Predicting the Level of Obesity of a Patient - Part 2 (Association Rules, Clustering, MLP)

שאלה 1 – חוקי הקשר

א. נבחר את האלגוריתם האפריורי. האלגוריתם מוצא את הקבוצות התדירות שעומדות בסף התמיכה המינימלי בצורה אטרקטיבית. מתחילים מהקבוצות התדירות בגודל 1 ובכל איטרציה מוצאים את הקבוצות התדירות הגדולות ב-1 מגודל הקבוצות התדירות באיטרציה הקודמת. כלומר באיטרציה ה-k מאחרים זוגות של קבוצות תדירות מגודל k-1 ומשאירים רק את הקבוצות שעומדות בסף התמיכה המינימלי, עד שלא נוכל לאחד יותר קבוצות תמיכה. בסוף התהליך הזה נייצר מקבוצות התדירות חוקי הקשר. החוקים האלה כבר עומדים בסף התמיכה המינימלי כי הקבוצות עומדות בסף התמיכה המינימלי. אם נשאיר רק את החוקים העומדים בסף הביטחון המינימלי נקבל את כל חוקי ההקשר החזקים.

האלגוריתם מתבסס על תכונת האנטי-מונוטוניות – אם קבוצה אינה שכיחה, אז גם כל קבוצה המכילה אותה אינה שכיחה. בזכות העקרון הזה האלגוריתם יכול לסנן קבוצות שאינן שכיחות בכל איטרציה של האלגוריתם, ולקצר משמעותית את זמן הריצה של חיפוש קבוצות תדירות (לעומת הזמן שהיה לוקח לאלגוריתם הנאיבי).

בחרתי להשתמש באלגוריתם הזה כי הוא קל להבנה ולמעקב, הוא יעיל, ומבטיח שימצא את כל חוקי ההקשר.

ב. אמצע את כל קבוצות התדירות תוך שימוש באלגוריתם אפריורי בעזרת פייתון. בשלב הכנת הנתונים, הורדתי את התכונות של משקל וגובה, כפי שעשיתי בממ״ן 21, כי תכונות אלה קובעות באופן כמעט ישיר את דרגת ההשמנה. לתכונת הגיל עשיתי דיסקרטיצזיה לפי ידע כללי על שלבי החיים ואורחי החיים של אנשים בארה״ב – למשל עד גיל 22 בערך הרוב נמצאים בלימודים, באיזור גיל 28 מתחילים להתייצב בחיים, וכו. במהלך הכנת הנתונים של המאמר, יצרו נתונים בצורה סינטטית ולכן נתונים בדידים הפכו לרציפים – החזרתי אותם לבדידים בהתאם לשאלון המקורי. כדי שיצאו יותר חוקים חזקים, איחדתי ערכים בחלק מהתכונות. בנוסף מאותה סיבה איחדתי ערכים בתכונת המטרה – את Underweight ו- Normal ביחד, את הרמות של Overweight ביחד, ואת הרמות של Obesity ביחד.

לבסוף הוספתי prefix לכל הערכים בבסיס הנתונים לפי שם התכונה של הערך, כדי שחוקי ההקשר יהיו ברורים.

אחרי הרצת האלגוריתם קיבלתי 366 קבוצות תדירות. מצורף קובץ הכולל את כל הקבוצות, אך להלן חלק מהן. ניתן לראות שכולן עומדות בסף תמיכה 40% כנדרש.

	support	itemsets
0	0.465656	(Age_[14-22))
1	0.860256	(CAEC_Never-Sometimes)
2	0.663667	(CALC_Sometimes)
3	0.525817	(CH2O_2)
4	0.883941	(FAVC_yes)
...	...	...
361	0.427286	(SMOKE_no, SCC_no, MTRANS_Public_Transportatio...
362	0.410232	(SMOKE_no, SCC_no, FHO_yes, MTRANS_Public_Tran...
363	0.419706	(SMOKE_no, CALC_Sometimes, SCC_no, FHO_yes, NC...
364	0.429180	(SMOKE_no, SCC_no, FHO_yes, MTRANS_Public_Tran...
365	0.403126	(SMOKE_no, CALC_Sometimes, SCC_no, FHO_yes, NC...

366 rows × 2 columns

ג. קיבלתי סה״כ 3344 חוקי הקשר חזקים. צירפתי במסמך הזה חלק מהם, ובקובץ נפרד אפשר לראות את כולם.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(Age_[14-22))	(MTRANS_Public_Transportation)	0.465656	0.748460	0.409758	0.879959	1.175692	0.061233	2.095452
1	(Age_[14-22))	(SCC_no)	0.465656	0.954524	0.428233	0.919634	0.963448	-0.016247	0.565861
2	(Age_[14-22))	(SMOKE_no)	0.465656	0.979157	0.459972	0.987792	1.008820	0.004021	1.707406
3	(CAEC_Never-Sometimes)	(CALC_Sometimes)	0.860256	0.663667	0.593558	0.689978	1.039646	0.022635	1.084870
4	(CALC_Sometimes)	(CAEC_Never-Sometimes)	0.663667	0.860256	0.593558	0.894361	1.039646	0.022635	1.322849
...	...	...	...	...	...	...	...	...	...
3339	(CAEC_Never-Sometimes, CALC_Sometimes)	(SMOKE_no, SCC_no, FHO_yes, NCP_3, FAVC_yes)	0.593558	0.570346	0.403126	0.679170	1.190804	0.064593	1.339196
3340	(FAVC_yes, CALC_Sometimes)	(SMOKE_no, SCC_no, FHO_yes, NCP_3, CAEC_Never-...	0.607769	0.557082	0.403126	0.663289	1.190649	0.064549	1.315426
3341	(NCP_3, FHO_yes)	(SMOKE_no, CALC_Sometimes, SCC_no, CAEC_Never-...	0.640928	0.532923	0.403126	0.628973	1.180232	0.061561	1.258875
3342	(NCP_3, CAEC_Never-Sometimes)	(SMOKE_no, CALC_Sometimes, SCC_no, FHO_yes, FA...	0.656087	0.489342	0.403126	0.614440	1.255647	0.082076	1.324461
3343	(CALC_Sometimes)	(SMOKE_no, SCC_no, FHO_yes, NCP_3, CAEC_Never-...	0.663667	0.520606	0.403126	0.607423	1.166761	0.057617	1.221146

3344 rows × 9 columns

ד. תוצאות השיטה (חוקי ההקשר החזקים שמנבים דרגת השמנה) מצורפים בהמשך:

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
(CAEC_Never-Sometimes, FHO_yes)	(NObeyesdad_Obesity)	0.740881	0.460445	0.450024	0.607417	1.319194	0.108888	1.374371
(FAVC_yes, CAEC_Never-Sometimes, FHO_yes)	(NObeyesdad_Obesity)	0.683089	0.460445	0.441971	0.647018	1.405201	0.127446	1.528562
(CAEC_Never-Sometimes, FHO_yes, SCC_no)	(NObeyesdad_Obesity)	0.726670	0.460445	0.449076	0.617992	1.342162	0.114485	1.412418
(SMOKE_no, CAEC_Never-Sometimes, FHO_yes)	(NObeyesdad_Obesity)	0.727144	0.460445	0.440076	0.605212	1.314405	0.105266	1.366694
(FAVC_yes, FHO_yes, SCC_no)	(NObeyesdad_Obesity)	0.733775	0.460445	0.447655	0.610071	1.324959	0.109792	1.383726
(FAVC_yes, CAEC_Never-Sometimes, FHO_yes, SCC_no)	(NObeyesdad_Obesity)	0.673614	0.460445	0.441023	0.654712	1.421910	0.130861	1.562621
(SMOKE_no, FAVC_yes, CAEC_Never-Sometimes, FHO...	(NObeyesdad_Obesity)	0.672667	0.460445	0.433444	0.644366	1.399441	0.123718	1.517164
(SMOKE_no, CAEC_Never-Sometimes, FHO_yes, SCC_no)	(NObeyesdad_Obesity)	0.714353	0.460445	0.439602	0.615385	1.336499	0.110681	1.402842
(SMOKE_no, FAVC_yes, FHO_yes, SCC_no)	(NObeyesdad_Obesity)	0.721459	0.460445	0.439602	0.609324	1.323336	0.107410	1.381079
(SMOKE_no, SCC_no, FHO_yes, CAEC_Never-Sometim...	(NObeyesdad_Obesity)	0.663667	0.460445	0.432970	0.652391	1.416870	0.127388	1.552189

ה. ניתן שלראות שכל חוקי ההקשר מסיקים את דרגת השמנה Obesity, שכל החוקים עוברים את סף התמיכה 40% וסף הביטחון 60%, וכי מדד ה lift של כולם מעל 1. נשים לב שלכל החוקים כמעט אותו ערך לכל המדדים כולל support, confidence, lift, leverage, conviction ובצד הגורר של חוקי ההקשר ישנם הרבה תכונות שחוזרות על עצמן. לכן מסיק שישנם תכונות מיותרות בהסקת דרגת השמנה מחוקי הקשר. למשל {CAEC_Never-Sometimes, FHO_yes} וגם {SMOKE_no, CAEC_Never-Sometimes, FHO_yes}. לכן נסיק ש SMOKE_no מיותר בהסקת דרגת ההשמנה. נקבל את החוקים הבאים:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	for_classification
206	(FHO_yes, CAEC_Never-Sometimes)	(NObeyesdad_Obesity)	0.740881	0.460445	0.450024	0.607417	1.319194	0.108888	1.374371	True
804	(FAVC_yes, FHO_yes, CAEC_Never-Sometimes)	(NObeyesdad_Obesity)	0.683089	0.460445	0.441971	0.647018	1.405201	0.127446	1.528562	True
1026	(SCC_no, FHO_yes, CAEC_Never-Sometimes)	(NObeyesdad_Obesity)	0.726670	0.460445	0.449076	0.617992	1.342162	0.114485	1.412418	True
1384	(FAVC_yes, FHO_yes, SCC_no)	(NObeyesdad_Obesity)	0.733775	0.460445	0.447655	0.610071	1.324959	0.109792	1.383726	True
2021	(FAVC_yes, SCC_no, FHO_yes, CAEC_Never-Sometimes)	(NObeyesdad_Obesity)	0.673614	0.460445	0.441023	0.654712	1.421910	0.130861	1.562621	True

התכונות הפחות שכיחות בחוקים הם SCC_no, FAVC_yes אבל אם נסיר אותם "נאבד" את החוק האחרון שהוא עם מדד ה lift הכי גבוה ולכן לא נצמצם יותר מכונות.

החוק הכי מובהק שקיבלנו הוא החוק האחרון, עם support כמעט הכי גבוה, ו confidence, lift, leverage, conviction הכי גבוהים. כלומר התכונות הבאות מנבות דרגת השמנה Obesity:

- Frequent consumption of high caloric foods
- No monitoring of caloric consumption
- A family member is overweight
- Consumption of foods between meals: Never/Sometimes

שלושת הסיבות הראשות נשמעות הגיוניות. הסיבה הרביעית פחות הגיונית. כפי שאני רואה זאת יש מספר אופציות: הראשונה היא שסיבה (CAEC_Never-Sometimes) זו משפיעה הרבה כי חוק ההקשר הרביעי בטבלה אינו כולל אותו ועדיין עם מדדים גבוהים. האופציה השנייה היא אכילה מופחת בין ארוחות גורמת לרעב מוגבר וכתוצאה מכך אכילת יתר ולא מבוקרת בארוחות, מה שגורם להשמנה. ייתכנו גם סיבות אחרות.

שאלה 2 – ניתוח אשכולות

א. ישנם מספר ממדים/שיטות כדי להעריך את האיכות האשכול. הראשון הוא הערכה מוקדמת ונקרא clustering tendency. הוא מנסה להעריך מראש אם שימוש בשיטת אשכול כדאי, כלומר אם הנתונים מחולקים לאשכולות באופן "טבעי". המדד השני זה בחירת מספר האשכולות k האופטימלי. אפשר למצוא את הערך הזה או ע"י ניסוי וטעייה או שיטה שנקראת elbow method בה מריצים את אלגוריתם האשכול מספר פעמים עם ערכים שונים ל-k ובכל הרצה מחשבים ערך לצפיפות ושונות האשכולות. מקבלים גרף של המדד הזה כפונקציה של k. אחרי יצירת האשכולות, אפשר לחשב מספר מדדים כדי לנסות למדוד את איכות האשכולות. ישנם מדדים המודדים את האיכות ההפרדה בין אשכולות ומדדים אחרים המודדים את מידת הקומפקטיות של האשכולות. ממד המדדים ה-Intrinsic נקראה Silhouette Coefficient.

ב. בחרתי באלגוריתם אשכול K-Means. בהנתן מספר אשכולות קבוע (k) האלגוריתם בוחר k נקודות באופן אקראי, האלה ה-centroids (מרכזים) ההתחלתיים של האשכולות. האלגוריתם מחשב לכל נקודה את מרחקה לכל אחד מ-k המרכזים, ומשבצת את הנקודה לאשכול עם המרכז הכי קרוב. כעט נקודת ה-centroid מתעדכנת לפי הממוצע של כל הנקודות ששובצו לאשכול. האלגוריתם חוזר על שלבי שיבוץ הנקודות לאשכולות ועדכון המרכז, עד שה-centroids לא משתנים יותר.

בחרתי באלגוריתם זה כי אפשר לשלוט בכמה אשכולות נרצה שייווצרו וכך יכולתי להתאים את מספר האשכולות למספר הסיווגים – Underweight/Normal, Overweight, Obese (איחדתי סיווגים בתכונת המטרה כפי שעשיתי להכנת הנתונים של חוקי ההקשר). בנוסף בספריית sklearn ניתן להשתמש בשיטות fit ו predict בנפרד ומנגד עם אלגוריתם DBSCAN קיימת רק השיטה של fit_predict כשיטה אחת, מה שלא מאפשר לבדוק את האשכולות על נתוני בדיקה ולהגיע ואחוזי דיוק (להשתמש באשכול כשיטת classification).

ג. השלבים הראשונים של הכנת הנתונים היו זהים לאלה עבור מציאת חוקי הקשר (ראו שאלה 1 סעיף ב). למרות הדמיון בהכנת הנתונים, נדרש לבצע נרמול לכל הנתונים. לשם כך נצרכתי להמיר את כל הערכים לנומריים בעזרת encoding. מכיוון שרוב הנתונים היו בדידים, השתמשתי בשיטת Min-Max. בנוסף מכיוון הערכים של התכונה MTRANS אינם אורדינליים (ordinal) והאלגוריתם K-Means מבוסס מרחק, הערכים של תכונה זו עלולים להטעות את תוצאות האלגוריתם, ולכן הסרתי תכונה זו. לאחר פיצול הדאטה לנתוני אימון ונתוני בדיקה, הרצתי את אלגוריתם האשכול.

התחלתי להעשות fit למודל עם הערך של הפרמטר k=n_clusters=3 כדי שיתאים לשלושת הסיווגים האפשריים - Underweight/Normal, Overweight, Obese. עשיתי predict עם נתוני הבדיקה וקיבלתי שיוך של כל רשומה בנתוני האימון לאשכול שהמרכז שלו הכי קרוב לאותה הרשומה. עשיתי groupby ו count לפי האשכולות, וראיתי שאין אשכול שרוב הנקודות בו הם מסיווג Overweight.

```
   y_valid                     y_pred
0  Obesity                     cluster_1    132
   Overweight                  cluster_1     51
   Insufficient/Normal_Weight  cluster_1     33
1  Obesity                     cluster_2    109
   Overweight                  cluster_2     75
   Insufficient/Normal_Weight  cluster_2     34
2  Insufficient/Normal_Weight  cluster_3     63
   Overweight                  cluster_3     29
   Obesity                     cluster_3      2
dtype: int64
```

העליתי את מספר האשכולות k אז שקיבלתי שלכל דרגת השמנה יש אשכול שרוב השיבוצים בו הם מסוג דרגת ההשמנה הזה. עם k=6 קיבלתי תוצאה כזאת.

```
   y_valid                     y_pred
0  Overweight                  cluster_1     39
   Obesity                     cluster_1     27
   Insufficient/Normal_Weight  cluster_1     11
1  Insufficient/Normal_Weight  cluster_2     23
   Overweight                  cluster_2     16
   Obesity                     cluster_2      2
2  Obesity                     cluster_3     53
   Overweight                  cluster_3     31
   Insufficient/Normal_Weight  cluster_3      8
3  Obesity                     cluster_4     56
   Overweight                  cluster_4     44
   Insufficient/Normal_Weight  cluster_4     26
4  Obesity                     cluster_5    105
   Insufficient/Normal_Weight  cluster_5     22
   Overweight                  cluster_5     12
5  Insufficient/Normal_Weight  cluster_6     40
   Overweight                  cluster_6     13
dtype: int64
```

איחדתי אשכולות שמסווגות לאותה דרגת השמנה כדי לקבל סה״כ 3 אשכולות – אשכול אחד לכל דרגת השמנה.

ד. לסיכום התהליך, יצרתי 3 אשכולות – אשכול אחד לכל דרגת השמנה. ניתן לראות להלן לכל אשכול אכן יש רוב של סיווגים אמתיים מדרגת השמנה אחת. למשל, לאשכול הנקרא Underweight/Normal אכן רוב הרשומות בו עם דרגת השמנה זו. עם נתוני בדיקה בגודל 25% מגודל בסיס הנתונים המקורי, נקבל את הסיווגים הבאים (מקובצים לפי שלושת האשכולות):

```
pred_cluster                y_valid
Insufficient/Normal_Weight  Insufficient/Normal_Weight     63
                            Overweight                     29
                            Obesity                         2
Obesity                     Obesity                       214
                            Overweight                     87
                            Insufficient/Normal_Weight     56
Overweight                  Overweight                     39
                            Obesity                        27
                            Insufficient/Normal_Weight     11
dtype: int64
```

נציג את אותם התוצאות מנתוני הבדיקה ב – confusion matrix:

		Truth		
Predicted		Insufficient/Normal_Weight	Overweight	Obesity
	Insufficient/Normal_Weight	63	29	2
	Overweight	11	39	27
	Obesity	56	87	214

ה. נציג את נתוני הדיוק של התוצאות מנתוני הבדיקה:

```
                                  precision    recall  f1-score   support

Insufficient/Normal_Weight             0.67      0.48      0.56       130
                   Obesity             0.60      0.88      0.71       243
                Overweight             0.51      0.25      0.34       155

                  accuracy                         0.60       528
                 macro avg             0.59      0.54      0.54       528
              weighted avg             0.59      0.60      0.57       528
```

מודל הסיווג בעזרת k-means מגיע ל precision ו-recall לא רעים בכלל – 59% ו 54% בממוצע. ערך ה-recall של סיווג לדרגת השמנה Obesity גבוה במיוחד עם 88%, וזה הגיוני כי זו דרגת ההשמנה הכי גבוהה, והגיוני שיהיה יותר קל לסווג מקרים שיותר "רחוקים" מהנורמה. מסיבה דומה, הגיוני שערכי precision ו-recall של דרגת ההשמנה Overweight הכי נמוכים מבין דרגות ההשמנה האחרות.

שאלה 3 – רשת נוירונים מלאכותית

א. ברשת מסוג MultiLayerPercepton ישנם מספר שכבות של נוירונים עם חיבורים בין כל זוג נוירונים בשכבות סמוכות. השכבה הראשה נקראת Input Layer האחרונה Output Layer ובינתיים לפחות שכבה אחרת מסוג Hidden Layer. במעבר בין שכבה לשכבה הבאה, עבור כל צומת/קודקוד/נוירון בשכבה הבאה מחושב סכום המשקלים בהתאם לערכי הקודקודים מהשכבה הקודמת וערכי המשקלים/חיבורים בין השכבות. מפעילים את פונקציית האקטיבציה על כל סכום משקלים כזה, מוסיפים ערך bias, וזה הערך של אותו הקודקוד בשכבה הבאה.

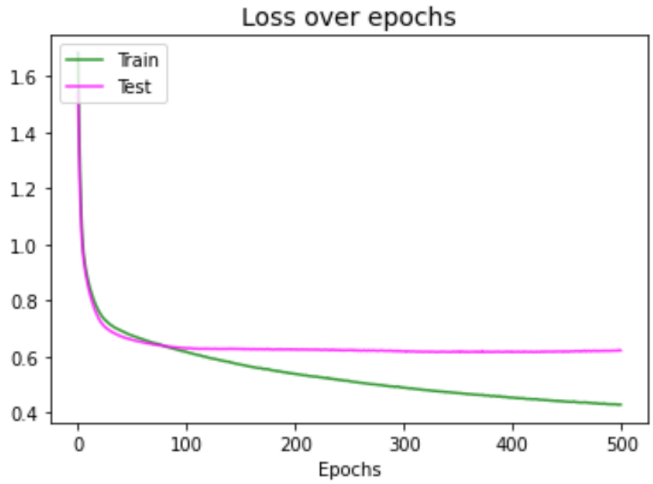
ב. נוכל לקבוע לרשת פרמטרים שונים אשר משפיעים אבל תוצאות ויעילות הרשת. כידוע, בדרך כלל קיים יחס הפוך בין איכות התוצאות ליעילות – כלל שנשמר יעילות האיכות תרד ולהפך.

לאחר העברת כל הנתונים ברשת (epoch) נוכל לחשב ערך שגיאה בעזרת נתוני הבדיקה ופונקציות שונות. עושים זאת כדי לקבל את ההבדל בין ה-predictions של הרשת לערכים האמתיים מנתוני הבדיקה. לאחר כל epoch מעדכנים את הערכים של המשקולות בניסיון להפחית את השגיאה. כאשר מעברים את כל הנתונים לרשת מספר פעמים, נוכל לקבל פונקציית/גרף שגיאה של הרשת כפונקציה של ה-epochs.

מכיוון שהרבה פעמים הפעלת כל הרשת בבת אחת תופסת יותר מידי זיכרון, מעבירים את הנתונים אל הרשת ב-batches (חלקים). ככל שה-batch יותר גדול כך הרשת יכולה ללמוד יותר טוב מהנתונים והדיוק של הרשת עולה.

קצב הלמידה זה גודל הצעדים בו הרשת מעדכנת את המשקולות. ככל שקצב הלמידה גדול/מהיר יותר, כך נסיים לאמן את הרשת יותר מהר, אך הרבה פעמים זה יפגע בדיוק הרשת.

ג. במהלך אימון הרשת לאורך 500 אפוקים, נקבל את הגרף הבא המראה את ערכי השגיאה של נתוני האימונים ונתוני השגיאה כפונקציה של האפוקים:



השתמשתי במודל MLPClassifier של Sklearn, ושתי הפונקציות מתארות את השגיאה עם המדד log-loss. ניתן לראות שמתקיימת דעיכה אקספוננציאלית גם של שגיאת נתוני האימון וגם של שגיאת נתוני הבדיקה, וזה מעיד על כך שהמודל מתכנס. הדעיכה של פונקציות שגיאת נתוני האימון מתכנסת ל-0.4, וזה מעיד על חיזוי כמעט מושלם של המודל. קיימת חפיפה חלקית בין שתי הפונקציות, כאשר אחרי 100 אפוקים נפתח פער בין פונקציית נתוני הבדיקה ונתוני האימון. זה יכול להיגרם מכך שנתוני האימון לא מספיק מייצגים את הבעיה הנתונה, למרות שגודל נתוני הבדיקה היה רק 15% מגודל בסיס הנתונים המקורי.

ד. נסתכל על המקרים בהם בוצע סיווג שגוי:

y_valid	y_pred	
0	1	12
	2	7
1	2	21
	0	16
2	1	12
	0	5

בנתונים לעיל, הספרה 0 מסמלת Underweight/Normal, הספרה 1 מסמלת Overweight, והספרה 2 מסמלת Obesity. ניתן לראות שבוצעו שהכי פחות סיווגים שגויים (17) לדרגת השמנה Obesity, וזה הגיוני כי זו דרגת ההשמנה הכי רחוקה מהנורמה. בוצעו הכי הרבה סיווגים שגויים (37) לדרגת השמנה Overweight, וזה הגיוני כי זו דרגת ההשמנה "שבאמצע" בין שתי דרגות השמנה אחרות.

ה. נציג את נתוני הדיוק של התוצאות מנתוני הבדיקה:

	precision	recall	f1-score	support
0	0.76	0.78	0.77	85
1	0.67	0.56	0.61	85
2	0.82	0.88	0.85	147
accuracy			0.77	317
macro avg	0.75	0.74	0.74	317
weighted avg	0.76	0.77	0.77	317

שוב, הספרה 0 מסמלת Underweight/Normal, הספרה 1 מסמלת Overweight, והספרה 2 מסמלת Obesity. מודל הסיווג בעזרת Multi-Layer Percepton מגיע ל precision ו-recall טובים עד טובים מאוד – 75% ו 74% בממוצע. בדומה לסיווג בעזרת אשכול, ערך ה-recall של סיווג לדרגת השמנה Obesity גבוה במיוחד עם 88%, וזה הגיוני כי זו דרגת ההשמנה הכי גבוהה, והגיוני שיהיה יותר קל לסווג מקרים שיותר "רחוקים" מהנורמה. בניגוד לסיווג בעזרת אשכול, ערך ה-precision של סיווג לדרגת השמנה Obesity גם גבוה במיוחד עם 82%. ושוב בדומה להסברים הקודמיים, הגיוני שערכי precision ו-recall של דרגת ההשמנה Overweight הכי נמוכים מבין דרגות ההשמנה האחרות, כי זו דרגת השמנה שהיא "באמצע" בין שתי דרגות השמנה אחרות. בנוסף יש לציין לחיוב כי שערכי precision ו-recall של דרגת ההשמנה Underweight/Normal גם כן גבוהים.

שאלה 4 – סיכום ומסקנות

אסכם בטבלה את נתוני הדיוק ותוצאות חשובות של כל אחד מהמודלים מממ"ן 21 והממ"ן הזה.

מודל	Weighted avg recall	Weighted avg precision
עץ החלטה	57%	57%
K-Nearest Neighbors	67%	66%
חוקי הקשר	--	--
ניתוח אשכולות	60%	59%
רשת נוירונים	77%	76%

בין המודלים שיש להם precision ו-recall, ניתן לראות שרשת נוירונים מגיעה לתוצאות הכי טובות, עם ציונים של 76-77% למדדים אלה. הציונים הגבוהים של מודל זה מחפים על כך שלא ניתן לתאר את ההחלטות שלו בצורה אנליטית, אלה רק להסתמך על התוצאה שלו כשנותנים לו רשומה חדשה של נתונים.

חוקי ההקשר אמנם לא נמדדים לפי המדדים precision ו-recall, אך התוצאות שקיבלתי עמדו בסף התמיכה, בטחון, ועם ערך lift>1. בנוסף המודל זיהה גורמים ספציפיים אשר חוזים את דרגת ההשמנה Obesity, בניגוד למודלים האשכול ורשת הנוירונים שינם מספקים "הסבר" לסיווג שלהם.

מודל החיזוי בעזרת אשכול אינו הגיע לתוצאות טובות במיוחד. יכול להיות שזה קורה כי לנתונים אין נטייה לאשכול, או כי בעצם השתמשנו בשיטת למידה לא מפוקחת למשימת חיזוי. זה הגביל את אלגוריתמי האשכול שהיה ניתן להשתמש בהם במקרה זה, בגלל דרך פעולה והיישום של האלגוריתמים השונים. למשל, לא היה ניתן להשתמש באשכול DBSCAN כי לא היה לו את המתודות fit ו-predict בנפרד המאפשר לבדוק את המודל על נתוני בדיקה, אלא היה רק fit_predict ביחד.