In this project, the pre-processing for the dataset needed to be determined. First, we use SVM as the classifier model for testing of the pre-processing method.

The validation method used is k-fold validation, k is chosen as 5.

Text

Description automatically generated

From the screenshot of testing result, the standardization out-performed the raw dataset and the normalized dataset. Therefore, standardization will be used as the pre-processing method for this project.

In machine learning there are multiple classifier model can be used. The default setting of these classifier model are tested. It is used to provide the basis for what decide which classifier model will be used for further testing.

The ensemble method of voting classifier is used. Therefore, the hyperparameters for these models will be tested.

For a simple implementation of the ensemble method of the voting classifier the scores is as follows.

A picture containing text

Description automatically generated

To improve the performance. The hyperparameters is tested.

For SVM, the kernel and the regularization parameter c are one of the important hyperparameters.

In this project, gridsearch is used to tested for the best hyperparameters.

Kernels is chosen from 'linear', 'poly', 'rbf', 'sigmoid', C is chosen from 0.001, 0.01, 0.1, 1, 10, 100, 1000



From the screen-shot, the rbf kernel and at C 10 obtain the highest scores

For knn

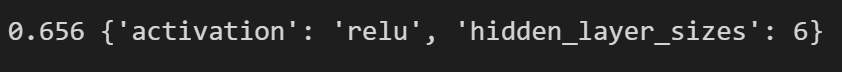
The n\_neighbors from 1 to 10 is tested



For MLP

The activation function of ('identity', 'logistic', 'tanh', 'relu')

The 'hidden\_layer\_sizes' from 3 to 15 is tested



For SGD

The loss function : hinge', 'log', 'modified\_huber', 'perceptron' is tested



Instead of decision tree, random forest is used

The 'max\_depth’ for 2 to 5

The n\_estimators':(80, 90, 100, 110, 120) is tested



We try to test some classifier and don’t use it in the ensemble method.

From the score the kNN and MLP will be eliminated in the voting classifier.

A picture containing graphical user interface

Description automatically generated

The score is improved.

.