Introduction to Data Science

Semester Project



Session: 2021-2025

**Submitted To:**

Dr. Irfan Yousaf

**Submitted by:**

Zain Ali

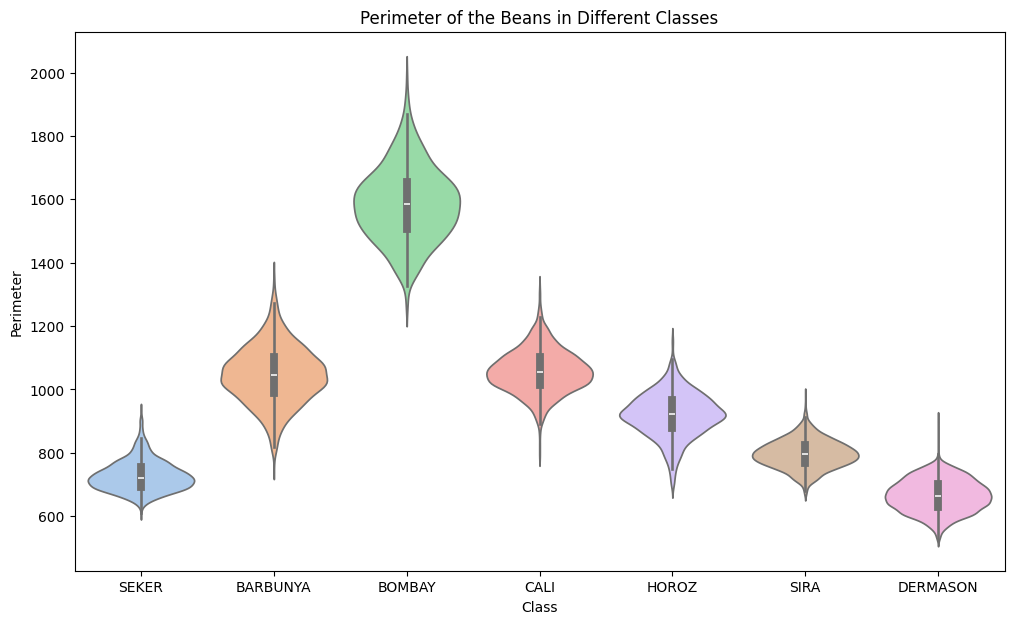
**Registration No:**

2021-SE-23

Department of Computer Science, New Campus **University of Engineering and Technology Lahore, Pakistan**

# Bivariate Analysis

## Violin Plot



2021-SE-23

Figure 1

The violin plot shows the visualization of perimeters of beans in different classes. The classes are labeled on the x-axis, and the perimeter is on the y-axis. The wider the body of the violin plot at a particular point on the x-axis, the more data points there are at that perimeter value for that class. The thin line in the middle of the violin plot shows the median perimeter value, and the quartiles are denoted by the ends of the wider section of the violin plot.

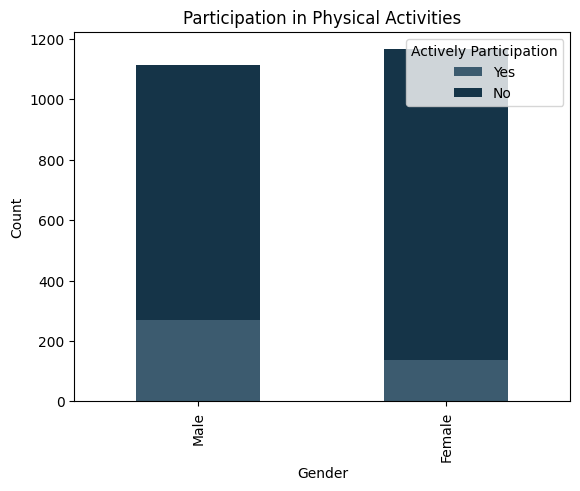
The violin plot for the “CALI” class shows that the median perimeter of beans in that class is around 1300, with a wider range of perimeters from 1100 to 1500. There are also a few data points that have perimeters outside this range.

“CALI” and “HOROZ” classes have narrower violins, indicating a more concentrated distribution of perimeters, with most beans having similar sizes.

By comparing the violin plots of the different classes, you can see how the distribution of perimeter values differs between the classes. For instance, the “DERMASON” class appears to have a wider range of perimeter values than the “CALI” class.

Overall the violin plot allows you to visually compare the distribution of bean perimeters across different classes. It reveals that some classes have a wider variety of bean sizes, while others tend to have more uniformly sized beans.

## stacked bar plot



2021-SE-23

Figure 2

This stacked bar chart illustrates the distribution of male and female participants engaging in moderate or vigorous-intensity sports, fitness, or recreational activities. Each bar is segmented into two sections representing "Participants" and "Non-participants".

The horizontal axis demonstrates gender (male and female), while the vertical axis depicts the count of individuals. For each gender, there are two segments: "Yes" indicating participants and "No" indicating non-participants.

Observing the chart, it becomes evident that there is a significant gender disparity in participation levels. Among males, over 250 individuals are actively involved in weekly recreational activities, with a notable portion, exceeding 800, not participating. Conversely, the participation rate among females is comparatively lower, with approximately 170 actively participating and around 1000 not engaged.

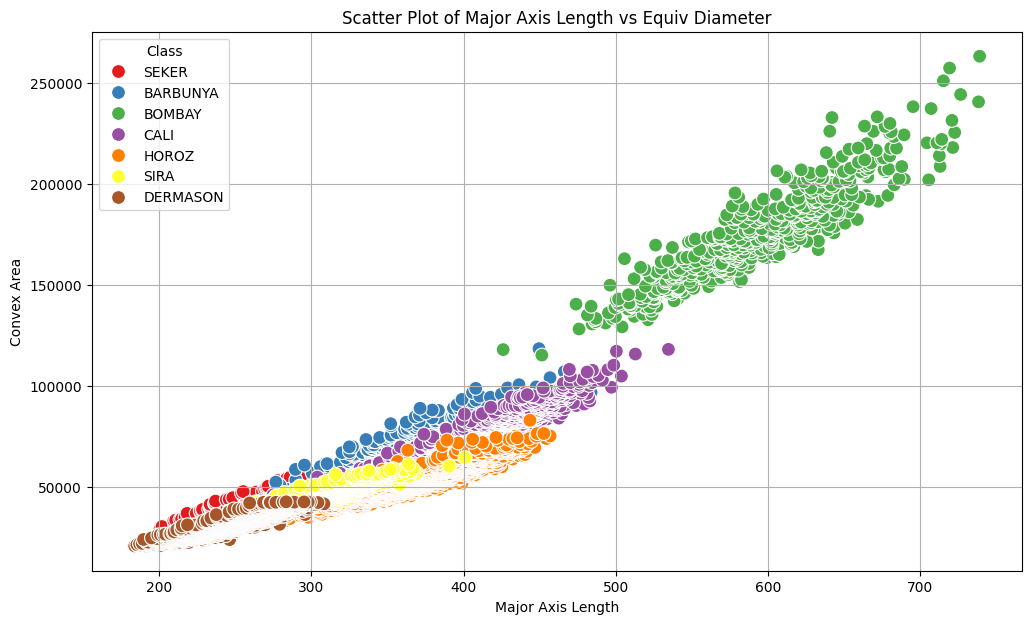
Upon closer examination, it's discernible that the overall participation rate in moderate and vigorous physical activity is notably low for both genders. The ratio of participants to non-participants stands at approximately 7:30, underscoring a substantial portion of the population not engaging in such activities.

Further analysis reveals a gender discrepancy within the participating demographic. Males exhibit a higher participation rate compared to females, with a ratio of 13:9 between male and female participants. Conversely, among non-participants, females outnumber males, with a ratio of 4:5, indicating a higher proportion of females not engaging in physical activities compared to their male counterparts.

In summary, the stacked bar chart highlights the disparity in participation levels between genders, emphasizing the need for targeted interventions to promote physical activity, particularly among females, and address the overall low participation rates in moderate and vigorous physical activities.

# Multivariate Analysis

## Scatter Plot



2021-SE-23

Figure 3

This scatter plot shows the classification of dry beans based on their Major Axis length and Convex Area. The x-axis shows the Major Axis Length of Dry Bean, while the y-axis represents the Convex Area of Dry Beans. The legend represents the different classes (i.e. SEKER, BARBUNYA, BOMBAY, CALI, HOROZ, SIRA, DERMASON) to which dry bean belongs. Each colored dot on agraph represents a data point.

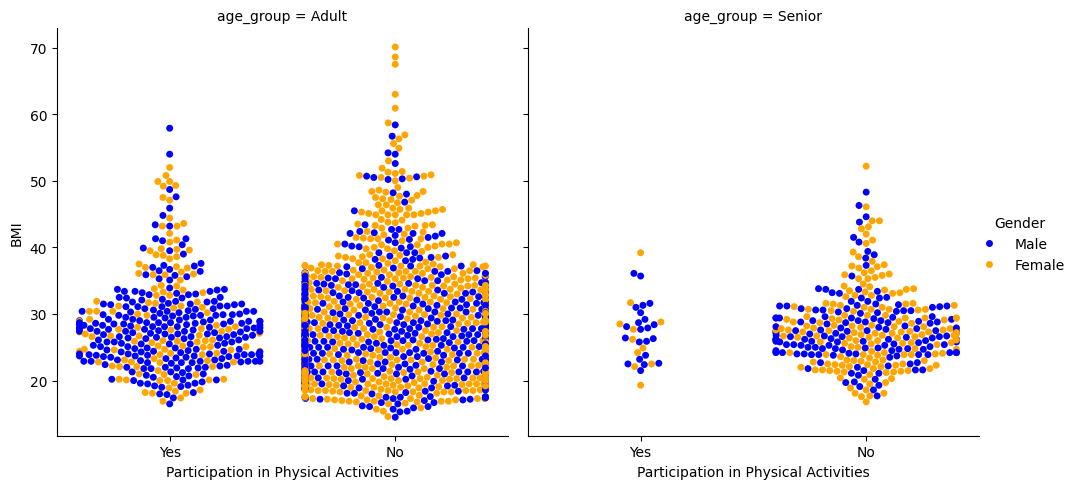
The graph displays the direct correlation between the equivalent diameter and the major axis length. The comparable diameter is growing in tandem with the major axis length. The dry beans with an equal main axis length ranging from 100 to 300 diameter ranging from 160 to 230 is classified as "DERMASON". The class "SEKER" includes dry beans with main axis lengths between 200 and 320 and comparable diameters between 200 and 270.

The graph displays the direct correlation between the equivalent diameter and the major axis length. The comparable diameter is growing in tandem with the major axis length. The dry beans with an equal main axis length ranging from 100 to 300 diameter ranging from 160 to 230 is classified as "DERMASON". The class "SEKER" includes dry beans with main axis lengths between 200 and 320 and comparable diameters between 200 and 270.

Dry beans fall into one of two classes: BOMBAY when the major axis length is 500 and the equivalent diameter is 400, and DERMASON when the major axis length is less than 200 and the equivalent diameter is less than 180.

In Conclusion, this scatter plot delineates distinct classifications of dry beans based on major axis length and convex area, highlighting their correlation with equivalent diameter. It offers valuable insights for dry beans and different areas, aiding in bean classification.

## Swarm Plot



2021-SE-23

Figure 4

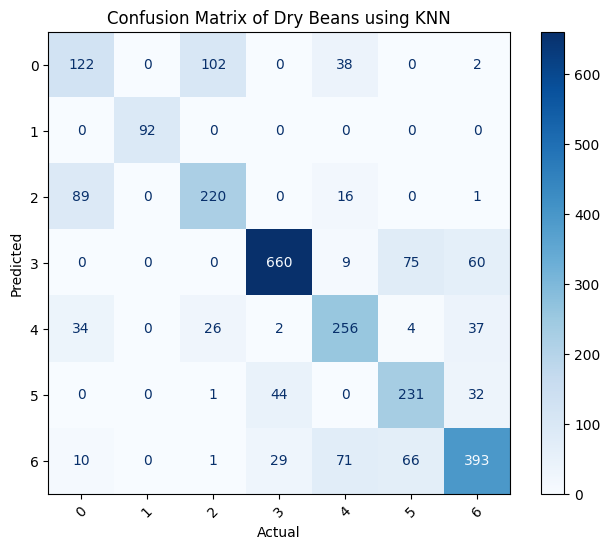
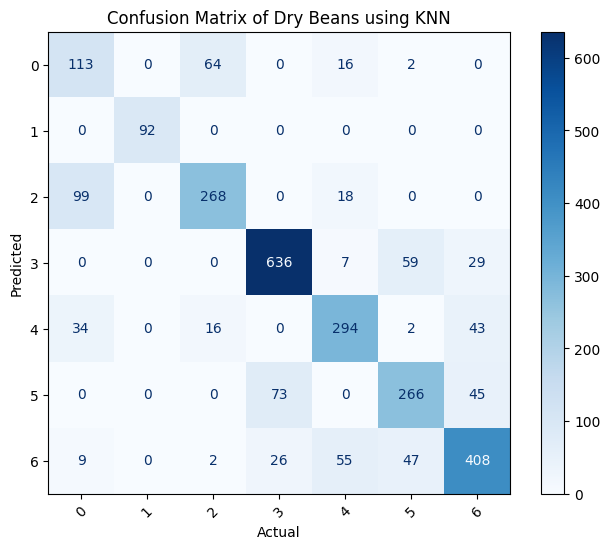
This swarm plot illustrates the Body Mass Index (BMI) across two age demographics: adults and seniors, categorized by their involvement in moderate or vigorous-intensity recreational activities. On the x-axis, participation in recreational activities is delineated, while the y-axis denotes the BMI of each individual. Males are depicted in a blue dots , while females are represented in a yellow color, with each dot symbolizing a data point.

Among adults, a notable proportion abstains from recreational activities, exhibiting a higher BMI compared to their active counterparts. Their BMI ranges from 5 to 75, with a concentration between 20 to 35. Similarly, among seniors, those not engaged in recreational activities tend to have higher BMIs, ranging from 10 to 50, with a focal point between 20 to 30.

Participation rates differ between genders and age groups. Overall, males demonstrate higher participation levels compared to females in both adult and senior cohorts. Conversely, female participation is notably lower across both demographics.

In summary, the plot reveals that inactive adults and seniors tend to have higher BMIs compared to their active counterparts. Moreover, while male participation surpasses female involvement across all age groups, participation rates are generally higher among adults compared to seniors.

# Dry\_Bean\_Dataset Results

2021-SE-23

2021-SE-23

Figure 5 Figure 6

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1-Score |
| For KNN Model | 0.725 | 0.730 | 0.718 | 0.722 |
| For Naïve Bayes Model | 0.763 | 0.763 | 0.759 | 0.759 |

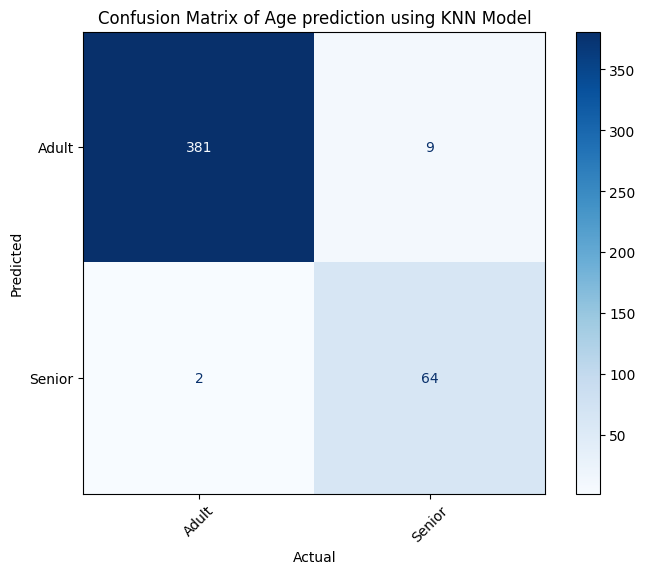
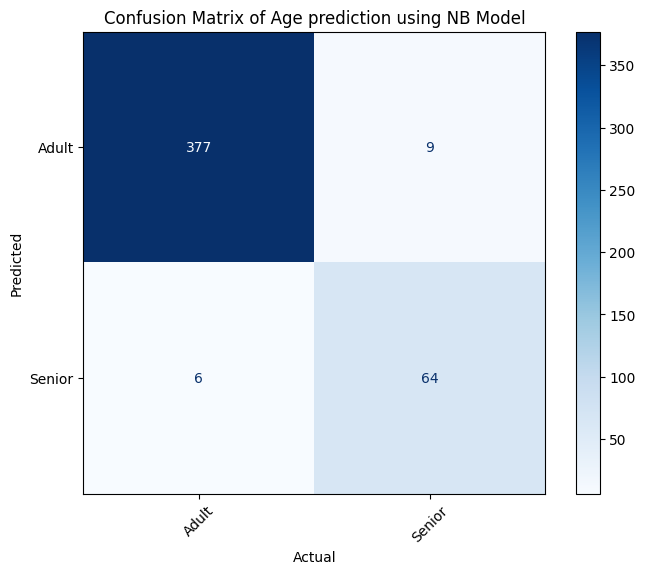
### Results:

The confusion matrix produced by using the K Nearest Neighbours Classifier on the Dry Bean Dataset (Dry\_Bean\_Dataset.csv) is shown in Figure 5, while the confusion matrix produced by using the Gaussian Naive Bayes Classifier on the same dataset is shown in Figure 6.

The diagonal parts in both figures show how many cases were correctly predicted, while the off-diagonal elements show how many cases were incorrectly classified.

The Gaussian Naive Bayes classifier performs better on this dataset, achieving an accuracy of 76%, than the K Nearest Neighbours classifier, which achieves an accuracy of 72%. While the Gaussian Naive Bayes Classifier displays an F1-score of 72%, recall of 75%, and accuracy of 73%, the K Nearest Neighbours Classifier displays a recall of 71%, F1-score of 75%, and precision of 73%.

# NHANES\_age\_prediction Results

2021-SE-23

2021-SE-23

Figure 7 Figure 8

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1-Score |
| For KNN Model | 0.976 | 0.973 | 0.936 | 0.953 |
| For Naïve Bayes Model | 0.967 | 0.945 | 0.938 | 0.931 |

### Results

The confusion matrix produced by the K Nearest Neighbours Classifier from the NHANES\_age\_prediction.csv dataset is shown in Figure 7. The following are some of the insights this matrix offers:

* True Positive (TP): 381 people who were accurately predicted and identified as adults.
* False Positive (FP): Nine people who were expected to be adults but were categorised as seniors.
* False Negative (FN): Two people categorised as adults who were expected to be seniors.
* True Negative (TN): 64 people correctly predicted and recognised as seniors.

The confusion matrix produced by the Gaussian Naive Bayes Classifier from the NHANES\_age\_prediction.csv dataset is shown in Figure 8. The following is expounded upon in this matrix:

* 377 people were successfully identified as adults and their predictions were true positives (TP).
* False Positive (FP): Nine people who were expected to be adults but were categorised as seniors.
* False Negative (FN): six people who were categorised as adults but were expected to be seniors.
* True Negative (TN): 64 people correctly predicted and recognised as seniors.But when it comes to this dataset, the K Nearest Neighbours Classifier performs better, with 97% accuracy as opposed to 96% for the Gaussian Naive Bayes Classifier.