ECGR 5105 Homework 3: Linear Classification

Lauren Bourque (Student ID: 801264851)

October 11, 2023

GitHub Link

Click here to view the code

Problem 1.

1. Diabetes Dataset Logistic Regression Results:

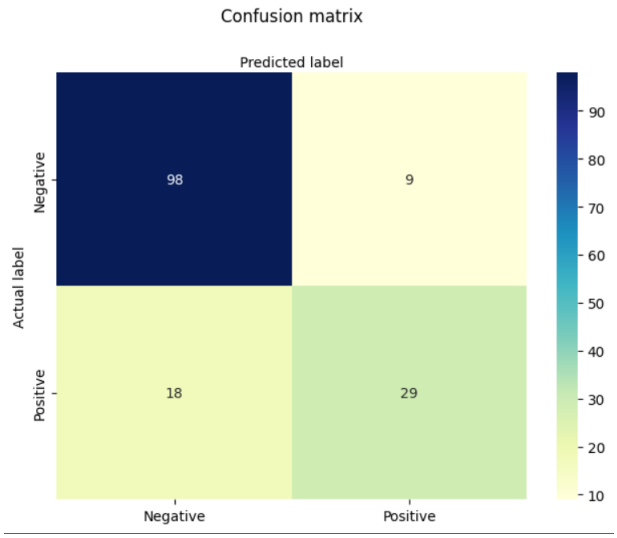
The result types and their respective values are listed in the table below:

Result Type	Value
Accuracy	0.8246753246753247
Precision	0.7631578947368421
Recall	0.6170212765957447
F1 Score	0.6823529411764706

2. Diabetes Dataset Confusion Matrix:

The confusion matrix containing the classes and their predicted values can be seen below:

Figure 1: Confusion Matrix for Diabetes Dataset



Problem 2a.

1. Cancer Dataset Logistic Regression Results:

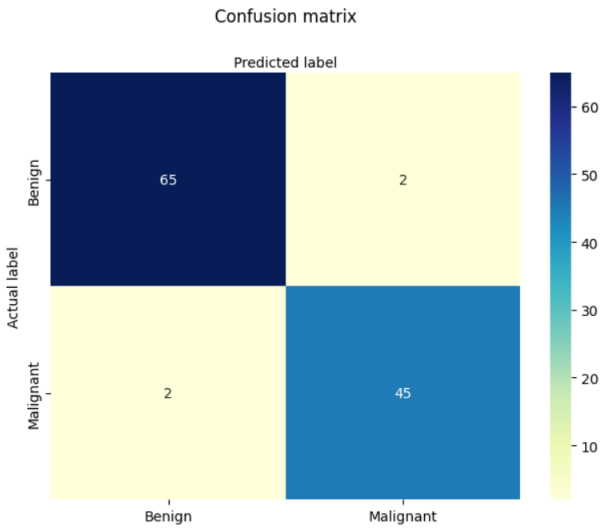
The result types and their respective values are listed in the table below:

Result Type	Value
Accuracy	0.9649122807017544
Precision	0.9574468085106383
Recall	0.9574468085106383
F1 Score	0.9574468085106385

2. Cancer Dataset Confusion Matrix:

The confusion matrix containing the classes and their predicted values can be seen below:

Figure 2: Confusion Matrix for Cancer Dataset



Problem 2b.

1. Testing Different Parameters Penalties:

The parameter values and the training and test accuracy results are listed in the table below:

Parameter Value	Training Accuracy	Test Accuracy
10	0.9890	0.9474
5	0.9890	0.9561
1	0.9912	0.9561
0.1	0.9758	0.9737
0.001	0.6374	0.5877

From these results, 0.1 was chosen as a parameter penalty because it had the highest test accuracy along with having a high training accuracy that was very similar to the training accuracies from the other penalties.

2. Cancer Dataset Logistic Regression with a Parameters Penalty Results:

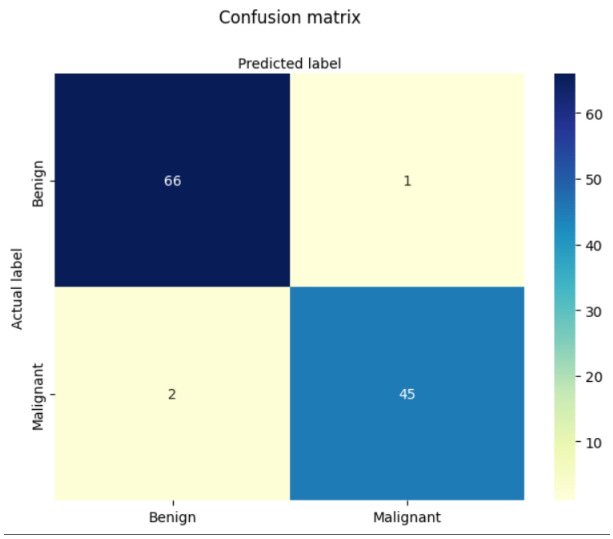
The result types and their respective values are listed in the table below:

Result Type	Value
Accuracy	0.9736842105263158
Precision	0.9782608695652174
Recall	0.9574468085106383
F1 Score	0.967741935483871

3. Cancer Dataset Confusion Matrix:

The confusion matrix containing the classes and their predicted values can be seen below:

Figure 3: Confusion Matrix for Cancer Dataset with Parameters Penalty



Problem 3.

1. Cancer Dataset Naive Bayesian Classifier Results:

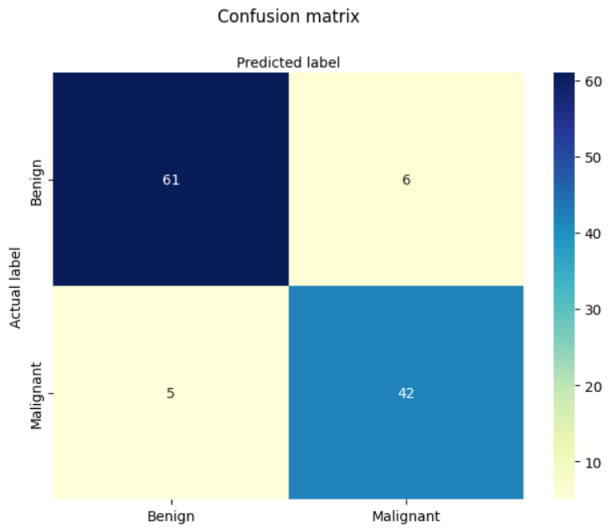
The result types and their respective values are listed in the table below:

Result Type	Value
Accuracy	0.9035087719298246
Precision	0.875
Recall	0.8936170212765957
F1 Score	0.8842105263157894

2. Cancer Dataset Naive Bayesian Confusion Matrix:

The confusion matrix containing the classes and their predicted values can be seen below:

Figure 4: Confusion Matrix for Cancer Dataset Using Naive Bayesian Classifier



The accuracy, precision, recall, and F1 score for the Naive Bayesian Classifier were lower than the values found using logistic regression. They were even lower than the values found using logistic regression with a parameters penalty.

Problem 4.

1. Cancer Dataset PCA Logistic Regression Results:

The result types and their respective values are listed in the table below:

Result Type	Value
Accuracy	0.9649122807017544
Precision	0.9574468085106383
Recall	0.9574468085106383
F1 Score	0.9574468085106385

The optimal K value was found to be 11 principal components. This achieved the highest accuracy, precision, recall, and F1 score among the different K values. K values from 1 to 15 were tested. These results are more accurate than the Naive Bayesian classifier without PCA, but less accurate than the logistic regression without PCA. This shows that PCA slightly decreases the accuracy of logistic regression.

Problem 5.

1. Cancer Dataset PCA Naive Bayes Classifier Results:

The result types and their respective values are listed in the table below:

Result Type	Value
Accuracy	0.9210526315789473
Precision	0.9318181818181818
Recall	0.8723404255319149
F1 Score	0.9010989010989012

The optimal K value was found to be 5 principal components. This achieved the highest accuracy, precision, recall, and F1 score among the different K values. K values from 1 to 15 were tested. These results are more accurate than the Naive Bayesian classifier without PCA, but less accurate than the logistic regression without PCA and with PCA. This shows that PCA slightly increase the accuracy of Naive Bayesian Classifier.