1. הגדרת הבעיה והכנת הנתונים

א. הגדרת מטרת כריית המידע

מטרת כריית המידע היא לנבא את רמת השמנה בקרב אנשים ממדינות מקסיקו, פרו וקולומביה על סמך הרגלי אכילה ומצב בריאותי. בפרויקט זה אבנה מודל שלימד מתוך סט הנתונים הנתון כיצד לחזות את רמת ההשמנה.

ב. הגדרת הנתונים

שם תכונה	תאור התכונה	סוג	תחום ערכים	ממוצע	סטיית תקן	ערכים לא חוקיים
Gender	מי	קטגורי	גבר/אישה	-	-	-
Age	גיל	נומרי -רציף	14-44	24.155	5.829	17
Height	גובה	נומרי -רציף	1.45-1.98	1.71	0.091	-
Weight	משקל	נומרי -רציף	39-173	88.952	26.884	-
Family_history_with_overweight	היסטוריית השמנה במשפחה	קטגורי	כן/לא	-	-	-
FAVC	תדירות צריכת מזון בעל ערך קלורי גבוה	קטגורי	כן/לא	-	-	-
FCVC	תדירות צריכת ירקות	נומרי -רציף	1-3	2.419	0.534	-
NCP	מס' הארוחות העיקריות ביום	נומרי -רציף	1-4	2.686	0.778	334
CAEC	צריכת מזון בין ארוחות	קטגורי	לא/לפעמים/ בתדירות גבוהה/ תמיד	-	-	-
SMOKE	האם מעשן	קטגורי	כן/לא	-	-	-
CH2O	צריכת נוזלים ביום (כמות)	נומרי -רציף	1-3	2.008	0.613	-
SCC	בקרה על צריכת הקלוריות	קטגורי	כן/לא	-	-	-
FAF	תדירות של פעילות גופנית	נומרי -רציף	0-3	1.01	0.851	-
TUE	זמן מסך	נומרי -רציף	0-2	0.658	0.609	-
CALC	צריכת אלכוהול	קטגורי	לא/לפעמים/ בתדירות גבוהה/ תמיד	-	-	-
MTRANS	אמצעי תחבורה		רכב/תחב"צ/ אופנוע/אופניים/ הליכה	-	-	-

משתנה המטרה אותו אנו צריכים לנבא הוא רמת ההשמנה (NObeyesdad) שמאופיין :

- Insufficient Weight
- Normal Weight
- Overweight Level I
- Overweight Level II
- Obesity Type I
- Obesity Type II
- Obesity Type III

ג. הגדרות ותיאור שלבי ה- KDD עבור הבעיה:

1. <u>איסוף ושמירת הנתונים:</u>
   - סט הנתונים הורד מ https://www.kaggle.com/mpwolke/obesity-levels-life-style/data
2. <u>ניקוי הנתונים:</u>
   - מעבר על הנתונים וטיפול ברשומות בעלות ערכים חסרים / קיצוניים (outliers).
   - בסט שקיבלנו לא היו ערכים חסרים. נתונים קיצוניים הוסרו.
3. <u>ביצוע טרנספורמציות על הנתונים:</u>
   - ביצוע שינויים על הסט הקיים (הסרת/הוספת/יצירת עמודות) שיתרמו למודל שלנו.
4. <u>בחירת שיטות לכריית מידע:</u>
   - בחינה של האפשרויות השונות כגון רגרסיה/עצי החלטה ומה מתאים למקרה שלנו. כל אחת הינה בעלת משפחות אלגוריתמים שונים ומיועדת לבעיות שונות.
5. <u>ביצוע דיסקרטיזציות וסיווג הנתונים:</u>
   - בחינה של עמודות עליהן נוכל לבצע דיסקרטיזציה כך שביצוע הפעולה על אותן התכונות, תועיל למודל.
6. <u>הרצת שיטות לכריית מידע שנבחרו</u>
   - חלוקת המידע לקבוצת אימון(training) וקבוצת מבחן (test).
   - מס' הרצות של האלגוריתמים שנבחרו על קבוצת האימון ובדיקת הדיוק על קבוצת המבחן לצורך מדידת ההתנהגות וקבלת ביצועים אופטימלים.
7. <u>ניתוח התוצאות והסקת מסקנות:</u>
   - בחינת התוצאות ורמת הדיוק של האלגוריתמים השונים.

ד. בחירת שיטות לכריית מידע

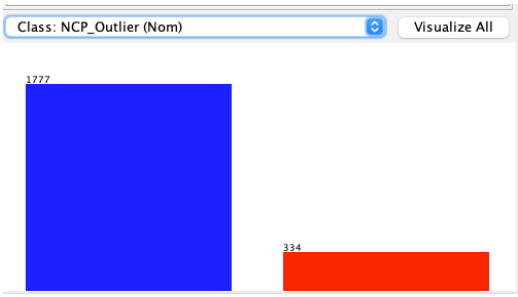
- **רגרסיה לינארית**
  שיטה זו מתאימה למשתנים נומריים כאשר הקשר למשתנה המטרה הוא לינארי.
- **רגרסיה לא לינארית**
  כמו רגרסיה לינארית, משתמשים בה כאשר הנתונים לא מתנהגים לפי מודל לינארי אלא פולינומי.
- **עץ החלטה** J48 – C4.5
  אלג' לבניית עץ החלטה שמבוסס על ID3. משתמש במדד Gain בשונה מIG בID3, מבצע גיזום במקרה וקיימת פגיעה משמעותית כתוצאה מהחלפה של צומת בעלה בעץ ותומך בערכים חסרים.
- **עץ החלטה** Cart – Classification And Regression Tree
  קבוצה של עצי החלטה שמאפשר לבחור את המדד שעל פיו העץ נבנה. משתמש במדד Gini, הפיצול הוא בינרי והאלגוריתם מבצע גיזום למניעת Overfitting .

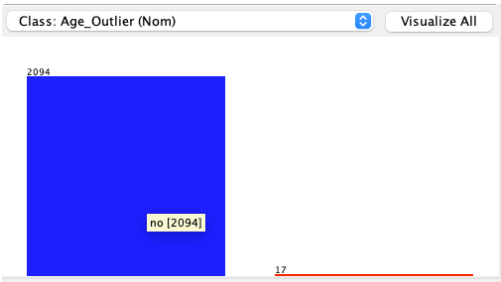
ה. שלבי הכנת הנתונים.

משלב זה ואילך, אתאר ואבצע את תהליך ה Data Preprocessing בעזרת Weka לאחר העמקה בחומר הלימוד ואופן השימוש ב Weka.

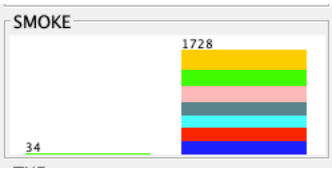
1. ניקוי הנתונים

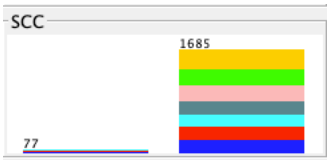
בשלב זה אטפל בערכים חסרים וערכים חריגים. כבר ידוע לנו מסט הנתונים שלא קיימים ערכים חסרים ולכן נעבור לבדיקת ערכים חריגים/קיצוניים. (outliers). את הבדיקה אבצע בעזרת פילטר למידה לא-מודרכת ב Weka בשם Interquartile Range שמאפשר לנו לסמן עבור כל תכונה, האם קיימים בה outliers/extreme values.
את הפלט נוכל למצוא תחת הקובץ outliers detection.arff .
כפי שניתן לראות באיורים מטה, קיימים 334 ערכים חריגים בתכונה NCP ו-17 ערכים חריגים בתכונה Age .
כדי להסיר אותם, אשתמש בפילטר מופע למידה-לא מודרכת בשם RemoveWithValues. בפילטר זה אציין את אינדקס התכונה שבה אני רוצה להסיר את הערכים החריגים ואת אינדקס של 'yes' כלומר הערכים החריגים.
במקרה שלנו, Age_Outlier מופיעה באינדקס 18 וNCP_Outlier מופיע ב- 26
לאחר הסרת הoutliers נקבל צמצום מ2111 ל-1762 רשומות סה"כ.
הקובץ העדכני לאחר צמצום בשם outliersRemoved.arff .

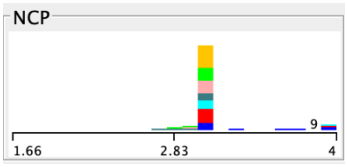




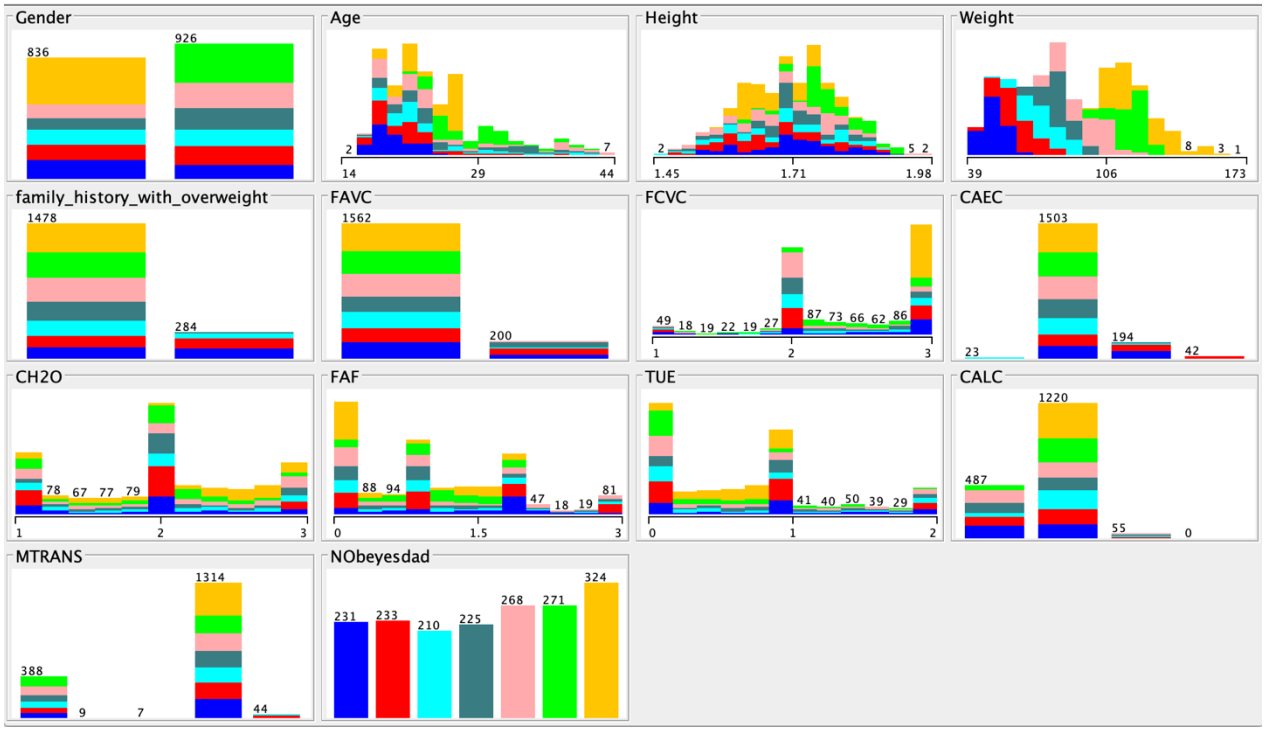
בנוסף, נשים לב שקיימות תכונות שיכולות לגרום לנו ליתר התאמה (overfitting) כתוצאה מהמון תשובות בעלות אותו הערך.
התכונות שהוסרו: NCP, SMOKE, SCC







כעת , זהו המצב הנוכחי של התפלגות רמת ההשמנה ביחס לכל אחת מהתכונות.

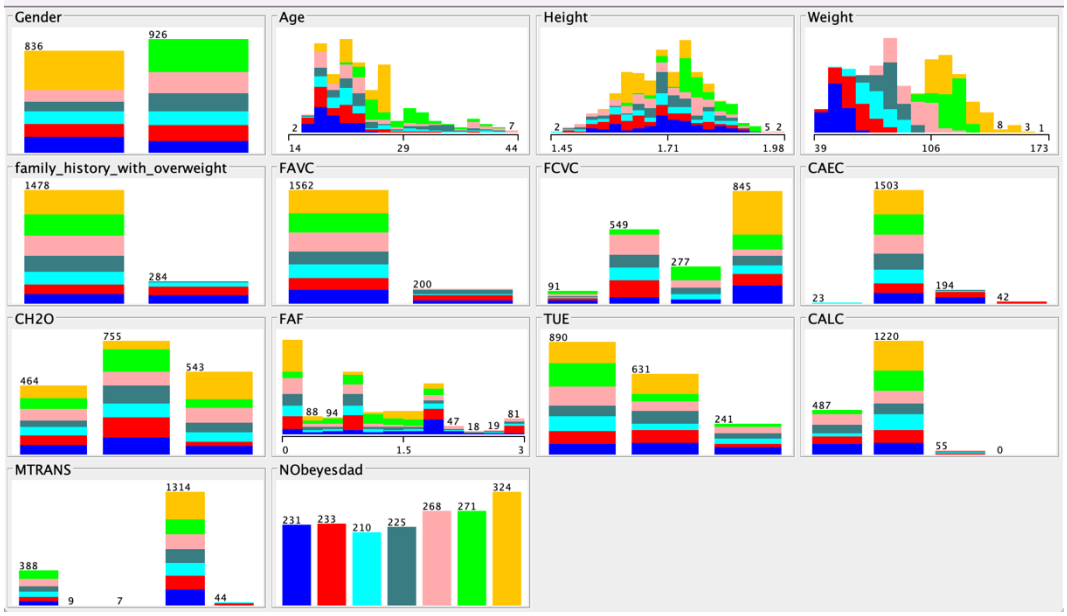


2. <u>ביצוע רדוקציה וטרנספורמציה של הנתונים:</u>

נבצע דיסקרטיזציה של הנתונים ע"י חלוקה למחלקות סדרי גודל שונות כך שכל מחלקה תכיל שילוב של רשומות מטווח מסוים. (שיטת Binning)

נחלק את המאפיינים CH2O,TUE ל-3 תחומים ואת FCVC ל-4 תחומים באמצעות equal depth binning כך שנוכל לסווג בצורה יותר טובה את סוג ההשמנה.

המידע לאחר הכנת הנתונים:



2. סיווג וחיזוי

א. נבחר בשיטות J48, SimpleCart . השיטות הלינאריות משמשות לחיזוי ואילו משתנה המטרה הנתון דורש שיטת סיווג ולכן לא נשתמש בהן.

ב.

J48

נבחן את האלגוריתם במס' הרצות כאשר בכל הרצה נשתמש בפרמטרים שונים. (מקדם ביטחון/מס' רשומות בעלה וכו'..) נבדוק את העץ שנבנה, מס' העלים, רמת הדיות והשגיאה.
אשתמש בשיטת k-fold כאשר k=10.

CART

נבחן את האלגוריתם במס' הרצות כאשר בכל הרצה נשתמש בפרמטרים שונים.(עם/בלי גיזום וכו..)
נבדוק את העץ שנבנה, מס' העלים, רמת הדיות והשגיאה.
אשתמש בשיטת k-fold כאשר k=10.

ג. תוצאות הדיווחים:

RUN	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Tree size	Number of Leaves/Nodes
J48 (Unpruned)	0.938	0.009	0.939	0.938	0.938	0.978	112	62
J48 (Binary Splits = True)	0.941	0.009	0.941	0.941	0.941	0.977	99	50
J48 (Pruned, Binary Splits = False)	0.943	0.009	0.943	0.943	0.943	0.976	146	85
J48 (Unpruned, Binary Splits = False, MinLeaf=4)	0.924	0.012	0.924	0.924	0.924	0.979	91	49
SimpleCart (Pruned)	0.936	0.010	0.936	0.936	0.936	0.979	101	51
SimpleCart (Unpruned)	0.935	0.010	0.935	0.935	0.935	0.977	147	74
SimpleCart (Pruned, MinLeaf=4)	0.921	0.012	0.921	0.921	0.921	0.981	91	46
SimpleCart (Unpruned, MinLeaf=4)	0.924	0.012	0.924	0.924	0.924	0.982	109	55

קיבלנו תוצאות ROC טובות כאשר:
J48 ללא גיזום וללא פיצול בינרי וגם כאשר minleaf=4
ו- Cart עם גיזום ו minleaf=4 .

ד. <u>הערכת מידת הדיוק</u>

נבנה confusion matrix עבור כל אחת מהשיטות ונציג את המדדים :

J48 (Unpruned, Binary Splits = False, MinLeaf=4)

Actual \ Predicted	a	b	c	d	e	f	g	Classified as
a	219	12	0	0	0	0	0	a = Insufficient_Weight
b	13	198	22	0	0	0	0	b = Normal_Weight
c	0	16	176	17	1	0	0	c = Overweight_Level_I
d	0	1	10	200	14	0	0	d = Overweight_Level_II
e	0	0	3	6	255	4	0	e = Obesity_Type_I
f	0	0	0	0	14	257	0	f = Obesity_Type_II
g	0	0	0	0	1	0	323	g = Obesity_Type_III
Total	232	227	211	223	285	261	323	

$$\text{Accuracy Rate } = \frac{219+198+176+200+255+257+323}{1762} = 92.395\%$$

$$\text{Error Rate} = 1 - \text{Accuracy Rate} = 1 - 0.92395 = 7.605\%$$

$$\text{Precision} = \frac{0.924}{0.924+0.012} = 98.717\%$$

SimpleCart (Pruned,MinLeaf=4)

Predicted / Actual	a	b	c	d	e	f	g	Classified as
a	222	9	0	0	0	0	0	a = Insufficient_Weight
b	23	185	25	0	0	0	0	b = Normal_Weight
c	0	24	176	10	0	0	0	c = Overweight_Level_I
d	0	1	4	209	11	0	0	d = Overweight_Level_II
e	0	0	4	8	248	8	0	e = Obesity_Type_I
f	0	0	0	0	11	260	0	f = Obesity_Type_II
g	0	0	0	0	1	0	323	g = Obesity_Type_III
Total	245	219	209	227	271	268	323	

$$\text{Accuracy Rate } = \frac{222+185+176+209+248+260+323}{1762} = 92.1112\%$$

$$\text{Error Rate} = 1 - \text{Accuracy Rate} = 1 - 0.92395 = 7.888\%$$

$$\text{Precision} = \frac{0.921}{0.921+0.012} = 98.713\%$$

ה. מסקנות

- הגדלת מספר מינימלי לעלה הניבה תוצאות טובות יותר בשני האלגוריתמים
- האלגוריתם Cart עם גיזום נותן תוצאות טובות יותר מאשר ללא גיזום
- האלגוריתם J48 עובד טוב יותר בלי פיצולים בינאריים וללא גיזום.
- לשניהם כמעט אותה רמת דיוק (92.111% , 92.395%) .
- ל J48 לקח 0.02sec לבנות את המודל ול- Cart לקח 0.24sec

לאור כל מה שצוין לעיל, נעדיף את J48 לסיווג רמת ההשמנה של אדם.