

ממ"ן 22 – תומר ריפין 322230608

שאלה 1 – חוקי הקשר

א. בחירת האלגוריתם – A-priori

האלגוריתם מוצא קבוצות תדירות באורכים הולכים וגדלים באופן איטרטיבי על ידי בניית קבוצות מועמדות ובדיקתן אל מול תנאי הסף של תמיכה מינימלית (Minimal Support). בסיום ריצת האלגוריתם ומציאת הקבוצות השכיחות, האלגוריתם מפיק את חוקי ההקשר החזקים שמצא. לאלגוריתם זה יתרון משמעותי – הוא קל מאוד למימוש וקל להבנה. החסרון המשמעותי שלו הוא היעילות זמן הריצה אך מכיוון שמסד הנתונים שלנו 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 שבצעתי לתכונות מסוימות בממ"ן 21 יצרו קבוצות אשר שכיחותן קטנה מ-40% ולכן אבצע דיסקרטיזציה מחדש לנתונים כדי שנוכל להסיק ממנה מסקנות בהינתן רף שכיחות זה. אפרט את ההכנה מחדש של הנתונים שעשיתי – עבור התכונות הנומריות בצעתי `Equal Frequency Discretization` עם שני `Bins` כדי להגיע לקבוצות שעוברות את הסף.

עבור התכונות הקטגוריות רציתי להגיע למצב שבו לפחות לקבוצה דומיננטית יש 40%, בתכונות הבינאריות זה כמובן התאפשר. ולהן נתתי שמות כדי שפלט האלגוריתם יהיה מובן יותר. בתכונת `smoking_status` כמות ה-`Unknowns` הייתה גבוהה ורציתי לצמצם אותם משמעותית כדי ששאר הקבוצות יעברו את הסף. דבר ראשון איחדתי בין `formerly smoked` ל-`smokes` ואת שאר ה-`Unknowns` מלאתי עם השלמה לקבוצה הדומה ביותר באמצעות אלגוריתם KNN. כך שבסופו של דבר נשארתי רק עם שתי קבוצות – `smokes` ו-`never_smoked`. אסכם את מצב העמודות עכשיו:

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

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

הקבוצות התדירות שנמצאו באמצעות הרצת האלגוריתם

השתמשתי באלגוריתם a-priori אשר ממומש מראש בחבילה הפייתונית mlxtend ולהלן תוצאות ההרצה שלו:

- נמצאו 81 קבוצות תדירות
- מתוכן, 15 קבוצות תדירות בגודל 1, 38 בגודל 2, ו-28 בגודל 3.
- כמובן שלא הייתה קבוצת תדירות עבור המשתנה stroke_Stroke כי היא מתחת לרף הביטחון.
- היו 34 קבוצות תדירות שהכילו את המשתנה stroke_Healthy מעמודת המטרה. זה הגיוני בעיקר מפאת חוסר האיזון של תכונה זו.
- ניתן לצפות בכל הקבוצות התדירות ב**נספח א**.

ג. הצגת חוקי ההקשר החזקים

נחפש תבניות שקשורות לתכונת הסיווג – stroke ולכן אציג רק את חוקי ההקשר שמכילים את stroke_Healthy כאחד מהערכים הנגזרים. אציג עמודות שעומדות ברף $\text{min_confidence} = 60\%$ בלבד. יצאו 99 חוקי הקשר שכאלו, גם הם יופיעו ב**נספח ב**. כדי לסנן את חוקי ההקשר המעניינים ביותר אציג את אלו עם המדדים הגבוהים ביותר - lift ו-confidence גבוהים ($\text{Lift} > 1.1$ וגם $\text{confidence} > 0.8$) (בדקתי גם $\text{lift} < 0.9$ אך זה לא הניב כלום).

```
['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 בטרנזקציות לבין עמודת המטרה – הסיכוי של המועמד לקבל שבץ. וכדי לדייק – הסיכוי של המועמד לא לקבל שבץ (כאשר הסיכוי לקבל שבץ הוא המאורע המשלים).

מדד $\text{lift} > 1$ מראה קורלציה חיובית בין בין הפריטים בטרנזקציה, $\text{lift}=1$ מראה שאין קשר בין הפריטים. מ-3 חוקים אלו אנחנו למדים שלגיל צעיר וללחץ דם נמוך יש את הקשר החזק ביותר לכך שלא חוטפים שבץ בדומה לתוצאות שקיבלנו בממ"ן 21. נתונים רבים יותר על חולי שבץ היו יכולים לתרום לנו בכך שהיינו יכולים לצפות גם בחוקים אשר הסיפא שלהם הייתה stroke_Stroke וכך לקבל מידע משמעותי על מה גורם למועמדים לקבל שבץ ולא רק מה גורם להם להיות בריאים.

שאלה 2 – ניתוח אשכולות

א. מדדי איכות לאשכולות

לתהליך ניתוח אשכולות יש מגוון רחב של מדדי איכות. למשל – יכולות התמודדות עם נתונים רועשים, נתונים רב מימדיים מטיפוסים שונים, סקלביליות, יכולת הסתגלות של המודל לנתונים חדשים ועוד.

למדידת האיכות של האשכולות עצמם, נעזר בשני מדדי האיכות הבאים:

1. הומוגניות ושלמות – מדד שעל פיו נוכל להחליט עד כמה "המרחק" בין עצמים בכל אשכול הוא קטן – עד כמה העצמי סבכל אשכול ואשכול דומים או שונים מעצמים באשכולות האחרים. ככל שעצמים בתוך האשכול דומים אחד לשני אך שונים מעצמים באשכולות אחרים כך החלוקה נחשבת איכותית יותר.
2. מגמתיות – מדד שעל פיו אפשר לבדוק האם קיימת מגמה או תופעה, שלא נתפסת בעין האדם שניתן ללמוד עליה מהחלוקה לאשכולות. מדד זה יכול להעיד על חלוקה טובה במידה והאשכולות שנוצרו הניבו מבנים לא אקראיים.

מכיוון שאני מבצע את העבודה בפיתון עם הספרייה sklearn אשתמש במדד האיכות silhouette-score. שמאפשר למדוד עד כמה חלוקת הנתונים בסט הנתונים לאשכולות 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 האשכולות. נשתמש גם במדד זה למדידת איכות האשכולות.

ב. בחירת גישה לניתוח אשכולות

בחרתי באלגוריתם k-means לניתוח האשכולות.
אלגוריתם k-means הוא אלגוריתם איטרטיבי המקבל ערך k אשר קובע את מספר האשכולות הנדרשים ופונקציית מרחק אשר מגדירה את המרחק בין שני אובייקטים בסט הנתונים. האלגוריתם מקבץ את n הפרטים שהתקבלו בסט הנתונים תוך מזעור ריבועי המרחקים (WCSS) מכל מרכז באשכול. באלגוריתם זה לא מתבצע אימון והתחזית מתבצעת על כלל סט הנתונים. חשוב לציין שקיים trade-off בין השאיפה למזער את ה-WCSS ו-k. ככל ש-k קטן יותר כך יקטן ה-WCSS. אך ככל שנגדיל את k נייצר מספר גדול של אשכולות ויהיה לנו קשה יותר לבצע אנליזה נוחה של המידע ולייצר מסקנות.
כדי למצוא את ה-k האידיאלי נשתמש ב-Elbow Method שעליה אפרט בהמשך.

נציג פסאודו-קוד של האלגוריתם:

1. בחר k פריטים רנדומליים מתוך סט הנתונים
2. מקם את סדרת מרכזי הכובד $C = \{c_1, \dots, c_k\}$ בצורה רנדומלית עבור k הפריטים
3. כל עוד לא התכנסו או לא הגענו למגבלת האיטרציות:
4. עבור כל פריט x_i :
5. מצא את מרכז הכובד הקרוב ביותר ל- x_i מכלל מרכזי הכובד C והכניסו ל- c_{x_i}
6. הכנס את הנקודה x_i לאשכול של מרכז הכובד c_{x_i}
7. עבור כל אשכול מ- 1 עד k :
8. מקם את מרכז הכובד המורכב ממוצע הפריטים שנמצאים באשכול זה מחדש
9. חזור את האשכולות שנוצרו

יתרונות וחסרונות של אלגוריתם k-means:

יתרונות:

1. אלגוריתם טבעי ופשוט להבנה
2. קל למימוש ויעיל
3. אלגוריתם זה טוב ע-ם ערכים נומריים ורציפים שמאפיינים מידע בריאותי של חולים וכמו שראינו כאן מאפיין את תכונות אשר משפיעות מאוד על חיזוי השבץ (גיל, BMI, AGL)

חסרונות:

1. חלוקה לא אופטימלית – מבוצעות בשלבים הראשונים של האלגוריתם בחירות רנדומליות שיכולות להוביל לחלוקה לא אופטימלית
2. תלות בפונקציית המרחק – פונקציית מרחק שלא מתאימה לפרמטרים יכולה לפגוע מאוד בתוצאות הניתוח בעיקר מפאת נוחות ניתוח הנתונים ושהוא עובד טוב עם ערכים רציפים שראינו שמשמעותיים מאוד לשבץ בחרתי להשתמש ב-k-means.

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

הכנת הנתונים – בחנתי מספר סטים שונים של נתונים להשתמש בהם, לבסוף לקחתי את אותו סט הנתונים לאחר הנקיונות והסדר שעשיתי בממ"ן 21 לנתונים אבל במקום לבצע דיסקרטיזציה לתכונות הנומריות הפעלתי עליהן את הפונקציה min-max עם הגבולות [0,1] כדי לנרמל את המרחקים ולא לתת יתרון משמעותי בפונקציית המרחק לתכונות אלו על פני תכונות קטגוריות שהן 0 או 1.

מטרת כריית המידע הזו היא לנסות לאפיין אשכולות של חולים ב-"סיכון גבוה" או "סיכון נמוך". לכן, כדי לא לייצר bias בחלוקה לאשכולות הורדתי את עמודת המטרה stroke לפני הרצת k-means. רק לאחר ההרצה ארצה להצמיד חזרה את התווית ולנסות לאפיין את האשכולות לפי פיזור ה-stroke באשכול.

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

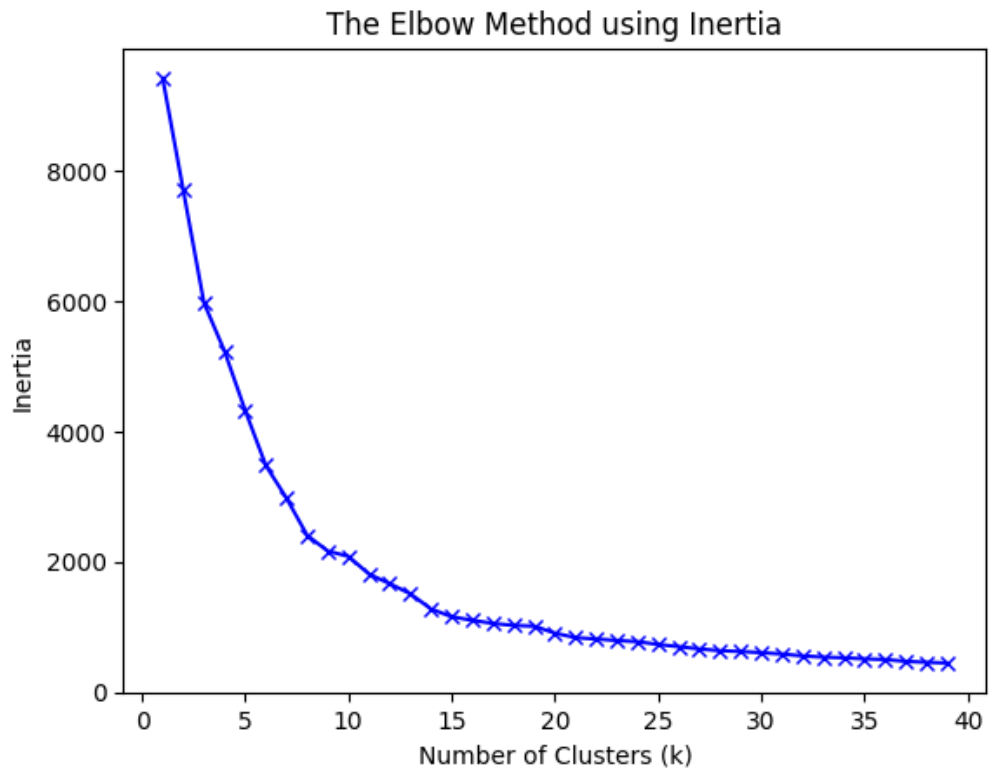
1. מספר האיטרציות המקסימלי (max_iter) - 300
2. פונקציית אתחול לחלוקת k הפריטים הראשונים (init) – יש שתי אופציות
 - a. Random – השיטה המוכרת המתוארת בפסאודו-קוד שהצגתי למעלה
 - b. K-means++ – שיטה שבחרת את k הפריטים הראשונים בצורה "חכמה" על מנת להתכנס במהירות. בחרתי בשיטה זו כי היא הניבה לי את התוצאות המדויקות יותר.
3. אלגוריתם לחישוב האשכולות – ישנן שתי אופציות אשר ממומשות תחת הספרייה
 - a. Lloyd – השיטה הקלאסית של k-means – מקצה נקודות לאשכול הקרוב ביותר אליה ומעדכן לאחר מכן את המרכזים
 - b. Elkan – משתמש בוריאנט Elkan של אלגוריתם k-means אשר משפר ביצועים לחישוב באמצעות שימוש ב-"אי שיוויון המשולש" כדי להמנע מחישובים מיותרים.

השתמשתי באלגוריתם Lloyd כי כמות הנתונים לא גבוהה במיוחד וזה האלגוריתם המומלץ והדיפולטי במידה ואין חשיבות למהירות הביצועים.

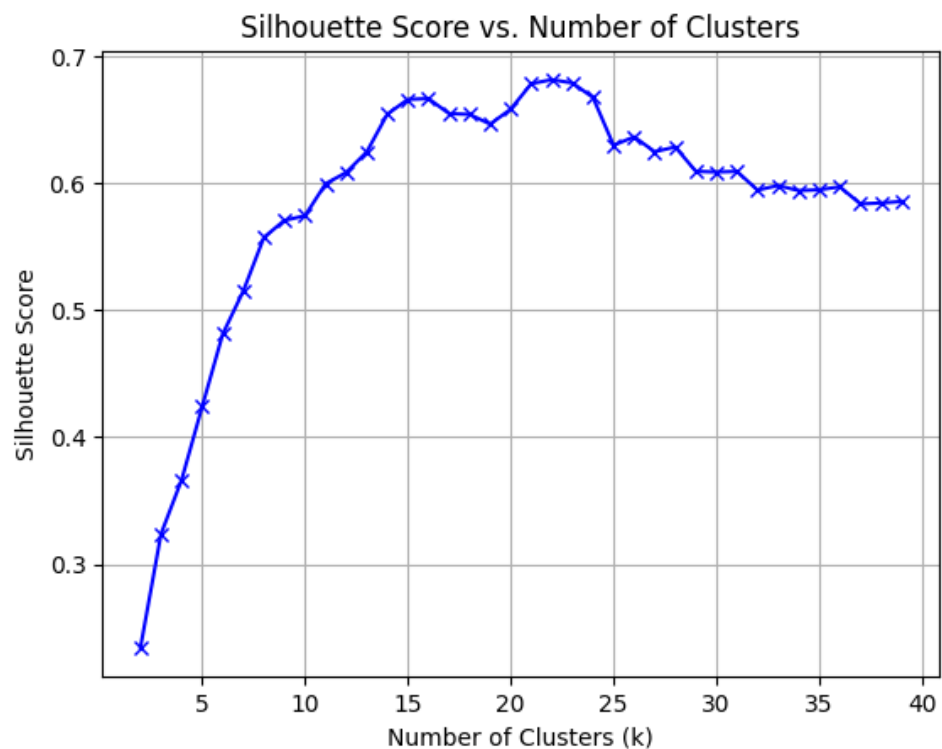
4. פונקציית המרחק – פונקציית המרחק בה משתמשת sklearn היא מרחק אוקלידי בין נקודות אשר מתואר כך: $d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$ לכל זוג נקודות x, y מכל מימד.

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

נעזרתי במדד ה-inertia של ספריית sklearn שמביע את ערך ה-WCSS עבור ריצה של k-means ובעזרתו בניתי גרף שישמש אותנו בבחירת הפרמטר עם Elbow Method:



בנוסף, ארצה למדוד מה ה-silhouette-score לכל k כדי לוודא שאני יוצר אשכולות איכותיים:

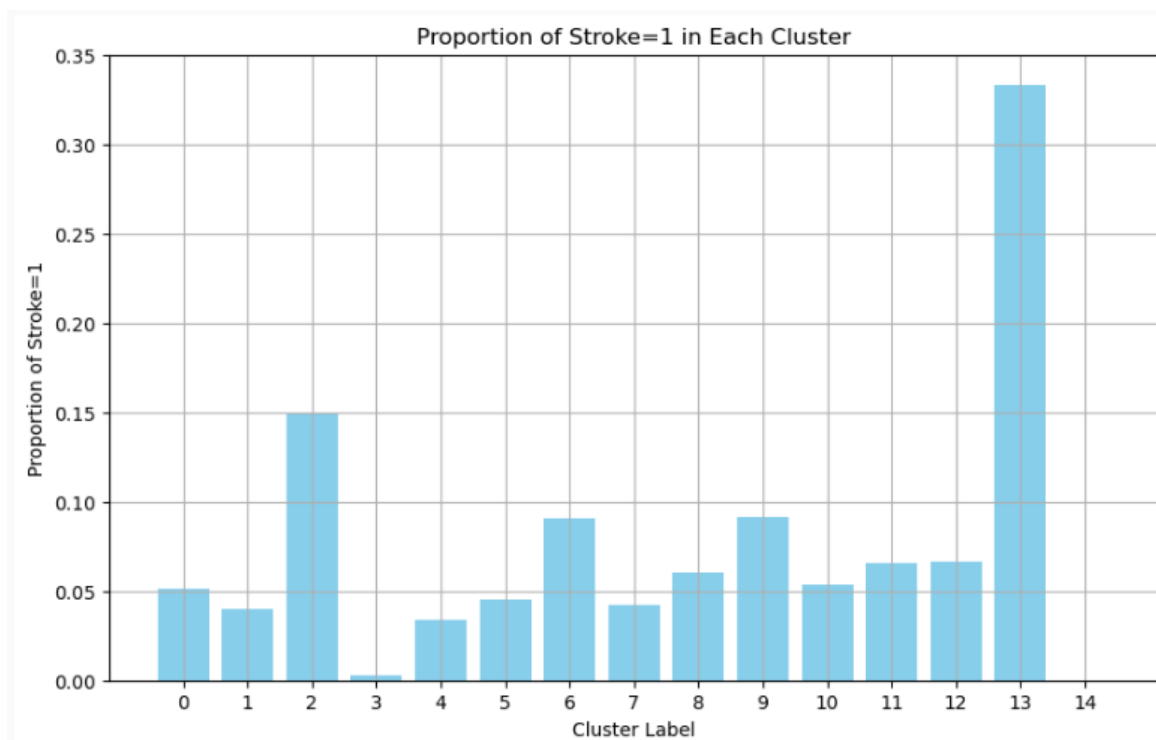


כפי שניתן לראות, סיום הירידה החדה ביותר בגרף ה-inertia מסתיים כאשר $k=9$, אך שם ה-silhouette-score עוד יחסית נמוך. ישנה עוד ירידה משמעותית בגרף ה-inertia שמסתיימת ב- $k=15$ ושם כבר ניתן לראות שהגענו כבר כמעט ל-silhouette-score מקסימלי לכן לשאר ניתוח האשכולות שלנו נבחר ב- $k=15$.

ד. דיווח התוצאות

לאחר הרצת אלגוריתם k-means על הנתונים ללא עמודת stroke עם $k=15$ קיבלנו 15 אשכולות. נמדוד את ההומוגניות של האשכולות, קיבלנו $\text{homogeneity-score} = 0.06$, תוצאה נמוכה לכל הדעות ומה שמראה שרוב האשכולות לא מאוד טהורים עם התייחסות ל-stroke ורובם מכילים תערובת של חולים ובריאים.

אציג את היסטוגרמה של פרופורציית החולים בכל אשכול:

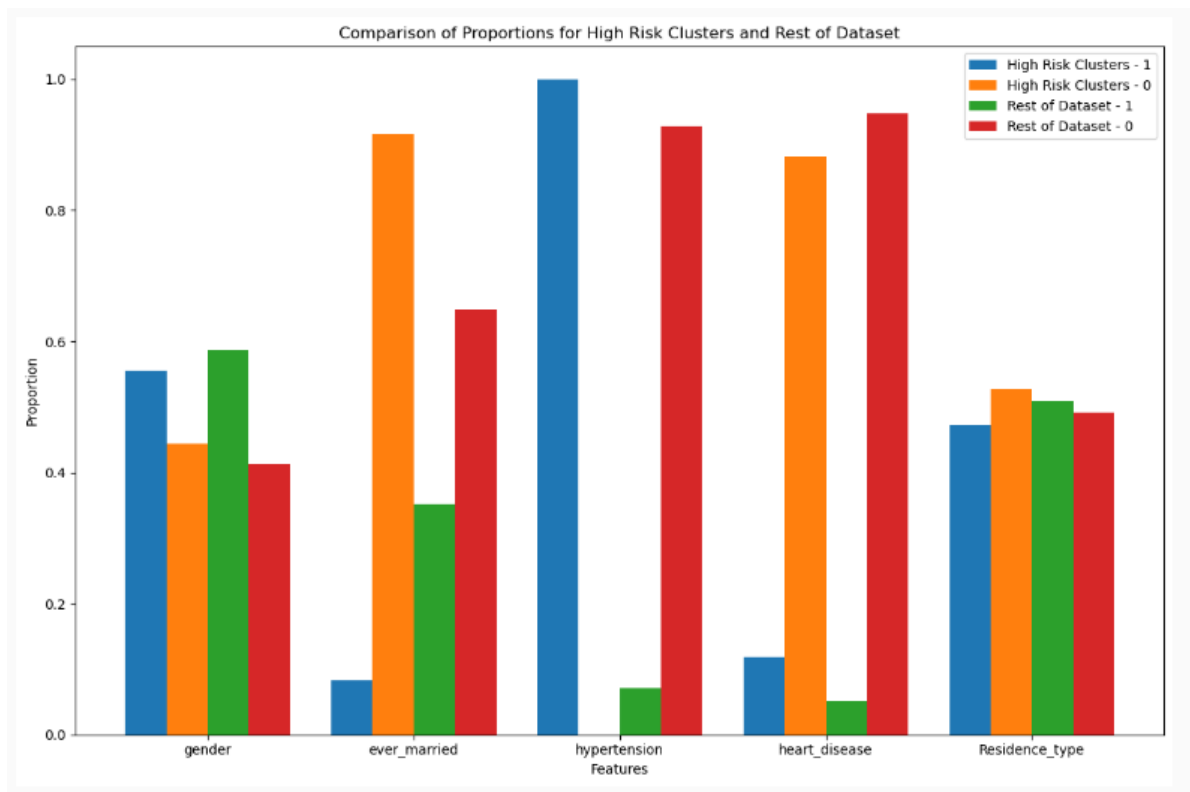


ניתן לראות שאשכול 13 הוא ב-"סיכון גבוה לשבץ" משמעותית משאר האשכולות. גם באשכול 2 יש אחוז גבוה יותר ולכן ננסה לנתח את המאפיינים של קבוצות אלו כדי לנסות למצוא את גורמי הסיכון שמאפיינים אותם.

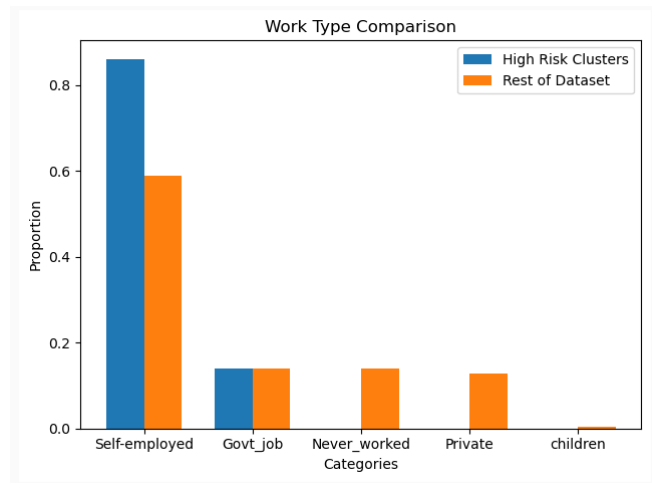
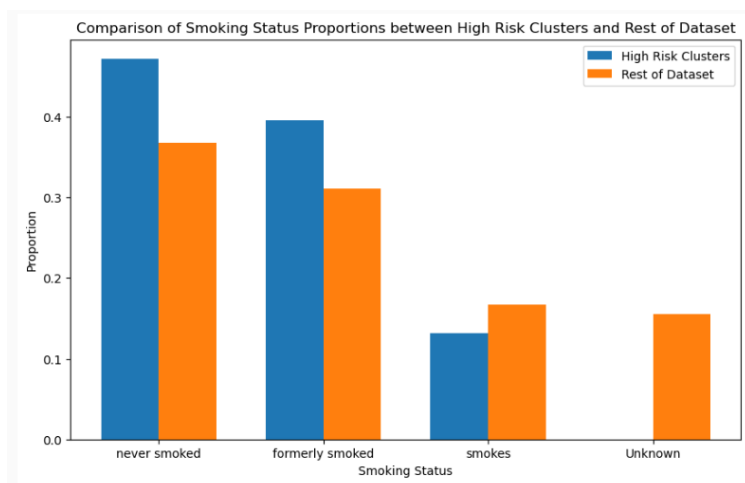
נסתכל על הממוצעים של האשכולות האלו בפרמטרים הנומריים לעומת הממוצעים של האשכולות האחרים:

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 בפרופורציית התכונות בין האשכולות עם הסיכון הגבוה לאשכולות של שאר הדאטא.



נעשה את זה גם לתכונות הקטגוריאליות:



ה. ניתוח ומסקנות

נסתכל על ההבדלים המרכזיים בין האשכולות עם "סיכון גבוה" לאלו שלא בסיכון גבוה וננסה ללמוד על התכונות המרכזיות שמאפיינות את האשכולות האלו.

שלושה הבדלים מרכזיים שקופצים לעין מיד –

1. קיים הבדל גדול מאוד בין ממוצע הגילאים של הקבוצות בסיכון לבין שאר האוכלוסיה – ממוצע הגילאים של האשכולות בסיכון היה 67 ושל שאר האוכלוסיה 43 בלבד!
2. באופן דומה, גם רמת הגלוקוז באוכלוסיה בסיכון גבוהה משמעותית משאר האוכלוסיה עם הבדל של כמעט 30.
3. נתון מדהים שניתן לשים לב אליו זה שכל האנשים בקבוצה בסיכון בניגוד מוחלט לפרופורציה שלהם באוכלוסיה סווגו עם לחץ דם גבוה – $hypertension=1$ בעוד באוכלוסיה הכללית אחוז האנשים עם $hypertension=1$ קטן מ-5%.

מכאן, נוכל להסיק שאלו פרמטרים שיכולים לאפיין חולים בסיכון בדומה מאוד לתוצאות שקיבלנו מניתוחים אחרים.

להבא, כדי לשפר את הניתוח, נוכל אולי לצמצם את המימד של הנתונים על ידי מחיקת תכונות אשר בקורלציה גבוהה אחת לשנייה כמו `age` ו-`ever_married` בכדי לנסות לנרמל את השפעתן. בנוסף, נוכל לנסות להשתמש ב-DBSCAN ולהשוות את התוצאות.

שאלה 3 – רשת נוירונים מלאכותית

א. הגדרת ארכיטקטורת הרשת והנתונים בהם אשתמש

הנתונים בהם אשתמש הם הנתונים בהם השתמשתי לניתוח האשכולות בשאלה 2, כמו אלגוריתמים לאשכול, כך גם אלגוריתמים של רשתות נוירונים רגילים ל-scale של הנתונים לכן חשוב מאוד לבצע להם סטנדרטיזציה מתאימה. אזכיר שבשאלה 2 השתמשתי ב-MinMax עם הגבולות [0-1] לנתונים הנומרים ו-one hot encoding (dummy encoding) לעמודות הקטגוריות.

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

רשת הנוירונים שנבחר הינה רשת הזנה קדמית (Feedforward Neural Network). אופן זרימת הנתונים ברשת הוא זרימה לכיוון אחד משכבת הקלט לשכבות החביות ולכיוון שכבת הפלט.

הרשת מורכבת מ-3 סוגי שכבות באופן בו בכל שכבה, כל נוירון מחובר בצלעות ממושקלות לכל אחד מהנוירונים בשכבה הבאה.

שכבת הקלט (input) – משמשת כקלט לרשת כך שכל קודקוד מייצג עמודה (feature) בנתונים שלנו. לאחר הפעלת OHE על הנתונים התקבלו 17 עמודות ולכן יהיו לנו 17 נוירונים בשכבה זו.

שכבות חביות (hidden) - נמצאות בין שכבת הקלט לפלט. לאחר ניסוי וטעייה רבים עם שתי שכבות מרובות נוירונים ((10,6), (32,16)) כדי לתפוס את מורכבות הבעיה או שכבה אחת בגדלים שונים (15,10,8). מצאתי שאת התוצאות הטובות ביותר קיבלתי עם שתי שכבות חביות, בשכבה הראשונה 10 נוירונים ובשכבה השנייה 6 נוירונים.

שכבת הפלט (output) – השכבה האחרונה ברשת המשמשת כשכבת יציאה שמציגה את תוצאות החיזוי הסופיות של הרשת. בשכבה זו יהיה נוירון אחד ויחיד, ההסתברות שתצא בנוירון זו תחולק כך שאם $p > 0.5$ אז נסווג ל-stroke=1, אחרת stroke=0.

אשתמש בשתי פונקציות הפעלה שונות, הראשונה היא סיגמויד (Sigmoid):

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

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

השנייה הינה Rectified Linear Unit (ReLU) –

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

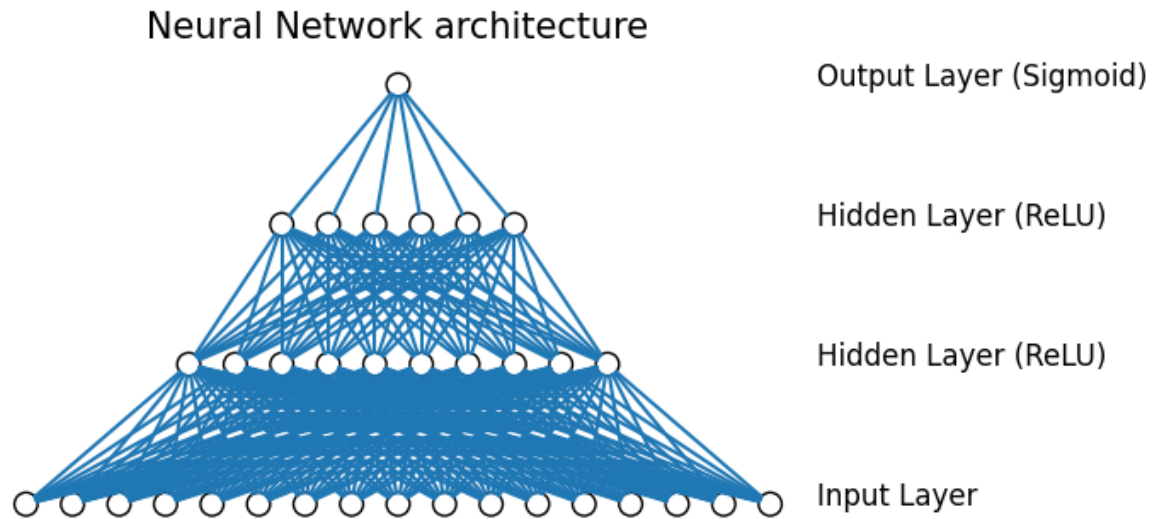
פונקצייה זו היא פונקצייה טובה מאוד לשכבות חביות ויש לה יתרונות רבים; היא יעילה חישובית (הרבה יותר מסיגמויד), ובניגוד לפונקצייה של סיגמויד היא אינה יוצרת את בעיית *Vanishing Gradient Problem*, בעיה זו נוצרת כאשר הגרדיאנט (נגזרת) שלה בסופו של דבר חסומה ב-0.25 ברשתות עמוקות מחשבים את הגרדיאנט על ידי כלל

השרשרת ושימוש בפונקציית סיגמויד יכול לאחר שכבות לא רבות לאיפוס של הגרדיאנט כי $0 \leq grad \leq \left(\frac{1}{4}\right)^k$

כאשר k הוא מספר השכבות ברשת. בעיה זו גורמת לכך שעדכונים למשקלים בשכבות קודמות כמעט לא מחלחלים לשכבות הבאות והרשת עוצרת מלמוד.

ReLU פותרת בעיה זו ע"י שעבור ערכים חיוביים הגרדיאנט יהיה תמיד 1 וכך יחלחל בקלות ברשת. למרות שבמקרה שלנו הרשת אינה עמוקה, ההבדלים בריצה בין סיגמויד ל-*ReLU* היו זניחים ולכן העדפתי להשתמש ב-*ReLU* כ- *best practice*.

להלן ציור של ארכיטקטורת הרשת:



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

פרמטרים של תהליך אופטימיזציה –

פונקציית העלות השגיאה שנשתמש בה היא Binary Crossentropy – פונקציה זו היא המתאימה ביותר לבעיות סיווג בינארי אשר אנו נדרשים לפתור. פונקציה זו מוגדרת כך:

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

כאשר y מייצגת את ה-label האמיתי של החולה (stroke=0 or stroke=1) ו- $p(y)$ זו תוצאת החיזוי של המודל (הסתברות שנעה בין 0 ל-1).

אלגוריתם האופטימיזציה שלנו יהיה (SGD) Stochastic Gradient Decent – אלגוריתם אופטימיזציה שמעדכן את המשקלים במודל באמצעות. חישוב הגרדיאנט באלגוריתם זה מבוסס על קבוצת תצפיות אקראיות קטנה. בחלק מהמקרים גישה זו מסייעת בהפחתת זמני הריצה והמנעות ממינימום מקומי.

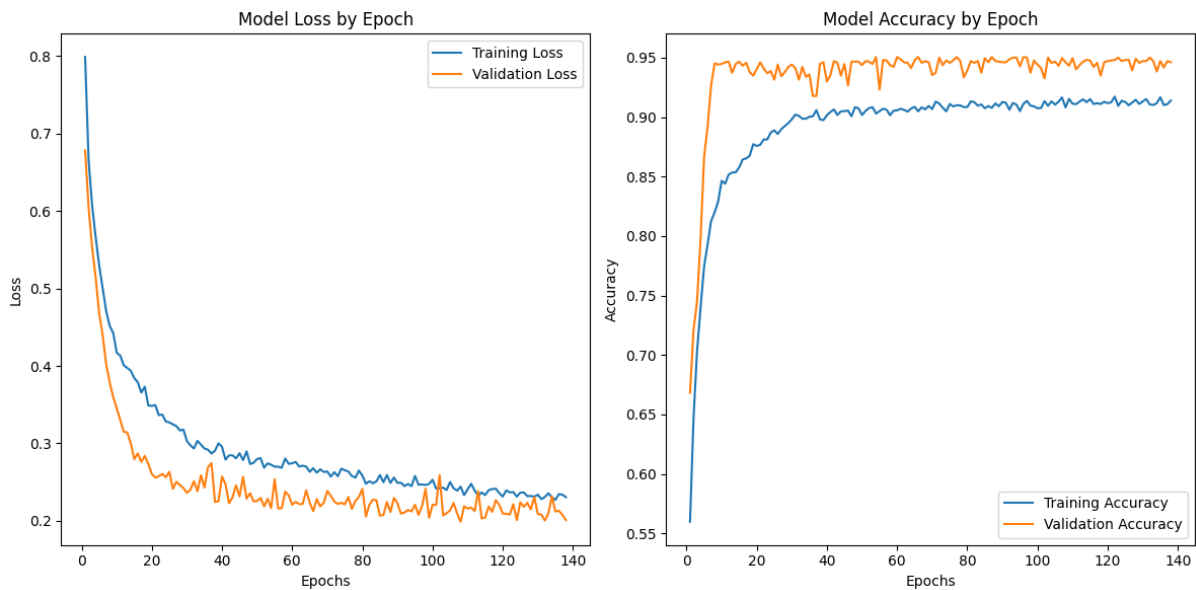
הוא פשוט ומתאים לכמות גדולה של נתונים (אלפים) אך חשוב לזכור שיש לו התכנסות יחסית איטית.

קצב הלמידה - בגלל שאנחנו משתמשים ב-SGD נצטרך להיות זהירים עם קצב הלמידה, SGD רגיש מאוד לשינויים בקצב הלמידה; קצב למידה גבוה מדי יכול לגרום לאלגוריתם לסטות (to diverge) בעוד קצב למידה נמוך מדי יגרור האטה בהתכנסות הלמידה. לאחר ניסוי וטעייה, ראינו שקצב הלמידה 0.01 היה הטוב ביותר מבחינת יציבות וקצב השיפור שניתן לראות בגרפים של הדיוק ופונקציית השגיאה לאורך ה-epochs.

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

ג. הרצת הרשת ודיווח Epochs גרף השגיאה עבור נתוני האימון והמבחן

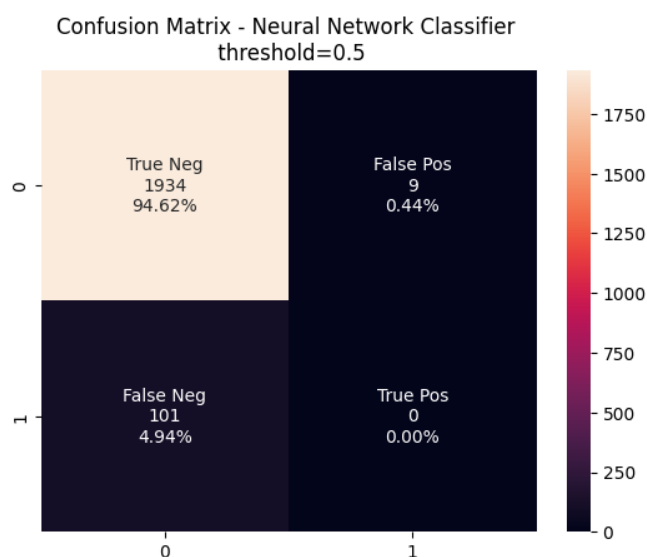
לאחר הרצה של הרשת, אציג את גרף ביצועי הרשת (accuracy) לאורך ה-epochs של האימון ביחד עם גרף השגיאה עבור נתוני האימון והמבחן בפונקציה של ה-epochs.



ניתן לראות שמאוד מהר המודל הצליח להתכנס מבחינת ה-accuracy (ב-epoch 40 בערך) וכך גם מבחינת ה-Loss אך עדיין אפשר לראות ירידה קלה ויציבה ב-Loss. כדי להחליט מתי לסיים ללמוד ומתי המודל הגיע למיצוי של הלמידה השתמשתי במיקרופרמטר EarlyStopping. פרמטר זה עוזר לזהות מתי המודל מגיע לסטגנציה בלמידה, הוא לוקח פרמטר patience שמגדיר תוך כמה epochs צריך לראות שיפור ב-Validation Loss לפני שהוא עוצר את הלמידה. לאחר שההבין שהלמידה עצרה הוא מחזיר את המודל להיות עם המשקלים של ה-Loss הנמוך ביותר לאורך הלמידה ולא הסופי. לאורך למידה זו הגענו למינמום ב-Epoch=108.

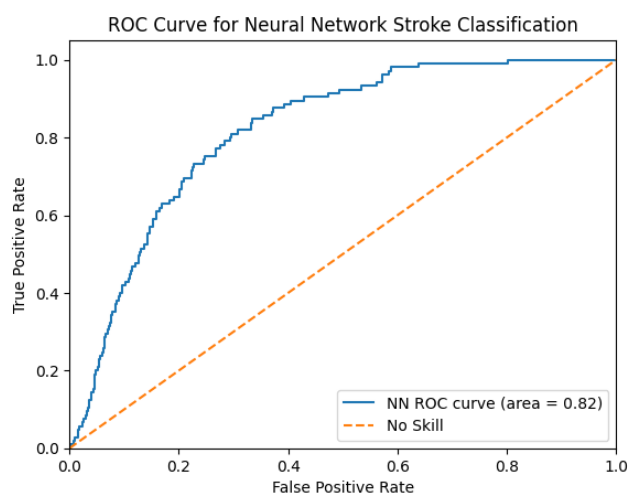
השתמשתי בנוסף במיקרופרמטר בשם ReduceLROnPlateau – מיקרופרמטר זה משמש בעיקר לעזור למודלים להתאים את קצב הלמידה שלהם ברגע שהם מגיעים לסטגנציה. השתמשתי בפרמטר זה כדי לסחוט עוד למידה מהמודל אחרי שהוא הגיע ל-epoch=60 בערך וזה אכן עזר להוריד עוד קצת את ה-Validation Loss.

נסתכל על תוצאות ה-Confusion Matrix הראשוניות –



זה מתאים מאוד למה שראינו בגרף ה-accuracy, לא הצלחנו לסווג אף מקרה אחד של stroke=1. ראינו כבר שבמודלים קודמים בממ"ן 21 היינו צריכים לשחק עם ה-threshold של המודל כדי להגדיל את הרגישות שלו למקרים שבהם הוא חושד אפילו קצת שמדובר במקרה של שבץ. אבל לפני הכל צריך לוודא שהמודל בכלל למד משהו ושהוא לא סתם מנחש תמיד 0.

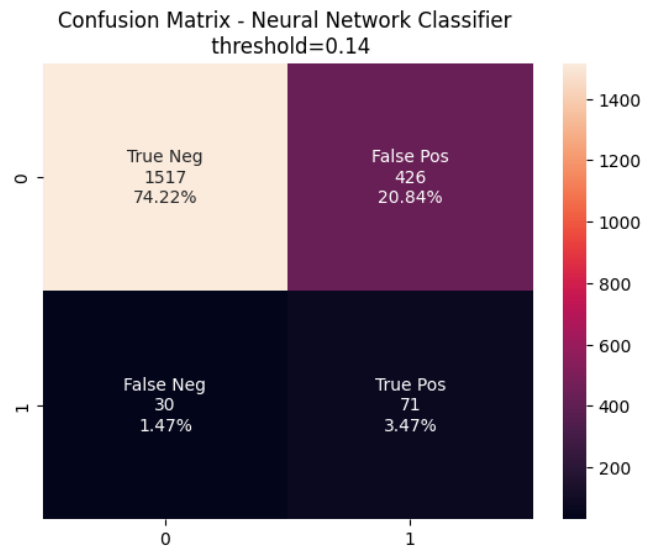
נסתכל על ה-ROC Curve כדי לוודא זאת:



נינוכח לראות ה-AUC שלנו הוא 0.82, תוצאה שנחשבת אפילו יותר טובה משני המודלים שהשתמשנו בהם בממ"ן 21. זה אומר שהמודל יש יכולת חזקה להבדיל בין מקרים של שבץ, ספציפית זה אומר שיש לו סיכוי של 82% לסווג מקרה רנדומלי חיובי מעל מקרה רנדומלי שלילי. זה אומר שהבעיה שנתקלתי בה שהמודל לא הצליח לסווג אף מקרה של stroke=1 נכון היא לא בגלל שהוא לא הצליח ללמוד אלא בגלל שה-threshold היה לו נמוך מדי.

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

לאחר כמה ניסיונות מצאתי שה-threshold האופטימלי יושב ב-0.14 (אם המודל מחזיק הסתברות גדולה מ-0.14 אז תסווג את החולה כ-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 מהתוצאות של המודלים בממ"ן 21.

ד. מקרים חריגים בהם היה ניתוח שגוי

התבצעו הרבה מקרים של סיווג שגוי (426 FP, 30 FN), נציג את מקרי ה-FN כי הם מעניינים הרבה יותר וחשוב להבין למה הם התפספסו דווקא. מקרי ה-FP יהיו בנספח ג'.

id	gender	age	hypertensi	heart_disea	ever_marr	Residence	avg_gluco	bmi	work_type	work_type	work_type	work_type	work_type	work_type	smoking_	smoking_	smoking_	smoking_	pred	stroke
211	62439	1	0.621582	0	0	0	0.223017	0.194731	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	0	1
86	3253	0	0.743652	0	1	0	0.261703	0.194731	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	0	1
225	39186	1	0.694824	0	1	0	1.0745361	0.237113	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	0	1
219	31421	0	0.890137	0	1	0	0.759902	0.209622	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	0	1
180	54567	1	0.560547	0	0	0	1.0106454	0.234822	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	0	1
182	39912	1	0.389648	0	0	0	0.09699	0.224513	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	0	1
140	20439	0	1	0	1	0	0.224171	0.168385	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	0	1
186	16077	0	0.768066	0	1	0	1.028423	0.277205	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	0	1
78	45805	1	0.621582	0	0	0	1.0508679	0.203895	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	0	1
97	56841	0	0.707031	0	1	0	0.8562	0.241695	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	0	1
153	12482	0	0.829102	0	0	0	1.0104792	0.197022	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	0	1
62	65842	1	0.816895	1	0	0	0.031484	0.171821	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	0	1
100	12363	0	0.780273	0	1	0	1.087619	0.211913	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	0	1
141	45965	1	0.719238	0	0	0	0.283076	0.154639	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	0	1
143	37651	1	0.841309	1	1	1	1.078709	0.303551	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	0	1
192	36255	0	0.719238	0	0	0	0.290416	0.28866	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	0	1
204	62019	0	0.658203	0	0	0	0.151094	0.238259	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	0	1
34	14248	0	0.584961	0	0	1	1.0134244	0.222222	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	0	1
81	26015	1	0.804688	0	0	0	1.0213877	0.203895	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	0	1
87	71796	1	0.853516	0	1	0	0.019527	0.252005	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	0	1
224	8899	0	0.597168	0	0	1	0.229619	0.247423	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	0	1
41	1261	0	0.658203	0	0	0	1.074324	0.208477	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	0	1
18	27458	1	0.731445	0	0	1	1.0157419	0.315006	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	0	1
31	33879	0	0.511719	0	0	0	0.130597	0.172967	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	0	1
162	69768	1	0.015137	0	0	1	1.0704	0.203895	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	0	1
93	37726	1	0.975586	1	0	0	1.062044	0.182131	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	0	1
239	32221	0	0.731445	0	1	0	1.0169883	0.293242	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	0	1
236	28493	0	0.694824	0	0	0	1.0143939	0.245132	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	0	1
161	16590	0	0.865723	0	1	0	1.012298	0.203895	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	0	1
83	66638	1	0.829102	1	0	1	1.0113886	0.222222	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	0	1

נוכל לראות שקבוצת ה-FN שלנו מורכבת מאנשים מבוגרים אבל לא מאוד (אחוזון 70 לעומת אחוזון 84 של החולים). הם היו באחוזון נמוך משמעותית ברמת ה-AGL (36 לעומת 22). הבדלים משמעותיים אלו, וכך גם הבדלים אחרים היו יכולים לגרום לכך שסווגו באופן לא מדויק אך קשה מאוד להסיק את זה מניתוח שטחי יחסית זה.

ה. ניתוח התוצאות והסקת מסקנות

בצעתי ניתוח לאורך כל הסעיפים למעלה אך אוסיף כאן סיכום ומסקנות עיקריות שלי משאלה זו.

גם כאן, כמו בניסיונות הסיווג האחרים שלי בממ"ן 21 נתקלתי בבעיה מרכזית והיא data imbalance. גם כאן המודל התקשה מאוד לחזות חולי שבץ בצורה טובה ויצר הרבה FP בשביל מעט FN. בניגוד למסווגים אחרים כמו עצי החלטה שנלמדו ביחידות קודמות, קשה יותר להבין לאחר האימון של המודל לפי מה הוא החליט לסווג וללמוד משם על התכונות הרלוונטיות כדי "לסמן" מועמדים עם פרמטרים רלוונטים לקבוצת סיכון כלשהי והרבה יותר צריך להסתמך על המודל כ-black box.

גם כאן, למרות שאכן קשה יותר להבין מה גורם למודל להתנהג כמו שהוא, ניתן היה לראות שהוא נטה לסמן חולים מבוגרים יותר ועם רמת גלוקוז גבוהה יותר כמועמדים לשבץ. התנהגות שדומה מאוד לתוצאות שראינו לאורך כל השאלות בממ"ן וגם בממ"ן הקודם.

סיכום ומסקנות

נשווה 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

נוכל לראות שלמרות השוני המהותי בין המודלים, לאחר אפטומם ועבודה אינטנסיבית לא היה הבדל גדול בתוצאות שהם נתנו לפחות על הנייר. נראה ש-NN ו-RF היו טובים יותר במעט מ-CART אך יחסית שקולים אחד לשני בכל שאר הפרמטרים.

באופן עקבי ולאורך כלל התוצאות של המודלים השונים, גם אלו ממשפחת ה-Supervised Learning וגם ממשפחת ה-Unsupervised Learning ראינו קורלציה כזו או אחרת בין תכונות מסוימות בסט הנתונים לבין עמודת המטרה. ראינו שלגיל הפציינט (age), רמת הגלוקוז הממוצעת בדם (AGL) ולחץ דם גבוה (hypertension) הובילו לסיכוי גבוה יותר לשבץ. ודווקא פרמטרים כמו BMI, smoking-status שחשבתי אולי בהתחלה שיכולה להיות להם קורלציה לשבץ לא ראינו שמופיעים בכלל כגורמים משפיעים.

ראינו את הפרמטרים האלו בצורה בולטת במיוחד ובנחות דרך תוצאות האשכול וחוקי ההקשר אך גם דרך צפייה בעץ ההחלטה שנוצר ב-CART וגם דרך ניתוח הדגימות שלא סווגו כראוי ברשת הניורונים יכולנו לראות את השפעת התכונות האלו על המודלים ועל התוצאה. הקונסיסטנטיות הזו מראה על כך שעבודת חקר הנתונים התבצעה כראוי.

כמה נקודות לסיכום שחשובות בשבילי –

איסוף הנתונים –

- למרות כל התוצאות שקיבלנו והיכולת שלנו להשתמש במודלים שייצרנו לחיזוי שבץ, המודלים לא הגיעו לרמת דיוק גבוהה. סיווג הבריאות של אדם זו משימה קשה מאוד וצריכה להתבסס על הרבה יותר נתונים, תכונות ופרמטרים כדי להגיע לדיוק גבוה וגם אז זה יהיה כמעט בלתי אפשרי לדייק.
- הסתמכנו על כמות זעומה של מקרים שבהם stroke=1 ולכן נאצלנו להשתמש באלגוריתמים ליצירת דגימות סינטטית שעזרה אך במעט לשפר את המודל.
- שיפור נוסף לאיסוף הנתונים היה יכול להיות באיסוף תכונות שימושיות יותר ולא קורלטיביות אחת לשנייה, למשל התכונה ever_married הייתה קורלטיבית מאוד לגיל ולא תרמה כלל לסיווג הנתונים.

הכנה וניקוי הנתונים –

- סידור וניקוי הנתונים יכול להשפיע דרסטית על תוצאות המחקר, חייבים לשים לב לכל פעולה שעושים ולהבין אותה טוב מאוד אחרת יכולים לפגוע בצורה דרסטית בתוצאות המודל מבלי לדעת זאת.
- פעולת המחקר צריכה לקרות באופן איטרטיבי שבו מבצעים ניסוי וטעייה, מנסים שיטה מסוימת רואים את התוצאות ואז חוזרים לומדים מתקנים ומשפרים כל הזמן את המודל ואת הלמידה.

בנימה אישית, אהבתי מאוד את ההתנסות הרבה בנושאים השונים, אני מרגיש שהלמידה דרך הידיים והמחשב משמעותית ובלתי נפרדת 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 כראוי



fp_nn_results.csv