Assignment 5: Train-Test Split

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.

e. Impact of outliers

- i. Identify top 50 female hights in train data, increase hight of these female samples by 10 cm each
  - 1. Observe change in mean and sd of train data after change in heights
- ii. Train the probability-based classification algorithm on this altered train data
  - 1. Estimate the classification accuracy on both the train and test data
  - 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 after outlier removal and estimate classification accuracy on both test and train data
  - 4. Observe the changes in test and train accuracy.

f. Impact of Trimming

- i. Consider the female train data including the 50 outliers for this section
- ii. 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
- iii. Observe impact of trimming on classification accuracy on train and test data sets

Submission Guidelines:

Submit your code (5 marks) and observations (5 marks) by 10 am on Thursday 27th Feb 2025