Stanly Gomes
Student ID: 801118166

**Homework 2**

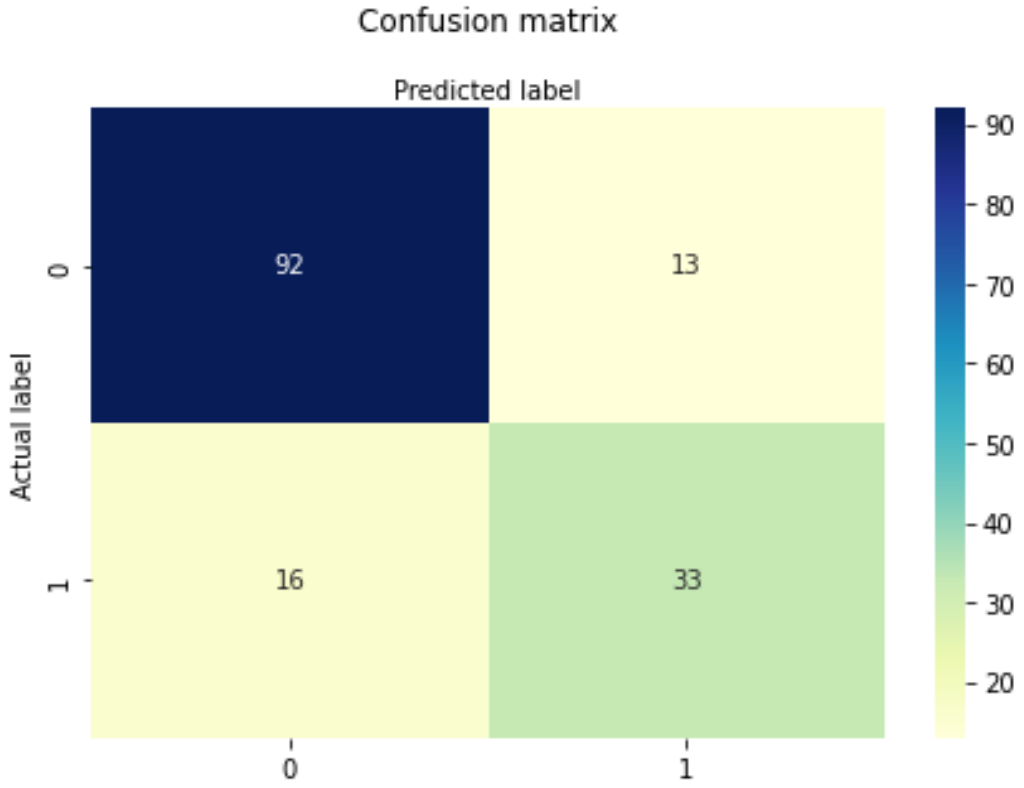
GitHub Repository Link: https://github.com/NaraPvP/IntroToML

**Problem 1:**

For feature scaling, standardization was used as it provided more balance between the inputs for this dataset. All calculations were done using the “metrics” library in “sklearn”, which provided the following values for the diabetes dataset using an 80/20 training split:

- Accuracy: 0.8116883116883117 = 81.17%
- Precision: 0.717391304347826 = 71.74%
- Recall: 0.673469387755102 = 67.35%

With the logistic regression function, the confusion matrix for the binary classifier of positive diabetes cases was generated as such:




**Problem 2:**

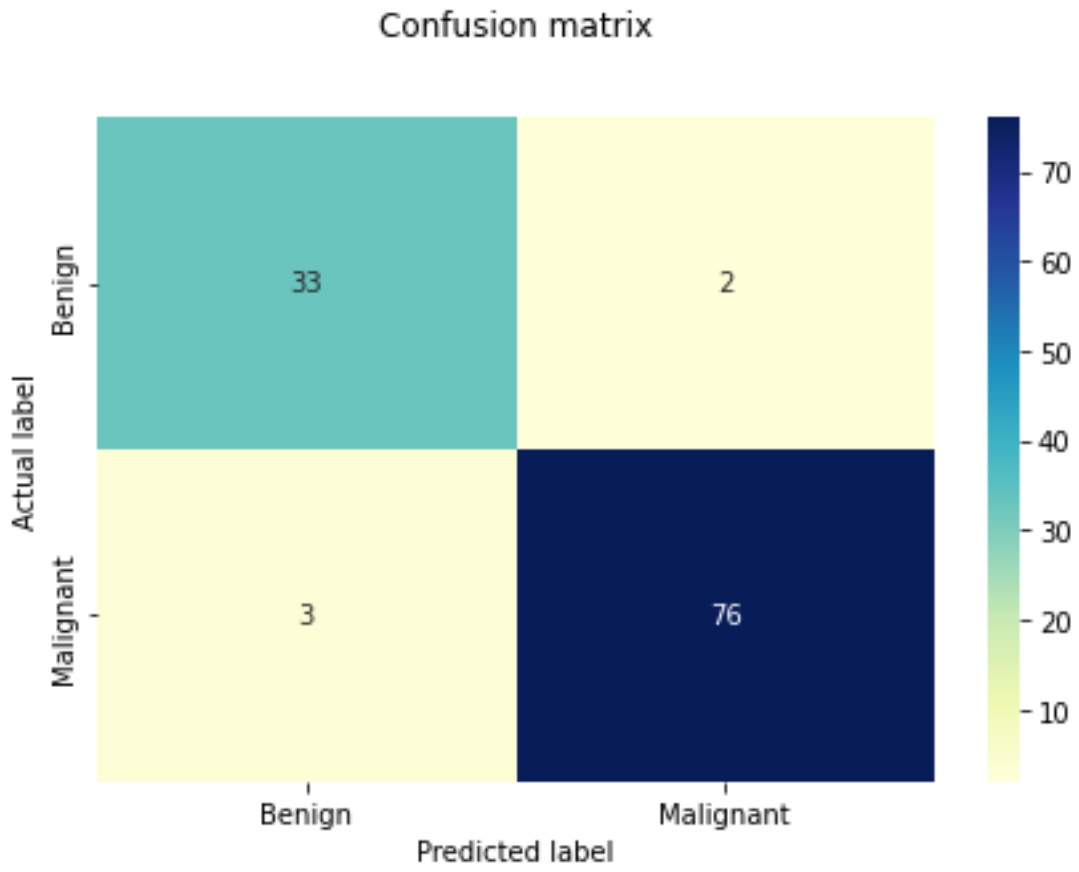
To generalize our model further, K-fold cross validation is used in place of a training split (along with standardization for feature scaling). This allows for a more diverse test set, which should provide a more accurate model for new data points. In this case though, the accuracy of the cross-validation method was lower than the training split as shown below:

- Average Accuracy of K-Fold validation with K = 5: 77.471% (2.893%)
- Average Accuracy of K-Fold validation with K = 10: 77.604% (5.036%)

Stanly Gomes
Student ID: 801118166

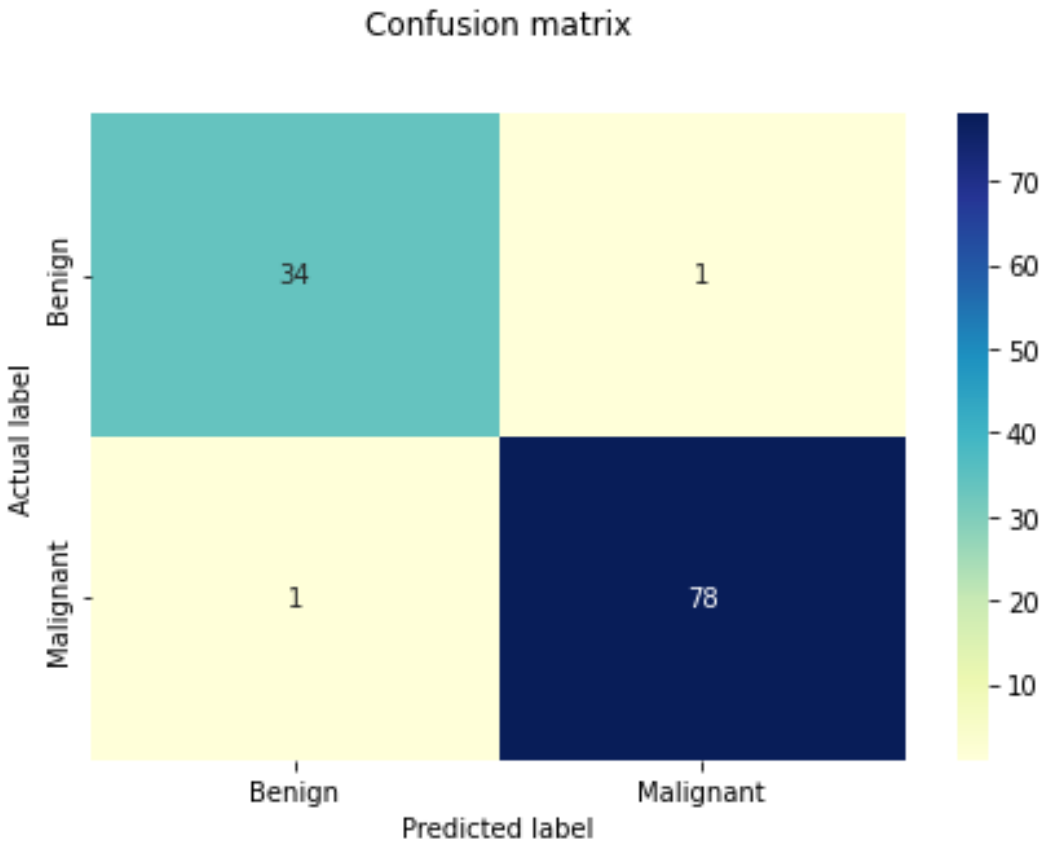
**Problem 3:**

1. Similarly to Problem 1, a 80/20 training split is done on the cancer dataset provided in the "sklearn" datasets. Standardization was used as the feature scaling for the same reasons as before. Using the functions imported from the "metrics" library, the results of the classifier are shown below:
    - Accuracy: 0.956140350877193 = 95.61%
    - Precision: 0.916666666666666 = 91.67%
    - Recall: 0.9302325581395349 = 93.02%




2. With weight penalties added, here are the results of the training performed in Part 1 of Problem 3:
    - Accuracy: 0.9824561403508771 = 98.25%
    - Precision: 0.9714285714285714 = 97.14%
    - Recall: 0.9714285714285714 = 97.14%

Stanly Gomes
Student ID: 801118166




**Problem 4:**

1. Problem 3 (Part 1) was repeated with K-Fold cross-validation instead of training split, which did yield better or consistent average accuracy than the split did. Here are the results of this classifier:
    - Average Accuracy of K-Fold validation with K = 5: 95.259% (2.115%)
    - Average Accuracy of K-Fold validation with K = 10: 95.611% (2.250%)
2. Similarly, Problem 3 (Part 2) was repeated with the K-Fold cross-validation to see how the average accuracy has changed in the weighed classifier. As expected, the average accuracy was the largest when this validation method was used with weighted parameters.
    - Average Accuracy of K-Fold validation with K = 5: 97.193% (1.701%)
    - Average Accuracy of K-Fold validation with K = 10: 97.895% (2.046%)