מבוא למערכות לומדות (236756) | תרגיל בית 1 גדול

ליאל פרבר | 214413437

ראובן טימסיט | 330083858

4 ביולי 2024

שאלה 1

יש 1250 שורות ו- 25 עמודות ב- $dataset$.

שאלה 2

```
conversations_per_day
3     218
2     204
5     179
4     168
1     108
6     107
7      94
8      54
9      42
10     29
11     16
13      8
12      7
14      6
16      5
15      3
17      1
29      1
Name: count, dtype: int64
```

אנו חושבים שפיצ'ר זה מייצג את מספר השיחות הפרונטליות שהיו לאדם כלשהו ביום יחיד. נשים לב שיש סדר טבעי המוגדר על פיצ'ר זה (על קבוצת הערכים האפשריים שלו) והוא הסדר הרגיל של $\mathbb{N}$. בנוסף, פיצ'ר זה יכול לשמש כמשתנה קטגורי כי הוא לא רציף אלא בדיד. לכן, פיצ'ר זה הוא מטיפוס אורדינל.

שאלה 3

feature type	feature description	feature name
אחר	תעודת הזהות של מטופל (מזהה ייחודי)	patient_id
רציף	גיל של מטופל	age
קטגורי	מגדר של מטופל	sex
רציף	משקל של מטופל (בק"ג)	weight
קטגורי	סוג דם של מטופל	blood_type
רציף	מיקום נוכחי של מטופל בכדור הארץ	current_location
רציף	מספר האחים והאחיות של מטופל	num_of_siblings
אורדינל	רמת השמחה של מטופל	happiness_score
רציף	הכנסה בייתית של מטופל	household_income
אורדינל	מספר שיחות ביום של מטופל	conversations_per_day
רציף	רמת הסוכר (בדם) של מטופל	sugar_levels
אורדינל	כמות הפעילות הגופנית שמטופל מבצע בפרק זמן מסויים	sport_activity
קטגורי	תאריך ביצוע בדיקת pcr של מטופל	pcr_date
רציף	תוצאת בדיקת pcr מספר 1 שביצע מטופל	PCR_01
רציף	תוצאת בדיקת pcr מספר 2 שביצע מטופל	PCR_02
רציף	תוצאת בדיקת pcr מספר 3 שביצע מטופל	PCR_03
רציף	תוצאת בדיקת pcr מספר 4 שביצע מטופל	PCR_04
רציף	תוצאת בדיקת pcr מספר 5 שביצע מטופל	PCR_05
רציף	תוצאת בדיקת pcr מספר 6 שביצע מטופל	PCR_06
רציף	תוצאת בדיקת pcr מספר 7 שביצע מטופל	PCR_07
רציף	תוצאת בדיקת pcr מספר 8 שביצע מטופל	PCR_08
רציף	תוצאת בדיקת pcr מספר 9 שביצע מטופל	PCR_09
רציף	תוצאת בדיקת pcr מספר 10 שביצע מטופל	PCR_10

שאלה 4

אנו רוצים להשתמש באותו $random\_state$ בכל הפעלה של הקוד שלנו כדי להבטיח שכל התוצאות שאנו מקבלים עקביות עם קבוצות האימון והמבחן שלנו. אילו היינו מבצעים חלוקה רנדומלית לקבוצות אימון ומבחן בכל פעם שנזדקק לכך התוצאות שהיינו מקבלים באנליזות השונות שלנו היו בהתאם לחלוקה הנ"ל ולא בהכרח מתאימות אחת לשנייה. כלומר, התוצאות לא היו מאפשרות לנו להסיק מידע על המודל.

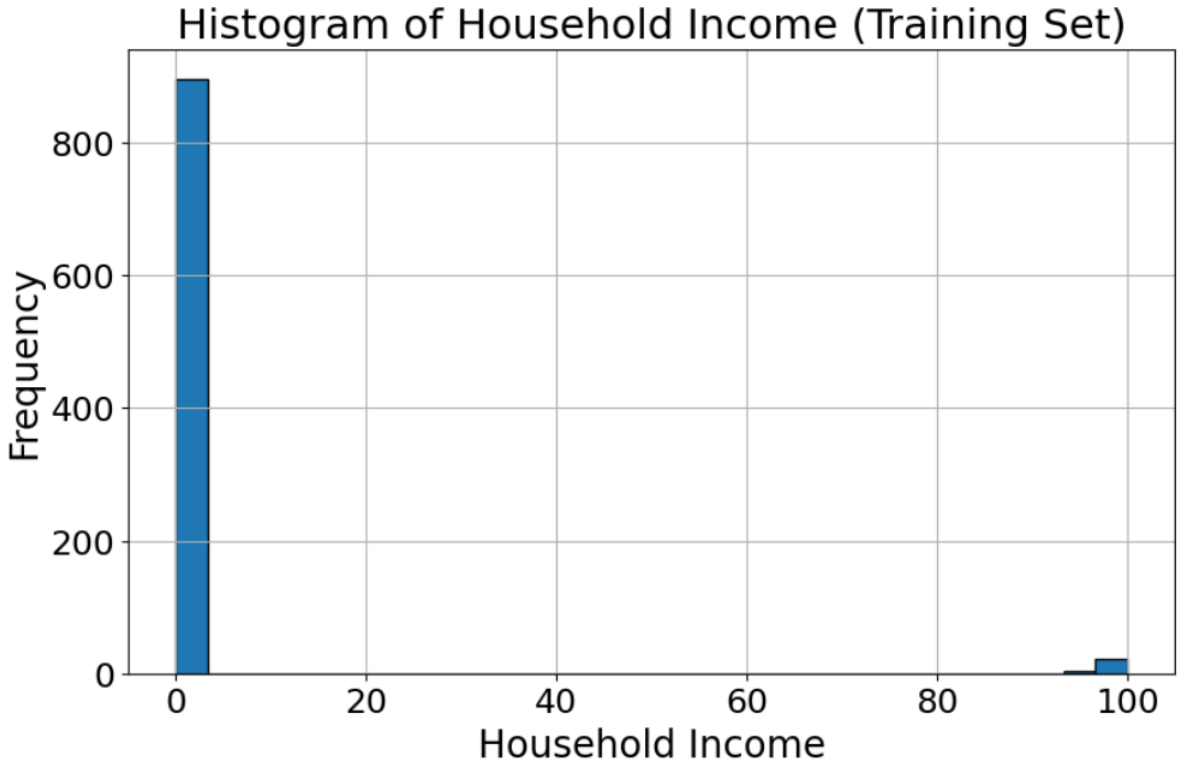
שאלה 5

```
Missing values in training set:
 household_income    80
dtype: int64

Missing values in test set:
 household_income    29
dtype: int64
```

יש 80 דוגמאות בקבוצת האימון עם שדה $household\_income$ ריק ויש 29 דוגמאות בקבוצת המבחן עם שדה $household\_income$ ריק.

שאלה 6



ניתן לזהות $outliers$ בצד ימין התחתון של ההיסטוגרמה - יש מספר נקודות ב- $data$ בעלות שדה $household\_income$ בסביבת 100, כאשר אצל רוב הנקודות ערך שדה זה הוא בסביבת 0.

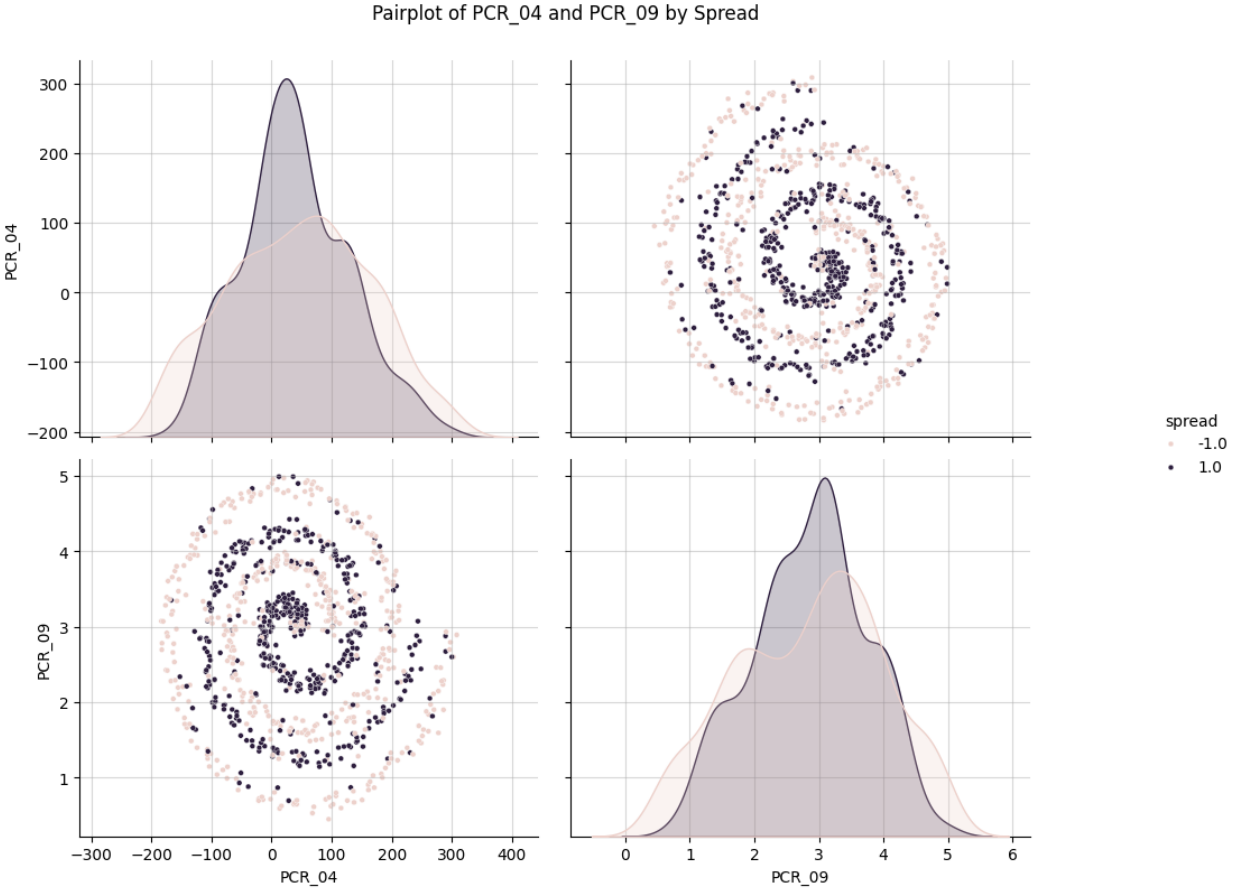
שאלה 7

```
Mean Household Income: 3.3169565217391304
Median Household Income: 0.7
```

ניתן לראות כי יש הבדל משמעותי בין ממוצע ההכנסות לבית של מטופל לחציון שלו. ה- $outliers$ בסעיף הקודם גורמים להטייה בממוצע כך שלא יהיה דומה לחציון. אבל כמות ה- $outliers$ קטנה מאוד ביחס לכמות הנקודות ב- $data$ עם שדה $household\_income$ בסביבת 0. לכן, החציון אינו מושפע כמעט מה- $outliers$ הללו. במקרה זה נעדיף להחליף את שדות ה- $household\_income$ הריקים בחציון שמייצג באופן איכותי יותר את ממוצע השדה $household\_income$ בקרב מטופלים.

שאלה 8

זוג הפיצ'רים השימושי ביותר לחיזוי פיצ'ר ה- $spread$ הוא (PCR_4, PCR_9). מצ"ב ה- $seaborm.pairplot$ של זוג זה:



ניתן לראות כי יש מצבורים שונים של נקודות עם $spread = 1.0$ ועם $spread = -1.0$ שהם נפרדים. כלומר, קל לסווג איזורים בגרף השמאלי התחתון לכאלו שבהם הנקודות הן עם $spread$ כ- 1.0 או כ- -1.0 וזאת כי יש בגרף תבנית ברורה של סיווגי נקודות (ספירלה). אומנם יש נקודות רועשות שנטעה בחיזוי ה- $spread$ שלהן כי הן בעלות סיווג הפוך מהסיווג של כל הנקודות בסביבתן אבל הן מעטות. עבור כל זוג פיצ'רים אחר נתקשה לסווג איזורים בהם רוב הנקודות בעלות $spread$ זהה משום שפיזור הנקודות המתקבל לפי זוג הפיצ'רים יהיה "רנדומלי מידי" (לא תבניתי).

שאלה 9

```python
def predict(self, X):
    X_test = np.array(X)
    distances = cdist(X_test, self.X_train, metric='euclidean') # Compute distance between each pair in X_test x self.X_train

    # finding the indices of the smallest self.n_neighbors elements from each row of a distances 2D array
    n_neighbors_indices = np.argpartition(distances, self.n_neighbors, axis=1)[:, :self.n_neighbors]
    # for each set of indices in a row, retrieve the corresponding labels from the training set based on those indices
    n_nearest_neighbor_labels = self.Y_train[n_neighbors_indices]
    # for each set of indices in a row, generate a prediction label for the corresponding test set element based on the majority of these labels
    n_nearest_neighbor_labels_sum = np.sum(n_nearest_neighbor_labels, axis=1)
    predictions = np.where(n_nearest_neighbor_labels_sum >= 0, 1, -1)
    return predictions
```

כאשר מפעילים את $predict$ על נקודת מבחן יחידה ($X = [x]$):

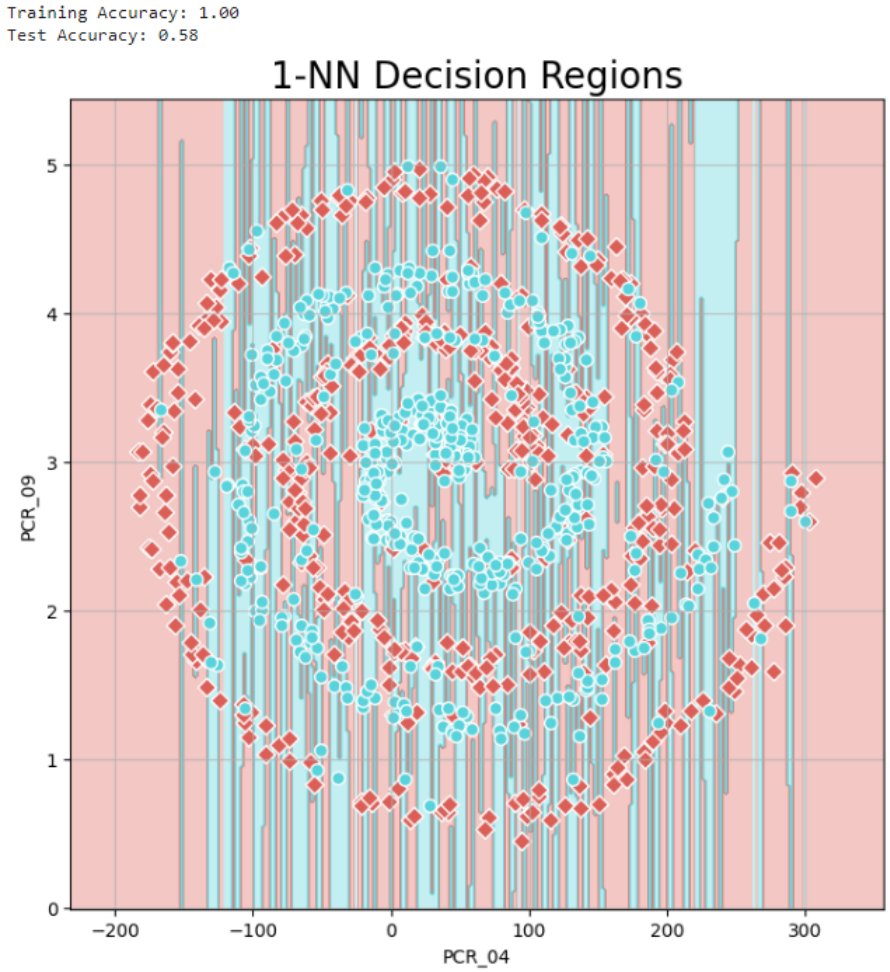
- יצירת מערך $numpy$ עם x לוקחת $O(d)$ זמן כי צריך להעתיק את $x \in values(feature_1) \times ... \times values(feature_d)$ (וקטור בגודל d) למערך.
- יצירת המטריצה $distances$ בגודל $1 \times m$ (מחזיקה מרחקים בין x ל- m דוגמאות האימון) לוקחת $O(m)$ זמן.
- מציאת $n\_nearest\_nighbor\_indices$ לוקחת $O(m \cdot log(m))$ זמן כי אנו מניחים ש- $self.n\_neighbors \leq m$ (כלומר, לא מתחשבים ביותר שכנים ממספר דוגמאות האימון) ולכן במקרה הגרוע נמיין את כל <u>השורה הראשונה</u> ב- $distances$ (בעלת m איברים), והמיון לוקח $O(m \cdot log(m))$.
- מציאת $n\_nearest\_nighbor\_sum$ לוקחת $O(m)$ זמן כי אנו סוכמים m איברים <u>בשורה הראשונה</u> ב- $distances$.
- מציאת $predictions$ לוקחת $O(1)$ כי מההסבר בנקודה הקודמת $n\_nearest\_nighbor\_sum$ הוא מערך בגודל 1 (וסיבוכיות $np.where$ היא כגודל המערך עליו היא פועלת).

סה"כ סיבוכיות $predict$ היא $O(d + m \cdot log(m))$.

הערה: אפשר לבצע את מציאת $n\_nearest\_nighbor\_indices$ ע"י מציאת האיבר הקטן ביותר הבא כל פעם מתוך m איברים. נעשה זאת d פעמים ונקבל סיבוכיות $O(md)$ עבור השלב הזה ובכללי עבור $predict$.

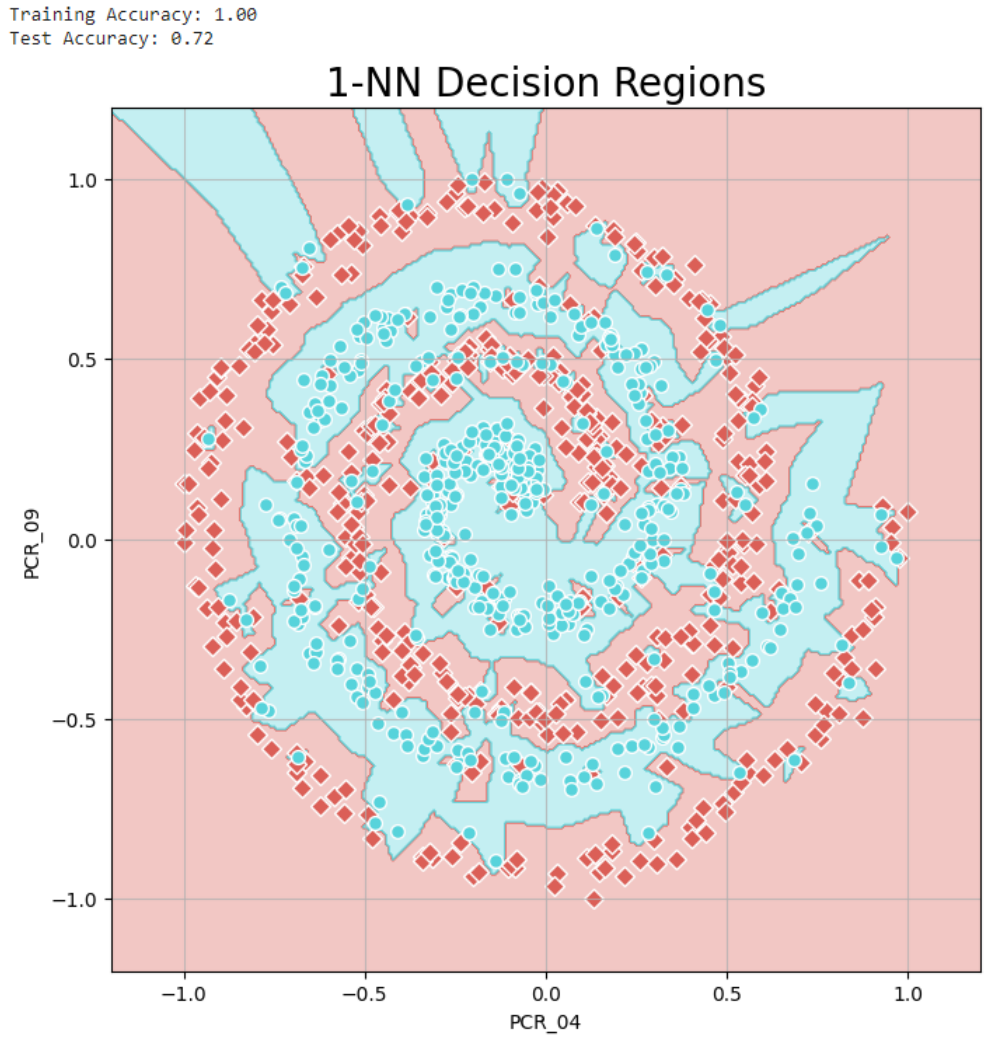
הסיבוכיות הסופית היא $O(min\{d + m \cdot log(m), md\})$.

שאלה 10



כפי שניתן לראות אחוזי הדיוק של דוגמאות האימון והמבחן הם 100%, 58% בהתאמה. (דוגמאות כחולות הן בעלות $spread = 1$ ואדומות בעלות $spread = -1$).

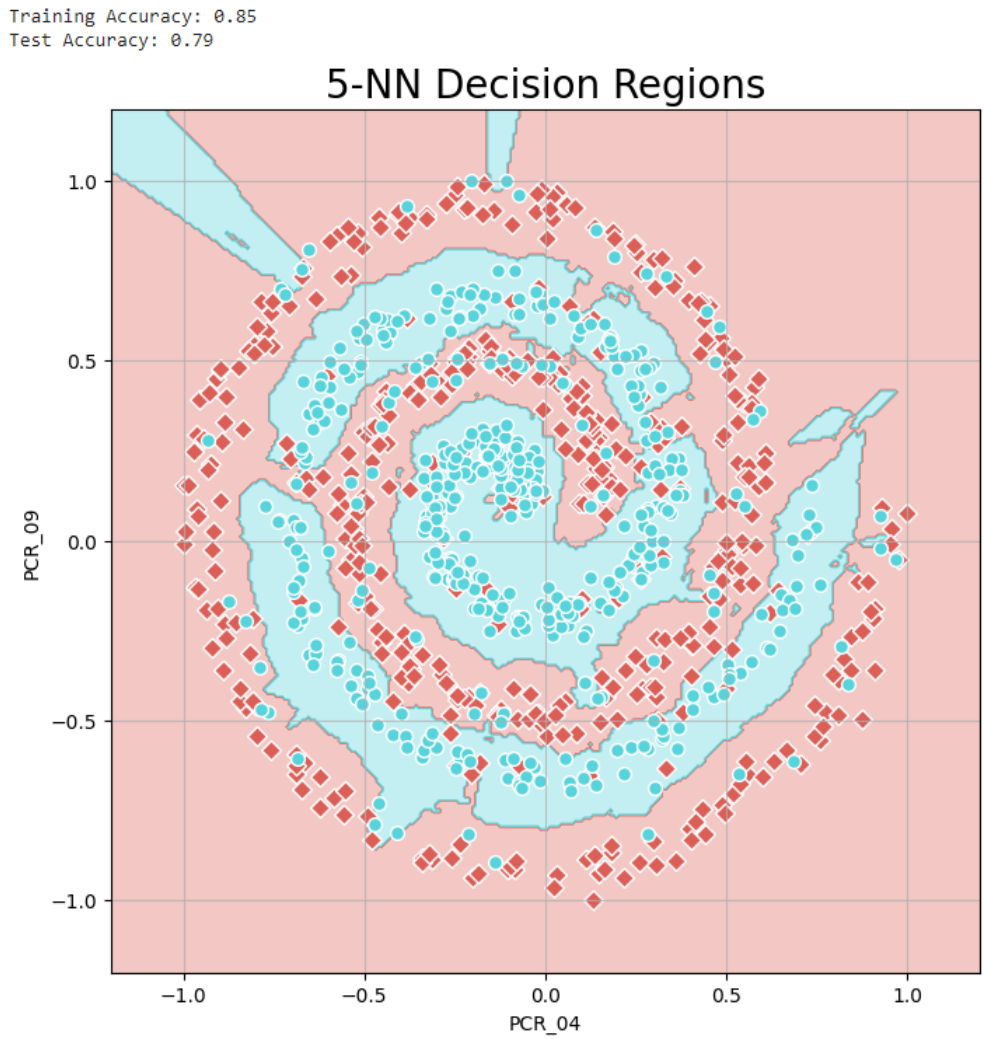
שאלה 11



כפי שניתן לראות אחוז הדיוק של דוגמאות האימון הוא 100% וזאת כי סיווג נקודה נקבע לפי סיווג נקודת האימון הקרובה ביותר אליה וברור כי נקודת האימון הכי קרובה לנקודת אימון x היא x. לכן, ללא תלות בנרמול הפיצ'רים כל נקודת אימון x תקבל את הסיווג של עצמה ולפיכך נקבל 100% דיוק בסיווג דוגמאות האימון.

בנוסף, אחוז הדיוק של דוגמאות המבחן הוא 72% - יותר גבוה מבשאלה 10. אנו מחשבים מרחקים בין נקודות ע"י מטריקה אוקלידית. לכן, פיצ'רים בעל טווח רחב משפיעים הרבה יותר על המרחק המתקבל מאשר כאלו עם טווח מצומצם. כתוצאה מכך, הם משפיעים יותר על סיווג נקודות ב- $data$. כאשר יש מעט מאוד פיצ'רים עם טווח רחב נקבל סיווגים מוטים מידי ולא מדויקים. אם מנרמלים את כל הפיצ'רים לאותו טווח אז כל פיצ'ר תורם תרומה באותו סדר גודל לחישוב המרחק. בדרך זו kNN מתחשב בכל הפיצ'רים באופן הוגן יותר ולא מעניק משקל גדול מידי לפיצ'ר מסויים עם טווח רחב. כתוצאה, נקבל דיוק גבוה יותר בסיווג עם נרמול פיצ'רים מאשר בסיווג ללא נרמול. בתמונה לעיל הנרמול של PCR_09, PCR_04 גורם להקטנת הטווח הגדול של PCR_09 לאותו טווח $[-1,1]$ של PCR_04.

שאלה 12



כפי שניתן לראות אחוז הדיוק של דוגמאות האימון הוא 85% וזאת כי סיווג נקודה כבר לא נקבע לפי סיווג נקודת האימון הקרובה ביותר אליה אלא לפי ה- 5 הכי קרובות (הסיווג הוא הסיווג הנפוץ ביותר ב- 5 הסיווגים של 5 הנקודות הנ"ל). ברור כי עתה יתכנו סיווגים שגויים עבור נקודות אימון מסויימות כי יש יותר $underfiting$ מאשר ב- $1-NN$ כי מתחשבים ביותר דוגמאות אימון בסיווג. כמו כן, העלאת ה- $underfiting$ גורמת לעליה באחוז הדיוק של דוגמאות המבחן ביחס לשאלה 11- 79%.

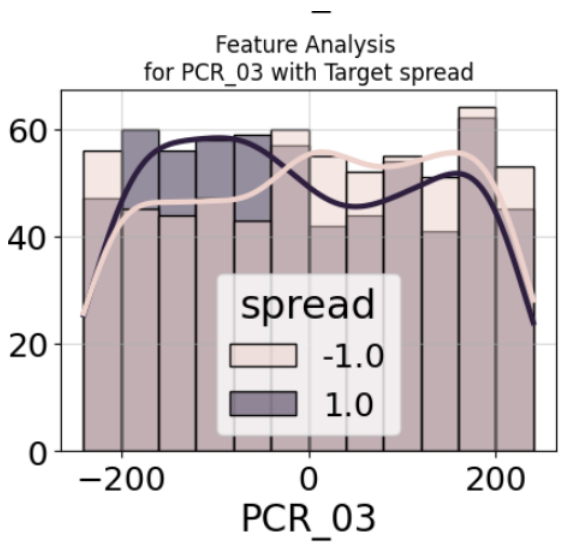
שאלה 13

נשים לב שהתפלגות אחידה מעל $[2,5]$ היא בעלת טווח $[2,5]$ והתפלגות כי בריבוע עם $k=2$ היא בעלת טווח $[0,\infty)$. נרמול 2 הפיצ'רים הנדגמים מ- 2 ההתפלגויות הנ"ל (בהתאמה) לאותו טווח $[-1,1]$ יצטרך להיות "חזק" (גדול מספיק) עקב ערכי ההתפלגות כי בריבוע הגדולים (נצפה שיהיו כאלה אם ניקח קבוצת דוגמאות מספיק גדולה). בהתאם, הוא יהיה חזק מידי עבור ערכי ההתפלגות האחידה ויגרום להם להיות זהים (כלומר ינטרל את היכולת לסווג לפי הפיצ'ר הראשון כי לכל הדוגמאות יש אותו ערך עבורו). עדיף

יהיה לנרמל את ערכי ההתפלגות כי בריבוע לפי נרמול סטנדרטי.

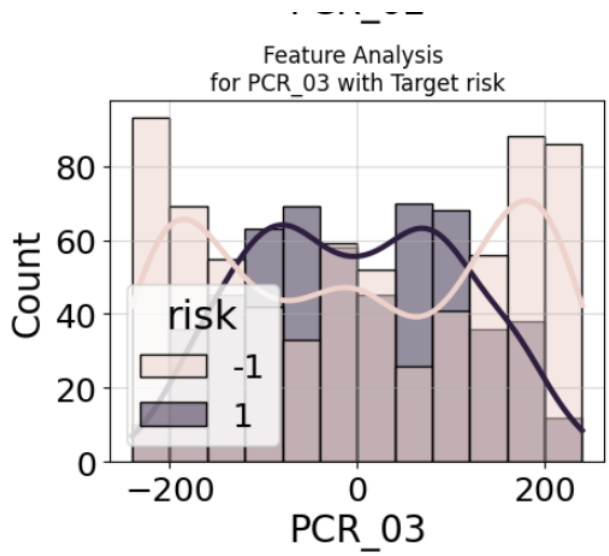
שאלה 14

נבחר את הפיצ'ר PCR_03. נשים לב שעבור ערכי PCR_03 "ממוצעים" ומעלה או קטנים ממש רוב הדוגמאות מסווגות כבעלות $spread = -1$. עבור ערכי PCR_03 קטנים עד "ממוצעים" רוב הדוגמאות מסווגות כבעלות $spread = 1$. לכן, הפיצ'ר הנ"ל נותן לנו אינפורמציה טובה יחסית לגבי החיזוי של $spread$.



שאלה 15

נבחר את הפיצ'ר PCR_03. נשים לב שעבור ערכי PCR_03 קטנים/גדולים רוב הדוגמאות מסווגות כבעלות $spread = -1$. עבור ערכי PCR_03 "ממוצעים" רוב הדוגמאות מסווגות כבעלות $spread = 1$. לכן, הפיצ'ר הנ"ל נותן לנו אינפורמציה טובה יחסית לגבי החיזוי של $spread$.



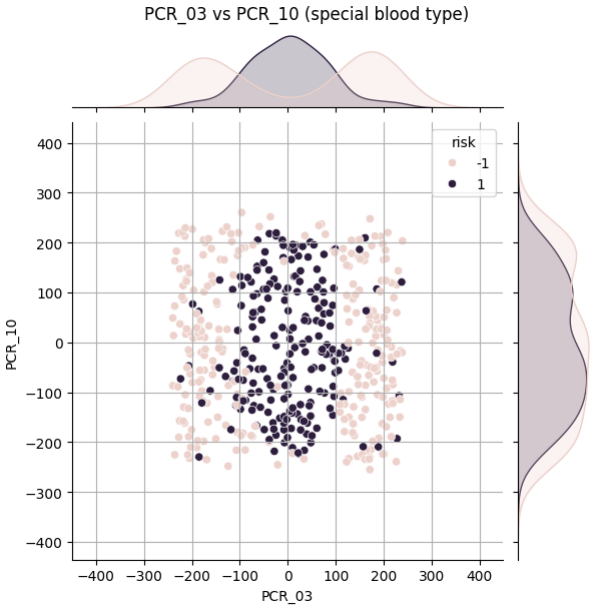
שאלה 16

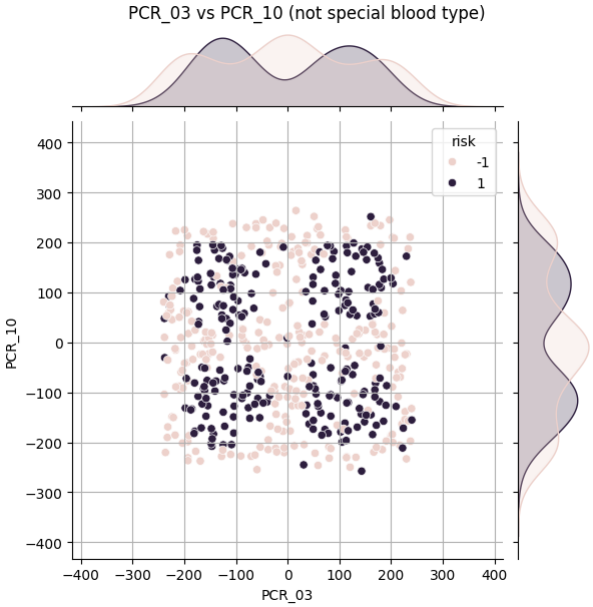
זוג הפיצ'רים השימושי ביותר לחיזוי פיצ'ר ה- $risk$ הוא (PCR_3, PCR_10).

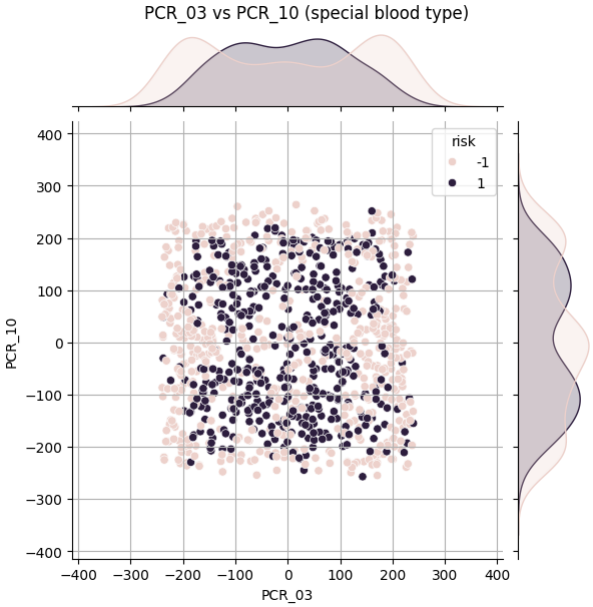
נסתכל על התרשימים המתקבלים במחברת פייתון שלנו לאחר חלוקת קבוצת האימון ל-2 לפי $SpecialProperty$. מתקיים ב-2 הקבוצות החדשות שהגרפים של הפיצ'רים PCR_3, PCR_10 מכילים מצבורים שונים של נקודות עם $risk = 1.0$ ועם $risk = -1.0$ שהם נפרדים. כלומר, קל לסווג איזורים נפרדים במרחב לפי המצבורים השונים בגרפים הללו (בכל אחד מהאיזורים כל הנקודות יהיו עם ערך $risk$ מסויים בהתאם לערך של רוב הנקודות באותו איזור בגרף).

הגרף של כל זוג פיצ'רים אחר יותר מידי "רנדומלי" ויש ערבוב בין נקודות עם ערכי $risk$ שונים (כך שסיווג כל ה- $data$ לפי כל זוג כזה יהיה פחות מדויק מאשר סיווג לפי PCR_3, PCR_10).

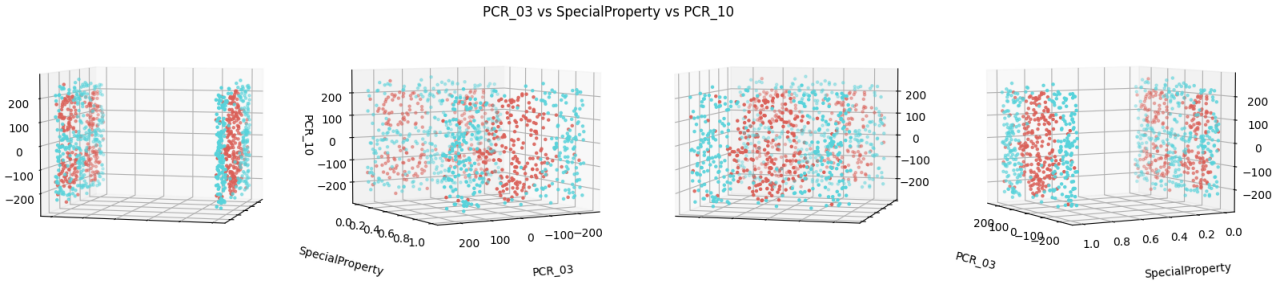
שאלה 17


PCR_03 vs PCR_10 (special blood type)
risk
-1
1
PCR_10
PCR_03
400
300
200
100
0
−100
−200
−300
−400


PCR_03 vs PCR_10 (not special blood type)
risk
-1
1
PCR_10
PCR_03
400
300
200
100
0
−100
−200
−300
−400


PCR_03 vs PCR_10 (special blood type)
risk
-1
1
PCR_10
PCR_03
400
300
200
100
0
−100
−200
−300
−400
−400
−300
−200
−100
0
100
200
300
400

שאלה 18

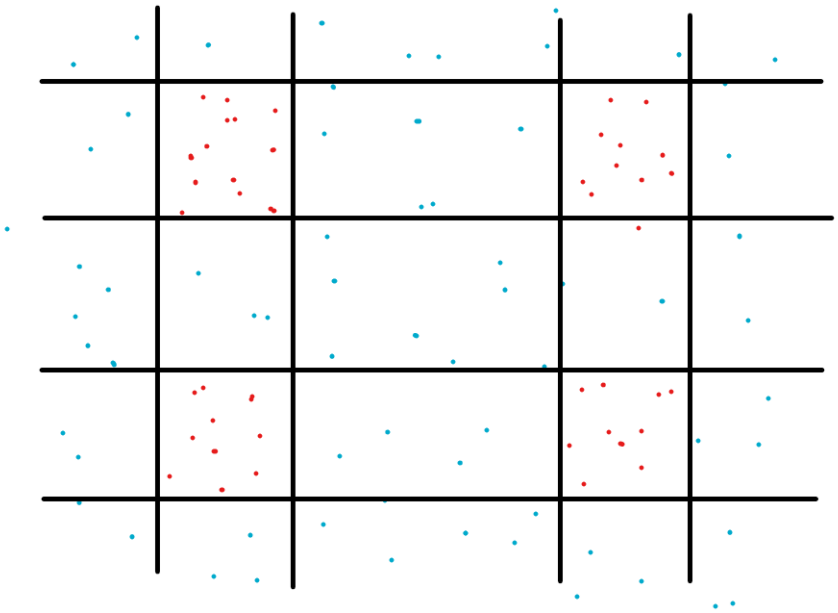


שאלה 19

נשים לב כי ה- $data$ שאנו צריכים לסווג מוצג בתמונות בשאלה 18.

מודל של עץ החלטה עם עומק מקסימלי 3 אינו מתאים לסיווג קבוצת האימון משום שהוא מאפשר סיווג מקסימלי לפי 3 מסווגים לינארים שונים (כי בקורס אנו עובדים עם עצי החלטה בינאריים בלבד כך שכל צומת בעץ יכול לשמש כמסווג לינארי אחר של ה- $data$, ביחס לפיצ'ר מסויים):

נשים לב לפי התמונה הימנית ביותר בשאלה 18 כי כל דוגמאות האימון עם $SpecialProperty = 0$ מסודרות במעין 4 ריבועים אדומים מוקפים בשוליים כחולים (עד כדי דוגמאות רועשות). כדי לסווג נכון את דוגמאות אלו אנו צריכים 8 מסווגים לינארים (8 הקווים השחורים בתמונה הבאה):



כך שכל ריבוע עם נקודות אדומות יסווג כאיזור עם $risk = 0$ (נניח שנקודות אדומות מייצגות תווית $risk = 0$) ויתר המלבנים יסווגו כאיזורים עם $risk = 1$.

לפיכך, עץ החלטה כנ"ל לא מספיק עבור סיווג מושלם (ואפילו לא עבור סיווג איכותי עם אחוז דיוק גבוה) של דוגמאות האימון.

שאלה 20

מודל של עץ החלטה עם עומק מקסימלי 30 מתאים לסיווג קבוצת האימון משום שהוא מאפשר סיווג מקסימלי לפי 30 מסווגים לינארים שונים:

נשים לב לפי התמונה הימנית ביותר בשאלה 18 כי את כל דוגמאות האימון עם $SpecialProperty = 1$ אפשר לסווג עם כ- 2 מסווגים לינארים (עד כדי דוגמאות רועשות שיסווגו לא נכון). את דוגמאות האימון עם $SpecialProperty = 0$ אפשר לסווג עם כ- 8 מסווגים לינארים (עד כדי דוגמאות רועשות שיסווגו לא נכון). לכן, בעזרת 30 מסווגים לינארים אפשר להשיג סיווג מדויק מאוד (ואפילו $overfitting$) של קבוצת האימון.

שאלה 21

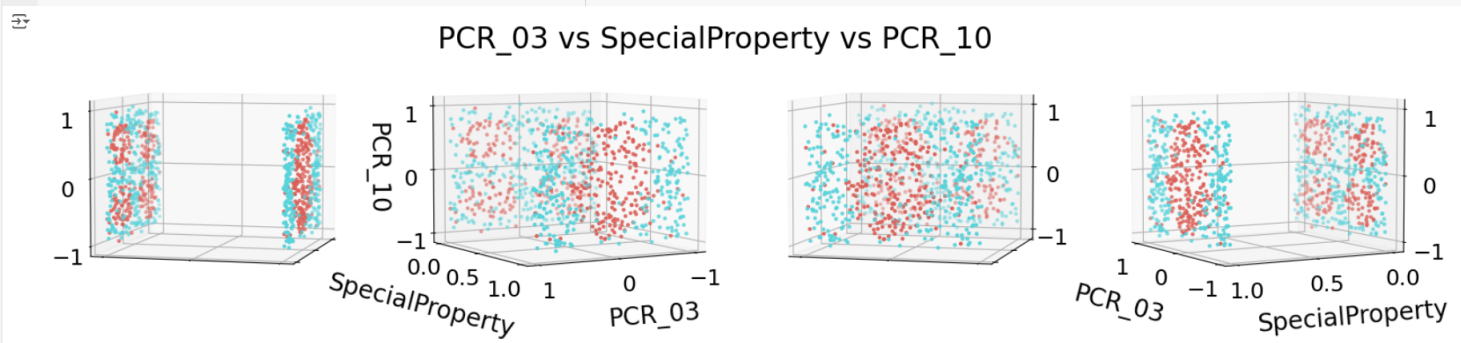
מודל של $1 - NN$ לא מתאים לסיווג קבוצת האימון משום שיהיו הרבה דוגמאות אימון שיסווגו לא נכון:

נסתכל על כל דוגמאות האימון עם $SpecialProperty = 1$ מרחקן מנקודות אימון עם $SpecialProperty = 0$ הוא יחידות בודדות. לעומת זאת, מרחקן מנקודות אימון אחרות עם $SpecialProperty = 1$ הוא עשרות רבות של יחידות (לפי ה- $scale$ של PCR_03 בתמונה הימנית בשאלה 18). אם כן, $1-NN$ יסווג כל דוגמה עם $SpecialProperty = 1$ בהתאם לצבע נקודה שנמצאת "בערך מולה" במישור $SpecialProperty = 0$ (באופן פורמלי- שנמצאת בסביבת נקודת החיתוך של ישר המאונך ל- 2 המישורים עם המישור $SpecialProperty = 0$). כלומר, ההטלות של כל אחד מהמישורים $SpecialProperty = 1, SpecialProperty = 0$ על השני קובעות את סיווג הנקודות במישורים. כמו כן, האיזורים האדומים במישור $SpecialProperty = 0$ יגרמו לסיווג לא נכון של חלק מהנקודות הכחולות ב- $SpecialProperty = 1$.

שאלה 22

נבדוק את השפעת הנורמליזציה של $PCR\_01 - PCR\_10$ (נוסיף קוד שמדפיס גרף $3D$ חדש כמו בשאלה 18):

```
plot3d(trainingSet, "PCR_03", "SpecialProperty", "PCR_10", title="PCR_03 vs SpecialProperty vs PCR_10", hue="risk", s=5)
```



כפי שניתן לראות:

- הנורמליזציה לא משפיעה על העובדה שעץ החלטה בעומק מקסימלי 3 לא מתאים לסיווג דוגמאות האימון שכן עתה הן מסודרות באותו אופן כמו בקבוצת האימון עם פיצ'רים לא מנורמלים.
- הנורמליזציה לא משפיעה על העובדה שעץ החלטה בעומק מקסימלי 30 מתאים לסיווג דוגמאות האימון שכן עתה הן מסודרות באותו אופן כמו בקבוצת האימון עם פיצ'רים לא מנורמלים.
- הנורמליזציה משפיעה על העובדה שמודל $1 - NN$ לא מתאים לסיווג דוגמאות האימון שכן עתה הוא כן מתאים לסיווג. זאת משום שכעת נקודת האימון y הקרובה ביותר לכל נקודת אימון x מקיימת $x.SpecialProperty = y.SpecialProperty$ ולכן רוב מוחלט של נקודת באיזור כחול/אדום יקבלו סיווג כחול/אדום בהתאמה (עד כדי נקודות בתפר בין אזורים צבועים בצבע שונה, אשר חלקן יסווגו באופן שגוי).

שאלה 23

$Normalization\ method$	New	$Keep$	$Feature\ name$
	X	V	$patient\_id$
	X	V	age
	X	V	sex
	X	V	$weight$
	X	X	$blood\_type$
	X	V	$current\_location$
	X	V	$num\_of\_siblings$
	X	V	$happiness\_score$
	X	V	$household\_income$
	X	V	$conversations\_per\_day$
	X	V	$sugar\_levels$
	X	V	$sport\_activity$
	X	V	pcr_date
$Standartization$	X	V	PCR_01
$Standartization$	X	V	PCR_02
$Min - Max$	X	V	PCR_03
$Standartization$	X	V	PCR_04
$Standartization$	X	V	PCR_05
$Standartization$	X	V	PCR_06
$Standartization$	X	V	PCR_07
$Standartization$	X	V	PCR_08
$Standartization$	X	V	PCR_09
$Min - Max$	X	V	PCR_10
	V	V	$SpecialProperty$