תרגיל בית 1

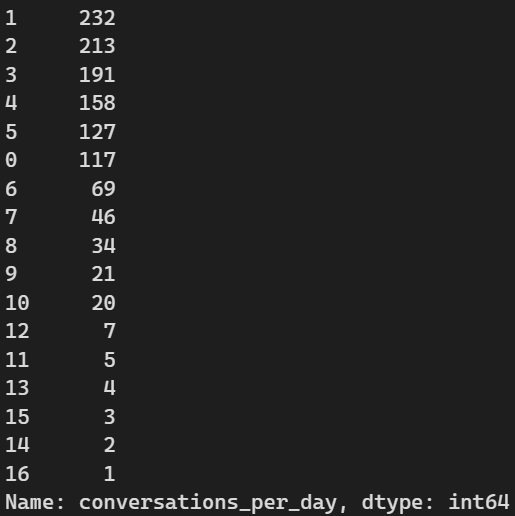
מגישים:  
אלמוג קריף 313151185  
אורי מינץ 314616897

# חלק 1

## שאלה 1

ב-dataset יש 1250 שורות ו-26 עמודות

## שאלה 2



כמות השיחות השונות שבוצעו ביום. ישנם 232 אנשים שביצעו שיחה אחת, 213 שבצעו שתיים, וכו'...  
כיוון שמספר שיחות הוא מספר טבעי יש לו יחס סדר של הטבעיים אבל הוא בדיד ולכן הוא מספר סידורי.

## שאלה 3

|  |  |  |  |
| --- | --- | --- | --- |
|  | שם התכונה | סוג | תיאור |
| 1 | patient\_id | Ordinal | מספר מזהה של הפציינט |
| 2 | age | Ordinal | גיל |
| 3 | sex | Categorical | מין |
| 4 | weight | Continues | משקל |
| 5 | blood\_type | Categorical | סוג הדם |
| 6 | current\_location | Continues | מיקום גאוגרפי של מיקום המגורים של המטופל |
| 7 | num\_of\_siblings | Ordinal | מספר האחים |
| 8 | happiness\_score | Ordinal | דרגת האושר של המטופל |
| 9 | household\_income | Ordinal | דרגת הכנסה של הבית |
| 10 | conversations\_per\_day | Ordinal | מספר שיחות של המטופל ביום |
| 11 | sugar\_levels | Ordinal | דרגת רמת הסוכר בדם של המטופל |
| 12 | sport\_activity | Ordinal | רמת פעילות גופנית |
| 13 | symptoms | Categorical | התסמינים שהמטופל חווה |
| 14 | pcr\_date | Ordinal | תאריך בו ביצע בדיקת PCR |
| 15 | PCR\_01 | Continues | נתון ראשון על ה-PCR |
| 16 | PCR\_02 | Continues | נתון שני על ה-PCR |
| 17 | PCR\_03 | Continues | נתון שלישי על ה-PCR |
| 18 | PCR\_04 | Continues | נתון רביעי על ה-PCR |
| 19 | PCR\_05 | Continues | נתון חמישי על ה-PCR |
| 20 | PCR\_06 | Continues | נתון שישי על ה-PCR |
| 21 | PCR\_07 | Continues | נתון שביעי על ה-PCR |
| 22 | PCR\_08 | Continues | נתון שמיני על ה-PCR |
| 23 | PCR\_09 | Continues | נתון תשיעי על ה-PCR |
| 24 | PCR\_10 | Continues | נתון עשירי על ה-PCR |

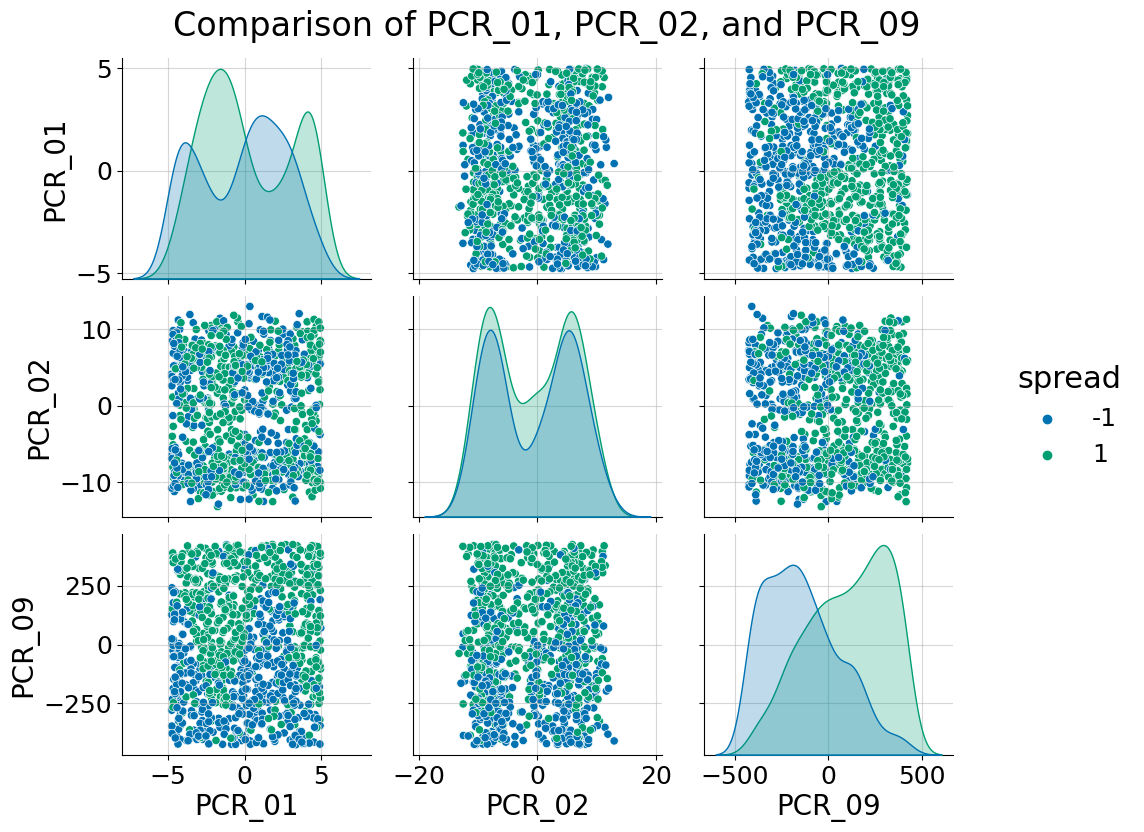
## שאלה 4

# חלק 2

## שאלה 5

|  |  |
| --- | --- |
|  | קורלציה |
| PCR\_01 |  |
| PCR\_02 |  |
| PCR\_03 |  |

## שאלה 6

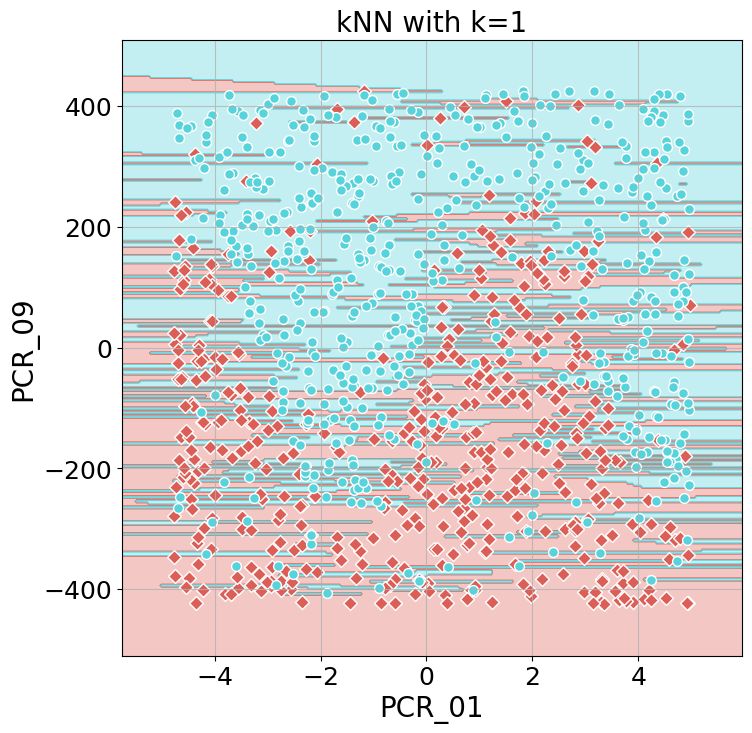


לפי דעתנו בחירה של PCR\_01 ו-PCR\_09 יביא את ההפרדה הטובה ביותר כיוון שבחירה של שניהם מביא את הגרף בעל ההפרדה הברורה ביותר.

## שאלה 7

נחלק את הפונקציה לשלבים.  
בשלב הראשון משתמשים ב-np.argpartition כאשר סיבוכיות הזמן שלה הוא כיוון שהפונקציה מבצעת quick sort ומסיימת כאשר היא מוצאת שיש שכנים לכן השוואות לכל בנוסף כל השוואה לוקחת לכן נקבל .  
אחר כך מוצאים את הלייבילים המתאימים לאינדקסים שמצאנו פעולה שלוקחת ולבסוף סוכמים את הלייבלים ומחזירים את הסימן, פעולה שלוקחת .  
לכן סה"כ נקבל וכיוון ש- נקבל .

## שאלה 8



הדיוק של ה-kNN על האימון הוא 1.0 בעוד הדיוק של ה-kNN על המבחן הוא 0.756.  
ניתן היה לצפות שהדיוק על האימון יהיה 1.0 כיוון שלכל נקודה, הנקודה הקרובה ביותר אליה שאימנו את המודל עליה היא הנקודה עצמה, לכן היא תהיה הנקודה הקרובה ביותר ו"החיזוי" יהיה מתאים.

## שאלה 9

Chart, scatter chart

Description automatically generated