**הסבר וורד מטלה 3 – COVID-19**

ליעד בן יחיאל – 207637414

שלב 0 – שאיבת הדאטה

כשלב ראשוני נכנסתי לקובץ הCVS שצורף ובעקבות הבנת משמעות העמודות וההוראות המתבקשות החלטתי איזה עמודות לקחת כמשתנים בלתי תלויים ואילו משתנים כמשתנים תלויים.

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

שלב 1 – ניקוי הדאטה

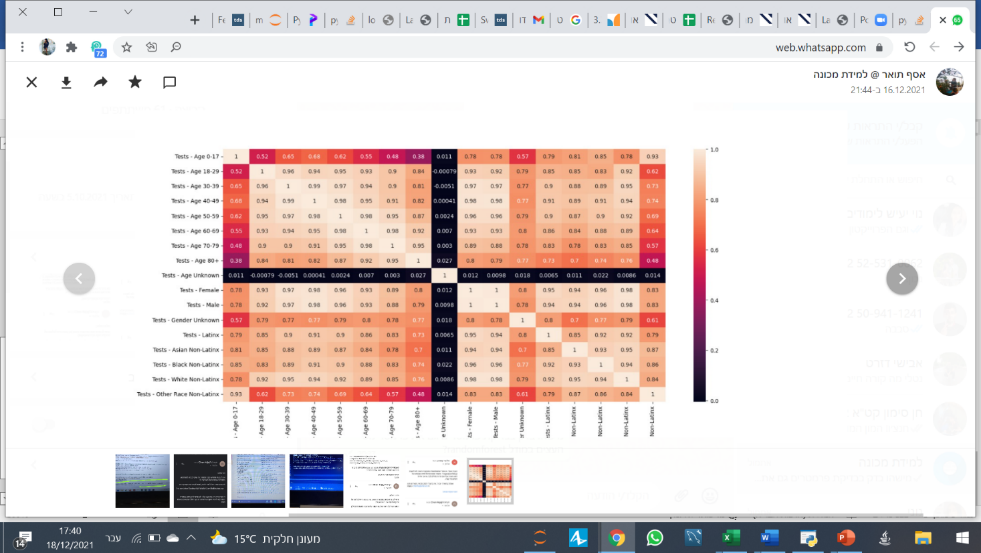
כחלק ראשוני לאחר קביעת המשתנים התלויים והבלתי תלויים למדנו שעלינו לנקות את הדאטה. בעקבות צפייה על הדאטה ובדיקה נצפה שישנם עמודות ושורות עם ערכים נורא גבוהים מכיוון שהיו באותו יום בדיקות רבות וימים עם ערכים נורא נמוכים.

ולכן נערך סקיילינג לכלל הנתונים כפי שנלמד.

תמונה שמכילה טקסט, כתום

התיאור נוצר באופן אוטומטי

כמו התבצע ניקוי לפיצ'ארים שביניהם ניכרת קורולצייה גבוהה (מ0.975).

החקירה על הקורולציה התבצעה רק על הTRAIN SET אך הושמטו גם העמודות בהתאם בTEST SET על מנת לשמור על אותו גודל מערך נתונים.

בעקבות מפת קורולציה שבה נצפה לרוב קורלציות גבוהות שהורידו פיצ'ארים רבים מידיי נקבע נקודת החלטה לקורולצייה גבוהה מ0.975

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

\*בעקבות חקר נתונים לא ניכרה סיבת להוצאת OUTLIERS הוצאת שורות/עמודות ריקות, מילוי תאים ריקים וכד'

שלב 2 - המודלים

2.0 פיצול דאטה

כל הדאטה פוצל לTrain set וTest set 20%-80%



ובנוסף בניתי פונקציה שמעריכה על המודל במספר פרמטרים שנלמדו.

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

KFOLD\* שנבדק בכל המודלים:



כל בנאי המודלים למעט רגרסיה לינארית נבנו לפני ההרצה והוכנסו לליסט יחד עם שם המודל ולאחר מכן בלולאה רצתי על כל המודלים כולל בדיקות הייפרמטרס ייחודיות.

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי



לאחר כל ריצה אופטימלית על מודל הוכנס ערך הr2 עם שם המודל לליסט לצורך השוואה בסוף. (בתוך הפונקציה הוגדר הליסט כGlobal).

2.1 – Linear regression

שלב א – לאחר ניקוי הדאטה בשלב 1 הוספתי עמודה של 1 לפני הCONSTANT כפי שנלמד על מנת "לתת" משקל אליו.

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

שלב ב – ביצוע בחירת פיצ'ארים רלוונטיים לפי p-values באמצעות שיטתbackward elimination. לולאה שמוריד כל פעם את הפיצ'אר בעל ה p-values הגבוה ביותר עד שכל הפיצ'ארים בעלי p-values קטן מ0.05. כמו קודם הבדיקה התבצעה על הTRAIN SET אך התוצאה הושלכה גם על הTEST SET.

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

שלב ג – התבצעה הפרדיקציה לפי בנאי רגרסיה לינארית הודפסו תוצאות כמה מדדים נבחרים.

דוגמא לפי מספר חיוביים:

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

2.2 - Polynomial regression

שלב א – נבנה בנאי למודל רגרסייה לינארית שעליו יתבסס הרגרסייה הפולינומית.



שבה נכנס בדיקה האם כדאי לבצע נורמליזצייה על הנתונים בנוסף או לא.



לאחר מכן התבצע בדיקה על המודל ונבדק על מעלה בגודל 1-5 מה הכי טוב והוחזר הפתרון הטוב ביותר עם המעלה המיטבית.

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

תוצאות הייפרמטרס עבור חיוביים: החזקה הטובה ביותר – 1.

תוצאות הייפרמטרס עבור שליליים: החזקה הטובה ביותר – 1.

2.3 - Ridge regression

נבנה בנאי ולאחר מכן התבצע חיפוש מיטבי לאלפא בין 0 ל0.1 בטווחים של 0.01 והוחלט על פי r2 על פיו נקבע האלפא האופטימלית ואומן המודל ולאחר מכן הוחזרו המדדים הנבדקים ותוצאת r2 הוכנסה לליסט לצורך השוואה סופית.

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

תוצאות הייפרמטרס עבור חיוביים: אלפא אופטימלית – 0.04 .

תוצאות הייפרמטרס עבור שליליים: אלפא אופטימלית – 0.05 .

2.4 - Lasso regression

נבדקו אלפא בטווח 0 עד 1 בטווח של 0.01.



כמו כן נבדק האלפא המיטבית, המודל אומן בהתאם והוחזרו ערכי המדידה של המודל.

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

תוצאות הייפרמטרס עבור חיוביים: אלפא אופטימלית – 0.16 .

תוצאות הייפרמטרס עבור שליליים: אלפא אופטימלית – 0.54 .

2.5 - Random Forest

נבדקו הפרמטרים המתבקשים – מספר העצים: 50-120 ברווח של 10 עצים.

ועומק העץ: 4-14 ברווח של 1.

המודל אומן בהתאם והוחזרו ערכי המדידה של המודל.

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

תוצאות הייפרמטרס עבור חיוביים: גודל עומק – 10, מספר עצים – 110.

תוצאות הייפרמטרס עבור שליליים: גודל עומק – 8, מספר עצים – 100.

2.6 – KneighborsRegressor

במודל זה כמתבקש לאחר הכנסת המשתנים למודל הוערכו מספר שכנים K אופטימלי לבדיקה והובחן באמצעות r2 הגדול ביותר.

בנוסף נבחן מדד p אשר בוחן האם עדיף למדוד את המרחקים על פי מרחק אוקלידי או מרחק מנהטן.

המודל אומן בהתאם והוחזרו ערכי המדידה של המודל.

תוצאות הייפרמטרס עבור חיוביים: מספר שכנים – 1 , מרחק מנהטן.

תוצאות הייפרמטרס עבור שליליים: מספר שכנים – 1, מרחק מנהטן.

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטי

שלב 3 – תוצאות

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטיבקוד הוצא כפלט לכל מודל מה הHYPERPARAMETERS המיטביים ותוצאות מדדים mse, rmse, mae, r2 score . לצורך העניין אציג את תוצאות r2 של כל המודלים.

תוצאות המודל לחיזוי מספר החיוביים

תוצאות המודל לחיזוי מספר השליליים

לצורך בחירת המודל חיזוי הטוב יותר עבור חיזוי שליליים וחיוביים נבחן באמצעות r2 score כך:

תמונה שמכילה טקסט

התיאור נוצר באופן אוטומטיתמונה שמכילה טקסט

התיאור נוצר באופן אוטומטיוהתוצאות:

שלב 4 – מסקנות

שלב ניקוי הדאטה חשוב ואך קריטי אך קודם כל יש לשים לב כיצד הדאטה מתנהג.

במהלך המטלה נאלצתי לחקור את הנתונים, לפלח את חשיבות הימים בגרפים ועוד. על מנת להבין כיצד כדאי לי לנקות את הדאטה.

למשל התפלאתי למצוא כיצד בגלל התנהגות הדאטה כלל החיזויים בתחום השלילי יצאו עם r2 גבוה.

לאחר מכן בעת אימון המודלים וחזייה בתוצאות הבנתי את החשיבות לHyperParameters השונים וכיצד הטווחים של החישוב שלהם צריכים להתבצע על מנת להקל על המעבד בזמן ריצה ולהגיע לתוצאות טובות יותר.

ובאילו מדדי הערכה כדאי להשתמש בחישובים השונים במודלים.