**מבוא למערכות לומדות- תרגיל בית 1- דו״ח עבודה**

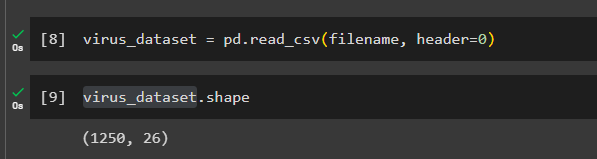
**מגישים:   
  
גל קסטן 316353176**

**חן פרי – 313283657**

(Q1)

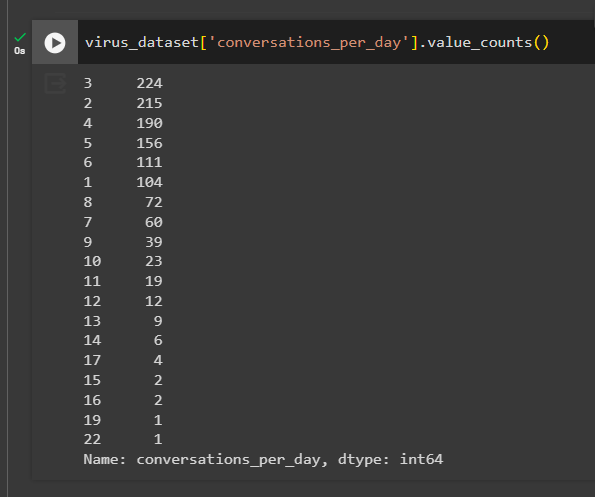
מספר השורות הוא 1250

מספר העמודות הוא 26



(Q2)

פלט:



בעולם האמיתי, המשתנה ״ מספר שיחות ליום״ מתייחס כנראה לממוצע מספר השיחות ביום שאדם מבצע פנים מול פנים. ככל שמטופל מקיים יותר שיחות ככה קיים סיכוי גבוה יותר שהווירוס יופץ למעגלים נוספים ומצד שני ככל שמטופל בודד יותר עולה הסיכוי שמצבו ידרדר (העדר טיפול, חוסר מצב רוח).

סוג המשתנה הוא ordinal כיוון שמצד אחד המשתנה קטגורי (הוא דיסקרטי, ויש מספר סופי של ערכים שהוא יכול לקבל) אבל מצד שני קיים סדר טבעי בין הערכים שהמשתנה יכול לקבל (מדובר במספרים טבעיים עם יחס סדר) ויש משמעות כמותית לערכים גדולים/קטנים של כמות שיחות ביום.

(Q3)

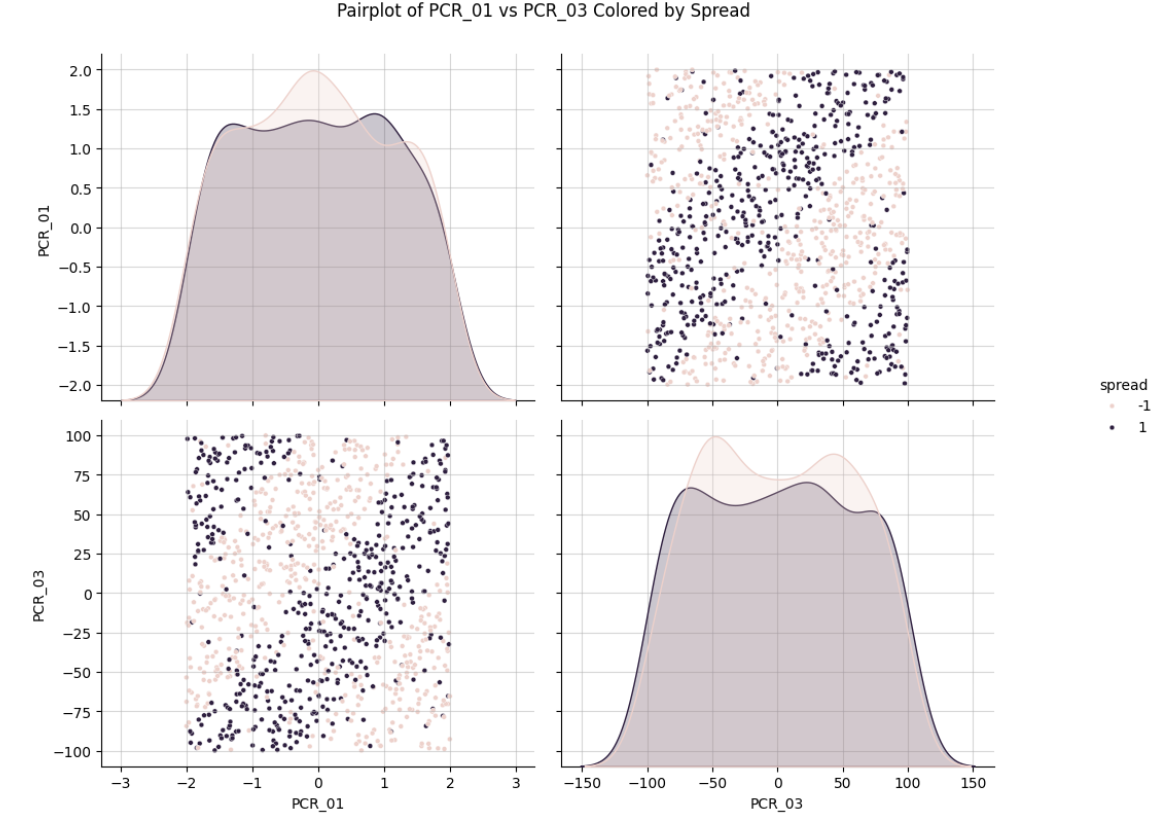
|  |  |  |
| --- | --- | --- |
| Type | Description | Feature name |
| Other | Incremental id of the patient | **patient\_id** |
| Ordinal | Age of the paitent | **age** |
| Categorical | Sex of the patient (Male/Female) | **sex** |
| Continuous | Weight of the patient in kilograms | **weight** |
| Categorical | The blood type of the patient (O/A/AB/B and +/-) | **blood\_type** |
| Other | Current location of the patient (latitude and longitude). | **current\_location** |
| Ordinal | The number of siblings the patient has | **num\_of\_siblings** |
| Ordinal | A Score describing the patient’s level of happiness | **happiness\_score** |
| Ordinal | A categorical representation of the patient’s household income (might be division to economy classes) | **household\_income** |
| Ordinal | Average number of conversations the patient has everyday | **conversations\_per\_day** |
| Ordinal | The sugar level measurements for the patient | **sugar\_levels** |
| Ordinal | The patient’s level of sport activity on scale from 0-5 | **sport\_activity** |
| Other | Textual description of any symptoms the patient may have reported | **symptoms** |
| Other | The date in which a PCR test was conducted | **pcr\_date** |
| Continuous | Numerical results of PCR Tests. Could be Measurements of different genetic sequences or pathogens. | **PCR\_xx** |

**(Q4)**

חשוב להשתמש באותו פיצול כאשר אנו מבצעים את תהליך ניתוח המידע מאחר ואנחנו רוצים שיהיה ביכולתנו לשחזר את התוצאות שקיבלנו הניסוי ושנוכל לקבל את אותם המודלים ( פיצול שונה יגרום לנו לקבל תת קבוצה שונה של נתונים בכל פעם ולכן אנו עשויים לקבל מודלים שונים).

בנוסף, כאשר אנו משווים אלגוריתמים של מודלי למידה, אנחנו רוצים שיהיה ביכולתנו לבצע השוואה הוגנת של אלגוריתמי למידה שונים תחת אותם תנאים ולכן חשוב שנשתמש באותם אותם תתי קבוצות של הנתונים שלנו.

**(Q5)**



על בסיס התמונה, הפיצ׳רים pcr\_01, pcr\_03 עשויים להיות שימושיים לחיזוי spread ביחד מאחר וניתן לראות כי הדאטה כמעט פריד ל4 אזורים שונים, כאשר בכל אזור יש הרבה דוגמאות מאותה מחלקה ומעט outliers. עם זאת, הדאטה לא פריד לינארית (מאחר ולא ניתן לפצל את הדאטה כך שבצד אחד נמצאת מחלקה אחת ובצד נמצאת מחלקה שניה) ולכן נצטרך מודל שיודע ללכוד קשרים מורכבים יותר.

כל פיצ׳ר בעצמו לא מספיק כדי לחזות את הדאטה- ההתפלגויות השוליות הן של pcr\_01 והן של pcr\_03 מראות חפיפה בטווח הערכים של המשתנה כאשר הspread הוא ממחלקה 1 וכאשר הspread ממחלקה -1. המשמעות של כך שהנתונים לא ניתנים להפרדה לינארית (בממד אחד – הפרדה על בסיס פונקציית סף כפי שראינו) על בסיס אחד מהמשתנים כדי לחזות את הspread .

**(Q6)**

**correlation between spread and PCR\_01**: 0.006

**correlation between spread and PCR\_03:** -0.004

מן הממצאים עולה כי :

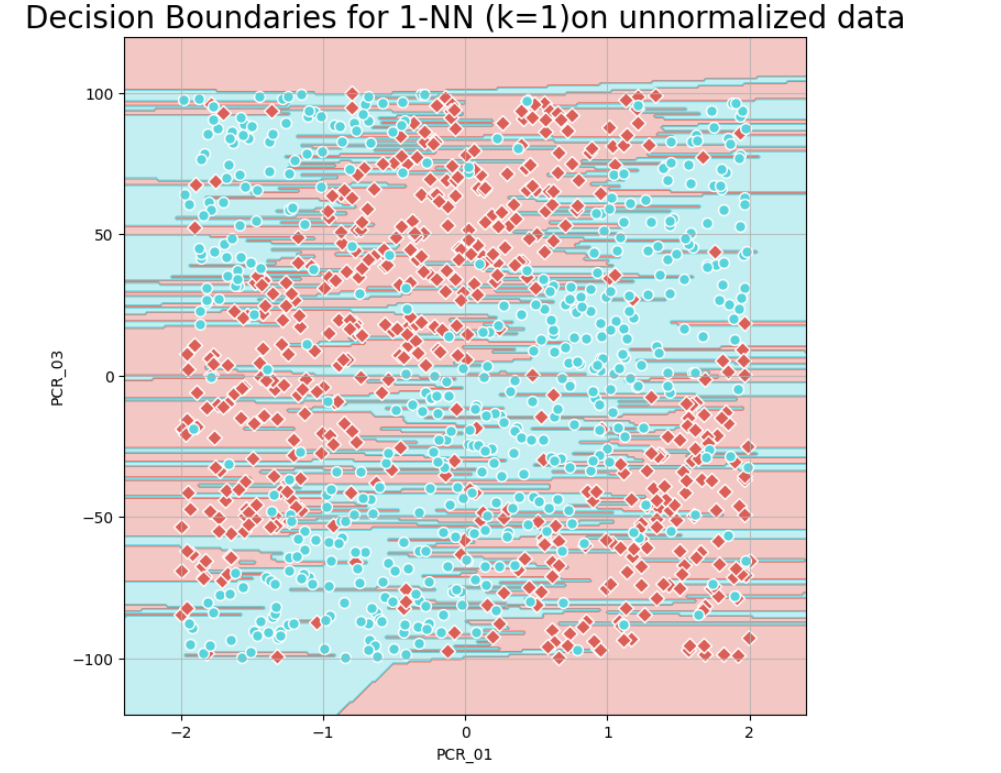
* הקורלציה בין pcr\_01 לspread מאד קרובה ל0, כלומר קשר לינארי חלש עם נטייה לקשר חיובי (כלומר אם משתנה אחד גדל גם השני)
* הקורלציה בין pcr\_01 לspread מאד קרובה ל0, כלומר קשר לינארי חלש עם נטייה לקשר שלילי (כלומר אם משתנה אחד קטן המשתנה השני גדל)

ממצאים אלו תומכים במה שמצאנו קודם לכן. על בסיס הקורלציה ניתן להסיק כי הקשר בין המשתנים לspread אינו לינארי, כלומר לא ניתן להצביע על כך שאם אחד מערכי הpcr עולה/יורד כך גם משתנה המחלקה של spread ולכן לא ניתן לחזות את spread רק על בסיס פונקציית סף כלשהי. בסעיף הקודם ראינו כי כאשר מסתכלים על ההתפלגות השולית של כל אחד מן המשתנים לבד בהפרדה לפי קטגוריות spread, הייתה חפיפה בין הערכים, כלומר לא הייתה הפרדה לינארית ברורה לעין בין טווח ערכים שpcr\_01 מקבל עבור spread=1 וטווח ערכים עבור spread = -1. באופן דומה גם עבור pcr\_03.

**(Q7)**

**זמן ריצה של פונקציית החיזוי על נקודה אחת P מממד d :**

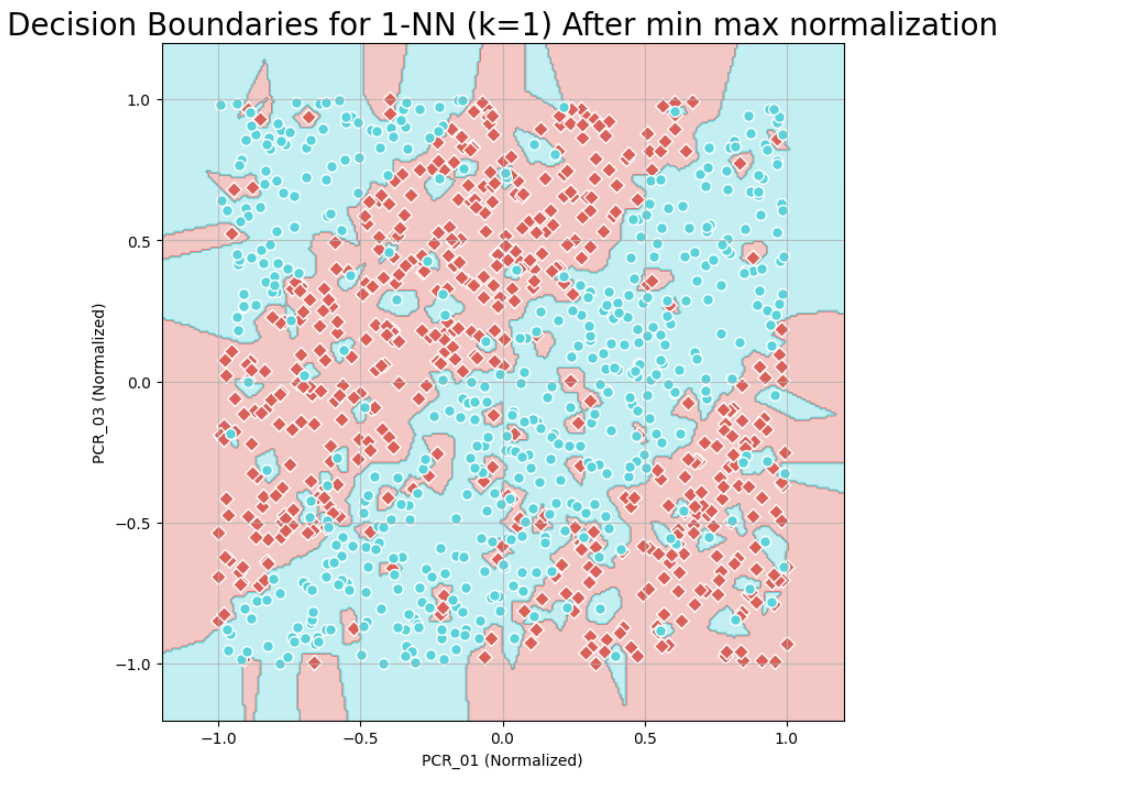
* שלב 1- חישוב המרחקים בין הנקודה p לשאר הנקודות: מעבר על כל הנקודות בסט האימון, כאשר לכל נקודה- מחשבים את המרחק האוקלידי בינה לבין הנקודה p. עלות חישוב המרחק בין 2 נקודות תלויה בממדי הנקודות. מאחר והנקודות מממד d לצורך חישוב המרחק האוקלידי דרושות d מכפלות, ולכן חישוב אחד של המרחק יעלה O(d). סה״כ נדרש ל O(m\*d) חישובים בשלב זה.
* שלב 2- חיפוש k השכנים הקרובים ביותר לנקודה p: לצורך החישוב אנו משתמשות בפונקציה argpartition על מערך המרחקים שקיבלנו בשלב הקודם , כאשר אנחנו מבצעות חלוקה כך ש k האינדקסים של האיברים הכי קטנים יופיעו בk האיברים הראשונים במערך הפלט. ניתן להעריך שהפונקציה משתמשת מאחורי הקלעים באלגוריתם partition שסיבוכיות הזמן שלו לינארית בגודל האיברים במערך. לכן עלות שלב זה- O(m).
* שלב 3: חישוב הפרדיקציה ע״פ החלטת הרוב: עבור k השכנים הכי קרובים אנו סוכמות את התיוגים של הנקודות ורואות האם הסכום חיובי (יותר שכנים עם תיוג 1) או שלילי (יותר שכנים עם תיוג -1) ועל פי סימן הסכום מחזירות את הפרדיקציה. סה״כ עלות שלב זה היא O(k) מאחר ועוברים רק על k הנקודות עם המרחק הקטן ביותר מנקודת הקלט
* סה״כ סיבוכיות O(md+m+k) מאחר וk<m סה״כ נקבל O(md).

(Q8)

מידת הדיוק של המודל על קבוצת האימון- 1 (כצפוי, מאחר והנקודה הכי קרובה לנקודה ב1-NN היא הנקודה עצמה).

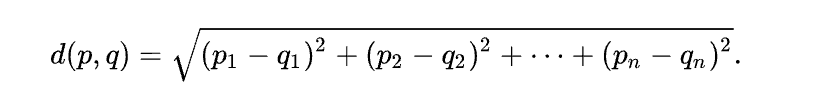
מידת הדיוק של המודל על קבוצת המבחן – 0.664

(Q9)



מידת הדיוק של המודל על קבוצת האימון- 1 (כצפוי, מאחר והנקודה הכי קרובה לנקודה ב1-NN היא הנקודה עצמה).

מידת הדיוק של המודל על קבוצת המבחן – 0.756

נורמליזציה של הנתונים חשובה לאלגוריתם הknn מאחר ואלגוריתם הknn מחפש את השכנים הקרובים ע״י פונקציית מרחק , ספציפית במקרה של התרגיל מרחק אוקלידי, וזוהי פונקציה שרגישה להבדלים בקנה המידה בו מודדים כל פיצ׳ר. ניזכר כי מרחק אוקלידי בין 2 נקודות נמדד באופן הבא: כאשר .

מהנוסחה ניתן לראות שפי׳צרים הנעים על טווח ערכים רחב (קנה מידה גדול יותר) יתרמו יותר לגודל המרחק האוקלידי מאחר וההפרשים בפיצ׳רים האלו יהיו גדולים יותר, על אף שאינם בהכרח חשובים יותר מפיצ׳רים אחרים.

כתוצאה מכך, פיצ׳רים אלו הופכים לדומיננטיים ומאפילים על פיצ׳רים עם קנה מידה נמוך.

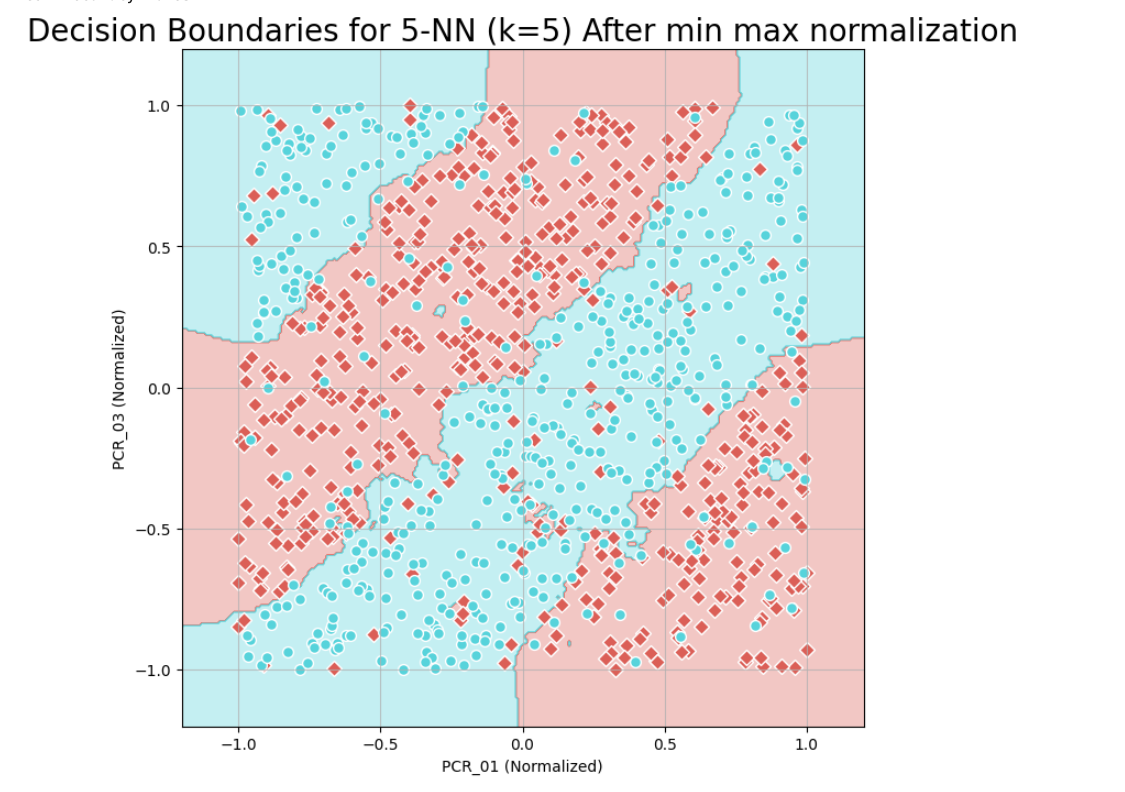
בשאלה 8, לפני שביצענו נורמליזציה על הנתונים, ניתן לראות שגבולות ההחלטה של knn נראים מוזר(פסים אופקיים כחולים על כל אזורי ההחלטה).

הפסים הכחולים האלה קשורים לעובדה שטווח הערכים של pcr\_03 [-100,100] וטווח הערכים של pcr\_01 [-2,2] כלומר פיצ׳ר pcr\_03 הרבה יותר דומיננטי בחישוב המרחק האוקלידי. זה גורם לתופעת ה״פסים האופקיים הכחולים״ גם באזורים בהם ריבוי נקודות אדומות.

מאחר ולפיצ׳ר pcr\_03 קנה מידה גדול יותר, מקבלים שעבור כל נקודה – הנקודות הקרובות יותר אליה הן נקודות בציר האופקי (שם הערך של pcr\_03 קבוע בין כל השכנים ולכן המרחק האוקלידי קטן יותר) ואילו הנקודות בציר האנכי רחוקות יותר (אפילו הפרשים קטנים בין ערכי pcr\_03 הן בקנה מידה גדול יותר). כתוצאה מכך , יש פסים כחולים אופקיים (נקודה כחולה רחוקה בציר האופקי היא דווקא יותר קרובה אוקלידית מאשר נקודות אדומות בציר האנכי).

לאחר שעשינו נרמול, קנה המידה של שני הפיצ׳רים היה באותה קנה מידה – [1,-1] ולכן לא היה פיצ׳ר יותר דומיננטי מהשני. ונראה אכן שקיבלנו שאזורי ההחלטה של מסווג הknn הרבה יותר קרובים למה שהיינו מצפים (מאוד קרובים למה שקיבלנו בpair plot).

(Q10)



מידת הדיוק של המודל על קבוצת האימון- 0.879

מידת הדיוק של המודל על קבוצת המבחן –0.852

**ההשפעה של k על אזורי ההחלטה של אלגוריתם knn הם :**

\*k נמוך גורם לoverfitting- כאשר מספר השכנים קטן, כפי שראינו עבור k=1, יש התאמה של המודל לכל הנקודות בקבוצת האימון, אפילו לנקודות הנחשבות רעש, outliert. כך ניתן לראות בתמונה של אזורי ההחלטה עבור k=1, אזורים קטנים כחולים סביב נקודות כחולות , גם כאשר כל הסביבה אדומה ולהפך.

\*כאשר k גדל, אזורי ההחלטה נהיים ״חלקים יותר״, פחות רגישים לרעש ולנקודות בודדות. K גדול יותר משפר את יכולת ההכללה של המודל. ניתן לראות שמידת הדיוק של knn עבור k=5 גדולה יותר (עם זאת מידת הדיוק על קבוצת האימון ירדה).

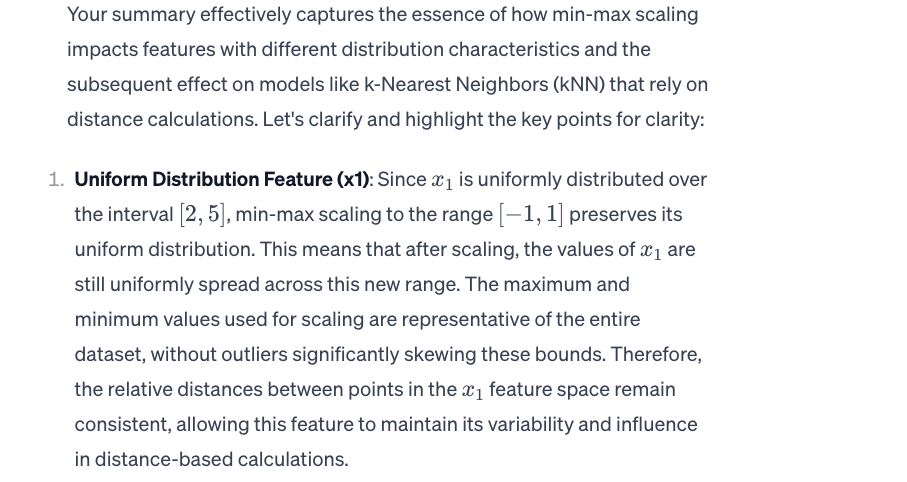
עם זאת, עבור k גדול מדי נקבל שאזורי ההחלטה כבר מתחילים להטשטש (הפשטה גדולה מדי של המודל וחוסר יכולת שלו לתפוס את התבניות בdata , underfitting מאחר וכאשר המסווג יחליט לסווג נקודה הוא יתחשב בתיוג של נקודות רחוקות מדי מהנקודה.

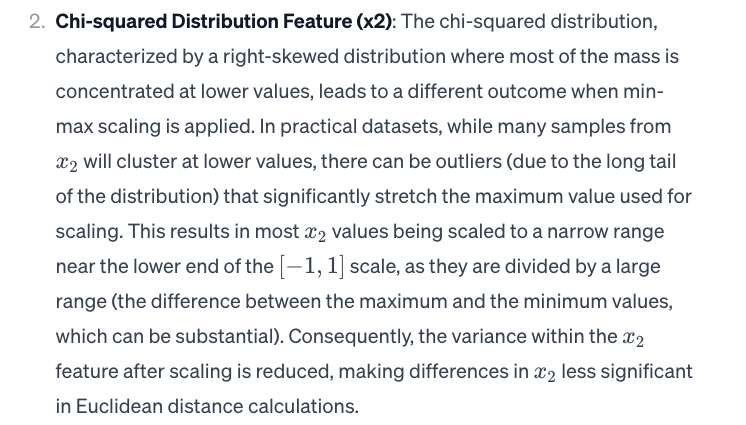
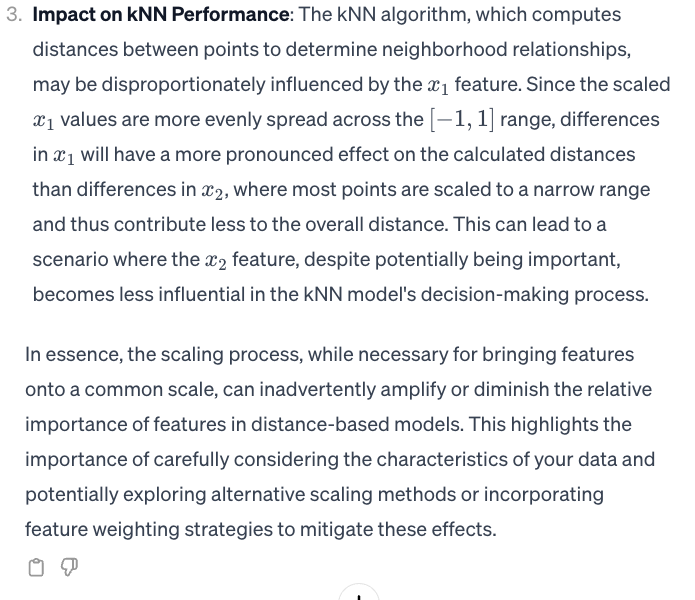
(Q11)

בקצרה-

מאחר והפיצ׳ר x1 מתפלג באופן אחיד על קטע [2,5] , גם לאחר min-max scale , הפיצ׳ר x1 החדש יתפלג גם באופן אחיד בין [1,-1] , כלומר גם אם נדגום n דגימות, הmax ,min לא יהיו outliers ביחד לשאר הנקודות.

לעומת זאת , רוב המסה של ההתפלגות כי בריבוע מרוכזת בטווחים נמוכים [ההסתברות לקבל ערך קטן שווה ל3 היא כמעט 0.8) מה שיכול לגרום לכך שעבור סדרה של הרבה דגימות – רוב הדגימות יהיו מרוכזות בערכים נמוכים אך מעט דגימות יהיו גדולות (outliers) ויביאו לכך שהmaximum גדול. כתוצה מכך בנירמול min-max רוב הנקודות יהיו מרוכזות בטווח קטן , קרוב ל0 (מאחר ומחלקים בהפרש בין המקסימום למינימום שככל הנראה יהיה גדול). דבר זה יגרום לכך שהפיצ׳ר x2 יהיה פחות חשוב במונחים אוקלידים, (בציר של x2 המרחק בין הפיצ׳רים יהיה קטן מאד וכמעט חסר חשיבות בחישוב המרחק האוקלידי, ואילו הפיצ׳ר x1 שדגימותיו מפוזרות יחסית אחיד , יהיה יותר דומיננטי החישוב המרחק האוקלידי. ושוב תקרה אותה תופעה שהסברנו עליה בשאלה 9.





(Q12)

מאחר ויש 8 סוגים שונים של סוגי דם (A+,A-,AB+,AB-,B+,B-,O+,O-) נצטרך 8 קטגוריות בוליאניות. (כעקרון מספיק גם להשתמש ב7 קטגוריות , כך שאם הערך עבור כל הפיצ׳רים האלה false זה אומר שהפיצ׳ר שייך לקטגוריה השמינית)

(Q13)

כן , אפשר להפיק מידע מהפיצ׳ר סימפטומים כדי לסייע בניבוי.   
כאשר חקרנו את הדאטה, שמנו לב שניתן לחלץ מהטקסט החופשי של כל חולה קבוצת של סימפטומים שונים ולייצר עבור כל סימפטום כזה פיצ׳ר בוליאני שהוא חיובי במידה ולחולה יש את הסימפטום הזה.

|  |  |
| --- | --- |
| פיצ׳ר לאחר טרנספומציה | פיצ׳ר מקורי |
| הוחלט להסירו מאחר ולא מסייע לפרדיקציה (מזהה ייחודי של דגימה בדאטהסט) | paitent\_id |
| שונה לפיצ׳רים בוליאנים בtask d | blood type |
| שונה לפיצ׳רים בוליאניים כפי שתיארנו | symptoms |
| תיוגים של M+F שונו ל0 ו1 | sex |
| פוצל ל Latitude, Longtitude | current\_location |
| פוצל לפיצ׳רים נומריים של שנה,יום,שבוע | Pcr\_date |