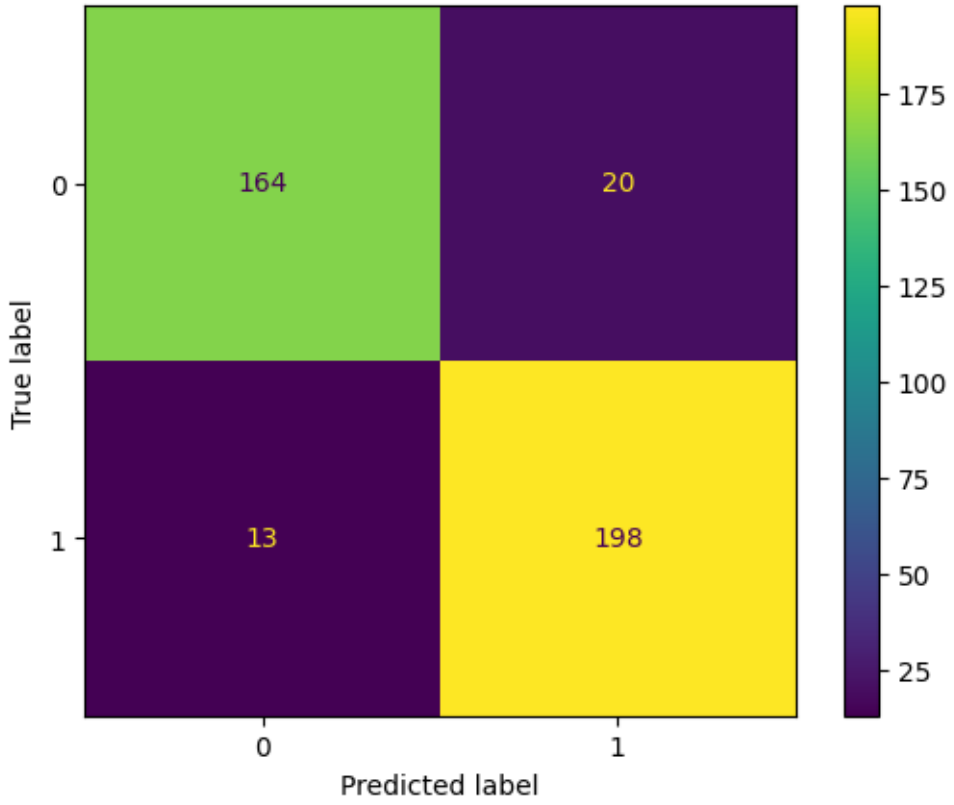


Figure 10: confusion matrix

*** As we can see the accuracy has indeed increased by removing outliers from 88 to 91.6**