תרגיל בית 2

Graphical user interface, text, application, email

Description automatically generated

Chart, scatter chart

Description automatically generated

Graphical user interface, text

Description automatically generated

לפי הגרף שיצא ניתן לראות ש – אנשים אשר היו בעלי ערכים גבוהים יותר ב 07\_PCR , 03\_PCR יחדיו ונמוכים ב 10\_PCR קוטלגו כבעלי spread גבוהה  
בנוסף ניתן לראות שהערכים שקוטלגו כ-spread נמוך מרוכזים יחדיו סביב ערכי 03\_PCR הקרובים ל-0 , ערכי 07\_PCR ששליליים וערכי 10\_PCR גדולים  
Graphical user interface, text, application

Description automatically generated

רמת הדיוק אותה קיבלנו היא 0.82

Graphical user interface, text

Description automatically generated

Z-score יודע להתמודד עם outliers מפני שהכל מוזז בהתאם לממוצע וסטיית תקן  
אך מכיוון שכל תכונה מושפעת רק מהממוצע וסטיית התקן שלה, התחום המתקבל לכל תכונה הוא לא בהכרח זהה

לדוגמה – לתכונה A עם mean = 20 ו סטיית תקן של 1 נקבל תחום של   
לעומת מקרה של תכונה B עם mean = 20 וסטיית תקן של 20 נקבל שכל הערכים מקיימים

לפי הנוסחה של z-score , לפיכך הערכים יהיו בין   
ותחום זה שונה מהתחום הקודם ולכן כל תכונה תשפיע על המרחק באופן אחר .

נבחר בתכונה זו במקרה שתחומי הערכים של התכונות זהים על מנת לאפשר טיפול ב outliers .

Min-max scaling : הנוסחה הזו מנרמלת את הערכים של כל תכונה לתחום , אך אינה יודעת להתמודד עם outliers , (מפני שהערכי הקיצון יקבלו את הערכים 1 או 0 דבר הישנה את ההתפלגות ויקשה עלינו לזהות את הoutliers )  
אך מכיוון שהנוסחה מעבירה את כל התכונות לתחום כל התכונות משפיעות באותו אופן על המרחק (לדוגמה ב KNN , קירבה תושפעה באופן שווה מכל התכונות ותיהיה מדוייקת יותר לעומת z-score שם תכונה אחת עלולה ליצור קרבה "מזוייפת" )  
ההיתרון הבולט של שיטה זו היא שחישוב מרחק של דוגמה מדוגמה אחרת נעשה באופן המתייחס באופן שווה לכל התכונות ובכך יכול למנוע טעויות.  
וכשאין outliers ברורים

Graphical user interface, text

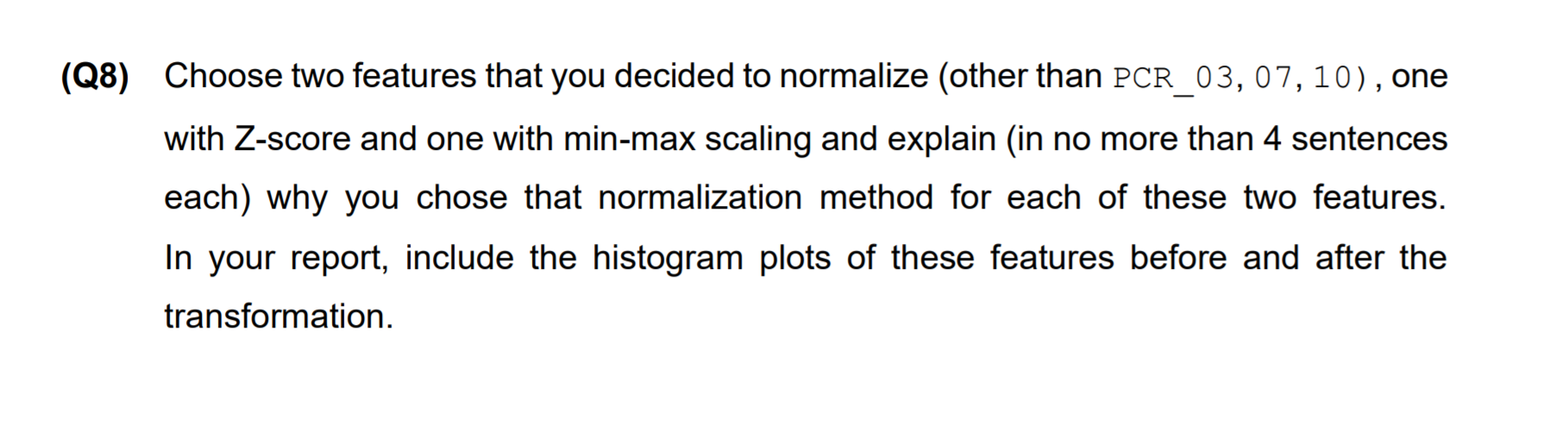
Description automatically generated

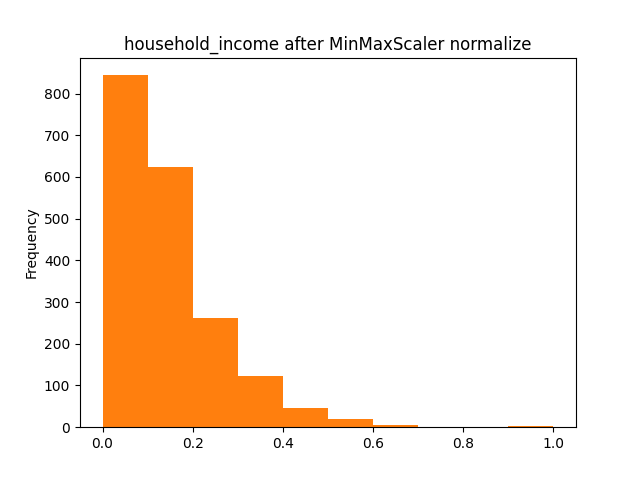
לאחר נירמול הערכים באופן הבא :

PCR\_03 בעזרת z-score  
PCR\_07 בעזרת z-score  
PCR\_10 בעזרת z-score

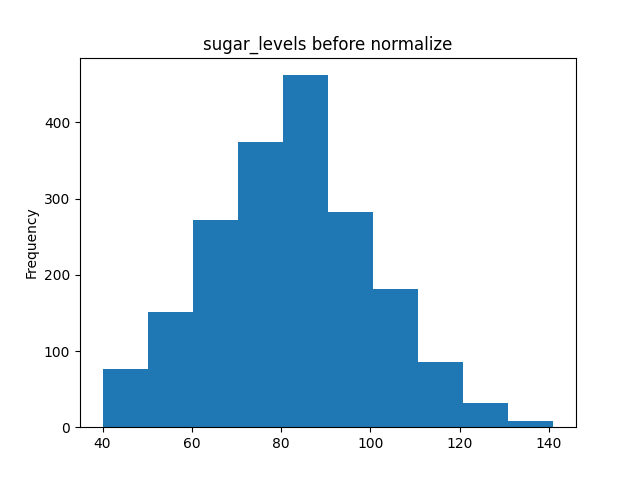
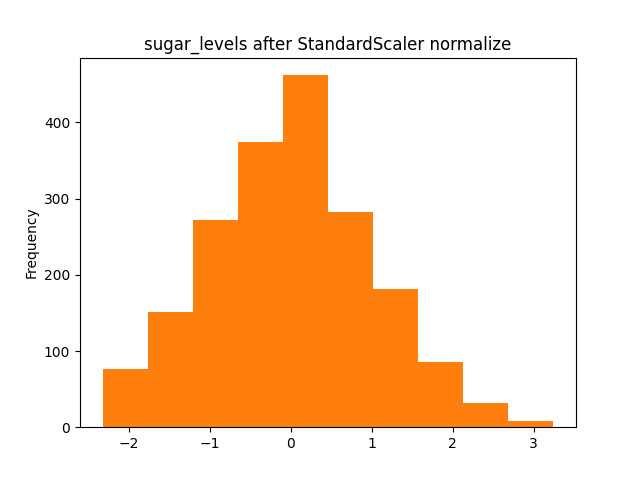
קיבלנו את התוצאה 0.92 על סט האימון

ההבדל בתוצאות הגיעה בשל הבדיקת מרחק מתוצאות שונות לפי הערכים, מכיוון שהבדל בערך אחד בצורה קיצונית יכל לגרום לקרבה לפי ערך זה למרות שקרבה זו לא באמת קיימת . (כי התוצאה מרחק היא בחד מימד ולא לפי כל תכונה בנפרד).  
ולכן נרמול הערכים, ובכך לתת לכל תכונה השפעה יחסית זהה על המרחק מאפשר להגיע לרמת דיוק גבוהה יותר.



Chart, histogram

Description automatically generatedבחרנו את השדות household\_income ו suger\_levels



כפי ששמנו לב השדה household\_income אינו דומה להתפלגות גוסיאנית, ולכן בחרנו לנרמל אותו לפי minmax ואכן ניתן לראות שההתפלגות נשמרת בצורה טובה

כמו כן השדה suger\_levels דומה להתפלגות גוסינית ולכן בחרנו לנרמל לפי strandScaler ואכן ניתן לראות שההפתלגות נשמרת בצורה טובה.

A screenshot of a computer

Description automatically generated

Chart, scatter chart

Description automatically generated

להוסיף הסבר מניח על הדעת

A screenshot of a computer

Description automatically generated

Chart, treemap chart

Description automatically generated

Graphical user interface, text

Description automatically generated

שאלה 11,

כפי שראינו בדוח הקודם שמנו לב שיש קורולציה בין risk לבין sugar\_level, ניתן לראות זאת על סמך הגרף הבא -

*Chart, histogram

Description automatically generated*

ועל סמך חישוב הקורולוציה שיוצא **0.2876366**

וכמו כן נשים לב כי –

Chart, scatter chart

Description automatically generated

כלומר , PCR\_01 וגם PCR\_02 יכולים לשמש לסיווג טוב של risk

לפיכך נבחר את הנתונים הבאים:

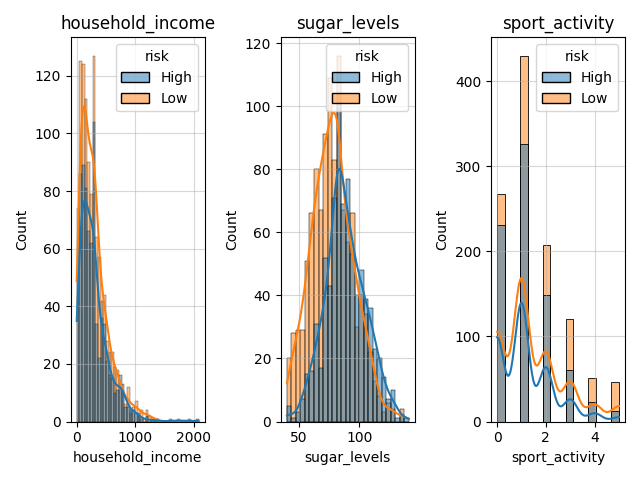
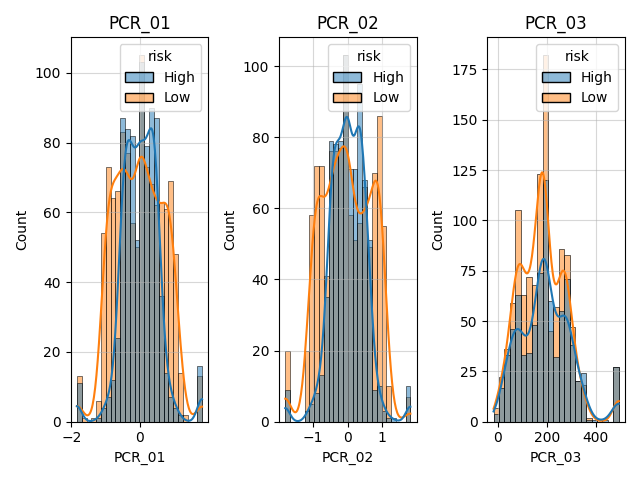
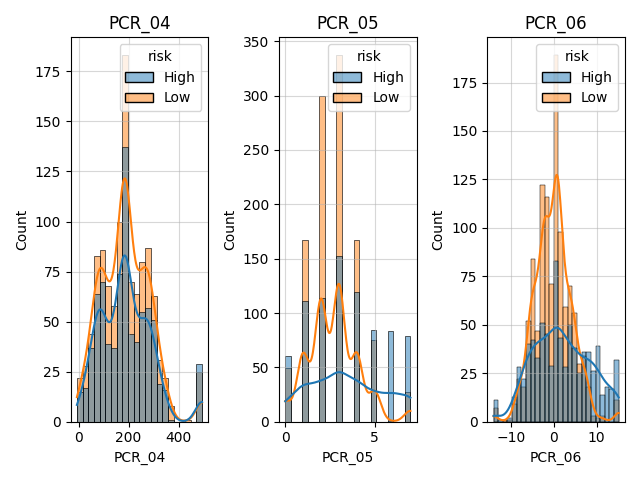
PCR\_1,PCR\_2, Sugar\_level

Text

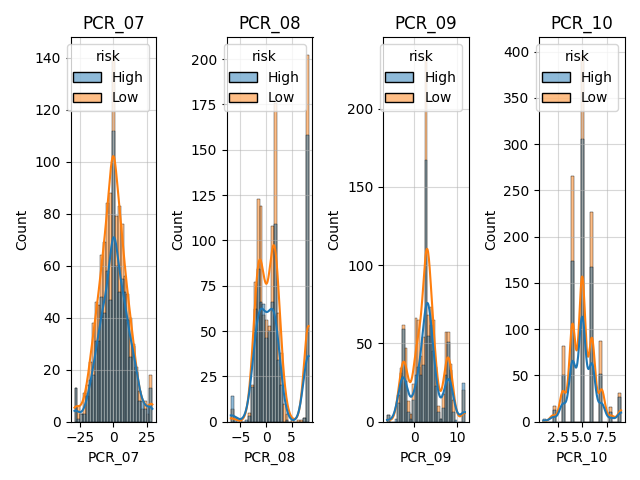
Description automatically generated with medium confidence

נחשב את הקורולציה ונקבל אותם –

להלן הגרפים:

Chart, histogram

Description automatically generated



וכעת חישובי הקורולציה, ונקבל:

|  |  |
| --- | --- |
|  |  |
| 0.2876366611320253 | sugar\_levels |
| 0.20683162673822178 | PCR\_05 |
| 0.1847972022180316 | PCR\_06 |
| 0.10559581603526123 | sport\_activity |
| 0.1017142169353659 | age |
| 0.053002364247947784 | PCR\_03 |
| 0.04375062889983171 | PCR\_07 |
| 0.03698906984068623 | PCR\_02 |
| 0.03261444239769013 | PCR\_01 |
| 0.02873376366907449 | blood\_type\_B- |

שאלה 13:

Graphical user interface, text

Description automatically generated

להלן העץ המתקבל

Diagram

Description automatically generated

ורמת הדיוק המתקבלת הינה 0.772

Graphical user interface, text, application, email

Description automatically generated

כן, נשים לב למשל כי sugar\_level מופיע גם בשאלה 11 וגם בשאלה 12.

כעת נבצע ניתוח קטן של כל השדות ואיפה הם נמצאים בעץ ההחלטה:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| באיזה רמה | מופיע בעץ החלטה | מופיע ב שאלה 13 | מופיע ב שאלה 12 | תכונה |
| רמה 1 רמה 3 רמה 4 | כן | כן | כן | sugar\_levels |
| רמה 4 | כן | כן | לא | PCR\_05 |
| רמה 1 רמה 3 | כן | כן | לא | PCR\_06 |
|  |  | כן | לא | sport\_activity |
|  |  | כן | לא | age |
|  |  | כן | לא | PCR\_03 |
|  |  | כן | לא | PCR\_07 |
| רמה 3 רמה 4 | כן | כן | כן | PCR\_02 |
| רמה 1 רמה 3 רמה 4 | כן | כן | כן | PCR\_01 |
|  |  | כן | לא | blood\_type\_B- |
| רמה 4 | כן | לא | לא | household\_income |

להלן מספר ממצאים:

* נשים לב כי כל התכונות שרשמנו בשאלה 12, מופיעות גם בעץ ההחלטה, ואכן ה sugar\_level מופיע בשורש, זאת משום שיש לו קורולציה מאוד חזקה עם risk .
* כמו כן נשים לב שחלק מן התכונות שמופיעות בשאלה 13,לא מופיעות בעץ ההחלטה, דבר זה ייתכן משום שיש קורלציה בין sugar\_level לבין sport\_activety ולכן לסנן עם שניהם לא יוסיף יותר מיידי מידע,
* כמו כן יתכן שבין age לבין חלק מן התכנונות המפויעות בעץ גם יש קורולציה המונעת ממנו להופיע גם בעץ ההחלטה
* לגבי pcr\_03 , pcr\_07 , ו blood\_type, יש להם קורולציה מאוד נמוכה יחד עם risk וכמו כן כנראה שילובם עם שאר ערכי המטרות לא מוסיף יותר מיידי מידע ולכן הם לא מופיעים בעץ ההחלטה
* להסביר למה יש אותם תכונות המופיעיות ברמות שונות בעץ ההחלטה

|  |  |
| --- | --- |
| 0.2876366611320253 | sugar\_levels |
| 0.20683162673822178 | PCR\_05 |
| 0.1847972022180316 | PCR\_06 |
| 0.10559581603526123 | sport\_activity |
| 0.1017142169353659 | age |
| 0.053002364247947784 | PCR\_03 |
| 0.04375062889983171 | PCR\_07 |
| 0.03698906984068623 | PCR\_02 |
| 0.03261444239769013 | PCR\_01 |
| 0.02873376366907449 | blood\_type\_B- |