תרגיל בית רטוב 1 – מבוא למערכות לומדות (02360766)

בן הייטנר – 213930175

לילך ביטון - 205764517

<u>חלק 1</u>:

(1) לפי פונקצית shape ניתן לקבוע כי לאוסף הנתונים יש 25 עמודות ו-1250 שורות.

(2) התכונה מייצגת את מספר השיחות (כנראה פנים אל פנים) שמטופל ניהל ביום. התכונה הזו היא אורדינלית בגלל שהיא יכולה להיות מספר טבעי בלבד (אין דבר כזה מספר שלילי של שיחות וכנ"ל שיחה לא שלמה).

```
conversations_per_day
2     220
4     207
3     201
5     153
6     125
1     115
7      77
8      52
10     33
9      23
11     14
12     10
13      7
16      4
14      4
17      3
21      2
Name: count, dtype: int64
```

(3)

שם תכונה	סוג	תיאור
patient_id	אורדינלי	מספר מזהה של המטופל במאגר
age	אורדינלי	גיל המטופל
sex	קטגורי	מין המטופל
weight	רציף	משקל המטופל
blood_type	קטגורי	סוג הדם של המטופל
current_location	קטגורי	מיקום המטופל (תמיד 0 משום מה)
num_of_sibilings	אורדינלי	מספר האחים ואחיות של המטופל

happiness_score	אורדינלי	מדד האושר של המטופל מ-1 ל-10 לפיו
houselhold_income	רציף	הכנסת משק בית (סדר גודל של רבבות?)
conversations_per_day	אורדינלית	מספר שיחות (פנים אל פנים) שניהל המטופל ביום
sugar_levels	אורדינלי	רמות סוכר בדם של המטופל
sport_activity	אורדינלי	רמת הפעילות של המטופל מ-1 ל-5 לפיו
pcr_date	אורדינלי	תאריך בו בוצעה בדיקת PCR על המטופל
PCR_01	רציף	תוצאת בדיקת PCR ראשונה
PCR_02	רציף	תוצאת בדיקת PCR שנייה
PCR_03	רציף	תוצאת בדיקת PCR שלישית
PCR_04	רציף	תוצאת בדיקת PCR רביעית
PCR_05	רציף	תוצאת בדיקת PCR חמישית
PCR_06	רציף	תוצאת בדיקת PCR שישית
PCR_07	רציף	תוצאת בדיקת PCR שביעית
PCR_08	רציף	תוצאת בדיקת PCR שמינית
PCR_09	רציף	תוצאת בדיקת PCR תשיעית
PCR_10	רציף	תוצאת בדיקת PCR עשירית

(4) הסיבה שחשוב לשמור על חלוקה קבועה לאורך כל האנליזה של המידע היא שככה ניתן לראות את השפעת המידול והאופטימיזציה על המידע. חלוקה קבועה נותנת ולידציה לשינוי בדיוק בכך שידוע שההבדל לא נובע משינוי בתנאיי האימון/בדיקה.

חלק 2:

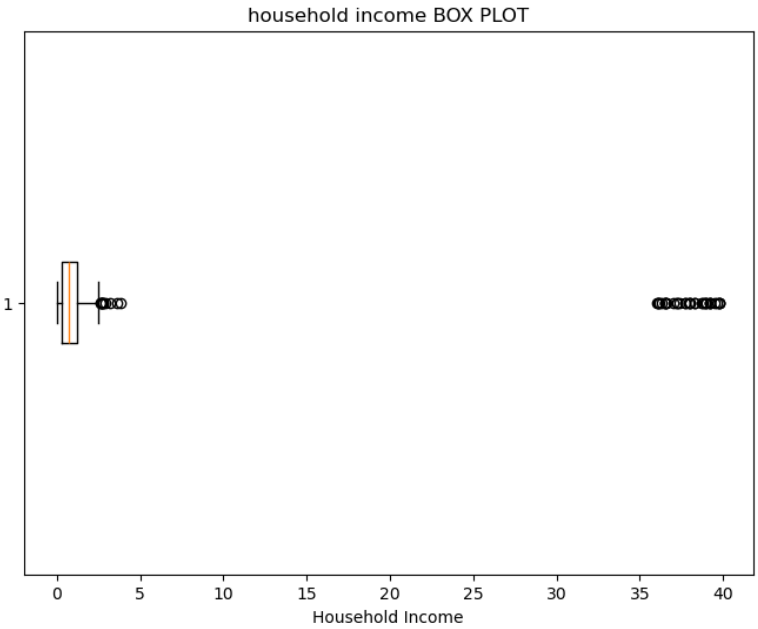
(5) עבור סט האימון:

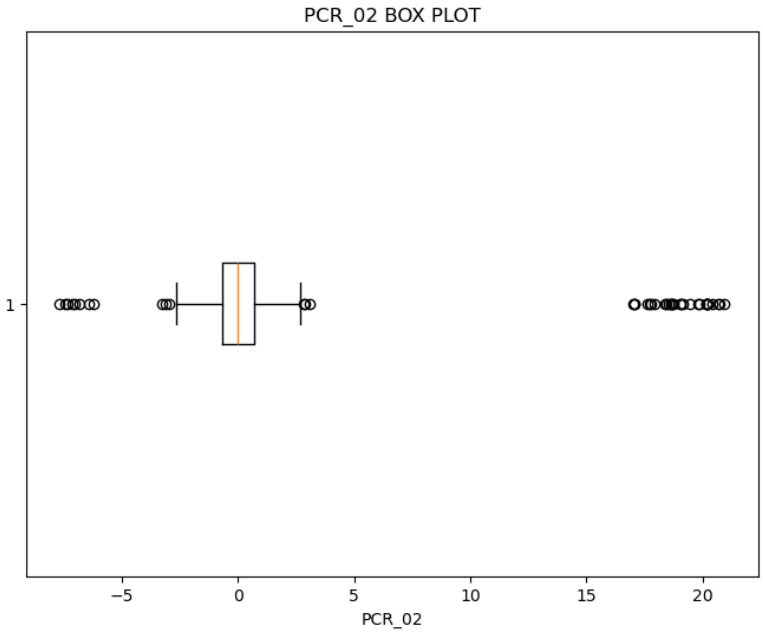
```
household_income        108
PCR_02                   58
```

עבור סט הבדיקה:

```
household_income         31
PCR_02                   16
```

(6)





חריגות הם נקודות בעלי ערכים גדולים מאוד או קטנים מאוד ביחס לשאר הנקודות באוסף הנתונים.
בדר"כ כלל מזהים מרחק גדול מאוד ביחס ע"י חלוקה לרבעים וסימון כחריגות נקודות שנמצאות מעל $Q3 + 1.5 \cdot IQR$ או מתחת ל-$Q1 - 1.5 \cdot IQR$. ב-box plot מסמנים את הספים שציינו קודם ובעזרתם מזהים את החריגות בנתונים.
כפי שניתן לראות מהשרטוטים לעיל, כל נקודה בכל אחד מהשרטוטים היא חריגה.

(7) ***household_income עבור:***

ממוצע: ~2.18

חציון: 0.7

PCR_02 עבור:

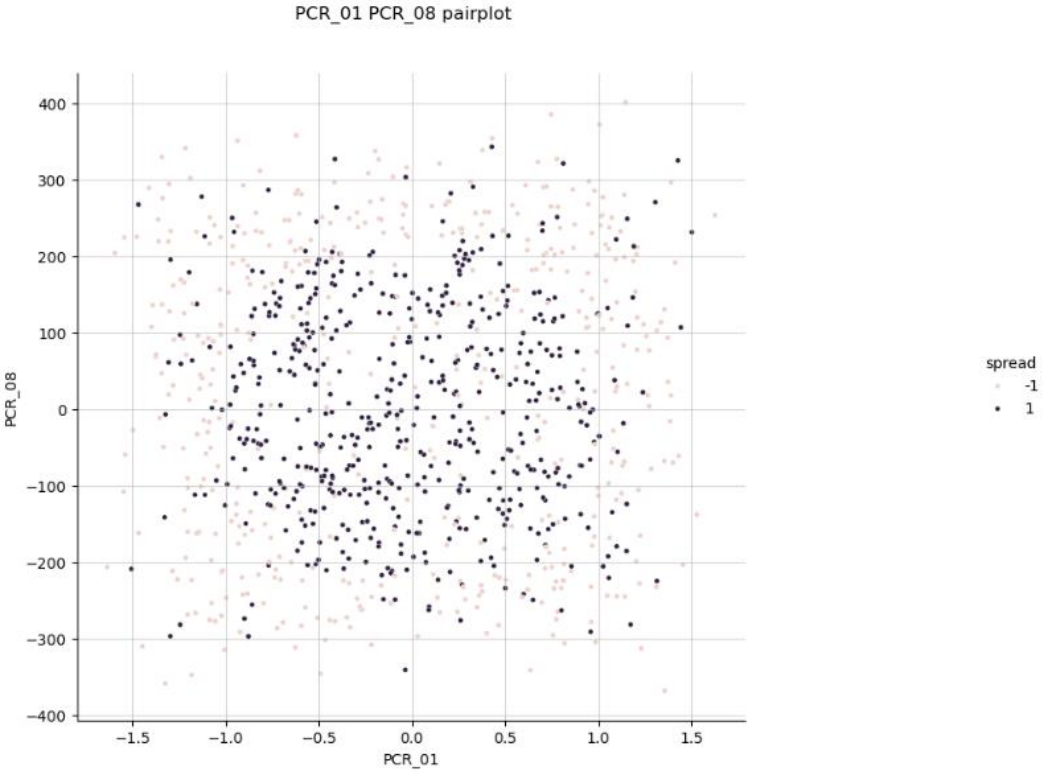
ממוצע: ~0.496

חציון: -0.005455

אכן יש הפרש מהותי בין הממוצע לחציון והסיבה לכך היא הכמות הגבוהה של חריגות (ממוצע רגיש לחריגות). במקרה שלנו נבחר להשתמש בשיטה b ולמלא בעזרת החציון. הסיבה לכך היא שכפי שציינו קודם, ממוצע רגיש לחריגות ועל כן מילוי באמצעותו סביר ליצור/לחזק הטיה.

חלק 3:

(8)



הצמד PCR_08 ו-PCR_01 שימושי לחיזוי spread כיוון יוצר הפרדה יחסית טובה של אזור מעגלי של בעיקר spread חיובי ואזור היקפי של בעיקר spread שלילי.

(9) לפי החישוב של ידידינו NumPy, המתאם בין PCR_01 ל-spread הוא 0.0015 ובין PCR_08 ל-spread הוא 0.0711-. זה לא סותר את הטענה שהבחירה שלנו טובה לחיזוי spread שכן הסיבה למתאם האפסי היא שבין לתכונות שבחרנו לבין spread אין קשר לינארי ברור. בכל זאת אין זה אומר שלא נוכל לחזות את spread לפיהן מכיוון שניתן לעשות זאת בעזרת מסווגים שלא מסתמכים על קשר לינארי (כמו שאנו עושים בהמשך עם kNN).

(10) נבחין כי הפעולות בפונקציה שלנו שישפיעו על הסיבוכיות (כלומר לא יתבצעו ב-O(1)) הן: cdist, argpartition, copy ו-mean.

cdist: מדידת מרחק בין זוג נקודות ממימד d מתבצע בסיבוכיות $O(d)$. נעשה זאת בין m נקודות אימון לנק' המבחן היחידה שלנו ונקבל בסה"כ: $O(md)$.

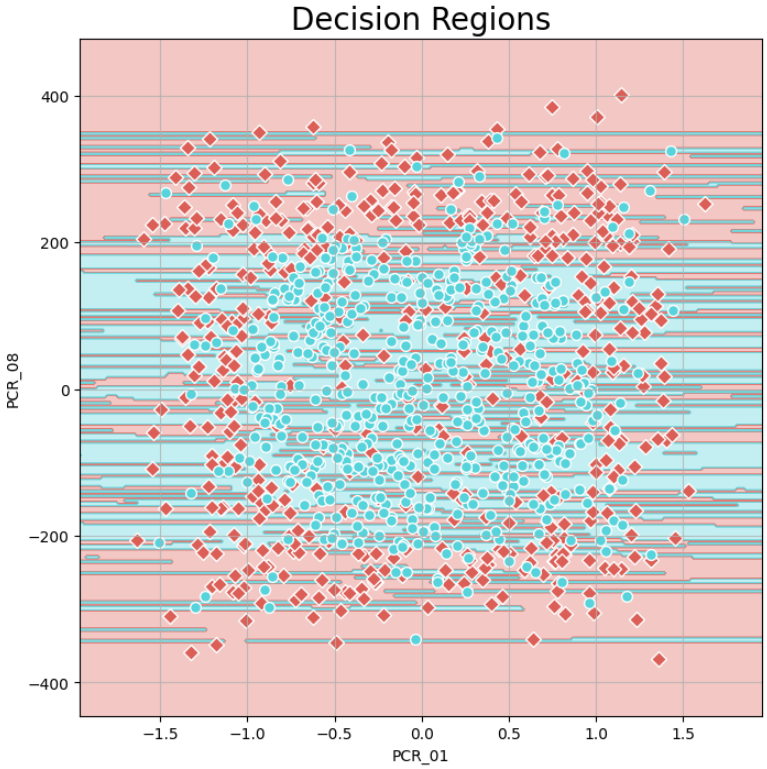
argpartition: הפונקציה הזו עושה מיון חלקי של m האינדקסים ובכך מוצאת את $k+1$ האלמנטים הקרובים ביותר. בהנחה שהמיון אכן טוב וחסכוני, זה יתבצע ב-$O(mlogk)$.

copy: פה נעתיק k משתנים וזה כמובן יתבצע ב-$O(k)$.

<u>mean</u>: זו למעשה מבצעת k פעולות חיבור ופעולת חילוק אחת ולכן מתבצעת ב-$O(k)$.

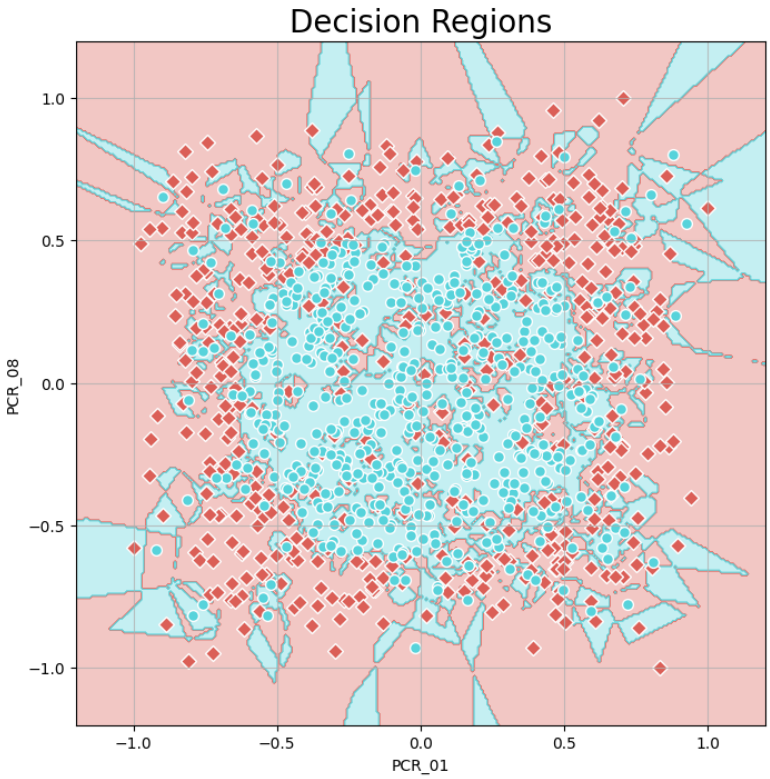
בסה"כ קיבלנו כי הפונקציה כולה תתבצע ב-$\boldsymbol{O(md)}$.

(11)



דיוק האימון הוא 63% ודיוק הבדיקה הוא 57.2%.

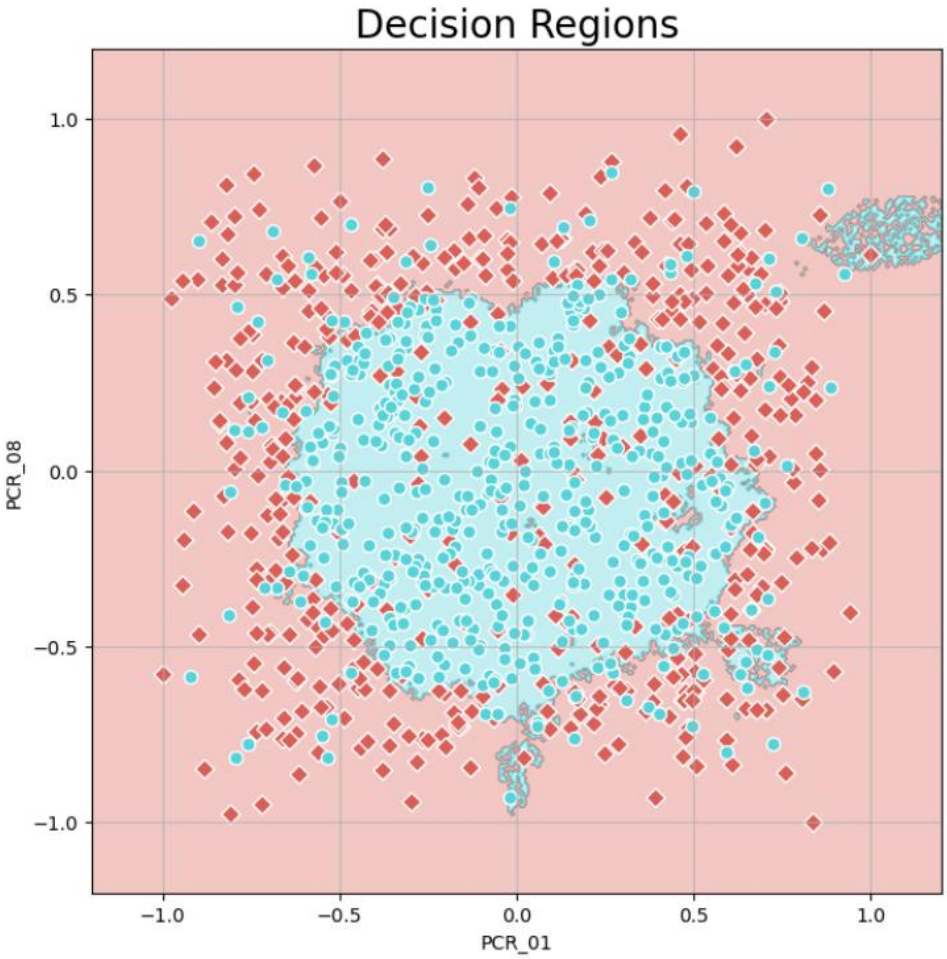
(12)



דיוק האימון הוא 69.3% ודיוק הבדיקה הוא 66%.

כפי שאפשר לראות משרטוט שקיבלנו, הנורמליזציה מקנה התגבשות יותר טובה של אזורי ההחלטה (לעומת הקוים הספורדים שקיבלנו קודם). זה נובע מכך שהנורמליצזיה מבטלת את ההשפעה שיש להבדלי הטווחים בין התכונות על המודל.

(13)



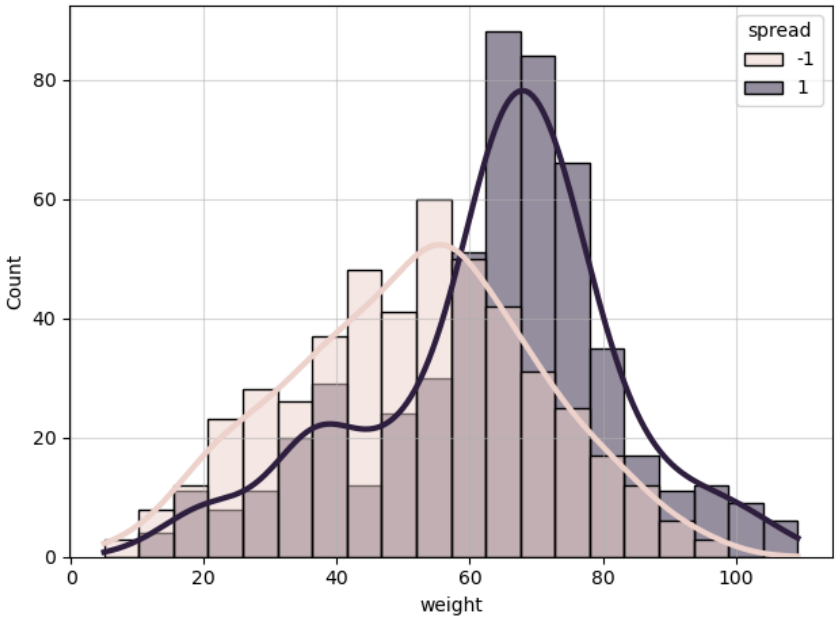
דיוק האימון הוא 80.6% ודיוק הבדיקה הוא 76.8%.

כפי שאפשר לראות שימוש ב-k גדול יותר מובילה להתהוות של גושים יותר אחידים שפחות רגישים לטעויות נקודתיות. זה נובע מההשפעה הקטנה של כל נקודה בממוצע שקובע את אזורי ההחלטה עם גדילת k. בפשטות, k קטן יתר על המידה גורם ל-overfitting ו-k גדול יתר על המידה גורם ל-underfitting.

(14) נבחין כי שתי ההתפלגויות שונות מאוד זו מזו ובפרט בקצוות שלהן (היוניפורמית אחידה לאורכה וכי-בריבוע אינה אחידה ובעלת זנב ארוך). כמו כן כי-בריבוע אינה חסומה מלמעלה, וזה רק מחזק את חוסר ההתאמה בנירמול שתי התכונה לפי min-max. כתוצאה מאלה, לא רק שנקבל עיוות מהותי המתבטא בכיווץ של הזנב ומתיחה של הסטיות של התכונה הנדגמת מכי-בריבוע, אלא לא נקבל שום עיוות משמעותי עבור התכונה הנדגמת מההתפלגות היוניפורמית. כלומר, בכך שהפעלנו על שתי התכונות נרמול min-max, סביר שרק נחמיר את ההבדלים בינהן.

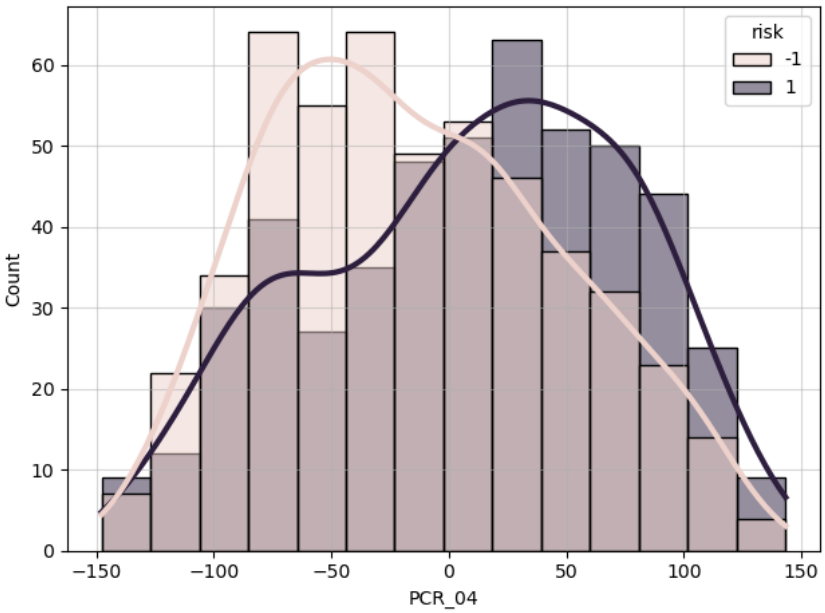
חלק 4:

(15)



הפיצ'ר הכי אינפורמטיבי (חוץ מהפיצ'רים PCR_08 ו-PCR_01 משאלה 8) הוא הפיצ'ר weight מכיוון שכפי שניתן לראות בגרף המצורף, הפילוגים של פיצ'ר זה לתיוג spread חיובי או שלילי, שונה במובן הזה שבערכים בהם הסבירות הגבוהה ביותר לתיוג חיובי (בפיק של הפילוג הכהה), מתקבל סבירות נמוכה יותר לתיוג שלילי ולהפך, ולכן על סמך פילוגים אלו ניתן להפריד את הנתונים במידה.

(16)

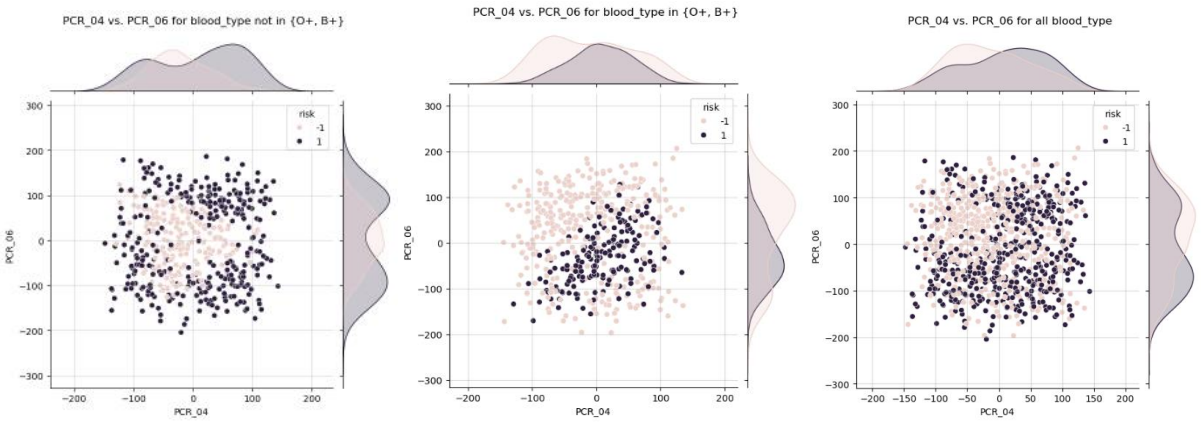


הפיצ'ר הכי אינפורמטיבי בעבור התיוג risk הוא הפיצ'ר PCR_04 מכיוון שכפי שניתן לראות בגרף המצורף, הפילוגים של פיצ'ר זה לתיוג risk חיובי או שלילי, שונה במובן הזה שבערכים בהם הסבירות

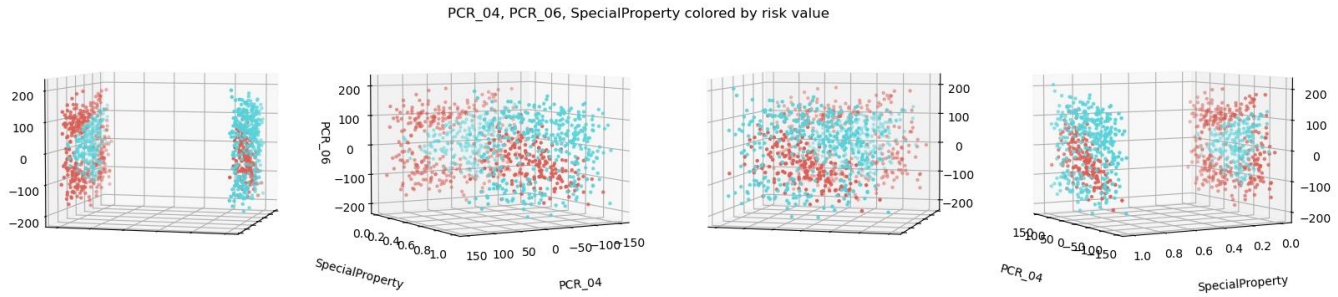
הגבוהה ביותר לתיוג חיובי (בפיק של הפילוג הכהה), מתקבל סבירות נמוכה יותר לתיוג שלילי ולהפך, ולכן על סמך פילוגים אלו ניתן להפריד את הנתונים במידה מסוימת.

(17) הצמד PCR_06 ו-PCR_04 שימושי לחיזוי risk (עם החלוקה לפי SpecialProperty) מאחר ויוצר הפרדה מרחבית יחסית טובה בין תיוג risk חיובי לתיוג risk שלילי בעבור כל קבוצת חלוקה. בקבוצה אחת ניתן לראות כי נתונים בעלי תיוג חיובי מקובצים במרכז והיתר מסביב, ואילו בקבוצה השנייה ניתן לראות כי נתונים בעלי תיוג שלילי מקובצים במרכז והיתר מסביב.

(18)



(19)



(20) עץ החלטה בעומק 3 יתאים את קבוצת האימון בצורה לא אופטימלית, מאחר והצמד PCR_06 ו-PCR_04 הינם רציפים ולא פרידים לינארית אחד עם השני, זאת אומרת שלאחר הפיצול של הפיצ'ר הבוליאני (חלוקה לפי SpecialProperty), נצטרך לבצע כמה פיצולים בעבור טווחים שונים לאותו פיצ'ר לסירוגין עם הפיצ'ר הרציף השני, על מנת לבצע התאמה אופטימלית, אשר בהכרח לא יתאפשר בעומק 3 בלבד.

(21) עץ החלטה בעומק 30 יתאים את קבוצת האימון בצורה טובה מאחר ולאחר הפיצול של הפיצ'ר הבוליאני (חלוקה לפי SpecialProperty), יהיה ניתן לבצע פיצולים רבים נוספים בעבור כל טווח לכל פיצ'ר לחילופין עד להפרדה אופטימלית.

(22) כפי שהראינו בשאלות 11-12, התאמה על סמך מודל 1-NN אינו מבצע התאמה טובה מאחר והחיזוי מתבצעת בעבור הנקודה הקרובה ביותר בלבד, ומאחר וה-Scale של כל פיצ'ר שונה ובעל פילוג שונה, במיוחד עבור הפיצ'ר הבוליאני. לכן יכול להיווצר עיוות במרחקים, בין שני הפיצ'רים הרציפים, ובמיוחד בשילוב עם הפיצ'ר הבוליאני. ראינו בניתוח הקודם שעשינו כי הפרדה על סמך פיצ'ר זה תחילה משפרת את יכולת ההפרדה בין שני הפיצ'רים הרציפים, ובהתאמה כזו פיצ'ר זה ככל הנראה לא ישפיע מאחר והמרחק המקסימלי בציר זה הינו 1 בלבד אשר זניח ביחס למרחקים בשאר הצירים.

(23) התשובות לשאלות 20 ו-21 לא ישתנו שכן העצים לא "ירוויחו" או "יפסידו" מהנורמליזציה. זאת מכיוון שזאת לא תיצור הפרדה לינארית ברורה שתקל על עץ בעומק שלוש להפריד בין PCR_06 ל-PCR_04 אך גם לא תעמיס קשיים על עץ בעומק שלושים שמסוגל לפצל למספר רב של טווחים קטנים. לעומת זאת, התשובה לשאלה 22 תשתנה שכן הנרמול מאזן בין ה-scales של התכונות השונות ומקל על ביצוע הפרדה בעזרת 1-NN.

(24)

שם תכונה	חדש	שיטת נרמול	אסטרטגית מילוי
age	X	-	-
sex	X	-	-
weight	X	-	-
SpecialProperty	V	-	-
current_location	X	-	-
num_of_sibilings	X	-	-
happiness_score	X	-	-
houselhold_income	X	-	לפי חציון
conversations_per_day	X	-	-
sugar_levels	X	-	-
sport_activity	X	-	-
pcr_date	X	-	-
PCR_01	X	Min-Max	-

לפי חציון	Standardization	X	PCR_02
-	Min-Max	X	PCR_03
-	Min-Max	X	PCR_04
-	Standardization	X	PCR_05
-	Min-Max	X	PCR_06
-	Standardization	X	PCR_07
-	Min-Max	X	PCR_08
-	Standardization	X	PCR_09
-	Standardization	X	PCR_10