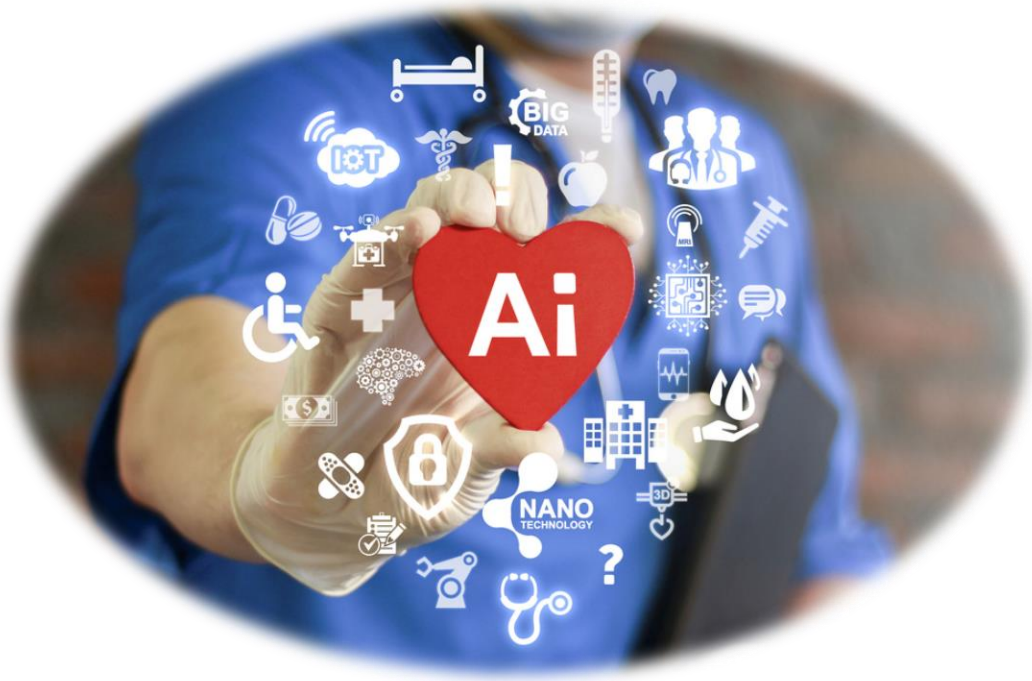
כריית מידע - פרוייקט סוף

גילוי מחלות לב



שמות הסטודנטים:

שון כמיליאן 313373359

בר לוי

גל תשובה 315534313

שם המרצה:

אוהד איזנהנדלר

גילוי מחלות לב

תקציר:

מחלות לב הן גורם התמותה השני בישראל. הן מחלות המעידות על תפקוד לא תקין של הלב, ויכולות להיגרם כתוצאה מאי אספקת הדם הדרושה ללב, הפרעות בקצב שריר הלב או כתוצאה מהיצרות של המסתמים. מחלות לב פוגעות בתפקוד התקין של הלב וביכולתו של הלב להזרים דם אל חלקי הגוף. כתוצאה מכך, הן עלולות לגרום במקרים מסויימים לנכות ואף למוות.

מטרתו העיקרית של המודל היא להוות כלי לחיזוי מחלות לב, ושיפור תפקודו של הלב במקרים בהם המודל מזהה סיכון לחלות בה. המודל יתבסס על נתונים מקיפים המכילים את הגורמים העיקריים למחלות לב, וישמש כגורם המשפיע על התפתחות מחלות לב ואף על שינוי אורח החיים של המטופלים.

מאגר הנתונים עליו הפרוייקט מתבסס [1] מכיל נתונים על מטופלים שונים ומציג את מצב הלב שלהם – כלומר אם מטופל חלה או לא במחלת לב. המודל יבצע קלאסיפיקציה תוך שימוש בנתונים שבמאגר ויחזה אם המטופל נמצא בסיכון למחלת לב. כמו-כן, בוצע ניקוי והפחתה של מידע, מתוך רצון לקבל את התוצאות הטובות ביותר.

מבוא:

הלב נמצא במרכז גוף האדם והוא מתפקד כמשאבה האחראית על זרימת הדם אל איברי הגוף השונים. התהליך בו החדר במצב רפוי ומתמלא בו דם נקרא דיאסטולה, והתהליך בו הלב מכווץ נקרא סיסטולה.

מחלות לב נחלקות לשני סוגים – מחלות לב מולדות ומחלות לב שנרכשות במהלך החיים (בד"כ כתוצאה מגורמי סיכון כמו סכרת או כולסטרול גבוה). הן יכולות להתגלות כשיש בעיה בהזרמת הדם אל הלב או בתפקודו של הלב. כאמור, הלב הינו האיבר המשמעותי ביותר בגוף האדם ולכן, פגיעה בתפקודו עלולה לגרום לבעיה בהזרמת הדם לחלקי הגוף, התקפי לב, אי-ספיקת לב ובמקרים מסויימים גם למוות.

היתרון בניתוחים מסוג זה הוא שהם מאפשרים גילוי גורמי הסיכון והשפעתם על התפתחות מחלות לב. יכולת חיזוי חכמה בהתאם לגורמי הסיכון יכולה למנוע את התדרדרות התפקוד של הלב, ולחסוך עלויות והוצאות על טיפולים (כמו ניתוחים או שעות עבודה) למערכת הבריאות.

באמצעות data mining חכם, מתקבלת האפשרות להעריך את מצבו ותפקודו של הלב דרך נתונים שמתקבלים מהמטופל, טיפול מקדים ושיפור המצב הקיים. בנוסף, מתאפשר גם לחזות את התוצאות העתידיות של המטופל. כך עשוי לסייע ה data mining בהיבטים נוספים כמו שמירה על איכות החיים של המטופל, והפקת תועלת מהטיפול בו.

המחקר באמצעות data mining יתבצע עבור החולים והמטופלים עצמם, וכן יאפשר גילוי מוקדם של אפשרות בה מטופל עלול לחלות במחלת לב. במידה של הימנעות ממחקר מסוג זה, מצבו הבריאותי של מטופל כלשהו עלול להמשיך להתדרדר. התדרדרות כזו ניתנת לעצירה באמצעות מתן תשומת לב לנתונים שמתקבלים מתוצאות בדיקותיו של המטופל, וכן שימוש ב- data mining חכם.

במטרה להפחית את התמותה כתוצאה ממחלות לב יש צורך לחזות את הסיכוי לחלות בהן תוך התייחסות לנתונים כמו רמת כולסטרול, משקל, גיל או מדד BMI. המודל בפרוייקט זה יציג ניסיון לחזות את סיכוי זה תוך שימוש בלמידה מפוקחת ובעזרת סיווג – קלאסיפיקציה.

חומרים ושיטות:

המאגר:

מאגר הנתונים הנבחר משקף את הבעיה העסקית שתתואר בפרוייקט, והוא מתבסס על 70,000 דגימות, 13 עמודות שהן המשתנים ועמודת מטרה אחת. הבעיה העסקית הינה בעיה מפוקחת, והיא זיהוי הסיכוי להיווצרות מחלת לב בהשפעת גורמי הסיכון, שהם המשתנים שנבדקו. המאגר מפרט את: מספר מזהה, גיל בימים, גיל בשנים, גיל בימים, מין, גובה, משקל, לחץ דם סיסטולי, לחץ דם דיאסטולי, כולסטרול, רמת ריכוז הסוכר, רמת האלכוהול בדם ורמת הפעילות האקטיבית שהמטופל מבצע.

ערך המטרה הנבחר: האם יש סיכוי למחלת לב או לא.

ערך זה הינו ערך נומינלי המכיל שני ערכים (True\False). ניתוח ערך המטרה ותוצאותיו יכול להוות מידע שימושי עבור אוכלוסיות הנמצאות בסיכון למחלה. איתור גורמי הסיכון יגרום לנמצאים בקבוצות אלו לשפר את אורח חייהם וגילוי מוקדם של מחלות לב.

Pre-processing:

לאחר בחינת מאגר הנתונים שמכיל כ70,000 רשומות (דגימות), הוחלט להשתמש במאגר כולו, כדי לקבל תוצאות בעלות דיוק מקסימלי אפשרי.

Data Cleaning:

בזמן סריקת הדגימות שהופיעו במאגר, לא הופיעו ערכים חסרים, אבל נמצאו ערכים חריגים שהיו קטנים מאוד עבור המשתנה שנמדד (כמו לחץ דם סיסטולי 30), או ערכים שליליים (גובה או משקל למשל). ערכים אלו נמחקו מהמאגר. בנוסף, נמצאו דגימות רבות בהן נתוני לחץ הדם היו הפוכים. במאגר הנתונים יש שתי עמודות המשקפות את לחץ הדם – האחת, לחץ גבוה (מתארת את הסיסטולה) והשנייה לחץ נמוך (מתארת את הדיאסטולה). במאגר הופיעו דגימות בהן ערך הלחץ דם הגבוה היה נמוך מלחץ הדם הנמוך (כנראה נבע מטעות אנוש), והן נמחקו מהמאגר.

בסוף התהליך נותרו דגימות בהן לחץ הדם הסיסטולי הוא בין 70 ל-200, לחץ הדם הדיאסטולי הוא בין 40 ל-120, וללא דגימות עם ערכים שליליים כלשהם. המטרה כאן, הייתה להשאיר גם דגימות בהן לחץ הדם הוא נמוך, תקין או גבוה [4].

ניקוי המאגר בוצע באמצעות סיפריית "pandas" ושימוש בקוד פייתון. בסופו של התהליך, נותרו 68,587 רשומות.

Data Reduction:

לאחר סריקת העמודות, נמצאו מספר שדות בהם בוצעו שינויים בכדי לייעל את המאגר ולהשתמש בנתונים שרלוונטיים לחיזוי.

בשלב ראשון, העמודה הראשונה שנמחקה היא המספר המזהה, משום שאין לו כל השפעה על ערך המטרה ואינו תורם בגילוי מחלות לב. לאחר מכן, הוחלט שהגיל בשנים מתאר בצורה הטובה ביותר את גיל הנבדק. לכן עמודת הגיל בימים הורדה כדי למנוע כפילויות. כמו-כן, הגיל בשנים מוצג במספרים שלמים בלבד, לשם דיוק המידע וביצוע הדיסקיטיזציה בהמשך.

Data Transformation:

לאחר ניקוי והורדת הנתונים, התקבלו 11 עמודות משתנים, ומשתנה מטרה אחד. עמודות הגובה והמשקל הומרו לעמודה אחת שתיארה את מדד ה-BMI. תחילה, בעזרת קוד פייתון הגובה הומר למטרים, ודרך הנוסחה $\frac{\text{משקל}}{\text{הגובה}^2}$ חושב מדד הBMI לכל מטופל. התוצאות שהתקבלו עברו דיסקריטיזציה, וחולקו לארבע קבוצות בהתאם למדדי הBMI המקובלים, כמפורט בנספח א':

- כל הערכים עד 18.5 קיבלו את הספרה 1 שתיארה מצב של תת משקל.
- כל הערכים מ-18.5 עד 24.5 (לא כולל) קיבלו את הספרה 2, שהצביעה על מצב תקין.
- כל הערכים מ-24.5 עד 29.5 (לא כולל) תיארו מצב של השמנה דרך הספרה 3.
- שאר הערכים קיבלו את הספרה 4 שהצביעה על מצב של השמנה חמורה.

בהמשך, באמצעות קוד פייתון הגיל בשנים עבר גם כן דיסקריטיזציה וחולק לשלוש קבוצות, כאשר בכל קבוצה מספר הנבדקים היה זהה (equal width):

- הקבוצה הראשונה קיבלה את הערך 1, והכילה נבדקים בגילאי 30 עד 41.667 (כולל).
- הקבוצה השנייה קיבלה את הערך 2, והכילה נבדקים בגילאי 41.667 עד 53.33.
- הקבוצה השלישית הכילה את הנבדקים המבוגרים ביותר וקיבלה את הערך 3.

בנוסף, נבחנה האפשרות לעשות דיסקריטיזציה עבור לחץ הדם, כדי להימנע מ-overfitting. הדיסקריטיזציה כללה את השלבים הבאים:

תחילה, העמודות המתארות את לחץ הדם (ap_hi, ap_lo) הומרו לעמודה אחת שתיארה את לחץ הדם הממוצע בעזרת הנוסחה: $MBP = \frac{ap\_hi}{3} + \frac{2 \cdot ap\_low}{3}$. לאחר מכן בוצעה דיסקריטיזציה בה חולק לחץ הדם לרמות [2,3]:

- הערכים שקטנים מ-70 קיבלו את הספרה 1 (לא כולל), ותיארו לחץ דם נמוך.
- הערכים שבין 70 ל-103.33 קיבלו את הספרה 2 ותיארו לחץ דם תקין.
- הערכים הגבוהים מ-103.33 קיבלו את הספרה 3 והצביעו על לחץ דם גבוה.

כתוצאה מכך, התקבלו שני מאגרי נתונים. אחד הכיל דיסקריטיזציה רק לפי ה-BMI והגיל (דאטה 1), והשני הכיל דיסקריטזציות לפי המתואר לעיל (דאטה 2).

Algorithms:

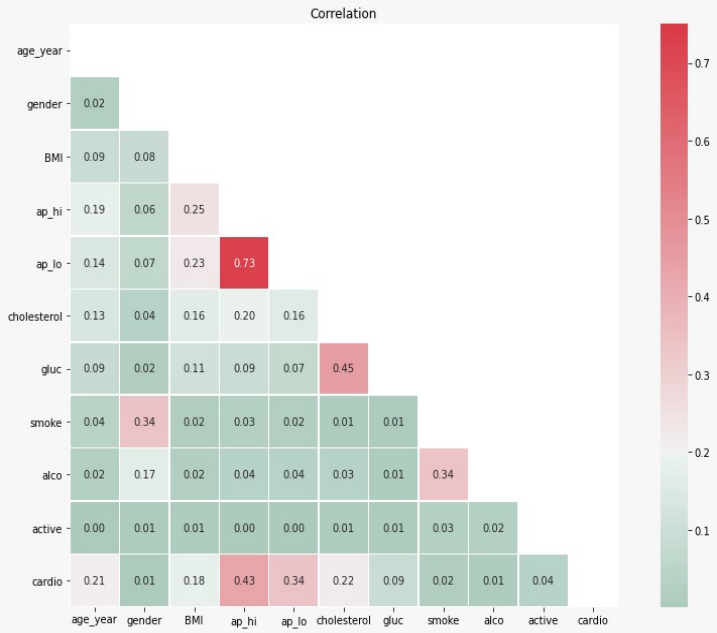
נבחנו ערכי ה- ROC וה-Accuracy דרך אלגוריתם עץ החלטות תוך הכנסת מאגרי הנתונים. בנוסף, נבדוק גם את מדד ה-Recall לכל אחד מהמאגרים.

- Decision tree (J-48) – אחת מהגישות לחיזוי תפציות, בעץ החלטה בודקים באיזה סדר יש להסתכל על המשתנים כדי לקבל כמה שיותר הומוגניות. כלומר, למצב שבו יש וודאות לגבי התחזית. נקודת ההתחלה היא משתנה מסביר מסויים, ממנו יוצאים ענפים, וכל ענף מגיע למשתנה מסביר אחר או להכרעה.

נרצה שערך ה- ROC יהיה כמה שיותר קרוב ל-1, והוא מתאר את השטח שנותר מתחת לגרף TPR vs. FPR. בנוסף, נרצה שערך הדיוק (Accuracy) יהיה גדול ככול האפשר, משום שהוא מראה את האחוז בהן המסווג צדק, ביחס לכל הדגימות. ונבחן את ה-Recall כאשר ערך המטרה הוא 1 (כלומר, בסיכון לחלות), ונרצה שיהיה כמה שיותר קרוב ל-1.

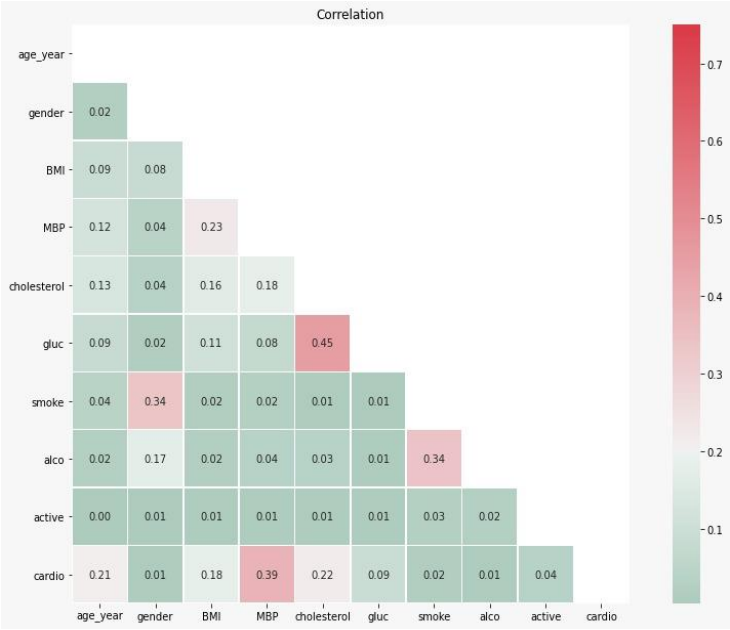
בנוסף, בדקנו את ערכי הקורולציה שהתקבלו עבור שתי הדטאות:

דאטה 1:



ניתן לראות שהקורולציה המקסימלית היא בין המשתנים לחץ דם סיסטולי (ap_hi) לבין לחץ דם דיאסטולי (ap_lo), ובאופן כללי, הקורולציה בין המשתנים נמוכה יחסית.

דאטה 2:

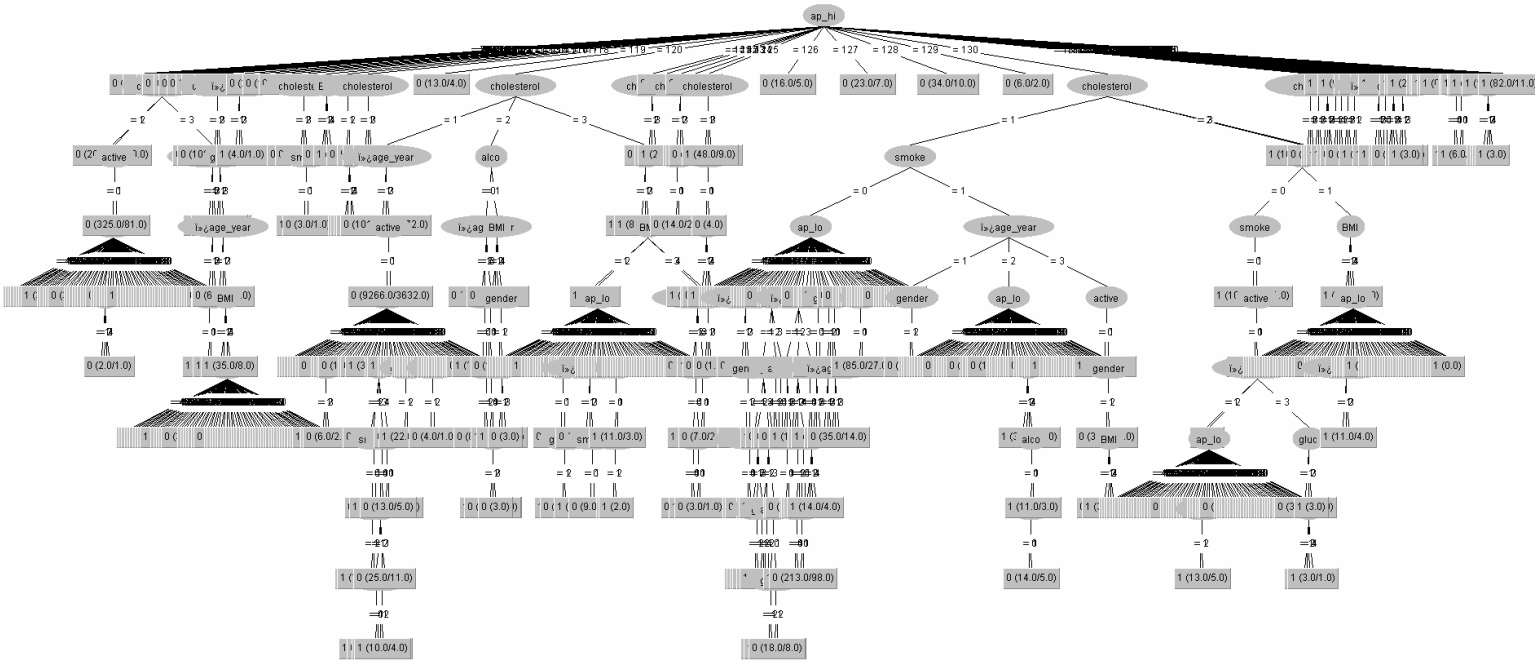


כאן, הקורולציה המקסימלית היא בין רמת הגלוקוז לכולסטרול. קורולציה זו זהה לקורולציה בדאטה 1, ובאופן כללי, הקורולציה בין המשתנים נמוכה יחסית.

תוצאות:

עבור דאטה 1:

להלן העץ שהתקבל:

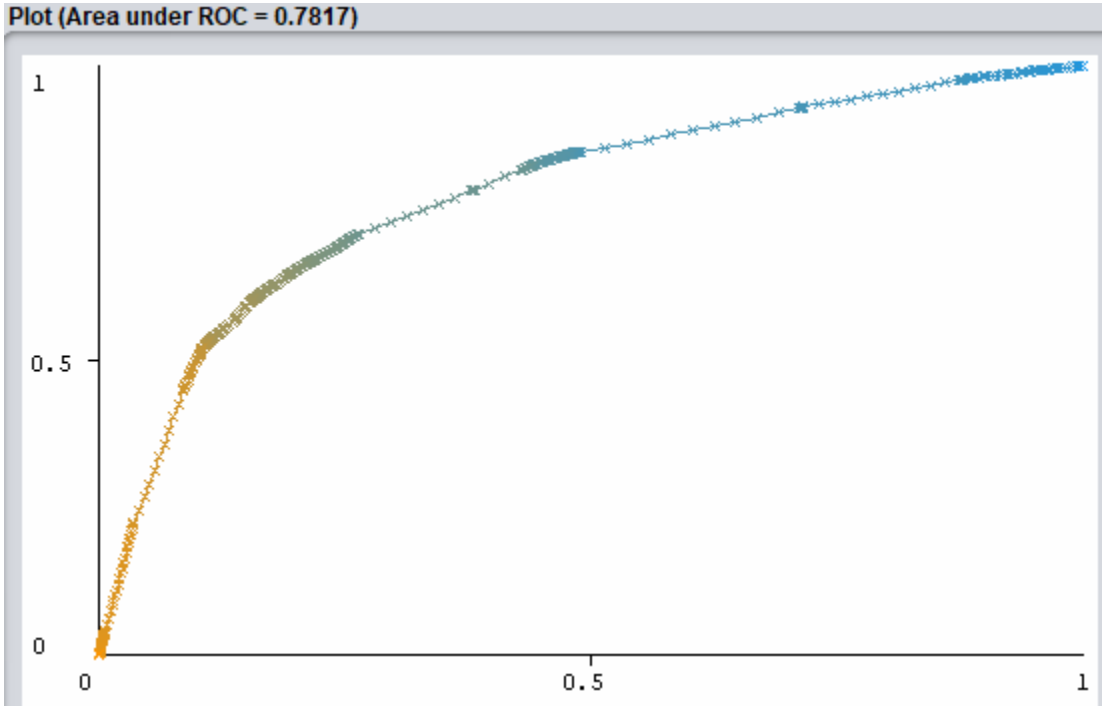


ניתן לראות שהמשתנה המסביר העיקרי שמקטין את אי הוודאות בצורה הטובה ביותר הוא לחץ הדם הסיסטולי (ap_hi) ממנו יוצאות המון הסתעפויות (משום שיש המון ערכים למשתנה). המשתנה המסביר השני אינו אחיד לכל ההסתעפויות: בחלקן יש קבלת החלטה, בחלקן עוברים למשתנה מסביר שני שהוא הכולסטרול (cholesterol), ובכמה מהן עוברים למשתנה הגיל.

ותוצאותיו של העץ הן:

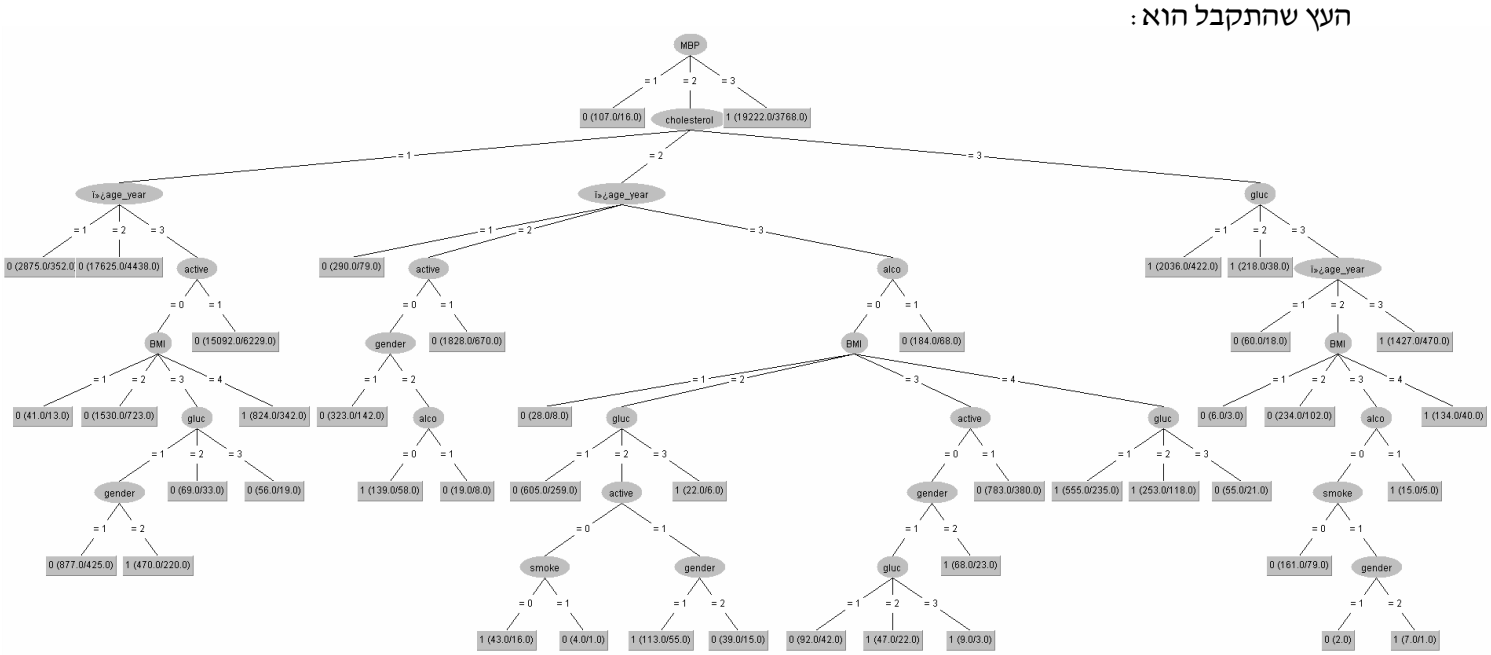
```
Correctly Classified Instances        49785               72.5866 %
Incorrectly Classified Instances      18802               27.4134 %
Kappa statistic                           0.4509
Mean absolute error                       0.3701
Root mean squared error                   0.4328
Relative absolute error                  74.0393 %
Root relative squared error              86.5623 %
Total Number of Instances             68587

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.790    0.340    0.704      0.790    0.745      0.454    0.782     0.748     0
                 0.660    0.210    0.755      0.660    0.704      0.454    0.782     0.762     1
Weighted Avg.    0.726    0.276    0.729      0.726    0.725      0.454    0.782     0.755

=== Confusion Matrix ===

    a     b   <-- classified as
27411  7273 |     a = 0
11529 22374 |     b = 1
```

גרף ROC:



עבור דאטה 2:

העץ שהתקבל הוא:

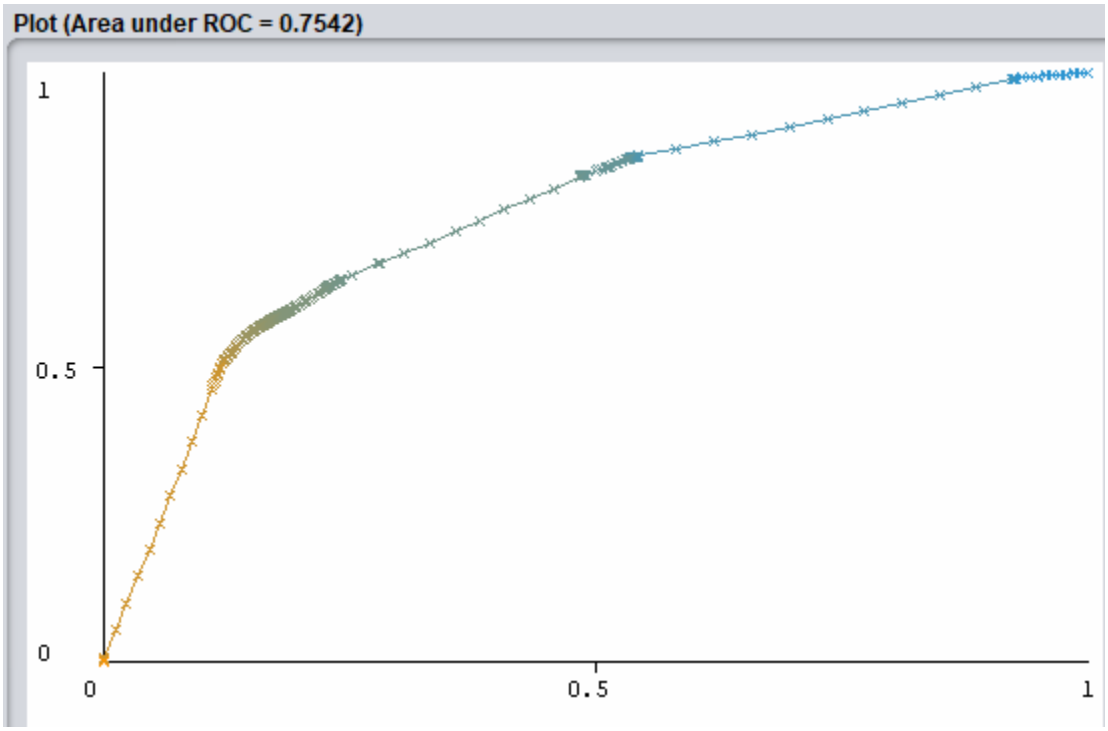


ניתן לראות שהמשתנה המסביר העיקרי שמקטין את אי הוודאות בצורה הטובה ביותר הוא לחץ הדם הממוצע (MBP) ממנו יוצאות שלוש הסתעפויות. המשתנה המסביר השני אחיד, והוא הכולסטרול (cholesterol).

ותוצאותיו של העץ הן:

```
Correctly Classified Instances        48362               70.5119 %
Incorrectly Classified Instances      20225               29.4881 %
Kappa statistic                           0.4086
Mean absolute error                       0.3919
Root mean squared error                   0.4431
Relative absolute error                  78.392  %
Root relative squared error              88.6345 %
Total Number of Instances             68587

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.823    0.416    0.669      0.823    0.739      0.420    0.754     0.727     0
                 0.584    0.177    0.764      0.584    0.662      0.420    0.754     0.728     1
Weighted Avg.    0.705    0.298    0.716      0.705    0.701      0.420    0.754     0.727

=== Confusion Matrix ===

    a     b   <-- classified as
28559  6125 |     a = 0
14100 19803 |     b = 1
```

גרף ROC:



ניתוח התוצאות ודיון:

מטרת המודל הייתה לחזות האם מטופל נמצא בסיכון לחלות במחלת לב תוך התחשבות בקריטריונים כמו BMI, לחץ דם, רמת כולסטרול בדם וכו'.

1. עצי ההחלטות: באופן כללי, עצי ההחלטות נותנים סיווג יחסית טוב, אך לא מיטבי. עבור דאטה 1, התקבל עץ החלטות מסועף ועמוס מאוד, שעלול היה להוביל לoverfitting, ולכן לא משקף בצורה מדוייקת את המצב האמיתי. לעומת זאת, העץ שהתקבל דרך דאטה 2 מתאר את המציאות בצורה מיטבית, בה התוצאות ממויינות לפי טווחים.
2. התוצאות: אחוז הדיוק בדאטה 1 הינו 72.6%, ובדאטה 2 הוא 70.5% - הבדל מזערי של בערך 2.1%. בנוסף, במטריצת הבלבול יש יתרונות וחסרונות לכל דאטה:
   - מספר המטופלים שאינם בסיכון למחלת לב, והמסווג סיווג אותם ככאלה, גבוה יותר בדאטה 2 מאשר בדאטה 1.
   - מספר המטופלים שאינם בסיכון למחלת לב, והמסווג סיווג אותם כמטופלים בסיכון, גבוה יותר בדאטה 1 מאשר בדאטה 2.
   - מספר המטופלים שנמצאים בסיכון למחלת לב, והמסווג סיווג אותם כמטופלים שאינם בסיכון, גבוה יותר בדאטה 2 מאשר בדאטה 1.
   - מספר המטופלים שנמצאים בסיכון למחלת לב, והמסווג סיווג אותם ככאלה, גבוה יותר בדאטה 1 מאשר בדאטה 2.

   מתוך תוצאות אלו, ניכר שהסיווג עבור דאטה 1 מוצלח וגבוה יותר. כלומר, הסיכוי שהמסווג יטעה, גבוה יותר דווקא בדאטה 2.
3. גרף ROC: כאמור, נרצה שהערך יהיה כמה שיותר קרוב ל-1. הערך הגבוה ביותר שהתקבל הוא 0.7817 עבור דאטה 1, והערך שהתקבל עבור דאטה 2 הוא 0.7542. גם כאן מתקבל הבדל מזערי - 0.0275.
4. Recall: נשווה בין התוצאות שהתקבלו עבור ערך זה בין שתי הדאטאות. דאטה 1, קיבלה את התוצאה 0.660 שהם 66%, לעומת דאטה 2 שקיבלה 0.584 שהם 58.4%. כאן מתקבל הבדל משמעותי יותר של כ-8%, שמצביע על כך שדאטה 1 סיווגה בצורה טובה יותר את כמות החולים שבסיכון מתוך אלו שבאמת בסיכון.

סיכום ומסקנות:

מסקנות שעלו מהפרוייקט:

- עבור עץ פשוט יותר לניתוח, קיבלנו תוצאות טובות פחות, אך הענפים בו תאמו בצורה טובה יותר את המציאות.
- בעץ מסועף יותר קיבלנו דיוק גבוה יותר. כלומר, פחות מידע נאבד.
- אם היה מתבצע עיבוד נתונים בדרך שונה, ייתכן שהתוצאות היו בעלות ערכים טובים יותר.
- סביר להניח שבמידה והמשתנים המסבירים היו שונים מאלו שנבחנו בפרוייקט, התוצאות היו טובות יותר.
- כמות גדולה יותר של משתנים, יכלה להוביל לתוצאות טובות יותר.
- ייתכן כי תוצאת ה Recall של דאטה 1 גבוהה יותר מבדאטה 2 בעקבות שינוי העמודות. כלומר בדאטה 2 המטופלים חולקו לטווחים לפי לחץ הדם, ובדאטה 1 השתמשנו בנתונים יותר מדוייקים לכל מטופל. לכן סביר לומר שהתקבל Recall יותר טוב בדאטה 1.
- בעולם האמיתי מדובר במאגרי מידע הרבה יותר גדולים, מאלו שעליהם התבסס המודל. לכן, כשהמודל של דאטה 1 מופעל על מאגרים אלו, התוצאות עלולות להגיע מאוחר מהצפוי בעקבות overfitting. אמנם, אחוזי הדיוק שלו מעט גבוהים יותר, אך בענייני בריאות המהירות היא פקטור מרכזי, כך שהמודל של דאטה 2 יותר אפקטיבי על אף שהוא קצת פחות מדוייק.

מחקר עתידי:

1. באופן כללי, פיתוח מודל לחיזוי מחלות לב מתבסס על גורמים רבים, רובם פיזיולוגים ומשתנים בין נבדק לנבדק. לכן, המודלים העתידיים שיפותחו, צריכים לכלול קבוצות וטווחים בהם גורמים אלו דומים.
2. פיתוח מודל נוסף שיצביע על הגורם העיקרי שבגללו אדם נמצא בסיכון לחלות במחלת לב, יתרום בשינוי אורח החיים בצורה המיטבית, ובהמנעות ממחלת לב.

הפניות:

1. https://www.kaggle.com/raminhashimzade/cardio-disease
2. https://hospitals.clalit.co.il/rabin/he/departments-and-clinics/medicine/Pages/low_blood_pressure.aspx
3. https://www.clalit.co.il/he/medical/medical_diagnosis/Pages/hypertension.aspx.
4. https://www.tasmc.org.il/Be-Well/calculators/Pages/Blood-Pressure.aspx

נספחים:

נספח א' – חלוקת טווחי מדד BMI