מבוא ללמידת מכונה – רטוב 1 report

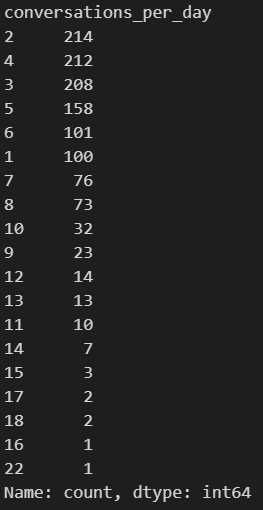
שקד עדי 327707709 אורי כשר חיטין

11.5.25

**חלק 1 – העלאת דאטה:**

**שאלה 1:**

ב – dataset יש 1250 שורות ו - 24 עמודות.

**שאלה 2:**

התכונה מתארת את **מספר האנשים** שקיימו **מספר מסוים של שיחות ביום** (ממוין בסדר יורד לפי מספר האנשים).

התכונה "מספר שיחות ביום" היא משתנה אורדינלי כי הערכים (מספרים שלמים) מייצגים קטגוריות עם סדר טבעי בעל משמעות — שתי שיחות הן יותר מאחת ופחות משלוש, גם אם לא מתייחסים להבדלים ביניהן בצורה רציפה.

**שאלה 3:**

|  |  |  |
| --- | --- | --- |
| **שם התכונה** | **תיאור** | **סוג** |
| תעודת זהות | מזהה ייחודי של הנבדק | אחר |
| גיל | גיל הנבדק | בדיד |
| מגדר | מין הנבדק (זכר / נקבה) | קטגורי |
| משקל | משקל הנבדק | רציף |
| סוג דם | סוג הדם של הנבדק | קטגורי |
| מספר אחים | מספר האחים של הנבדק | רציף |
| מדד אושר | מדד האושר של הנבדק | בדיד |
| שיחות ביום | מספר השיחות שביצע הנבדק ביום | אורדינלי |
| רמות סוכר | רמת הסוכר של הנבדק | בדיד |
| רמת פעילות גופנית | רמת עיסוק בפעילות ספורטיבית | אורדינלי |
| תאריך הבדיקה | תאריך ביצוע בדיקת ה - PCR | אחר |
| הכנסה | הכנסות חודשיות של משק הבית | רציף |
| PCR 01 | תוצאה של בדיקת PCR 01 | רציף |
| PCR 02 | תוצאה של בדיקת PCR 02 | רציף |
| PCR 03 | תוצאה של בדיקת PCR 03 | רציף |
| PCR 04 | תוצאה של בדיקת PCR 04 | רציף |
| PCR 05 | תוצאה של בדיקת PCR 05 | רציף |
| PCR 06 | תוצאה של בדיקת PCR 06 | רציף |
| PCR 07 | תוצאה של בדיקת PCR 07 | רציף |
| PCR 08 | תוצאה של בדיקת PCR 08 | רציף |
| PCR 09 | תוצאה של בדיקת PCR 09 | רציף |
| PCR 10 | תוצאה של בדיקת PCR 10 | רציף |

**שאלה 4:**

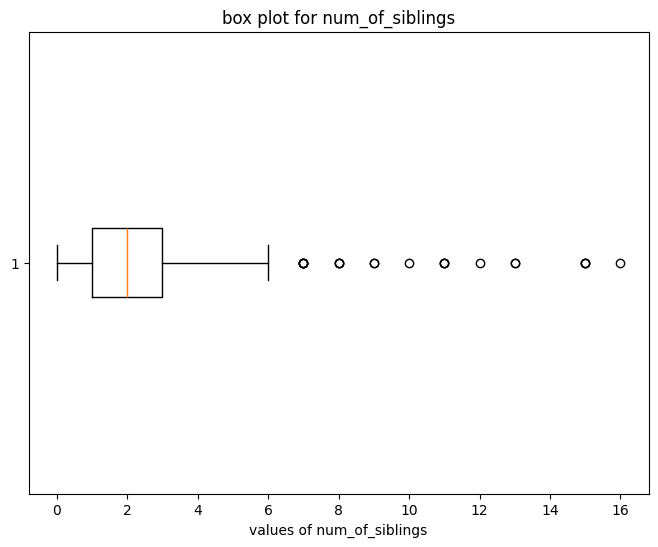
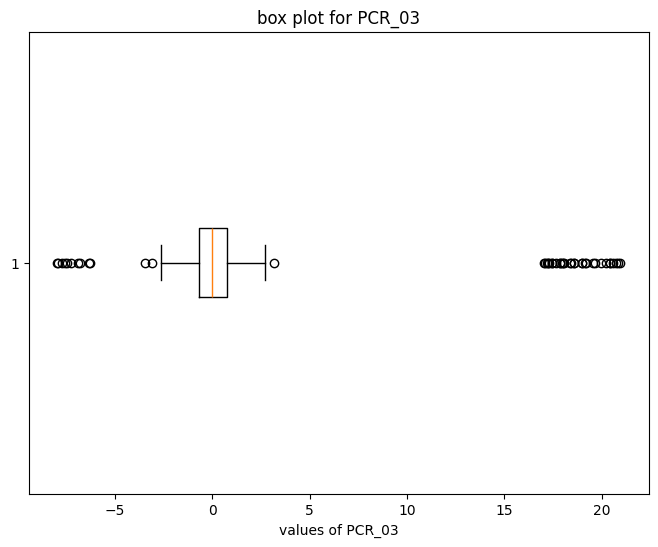
חשוב להשתמש בדיוק באותו פיצול של הנתונים בכל הניתוחים של התוצאות שלנו כדי שנוכל **להשוות בצורה הוגנת בין מודלים.**אם כל פעם נפצל בצורה שונה, ייתכן שהבדלים בביצועים של המודל יהיו בעקבות שינויים בדאטה ולא בגלל האלגוריתם של המודל עצמו. לכן, שימוש בפיצול קבוע מבטיח לנו עקביות והערכת תוצאות שמבוססת על **אותם נתונים בדיוק** בכל פעם.

**חלק 2 – ערכים חסרים וחריגים:**

**שאלה 5:**

|  |  |  |
| --- | --- | --- |
|  | **שם השדה** | **מספר הערכים החסרים** |
| **Training set** | מספר אחים | 108 |
|  | PCR 03 | 53 |
| **Test set** | מספר אחים | 27 |
|  | PCR 03 | 21 |

**שאלה 6:**



Outliners הם ערכים חריגים, כלומר נתונים שמסתמנים כחריגים או שונים בצורה משמעותית מהתבנית הכללית של שאר הנתונים. הם יכולים להטות את תוצאות הניתוח הסטטיסטי ולהקשות על הסקת מסקנות מדויקות.​

Box plot הוא כלי גרפי המייצג את התפלגות הנתונים באמצעות חמש סטטיסטיקות סיכום: מינימום, רבעון ראשון (Q1), חציון (Q2), רבעון שלישי (Q3) ומקסימום. התרשים מחולק לארבעה חלקים, כל אחד מייצג 25% מהנתונים.

התרשים מזהה ערכים חריגים באופן הבא:

1. נמיין את הנתונים בסדר עולה.
2. נחשב רבעון ראשון (Q1), חציון (Q2) ורבעון שלישי (Q3).
3. נקבע את טווח הרבעונים (IQR) על ידי החישוב IQR = Q3 – Q1.
4. נחשב את הגבול העליון והתחתון לערכים חריגים:

גבול תחתון: Q1 – 1.5 \* IQR.

גבול עליון: Q3 + 1.5 \* IQR.

כל נתון שנמצא מחוץ לטווח בין הגבול העליון והתחתון יסומן כערך חריג.

עבור התמונות המצורפות לעיל:

בשדה מספר האחים ישנם ערכים חריגים בערכים {7 , 8 , … , 13 , 15 , 16}.

ובשדה PCR 03 ישנם ערכים חריגים בטווחים [-9 , -6] , [17 , 22] ובאזור הערך 3-.

**שאלה 7:**

|  |  |  |
| --- | --- | --- |
|  | **מספר האחים** | **PCR 03** |
| **ממוצע** | 1.9417 | 0.5212 |
| **חציון** | 2 | 0.007- |

נשים לב כי עבור מספר האחים הממוצע קרוב לחציון מה שמעיד על פיזור נתונים סימטרי יחסית סביב הממוצע. אין הבדל משמעותי בין הממוצע והחציון וככל הנראה הנתונים לא נתונים להשפעה מרובה של ערכים חריגים.

לעומת זאת, בשדה PCR 03 החציון הוא שלילי בעוד שהממוצע חיובי, כלומר יש ביניהם הבדל משמעותי מה שיכול להצביע על כך שקיימים ערכים חריגים חיוביים גדולים מאוד שגורמים להטיית הממוצע כלפי מעלה (כי הממוצע חיובי).

במקרה שלנו, נעדיף להשתמש בחציון מכיוון שהוא פחות רגיש לערכים קיצוניים. הראנו כי קיים הבדל משמעות בין החציון לממוצע עבור השדה PCR 03, כלומר אם נבחר בממוצע ייתכן שנגדיל את כמות הערכים החריגים במדגם ולכן הבחירה הטובה יותר למילוי הערכים החסרים תהיה החציון.

**חלק 3 – חימום:**

**שאלה 8:**

תשובה

**שאלה 9:**

תשובה

**שאלה 10:**

ננתח את סיבוכיות הזמן של פונקציית החיזוי עבור נקודת בדיקה אחת:

1. distance.cdist( ): מחשב מרחקים בין נקודת בדיקה אחת לבין כל m נקודות האימון. כל חישוב מרחק אוקלידי דורש O(d) לכן בסך הכל נקבל O(md).
2. argpartition( ): בוחר את מיקומי k הקטנים ביותר מתוך וקטור באורך m בסיבוכיות O(m).
3. mean( ) , sign( ): החישובים מתבצעים על קלט בגודל k (k השכנים הקרובים ביותר) ולכן מתבצעים בסיבוכיות O(k).

**שאלה 11:**

\*תמונה\*

**שאלה 12:**

\*תמונה\*

נסביר למה נרמול חשוב ב – nearest neighbor models:

במודלים כמו kNN החיזוי מתבצע על בסיס מדידת מרחקים בין הנקודה הנבחרת לבין נקודות האימון. כאשר התכונות של הנתונים אינן נמצאות באותו סדר גודל (למשל אם תכונה אחת בין 0-1 ותכונה אחרת בין 0-1,000,000) התכונה בעלת הטווח הרחב יותר תשתלט על חישובי המרחק.

לכן, נרמול הוא קריטי לאיזון של התכונות כך שכולן ישפיעו באופן דומה על מדידת המרחק. הדבר משפר את הדיוק, איכות החיזוי ותורם בקבלת החלטות.

**שאלה 13:**

\*תמונה\*  
נסביר את המשמעות של בחירת הערך k:

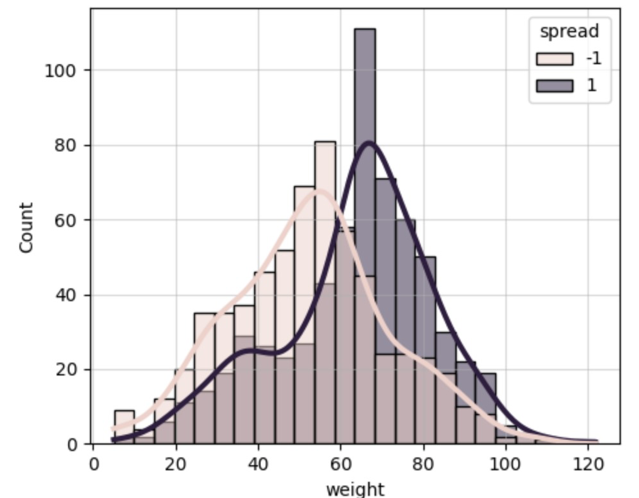
הפרמטר k קובע את מספר השכנים המשתתפים בהחלטת הסיווג של נקודה חדשה. כאשר ערך k קטן (למשל k = 1) המודל עלול להיות רגיש מאוד לנקודות חריגות, מה שעלול להוביל ל – overfitting.

לעומת זאת כאשר ערך k גדול יותר (למשל k = 5) ההחלטה מתבססת על ממוצע של כמה שכנים. כתוצאה מכך, המודל פחות רגיש לרעש, מתבצעת הכללה טובה יותר על הנתונים ונקבל אזורי החלטה חלקים יותר.

**שאלה 14:**

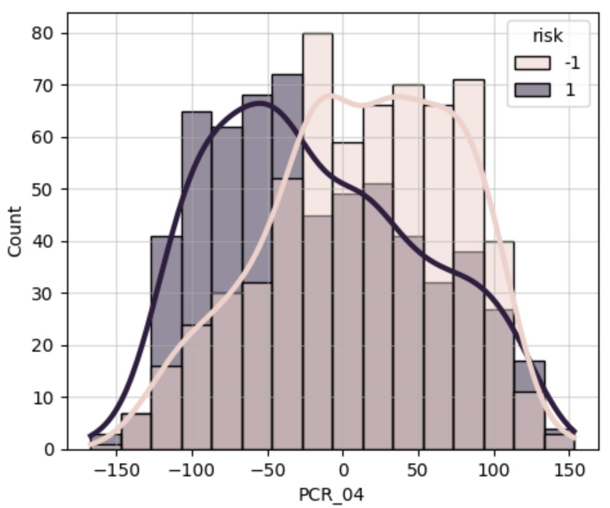
נרמול של שתי התכונות לפי min-max scaling לטווח [-1, 1] הוא רעיון רע כאשר אחת מהתכונות מתפלגת χ2 (𝑘 = 2). מהגרף המתאר את ההתפלגות ניתן לראות כי רוב הדגימות מרוכזות סביב ערכים קטנים, כלומר ההתפלגות אינה סימטרית. לכן אם נבצע נרמול כזה, הערך החריג הגבוה ביותר בעצם יקבע את פריסת הנתונים. התוצאה תהיה שרוב הדגימות (הקטנות) יידחסו לטווח מאוד צר וההבדלים ביניהן כמעט ייעלמו. במצב כזה מודל תלוי במרחק כמו kNN לא יוכל להבחין בצורה מדויקת בין דגימות שונות והתכונה מאבדת מחשיבותה.

**חלק 4 – חקר הדאטה:**

**שאלה 15:**

בחרנו את המשתנה weight.

זוהי התכונה הכי אינפורמטיבית מכיוון שבגרף המצורף ניתן לראות כי ההתפלגויות של שתי הקבוצות מופרדות בצורה הטובה ביותר. ניתן לומר כי הקבוצה עבורה spread = 1 מרוכזת במשקלים גבוהים יותר בעוד שהקבוצה spread = -1 מרוכזת במשקלים נמוכים יותר ולכן ניתן לבצע סיווג יותר בקלות.

**שאלה 16:**

בחרנו את המשתנה PCD\_04.

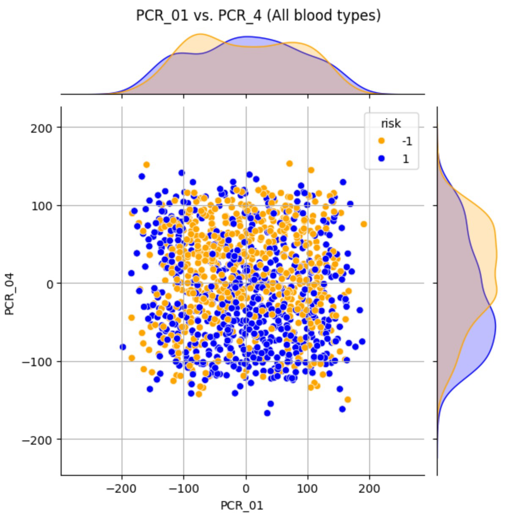
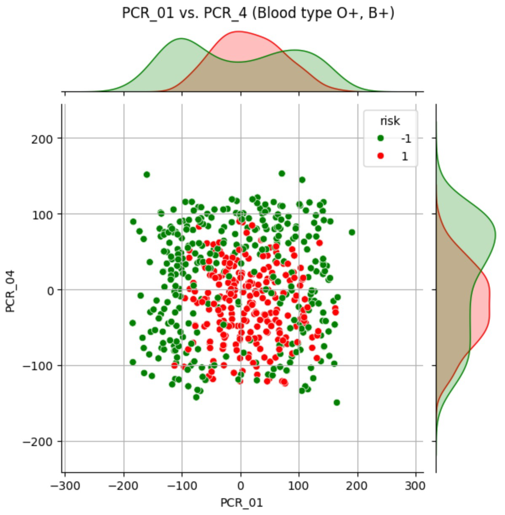
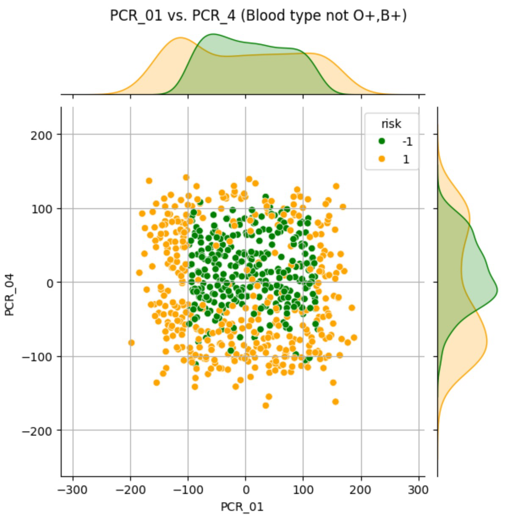
זוהי התכונה הכי אינפורמטיבית מכיוון שבגרף המצורף ניתן לראות כי ההתפלגויות של שתי הקבוצות מופרדות בצורה הטובה ביותר. ניתן לומר כי הקבוצה עבורה risk = 1 מרוכזת בתוצאות גבוהות יותר בעוד שהקבוצה risk = -1 מרוכזת בתוצאות נמוכות יותר ולכן ניתן לבצע סיווג יותר בקלות.

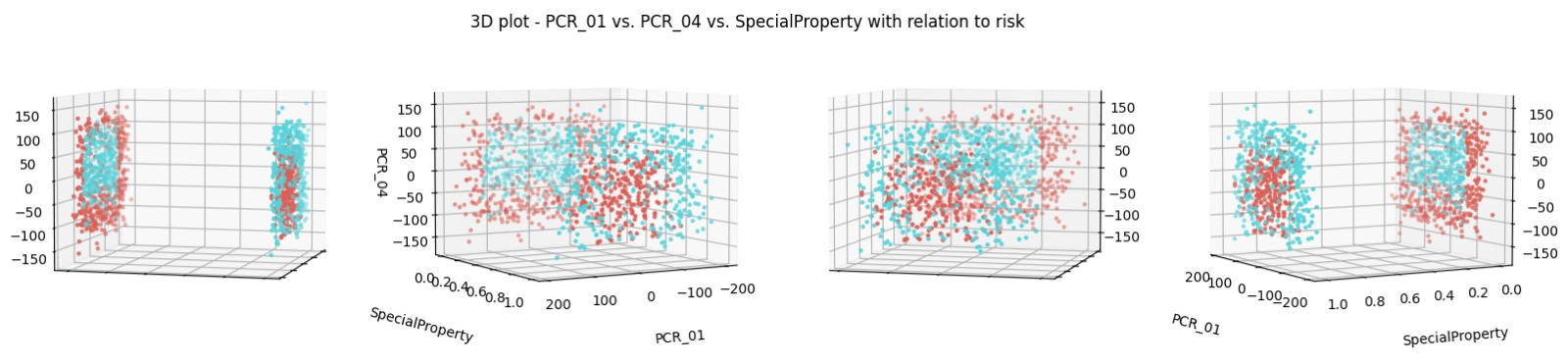
**שאלה 17:**

בחרנו את המשתנים PCR\_01, PCR\_04.

הסיבה לכך היא שבגרף ניתן לראות כי הם מציגים את ההפרדה הברורה ביותר בין הקבוצות לפי SpecialProperty כלומר, הפיזור של הדגימות משתי הקבוצות שונה לאורך רכיבים אלו. הפרדה כזו מצביעה על כך שרכיבים אלו בעלי מידע רלוונטי לגבי הפרמטר risk ולכן עשויים להיות שימושיים בחיזוי.

**שאלה 18:**



**שאלה 19:**

**שאלה 20:**

עץ החלטה בעומק מקסימלי 3 מייצר לכל היותר 8 עלים (2³), כלומר הוא מחלק את המרחב ל - 8 אזורים שונים. כאשר מסתכלים על הגרפים (משאלה 19), ניתן לראות שהנתונים אינם מופרדים בצורה חד משמעית לאורך הצירים PCR\_01 ו - PCR\_04, אך כן קיימת מגמה כללית - ישנן יותר תצפיות בצבע מסוים במרכז ומהצבע הנוסף מסביב למרכז. לכן, עץ בעומק מרבי 3 יוכל לתפוס מגמות כלליות בנתוני האימון אך לא יספיק כדי לבצע התאמה מדויקת בין הקבוצות.

**שאלה 21:**

עץ החלטה בעומק מקסימלי 30 יוכל לבצע כמות עצומה של פיצולים כלומר, לחלק את המרחב להמון תתי אזורים קטנים. עומק כזה נותן לנו את האפשרות להתאים את הסיווג כמעט אחד לאחד לנתוני האימון, גם אם הפרדה בין הקבוצות תהיה לא פשוטה. עומק גבוה עשוי לסייע בזיהוי תבניות מורכבות בהתפלגות ולכן במונחי התאמה ל - training data מדובר בהתאמה כמעט מושלמת.

**שאלה 22:**

תשובה

**חלק 5 – נרמול של הדאטה:**

**שאלה 23:**

תשובה

**חלק 6 –:**

**שאלה 24:**

|  |  |  |  |
| --- | --- | --- | --- |
| שם התכונה | חדש | שיטת נרמול | שיטה למילוי ערכים חסרים |