

ממ"ג 22 – תומר ריפין 322230608

שאלה 1 – חוקי הקשר

A. בחירת האלגוריתם – A-priori

האלגוריתם מוצא קבוצות תדירות באורכים הולכים וגדלים באופן איטרטיבי על ידי בניית קבוצות מעמדות ובדיקה אל מול תנאי הסף של תמייה מינימלית (Minimal Support). בסיום ריצת האלגוריתם ומציאת הקבוצות השכיחות, האלגוריתם מפיק את חוקי ההקשר החזקים שמצוין.

לאלגוריתם זה יתרון משמעותי – הוא קל לימוש וקל להבנה. החסרון המשמעותי שלו הוא הייעילות וזמן הריצה אף מכיוון שמדד הנתונים שלו לא גדול (5000 שורות) זה אינו פקטור משמעותי מספיק.

ב. נניח – Min_confidence 60%, Min_support 40%

נשים לב שכדי לעמוד בסף של Min_support של 40%, ערך בעל שכיחות נמוכה מ-2044 שורות לעולם לא יופיע בפלט של האלגוריתם. נשים לב שהערך של תכונת הסיווג שלנו stroke=1 אינו עומד בסף השכיחות הנ"ל. כמו כן, הדיסקרטיזציה שבצתעה לתכונות מסוימות בממ"ג 21 יוצר קבוצות אשר שכיחותן קטנה מ-40% וכן אבצע דיסקרטיזציה חדש לנתחים כדי שנוכל להסיק ממנה מסקנות בהינתן רף שכיחות זהה. אפרט את ההבנה מחדש של הנתונים שעשיתי –

עבור התכונות הnumerical בצעתי Frequency Discretization עם שני bins כדי להגיע לקבוצות שעבורו הטעינה את הסף.

עבור התכונות הקטגוריאליות רציתי להגיע למצב שבו לפחות לקבוצה דומיננטית יש 40%, בתכונות הבינאריות זה כਮון התאפסה. ולהן נתתי שמות כדי שפלט האלגוריתם יהיה מובן יותר.

בתכונת smoking_status במודול Unknown היה גבולה ורציתי לצמצם אותו משמעותית כדי שלאורח הקבוצות יעברו את הסף. דבר ראשון איחדתי בין smokedformerly smoked לבין smokers. ואחר שאר המודולים מלאתי עם השלמה לקבוצה הדומה ביותר באמצעות אלגוריתם KNN. כך שבסופו של דבר נשארתי רק עם שתי קבוצות – never_smoked ו-smokes.

אסכם את מצב העמדות עבשי:

הטפלות	עמדוֹת
Young – 2618, Old – 2492	Age
Female – 2995, Male - 2115	Gender
Low – 2684, High – 2426	Bmi
Low – 2555, High – 2555	Agl
Never_smoked – 2940, smokes - 2170	Smoking_status
Urban- 2596, Rural - 2514	Residence_type
Private – 2925, self_employed – 819, children – 687, govt_job – 657, never_worked – 22	Work_type
No – 4834, Yes - 276	Heart_disease
Low – 4612, High - 498	Hypertension
Yes – 3353, No - 1757	Ever_married
Healthy – 4861, stroke - 249	Stroke

לאחר מכן, השתמשתי בפונקציה (`pd.get_dummies()` שתווארה בממ"ג 21 ובאמצעותה הפכתי את כל הערכים הקטגוריאליים לעמדות כאשר אם הערך הזה קיים יהיה `True`, אחרת יהיה `False`.

הקבוצות המתדרות שנמצאו באמצעות הרצת האלגוריתם השתמשתי באlgorthm prior-a אשר ממומש מראש בחבילת הפיתוחית mlxtend ולהן תוצאות ההרצה如下:

- נמצאו 81 קבוצות תדרות
- מתחון, 15 קבוצות תדרות בגודל 1, 38 בגודל 2, ו-28 בגודל 3.
- כМОון שלא הייתה קבוצה תדרות עבור המשטנה stroke_Stroke כי היא מתחת לף הביטחון.
- היו 34 קבוצות תדרות שהובילו את המשטנה stroke_Healthy מעמודת המטרה. זה הגיוני בעקבות מפאת חוסר האיזון של תכונה זו.
- ניתן לצפות בכל הקבוצות המתדרות בנספח א.

ג. הציגת חוקי ההקשר חזקים

נחפש תבניות הקשורות לתבונת הסיווג – stroke ולכן אציג רק את חוקי ההקשר שמקלים את stroke_Healthy באחד מהערכיהם הנגררים. אציג עמודות שעומדות ברף $\text{confidence} = 60\% \text{ min_confidence}$ בלבד. יצאו 99 חוקי הקשר שכאלו, גם הם יופיעו בנספח ב.

כדי לסנן את חוקי ההקשר המעוניינים ביותר אמל' עם המדדים הגבוהים ביותר - lift ו- confidence.

גובהים ($\text{confidence} > 0.8$ ו- $\text{lift} > 1.1$) וגם ($\text{confidence} > 0.9$ ו- $\text{lift} < 0.9$ לא נקבע בolumn).

```
['age_Young'] => ['hypertension_Low', 'stroke_Healthy'], lift=1.12, confidence=0.97
['age_Young', 'heart_disease_No'] => ['hypertension_Low', 'stroke_Healthy'], lift=1.12, confidence=0.97
['age_Young'] => ['hypertension_Low', 'heart_disease_No', 'stroke_Healthy'], lift=1.16, confidence=0.97
```

ד. הרצת האלגוריתם ודיווח התוצאות נעשו בסעיפים ב' וג'

ה. ניתוח התוצאות והסקת מסקנות

מחוקים אלו אנו למדים על מידת הקשר בין צירופי ערכים של פריטים כאלו ואחרים הנמצאים בטרנסקיוזות בין עמודת המטרה – הסיבוי של המועמד לקבל שבע. וכדי לדין – הסיבוי של המועמד לא לקבל שבע (באשר הסיבוי לקבל שבע הוא המאורע המשלים).

מדד $\text{lift} > 1$ מראה קוליזיה חיובית בין הפריטים בטרנסקיוזה, $\text{lift}=1$ מראה שאין קשר בין הפריטים. מ-3 חוקים אלו אנחנו למדים שלגיל צער ולהז דם נמוך יש את הקשר חזק ביותר לכך שלא חוטפים שבע בדומה לתוצאות שקיבלנו בממ"ז 21. נתונים רבים יותר על חול שבע היו יכולים לתורם לנו בכך שהם יכולים לצפות גם בחוקים אשר הסיפה שלהם הייתה stroke_Stroke וכך לקבל מידע שימושית על מה גורם למועמדים לקבל שבע ולא רק מה גורם להם להיות בריאים.

שאלה 2 – ניתוח אשכבות

א. מדריך אינטואיטיבי לאשכבות

לتحלין ניתוח אשכבות יש מגוון רחב של מדריך אינטואיטיבי. למשל – יכולות התרמודודות עם נתונים חישומיים, נתונים וביומדדים מטיפוסים שונים, סקלביולוגיות, יכולת הסתגלות של המודל לנתונים חדשים ועוד.

למדידת האיכות של האשכבות עצמן, נעזר בשני מדריך האיכות הבאים:

1. הומוגניות ושלמות – מדריך שועל פיו נוכל להחליט עד כמה "המრחיק" בין עצמים בכל אשכול הוא קטן – עד כמה העצמי סבכל אשכול ואשכול דומים או שונים מעיצמים באשכבות אחרים. ככל שעיצמים בתוך האשכול דומים אחד לשני אף שונים מעיצמים באשכבות אחרים כך החלקה נחשבת איכותית יותר.
2. מגמתיות – מדריך שועל פיו אפשר לבדוק האם קיימת מגמה או תופעה, שלא נתפסת בעין האדם שניתן ללמידה עליה מהחלקה לאשכבות. מדריך זה יכול להעיד על חלוקה טובה במידה והאשכבות שנוצרו הינם לא אקרראיים.

מכיוון שאנו מבצע את העבודה בפייתון עם הספרייה `sklearn` אפשר להשתמש במדד האיכות `silhouette-score`. שמאפשר למדוד עד כמה חלוקת הנתונים בסט הנתונים לאשכבות מתאימה. פונקציית המטריקה של מדד זה מחזירה ערכים בתחום [−1, 1] כאשר 1 הוא הערך הטוב ביותר, −1 הוא הרע ביותר ואפס מייצג חפיפה בין האשכבות. השתמש גם במדד זה למדידת איכות האשכבות.

ב. בחירת גישה לניתוח אשכבות

בחorthy באלגוריתם k -means לניתוח האשכבות. אלגוריתם k -means הוא אלגוריתם איטרטיבי המקבל ערך k אשר קובע את מספר האשכבות המדדרים ופונקציית מרחיק אשר מגדרה את המרחק בין שני אובייקטים בסט הנתונים. האלגוריתם מקבץ את x הפריטים שהתקבלו בסט הנתונים תוך מינימיזציה המורחיקם (WCSS) מכל מרכז באשכול. באלגוריתם זה לא מtabצע אימון והתוצאות מtabצעת על כל סט הנתונים. חשוב לציין שקיים trade-off בין השאייה למינימיזציה $WCSS$ ו- k . ככל ש- k קטן יותר כך יקטן $WCSS$. אך ככל שנגדיל את k ניצור מספר גדול של אשכבות ויהיה לנו קשה יותר לבצע אנליזה נוחה של המידע וליציר מסקנות. כדי למצוא את k האידיאלי משתמש ב-*Elbow Method* שעליה אפרט בהמשך.

מציג פסאודו-קוד של האלגוריתם:

1. בחר k פריטים רandomים מתוך סט הנתונים
2. מוקם את סדרת מרכזים הנקובד $C = \{c_1, \dots, c_k\}$ בזרהRndomilit עברו k הפריטים
3. **לעוז לא התכנסנו או לא הגיעו למוגבלת האיטרציות:**
4. עבור כל פריט x_i :
5. מצא את מרכז הנקובד הקרוב ביותר ל- x_i מכל מרכז הנקובד C והכניסו ל- c_{x_i}
6. הכנס את הנקובה x_i לאשכול של מרכז הנקובד c_{x_i}
7. עבור כל אשכול מי- 1 עד k :
8. מוקם את מרכז הנקובד המורכב ממוצע הפריטים שנמצאים באשכול זה מחדש
9. **החזיר את האשכבות שנוצרו**

תירונות וחסרונות של אלגוריתם k -means :

יתרונות:

1. אלגוריתם טبعי ופשוט להבנה
2. קל לימוש ויעיל
3. אלגוריתם זה טוב עם ערכי נורמיים ורציפים שמאפיינים מידע בריאוטי של חולים ובמו שראינו כאן לפחות את תכונות אשר משפיעות מאוד על חיזוי השבץ (גיל, BMI, AGL)

חסרונות:

1. חלוקה לא אופטימלית – מבוצעות בשלבים הראשונים של האלגוריתם בחירותRndומליות שיכולות להוביל לחלוקת לא אופטימלית
2. תלות בפונקציית המרחק – פונקציית מרחק שלא מתאימה לפרמטרים יכולה לפגוע מאוד בתוצאות הבינות בעקבות מפתת נוחות הנתונים ושהוא עובד טוב עם ערכי רציפים שראינו שימושיים מאוד לשבע בחרתי להשתמש ב-k-means.

ג. שלבי ניתוח האשכולות – הכנת הנתונים – פרמטרים – ערכי הפרמטרים

הכנת הנתונים – בchnerתי מספר סטיטים שונים של נתונים להשתמש בהם, לבסוףלקח את אותו סט הנתונים לאחר הניקיון והסדר שעשייתי בממ"ן 21 נתונים אבל במקום לבצע דיסקרטיזציה לתכונות הנומריות ה فعلתי עלייהן את הפונקציה max-min עם הגבולות [0,1] כדי לנормל את המרחקים ולא לחתת יתרון ממשמעותי בפונקציית המרחק לתוכנות אלו על פני תכונות קטגוריאליות שהן או 0 או 1.

מטרת ברירת המידע זו היא לנסות לאפיין אשכולות של חולים-ב- "סיכון גובה" או "סיכון נמוך". لكن, כדי לא ליצר bias בחלוקת לאשכולות הורדתי את עמודת המטרה stroke לפניה הרצת k-means. רק לאחר ההרצה ארצה להציגן חוזרת את התווית ו לנסות לאפיין את האשכולות לפי פיזור stroke באשכול.

בחירת הפרמטרים – באלגוריתם k-means הפרמטר המרכזי שצריך לבחור הוא פרמטר k שמייצג את מספר האשכולות המבוקש לחילוקה. פרמטרים נוספים אשר משתמשים בהם לטובת הרצת האלגוריתם בספריית sklearn הם:

1. מספר האיטרציות המקסימלי (max_iter) - 300
2. פונקציית אתחול לחילוקת k הפריטים הראשונים (init) – יש שתי אופציות
 - a. Random – השיטה המוכרת המתווארת בפסאודו-קוד שהציגתי לעלה
 - b. K-means++ – שיטה שבחרת את k הפריטים הראשונים בצוות "חכמה" על מנת להוכנס במהירות. בחרתי בשיטה זו כי היא הניבה לי את התוצאות המדויקות יותר.
3. אלגוריתם לחישוב האשכולות – ישנן שתי אופציות אשר ממומשות תחת הספרייה employed
 - a. Elkan – השיטה הקלאסית של k-means – מקצת נקודת לאשכול הקרוב ביותר אליה וمعدכן לאחר מכן את המרכדים
 - b. Elkan – משתמש בוריאנט Elkan של אלגוריתם k-means אשר משפר ביצועים לחישוב באמצעות שימוש ב-"אי שיויון המשולש" כדי להמנע מחישובים מיותרם.

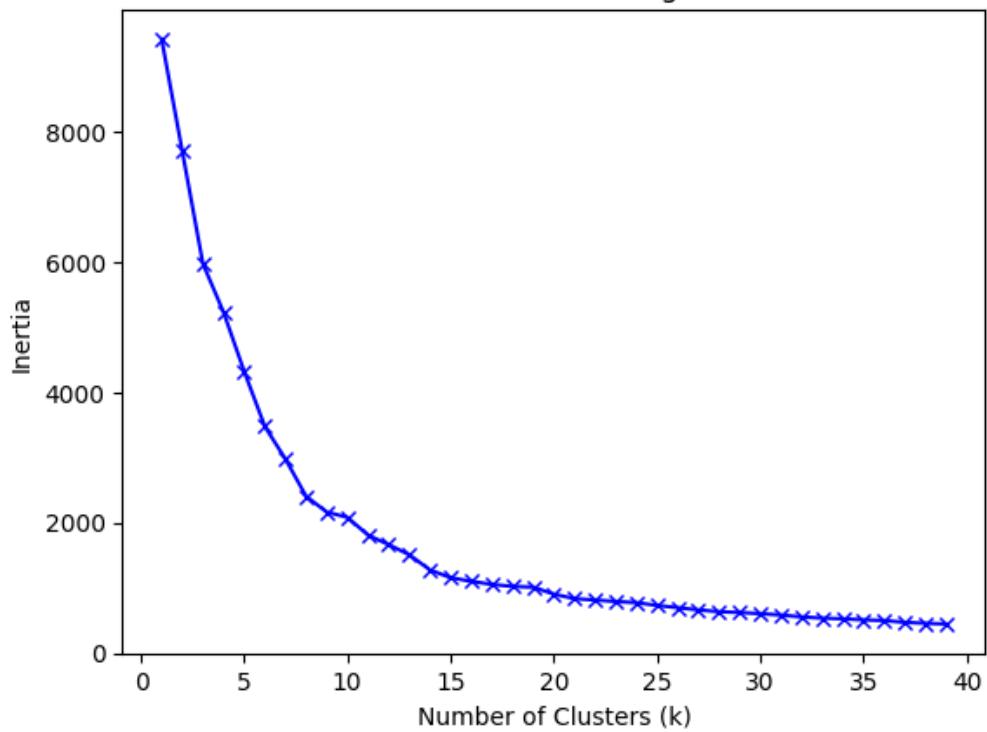
השתמשתי באלגוריתםloyd כיוון שהוא ייחודי וזה האלגוריתם המומלץ והDİPOLITY. במידה ואין חשיבות למהירות הביצועים.

4. פונקציית המרחק – פונקציית המרחק בה משתמשת sklearn היא מרחק אוקלידי בין נקודות אשר מתואר כז': $\sqrt{\sum_{i=1}^n (x_i - y_i)^2}$ לכל זוג נקודות x, y בלבד מימד.

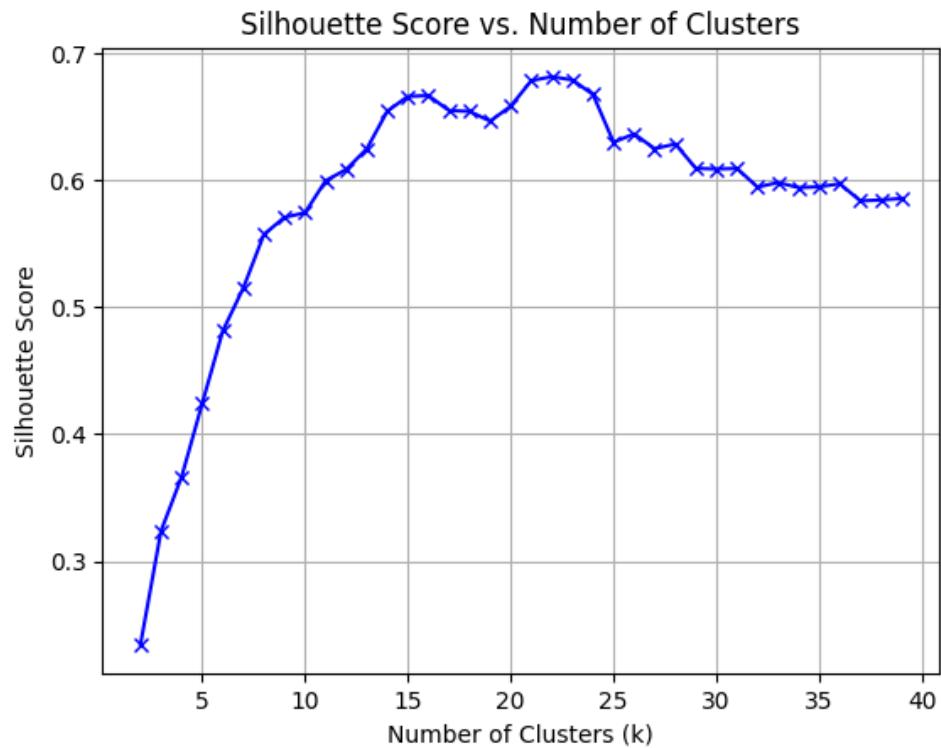
בחירה k על ידי שימוש ב-method elbow ועל ידי Silhouette Score:

נעזרתי במדד inertia של ספריית sklearn שמביע את ערךWCSSüber ריצה של k-means ובעזרתו בניתי גраф שישמש אותנו בבחירה הפרמטר עם Elbow Method

The Elbow Method using Inertia



בנוסף, ארצה למדוד מה ה-sillouhete-score לכל k כדי לבדוק שאני יוצר אשכולות איותיים:

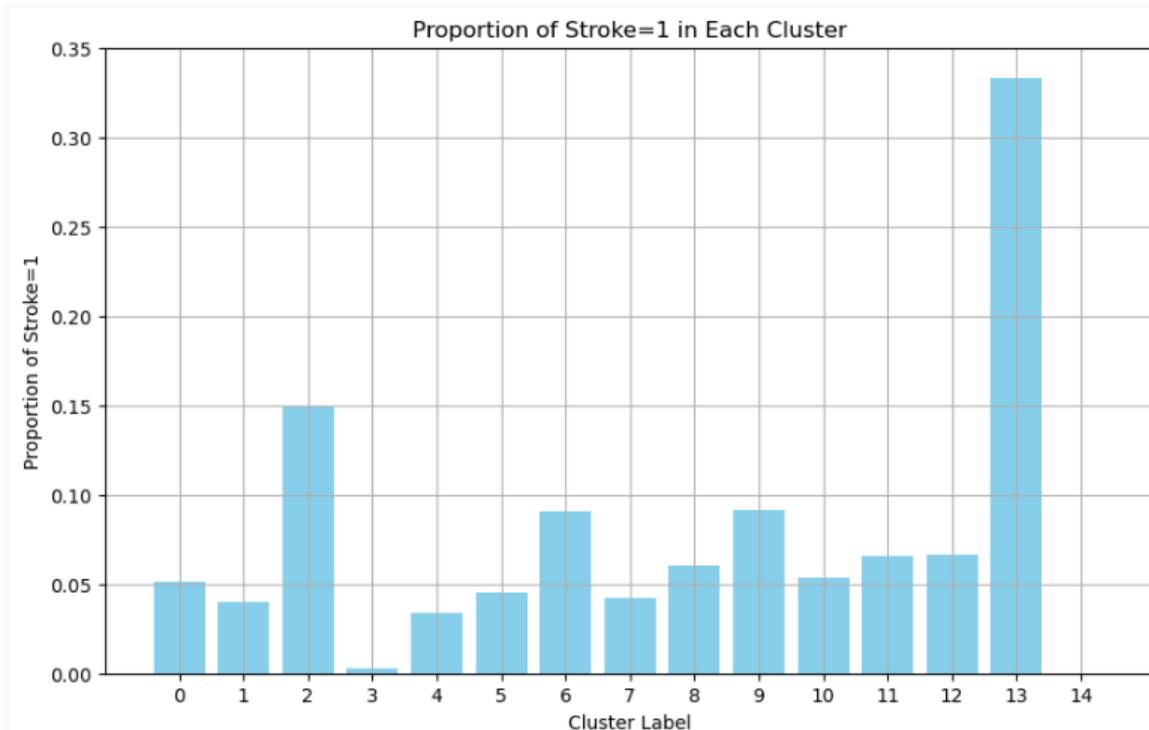


כפי שניתן לראות, סיום הירידה החדה ביותר בגרף ה-inertia מוסתיים כאשר $k=9$, אך שם ה-silhouette-score עוד יחסית נמוך. ישנה עוד ירידה משמעותית בגרף ה-inertia שמוסתיימת ב- $k=15$ ושם כבר ניתן לראות שהגענו כבר כמעט לגמרי ל-silhouette-score מקסימלי שכן לשאר ניתוח האשכולות שלנו נבחר ב- $k=15$.

ד. דיווח התוצאות

לאחר הרצת אלגוריתם k-means על הנתונים ללא עמודת stroke עם $k=15$ קיבלנו 15 אשכבות. נמדד את ההומוגניות של האשכבות, קיבלנו homogeneity-score = 0.06, תוצאה נמוכה לכל הדעתות ומה שקרה שרוב האשכבות לא מאוד טהורין עם התייחסות ל-stroke ורובם מכילים תערובת של חולים ובריאים.

אציג את היסטוגרמה של פרופורציית החולים בכל אשכול:

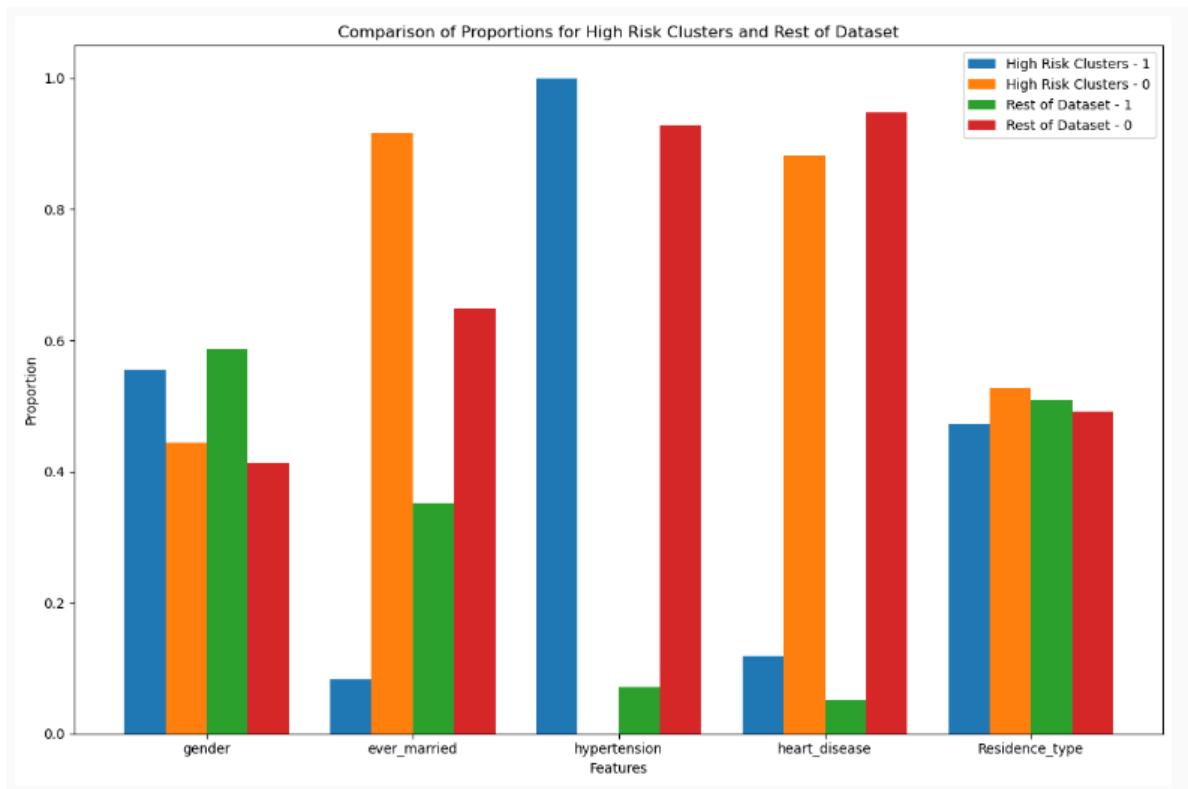


ניתן לראות אשכול 13 הוא ב-"סיכון גבוה לשbez" משמעותית משאר האשכבות. גם באשכול 2 יש אחוז גבוה יותר ולכן ננסה לנתח את המאפיינים של קבוצות אלו כדי לנסוט למצוא את גורמי הסיכון שמאפיינים אותם.

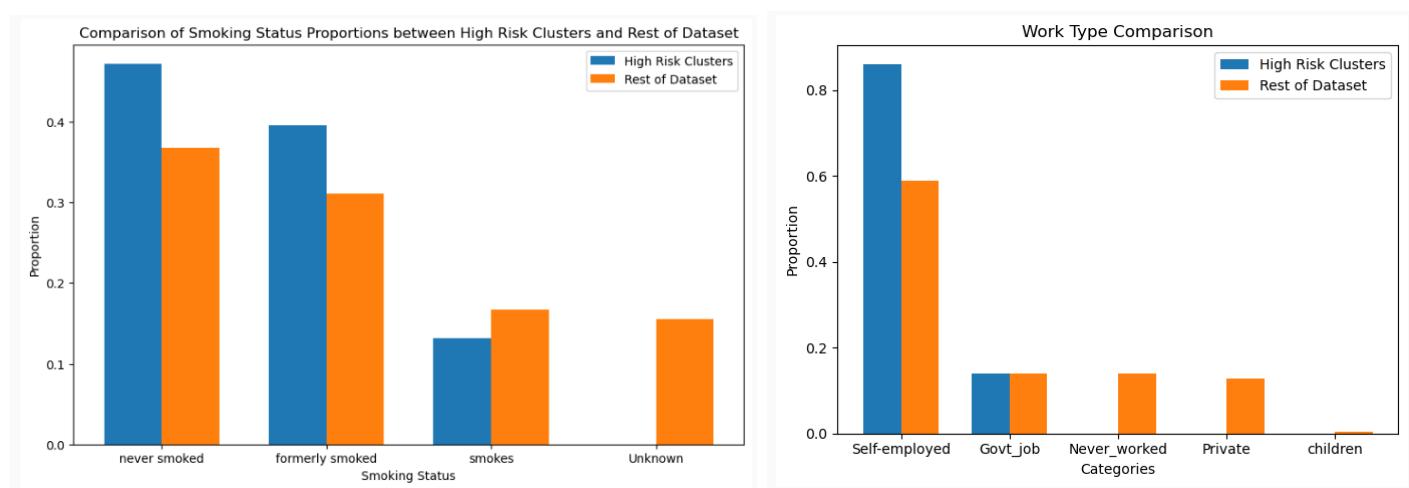
נסתכל על הממצאים של האשכבות הללו בפרמטרים הנומריים לעומת הממצאים של האשכבות האחרים:

column	High Risk Mean	Low Risk Mean	Diff
age	67.097	42.53	24.56
BMI	31.82	28.77	3.04
AGL	134.05	105.33	28.71

נסתכל על הקבוצות הבינאריות שנותרו וננסה לראות האם יש הבדלים בפרופורצית התכונות בין האשכולות עם הסיכון הגבוה לאשכולות של שאר הדטא.



נעsha את זה גם לתחנות הקטגוריאליות:



ה. ניתוח ומסקנות

נסתכל על ההבדלים המרכזיים בין האשכולות עם "סיכון גבוה" לאלו שלא בסיכון גבוה וננסה ללמידה על התוכנות המרכזיות שמאפיינאות את האשכולות האלה.

שלושה הבדלים מרכזיים שקופצים לעין מיד –

1. קיים הבדל גדול מאוד בין ממוצע הגילאים של הקבוצות בסיכון לבין שאר האוכלוסייה – ממוצע הגילאים של האשכולות בסיכון היה 67 ושל שאר האוכלוסייה 43 בלבד!
2. באופן דומה, גם רמת הגלוקוז באוכלוסייה בסיכון גבוה משמעותית משאר האוכלוסייה עם הבדל של כמעט .30
3. נתון מדהים שנייתן לשים לב אליו זה שככל האנשים בקבוצה בסיכון בניגוד מוחלט לפורופורציה שלהם באוכלוסייה סוגו עם לחץ דם גבוה – $hypertension = 1$ לעומת $hypertension = 0$ בעוד הכללית אחוז האנשים עם .5%

מכאן, נוכל להסיק שאלה פרמטרים שיכולים לאפיין חולים בסיכון בדומה מאוד לתוצאות שקיבלנו מכנתוחים אחרים.

lehava, כדי לשפר את הכנתוח, נוכל אולי לצמצם את המימד של הנתונים על ידי מחיקת תכונת אשר בקורסציה גבוהה אחת לשניה כמו `ever_married` ו-`age` בכך לנרטמל את השפעתן. בנוסף, נוכל לנסות להשתמש ב-DBSCAN ולהשווות את התוצאות.

שאלה 3 – רשת נירוניים מלאכותית

A. הגדרת ארכיטקטורת הרשת והנתונים בהם משתמש

הנתונים בהם משתמש הם הנתונים בהם השתמשתי לנוכח האשכולות בשאלה 2, כמו אלגוריתמים לאשכול, אך גם אלגוריתמים של רשותות נירוניים רגילים לשאלה scale של הנתונים לכן חשוב מאוד לבצע להם סטנדרטיזציה מתאימה. אזכיר שבשאלה 2 השתמשתי ב-*MinMax* עם הגבולות [0-1] לנתונים הnumerical ו-*one hot encoding* (encoding) לעמדות הקטגוריאליות.

בנוסף, השתמשתי באלגוריתם SMOTE כמו במ"ן 21 אשר מייצר דוגמאות סיניות של נתונים מה-class minority (במקורה שלנו stroke=1) כדי לאזן את נתוני הלמידה. במובן שלא נריץ את SMOTE על נתונים המבחן שלנו כדי לא לייצר *data-leakage*. השתמשתי באלגוריתם זה כדי לייצר מספר דוגמאות זהה בין stroke=1 ל-0. פיצול נתונים המבחן והלמידה היה 0.6 ללמידה ו-0.4 למבחן.

רשת הנירוניים שנבחרה הינה רשת הזנה קדמית (Feedforward Neural Network). אופן זרימת הנתונים ברשת הוא זרימה לבירן אחד משכבה הילילית לשכבות החבויות ולכיוון שכבת הפלט.

רשת מורכבת מ-3 סוגי שכבות באופן בו בכל שכבה, כל נירון מחובר בצלעות ממושקלות לכל אחד מהנירונים בשכבה הבאה.

שכבת הילילית (input) – משמשת בקלט לרשת כך שכל קודקוד מייצג עמודה (feature) נתונים שלנו. לאחר הפעלת *AE* על הנתונים התקבלו 17 עמודות וכן יהו לנו 17 נירונים בשכבה זו.

שכבות חבויות (hidden) – נמצאות בין שכבת הילילית לפט. לאחר ניסוי וטעייה רבים עם שתי שכבות מרובות נירונים ((10,6), (32,16)) כדי לתפוס את מרכיבות הבעיה או שכבה אחת בגודלים שונים (15,10,8). מצאתי שאת התוצאות הטובות ביותר ביוטר קיברתי עם שתי שכבות חבויות, בשכבה הראשונה 10 נירונים ובשכבה השנייה 6 נירונים.

שכבת הפלט (output) – השכבה האחורנית ברשת המשמשת לשכבת ציאה שמצויה את תוצאות החיזוי הסופיות של הרשת. בשכבה זו יהיה נירון אחד ויחיד, ההסתברות שתצא בנירון זו תחולק כך שאם $x > 0$ אז נסואג $stroke=1$ אחרת, $stroke=0$.

אستخدم בשתי פונקציות הפעלה שונות, הראשונה היא סיגמוד (sigmoid):

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

הפלט של פונקציה זו תמיד נמצא בין 0 ל-1 מה שהופך אותה מתאימה לסיווג ביןארי כאשר צריך לתת לכל תוצאה הסתברות מסוימת וכן משתמש בה לשכבת הפלט.

– Rectified Liner Unit (ReLU)

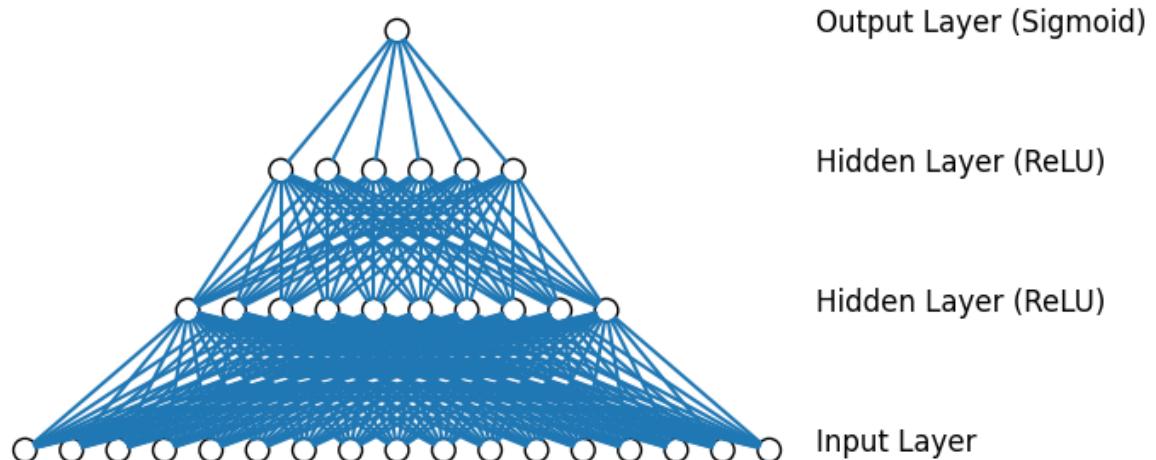
$$f(x) = \max(0, x)$$

פונקציה זו היא פונקציה טובה מאוד לשכבות חבויות ויש לה יתרונות רבים; היא עיליה חישובית (הרבבה יותר מסיגמוד), ובניגוד לפונקציה של סיגמוד היא אינה יוצרת את בעיית *Vanishing Gradient Problem*, בעיה זו נוצרת כאשר הגרדיינט (נגזרת) שלה בסופו של דבר חסומה ב-0.25 ברשותות עמוקות מחשבים את הגרדיינט על ידי כל השרשרת ושימוש בפונקציית סיגמוד יכול לאחר שכבות לא רבות לאיפוס של הגרדיינט כי $\left(\frac{1}{4}\right)^k \leq grad \leq 0$ כאשר k הוא מספר השכבות ברשת. בעיה זו גורמת לכך שעדכונים למשקלים בשכבות קדמות כמעט לא מחליפים לשכבות הבאות והרשת עצמה מלמוד.

ReLU פותרת בעיה זו ע"י שubahו ערכים חיוביים הגדריאנט יהיה תמיד 1 וכן יחלחל בקלות ברשות. למרות שבמקרה שלנו הרשת אינה עמוקה, ההבדלים בריצה בין סיגמוד-*ReLU* היו זניחים ולכן העדפתו להשתמש ב-*ReLU*-*best practice*.

להלן צור של ארכיטקטורת הרשת:

Neural Network architecture



ב. פרמטרים של תהליכי האופטימיזציה – פונקציית השגיאה, גודל ה-batch, קצב הלמידה

פרמטרים של תהליכי אופטימיזציה –

פונקציית העלות השגיאה שנשתמש בה היא $\text{Binary Crossentropy}$ או היא המתאימה ביותר לביעות סיווג ביןארי אשר אנו נדרשים לפתור. פונקציה זו מוגדרת כר:

$$BCE = -\frac{1}{N} \sum_{i=1}^N [y_i * \log(p(y_i)) + (1 - y_i) * \log(1 - p(y_i))]$$

כאשר y מייצגת את ה-label האמתי של החולה ($\text{stroke}=0$ or $\text{stroke}=1$) ו- $p(y)$ זו תוצאה החיזוי של המודל (הסתברות שנעה בין 0 ל-1).

אלגוריתם האופטימיזציה שלנו יהיה SGD (Stochastic Gradient Decent) – אלגוריתם אופטימיזציה שמעדכן את המשקלים במודול באמצעות חישוב הגרדיינט באלגוריתם זה מבוסס על קבוצת תוצאות אקרואיות קטנה. בחלק מהמקרים גישה זו מסייעת בהפחיתת זמני הריצה והמנועות ממינימום מקומי.

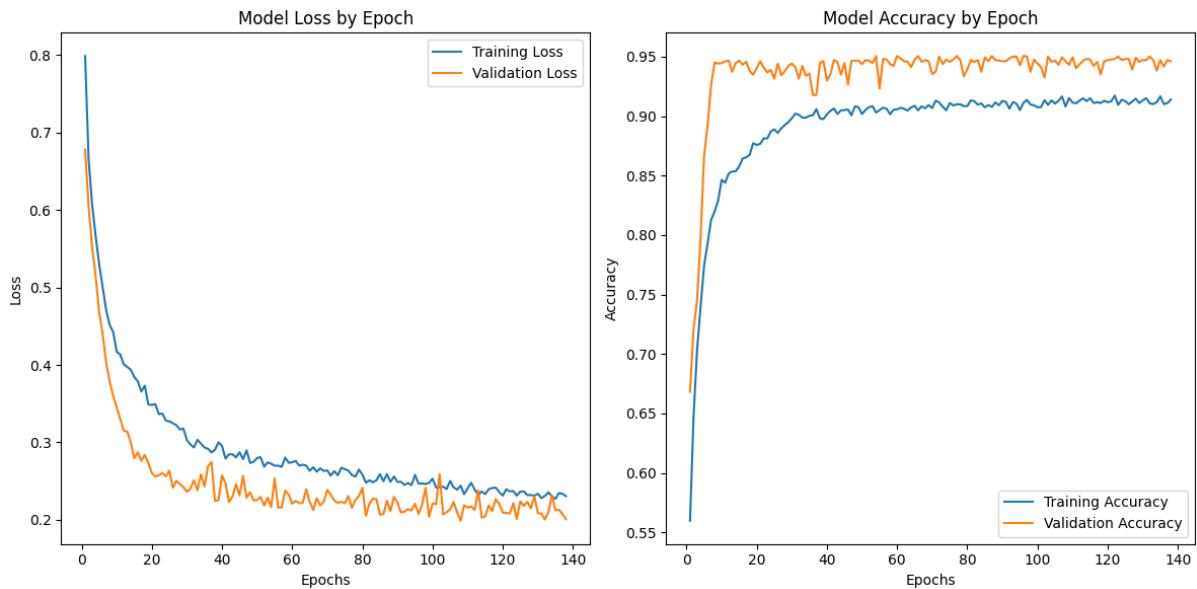
הוא פשוט ומתחאים לכמות גדולה של נתונים (אלפים) אך חשוב לציין לו התכונות יחסית איטית.

קצב הלמידה – בגלל שאנו משתמשים ב-SGD נצורך להיות זהירים עם קצב הלמידה, SGD וגישה מאוד לשוניים בקצב הלמידה; קצב למידה גבוה מדי יכול לגרום לאלגוריתם לסתות (to diverge) בעוד קצב למידה נמוך מדי יגרור האטה בהתקנסות הלמידה. לאחר ניסוי וטעייה, ראיינו שקצב הלמידה 0.01 היה הטוב ביותר מבחינה יציבות וקצב השיפור שניתן לדראות בגרפים של הדיק ופונקציית השגיאה לאורך ה-epochs.

גודל ה-batch שנבחר הינו 100, זה הגודל הדיפולטי בחבילה שהשתמשתי בה ושינויו שלו לא נתן שיפור משמעותי בתוצאות.

ג. הרצת הרשת וдиוקן Epochs

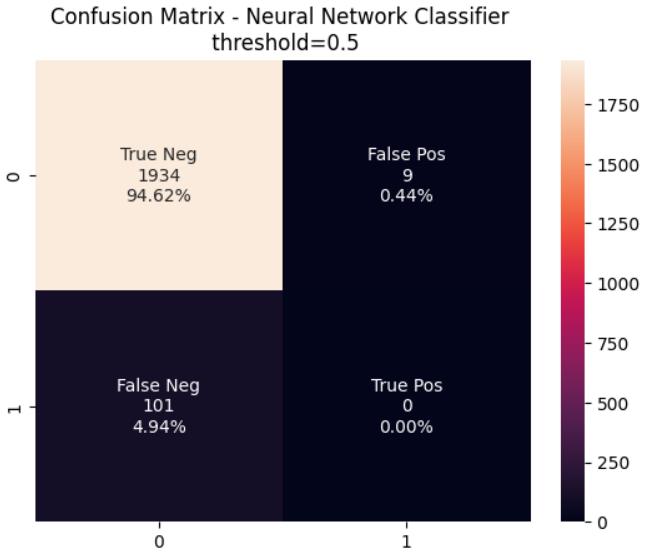
לאחר הרצה של הרשת, אציג את גרף ביצועי הרשת (accuracy) לאורך ה-epochs של האימון ביחד עם גרף השגיאה עבור נתונים האימון וה מבחן בפונקציה של ה-epoches.



ניתן לראות שמאוד מהר המודל הצליח להתקנס מבחינת-loss (ב-epoch 40 בערך) וכן גם מבחינת-loss Accuracy אף עדין אפשר לראות ירידה קלה ויציבה בLoss. כדי להחליט מתי לסיים ללמידה ומתי המודל הגיע למיצוי של הלמידה השתמשתי במיקורופרמטר EarlyStopping. פרמטר זה עוזר לזהות מתי המודל מגע לסתGANICA בלמידה, הוא לוקח פרמטר patience שגדיר תוקן כמה epochs צריך לראות שיפור ב-lossathan Validation Loss לפני שהוא עצור את הלמידה. לאחר שההBIN שהלמידה עזרה הוא מחזיר את המודל להיות עם המשקלים של lossathan Loss הנמשך יותר לאורך הלמידה ולא הסופי. לאורך למידה זו הגיעו למינום ב-epoch=108.

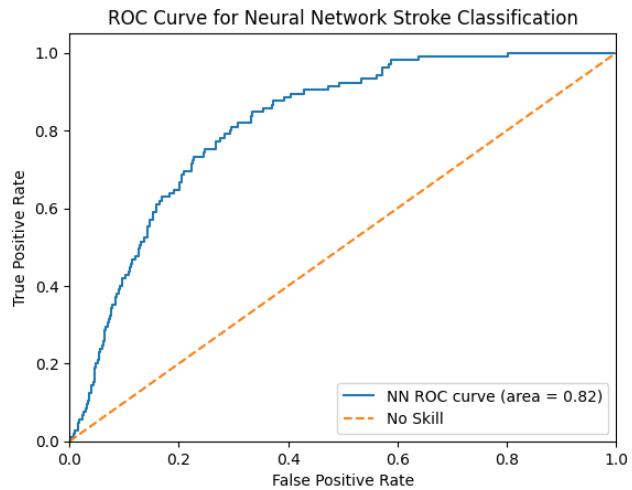
השתמשתי בנוסף במיקורופרמטר בשם ReduceLROnPlateau – מיקורופרמטר זה משמש בעיקר לעזר למודלים להتاימים את קצב הלמידה שלהם ברגע שהם מגעים לסתGANICA. השתמשתי בפרמטר זה כדי לסחוט עוד למידה מהמודל אחרי שהוא הגיע ל-epoch=60 בערך וזה אכן עזר להורד עוד קצת את lossathan Validation Loss.

נסתכל על תוצאות ה-Confusion Matrix הראשונית –



זה מתאים מאד למה שראינו בגרף ה- accuracy , לא הצלחנו לסווג אף מקרה אחד של $\text{stroke}=1$. ראיינו בבר שבמודלים קודמים בממ"נ 21 היונו צריכים לשחק עם threshold של המודל כדי להגדיל את הרגישות שלו למקרים שבהם הוא חושד אפיו קצת שמדובר במקרה של שbez. אבל לפני הכל צריך לוודא שהמודל בכלל לימד משהו ושהוא לא סתם מנחש תמיד 0.

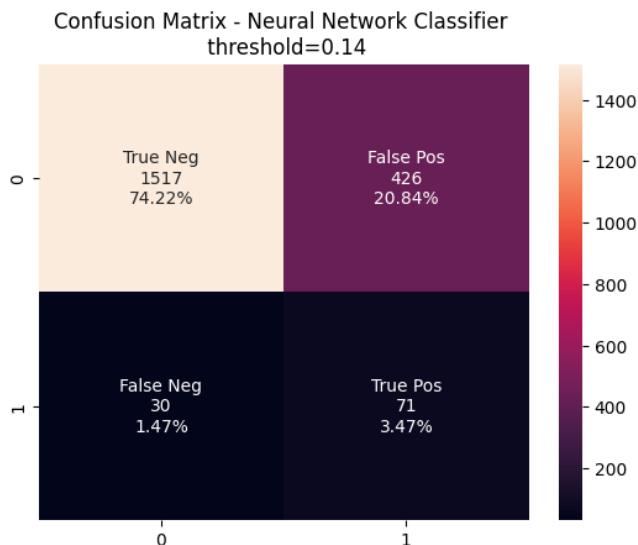
נסתכל על ה-ROC Curve כדי לוודא זאת:



פיעוכי לראות ה-AUC שלנו הוא 0.82, תוצאה שנחשבת אפיו יותר טובה מאשר המודלים שהשתמשנו בהם בממ"נ 21. זה אומר שהמודל יש יכולת חזקה להבדיל בין מקרים של שbez, ספציפית זה אומר שיש לו סיכוי של 82% לסווג מקרה ונדומלי חיובי מעיל מקרה ונדומלי שלילי. זה אומר שהבעיה שנטקלתי בה שהמודל לא הצליח לסווג אף מקרה של $\text{stroke}=1$ נכון היא לא בഗל שהוא לא הצליח למדוד אלא בගל שה-threshold היה לו נמוך מדי.

נחפש עבשיו את ה-threshold עבורו נקבל את ההבדל הטוב ביותר בין המקרים.

לאחר כמה ניסיונות מצאתי שה-threshold האופטימלי ישב ב-0.14 (אם המודל מחזיק הסתברות גדולה מ-0.14 אז תסוויג את החוליה $\text{stroke}=1$).



אלו תוצאות מרשימות, נציג אותן במלואן:

Metric	Score
Precision	0.143
Recall	0.703
ROC_AUC	0.825
Accuracy	0.777

את ההשוואה למודלים האחרים נעשה רק בפרק המסקנות אבל נראה שהמודל לא סיפק תוצאות משמעותית טובות יותר מהතוצאות של המודלים בממ"נ .21

ד. מקרים חריגים בהם היה ניתוח שגוי

התבכעו הרבה מקרים של סיוג שגוי (426 FP, 30 FN) כי הם מעוניינים הרבה יותר וחשוב לה宾ி מהם התפססו דוקא. מקרי ה-FP יהיו [בנוסף ג'](#).

	<u>id</u>	<u>gender</u>	<u>age</u>	<u>hypertens heart_dise ever_marr</u>	<u>Residence</u>	<u>avg_gluco</u>	<u>bmi</u>	<u>work_type</u>	<u>work_type</u>	<u>work_type</u>	<u>work_type</u>	<u>work_type</u>	<u>smoking_ smoking_ smoking_ smoking_ pred</u>	<u>stroke</u>
211	62439	1	0.621582	0 0 0	0 0.223017	0.194731	TRUE FALSE	FALSE FALSE	FALSE FALSE	FALSE FALSE	TRUE FALSE	FALSE FALSE	0 1	
86	3253	0	0.743652	0 1 0	0 0.261703	0.194731	FALSE TRUE	TRUE FALSE	FALSE FALSE	FALSE FALSE	FALSE FALSE	TRUE FALSE	0 1	
225	39186	1	0.694824	0 1 0	1 0.745361	0.237113	FALSE FALSE	TRUE FALSE	FALSE FALSE	TRUE FALSE	FALSE FALSE	FALSE FALSE	0 1	
219	31421	0	0.890137	0 1 0	0 0.759902	0.209622	TRUE FALSE	FALSE FALSE	FALSE FALSE	FALSE FALSE	TRUE FALSE	TRUE FALSE	0 1	
180	54567	1	0.560547	0 0 0	1 0.106454	0.234822	FALSE FALSE	TRUE FALSE	FALSE FALSE	FALSE FALSE	FALSE FALSE	TRUE FALSE	0 1	
182	39912	1	0.389648	0 0 0	0 0.09693	0.224513	FALSE FALSE	TRUE FALSE	FALSE FALSE	FALSE FALSE	FALSE FALSE	TRUE FALSE	0 1	
140	20439	0	1	0 1 0	0 0.224171	0.168385	TRUE FALSE	FALSE FALSE	FALSE FALSE	FALSE FALSE	TRUE FALSE	TRUE FALSE	0 1	
186	16077	0	0.768066	0 1 0	1 0.28423	0.277205	FALSE FALSE	TRUE FALSE	FALSE FALSE	TRUE FALSE	TRUE FALSE	FALSE FALSE	0 1	
78	45805	1	0.621582	0 0 0	1 0.508679	0.203895	FALSE FALSE	TRUE FALSE	FALSE FALSE	TRUE FALSE	TRUE FALSE	TRUE FALSE	0 1	
97	56841	0	0.707031	0 1 0	0 0.8562	0.241695	FALSE FALSE	TRUE FALSE	FALSE FALSE	FALSE FALSE	FALSE FALSE	TRUE TRUE	0 1	
153	12482	0	0.829102	0 0 0	1 0.104792	0.197022	FALSE FALSE	TRUE FALSE	FALSE FALSE	FALSE FALSE	FALSE FALSE	TRUE TRUE	0 1	
62	65842	1	0.816895	1 0 0	0 0.031484	0.171821	FALSE FALSE	TRUE FALSE	FALSE FALSE	FALSE FALSE	FALSE FALSE	TRUE TRUE	0 1	
100	12363	0	0.780273	0 1 0	1 0.087619	0.211913	TRUE FALSE	FALSE FALSE	FALSE FALSE	TRUE FALSE	FALSE FALSE	FALSE FALSE	0 1	
141	45965	1	0.719238	0 0 0	0 0.283076	0.154639	FALSE FALSE	TRUE FALSE	FALSE FALSE	TRUE FALSE	FALSE FALSE	TRUE TRUE	0 1	
143	37651	1	0.841309	1 1 1	1 0.078709	0.303551	FALSE FALSE	TRUE TRUE	FALSE FALSE	FALSE FALSE	TRUE FALSE	TRUE FALSE	0 1	
192	36255	0	0.719238	0 0 0	0 0.290416	0.28866	FALSE FALSE	TRUE FALSE	FALSE FALSE	FALSE FALSE	TRUE FALSE	TRUE TRUE	0 1	
204	62019	0	0.658203	0 0 0	0 0.151094	0.238259	TRUE FALSE	FALSE FALSE	FALSE FALSE	FALSE FALSE	FALSE FALSE	TRUE TRUE	0 1	
34	14248	0	0.584961	0 0 1	1 0.134244	0.222222	TRUE FALSE	FALSE FALSE	FALSE FALSE	FALSE FALSE	TRUE FALSE	FALSE FALSE	0 1	
81	26015	1	0.804688	0 0 0	0 0.213877	0.203895	FALSE FALSE	TRUE TRUE	FALSE FALSE	TRUE FALSE	FALSE FALSE	FALSE FALSE	0 1	
87	71796	1	0.853516	0 1 0	0 0.019527	0.252005	FALSE FALSE	TRUE FALSE	FALSE FALSE	TRUE FALSE	FALSE FALSE	FALSE FALSE	0 1	
224	8899	0	0.597168	0 0 1	0 0.229619	0.247423	FALSE FALSE	TRUE FALSE	FALSE FALSE	FALSE FALSE	FALSE FALSE	TRUE TRUE	0 1	
41	1261	0	0.658203	0 0 0	1 0.074324	0.208477	FALSE FALSE	TRUE FALSE	FALSE FALSE	TRUE FALSE	TRUE FALSE	TRUE FALSE	0 1	
18	27458	1	0.731445	0 0 1	1 0.157419	0.315006	FALSE FALSE	TRUE FALSE	FALSE FALSE	FALSE FALSE	TRUE FALSE	TRUE FALSE	0 1	
31	33879	0	0.511719	0 0 0	0 0.130597	0.172967	FALSE FALSE	TRUE TRUE	FALSE FALSE	TRUE FALSE	FALSE FALSE	FALSE FALSE	0 1	
162	69768	1	0.015137	0 0 1	1 0.0704	0.203895	FALSE FALSE	TRUE TRUE	FALSE FALSE	TRUE FALSE	FALSE FALSE	0 1		
93	37726	1	0.975588	1 0 0	1 0.062044	0.182131	FALSE FALSE	TRUE TRUE	FALSE FALSE	TRUE FALSE	FALSE FALSE	FALSE FALSE	0 1	
239	32221	0	0.731445	0 1 0	1 0.169883	0.293242	FALSE FALSE	TRUE FALSE	FALSE FALSE	TRUE FALSE	FALSE FALSE	TRUE TRUE	0 1	
236	28493	0	0.694824	0 0 0	1 0.143932	0.245132	FALSE FALSE	TRUE FALSE	FALSE FALSE	TRUE FALSE	FALSE FALSE	FALSE FALSE	0 1	
161	16590	0	0.865723	0 1 0	1 0.12298	0.203895	FALSE FALSE	TRUE FALSE	FALSE FALSE	TRUE FALSE	FALSE FALSE	TRUE TRUE	0 1	
83	66638	1	0.829102	1 0 1	1 0.113886	0.222222	FALSE FALSE	TRUE TRUE	FALSE FALSE	TRUE FALSE	TRUE TRUE	FALSE FALSE	0 1	

ונכל לראות שקבוצת ה-FN שלו מורכבת מאנשים מבוגרים אבל לא מאוד (אחוזון 70 לעומת אחוזון 84 של החוליםים). הם היו באחוזון נמוך ממשמעותית ברמת ה-AGL (36 לעומת 22). הבדלים משמעותיים אלו, וכן גם הבדלים אחרים היו יכולים לגרום לכך ששסוגו באופן לא מדויק אך קשה מאד להסביר את זה מנקודת שיחית יחסית זו.

ה. ניתוח התוצאות והסקת מסקנות

בצערנו ניתוח לאורך כל הסעיפים לעיל אף אוסף בעוד סיבום ומסקנות עיקריות של השאלה זו.

גם כאן, כמו בנסיבות הסיוג האחרים של בממ"ן 21 נתקלתי בבעיה מרובנית והיא `data imbalance`. גם כאן המודל התקשה מאוד לחזות חולי שבעצם טוביה ויצר הרבה FP בשבייל מעט FN. בגין מסוגים אחרים כמו עצי החלטה שנלמדו ביחידות קודמות, קשה יותר להבין לאחר האימון של המודול לפי מה הוא החליט לסוג וללמוד משם על התכונות הרלוונטיות כדי "לסמן" מועמדים עם פרמטרים רלוונטיים לקבוצה מסוימת בשלושה ורבות יותר צריך להסתמך על המודל `black-box`.

גם כאן, למרות שakan קשה יותר להבין מה גורם למודל להתנהג כמו שהוא, ניתן היה לראות שהוא נתן למספר חולים מבוגרים יותר ועם רמת גלוכוז גבוהה יותר במועדים רבים. התנהגות שדומה מאוד לתוצאות שראינו לאורך כל השאלות בממ"ן וגם בממ"ן הקודם.

סיכום ומסקנות

בשווה את תוצאות מודלי הסיווג שעבדנו איתם בשני הממ"נים –

Metric	RandomForest	CART gini-index	Neural-Network
Accuracy	0.757	0.691	0.777
Recall	0.767	0.692	0.703
ROC AUC	0.824	0.815	0.825
Precision	0.134	0.129	0.143

ונכל לראות של מרבית השווי המהותי בין המודלים, לאחר אפטום ועבודה אינטנסיבית לא היה הבדל גדול בתוצאות שהם נתנו לפחות על הנייה. נראה ש-NN ו-RF היו טובים יותר במעט מ-CART אך יחסית שקולים אחד לשני בכל שאר הפרמטרים.

באופן עקי ולאורך כל התוצאות של המודלים השונים, גם אלו ממשפחת ה-Supervised Learning וגם ממשפחת Supervised Learning ראיינו קורלציה כזו או אחרת בין תכונות מסוימות בסט הנתונים לבין עמודת המטרה. ראיינו שלגיל הפסיכינט (age), רמת הגלוקוז הממוצעת בدم (AGL) ולחץ דם גבוה (hypertension) הובילו לסיכון גבוה יותר לשbez. ודוקא פרמטרים כמו BMI, smoking-status שחשבתי אולי בהתחלה שיכולה להיות להם קורלציה לשbez לא ראיינו שמשמעותם בכלל גורמים משפיעים.

ראיינו את הפרמטרים האלו בצורה בולטת במיוחד ובנוחות דרך תוצאות האשכול וחוקי ההקשר אף גם דרך צפיה בעז ההחלטה שנוצר ב-CART וגם דרך ניתוח הדגימות שלא סוגו בראוו בראשת הנזירונים יכולים לראות את השפעת התכונות הללו על המודלים ועל התוצאה. הקונסיסטנטיות חז מראה על כך שעבודות חקר הנתונים ה证实עה בראוו.

במה נקודת לסיכון שחשובות בשבייל –

סיכום הנתונים –

- למרות כל התוצאות שקיבלו והיכולת שלנו להשתמש במודלים שייצרנו לחיזוי שbez, המודלים לא הגיעו לרמת דיקון גבוהה. סיווג הבדיקות של אדם זו מושימה קשה מאוד וצריכה להתבסס על הרובה יותר נתונים, תכונות ופרמטרים כדי להגיע לדיקון גבוה וגם אז זה יהיה כמעט בלתי אפשרי לדיקון.
- הסתמכו על במota צעומה של מקרים שהם *stroke=1* וכן נאצלנו להשתמש באלגוריתמים ליצירת דגימות סינטטיות שעזרה אף במעט לשפר את המודל.
- שיפור נוסף לאיסוף הנתונים היה יכול להיות באיסוף תכונות שימושיות יותר ולא קורלטיביות אחת לשניה, למשל התכונה *ever_married* הייתה קורלטיבית מאוד לגיל ולא תרמה כלל לשינוי הנתונים.

הבנה וניתוח הנתונים –

- סידור ונקיי הנתונים יכול להשפיע דרסטית על תוצאות המחקר, חיבורים לשיטים לב לכל פעולה שעושים ולהבין אותה טוב מאוד אחרת יכולים לפגוע בצורה דרסטית בתוצאות המודל מבלתי לדעת זאת.
 - פעולות המחקר צריכה לקרו באופן איטרטיבי שבו מבצעים ניסוי וטעייה, מנסים שיטה מסוימת רואים את התוצאות ואז חוזרים למדדים מתקנים ומשפרים כל הזמן את המודל ואת הלמידה.
- בנימה אישית, אהבתן מאוד את התנסות הרבה בנושאים השונים, אני מרגיש שהלמידה דרך הידים והמחשב משמעותית ובלתי נפרד מהלמידה מההרצאות וממדריך הלמידה ואני שמח שיצא לי להתנסות בה.

נספח א' – קבוצות תדירות וחוקים לשאלה 2

קבוצות תדירות בגודל 1:

```
[['hypertension_Low'],
['gender_Female'],
['heart_disease_No'],
['stroke_Healthy'],
['age_Old'],
['ever_married_Yes'],
['age_Young'],
['work_type_Private'],
['Residence_type_Rural'],
['Residence_type_Urban'],
['avg_glucose_level_High'],
['avg_glucose_level_Low'],
['bmi_High'],
['bmi_Low'],
['smoking_status_never smoked']]
```

קבוצות תדירות בגודל 2:

```
['heart_disease_No', 'hypertension_Low']
['gender_Female', 'hypertension_Low']
['heart_disease_No', 'gender_Female']
['hypertension_Low', 'stroke_Healthy']
['gender_Female', 'stroke_Healthy']
['heart_disease_No', 'stroke_Healthy']
['age_Old', 'heart_disease_No']
['age_Old', 'ever_married_Yes']
['ever_married_Yes', 'heart_disease_No']
['age_Old', 'stroke_Healthy']
['ever_married_Yes', 'stroke_Healthy']
['age_Young', 'heart_disease_No']
```

[*'age_Young'*, *'hypertension_Low'*]
[*'age_Young'*, *'stroke_Healthy'*]
[*'ever_married_Yes'*, *'hypertension_Low'*]
[*'work_type_Private'*, *'heart_disease_No'*]
[*'work_type_Private'*, *'hypertension_Low'*]
[*'heart_disease_No'*, *'Residence_type_Rural'*]
[*'Residence_type_Rural'*, *'hypertension_Low'*]
[*'Residence_type_Urban'*, *'heart_disease_No'*]
[*'Residence_type_Urban'*, *'hypertension_Low'*]
[*'heart_disease_No'*, *'avg_glucose_level_High'*]
[*'avg_glucose_level_High'*, *'hypertension_Low'*]
[*'avg_glucose_level_Low'*, *'heart_disease_No'*]
[*'avg_glucose_level_Low'*, *'hypertension_Low'*]
[*'bmi_Low'*, *'heart_disease_No'*]
[*'bmi_Low'*, *'hypertension_Low'*]
[*'smoking_status_never smoked'*, *'heart_disease_No'*]
[*'smoking_status_never smoked'*, *'hypertension_Low'*]
[*'work_type_Private'*, *'stroke_Healthy'*]
[*'Residence_type_Rural'*, *'stroke_Healthy'*]
[*'Residence_type_Urban'*, *'stroke_Healthy'*]
[*'avg_glucose_level_High'*, *'stroke_Healthy'*]
[*'avg_glucose_level_Low'*, *'stroke_Healthy'*]
[*'bmi_Low'*, *'stroke_Healthy'*]
[*'smoking_status_never smoked'*, *'stroke_Healthy'*]
[*'heart_disease_No'*, *'bmi_High'*]
[*'bmi_High'*, *'stroke_Healthy'*]
קבוצות תדירות בגודל 3:
[*'hypertension_Low'*, *'heart_disease_No'*, *'stroke_Healthy'*]
[*'hypertension_Low'*, *'gender_Female'*, *'stroke_Healthy'*]
[*'heart_disease_No'*, *'gender_Female'*, *'stroke_Healthy'*]

['heart_disease_No', 'gender_Female', 'hypertension_Low']
['age_Young', 'heart_disease_No', 'stroke_Healthy']
['age_Young', 'hypertension_Low', 'stroke_Healthy']
['age_Young', 'heart_disease_No', 'hypertension_Low']
['ever_married_Yes', 'heart_disease_No', 'stroke_Healthy']
['hypertension_Low', 'ever_married_Yes', 'stroke_Healthy']
['ever_married_Yes', 'heart_disease_No', 'hypertension_Low']
['work_type_Private', 'heart_disease_No', 'stroke_Healthy']
['work_type_Private', 'hypertension_Low', 'stroke_Healthy']
['work_type_Private', 'heart_disease_No', 'hypertension_Low']
['heart_disease_No', 'Residence_type_Rural', 'hypertension_Low']
['stroke_Healthy', 'Residence_type_Rural', 'hypertension_Low']
['heart_disease_No', 'Residence_type_Rural', 'stroke_Healthy']
['Residence_type_Urban', 'heart_disease_No', 'stroke_Healthy']
['Residence_type_Urban', 'hypertension_Low', 'stroke_Healthy']
['Residence_type_Urban', 'heart_disease_No', 'hypertension_Low']
['avg_glucose_level_Low', 'heart_disease_No', 'stroke_Healthy']
['avg_glucose_level_Low', 'hypertension_Low', 'stroke_Healthy']
['avg_glucose_level_Low', 'heart_disease_No', 'hypertension_Low']
['bmi_Low', 'heart_disease_No', 'stroke_Healthy']
['hypertension_Low', 'bmi_Low', 'stroke_Healthy']
['bmi_Low', 'heart_disease_No', 'hypertension_Low']
['smoking_status_never smoked', 'heart_disease_No', 'stroke_Healthy']
['smoking_status_never smoked', 'hypertension_Low', 'stroke_Healthy']
['smoking_status_never smoked', 'heart_disease_No', 'hypertension_Low']

נספח ב' – חוקי הקשר חזקים עם Stroke_Healthy בצד אחד שלהם.

['gender_Female'] => ['stroke_Healthy'], lift=1.0, confidence=0.95
['age_Old'] => ['stroke_Healthy'], lift=0.95, confidence=0.91
['age_Young'] => ['stroke_Healthy'], lift=1.05, confidence=1.0
['hypertension_Low'] => ['stroke_Healthy'], lift=1.01, confidence=0.96
['heart_disease_No'] => ['stroke_Healthy'], lift=1.01, confidence=0.96
['ever_married_Yes'] => ['stroke_Healthy'], lift=0.98, confidence=0.93
['work_type_Private'] => ['stroke_Healthy'], lift=1.0, confidence=0.95
['Residence_type_Rural'] => ['stroke_Healthy'], lift=1.0, confidence=0.95
['Residence_type_Urban'] => ['stroke_Healthy'], lift=1.0, confidence=0.95
['avg_glucose_level_High'] => ['stroke_Healthy'], lift=0.99, confidence=0.94
['avg_glucose_level_Low'] => ['stroke_Healthy'], lift=1.01, confidence=0.96
['bmi_High'] => ['stroke_Healthy'], lift=1.0, confidence=0.95
['bmi_Low'] => ['stroke_Healthy'], lift=1.0, confidence=0.95
['smoking_status_never smoked'] => ['stroke_Healthy'], lift=1.01, confidence=0.96
['gender_Female', 'hypertension_Low'] => ['stroke_Healthy'], lift=1.01, confidence=0.96
['gender_Female'] => ['hypertension_Low', 'stroke_Healthy'], lift=1.01, confidence=0.87
['heart_disease_No', 'gender_Female'] => ['stroke_Healthy'], lift=1.01, confidence=0.96
['gender_Female'] => ['heart_disease_No', 'stroke_Healthy'], lift=1.02, confidence=0.92
['age_Old', 'ever_married_Yes'] => ['stroke_Healthy'], lift=0.96, confidence=0.91
['age_Old'] => ['ever_married_Yes', 'stroke_Healthy'], lift=1.37, confidence=0.84
['ever_married_Yes'] => ['age_Old', 'stroke_Healthy'], lift=1.41, confidence=0.62
['age_Young', 'hypertension_Low'] => ['stroke_Healthy'], lift=1.05, confidence=1.0
['age_Young'] => ['hypertension_Low', 'stroke_Healthy'], lift=1.12, confidence=0.97
['age_Young', 'heart_disease_No'] => ['stroke_Healthy'], lift=1.05, confidence=1.0
['age_Young'] => ['heart_disease_No', 'stroke_Healthy'], lift=1.09, confidence=0.99
['heart_disease_No', 'hypertension_Low'] => ['stroke_Healthy'], lift=1.02, confidence=0.97
['heart_disease_No'] => ['hypertension_Low', 'stroke_Healthy'], lift=1.01, confidence=0.88
['hypertension_Low'] => ['heart_disease_No', 'stroke_Healthy'], lift=1.02, confidence=0.92
['ever_married_Yes', 'hypertension_Low'] => ['stroke_Healthy'], lift=0.99, confidence=0.94

['ever_married_Yes'] => ['hypertension_Low', 'stroke_Healthy'], lift=0.94, confidence=0.82
['work_type_Private', 'hypertension_Low'] => ['stroke_Healthy'], lift=1.01, confidence=0.96
['work_type_Private'] => ['hypertension_Low', 'stroke_Healthy'], lift=1.0, confidence=0.86
['Residence_type_Rural', 'hypertension_Low'] => ['stroke_Healthy'], lift=1.01, confidence=0.96
['Residence_type_Rural'] => ['stroke_Healthy', 'hypertension_Low'], lift=1.0, confidence=0.87
['Residence_type_Urban', 'hypertension_Low'] => ['stroke_Healthy'], lift=1.01, confidence=0.96
['Residence_type_Urban'] => ['hypertension_Low', 'stroke_Healthy'], lift=1.0, confidence=0.87
['avg_glucose_level_High', 'hypertension_Low'] => ['stroke_Healthy'], lift=1.0, confidence=0.95
['avg_glucose_level_High'] => ['hypertension_Low', 'stroke_Healthy'], lift=0.96, confidence=0.83
['avg_glucose_level_Low', 'hypertension_Low'] => ['stroke_Healthy'], lift=1.02, confidence=0.97
['avg_glucose_level_Low'] => ['hypertension_Low', 'stroke_Healthy'], lift=1.04, confidence=0.9
['bmi_Low', 'hypertension_Low'] => ['stroke_Healthy'], lift=1.01, confidence=0.96
['bmi_Low'] => ['hypertension_Low', 'stroke_Healthy'], lift=1.04, confidence=0.9
['smoking_status_never smoked', 'hypertension_Low'] => ['stroke_Healthy'], lift=1.02, confidence=0.97
['smoking_status_never smoked'] => ['hypertension_Low', 'stroke_Healthy'], lift=1.03, confidence=0.89
['ever_married_Yes', 'heart_disease_No'] => ['stroke_Healthy'], lift=0.99, confidence=0.94
['heart_disease_No'] => ['ever_married_Yes', 'stroke_Healthy'], lift=0.99, confidence=0.61
['ever_married_Yes'] => ['heart_disease_No', 'stroke_Healthy'], lift=0.96, confidence=0.87
['work_type_Private', 'heart_disease_No'] => ['stroke_Healthy'], lift=1.01, confidence=0.96
['work_type_Private'] => ['heart_disease_No', 'stroke_Healthy'], lift=1.0, confidence=0.91
['heart_disease_No', 'Residence_type_Rural'] => ['stroke_Healthy'], lift=1.01, confidence=0.96
['Residence_type_Rural'] => ['heart_disease_No', 'stroke_Healthy'], lift=1.0, confidence=0.91
['Residence_type_Urban', 'heart_disease_No'] => ['stroke_Healthy'], lift=1.0, confidence=0.96
['Residence_type_Urban'] => ['heart_disease_No', 'stroke_Healthy'], lift=1.0, confidence=0.9
['heart_disease_No', 'avg_glucose_level_High'] => ['stroke_Healthy'], lift=1.0, confidence=0.95
['avg_glucose_level_High'] => ['heart_disease_No', 'stroke_Healthy'], lift=0.97, confidence=0.88
['avg_glucose_level_Low', 'heart_disease_No'] => ['stroke_Healthy'], lift=1.02, confidence=0.97
['avg_glucose_level_Low'] => ['heart_disease_No', 'stroke_Healthy'], lift=1.03, confidence=0.93

['heart_disease_No', 'bmi_High'] => ['stroke_Healthy'], lift=1.01, confidence=0.96

['bmi_High'] => ['heart_disease_No', 'stroke_Healthy'], lift=0.99, confidence=0.9

['bmi_Low', 'heart_disease_No'] => ['stroke_Healthy'], lift=1.01, confidence=0.96

['bmi_Low'] => ['heart_disease_No', 'stroke_Healthy'], lift=1.01, confidence=0.91

['smoking_status_never smoked', 'heart_disease_No'] => ['stroke_Healthy'], lift=1.02, confidence=0.97

['smoking_status_never smoked'] => ['heart_disease_No', 'stroke_Healthy'], lift=1.03, confidence=0.94

['heart_disease_No', 'gender_Female', 'hypertension_Low'] => ['stroke_Healthy'], lift=1.02, confidence=0.97

['gender_Female', 'hypertension_Low'] => ['heart_disease_No', 'stroke_Healthy'], lift=1.03, confidence=0.94

['heart_disease_No', 'gender_Female'] => ['hypertension_Low', 'stroke_Healthy'], lift=1.02, confidence=0.88

['gender_Female'] => ['hypertension_Low', 'heart_disease_No', 'stroke_Healthy'], lift=1.02, confidence=0.85

['age_Young', 'heart_disease_No', 'hypertension_Low'] => ['stroke_Healthy'], lift=1.05, confidence=1.0

['age_Young', 'heart_disease_No'] => ['hypertension_Low', 'stroke_Healthy'], lift=1.12, confidence=0.97

['age_Young', 'hypertension_Low'] => ['heart_disease_No', 'stroke_Healthy'], lift=1.09, confidence=0.99

['age_Young'] => ['hypertension_Low', 'heart_disease_No', 'stroke_Healthy'], lift=1.16, confidence=0.97

['ever_married_Yes', 'heart_disease_No', 'hypertension_Low'] => ['stroke_Healthy'], lift=1.0, confidence=0.95

['ever_married_Yes', 'heart_disease_No'] => ['hypertension_Low', 'stroke_Healthy'], lift=0.96, confidence=0.83

['ever_married_Yes', 'hypertension_Low'] => ['heart_disease_No', 'stroke_Healthy'], lift=0.98, confidence=0.89

['ever_married_Yes'] => ['hypertension_Low', 'heart_disease_No', 'stroke_Healthy'], lift=0.93, confidence=0.77

['work_type_Private', 'heart_disease_No', 'hypertension_Low'] => ['stroke_Healthy'], lift=1.01, confidence=0.96

['work_type_Private', 'heart_disease_No'] => ['hypertension_Low', 'stroke_Healthy'], lift=1.01, confidence=0.88

['work_type_Private', 'hypertension_Low'] => ['heart_disease_No', 'stroke_Healthy'], lift=1.01, confidence=0.92

['work_type_Private'] => ['hypertension_Low', 'heart_disease_No', 'stroke_Healthy'], lift=1.0, confidence=0.83

['heart_disease_No', 'Residence_type_Rural', 'hypertension_Low'] => ['stroke_Healthy'], lift=1.02, confidence=0.97

['Residence_type_Rural', 'hypertension_Low'] => ['heart_disease_No', 'stroke_Healthy'], lift=1.02, confidence=0.93

['heart_disease_No', 'Residence_type_Rural'] => ['stroke_Healthy', 'hypertension_Low'], lift=1.02, confidence=0.88

['Residence_type_Rural'] => ['stroke_Healthy', 'heart_disease_No', 'hypertension_Low'], lift=1.0, confidence=0.83

['Residence_type_Urban', 'heart_disease_No', 'hypertension_Low'] => ['stroke_Healthy'], lift=1.01, confidence=0.96

['Residence_type_Urban', 'heart_disease_No'] => ['hypertension_Low', 'stroke_Healthy'], lift=1.01, confidence=0.88

['Residence_type_Urban', 'hypertension_Low'] => ['heart_disease_No', 'stroke_Healthy'], lift=1.01, confidence=0.92

['Residence_type_Urban'] => ['hypertension_Low', 'heart_disease_No', 'stroke_Healthy'], lift=1.0, confidence=0.83

['avg_glucose_level_Low', 'heart_disease_No', 'hypertension_Low'] => ['stroke_Healthy'], lift=1.02, confidence=0.97

['avg_glucose_level_Low', 'heart_disease_No'] => ['hypertension_Low', 'stroke_Healthy'], lift=1.05, confidence=0.91

['avg_glucose_level_Low', 'hypertension_Low'] => ['heart_disease_No', 'stroke_Healthy'], lift=1.04, confidence=0.94

['avg_glucose_level_Low'] => ['hypertension_Low', 'heart_disease_No', 'stroke_Healthy'], lift=1.05, confidence=0.87

['bmi_Low', 'heart_disease_No', 'hypertension_Low'] => ['stroke_Healthy'], lift=1.02, confidence=0.97

['bmi_Low', 'heart_disease_No'] => ['hypertension_Low', 'stroke_Healthy'], lift=1.05, confidence=0.91

['bmi_Low', 'hypertension_Low'] => ['heart_disease_No', 'stroke_Healthy'], lift=1.03, confidence=0.93

['bmi_Low'] => ['hypertension_Low', 'heart_disease_No', 'stroke_Healthy'], lift=1.04,
confidence=0.87

['smoking_status_never smoked', 'heart_disease_No', 'hypertension_Low'] => ['stroke_Healthy'],
lift=1.03, confidence=0.98

['smoking_status_never smoked', 'heart_disease_No'] => ['hypertension_Low', 'stroke_Healthy'],
lift=1.04, confidence=0.9

['smoking_status_never smoked', 'hypertension_Low'] => ['heart_disease_No', 'stroke_Healthy'],
lift=1.05, confidence=0.95

['smoking_status_never smoked'] => ['hypertension_Low', 'heart_disease_No', 'stroke_Healthy'],
lift=1.04, confidence=0.87

נספח ג' – מקרים של FP שלא סווגו בראוי



fp_nn_results.csv