

### Assignment 3: Univariate Outlier Detection and Analysis

- a. Identify top 50 female heights in distributions generated in assignment 1, increase height of these female samples by 10 cm each
- b. Observe change in sample mean and sd after change in heights
- c. Run Classification algorithms developed in assignment 1.c on this altered dataset and note change in classification accuracy in each case
- d. Design strategies to detect outliers in female sample set
  - i. Visual –
    1. plot the data histogram and observe gaps, elbow etc
    2. box and whisker plot. Use of whiskers to find outliers
  - ii. Parametric –
    1. convert heights into z score,
    2. experiment with z score cutoffs such as 2 and 3 ( on both sides)
  - iii. Non parametric
    1. Detection and removal based on inter quartile range
    2. Detection of outliers based on MAD
    3. Experiment with cutoffs such as 1.5, 2, 3 etc ( on both sides)
- e. Remove data labelled as outliers using z score or iqr or MAD cutoffs
- f. Run again the classification methods from assignment 1.c and document impact on mean, sd and classification accuracy
- g. Data trimming- drop lower and upper k% data (vary k between 1% to 15% in increments of 1%) from 1.a and run classification algorithms. Observe impact on accuracy via scatter plot

#### ML Assignment 3 Submission Guidelines :

1. Submission closes by 10:00am Monday 10<sup>th</sup> Feb 2025
2. Submit your code and observations in separate python and doc file respectively
3. File Name Convention for python, doc and wrapping Zip file – BatchName\_PRN\_A3. Substitute appropriate values of BatchName and PRN
4. Please note the following
  - a. Max score for the assignment will be 10.
  - b. Copied code shall lead to disqualification of submission.
  - c. Code and Observations will carry 5 marks each.