עבודה מסכמת

תכנות מדעי בשפת פיתון

מגישה: עדי פלד
תעודת זהות: 204818678

חלק א: הקדמה

1. הנושא

הדאטאסט מציג מדדי גוף של אנשים לצד תוצאות תרגילי ספורט שלהם, ולצד זה מציג סיווג שלהם לרמות ביצועים A B C D , כך שA היא הרמה הטובה ביותר.

רמת הביצועים של בן אדם נקבעת ע"י מדדי גוף ותוצאות של תרגילי ספורט:

- מדדי הגוף: גובה, משקל, אחוז שומן בגוף, לחץ דם סיסטולי, לחץ דם דיאסטולי.

התרגילים: כח אחיזה, ישיבה והתמתחות קדימה, כפיפות בטן, קפיצה לרוחק.

תוצאות תרגילי הספורט נמדדות למשל בכמות חזרות\מרחק בס"מ בהתאמה לתרגיל.

2. פירוט על הפיצ'רים וסוגיהם

שם פיצ'ר	מידע על הפיצ'ר	סוג
age		נומרי , float
gender	ערכים: M,F	קטגורי, נומינלי
height_cm	נמדד בס"מ	נומרי , float
weight_kg	נמדד בק"ג	נומרי , float
body fat_%	נמדד באחוזים	נומרי , float
diastolic	לחץ דם דיאסטולי הוא הלחץ הנמוך ביותר בזמן הרפיית חדרי הלב.	נומרי , float
systolic	לחץ דם סיסטולי הוא לחץ השיא בזמן התכווצות חדרי הלב.	נומרי , float
gripForce	גודל כח האחיזה	נומרי , float
sit and bend forward_cm	בעת ישיבה על הרצפה ויישור הרגליים קדימה, כמה אתה מגיע קדימה עם הידיים. כפות הרגליים מהוות את קו ה0. אם לא עברת אותם תהיה בערך שלילי, עברת- ערך חיובי.	נומרי , float
sit-ups counts	כמות כפיפות בטן	נומרי , float
broad jump_cm	קפיצה לרוחק, המרחק נמדד בס"מ	נומרי , float
class	ערכים : A,B,C,D – A הכי טוב	קטגורי, אורדינלי

3. מבט ראשוני על נתוני הדאטאסט

	age	gender	height_cm	weight_kg	body fat_%	diastolic	systolic	gripForce	sit and bend forward_cm	sit-ups counts	broad jump_cm	class
0	27.0	M	172.3	75.24	21.3	80.0	130.0	54.9	18.4	60.0	217.0	C
1	25.0	M	165.0	55.80	15.7	77.0	126.0	36.4	16.3	53.0	229.0	A
2	31.0	M	179.6	78.00	20.1	92.0	152.0	44.8	12.0	49.0	181.0	C
3	32.0	M	174.5	71.10	18.4	76.0	147.0	41.4	15.2	53.0	219.0	B
4	28.0	M	173.8	67.70	17.1	70.0	127.0	43.5	27.1	45.0	217.0	B

4. גודל הדאטאסט

שורות: 13,393

עמודות: 12

5. כמות חוסרים

- לא קיימים ערכי null בכל העמודות.
- בעמודות: systolic, diastolic, gripForce, broad jump, sit-ups, קיימות דגימות המכילות את הערך 0, בדקתי האם יש הגיון במדידה כזו בעמודות אלה ונראה כי אלו ערכים חסרים, בהם אטפל בשלב העיבוד המקדים.

6. מקורות מידע

- מהו gripForce
- איך מעריכים את המדד 'sit and bend forward'
- ההבדל בין לחץ דם סיסטולי לדיאסטולי

חלק ב: ניתוח נתונים ראשוני

1. ניתוח סטטיסטי של הפיצ'רים

עבור פיצ'רים נומריים:

	count	mean	std	min	25%	50%	75%	max
age	13393.0	36.775106	13.625639	21.0	25.0	32.0	48.0	64.0
height_cm	13393.0	168.559807	8.426583	125.0	162.4	169.2	174.8	193.8
weight_kg	13393.0	67.447316	11.949666	26.3	58.2	67.4	75.3	138.1
body fat_%	13393.0	23.240165	7.256844	3.0	18.0	22.8	28.0	78.4
diastolic	13393.0	78.796842	10.742033	0.0	71.0	79.0	86.0	156.2
systolic	13393.0	130.234817	14.713954	0.0	120.0	130.0	141.0	201.0
gripForce	13393.0	36.963877	10.624864	0.0	27.5	37.9	45.2	70.5
sit and bend forward_cm	13393.0	15.209268	8.456677	-25.0	10.9	16.2	20.7	213.0
sit-ups counts	13393.0	39.771224	14.276698	0.0	30.0	41.0	50.0	80.0
broad jump_cm	13393.0	190.129627	39.868000	0.0	162.0	193.0	221.0	303.0

עבור פיצ'רים קטגוריאליים:

	count	unique	top	freq
gender	13393	2	M	8467
class	13393	4	C	3349

מסקנות :

- Age:
  - גיל ממוצע 36 , והחציון 32
  - טווח הגילאים: 21-64
- Gender:
  - רוב האנשים הם זכרים ומספרם 8467 , שזה בערך 63% מהדאטאסט.
- Height_cm
  - גובה ממוצע 168 ס"מ, והחציון 169.
  - טווח הגבהים: 125-193- מאוד נמוך ומאוד גבוה.
- Weight_cm:
  - משקל ממוצע 67, והחציון 67 גם.
  - טווח המשקלים: 26.3-138 , מאוד נמוך בטווחי גילאים אלו ומאוד גבוה.
- Body_fat:
  - אחוז שומן ממוצע הוא 23 והחציון גם כמעט 23.
  - טווח אחוזי השומן: 3-78, מאוד נמוך ומאוד גבוה- נראה אף לא הגיוני.

- diastolic:
  - לחץ דם דיאסטולי ממוצע הוא 78.7 וגם החציון קרוב 79.
  - טווח אחוזי לחץ דם זה: 0-156, הטווח נראה לא הגיוני, 156 זה מאוד גבוה, ו-0 זה לא הגיוני מבחינה רפואית.
- systolic:
  - לחץ דם סיסטולי ממוצע הוא 130 וגם החציון 130.
  - טווח אחוזי לחץ דם זה: 0-201 , הטווח נראה לא הגיוני, 200 זה מאוד גבוה, ו0 זה ערך לא הגיוני מבחינה רפואית.
- gripForce
  - תוצאה ממוצעת לתרגיל זה : 37 והחציון 38.
  - טווח התוצאות: 0-70.5, 0 נראה ערך לא הגיוני בתרגיל זה.
- Sit and bend forward
  - תוצאה ממוצעת לתרגיל זה 15 ס"מ.
  - טווח התוצאות: 213 – 25-
- Sit-ups counts
  - תוצאה ממוצעת לתרגיל זה 39.7 והחציון 41.
  - טווח התוצאות: 0-80, 0 נראה ערך לא הגיוני לתרגיל זה.
- Broad jump
  - תוצאה ממוצעת לתרגיל זה 190 ס"מ והחציון 193.
  - טווח התוצאות: 0-303, 0 נראה ערך לא הגיוני לתרגיל זה.
- class
  - הקלאס שמופיעה הכי הרבה זו C (ב1 יותר מהשאר כך שאין לזה משמעות)

מסקנה גורפת:

לאחר בדיקה, אין משמעות לערך מינימום 0 בפיצ'רים בהם קיימים ערכים כאלו ולכן אבצע תיקון זה בהמשך בעמודות:

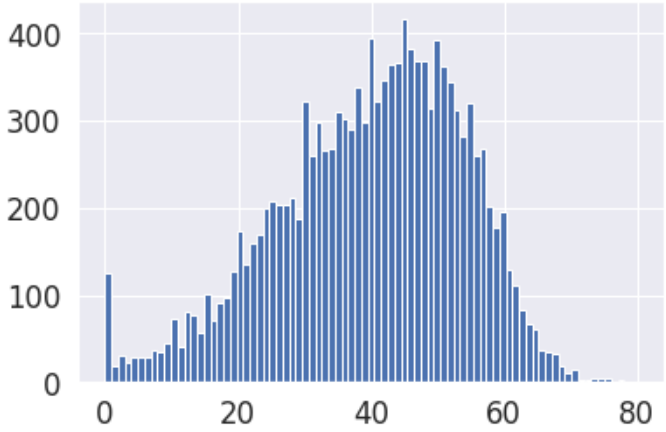
- Diastolic
- Systolic
- gripForce
- Sit-ups counts
- Broad jump

2. סיכום תיקוני הנתונים

א. שינוי סוג העמודות שמסוג object לסוג Categorical .

ב. בדיקה ושינוי משמעות הערך 0 בעמודות systolic,gripForce, sit ups counts ,broad jump ו diastolic

- עבור עמודות לחץ הדם systolic וdiastolic: קיימת מדידה אחת שבה לחץ הדם בשני סוגיו הוא 0 ולכן אשלים ערכים אלו לחציון.
- עבור gripforce ו broad jump: הערך 0 מופיע מעט מאוד פעמים(3 ו10 בהתאמה) ולכן השלמה שלהם תשפיע בצורה מזערית על התפלגות הערכים בפיצ'רים האלה.
- עבור sit-ups counts קצת פחות מ1% מהדאטא מכיל את הערך 0, עם זאת התפלגות ערכי הפיצ'ר מצביעה על חריגה, ולכן אשלים גם ערכים אלו לחציון.



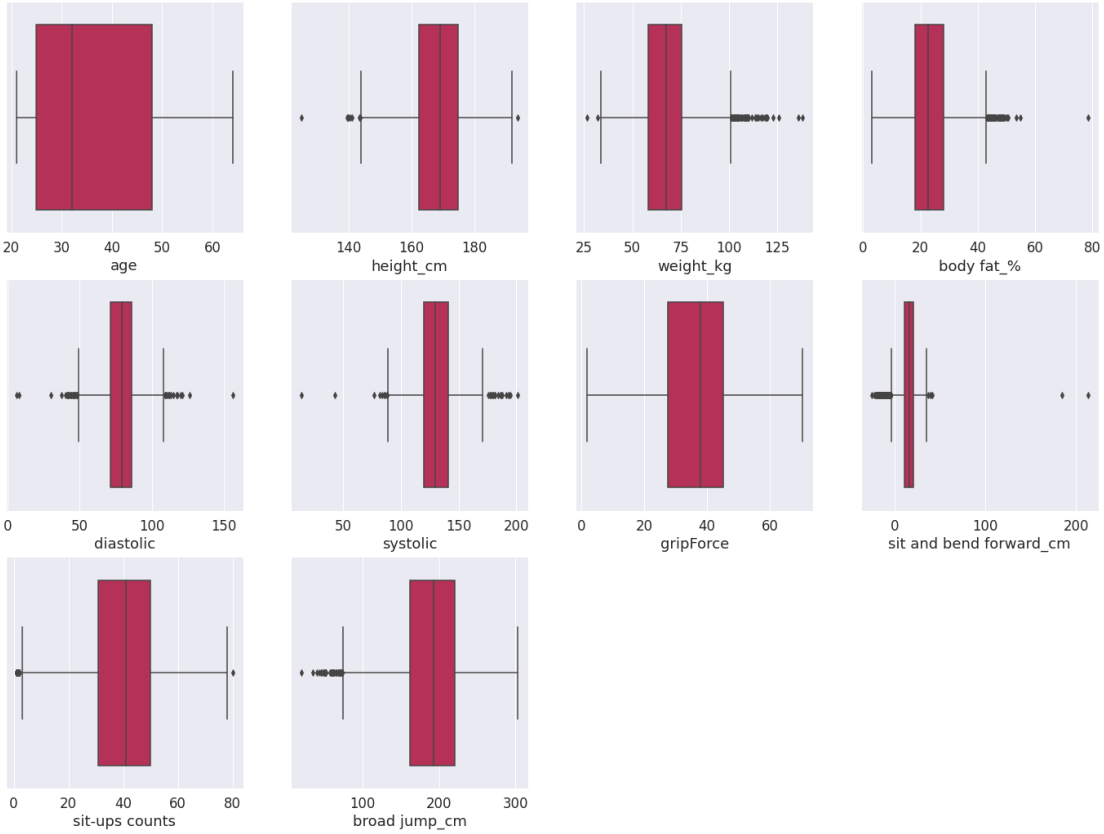
השינוי שביצעתי:
בכל מקום בעמודות אלו בהם הופיע הערך 0, הוא הוחלף בערך החציון של אותה עמודה. ההחלטה להחליף בחציון ולא בממוצע נבעה מ2 סיבות, ערכי הממוצע והחציון בכל אחד מהפיצ'רים היו קרובים מאוד (ביחס לקנה המידה של כלל הערכים) , וכיוון שלא בדקתי באופן מדויק מאוד האם הערכים מתפלגים נורמלית(לא בדקתי כי זה דורש שימוש באמצעים שלא נלמד).

ג. המרה מfloat לint לעמודת שאין להן באמת שבר לאחר הנקודה אלא המספר בצורה של num.0 לעמודות: diastolic, systolic, broad jump , sit-ups counts.

ד. הסרת blatant outliers מהפיצ'רים:

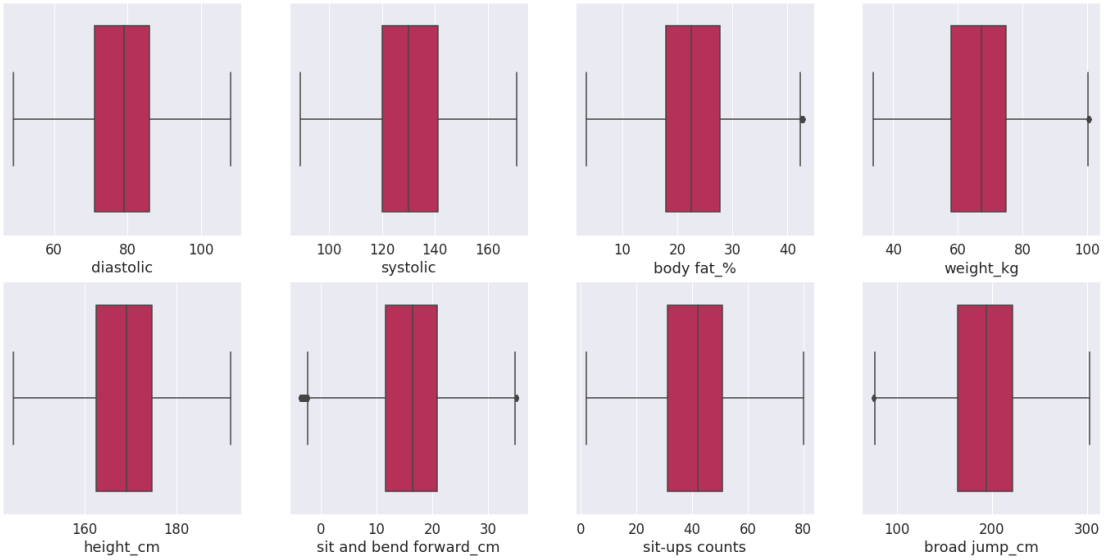
שלבים:

- בדיקה האם יש outliers ואם כן לאיזה פיצ'רים:



- לפיצ'רים: age ו gripForce אין outliers לכן נוריד אותם מהרשימה של הפיצ'רים שנוריד להם את הoutliers.

- המצב לאחר הסרת ה outliers מהפיצ'רים שהגדרתי, עבור כל המחלקות:



מסקנות לאחר הורדת הoutliers מכל המחלקות לפיצ'רים נבחרים:

נתונים סטטיסטים לפני הסרה:

	age	height_cm	weight_kg	body fat_%	diastolic	systolic	gripForce	sit and bend forward_cm	sit-ups counts	broad jump_cm
count	13393.000000	13393.000000	13393.000000	13393.000000	13393.000000	13393.000000	13393.000000	13393.000000	13393.000000	13393.000000
mean	36.775106	168.559807	67.447316	23.240165	78.802583	130.244456	36.972157	15.222896	40.135145	190.271485
std	13.625639	8.426583	11.949666	7.256844	10.720358	14.671247	10.610447	8.444402	13.745460	39.527754
min	21.000000	125.000000	26.300000	3.000000	6.000000	14.000000	1.600000	-25.000000	1.000000	20.000000
25%	25.000000	162.400000	58.200000	18.000000	71.000000	120.000000	27.500000	10.900000	31.000000	162.000000
50%	32.000000	169.200000	67.400000	22.800000	79.000000	130.000000	37.900000	16.200000	41.000000	193.000000
75%	48.000000	174.800000	75.300000	28.000000	86.000000	141.000000	45.200000	20.700000	50.000000	221.000000
max	64.000000	193.800000	138.100000	78.400000	156.000000	201.000000	70.500000	213.000000	80.000000	303.000000

נתונים סטטיסטים לאחר הסרה:

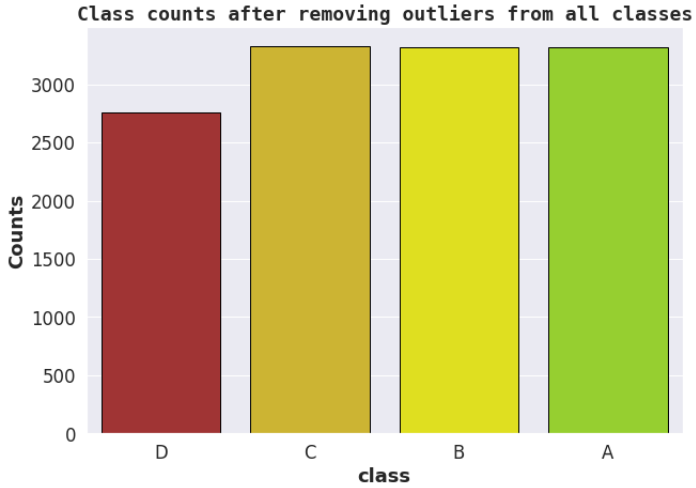
	age	height_cm	weight_kg	body fat_%	diastolic	systolic	gripForce	sit and bend forward_cm	sit-ups counts	broad jump_cm
count	12717.000000	12717.000000	12717.000000	12717.000000	12717.000000	12717.000000	12717.000000	12717.000000	12717.000000	12717.000000
mean	36.662578	168.524951	67.021409	22.937001	78.782889	130.189825	37.070916	16.007798	40.611937	191.493670
std	13.569631	8.280444	11.403543	6.975733	10.424716	14.361776	10.574119	6.961553	13.532980	38.604432
min	21.000000	144.500000	33.700000	3.500000	49.000000	89.000000	1.600000	-3.700000	2.000000	75.000000
25%	25.000000	162.500000	58.000000	17.900000	71.000000	120.000000	27.600000	11.600000	31.000000	164.000000
50%	32.000000	169.100000	67.200000	22.600000	79.000000	130.000000	38.000000	16.500000	42.000000	194.000000
75%	48.000000	174.600000	74.900000	27.700000	86.000000	141.000000	45.300000	20.900000	51.000000	222.000000
max	64.000000	191.800000	100.800000	42.900000	108.000000	171.000000	70.500000	35.200000	80.000000	303.000000

- המינימום גדל והמקסימום קטן ברוב הפיצ'רים, וכן סטיית התקן קטנה בכולם.
- עדין נשארו מעט outliers, הסיבה לכך היא שביצעתי מעבר על כל פיצ'ר ומיד לאחריו הוסר הדאטא שהוא outliers, ובכל פעם האיזון בעמודות משתנה כשאני מורידה שורות בגלל outliers מעמודה מסוימת.
- כמות השורות שהורדו: 675
- כמות הערכים שהורדו: בדקתי את האיזון מבחינה מספרית בכל מחלקה לפני ההורדה ולאחר ההורדה:

לפני: כמות שווה (כמעט) בין כל המחלקות

```
class
D    3349
C    3349
B    3347
A    3348
```

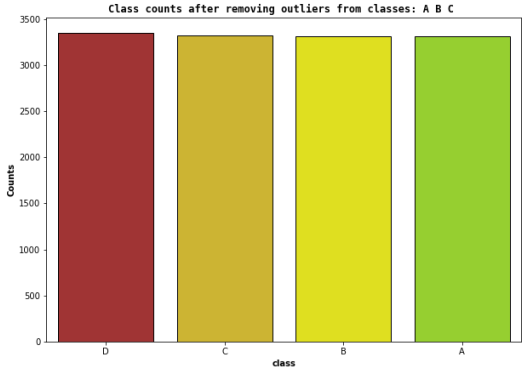
אחרי: ירידה מועטה במחלקות A B C וירידה משמעותית במחלקה D.




```
class
D    2756
C    3328
B    3315
A    3318
```

- לאחר השוואת כמות האנשים במחלקה D לפני ואחרי:
  - ניתן לראות ש90% מהoutliers היו שייכים למחלקה D.
  - ניתן להסיק כי אלו אינם ערכים חריגים, אלא סממנים המתאימים לקטגוריה D.
  - ההחלטה שלי היא לא לנקות את הoutliers שקשורים למחלקה D, אלא רק מהמחלקות האחרות.

- התפלגות המחלקות לאחר הורדת חריגים ממחלקות A B C בלבד:




```
class
D    3349
C    3328
B    3315
A    3317
```

ניתן לראות כי התפלגות המחלקות מאוד אחידה בקירוב.
סה"כ שורות שנותרו לאחר סעיף זה: 13309.

ה. <u>הוספת עמודת bmi מספרית והוספת עמודה bmi_cat קטגורית לפי המספרית</u>
עמודת BMI משקללת משקל וגובה, וכך אפשר להבין מה תקינות המשקל שלך , עיגול התוצאות עד 2 ספרות לאחר הנקודה.
הנוסחא שעל בסיסה אצור את עמודת הbmi: $weight(kg)/ height^2(meter)$

עמודה bmi קטגורית: הקטגוריות-תת משקל, משקל תקין, עודף משקל.
חלוקה ל3 טווחים (על בסיס מידע מהימן)
- מתחת ל־18.5 מוגדר כתת משקל
- 18.5-25 מוגדר תקין
- 25-29.9 מוגדר עודף משקל

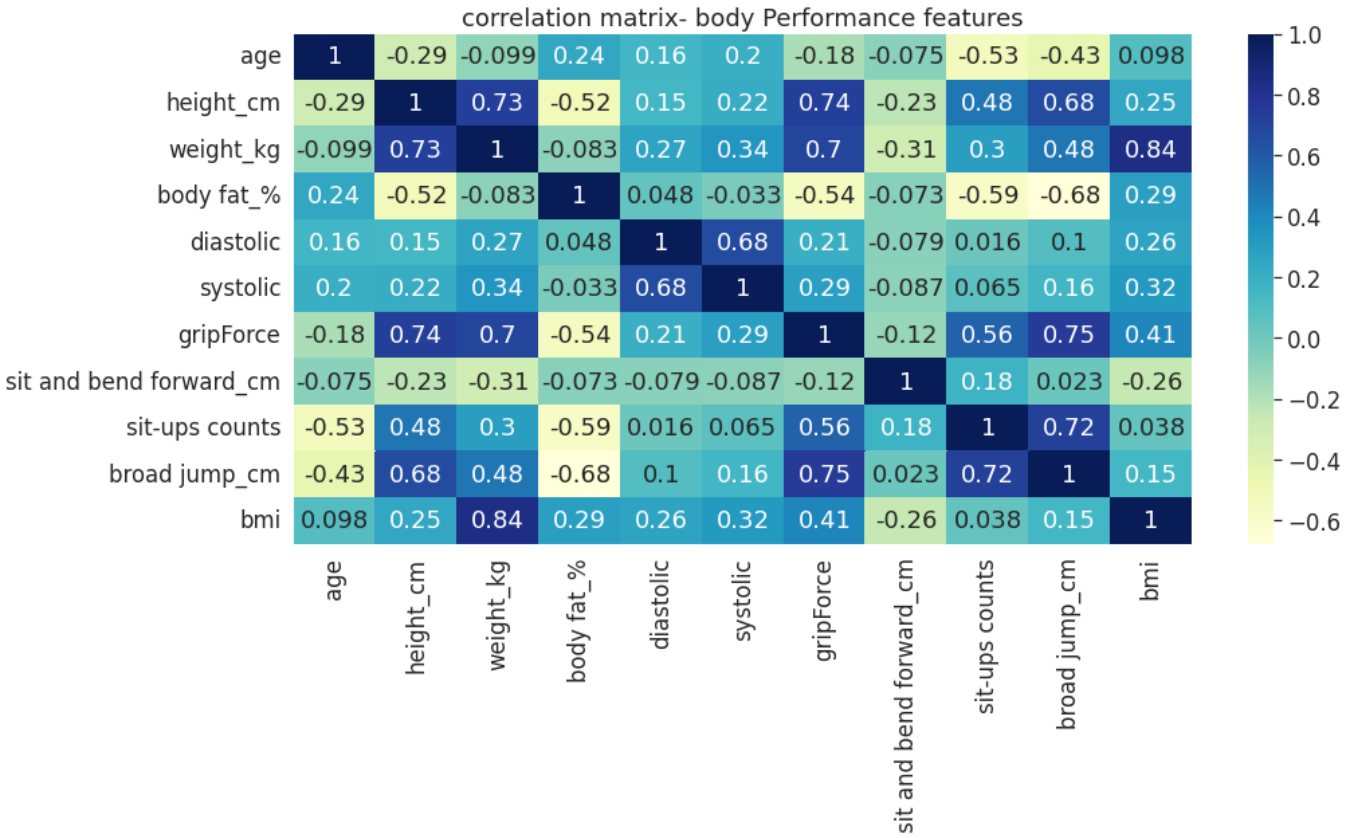
📎קובץ דאטא מעובד –מצורף להגשה: bodyPerformance_manipulated_after_section_2.csv

חלק ג: ניתוח נתונים באופן מחקרי

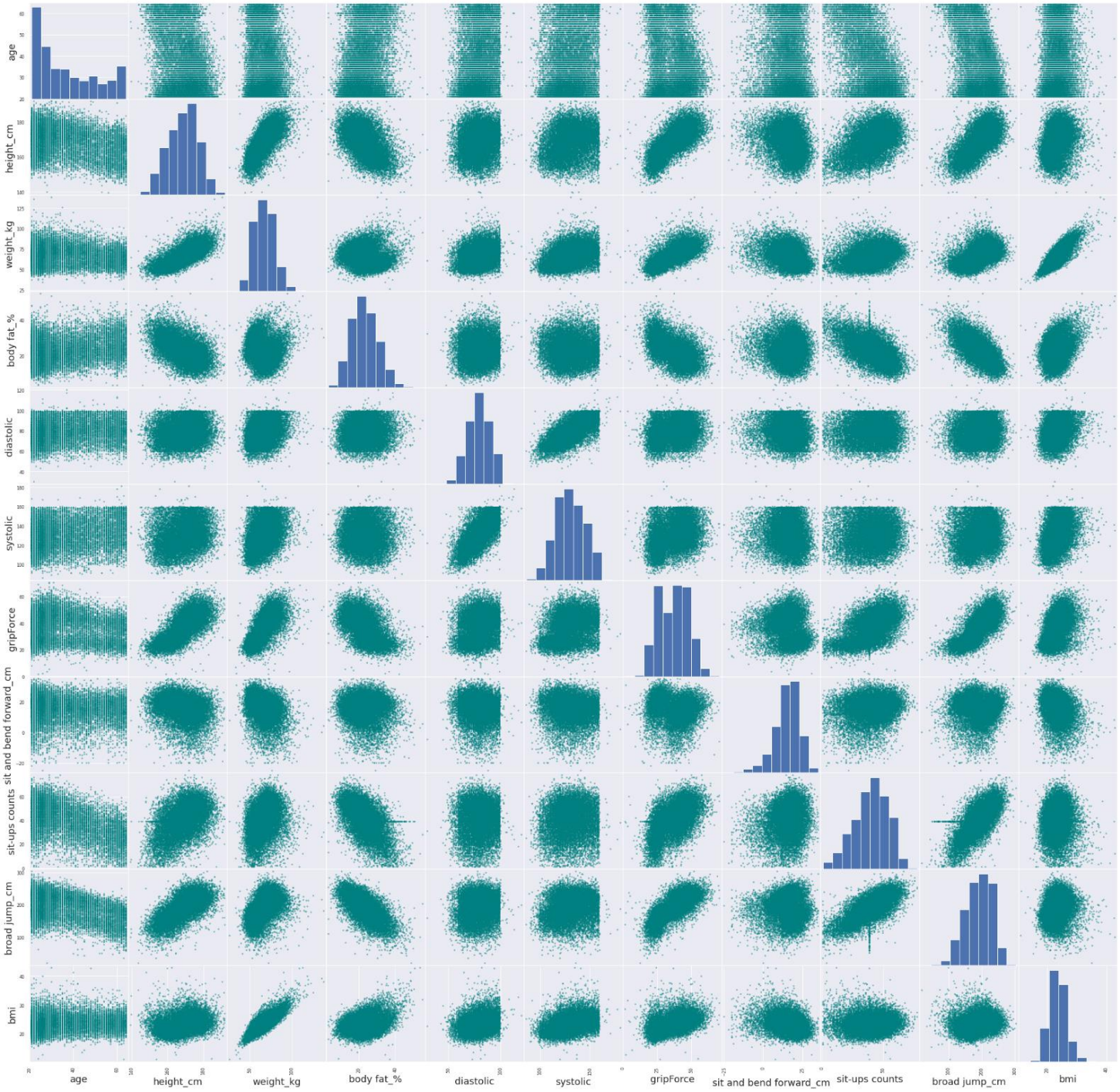
1. קורלציית פיצ'רים

א. מטריצת קורלציה לצד scatter matrix

בחרתי להראות לצד מטריצת הקורלציה את scatter_matrix כדי להציג באופן ויזואלי את פיזור ערכי זוגות הפיצ'רים האפשריים.



scatter_matrix- body Performance features



מסקנות ממטריצת הקורלציה ומscatter matrix:

א. weight X height
- קורלציה לינארית חיובית, מגמת עליה
- ככל שאתה גבוה המשקל שלך גבוה.
- במטריצת הקורלציה: 0.73

ב. systolic X diastolic
- קורלציה לינארית חיובית, מגמת עליה
- ככל שהלחץ הדם הסיסטולי גבוה, כך הלחץ דם הדיאסטולי גבוה.
- במטריצת הקורלציה: 0.68

ג. gripForce X height
- קורלציה לינארית חיובית, מגמת עליה
- ככל שהגובה גבוה כך כח האחיזה חזק.
- במטריצת הקורלציה: 0.74

ד. gripForce X weight
- קורלציה לינארית חיובית, מגמת עליה
- ככל שהמשקל גבוה כך כח האחיזה חזק.
- במטריצת הקורלציה: 0.7

ה. gripForce X broad jump
- קורלציה לינארית חיובית, מגמת עליה
- ככל שכח האחיזה חזק, כך הקפיצה לרוחק רחוקה יותר.
- במטריצת הקורלציה: 0.75

ו. situps X broad jump
- קורלציה לינארית חיובית, מגמת עליה
- ככל שאתה עושה יותר כפיפות בטן, כך הקפיצה לרוחק שלך יותר רחוקה.
- במטריצת הקורלציה: 0.72

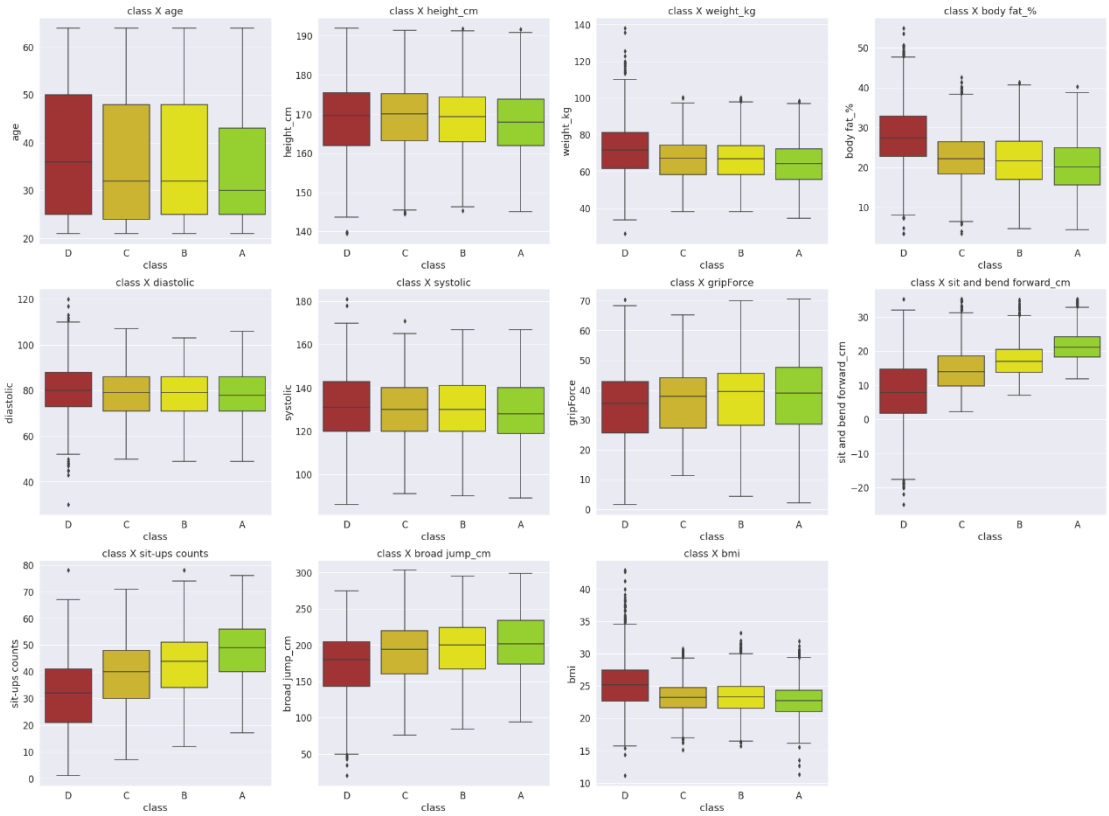
ז. height X broad jump
- קורלציה לינארית חיובית, מגמת עליה
- ככל שאתה יותר גבוה כך קפיצתך רחוקה יותר.
- במטריצת הקורלציה: 0.68

ח. Weight X bmi
- קורלציה לינארית חיובית חזקה, מגמת עליה
- ככל שיש לך משקל גבוה, כך הbmi שלך גבוה- זו תוצאה צפויה אם מתייחסים לנוסחא לחישוב bmi.
- במטריצת הקורלציה: **0.84**
- הקשר חזק מאוד באופן יחסי לשאר, אתייחס לכך בהמשך מבחינת טיפול.

ט. bodyfat X broad jump
- קורלציה לינארית שלילית, מגמת ירידה
- ככל שאחוז השומן שלך נמוך כך תקפוץ רחוק יותר.
- במטריצת הקורלציה: -0.68

ב. קשר פיצרים קטגורים לנומריים

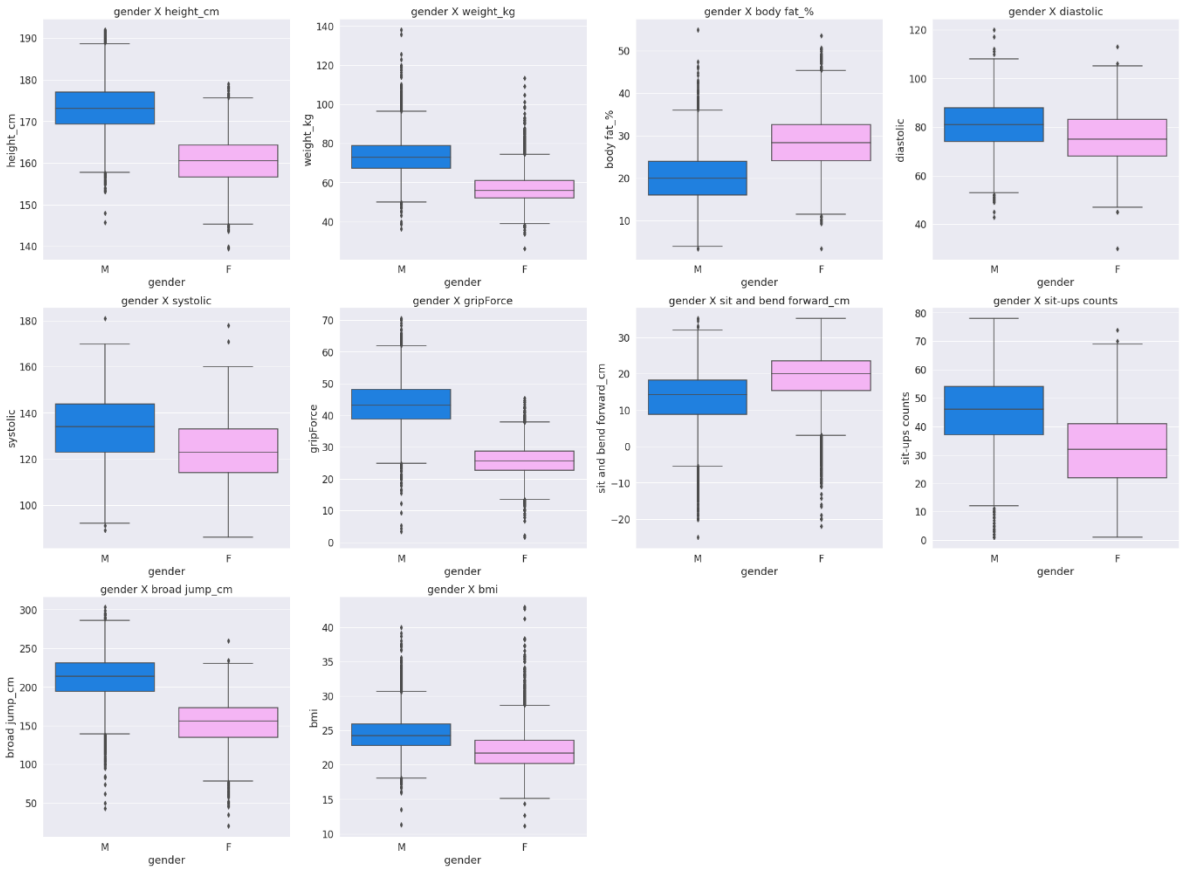
פיזור ערכי הפיצ'רים הנומריים כתלות בclass



מסקנות:

קיימים פיצ'רים כמו age ו systolic שנראה כי פיזור הערכים על פני הקלאסים אחיד יחסית.
לעומת זאת, ניתן לראות בפיצ'רים sit and bend forward ו sit-ups counts כי קיימת מגמה עליה .
נראה כי לאף אחד מהפיצ'רים אין יכולת הבחנה בין הקלאסים בכוחות עצמו בלבד, ולכן כצפוי נרצה להשתמש ביותר מאחד מהם על מנת להפיק prediction טוב.
בנוסף, ניתן לראות כי בכל המחלקות יש פיזור על טווח גילאים דומה, אך מחלקה A מתאפיינת בכך שגילאים מבוגרים (סביב 45+) אינם מופיעים בה , בעוד שבשאר המחלקות כן.

פיזור ערכי הפיצ'רים הנומריים כתלות בgender



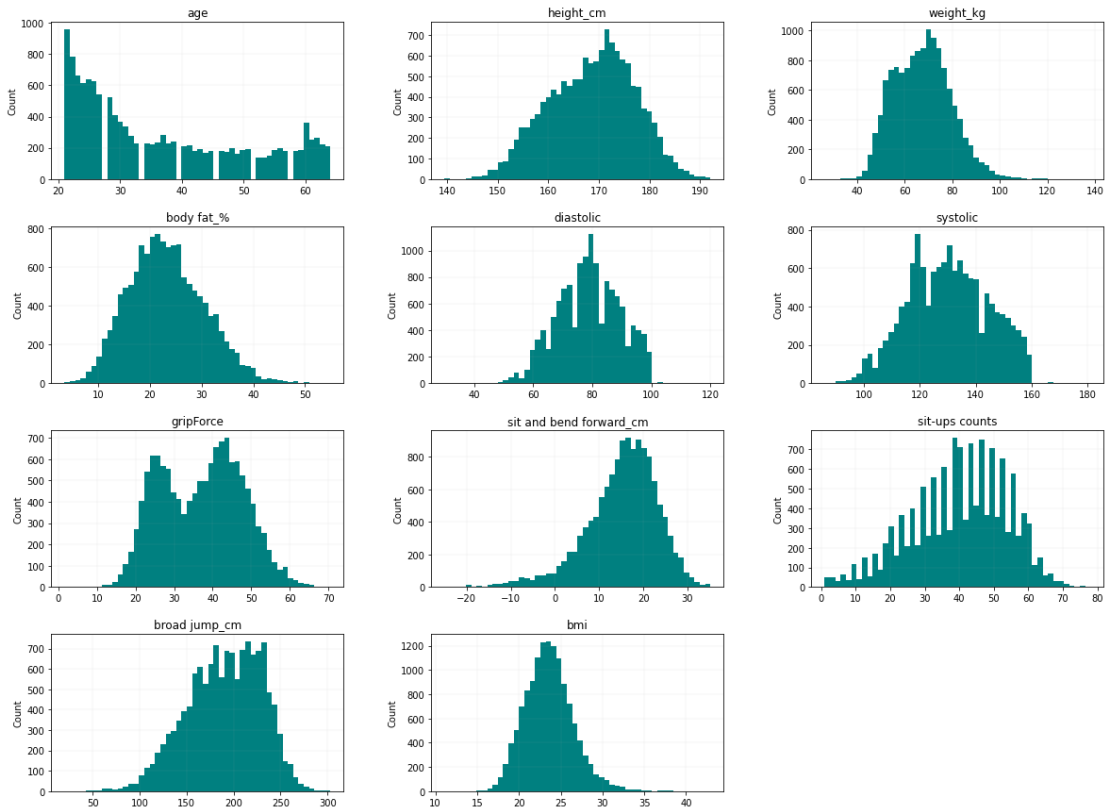
מסקנות:

ברוב הפיצ'רים קיימת הפרדה נראית לעין עבור גברים ונשים.

ניתן לראות כי קיימת הפרדה בטווח הבין רבעוני (בין הרבעון הראשון לשלישי - "הקופסא") למשל בפיצ'רים gripForce, height_cm, weight_kg בין 2 המגדרים.

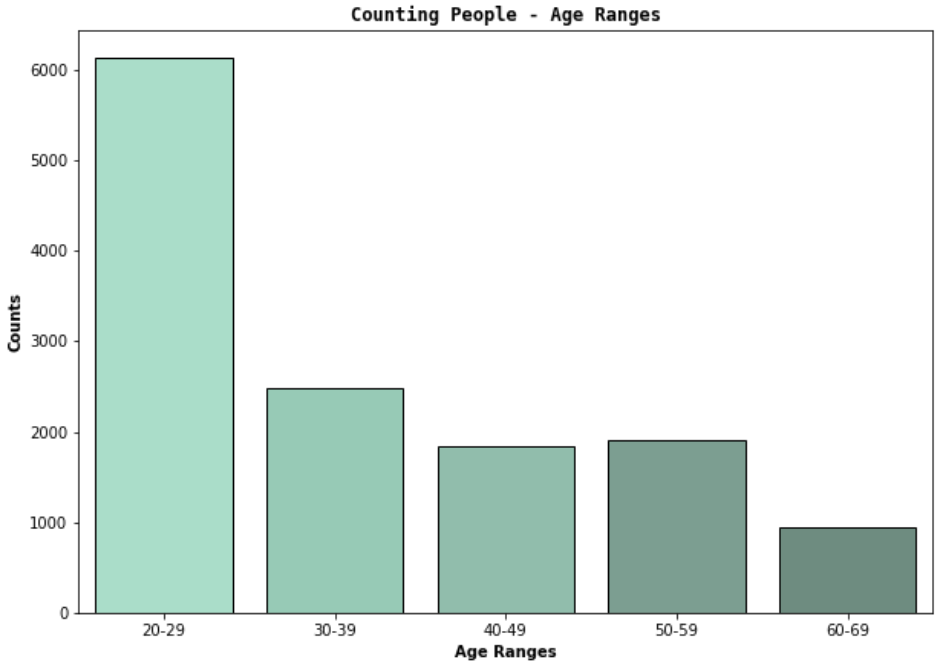
2. ניתוח כל פיצ'ר

א. ניתוח פיצ'רים נומריים ע"י היסטוגרמות



1) Age

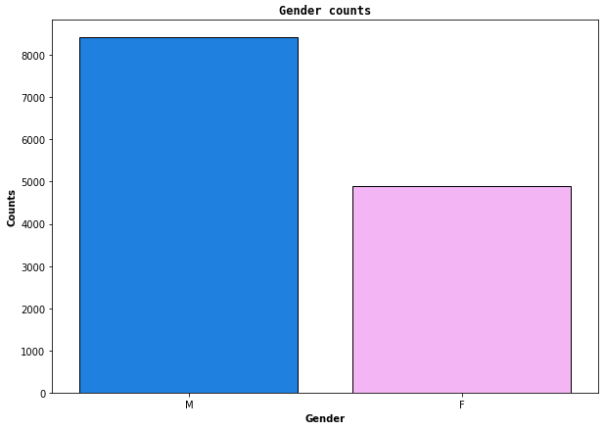
- לפי ההיסטוגרמה למעלה- רוב האנשים צעירים, מתחת לגיל 50.
- טווח הגילאים הוא 21-64
- ביצעתי חלוקה של הגילאים, לטווחים כדי לראות לפי קטגוריות טווחי גילאים באופן ויזואלי. ניתן לראות לפי החלוקה, כי גילאים 20-29 היא הקטגוריה הדומיננטית.



2) Height_cm
   - התפלגות נורמלית
   - גובה ממוצע סביב 169 ס"מ.
3) Weight_kg
   - משקל ממוצע סביב 70 ק"ג.
4) Body_fat%
   - ממוצע סביב 25%.
5) diastolic
   - התפלגות נורמלית
   - ממוצע סביב 80.
6) Systolic
   - ממוצע סביב 125.
7) gripForce
   - ממוצע סביב 40.
8) Sit and bend forward
   - ממוצע סביב 15 ס"מ.
9) Sit-ups
   - ממוצע סביב 40.
10) Broad jump
   - ממוצע סביב 190 ס"מ
11) bmi
   - התפלגות נורמלית
   - ממוצע סביב 22 שאומר משקל תקין.

ב. ניתוח פיצ'רים קטגורים :

Gender.1 - השוואה בין גברים ונשים



עבור גברים:

	age	height_cm	weight_kg	body fat_%	diastolic	systolic	gripForce	sit and bend forward_cm	sit-ups counts	broad jump_cm	bmi	class_numeric
count	8416.000000	8416.000000	8416.000000	8416.000000	8416.000000	8416.000000	8416.000000	8416.000000	8416.000000	8416.000000	8416.000000	8416.000000
mean	36.125594	173.262001	73.569788	20.186285	80.686074	133.832818	43.444764	13.033948	44.955561	211.608365	24.475605	1.430490
std	13.083389	5.794438	9.433212	5.913664	10.133011	13.272104	7.133429	7.913728	11.668899	27.908613	2.626966	1.098442
min	21.000000	145.800000	36.500000	3.500000	43.000000	89.000000	3.500000	-25.000000	1.000000	43.000000	11.314973	0.000000
25%	25.000000	169.300000	67.300000	16.000000	74.000000	123.000000	38.800000	8.800000	37.000000	194.000000	22.782523	0.000000
50%	32.000000	173.100000	72.800000	20.000000	81.000000	134.000000	43.300000	14.200000	46.000000	214.000000	24.285077	1.000000
75%	45.000000	177.100000	79.000000	24.000000	88.000000	144.000000	48.100000	18.300000	54.000000	231.000000	25.934162	2.000000
max	64.000000	192.000000	138.100000	54.900000	120.000000	181.000000	70.500000	35.200000	78.000000	303.000000	39.949756	3.000000

עבור נשים:

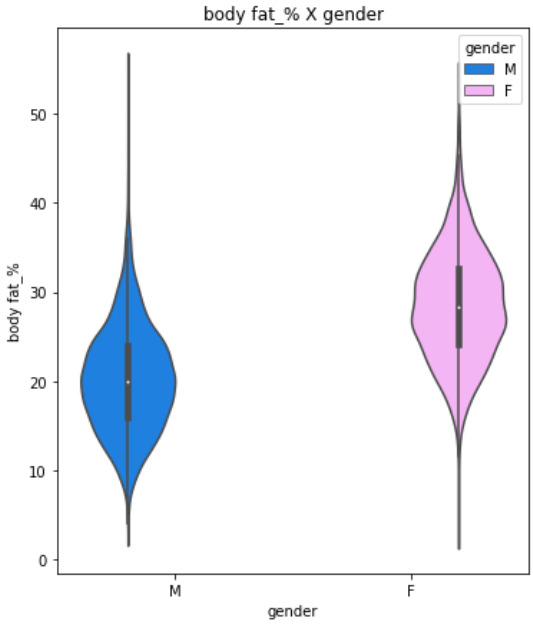
	age	height_cm	weight_kg	body fat_%	diastolic	systolic	gripForce	sit and bend forward_cm	sit-ups counts	broad jump_cm	bmi	class_numeric
count	4893.000000	4893.000000	4893.000000	4893.000000	4893.000000	4893.000000	4893.000000	4893.000000	4893.000000	4893.000000	4893.000000	4893.000000
mean	37.817699	160.522481	56.939064	28.475241	75.642959	124.051298	25.837871	18.843846	31.846311	153.684243	22.106918	1.608420
std	14.397333	5.602971	7.630039	6.217572	10.389622	14.212182	4.687914	7.150364	13.058789	27.722876	2.834536	1.144482
min	21.000000	139.500000	26.300000	3.500000	30.000000	86.000000	1.600000	-22.000000	1.000000	20.000000	11.103976	0.000000
25%	24.000000	156.700000	52.000000	24.100000	68.000000	114.000000	22.700000	15.300000	22.000000	135.000000	20.189666	1.000000
50%	34.000000	160.500000	56.000000	28.300000	75.000000	123.000000	25.700000	20.000000	32.000000	156.000000	21.740297	2.000000
75%	51.000000	164.300000	61.000000	32.600000	83.000000	133.000000	28.800000	23.500000	41.000000	173.000000	23.601226	3.000000
max	64.000000	179.000000	113.300000	53.500000	113.000000	178.000000	45.500000	35.200000	74.000000	260.000000	42.906509	3.000000

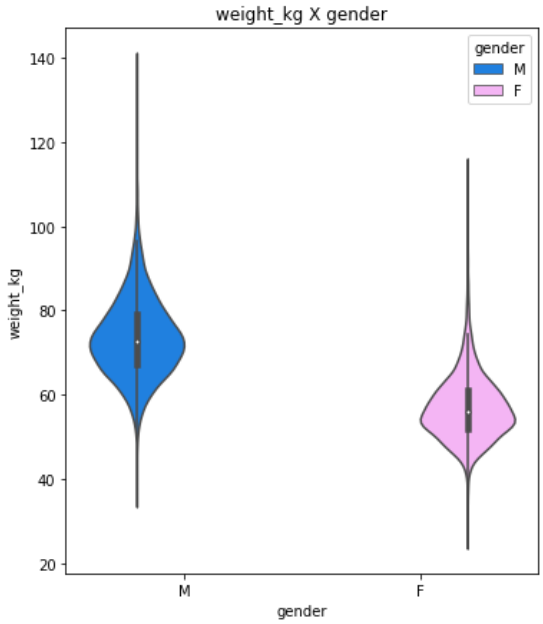
מסקנות פיצ'ר gender:

- יש יותר גברים מנשים . (63% גברים, ו37% נשים)
- הגובה הממוצע של גברים גדול משל הנשים.
- המשקל הממוצע של גברים גדול משל נשים.
- הbmi הממוצע של גברים גדול משל נשים.
- אחוז השומן של הנשים גדול משל הגברים.
- לחץ הדם הסיסטולי והדיאסטולי בממוצע של גברים גבוה משל נשים.
- תוצאות התרגיל grip force של גברים גבוהות יותר מאשר נשים.
- תוצאות התרגיל sit and bend forward של נשים גבוהות יותר מאשר גברים.
- תוצאות התרגיל sit-ups count של גברים גבוהות יותר מאשר נשים.
- תוצאות התרגיל broad jump של גברים גבוהות יותר מאשר נשים.
- גברים טובים יותר ב: sit-ups, broad jump, grip force (לפי ממוצע)
- נשים טובות יותר ב: sit and bend forward (לפי ממוצע)

אציג בגרף violin ויזואליזציה של 2 ממסקנותיי: אחוזי השומן בגוף של נשים גדולים יותר מגברים, ומשקל הגברים גדול יותר מנשים.

- בגרף השמאלי: לגברים אחוזי שומן נמוכים יותר, הממוצע הוא בסביבות 20%, החציון הוא סביב 19%. לנשים יש אחוזי שומן גבוהים יותר, הממוצע הוא בסביבות 27%, החציון הוא בסביבות 29%.
- בגרף הימני: לגברים משקל גבוה יותר, הממוצע הוא בסביבות 70 ק"ג, החציון הוא בסביבות 72 ק"ג. לנשים משקל נמוך יותר, הממוצע הוא בסביבות 50 ק"ג, החציון הוא בסביבות 55 ק"ג.





Class.2

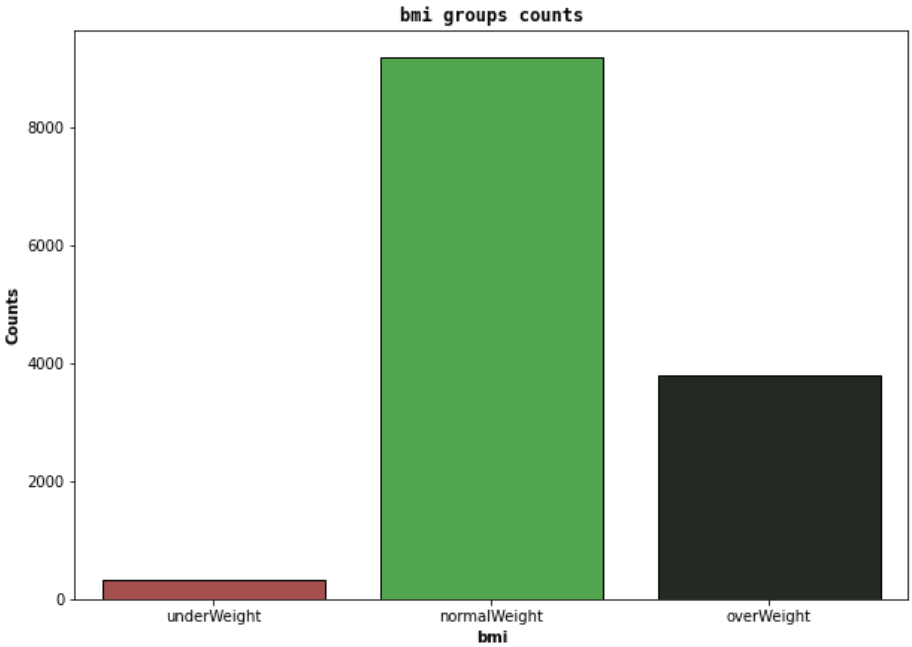
הצגתי בסעיף ניקוי הדאטא את התפלגות הנתונים בין המחלקות לאחר ניקוי הדאטא ממחלקות A B C.
בטבלה הבאה אציג את הממוצע עבור כל מחלקה לכל פיצ'ר :

class	age	height_cm	weight_kg	body fat_%	diastolic	systolic	gripForce	sit and bend forward_cm	sit-ups counts	broad jump_cm	bmi	class_numeric
D	38.058226	168.625530	71.996948	27.735432	80.099731	131.123320	34.758783	7.644757	31.336817	174.161242	25.181198	0.0
C	36.659555	169.190475	66.773218	22.628168	78.563702	129.898137	36.614124	14.383404	38.757212	189.020433	23.194850	1.0
B	37.033183	168.625551	66.609107	21.994726	78.720664	130.641931	37.943196	17.395297	42.680543	195.503469	23.305765	2.0
A	35.227615	167.869460	64.400947	20.534250	77.932469	129.276153	38.593663	21.332822	47.860416	202.729273	22.723225	3.0

מסקנות פיצ'ר class:

- הגיל הממוצע נמוך יותר ככל שהקלאס טוב יותר.
- המשקל הממוצע נמוך יותר ככל שהקלאס טוב יותר.
- אחוז השומן הממוצע נמוך יותר ככל שהקלאס טוב יותר.
- תוצאת התרגיל sit end bend forward הממוצע גבוה יותר ככל שהקלאס טוב יותר.
- תוצאת התרגיל sit ups הממוצע גבוה יותר ככל שהקלאס טוב יותר.
- תוצאת התרגיל broad jump הממוצע גבוה יותר ככל שהקלאס טוב יותר.
- הbmi הממוצע נמוך יותר ככל שהקלאס טוב יותר.

3.ניתוח פיצ'ר bmi באופן קטגורי: bmi_cat

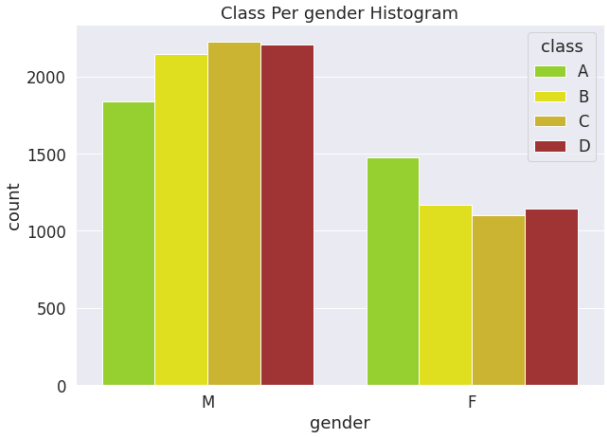


מסקנה

ניתן לראות כי מרבית האנשים בדאטאסט בעלי BMI תקין .

3. ויזואליזציה מעניינת

א. Class מול gender



מסקנות :

- גברים: המחלקה הדומיננטית היא C .
- נשים: המחלקה הדומיננטית היא A.
- בשני המגדרים מחלקות B C D במספרן מאוד קרובות, מחלקה A נראית יותר חריגה כמותית.
  אצל הנשים: מחלקה A גבוהה מכולן, אצל הגברים: מחלקה A נמוכה מכולן.
  *חשוב לזכור כי כמות הנשים קטנה משל הגברים .

נסתכל על הנתונים הסטטיסטיים בחלוקה זו :לפי מין ומחלקה, ונסיק מסקנות ביחס לממוצע, מסקנות אלה יכולות סממן עזר כיצד להשתייך למחלקה A לפי מגדר:

gender	class	age	height_cm	weight_kg	body fat_%	diastolic	systolic	gripForce	sit and bend forward_cm	sit-ups counts	broad jump_cm	bmi
M	D	35.734574	173.561162	77.974492	24.794975	82.156987	134.342105	40.692559	5.557316	36.508621	196.148820	25.844120
	C	35.574573	173.579245	72.239879	19.577516	80.184636	133.148697	42.469561	12.671631	44.081761	209.397125	23.944802
	B	36.988811	172.902704	72.197310	18.941704	80.376690	133.989277	44.283074	15.708559	47.560373	215.724476	24.129161
	A	36.254210	172.938892	71.503726	16.855048	79.891907	133.868007	46.942042	19.306611	53.089625	227.994025	23.882662
F	D	42.531004	159.124978	60.490830	33.395492	76.139738	124.927511	23.336912	11.662854	21.381659	131.837555	23.904795
	C	38.851180	160.325318	55.730762	28.790376	75.289474	123.332123	24.786352	17.841125	28.001815	147.860254	21.680100
	B	37.114530	160.784103	56.364068	27.591932	75.684615	124.505128	26.320085	20.487650	33.734188	158.431624	21.796197
	A	33.947154	161.546409	55.541721	25.123281	75.488482	123.548780	28.180813	23.860095	41.338076	171.216802	21.277188

נשים:

- גיל –נמוך יחסית, ממוצע 34
- גובה -גבוה, ממוצע 161
- משקל- אין משמעות ביחס לשאר המחלקות והסיווג
- נמוך body fat_% - ממוצע 25%
- לחץ דם דיאסטולי נמוך- אין משמעות ביחס לשאר המחלקות והסיווג
- לחץ דם סיסטולי - אין משמעות ביחס לשאר המחלקות והסיווג
- grip force- גבוה, ממוצע 28
- sit and bend forward_cm- גבוה, ממוצע 24
- sit-ups counts- גבוה, ממוצע 41
- broad jump_cm- קפיצה רחוקה, ממוצע 171 ס"מ
- bmi- אין משמעות ביחס לשאר המחלקות והסיווג

שילוב של תכונות אלה יכולות להוביל אישה למחלקה A.

גברים:

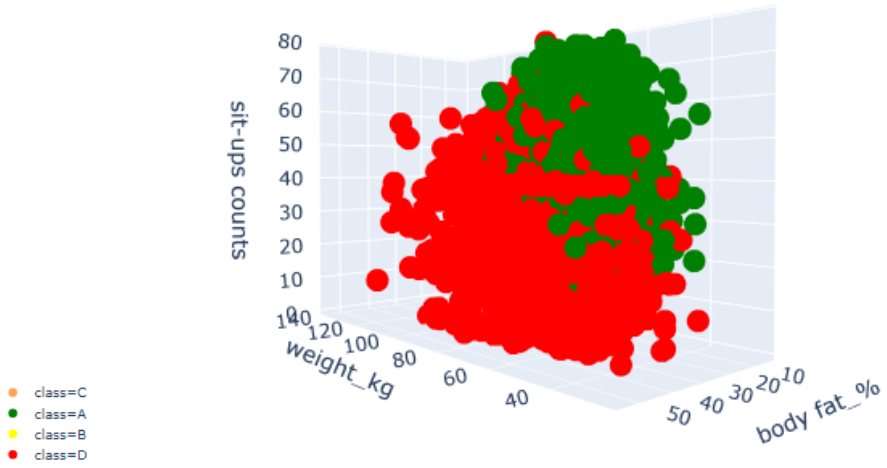
- גיל - אין משמעות ביחס לשאר המחלקות והסיווג
- גובה- אין משמעות ביחס לשאר המחלקות והסיווג
- משקל -נמוך, ממוצע 71 ק"ג
- %_body fat - נמוך, ממוצע 16%
- לחץ דם דיאסטולי- נמוך, ממוצע 79
- לחץ דם סיסטולי - נמוך, ממוצע 133
- grip force -גבוה, ממוצע 46
- sit and bend forward_cm -גבוה, ממוצע 19
- sit-ups counts- גבוה, ממוצע 53
- broad jump_cm- קפיצה רחוקה, ממוצע 227 ס"מ
- bmi- אין משמעות ביחס לשאר המחלקות והסיווג

שילוב של תכונות אלה יכולות להוביל גבר למחלקה A.

מסקנה כוללת לסעיף זה:

- אצל נשים המדדים הדומיננטים לקביעה: sit and bend ,gripForce ,bodyfat ,height ,age broad jump_cm, sit-ups counts,forward_cm
- אצל גברים המדדים הדומיננטים לקביעה: sit ,gripForce ,systolic ,diastolic ,bodyfat ,weight broad jump_cm, sit-ups counts,and bend forward_cm
- ניתן להבחין כי המדדים מבחינת הממוצע בין גברים ונשים במחלקה A שונים בחלקם.

ב. weight X bodyfat_% X sit-ups של מחלקה A מול מחלקה D



מסקנה:

בכל שאחוז השומן גבוה, תוצאות כפיפות הבטן נמוכות, המשקל גבוה, תשויך למחלקה D.

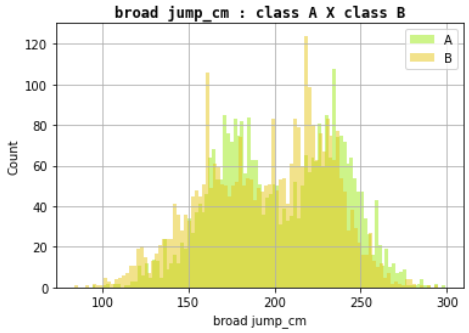
בכל שאחוז השומן נמוך, תוצאת כפיפות הבטן נמוכה, המשקל נמוך תשויך למחלקה A.

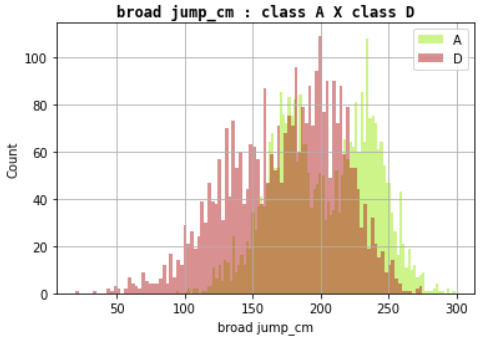
ניתן להבין כי אלו פיצ'רים שיכולים לסייע לנו בלהבדיל בין מחלקות A ו D (לא באופן מוחלט), אך בין B לC הערכים מאוד מעורבבים ,ולכן תצוגה זו עבור מחלקות B C לא מועילה.

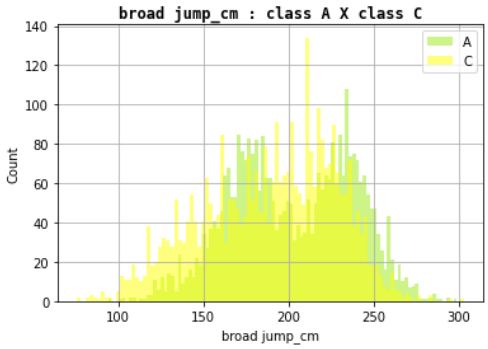
ג. Class X 2 sports exercises

תרגיל 1: broad jump

השוואה בין מחלקה A לכל מחלקה עבור תרגיל קפיצה לרוחק:







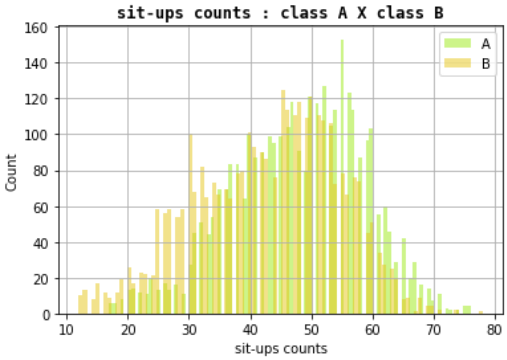
מסקנה:

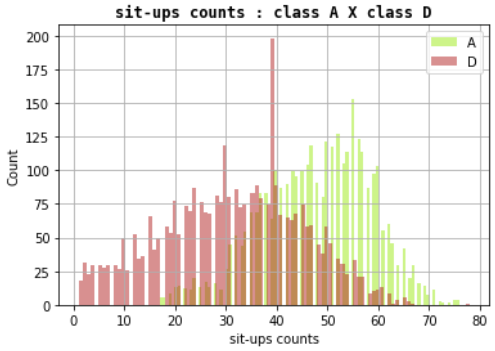
לאחר ביצוע השוואה בין מחלקה A לבין כל שאר המחלקות מבחינת תרגיל broad jump ניתן לראות כי :

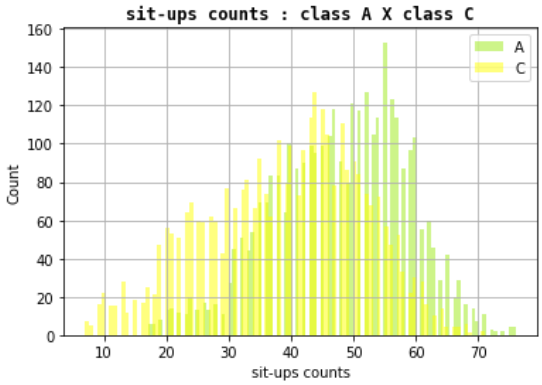
- מחלקה A בולטת קדימה בציר x באופן יחסי בתוצאות גבוהות יותר משאר המחלקות.
- מחלקה D בולטת לעומת A בעריכה הנמוכים.
- מחלקות B C מתנהגות באופן יחסי בצורה דומה.

תרגיל 2: sit-ups

השוואה בין מחלקה A לכל מחלקה עבור תרגיל כפיפות בטן:





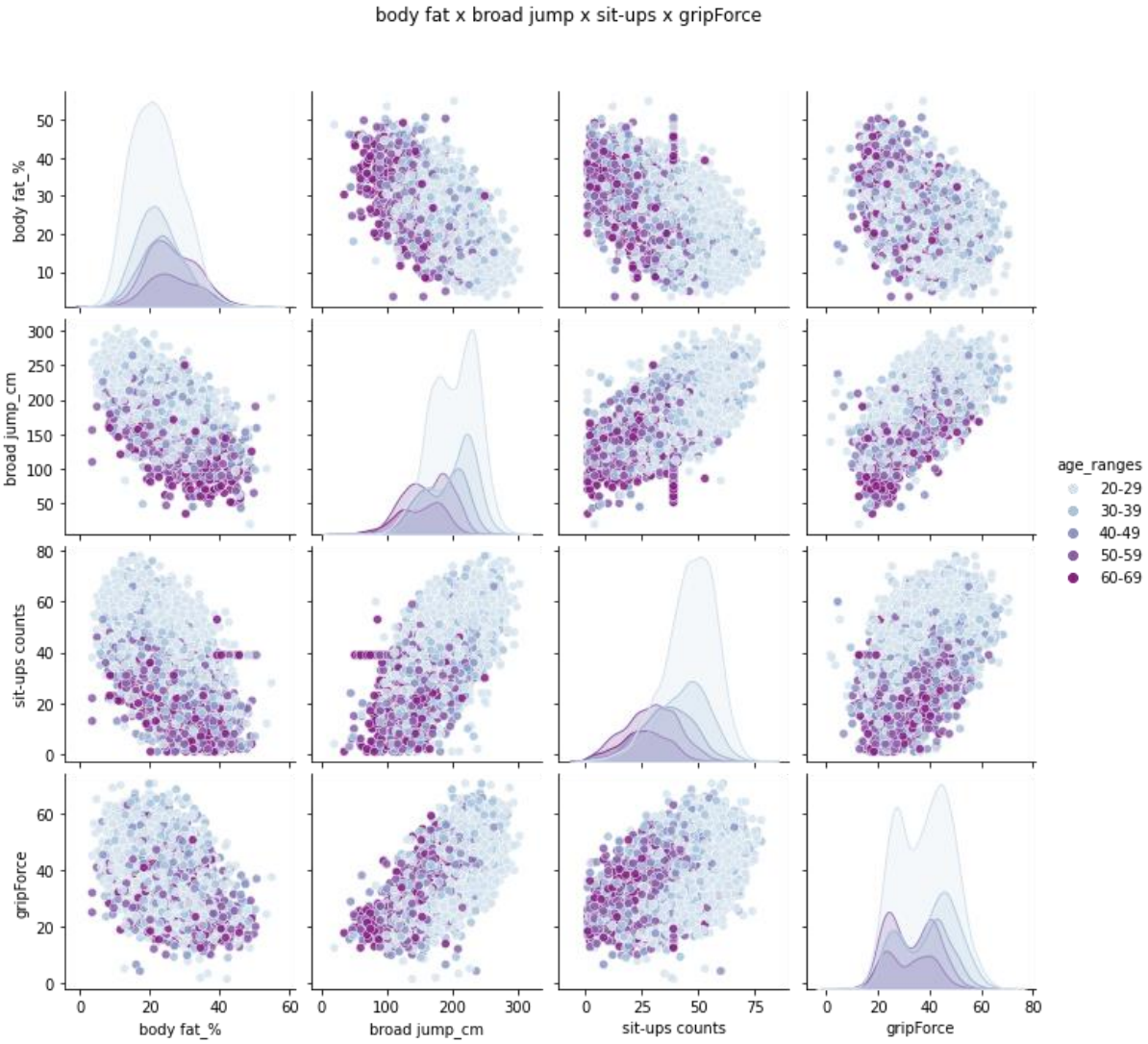


מסקנה

לאחר ביצוע השוואה בין מחלקה A לבין כל שאר המחלקות מבחינת תרגיל sit-ups ניתן לראות כי :

- מחלקה A בולטת באופן יחסי בתוצאות גבוהות יותר משאר המחלקות, גם בערכיה הגבוהים וגם בכמות הגבוהה לערכים גבוהים.
- מחלקות D וגם C בולטות לעומת A בעריכה הנמוכים, גם בכמות .

ד. יחסים בין הפיצ'רים : body fat x broad jump x sit-ups x gripForce בשילוב טווח גילאים



מסקנות:

- Broad jump X bodyfat
ניתן לראות כי אחוזי שומן נמוכים מאפיינים בעיקר צעירים(20-29) שלהם קפיצה למרחק רחוקה, בעוד שאחוזי שומן גבוהים מאפיינים יותר מבוגרים, שאחוזי השומן שלהם נמוכים.
- Broad jump X sit ups
ניתן לראות כי קפיצה למרחק קצר וכפיפות בטן מועטות מאפיינות אנשים מבוגרים, בעוד שקפיצה למרחק רחוק וכפיפות בטן מרובות מאפיינות אנשים צעירים.
- Sit-ups X gripForce
ניתן לראות כי gripForce נמוך, מאפיין אנשים מבוגרים בשילוב כפיפות בטן יחסית מעטות, בעוד שgripForce גבוה מאפיין בעיקר אנשים צעירים יותר בשילוב כפיפות בטן מרובות.

4. ניקיון נוסף לאחר החקירה

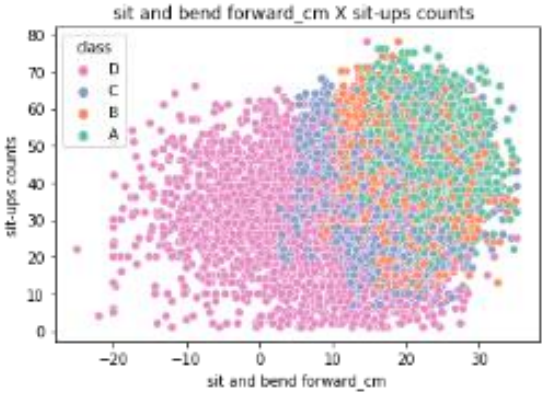
כיוון שהBMI מחושב באמצעות המשקל ותלוי בו לינארית, קיבלנו קורלציה גבוהה (0.84)בין השניים. לכן, בשלב זה אחליט להסיר פיצ'ר זה(וכן גם את הפיצ'ר הקטגורי שמחזיק בטווחים).
אציין, כי בשלב זה בחנתי בדיקה, האם פיצ'ר זה ישנה את הprediction שלי בעץ ההחלטה על הדאטא המעובד עבור כל הפיצ'רים שלי וגיליתי כי הוא יניב את אותה תוצאה בדיוק בaccuracy לולא הייתי מורידה עמודה זו וזו היתה סיבה נוספת להחלטה להסיר עמודה זו כיוון שאינה תורמת למודל .
(ניסוי זה נעשה בשלב הpredication , פעם אחת ללא drop של עמודת bmi הרגילה והקטגורית ופעם עם drop לעמודת bmi הרגילה והקטגורית).

חלק ד: מודל סיווג

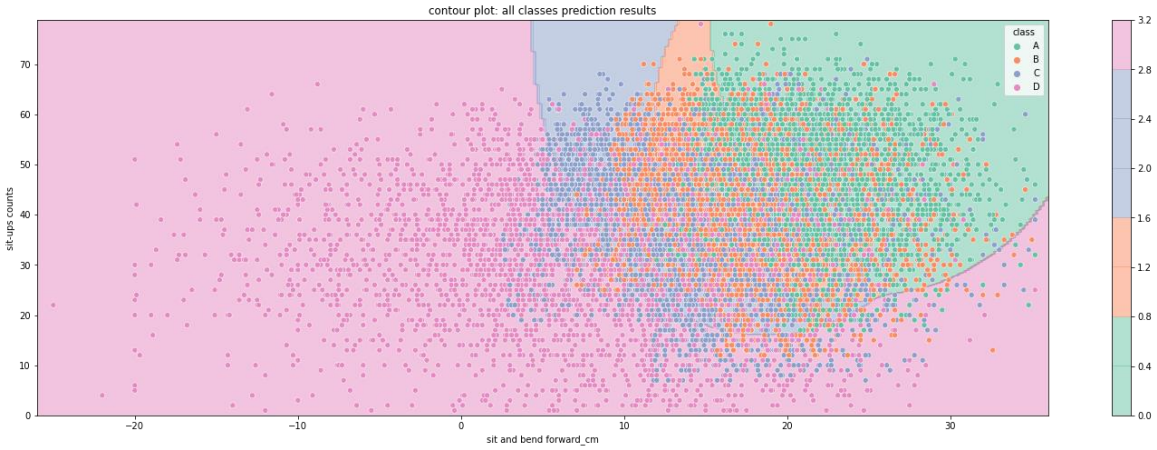
חלק ראשון: Naïve baise בבחירת 2 פיצ'רים על הדאטא המעובד

שלב מקדים:
הפיכת הפיצ'רים לנומריים.

א.בחירת 2 פיצ'רים שיתנו את התוצאות הטוב ביותר עבור מסווג נאיב בייס:
לאחר צפייה בגרף pairplot שמציג פיצ'ר מול פיצר והצבעים לפי המחלקות, ניתן לראות כי 2 הפיצ'רים שנותנים את התוצאות הטובות ביותר הם: sit and bend forward_cm ו sit-ups counts.
אופן החיפוש נעשה ע"י חיפוש הפרדה מקסימלית בין הקבוצות, פיצרים אלו איפשרו הפרדה זו בצורה מקסימלית לעומת השאר.
#קטע זה מופיע בהערה בקוד מאחר ויצירת pairplot עם כמות גדולה של פיצ'רים לוקח זמן .



ב. ויזואליזציה של שני הפיצ'רים בגרף 2D:



*הערה: 2 גרפים אלה נמצאים בקוד בסעיף B

- גרף זה ממפה עבור זוג הפיצ'רים הנבחר מה תהיה הprediction . ניתן לראות כי עבור class D השטח הוא מירבי כלומר יש נטייה אליה.
- הנקודות בגרף מייצגות דגימות מהtest set, עם צבעים בהתאמה. כאשר יש התאמה בין צבע הנקודה לצבע הרקע מדובר בחיזוי נכון ולהפך אחרת.

מסקנות

- ניתן לראות כי קיים שטח גדול מאוד (מעל 50%) ממרחב הפיצ'רים שנוצר מזוג הפיצ'רים שבחרתי שמתאים ל class D.
  לכן אצפה שהחיזוי ל class D יהיה מדויק יותר.
- באזורים בגרף בהם החיזוי הוא ל class A B C ניתן לראות כי קיימות נקודות מכל הקלאסים האחרים בכמות ניכרת , אך ניתן להבחין גם כי ריכוז הנקודות המתאימות לקלאס החיזוי גבוה יותר.
- מהדוח קלסיפיקציה:
  - 0.55 Accuracy –נמוך.
  - המדדים האחרים נמוכים יחסית, הם הכי גבוהים עבור D והכי נמוכים עבור B. בהתאמה למה שנצפה בגרף.

	precision	recall	f1-score	support
A	0.59	0.71	0.64	1012
B	0.37	0.41	0.39	958
C	0.49	0.41	0.45	1021
D	0.75	0.66	0.70	1002
accuracy			0.55	3993
macro avg	0.55	0.55	0.55	3993
weighted avg	0.55	0.55	0.55	3993

חלק שני: Decision tree, א – על דאטא מקורי, ב-על דאטא מעובד

א. עץ החלטה בסיסי בשימוש בדאטא המקורי וכל הפיצ'רים:

דוח הקלסיפיקציה:

Classification Report - Decision Tree with all features

	precision	recall	f1-score
A	0.68	0.67	0.68
B	0.52	0.52	0.52
C	0.58	0.6	0.59
D	0.8	0.78	0.79
accuracy	0.64	0.64	0.64
macro avg	0.64	0.64	0.64
weighted avg	0.65	0.64	0.64

מסקנות:

- ניתן לראות כי הaccuracy הוא 0.64 .
- המחלקה שצדקנו בה הכי הרבה היא D ולאחריה A.
- המחלקה שטעינו בה הכי הרבה היא B.

ב.עץ החלטה על דאטא מעובד עם פיצ'רים נבחרים ותנאי עצירה:

1. <u>יצירת עץ מקדים מהדאטא המעובד וכל הפיצ'רים לחישוב importance:</u>

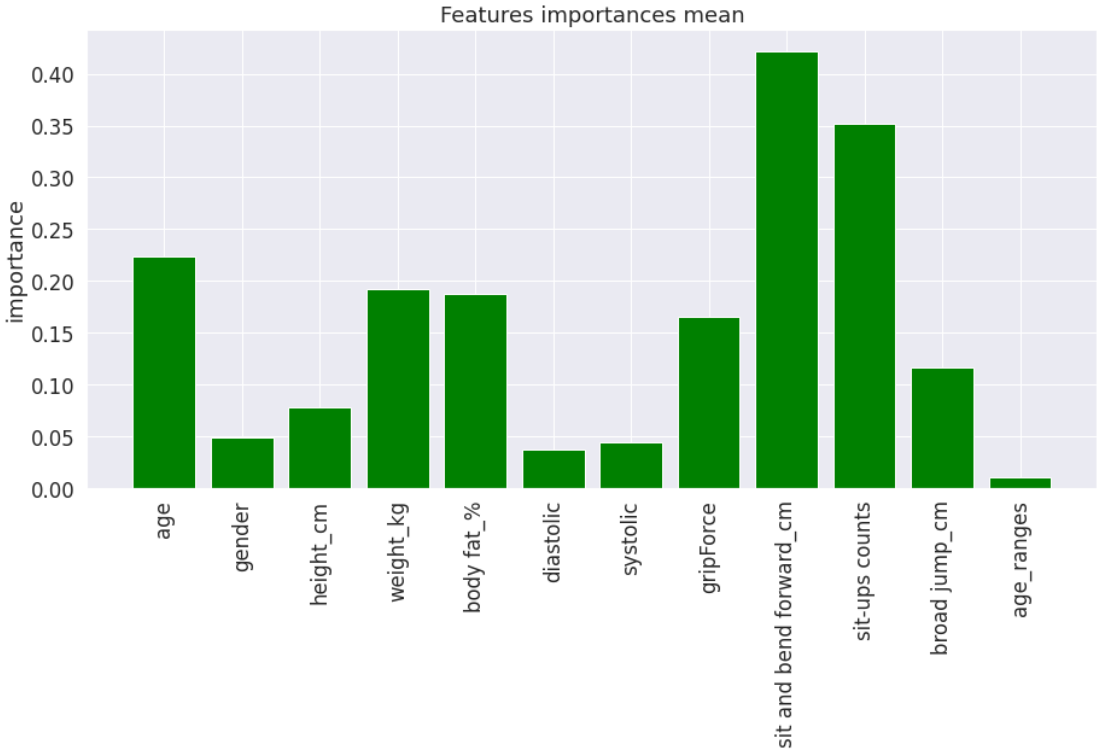
לחישוב importance על הפיצ'רים, יצרתי עץ מכל הפיצ'רים על הדאטא המעובד, התוצאה יצאה בדומה לתוצאה בסעיף הקודם מבחינת accuracy , כלומר הדאטא המקורי והמעובד מאוד דומים בתוצאות. כיוון שהשלמנו מעט מאוד ערכים חסרים (החלפת 0 בחציון בעמודות שתוארו לעיל) זו תוצאה צפויה.

Classification Report - Decision Tree

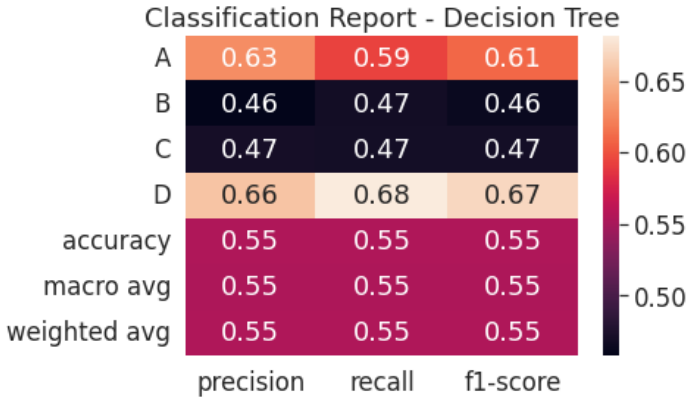
	precision	recall	f1-score
A	0.68	0.68	0.68
B	0.51	0.51	0.51
C	0.58	0.57	0.58
D	0.77	0.78	0.78
accuracy	0.64	0.64	0.64
macro avg	0.64	0.64	0.64
weighted avg	0.64	0.64	0.64

2. חישוב importance :

ניתן לראות כי הפיצ'רים הכי רלוונטים הם: sit and bend forward_cm ו sit-ups count , אפשר גם להגיד שהפיצ'ר age בולט ויתכן גם שweight_kg.



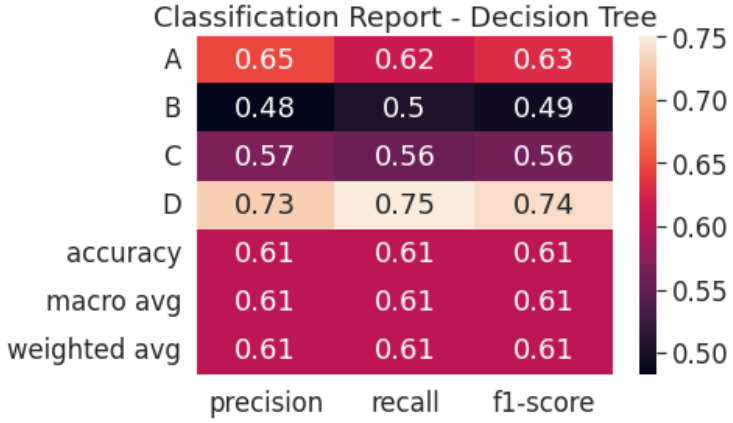
3. יצירת עץ ראשוני עם 3 הפיצ'רים בעלי ה importance הגבוה ביותר: sit and bend_forward_cm ו-sit age,ups count הדוח שהתקבל:



מסקנה: ניתן לראות כי כל אחד ממדדי הדיוק של המודל ירדו.

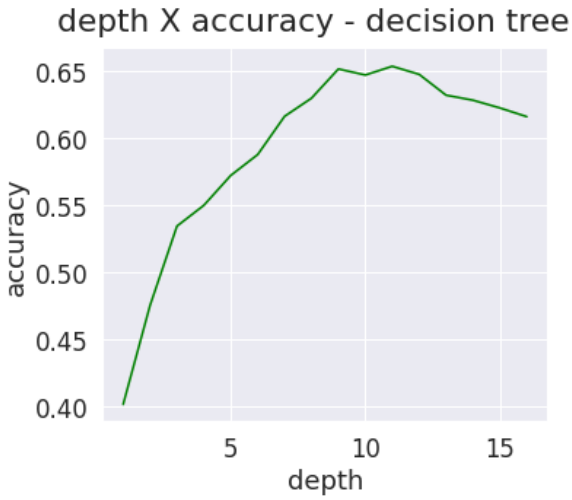
4. ננסה לשפר את העץ. כל שינוי יבחן בפני עצמו, במידה ויהיה שיפור בתוצאות חיזוי המודל, השינוי יישמר, אחרת נעבור לבדוק שינוי אחר עם אותם הנתונים.
השיפורים שבדקתי, בסדר הזה, הם:

ניסיון לשיפור 1- הוספת פיצ'ר נוסף: הוספתי את הפיצ'ר weight_kg שגם לו יש important יחסית גבוה.

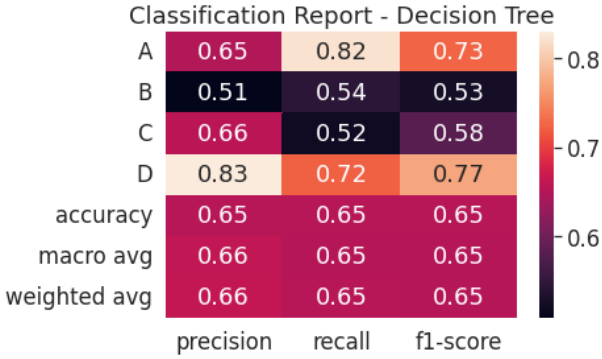


מסקנה: ניתן לראות כי כל אחד ממדדי הדיוק של המודל עלו, לכן נבחר לשמור את הפיצ'ר החדש ברשימת הפיצ'רים לחיזוי.

ניסיון לשיפור 2: (בשימוש ב4 הפיצרים) תנאי עצירה max_depth -עומק מקסימלי



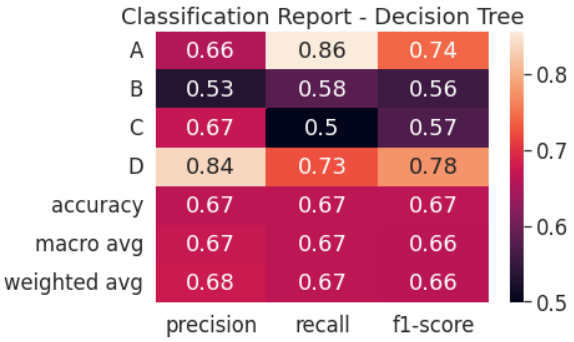
ניתן לראות כי עומק 11, יניב את הaccuracy הכי גבוה 0.65.



מסקנה: משפר את סעיף קודם. לכן נמשיך עם תנאי עצר זה אלא אם נמצא טוב ממנו.

ניסיון לשיפור 3: תנאי עצירה:max_leaf_nodes

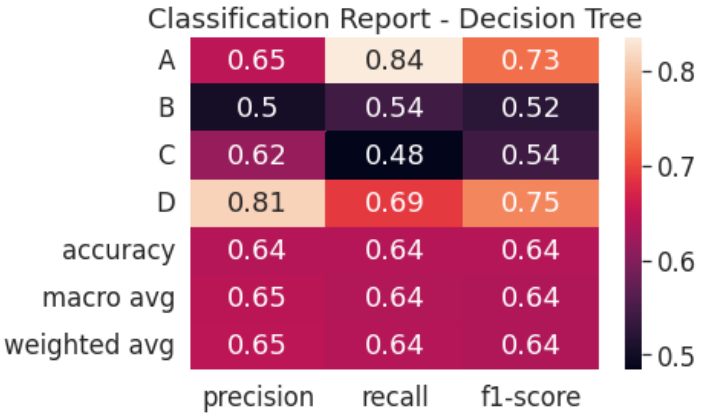
הגבלת כמות עלים: ל185 ונקבל : 0.67 accuracy



מסקנה: משפר את סעיף קודם. לכן נמשיך עם תנאי עצר זה במקום הקודם.

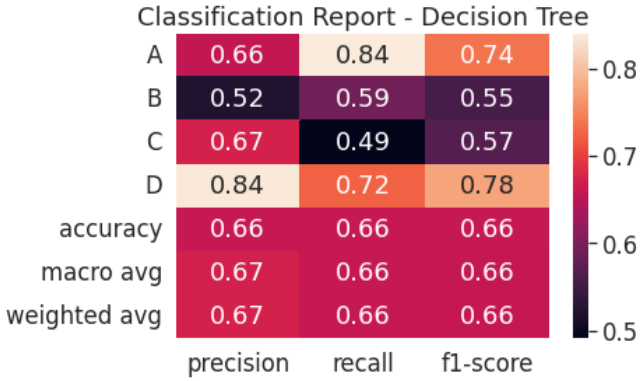
ניסיון לשיפור 4: min_samples_split

כאשר נגדיר תנאי עצירה זה ל0.01 נקבל accuracy של 0.64- הכי גבוה מכל הניסיונות.



מסקנה : לא נשתמש בו, שיפור max_leaf_nodes עדיף.

ניסיון לשיפור 5: min_impurity_decrease



מסקנה : לא נשתמש בו, שיפור max_leaf_nodes עדיף.

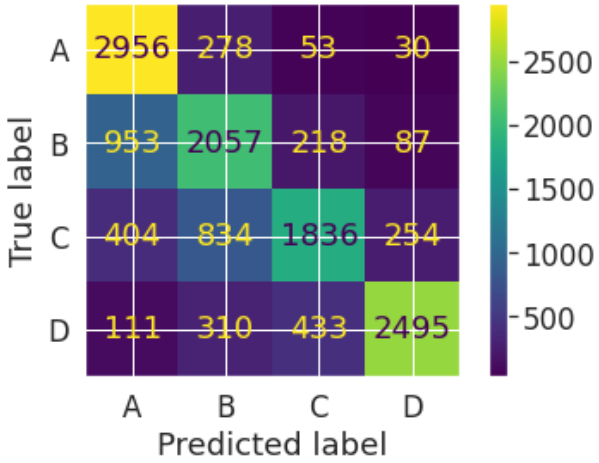
ניסיון לשיפור כללי: עבור כל הנסיונות ביצעתי שינויים במדד הcriterion : gini וentropy, וגיליתי שentropy סיפק תוצאות טובות יותר עבור דאטא זה, או שבחלק מהמקרים הם זהים.

הערה: קביעת המספר לתנאי העצירה נעשה לאחר הרבה ניסיון של בדיקה לכל מדד מה המספר שמניב את התוצאה הגבוה ביותר .

<u>מטריצה לסיכום טעויות המודל הסופי</u>

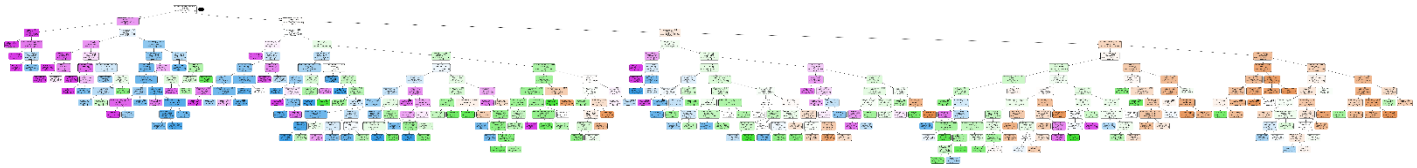
ניתן לראות במטריצה זו באלכסון עבור כל מחלקה כמה פעמים צדקנו עבורה .
בכל מקום אחר שלא על האלכסון ניתן לראות כמה פעמים חשבנו שזו המחלקה X אבל היא היתה תויגה למשהו אחר.

- עבור A: עבור 2956 פעמים שהמודל זיהה שזו מחלקה A הוא צדק .
  278 פעמים, זו היתה מחלקה A אבל המודל תייג אותה לB.
  כלומר היתה נטיה למחלקה A כאשר המודל טועה עבורה להיות מתויגת למחלקה B .
- עבור B: עבור 2057 פימים שחשבנו שזו מחלקה B צדקנו.
  953 פעמים, זו היתה מחלקה B אבל המודל תייג אותה לA.
  כלומר היתה נטיה למחלקה B כאשר המודל טועה עבורה להיות מתויגת למחלקה A.
- עבור C: עבור 1836 פעמים שהמודל זיהה שזו מחלקה C הוא צדק.
  834 פעמים, זו היתה מחלקה C אבל המודל תייג אותו לB .
  כלומר היתה נטיה למחלקה C כאשר המודל טועה עבורה להיות מתויגת למחלקה B.
- עבור D: עבור 2495 פעמים שהמודל זיהה שזו מחלקה D הוא צדק.
  433 פעמים, זו היתה מחלקה D אבל המודל תייג אותו לC .
  כלומר היתה נטיה למחלקה D כאשר המודל טועה עבורה להיות מתויגת למחלקה C.



העץ הסופי:

1. עץ שנבנה לפי 4 פיצ'רים : sit and bend forward, sit-ups counts, weight_kg,age
2. תנאי עצירה: max_leaf_nodes , נגביל ל185, יניב תוצאת accuracy של 0.67



📎 תמונה מצורפת להגשה.

סיכום

- במהלך החקירה של הדאטא, גיליתי כי הפיצ'רים sit and bend forward וsit-ups counts הם בעלי קורלציה גבוהה לקלאס, מה שבא לידי ביטוי בבחירתם לקלסיפיקציה ב2 סוגי המודלים. ניתן לראות זאת גם בfeature importance.
- המסווג נאיב בייס, שנבנה על הדאטא המעובד, בעל accuracy: 0.55, כלומר אינו מדוייק.
- המסווג עץ החלטה שנבנה על הדאטא המקורי עם כל הפיצ'רים , בעל accuracy : 0.64
- המסווג עץ החלטה שנבנה על הדאטא המעובד ועם 4 פיצ'רים בולטים: sit and bend forward וsit-ups counts, age, grip-force בעל accuracy 0.67 בהוספת תנאי העצירה שמגביל את כמות העלים.
- ניתן לראות שבכל המסווגים האלו גם בnaive baise וגם בעץ ההחלטה הדיוק אינו גבוה 64%-68% ולכן יש טעם לבחון מסווגים נוספים לשיפור תוצאות החיזוי.
- ניתן לראות כי קבוצה D היא הקבוצה עם רמת הדיוק הכי גבוהה, לאורך הדרך היא הייתה בעלת מדדי הדיוק הגבוהים ביותר, לעומת קבוצה B שקיבלה תוצאות נמוכות משאר המחלקות.

- כאמור, יש מקום לבחינת אלגוריתמים נוספים לשיפור תוצאות הprediction, אך ראינו ביצד שינויים בפרמטרים של המודל הספציפי שבחרנו יכולים לשפר את תוצאות החיזוי על אותו הדאטא.