Johnny Nguyen

Student ID: 801119047

Homework #2

GitHub Link: https://github.com/Norumai01/Intro_Machine_Learning/tree/main/HW_2

Problem 1:

- Standardization scaling was uses in the diabetes' dataset to obtain balanced scoring of the precision and recall for more accuracy. The precision and recall at negative test for diabetes is 84 percent for precision and 88 percent for recall. While at the positive test, it is 69 percent for precision and 62 percent for recall. The overall accuracy is 80 percent, which might indicate a sufficient model for determining diabetes.

Problem 2:

- Comparing the overall accuracy of the dataset to the K-fold cross validation accuracy, the validation sets show a slightly lower accuracy when compared to the training model. The training set look suitable for determining diabetes generally, might need adjustment if it isn't accurate with real datasets.

Problem 3:

- Min-Max scaling was used in the breast cancer's dataset to obtain optimized training model. The precision is 84 percent and recall are 100 percent for the negative testing. While for positive testing, the precision is 100 percent and recall are 87 percent. The overall accuracy of the testing is 92 percent, which might sound good on paper, but it is only accurate with the training model. The model will need more given dataset and testing for accuracy.
- With the parameter penalties, the training accuracy does improve with a learning rate of between 50-100. As a result, the user will see sufficient test's accuracy that can be considered for real datasets.

Problem 4:

- With the K-fold cross validation accuracy, the validation sets show slightly higher accuracy when compared to the training model. While the accuracy is high with the validation sets, the accuracy might be too ideal and need adjustment to be use for general datasets.

- With the parameter penalties, the training accuracy shows a sufficient accuracy with the test’s accuracy. With the accuracy and parameter penalties, the training model might be sufficient for real datasets.