מבוא למערכות לומדות - תרגיל בית 1 - דו"ח עבודה

מגישים - עומר שמחי - 316572593, חן פרץ - 204219638

5 במאי 2021

חלק 1 - טעינת הנתונים

1. טעינת הנתונים מקובץ ה-csv - נעשה בקובץ הסקריפט.

2 + 3 הטבלה להלן מכילה את הפיצ'רים, הטיפוס של כל פיצ'ר (מבחינת טיפוס בקוד) וכן תיאור של מה הוא מודד/מתאר לפי הבנתנו ולפי בדיקה באינטרנט.

שם הפיצ'ר	טיפוס	תיאור
ID	$int64$	תעודת זהות
$Address$	$string$	כתובת מגורים
$AgeGroup$	$float64$	קבוצת הגיל של האדם הנבדק
BMI	$float64$	מדד ה־BMI
$BloodType$	$Category: A\pm, AB, B\pm, O\pm$	סוג הדם
$ConversatiosPerDay$	$float64$	מספר השיחות "פנים מול פנים" שהנבדק מבצע ליום
$CurrentLocation$ $(*)$	$(float64, float64)$	מיקום נוכחי
$DateOfPCRTest$ $(**)$	$Date$	תאריך בדיקת ה־PCR
$DisciplineScore$	$float64$	מדד של משמעת להנחיות משרד הבריאות
$HappinessScore$	$float64$	מדד אושר/שמחה
$HouseholdExpense$ $OnPresents$	$float64$	סך הוצאות משק הבית על מתנות
$HouseholdExpense$ $OnSocialGames$	$float64$	סך הוצאות משק הבית על משחקי חברה
$HouseholdExpense$ $ParkingTicketsPerYear$	$float64$	סך הוצאות משק הבית על כרטיסי חניה בשנה
Job	$string$	העבודה בה מועסק האדם
$MedicalCarePerYear$	$float64$	הוצאות על טיפול רפואי בשנה
$NrCousins$	$int64$	מספר בני הדודים
PCR_10	$float64$	בדיקת PCR של 10 סיבובים
PCR_11	$float64$	בדיקת PCR של 11 סיבובים
PCR_15	$float64$	בדיקת PCR של 15 סיבובים
PCR_17	$float64$	בדיקת PCR של 17 סיבובים
PCR_19	$float64$	בדיקת PCR של 19 סיבובים
PCR_32	$float64$	בדיקת PCR של 32 סיבובים
PCR_45	$float64$	בדיקת PCR של 45 סיבובים
PCR_46	$float64$	בדיקת PCR של 46 סיבובים
PCR_7	$float64$	בדיקת PCR של 7 סיבובים

בדיקת PCR של 72 סיבובים	$float64$	PCR_72
בדיקת PCR של 76 סיבובים	$float64$	PCR_76
בדיקת PCR של 8 סיבובים	$float64$	PCR_8
בדיקת PCR של 83 סיבובים	$float64$	PCR_83
בדיקת PCR של 89 סיבובים	$float64$	PCR_89
בדיקת PCR של 9 סיבובים	$float64$	PCR_9
בדיקת PCR של 93 סיבובים	$float64$	PCR_93
בדיקת PCR של 95 סיבובים	$float64$	PCR_95
הצהרה עצמית על התסמינים של הנבדק	$string$	$Self\_declaration\_$ $of\_illness\_form$
מגדר/מין	$Category : M, F$	Sex
זמן המוקדש לפעילות חברתית ליום (בדקות)	$float64$	$SocialActivitiesPerDay$
זמן השהות ברשת חברתית ליום (בדקות)	$float64$	$SocialMediaPerDay$
זמן ביצוע פעילויות ספורט ליום (בדקות)	$float64$	$SportsPerDay$
מספר הצעדים לשנה (במיליונים)	$float64$	$StepsPerYear$
מספר שעות העמידה ביום (בדקות)	$float64$	$StudingPerDay$
סוג הנגיף	$string$	$Virus$
רמת ההדבקה	$Category : low, medium, high$	$SpreadLevel$
רמת סיכון	$Category : low, medium, high$	$Risk$

$(*)$ - במקור משתנה זה היה בפורמט של $string$ (כלומר המיקום על בסיס שתי הקורדינטות). כמובן, טיפוסו המדוייק יותר הוא זוג סדור.

$(**)$ - במקור התאריך הוצג בפורמט של $string$. אנו שינינו אותו לטיפוס $Date$ במוכר בפייתון (מהספרייה $datetime$).

חלק 2 - שיוך, ניקוי ונרמול הנתונים

חלוקת הנתונים

4. חלוקת הנתונים ל־3 סטים בצורה רנדומית - סט למידה ($training\ set$), סט ולידציה ($validation\ set$) וסט של בדיקה ($testing\ set$). הסטים חולקו ביחס של $60/20/20$ (כאשר ה־$training\ set$ הוא ה־60%). מעתה כל המניפולציה, טרנספורמציות וכ"ו יתבצעו על ה־$training\ set$ ולאחר מכן, יחולו על שני הסטים האחרים.

הבנה וחקר הנתונים

5. החלפנו **בקוד** את הטיפוסים של הפיצ'רים $CurrentLocation, DateOfPCRTest$ שקודם לכן יוצגו בפורמט $string$. את המיקום החלפנו בזוג סדור של המיקום המספרי ($(float, float)$) ואת תאריך הבדיקה שינינו לפורמט $Date$ בפייתון מפורמט $string$ בו היה מוצג קודם לכן.

6. שינוי/יצירת פיצ'רים חדשים עקב הצגתם בפורמט מטיפוס $string$ שאינו מועיל לצרכי למידה -

(א) $currentLocation$ - המיקום הנוכחי. הפיצ'ר הוצג במקור בפורמט $string$. כדי להפוך אותו לנוח יותר לניתוח, פרסרנו ($parse$) את המיקום **ויצרנו שני פיצ'רים נומריים חדשים** - אחד לסימון קו האורך ($CurrentLocationLongitude$) והשני לסימון קו הרוחב ($CurrentLocationLatitude$). לאחר מכן **הסרנו** את הפיצ'ר המקורי.

(ב) $DateOfPCRTest$ - התאריך של ביצוע בדיקת ה־PCR. הפיצ'ר היה מוצג בפורמט מטיפוס $string$ במקור. כדי שיהיה לנו יותר שימושי, החלטנו **ליצור פיצ'ר נומרי חדש** שבמקום להציג את תאריך הבדיקה, יציג את מספר הימים שעברו מאז $1.1.2020$ ועד תאריך הבדיקה (לאחר שבחנו פיצ'ר זה, שמנו לב שהבדיקות התבצעו החל מ־$1.1.2020$ ואילך). לאחר מכן הסרנו את הפיצ'ר הישן כיוון שאין לנו צורך בו יותר.

(ג) $Risk$ - **שינינו את הפיצ'ר הנ"ל והפכנו אותו מקטגורי לנומרי** בהתאם לחוקיות - $low = 0$, $medium = 1$ ו־$high = 2$.

(ד) $SpreadLevel$ -**שינינו את הפיצ'ר הנ"ל והפכנו אותו מקטגורי לנומרי** - $low = 0$, $medium = 1$ ו־$high = 2$.

(ה) $Virus$ - בתרגיל עלינו לזהות הימצאות של נגיף הקורונה אצל הנבדקים. על כן, **שינינו את הפיצ'ר הנ"ל מקטגורי לבינארי** - 0 אם הנגיף איננו נגיף הקורונה ו־1 אם הנגיף בו נדבק האדם הוא נגיף הקורונה.

(ו) $Self\_declaration\_of\_Illness\_Form$ - כדי להפוך את הפיצ'ר הזה לשימושי עבור זיהוי של חולים בקורונה, לקחנו מבין כל התסמינים שהנבדקים העידו עליהם, את אלו שהם תסימני קורונה (עפ"י בדיקה שביצענו באינטרנט). לאחר מכן, **שינינו את הפיצ'ר לפיצ'ר בינארי.** מעתה, הפיצ'ר יקבע אם הנבדק, עפ"י התסמיניו, חשוד כנשא של הנגיף (במידה והיו לו 3 או יותר תסמינים מבין הרשימה שהגדרנו, קבענו שהוא חשוד כנשא, אחרת איננו חשוד). כלומר הפכנו את הפיצ'ר לפיצ'ר בינארי.

7. (א) הסרנו את הפיצ'רים הבאים:

i. $currentLocation$ - הסרת פיצ'ר זה עקב שינויו (הפרדת המיקום לשתי קורדינטות במקום מחרוזת, הוסבר קודם לכן).

ii. PCR_11, PCR_15, Job - הסרת פיצ'רים אלו היות שערכים רבים היו חסרים בהם (Job - 27% נתונים חסרים, PCR_11, PCR_15 - יותר מ-82% ערכים חסרים).

iii. הסרת פיצ'רים אלו בעקבות היותם לא קשורים בעינינו לנשאות/אי נשאות של נגיף הקורונה או שאין לנו רצון לגלות קשר ביניהם לבין נשאות (למשל מספר בין דודים ($NrCousins$) - אם האלגוריתם הלומד יסיק כי יש קשר קורלציה בין נשאות נגיף הקורונה לבין מספר בני הדודים של הנבדק, כמובן לא נרצה להציג קשר זה, היות שאין לו הקשר מציאותי סביר והוא כנראה הראה קורלציה בגלל אוסף הנתונים המאד ספציפי שבחנו). אלו הפיצ'רים שבחרנו להסיר -

$$NrCousins, HouseholdExpenseOnPresents,$$
$$HouseholdExpenseOnSocialGames,$$
$$HouseholdExpenseParkingTicketsPerYear, SocialMediaPerDay,$$
$$StudingPerDay, Address, ID$$

iv. הסרת חלק מבדיקות ה-PCR (יפורט בהמשך). ההסרה היא בגלל $Outliers$ רבים מאד (יחסית לכמות הכוללת של הערכים).

ערכים חסרים

8. **מילוי ערכים חסרים -**

(א) $Risk$ - רק מדידה אחת חסרה ("nan") - את המדידה הזו החלטנו למלא עם הערך 2 ("$high$") - החלטנו לסווג אותה "באופן מחמיר".

(ב) $Virus$ - רק מדידה אחת חסרה ("nan") - את המדידה הזו החלטנו למלא עם הערך 1 (= חשוד בנשאות) - החלטנו לסווג אותה "באופן מחמיר".

(ג) $SpreadLevel$ - רק מדידה אחת חסרה ("nan") - את המדידה הזו החלטנו למלא עם הערך 2 ("$high$") - החלטנו לסווג אותה "באופן מחמיר".

(ד) $CurrentLocationLatitude, CurrentLocationLongitude$ -את הערכים החסרים שהיו מילינו ב--1 היות שאין משמעות למילוי ערכים אלו עם ערך ממוצע/ ערכים לפי ההתפלגות הקיימית (אין משמעות, משום שקווי האורך/רוחב שיצאו עבור השלמת הערכים חסרים, יכולים להיות קווי אורך ורוחב של מיקומים שלא גרים בהם אנשים כלל!).

(ה) $Self\_declaration\_of\_Illness\_Form$ - עבור אנשים שלא היו להם סימפטומים שעליהם הצהירו (ולכן לא מולא להם הערך "1" או הערך "0" בתור חשודים בנשאות או לא) מילינו את הערך "0" - נתייחס כאילו הם אינם חשודים כנשאים לפי פרמטר זה (כפי שסביר לעשות - אם אדם לא הצהיר על סימפטומים, אין סיבה לחשוד בו כנשא של נגיף הקורונה).

(ו) **עבור שאר הפיצ'רים בהם היות נתונים חסרים -** מילינו את הערכים החסרים שהיו לפי **ההתפלגות** של **הערכים הקיימים** (כלומר הערכים החסרים מולאו בהתאם לחלק היחסי הקיים שלהם מהכלל). הסיבה למילוי הערכים בצורה הזו היא מפני שזו השיטה הטובה ביותר שחשבנו שתייצג את הנתונים שכן התקבלו. הערכים החסרים לא מהווים אחוז ניכר מכלל הדגימות (כי הורדנו כבר בשלב הקודם פיצ'רים בהם חסרות הרבה דגימות). לכן, סביר שאילו לא היו חסרים, היו מזדהים בצורה מקורבת להתפלגות של הנתונים הקיימים.

זיהוי של נתונים יוצאים מן הכלל ($Outliers$)

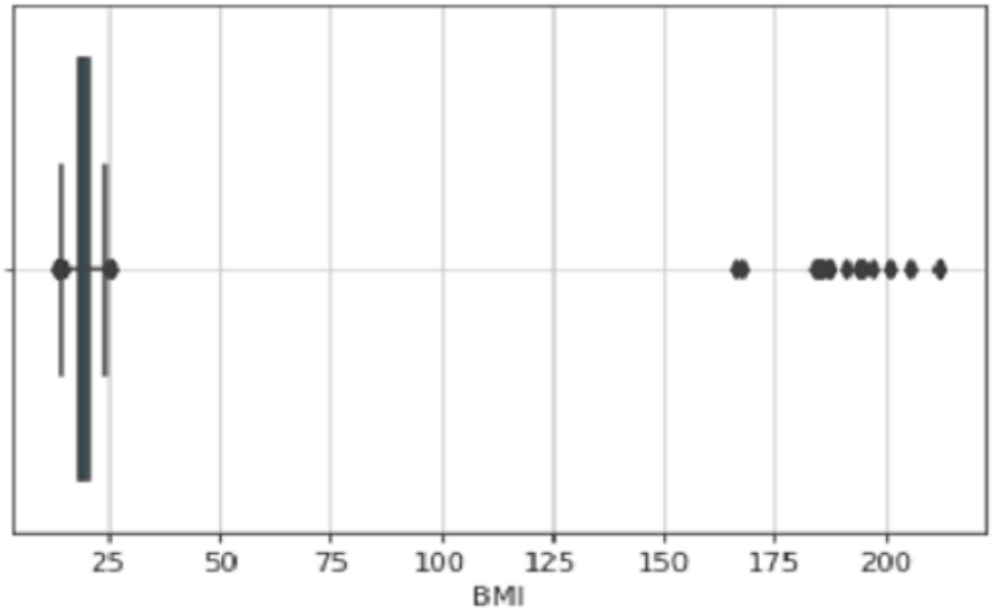
9. בחרנו בפיצ'רים הבאים כדי לבצע זיהוי של יוצאים מן הכלל ("$Outlier\ Detection$") - BMI ו-$SocialActivitiesPerDay$. בחרנו בפיצ'רים הללו מפני שהיוצאים מן הכלל בהם היות מאד מעטים מספרית (כלומר ביחס לסך הדגימות) וכן שההשפעה שלהן יחסית למספרם היה גדול יחסית. **הסבר ביחס לאיורים -**

(א) BMI - כמעט כל הערכים רוכזו בין טווח הערכים 14 ל-25 וכמה ערכים בודדים היו מעל 150, דבר שכמובן מהווה סטייה ברורה משאר הערכים הנדגמים וכן מטווח הערכים האמיתי שפיצ'ר זה יכול לקבל במציאות).

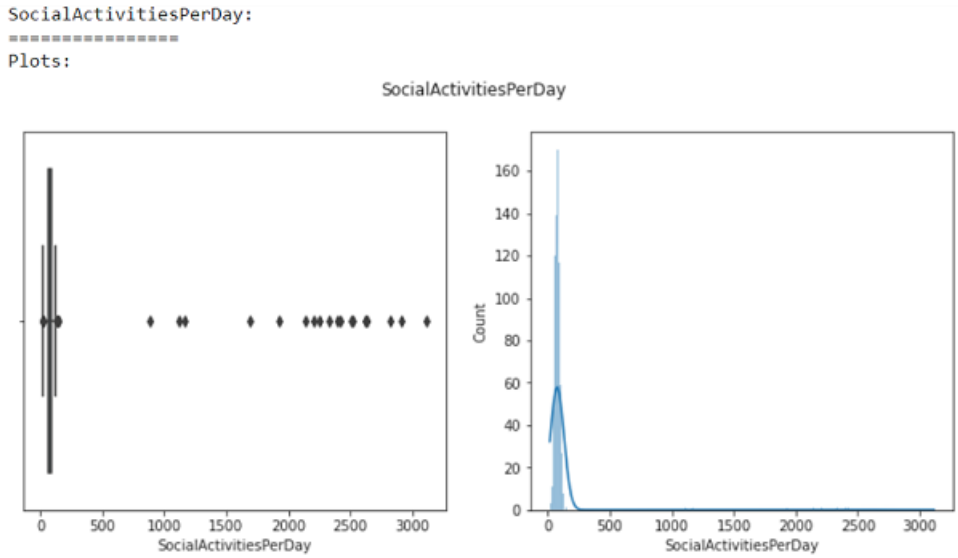
(ב) $SocialActivitiesPerDay$ - כמעט כל הערכים רוכזו בין טווח הערכים 0 ל-100 וכמה ערכים בודדים היו מעל 500, דבר שכמובן מהווה סטייה ברורה משאר הערכים הנדגמים.

איור 1: BMI לפני טיפול ב־$Outliers$

```
BMI Outlier detection:
================
Boxplot:
```



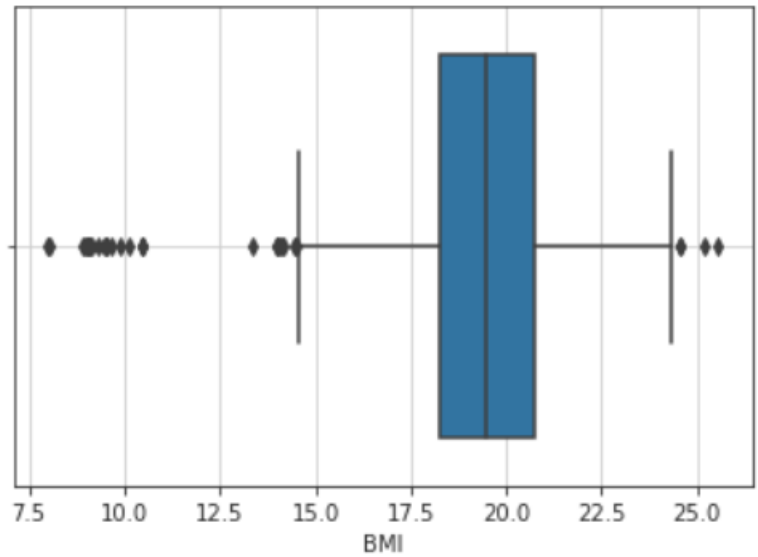
איור 2: $SocialActivitiesPerDay$ לפני טיפול ב־$Outliers$



10. **ניתוח, זיהוי והורדת היוצאים מן הכלל (**$Outliers$**)** בכל אחד משני הפיצ'רים שצויינו בסעיף הקודם:

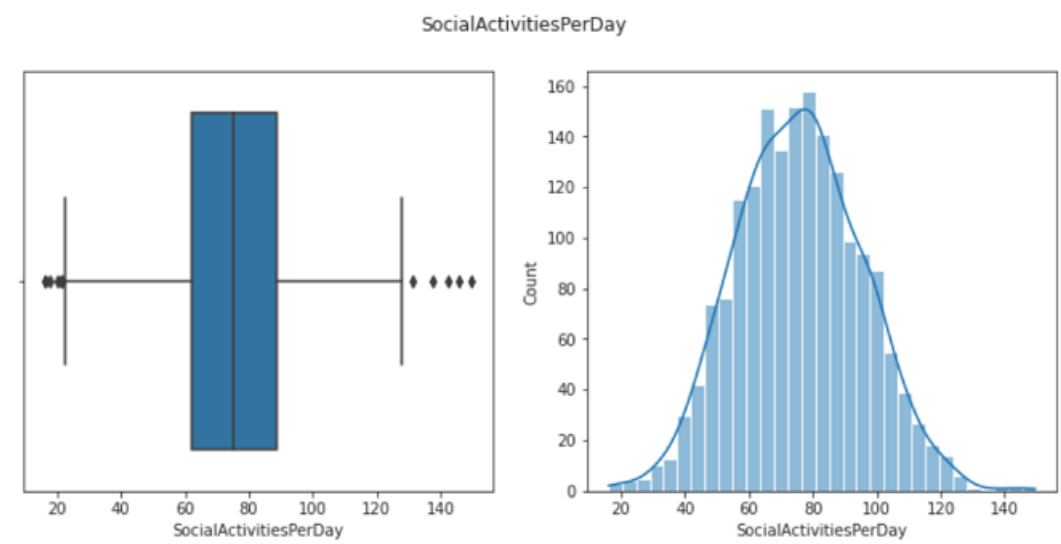
איור 3: BMI לאחר טיפול ב־$Outliers$

```
BMI after removing Outliers:
Boxplot:
```



איור 4: $SocialActivitiesPerDay$ לאחר טיפול ב־$Outliers$

```
SocialActivitiesPerDay after removing Outliers:
Plots:
```



(א) BMI -

i. <u>זיהוי היוצאים מן הכלל</u> - לאחר שהצגנו את הערכים בהיסטוגרמת $Boxplot$ (ראו איור מהסעיף הקודם), ראינו שכמעט כל הערכים רוכזו בין טווח הערכים 14 ל־25 וכמה ערכים בודדים היו מעל 150, דבר שכמובן מהווה סטייה ברורה משאר הערכים הנדגמים וכן מטווח הערכים האמיתי שפיצ'ר זה יכול לקבל במציאות).

ii. <u>טיפול ביוצאים מן הכלל</u> - השתמשנו ב־$Zscore$ על מנת להסיר יוצאים מן הכלל - קבענו שכל מי שיש לו BMI שגבוה מ־50 (מדד שאיננו סביר) הוא $Outlier$ ועל כן נחליף את ערכו ע"י שימוש ב־$Zscore$ כדי "למרכז" את רוב הערכים במרכז הדגימות וכדי להוריד את ההשפעה של היוצאים מן הכלל (כעת, כפי שרואים מהגרף בסעיף הנ"ל, היוצאים מן הכלל לא חורגים מרוב הערכים המרוכזים במרכז ביותר מ10 יחידות).

(ב) $SocialActivitiesPerDay$ -

i. <u>זיהוי היוצאים מן הכלל</u> - לאחר שהצגנו את הערכים בהיסטוגרמת $Boxplot$, ראינו שכמעט כל הערכים רוכזו בטווח שבין 0 ל100 (באופן גס) וכמה ערכים בודדים היו מעל 500, דבר שכמובן מהווה סטייה ברורה משאר הערכים הנדגמים).

ii. <u>טיפול ביוצאים מן הכלל</u> - השתמשנו **בהתפלגות של רוב הערכים** (כל הערכים שלא עולים על 500) על מנת להסיר יוצאים מן הכלל - קבענו שכל מי שיש לו ערך שגבוה מ־500 (מדד שאיננו סביר) הוא $Outlier$ ועל כן נחליף את ערכו ע"י שימוש בהתפלגות של רוב הערכים (שהינם בטווח סביר) כדי "למרכז" את רוב הערכים במרכז הדגימות וכדי להוריד את ההשפעה של היוצאים מן הכלל (כעת, כפי שרואים מהגרף שצורף בסעיף, היוצאים מן הכלל לא חורגים **מרוב הערכים** ביותר מ40 יחידות).

(ג) PCR_i - כאשר $i \notin \{19, 7, 72, 89, 95\}$ וכן i הוא מספר של בדיקה שמוצגת בנתונים -

i. <u>זיהוי היוצאים מן הכלל</u> - לאחר שהצגנו את הערכים בהיסטוגרמת $Boxplot$, ראינו בבדיקות PCR רבות שיש ערכים רבים עם הפרשים גדולים אחד מהשני.

ii. <u>טיפול ביוצאים מן הכלל</u> - עבור על בדיקת PCR_i אשר בה השונות (std) עלתה על הערך 4 (עליו החלטנו) החלטנו להסירו, היות שהערכים שם לא מעידים, לדעתינו, על ממצאים חד משמעיים בגלל הכמות הרבה של היוצאים מן הכלל ביחס לשאר הערכים.

לסיכום חלק זה - בשני הפיצ'רים נראו חריגות ברורות של מספר בודד של ערכים **מרוב הערכים שנדגמו** ולכן הם זוהו כיוצאים מן הכלל ותוקנו בהתאם (כפי שהוסבר, באופן שונה,

בהתאם לכל אחד משני הפיצ'רים).

טרנספורמציות לנתונים

11. נרמלנו את הפיצ'ר $StepsPerYear$ לפי שיטת הנרמול $Zscore$. הנרמול נבחר להיות כזה היות שאין כמעט יוצאים מן הכלל ($Outliers$) ולכן אין צורך לבצע $min\_max\ scaling$ כי השפעתם זניחה. בנוסף, הנרמול הנ"ל לא משנה את המהות של הפיצ'ר (שמהותו היא כמה, באופן יחסי לנבדקים האחרים, אותו נבדק צעד בשנה ולכן מספר הצעדים הספציפים לא משנה).

12. להלן הפיצ'רים שבוצעה עליהם טרנספורמצייה בהתאם ל־4 השיטות שהוצגו:

(א) $Categorial\ to\ Numeric\ Conversion$ -

i. $BloodType$ - עבור כל אחד מסוגי הדם שנבדקו, התאמנו מספר עולה ($0, 1, ..$) שייצג אותו. הסיבה לייצוג הנ"ל היא היכולת לנתח אותו בהמשך בנוחות, ובנוסף למלא את הערכים החסרים בהתאם להתפלגות של הערכים שנדגמו (דבר שלא אפשרי כמובן אם הפיצ'ר איננו נומרי).

ii. Sex - עבור המגדרים התאמנו מספרים - זכר - 0 , נקבה - 1. הסיבה לייצוג הנ"ל היא היכולת לנתח אותו בהמשך בנוחות, ובנוסף למלא את הערכים החסרים בהתאם להתפלגות של הערכים שנדגמו (דבר שלא אפשרי כמובן אם הפיצ'ר איננו נומרי).

iii. $Risk$ - עבור כל אחת משלושת רמות הסיכון התאמנו מספר עולה ($low = 0, medium = 1, high = 2$). הסיבה לייצוג הנ"ל היא היכולת לנתח אותו בהמשך בנוחות, ובנוסף למלא את הערכים החסרים בהתאם להתפלגות של הערכים שנדגמו (דבר שלא אפשרי כמובן אם הפיצ'ר איננו נומרי).

iv. $SpreadLevel$ - עבור כל אחת משלושת רמות ההדבקה התאמנו מספר עולה ($low = 0, medium = 1, high = 2$). הסיבה לייצוג הנ"ל היא היכולת לנתח אותו בהמשך בנוחות, ובנוסף למלא את הערכים החסרים בהתאם להתפלגות של הערכים שנדגמו (דבר שלא אפשרי כמובן אם הפיצ'ר איננו נומרי).

v. $DateOfPCRTest$ - הפיצ'ר היה מוצג בפורמט מטיפוס $string$ במקור. כדי שיהיה לנו יותר שימושי, החלטנו **ליצור פיצ'ר נומרי חדש** שבמקום להציג את תאריך הבדיקה, יציג את מספר הימים שעברו מאז 1.1.2020 ועד תאריך הבדיקה (לאחר שבחנו פיצ'ר זה, שמנו לב שהבדיקות התבצעו החל מ־1.1.2020 ואילך). הסיבה לשינוי בייצוג היא שהייצוג היה כמחרוזת, דבר שלא היה נוח לעבוד עימו. בנוסף, היות שרצינו לייצג ע"י מספר אחד את תאריך הבדיקה, והיות שבסה"כ הבדיקות התפרסו על פני שנתיים

(2020 – 2021) יכולנו בצורה נוחה לוותר על חודש ויום הבדיקה ולייצג אותה רק ע"י מספר הימים מ־1.1.2020.

(ב) $Attribute/Feature\ Construction$ -

i. $Virus$ - שינינו את התכונה של הפיצ'ר - במקום לרשום מה הוירוס בו נדבק האדם, ייצגנו בייצוג בינארי (0 - כן, 1 - לא) האם האדם נדבק בנגיף הקורונה או לא. הסיבה שבחרנו בייצוג זה הוא שמטרת הניתוח והלמידה של הנתונים הללו היא למעשה לזהות **נשאות בקורונה** ולכן זה המידע שמעניין אותנו לייצג בנוגע לוירוס שהנבדקים נושאים.

ii. $Self\_declaration\_of\_illness\_form$ - שינינו את תכונת הפיצ'ר - במקום לתאר את הסימטומים של הנבדק במילים, בדקנו באינטרנט מה התסמינים של מחלת הקורונה, ובהתאם לכך, החלטנו כי לנבדק שיש לפחות 3 תסמיני קורונה, נצמיד את הספרה 1 (כלומר חשוד כנשא של קורונה) ואילו לנבדק שאין לו סימפטומים/ אין מספרי סימפטומים של מחלקת הקורונה נצמיד את הספרה 0 (כלומר הוא איננו חשוד בנשאות הנגיף). בסה"כ הפכנו את הפיצ'ר מפיצ'ר שהוצג כמחרוזת לפיצ'ר בינארי. הסיבה לייצוג הנ"ל היא היכולת לנתח אותו בהמשך בנוחות, ובנוסף למלא את הערכים החסרים עם 0־ים (כיוון שלנבדק שלא הצהיר על סימפטומים כלל, בפרט לא נחשוד שהוא נשא של נגיף הקורונה).

iii. $currentLocation$ - הורדנו פיצ'ר זה ובמקומו יצרנו **קבוצה של שני פיצ'רים** שיתארו את קו האורך והרוחב של המיקום הנוכחי. הסיבה לייצוג הנ"ל היא היכולת לנתח אותו בהמשך בנוחות.

(ג) $Scailing$ -

i. $StepsPerYear$ - תואר הנרמול בסעיף 11 בפירוט. הסיבה לטרנספורמציה ע"י נרמול היא שהיא איננה משנה את המהות של הפיצ'ר שחשובה לנו (שמהותו היא כמה, באופן יחסי לנבדקים האחרים, אותו נבדק צעד בשנה ולכן מספר הצעדים הספציפים לא משנה).

(ד) $Generalization$ ובנוסף $Discretization$ -

i. BMI - הפכנו את הפיצ'ר לפיצ'ר **דיסקרטי** במקום פיצ'ר **רציף** - סיווגנו את מדדי ה־BMI לפי הקבוצות המוגדרות שלו (תת משקל, משקל תקין ומשקל יתר). קבענו כי BMI לפי הטווחים הבאים יסווג לקבוצות עם המשמעות המתאימה (לפי מידע שמצאנו בויקיפדיה) -

א'. $BMI < 18.5$ - תת משקל - ייוצג ע"י הספרה 0.

ב'. $18.5 \leq BMI < 25$ - משקל תקין, ייוצג ע"י הספרה 1.

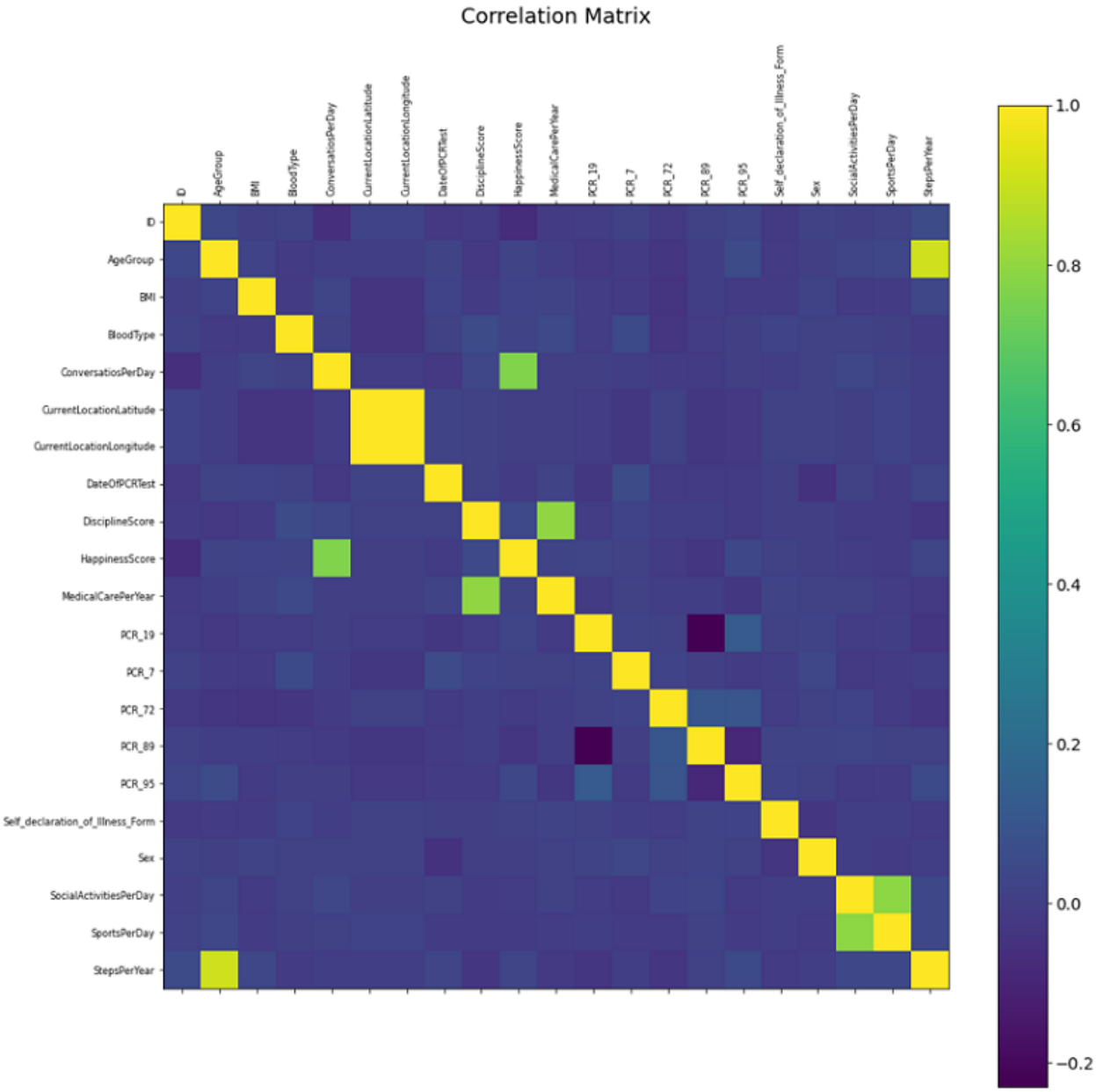
ג'. $BMI \geq 25$ - משקל יתר.

הסיבה לשינוי בייצוג היא שהעניין העיקרי שלנו במדד ה־BMI היא האם הוא מייצג נבדק שבמשקל יתר, תת משקל או במשקל תקין ולא במדד הספציפי שלו. בצורה כזו נקבל אינדיקציה יותר ברורה האם ה־BMI משפיע על הסיכוי לנשאות נגיף הקורנה.

3. בחירה של פיצ'רים

13. להלן טבלת הקורלציות של הפיצ'רים -

איור 5: טבלת הקורלציה בין הפיצ'רים



לפי טבלה זו, קבענו שקורלציה שבעינינו מספיק הדוקה היא קורלציה עם ערך של

לפחות 0.75. על כן, כפי שניתן לראות בטבלה (לפי הצבעים) נקבל את זוגות המשתנים הבאים כמשתנים קורלטיביים:

(א) $AgeGroup - StepsPerYear$ - קורלציה עם ערך 0.897.

(ב) $ConversatiosPerDay - HappinessScore$ - קורלצייה עם ערך 0.766.

(ג) $CurrentLocationLatitude - CurrentLocationLongitude$ - קורלצייה עם ערך 1.

(ד) $DiciplineScore - MedicalCarePerYear$ - קורלצייה עם ערך 0.759.

(ה) $SportsPerDay - SocialActivitiesPerDay$ - קורלצייה עם ערך 0.808.

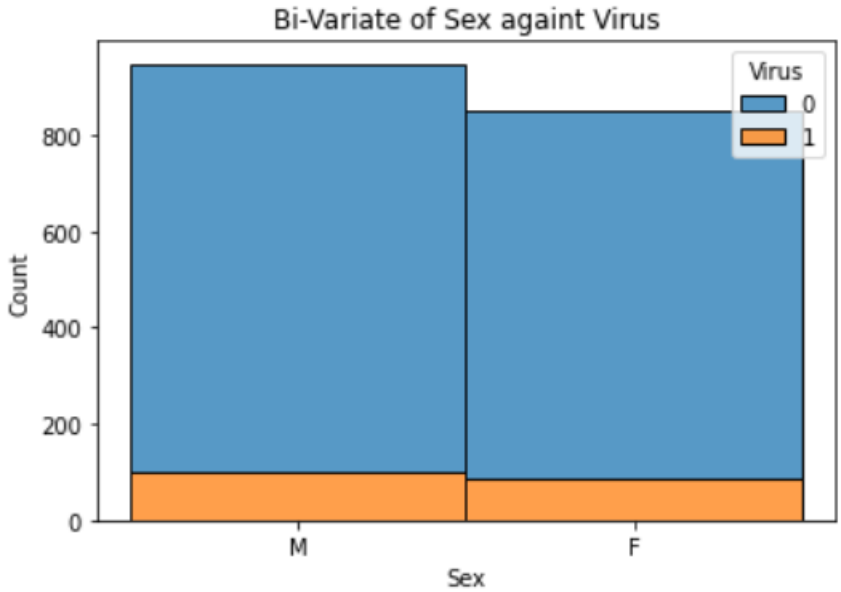
$(*)$ - ההיסטוגרמות הממחישות גרפית את הקשרים בין כל זוג מן הזוגות הנ"ל מוצגים באופן מסודר בקובץ הסקריפט המצורף.

14. (א) הורדנו פיצ'רים נוספים ע"י שתי שיטות של $wrapper\ method$: $Decision\ trees$ ו-kNN. באמצעות סט הנתונים עבור הלמידה ($training\_set$) נתנו לשני האלגוריתמים הללו ללמוד את הנתונים כך שכל אחד הוציא מספר פיצ'רים משמעותיים. לקחנו את איחוד הפיצ'רים המשמעותיים מריצת שני האלגוריתמים ואת הפיצ'רים הנותרים הסרנו. **נעיר כאן שגילינו שפעולת האלגוריתמים הללו איננה דטרמיניסטית ולכן הסרנו פיצ'רים ע"ס כמה ריצות שבהן קיבלנו אותן תוצאות.** בריצות הללו הסרנו את הפיצ'רים הבאים:

$$ID, BloodType, CurrentLocationLatitude, PCR\_95, DateOfPCRTest$$

(ב) בסעיף זה ביצענו $bevariate\ analysis$ על הפיצ'ר Sex. גילינו כי אין מגמה ניכרת בין זכר (M תרשים המצורף) לנקבה (F בתרשים המצורף) בתיוג של הנבדקים כנשאי קורונה ולכן החלטנו להסיר פיצ'ר זה היות שהוא לא תורם מידע נוסף. להלן מצורף התרשים -

איור 6: $bevariate\ analysis\ on\ the\ feature\ Sex$



15. להלן טבלה של כל הפיצ'רים (אלא שהוסרו, אלו שהתווספו ואלא שנשארו מההתחלה). הפיצ'רים שנשארו הם כל הפיצ'רים שלא הורדו בשלבים הקודמים (שתוארו לאורך הדו"ח ובקוד) וכן הפיצ'רים שמהווים משתני מטרה ($Risk, SpreadLevel, Virus$):

שם הפיצ'ר	נשאר/הוסר	הסבר קצר מדוע הוסרו
ID	הוסר	אין קשר סביר לנשאות הנגיף
$Address$	הוסר	אין קשר סביר לנשאות הנגיף
$AgeGroup$	נשאר	
BMI	נשאר	
$BloodType$	הוסר	*wrapper method*
$ConversatiosPerDay$	הוסר	קורלציה גבוהה עם משתנה אחר שנשאר
$CurrentLocation$	הוסר	הוחלף ע"י פיצ'רים אחרים
$DateOfPCRTest$	הוסר	*wrapper method*
$DisciplineScore$	נשאר	
$HappinessScore$	נשאר	
$HouseholdExpense$ $OnPresents$	הוסר	אין קשר סביר לנשאות הנגיף
$HouseholdExpense$ $OnSocialGames$	הוסר	אין קשר סביר לנשאות הנגיף
$HouseholdExpense$ $ParkingTicketsPerYear$	הוסר	אין קשר סביר לנשאות הנגיף
Job	הוסר	יותר מידי ערכים חסרים
$MedicalCarePerYear$	הוסר	קורלציה גבוהה עם משתנה אחר שנשאר
$NrCousins$	הוסר	אין קשר סביר לנשאות הנגיף
PCR_10	הוסר	ריבוי *Outliers*
PCR_11	הוסר	יותר מידי ערכים חסרים
PCR_15	הוסר	יותר מידי ערכים חסרים
PCR_17	הוסר	ריבוי *Outliers*
PCR_19	נשאר	
PCR_32	הוסר	ריבוי *Outliers*
PCR_45	הוסר	ריבוי *Outliers*
PCR_46	הוסר	ריבוי *Outliers*
PCR_7	נשאר	

PCR_72	נשאר	
PCR_76	הוסר	ריבוי $Outliers$
PCR_8	הוסר	ריבוי $Outliers$
PCR_83	הוסר	ריבוי $Outliers$
PCR_89	נשאר	
PCR_9	הוסר	ריבוי $Outliers$
PCR_93	הוסר	ריבוי $Outliers$
PCR_95	הוסר	*wrapper method*
$Self\_declaration\_ of\_illness\_form$	נשאר	
Sex	הוסר	*bevariate analysis*
$SocialActivitiesPerDay$	הוסר	קורלציה גבוהה עם משתנה אחר שנשאר
$SocialMediaPerDay$	הוסר	אין קשר סביר לנשאות הנגיף
$SportsPerDay$	נשאר	
$StepsPerYear$	הוסר	קורלציה גבוהה עם משתנה אחר שנשאר
$StudingPerDay$	הוסר	אין קשר סביר לנשאות הנגיף
$Virus$	נשאר	משתנה המטרה המרכזי שלנו
$SpreadLevel$	נשאר	משתנה מטרה
$Risk$	נשאר	משתנה מטרה
$CurrentLocationLongitude$	הוסר	קורלציה גבוהה עם משתנה אחר שנשאר
$CurrentLocationLatitude$	הוסר	*wrapper method*

16. בקובץ הסקריפט.