תרגיל בית 1

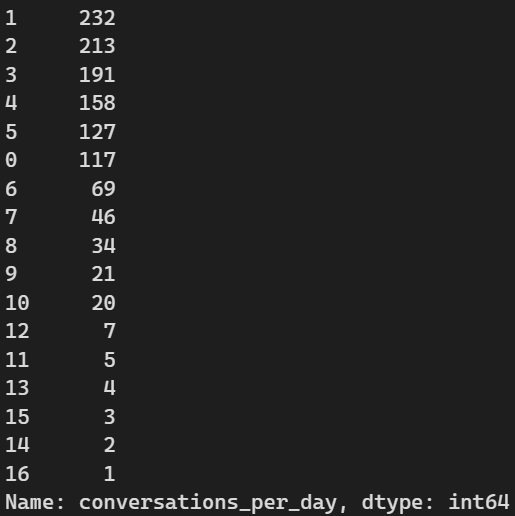
מגישים:  
אלמוג קריף 313151185  
אורי מינץ 314616897

# 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 | Other | 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 | Other | 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

Shape

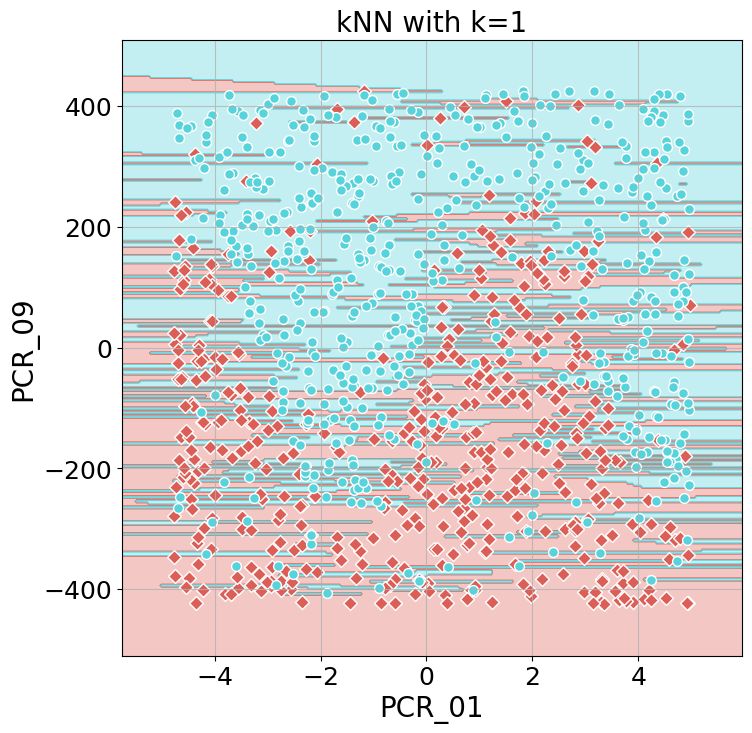
Description automatically generated with low confidence

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 נחפש מרחקים בין הנקודות, כל נקודה היא מממד ֹלכן הסיבוכיות היא וקיבלנו וקטור מרחקים בגודל שלב שני נשתמש ב argpartition על הוקטור שקיבלנו בשלב 1 סה״כ (מציאת נקודות הקרובות ביותר).  
לכן סה"כ נקבל וכיוון ש- זניח 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

הדיוק של ה-kNN על האימון הוא 1.0

הדיוק של ה-kNN על המבחן הוא 0.78

ניתן לראות שיש שיפור בתוצאות הדיוק של המבחן, כלומר הנורמליזציה שיפרה את הדיוק.

הוכחה מתמטית: נניח כי יש לנו dataset עם 2 פיטצ׳רים, ונניח שיש לנו 2 נקודות data

המרחק האוקלידי ביניהם יהיה : אם ל יש סט גדול יותר של ערכים מ אז ל תהיה השפעה גדולה יותר על המרחק האוקלידי, ולכן אם לא ננרמל המרחק האוקלידי יהיה משוחד לכיוון .

## שאלה 10

Chart, scatter chart

Description automatically generated  
הknn עבור הMinMax המנורמל עם k=7 של training הוא 0.884

הknn עבור הMinMax המנורמל עם k=7 של test הוא 0.872

עבור k=1 קיבלנו overfitting ועבור k=7 קיבלנו תוצאה טובה יותר באופן כללי.

ניתן לראות גם שבגרף עבור k=7 חלק יותר ויש בו פחות ״איים״ כי הוא פחות מושפע מדוגמאות בודדות.

## שאלה 11

לבצע נרמול של שני המאפיינים ל-[1-,1] עם MinMaxScaler הוא בעייתי בגלל תכונות ההתפלגות של

הערכים שהתפלגות זו מקבלת הם לכן, כיוון שאין חסם עליון לערך של x, ביצוע נרמול על ה-training data set יכול לגרום לבעיות, הראשונה היא שהערך המקסימלי שנמצא במידע זה אינו המקסימום האפשרי, לכן יתכן ונקבל ערכים במבחן שאחרי הנרמול יהיו גבוהים מ-1. שנית, אם נניח שנקבל ערך x גבוהה מאוד, דבר זה יגרום לכל הערכים "להתכווץ" לכיוון ה-0 ובכך המאפיין הזה ישפיע הרבה פחות על המרחק מאשר המאפיין המפולג יוניפורמי. לכן ביצוע פעולה זאת מוסיפה חשיבות גדולה יותר עבור דוגמאות קצה. ולכן אם ננסה לחשב מרחקים של השכנים הכי קרובים עבור KNN כלשהו נצטרך להשתמש בנוסחה מהצורה אך למשתנה היוניפורמי יש סט גדול רחב וגדול יותר של ערכים מהמשתנה ולכן הפתרון יהיה ״משוחד״ יותר לכיוון של המשתנה היוניפורמי בדומה להסבר של שאלה 9.

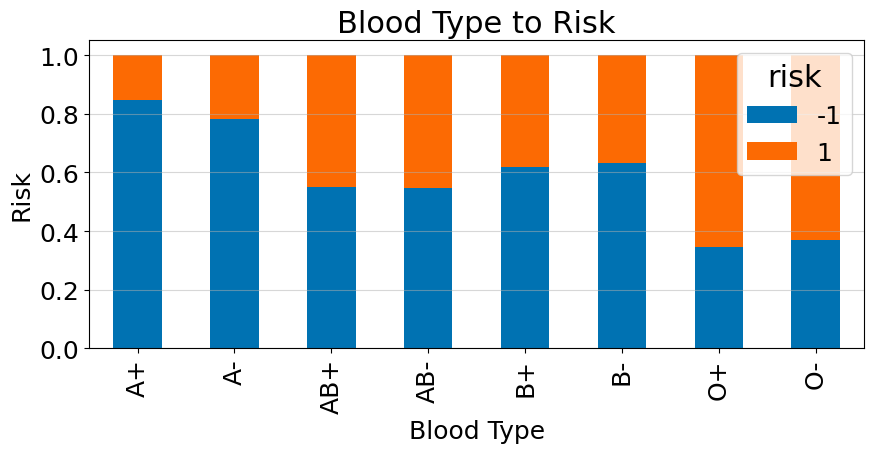
# חלק 3

## שאלה 12

נצטרך 8 boolean types עבור שמונת הקטגוריות בצורה הבאה: A+=(0,0,0,0,0,0,0,1)

A-=(0,0,0,0,0,0,1,0) , B+=(0,0,0,0,0,1,0,0) , וכו׳ וקטור בעל 8 כניסות אחת לכל סוג.

## שאלה 13



הגיוני לחלק את הקבוצות כמו שתואר בשאלה כי זה יוצר 3 קבוצות סיכון, סיכון נמוך (מתחת ל-30 אחוז), סיכון בינוני (40-60 אחוז), וסיכון גבוה (מעל 75 אחוז).

## שאלה 14

ניתן לפצל את הfeatures שהם לא numeric, פיצלנו data את:

blood\_type – שונה ל- לפי רמות סיכון בהתאם לשאלה 13.

current\_location – שונה ל2 features חדשים, אחד לפי x ואחד לפי y.

sex – שונה ל עבור female ו- עבור male

Symptoms – לכל סימפטום יצרנו features חדש שבו מסומן אם קיים לחולה הסימפטום ו- אם לא.

Pcr\_date - שונה לזמן בשניות מרגע קבלת החיסון.

## Chart, bar chart, histogram Description automatically generatedשאלה 15

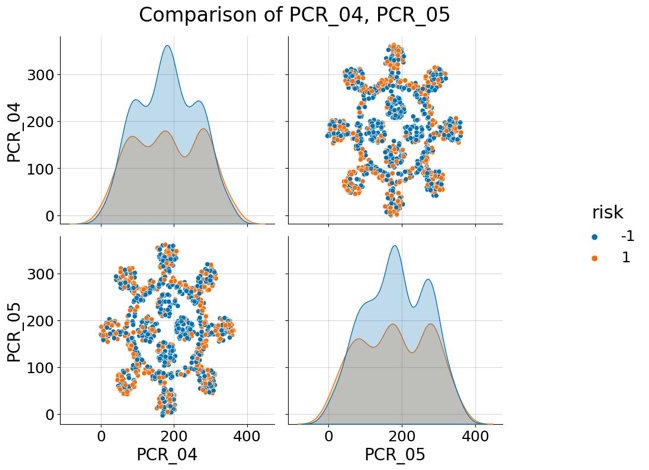
נבחר בfeature sugar\_level בגלל שכמו שניתן לראות בגרף מעל אזור ה90 האנשים בעלי spead חיובי גבוה יותר לעומת אחוז האנשים בעלי spead שלילי וניתן גם לראות כי מתחת ל90 באופן יחסי אחוז הspead השלילי גבוה יותר מאחוז ה-spead החיובי.

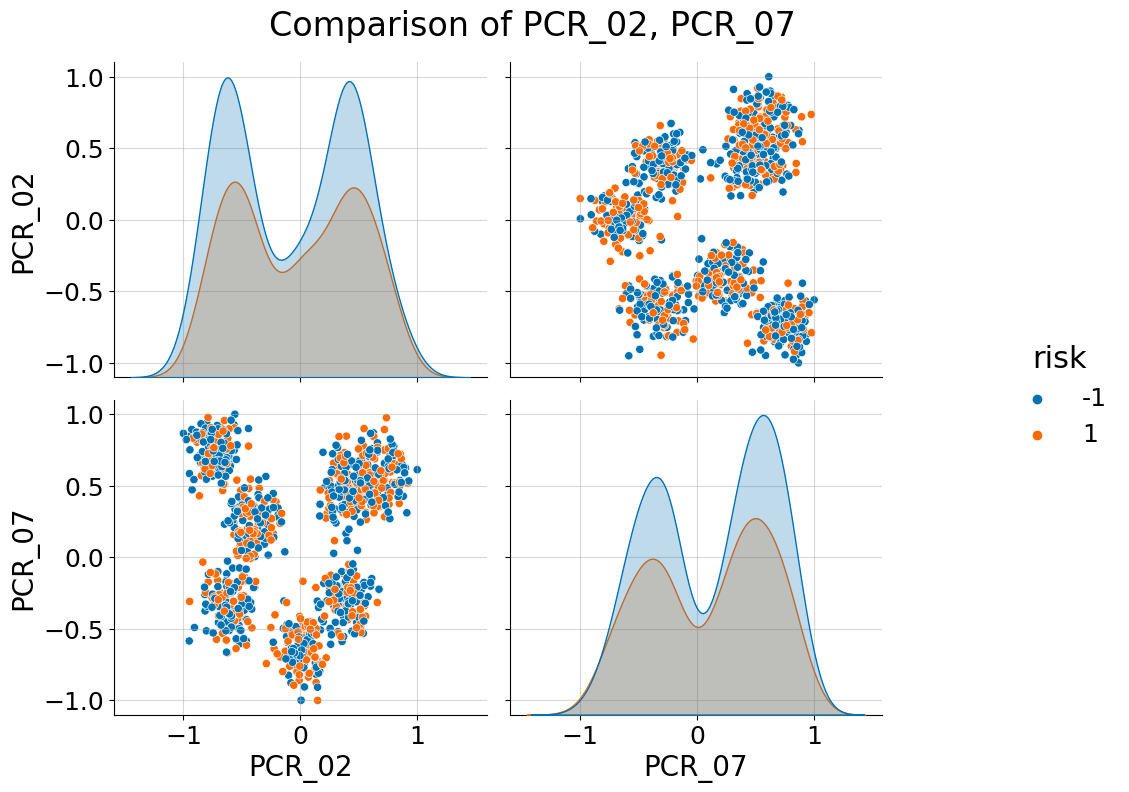
## Chart, bar chart, histogram Description automatically generated

## שאלה 16

גם כאן נבחר בfeature sugar\_level בגלל שכמו שניתן לראות בגרף מעל אזור ה75 האנשים בעלי risk חיובי גבוה יותר לעומת אחוז האנשים בעלי risk שלילי וניתן גם לראות כי מתחת ל75 באופן יחסי אחוז הrisk השלילי גבוה יותר מאחוז הrisk החיובי.

## שאלה 17

בחרנו את pcr\_04 עם pcr\_05 שנראה כמו וירוס הקורונה, ואת pcr\_02 עם pcr\_07 שנראה כמו גושים:



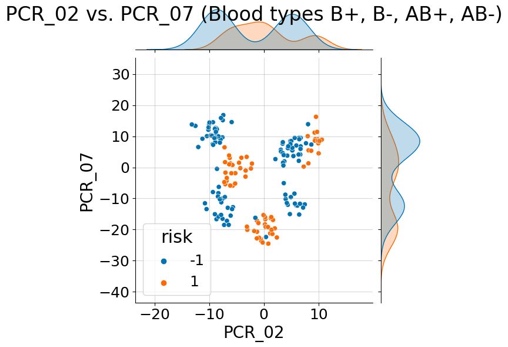
לא ניתן לזהות שום שינוי בrisk בהיקף גדול לפי 2 הדוגמאות שלעיל, ניתן לראות שלא ניתן למצוא אזור שבו החלק השולט יותר הוא 1+ או 1- (כחול כתום) עבור risk

## Chart, scatter chart Description automatically generatedChart, scatter chart Description automatically generatedשאלה 18

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

## שאלה 19

לפי התכונות שבחרנו לעיל (נסתכל לדוגמא על PCR\_07 ו- PCR\_02) ניתן לראות שקיימת חלוקה ל3 מקבצים שונים עבור ההתאמה לrisk, אם נבחר להשתמש ב- נוכל להוסיף עוד מימד (נסמנו ב-z), נגדיר לכל הנקודות בעלות סיכון נמוך {{A+,A- ערך z של , עבור הנקודות בעלות סיכון גבוה {O+,O-} נגדיר ערך z של , ועבור הנקודות בעלות סיכון ממוצע נשאר בציר ה-. ניתן לראות שאם נבצע את אלגוריתם ה- הנקודות בקבוצות הדם השונות לא ישפיעו על הסיווג. ובכך למרות שהמרחק בניהם בצירים המקוריים קרוב לאחר הוספת המימד נכון לסווג אותם בצורה יעילה.

עבור עץ החלטה, ניתן לחלק לפי סוגי הקבוצות בחלוקה הראשונה נבחר האם זה שייך לקבוצה בעלת סיכון נמוך {{A+,A- או לשאר הקבוצות. במידה והוא שייך לשאר הקבוצות נבחר החלטה שניה בעץ, סיכון גבוה {O+,O-} או לא. ובכך נוכל לחלק לכל קבוצות הדם השונות. לדוג׳ עבור ההחלטה {{A+,A- נחזיר risk=-1. ועבור סיכון בינוני נוכל לבצע עוד חלוקות, לפי PCR\_02, PCR\_07 .

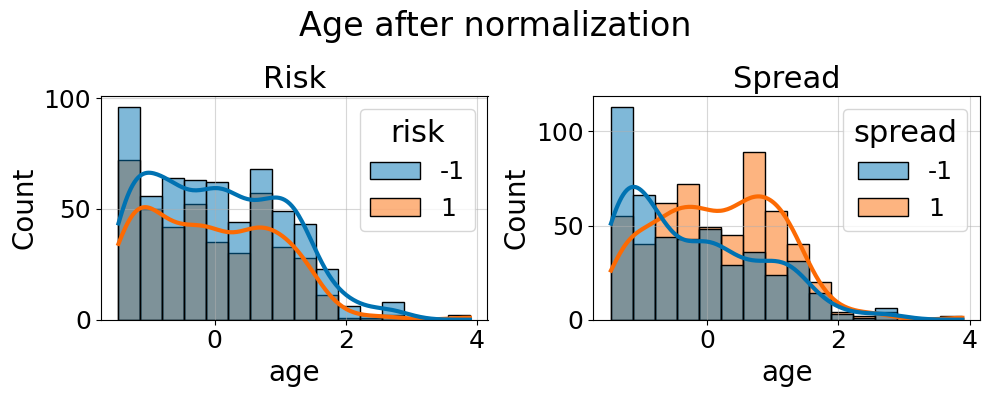
עבור מודל לינארי, כיוון שלא ניתן להוסיף מימד, לא נוכל לסווג את הנקודה בצורה טובה. ניתן לראות לדוגמא שאם נבחר את PCR\_02 ו- PCR\_07 לא נוכל לפריד בעזרת קו לינארי את ה-data בצורה מספקת.

# חלק 4

## שאלה 20

Chart, bar chart

Description automatically generated



# חלק 5

## שאלה 21

*עבור* forward feature selection *בכל איטירציה כאשר מחשבים מ- תכונות ל- תכונות עבור תכונות של ה-*dataset *הפעולה תיקח בדיקת תכונות*

חישוב של SFS במצב של backward selection עם תכונות ל- תכונות היא לכן חישוב של backward selection מ- תכונות ל- תכונות תיקח:

## שאלה 23

התכונות שנבחרו על ידי אלגו' SFS הינן: weight, PCR\_01 ו-PCR\_09. גם PCR\_01 וגם PCR\_09 נבחרו על ידנו בסעיף 6 אבל בחרנו בסעיף 15 את sugar\_level, בפתרון סעיף זה התלבטנו בינו לבין weight שכן גם weight נראה לנו מאוד אינפורמטיבי.

## שאלה 24

נרמול משפיע על התכונות שיבחרו. ראינו בשאלה 9 בבחירה של ה-kNN שנרמול התכונות הביא לדיוק יותר טוב, ובחירת נקודות קרובות אחרות. לכן נרמול של התכונות יכול להביא לשינוי של היכולת של התכונות לסווג ולכן האלגוריתם יבחר תכונות אחרות.

# חלק 6

## שאלה 25

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | שם התכונה | נשארו | חדשים | סוג נרמול | הסבר |
| 1 | patient\_id | V | X | Min max | כיוון שלפי ה-dataset ההתפלגות היא פחות או יותר יוניפורמית, עשינו נרמול זה כדי לשמור על ההתפלגות הזאת (אין מקרי קצה) |
| 2 | age | V | X | Standard scaler | כיוון שיש מקרי קצה ב-dataset כדי שהם לא ישפיעו הרבה על הנרמול עשינו את הנרמול הזה (יש מעט מאוד נתונים עבור גיל גדול מ-75 בעוד רוב ה-data נמצא מתחת ל-50) |
| 3 | sex | V | X |  | שינינו את הערכים של sex להיות ו- ולכן לא היה צריך לנרמל |
| 4 | weight | V | X | Standard scaler | כיוון שיש מקרי קצה ב-dataset כדי שהם לא ישפיעו הרבה על הנרמול עשינו את הנרמול הזה (יש מעט מאוד נתונים עבור משקל גדול מ-90 ומתחת ל-25 בעוד רוב ה-data נמצא בין 25 ל-90) |
| 5 | blood\_type | X | X |  | כיוון ש-blood\_type הכיל מחרוזות המרנו את הערכים שלו להיות בקבוצות והעברנו את זה ל-blood\_goup |
| 6 | current\_location | X | X |  | פיצלנו את current\_location לקואורדינטות ו- |
| 7 | num\_of\_siblings | V | X | Standard scaler | כיוון שיש מקרי קצה ב-dataset כדי שהם לא ישפיעו הרבה על הנרמול עשינו את הנרמול הזה (יש מעט מאוד נתונים עבור מספר גדול מ-3 ויש מקרה קצה של 7 שזה מקרה קצה רחוק) |
| 8 | happiness\_score | V | X | Standard scaler | כיוון שיש מקרי קצה ב-dataset כדי שהם לא ישפיעו הרבה על הנרמול עשינו את הנרמול הזה (יש מעט מאוד מקרים של פחות מ-5 ויש מקרים של 2) |
| 9 | household\_income | V | X | Standard scaler | כיוון שיש מקרי קצה ב-dataset כדי שהם לא ישפיעו הרבה על הנרמול עשינו את הנרמול הזה (יש מעט מאוד מקרים של יותר מ-4 ויש מקרים של 8) |
| 10 | conversations\_per\_day | V | X | Standard scaler | כיוון שיש מקרי קצה ב-dataset כדי שהם לא ישפיעו הרבה על הנרמול עשינו את הנרמול הזה (יש מעט מאוד מקרים של יותר מ-5 ויש מקרים של 15) |
| 11 | sugar\_levels | V | X | Standard scaler | כיוון שיש מקרי קצה ב-dataset כדי שהם לא ישפיעו הרבה על הנרמול עשינו את הנרמול הזה (יש מעט מאוד מקרים מחוץ לטווח של 70 עד 110 ויש דוגמאות ב150 וב-50) |
| 12 | sport\_activity | V | X | Min max |  |
| 13 | symptoms | X | X |  | הסרנו את symptoms כיוון שהוא הכיל מחרוזות ולא מספרים |
| 14 | pcr\_date | V | X | Min max | כיוון שלפי ה-dataset ההתפלגות היא פחות או יותר יוניפורמית, עשינו נרמול זה כדי לשמור על ההתפלגות הזאת (אין מקרי קצה) |
| 15 | PCR\_01 | V | X | Min max | כיוון שלפי ה-dataset ההתפלגות היא פחות או יותר יוניפורמית, עשינו נרמול זה כדי לשמור על ההתפלגות הזאת (אין מקרי קצה) |
| 16 | PCR\_02 | V | X | Min max |  |
| 17 | PCR\_03 | V | X | Min max | כיוון שלפי ה-dataset ההתפלגות היא פחות או יותר יוניפורמית, עשינו נרמול זה כדי לשמור על ההתפלגות הזאת (אין מקרי קצה) |
| 18 | PCR\_04 | V | X | Min max |  |
| 19 | PCR\_05 | V | X | Min max |  |
| 20 | PCR\_06 | V | X | Standard scaler |  |
| 21 | PCR\_07 | V | X | Min max |  |
| 22 | PCR\_08 | V | X | Standard scaler |  |
| 23 | PCR\_09 | V | X | Min max | כיוון שלפי ה-dataset ההתפלגות היא פחות או יותר יוניפורמית, עשינו נרמול זה כדי לשמור על ההתפלגות הזאת (אין מקרי קצה) |
| 24 | PCR\_10 | V | X | Min max |  |
| 25 | blood\_group | V | V |  | חילקנו את שלושת הקבוצות של blood\_group להיות מתאימים ל-0,1 ו- ולכן לא היה צריך לנרמל |
| 26 | fever | V | V |  | כיוון שזה ערך בוליאני (של ו-) לא היה צריך לנרמל, הוספנו אותו כחלק מפיצול של symptoms |
| 27 | cough | V | V |  | כיוון שזה ערך בוליאני (של ו-) לא היה צריך לנרמל, הוספנו אותו כחלק מפיצול של symptoms |
| 28 | shortness\_of\_breath | V | V |  | כיוון שזה ערך בוליאני (של ו-) לא היה צריך לנרמל, הוספנו אותו כחלק מפיצול של symptoms |
| 29 | sore\_throat | V | V |  | כיוון שזה ערך בוליאני (של ו-) לא היה צריך לנרמל, הוספנו אותו כחלק מפיצול של symptoms |
| 30 | smell\_loss | V | V |  | כיוון שזה ערך בוליאני (של ו-) לא היה צריך לנרמל, הוספנו אותו כחלק מפיצול של symptoms |
| 31 | x\_location | V | V | Standard scaler |  |
| 32 | y\_location | V | V | Standard scaler |  |