**Machine Learning**

**Assignment 5: Train-Test Split**

Name: Jasmine I. Sayyad

PRN: 22510033

Batch: T2

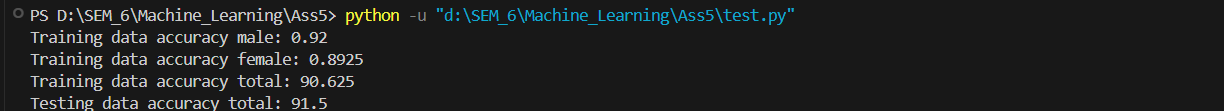
a. Generate 1000 male heights - mean 166, sd = 5.5

b. Generate 1000 female heights – mean 152, sd =4.5

c. Use test train split to set aside random 200 male and random 200 female data points as test set

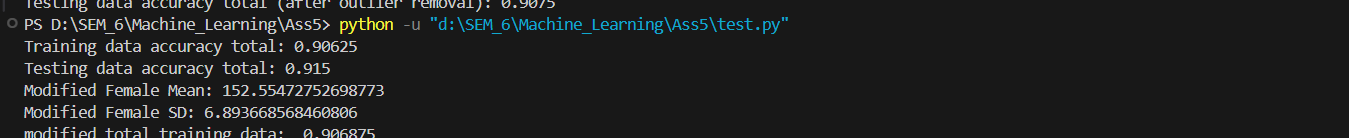
d. Use train data set of remaining 800 male and 800 female heights to train Probability based classifier. Calculate classification accuracy on both train and test data points.

Output:



e. Impact of outliers

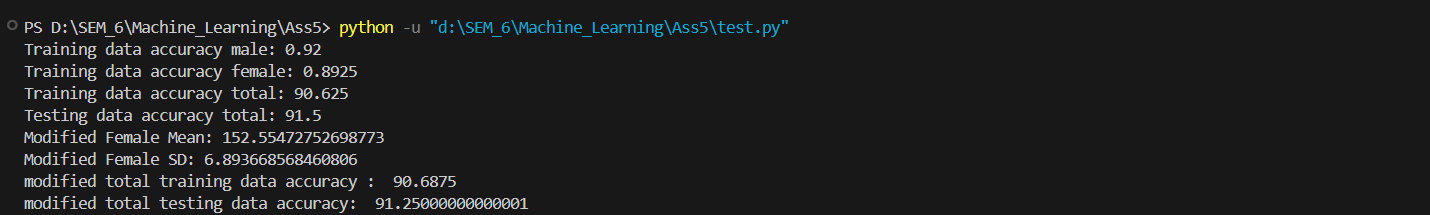
* Identify top 50 female hights in train data, increase hight of these female samples by 10 cm each. Observe change in mean and sd of train data after change in heights



The mean and SD of modified female heights will slightly increase

* Train the probability-based classification algorithm on this altered train data

1. Estimate the classification accuracy on both the train and test data



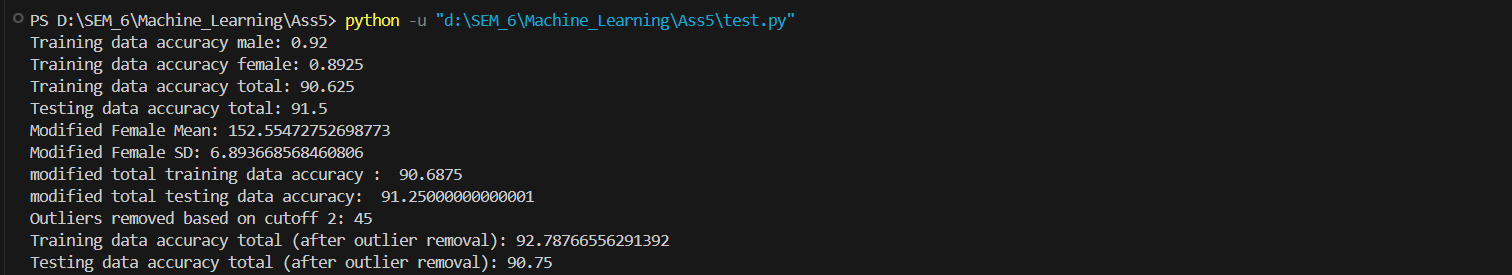
When we modify the female heights data the accuracy of training data won’t be that much affected because the overall data distribution in training does not shift drastically.

The testing data’s accuracy slightly decreased because of the removal of useful extreme value.

2. Remove outliers from the train data using z-score method on female data

3. Again, train the probability-based classification on the train data a er outlier removal and estimate classification accuracy on both test and train data

4. Observe the changes in test and train accuracy.



1. Effect on Training Accuracy: increases slightly

Reason**:** Removing outliers makes the training data more representative of the general population. Outliers can introduce noise and distort probability-based models. By eliminating outliers, the model learns a morestabledistribution, improving training accuracy.

2. Effect on Testing Accuracy:

The test accuracy dropped very slightly this is may be due to fewer outliers were removed from the train set, which means it may still contain values that don’t match the cleaned train set.

f. Impact of Trimming

For k in range (1:15)

1. Trim upper and lower k% of female train data set

2. Train probability based on classifier on female trimmed train dataset and male train data set

3. Calculate accuracy of classification on both train and test data set

Observe impact of trimming on classification accuracy on train and test data sets

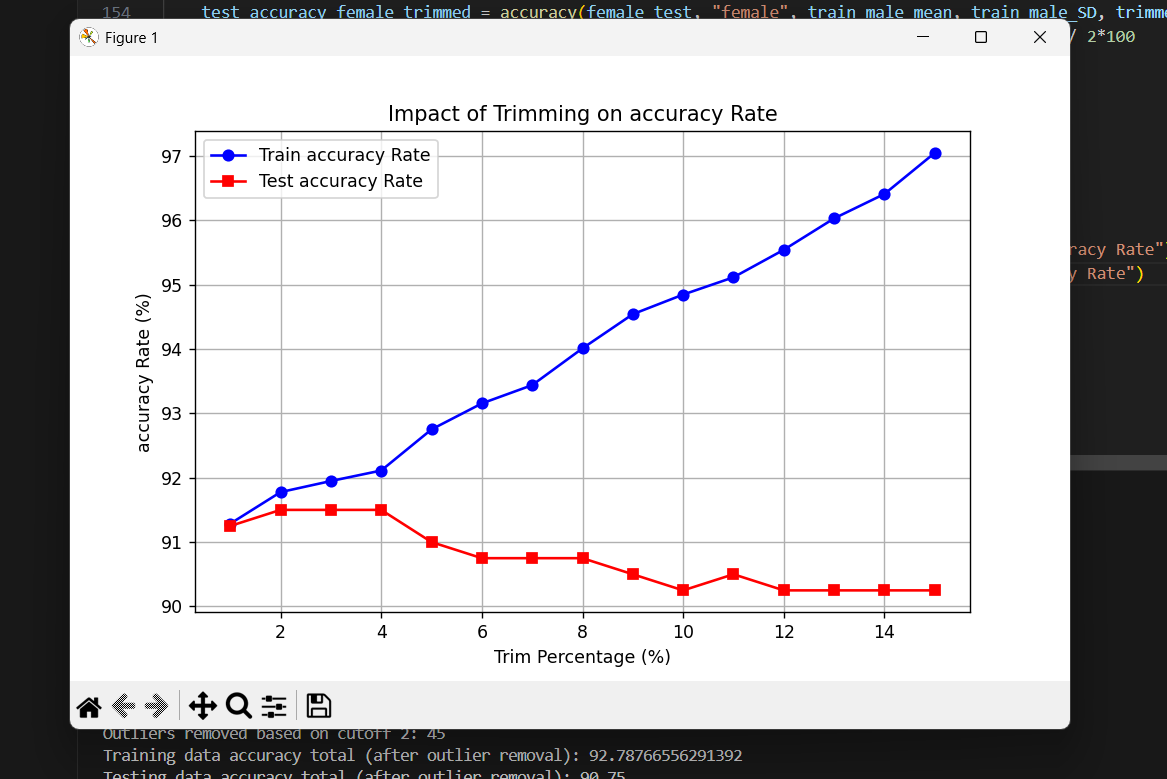
Observations:

Training Accuracy Increases with More Trimming

* As the trim percentage increases, training accuracy improves consistently.
* This is because outlier removal makes the training data cleaner

Test Accuracy Remains Almost Constant or Slightly Declines

* Initially, test accuracy remains stable with increasing trimming.
* However, as the trim percentage incre\test accuracy **begins to drop**.
* This suggests that too much trimming removes valuable information, leading to a less generalizable model.



Moderate outlier removal (4-6%)helps improve training accuracy without harming test accuracy**.**

Excessive trimming (>10%) leads to overfitting, where train accuracy is high, but test accuracy drops.