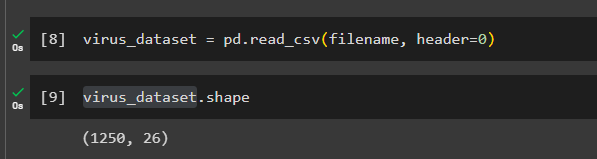
**Part 1: Data Loading and First Look**

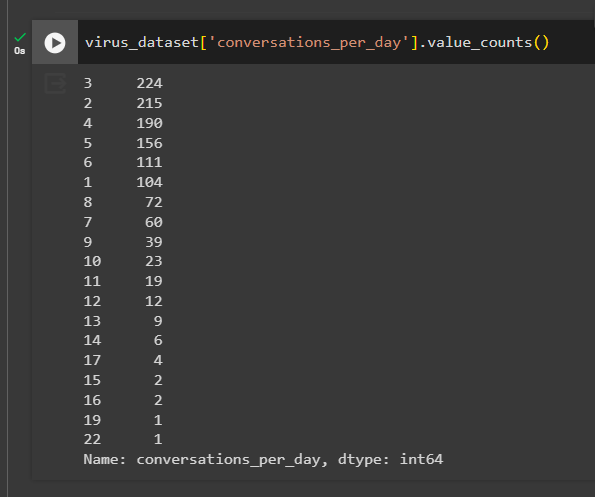
**Q1**

מספר השורות הוא 1250

מספר העמודות הוא 26.



**Q2**



התכונה 'שיחות ליום' כלומר כמה שיחות כל מטופל מבצע ביום.

בעולם האמיתי, תכונה זאת מצד אחד מתייחסת להפצת הוירוס, כלומר ככל שמטופל מקיים יותר שיחות ככה הסיכוי שהוירוס יופץ לעוד מעגלים של אנשים גדלה.

מצד שני, מתייחסת לסיכון לחלות במחלות חמורות יותר, אם מטופל הוא בודד ככה עולה הסיכוי שמצבו ידרדר בהיעדר טיפול או עזרה ובהעדר מצב רוח.

התכונה הזאת היא “ordinal” variable type כי תכונה זאת categorical עם סידור טבעי.

Categorical כי הוא מכיל מספר סופי של קטגוריות או קבוצות נפרדות, כל קבוצה מגדירה את כל מי מי שביצע X שיחות ביום.

בנוסף עם סידור טבעי כי יש סדר סדר מלא, ניתן להשוות בין כל שני ערכים, מי יותר גדול/קטן מהשני.

Q3

|  |  |  |
| --- | --- | --- |
| Type | Description | Feature name |
| other | מספר מזהה של המטופל  למשל יכול להיות תעודת זהות, מספר דרכון וכל מזהה ייחודי אפשרי. | patient\_id |
| Ordinal | גיל של המטופל | age |
| categorical | מין המטופל(זכר או נקבה) | sex |
| Continuous | משקל המטופל | Weight |
| categorical | סוג דם של המטופל | blood\_type |
| other | מיקום של המטופל לפי קורדינטות | current\_location |
| Ordinal | כמה אחים יש למטופל | num\_of\_siblings |
| Ordinal | רמת האושר של האדם,  כנראה לפי שאלות ששואלים את המטופל נותנים ציון לרמת האושר בטווח מספרים כלשהו. | happiness\_score |
| Ordinal | הכנסה כוללת של כל בני משק הבית בו נמצא המטופל.(מעוגל) | household\_income |
| Ordinal | כמה שיחות מבצע כל מטופל בכל יום | conversations\_per\_day |
| Ordinal | רמת הסוכר בדם של המטופל  (מעוגל) | sugar\_levels |
| Ordinal | פעילות גופנית של המטופל בטווח מספרים 0-5 | sport\_activity |
| categorical | סימפטומים של המטופל | symptoms |
| other | היום שבו המטופל עשה את בידקת הPCR שלו. | pcr\_date |
| Continuous | TODO | PCR\_01 |
| PCR\_02 |
| PCR\_03 |
| PCR\_04 |
| PCR\_05 |
| PCR\_06 |
| PCR\_07 |
| PCR\_08 |
| PCR\_09  PCR\_10 |

**Partitioning the data**

Q4

חשוב להשתמש באותו פיצול בניתוח המידע כי אנחנו רוצים לעבוד על אותו תת קבוצה של הdataframe שלנו כי אחרת נקבל כל פעם ניתוח שונה של המידע כי אנחנו נעבוד על תת קבוצה אחרת כל פעם שיגרום לקבלת מודלים שונים.

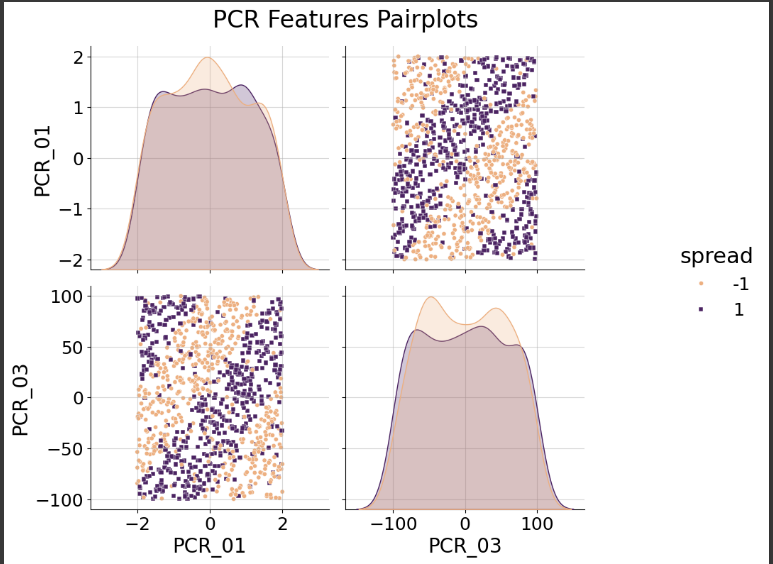
חשוב שנוכל לשחזר כל פעם את הניסוי שאותו ביצענו ולהשוות את אלגורתמי הלמידה על אותן תת קבוצות של הנתוונים שלנו.

**Part 2: Warming up with k-Nearest Neighbors**

**Basic data exploration**

Q5

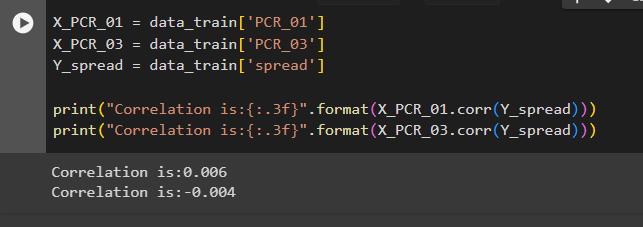
ניתן לראות שאפשר להפריד את הdata למקטעים בצורה די טובה, כמות המקטעים לא גדולה וכל מקטע עם כמות גדולה של נקודות ולכן ההפרדה, כפי שזה נראה, טובה ויכולה לתת תחזית די טובה עם שגיאה נמוכה.



Q6

Correlation הוא מדד סטטיסטי שניתן לראות אם שני משתנים יש ביניהם יחס ישר, כלומר מתשנים ביחד בקצב קבוע.

התשובות שקיבלנו :



הcorrelation בין PCR\_01 ו-spread הוא מאוד קרוב לאפס זה אומר שהקשר הלינארי חלש, אבל חיובי כלומר הם גדלים ביחד .

הcorrelation בין PCR\_03 ו-spread הוא קרוב מאוד לאפס ולכן אומר שהקשר הלינארי חלש והוא גם שלילי כלומר ערכים של אחד נוטים לגדול כשהערכים של השני קטנים.

אנחנו רואים שהתכונות הללו מתנהגות באופן שונה עם התכונה spread ולכן אלו שתי תכונות שיכולות לתת לנו מידע טוב על הdatabase שלנו בהקשר של התכונה spread, זה גם מדגיש את מה שראינו בסעיף קודם ששתי התכונות האלו הם שונות וניתנות להפרדה על הdatabase שלנו ולכן יכולות לנתת לנו prediction טוב.

בתשובה שקיבלנו קודם לכן, קיבלנו מודל שניתן להפרדה באמצעות הסתכלות על שתי תכונות בלבד ואת תכונת מטרה spread, כאן קיבלנו שכל תכונה מתנהגת באופן שונה עם spread - מתנהגת בגדילה שונה ולא לינארית עם תכונת המטרה, אחת גדלה יחד איתה והשנייה גדלה הפוך ממנה ולכן TODO

**k-NN implementation**

Q7

Predict מקבל datapoint ממימד d, נסמנה בx.

שלב 1 : מחשבים את המרחק של הנקודה מכל הנקודות בdatabase כלומר m נקודות ממימד d, מחשבים את המרחק האוקלידי L2, יש m מרחקים, מרחק של כל נקודה x, כל מרחק מחושב במימד d ולכן לוקח d צעדים וסהכ O(md).

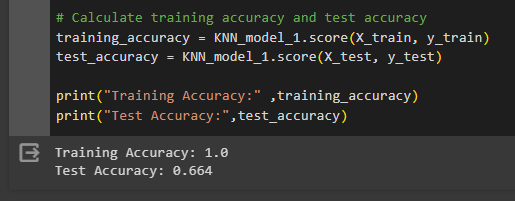
שלב 2 : מבצעים sort על מערך המרחקים שקיבלנו בשלב 1, בשביל לקבל את המרחקים הקטנים ביותר מהנקודה, בעצם אנחנו מקבלים קבוצת אינדקסים של הנקודות עם המרחק הקטן ביותר מהנקודה ונרצה רק את הk הכי קרובים. מיון זה מבוצע בo(mlogm) כגודל הdatabase שלנו.

שלב 3 : סוכמים את הלייבלים של הנקודות הללו ורואים האם יש יותר +1 או -1 בחישוביות של O(k) כי עוברים רק על k הנקודות עם המרחק הקטן ביותר מנקודת הקלט.

סהכ נקבל סיבוכיות : O(md+mlogm+k), k<m ולכן : O(md+mlogm)

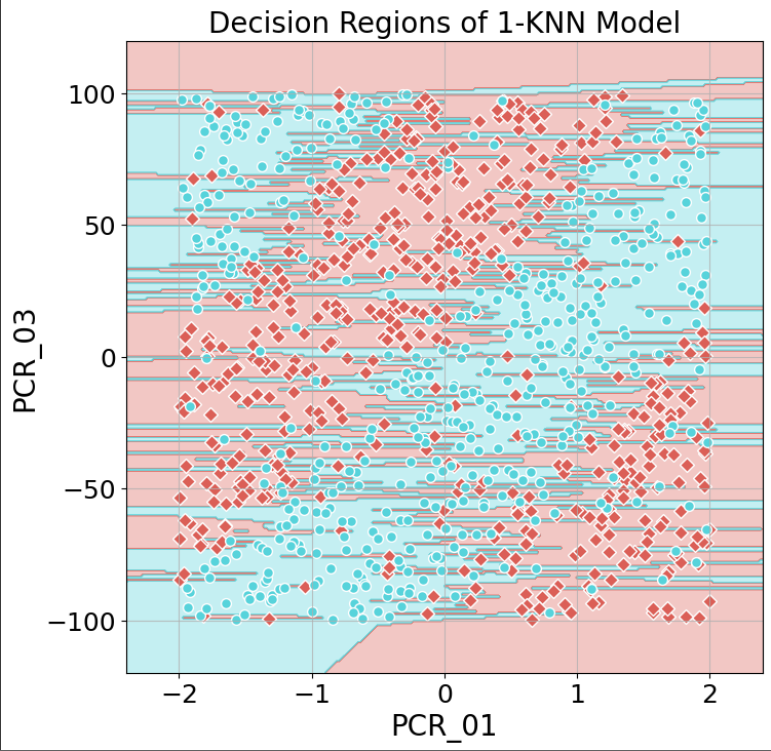
למה בהצראה ראינו שזה O(md) TODO

Q8



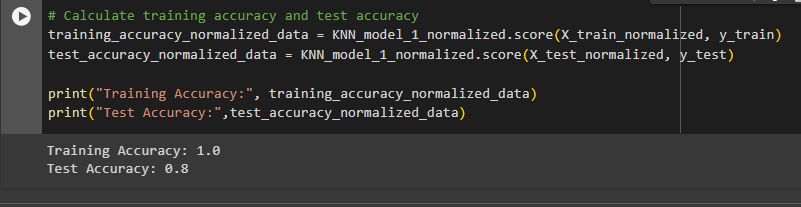
Training\_accuarcy = 1.0

Test\_accuracy = 0.664

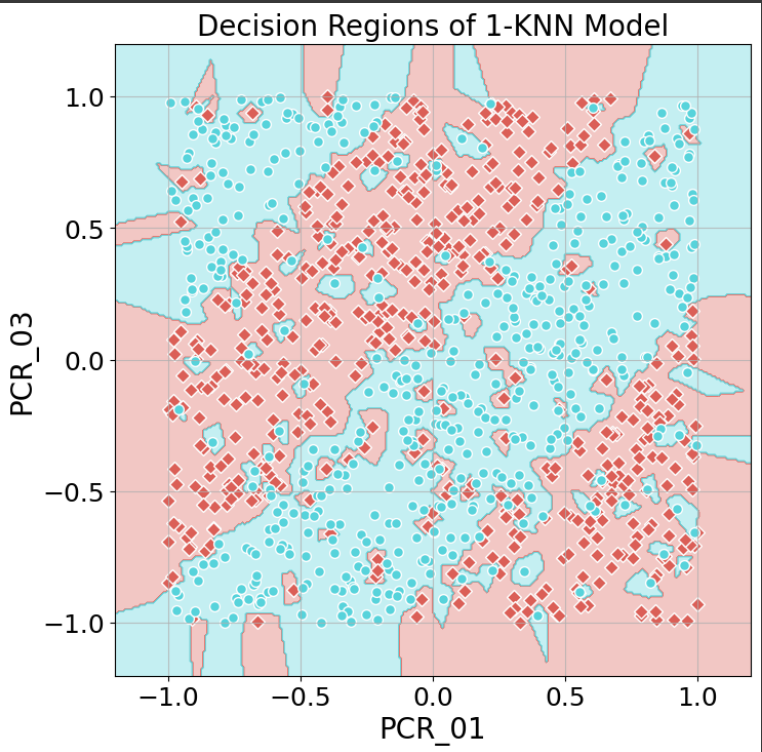


**Data Normalization**

Q9



TODO -יצא לנו תשובות שונות פה

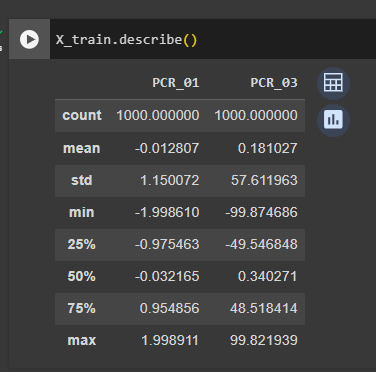


ניתן לראות בסעיף הקודם שdatapoints מאוד קרובות אחת לשנייה ואפילו חופפות למרות שיש להם לייבלים שונים, בעצם הנורמליזציה תיקנה טיפה את בעיית "המעיכה" בשני הצירים מה שגרם לdatapoints להתרחק אחד מהשנייה ולאפשר בחירת את השכן הקרוב ביותר בצורה יותר טובה. ניתן לראות זאת בdecision regions of the model, בסעיף הקודם הצבע האדום "נמעך" על איזורים רבים לעומת בסעיף הזה שיש הפרדה טובה יותר בין הצבע הכחול לאדום.

התוצאות בשני הסעיפים לdata training מראים דיוק של 1, כי מדובר על k=1 ולכן מתאימים לנקודה את הלייבל שלה.

לעומת זאת אנחנו רואים דיוק יותר טוב בdata testin בסעיף הזה, כי ההפרדה היא יותר טובה ולא "נמרחת" על איזורים כמו בסעיף הקודם וגורמת לoverfitiing מיותר.

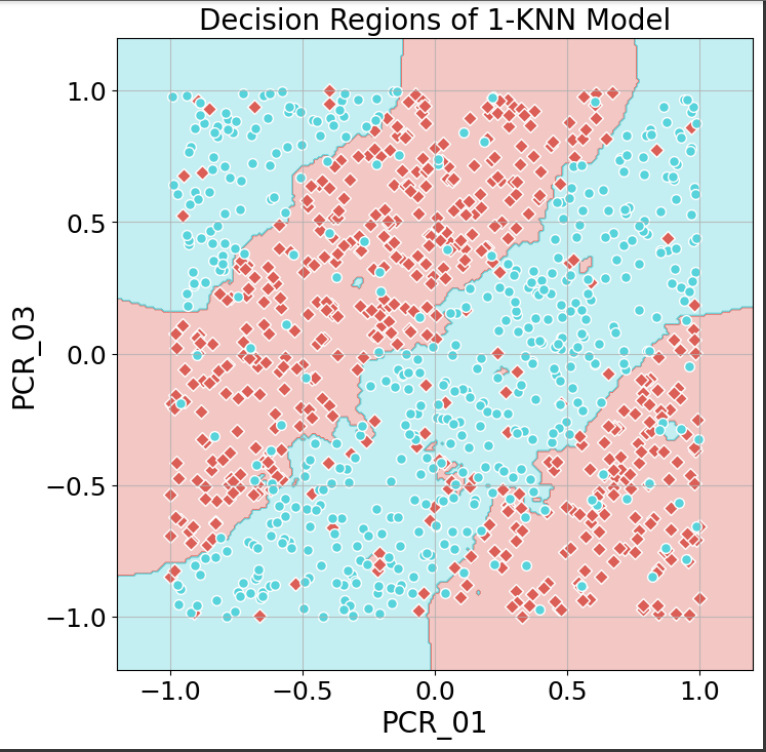
ניתן לראות שהטווח של כל אחד מהפיצרים הוא שונה :



PCR\_01 בטווח מספרים (-2,2) לעומת PCR\_03 בטווח מספרים (-100,100), לכן בחישוב של המרחק (L2) התכונה PCR\_03 שלה scale גדול יותר יהיה לה משקל גדול יותר על התוצאה.

TODO הוכחה מתמטית

Q10



A screen shot of a computer

Description automatically generated