جـــامـــعـــة
الملك سعود
King Saud University

Body Performance

IT326| Data Mining project

Rawan Alsaif	443200449
Hadeel Alsaleh	441201424
Razan Alkhaluqy	443204373
Lama Alnahabi	443201417

Section:71680

Supervised by: I.Hanan AlTamimi

OUTLINE

01. **PROBLEM**

02. **DATA INFORMATION**
Original data, graphs

03. **DATA PREPROCESSING**
Data cleaning and transformation

04. **DATA MINING TASK**
Classification , clustering

05. **FINDINGS**

01. Problem & Goal

Problem:

- Increasing prevalence of smartphones impacting how we interact with our bodies.
- There is a lack of awareness regarding the factors that influence body performance.

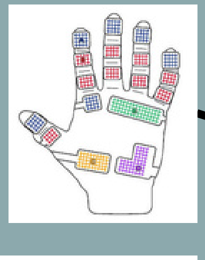
Goal:

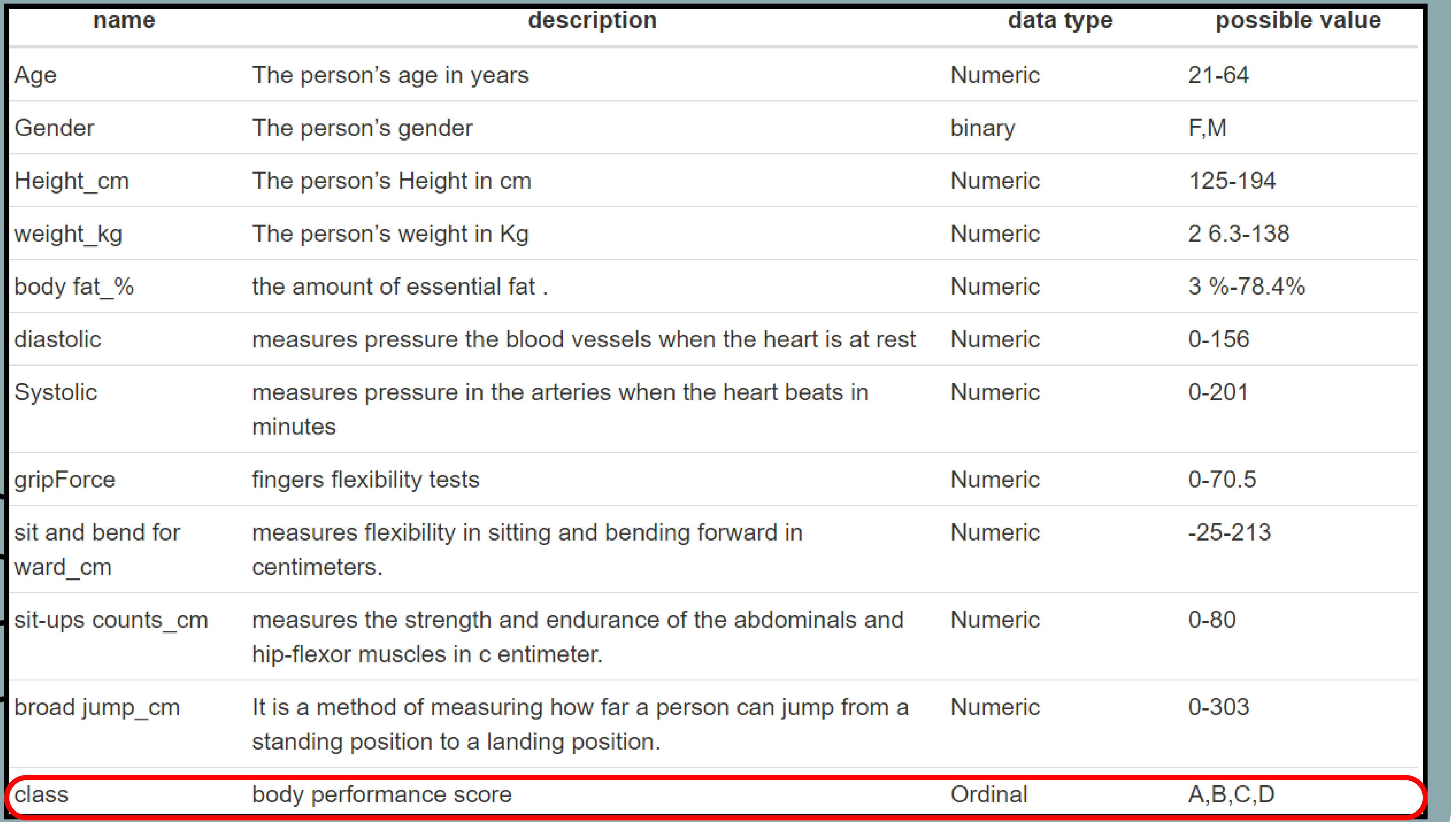
- Analyze and extract insights from body performance data.

02. DATA INFORMATION

#rows: 13393
#columns: 12

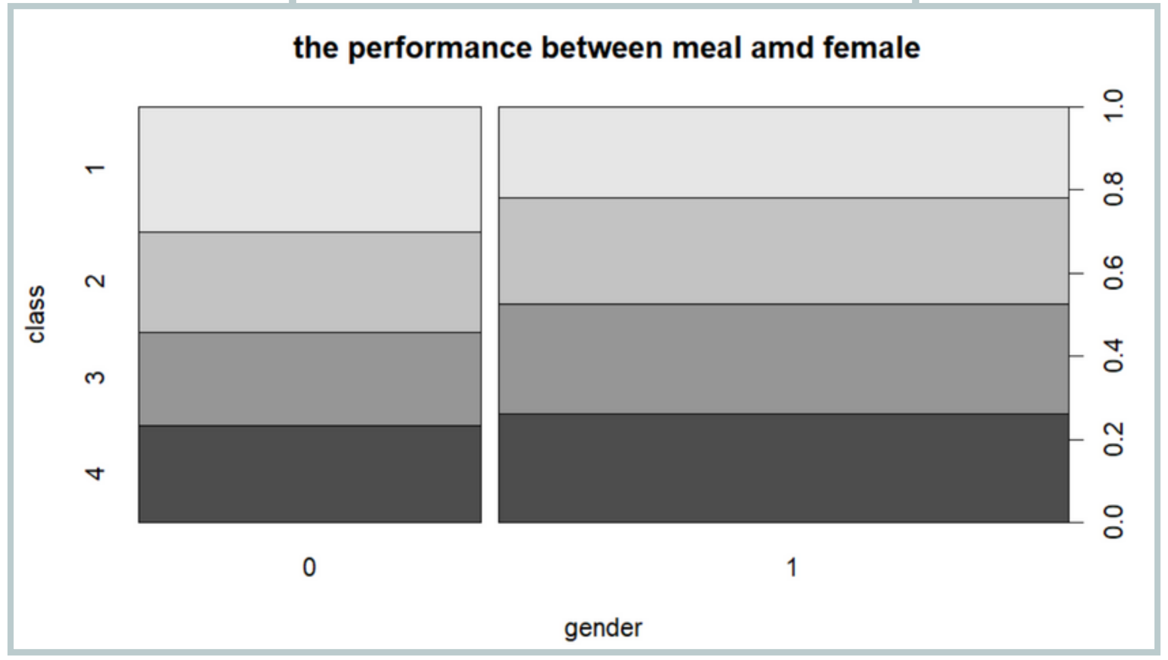
class labe: class



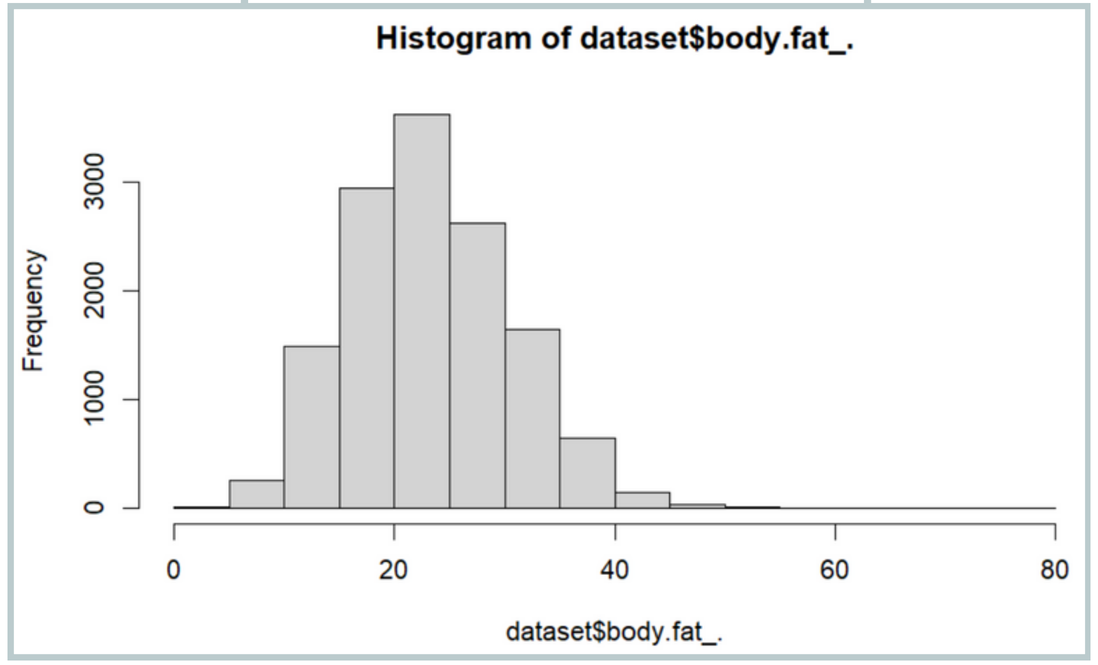


name	description	data type	possible value
Age	The person's age in years	Numeric	21-64
Gender	The person's gender	binary	F,M
Height_cm	The person's Height in cm	Numeric	125-194
weight_kg	The person's weight in Kg	Numeric	2 6.3-138
body fat_%	the amount of essential fat .	Numeric	3 %-78.4%
diastolic	measures pressure the blood vessels when the heart is at rest	Numeric	0-156
Systolic	measures pressure in the arteries when the heart beats in minutes	Numeric	0-201
gripForce	fingers flexibility tests	Numeric	0-70.5
sit and bend for ward_cm	measures flexibility in sitting and bending forward in centimeters.	Numeric	-25-213
sit-ups counts_cm	measures the strength and endurance of the abdominals and hip-flexor muscles in c entimeter.	Numeric	0-80
broad jump_cm	It is a method of measuring how far a person can jump from a standing position to a landing position.	Numeric	0-303
class	body performance score	Ordinal	A,B,C,D

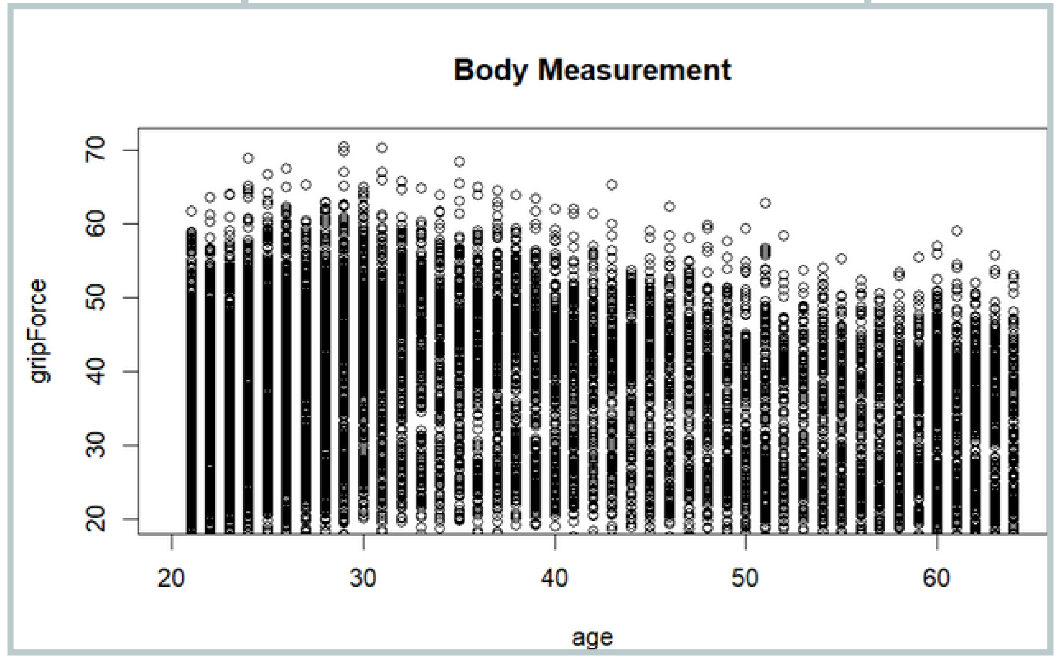
Body performance | from kaggle



The relationship between gender and class lable

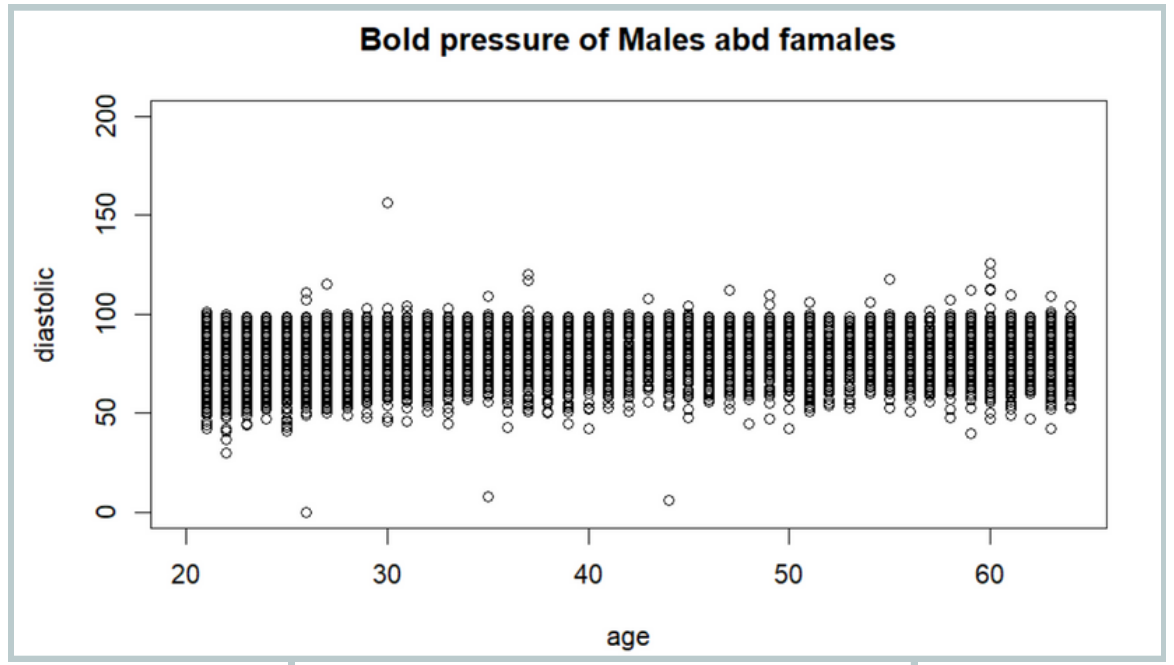


The frequency distribution of body fat values.

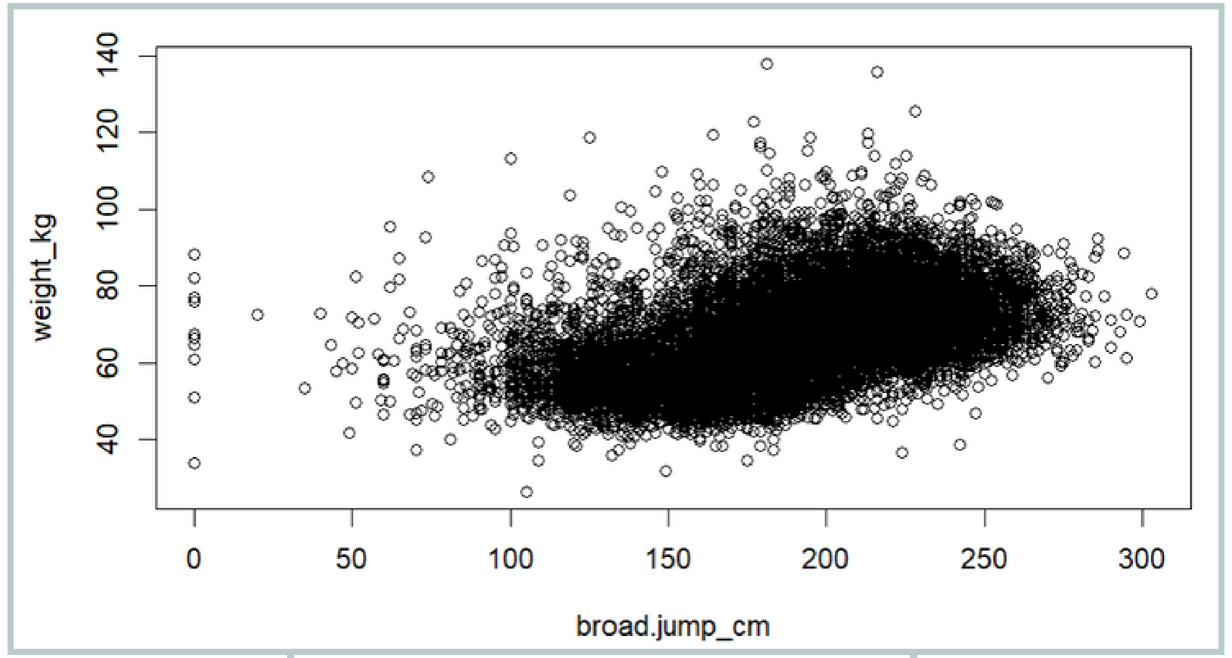


The relationship between grip force and age

correlation between Age and diastolic.



The relationship between bord jumb and weight(Kg)



03. DATA PREPROCESSING

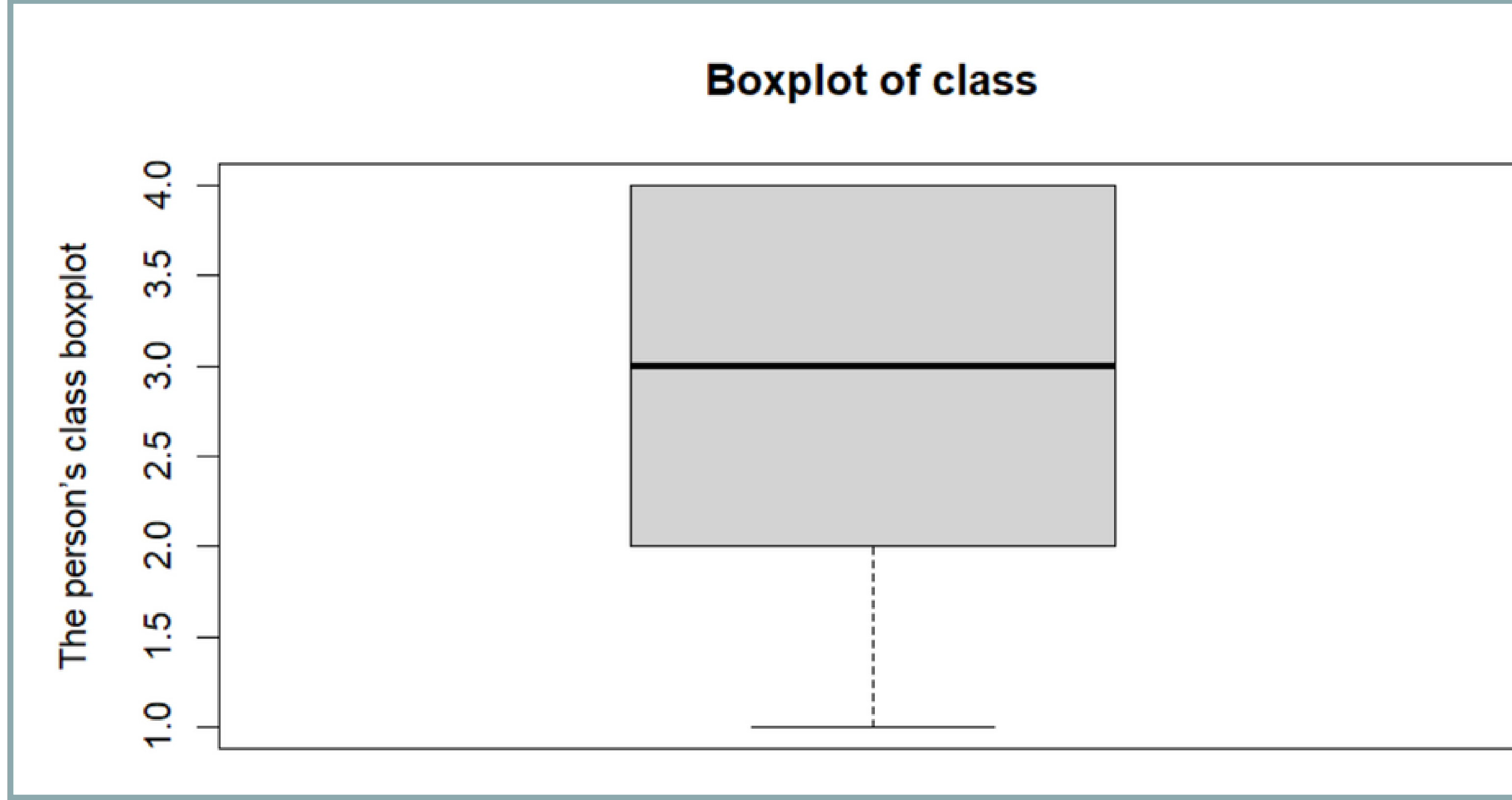
1- Delete nulls, duplicate values

We detect: 0 nulls, 1 duplicate

2- Detect and delete outlets

3- imbalance dataset problem

body performance dataset was balanced



Balance dataset

03. DATA PREPROCESSING

4- encoding some attributes (gender, class)

Gender to(0,1), class (1,2,3,4)

5- Normalization

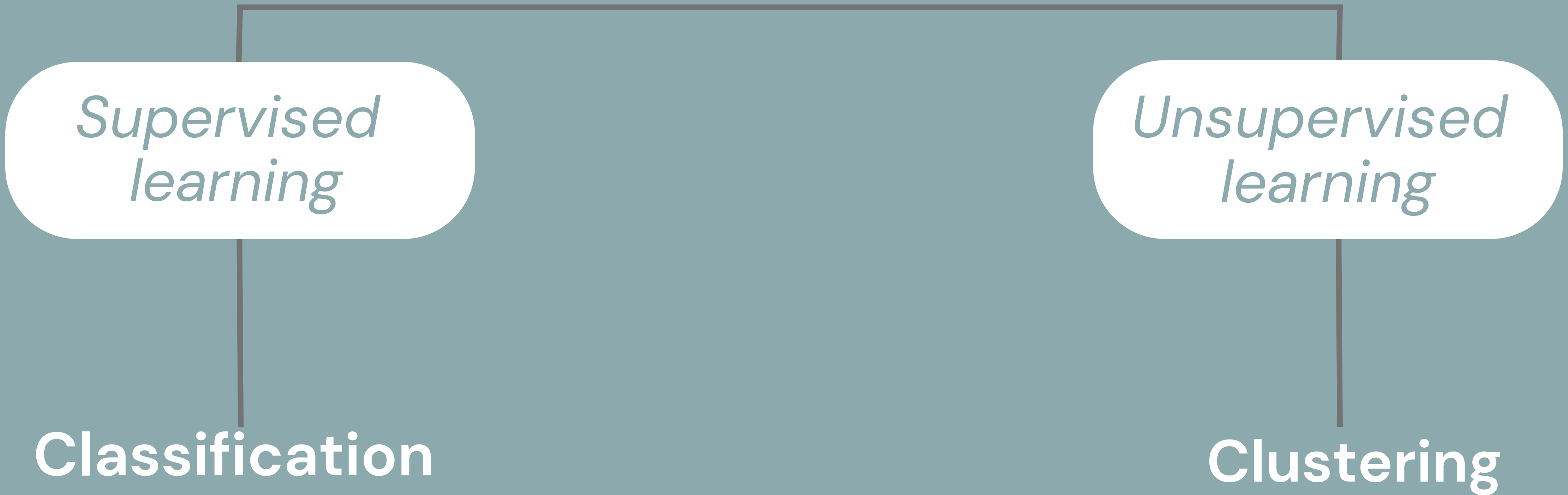
Fot the numeric attributes to let them have equal weight.

6-Discretization

We apply it on “age” (21, 35],(35, 49] and (49, 63] .

04. DATA MINING TASKS

We applied two data mining tasks to help us predict a person fitness level



Supervised learning

Classification

Unsupervised learning

Clustering

04.1 CLUSTERING

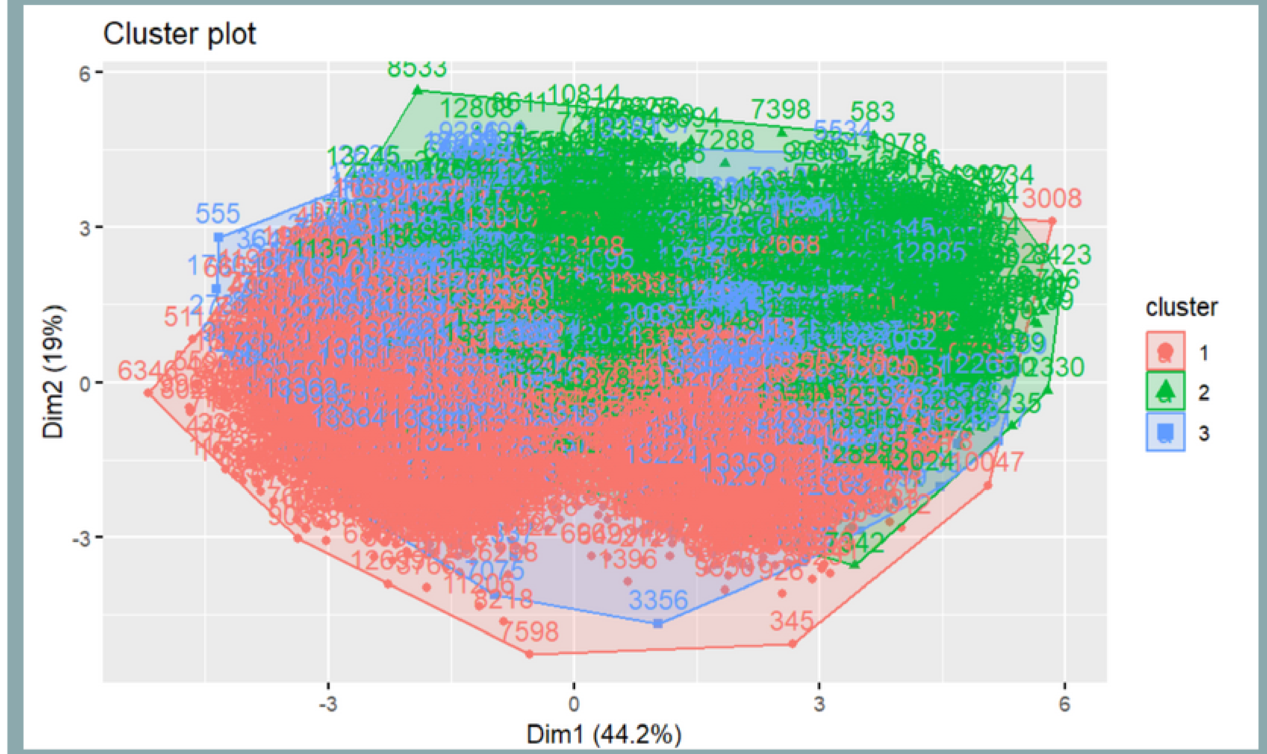
1- Preprocessing before Clustering:

Delete Class label, checking data type

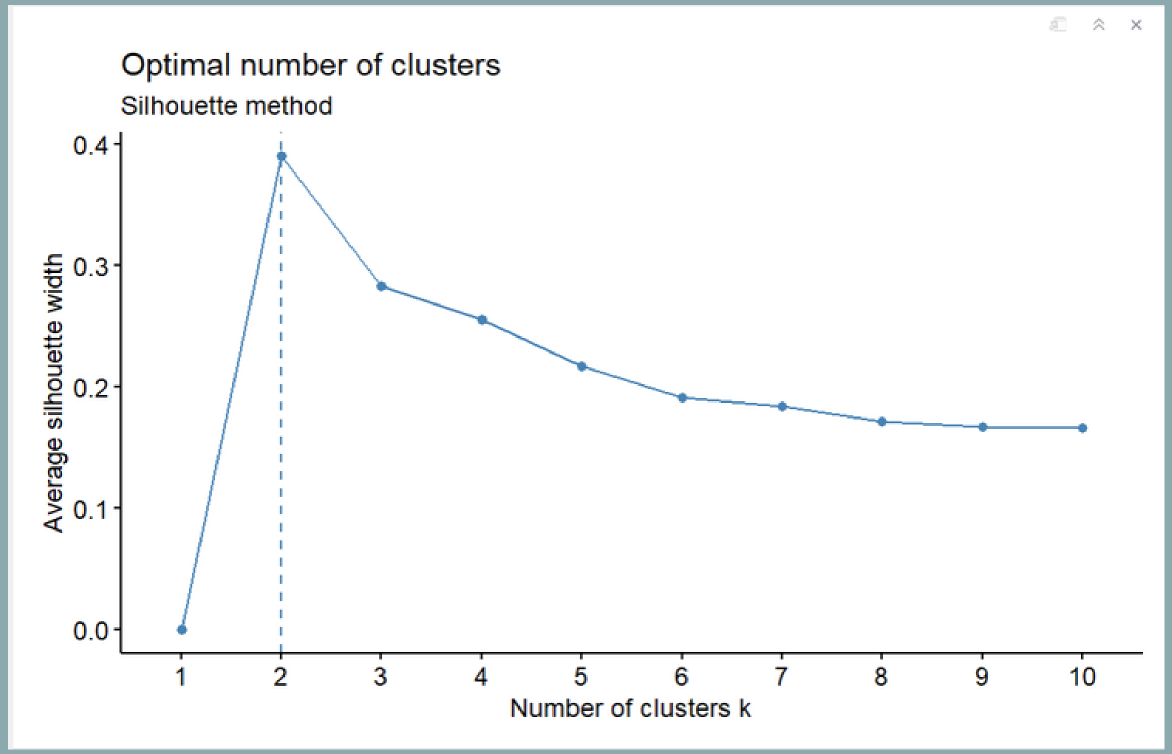
2- Determine optimal number of clusters :

Using 3 different methods:

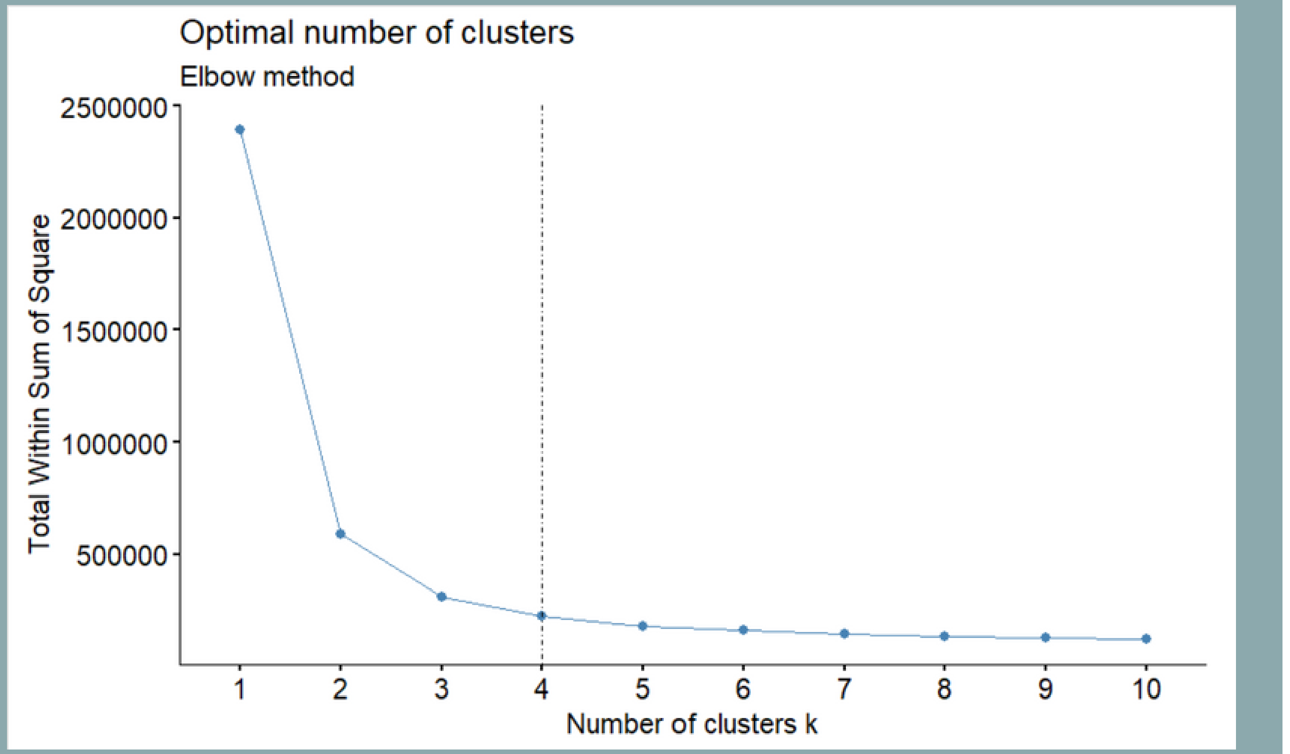
1- kmeansrun



2- Average silhouette method



3- Elbow method



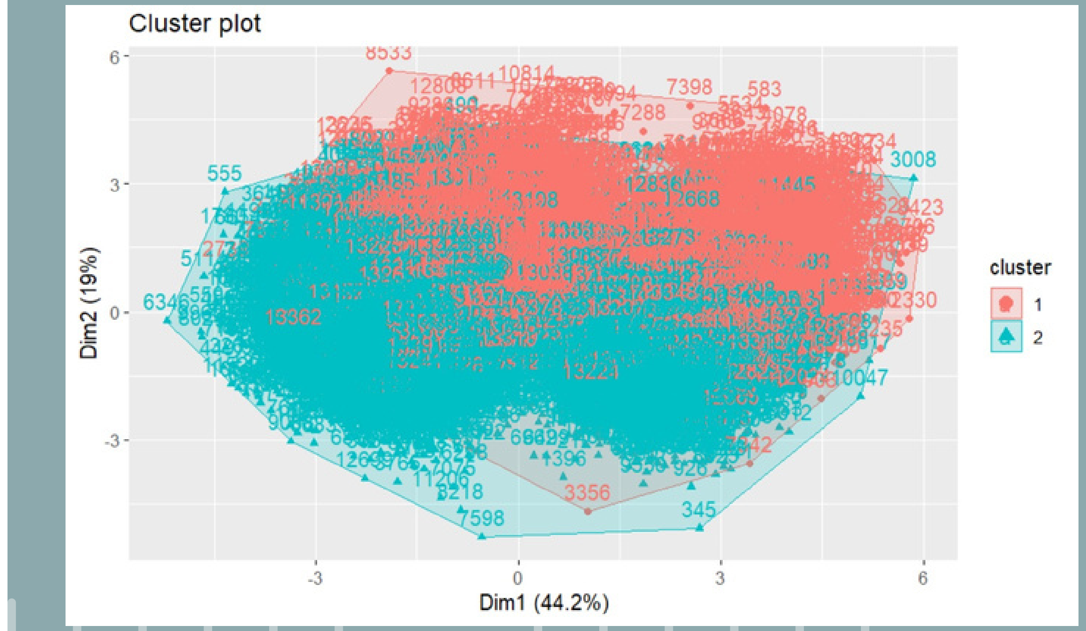
04.1 Clustering

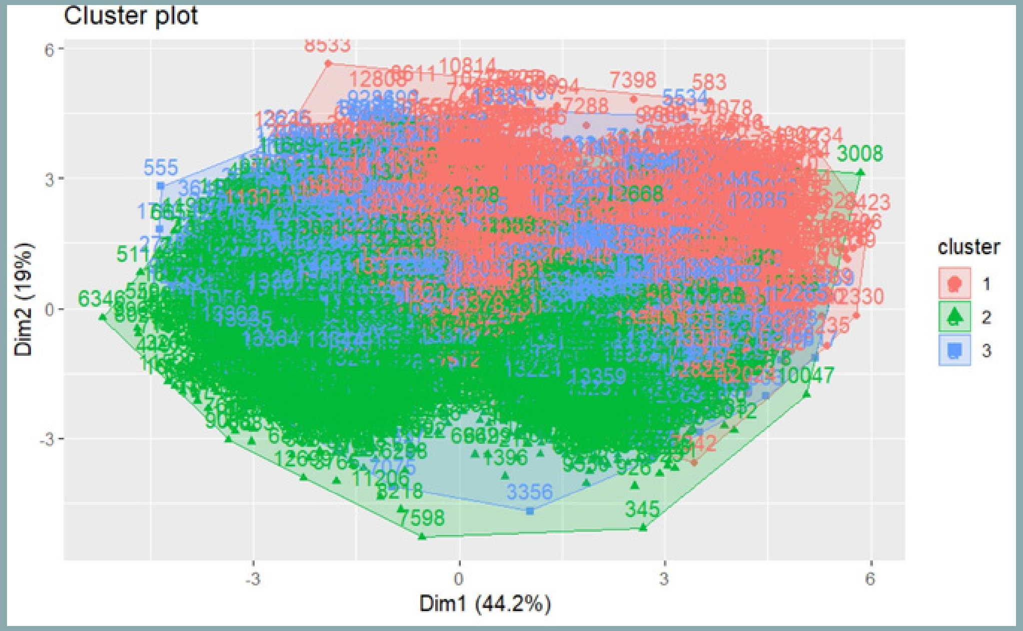
1. Partition data using k-mean algorithm

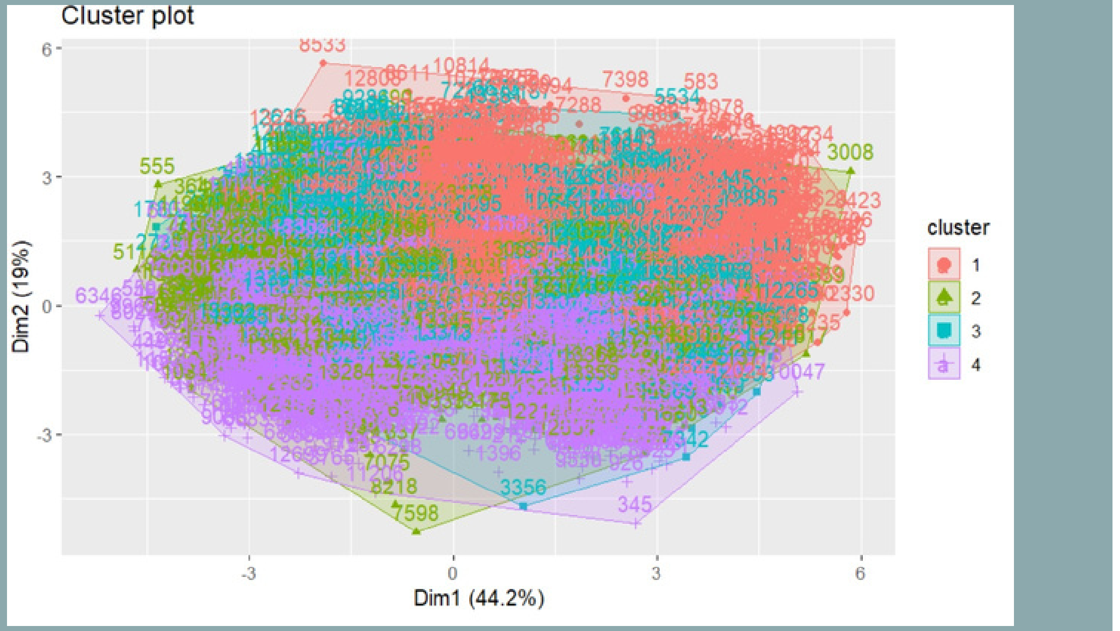
k=2

k=3

k=4







04.1 Clustering

1. Cluster evaluation:

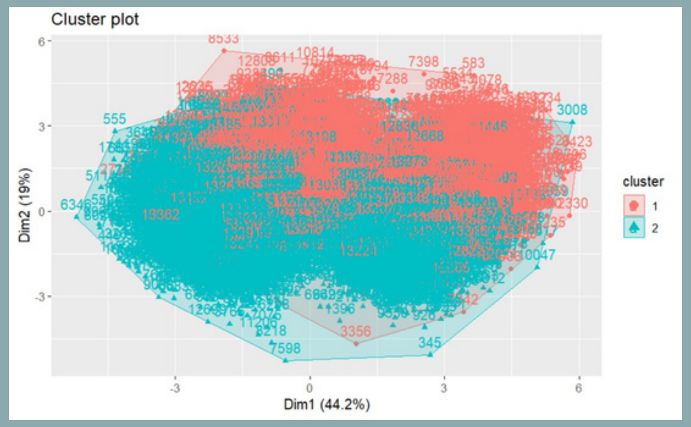
⬆Average Sillhouette width= 0.63

⬇Total within-cluster sum of square=5 91064.6

⬆BCubed precision=0.2516112

⬆BCubed recall=0.5454527

k=2



CLASSIFICATION

- Our classification objective is to develop a predictive model for our dataset with class label A, B, C, and D based on various attributes
- We divided the dataset into two groups: training set and testing set .
- We applied the classification methods on three different partitions :

70% training, 30% testing

75% training, 25% testing

80% training, 20% testing

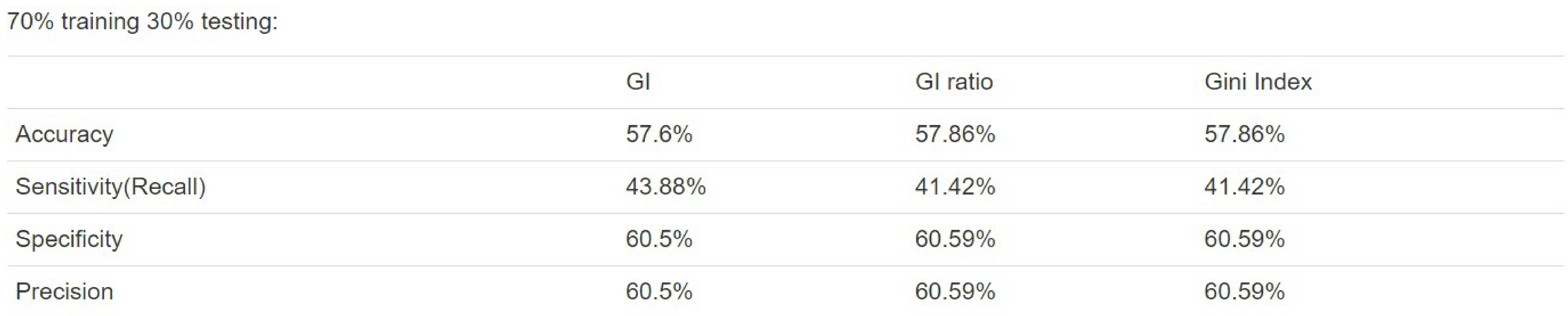
CLASSIFICATION

We applied three different methods for each partition:

- ID3(information gain)
- C4.5 (gain ratio)
- CART (gini index)

CLASSIFICATION

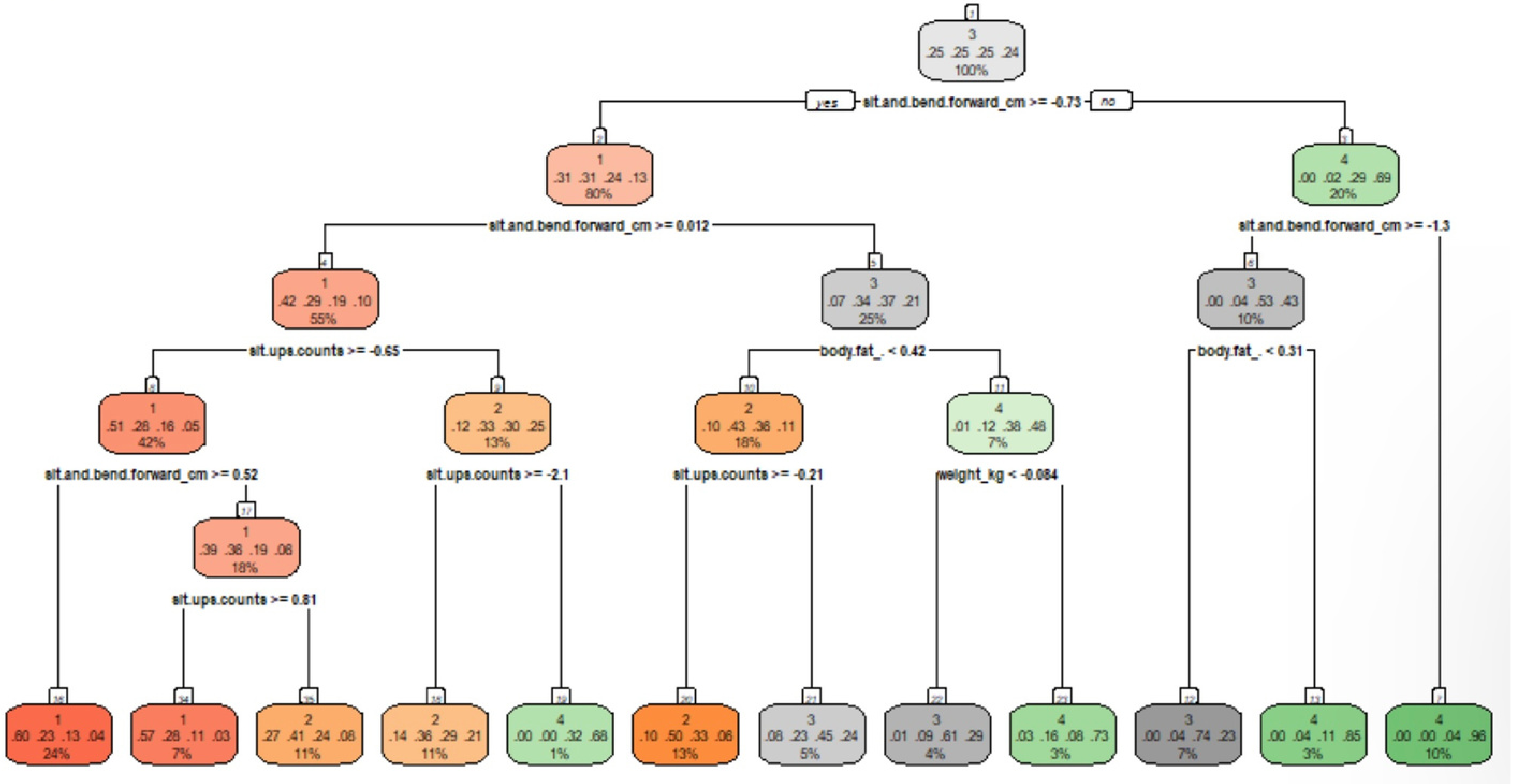
The 70/30 partition has the highest accuracy:



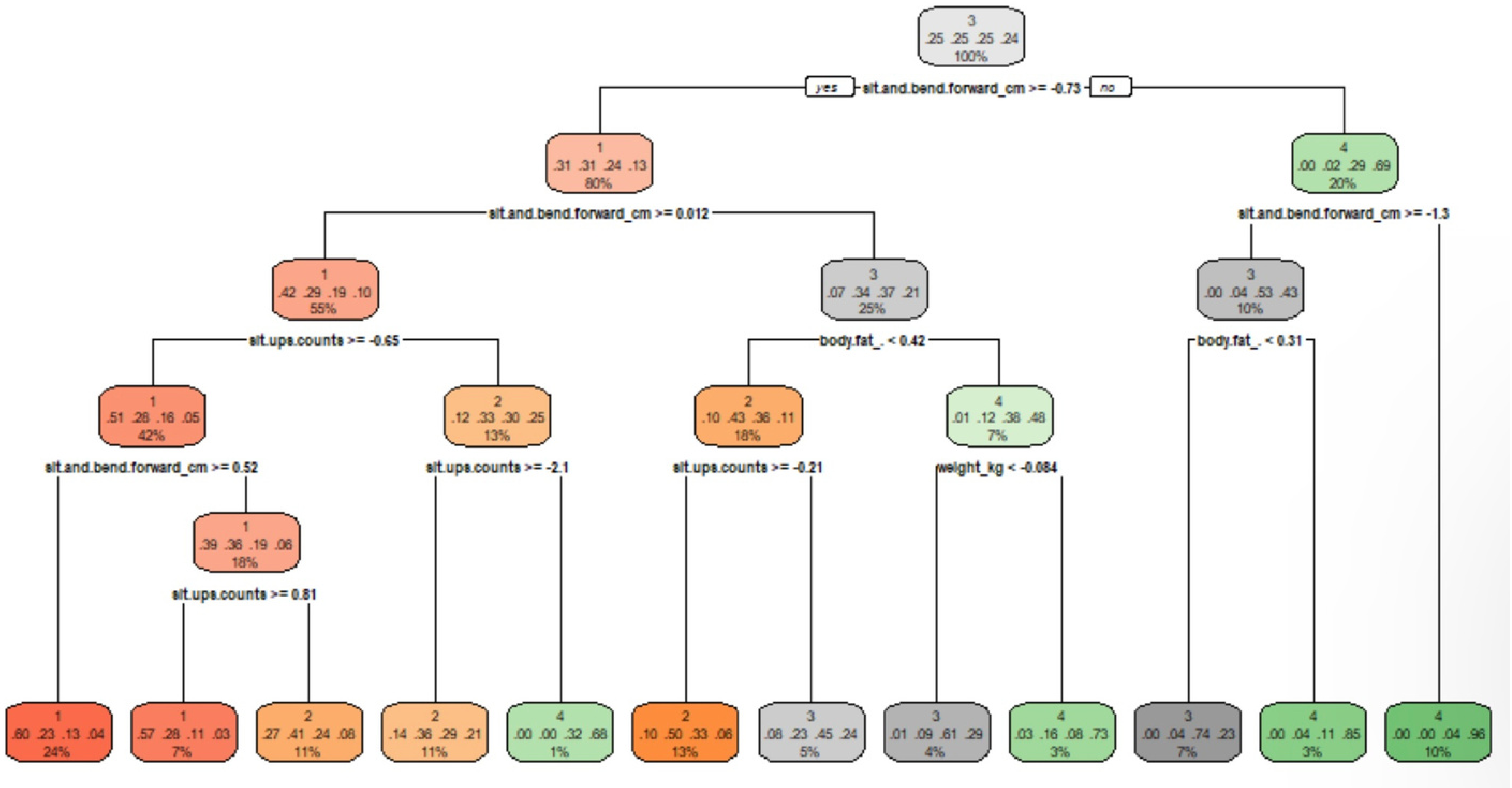
70% training 30% testing:

	GI	GI ratio	Gini Index
Accuracy	57.6%	57.86%	57.86%
Sensitivity(Recall)	43.88%	41.42%	41.42%
Specificity	60.5%	60.59%	60.59%
Precision	60.5%	60.59%	60.59%

CLASSIFICATION



CLASSIFICATION



FINDINGS

considering the similar accuracy achieved by both classification and clustering approaches, we made the decision to choose classification due to its direct suitability to our specific problem and its effective ability to accurately classify the data.

THANK YOU

REFRENCES:

[1] ["Body performance Data," Kaggle, Jun. 29, 2022.]
(https://www.kaggle.com/datasets/kukuroo3/body-performance-data)

[2].[Codeguyas, "Body Performance Data EDA," Kaggle, Dec. 19, 2021.]
(https://www.kaggle.com/code/codeguyas101/body-performance-data-eda)

[3]."discretize function - RDocumentation."]
(https://www.rdocumentation.org/packages/arules/versions/1.6-4/topics/discretize)

[4].["RPubs - Classification and Regression Trees (CART) in R."]
(https://rpubs.com/camguild/803096)

[5]. ["RPubs - Cluster Analysis in R."](https://rpubs.com/odenipinedo/cluster-analysis-in-R)

[6].[Shivanirana, "📌🧐Guide to Complete Statistical Analysis📊✅," Kaggle, Mar. 11, 2022.]
(https://www.kaggle.com/code/shivanirana63/guide-to-complete-statistical-analysis)

[overall].[Body performance note
book](https://rpubs.com/TeraPutera/LBB-CM-2)