Johnny Nguyen

Student ID: 801119047

Homework #3

GitHub Link: https://github.com/Norumai01/Intro_Machine_Learning/tree/main/HW_3

Problem 1:

- Standardization scaling was uses for the following breast cancer dataset for optimized training model. Comparing the precision, recall and accuracy of the Naive Bayes' and logistic regression's training model, they have a 2-3 percent difference. However, they are generally the same range of accuracy so does not make any difference to use either model.

Problem 2:

- Using the PCA extraction, the dataset's features are reduced, and standardization scaling was used for the logistic regression's training model. Due to variety of independent features and the scales of values, the accuracy seems to vary up and down as shown in Naïve Bayes' training model. However, for the logistic regression, the results seem to stay consistent.
- K = 10 (left picture) and K = 5 (Right Picture):

```
Confusion Matrix:
 [[44  3]
 [ 2 65]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.96      0.94      0.95        47
         1.0       0.96      0.97      0.96        67

    accuracy                           0.96       114
   macro avg       0.96      0.95      0.95       114
weighted avg       0.96      0.96      0.96       114
```

```
Confusion Matrix:
 [[43  4]
 [ 2 65]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.96      0.91      0.93        47
         1.0       0.94      0.97      0.96        67

    accuracy                           0.95       114
   macro avg       0.95      0.94      0.95       114
weighted avg       0.95      0.95      0.95       114
```

Problem 3:

- Using the PCA extraction, the dataset's features are reduced, and standardization scaling was used for the Naïve Bayes' training model. The classification shows varied percentage of the accuracy, precision and recall at different numbers of components. For example, when the number of components is reduced to 2, the result significantly increases compared to 10 components.
- K = 10 (left picture) and K = 2 (right picture):

Confusion Matrix:
```
 [[42  5]
 [10 57]]
```

Classification Report:

	precision	recall	f1-score	support
0.0	0.81	0.89	0.85	47
1.0	0.92	0.85	0.88	67
accuracy			0.87	114
macro avg	0.86	0.87	0.87	114
weighted avg	0.87	0.87	0.87	114

Confusion Matrix:
```
 [[39  8]
 [ 4 63]]
```

Classification Report:

	precision	recall	f1-score	support
0.0	0.91	0.83	0.87	47
1.0	0.89	0.94	0.91	67
accuracy			0.89	114
macro avg	0.90	0.89	0.89	114
weighted avg	0.90	0.89	0.89	114