**תרגיל בית רטוב 1 – מבוא למערכות לומדות (02360766)**

*בן הייטנר – 213930175*

*לילך ביטון - 205764517*

חלק 1:

1. לפי פונקצית shape ניתן לקבוע כי לאוסף הנתונים יש 25 עמודות ו-1250 שורות.
2. A screenshot of a black screen

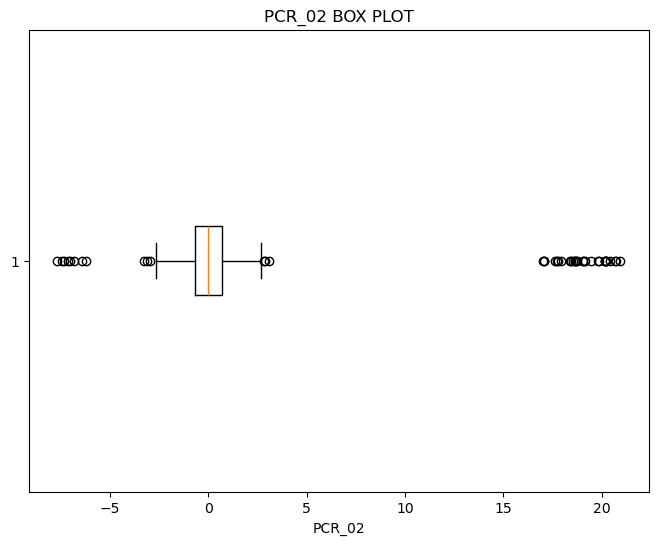
   Description automatically generatedהתכונה מייצגת את מספר השיחות (כנראה פנים אל פנים) שמטופל ניהל ביום. התכונה הזו היא אורדינלית בגלל שהיא יכולה להיות מספר טבעי בלבד (אין דבר כזה מספר שלילי של שיחות וכנ"ל שיחה לא שלמה).

|  |  |  |
| --- | --- | --- |
| **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 |
|  |  |  |

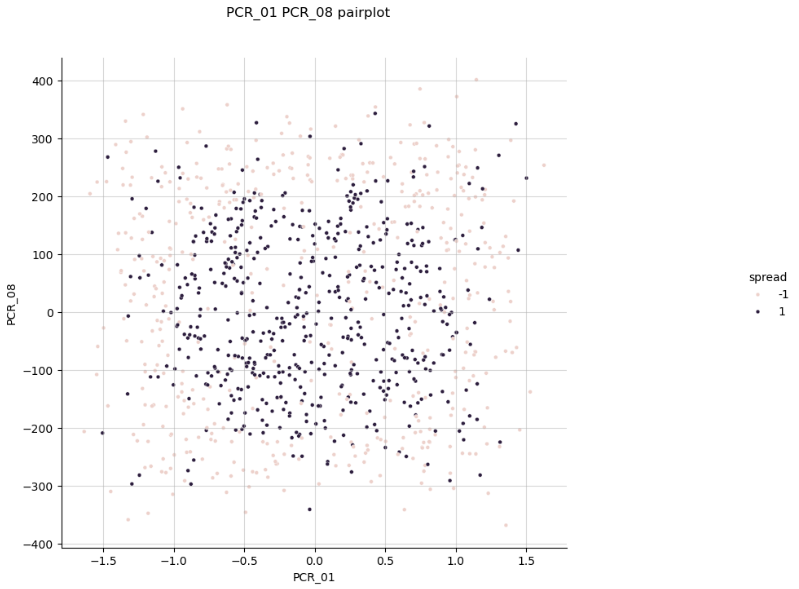
1. הסיבה שחשוב לשמור על חלוקה קבועה לאורך כל האנליזה של המידע היא שככה ניתן לראות את השפעת המידול והאופטימיזציה על המידע. חלוקה קבועה נותנת ולידציה לשינוי בדיוק בכך שידוע שההבדל לא נובע משינוי בתנאיי האימון/בדיקה.

חלק 2:

1. עבור סט האימון:  
     
   עבור סט הבדיקה:
2. A graph of a diagram

   Description automatically generated with medium confidence  
     
   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 או מתחת ל-. 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.
3. ***עבור household\_income:***  
   ממוצע: 2.18~  
   חציון: 0.7  
     
   ***עבור PCR\_02:***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:

1.   
     
   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.
2. לפי החישוב של ידידינו NumPy המתאם עבור הצמד שבחרנו הוא 0.018. זה לא סותר את הטענה שהבחירה שלנו טובה לחיזוי spread שכן הסיבה למתאם הנמוך היא שתכונות הצמד אינן פרידות ועל כן לא מקיימות קשר לינארי ברור. בכל זאת אין זה אומר שלא נוכל לחזות את spread לפיהן.
3. 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 . נעשה זאת בין m נקודות אימון לנק' המבחן היחידה שלנו ונקבל בסה"כ: .  
   argpartition: הפונקציה הזו עושה מיון חלקי של m האינדקסים ובכך מוצאת את האלמנטים הקרובים ביותר. בהנחה שהמיון אכן טוב וחסכוני, זה יתבצע ב-.copy: פה נעתיק k משתנים וזה כמובן יתבצע ב-.mean: זו למעשה מבצעת k פעולות חיבור ופעולת חילוק אחת ולכן מתבצעת ב-.  
   בסה"כ קיבלנו כי הפונקציה כולה תתבצע ב-.
4. A diagram of a graph

   Description automatically generated with medium confidence  
   דיוק האימון הוא 63% ודיוק הבדיקה הוא 57.2%.
5. A diagram of a map

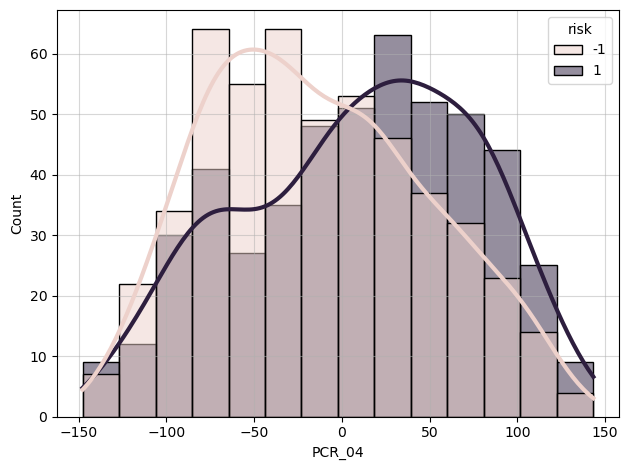
   Description automatically generated  
   דיוק האימון הוא 69.3% ודיוק הבדיקה הוא 66%.  
   כפי שאפשר לראות משרטוט שקיבלנו, הנורמליזציה מקנה התגבשות יותר טובה של אזורי ההחלטה (לעומת הקוים הספורדים שקיבלנו קודם). זה נובע מכך שהנורמליצזיה מבטלת את ההשפעה שיש להבדלי הטווחים בין התכונות על המודל.
6.   
   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.
7. 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:

1. תמונה שמכילה תרשים, עלילה, קו

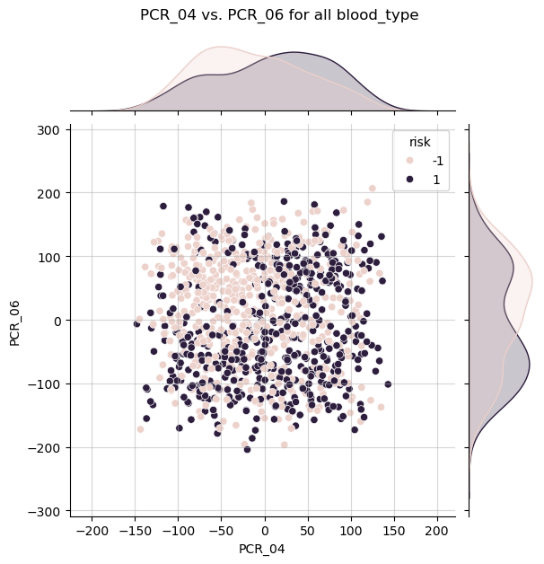
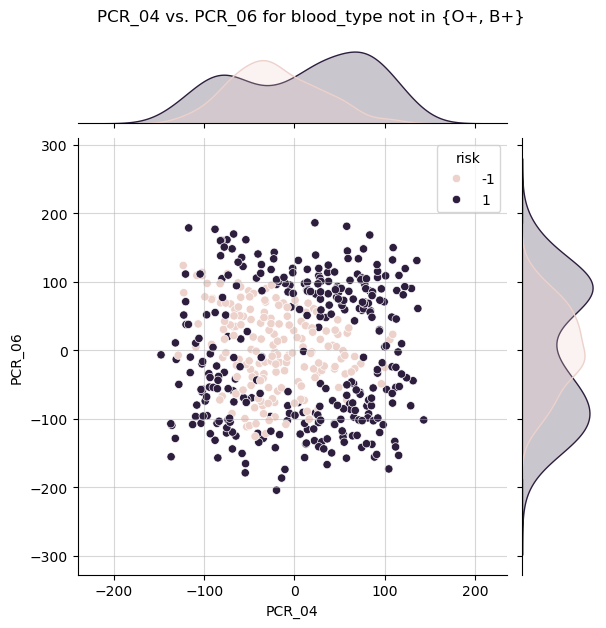
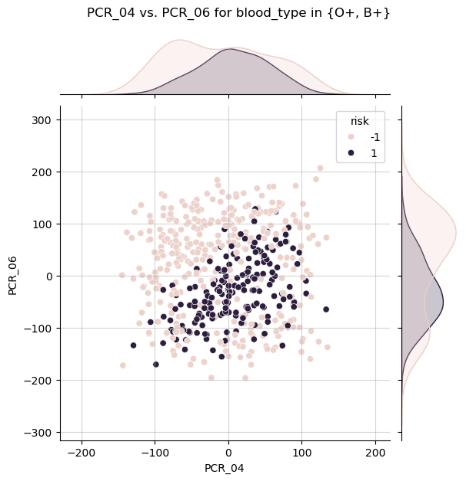
   התיאור נוצר באופן אוטומטי

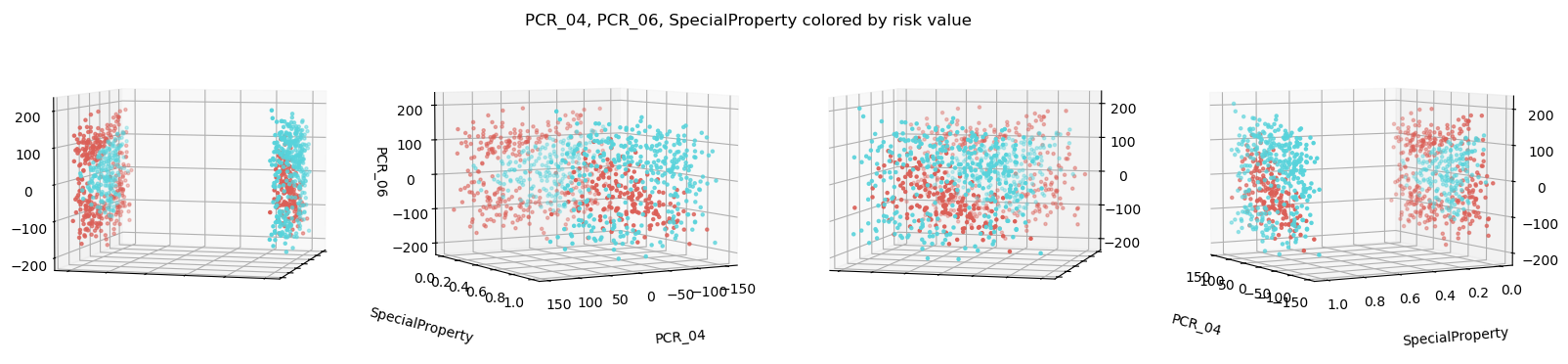
הפיצ'ר הכי אינפורמטיבי (חוץ מהפיצ'רים PCR\_08 ו-PCR\_01 משאלה 8) הוא הפיצ'ר weight מכיוון שכפי שניתן לראות בגרף המצורף, הפילוגים של פיצ'ר זה לתיוג spread חיובי או שלילי, שונה במובן הזה שבערכים בהם הסבירות הגבוהה ביותר לתיוג חיובי (בפיק של הפילוג הכהה), מתקבל סבירות נמוכה יותר לתיוג שלילי ולהפך, ולכן על סמך פילוגים אלו ניתן להפריד את הנתונים במידה.

1. 

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.

1. 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.



1. 
2. 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.
3. עץ החלטה בעומק 30 יתאים את קבוצה 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.
4. 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.