ממ"ן 21 - כריית מידע

דניאל לייבנר - 208271775

קורס - כריית מידע – 20595
סמסטר 2022ב
23–אפריל–22

ממן 21

ממן 21

שאלה 1:

סעיף א'

הגדירו את מטרות כריית המידע. ציינו את ההנחות וההפשטות בהן השתמשתם.

מטרת כריית המידע היא לחזות את דרגת ההשמנה של נבדק מהאוכלוסייה שבאזור מקסיקו פרו וקולומביה באמצעות התכונות מבסיס הנתונים שניתן לנו.

התכונה שאנו רוצים חזות(לכל רשומה) היא העמודה NObeyesdad (רמת השמנת היתר) .

בסיס הנתונים שלנו כולל 2111 רשומות כאשר לכל רשומה יש 17 תכונות סב"ה (כולל היעד לחיזוי).

```
In [98]: df.shape
Out[98]: (2111, 17)
```

איור 1

ההנחות שהשתמשתי בהם במהלך הפרויקט:

1. יש תלות בין המשתנים שניתנו במאגר המידע לבין רמת השמנה.
2. בהתאם לנאמר במאמר, הנתונים נאספו ע"י מקור אמין שאין בכוונתו "לחבל" במטרות הכרייה.
3. ניסיתי לבצע הנחה שאומרת שרמת ההשמנה שקוטלגה לאדם היא לפי נוסחת ה bmi כפי שמובא במאמר אבל לאחר ניתוח הנתונים גיליתי שהנחה זו איננה נכונה ולכן לא אוכל להשמיט את הנתונים הללו מהמאגר.

ממן 21

סעיף ב'

הגדירו את הנתונים בהם השתמשתם בפרויקט כדוגמת: תכונות, סוג הנתונים, נתונים חסרים, תחומי ערכים ועוד.

הנתונים שהתקבלו, (כלומר מאגר המידע הגולמי) כלל 2111 רשומות כאשר לכל רשומה יש 17 תכונות (לא כולל תכונת האינדקס, שכמובן לא מהווה גורם השפעה אך היא יחודית לכל רשומה.)

בהתאם לנאמר במאמר ההסבר על הנתונים אפשר לחלק את התכונות(16) שאינן עמודת היעד ל3 קטגוריות אב מכלילות.

הרגלי אכילה :

1. FAVC - צריכה תכופה של מזון עם ערך קלורי גבוה
2. FCVC - תדירות צריכת הירקות
3. NCP - מספר הארוחות העיקריות
4. CAEC - צריכת מזון בין הארוחות
5. CH20 - צריכת מים מדי יום
6. CALC - צריכת אלכוהול

הרגלים משפיעים שאינם קשורים באכילה/שתייה :

1. SCC - ניטור צריכת קלוריות
2. FAF - תדירות פעילות גופנית
3. TUE - זמן במכשירים טכנולוגיים
4. MTRANS - סוג שימוש בתחבורה
5. SMOKE - עישון

תכונות פיזיות של נבדק :

1. GENDER - מגדר
2. AGE - גיל
3. HEIGHT - גובה
4. WEIGHT - משקל
5. family_history_with_overweight – היסטוריה משפחתית עם משקל עודף

בעזרת השימוש בפונקציה info נוכל לראות שאין נתונים חסרים במאגר המידע.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2111 entries, 0 to 2110
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Gender                          2111 non-null   object
 1   Age                             2111 non-null   float64
 2   Height                          2111 non-null   float64
 3   Weight                          2111 non-null   float64
 4   family_history_with_overweight  2111 non-null   object
 5   FAVC                            2111 non-null   object
 6   FCVC                            2111 non-null   float64
 7   NCP                             2111 non-null   float64
 8   CAEC                            2111 non-null   object
 9   SMOKE                           2111 non-null   object
 10  CH2O                            2111 non-null   float64
 11  SCC                             2111 non-null   object
 12  FAF                             2111 non-null   float64
 13  TUE                             2111 non-null   float64
 14  CALC                            2111 non-null   object
 15  MTRANS                          2111 non-null   object
 16  NObeyesdad                      2111 non-null   object
dtypes: float64(8), object(9)
memory usage: 280.5+ KB
```

איור 2

ממן 21

בדומה לכך, פרטים על הנתונים אפשר לקבל בפשטות ע״י הפונקציה describe :

```
df.describe()
```

	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE
count	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000
mean	24.312600	1.701677	86.586058	2.419043	2.685628	2.008011	1.010298	0.657866
std	6.345968	0.093305	26.191172	0.533927	0.778039	0.612953	0.850592	0.608927
min	14.000000	1.450000	39.000000	1.000000	1.000000	1.000000	0.000000	0.000000
25%	19.947192	1.630000	65.473343	2.000000	2.658738	1.584812	0.124505	0.000000
50%	22.777890	1.700499	83.000000	2.385502	3.000000	2.000000	1.000000	0.625350
75%	26.000000	1.768464	107.430682	3.000000	3.000000	2.477420	1.666678	1.000000
max	61.000000	1.980000	173.000000	3.000000	4.000000	3.000000	3.000000	2.000000

איור 3

ממן 21

ריכזתי את כלל הנתונים הללו לטבלה הבאה כמו שנתבקש בדוגמה לפתרון שנמצאת באתר הקורס.

שם תכונה	תיאור	סוג נתון	יחידות מידה	תחום ערכים	ממוצע	סטיית תקן	ערכים לא חוקיים
Gender	מגדר	category	Male - גברים Female - נשים	Male 1068 Female 1043			ללא
Age	גיל	numeric	שנים עם נקודה עשרונית - הומר לשנים ללא נקודה עשרונית	מ - 14 עד -61	24.3126	6.3459683	ללא
Height	גובה	numeric	מטרים	מ - 1.45 עד -	1.701677	0.0933048	ללא
Weight	משקל	numeric	קילוגרמים	מ - 39 עד -173	86.58606	26.191172	ללא
family_history_with_overweight	משפחתית עם משקל	category	yes\no	yes 1726 no 385			ללא
FAVC	האם אוכל מזון רב	category	yes\no	yes 1866 no 245			ללא
FCVC	תדירות צריכת ירקות	רציף		מ - 1 עד -3	2.419043	0.5339266	אבל מכיוון שהנתונים נאספו ובוצע עליהם כבר ניתוח מקדים לא נשנה את
NCP	מספר הארוחות העיקריות	רציף	מספר פעמים	מ - 1 עד -4	2.685628	0.7780386	אבל מכיוון שהנתונים נאספו ובוצע עליהם כבר ניתוח מקדים לא נשנה את
CAEC	צריכת מזון בין הארוחות	category		1765 Frequently 242 Always 53			ללא
SMOKE	האם מעשן	category		no 2067 yes 44			ללא
CH2O	צריכת מים יומית	רציף	ליטר	מ - 1 עד -3	2.008011	0.6129535	אבל מכיוון שהנתונים נאספו ובוצע עליהם כבר ניתוח מקדים לא נשנה את
SCC	ניטור צריכת קלוריות	category		no 2015 yes 96			ללא
FAF	תדירות פעילות גופנית	רציף	ימים	מ - 0 עד -3	1.010298	0.8505924	אבל מכיוון שהנתונים נאספו ובוצע עליהם כבר ניתוח מקדים לא נשנה את
TUE	זמן שימוש במכשירי טכנולוגיה	רציף	ערך מייצג עבור קטגוריות(מספר שעות ביחס לקטגוריה כאשר biningה לשעות הוא [0-2,3-5,5+]	מ - 0 עד -2	0.657866	0.6089273	אבל מכיוון שהנתונים נאספו ובוצע עליהם כבר ניתוח מקדים לא נשנה את
CALC	תדירות צריכת אלכוהול	category		1401 no 639 Frequently 70			ללא
MTRANS	שימוש בתחבורה	category		tation 1580 Automobile 457 Walking 56			ללא
NObeyesdad	קטלוג סופי	category	בהתאם למתואר במאמר	351 Obesity_Type_III 324 Obesity_Type_II 297 Overweight_Level_I 290			ללא

סעיף ג'

בהמשך לסעיפים א ו-ב , הגדירו ותארו את שלבי ה- KDD עבור הבעיה הנתונה.

שלבי ה-KDD :

1. הגדרת המטרות של כריית המידע –
   א. מטרת כריית המידע היא לחזות את דרגת ההשמנה של נבדק מהאוכלוסייה שבאזור מקסיקו פרו וקולומביה באמצעות התכונות מבסיס הנתונים שניתן לנו.

2. איסוף ושמירת הנתונים-
   ב. בחירת סט הנתונים עליו יבוצע התהליך.
   ג. הנתונים התקבלו מ- https://www.kaggle.com/mpwolke/obesity-levels-life-style/data וקובץ ההסבר על הנתונים שבמאגר המידע התקבלו מ - https://www.sciencedirect.com/science/article/pii/S2352340919306985?via%3Dihub
   ד. את הקובץ טענתי ל python בעזרת pandas .

3. ניקוי הנתונים –
   ה. בדיקה האם קיימות רשומות בעלות ערכים חסרים .
   ו. בדיקה האם קיימות רשומות בעלות ערכים לא הגיוניים
   ז. אפיון הערכים שהתקבלו בא' וב' והחלטה על אופן הטיפול בהם (ממוצע, החלקה, הורדת הרשומה וכו')
   ח. במאגר הנתונים שהתקבלו לא היו ערכים חסרים(להלן איור 2) או ערכים לא חוקיים/חריגים וקיצוניים(יפורט בהמשך).

4. ביצוע טרנספורמציות על הנתונים –
   ט. שינוי דרך השמירה של המידע.
   א. שינוי ערכי הרשומות לצורה אחרת
   ב. יצירת מאפיינים חדשים בהתאם למטרת הכרייה.

5. בחירת שיטות לכריית מידע (כגון רגרסיה, עצי החלטה וכו)
   י. השוואת אלגוריתמים לפתרון הבעיה (שקף 18 כהדרכה כללית לדרך הפיתרון הרלוונטית.)
   א. בנוגע לבעיה הנ"ל יש לבחון מודלים לסיווג קטגורי וכן מודלים לסיווג רציפים מכיוון שלמרות שהקבוצות ניתנו לנו בשם הערך שהן מייצגות הוא רציף.

6. ביצוע דיסקרטיזציות וסיווג הנתונים
   יא. בחינת אפשרות לביצוע דיסקרטיזציה בהתאם למטרות הכרייה והאלגוריתמים שנשקלים להתבצע (נתונים דיסקרטיים או רציפים וכו'.)

7. הרצת שיטות לכריית מידע שנבחרו
   יב. חלוקה המידע ל-training ו-test בהתאם לגודל מאגר המידע הנתון(1/3, k-fold ,bootstrap)
   יג. יצירת מודלים ע"י הרצת האלגוריתמים מסעיף 4 לעיל עם ורסיות שונות שלהם (גיני / אנטרופיה וכו') על training

8. ניתוח התוצאות :
   יד. בחינת הביצועים של האלגוריתמים בוורסיות השונות שלהם.
   טו. בחינה בעזרת מטריצת ערפול.
   טז. בחינה בעזרת מדדי הערכה.
   יז. בחינת פשטות ודיוק של האלגוריתמים.
   יח. בחינת יעילות המודלים והמסווגים השונים בעזרת עקומת roc .
   יט. החלטה האם המודלים שנבחנו מספקים תוצאה ראויה ואפשר להסיק בעזרתם מסקנות או שנצרך לחזור לשלב טיוב הנתונים/בחירת שיטות כרייה וכו' .

9. הסקת מסקנות :
   כ. החלטה על מודל וורסיה שבעזרתו הסיווג של קריאה חדשה יהיה אופטימלי ככל שניתן
   כא. במקרה שלנו סיווג של קריאה חדשה כך שנדע להעריך את רמת ההשמנה של הקריאה החדשה .

סעיף ד'

בהמשך לסעיפים א ו-ב ערכו סקירה השוואתית לכלל החלופות האפשריות (לפחות 4 חלופות) לביצוע כריית מידע. בתשובתכם יש להתייחס ליתרונות/חסרונות כל אחת מהחלופות בהקשר לבעיה הנתונה.

	יתרונות הסוג	חסרונות הסוג	מדד הפיצול		הסבר כללי	יתרונות	חסרונות	ספציפית בנוגע לבעיה הנתונה
עץ החלטה	• יכול לקבל ערכים קטגוריאלים גם ערכים רציפים. • לא דורש נרמול של data • קל להשבה לנורמים לא מהתחום ולכן לדוגמה במקרה שלנו יהיה קל להעביר מידע זה לצבור ולמנוע השמנת יתר באוכלוסיה. • לא מתחייב שהקשר בין ב' ערכים יהיה ליניארי. • בעזרת לינארי תוצאות תוצאות לא משמעות לחלוטין.	• זמן בביצוע ההחלטות שלו • זמן האימון הנדרש ארוך יחסית גם מרחוק אחרים. • בדרך כלל המסקנות דורשות חישוב ארוך בשביל יתר בגלל לגיע אליהם. • יש נטייה ל overFitting בעיקר בשביל חיזוי ליניארי כים אחרים. • לא מספיק טוב בשביל יחסית גדול לגיע לחיזוי בשביל יחסית גדול מאוד סט אימון • מצריך סט אימון גדול לגיע לתוצאות טובות	**C4.5**	**Gain Ratio**	1. בכל צומת של העץ, בוחר את התכונה שמפצלת בצורה היעילה ביותר את קבוצת הסיווג לקבוצות משנה המפורטות ביותר לסיווג הנוכחי או לאחר. קריטריון הפיצול הוא רווח המידע המנורמל (הפרש באנטרופיה). 2. התכונה עם רווח המידע הנורמלי הגבוה ביותר נבחרת כדי לקבל את ההחלטה. 3. חזרה עבור העלים שנוצרו לעץ לעיל.	- מבצע גיזום ולכן פחות סבירות לoverfitting - מסוגל להתמודד עם ערכים חסרים ופחות נוקשה בהכנת הנתונים שלו. - יכול לבצע חלוקה של העץ עם תכונות שלהם יש עלויות שונות	- יוצר עצי החלטה לא מאוזנים בגלל שתכונותיהם מביאות ל"רווח" הטוב ביותר.	• הגיזום יביא לחיזוי טוב יותר לעומת id3 • יקל עלינו בהכנת המידע שכן לא נדרש להעביר את הנתונים הרציפים לבדידים ולכן יתן חיזוי טוב יותר.
			ID3	**Information gain**	1. חישוב האנטרופיה של כל תכונה של ערכת הנתונים 2. חלוקה לקבוצות משנה באמצעות התכונה שעבורה האנטרופיה המתקבלת היא אידיאלית 3. יצירת עץ החלטה המכיל תכונה זו. 4. חזרה על כל קבוצות המשנה עם התכונות הנותרות		- פתרון לא אופטימלי - אין גיזום בניגוד ל- c4.5 ול- cart - לא תומך בערכים חסרים/רציפים	• יקשה עלינו מאוד את הביצוע מכיוון שלא תומך בערכים רציפים, ולנו בנתונים יש כמה וכמה ערכים כאלו(גיל, משקל וכו')
			Cart	**GINI index**	1. בכל צומת של העץ, בוחר את התכונה שמפצלת בצורה היעילה ביותר את קבוצת הסיווג לקבוצות משנה המפורטות ביותר לסיווג הנוכחי או לאחר. קריטריון הפיצול הוא לפי מדד גיני 2. התכונה עם רווח המידע הנורמלי הגבוה ביותר נבחרת כדי לקבל את ההחלטה. 3. חזרה עבור העלים שנוצרו לעץ לעיל.	- בניגוד ל c4.5 תומך בחיזוי רציף. - לא מחשב ערכות כללים - מבצע גיזום (פחות סבירות לoverfitting) - חסין לערכים חסרים ופחות נוקשה בהכנת הנתונים שלו. - יכול לבצע חלוקה לפי עלויות תכונה שונות	- עץ ההחלטה המתקבל הוא בינארי	• הגיזום יביא לחיזוי טוב לעומת id3 • יקל עלינו בהכנת המידע שכן לא נדרש להעביר את הנתונים הרציפים לבדידים ולכן יתן חיזוי טוב יותר. • אפשרי לבצע חיזוי יותר מדויק מאשר זה שאנו מנסים ליישם בכך שנחזה השמנת יתר שתלוי בbmi וכן בנתונים נוספים שניתנו לנו, ולהגיע למסקנות רלוונטיות יותר(רציפות)
רגרסיה	• אמנון מהיר יותר בחם לעצים • נתון תוצאות בקלקולים לחלקים	• מניח שיש התפלגות נורמלית • רגיש לחריגים בחם בחם לינארי	**לינארית**	**פונקציית השגיאה בריבוע**		- עובד רק אם מדובר בקשר לינארי.	- רגיש מאוד לחריגים	• מכיוון שהבעיה הנתונה היא בדידה הישום של כלי זה יכול להיות אפשרי אבל יקשה עלינו מאוד את התהליך. • מכיוון שיש לנו ערכים רציפים ובדידים התוצאות של עץ החלטה ינטו להיות טובות יותר.
יער אקראי	**יתרונות:** • יכול להתמודד עם נתונים לא מאוזנים. וכן עם מידע חסר מונע overfitting		**חסרונות**: זמן חישוב ארוך יחסית לשאר האלגוריתמים		מטא מסווג (כמה עצי סיווג אקראיים) כאשר כל עץ סיווג תוך שימוש בתת רשימה של מאפיינים(אקראיים) מתוך כלל המאפיינים			יכול להיות יעיל מאוד לבעיה הנתונה בעקבות רמת הפירוט שיש לכל רשומה וכך שבעץ אקראי יבחנו בכל שלב רק מספר תכונות לסיווג ולא כלל התכונות ובעזרת כך להביא לתוצאות מיטביות

ממן 21

סעיף ה'

תארו את שלבי הכנת הנתונים. בתשובתכם יש להתייחס לבעיות באיכות הנתונים כדוגמת טיפול בערכים חסרים, תצוגה גרפית של הנתונים, ניקוי הנתונים, שילוב והמרה של נתונים ועוד.

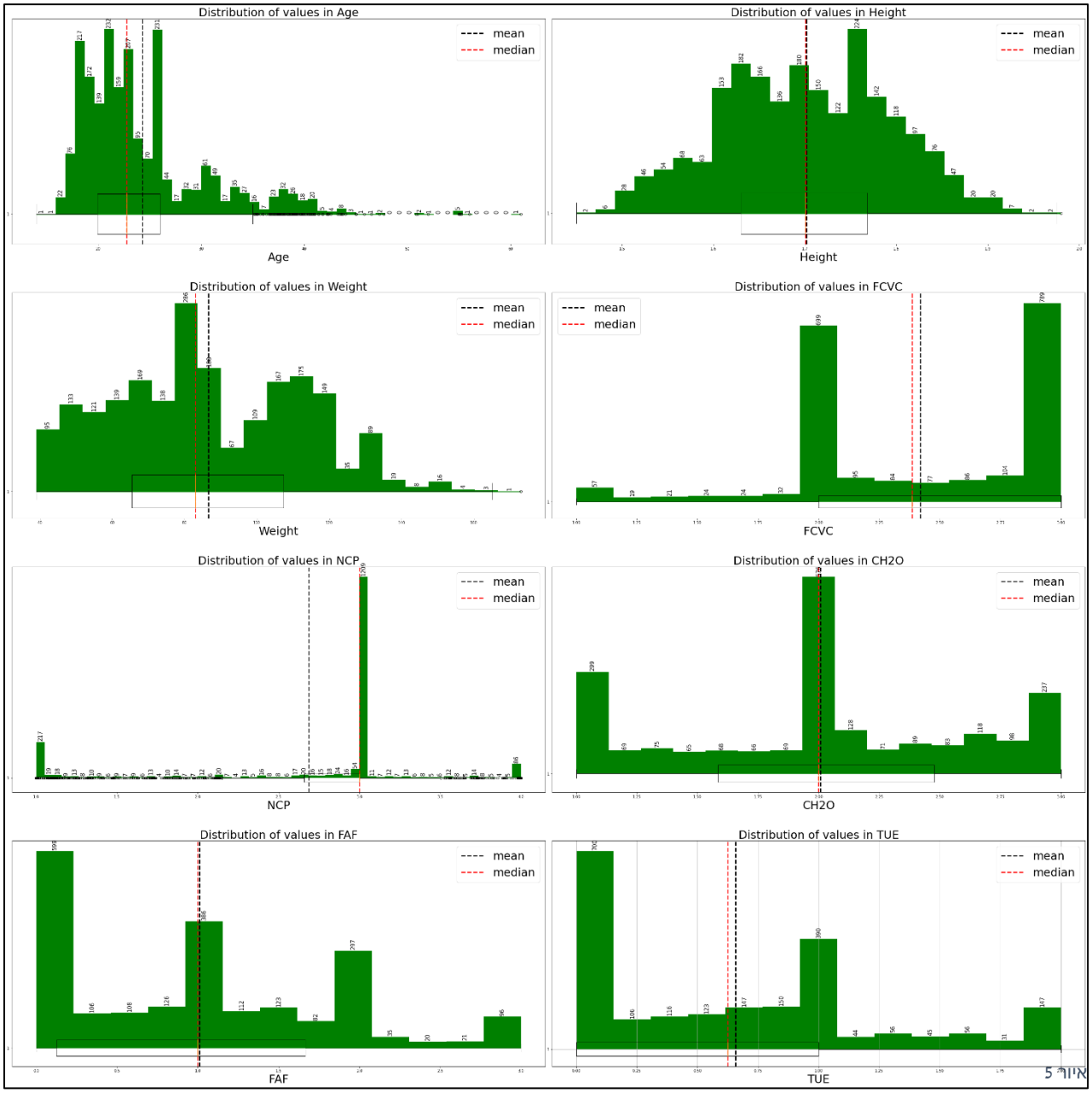
ניקוי הנתונים –

בדיקה האם קיימות רשומות בעלות ערכים חסרים:

בעזרת הפקודה info שניתן להכיל על data frames בpandas קל לראות שאין נתונים חסרים במאגר הנתונים. (להלן איור 2)

בדיקה האם קיימות רשומות בעלות ערכים חריגים:

בעזרת הפקודה describe שניתן להכיל על data frames בpandas נקבל הסתכלות כללית על מאגר המידע והנתונים בו (סטיית תקן, ממוצע, רבעונים וכו' (להלן איור 3) לאחר מכן ע"י שימוש בסולם מדידה של טווח בין-רבעוני (IQR – Interquartile Range) נאתר את הערכים החריגים והקיצוניים בתכונות הרציפות שנאספו ונציג גרפית את הנתונים של IQR(איור 6) על סקאלה אחידה עם גרף ההתפלגות של הנתונים.(איור 5 כללי ואיורים 7 - 11 מפורטים)



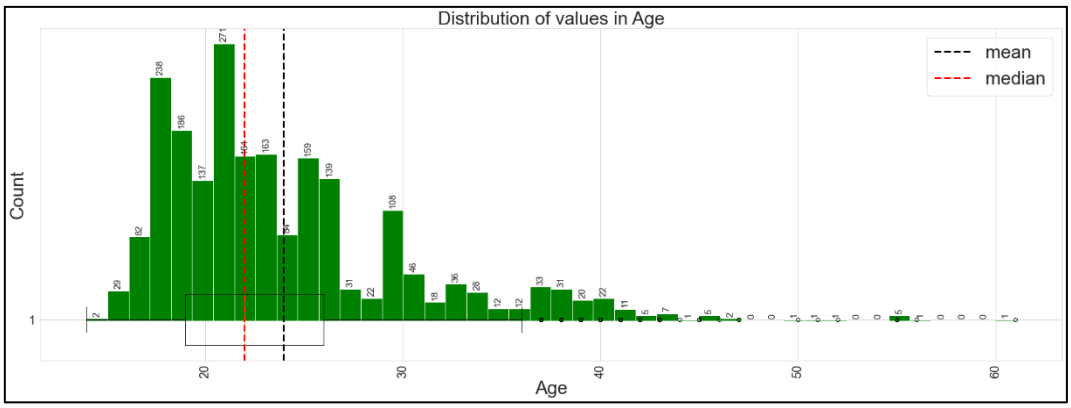
איור 5


Age boxplot
Height boxplot
Weight boxplot
FCVC boxplot
NCP boxplot
CH2O boxplot
FAF boxplot
TUE boxplot

איור6

עבור Age:

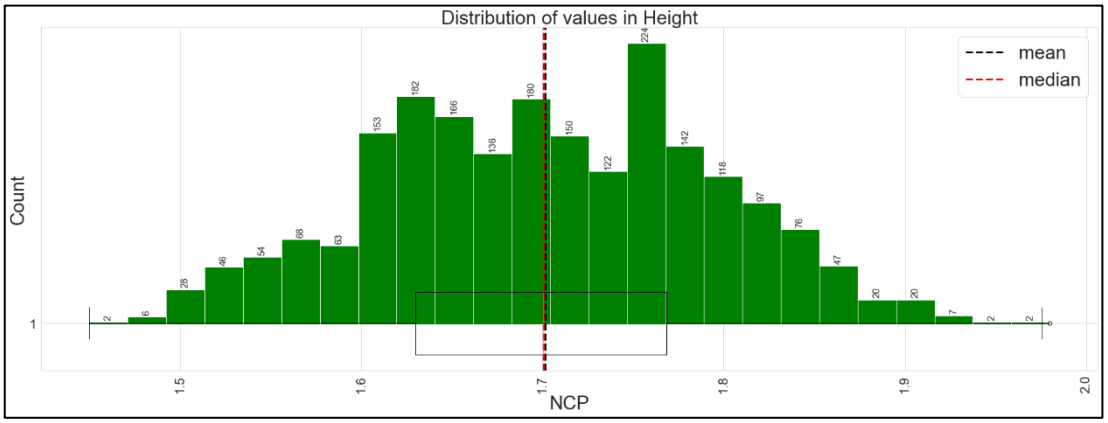
אפשר לראות בקלות שישנה הטייה בהתפלגות הנבדקים וזאת לפי האחוזונים של הנבדקים ומכיוון שבעוד שהחציון הוא 22 הממוצע הוא 23.97 וזאת כאשר אפשר לראות שהאדם המבוגר ביותר בן 61 והצעיר בן 14. אך לא ננקה נתונים אלו מכיוון שהגיוני שמרבית הנבדקים הם "צעירים" אבל נזכור זאת בשביל ההטיה שיכולה לבוא בהמשך בגלל כך)



איור 7

עבור Height :

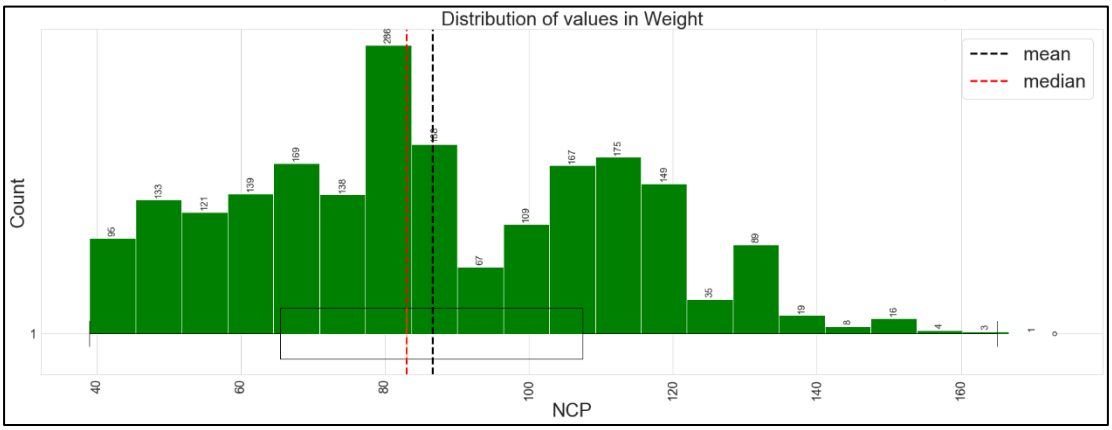
לפי סטיית התקן קל לראות שהנתונים בהתפלגות טובה.



איור 8

עבור Weight:

אפשר לראות שמדובר בערכים בטחום ההגיוני ושיש התפלגות דו מודאלית (סבירה) בנתונים. מכיוון שאנו עתידים להשתמש בעצי החלטה לא נמצע ניקוי נתונים לכך.

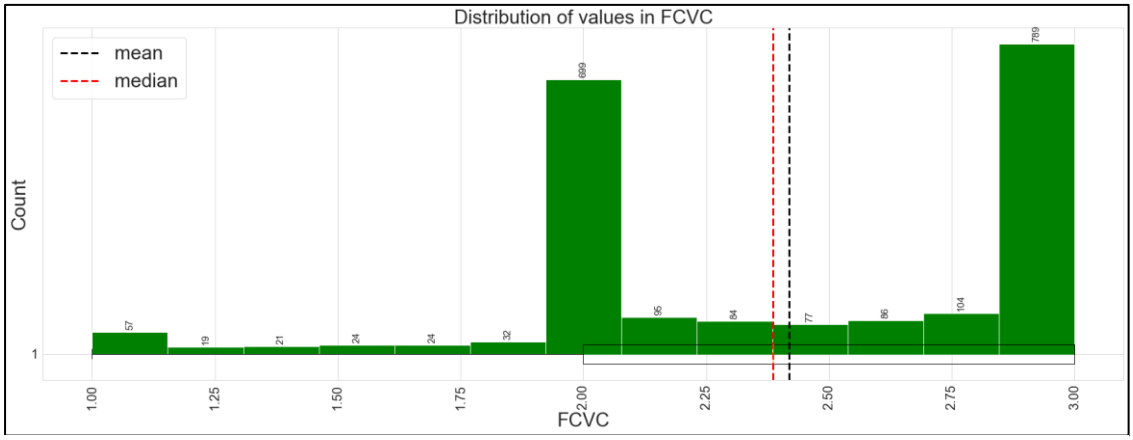


איור 9

ממן 21

עבור Fcvc (תדירות צריכת ירקות):

לפי המאמר המצורף לנתונים התשובות ניתנו בצורה קטגוריאלית ולא בצורה רציפה ולכן הערכים שאינם שלמים הם בעייתים, אולם מפאת היחס הגדול שיש להם במאגר הנתונים נוכל להניח שלא יתכן שמדובר בטעות אנוש וככל הנראה הנתונים שניתנו לנו בעמודה הנ"ל עברו נרמול או תהליך כל שהוא של טיוב. ולכן לא נורידם.



איור 10

לגבי שאר הנתונים שניתנו במאגר המידע כערכים רציפים(איורים 6-5) המקרה דומה מאוד ל FCVC לעיל ולכן בגלל שמקור המאמר הינו אמין ביותר ובגלל שככל הנראה לא מדובר בטעות אנוש ובגלל שכמות המידע שנאלץ להוריד בעקבות הסרת קריאות אלו הוא גדול מאוד, נצא מנקודת הנחה שהנתונים שניתנו לנו בעמודה הנ"ל עברו נרמול או תהליך הכנת נתונים מקדים להכנה שלנו ולכן הם ככל הנראה כבר מ טיובים. ולכן לא נורידם.

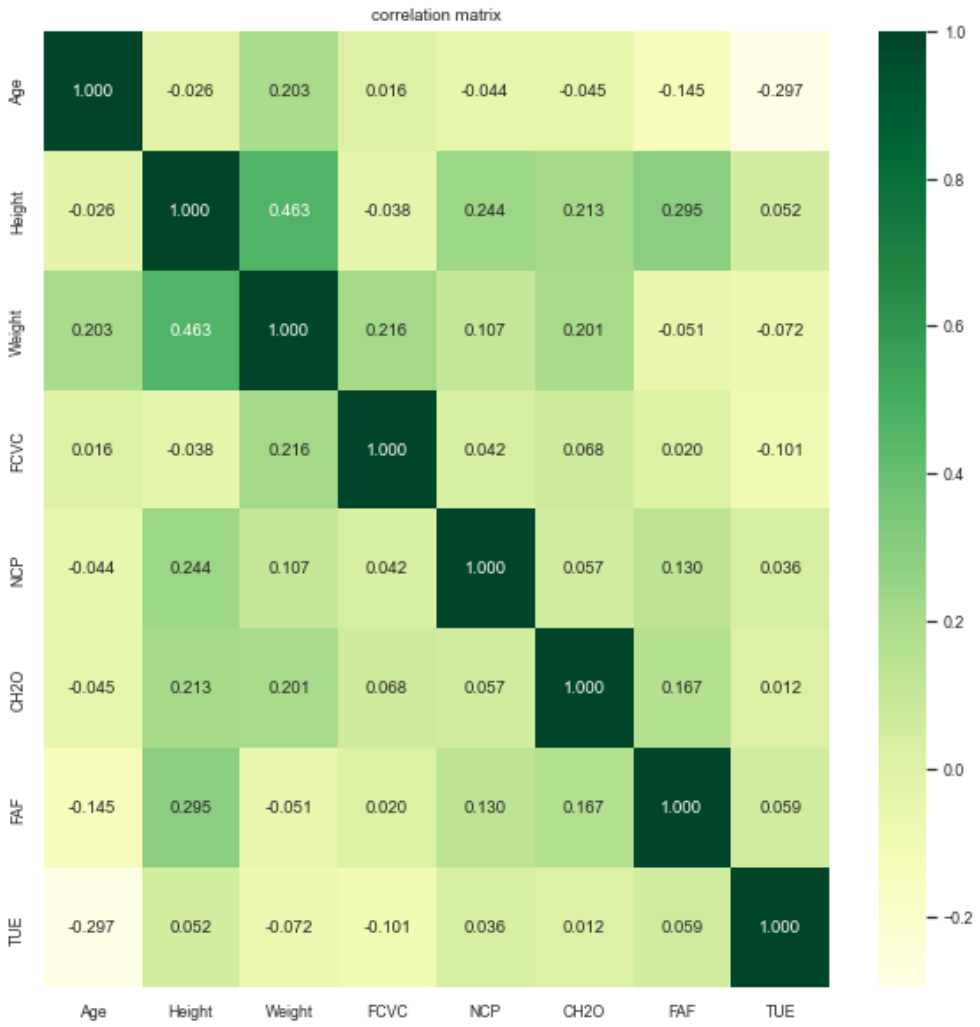
במיוחד מוזר הישום של השאלה אודות תדירות ביצוע הספורט, וכך שאין פירוט אודות אופן רישום הקריאות. שהרי 0 פעילות גופנית הרבה יותר רחוק רעיונית מ כל 5 ימים במרחקם מ כל יום אחד אך במספרים רציפים 0 קרוב יותר ל1 בעוד ש5 רחוק ממנו לאין שיעור, בהיותו התדירות האחרונה שאפשר להזין(ללא אפשרות ה0) .

ממן 21

בדיקה של תכונות מיותרות:

בשביל לבחון האם אפשר לוותר על חלק מהתכונות של מאגר המידע, הדפסתי מטריצת קורלציה כאשר שיטת המקדם היא לפי pearson (איור 11)

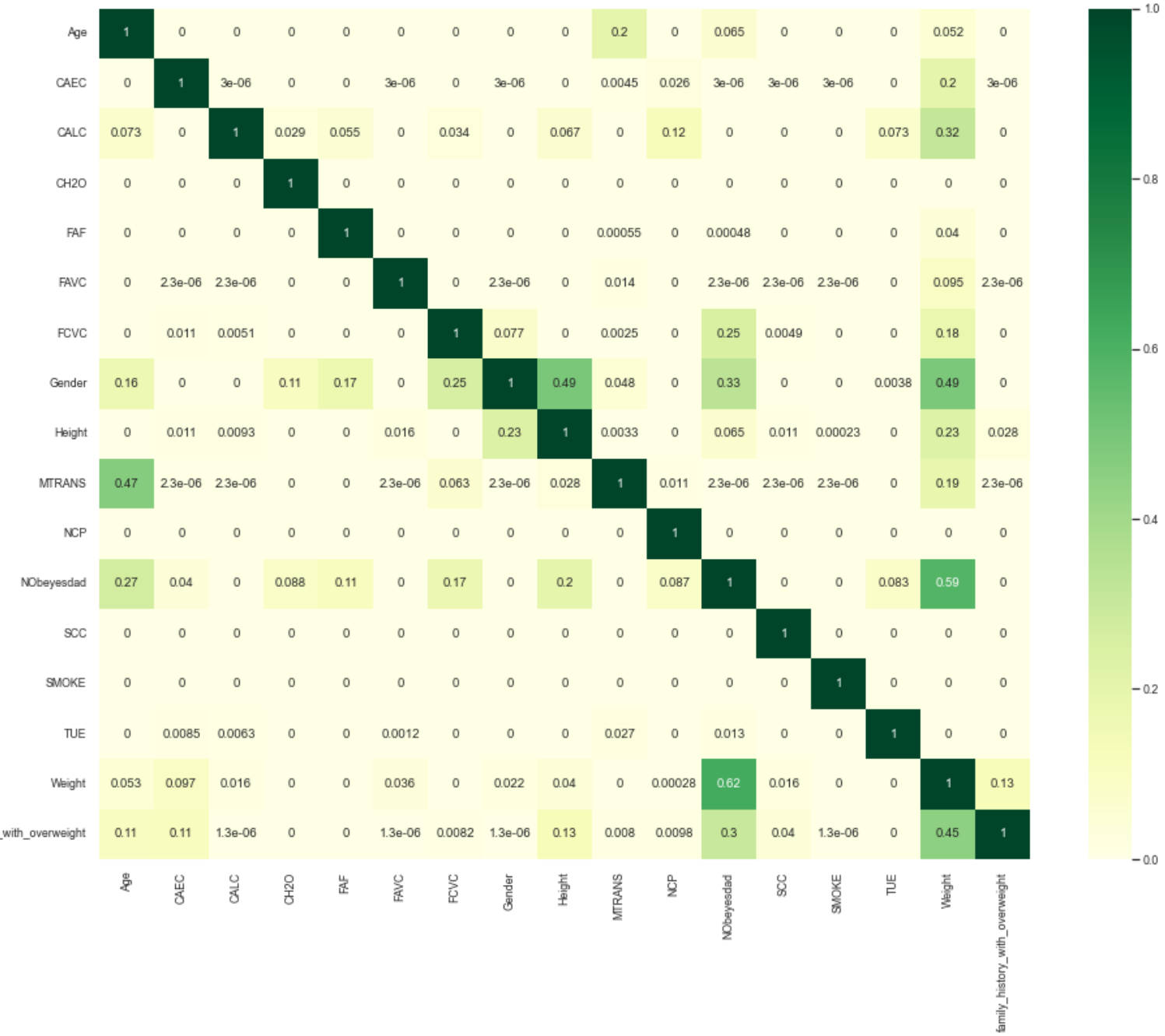
כמו כן ביצעתי בדיקה בעזרת מטריצת "pps" שמבצעת בנוסף לבדיקת ההקשר הרגילה גם בדיקת הקשר על ערכים קטגורים, ויתרונה הנוסף הוא שהיא בודקת קשרים לא לינארים בין ערכים במאגר הנתונים. השימוש בוצע בעזרת python והספריה שמצויה כ"קוד פתוח" בgithub (איור 12)



איור 11

במטריצת הקורלציה אפשר לראות שלמעט הקשר החזק שיש בין משקל לגובה כלל הערכים המספריים אינם תלוים במידה רבה מאוד.

כמו כן אפשר להבין בזכות מטריצת הקורלציה את הגדרת "עודף משקל" של המחקר, וכך שהיא איננה יושבת בקנה אחד עם ההגדרה של BMI .



איור12

במטריצת ה pps מעניין לראות את הקשר החזק של התכונות מגדר/הסטוריה משפחתית של השמנה ותדירות צריכת הירקות להשמנה.

בהקשר לתדירות צריכת הירקות הממצא הנ"ל איננו חד משמעי אך הוא מעניין במיוחד מכיוון שנראה לפי מטריצת הpps שאין קשר בכלל בין משקל האדם לבין היותו צורך ירקות(בזה הסדר) אך ישנו קשר בין היותו של אדם מסווג כבעל משקל עודף(כהגדרת המחקר) להיותו צורך ירקות.

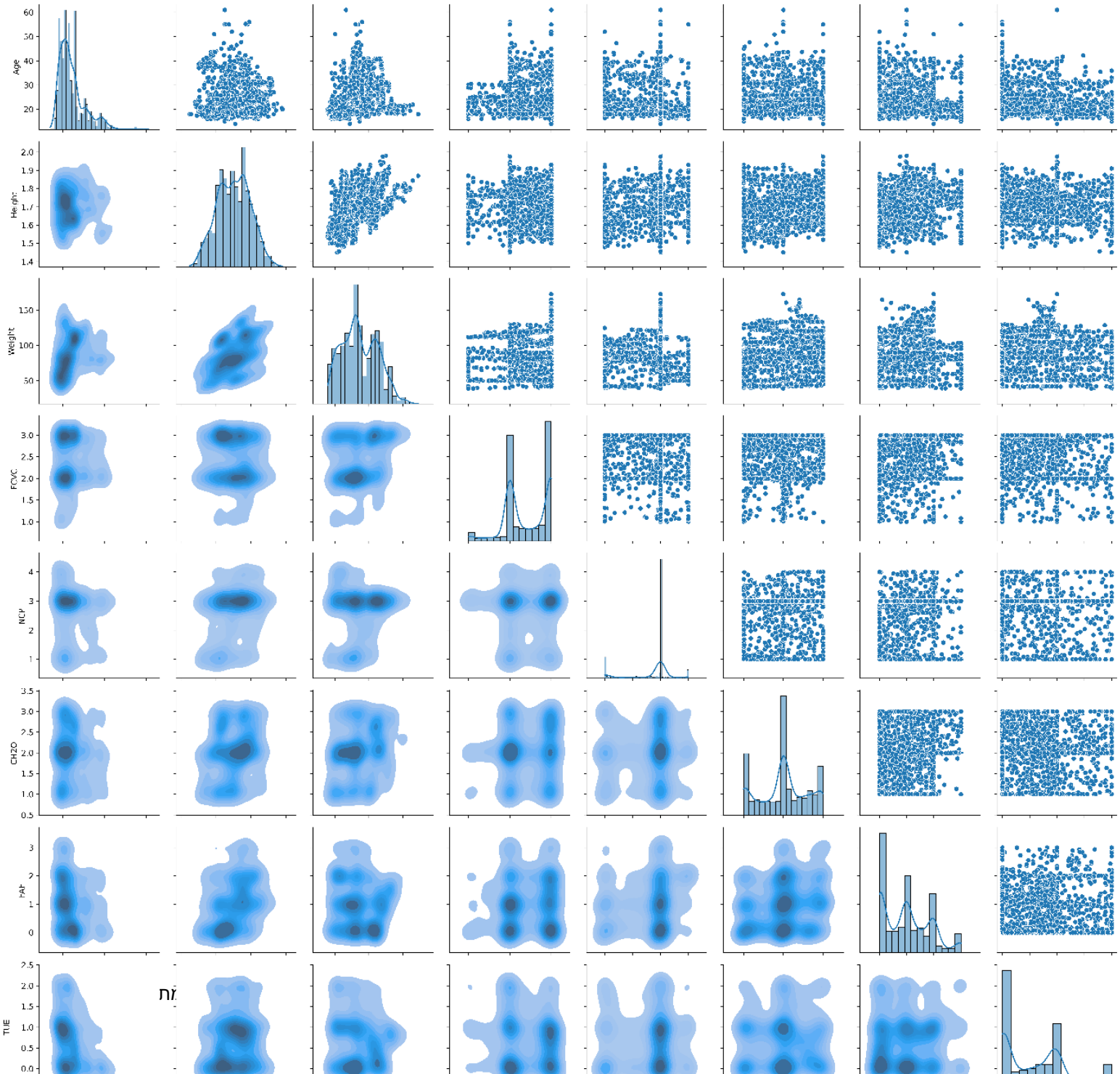
בגלל שכמות הנתונים שיש על "מבוגרים" איננה גדולה קשה להגיד שיש קשר חד משמעי אודות תכונות שנבדקו במחקר אבל אם המידע אודות הגילאים השונים היה רחב יותר כנראה שהיינו יכולים לסמלץ גורם שלישי שייצג את ההשפעה של חלק מהתכונות שהולכות בצמוד לגיל וכך לצמצם את הנתונים שאנו מנתחים. אולם בגלל המחסור בדאטא לא נשתמש בכך.

ממן 21

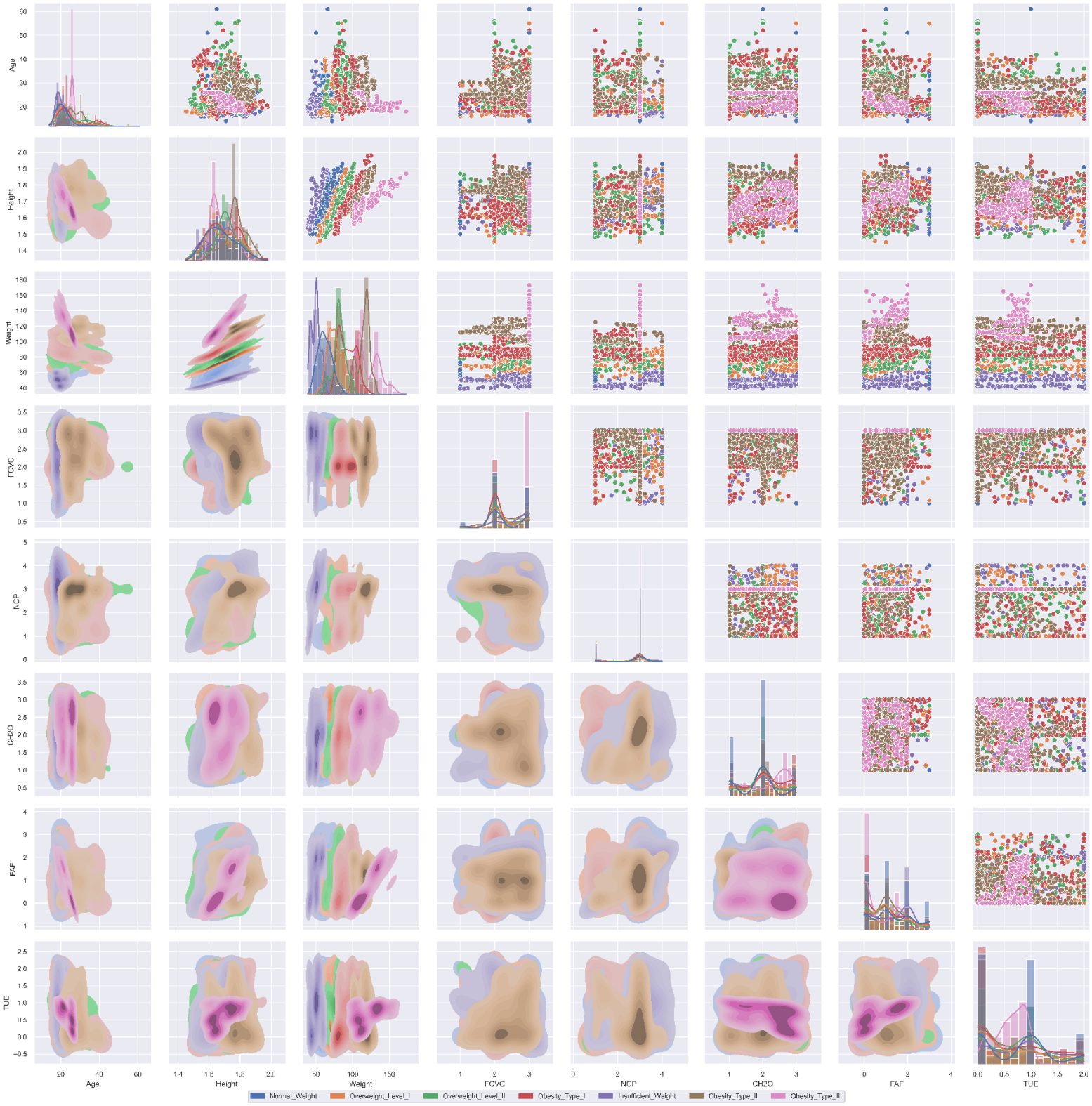
כמו כן בשביל לזהות תתי קבוצות של התפלגויות המידע וזיהוי ערכים חריגים הודפסה "קוביות נתונים" (מימוש של הרעיון של הצגת קוביית נתונים מהרצאותיו של פרופ' מרק לסט).

הסבר אודות "קוביות נתונים"(מימוש של הרעיון של הצגת קוביית נתונים מהרצאותיו של פרופ' מרק לסט) :

באלכסון יש את היסטוגרמות כגרף עמודות וכן הערכת צפיפות עם החלקה שלהם בעזרת ליבה גאוסית שיוצרת עקומה רציפה (KDE), במשולש התחתון יש את התפלגות הצפיפות עם החלקה שלהם בעזרת ליבה גאוסית (kde) כתלות ב2 משתנים. במשולש העליון יש גרף פיזור של הנתונים כתלות x ב y בהתאם לעמודות ולשורות.
בקבוצת הגרפים השנייה הדבר חוזר על עצמו עם הפרדה לפי רמות השמנה שדווחו כחלק מהמחקר.



גם בגרף הבא יש את אותו ניתוח של הנתונים עם חלוקה לפי רמות השמנה (מקרא צבעים בתחתית התרשים) :



כמו כן ננבחנה האופציה לוותר על נתונים בעזרת הפונקציות שמצויות ב feature selection שבספריה sklearn. אולם בשל הסיבות לעיל (על איתור החריגים) ומכיוון שבראיה כוללת, סט הנתונים איננו גדול כך שהדבר ישפיע על מהירות הריצה של האלגוריתמים במטלה זו החלטתי להשאיר את הנתונים.

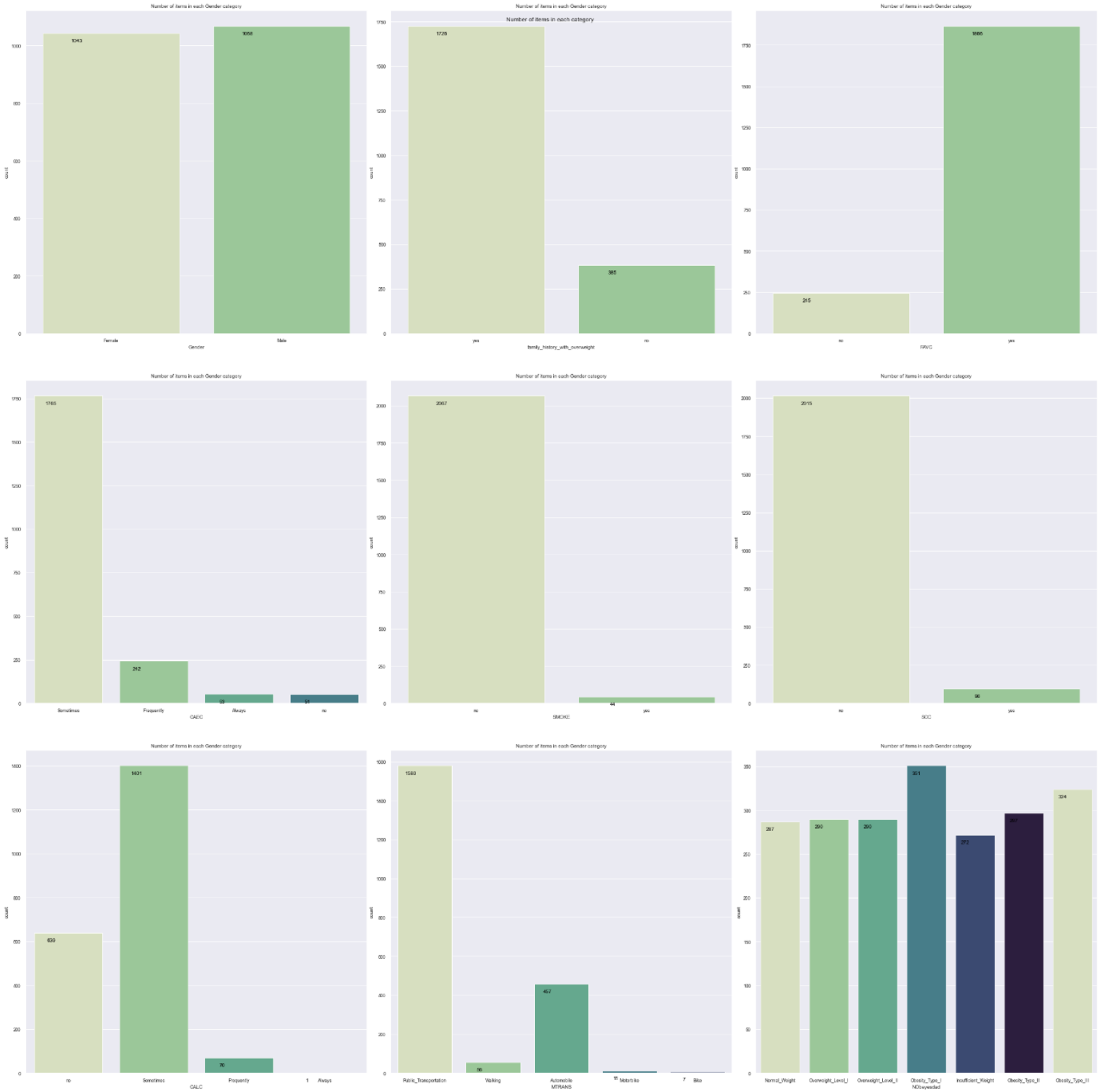
בדיקה האם קיימות רשומות בעלות ערכים לא הגיוניים:

את הבדיקה לקיום רשומות עם ערכים לא חוקיים ביצעתי במקביל לבדיקות לעיל. כאשר הדרך הנוחה ביותר לבדיקה הייתה בעזרת הפקודה describe (להלן איור 3)
לא נמצאו במאגר המידע רשומות עם תכונות לא הגיוניות למעט מה שכבר פירטתי לעיל.

ביצוע טרנספורמציות על הנתונים –

שינוי ערכי הרשומות לצורה אחרת

להלן גרף אודות חלוקת התכונות לקטגוריות פנימות לפני שינוי הרשומות (רק לתכונות דיסקרטיות)



ולכן :

עבור התכונות
family_history_with_overweight, FAVC, SMOKE SCC
ביצעתי המרה של דרך הרישום של המידע שלהן כך ש :
Yes : 1
No : 0
וזאת בשביל לחסוך מקום וניתוח של הנתונים בהמשך.

עבור התכונה Gender:
בוצע שינוי של אופן השמירה של הנתונים כך ש:
Male : 1
Female : 0
וזאת בשביל לחסוך מקום וניתוח של הנתונים בהמשך.

בתכונה MTRANS
ביצעתי טרנספורמציה כך שהערכים Walking , Motorbike , Bike ו- Automobile ירשמו כמחלקה אחת מכיוון שכמות הרשמות שיש לערכים הללו היא אפסית ביחס ל Public_Transportation והשארה שלהם כקטגוריה יכולה ליצור overfiting.
ולכן מעטה והלאה ערכי התכונה MTRANS ייוצגו כך ש :

Public_Transportation : 1

walking,Motorbike, Bike : 0

בתכונה CALC
ביצעתי טרנספורמציה כך שהערך Always עם Frequently. בגלל שלערך Always יש רק אזכור אחד והשארה שלו כקטגוריה יכולה ליצור overfiting (לא ביצעתי איחוד עם No מכיוון שהם מייצגים ערכים קטגוריאלים סדורים ולכן למרות ההתפלגות לא נבצע דיסקרטיזציה לאיחודם).

בתכונה Age
ביצעתי טרנספורמציה כך שהערך יעוגל לערך השלם הקרוב ביותר(בעזרת numpy around)

בתכונת Weight ו- Height :
ביצעתי טרנספורמציה כך שהערך יעוגל לערך עם דיוק של 2 ספרות אחרי הנקודה (בעזרת numpy around)

ביצעתי המרה של כלל התכונות(ללא הסיווג) לערך שהוא דצימלי בשביל טיוב של האלגוריתמים ומניעת הטייה של האלגוריתמים בעקבות טווחים שונים של נתונים.

הסיבה שלא בוצעה דיסקרטיזציה של התכונות גיל, גובה, משקל וכו' היא הממצאים שנצפו בעזרת מטריצת הpps שמראה שישנו הבדל בין ההשפה של כל אחת מהתכונות על הקריאה של ערך ביניים שיכול להשפיע על הערך לחיזו(קשר לא לינארי)

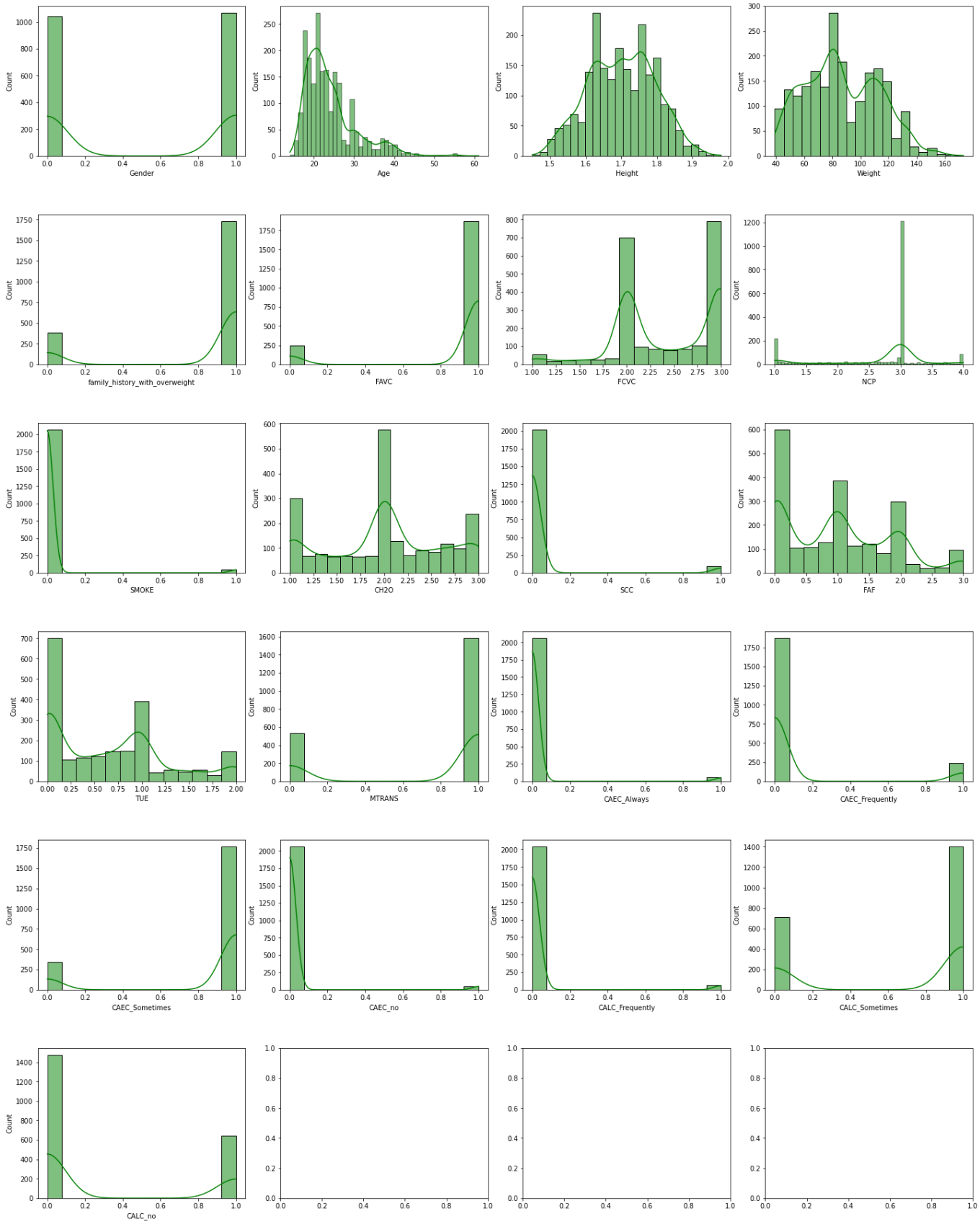
ולכן הדיסרטיזציה המינימלית הייתה עיגול הנתונים (+-0.5 לכל ערך) והשארתם בצורתם המקורית.

ממן 21

שינוי דרך השמירה של המידע.

בשביל לסווג נכונה את הערכים הנומינלים(קטגוריאלים) שיש לנו במאגר המידע, ביצעתי המרה של התכונות הקטגוריאליות הללו לתכונות בינאריות (שייוצגו 0 כאשר אין התכונה מתקיימת ו1 כאשר התכונה מתקיימת.)

ולכן כעת נקבל :

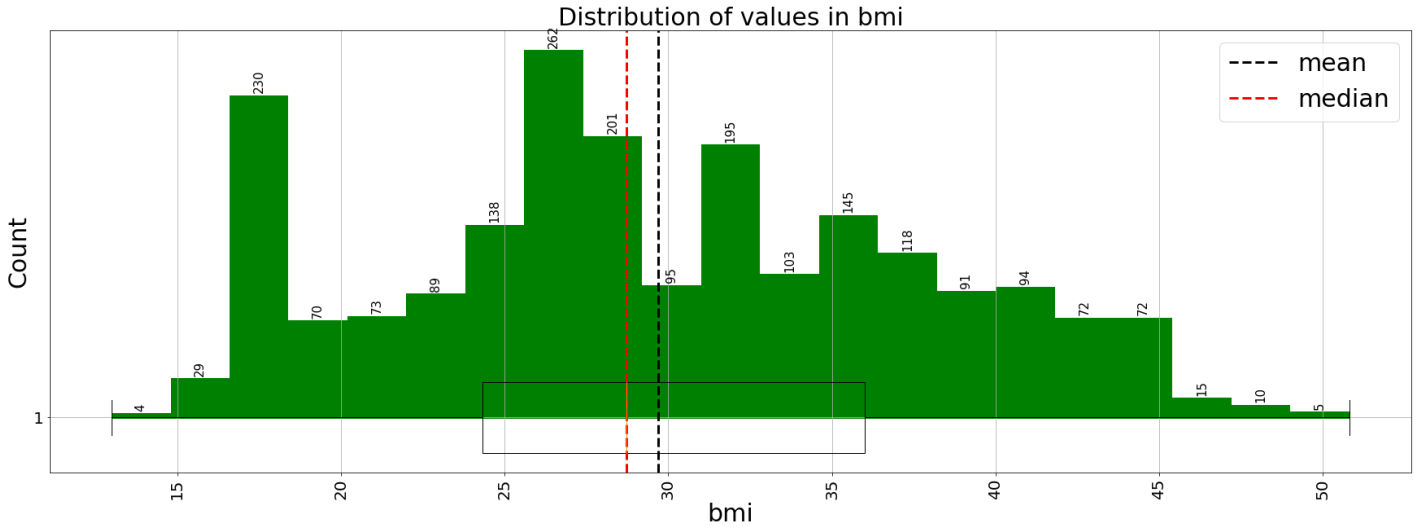


ממן 21

יצירת מאפיינים חדשים בהתאם למטרת הכרייה:

בהתאם להגדרת המחקר הוספתי משתנה חדש שיהיה החישוב של bmi כהגדרתו במחקר. לעניות דעתי ערך זה יצליח לחזות בצורה טובה את קטגורית היעד

התפלגות הbmi לפי רוחב יכולה לתת טיפה מושג על התפלגות הנתונים במאגר. (להלן בגרף המצורף)



שאלה 2:

סעיף א'

<u>בחרו שתי שיטות לסיווג הנתונים. הסבירו את השיטות ונמקו את בחירתכם.</u>

השיטות לכריית המידע שנבחרו הן מימוש עץ החלטה מסוג Cart ומימוש יער אקראי .

מיותר לציין שאם נכון לבחור כאלגוריתם לכרייה אלגוריתם שמשתמש בעץ החלטה, אז בהכרח השימוש ביער אקראי שלוקח בחשבון מספר עצי החלטה היא ככל הנראה הבחירה הטובה ביותר אם אל מול עינניו רק הרצון לקבל חיזוי נכון (בנוגע לזמן החישוב אכן זהו חיסרון שיש לקחת בחשבון אך מכיוון שישנו זמן רב יחסית למימוש המטלה אין זה טיעון לחוסר מימושו)

בנוסף לכך, מכיוון שמבין האלגוריתמים של עצי החלטה לאלגוריתם של cart יש מימוש מובנה בחבילה sklearn אז ה"פשטות" שתהיה במימושו היא מרכיב משמעותי בבחירת אלגוריתם לשימוש במטלה שעלינו להגיש, שהרי הסבר השיטה יהיה פשוט ביחס לאלגוריתם שאצטרך לממש בעזרת ספריות צד שאינן בליבת הקורס, ויכול להיות שישפיעו על הבנת בודק המטלה.

לא השתמשתי ברגרסיה לינארית לפתרון הבעיה הנל מכיוון שבעיית החיזוי שלנו הינה בעיית חיזוי בדידה, וזאת בעוד שרגרסיה רלוונטית לנתונים נומריים רציפים ואצלינו בסט המידע מרבית הערכים בדידים, בנוסף לכך, ראינו בעזרת מריצת ההתאמה – pps שישנם קשרים בין משתנים בקובית המידע אך הקשרים הללו(רובם ככולם) אינם קשרים לינארים(שהרי הם לא הופיעו במטריצת הקורלציה הסמטרית שהובאה גם כן לעיל אך כן הופיעו בpps).

הסיבה שבגללה נפסל השימוש באלגוריתם id3 ובאלגוריתם c4.5 היא חוסר התאמתם לחיזוי ערכים רציפים, וההבנה שלמרות שכרגע נדמה שבעיית הסיווג היא בדידה מאחורי המשתנה הבדיד של רמת ההשמנה עומד משתנה רציף שעשו עליו דיסקרטיזציה, ולכן מכיוון שאיננו יודעים עדיין אם יצליח המודל שלנו לפתור את הבעיות נעדיף לממש מודל שיהיה אפשר לבצע העברת ידע בין המודלים כך שמודל הבסיס הינו מודל עם תמיכה בחיזוי רציף.

סעיף ב'

תארו את שלבי השיטות שבחרתם בסעיף א.

ראשית חילקתי את הדאטא בעזרת train_test_split מהספריה sklearn.model_selection לפי יחס של 70 :30 בהתאם לשיטה המקובלת ובהתאם לכך שראינו שניתן לנו שמאגר הנתונים הוא מאגר מיצג ואפשר ללמוד ממנו בצורה כללית על האוכלוסיה הנגדמת כולה. כמו כן מחיפוש באינטרנט אחר המחקר הסקתי שכ70 אחוז מהדטא שעליו מבוצע את האימון הוא דאטא סינטטי ולכן לאחר ניסוי וטעייה הוחלט לחלק את המידע באחוזים האלו וכך להימנע במידת האפשר מoverfit אבל גם להגיע לתוצאות טובות.)

אלגוריתם CART:

- בכל צומת של העץ, בוחר את התכונה שמפצלת בצורה היעילה ביותר את קבוצת הסיווג לקבוצות משנה המפורטות ביותר לסיווג הנוכחי או לאחר.
- קריטריון הפיצול הוא לפי מדד גיני
- התכונה עם רווח המידע הנורמלי הגבוה ביותר נבחרת כדי לקבל את ההחלטה.
- חזרה עבור העלים שנוצרו לעץ לעיל

- **אלגוריתם יער אקראי:**
  - בונה קבוצה של עצי החלטה אקראיים על חלק מהתכונות(מוגדר על ידינו) .
  - מקבל החלטה בהתאם לתוצאות שהתקבלו מעצי ההחלטה האקראיים שנוצרו.
  - חוזר על התהליך עד שמגיע לתנאי הסף שהגדרנו לו.

סעיף ג' + ד' :

עבור כל שיטה דווחו את תוצאות הניתוחים.

אלגוריתם cart :

מימשתי את אלגוריתם cart בעזרת GridSearchCV שמגיעה עם ספריית sklearn.

התוצאות שהוחזרו מGridSearchCV היו שהפרמטרים הטובים ביותר להרצה הם :

{'criterion': 'entropy', 'splitter': 'best'}

כלומר בכלל לבצע את האלגוריתם של cart עם אנטרופיה כפונקציית הערכה וכאשר

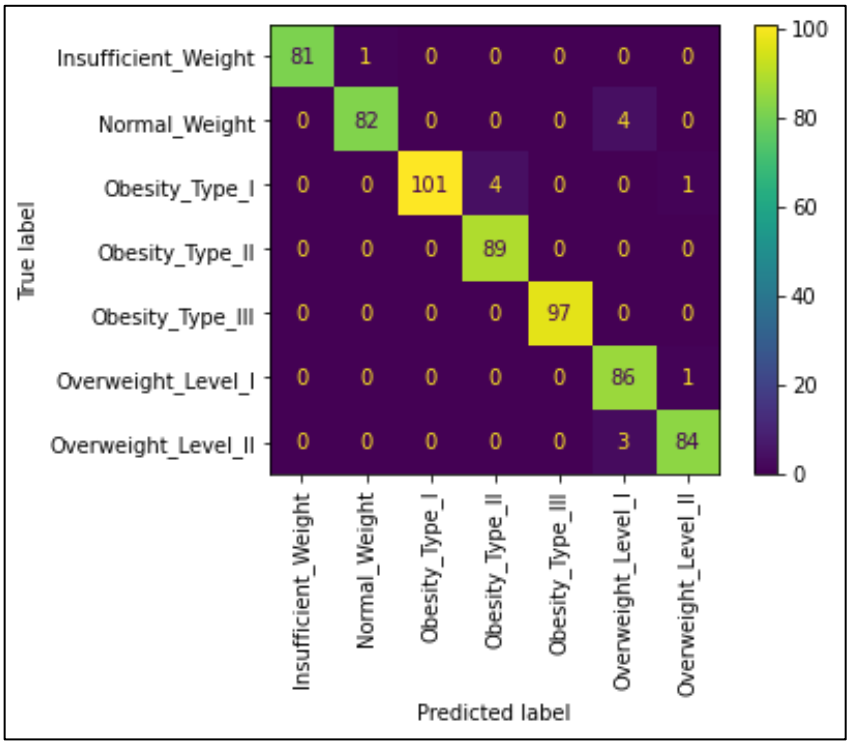
והציון של המודל עם הפרמטרים האלה הוא 0.9687

התוצאות המדויקות של המודל היו :

	precision	recall	f1-score	support
Insufficient_Weight	1.00	0.99	0.99	82
Normal_Weight	0.99	0.95	0.97	86
Obesity_Type_I	1.00	0.95	0.98	106
Obesity_Type_II	0.96	1.00	0.98	89
Obesity_Type_III	1.00	1.00	1.00	97
Overweight_Level_I	0.92	0.99	0.96	87
Overweight_Level_II	0.98	0.97	0.97	87
accuracy			0.98	634
macro avg	0.98	0.98	0.98	634
weighted avg	0.98	0.98	0.98	634

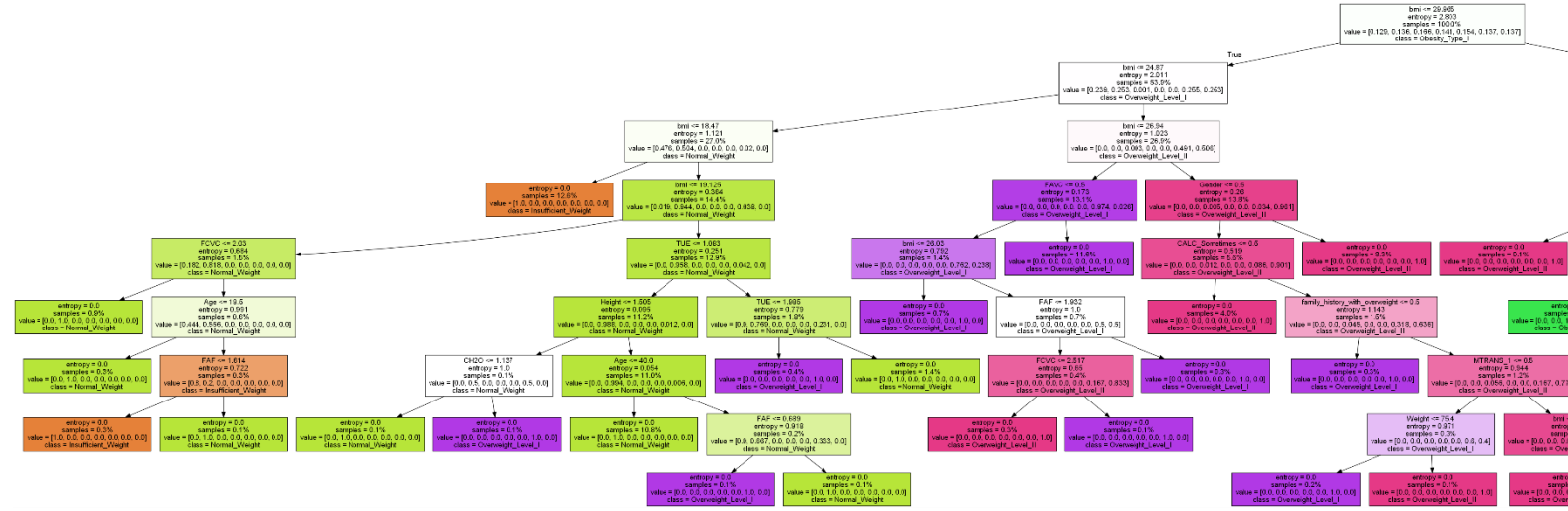
)Precision, Sensitivity, Specifity, Accuracy(

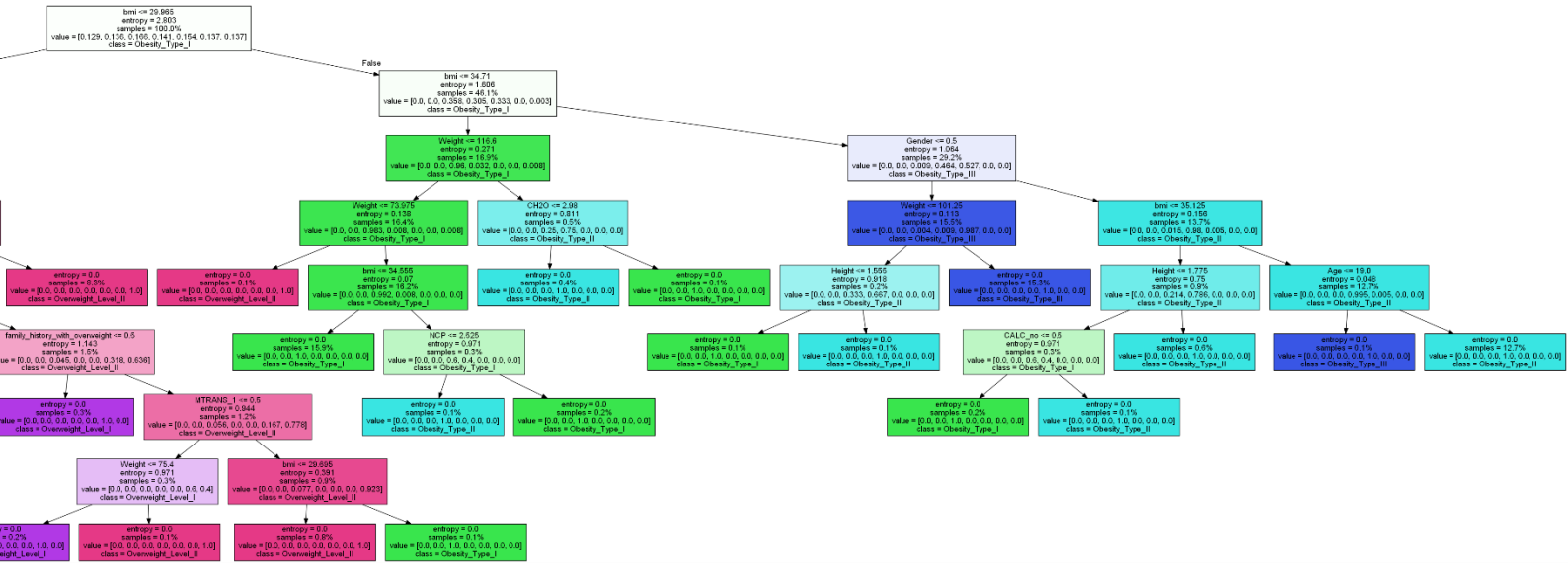
ומטריצת הערפול שהתקבלה היא:



ממן 21

והעץ שהתקבל(מפאת הדף חתכתי אותו לשניים) הוא:





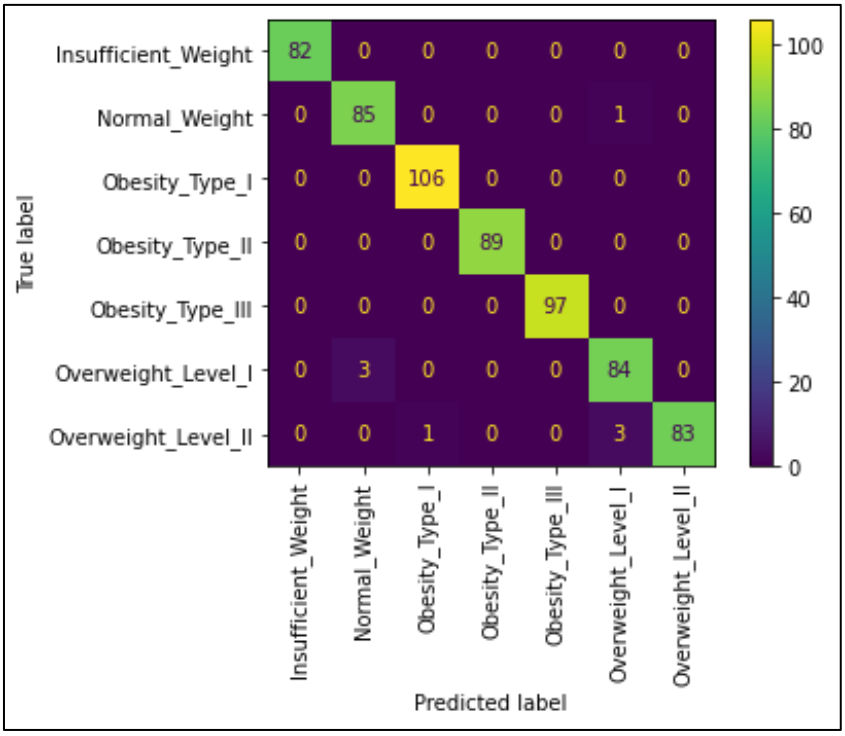
יער אקראי

עבור יער אקראי הגבלתי את כמות הפיצרים בכל החלטה להיות שורש של הכמות הכוללת של הפיצרים שיש לעץ . ובעזרת שימוש בשיטת k-fold Cross Validation (k=10) הציון של המודל היה : **0.9842**

והציונים המדויקים היו :

	precision	recall	f1-score	support
Insufficient_Weight	1.00	1.00	1.00	82
Normal_Weight	0.97	0.99	0.98	86
Obesity_Type_I	0.99	1.00	1.00	106
Obesity_Type_II	1.00	1.00	1.00	89
Obesity_Type_III	1.00	1.00	1.00	97
Overweight_Level_I	0.95	0.97	0.96	87
Overweight_Level_II	1.00	0.95	0.98	87
accuracy			0.99	634
macro avg	0.99	0.99	0.99	634
weighted avg	0.99	0.99	0.99	634

ומטריצת הערפול שהתקבלה היא:



מכיוון שיער אקראי בנוי מהרבה עצים וניתוחו קשה נעזרתי בתכונה feature_importances מ- sklearn

אפשר לראות שהדבר שהכי משפיע על הסיווג של אדם עם עודף משקל/תת משקל הוא ה bmi ואחריו המשקל. מעניין לגלות שתדירות אכילת ירקות משפיעה יותר על היותו של אדם מסווג לקבוצה מסוימת מאשר גובהו של האדם. וכאמור נראה בבירור שהוספת התכונה bmi תרמה רבות למודל.

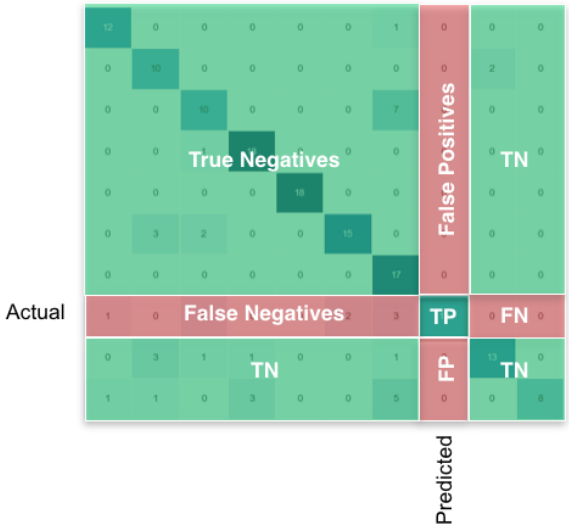
bmi	0.372728
Weight	0.189370
FCVC	0.063176
Height	0.051699
Age	0.050795
Gender	0.040497
NCP	0.036412
FAF	0.028002
TUE	0.027722
CH2O	0.026370
family_history_with_overweight	0.021610
CALC_Sometimes	0.014099
CAEC_Sometimes	0.013939
CALC_no	0.012820
CAEC_Frequently	0.011554
FAVC	0.010899
MTRANS_0	0.008845
MTRANS_1	0.006749
SCC	0.004182
CAEC_no	0.003266
CALC_Frequently	0.003026
CAEC_Always	0.001274
SMOKE	0.000966

סעיף ה'

<u>נתחו השוואתית את התוצאות והסיקו מסקנות כולל הצעות לשיפורים.</u>

ראשית (תוצאות האלגוריתמים מצויות לעיל,) נתחיל בהשוואת הaccuracy בין 2 המודלים שמומשו.
מתקבל שבעשירית האחוז יש יתרון ליער אקראי כאשר לפי cart נגיע ל 98 אחוז הצלחה ולפי יער אקראי נגיע לדיוק עם 99 אחוזי הצלחה.

שנית מכיוון שכאמור במאמר המלווה למאגר המידע, שחלק ממטרות המחקר זה למנוע השמנת יתר / תת משקל באוכלוסיית הבדיקה אז אנו נעדיף אבחנה מחמירה על פני מקלה(עדיף שאדם יהיה בדיאטה מאשר שלא יהיה מודע). ולכן, המדד הכי רלוונטי הוא (Sensitivity) recall, המעיד כמה פעמים צדק המודל שלנו וחזה תשובה טובה של אמת נכונה (true positive rate) מכיוון שמדובר בסיווג רב מחלקתי הסיווג בוצע בהתאם לתמונה הבאה (נלקחה מ- - stackoverflow) :



ולפי כך גם לפי מדד זה נעדיף את היער האקראי על פני cart.

בניגוד לשנתי הסיבות הראשונות, חשוב לזכור שאין בהשוואת התוצאה הזו את הכל כי יער אקראי קשה להסברה וגם המשאב החישובי שהוא משתמש בו רב יותר ולכן לא בהכרח שדיוק בפחות מעשירית האחוז מצדיק את ההשקעה החישובית , ובטח אם על בסיס פרויקט זה הולכת להיות הסברה לציבור הרחב אודות אכילה נבונה שאז יש עדיפות לעץ לפי cart שאותו קל להסביר.

לסיכום לכל אחת מהבחירות יש על מה לסמוך והבחירה באחת מהן תלויה במשתנים רבים בהתאם למטרות הכרייה הנוספות.