

**לימוד מכונה**

364-1-1811

**חלק א'**

מגישים:

בוגוסלבסקי דן – 316450253

דולב אמיר – 032670838

מרצה הקורס:

ד"ר ניר ניסים

עוזר הוראה:

בן חדד

תוכן עניינים:

**הגדרת הבעיה**

תיאור עולם התוכן הנחקר.................................................................................................3

הגדרת שאלת המחקר......................................................................................................3

**הבנת הנתונים**

מקור הנתונים ומשמעותם..................................................................................................3

הסתברות אפריורית וקשרים עם משתנה המטרה..................................................................4

בחינת סט האימון מול סט הבדיקה......................................................................................5

קשרים בין משתנים..........................................................................................................7

איכות הנתונים – מחיקה והשלמת תצפיות............................................................................8

**הכנת הנתונים**

השמטת מאפיינים............................................................................................................9

טיפול פרטני במאפיינים.....................................................................................................9

**הגדרת הבעיה:**

1. תיאור עולם התוכן הנחקר: הבעיה עליה אנו עובדים היא מציאת דרך לחזות מקרה בו אדם חווה התקף לב, על בסיס פיצ'רים שונים הניתנים למדידה. קיימים מחקרים רבים בנושא וישנם אלגוריתמים שפותחו, אשר מצליחים לחזות מצב שכזה ברמת דיוק גבוהה (כ-90 אחוז).
2. הגדרת שאלת המחקר: אנו מנסים לאמן מודל אשר יצליח לחזות ברמת דיוק טובה מצב של אדם החווה התקף לב. על ידי שימוש בכלים ושיטות של מערכות לומדות אנו יכולים לוודא את איכות ומהימנות הנתונים, לעבדם ולאמן בעזרתם את המודל כך שייתן את החיזוי הטוב ביותר.

**הבנת הנתונים:**

1. מקור הנתונים ומשמעותם: סט הנתונים נקרא “The Cleveland heart disease data”, אשר סופק על ידי אוניברסיטת קליפורניה ומקורו באתר UCI Machine Learning Repository.  
   בסיס הנתונים מכיל 303 תצפיות אשר 212 מתוכם סווגו לאימון המודל והשאר לבחינתו. כמו כן, הוא מכיל במקור 76 פיצ'רים אך אנו משתמשים ב13 מהם בלבד, כאשר חלקם מספריים וחלקם קטגוריים:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | הסבר | סוג | טווח | משמעות | חישה |
| age | גיל המטופל בשנים | מספרי | 34-80 1 |  | ידני |
| gender | מין המטופל | קטגורי | 0-1 | 0-נקבה; 1-זכר | ידני |
| cp | סוג כאב בחזה | קטגורי | 0-3 | 0-תעוקת לב טיפוסית  1-תעוקת לב לא טיפוסית  2-לא נובע מתעוקת לב  3-אסימפטומתי | מומחה |
| trestbps | לחץ דם במנוחה | מספרי | 94 - 200 |  | מדידה |
| chol | רמת כולסטרול | מספרי | 126-409 |  | מדידה |
| fbs | רמת סוכר בדם | קטגורי | 0-1 | 0-עד 120 מ"ג/דצ"ל ; 1 - אחרת | מדידה |
| restecg | תוצאות אק"ג | קטגורי | 0-2 | 0-תקין  1-עיוות בגל T במקטע ST  2-סיכוי להיפרטרופיה בחדר שמאלי | מומחה |
| thalach | דופק לב מקסימלי | מספרי | 88-195 |  | מדידה |
| exang | תעוקת חזה עקב מאמץ | קטגורי | 0-1 | 0-לא, 1-כן | מומחה |
| oldpeak | צניחה בסגמנט ST באק"ג, במאמץ לעומת מנוחה | מספרי | 0-5.6 |  | מומחה |
| slope | שיפוע בסגמנט ST בשיא מאמץ | קטגורי | 0-2 | 0-חיובי, 1-שטוח, 2-שלילי | מומחה |
| ca | מספר כלי דם גדולים | קטגורי | 0-4 | 0-3-כמות, 4-לא ידוע | מומחה |
| thal | מחלת תלסמיה | קטגורי | 0-4 | 0-ערך חסר, 1- פגם קבוע  2-נורמלי, 3- פגם ניתן לריפוי | מומחה |
| y | אירוע התקף לב | קטגורי | 0-1 | 1-התקף לב, 0-אחרת | מומחה |

1 לאחר מחיקת תצפיות לא הגיוניות מבחינת ערכים.

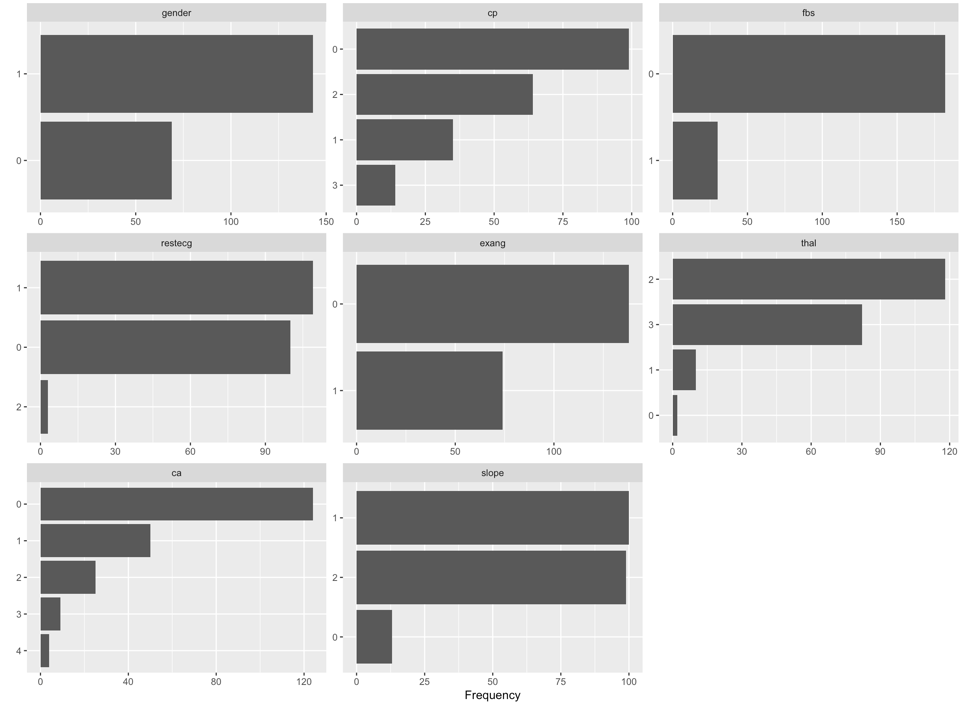
A picture containing screenshot

Description automatically generated

1. הסתברות אפריורית וקשרים:

ראשית נבחן את ההסתברות האפריורית של משתנה המטרה: בסט האימון קיימת הסתברות של 53% לקבל את הערך 1. כלומר ניתן להגיד שמשתנה המטרה מתפלג פחות או יותר 50-50 **במדגם שלנו**.

תרשים 1

באופן דומה ניתן לראות את התפלגות שאר הפיצ'רים וההסתברויות האפריוריות שלהם לקבלת כל ערך. בבחינת הפיצ'רים הקטגוריים (תרשים 2) ניתן גם לראות שאין פיצ'רים בעלי קרדינליות גבוהה (מס' רב של קטגוריות, מה שיכול להוות בעיה באימון מודל).

תרשים 2

A picture containing sitting, photo, white, table

Description automatically generatedבבחינת הפיצ'רים הנומריים, ניתן לראות כי לפיצ'ר age יש ערכים בלתי הגיוניים ונרצה לתקן זאת לפני שנמשיך בניתוחים (פירוט התיקון יפורט בסעיף 3).

תרשים 3

A close up of a map

Description automatically generatedבחינת סט האימון מול סט הבדיקה:

A screenshot of a cell phone

Description automatically generatedנרצה לבחון האם הדגימה שלנו מייצגת את ה"מציאות" עליה אנחנו מעוניינים להיבחן בבחינת ערכי הפיצ'רים השונים. ניתן לראות כי הפיצ'רים מתפלגים בצורה יחסית דומה בסט האימון ובסט הבדיקה, לכן ניתן לומר שנתוני האימון רלוונטיים לסט הבדיקה, והם מייצגים את אותה "מציאות".

תרשים 4

בחינת הפיצ'רים מול משתנה המטרה

כעת נרצה לבחון כמה הפיצ'רים השונים יכולים לעזור לנו במשימת ניבוי הסיכוי להתקף לב. לצורך כך אנחנו מציגים את ההתפלגות של כל משתנה עבור כל אחת מה-classes, וכמו כן את אחוז התצפיות שמקבלות y = 1 (התקף לב) עבור כל קבוצת ערכים.

לדוגמא ניתן לראות שבקרב תצפיות בטווח הגילאים 34-42, 80% מהתצפיות שייכות לתצפיות בהם המטופלים קיבלו התקף לב, לעומת 20% בלבד מהתצפיות בגילאים 59-61. ניתן לראות שפיצ'ר הגיל מפריד יפה בין ה-classes לעומת פיצ'ר לחץ הדם במנוחה שמפריד משמעותית פחות טוב.

A picture containing text, map

Description automatically generatedA picture containing text, map

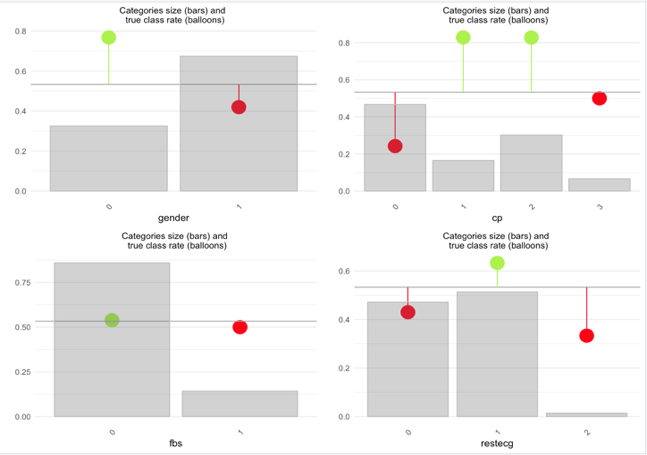
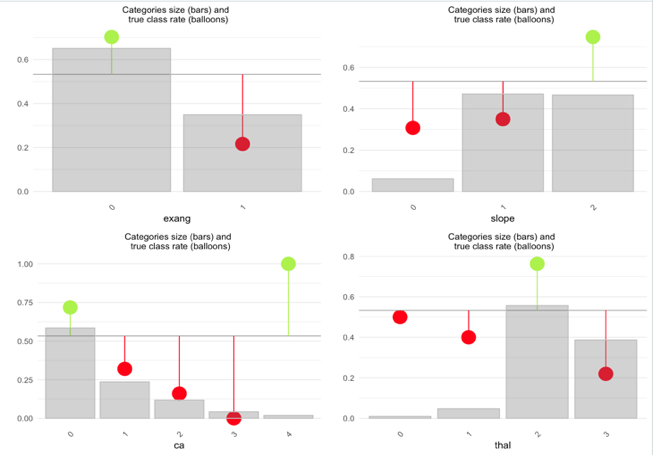
Description automatically generatedבאופן דומה, כולסטרול מפריד פחות טוב לעומת בדיקת האק"ג ודופק לב המקסימלי.

A close up of a map

Description automatically generated

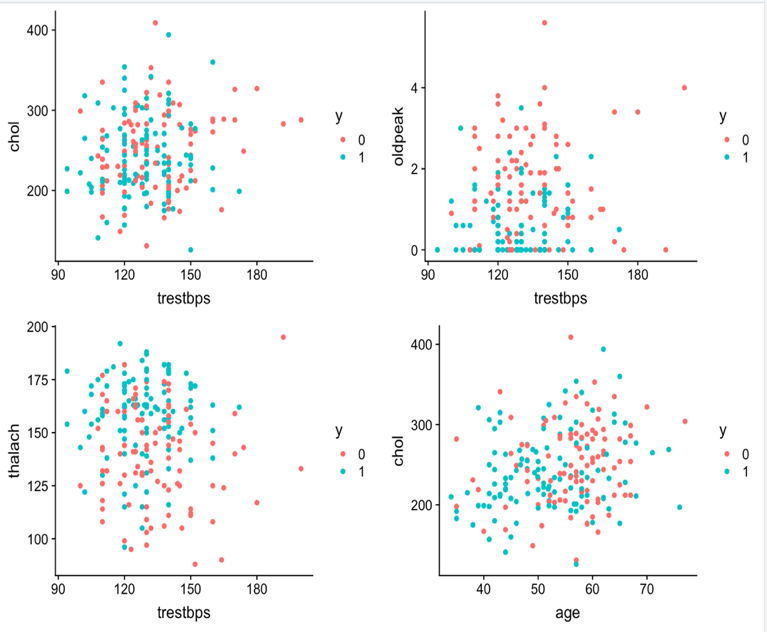
תרשים 5

בקרב הפיצ'רים הקטגוריים (תרשים 6), ניתן לבדוק באופן דומה איזה פיצ'רים מפרידים היטב בין הקטגוריות השונות, תכונה שתעזור להם להיות אפקטיביים יותר עבור אימון המודל. נראה כי פיצ'ר ה-fbs שמודד את רמת הסוכר בדם לא מפריד טוב בין המחלקות, לעומת הפיצ'רים האחרים שעושים עבודה לא רעה.



תרשים 6

בדיקת כל פיצ'ר בפני עצמו מול משתנה המטרה איננה מספקת, שכן ייתכנו פיצ'רים שהצטלבויות הערכים שלהם מפיקים מידע בעל ערך. לצורך כך ניתן לבחון תרשימי פיזור, ובדוגמאות שלפנינו (תרשים 7) ניתן לראות מספר דוגמאות לקורלציות משותפות מעניינות בין שני משתנים למשתנה המטרה.



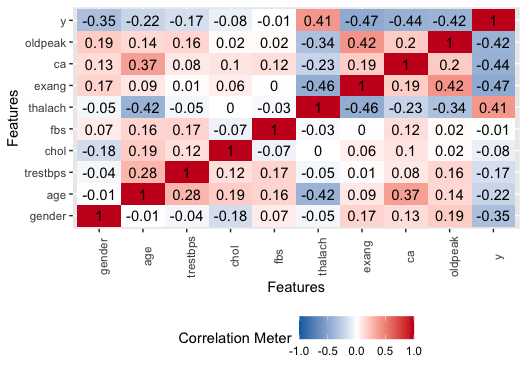
תרשים 7

כאשר בחנו את משתנה לחץ הדם (trestbps) בפני עצמו מול משתנה המטרה הוא לא נראה משמעותי במיוחד, אבל כאשר רואים לחץ דם גבוה בשילוב עם כולסטרול גבוה, או צניחה גבוהה בסגמנט הST קיים סיכוי נמוך יותר להתקף לב (בקרב הנתונים). כמו כן דופק לב גבוה ולחץ דם נמוך גם הם אינדיקטיביים במשותף לאי התקף לב בקרב הנתונים.

A screenshot of a cell phone

Description automatically generatedלבסוף רצינו להוסיף גם בדיקה כמותנית (quantitative) של הפיצ'רים השונים, ולצורך כך בחרנו בבדיקת ה-Information Gain. ערך גבוה של IG מעיד על פיצ'ר שסביר מאד שיבחר על ידי מודל לומד שכן הוא מקטין את האנטרופיה (אי הודאות, חוסר סדר) בdata באופן משמעותי.

ניתן לשים לב שמשתנה רמת הסוכר בדם קיבל ערך IG נמוך מאד, מה שתואם את הבדיקה האיכותנית מהחלק הקודם.

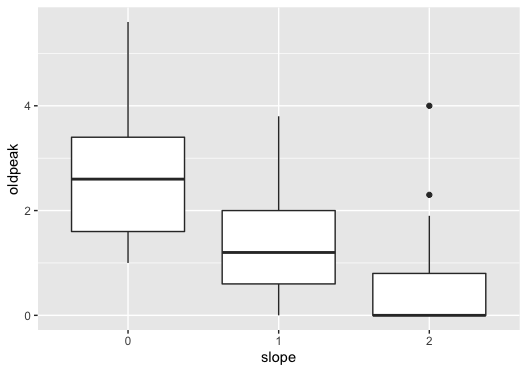
קשרים בין משתנים:

ראשית, נבחן האם קיימים קשרים בין הפיצ'רים הנומריים והאורדינליים בסט הנתונים בעזרת מטריצת קורלציות של פירסון.

לא ניתן להגיד שקיים מתאם גבוה שמצדיק התייחסות בין המשתנים שנבדקו.

תרשים 8

ניתן לראות מתאם קל (|0.35| - |0.46|) בין שלושה פיצ'רים: thalach, exang, oldpeak.  
המתאם בין הפיצ'רים הללו צפוי, במידת מה, משום שידוע שאדם בעל דופק מקסימלי גבוה יותר הוא בריא יותר, על כן, לא נצפה ממנו לחוות תעוקת חזה במאמץ ולא נצפה לראות צניחות בסגמנט  
 ה-ST בבדיקות האק"ג שלו (זאת לאחר בדיקה כי פיצ'ר זה נמצא קשור למקרים של אירועי ומחלות לב).



תרשים 9

ננסה לבדוק גם את המתאם בין פיצ'רים קטגוריים לנומריים על ידי תרשים בוקס פלוט (תרשים 9).

ניתן לראות כי קיים מתאם כלשהו בין הצניחה לשיפוע בסגמנט ST בשיא מאמץ.

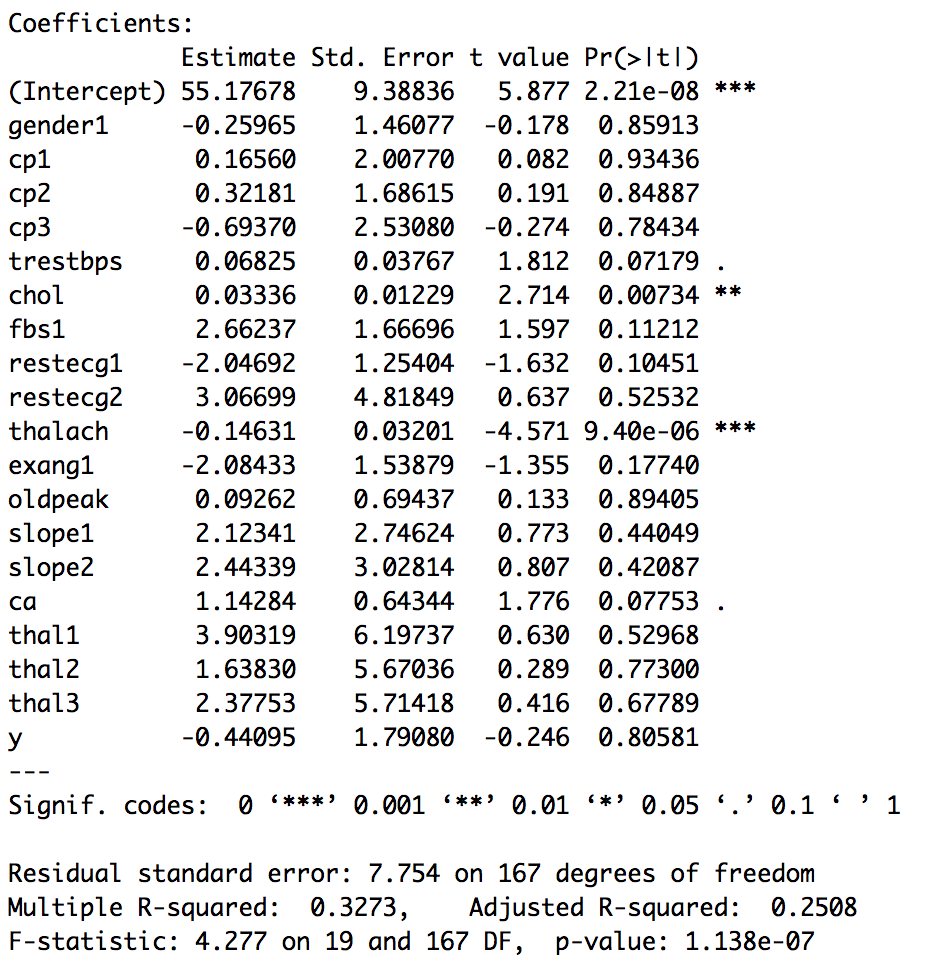
גם לקשר זה יש היגיון רב שכן הנתונים נמדדים במקביל על האק"ג באותו האזור.

מעבר לקשר זה לא מצאנו קשרים מעניינים ומשמעותיים נוספים.

1. איכות הנתונים

בסט הנתונים לא נצפו נתונים חסרים של ממש – כלומר, אין ערך שאינו מוזן. ישנן תצפיות בפיצ'ר ‘ca’ ו ’thal’אשר מייצגים מצב של ערך לא ידוע, אך מאחר וערך זה מופיע גם בסט הבדיקה נשאיר אותו גם כחלק מסט האימון.

בנוסף, כפי שציינו קודם, בפיצ'ר age ישנם נתונים לא הגיוניים של ערכי גיל (מעל 1000 שנים).  
במקרה זה ישנן מספר אפשרויות:

* מחיקת התצפיות: יתרון הפעולה היא קלות הביצוע. עם זאת, מדובר בכ-25 תצפיות המהוות יותר מ10% מסט האימון.
* הורדת הפיצ'ר: היתרון הוא קלות הביצוע, אך גם פעולה זו גורמת לנו לאבד מידע שימושי עבור המודל.
* השלמת התצפיות: פעולה זו היא מורכבת יותר. אנו בחרנו לעשות השלמה של התצפיות עבור הסעיף הקודם כדי לראות את האיזון בין בסיסי הנתונים.  
  הדרך בה בחרנו להשלים את הנתונים היא על ידי חיזוי של רגרסיה ליניארית מול שאר הפיצ'רים.  
  למרות שבמודל הלינארי רוב הפיצ'רים לא הראו אפקט מובהק, וכמו כן יש לו R בריבוע נמוך יחסית, בחרנו להשתמש בו כדי להתקרב בצורה הטובה ביותר שיכלנו, ראינו כי פיצ'ר הגיל מפריד בצורה טובה בין ה-classes ולכן לא נרצה לוותר עליו או לאבד את הנתונים של הפיצ'רים האחרים.

ייתכן וקיימת דרך טובה יותר לנסות לחזות את גילאי הנבדקים. אופציה מורכבת יחסית היא פריסה של חלק מהמשתנים לסגמנטים והפעלת רגרסיה ליניארית בחתכים שונים של המשתנה בניסיון לחזות את הגיל.

**הכנת הנתונים:**

1. השמטת מאפיינים:

הן בבחינה האיכותנית והן בבחינה הכמותנית, פיצ'ר רמת הסוכר בדם (trestbps) נראה כלא אינדיקטיבי ולכן ניתן להוריד אותו מהמודל. ישנם פיצ'רים נוספים שקיבלו ערך IG נמוך כמו לדוגמא משתנה לחץ הדם (trestbps), אבל ראינו בבחינה האיכותנית שמשתנה זה מציג אינטראקציה יפה עם משתנים אחרים ולכן נשקול להשאיר אותו לשלב המידול.

1. טיפול פרטני במאפיינים:

* ראינו שפיצ'ר הכולסטרול מקבל IG נמוך וכמו כן שהוא לא מפריד בצורה טובה מאד בין הclasses בצורתו הנוכחית. בעולם הרפואה נהוג לשייך רמות שונות של כולסטרול (נמוך\בינוני\גבוה) לסיכוי לחלות בהתקף לב. לכן נבצע דיסקרטיזציה למשתנה ונהפוך אותו לפיצ'ר עם שלוש רמות בלבד.
* באופן דומה ניתן לעשות דיסקרטיזציה למשתנה לחץ הדם (trestbps) ולהפכו למשתנה בעל 4 רמות בלבד שמייצגות רמת סיכון מעולם הרפואה.  
  כאשר אנחנו עושים דיסקריטזציה למשתנים, אנחנו מאבדים מידע (ולכן גם מדד הIG שלהם יהיה נמוך יותר), אבל יכול להיות שלאחר צמצום מרחב הקטגוריות של המשתנה, המודל ידע לבחור נקודות פיצול מדויקות יותר בעת האימון שישפרו את החיזוי.