חלק 1

# שאלה 1:

בטבלה 1250 שורות ו26 עמודות

# שאלה 2:

יתכן ש- num\_of\_siblings הוא משתנה המייצג את כמות קרובי המשפחה של אותו אדם אשר חלו בקורונה. זוהי תכונה אורדינאלית כיוון שהיא בנויה מקטגוריות בדידות- כמות הקרובים. לקטגוריות אלו סדר טבעי- ניתן לסדר אותן בסדר עולה לפי כמות הקרובים.



# שאלה 3:

|  |  |  |
| --- | --- | --- |
| **שם הפיצ'ר** | **סוג** | **הסבר קצר** |
| patient\_id | ordinal | תעודת הזהות של הנבדק. |
| age | ordinal | גיל הנבדק. |
| sex | categorial | מין הנבדק. |
| weight | continuous | משקל הנבדק. |
| blood\_type | categorial | סוג הדם של הנבדק. |
| current\_location | other | המיקום הנוכחי של הנבדק בכל רגע נתון. |
| num\_of\_siblings | ordinal | מספר האחים/אחיות של הנבדק. |
| happiness\_score | ordinal | מדד אושר של הנבדק בסקאלה מ-1 עד 9. |
| household\_income | continuous | הכנסה ממוצעת של הנבדק (שנתית ובאלפי דולרים). |
| conversations\_per\_day | ordinal | כמות השיחות הממוצעת (פנים אל פנים) שהנבדק מקיים ביום. |
| sugar\_levels | ordinal | רמת הסוכר בדם של הנבדק. |
| sport\_activity | ordinal | רמת הפעילות הגופנית של הנבדק מרמה 0 עד רמה 4. |
| symptoms | other | הסימפטומים שהרגיש הנבדק.  נתון זה יכול להיות קבוצה של סימפטומים ולא רק סימפטום יחיד. |
| pcr\_date | other | תאריך בדיקת ה-PCR |
| PCR\_01 | Continuous | מאפיין 1 של בדיקת ה-PCR |
| PCR\_02 | Continuous | מאפיין 2 של בדיקת ה-PCR |
| PCR\_03 | Continuous | מאפיין 3 של בדיקת ה-PCR |
| PCR\_04 | Continuous | מאפיין 4 של בדיקת ה-PCR |
| PCR\_05 | Continuous | מאפיין 5 של בדיקת ה-PCR |
| PCR\_06 | Continuous | מאפיין 6 של בדיקת ה-PCR |
| PCR\_07 | Continuous | מאפיין 7 של בדיקת ה-PCR |
| PCR\_08 | Continuous | מאפיין 8 של בדיקת ה-PCR |
| PCR\_09 | Continuous | מאפיין 9 של בדיקת ה-PCR |
| PCR\_10 | Continuous | מאפיין 10 של בדיקת ה-PCR |

# שאלה 4:

חשוב להשתמש באותה חלוקה של הדאטה בכל פעם כדי לוודא שבאופן קבוע יש הפרדה בין הדוגמאות המשמשות לאימון המודל והדוגמאות. אם נפצל בכל פעם בצורה שונה את הדוגמאות, ייתכן שדוגמאות מסוימות ישמשו לאימון המודל ולאחר מכן גם לבדיקתו, אך אם נבחן את המודל על דוגמאות ששימשו אותו לאימון, לא נוכל לצפות כיצד יתמודד המודל עם דוגמאות אמיתיות שלא פגש בהן עדיין.

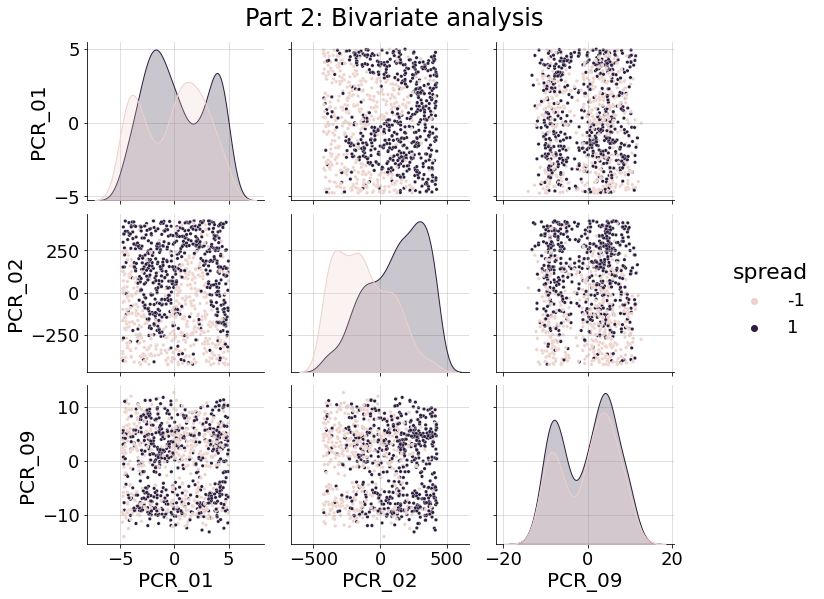
כמו כן, בעת חקירת הדאטה ובחירת התכונות הרלוונטיות עבור המודל שלנו אנו רוצים להסתכל רק על הדוגמאות שישמשו אותנו לאימון המודל. בחירת התכונות משפיעה על המודל שלנו, ועל המודל להיות בלתי תלוי בסט הבדיקה.

חלק 2

# שאלה 5:

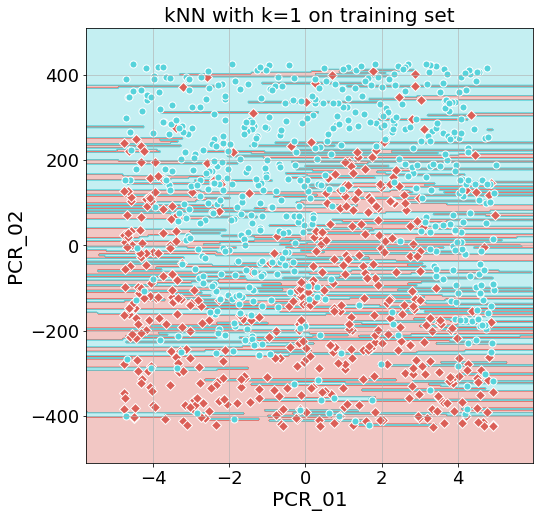
* קורלציה בין PCR\_01 וspread: 0.087066
* קורלציה בין PCR\_02 וspread: 0.514825
* קורלציה בין PCR\_09 וspread: -0.031038

# שאלה 6:



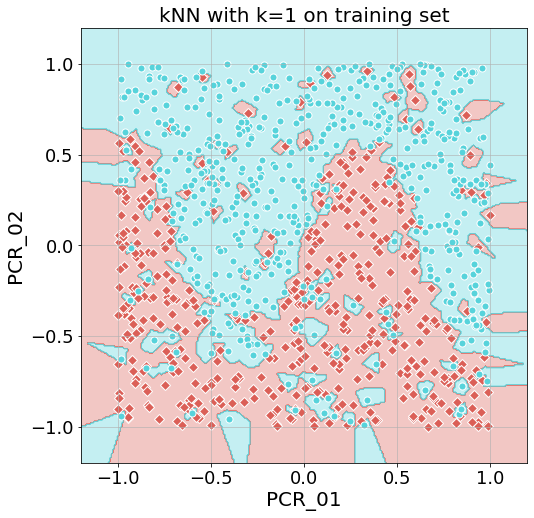
* PCR\_01 וPCR\_02 הן התכונות הטובות ביותר לניבוי spread מכיוון שהן מפרידות את הדוגמאות בצורה הכי טובה.

# שאלה 7



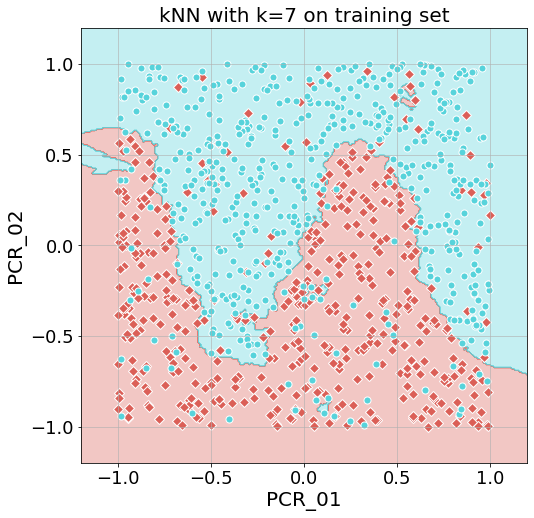
* הדיוק של סט האימון הוא 1. הסיבה לכך היא הבחירה של k=1, כאשר המודל מסווג נקודה בסט האימון הוא ימצא שהנקודה הכי קרובה אליה היא הדוגמה עצמה, לכן בהכרח יסווג אותה נכון.
* הדיוק עבור סט הבדיקה הוא 0.72

# שאלה 8



* הדיוק עבור סט הבדיקה השתפר וכעת הוא 0.8. הדיוק עבור סט האימון נשאר זהה כמצופה.
* KNN מודל מבוסס על מרחק בין שכנים. המרחק מושפע באופן ישיר מהערכים האבסולוטיים של התכונות. לכן יש לנרמל את ערכי התכונות כדי לתת את אותו המשקל לכל אחת מהן.

# שאלה 9



* הדיוק של המודל עבור סט האימון הוא: 0.881 ועבור סט המבחן הוא: 0.892
* ניתן לראות שאזורי ההחלטה של המודל רציפים יותר וקיים פחות רעש במודל עם k=7 ביחס למודל עם k=1. כמו כן, הדיוק עבור סט המבחן גבוהה יותר והדיוק עבור סט האימון נמוך יותר. בחירת k=1 גרמה לoverfitting במודל, בעקבות כך ראינו רעש באיזורי ההחלטה ודיוק מאוד גבוה עבור סט האימון. כאשר הגדלנו את k ראינו פחות רעש באיזורי ההחלטה, ודיוק גבוה יותר עבור סט המבחן, אך ירידה בהתאמת המודל לסט האימון.

# שאלה 10

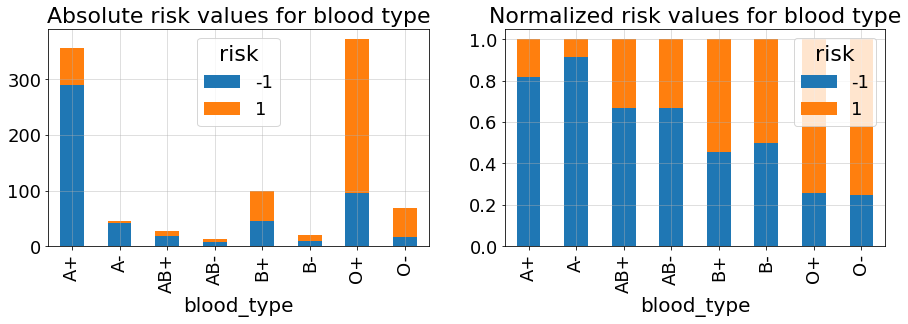
Normalizing a variable sampled from a normal distribution by min-max is very problematic. This function is based on the maximum and minimum value, these values will be very uncommon, the absolute majority of the values will be distributed around zero. As a result, a very large density of values will be created, which will make classification difficult and harm the results.

חלק 3

# שאלה 11

* כדי לייצג את blood\_type יהיה לנו צורך ב8 משתני boolean.

# שאלה 12

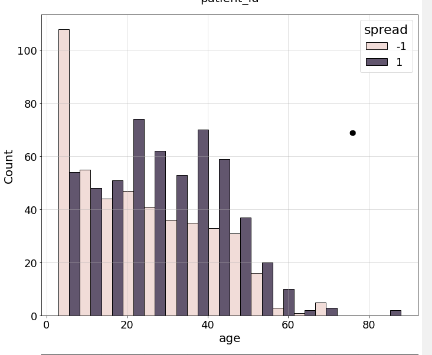


* חילקנו את סוגי הדם לשלוש קבוצות בהתאם לחלק היחסי בקבוצה שתוייג 1 בrisk, כמו כן, רצינו לחלק לשלוש קבוצות דומות בגודלן.
* קבוצה ראשונה: A+ וA- ניתן לראות שבשתי הקבוצות רוב הדוגמאות תויגו ב-1 .
* קבוצה שניה: O+ ו O- ניתן לראות שבשתי הקבוצות רוב הדוגמאות תויגו ב-1.
* קבוצה שלישית: שאר סוגי הדם, ניתן לראות שבשאר הקבוצות אין נטייה משמעותית של התיוגים ל-1 או ל1.

# שאלה 13

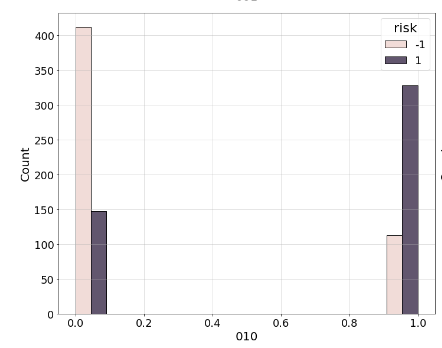
* את רשימת הסימפטומים המרנו לעמודות בטבלה, כך שכל עמודה מייצגת סימפטום. עבור כל דוגמא יופיע המספר 1 בעמודות של הסימפטומים שהופיעו ברשימת הסימפטומים שלה ו-0 עבור סימפטומים שלא היו ברשימת הסימפטומים.

# שאלה 14



* ניתן לראות שקיימים תחומי גילאים בהם יש הפרשים משמעותיים בין כמות הדוגמאות שתוייגו ב-1 לבין הדוגמאות שתויגו ב-1. לדוגמא, בגילאים נמוכים (העמודה השמאלית ביותר) , ניתן להסיק מהגרף, שהסיכוי להיות מתוייג ב-1 הוא גבוה יותר (יותר מפי 2 מקרים).

# שאלה 15



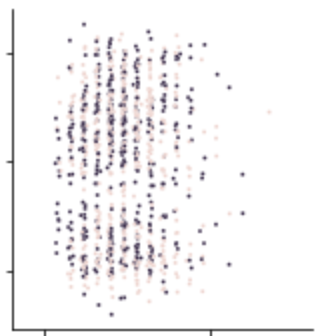
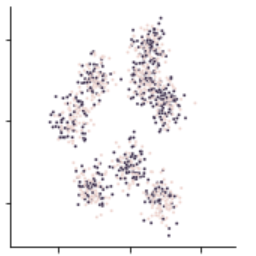
* ניתן לראות שעבור דוגמאות מקבוצת הדם 010, כלומר "010"=1 סיכוי גבוהה יותר לתיוג 1 ביחס לדוגמאות שאינן בקבוצה.

# שאלה 16

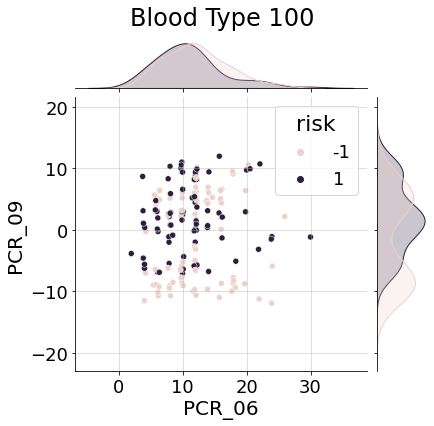
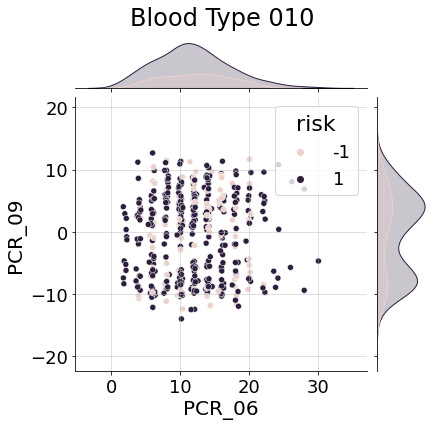
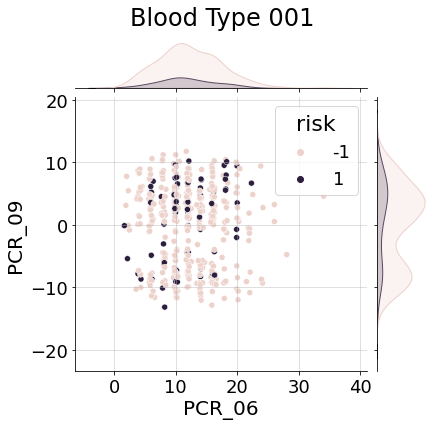
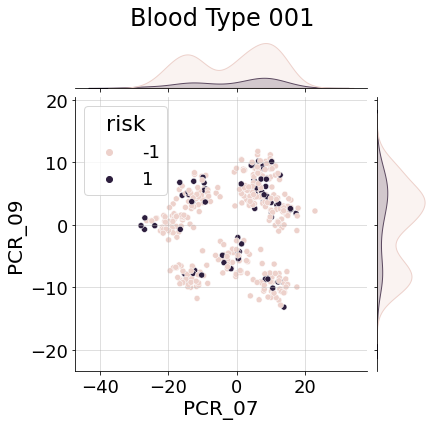
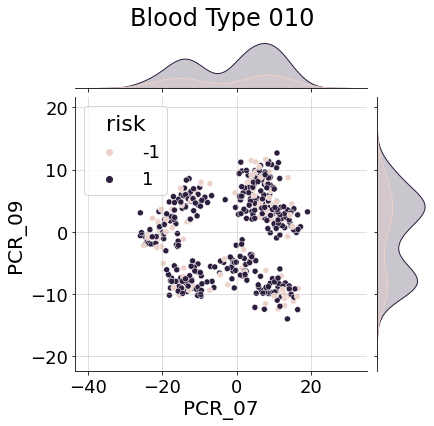
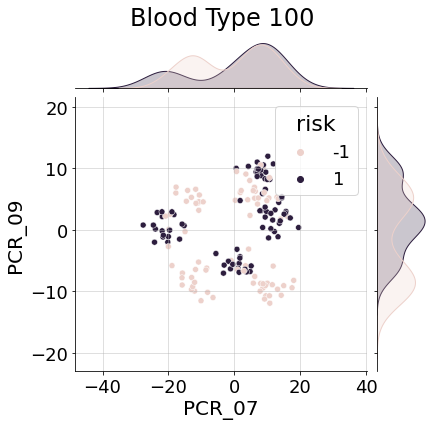
* עשרת התכונות המתואמות ביותר לrisk הן:

001,010, shortness\_of\_breath, household\_income, cough, weight, PCR\_03, PCR\_06, age, PCR\_05

# שאלה 17

* x= PCR\_06 y=PCR\_09
* X=PCR\_07 Y=PCR\_09
* לא ניתן לזהות שהזוגות האלו בפני עצמם מאפשרים להסיק מסקנות לגבי risk.

# שאלה 18



# שאלה 19

* התכונות האינפורמטיביות אינן בעלות קורלציה גבוהה עם risk.

# שאלה 20

בהתאם לחקירת הדאטה שעשינו עד כה, אנו משערים שהמודל שהכי יתאים הוא עץ החלטה. ראינו שאין הפרדה לינארית או מקומית של הדאטה בשאלה 17, אך גילינו שלאחר חלוקת הדאטה לקבוצות לפי קבוצות סוג הדם הדאטה פריד יותר, עץ החלטה יכול לעשות הפרדה כזאת.

# שאלה 21

# שאלה 22

* בforward feature selection נצטרך לאמן O(d2\*d1) מודלים.
* בbackward feature selection נצטרך לאמן O((d1-d2)\*d1) מודלים.

# שאלה 23

* התכונות שהתקבלו הן: ['weight', 'PCR\_02', 'PCR\_09']
* התכונות כוללות את PCR\_02 שבחרנו בשאלה 6, התכונות לא כוללות את התכונות שבחרנו באופן ידני בשאלה 14.

# שאלה 24

* הsequential feature selection מייצר מודלים על מנת למדוד ביצועים בעת הוספה\החסרה של תכונה. מודלים עלולים להיות מושפעים מהנורמליזיציה וכך גם ביצועיהם, לדוגמא, אנחנו משתמשים בkNN וראינו שתוצאות מודל זה מושפעות מנורמליזציה. מכיוון שביצועי המודל מושפעים מהנורמליזציה, גם בחירת התכונות מושפעת מהנורמליזציה. לכן חשוב לעשות את הנורמליזציה לפני בחירת התכונות.

# שאלה 25

* בחירת המודל של sequential feature selection אכן משפיעה על תהליך הבחירה. בחירת התכונות מבוססת על ביצועי המודל, על כן במהלך התהליך ייבחרו התכונות האופטימליות עבור המודל הספציפי שבחרנו.

# שאלה 26

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **שם הפיצ'ר** | **Keep** | **New** | **Normalization method** | **הסבר** |
| patient\_id | V | X | MinMax | *מצאנו כי משפיע על חיזוי* |
| age | V | X | MinMax | השארנו אותו מכיוון שהוא מועיל בחיזוי של covid |
| sex | V | X | בוליאני | *מצאנו כי משפיע על חיזוי* |
| weight | V | X | Standard | *מצאנו כי משפיע על חיזוי* |
| blood\_type | X | X | - | *לא נומרי, יצרנו מתכונה זו מס' תכונות בוליאניות* |
| current\_location | X | X | - | *לא מצאנו כי משפיע על החיזוי* |
| num\_of\_siblings | V | X | Standard | *מצאנו כי משפיע על חיזוי* |
| happiness\_score | V | X | Standard | השארנו אותו מכיוון שהוא מועיל בחיזוי של covid |
| household\_income | V | X | Standard | *מצאנו כי משפיע על חיזוי* |
| conversations\_per\_day | V | X | Standard | *מצאנו כי משפיע על חיזוי* |
| sugar\_levels | V | X | Standard | השארנו אותו מכיוון שהוא מועיל בחיזוי של covid |
| sport\_activity | V | X | MinMax | *מצאנו כי משפיע על חיזוי* |
| symptoms | X | X | - | *לא נומרי. יצרנו מתכונה זו מס' תכונות בוליאניות* |
| pcr\_date | X | X | - | *לא מצאנו כי בעל השפעה על החיזוי* |
| PCR\_01 | V | X | MinMax | *מצאנו כי משפיע על חיזוי* |
| PCR\_02 | V | X | MinMax | *מצאנו כי משפיע על חיזוי* |
| PCR\_03 | V | X | MinMax | *מצאנו כי משפיע על חיזוי* |
| PCR\_04 | V | X | MinMax | השארנו אותו מכיוון שהוא מועיל בחיזוי של covid |
| PCR\_05 | V | X | MinMax | *מצאנו כי משפיע על חיזוי* |
| PCR\_06 | V | X | Standard | *מצאנו כי משפיע על חיזוי* |
| PCR\_07 | V | X | Standard | השארנו אותו מכיוון שהוא מועיל בחיזוי של covid |
| PCR\_08 | V | X | Standard | *מצאנו כי משפיע על חיזוי* |
| PCR\_09 | V | X | Standard | *מצאנו כי משפיע על חיזוי* |
| PCR\_10 | V | X | Standard | *מצאנו כי משפיע על חיזוי* |
| 001 | V | V | בוליאני | יצרנו נתון זה מתוך blood type, והשארנו אותו מכיוון שהוא מועיל בחיזוי של covid. |
| 010 | V | V | בוליאני | יצרנו נתון זה מתוך blood type, והשארנו אותו מכיוון שהוא מועיל בחיזוי של covid. |
| 100 | V | V | בוליאני | יצרנו נתון זה מתוך blood type, והשארנו אותו מכיוון שהוא מועיל בחיזוי של covid. |
| cough | V | V | בוליאני | יצרנו נתון זה מתוך blood type, והשארנו אותו מכיוון שהוא מועיל בחיזוי של covid. |
| fever | V | V | בוליאני | יצרנו נתון זה מתוך symptoms, והשארנו אותו מכיוון שהוא מועיל בחיזוי של covid. |
| low\_appetite | V | V | בוליאני | יצרנו נתון זה מתוך symptoms, והשארנו אותו מכיוון שהוא מועיל בחיזוי של covid. |
| shortness\_of\_breath | V | V | בוליאני | יצרנו נתון זה מתוך symptoms, והשארנו אותו מכיוון שהוא מועיל בחיזוי של covid. |
| sore\_throat | V | V | בוליאני | יצרנו נתון זה מתוך symptoms, והשארנו אותו מכיוון שהוא מועיל בחיזוי של covid. |