Claude Shyaka
ID#: 801326243
ECGR-5105

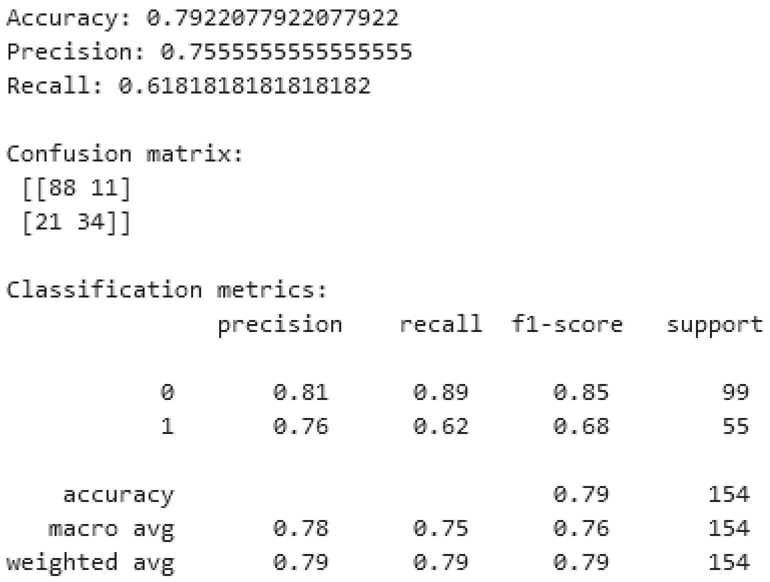
Homework 2: Logistic Regression and K-fold Cross Validation

GitHub Repo: https://github.com/claudeshyaka/ml

In this exercise, logistic regression binary classifiers were trained on the Diabetes dataset and Cancer dataset. Logistic regression models were trained using a train/test split of 80% and 20% respectively, then the models were retrained using the k-fold cross validation method. Results obtained in each scenario are discussed in this report.

1. In this section, a logistic regression binary classifier was trained on the Diabetes dataset. The dataset was split into 80% for training and 20% for testing. In addition, both train and test sets were scaled using the MinMaxScaler from Sklearn. The optimal results were obtained based on the following parameters:
   a. Penalty: ‘l2’
   b. C: 1.2
   c. Solver: ‘liblinear’
   d. Random_state: 42

   Results are shown below:



```
Accuracy: 0.7922077922077922
Precision: 0.7555555555555555
Recall: 0.6181818181818182

Confusion matrix:
 [[88 11]
 [21 34]]

Classification metrics:
              precision    recall  f1-score   support

           0       0.81      0.89      0.85        99
           1       0.76      0.62      0.68        55

    accuracy                           0.79       154
   macro avg       0.78      0.75      0.76       154
weighted avg       0.79      0.79      0.79       154
```

   The confusion matrix plot is shown below:

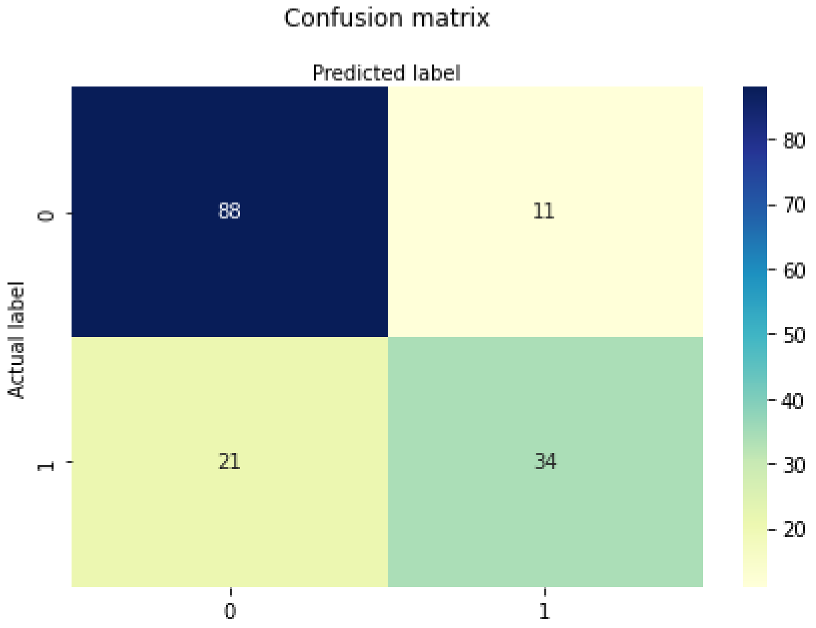


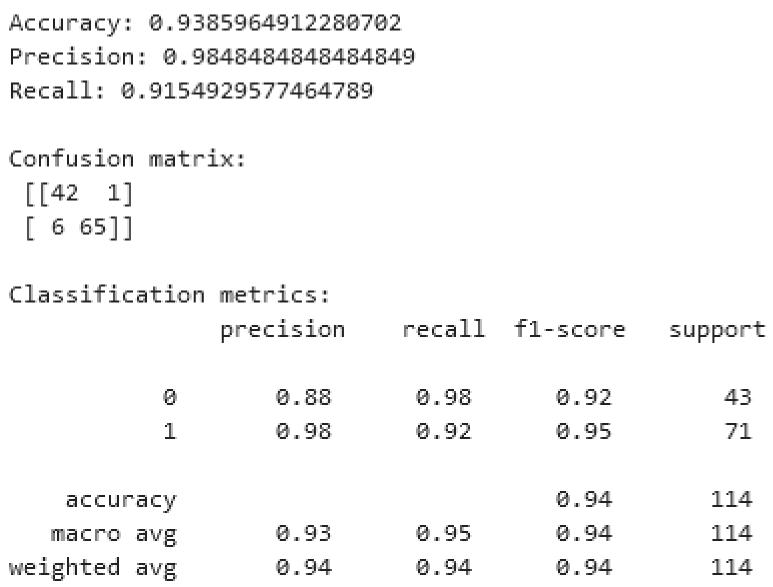
Figure 1: Confusion matrix of the Diabetes dataset

2. The binary classifier was retrained using the K-fold cross validation. K values of 5 and 10 were used in the KFold model. The results are as follows:
    a. For K = 5, an **average accuracy of 0.764** and a standard deviation of 0.021 were reported, and
    b. For K = 10, an **average accuracy of 0.765** was reported a standard deviation of 0.056 were reported.

    Comparing accuracy results of problem 1 and 2, it can be concluded that model was not overfitted on the training set.

3. Training a logistic regression classifier on the cancer dataset
    a. In this section, a logistic regression binary classifier was trained on the Cancer dataset without weight penalty. The dataset was split into 80% for training and 20% for testing. In addition, both train and test sets were scaled using the StandardScaler from Sklearn. The following parameters were used:
        i. Penalty: ‘none’
        ii. Solver: ‘lbfgsr’
        iii. Random_state: 42

        Results are presented here:



```
Accuracy: 0.9385964912280702
Precision: 0.9848484848484849
Recall: 0.9154929577464789

Confusion matrix:
 [[42  1]
 [ 6 65]]

Classification metrics:
               precision    recall  f1-score   support

           0       0.88      0.98      0.92        43
           1       0.98      0.92      0.95        71

    accuracy                           0.94       114
   macro avg       0.93      0.95      0.94       114
weighted avg       0.94      0.94      0.94       114
```

The confusion matrix plot is shown below:

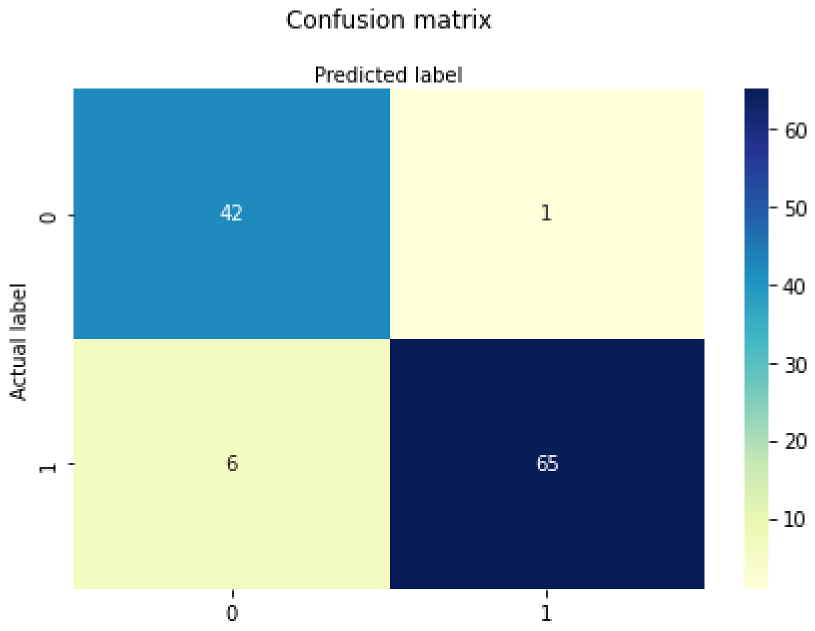
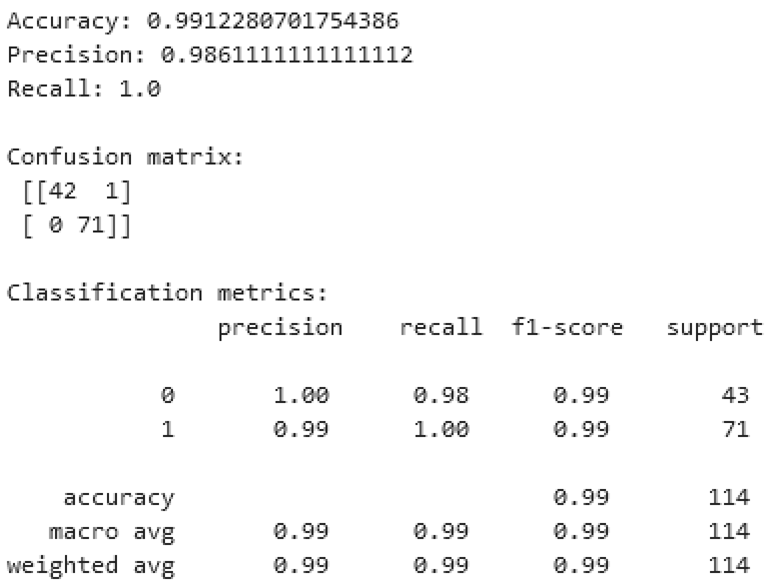


Figure 2: Confusion matrix for cancer dataset without weight penalty

b. In this section, a logistic regression binary classifier was trained on the Cancer dataset weight penalty added to the model. The following parameters were used:

    i. Penalty: ‘l2’
    ii. Solver: ‘liblinear’
    iii. C: ‘0.04’
    iv. Random_state: 42

Results as follows:



```
Accuracy: 0.9912280701754386
Precision: 0.9861111111111112
Recall: 1.0

Confusion matrix:
 [[42  1]
 [ 0 71]]

Classification metrics:
               precision    recall  f1-score   support

           0       1.00      0.98      0.99        43
           1       0.99      1.00      0.99        71

    accuracy                           0.99       114
   macro avg       0.99      0.99      0.99       114
weighted avg       0.99      0.99      0.99       114
```

The confusion matrix plot is shown below:

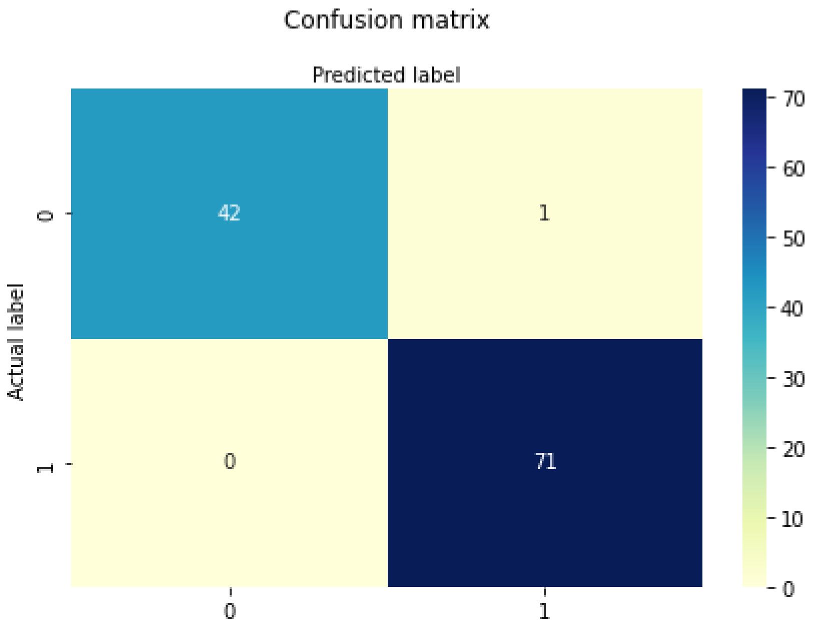


Figure 3: Confusion matrix for cancer dataset with weight penalty

4. Training a logistic regression classifier on the cancer dataset using k-fold cross validation
    a. The binary classifier from problem 3.a was retrained using the K-fold cross validation. K values of 5 and 10 were used in the KFold model. The results are as follows:
        i. For K = 5, an **average accuracy of 0.954** and a standard deviation of 0.020 were reported, and
        ii. For K = 10, an **average accuracy of 0.954** was reported a standard deviation of 0.031 were reported.

Note, in addition, the model did not converge for all fold and a ConvergenceWarning of TOTAL NO. of ITERATIONS REACHED LIMIT was reported. Please refer to the jupyter notebook for more details.

b. The binary classifier from problem 3.b was retrained using K-fold cross validation. K values of 5 and 10 were used in the KFold model. The results are as follows:

   i. For K = 5, an **average accuracy of 0.977** and a standard deviation of 0.0089 were reported, and

   ii. For K = 10, an **average accuracy of 0.977** was reported a standard deviation of 0.0157were reported.