**Assignment 3**

**Machine Learning**

**Name: Jasmine I. Sayyad**

**PRN: 22510033**

**Batch: T-2**

Q1. Identify top 50 female hights in distributions generated in assignment 1, increase hight of these female samples by 10 cm each b. Observe change in sample mean and sd a er change in heights c. Run Classification algorithms developed in assignment1.c on this altered dataset and note change in classification accuracy in each case

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

from scipy.stats import norm

male\_mean = 166

female\_mean = 152

SD = 5

female\_ht = pd.Series(np.random.normal(female\_mean,SD,1000))

male\_ht = pd.Series(np.random.normal(male\_mean,SD,1000))

total = female\_ht.size + male\_ht.size

plt.boxplot(female\_ht,  patch\_artist=True, boxprops=dict(facecolor="lightblue"))

plt.title("original female heights")

plt.show()

plt.hist([female\_ht, male\_ht], bins = 100 , label = ['Female','Male'])

plt.title("original distribution")

plt.legend(loc = 'upper right')

plt.show()

# top 50 female heights modified

top\_indices = np.argsort(female\_ht)[-50:]

female\_ht[top\_indices] += 10

changed\_mean = female\_ht.mean()

print(changed\_mean)

changed\_SD = female\_ht.std()

print(changed\_SD)

plt.hist([female\_ht, male\_ht], bins = 100 , label = ['Female','Male'])

plt.title("distribution after changing female heights")

plt.legend(loc = 'upper right')

plt.show()

#calculating probability

F\_mean = female\_ht.mean()

M\_mean = male\_ht.mean()

F\_SD = female\_ht.std()

M\_SD = male\_ht.std()

misclassifiedmale = 0

for curr\_ht in male\_ht:

    female\_prob = norm.pdf(curr\_ht,F\_mean,F\_SD)

    male\_prob = norm.pdf(curr\_ht,M\_mean,M\_SD)

    if male\_prob < female\_prob:

        misclassifiedmale += 1

misclassifiedfemale = 0

for curr\_ht in female\_ht:

    female\_prob = norm.pdf(curr\_ht,F\_mean,F\_SD)

    male\_prob = norm.pdf(curr\_ht,M\_mean,M\_SD)

    if male\_prob > female\_prob:

        misclassifiedfemale += 1

print("MisplacedMales & misplacedFemales count : ", misclassifiedmale,misclassifiedfemale)

total\_misclassification = misclassifiedfemale + misclassifiedmale

rate = (total\_misclassification/total) \* 100

print("Misclassification rate : ", rate)

#Deriving threshold height to seperate male and female

max\_female\_height = float(female\_ht.max())

min\_male\_height = float(male\_ht.min())

threshold = min\_male\_height

min\_misclassification = 10\*\*12

for curr\_height in np.arange(min\_male\_height, max\_female\_height + 1.0, 1.0):

    misclassification = 0

    misclassification += sum(male\_ht < curr\_height) + sum(female\_ht > curr\_height)

    if(misclassification < min\_misclassification ):

        min\_misclassification = misclassification

        threshold = curr\_height

print("Threshold Height: ", threshold)

#quantize the  data at scale of 0.5 cm and empirically estimate the likelihood of male female in each segment based on majority

def frequency(scale, height):

    quantized\_val = np.floor(height/scale)

    no\_of\_ppl = quantized\_val.value\_counts()

    return no\_of\_ppl

interval = 0.5

new\_female\_ht = frequency(interval, female\_ht)

new\_male\_ht = frequency(interval, male\_ht)

min\_overlap\_height = int(new\_male\_ht.index.min())

max\_overlap\_height = int(new\_female\_ht.index.max())

total\_misplaced = 0

for ht in range(min\_overlap\_height, max\_overlap\_height+1):

    # ht = float(ht)

    Female\_count = new\_female\_ht.get(ht, 0)  # Use .get() for safe access

    Male\_count = new\_male\_ht.get(ht, 0)

    misplaced = min(Female\_count,Male\_count)

    total\_misplaced += misplaced

print("Total Misplaced: ", total\_misplaced)

misplaced\_rate = (total\_misplaced/total)\*100

print("rate of misplaced : ", misplaced\_rate)

#box and whisker plot. User of whiskers to find outliers

plt.boxplot(female\_ht,  patch\_artist=True, boxprops=dict(facecolor="lightblue"))

plt.title("modified female height")

plt.show()

#Parametric –

#1. convert heights into z score

Female\_ht\_zscore  = (female\_ht-changed\_mean)/changed\_SD

#experiment with z score cutoffs such as 2 and 3 ( on both sides)

outliers\_with2 = 0

outliers\_with3 = 0

for curr\_ht in Female\_ht\_zscore:

    if curr\_ht > 2 or curr\_ht < -2:

        outliers\_with2 += 1

    if curr\_ht >3 or curr\_ht < -3:

        outliers\_with3 += 1

print("Outliers based on cutoff 2: ",outliers\_with2)

print("Outliers based on cutoff 3: ", outliers\_with3)

# Detection and removal based on inter quartile range

Q1 = female\_ht.quantile(0.25)

Q3 = female\_ht.quantile(0.75)

IQR = Q3 - Q1

upper\_bound = Q3 + 1.5\*IQR

lower\_bound = Q1 - 1.5\*IQR

IQR\_outliers = []

for curr\_height in female\_ht:

    if curr\_height < lower\_bound or curr\_height > upper\_bound:

        IQR\_outliers.append(curr\_height)

# print(IQR\_outliers)

# Detection of Outliers Based on Median Absolute Deviation (MAD)

median = female\_ht.median()

MAD = np.median(np.abs(female\_ht - median))

MAD\_cutoff = 1.5

modified\_zscore = 0.6745 \* (female\_ht - median) / MAD

MAD\_outliers = []

for i in range(len(female\_ht)):

    if abs(modified\_zscore[i]) > MAD\_cutoff:

        MAD\_outliers.append(female\_ht[i])

# print(MAD\_outliers)

# Remove data labelled as outliers using z score or iqr or MAD cutoffs

IQR\_female\_ht = []

for curr\_height in female\_ht:

    if (curr\_height >= lower\_bound) and (curr\_height <= upper\_bound):

        IQR\_female\_ht.append(curr\_height)

IQR\_female\_ht = pd.Series(IQR\_female\_ht)

MAD\_female\_ht = []

for i in range(len(female\_ht)):

    if abs(modified\_zscore[i]) <= MAD\_cutoff:

        MAD\_female\_ht.append(female\_ht[i])

MAD\_female\_ht = pd.Series(MAD\_female\_ht)

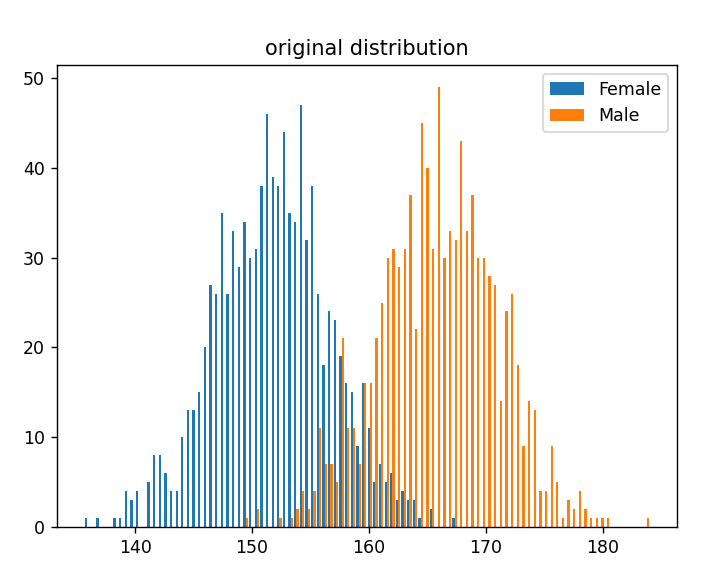
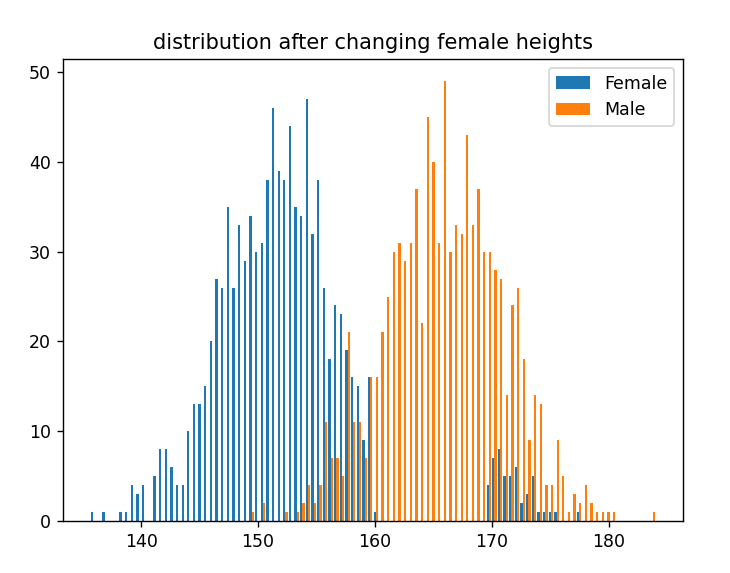
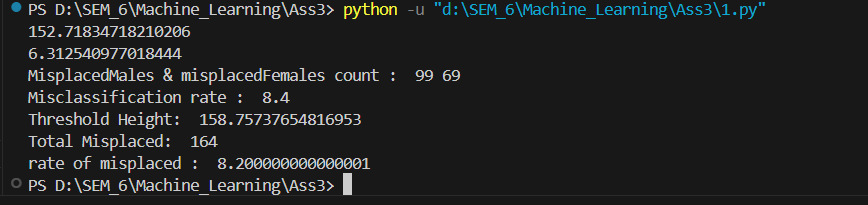
plt.boxplot([female\_ht, IQR\_female\_ht, MAD\_female\_ht],

            patch\_artist=True, boxprops=dict(facecolor="lightblue"))

plt.xticks([1, 2, 3], ["Original", "IQR Filtered", "MAD Filtered"])

plt.title("Comparison of Outlier Removal Methods using Boxplots")

plt.show()

Output:  
  


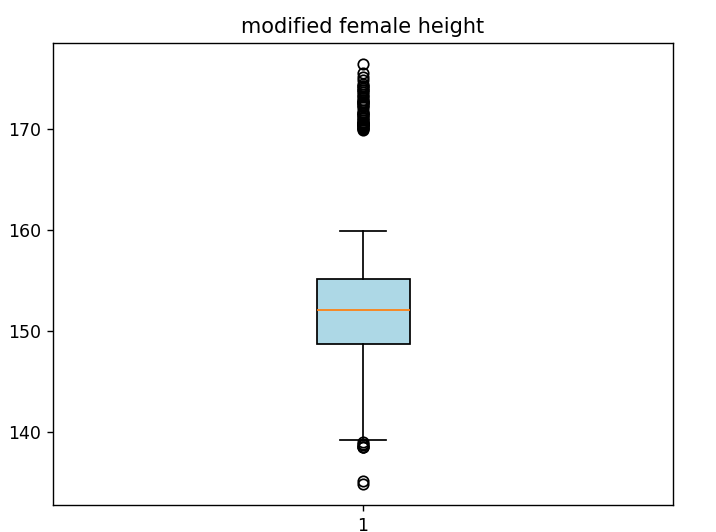
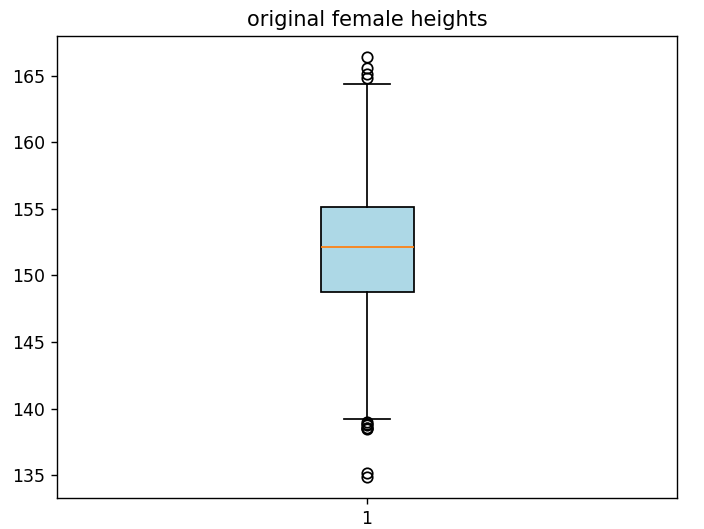
**Observations:**  
When we modify top 50 female heights by adding 10cm, some tall females will be misclassified as male. SD of females will increase and the change in mean is not that significant.

Misclassification rate will also increase.

Accuracy won’t be that much affected because we are just modifying 50 records, there will be slight decrease in accuracy. So only the misclassified females can increase.

Q2. Design strategies to detect outliers in female sample set i.

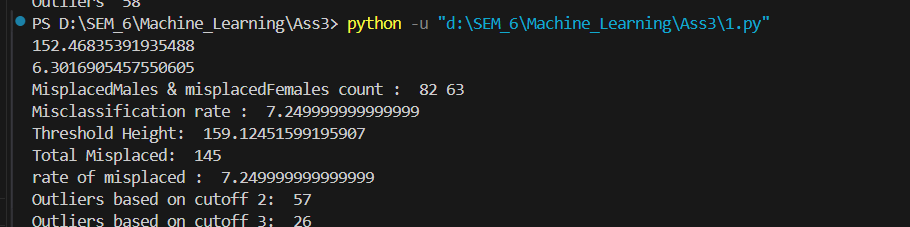
Visual – 1. plot the data histogram and obverse gaps, elbow etc

2. box and whisker plot. Use of whiskers to find outliers

**Observation:** Increased the number of outliers at upper whisker. The IQR is reduced.

ii. Parametric – 1. convert heights into z score,

2. experiment with z score cutoffs such as 2 and 3 (on both sides)



iii. Nonparametric-

1. Detection and removal based on interquartile range

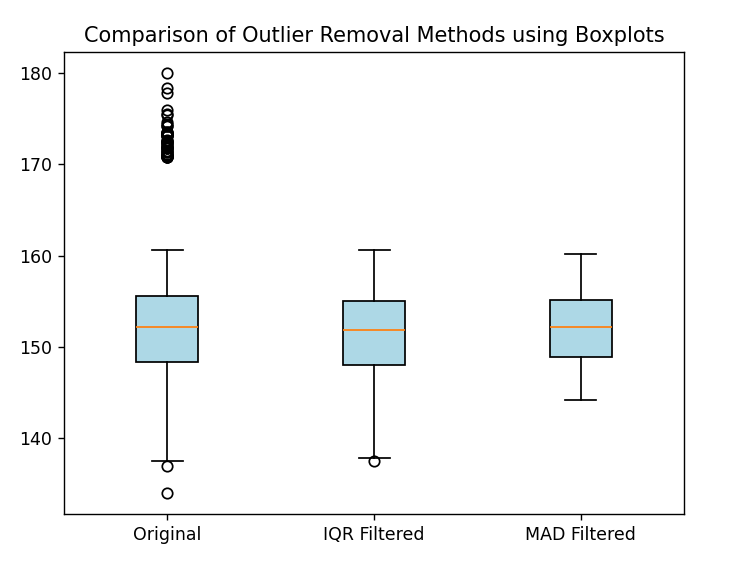
observation: removes extreme values at the tails of the distribution.

2. Detection of outliers based on MAD

3. Experiment with cutoffs such as 1.5, 2, 3 etc (on both sides)

Higher cutoffs detects only extreme values and lower detect more outliers.

e. Remove data labelled as outliers using z score or iqr or MAD cutoffs



**Observations**: The Median Absolute Deviation (MAD) method usually detects more outliers than the Interquartile Range (IQR) method. Because IQR considers extreme points and in case of MAD, it is based on the deviation from median.

If we increase MAD cutoff it will decrease the number of outliers detected.

f. Run again the classification methods from assignment 1.c and document impact on mean, sd and classification accuracy

**Observation:**

1) The mean of modified data will increase since we are increasing the top 50 female heights by 10cm

2) The SD will also increase   
3) More females will be misclassified as males due to larger heights thus classification accuracy will decrease.

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.

**Observation:** the misclassification rate will decrease with increased data trimming rate but after certain percentage trimming the misclassification rate gets saturated.

