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

```
[8] virus_dataset = pd.read_csv(filename, header=0)

[9] virus_dataset.shape

    (1250, 26)
```

(Q2)
פלט:

```
virus_dataset['conversations_per_day'].value_counts()

3     224
2     215
4     190
5     156
6     111
1     104
8      72
7      60
9      39
10     23
11     19
12     12
13      9
14      6
17      4
15      2
16      2
19      1
22      1
Name: conversations_per_day, dtype: int64
```

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.

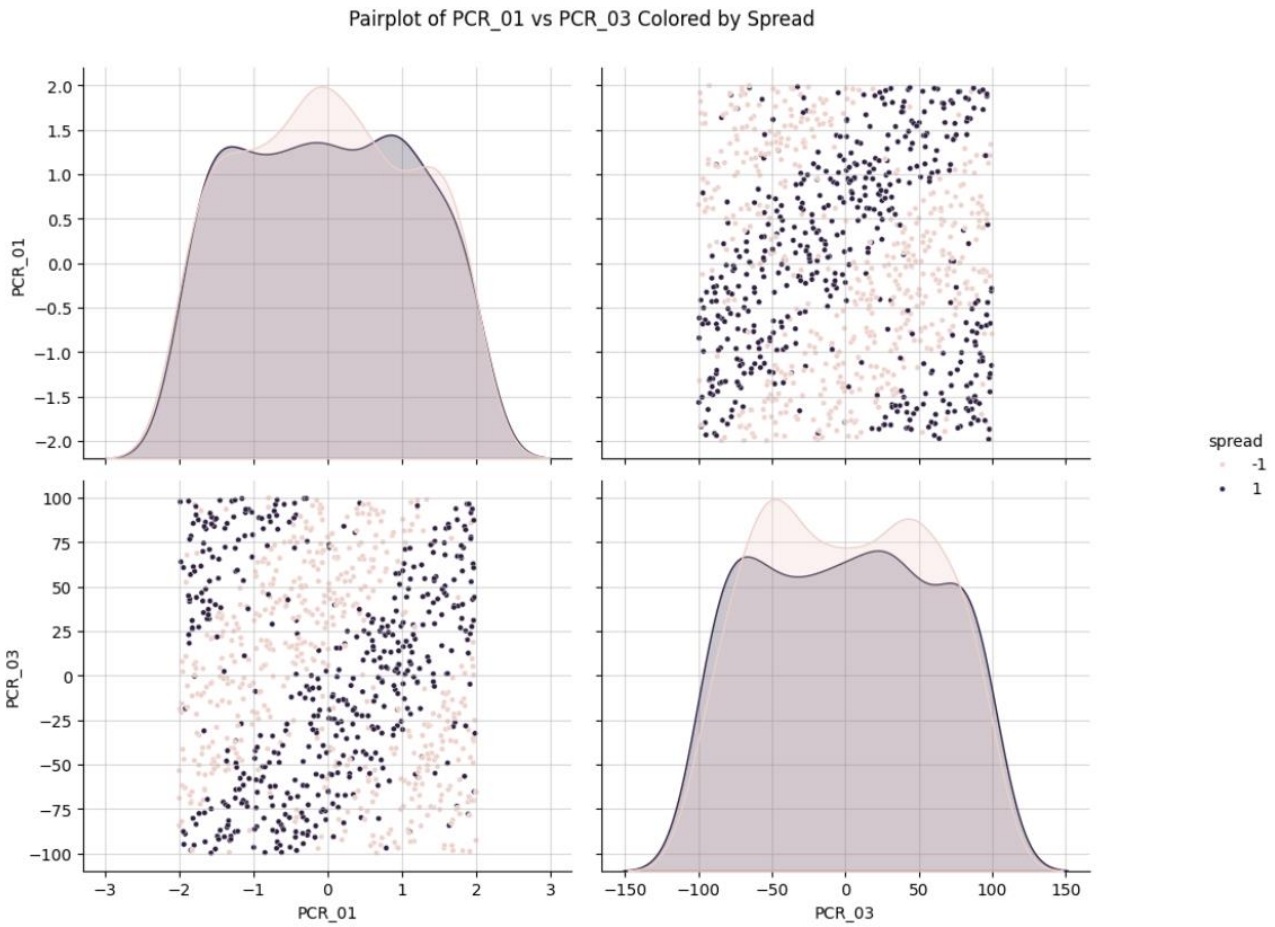
(Q3)

Feature name	Description	Type
patient_id	Incremental id of the patient	Other
age	Age of the paitent	Ordinal
sex	Sex of the patient (Male/Female)	Categorical
weight	Weight of the patient in kilograms	Continuous
blood_type	The blood type of the patient (O/A/AB/B and +/-)	Categorical
current_location	Current location of the patient (latitude and longitude).	Other
num_of_siblings	The number of siblings the patient has	Ordinal
happiness_score	A Score describing the patient's level of happiness	Ordinal
household_income	A categorical representation of the patient's household income (might be division to economy classes)	Ordinal
conversations_per_day	Average number of conversations the patient has everyday	Ordinal
sugar_levels	The sugar level measurements for the patient	Ordinal
sport_activity	The patient's level of sport activity on scale from 0-5	Ordinal
symptoms	Textual description of any symptoms the patient may have reported	Other
pcr_date	The date in which a PCR test was conducted	Other
PCR_xx	Numerical results of PCR Tests. Could be Measurements of different genetic sequences or pathogens.	Continuous

(Q4)

חשוב להשתמש באותו פיצול כאשר אנו מבצעים את תהליך ניתוח המידע מאחר ואנחנו רוצים שיהיה ביכולתנו לשחזר את התוצאות שקיבלנו בניסוי 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 עשוים להיות <u>ביחד</u> שימושיים לחיזוי spread מאחר וניתן לראות כי הדאטה כמעט פריד ל4 אזורים שונים, כאשר בכל אזור יש הרבה דוגמאות מאותה מחלקה.
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_03 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.

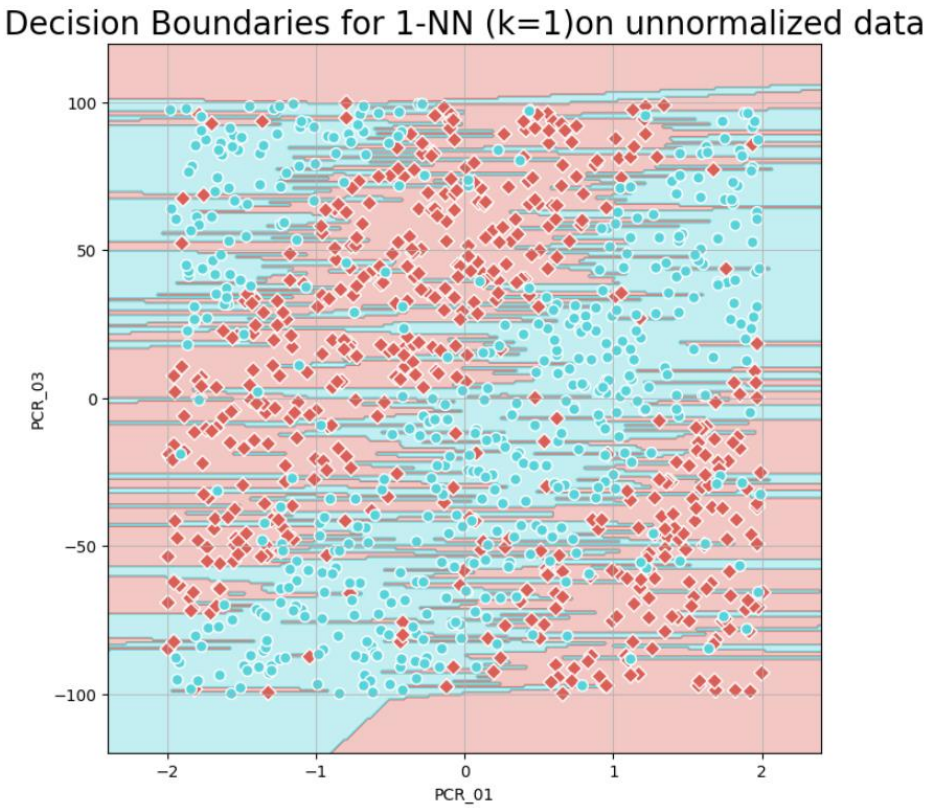
(Q7)

<u>זמן ריצה של פונקציית החיזוי על נקודה אחת P מממד d :</u>

- <u>שלב 1- חישוב המרחקים בין הנקודה p לשאר הנקודות:</u> 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.
- <u>שלב 2- חיפוש k השכנים הקרובים ביותר לנקודה p:</u> 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).
- <u>שלב 3: חישוב הפרדיקציה ע״פ החלטת הרוב:</u> 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).

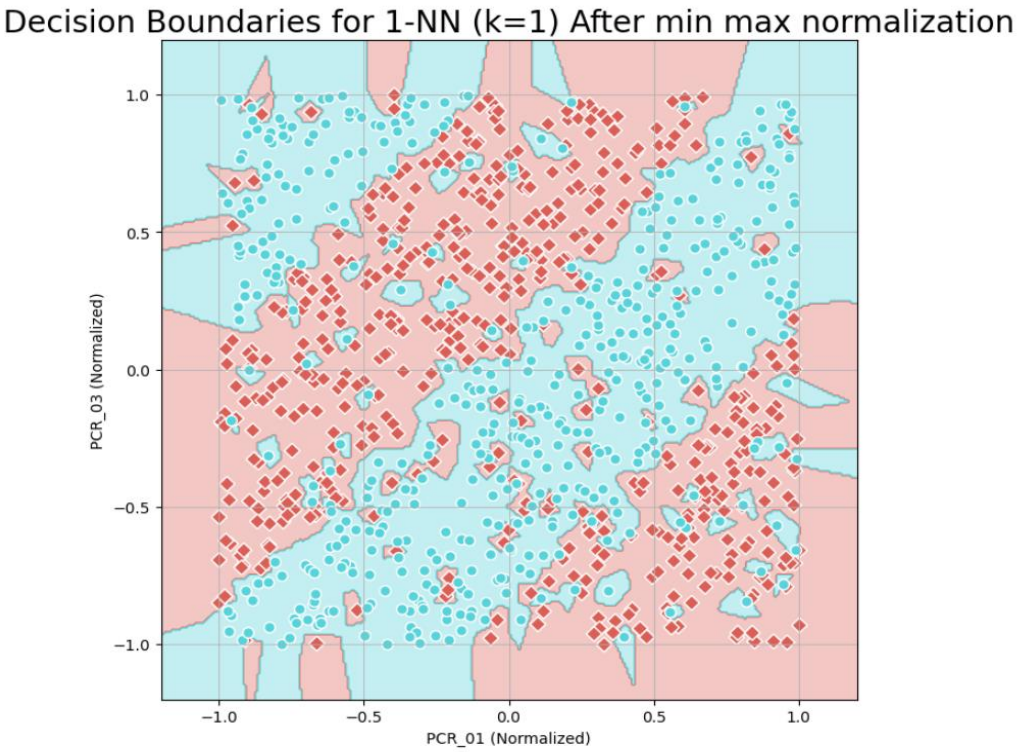
(Q8)



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:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}.$$

כאשר $p, q \in R^n$.

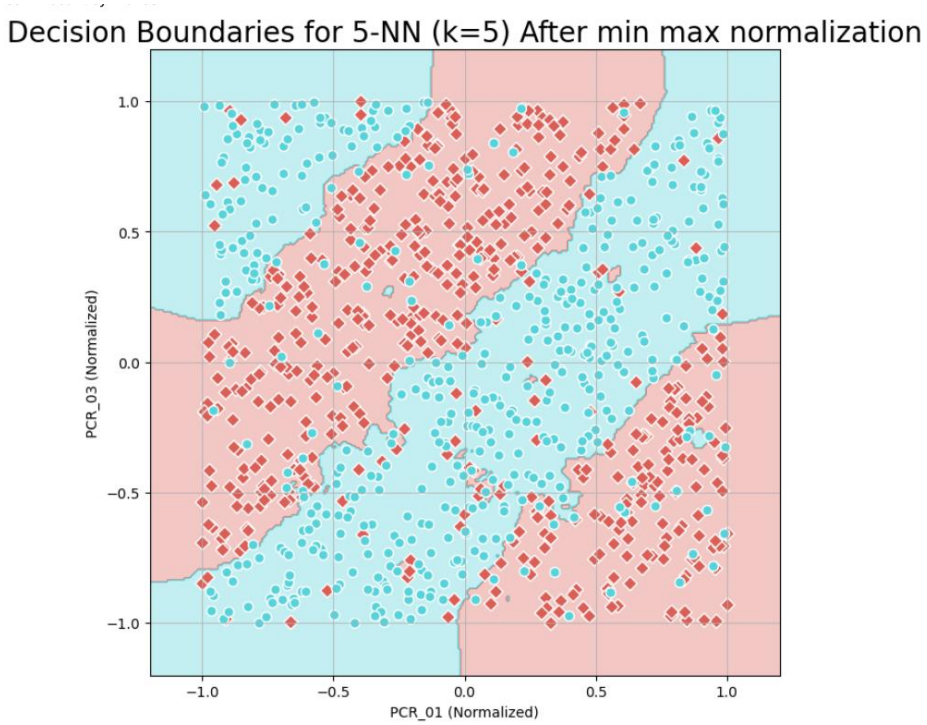
מהנוסחה ניתן לראות שפי׳צרים הנעים על טווח ערכים רחב (קנה מידה גדול יותר) יתרמו יותר לגודל המרחק האוקלידי מאחר וההפרשים בפיצ׳רים האלו יהיו גדולים יותר, על אף שאינם בהכרח חשובים יותר מפיצ׳רים אחרים.
כתוצאה מכך, פיצ׳רים אלו יותר משפיעים על המרחק האוקלידיs.

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 שיכולות להיחשב כרעש, outlier. 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)
מאחר והפיצ'ר 1 מתפלג באופן אחיד על קטע [2,5] , גם לאחר min-max scale , פיצ'ר זה יתפלג גם באופן אחיד בין [-1,1] .
לעומת זאת, ההתפלגות של פיצ'ר 2, כי בריבוע, היא התפלגות לא חסומה (כלומר המשתנה המקרי יכול לקבל כל ערך) אך 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 דגימות – רוב הדגימות יהיו מרוכזות בערכים נמוכים אך מעט דגימות יהיו בערכים גדולים.
כתוצה מכך בנירמול min-max רוב הנקודות יהיו מרוכזות בטווח קטן, קרוב ל0, דבר זה יגרום לכך שהפיצ'ר 2 יהיה פחות חשוב במונחים אוקלידים, בציר של פיצ'ר 2 המרחק בין הדגימות יהיה קטן מאד וכמעט חסר חשיבות בחישוב המרחק האוקלידי, ואילו הפיצ'ר 1 שדגימותיו מפוזרות יחסית אחיד, יהיה יותר דומיננטי בחישוב המרחק האוקלידי ותקרה אותה תופעה בKNN כמו שראינו בשאלה 9.

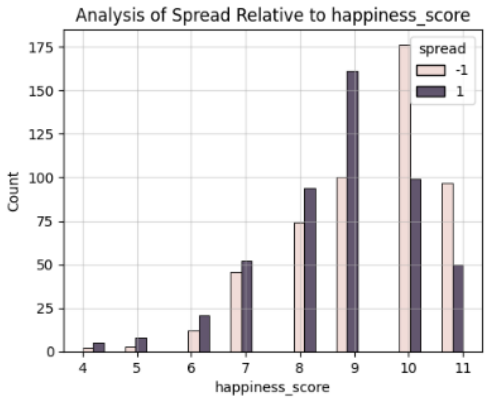
(Q12)
מאחר ויש 8 סוגים שונים של סוגי דם (A+,A-,AB+,AB-,B+,B-,O+,O-) נצטרך 8 קטגוריות בוליאניות.
(כעקרון מספיק גם להשתמש ב7 קטגוריות , כך שאם הערך עבור כל הפיצ'רים האלה false זה אומר שהפיצ'ר שייך לקטגוריה השמינית)

(Q13)
כן , אפשר להפיק מידע מהפיצ'ר סימפטומים כדי לסייע בניבוי.
כאשר חקרנו את הדאטה, שמנו לב שניתן לחלץ מהטקסט החופשי של כל חולה את מספר הסימפטומים עבור כל חולה ולכן הוספנו עמודה חדשה במקום עמודת הסימפטומים שסופרת כמה סימפטומים היו לכל חולה (ייתכן גם 0).
עבור כל חולה פיצלנו את המחרוזת של הפיצ'ר symptoms לפי ';' וספרנו כמה סימפטומים קיבלנו.

פיצ׳ר מקורי	פיצ׳ר לאחר טרנספומציה
paitent_id	הוחלט להסירו מאחר ולא מסייע לפרדיקציה (מזהה ייחודי של דגימה בדאטהסט)
sex	שונו ל0 ו-1, במקום F ו-M, כלומר שונה לnumeric פיצ'ר.
blood type	שונה לפיצ׳רים בוליאנים בd task
current_location	פוצל ל Latitude, Longitude לערכים מספריים.
symptoms	שונה לפיצ׳רים בוליאניים כפי שתיארנו
Pcr_date	פוצל לפיצ׳רים נומריים של שנה,יום,שבוע

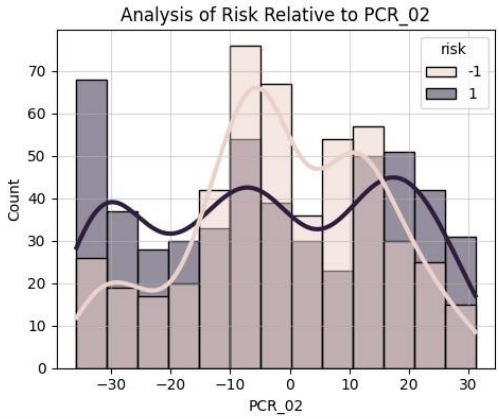
(Q14)

התכונה שבחרנו בשביל לנבא את המטרה spread היא happiness_score. בחרנו בתכונה זו מאחר שראינו כי בטווח של happiness_score 4-9 יש יותר מטופלים מקטגוריית spread=-1 (כלומר יש יותר הסתברות להיות בקטגוריה זו עם ציון הhappiness_score הוא בין 4-9) ולעומת זאת בטווח בין 10-11 יש יותר אנשים מקטגוריית spread=1 (כלומר יש יותר הסתברות להיות בקטגוריה זו אם ציון happiness הוא בין 10 ל11, או גדול מ9).



(Q15)

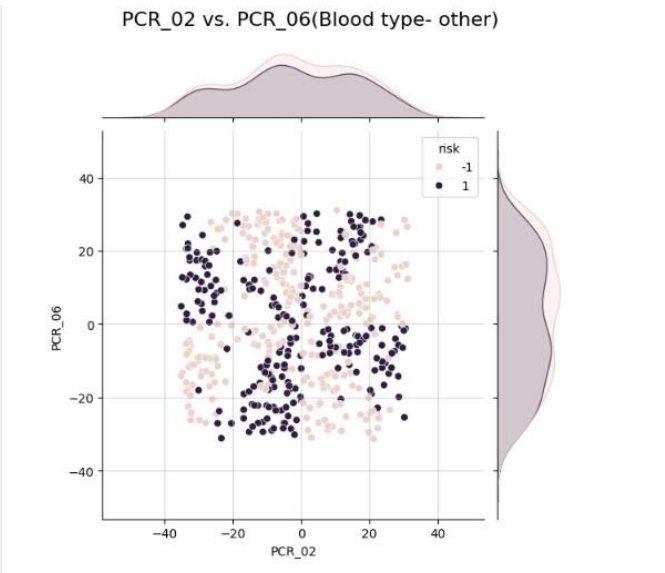
התכונה שבחרנו בשביל לנבא את המטרה Risk היא PCR_02. זאת משום שניתן לראות כי בטווח [-35 to -15] ובטווח [15 to 30] יש יותר סיכוי להיות עם risk=1, ואילו בטווח [-15 to 15] יש יותר סיכוי להיות עם risk=-1

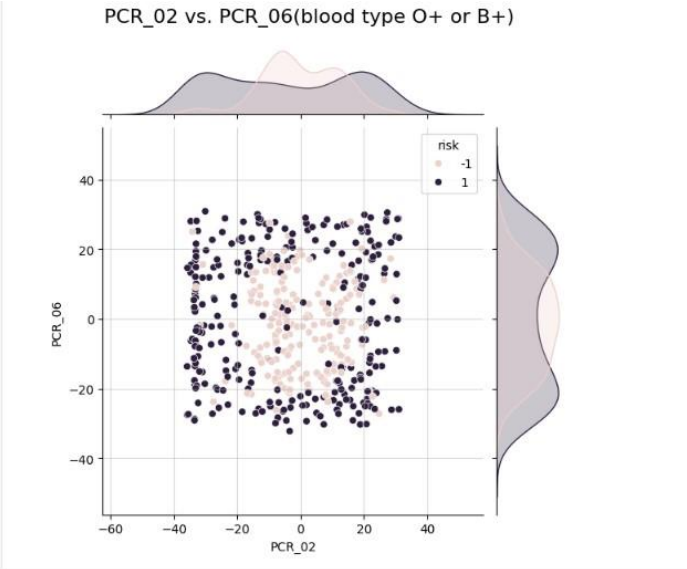


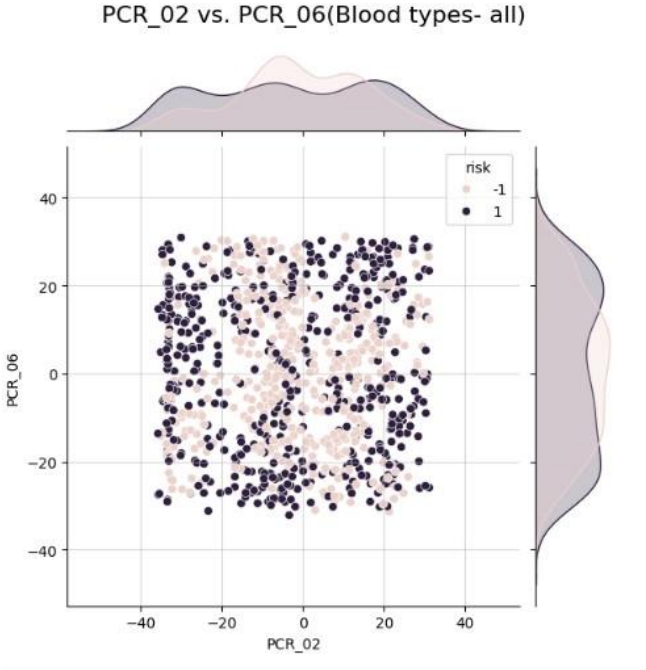
(Q16)

בחרנו את שתי התכונות : PCR_02 ו-PCR_06 . בחרנו תכונות אלו מאחר ורק עבור זוג פיצ׳רים זה ראינו כי קיימת הפרדה (עד כדי ״רעש״ קל) בין המחלקות השונות של risk.
ניתן לראות כי ההפרדה בין הנקודות אינה לינארית עבור 2 הקבוצות שקיבלנו לאחר חלוקה, לפי special property.
עבור שאר הזוגות, לא קיבלנו הפרדה בין הנקודות אלא הן היו נראות מעורבות.

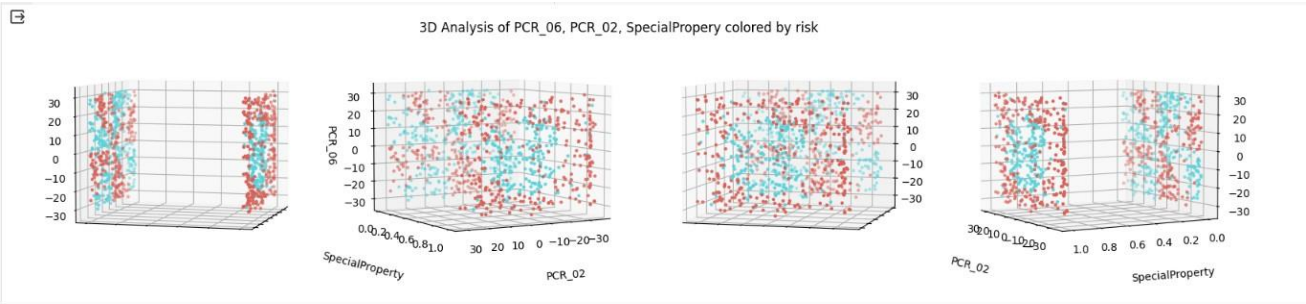
(Q17)







(Q18)

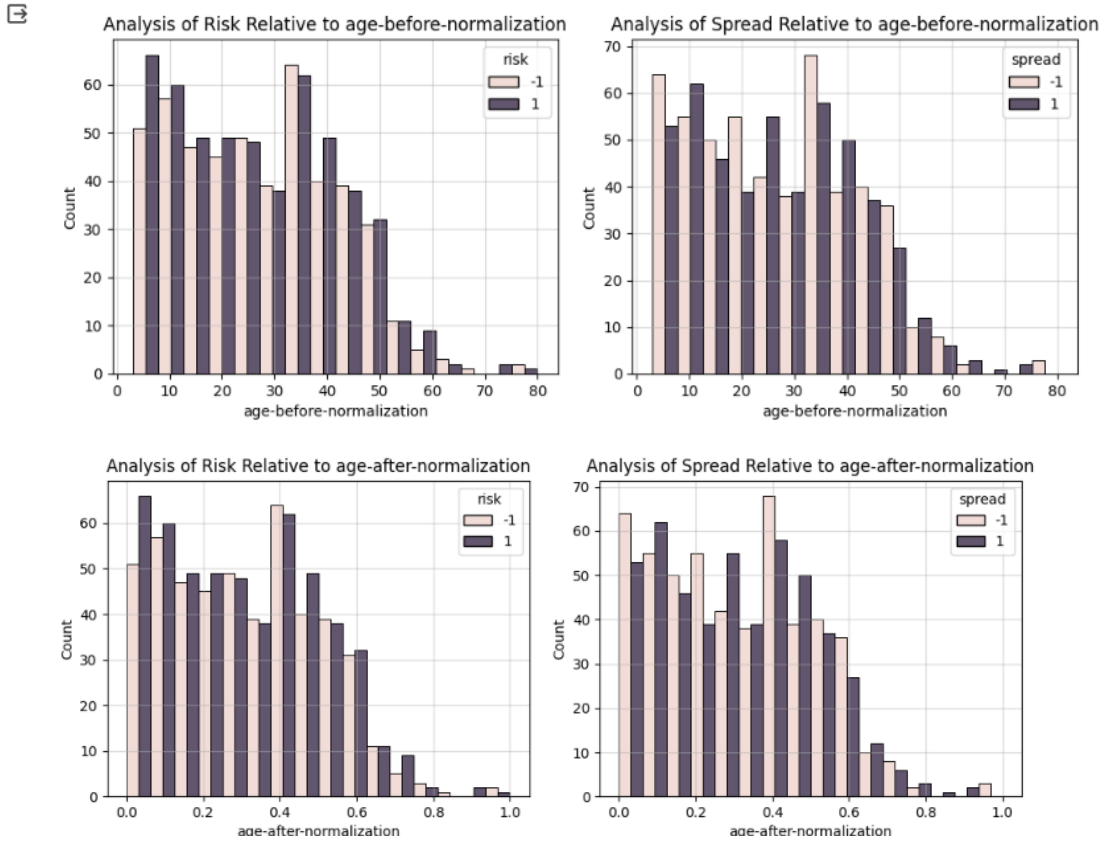


(Q19)
לעץ החלטה מעומק 3 תהיה יכולת מוגבלת להתאים את עצמו לקבוצת האימון. עץ החלטה מעומק 3 יכול להפריד את קבוצת האימון לכל היותר ל8 קבוצות שונות במרחב הפי׳צרים. עם זאת, על פי הplot ניתן לראות שהדאטה אומנם פריד, אך הוא אינו פריד לינארית ויש יותר מ8 אזורי החלטה שונים של קטגוריית risk.

(Q20)
עץ החלטה מעומק 30 יצליח להתאים אל הנתונים מקבוצת האימון בצורה טובה יותר. לעץ החלטה מעומק 30 יש יכולת להפריד את הדאטה ל 2^30 קבוצות שונות ולכן הוא יכול להגיע להפרדה מדויקת מאד של קבוצת האימון. עם זאת, עץ בעומק כזה עשוי להביא גם להתאמת יתר לנתוני האימון וoverfitting (עשויים להיווצר אזורי החלטה על סמך רעשים)

(Q21)
מודל knn עם k=1 לא יוכל להתאים לקבוצת האימון בצורה טובה. נשים לב כי טווח הערכים שמשתני הPCR יכולים לקבל הרבה יותר רחב מאשר טווח הערכים של special property (0 או 1). כתוצאה מכך, ההשפעה של משתני הpcr על גודל המרחק האוקלידי תהיה הרבה יותר דומיננטית, יש להם יותר משקל. שינויים קטנים בערכים של הpcr יכולים להגדיל יותר את גודל המרחק האוקלידי מאשר המשתנה הבוליאני ולכן מודל הknn עשוי להעדיף נקודות קרובות ביותר מקטגוריה הפוכה של special property ועם ערכי pcr כמה שיותר זהים.
לפיכך, מודל זה יפעל בניגוד למה שראינו בסעיף הקודם, שעדיף לפצל את הנתונים לפי special property ואז הנתונים פרידים (כי בעצם המודל שלנו יעדיף לבחור לנקודה מסוימת נקודה קרובה ביותר מהקטגוריה השנייה של special property).

(Q22)



(Q23)

התשובות בשאלה 19 ובשאלה 20 לא ישתנו מאחר ועצים אינם רגישים לscale של הפיצ׳רים (כלל ההחלטה המתקבל מעצים אינו נסמך על מרחק בין נקודות) ועל כן התשובות שלנו לשאלות אלו לא ישתנו.

התשובות בשאלה 21 תשתנה מאחר וknn אכן רגיש לקנה המידה של המשתנים כפי שהסברנו בשאלה זו, לאחר נרמול הנתונים , משתני הPCR יהיו בקנה מידה דומה לזה של המשתנה הבוליאני special property ולכן משתני הPCR ישפיעו פחות על גודל המרחק האוקלידי לעומת קודם. לפיכך אנו נקבל נקודות קרובות ביותר שונות, שמושפעות מ3 הפיצ׳רים במידה יותר אחידה. מידת הדיוק של knn אמורה לעלות לאחר נרמול הפיצ׳רים.

(Q24)

נציין שהשיקולים שלנו לבחירת שיטת נרמול בשאלה הזו היו:

- אם ההתפלגות של ערכי המשתנה היא מפוזרת על הטווח שהיא נמצאת בו, מזכירה התפלגות אחידה – עדיף להשתמש בשיטת min-max. זאת מאחר ושיטה זו מעבירה את הערכים לטווח חסום, וכאשר ההתפלגות היא אחידה, לאחר הטרנספורמציה ערכי המשתנה יהיו מפוזרים על כלל הטווח בין [0,1] ולא יהיו מצומצמים באזור קטן בטווח הזה. ראינו בסעיף 11, שכאשר מעבירים משתנה עם התפלגות צפופה משמאל למשל לטווח [0,1] , רוב הערכים עוברים לטווח מצומצם מאד בין 0 ל1 , מאחר וערכי המקסימום קיצוניים לעומת רוב הערכים בטווח.
- אם ההתפלגות של ערכי המשתנה מזכירה התפלגות נורמלית/התפלגות מצודדת אחרת שאינה חסומה אז עדיף להשתמש בשיטת נרמול standard scaler כי לווא דווקא שנרצה לדחוס טווח אינסופי של ערכים לתוך טווח מצומצם בין 0 ל1.
- אנו לא יכולים באמת לדעת אם משתנים מסוימים אכן חסומים בטווח ערכים מסוים מאחר והדאטה סט שקיבלנו הוא סופי ואין לנו את ההתפלגות המדויקת של כל משתנה, על כן קבענו את רוב ההחלטות שלנו על סמך הצורה של ההתפלגות.

הטבלה בעמוד הבא.

Feature name	keep	new	Normalization method	explanation
patient_id	X	X	-	בחרנו להסיר משתנה זה מאחר וזהו מספר מייצג ייחודי לכל חולה שאין לו משמעות בחיזוי risk או spread
age	V	X	Min-max	לגיל של החולה עשויה להיות השפעה על הspread והrisk של החולה לכן בחרנו להשאיר פיצ׳ר זה. בחרנו לנרמל עם שיטה זו מאחר וההיסטוגרמה של הפיצ׳ר קרובה להתפלגות אחידה (נראה שרוב הערכים בטווח מוגדר ומפוזרים עליו)
sex	X	X		גברים ונשים עשויים להגיב אחרת לווירוס הקורונה ולכן חשוב לקחת את המין בחשבון ובחרנו להשאיר משתנה זה. מאחר וזה משתנה קטגורי- הפכנו את הערכים M,F ל0 ו1 בהתאמה
sex (after change)	V	V	-	אין צורך בנרמול מאחר והמשתנה כבר בטווח הערכים הרצוי
blood type	X	X		משתנה זה הוסר במשימה D והוחלף במשתנה special property
Special Property	V	V	-	התבקשנו לייצר משתנה זה במשימה D. משתנה זה הוא בוליאני ומקבל ערך 1 כאשר המטופל משתייך לקבוצה של סוגי דם מסוימת. באופן כללי עשוי להיות קשר בין סוג הדם לבין מחלת הקורונה ולכן יכולה להיות השפעה על משתני הניבוי. בחרנו להשאיר משתנה זה. אין צורך לנרמל משתנה זה מאחר והוא בטווח הערכים הרצוי.
current_location	X	X		משתנה זה הוסר והוחלף בשני משתנים – longitude and latitude (המשתנה לא היה נומרי בצורתו המקורית אז החלטנו להפרידו ל2 משתנים רציפים)

Latitude	V	V	Standardization	חלק מקואורדינטת המיקום של האדם, למיקום הגיאוגרפי של האדם עשויה להיות השפעה על מידת הspread והrisk(אזור מדבק יותר/וירוס מסוכן יותר) ולכן בחרנו להשאיר משתנה זה. ההתפלגות של משתנה זה מזכירה התפלגות לא חסומה עם צידוד לימין לכן בחרנו להשתמש בשיטת נרמול זו.
Longitude	V	V	Standardization	חלק מקואורדינטת המיקום של האדם, למיקום הגיאוגרפי של האדם עשויה להיות השפעה על מידת הspread והrisk(אזור מדבק יותר/וירוס מסוכן יותר) ולכן בחרנו להשאיר משתנה זה. ההתפלגות של המשתנה מזכירה בצורתה התפלגות יחסית נורמלית, ולכן בחרנו להשתמש בz-score
weight	V	X	Standardization	למשקל יש השפעה על הבריאות של האדם ולכן עשויה השפעה של המשקל על המשתנים שאנו רוצים לנבא. למשתנה זה יש התפלגות הדומה להתפלגות נורמלית, ומאחר שזו ההתפלגות לא חסומה עדיף להשתמש בשיטת z-score לצורך נרמול
num_of_siblings	V	X	Standardization	למספר האחים של המטופל עשויה להיות השפעה על ניבוי הspread,לכן בחרנו להשאיר משתנה זה. ההיסטוגרמה של משתנה זה מזכירה היסטוגרמה של התפלגות נורמלית עם צידוד לשמאל ולכן בחרנו להשתמש בשיטת z score
happiness_score	V	X	Standardization	למידת האושר של המטופל יש השפעה על הסביבה שהוא נמצא בה (ייתכן שיהיה עם יותר אנשים) ולכן יכולה להשפיע על משתני הניבוי. לכן בחרנו להשאירו. ההיסטוגרמה של משתנה זה מזכירה היסטוגרמה של התפלגות נורמלית עם צידוד לימין ולכן בחרנו להשתמש בשיטת z score

household_income	V	X	Standardization	לרמת ההכנסה של המטופל עשויה להיות השפעה על הסביבה בה הוא נמצא ולכן יש קשר בין משתנה זה למשתני הניבוי. לכן בחרנו להשאירו. ההיסטוגרמה של משתנה זה מזכירה היסטוגרמה של התפלגות נורמלית עם צידוד לשמאל ולכן בחרנו להשתמש בשיטת z score
Conversations _per_day	V	X	Standardization	מספר השיחות שאדם מקיים ביום עשויים להשפיע על משתנה הspread ולכן בחרנו להשאירו. ההיסטוגרמה של משתנה זה מזכירה היסטוגרמה של התפלגות נורמלית עם צידוד לשמאל ולכן בחרנו להשתמש בשיטת z score
sugar_levels	V	X	Standardization	לרמת הסוכר בדם של האדם יש קשר לבריאות האדם ,דבר שעשוי להשפיע על משתני הניבוי ולכן בחרנו להשתמש במשתנה זה. ההתפלגות של משתנה זה דומה להתפלגות נורמלית ואינה חסומה, לכן בחרנו להשתמש בשיטת z-score לצורך נרמול
sport_activity	V	X	Standardization	לרמת הפעילות הגופנית של האדם יש קשר לבריאותו ולכן עשויה להשפיע על משתנה הניבוי. בחרנו להשאיר משתנה זה. ההיסטוגרמה של משתנה זה מזכירה היסטוגרמה של התפלגות נורמלית עם צידוד לשמאל ולכן בחרנו להשתמש בשיטת z score
symptoms	X	X		בחרנו להסיר משתנה זה ולגזור ממנו משתנה אחר בשם num_symptoms
num_of_symptoms	V	V	Standardization	מספר הסימפטומים שיש למטופל יכול להשפיע הן על הspread והן על הrisk ולכן בחרנו להשאירו. ההיסטוגרמה של משתנה זה מזכירה היסטוגרמה של התפלגות נורמלית עם צידוד לשמאל ולכן בחרנו להשתמש בשיטת z score
pcr_date	X	X		בחרנו להסיר משתנה זה ולגזור ממשתנה זה שלושה משתנים אחרים: pcr_date,pcr_month,pcr_year

pcr_year	V	V	Min-Max	לשנה שבה נערכו בדיקות הPCR עשוי להיות קשר למשתני הניבוי (יכול להיות שבשנה מסוימת היה וירוס מסוכן יותר/ מדבק יותר). לכן בחרנו להשתמש במשתנה זה . ההתפלגות של משתנה זה נראית יחסית אחידה לכן בחרנו בשיטת נרמול זו
pcr_month	V	V	Min-Max	לחודש נערכו בדיקות הPCR עשוי להיות קשר למשתני הניבוי (יכול להיות שבחודש מסוים היה וירוס מסוכן יותר/ מדבק יותר). לכן בחרנו להשתמש במשתנה זה. ההתפלגות של משתנה זה נראית יחסית אחידה לכן בחרנו בשיטת נרמול זו
pcr_day	V	V	Min-Max	ליום בו נערכו בדיקות הPCR עשוי להיות קשר למשתני הניבוי (יכול להיות שנגלה שבטווח מסוים של ימים בחודש התפרץ וירוס מסוכן יותר) לכן בחרנו להשתמש במשתנה זה. ההתפלגות של משתנה זה נראית יחסית אחידה לכן בחרנו בשיטת נרמול זו
PCR_01	V	V	Min-Max	משתנה זה מייצג תוצאה של בדיקת PCR ולכן עשויה להיות לו השפעה על משתנה הניבוי. בחרנו להשאירו. על פי הניתוח שעשינו למשתנה זה, ההתפלגות שלו מזכירה התפלגות אחידה בטווח חסום ולכן בחרנו להשתמש בשיטת נירמול זו
PCR_02	V	X	Min-Max	משתנה זה מייצג תוצאה של בדיקת PCR ולכן עשויה להיות לו השפעה על משתנה הניבוי. בחרנו להשאירו. על פי הניתוח שעשינו למשתנה זה, ההתפלגות שלו מזכירה התפלגות אחידה בטווח חסום ולכן בחרנו להשתמש בשיטת נירמול זו
PCR_03	V	X	Min-Max	משתנה זה מייצג תוצאה של בדיקת PCR ולכן עשויה להיות לו השפעה על משתנה הניבוי. בחרנו להשאירו. על פי הניתוח שעשינו למשתנה זה, ההתפלגות שלו מזכירה התפלגות אחידה בטווח חסום ולכן בחרנו להשתמש בשיטת נירמול זו
PCR_04	V	X	Standardization	משתנה זה מייצג תוצאה של בדיקת PCR ולכן עשויה להיות לו השפעה

				על משתנה הניבוי. בחרנו להשאירו. על פי הניתוח שעשינו למשתנה זה, ההתפלגות שלו מזכירה התפלגות נורמלית שאינה חסומה ולכן בחרנו להשתמש בשיטת נרמול זו
PCR_05	V	X	Min-Max	משתנה זה מייצג תוצאה של בדיקת PCR ולכן עשויה להיות לו השפעה על משתנה הניבוי. בחרנו להשאירו. על פי הניתוח שעשינו למשתנה זה, ההתפלגות שלו מזכירה התפלגות אחידה ולכן בחרנו להשתמש בשיטת נרמול זו
PCR_06	V	X	Min-Max	משתנה זה מייצג תוצאה של בדיקת PCR ולכן עשויה להיות לו השפעה על משתנה הניבוי. בחרנו להשאירו. על פי הניתוח שעשינו למשתנה זה, ההתפלגות שלו מזכירה התפלגות אחידה בטווח חסום ולכן בחרנו להשתמש בשיטת נירמול זו
PCR_07	V	X	Standardization	משתנה זה מייצג תוצאה של בדיקת PCR ולכן עשויה להיות לו השפעה על משתנה הניבוי. בחרנו להשאירו. על פי הניתוח שעשינו למשתנה זה, ההתפלגות שלו מזכירה התפלגות נורמלית שאינה חסומה ולכן בחרנו להשתמש בשיטת נרמול זו
PCR_08	V	X	Standardization	משתנה זה מייצג תוצאה של בדיקת PCR ולכן עשויה להיות לו השפעה על משתנה הניבוי. בחרנו להשאירו. על פי הניתוח שעשינו למשתנה זה, ההתפלגות שלו מזכירה התפלגות שאינה חסומה עם צידוד לימין ולכן בחרנו להשתמש בשיטת נרמול זו
PCR_09	V	X	Standardization	משתנה זה מייצג תוצאה של בדיקת PCR ולכן עשויה להיות לו השפעה על משתנה הניבוי. בחרנו להשאירו. על פי הניתוח שעשינו למשתנה זה, ההתפלגות שלו מזכירה התפלגות נורמלית שאינה חסומה ולכן בחרנו להשתמש בשיטת נרמול זו
PCR_10	V	X	Standardization	משתנה זה מייצג תוצאה של בדיקת PCR ולכן עשויה להיות לו השפעה על משתנה הניבוי. בחרנו להשאירו. על פי הניתוח שעשינו למשתנה זה, ההתפלגות שלו מזכירה התפלגות

נורמלית שאינה חסומה ולכן בחרנו להשתמש בשיטת נרמול זו				