תרגיל בית 1 – למידת מכונה

מגישים: ששון שמואל למעי (325172351) ויצחק גרינבוים (318837317)

1. מספר שורות – 1250, מספר עמודות – 26.
2. הפיצ'ר הזה מדבר על כמה שיחות ביום היה לחולה הקורונה עם המשפחה שלו או עם הרופאים. הטיפוס שלו אורדינלי מכיוון שטיפוס מספר השיחות שייך למספרים הטבעיים ולא לרציונליים/ממשיים, כלומר זה לא רציף (אין כזה דבר חצי שיחה).

	Conversations per Day	Count
0	3	224
1	2	215
2	4	190
3	5	156
4	6	111
5	1	104
6	8	72
7	7	60
8	9	39
9	10	23
10	11	19
11	12	12
12	13	9
13	14	6
14	17	4
15	15	2
16	16	2
17	19	1
18	22	1

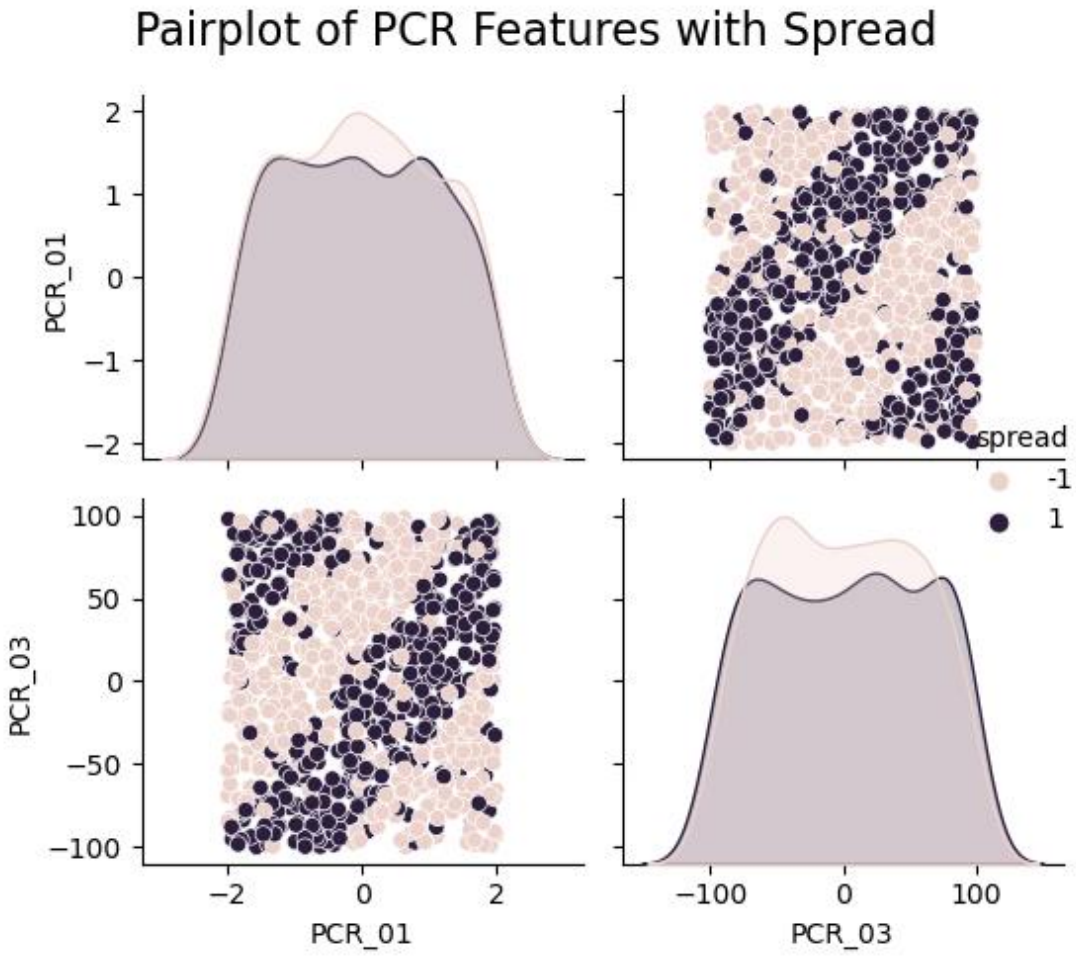
3. טבלת כל פיצ'ר:

	שם	תיאור	סוג
1	patient_id	מזהה מטופל	ordinal
2	age	גיל	ordinal
3	sex	מין	Categorical
4	weight	משקל	Continuous
5	blood_type	סוג דם	Categorical
6	current_location	מיקום נוכחי	other
7	num_of_siblings	מספר אחים	Ordinal
8	happiness_score	ציון אושרו של המטופל בסולם 3 עד 11	Continuous
9	household_income	גודל הכנסה של המשפחה בסולם 0 עד 8	Ordinal
10	conversations_per_day	מספר שיחות של המטופל עם בני אדם	Ordinal
11	sugar_levels	מידת הסוכר בדם	Ordinal
12	sport_activity	מידת פעילות ספורט של	Ordinal

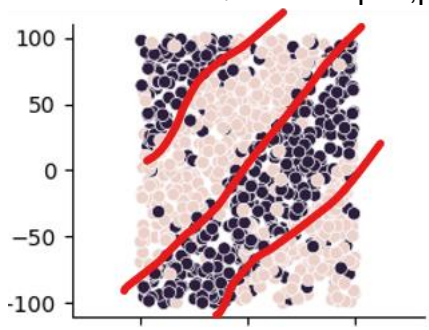
		המטופל בסולם 0 עד 5	
13	symptoms	הסיפטומים שהמטופל דיווח עלינם	other
14	pcr_date	תאריך בדיקת PCR	other
15	PCR_01	מאפיין 1 של בדיקת הPCR	Continuous
16	PCR_02	מאפיין 2 של בדיקת הPCR	Continuous
17	PCR_03	מאפיין 3 של בדיקת הPCR	Continuous
18	PCR_04	מאפיין 4 של בדיקת הPCR	Continuous
19	PCR_05	מאפיין 5 של בדיקת הPCR	Continuous
20	PCR_06	מאפיין 6 של בדיקת הPCR	Continuous
21	PCR_07	מאפיין 7 של בדיקת הPCR	Continuous
22	PCR_08	מאפיין 8 של בדיקת הPCR	Continuous
23	PCR_09	מאפיין 9 של בדיקת הPCR	Continuous
24	PCR_10	מאפיין 10 של בדיקת הPCR	Continuous

4. החשיבות של שמירת אותו הפיצול בכל המחקר הוא שאנו מוודאים שהתוצאות שלנו ניתנות לייצור מחדש. אחרת ההיסקים שאנו עלולים להסיק עשויים להשתנות.
כמו כן נרצה הפרדה בין הדוגמאות לאימון והדוגמאות לבחינה, כי אחרת עלולים להיות דוגמאות שישמשו גם גם לאימונו וגם לבדיקתו. כי אחרת לא נוכל למדוד את מידת דיוק המודל בדוגמאות שאינו פגש מראש.

5. הגרף:

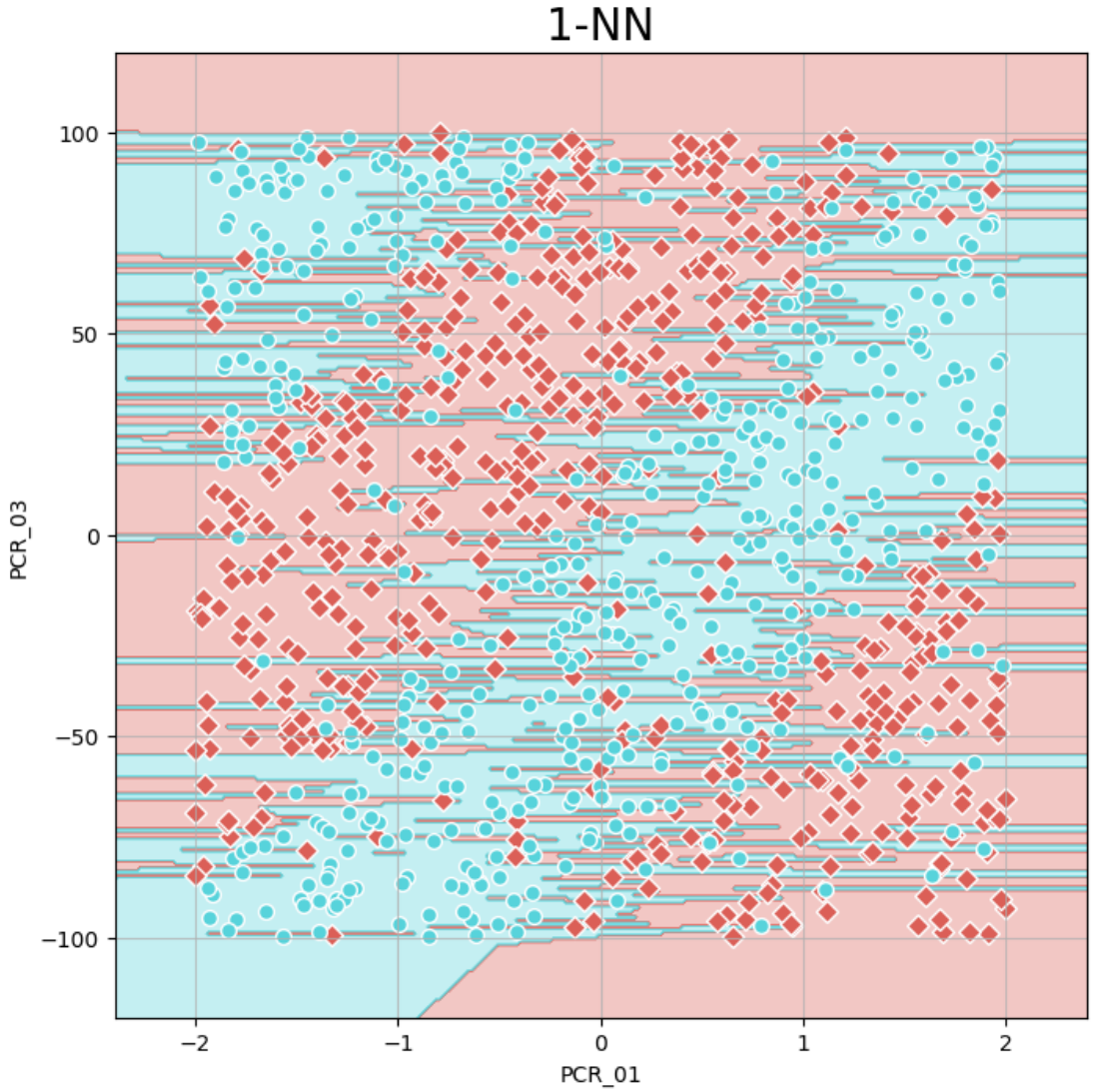


כן, ניתן לראות כי על ידי שלושה ישרים מקבילים ניתן לייצר הפרדה כמעט מוחלטת.

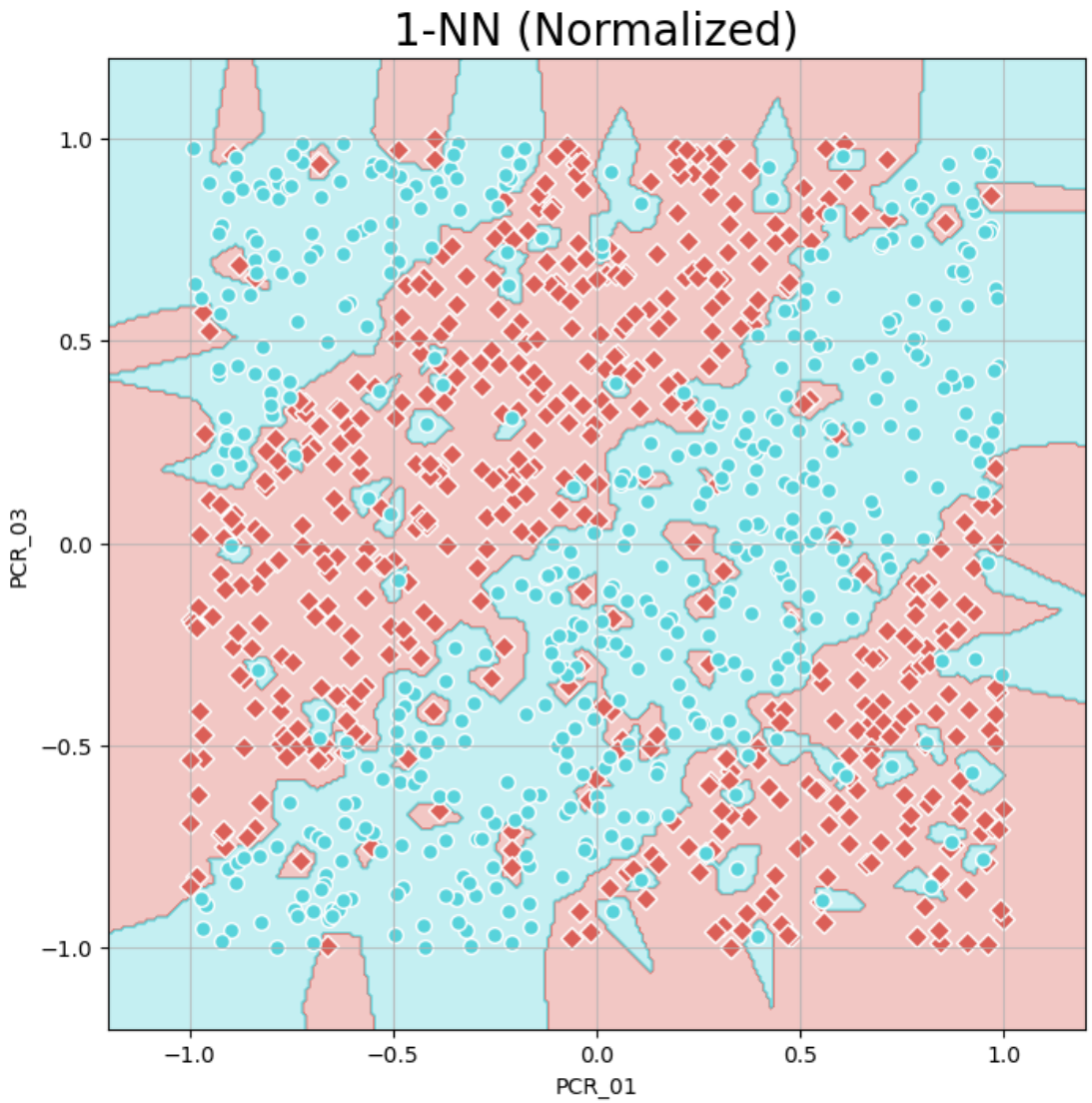


6. קורלציה בין PCR_01 לspread: 0.000442-.
קורלציה בין PCR_03 לspread: 0.015187.
למעשה זה אומר שכל מאפיין בנפרד לא יכול לעזור לנו לקבל פרדיקציה טובה. אבל בסעיף קודם, נטען כי כאשר משלבים בין שני המאפיינים ניתן לקבל הפרדה.
7. סיבוכיות הזמן:
אנו יוצרים אובייקט של מרחקים בין כל אובייקט אימון לאובייקט הטסט; המימד הוא d ולכן חישוב המרחק הוא O(d) עבור שני אובייקטים. מכיוון ואנו מחשבים את זה עבור כל אובייקטי האימון הסיבוכיות היא O(dm).
כעת אנו משתמשים בargpartition כדי למצוא את k האיברים הקטנים ביותר, סיבוכיות זמן: O(m). לאחר מכן אנו מחשבים את הסימן של האיברים הקרובים ביותר בסיבוכיות של O(k).
לכן בסך הכל: O(dm+m+k)=O(dm+k)=O(dm). מעבר שני כי k קטן מm.

8. דיוק על האימון – 1.0, דיוק על המבחן – 0.7



9. דיוק על האימון – 1.0, דיוק על המבחן – 0.796



ניתן לראות כי רמת הדיוק על המבחן גדלה בעשרה אחוזים.
מכיוון שKNN מבוסס על מרחק, עלינו לנרמל את הפיצ'רים כדי שניתן את אותו המשקל לכל פיצ'ר.
מרחק אוקלידי לא מנורמל: בין שני וקטורים מעל R^n.

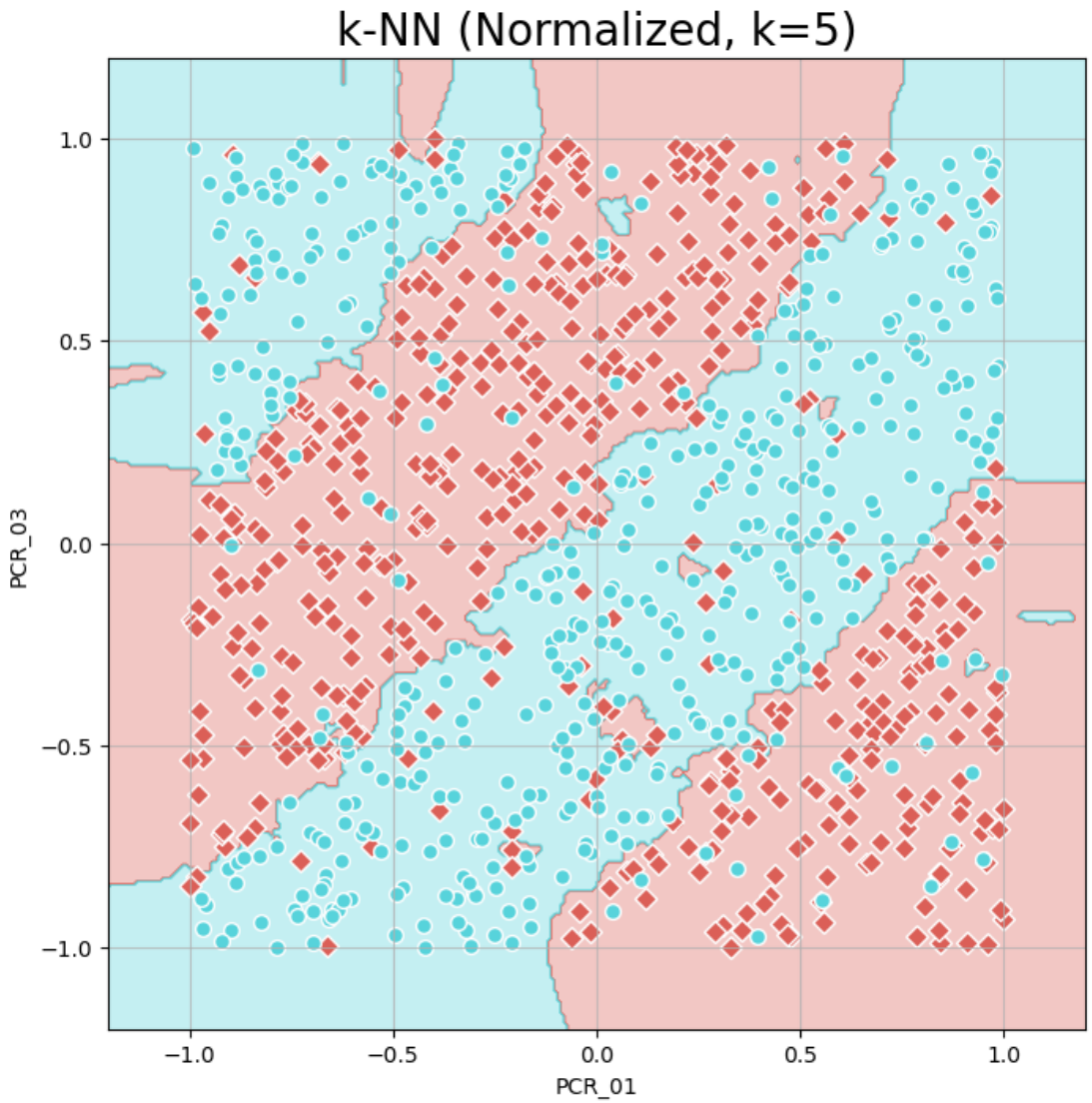
$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

מרחק מנורמל:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}\left(\frac{x_i-\mu_i}{\sigma_i} - \frac{y_i-\mu_i}{\sigma_i}\right)^2}$$

ניתן לראות כי במרחק לא מנורמל עבור פיצ'ר שהוא גדול משמעותית יכול להיות שההפרש בין xi ל yi הוא קטן במונחים של הפיצר, אבל כאשר סוכמים את ההפרשים, ההפרש הגדול יקבל דומיננטיות על ההפרשים הקטנים, למרות שיכול להיות שבהפרשים הקטנים יש מרחק יחסי גדול יותר ביחס לפיצ'ר.
לכן עלינו לנרמל ולתת לכל פיצ'ר בביטוי הסכום של המרחקים, את אותו המשקל כמו שאר הפיצ'רים.

10. דיוק על האימון – 0.885, דיוק על המבחן – 0.844

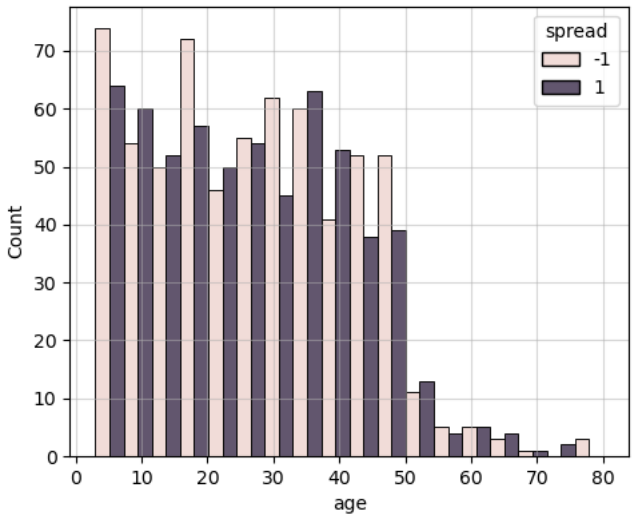


ניתן לראות שאזורי ההחלטה רציפים יותר מאשר בNN-1.
רמת הדיוק של המבחן ירדה, כי בNN-1 יש לנו את מקרה הקצה של OVERFITTING, לכן גם היה לנו יותר רעש באיזורי ההחלטה ולכן גם הדיוק על המבחן היה נמוך יותר כי הושפענו מהרעש הזה.
בכך שהגדלנו את מספר השכנים, קיבלנו יותר דיוק עבור המבחן אך פחות עבור האימון כי הוא מתעלם מהרעש.

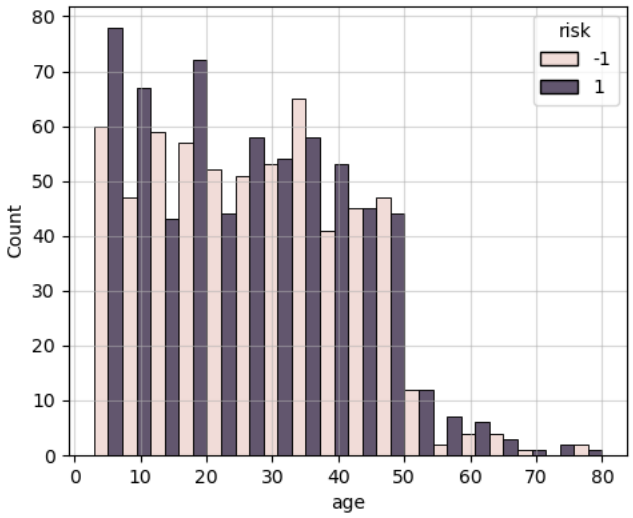
11. $Given\ that\ one\ of\ the\ features\ is\ a\ \chi^2\ variable, this\ normalization\ is\ a\ bad\ idea,$
$because\ \chi^2\ has\ a\ long\ tail\ which\ means\ a\ possibility\ of\ very\ big\ outliers.$
$If\ there\ are\ big\ outliers, it\ means\ that\ most\ of\ the\ reliable\ data\ is\ proportionally$
$close\ to\ the\ smaller\ values, and\ after\ normallization\ will\ be\ closer\ to\ -1,$
$so\ it\ means\ that\ almost\ all\ data\ points\ will\ be\ very\ similiar\ in\ their\ \chi^2\ parameter,$
$making\ it\ impossible\ to\ distinguish\ between\ them\ and\ virtually\ make\ them\ all\ in$
$the\ same\ distance\ from\ the\ test\ point, thus\ totally\ losing\ the feature's\ significance.$

12. נצטרך 4 (A, B, AB, O) כפול 2 (+,-) = 8 פיצ'רים בוליאניים.
13. רשימת הסיפטומים האפשריים הינה: {'smell_loss', 'sore_throat', 'fever', 'shortness_of_breath', 'cough', nan}. עבור כל סימפטום ניצור עמודה חדשה בטבלה ונעדכן לכל שורה את הערך של העמודות החדשות (1 אם הסימפטום קיים ו 1- אם אינו).

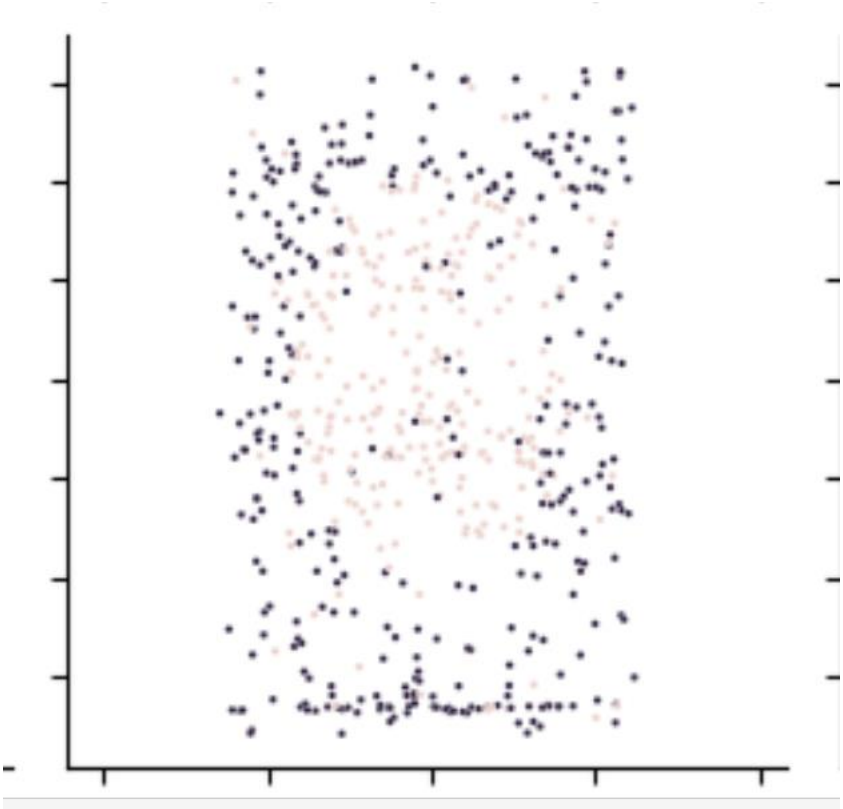
14. ניתן לראות כי עבור גילאים מסוימים כגון 50-40 סיכוי ההפצה קטנים ב20 אחוזים.



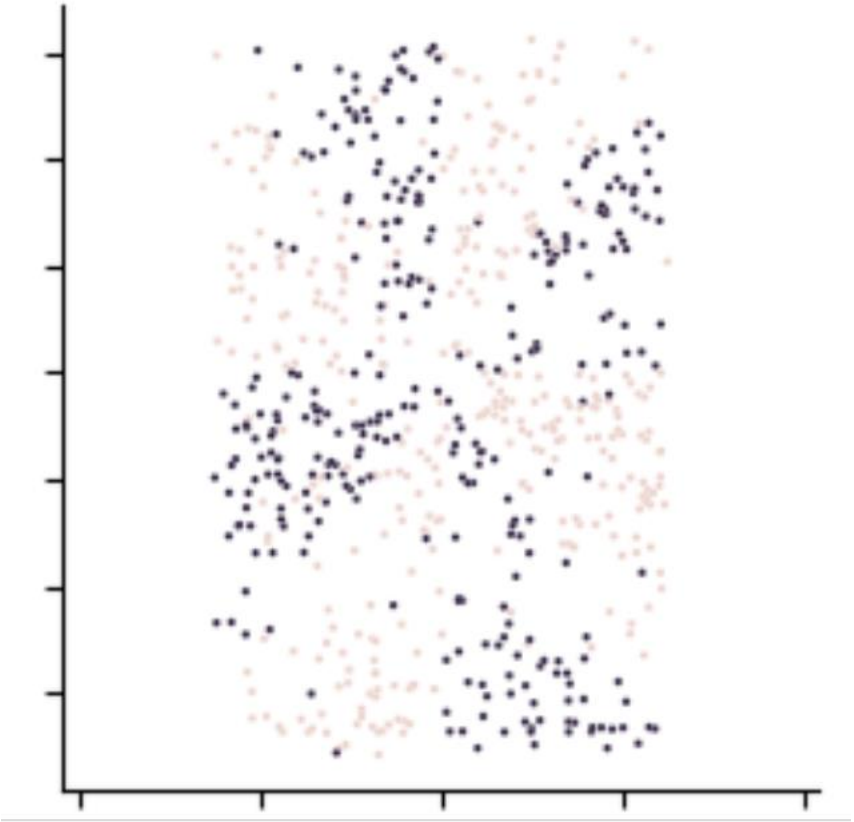
15. ניתן לראות כי עבור גילאים מסוימים, כגון גילאים קטנים שקטנים מ10, הסיכון גדול ב25 אחוזים.



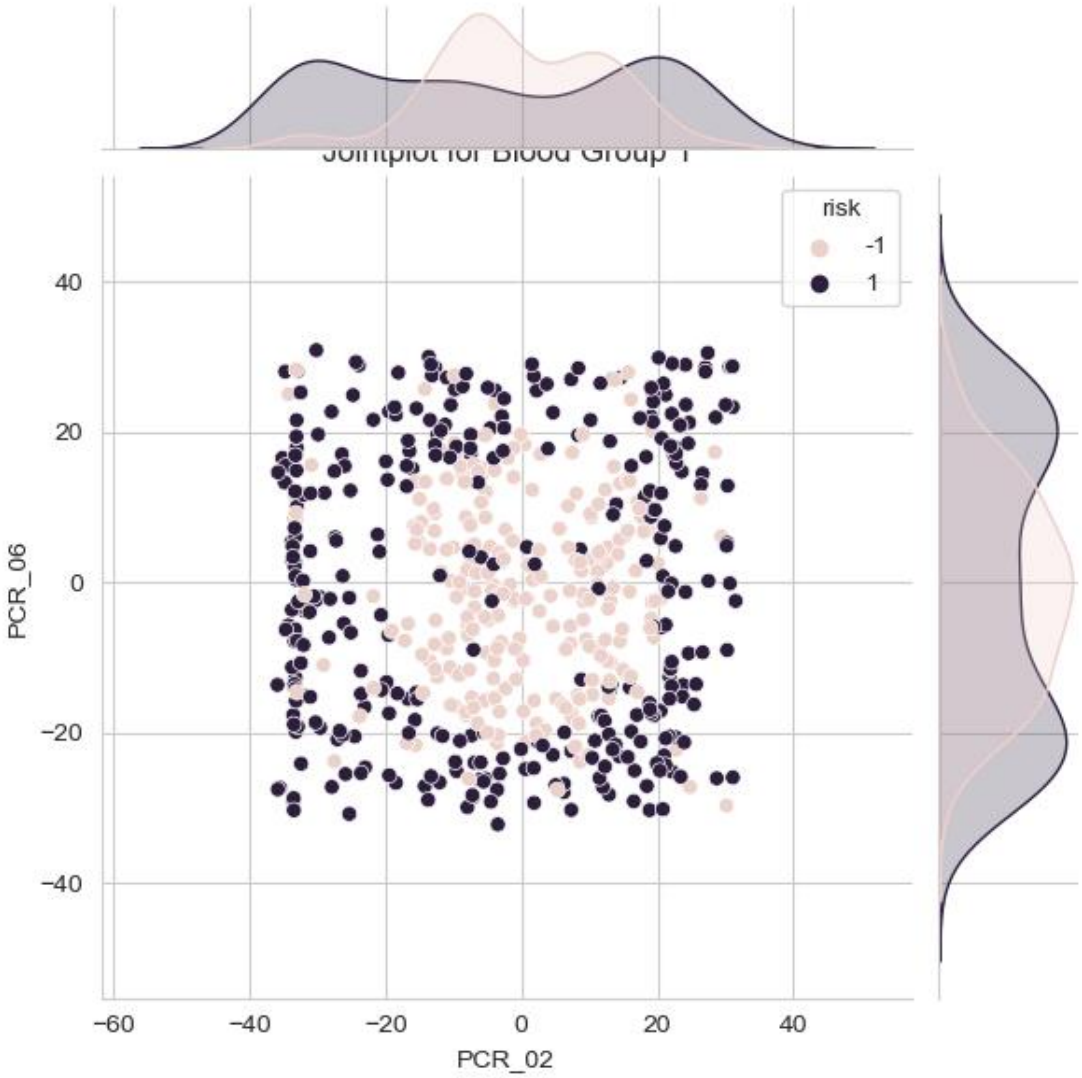
16. עבור שני הפיצולים בחרנו בPCR_02 וPCR_06. בחרנו זאת כי ניתן לראות שאפשר ליצור הפרדה על ידי מעגל באמצע, או למעשה כמו שראינו בתרגול ע"י הטלה למרחב תלת מימד ואז יצירת מפריד מישורי.

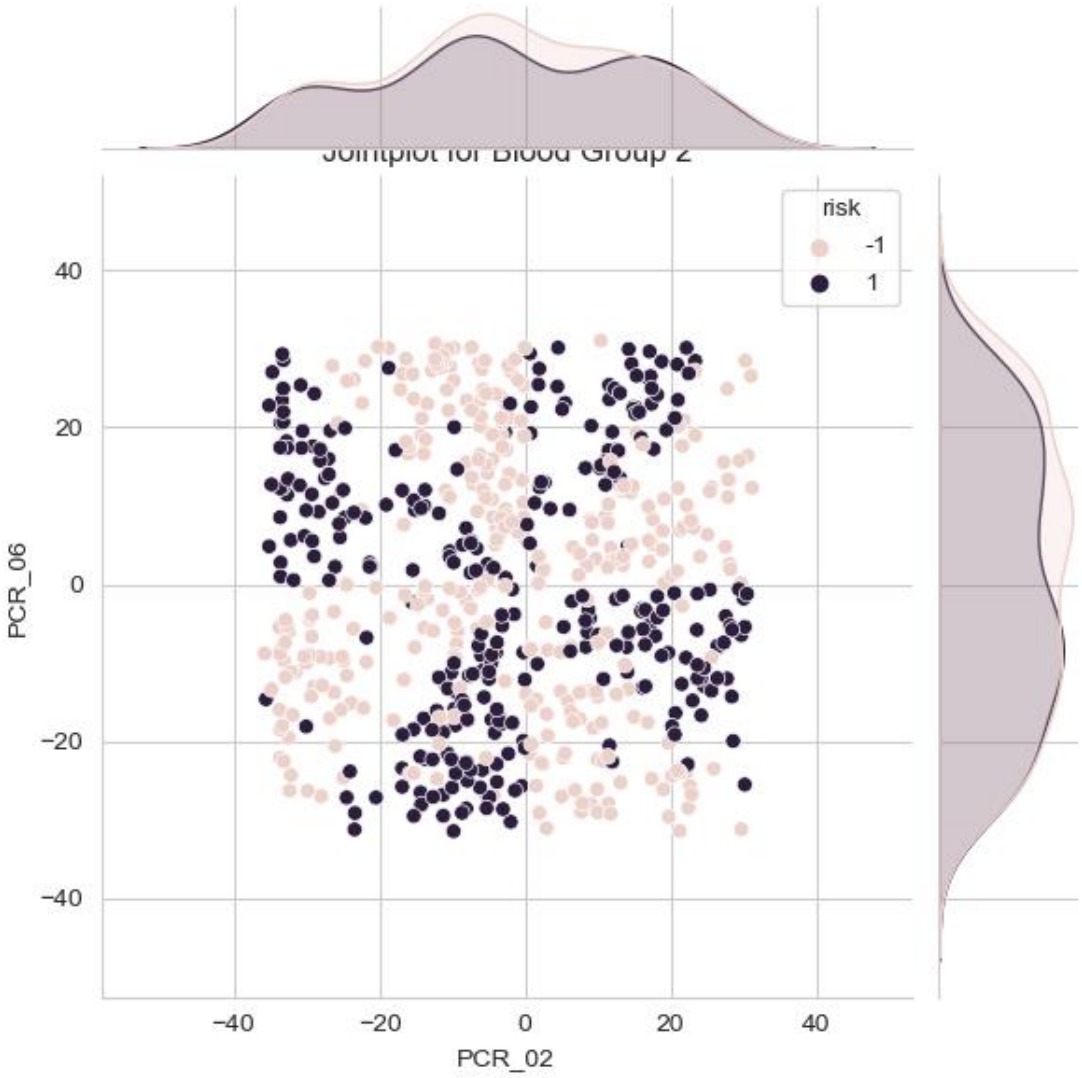


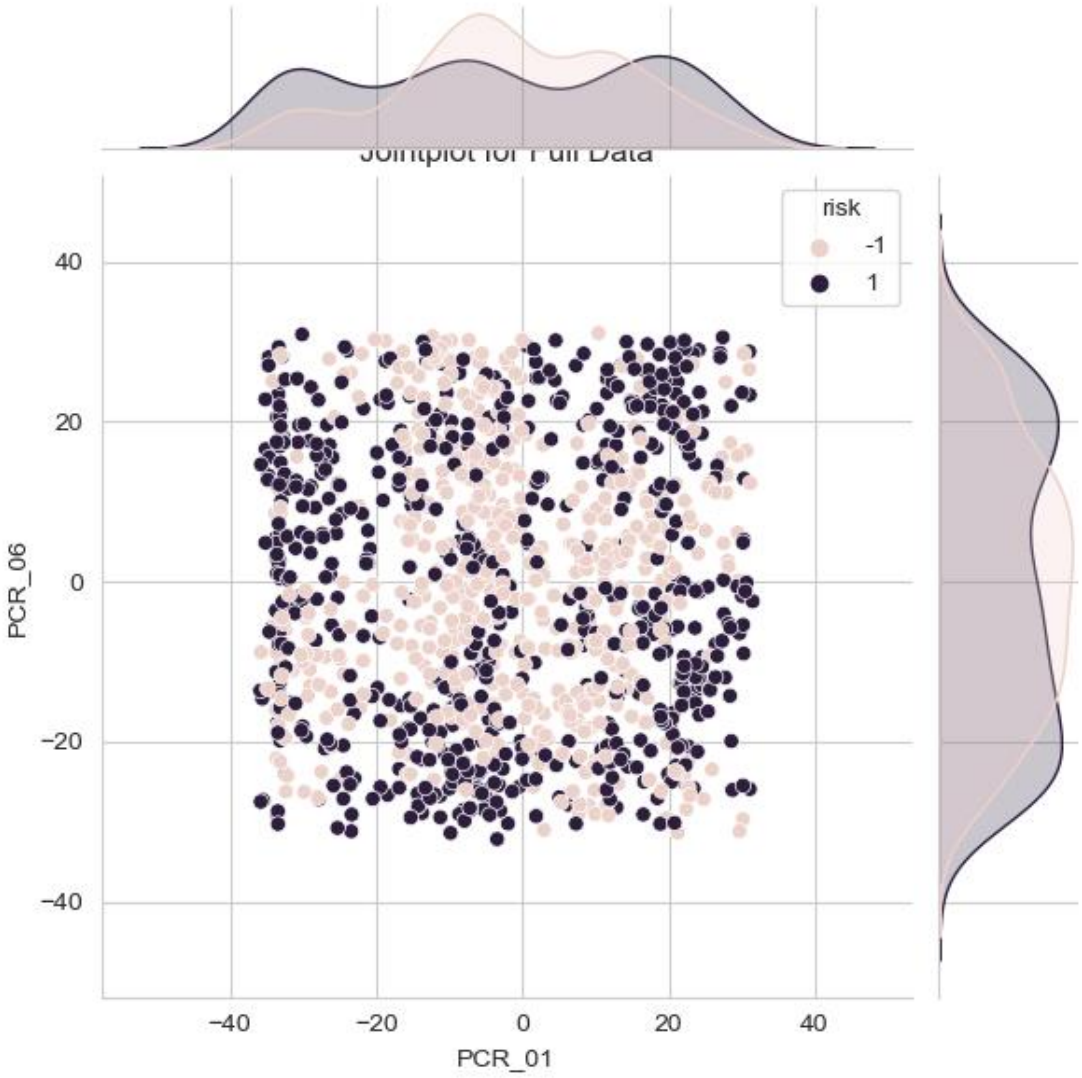
בקבוצה השנייה נראה כי אפשר ליצור מפריד עי ידי חלוקה לשמונה חלקים.



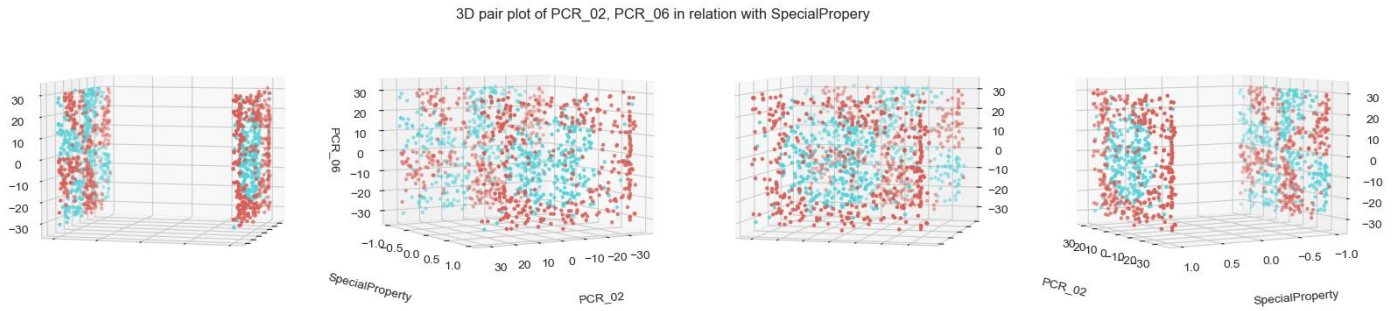
17. הגרפים:


Jointplot for Blood Group 1
risk
-1
1
40
20
0
−20
−40
PCR_06
−60
−40
−20
0
20
40
PCR_02


Jointplot for Blood Group 2
risk
-1
1
40
20
0
−20
−40
PCR_06
−40
−20
0
20
40
PCR_02

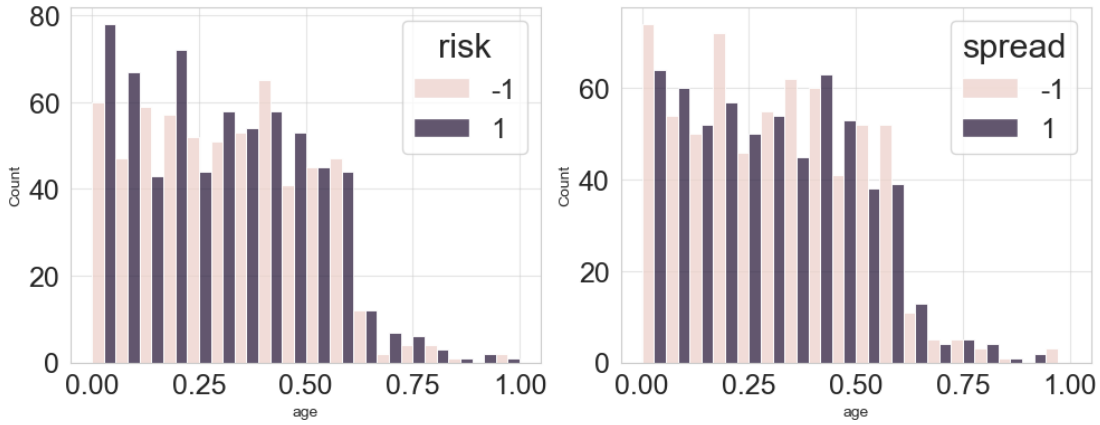


18. הגרף:

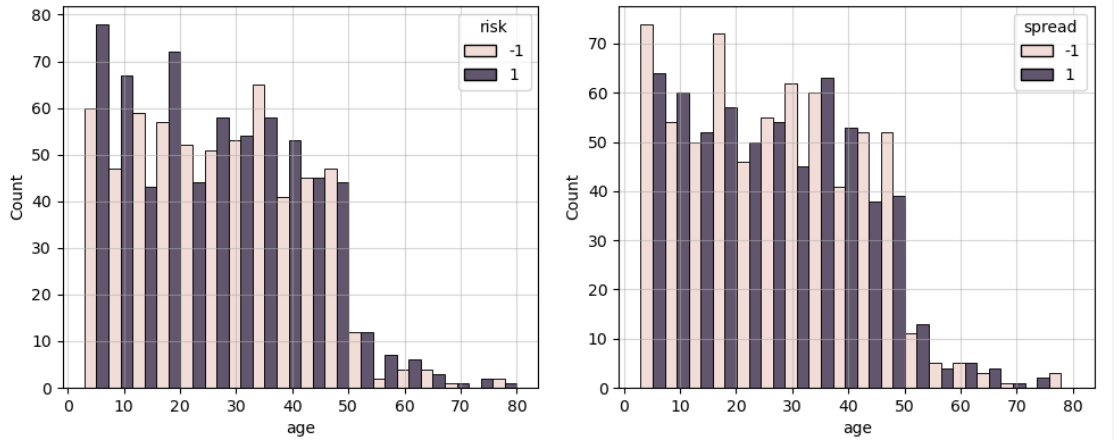


19. לא מספיק טוב כי אם הוא נמצא בקבוצת הדם שבה ראינו שיש הפרדה לשמונה חלקים לא({+O+, B}), אז דבר ראשון נפריד לפי קבוצת הדם ויישארו לנו עוד 2 הפרדות אבל זה לא מספיק.
20. זה מספיק, נפריד לפי קבוצת הדם. אם קיבלנו ששייכים ל: {+O+, B}. אז נבדוק האם נמצאים בתוך מעגל שרדיוסו 20 ומרכזו בראשית הצירים, אם כן אז הוא לא בסיכון ואם לא אז הוא כן בסיכון.
אם נמצאים בקבוצת סוגי הדם השנייה: נעשה הפרדה לפי כל רביע מתוך ארבעת הרביעים במערכת צירים PCR2 וPCR6. על כל רביע אפשר להחליט לפי נמצאים מעל הישר y=x או מתחתיו (או הישר y=-x).
21. הוא יצליח לאמן את זה בצורה טובה אבל לא לגמריי, כי עברו הדוגמאות שנמצאים על המפרידים שתיארנו (המעגל או הישרים), יכול להיות שהמודל יסווג אותם בצורה שגויה. אך עבור רוב הדוגמאות, זה יעבוד טוב.

22. אחרי:



לפני:



23. על 19 ו20 זה לא ישפיע כי למרות ששינינו את הSCALE עדיין צריך את אותן ההחלטות עד כדי קבוע.
כמו כן ב21 זה לא ישפיע כי עדיין רוב הדוגמאות לא יהיו על המפרידים ולכן NN-1 יסווג אותם נכון.

24. הטבלה:

הסבר	Normalization	New	Keep	שם
יוניפורמי	MinMax	X	V	patient_id
יוניפורמי	MinMax	X	V	age
נשאר לאבחנת ההשפעה על מין שונה	*בוליאני*	X	V	sex
נורמלי	Standard	X	V	weight
הומר לפיצ'רים אחרים	-	X	X	blood_type
לא רלוונטי	-	X	X	current_location
נורמלי	Standard	X	V	num_of_siblings
נורמלי	Standard	X	V	happiness_score
נורמלי	Standard	X	V	household_income
נורמלי	Standard	X	V	conversations_per_day
נורמלי	Standard	X	V	sugar_levels
נורמלי	Standard	X	V	sport_activity
הומר לפיצ'רים אחרים	-	X	X	symptoms
לא רלוונטי	-	X	X	pcr_date
יוניפורמי	MinMax	X	V	PCR_01
יוניפורמי	MinMax	X	V	PCR_02
יוניפורמי	MinMax	X	V	PCR_03

נורמלי	Standard	X	V	PCR_04
יוניפורמי	MinMax	X	V	PCR_05
יוניפורמי	MinMax	X	V	PCR_06
נורמלי	Standard	X	V	PCR_07
נורמלי	Standard	X	V	PCR_08
נורמלי	Standard	X	V	PCR_09
נורמלי	Standard	X	V	PCR_10
התקבל מהמרה של פי'צר אחר	*בוליאני*	V	V	SpecialProperty
התקבל מהמרה של פי'צר אחר	*בוליאני*	V	V	cough
התקבל מהמרה של פי'צר אחר	*בוליאני*	V	V	fever
התקבל מהמרה של פי'צר אחר	*בוליאני*	V	V	low_appetite
התקבל מהמרה של פי'צר אחר	*בוליאני*	V	V	shortness_of_breath
התקבל מהמרה של פי'צר אחר	*בוליאני*	V	V	sore_throat
התקבל מהמרה של פי'צר אחר	*בוליאני*	V	V	No Symptons