Claude Shyaka

ID#: 801326243

ECGR-5105

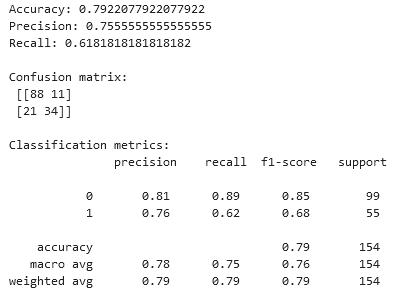
**Homework 2: Logistic Regression and K-fold Cross Validation**

**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:
   1. Penalty: ‘l2’
   2. C: 1.2
   3. Solver: ‘liblinear’
   4. Random\_state: 42

Results are shown below:



The confusion matrix plot is shown below:

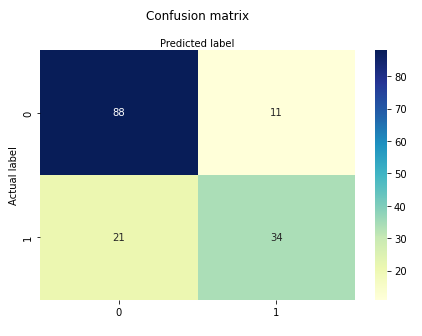


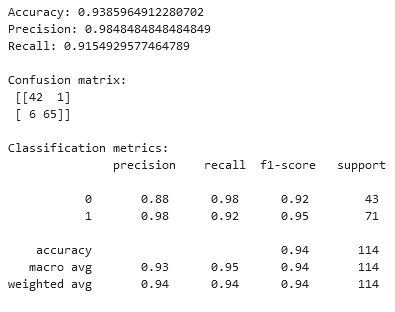
Figure 1: Confusion matrix of the Diabetes dataset

1. 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:
   1. For K = 5, an **average accuracy of 0.764** and a standard deviation of 0.021 were reported, and
   2. 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.

1. Training a logistic regression classifier on the cancer dataset
   1. 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:
      1. Penalty: ‘none’
      2. Solver: ‘lbfgsr’
      3. Random\_state: 42

Results are presented here:



The confusion matrix plot is shown below:

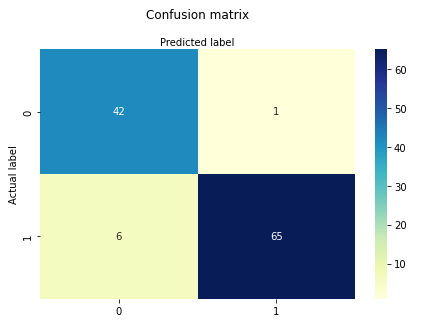
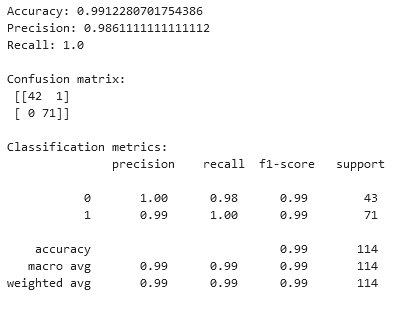


Figure 2: Confusion matrix for cancer dataset without weight penalty

* 1. 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:
     1. Penalty: ‘l2’
     2. Solver: ‘liblinear’
     3. C: ‘0.04’
     4. Random\_state: 42

Results as follows:



The confusion matrix plot is shown below:

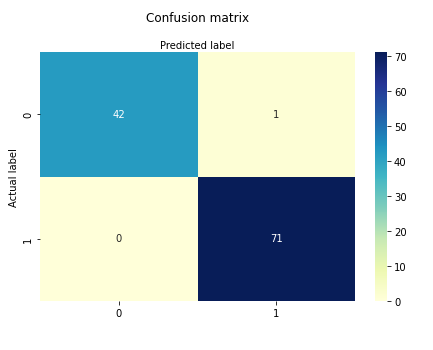


Figure 3: Confusion matrix for cancer dataset with weight penalty

1. Training a logistic regression classifier on the cancer dataset using k-fold cross validation
   1. 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:
      1. For K = 5, an **average accuracy of 0.954** and a standard deviation of 0.020 were reported, and
      2. 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.

* 1. 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:
     1. For K = 5, an **average accuracy of 0.977** and a standard deviation of 0.0089 were reported, and
     2. For K = 10, an **average accuracy of 0.977** was reported a standard deviation of 0.0157were reported.