חלק רטוב – פירוט תהליך הטיפול בנתונים

הכנת הנתונים

בשלב הכנת הנתונים העברנו את הנתונים במספר מאפיינים ופונקציות התאמה שונות, כך שכל פונקציה כזאת מתנהגת לפי הסטנדרט של sklearn (מממשת מטודות fit ו transform), וכך ניתן להתאים את הפונק' על נתוני האימון ולהפעיל אותן על שאר הנתונים. כמו כן ניתן להשתמש בpipeline של sklearn כדי להפעיל את המניפולציות בקלות.

בשלב הראשון הפרדנו את הנתונים לשלוש קבוצות – אימון, ואלידציה ומבחן. החלוקה היא 15% קבוצת מבחן, 15% קבוצת ואלידציה ו 70% קבוצת אימון.

<u>בחירת טיפוסי נתונים מתאימים:</u>

בשלב הבא לכל מאפיין בנתונים קבענו סוג נתונים מתאים:

Feature Name	Data Type
Address	String *
AgeGroup	Numeric
AvgHouseholdExpenseOnPresents	Numeric
AvgHouseholdExpenseOnSocialGames	Numeric
AvgHouseholdExpenseParkingTicketsPerYear	Numeric
AvgMinSportsPerDay	Numeric
AvgTimeOnSocialMedia	Numeric
AvgTimeOnStuding	Numeric
BMI	Numeric
BloodType	Categorical
CurrentLocation	String **
DateOfPCRTest	DateTime
DisciplineScore	Numeric
HappinessScore	Numeric
Job	String *
NrCousins	Numeric
SelfDeclarationOfIllnessForm	String ***
Sex	Categorical
StepsPerYear	Numeric
SyndromeClass	Categorical
TimeOnSocialActivities	Numeric
pcrResult1	Numeric
pcrResult10	Numeric
pcrResult11	Numeric
pcrResult12	Numeric
pcrResult13	Numeric
pcrResult14	Numeric
pcrResult15	Numeric
pcrResult16	Numeric
pcrResult2	Numeric
pcrResult3	Numeric
pcrResult4	Numeric
pcrResult5	Numeric

pcrResult6	Numeric
pcrResult7	Numeric
pcrResult8	Numeric
pcrResult9	Numeric

(*) בבדיקה של המאפיינים Address ו Job נראה כי אלו עמודות עם מספר גדול מאוד של ערכים יחודיים, ומספר גדול של ערכים חסרים. מכיוון שאין בין הנתונים יחס סדר כלשהו, ויש כמות גדולה מאוד של ערכים יחודיים הנחנו כי לא סביר שמאפיינים אלו יתרמו ללמידה, ולכן לא עיבדנו אותם מעבר לשלב הזה.

(**) המאפיין CurrentLocation למעשה מורכב משני ערכים, מעלות רוחב ומעלות אורך, של נ.צ גיאוגרפי. בשלב הראשון הגדרנו את המאפיין כמחרוזת, ובהמשך עיבדנו אותו לשני שדות – CurrentLocation_Lat ו CurrentLocation_Long נומריים בהתאמה.

(***) המאפיין SelfDeclarationOfIllnessForm מכיל למעשה רשימה של ערכי מחרוזת, בסדר משתנה, שחוזרים על עצמם בין שורות בנתונים. בשלב הראשון הגדרנו את המאפיין כמחרוזת, ובהמשך עיבדנו אותו לרשימה של מאפיינים בינאריים שמצביעים על הימצאות של סימפטום נתון בדיווח העצמי של הנדגם. רשימת המאפיינים שנוספו היא:

Diarrhea, Nausea_or_vomiting, Shortness_of_breath, Congestion_or_runny nose, Headache, No_Symptoms, Fatigue, Muscle_or_body_aches, Chills, Skin_redness, New_loss_of_taste_or_smell, Sore_throat.

בחרנו להגדיר בנפרד את המאפיין No_symptoms עבור שורות בהן מאפיין SelfDeclarationOfIllnessForm תחת ההנחה שזהו מידע יחודי – יתכן שהמחסור בכל סימפטום הוא מידע חשוב, ומכיוון שלא נעשה שימוש בכל המאפיינים בסופו של דבר, ויתור על חלק מהמאפיינים המתארים סימפטומים יכול לגרום לאובדן של המידע הזה.

בשלב הבא המרנו את המאפיינים הקטגוריים (SyndromeClass ו BloodType) לרשימה של מאפיינים בינאריים על ידי קידוד oneHot. רשימת המאפיינים שנוספו היא:

BloodType_AB-, BloodType_A+, BloodType_AB+, BloodType_A-, BloodType_B-, BloodType_O-, BloodType_B+, BloodType_O+, SyndromeClass_1, SyndromeClass_2, SyndromeClass_3, SyndromeClass_4.

השלמת נתונים חסרים

בסקירה של הנתונים נראה כי בכל עמודה קיימים נתונים חסרים רבים, וויתור על שורות עם נתונים חסרים לא בא בחשבון. תחילה ניסינו לבצע השלמת נתונים בערכים קבועים – ערך ממוצע עבור המאפיין במשתנים מספריים רציפים, כמו גם במשתנים מספריים בדידים בעלי יחס סדר (לדוגמא HappinessScore שהוא דירוג על סולם סדר), וערך נפוץ ביותר עבור משתנים קטגוריים.

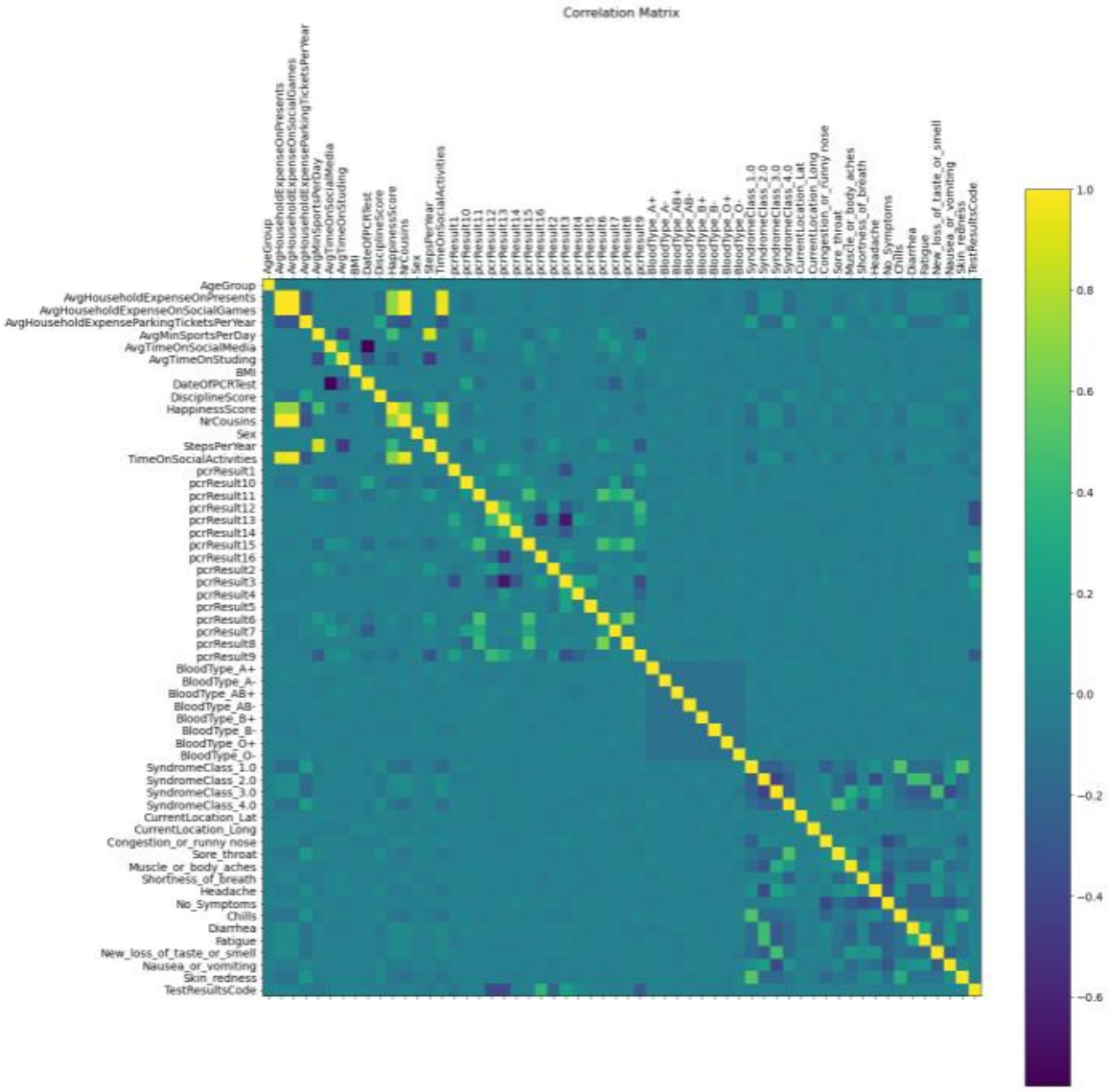
לא ביצענו השלמה עבור משתנים קטגוריים שקודדו בOneHot בכדי לא להטות את ההתפלגות, ומכיוון שקידוד OneHot מייצג בצורה שלמה מידע חסר.

לפני בחירת המאפיינים שלנו החלטנו להחליף את שיטת השלמת הנתונים החסרים, מכיוון שישנם נתונים חסרים רבים מאוד השלמה בערך קבוע מטה מאוד את ההתפלגות. לכן בחרנו במקום להשלים מאפיינים נומריים רציפית באמצעות אינטרפולציה איטרטיבית (עם מדיניות ראשונית של השלמה לחציון), ומאפיינים קטגוריים וסדורים בעזרת השלמה משכנים (KNN).

איתור והסרת ערכי קיצון

ראינו כי במספר מאפיינים, לדוגמא StepsPerYear קיימים ערכים שסוטים בצורה משמעותית מאוד מההתפלגות של שאר הנתונים, ולכן קטמנו את הנתונים במאפיינים כאלו לפי קצוות עליון ותחתון של box-plot, כאשר נתונים חורגים מעלה הוחלפו בערך Upper Whisker ונתונים חורגים מטה הוחלפו בערך Lower Whisker.

לאחר קטימת ערכי קיצון ניתן היה כבר לראות התאמה מסויימת בין מאפיינים שונים



נירמול של הנתונים

בשלב הבא ביצענו נירמול לתחום [-1,1] של הנתונים לשם טיוב ריצת האלגוריתמים. מאפיין Sex נקבע מראש להיות מיוצג ע"י {-1,1} מכיוון שייצג שני ערכים אפשריים.

מאפיינים מספריים רציפים נורמלו לתחום [-1,1] אם הכילו ערכים חיוביים ושליליים כאחד, ע"י נרמול max_abs כדי לשמור על חיוביות/שליליות של ערכים.

מאפיינים חיוביים בלבד, או מאפיינים מספריים בדידים נורמלו לתחום [0,1].

התייחסות מיוחדת ניתנה למאפיין CurrentLocation, שמייצג מעלות רוחב/אורך. כדי לשמור על יחס בין ערכי כל נ.צ ערכי מעלות אורך ורוחב נורמלו בפקטור 180.

בחירת מאפיינים

בשלב הראשון בפעלנו אלגוריתם Relief (פילטר) איטרטיבי כדי לבחור מאפיינים מיטביים על בסיס מקסימיזציה של מידע. לאחר מכן הפעלנו שיטת Wrapper על בסיס מספר מסווגים, כשלבסוף עשינו שימוש ב Random Forest Classifier שנתן את הציון הגבוה ביותר על סט הואלידציה. הפעלנו את המסווג תחת מעטפת של שלSFS מפני שנראה כי היה לו יתרון משמעותי על פני SBS במקרה הזה מפני אי התאמה של מאפיינים לתיוג, ויתירות גדולה של מאפיינים.

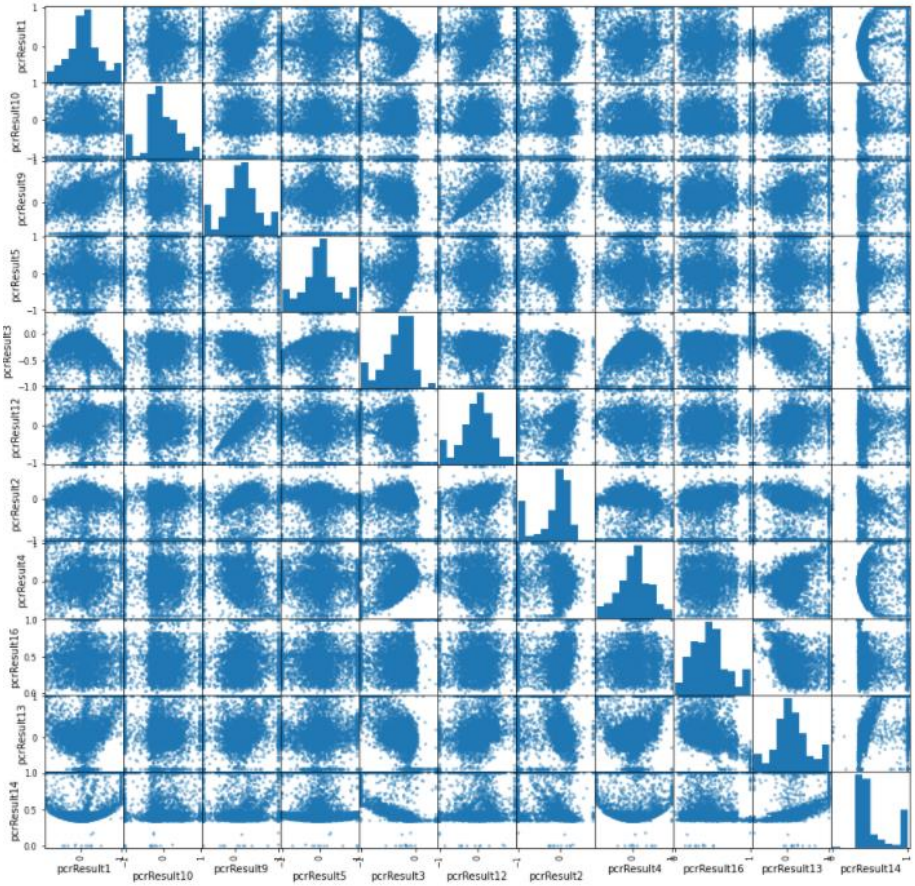
תוצאות אלגוריתם Relief:

AvgHouseholdExpenseOnPresents, AvgHouseholdExpenseOnSocialGames, AvgHouseholdExpenseParkingTicketsPerYear, AvgMinSportsPerDay, HappinessScore, NrCousins, StepsPerYear, TimeOnSocialActivities, pcrResult1, pcrResult10, pcrResult12, pcrResult13, pcrResult14, pcrResult16, pcrResult2, pcrResult3, pcrResult4, pcrResult5, pcrResult9, SyndromeClass_2.0

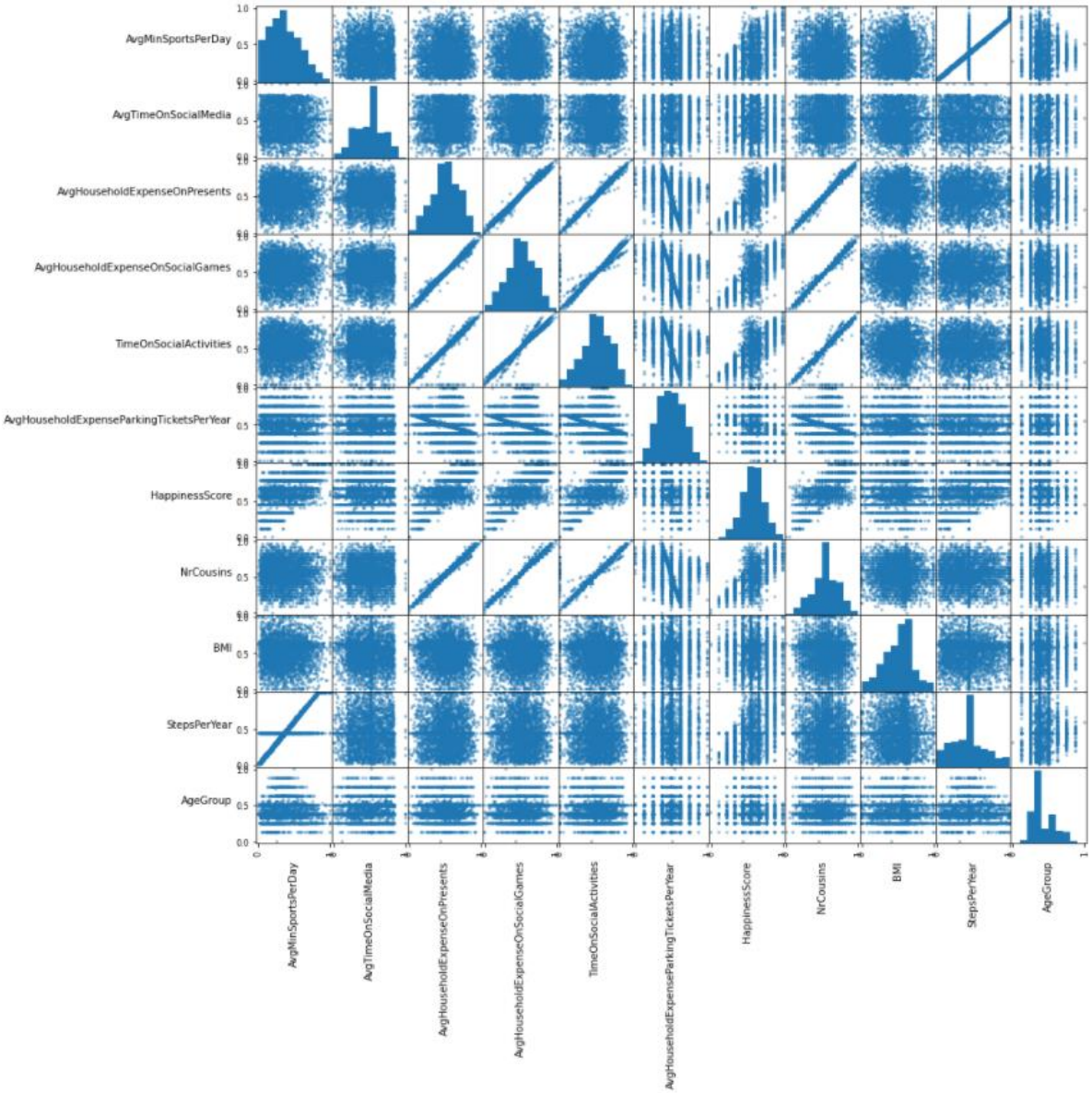
תוצאות Wrapper על מסווג Random Forest:

AgeGroup, AvgHouseholdExpenseOnSocialGames, AvgHouseholdExpenseParkingTicketsPerYear, AvgMinSportsPerDay, AvgTimeOnSocialMedia, BMI, pcrResult1, pcrResult12, pcrResult14, pcrResult16, pcrResult2, pcrResult4, pcrResult9, BloodType_AB+, BloodType_O-, SyndromeClass_1.0, CurrentLocation_Long, Muscle_or_body_aches, Shortness_of_breath, New_loss_of_taste_or_smell

לאחר מכן בחנו קורלציה והתפלגות משותפת של כלל המאפיינים שנבחרו. מאפייני בדיקות PCR נבחנו אחד כנגד השני, ופרט להתאמה שנצפתה עוד קודם בין בדיקות מסוג 3 ו 16 לבדיקה מסוג 13, לא נצפו ממצאים חדשים.



שאר המאפיינים נבחנו אחד כנגד השני ונראה שקיימים מספר מאפיינים עם קשר לינארי חזק ביניהם. מכל קבוצה כזו נבחר מאפיין אחד בלבד:



את האיחוד של קבוצות מאפיינים אלו הרצנו פעם נוספת באלגוריתם בחירת מאפיינים, הפעם SBS מפני שבחרנו להשאיר 18 מאפיינים מתוך הקבוצה המצומצמת, לכן הסרה של מאפיינים היתה יעילה יותר.

ובסופו של דבר רשימת המאפיינים שהגענו אליה היא:

AvgMinSportsPerDay, AvgTimeOnSocialMedia, AvgHouseholdExpenseParkingTicketsPerYear, BMI, pcrResult1, pcrResult12, pcrResult14, pcrResult16, pcrResult2, pcrResult4, pcrResult9, CurrentLocation, AvgHouseholdExpenseOnPresents, HappinessScore, pcrResult10, BloodType, SyndromeClass, SelfDeclarationOfIllnessForm.