Chunyuan Shen

ID:801322013

Homework 2

<https://github.com/cryyin/ECGR5105/tree/main/homework2>

**Problem 1 (20 points)**

Usingthe diabetes dataset, build a logistic regression binary classifier for positive diabetes. Please use 80% and 20% split between training and evaluation (test). Make sure to perform proper scaling and standardization before your training. Report your results, including accuracy, precision, and recall. At the end, plot the confusion matrix representing your binary classifier.

80% and 20% split between training and evaluation：



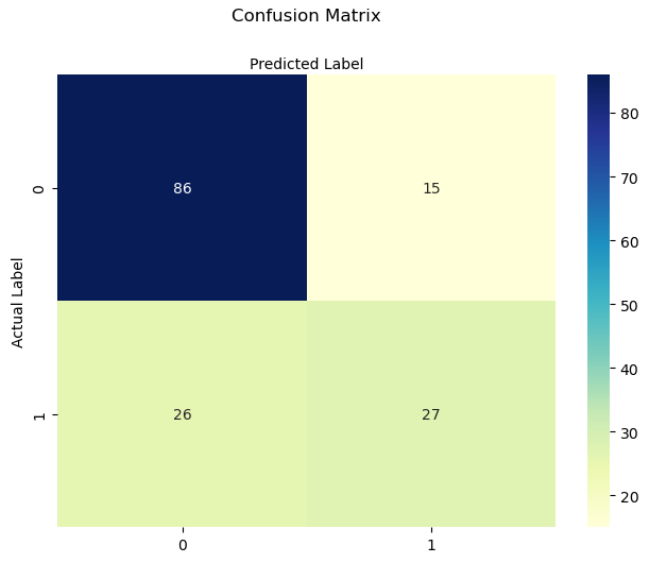
Results:

Accuracy: 0.7337662337662337

Precision: 0.6428571428571429

Recall: 0.5094339622641509

confusion matrix:

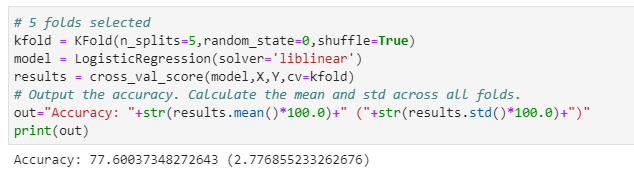


**Problem 2 (20 points)**

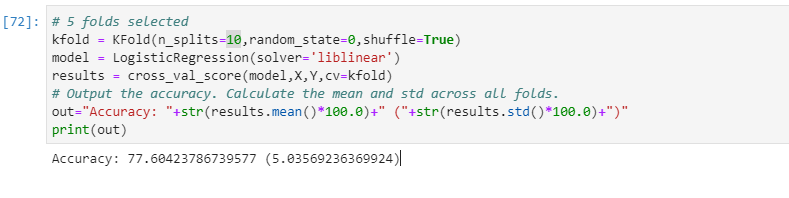
Repeat problem 1, and this time use K-fold cross-validation for your training and validation. Perform the training two times for K=5 and K=10. Analyze and compare your average accuracy against problem 1.

K-fold cross-validation:

K=5:



K=10:



As we can see, when K=5, the fluctuation range is smaller than k=10, and both of them has a better Accuracy than problem 1.

**Problem 3 (30pts):**

1. Use the cancer dataset to build a logistic regression model to classify the type of cancer (Malignant vs. benign). First, create a logistic regression that takes all 30 input features for classification. Please use 80% and 20% split between training and evaluation (test). Make sure to perform proper scaling and standardization before your training. Report your results, including accuracy, precision, and recall. At the end, plot the confusion matrix representing your binary classifier.

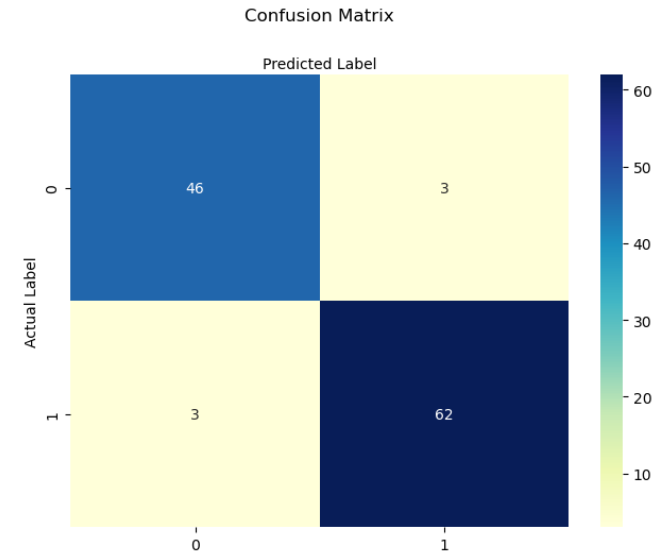
Results:

Accuracy: 0.9473684210526315

Precision: 0.9538461538461539

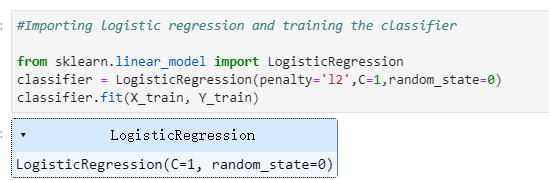
Recall: 0.9538461538461539

confusion matrix:



1. How about adding weight penalty here, considering the number of parameters. Add the weight penalty and repeat the training and report the results.

adding weight penalty:



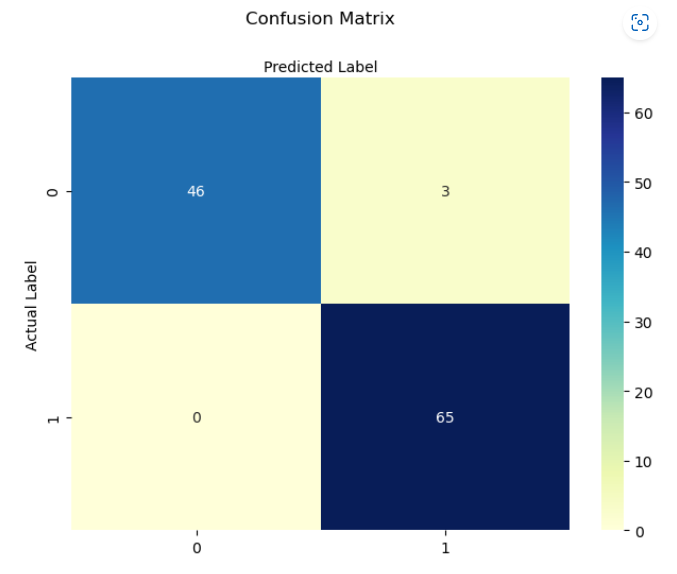
result:

Accuracy: 0.9736842105263158

Precision: 0.9558823529411765

Recall: 1.0

confusion matrix:

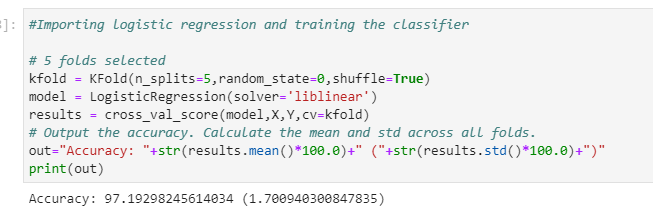


**Problem 4 (30pts):**

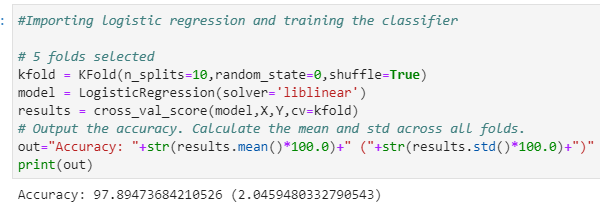
1. Repeat problem 3, and this time use K-fold cross-validation for your training and validation. Perform the training two times for K=5 and K=10. Analyze and compare your average accuracy against problem 3.

K-fold cross-validation:

K=5:

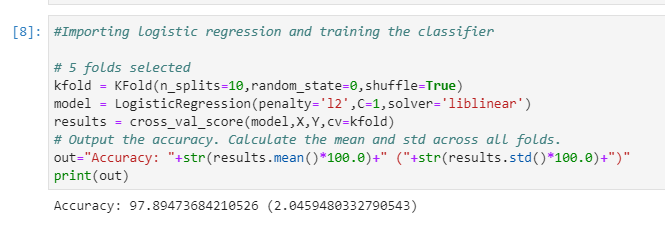


K=10:



This time both 2 of them are better than problem3, but K=10’s accuracy is better than K=5

2. How about adding weight penalty here, considering the number of parameters. Add the weight penalty and repeat the training and report the average accuracy.



There is no difference when penalty λ=1