פרויקט ML חלק א



מגישים:

אור עזר 205781396

שוהם שבת 308160456

**תוכן ענייניים**

[**הגדרת הבעיה** 3](#_Toc39838243)

[1. תיאור כללי של עולם התוכן הנחקר 3](#_Toc39838244)

[2. הגדרת שאלת המחקר 3](#_Toc39838245)

[**הבנת הנתונים** 3](#_Toc39838246)

[1. תיעוד מקורות הנתונים ומשמעותם 3](#_Toc39838247)

[2. הסתברויות אפריוריות וקשרים בין מאפיינים 4](#_Toc39838248)

[3. איכות הנתונים 9](#_Toc39838249)

[**הכנת הנתונים** 10](#_Toc39838250)

[1. בחירת מאפיינים 10](#_Toc39838251)

[2. טיפול פרטני במאפיינים 10](#_Toc39838252)

[**נספחים** 11](#_Toc39838253)

[נספח 1: Y- heart attack- האם הנבדק חווה התקף לב 11](#_Toc39838254)

[נספח 2: X1- age- גיל הנבדק 11](#_Toc39838255)

[נספח 3: X2- gender- מין הנבדק 12](#_Toc39838256)

[נספח 4: X3- cp- סוג הכאבים בחזה שחווה הנבדק 12](#_Toc39838257)

[נספח 5: X4- trestbps- לחץ דם במנוחה 13](#_Toc39838258)

[נספח 6: X5- chol- רמת כולסטרול בדם 13](#_Toc39838259)

[נספח 7: X6- fbs- רמת סוכר בדם לאחר צום 14](#_Toc39838260)

[נספח 8: X7- restecg- תוצאות אק"ג בזמן מנוחה 14](#_Toc39838261)

[נספח 9: X8- thalach- דופק מקסימלי שנמדד 15](#_Toc39838262)

[נספח 10: X9- exang- האם קיימים כאבים בחזה בזמן פעילות גופנית 15](#_Toc39838263)

[נספח 11: X10- oldpeak- צניחות במקטע ST באק"ג בזמן מאמץ ביחס למנוחה 16](#_Toc39838264)

[נספח 12: X11- slope- שיפוע מקטע ST באק"ג בזמן פעילות השיא 16](#_Toc39838265)

[נספח 13: X12- ca- מספר כלי הדם הגדולים שנצבעו בבדיקת פלורוסקופיה 17](#_Toc39838266)

[נספח 14: X13- thal- רמת הנזק לרקמת הלב חזרה לערך המתאים 17](#_Toc39838267)

[נספח 15: מטריצת קורלציה 18](#_Toc39838268)

# **הגדרת הבעיה**

## תיאור כללי של עולם התוכן הנחקר

עולם התוכן הנחקר בפרויקט זה הוא זיהוי התקפי לב אצל נבדקים. אירוע לב הוא אירוע רפואי חמור שיכול לגרום לפגיעה קשה באורח החיים של המטופל. אלמנט מרכזי בטיפולים כיום הוא זיהוי מוקדם ככל הניתן של התקף הלב ותחילת הטיפול. לשם כך הצוותים המטפלים נדרשים לנתח את המידע שמציג המטופל- גורמי הסיכון הכרוניים לצד סימנים וסימפטומים מיידים- ולקבל החלטה האם מדובר בהתקף לב או לא.

מחקרים קודמים שעסקו בנושא ניסו גם כן לאמן מודל שיחזה האם המטופל חווה התקף לב או לא. הכלים המרכזיים בהם נעשה שימוש הם עצי החלטה (decision trees) ורשתות נוירונים (neural network).

## הגדרת שאלת המחקר

מטרת המודל היא חיזוי- האם מטופל מסוים חווה אירוע לב בעזרת יצירת פרופיל מטופל שעלול לחוות אירוע לב. באמצעות השימוש במערכות לומדות ננסה לסווג כל נבדק לאחת משתי קבוצות- אלו שאנו מעריכים כי הם חווים אירוע לב ואלו שלא.

# **הבנת הנתונים**

## תיעוד מקורות הנתונים ומשמעותם

**מקור בסיס הנתונים:**

מקור בסיס הנתונים נוצר ע"י ארבעה רופאים וחוקים ממרכזים רפואיים שונים בעולם[[1]](#footnote-1): המכון הקרדיולוגי ההונגרי, בית החולים האוניברסיטאי בציריך (שוויץ), בית החולים האוניברסיטאי בבאזל (שוויץ) ומרכז רפואי בלונג ביץ' וקליבלנד (ארה"ב). במקור בסיס הנתונים מכיל 76 מאפיינים אך מחקרים רבים שעשו שימוש בבסיס נתונים זה השתמשו באותם 14 מאפיינים שמשמשים גם אותנו בפרויקט זה.

**מאפייני סט הנתונים:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **מס' משתנה** | **שם משתנה** | **משמעות** | **סוג משתנה** | **טווח הערכים בסט הנתונים** |
| X1 | age | גיל הנבדק (שנים) | רציף | 34-77 |
| X2 | gender | מין הנבדק | קטגוריאלי | 0- אישה, 1- גבר |
| X3 | cp | סוג הכאבים בחזה שחווה המטופל | קטגוריאלי | 0- כאבים אנגינוטים 'קלאסיים'  1- כאבים אנגינוטים שאינם 'קלאסיים'  2- כאבים שאינם אנגינוטים  3- ללא כאבים |
| X4 | trestbps | לחץ דם במנוחה (mm/Hg) | רציף | 94-200 |
| X5 | chol | רמת כולסטרול בדם (mg/dl) | רציף | 126-409 |
| X6 | fbs | רמת סוכר בדם לאחר צום (mg/dl) | קטגוריאלי | 0- רמת סוכר של עד 120mg/dl (תקינה)  1- רמת סוכר מעל 120 mg/dl (לא תקינה) |
| X7 | restecg | תוצאות אק"ג במנוחה | קטגוריאלי | 0- נורמלי  1- אק"ג המציג היפוכי גל T ו/או עליות/צניחות של מקטע ST של יותר מ- 0.05 mV.  2- אק"ג המציג היפרטרופיה של חדר שמאל (לפי Este's criteria) |
| X8 | thalach | דופק מקסימלי שנמדד | רציף | 88-195 |
| X9 | exang | האם קיימים כאבים בחזה בזמן פעילות גופנית | קטגוריאלי | 0- אין כאבים, 1 - יש כאבים |
| X10 | oldpeak | צניחות במקטע ST באק"ג בזמן מאמץ ביחס למנוחה | רציף | 0-5.6 |
| X11 | slope | שיפוע מקטע ST באק"ג בזמן פעילות השיא | קטגוריאלי | 0- עולה (upsloping)  1- שטוח (flat)  2- יורד (downsloping) |
| X12 | ca | מספר כלי הדם הגדולים שנצבעו בבדיקת פלורוסקופיה | בדיד | 0-3- מספר כלי הדם הגדולים שנצבעו  4- לא ידוע |
| X13 | thal | רמת הנזק לרקמת הלב | קטגוריאלי | 0- ערך חסר ,1 - נזק בלתי הפיך  2- נורמלי, 3- נזק הפיך |
| Y | heart attack | האם המטופל חווה אירוע לב | קטגוריאלי | 0- לא  1- כן |

## הסתברויות אפריוריות וקשרים בין מאפיינים

**הסתברויות אפריוריות ואיזון סט הנתונים:**

נציג הסתברויות אפריוריות של משתנה המטרה וכלל המאפיינים לפי סט האימון. עבור משתנים קטגוריאליים נציג את ההסתברויות באחוזים בעזרת גרף עמודות ועבור משתנים רציפים נציג היסטוגרמה.

**Y- heart attack- האם הנבדק חווה התקף לב:** **[(נספח 1)](#_נספח_1:_Y-)**

ניתן להסיק מהגרף כי משתנה זה איננו מאוזן. בבסיס הנתונים שלנו ההסתברות האפריורית לאירוע לב היא 53.3%. אולם, לפי מחקרים של המרכז לבקרת מחלות ומניעתן האמריקאי (CDC) רק כ- 0.25% מאוכלוסיית ארה"ב חווים התקף לב בשנה (805,000 מתוך כ- 328 מיליון אזרחים)[[2]](#footnote-2). גם אם נצמצם את גודל האוכלוסייה לגילאי 30 ומעלה אחוז זה לא עולה על 0.5%. יתר על כן, לפי משרד הבריאות האוסטרלי סיכוייו של אדם שנמצא בדרגת הסיכון הגבוהה ביותר לחוות התקף לב בחמש השנים הקרובות (בעקבות גורמי סיכון שחלקם נמדדים בסט נתונים זה, כמו: לחץ דם גבוה, סכרת, גיל ומין) הם כ-14% בלבד[[3]](#footnote-3).

**X1- age- גיל הנבדק:**[**(נספח 2)**](#_נספח_2:_X1-)

ראשית נשים לב כי ישנם מספר ערכים לא הגיוניים בנתונים. ביצענו תיקון לערכים אלו כפי שיוסבר בהמשך. הגרף המוצג הינו התפלגות הגילאים לאחר התיקון. על סמך מחקרים של ה- American Heart Association (AHA) אנו יודעים כי מרבית האנשים שחווים אירוע לבבי הינם בגילאי 40 עד 60 ונדיר לצפות בהתרחשות אירוע לבבי עבור מטופלים צעירים. כפי שניתן לראות בגרף מרבית הנתונים מתרכזים בטווח גילאים זה ולכן משתנה זה מייצג את המציאות.

**X2- gender- מין הנבדק:**[**(נספח 3)**](#_נספח_3:_)

מאותם מחקרים של ה- AHA אנו יודעים כי לגברים יש נטייה גדולה יותר לחוות אירוע לבבי מאשר נשים. לכן, הגיוני כי מרבית הנבדקים בסט הנתונים שלנו יהיו גברים- כפי שניתן לראות בגרף (67.5% גברים).

**X3- cp- סוג הכאבים בחזה שחווה הנבדק:** **[(נספח 4)](#_נספח_4:_X3-)**

ניתן לראות כי כמעט מחצית מהנבדקים חוו כאב שמוגדר ככאב אנגינוטי קלאסי- כלומר כאב ממקור לבבי שתואם את התיאורים בספרות המקצועית בנושא. דבר זה תואם את המציאות שכן- התיאור הקלאסי של הכאבים נבנה על סמך זאת שזהו סוג הכאבים שחוו מרבית המטופלים.

**X4- trestbps- לחץ דם במנוחה:** **[(נספח 5)](#_נספח_5:_X4-)**

משתנה זה מציג את לחץ הדם (סיסטולי) שנמדד למטופל בזמן מנוחה. אנו יודעים כי רמות תקינות של ערך זה הם 90-140 mm/Hg ונראה סביר כי שם יימצאו מרבית הנתונים- כפי שניתן לראות בגרף. נשים לב כי ישנם מעט ערכים קיצוניים (מעל 180 mm/Hg) ונשקול בהמשך להסירם/לתקנם.

**X5- chol- רמת כולסטרול בדם:** **[(נספח 6)](#_נספח_6:_X5-)**

מהגרף ניתן לראות כי מרבית הנתונים נמצאים מעבר לערך של 200 mg/dl-הגבול העליון התקין של בדיקה זו. נראה כי התפלגות זו איננה תואמת לגמרי את המציאות, שכן לפי ה- CDC האמריקאי לכ-30% מהאזרחים האמריקאים מעל גיל 20 יש רמת כולסטרול גבוהה בעוד בסט הנתונים שלנו זהו איננו הדבר.

**X6- fbs- רמת סוכר בדם לאחר צום:** **[(נספח 7)](#_נספח_7:_X6-)**

מהנתונים שלנו נראה כי לרוב המוחלט של הנבדקים נמדדה רמת סוכר נמוכה/שווה ל- 120 mg/dl- תוצאה הנחשבת לתקינה. תוצאה זו הולמת את המציאות שכן רק במקרים חריגים (כגון סכרת) אנו מצפים לראות תוצאה מעבר ל- 120 mg/dl.

**X7- restecg- תוצאות אק"ג בזמן מנוחה:** **[(נספח 8)](#_נספח_8:_)**

תוצאות אק"ג במנוחה הם מאפיין חשוב באבחנה של התקף לב ומהווים השוואה לתוצאות האק"ג בזמן ההתקף. מהנתונים עולה כי כמחצית מהנבדקים מציגים שינויים במקטע ST במנוחה וקצת פחות ממחצית מציגים אק"ג תקין. מעט מאוד של נבדקים מציגים הגדלה לא תקינה (היפרטרופיה) של חדר שמאל בלב.

**X8- thalach- דופק מקסימלי שנמדד:** **[(נספח 9)](#_נספח_9:_)**

ה- AHA פיתח נוסחה כללית לפיה דופק מקסימלי בדקה של אדם הוא בסביבות 220 פחות גיל המטופל[[4]](#footnote-4). מרבית המטופלים בסט הנתונים שלנו נמצאים בגילאים 45-65 ולכן הגיוני לראות כי הדופק המקסימלי של מרבית המטופלים נע סביב 140-175 פעימות בדקה.

**X9- exang- האם קיימים כאבים בחזה בזמן פעילות גופנית:** **[(נספח 10)](#_נספח_10:_X9-)**

הופעה של כאבים בחזה בזמן פעילות גופנית יכול לסייע לאבחנה של התקף לב. הופעת הכאבים יכולה לנבוע ממצב מצוקה בו נמצא הלב ולכן יכול להעיד כי ישנה חסימה חלקית של כלי דם בלב. בנתונים שלנו, נראה כי כשני שליש מהנבדקים לא חווים כאבים בחזה בזמן מאמץ. נתון זה תואם את תוצאות מאפיין X13- thal שמראה כי למרבית המטופלים אין לנזק לרקמת שריר הלב.

**X10- oldpeak- צניחות במקטע ST באק"ג בזמן מאמץ ביחס למנוחה:**[**(נספח 11)**](#_נספח_11:_X10-)

מהגרף ניתן לראות כי למרבית המטופלים אין שינוי במקטע ST בזמן מאמץ לעומת מנוחה. אנו יודעים כי צניחות ST מתרחשות אצל אדם בריא במאמץ קיצוני או אצל אדם עם כשל לבבי קבוע. מכיוון שאלו הם מצבים מועטים הגיוני לראות כי למרבית הנבדקים אין שינוי במקטע ST בזמן מאמץ ביחס למנוחה.

**X11- slope- שיפוע מקטע ST באק"ג בזמן פעילות השיא:** **[(נספח 12)](#_נספח_12:_X11-)**

שיפוע יורד (downsloping) במקטע ST בזמן מאמץ גופני מעיד על חוסר חמצן ומצוקה של הלב בזמן הפעילות. לפי הנתונים, חלק קטן מאוד מהנבדקים הראה סימן זה, בעוד מרביתם הציגו ערך תקין- שיפוע ישר (flat) או עולה (upsloping). תוצאות אלו תואמות את המציאות שכן הן עולות בקנה אחד עם תוצאות מאפיין oldpeak.

**X12- ca- מספר כלי הדם הגדולים שנצבעו בבדיקת פלורוסקופיה:**[**(נספח 13)**](#_נספח_13:_)

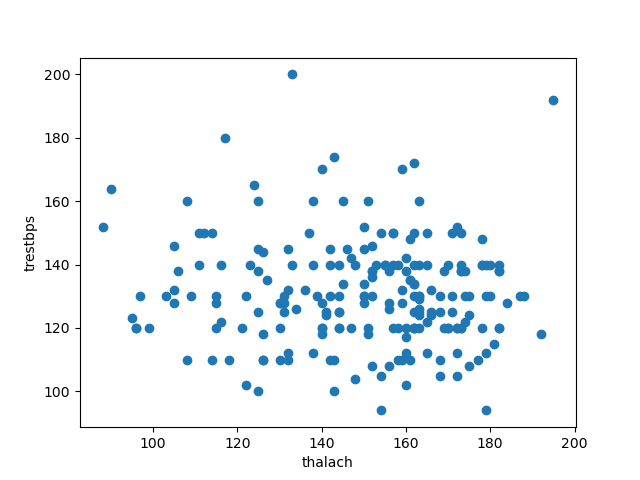
בבדיקה זו, כל כלי דם שנצבע משמעותו שכלי הדם איננו חסום (כך למשל ערך 0 מעיד כי כל כלי הדם חסומים). ניתן לראות כי למטופלים רבים יש לפחות כלי דם אחד חסום- 93.9% מהנבדקים קיבלו ערך 0-2 בבדיקה זו. לכן הגיוני לצפות כי עבור נבדקים אלו נראה כי הם אכן חוו התקף לב וערך משתנה המטרה הוא 1. בנוסף, נשים לב כי ב- 1.9% מהרשומות יש ערך לא ידוע שנטפל בו בהמשך (ערך 4).

**X13- thal- רמת הנזק לרקמת הלב: [(נספח 14)](#_נספח_14:_X13-)**

מאפיין זה בודק את רמת הנזק שנגרמה לרקמת שריר הלב. מהנתונים עולה כי למרבית המטופלים אין נזק כלל (normal) או ישנו נזק הפיך- כלומר הרקמה תחלים לאחר טיפול מתאים. בנוסף, נשים לב כי ב- 0.9% מהרשומות יש ערך חסר שיטופל בהמשך (ערך 0).

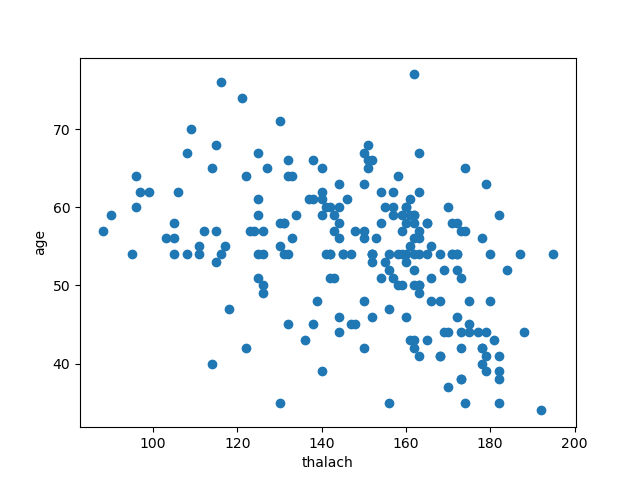
**קשרים בין משתנים רציפים :**

על מנת לנתח קשרים בין המשתנים המסבירים הרציפים לבין עצמם ובין המשתנים המסבירים למשתנה המוסבר נשתמש במטריצת קורלציה (לפי פירסון). [**(נספח 15)**](#_נספח_15:_מטריצת)

כעת נציג מספר קשרים בין משתנים שלדעתנו מרכזיים במפת החום:

ניתן לראות ממפת החום שישנה קורלציה שלילית יחסית נמוכה (0.074-) בין המשתנים trertbps, thalach כלומר הם אינם מוסברים אחד על ידי השני. נחזק עובדה זו על ידי הצגת הגרף בין שני המשתנים:

בעזרת גרף זה אנו יכולים לראות שאין מגמה בין שני הנתונים ושלא נוכל לקבל מידע ממשתנה אחד לגבי המשתנה השני.

****

קשר נוסף ממפת החום שניתן לראות הוא קורלציה שלילית (0.4-) בין המשתנים age, thalach כלומר הם יכולים להיות מוסברים אחד על ידי השני בקשר הפוך. נחזק עובדה זו על ידי הצגת הגרף בין שני המשתנים:

בעזרת גרף זה אנו יכולים לראות שישנה מגמה הפוכה יחסית קלה כלומר, כאשר הגיל יורד ה thalach עולה וההפך.

**קשרים בין משתנים קטגוריאליים :**

על מנת לבדוק קשר בין המשתנים הקטגוריאליים לבין עצמם, נבצע מבחני חי-בריבוע:

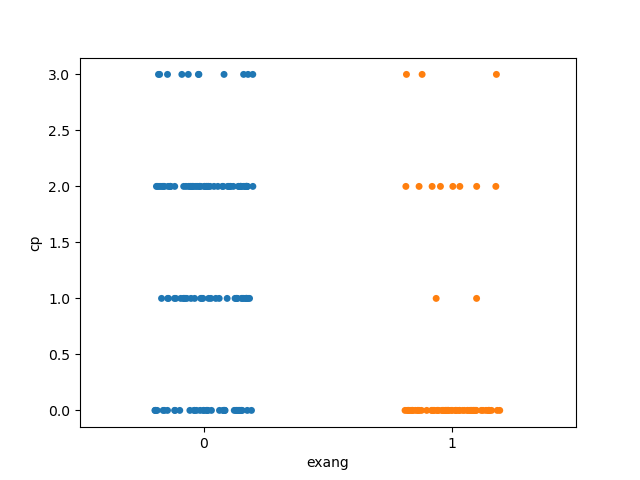
נניח B,A משתנים קטגוריאליים, ההשערות הן:

H0: A and B are independent  
H1: A and B are not independent

לקחנו את המשתנים הקטגוריאליים שיצאו מובהקים ושיש להם הסבר לוגי שיכול להסביר את התופעה:

**המשתנים: cp , exang:**

תוצאה:

ניתן לראות שכל ערכי p-value יצאו מובהקים מאוד (קטנים מאלפא = 0.05) ולכן ניתן לומר שיש מספיק נתונים לדחות את השערת ה-0 ולומר שהמשתנים הקטגוריאליים cp , exang תלויים זה בזה כלומר יש ביניהם קשר מובהק. תופעה זו יכולה להיות מוסברת בכך שכאב טיפוסי יכול להצביע על כאב איסכמי ולכן תהיה תעוקה בזמן מאמץ.

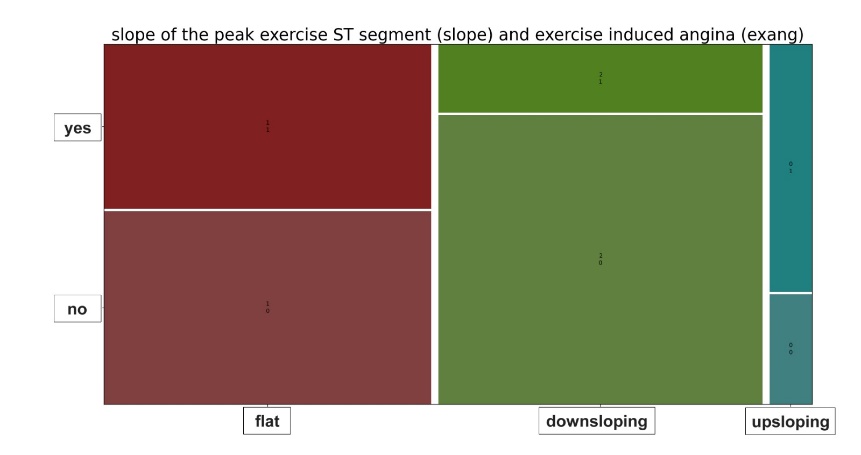
גרף בין שני המשתנים :

ניתן לראות ש EXANG = 1 אז סיכוי גבוה ש CP = 0

וכאשר EXANG = 0 אז בסיכוי גבוה ש CP = 2 or 1

תואם את ההסבר הלוגי לתופעה.

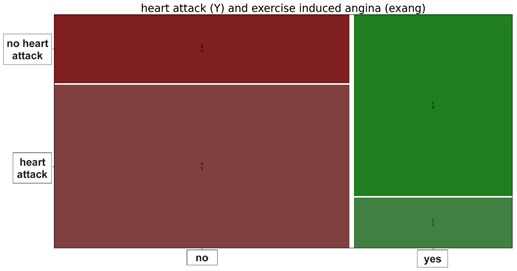
**המשתנים: slope , exang:**

תוצאה:   
ניתן לראות שכל ערכי p-value יצאו מובהקים מאוד (קטנים מאלפא = 0.05) ולכן ניתן לומר שיש מספיק נתונים לדחות את השערת ה-0 ולומר שהמשתנים הקטגוריאליים slope , exang תלויים זה בזה כלומר יש ביניהם קשר מובהק. תופעה זו יכולה להיות מוסברת בכך slope הינו ביטוי לשיפוע מקטע ST באק"ג לתעוקת לב שזה יכול להסביר תעוקה בזמן פעילות גופנית exang .

גרף בין שני המשתנים :

ניתן לראות שכשאר כאשר Slope = 0 אז בסיכוי גבוה EXANG = 1 ,כאשר Slope = 1 אז סיכויים שווים יחסית ל EXANG = 1 or 0 , כאשר Slope = 2 אז בסיכוי גבוה ש EXANG = 0 .

תואם את ההסבר הלוגי לתופעה.

**קשרים בין מאפיינים למשתנה המטרה:**

**קשר בין exang למשתנה המטרה:**

לדעתנו, מאפיין exang בעל השפעה על משתנה המטרה.

וראינו שיש לו קשר לשאר המשתנים.

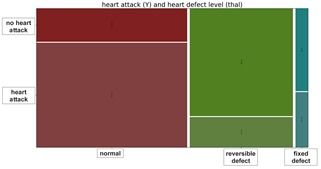
הסבר לוגי לתופעה הוא שכאבים בחזה בזמן פעילות גופנית יכול לסייע לאבחנה של התקף לב

גרף בין שני המשתנים :

ניתן לראות שכשאר כאשר y = 1 אז בסיכוי גבוה EXANG = 0 וכאשר y = 0 אז בסיכוי גבוה EXANG = 1

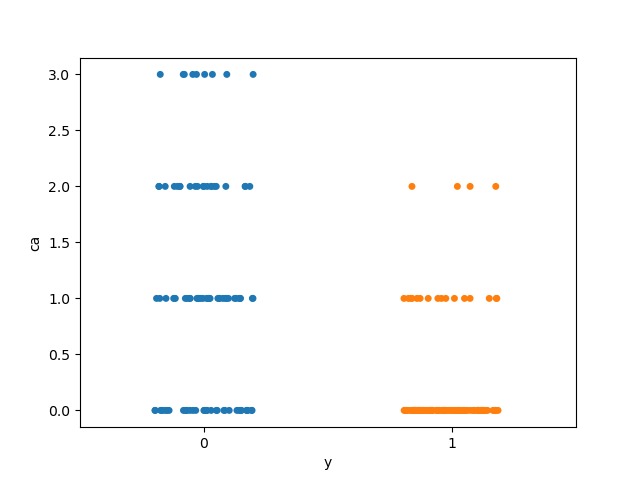
דבר הסותר את ההסבר הלוגי לתופעה (בדיוק הפוך).

**קשר בין thal למשתנה המטרה:**

לדעתנו, מאפיין thal, המציין את רמת הפגיעה בלב, בעל השפעה על משתנה המטרה. אנו רואים בגרף כי ערכים שונים של המאפיין משפיעים באופן שונה על ההסתברות לקיומו של התקף לב. למשל- עבור thal=2 (normal) ניתן לראות כי מרבית הערכים של משתנה המטרה הם y=1 (heart attack) ואילו עבור thal=3 (reversible defect) ניתן לראות כי מרבית הערכים של משתנה המטרה הם y=0 (no heart attack). עבור thal=1 (fixed defect) החלוקה איננה ברורה כמו שתי הקטגוריות האחרות- אולם אלו מיעוט של תצפיות (4.7%).

* הגרף המוצג הוא הגרף לאחר שינוי הערכים החסרים

**קשר בין ca למשתנה המטרה:**

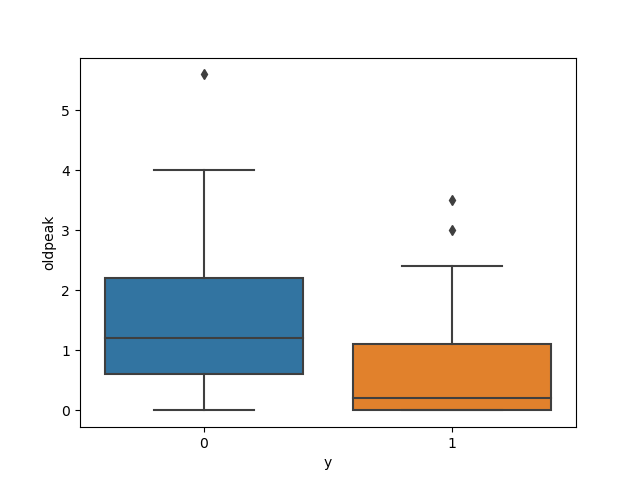
משתנה ca גם הוא יכול להיות חשוב ומשמעותי למודל. הסבר לוגי לכך - ca הינו מספר כלי דם גדולים שנצבעים בבדיקת פלואורסקופיה, וככל שמספר כלי הדם גדול יותר הבן אדם יותר בריא.

מהגרף ניתן לראות שככל שca - עולה כך בסיכוי יותר גבוה ש Y= 0 וככל שca יורד- הסיכוי ש Y = 1 גדל.

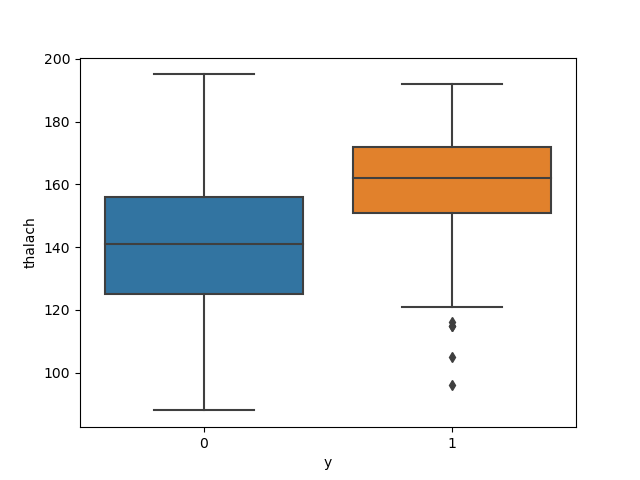
בצורה טובה יותר. מקרה זה תומך בהסבר הלוגי שהצגנו מקודם.

* הגרף המוצג הוא הגרף לאחר שינוי הערכים החסרים

**קשר בין oldpeak למשתנה המטרה:**

ניתן לראות במפת החום קורלציה שלילית גדולה יחסית ( 0.42- ) בין המשתנה oldpeak לבין משתנה המטרה. משתנה זה עשוי להיות משמעותי וחשוב למודל.

ניתן לראות מהגרף שכאשר ה oldpeakנמוך יותר (באזור ה- 0) הוא יותר מקושר לY שווה 1, והאחוזון ה25 ועד ה75 נעים בין 0-1, וכאשר ה oldpeak גבוה יותר הוא יותר מקושר לY ששווה ל-0. והאחוזון ה25 ועד ה75 נעים בין 0.6 – 2.2.

**קשר בין thalach למשתנה המטרה:**

ניתן לראות במפת החום קורלציה חיובית גדולה יחסית (0.44) בין המשתנה thalach לבין משתנה המטרה. משתנה זה עשוי להיות משמעותי וחשוב למודל.

ניתן לראות על פי הגרף שכאשר ה thalach גבוהה יותר (באזור ה- 160 ) הוא יותר מקושר לY שווה 1, החציון נמצא 160 והאחוזון ה25 ועד ה75 נעים בין 150-170, וכאשר ה thalach נמוך יותר (באזור ה-140) הוא יותר מקושר לY ששווה ל-0. החציון נמצא ב- 140 והאחוזון ה25 ועד ה75 נעים בין 110-155.

## איכות הנתונים

**נתונים חסרים:**

בסט הנתונים יש מעט נתונים חסרים- עבור מאפיין ca ישנם 1.9% ערכים חסרים (4 רשומות) ועבור מאפיין thal יש 0.9% ערכים חסרים (2 רשומות). 2 מאפיינים אלו הם משתנים קטגוריאליים ולכן לא נוכל להזין את הערך הממוצע. במקום זאת נחליף את הערכים החסרים לערך השכיח של אותו משתנה. מכיוון שמדובר במספר קטן של רשומות אנו מאמינים כי שינוי זה לא יביא ל- overfitting של המודל.

עבור ca- נחליף את הערכים החסרים ל-0, ועבור thal- נמלא את הערכים החסרים בערך 2.

**נתונים שאינם הגיוניים:**

בבסיס הנתונים ישנן 25 רשומות (כ-12% מהרשומות) בהן בגיל המטופל נרשמו ערכים לא הגיוניים (הרבה מעבר ל-120). בכדי לטפל בשגיאות אלו החלטנו לחשב את ממוצע וסטיית התקן הגילאים לפי הרשומות התקינות בלבד- ולאחר מכן להגריל לרשומות הלא תקינות ערך מהתפלגות נורמלית לפי ערכים אלו (ניתן להניח התפלגות נורמלית מכיוון שמדובר במספר גדול של תצפיות).

# **הכנת הנתונים**

## בחירת מאפיינים

החלטנו להסיר מהמודל את המאפיין fbs מכיוון שלדעתנו אינו מוסר מידע רב על משתנה המטרה לפי הידע שרכשנו עד כה. מההסתברויות האפריוריות אנו יודעים כי למרבית הנבדקים ישנה רמת סוכר תקינה (85.8%). מהגרף המצורף ניתן לראות כי המשתנה לא נותן הפרדה טובה בין הערכים של משתנה המטרה Y מכיוון שקווי ההפרדה בין ערכי Y השונים מאוד קרובים אחד לשני. בנוסף, ביצענו מבחן חי בריבוע לבדיקת אי תלות בין המשתנים והתקבל כי המשתנים אינם תלויים זה בזה (Pvalue=0.846).

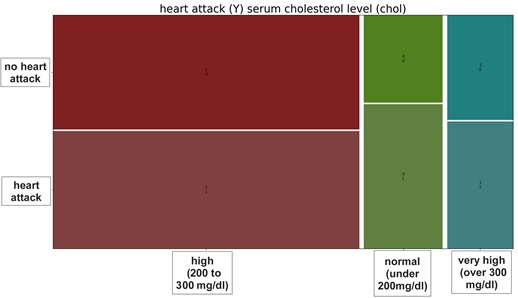
## טיפול פרטני במאפיינים

לאחר הסתכלות על הנתונים וניסיון להבינם ראינו שאנו לא מצליחים להסיק מידע על התנהגות המשתנה לעומת משתנה המטרה.

ניסינו להסתכל על הגרף בין שני המשתנים אך בגרף זה רוב הנקודות מפוזרות יחסית באופן שוויוני דבר היוצר קושי בהבנת הנתונים לדוגמא אם ניקח ערך ספציפי של המשתנה chol יהיה לנו קשה להגיד אם הוא מתקשר לרוב לY = 0 או לY = 1 דבר היכול ליצור בעיתיות בהבנת המודל .

לכן לאחר חשיבה החלטנו לנסות להפוך את משתנה chol לקטגוריאלי על פי הקטגוריות הבאות:

* 1= normal- ערכים של עד 200 mg/dl כולל
* 2= high- ערכים שבין 200 mg/dl ל- 300 mg/dl כולל
* 3= very high- ערכים מעל 300 mg/dl



לאחר המעבר למשתנה קטגוריאלי בדקנו אם כעת נוכל להסיק יותר מידע על התנהגות המשתנה לעומת משתנה המטרה.

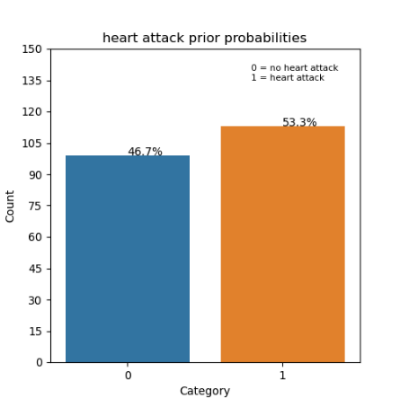
ניסינו להסתכל על הגרף בין שני המשתנים אך בגרף זה רוב השטחים (הפלגות הנקודות) מסודרים באופן שוויוני ויהיה לנו קשה להסיק מידע על משתנה המטרה .

לכן החלטנו להשמיט משתנה זה כי גם לאחר השינוי קשה להסיק מידע על משתנה המטרה עם הידע שיש לנו כעת.

# **נספחים**

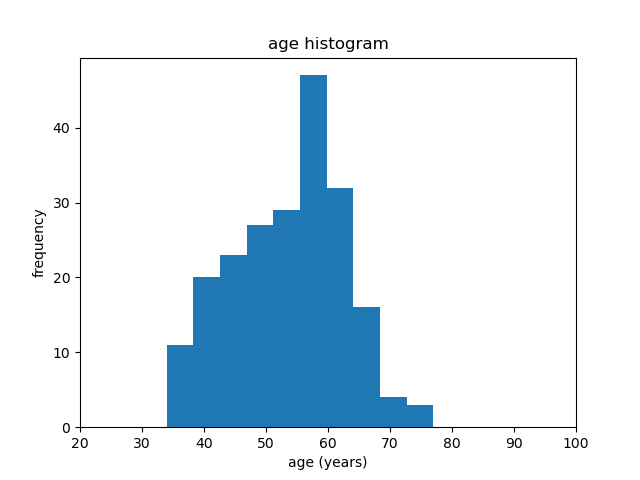
## נספח 1: Y- heart attack- האם הנבדק חווה התקף לב

[חזרה לערך המתאים](#_(נספח_1))



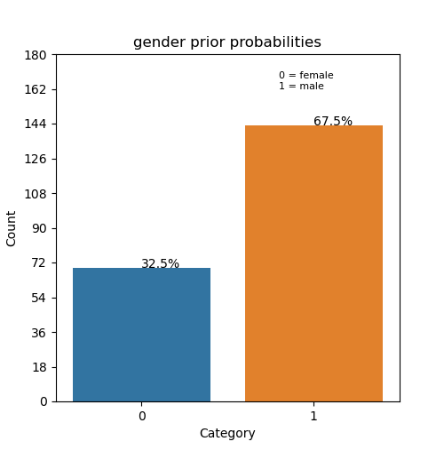
## נספח 2: X1- age- גיל הנבדק

[חזרה לערך המתאים](#_(נספח_2))

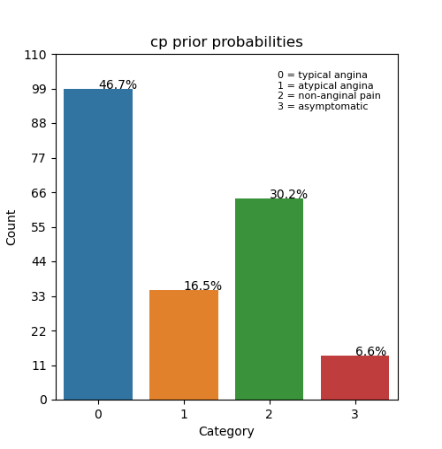


## נספח 3: X2- gender- מין הנבדק

[חזרה לערך המתאים](#_(נספח_3))

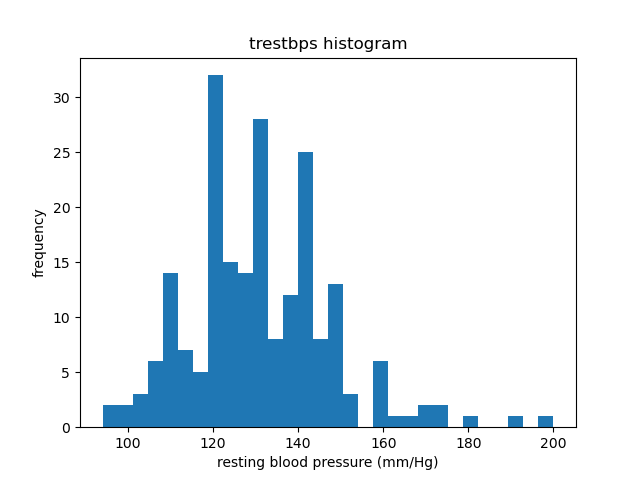


## נספח 4: X3- cp- סוג הכאבים בחזה שחווה הנבדק

[חזרה לערך המתאים](#_(נספח_4))

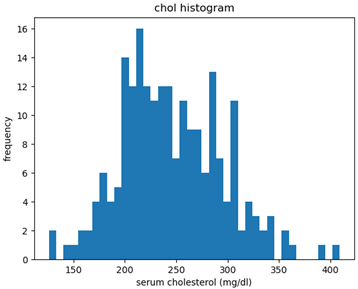
## 

## נספח 5: X4- trestbps- לחץ דם במנוחה

[חזרה לערך המתאים](#_(נספח_5))

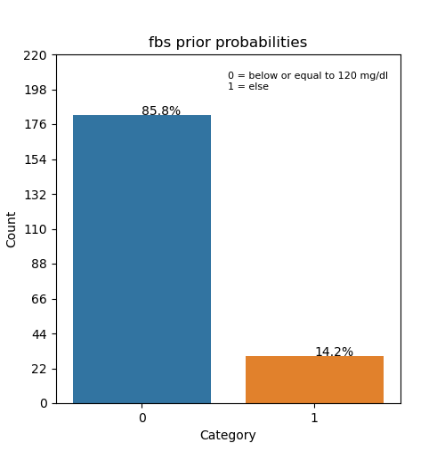
## נספח 6: X5- chol- רמת כולסטרול בדם

[חזרה לערך המתאים](#_(נספח_6))

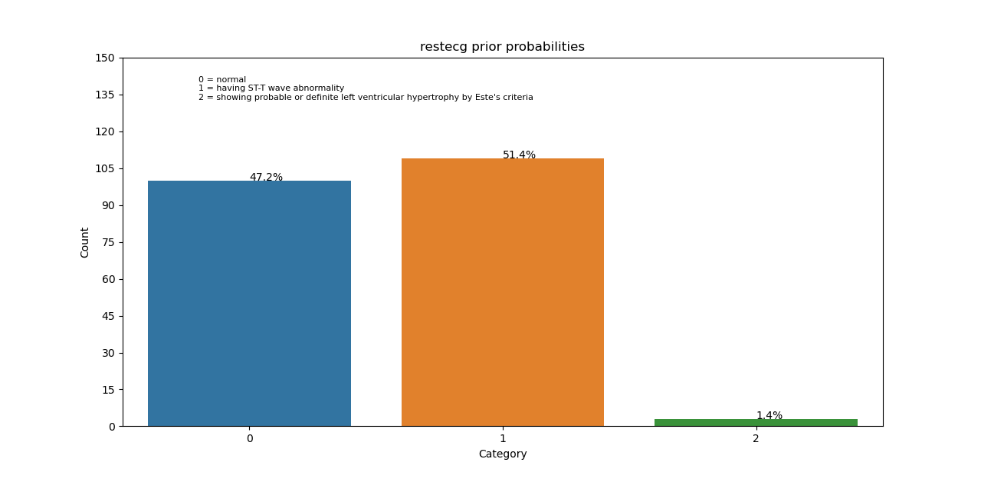


## נספח 7: X6- fbs- רמת סוכר בדם לאחר צום

[חזרה לערך המתאים](#_(נספח_7))



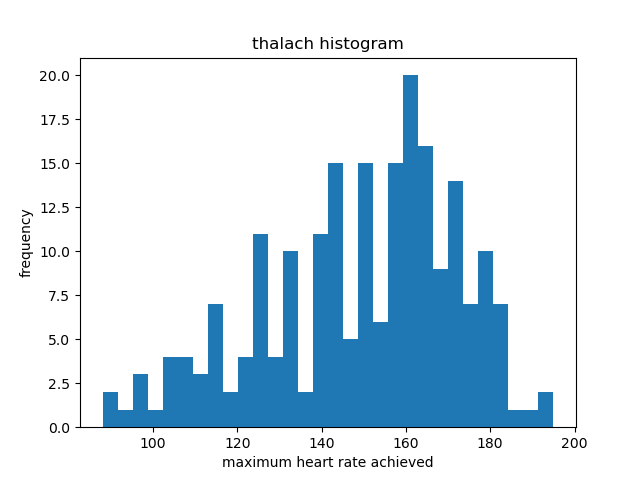
## נספח 8: X7- restecg- תוצאות אק"ג בזמן מנוחה

[חזרה לערך המתאים](#_(נספח_8))

## 

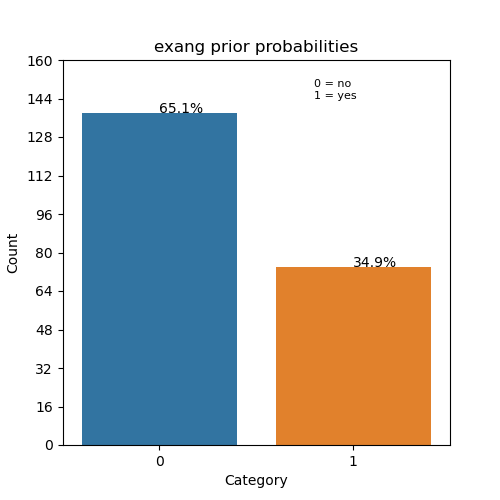
## נספח 9: X8- thalach- דופק מקסימלי שנמדד

[חזרה לערך המתאים](#_(נספח_9))

[](#_(נספח_9))

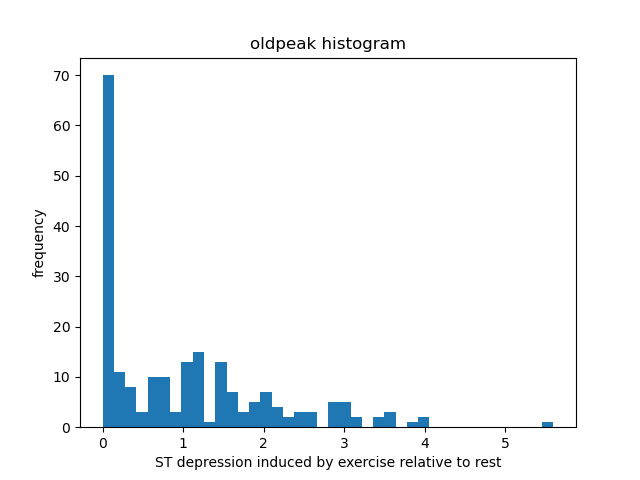
## נספח 10: X9- exang- האם קיימים כאבים בחזה בזמן פעילות גופנית

[חזרה לערך המתאים](#_(נספח_10))

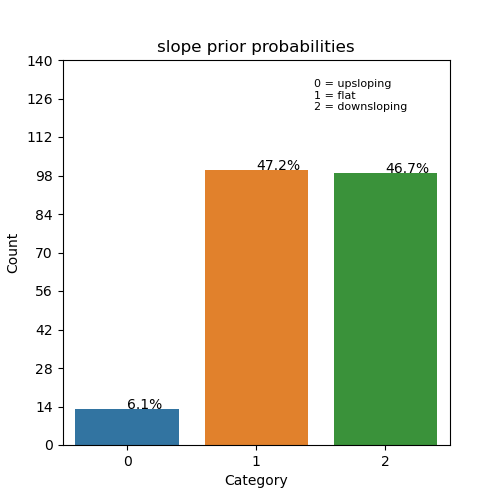
[](#_(נספח_10))

## נספח 11: X10- oldpeak- צניחות במקטע ST באק"ג בזמן מאמץ ביחס למנוחה

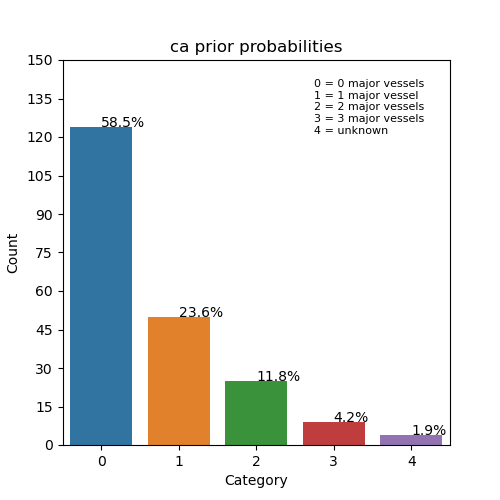
[חזרה לערך המתאים](#_(נספח_11))



## נספח 12: X11- slope- שיפוע מקטע ST באק"ג בזמן פעילות השיא

[חזרה לערך המתאים](#_(נספח_12))

## נספח 13: X12- ca- מספר כלי הדם הגדולים שנצבעו בבדיקת פלורוסקופיה

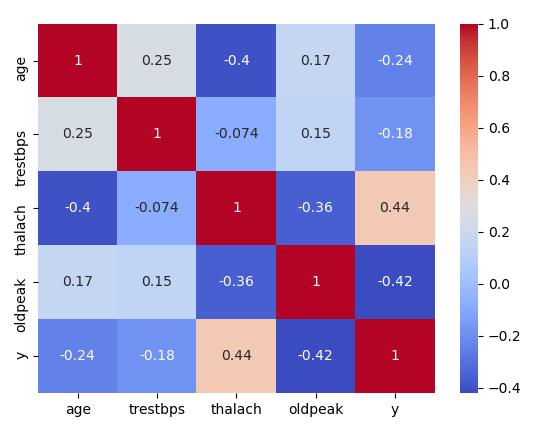
[חזרה לערך המתאים](#_X12-_ca-_מספר)

## 

## נספח 14: X13- thal- רמת הנזק לרקמת הלב [חזרה לערך המתאים](#_X13-_thal-_רמת)

## נספח 15: מטריצת קורלציה

[חזרה לערך המתאים](#_(נספח_15))



1. קישור להסבר על בסיס הנתונים- <http://archive.ics.uci.edu/ml/datasets/Heart+Disease> [↑](#footnote-ref-1)
2. <https://www.cdc.gov/heartdisease/facts.htm> [↑](#footnote-ref-2)
3. <https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/heart-disease-and-stroke-your-risk-score> [↑](#footnote-ref-3)
4. <https://www.heart.org/en/healthy-living/fitness/fitness-basics/target-heart-rates> [↑](#footnote-ref-4)