



אוניברסיטת תל אביב

הפקולטה להנדסה ע"ש איבי ואלדר פליישמן

מדעים להיי-טק

עבודה מסכמת בקורס מבוא ללמידת מכונה

מרצה: דור בנק

מתרגל: אילן וסילבסקי

מגישות : נועה ריפינסקי 206572018, מאי מגן 209373356

30/06/2023, סמסטר ב' תשפ"ג

מבוא -

במסגרת עבודה זו עסקנו בסיווג בינארי של קבצי הרצה (*exe*). ומטרתנו הייתה לחזות האם הקבצים הם זדוניים (1) או לא זדוניים (0) תוך הרצת מודלים שונים על סט הנתונים. במהלך עבודתנו בחנו את הנתונים הקיימים ואת אופן התפלגותם, טיפלו בערכים החסרים ובערכים הקיצוניים, חקרנו את ממדיות הבעיה וביצענו מספר מניפולציות על הפיצ'רים הקיימים. בסופו של דבר ערכנו השוואה בין ארבעה מודלים אפשריים לבעיית הסיווג, תוך חלוקת מערך הנתונים לסט אימון וסט ולידציה בו השתמשנו כדי לדמות את סט המבחן אותו נרצה לחזות. עבור כל מודל בדקנו אילו היפר-פרמטרים הם האופטימליים כדי לקבל את הציון הטוב ביותר ולאחר הרצתם ערכנו השוואה בין ביצועיהם על סט האימון וסט הוולידציה, בדקנו את פערי הביצועים וחקרנו האם יש *overfitting* (התאמת יתר של המודל). בסופו של דבר, לאחר בחינת כלל המודלים והתוצאות, בחרנו במודל יער רנדומלי (*Random Forest*) בכדי לסווג באמצעותו את סט המבחן.

חלק ראשון - אקספלורציה

במסגרת חלק זה עיקר עבודתנו הייתה חקירת הנתונים והפיצ'רים השונים – מהם סוגי הפיצ'רים (משתנים בינאריים, קטגוריאליים ורציפים), עבור כל פיצ'ר רציף הסתכלנו על סטטיסטיים וראינו מה ניתן ללמוד עליו. בשלב הבא, בחנו בעזרת היסטוגרמה אם הדאטה שקיבלנו מאוזן, כלומר האם סוג הקובץ (זדוני או לא) מחולק באופן שווה לסוגים השונים. בדיקה זו עוזרת למנוע מהמודל שבבנה בהמשך להיות מוטה כלפי מחלקה אחת - מחלקת הרוב, רק בגלל שהיא מכילה יותר נתונים. ראינו שהדאטה אכן מאוזן (נספח 2). לאחר מכן, רצינו לראות כיצד כל פיצ'ר מתפלג, הן בפני עצמו והן כתלות בלייבל (נספח 3). בעזרת היסטוגרמה עבור משתנה מספרי או גרף ברים עבור משתנים קטגוריאליים ובינאריים חיפשנו אירועים "מעניינים". כך למשל הבחנו שעבור כל קטגוריה של המשתנה C יש יחס דומה של הלייבלים 0 ו-1. דבר זה למעשה מלמד אותנו שמשתנה זה לא מסייע בהפרדה בין זדוני ולא זדוני ולכן כפי שניתן לראות בהמשך, בחרנו להסירו. ניתוח דומה ניתן לעשות למשתנה A . דבר מעניין נוסף אליו שמנו לב הוא המשתנה הקטגוריאלי *file_type_trid* שמחולק למספר גדול מאוד של קטגוריות וחלקן עם כמות דוגמאות קטנה ביותר. בסיום חלק זה יצרנו מטריצת קורלציה (נספח 4) של כלל המשתנים, ממנה למדנו על היחסים בין הפיצ'רים השונים. יחס מעניין לדוגמא הוא קורלציה גבוהה בין המשתנה *size* למשתנה *numstrings*. ניגע בקשר זה בהמשך.

חלק שני- עיבוד מקדים

בחרנו בסדר הפעולות הבא על מנת לעבד את הדאטה שלנו -

1. בניית פיצ'רים חדשים וביצוע מניפולציות מתמטיות על פיצ'רים קיימים - זהו השלב הראשון לאור ההבנה שפיצ'רים אלה הופכים להיות "פיצ'רים מן המניין" וגם עליהם נרצה לבצע את העיבוד המקדים
2. התמודדות עם משתנים קטגוריאליים - זה יהיה השלב הבא כי על משתנה הקטגוריאלי לא יתאפשר לבצע מילוי ערכים חסרים וסטנדרטיזציה, אך עבור משתנה ה-*dummy* שלו זה אפשרי. בשלב זה העתקנו ופיצלנו כל עותק לסט אימון וסט ולידציה - עותק אחד ישמש למודלים *KNN* ורגרסיה לוגיסטית והעותק השני ישמש למודלים מבוססי עצים - (נספח 9) מציג איזה עיבוד בוצע על איזה עותק.
3. ביצוע סטנדרטיזציה לנתונים - חשוב לבצע שלב זה לפני מילוי הערכים החסרים ולפני התמודדות עם ערכים חריגים על מנת למלא ולאזן את הערכים המדוברים בערכים לאחר שינוי קנה המידה .
4. הורדת ממדיות הבעיה ע"י מחיקת פיצ'רים וביצוע PCA - נבצע זאת לפני ההתמודדות עם ערכים "מיוחדים" על מנת לא לבצע עבודה על פיצ'רים שגם ככה נוריד.

5. התמודדות וטיפול בערכים חריגים - לאחר שינוי קנה המידה של הדאטה נכון יותר לטפל בערכים החריגים שכן אולי כעת חלקם כבר לא חריגים.
6. התמודדות וטיפול בנתונים חסרים - לאחר שינוי קנה המידה של הדאטה נכון יותר לטפל בערכים החסרים שכן כעת המילוי יהיה נכון יותר.
7. החלת העיבוד המקדים על סט המבחן - במקרים בהם יש הסתמכות על סטטיסטיים (לדוגמא ממוצע וסטיית תקן עבור סטנדרטיזציה) השתמשנו בערכים שמצאנו בסט האימון והחלנו אותן על סט המבחן על מנת למנוע זליגה של מידע מסט המבחן שעשוי להשפיע לנו על התוצאות.

בניית פיצ'רים חדשים - חשבנו שיהיה מעניין לבחון מספר קשרים בין הפיצ'רים השונים בדאטה. היחס בין *imports* ל- *exports*; היחס בין *printables* ל- *symbols*; והיחס בין מספר פיצ'רים לבין הפיצ'ר *size*, כך למעשה נכניס סוג של קנה מידה לפיצ'רים אלו. לדוגמא נצפה שעבור מספר גדול של סימבולים גם גודל הקובץ יהיה גדול ולהיפך. במידה ונראה יחס גדול מאוד/קטן מאוד (מספר הסימבולים מאוד גדול וגודל הקובץ ממש קטן) ייתכן ונסיק כי מדובר בתצפית חריגה. מכמו כן, הצגנו את התפלגות הפיצ'רים החדשים ואת ההתפלגות שלהם כתלות בלייבל ([נספח 5](#)) במקרים בהם ראינו שההתפלגות של הפיצ'ר על גבי הלייבל היא דומה, הנחנו כי ניתן להסיר את הפיצ'ר.

התמודדות עם משתנים קטגוריאליים - בחרנו לטפל במשתנה קטגוריאלי *file_type_trid* המכיל 89 קטגוריות באופן הבא - בחרנו את הקטגוריות הנדירות- כלומר אלו שמכילות אחוז קטן מאוד מהתצפיות (1% כלל אצבע) וצמצמנו אותן לקטגוריה אחת בשם *other_type*. כעת נשארו עם 21 קטגוריות ([נספח 6](#)) לאחר מכן, היה לנו רעיון לאחד את הקטגוריות לפי המילה הראשונה במחרוזת של שם הקטגוריה (לדוגמא- רצינו לאחד את כל הקטגוריות שמתחילות עם 'Win32') אבל לפני הביצוע חקרנו וגילינו שההתפלגות המשותפת של כל קטגוריה כזו עם הלייבל שונה ([נספח 6](#)), ועל כן הסקנו שאם נאחד אותן, נאבד מידע שיכול לעזור לפרדיקציה. לכן לבסוף בחרנו שלא לאחד קטגוריות נוספות. בשלב הבא, פיצלנו את המשתנה הקטגוריאלי בעזרת *OneHotEncoder*, יכולת המשמשת להמרת משתנים קטגוריים לייצוג מספרי. עשינו זאת כדי שבהמשך יהיה ניתן להשתמש בפיצ'ר למודלים מסוימים שלא יודעים להתמודד עם משתנים קטגוריאליים. ביצענו *OneHotEncoder* 'סטנדרטי' בו כל קטגוריה ספציפית מקבלת 0 אם התצפית היא לא מקטגוריה זו ו-1 אם היא משתייכת לאותה קטגוריה. לבסוף, ביצענו מניפולציה על הפיצ'רים שיצאו מה- *OneHotEncoder* והכפלנו אותם בפיצ'ר *file_type_prob_trid* (ההסתברות לסוג הקובץ) כך שלמעשה במקום הערך 1 תחת סוג הקובץ הרלוונטי, תופיע ההסתברות לסוג קובץ זה.

ביצוע סטנדרטיזציה לנתונים - כפי שראינו בחלק הראשון ([נספח 3](#)) הפיצ'רים אינם מנורמלים והם לא בקנה מידה שהוא בר השוואה. על מנת לטפל בזה שלפיצ'רים יש קני מידה שונים, בחרנו לבצע סטנדרטיזציה. מכיוון ומדובר בפעולה מתמטית קונסיסטנטית, ההשפעה שלה תהיה בעיקר במודלים כגון רגרסיה לינארית ו-*KNN* בהם יש חשיבות לערך המספרי של הפיצ'ר ולא רק לתחום שבו הוא נמצא (כמו בעצים שבדקים האם הערך גדול מערך מסוים או קטן ממנו) לכן בשלב זה, בנינו שני דאטה סט שונים כאשר כל אחד יתאים למודלים אחרים שנבחר להריץ.

הורדת ממדיות הבעיה - נסתכל על היחס בין כמות התצפיות שיש לנו לכמות הפיצ'רים שיש לנו - באופן כללי כמות הפיצ'רים לעומת גודל הדאטה היא לא גדולה במיוחד אך בכל זאת בחרנו להסיר כמה פיצ'רים ידנית - פיצ'ר A – התפלגות התצפיות עם לייבל 0 דומה מאוד להתפלגות התצפיות עם לייבל 1 (נספח 3) ולכן אנו מצפות שפיצ'ר זה לא יוסיף לנו מידע הכרחי לסיווג הבעיה. מסיבה דומה הסרנו את פיצ'ר C – תחת כל קטגוריה מספר התצפיות עם לייבל 0 דומה מאוד לתצפיות עם לייבל 1 (נספח 3). כך ביצענו גם עבור $numstrings$ משתנה $symbols/printables$ ו- $symbols/size$. בחרנו גם להסיר את משתנה $numstrings$ משום שהוא נמצא בקורלציה גבוהה עם המשתנה $size$ (נספח 4) ומכיל ערכים חסרים רבים ביחס ל- $size$, שלא מכיל ערכים חסרים כלל (נספח 7). בנוסף, ביצענו PCA להורדת ממדיות הבעיה תוך שמירה על 99% מהשונות וירדנו מ-45 פיצ'רים (לאחר איחוד קטגוריות של $file_type_trid$ ל-26. לפני ביצוע ה-PCA ביצענו סטנדרטיזציה לכל הדאטה של ה- $train$ שכן חייבים לתת לאלגוריתם זה דאטה מנורמל, אחרת הוא ייתן חשיבות רבה יותר לערכים בקנה מידה גדול יותר. לבסוף שמרנו את הפיצ'רים הנבחרים של ה-PCA, איתם נמשיך לחלק מהמודלים.

ערכים חריגים - בחרנו לטפל בערכים החריגים על פי כלל אצבע ולפי הסתכלות בגרפים השונים עבור כל פיצ'ר. עבור כל פיצ'ר החלטנו על $threshold$ שמעבר לו הדוגמאות נחשבות חריגות (לדוגמא אחוזון 98) לפי הסתכלות בהתפלגות ולאחר מכן לכל הערכים החריגים באותו פיצ'ר נתנו את הערך של ה- $threshold$ (נספח 8) העדפנו שיטה זו על פני מחיקת דוגמאות כדי לא לאבד מידע ודוגמאות רבות ובאותו זמן לשמור על החריגות של הדוגמאות אך רק להקטינה.

נתונים חסרים - בדקנו כמה נתונים חסרים יש בכל קטגוריה (נספח 7) ולאחר מכן בחרנו כמה גישות למילוי ערכים חסרים. כפי שראינו במטריצת הקורלציה (נספח 4) יש קורלציה גבוהה בין המשתנים $size$ ו- MZ ולכן בחרנו למלא את הערך של $size$ במקומות במקום הערכים החסרים של MZ עבור כל תצפית. את הערכים החסרים במשתנים הבינאריים בחרנו למלא עם הערך השכיח של כל פיצ'ר. את הערכים החסרים במשתנים המספריים בחרנו למלא עם הערך הממוצע של כל משתנה.

השלב האחרון של העיבוד המקדים הוא החלתו גם על סט המבחן. כפי שנראה בהמשך, המודל שבחרנו הוא *Random Forest* לכן נבצע על סט המבחן רק את העיבוד שרלוונטי למודל זה.

חלק שלישי – הרצת המודלים

המודלים אותם בחרנו להריץ הם - *KNN*, *Logistic Regression*, *Decision Tree*, *Random Forest* עבור כל מודל הרצנו *GridSearchCV* עם $cv = 5$ על מנת לבחור את ההיפר-פרמטרים הטובים ביותר כדי למקסם את ציון ה- roc_auc .

1. *KNN* – למודל זה הכנסנו את סט האימון לאחר ביצוע *PCA*, מילוי ערכים חסרים וסטנדרטיזציה מתוך הידיעה כי מודל זה דורש כל אחד מאלו. (נספח 12 מציג את ההיפר-פרמטרים שנבחרו ואת השפעתם על המודל בהיבטי שונות והטיה).

2. *Logistic Regression* - למודל זה הכנסנו את סט האימון לאחר ביצוע *PCA*, מילוי ערכים חסרים וסטנדרטיזציה מתוך הידיעה כי מודל זה דורש כל אחד מאלו. (נספח 13 מציג את ההיפר-פרמטרים שנבחרו ואת השפעתם על המודל בהיבטי שונות והטיה).

3. *Decision Tree* - למודל זה הכנסו את סט האימון לאחר ביצוע מילוי ערכים חסרים וסטנדרטיזציה. ידוע כי מודלים מבוססי עץ יכולים להתמודד עם מספר פיצ'רים גדול לכן לא ביצעו *PCA* עבור מודל זה. כמו כן, סטנדרטיזציה אינה נדרשת למודל זה אך גם אינה מזיקה ולאחר בדיקה היא אף תרמה במעט. **נספח 15** מציג את ההיפר-פרמטרים שנבחרו ואת השפעתם על המודל בהיבטי שונות והטיה.
4. *Random Forest* - למודל זה הכנסו את סט האימון לאחר ביצוע מילוי ערכים חסרים וסטנדרטיזציה. ידוע כי מודלים מבוססי עץ יכולים להתמודד עם מספר פיצ'רים גדול לכן לא ביצעו *PCA* עבור מודל זה. כמו כן, סטנדרטיזציה אינה נדרשת למודל זה אך גם אינה מזיקה ולאחר בדיקה היא אף תרמה במעט. **נספח 16** מציג את ההיפר-פרמטרים שנבחרו ואת השפעתם על המודל בהיבטי שונות והטיה. בעזרת feature importance בחנו את התרומה של כל פיצ'ר להצלחת המודל (**נספח 14**). כפי שניתן לראות לפיצ'רים *avlength*, *imports/exports*, *B* הייתה את ההשפעה הכי גדולה על המודל. בחלק 6 נמשיך לדון בתרומה של כל פיצ'ר.

חלק רביעי- הערכת מודלים

לאחר הרצת כל אחד מארבעת המודלים לעיל וחקירת התוצאות לעומק – המודל הנבחר הוא *Random Forest* בעל הביצועים הטובים ביותר. בשלב הראשון ננתח את ה-*Confusion Matrix* עבור מודל זה (**נספח 11**)

- "חיובי אמיתי" (*TP*) - חזינו 1 והערך האמיתי הוא 1 (צדקנו) - חזינו שהקובץ זדוני והוא אכן זדוני
- "חיובי שגוי" (*FP*) - חזינו 1 והערך האמיתי הוא 0 (טעינו) - חזינו שהקובץ זדוני והוא למעשה לא
- "שלילי שגוי" (*FN*) - חזינו 0 והערך האמיתי הוא 1 (טעינו) - חזינו שהקובץ לא זדוני והוא למעשה זדוני
- "שלילי אמיתי" (*TN*) - חזינו 0 והערך האמיתי הוא 0 (צדקנו) - חזינו שהקובץ לא זדוני והוא אכן לא

מסקנות עיקריות –

1. *Accuracy* = כמה קבצים תויגו נכון מתוך כל הקבצים בנתונים. נרצה שהערך הזה יהיה גבוה ככל האפשר כדי לחזות נכון אחוז גדול ככל האפשר מהדוגמאות החדשות $\frac{(TP + TN)}{(TP + FP + FN + TN)} \approx 0.932$
 2. *Precision* = כמה מהקבצים המתויגים כקבצים זדוניים הם באמת זדוניים. נרצה שהערך הזה יהיה גבוה ככל האפשר, כדי שהתחזית שלנו תהיה טובה ככל האפשר וללא שגיאות. $\frac{TP}{TP+FP} \approx 0.949$
 3. *Recall/Sensitivity* = מכל הקבצים הזדוניים, כמה מהם חזינו נכון (אמרנו שהם זדוניים). נרצה שהערך הזה יהיה גבוה ככל האפשר, כדי שהתחזית תהיה טובה ככל האפשר וללא שגיאות. $\frac{TP}{TP+FN} \approx 0.919$
 4. *Specificity* = מכל הקבצים הלא זדוניים, כמה מהם חזינו נכון (אמרנו שהם לא זדוניים). נרצה שהערך הזה יהיה גבוה ככל האפשר, כדי שהתחזית תהיה טובה ככל האפשר וללא שגיאות. $\frac{TN}{TN+FP} \approx 0.947$
 5. *False alarm* = מכל הקבצים הלא זדוניים, כמה מהם חזינו שגוי (אמרנו שהם זדוניים). נרצה שהערך הזה יהיה נמוך ככל האפשר, כדי שהשגיאה שלנו תהיה קטנה. $\frac{FP}{TP+FP} \approx 0.05$
- כפי שניתן לראות, עבור כל סוג מדידה קיבלנו את התוצאה לה קיוונו, כלומר המודל שלנו חוזה בצורה טובה את הלייבל.

את פערי הביצועים בין הרצת המודל על סט האימון ולאחר מכן על סט הוולידציה ניתן לראות ב-[נספח 19](#) עבור המודל הנבחר, הפער בין ציון ה- AUC הוא כ-0.019125. על מנת להגדיל את יכולת הכללת המודל השתמשנו ב- $GridSearchCV$ על מנת לבחור את ההיפר-פרמטרים הטובים ביותר לצורך הניתוח. ניתן להסיק שהשימוש ב- $random\ forest$ לא סובל מהתאמת יתר לאור העובדה שההפרש בין $roc - auc$ הממוצע בין סט האימון לבין סט הוולידציה נמוך מאוד.

חלק שישי- שימוש בכלים שלא נלמדו בקורס

בחלק זה בחרנו להשתמש ב- $SHAP$. הרעיון הגיע כשרצינו לראות את טעויות המודל $random - forest$ (בהרצתו על סט האימון ולאחר מכן על סט הוולידציה). חקרנו את השימוש בכלי מתמטי זה, המשמש על מנת להסביר איך המודל שהרצנו עובד עבור כל דוגמא, ומאפשר לראות את תרומתו של כל פיצ'ר לחיזוי של המודל. את ההרצה נבצע על הדוגמאות בהן המודל הנבחר שלנו טעה FP -ו- FN כדי לנסות להבין איזה מהפיצ'רים ומהערכים שלהם גרמו לטעויות. הרעיון מאחורי $SHAP$ הוא להקצות ערך חשיבות, המכונה ערך $Shapley$, לכל פיצ'ר בתחזית. ערך $Shapley$ הוא התרומה השולית הממוצעת הצפויה של פיצ'ר מסוים לתחזית, תוך בחינת כל השילובים האפשריים שלו ביחד עם שאר הפיצ'רים.

דוגמה – נרצה לבדוק את התרומה של הפיצ'ר $size$. ערך ה- $shapley$ של הפיצ'ר יחושב כך – נניח לשם הדוגמה כי יש עוד 3 פיצ'רים מלבד $size$ (אחר כך נכליל זאת על כל המודל)

1. הרצת המודל עבור כל אחת מהקומבינציות של שלושת הפיצ'רים האלו (2^3 אפשרויות)
2. הרצת כל אחד מהמודלים של סעיף 1, הפעם בתוספת הפיצ'ר $size$
3. חישוב ההפרש בין כל מודל של סעיף 1 (ללא $size$) לבין המודל המתאים לו מסעיף 2 (עם $size$)
4. חישוב ממוצע ההפרשים – זהו ערך ה- $shapley$.

כך יבוצע החישוב עבור כל פיצ'ר ביחס לכל שאר הפיצ'רים בדאטה.

התרומה של כל פיצ'ר והמסקנות לגביו מוצגות ב- [נספח 21](#)

סיכום -

מטרת העבודה הייתה לחזות האם קבצי הרצה שונים הם זדוניים (1) או לא זדוניים (0). במהלך העבודה חקרנו את הנתונים הקיימים וביצענו עליהם עיבוד על מנת להכניס למודל שלנו את הדאטה המתאים ביותר. לאחר מכן בחנו ארבעה מודלים אפשריים לבעיית הסיווג, ועבור כל מודל בדקנו אילו פרמטרים הם האופטימליים כדי לקבל את הציון הטוב ביותר. ערכנו השוואה בין ביצועי המודלים על סט האימון וסט הוולידציה. בשלב הבא, לאחר בחינת כלל המודלים והתוצאות, בחנו (בעזרת $SHAP$) את טעויות המודל בעל התוצאות הגבוהות ביותר ולבסוף בחרנו במודל יער רנדומלי ($Random Forest$) בכדי לסווג באמצעותו את סט המבחן שלנו.

נספחים

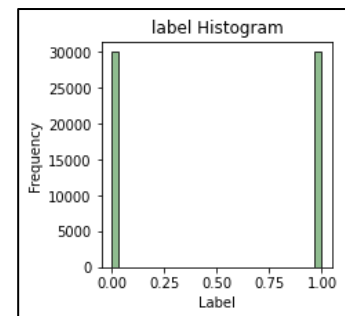
נספח 1 - חלוקת תחומי אחריות

- מאי – טיפול בערכים חסרים, סטנדרטיזציה, הורדת מימדיות, *decision tree*, *random forest*, הערכת המודלים בעזרת $k - fold$
 - נועה – טיפול בערכים חריגים, טיפול במשתנים קטגוריאליים, בניית פיצ'רים חדשים, *KNN*, *Logistic Reg*, בניית *Confusion – matrix*
- מעבר לכך, נעשתה עבודה משותפת בכתיבת הקוד, אקספלורציה של המשתנים, ניתוח הפלטים השונים והתוצאות עבור המודלים השונים, שימוש ב- *SHAP* וכתיבת הדו"ח והמסקנות.

נספח 2 – התפלגות הלייבל



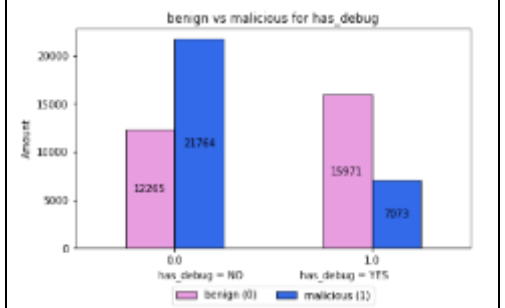
ניתן לראות שהנתונים מאוזנים לחלוטין (30,000 מכל סוג) מסקנה זו תעזור לנו בהמשך בניית המודלים.



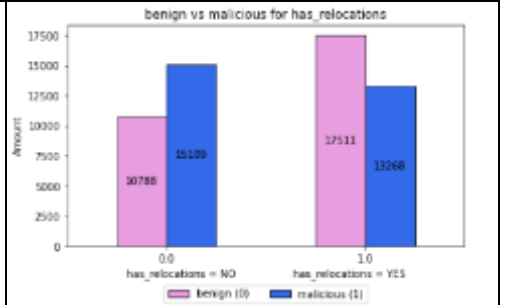
נספח 3 – התפלגות כל פיצ'ר כתלות בלייבל

פיצ'רים בינאריים -

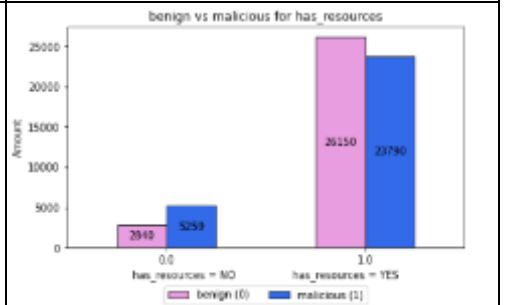
הפיצ'ר *has_debug* אינו מאוזן מבחינת כמות הדוגמאות שיש להן באגים ואלו שלא. יש רוב מוחלט לדוגמאות ללא איתור באגים. כמו כן, גם בתוך כל סוג (עם או בלי באגים) החלוקה לקבצים דדוניים או לא אינה מאוזנת. בנוסף, ניתן לראות שלדוגמאות ללא באגים יש יותר תוויות דדוניות, אבל לדוגמאות עם איתור באגים, רוב התוויות הן לא דדוניות.



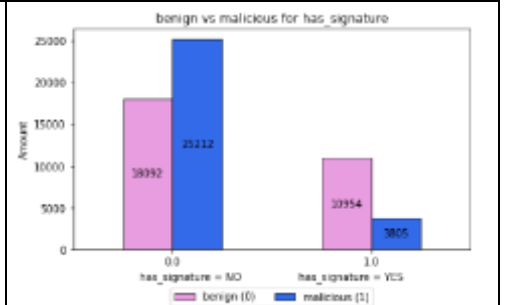
הפיצ'ר *has_relocations* אינו מאוזן מבחינת כמות הדוגמאות שיש להן רילוקיישן ואלו שאין. רוב הדוגמאות הן עם רילוקיישן. כמו כן, גם בתוך כל סוג (עם או בלי רילוקיישן) החלוקה לקבצים דדוניים או לא אינה מאוזנת. בנוסף, ניתן לראות שלדוגמאות ללא רילוקיישן יש יותר תוויות דדוניות לעומת דוגמאות עם רילוקיישן בהן רוב התוויות הן לא דדוניות.



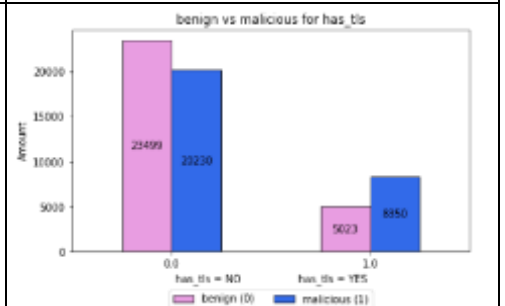
הפיצ'ר *has_resources* אינו מאוזן כלל מבחינת כמות הדוגמאות, הרוב המכריע של הדוגמאות הן עם משאבים. כמו כן, גם בתוך כל סוג (עם או בלי משאבים) החלוקה לקבצים דדוניים או לא אינה אחידה. ניתן לראות שלדוגמאות ללא משאבים יש יותר תוויות דדוניות לעומת דוגמאות עם משאבים בהן רוב התוויות הן לא דדוניות אך הפער בין הכמויות די קטן.



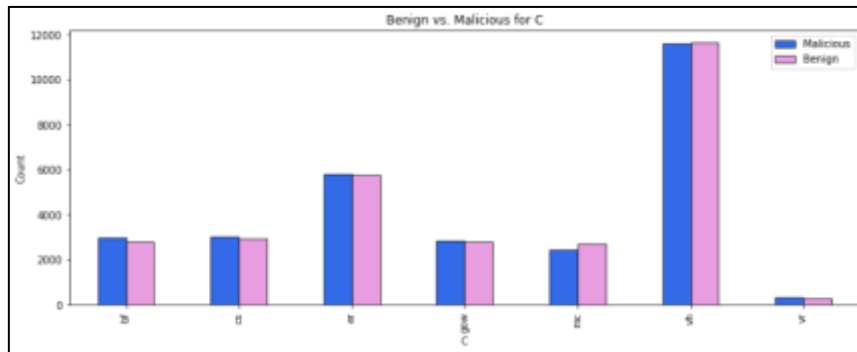
הפיצ'ר *has_signature* אינו מאוזן מבחינת כמות הדוגמאות, הרוב המוחץ של הדוגמאות הן ללא חתימה. כמו כן, גם בתוך כל סוג (עם או בלי חתימה) החלוקה לקבצים דדוניים או לא אינה מאוזנת. ניתן לראות שלדגימות ללא חתימה יש יותר תוויות דדוניות לעומת דוגמאות עם חתימה שבהן רוב התוויות לא דדוניות ושבשתייהן יש פער די גדול בין הכמויות.



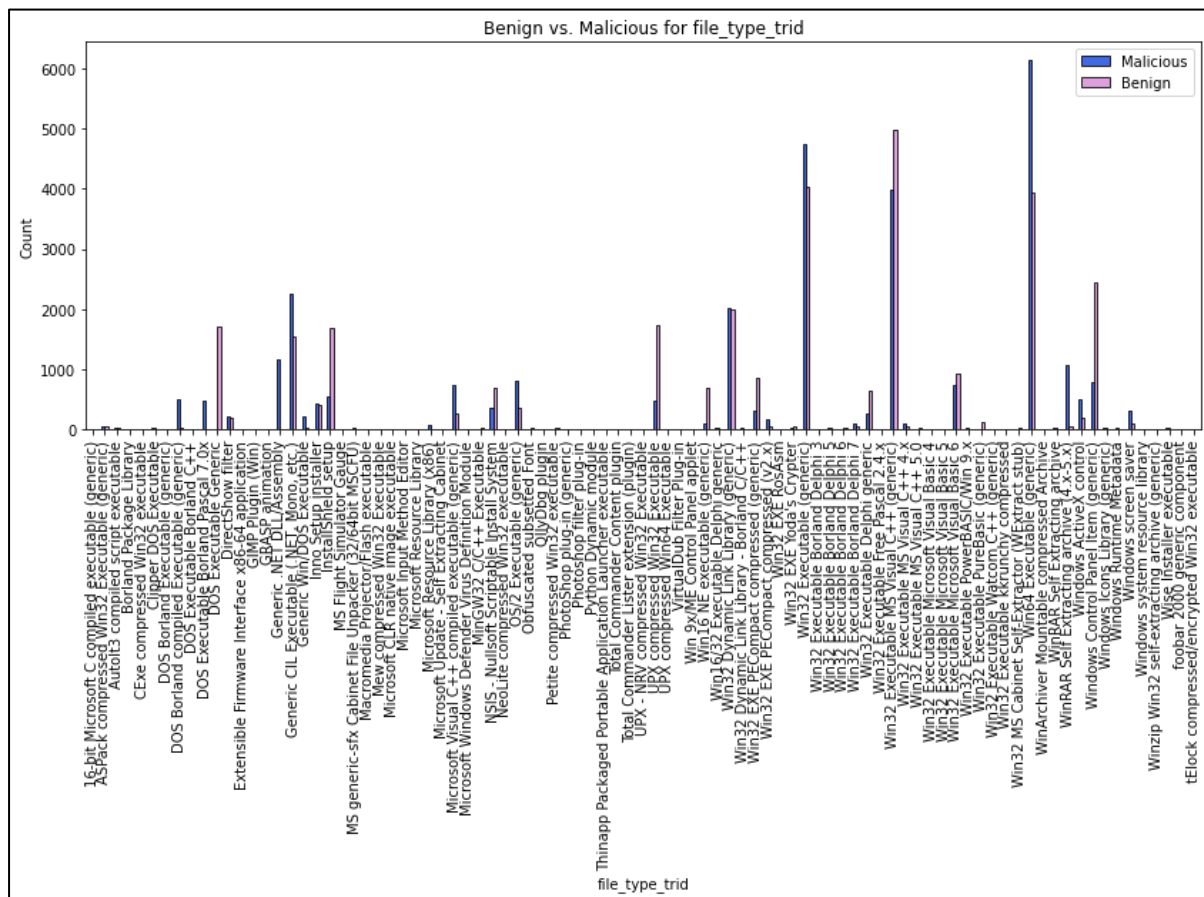
הפיצ'ר *has_tls* אינו מאוזן מבחינת כמות הדוגמאות, הרוב המוחץ של הדוגמאות הן ללא tls. כמו כן, גם בתוך כל סוג (עם או בלי tls) החלוקה לקבצים דדוניים או לא אינה מאוזנת ניתן לראות שלדוגמאות ללא tls יש יותר תוויות לא דדוניות אך הפער בכמויות לא גדול במיוחד, לעומת דוגמאות עם tls בהן רוב התוויות דדוניות.



פיצ'רים קטגוריאליים -



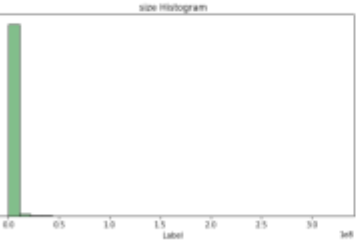
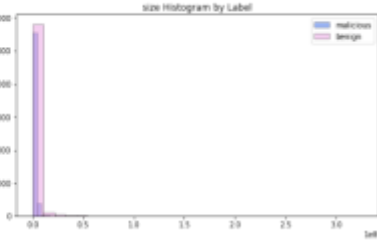
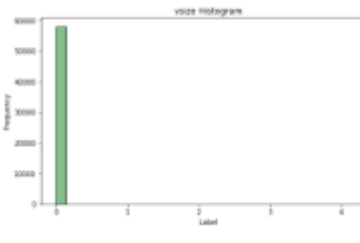
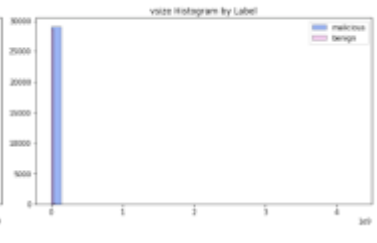
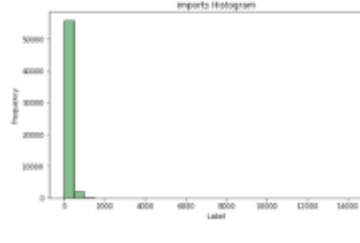
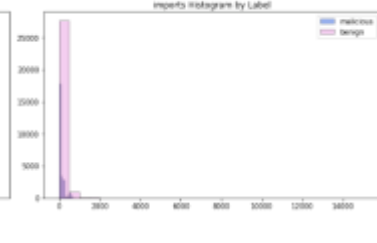
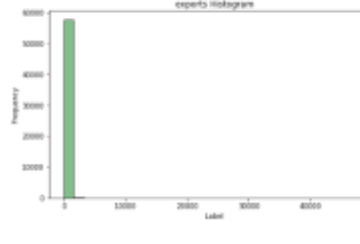
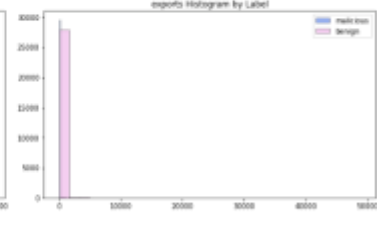
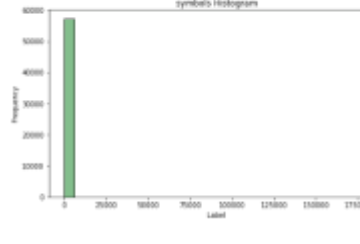
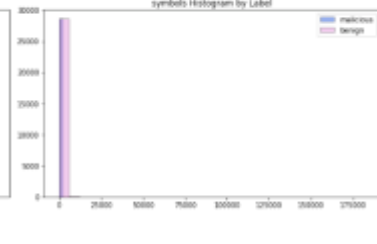
ניתן לראות שהפיצ'ר C מאוזן מאוד בחלוקה של כל אחת מהקטגוריות שלו לזדוני ולא זדוני. בשל מסקנה זו החלטנו להוריד את פיצ'ר זה מכיוון שהוא כנראה לא יתרום לתחזית.

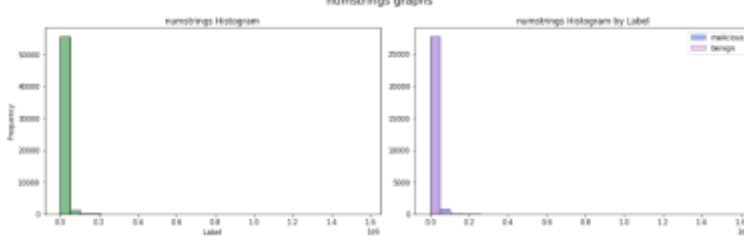
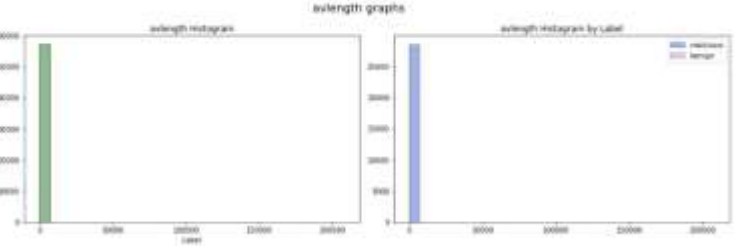
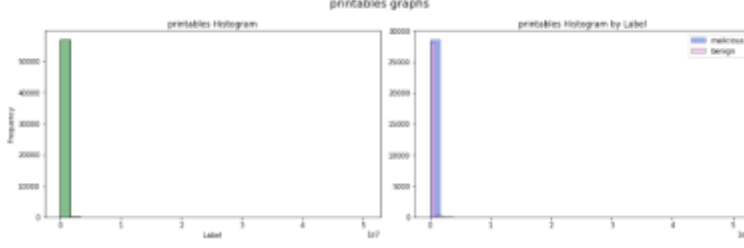
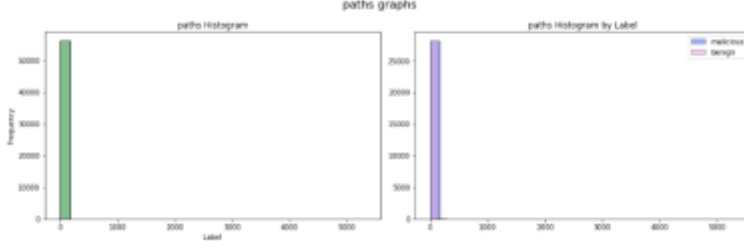
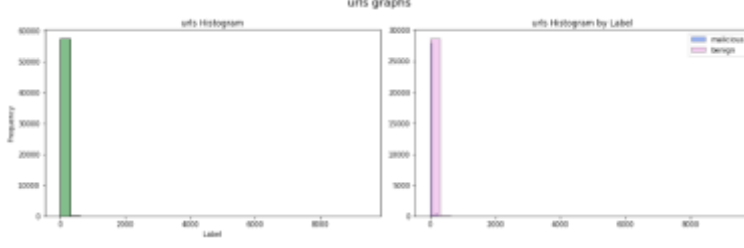


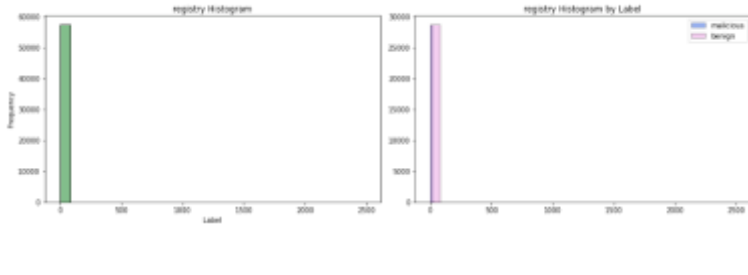
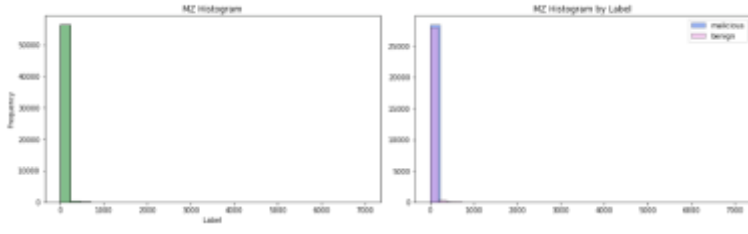
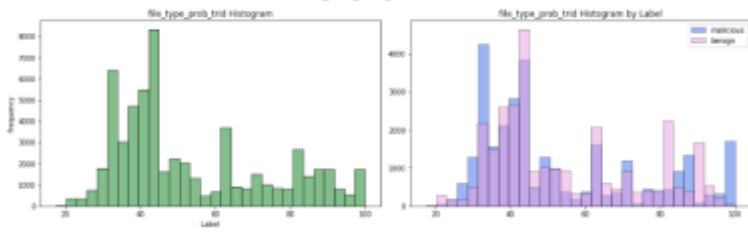
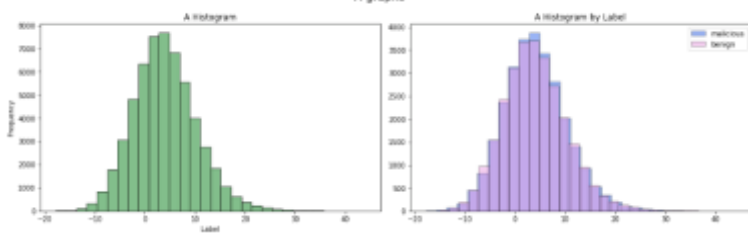
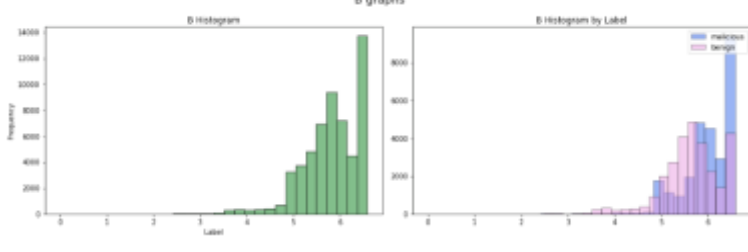
ניתן לראות שלפיצ'ר `file_type_trid` יש כמות גדולה מאוד של קטגוריות ולחלקן יש מספר רב של דוגמאות ולחלק (הגדולות ביותר) יש מספר קטן יותר של דוגמאות.

בהמשך נבחן האם כדאי לאחד קטגוריות "קטנות" על מנת לצמצם את מספר הממדים של הפיצ'ר.

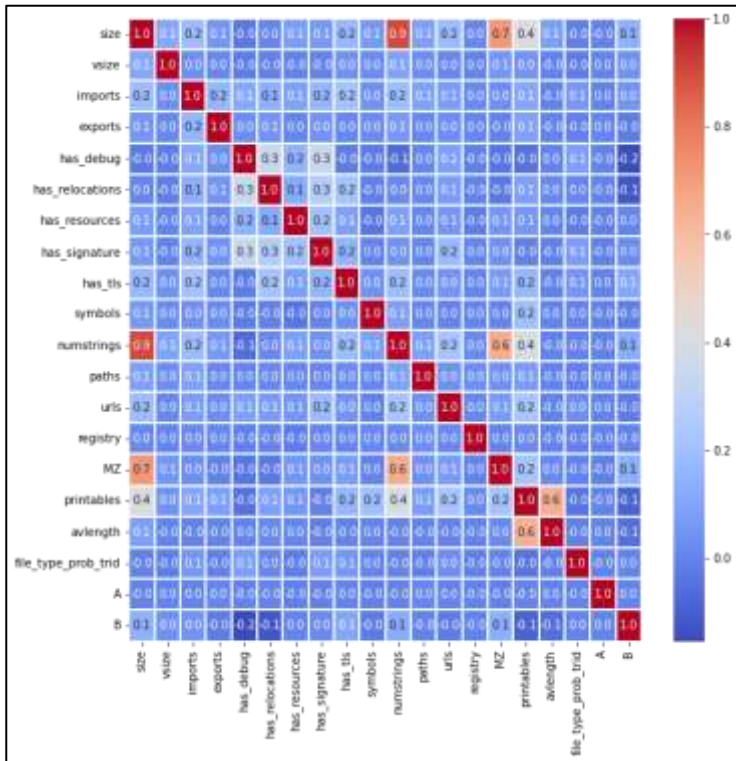
פיצ'רים מספריים -

<p>לפי הגרף ניתן לראות שלרוב הדגימות יש גדלים דומים, אך לפי קנה המידה של ציר ה-X ניתן לראות שיש ערכים חריגים (נעסוק בזה בהמשך). בנוסף, ניתן לראות שיש הבדל בין התפלגות הגדלים בין דוגמאות זדוניות לדוגמאות לא זדוניות. ניתן ללמוד שככל שהגודל גדול יותר, הדוגמאות הן לא זדוניות. מהתבוננות בגרפים הללו, ניתן להניח שהפיצ'ר אינו מתפלג נורמלית.</p>	<p style="text-align: center;">size graphs</p> <div style="display: flex; justify-content: space-around;">   </div>
<p>ההתפלגות של $vsize$ דומה לזו של $size$. על פי הגרף ניתן לראות שלרוב הדגימות יש גדלים דומים, אך לפי קנה המידה של ציר ה-X ניתן לראות שיש ערכים חריגים (נעסוק בזה בהמשך). בנוסף, ניתן לראות שיש הבדל בין התפלגות $vsize$ בין דוגמאות זדוניות לכאלה לא זדוניות. ניתן ללמוד שככל שה-$vsize$ גדול יותר, הדוגמאות הן זדוניות. מהתבוננות בגרפים הללו, ניתן להניח שהפיצ'ר אינו מתפלג נורמלית.</p>	<p style="text-align: center;">vsize graphs</p> <div style="display: flex; justify-content: space-around;">   </div>
<p>לפי הגרף ניתן לראות שלרוב הדגימות יש מספרים דומים וקטנים של יבוא, אך לפי קנה המידה של ציר ה-X ניתן לראות שיש ערכים חריגים (נעסוק בזה בהמשך). לאחר פיצול לפי סוג התווית, ניתן לראות את ההבדלים הגדולים בהתפלגות בין דוגמאות זדוניות לרגילות. ניתן ללמוד שככל שמספר הייבוא גדול יותר, כך נקבל יותר דוגמאות לא זדוניות. מהתבוננות בגרפים הללו, ניתן להניח שהפיצ'ר אינו מתפלג נורמלית.</p>	<p style="text-align: center;">imports graphs</p> <div style="display: flex; justify-content: space-around;">   </div>
<p>לפי הגרף ניתן לראות שלרוב הדגימות יש מספרים דומים וקטנים של יצוא, אך לפי קנה המידה של ציר ה-X ניתן לראות שיש ערכים חריגים (נעסוק בזה בהמשך). לאחר פיצול לפי לסוג התווית, אפשר לראות את ההבדלים הגדולים בהתפלגות בין דוגמאות זדוניות לרגילות. ניתן ללמוד שככל שמספר הייצואים גדול יותר, נקבל יותר דוגמאות לא זדוניות. מהתבוננות בגרפים הללו, ניתן להניח שהפיצ'ר אינו מתפלג נורמלית.</p>	<p style="text-align: center;">exports graphs</p> <div style="display: flex; justify-content: space-around;">   </div>
<p>לפי הגרף ניתן לראות שלרוב הדגימות יש מספרים דומים וקטנים של סמלים, אך לפי קנה המידה של ציר ה-X ניתן לראות שישנם ערכי חריגים (נעסוק בזה בהמשך). לאחר פיצול לפי סוג התווית, ניתן לראות את ההבדלים הגדולים בהתפלגות בין דוגמאות זדוניות לרגילות. אנו יכולים ללמוד שככל שמספר הסמלים גדול יותר, כך נקבל יותר דוגמאות לא זדוניות. מהתבוננות בגרפים הללו, ניתן להניח שהפיצ'ר אינו מתפלג נורמלית.</p>	<p style="text-align: center;">symbols graphs</p> <div style="display: flex; justify-content: space-around;">   </div>

<p>לפי הגרף ניתן לראות שלרוב הדגימות יש ערכי מספרים דומים, אבל לפי קנה המידה של ציר ה-X ניתן לראות שיש ערכים חריגים (נעסוק בזה בהמשך). לאחר פיצול לפי סוג התוויות, אפשר לראות שההתפלגות בין דוגמאות זדוניות ללא זדוניות ממש דומה. מהתבוננות בגרפים הללו, ניתן להניח שהפיצ'ר אינו מתפלג נורמלית.</p>	
<p>לפי הגרף ניתן לראות שלרוב הדגימות אורך ממוצע דומה, אך לפי קנה המידה של ציר ה-X ניתן לראות שיש ערכים חריגים (נעסוק בזה בהמשך). לאחר פיצול לפי סוג התוויות, נוכל לראות שרוב הדגימות הן זדוניות. (לא ניתן לראות את הדגימות המגוונות בגרף כנראה בגלל שהן חריגות או שיש להן כמות קטנה מאוד). מהתבוננות בגרפים הללו, ניתן להניח שהפיצ'ר אינו מתפלג נורמלית.</p>	
<p>לפי הגרף אנחנו יכולים לראות שלרוב הדוגמאות יש ערכים דומים וקטנים של ערכים ניתנים להדפסה, אבל לפי קנה המידה של ציר ה-X ניתן לראות שיש ערכים חריגים (נעסוק בזה בהמשך). לאחר פיצול לפי סוג התוויות, נוכל לראות את ההבדלים בהתפלגות בין דוגמאות זדוניות ורגילות. ניתן ללמוד שככל שמספר פריטי ההדפסה גדול יותר, כך רואים יותר דוגמאות זדוניות. מהתבוננות בגרפים הללו, ניתן להניח שהפיצ'ר אינו מתפלג נורמלית.</p>	
<p>לפי הגרף ניתן לראות שלרוב הדגימות יש ערכים דומים וקטנים של נתיבים, אך לפי קנה המידה של ציר ה-X ניתן לראות שיש ערכים חריגים (נעסוק בזה בהמשך). לאחר פיצול לפי סוג התוויות, אנו יכולים לראות שאין הבדל גדול בהתפלגות בין דוגמאות זדוניות לרגילות. מהתבוננות בגרפים הללו, ניתן להניח שהפיצ'ר אינו מתפלג נורמלית.</p>	
<p>לפי הגרף ניתן לראות שלרוב הדגימות יש מספרים דומים וקטנים של כתובות אתרים, אך לפי קנה המידה של ציר ה-X ניתן לראות שיש ערכים חריגים (נעסוק בזה בהמשך). לאחר פיצול לפי סוג התוויות, ניתן לראות את ההבדלים הגדולים בהתפלגות בין דוגמאות זדוניות ללא זדוניות. ניתן ללמוד שככל שכתובת האתר גדולה יותר, כך יש יותר דוגמאות לא זדוניות. מהתבוננות בגרפים הללו, ניתן להניח שהפיצ'ר אינו מתפלג נורמלית.</p>	

<p>לפי הגרף ניתן לראות שפיצ'ר זה מקבל את הערך 0 ברוב הדוגמאות, אבל לפי קנה המידה של ציר ה-X ניתן לראות שיש ערכים חריגים (נעסוק בזה בהמשך). לאחר פיצול לפי סוג התוויות, ניתן לראות את ההבדלים הגדולים בהתפלגות בין דוגמאות זדוניות ללא זדוניות. ניתן ללמוד שככל שמספר הרישום גדול יותר, כך נקבל יותר דוגמאות לא זדוניות. מהתבוננות בגרפים הללו, ניתן להניח שהפיצ'ר אינו מתפלג נורמלית.</p>	<p>registry graphs</p> 
<p>לפי הגרף ניתן לראות שלרוב הדגימות יש ערכים דומים וקטנים של MZ, אבל לפי קנה המידה של ציר ה-X ניתן לראות שיש ערכים חריגים (נעסוק בזה בהמשך). לאחר פיצול לפי סוג התוויות, ניתן לראות שאין הבדל גדול בהתפלגות בין דוגמאות זדוניות לרגילות. מהתבוננות בגרפים הללו, ניתן להניח שהפיצ'ר אינו מתפלג נורמלית.</p>	<p>MZ graphs</p> 
<p>$file_type_prob_trid$ מקבל ערכים רבים בטווח שבין 0 ל-100 (כאמור אלו הסתברויות). הפיצ'ר אינו מתפלג נורמלית. לאחר פיצול לפי סוג התוויות, ניתן לראות בבירור את ההבדל בין הכמויות לכל סוג, מה שעשוי להצביע על חשיבותה של תכונה זו.</p>	<p>file_type_prob_trid graphs</p> 
<p>נראה שהפיצ'ר A מתפלג באופן נורמלי. מעבר לכך, אפשר לראות שהחלוקה לדוגמאות זדוניות ולא זדוניות היא ממש דומה, ובהמשך נדון אם להוריד פיצ'ר הזו, מכיוון שהוא כנראה לא יתרם רבות לניבוי.</p>	<p>A graphs</p> 
<p>התפלגות פיצ'ר B נראית נורמלית בערך, אבל לא נניח זאת. ניתן להבחין בהבדלים בין ההתפלגות של דוגמאות זדוניות לבין דוגמאות לא זדוניות.</p>	<p>B graphs</p> 

נספח 4 – מטריצת קורלציה של כלל המשתנים



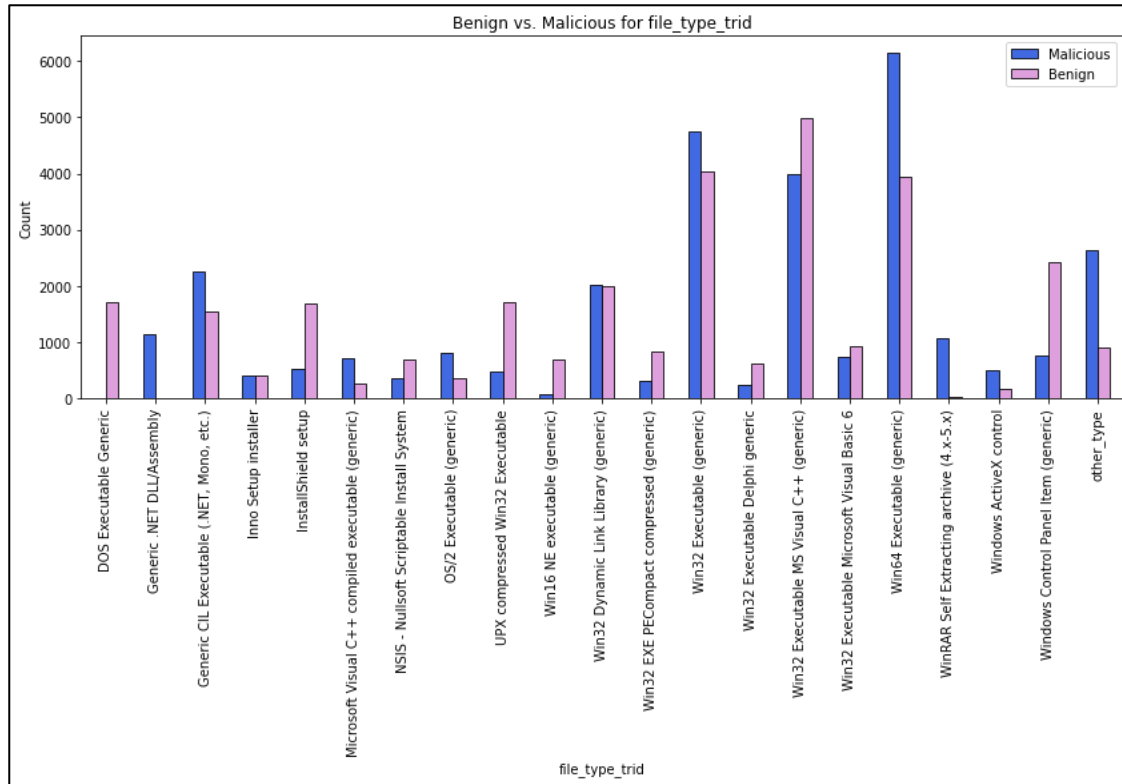
מטבלת המתאמים ניתן להסיק שרוב התכונות אינן תלויות זו בזו כי קיים מתאם נמוך מאוד בין רובן, עד כדי לא קיים. עם זאת, גודל הפיצ'ר נמצא בקשר חיובי חזק מאוד עם *numstrings* - דבר שנשמע אינטואיטיבי מכיוון שבכל שיש יותר מחרוזות, כך גודל הקובץ יהיה גדול יותר. כמו כן, גודל התכונה נמצא בקשר חיובי חזק עם *MZ* ו-*MZ* נמצא גם בקשר די חזק עם *numstring*. כאמור, שלוש התכונות הללו נמצאות בקורלציה זו עם זו ובהמשך נשתמש בטענה למטרות שונות.

נספח 5 - התפלגות הפיצ'רים החדשים; התפלגות כתלות בלייבל

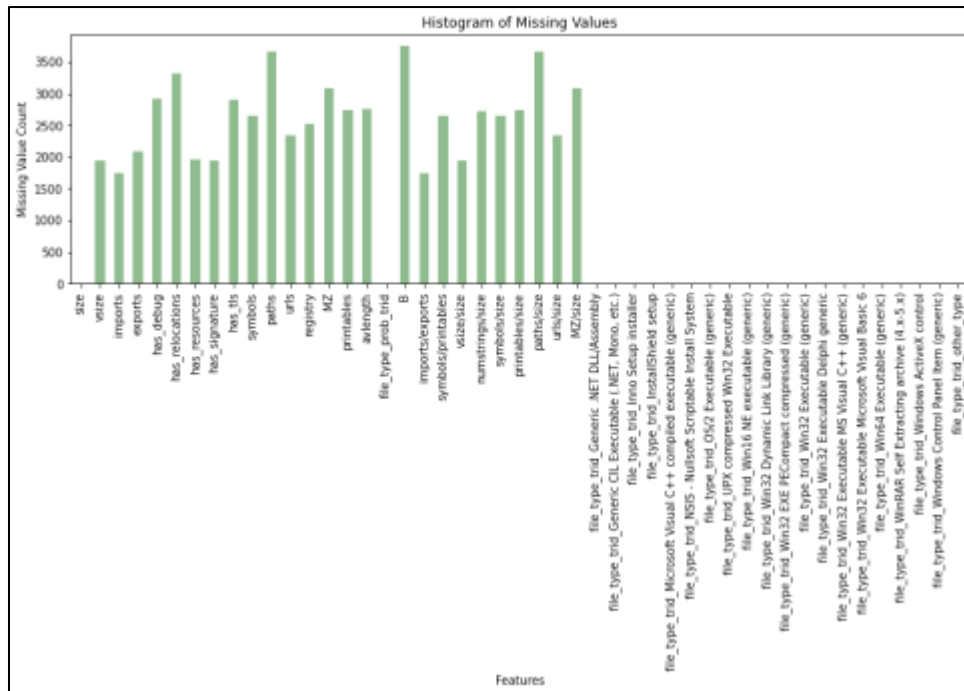
<p>יחס זה מלמד אותנו על הקשר בין כמות היבוא לכמות הייצוא שיש לקובץ. לפי הגרף ניתן לראות שלרוב הדגימות יש ערכים דומים של יבוא/יצוא, אך לפי קנה המידה של ציר ה-X ניתן לראות שיש ערכים חריגים (נעסוק בזה בהמשך). בנוסף, ניתן לראות שיש הבדל בהתפלגות הפיצ'ר בין דוגמאות זדוניות לדוגמאות לא זדוניות. ניתן ללמוד שככל שהערך של יבוא/יצוא גדול יותר, כך גדל הסיכוי שהדוגמאות לא זדוניות. מהתבוננות בגרפים הללו, ניתן להניח שהפיצ'ר אינו מתפלג נורמלית.</p>	
<p>יחס זה מלמד אותנו על הקשר בין כמות הסמלים למספר התווים הניתנים להדפסה. לפי הגרף ניתן לראות שלרוב הדגימות יש ערכים דומים, אך לפי קנה המידה של ציר ה-X ניתן לראות שיש ערכים חריגים (נעסוק בזה בהמשך). לאחר פיצול לפי סוג התוויות, ניתן לראות שאין הבדל כמעט בהתפלגות בין דוגמאות זדוניות לרגילות (נעסוק בזה בהמשך). מהתבוננות בגרפים הללו, ניתן להניח שהפיצ'ר אינו מתפלג נורמלית.</p>	
<p>יחס זה מלמד אותנו על הקשר בין גודל הקובץ לגודל תמונת הקובץ בעת טעינת הזיכרון. לפי הגרף ניתן לראות שלרוב הדגימות יש ערכים דומים, אך לפי קנה המידה של ציר ה-X ניתן לראות שיש ערכים חריגים (נעסוק בזה בהמשך). בנוסף, ניתן לראות שיש הבדל בהתפלגות הפיצ'ר בין דוגמאות זדוניות לדוגמאות לא זדוניות. רוב הדוגמאות הן לא זדוניות. מהתבוננות בגרפים הללו, ניתן להניח שהפיצ'ר אינו מתפלג נורמלית.</p>	
<p>יחס זה מלמד אותנו על הקשר בין מספר המחרוזות לגודל הקובץ. לפי הגרף ניתן לראות שלרוב הדגימות יש ערכים דומים, אך לפי קנה המידה של ציר ה-X ניתן לראות שיש ערכים חריגים (נעסוק בזה בהמשך). בנוסף, ניתן לראות שיש הבדל בהתפלגות הפיצ'ר בין דוגמאות זדוניות לדוגמאות לא זדוניות. מהתבוננות בגרפים הללו, ניתן להניח שהפיצ'ר אינו מתפלג נורמלית.</p>	
<p>יחס זה מלמד אותנו על הקשר בין מספר הסמלים לגודל הקובץ. לפי הגרף ניתן לראות שלרוב הדגימות יש ערכים דומים, אך לפי קנה המידה של ציר ה-X ניתן לראות שיש ערכים חריגים (נעסוק בזה בהמשך). לאחר פיצול לפי סוג התוויות, ניתן לראות שאין הבדל כמעט בהתפלגות בין דוגמאות זדוניות לרגילות (נעסוק בזה בהמשך). מהתבוננות בגרפים הללו, ניתן להניח שהפיצ'ר אינו מתפלג נורמלית.</p>	

<p>יחס זה מלמד אותנו על הקשר בין מספר התווים שניתן להדפיס לבין גודל הקובץ. לפי הגרף ניתן לראות שלרוב הדגימות יש ערכים דומים, אך לפי קנה המידה של ציר ה-X ניתן לראות שיש ערכים חריגים (נעסוק בזה בהמשך). בנוסף, ניתן לראות שיש הבדל בהתפלגות הפיצ'ר בין דוגמאות זדוניות לדוגמאות לא זדוניות. ככל שהערך של פיצ'ר זה קטן יותר, ההסתברות שהדוגמאות יהיו זדוניות גדולה יותר; עבור הערכים הגדולים הסיכוי שהדוגמאות לא זדוניות גדול יותר. מהתבוננות בגרפים הללו, ניתן להניח שהפיצ'ר אינו מתפלג נורמלית.</p>	<p>printables/size graphs</p> 
<p>יחס זה מלמד אותנו על הקשר בין מספר המחרוזות שמתחילות ב-C: שיכולות לציין את הנתיב לגודל הקובץ. לפי הגרף ניתן לראות שלרוב הדגימות יש ערכים דומים, אך לפי קנה המידה של ציר ה-X ניתן לראות שיש ערכים חריגים (נעסוק בזה בהמשך). בנוסף, ניתן לראות שיש הבדל בהתפלגות הפיצ'ר בין דוגמאות זדוניות לדוגמאות לא זדוניות. ככל שהערך של הפיצ'ר גדול יותר, כך נקבל דוגמאות לא זדוניות. מהתבוננות בגרפים הללו, ניתן להניח שהפיצ'ר אינו מתפלג נורמלית.</p>	<p>path/size graphs</p> 
<p>יחס זה מלמד אותנו על הקשר בין מספר המופעים של $http$ שיכולים לציין כתובת אתר לבין גודל הקובץ. לפי הגרף ניתן לראות שלרוב הדגימות יש ערכים דומים, אך לפי קנה המידה של ציר ה-X ניתן לראות שיש ערכים חריגים (נעסוק בזה בהמשך). בנוסף, ניתן לראות שיש הבדל בהתפלגות הפיצ'ר בין דוגמאות זדוניות לדוגמאות לא זדוניות. ככל שהערך של הפיצ'ר גדול יותר, כך נקבל יותר דוגמאות לא זדוניות. מהתבוננות בגרפים הללו, ניתן להניח שהפיצ'ר אינו מתפלג נורמלית.</p>	<p>url/size graphs</p> 
<p>יחס זה מלמד אותנו על הקשר בין מספר המופעים של המחרוזת הקצרה MZ לבין גודל הקובץ. לפי הגרף ניתן לראות שלרוב הדגימות יש ערכים דומים, אך לפי קנה המידה של ציר ה-X ניתן לראות שיש ערכים חריגים (נעסוק בזה בהמשך). בנוסף, ניתן לראות שיש הבדל בהתפלגות הפיצ'ר בין דוגמאות זדוניות לדוגמאות לא זדוניות. ככל שהערך של הפיצ'ר גדול יותר, כך נקבל יותר דוגמאות לא זדוניות. מהתבוננות בגרפים הללו, ניתן להניח שהפיצ'ר אינו מתפלג נורמלית.</p>	<p>MZ/size graphs</p> 

נספח 6 - משתנה *file_type_trid* אחרי איחוד קטגוריות נדירות

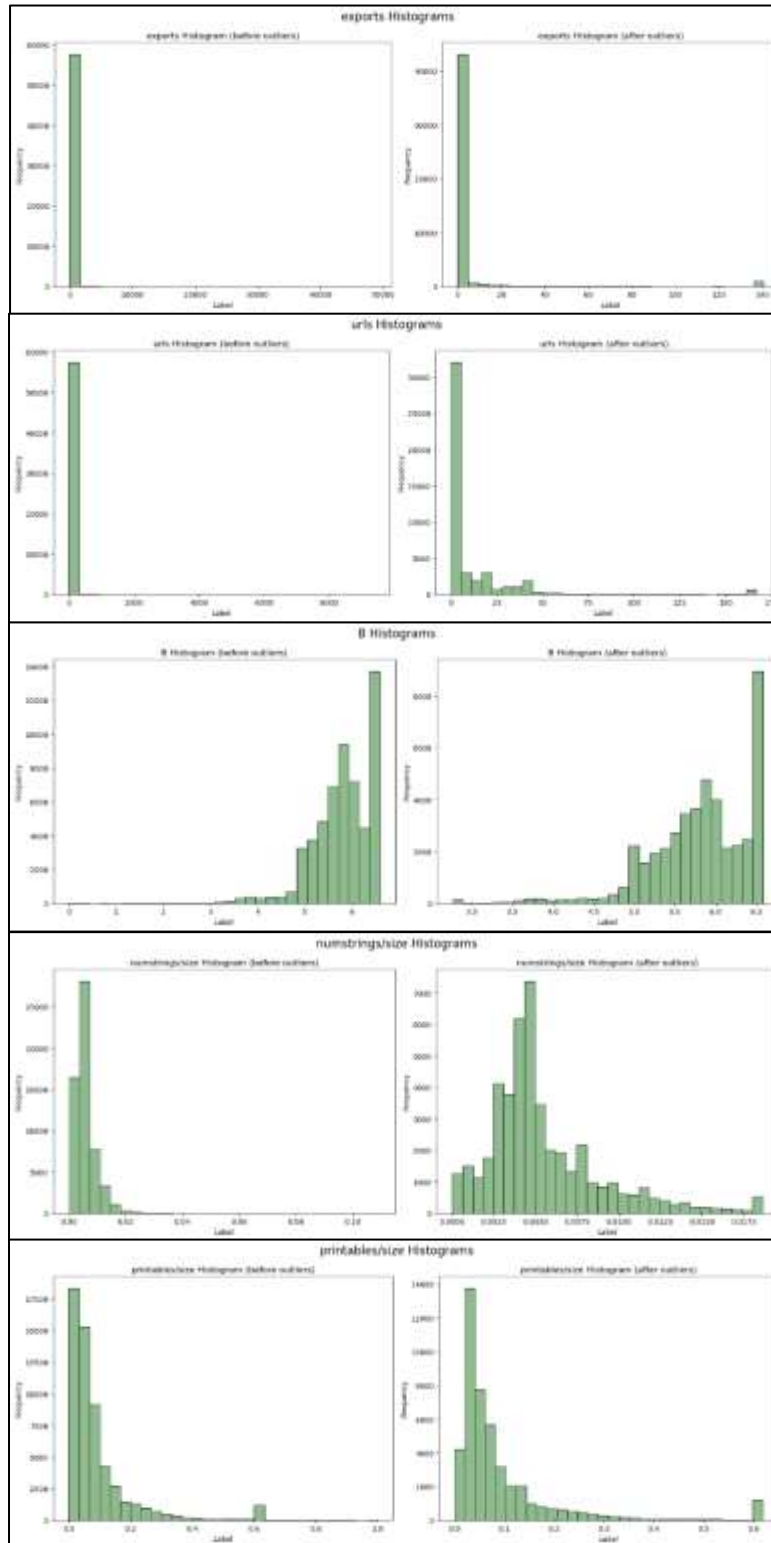


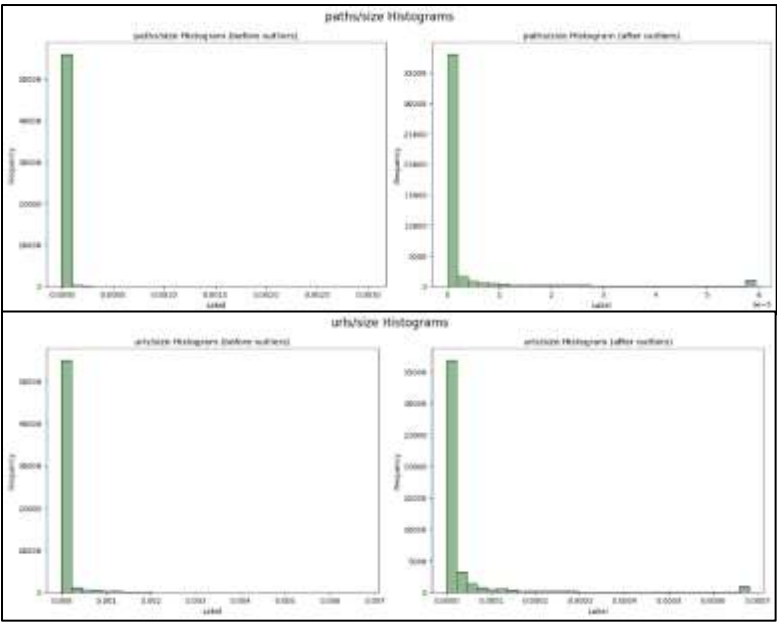
נספח 7 - ערכים חסרים לכל פיצ'ר



נספח 8 – צמצום ערכים קיצוניים

(משמאל מוצגת התפלגות הפיצ'ר לפני טיפול בערכים חסרים ומימין (אחריו)





נספח 9 – סיכום עיבוד מקדים

validation set for decision tree & random forest	train set for decision tree & random forest	validation set for KNN & logistic reg	train set for KNN & logistic reg	
ביצוע שלב זה על סט הנתונים לפני החלוקה לאימון וולידציה	ביצוע שלב זה על סט הנתונים לפני החלוקה לאימון וולידציה	ביצוע שלב זה על סט הנתונים לפני החלוקה לאימון וולידציה	ביצוע שלב זה על סט הנתונים לפני החלוקה לאימון וולידציה	בניית פיצ'רים חדשים
ביצוע שלב זה על סט הנתונים לפני החלוקה לאימון וולידציה	ביצוע שלב זה על סט הנתונים לפני החלוקה לאימון וולידציה	ביצוע שלב זה על סט הנתונים לפני החלוקה לאימון וולידציה	ביצוע שלב זה על סט הנתונים לפני החלוקה לאימון וולידציה	התמודדות עם משתנים קטגוריאליים
מחיקת הפיצ'רים C,A,numstrings	מחיקת הפיצ'רים C,A,numstrings לפני החלוקה	מחיקת הפיצ'רים C,A,numstrings לפני החלוקה	מחיקת הפיצ'רים C,A,numstrings לפני החלוקה	הורדת מימדיות הבעיה
ביצוע סטנדרטיזציה לסט זה על בסיס הממוצע והשונות של סט האימון	שמירת הממוצע והשונות של סט זה – בעזרתם עושים סטנדרטיזציה לסט הוולידציה ולאחר מכן לסט האימון	ביצוע סטנדרטיזציה לסט זה על בסיס הממוצע והשונות של סט האימון	שמירת הממוצע והשונות של סט זה – בעזרתם עושים סטנדרטיזציה לסט הוולידציה ולאחר מכן לסט האימון	ביצוע סטנדרטיזציה לנתונים
טיפול בערכים החריגים של כל פיצ'ר לפי ה-thresholds של סט האימון	שמירת ה-thresholds של כל פיצ'ר עבור סט האימון – בעזרתם מטפלים בערכים החריגים בסט הוולידציה ולאחר מכן בסט האימון	טיפול בערכים החריגים של כל פיצ'ר לפי ה-thresholds של סט האימון	שמירת ה-thresholds של כל פיצ'ר עבור סט האימון – בעזרתם מטפלים בערכים החריגים בסט הוולידציה ולאחר מכן בסט האימון	טיפול בערכים חריגים
טיפול בערכים החסרים של כל פיצ'ר לפי הממוצע/ערכים נפוצים של סט האימון	שמירת הממוצע/ערכים נפוצים של כל פיצ'ר עבור סט האימון – בעזרתם מטפלים בערכים החסרים בסט הוולידציה ולאחר מכן בסט האימון	טיפול בערכים החסרים של כל פיצ'ר לפי הממוצע/ערכים נפוצים של סט האימון	שמירת הממוצע/ערכים נפוצים של כל פיצ'ר עבור סט האימון – בעזרתם מטפלים בערכים החסרים בסט הוולידציה ולאחר מכן בסט האימון	טיפול בנתונים חסרים

*כמובן שביצענו עוד בדיקות שונות כמו – הרצת המודלים מבוססי העצים גם על הדאטה לאחר ביצוע PCA או ללא סטנדרטיזציה אך בסופו של דבר בחרנו את הדאטה שהוביל לתוצאות הטובות ביותר (כמתואר בטבלה לעיל)

* בכלל המודלים כשבדקנו היפר פרמטרים רצנו על טווח גדול כדי להתקבע על ההיפר-פרמטרים האידיאליים, בסוף צמצמנו את הטווחים לאלו שמתוארים בטבלאות על מנת לקצר את זמן הריצה.

נספח 10 – משמעות היפר-פרמטרים עבור KNN

KNN				
משמעות	ערכים נבדקים	ערך נבחר	השפעת הפרמטר על השונות וההטיה	
<i>n_neighbors</i>	קובע את מספר נקודות הנתונים השכנות לצורך הסיווג	[15, 16, ... 29] 26	מספר גבוה יותר של שכנים - גבול החלטה פחות מורכב, הטיה גבוהה יותר, שונות נמוכה יותר מספר נמוך יותר של שכנים - גבול החלטה מורכב יותר, הטיה נמוכה יותר, שונות גבוהה יותר	
<i>weights</i>	קובע את המשקל המוקצה לכל נקודת נתונים שכנה במהלך הסיווג	<i>uniform, distance</i>	<i>distance</i> <i>uniform</i> - מניח שלכל השכנים חשיבות שווה. זה יכול לגרום להטיה גבוהה יותר מכיוון שהמודל מתייחס לכל השכנים באופן שווה, ללא קשר לקרבתם לנקודה שברצונך לבדוק. משקלים אחידים - הטיה גבוהה יותר, פוטנציאל וריאציה נמוך יותר. <i>distance</i> – מניח שתרומתו של שכן עומדת ביחס הפוך למרחק שלו מהנקודה אותה רוצים לבדוק. לשכנים קרובים יותר יש השפעה גבוהה יותר, בעוד שלשכנים רחוקים יש השפעה נמוכה יותר. זה יכול להפחית את ההטיה, מכיוון שהמודל נותן משקל רב יותר לנקודות הקרובות לנקודת העניין משקלי מרחק - הטיה פוטנציאלית נמוכה יותר, אין השפעה ישירה על השונות	

נספח 11 – משמעות היפר-פרמטרים עבור רגרסיה לוגיסטית

<i>logistic reg</i>				
משמעות	ערכים נבדקים	ערך נבחר	השפעת הפרמטר על השונות וההטיה	
<i>penalty</i>	$l1, l2, elasticent$	$l2$	<p>שולט בטכניקת הרגולציה המשמשת למניעת התאמת יתר ואיזון פער ההטיה והשונות</p> <p>עונש L1 (Lasso) - מוסיף את הערכים האבסולוטיים של המקדמים לפונקציית ההפסד. זה יכול להפחית את מספר הפיצ'רים בשימוש במודל ולהגביר את ההטיה. גבוהה יותר, שונות נמוכה יותר.</p> <p>עונש L2 (ridge) - מוסיף את הגדלים בריבוע של המקדמים לפונקציית ההפסד. זה עוזר לשלוט בגודל המקדמים ולהפחית התאמת יתר. זה יכול לאזן את הפער בין הטיה לשונות על ידי ענישה של משקלים גדולים והקטנת השונות.</p> <p>עונש <i>elasticent</i> - משלב את L1 ו-L2, מוסיף לפונקציית ההפסד הן את הערכים האבסולוטיים והן את הערכים בריבוע של המקדמים.</p>	
C	$[10^{-10}, \dots, 1]$	0.1	<p>שולט בעוצמת העונש</p>	<p>C קטן יותר (ענישה חזקה יותר) - מגביר את עוצמת העונש, מה שמוביל להטיה גבוהה יותר ופוטנציאל לשונות נמוכה יותר.</p> <p>C גדול יותר (ענישה חלשה יותר) - מפחית את עוצמת העונש, זה עשוי להגביר את הסיכון להתאמת יתר ולשונות גבוהה יותר, אך ההטיה הפוטנציאלית נמוכה יותר.</p>
<i>solver</i>	$lbfgs, newton - cg, saga, sag$	$newton - cg$	<p>משמש לייעול פרמטרי המודל</p>	<p>לא משפיע ישירות על הפשרה בין שונות להטיה. בחירת הפותר יכולה להשפיע על תהליך הרגולריזציה, וכתוצאה מכך, על ההטיה-שונות.</p>

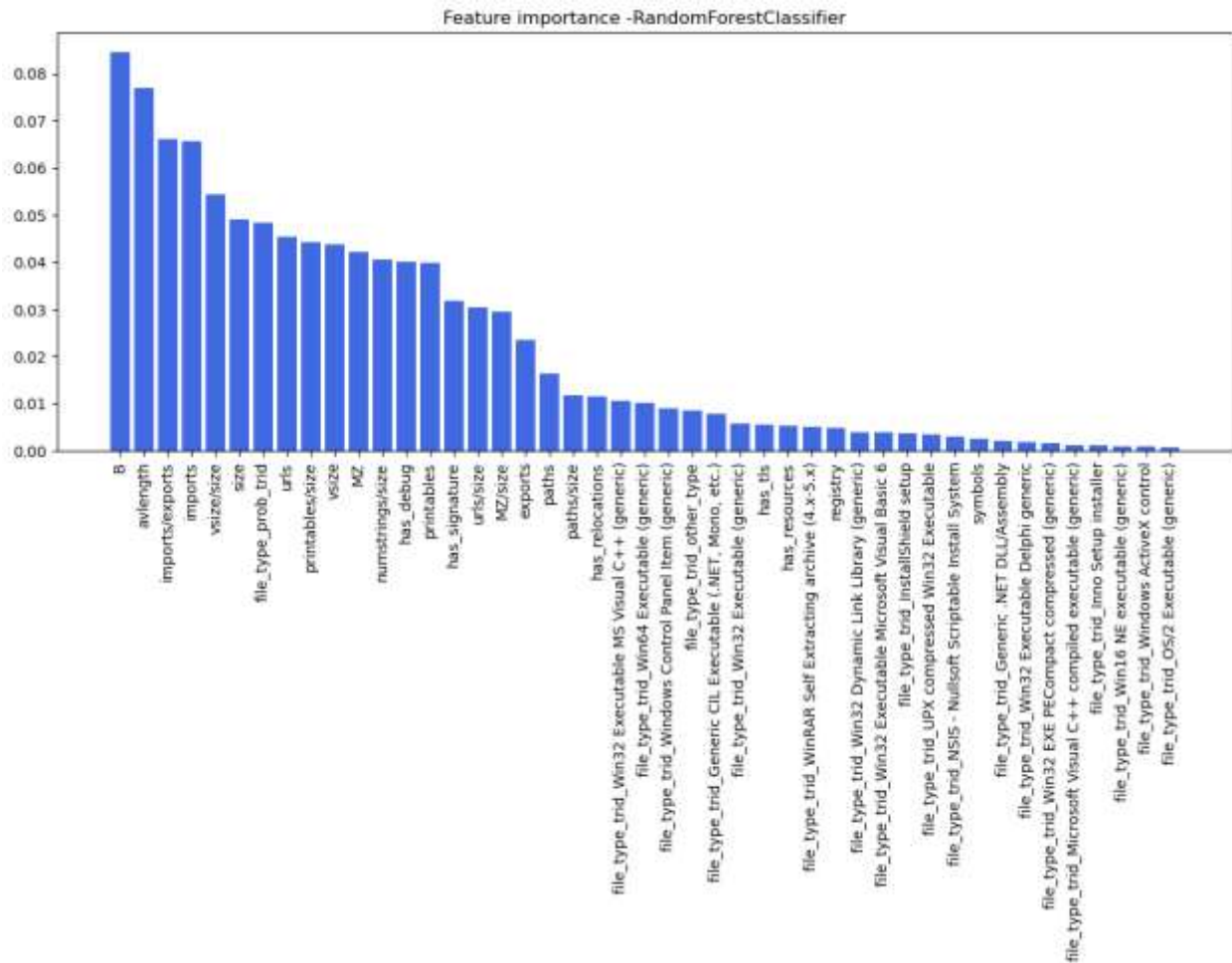
נספח 12 – משמעות היפר-פרמטרים עבור עץ החלטות

decision tree				
משמעות	ערכים נבדקים	ערך נבחר	השפעת הפרמטר על השונות וההטיה	
משמש להערכת איכות הפיצול במהלך בניית העץ	entropy, gini	entropy	<p><i>Gini</i> - מודד את ההסתברות לסיווג שגוי של אלמנט שנבחר באקראי מצומת נתון. המטרה שלו היא למזער סיווגים שגויים ולבנות עץ שמפריד בין הכיתות בצורה נקייה ככל האפשר. נוטה ליצור עצים מורכבים יותר עם פחות פיצולים אך עמוקים יותר, שיש להם שונות גבוהה יותר.</p> <p><i>entropy</i> - רוצה למקסם את המידע ולבנות עץ המספק את מירב המידע על הלייבל. נוטה לייצר עצים מאוזנים ורדודים יותר עם יותר פיצולים בכל רמה. עצים אלו יכולים להיות פחות מועדים להתאמות יתר ובעלי שונות נמוכה יותר בהשוואה לג'יני.</p>	criterion
שולט על העומק המרבי או המספר המרבי של רמות בעץ ההחלטות	[5, 10, 15, ..., 30]	20	<p>הגדלת <i>max_depth</i> מאפשרת לעץ לצמוח עמוק יותר, ובתוצאה מכך למודל מורכב יותר עם שונות גבוהה יותר. עץ עמוק יותר יכול ללכוד פרטים עדינים ואינטראקציות מורכבות בין תכונות, מה שעלול להוביל להתאמת יתר.</p> <p>הפחתת <i>max_depth</i> מגבילה את עומק העץ, ובתוצאה מכך מביאה למודל פשוט יותר עם שונות נמוכה יותר ופחות סיכוי להתאים יתר על המידה.</p>	max_depth
מגדיר את המספר המינימלי של דגימות הנדרש לפיצול צומת פנימי	[50, 55, ..., 100]	95	<p>הגדלת מספר הדגימות מובילה לפיצולים שמרניים יותר בעץ ההחלטות, ובתוצאה מכך למודל פשוט יותר עם שונות נמוכה יותר. עץ החלטות כזה הופך פחות מסוגל ללכוד דפוסים עדינים ורעש בנתוני האימון, מה שמוביל לשונות מופחתת.</p> <p>הפחתת הערך הזה מאפשרת יותר פיצולים ועץ מורכב יותר עם שונות גבוהה יותר. עץ כזה דורש פחות דגימות לצומת כדי שהוא יהיה כשיר לפיצול, דבר שעלול להוביל להתאמת יתר</p>	min_sample_split

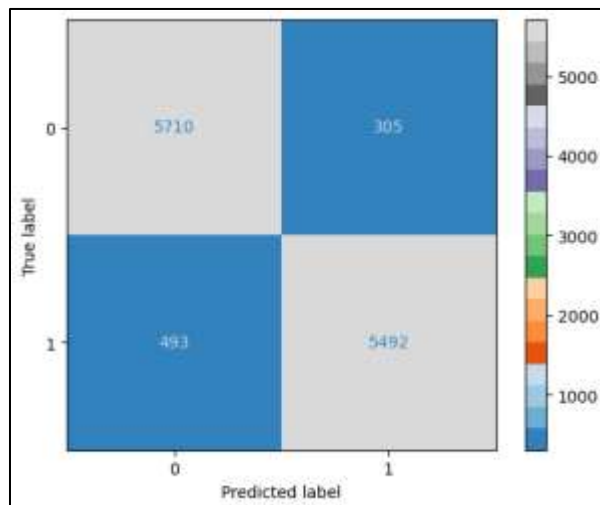
נספח 13 – משמעות היפר-פרמטרים עבור יער רנדומלי

random forest				
משמעות	ערכים נבדקים	ערך נבחר	השפעת הפרמטר על השונות וההטיה	
<i>n_estimators</i>	[100, 110, ..., 150]	140	הגדלת הערך של מספר העצים מובילה ליער מורכב יותר עם שונות גבוהה יותר. עם יותר עצי החלטה, יש פוטנציאל גדול יותר ללכידת דפוסים מורכבים בנתוני האימון. כתוצאה מכך, הגדלת כמות העצים יכולה להגביר את הסיכון להתאמת יתר. הפחתת הערך של מספר העצים מפחיתה את המורכבות של היער ומקטינה את השונות שלו. עם פחות עצי החלטה, היער מסתמך פחות על דפוסים מורכבים ומתמקד יותר בדפוסים כלליים בנתונים.	
<i>max_depth</i>	[30, 35, 40]	30	הגדלת <i>max_depth</i> מאפשרת לעץ לצמוח עמוק יותר, וכתוצאה מכך למודל מורכב יותר עם שונות גבוהה יותר. עץ עמוק יותר יכול ללכוד פרטים עדינים ואינטראקציות מורכבות בין תכונות, מה שעלול להוביל להתאמת יתר. הפחתת <i>max_depth</i> מגבילה את עומק העץ, וכתוצאה מכך מביאה למודל פשוט יותר עם שונות נמוכה יותר ופחות סיכוי להתאים יתר על המידה.	
<i>min_sample_split</i>	[2, 4, 6]	2	הגדלת מספר הדגימות מובילה לפיצולים שמרניים יותר בעץ ההחלטות, וכתוצאה מכך למודל פשוט יותר עם שונות נמוכה יותר. עץ החלטות כזה הופך פחות מסוגל ללכוד דפוסים עדינים ורעש בנתוני האימון, מה שמוביל לשונות מופחתת. הפחתת הערך הזה מאפשרת יותר פיצולים ועץ מורכב יותר עם שונות גבוהה יותר. עץ כזה דורש פחות דגימות לצומת כדי שהוא יהיה כשיר לפיצול, דבר שעלול להוביל להתאמת יתר	

Feature Importance - 14 נספח



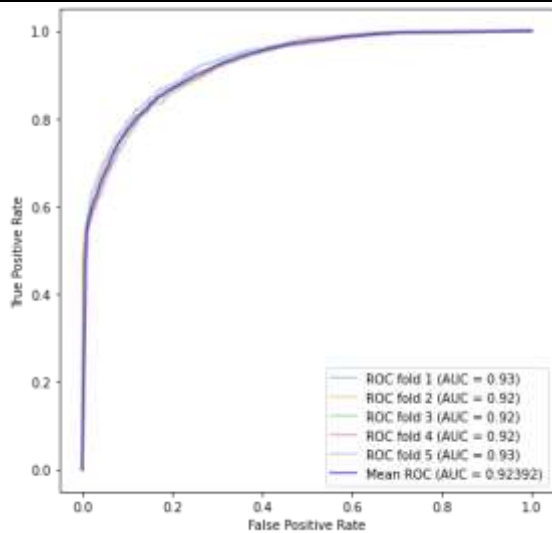
random forest confusion matrix – 15 נספח



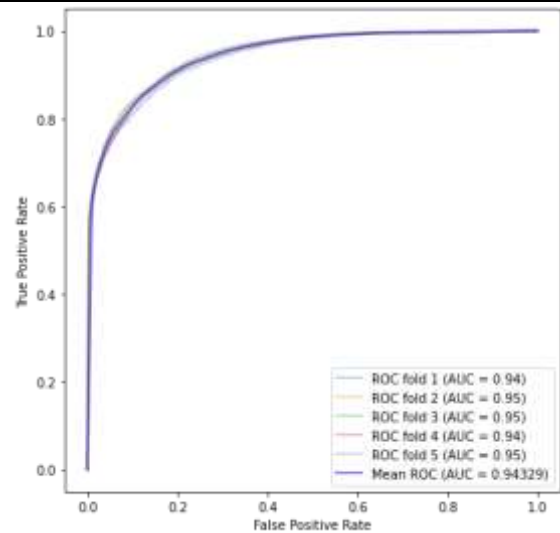
נספח 16 – ROC עבור כל $KFOLD$ עבור כל מודל

KNN

ולידציה



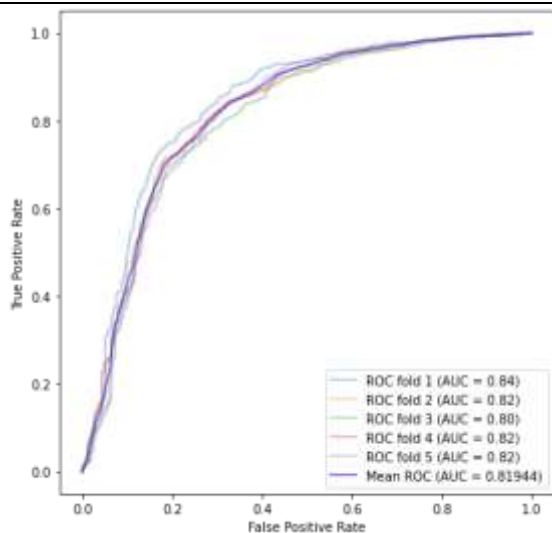
אימון



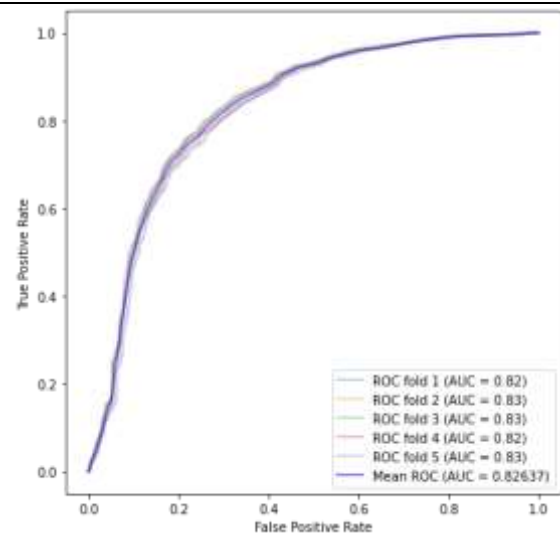
ניתן לראות כי ההפרש בין ה- roc – $mean$ של סט האימון ושל סט הולידציה גדול, דבר המעיד על $overfitting$ פוטנציאלי. חשוב לציין כי בחנו את המקרה ושינינו היפר-פרמטרים על מנת להתמודד עם בעיה זו, ולבסוף החלטנו לא להשתמש במודל זה על סט המבחן.

Logistic regression

ולידציה



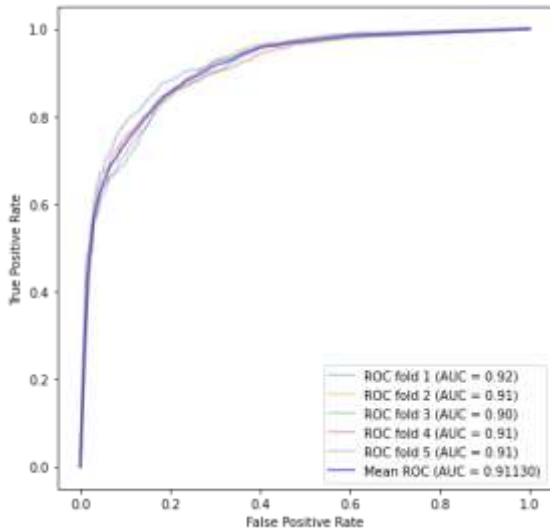
אימון



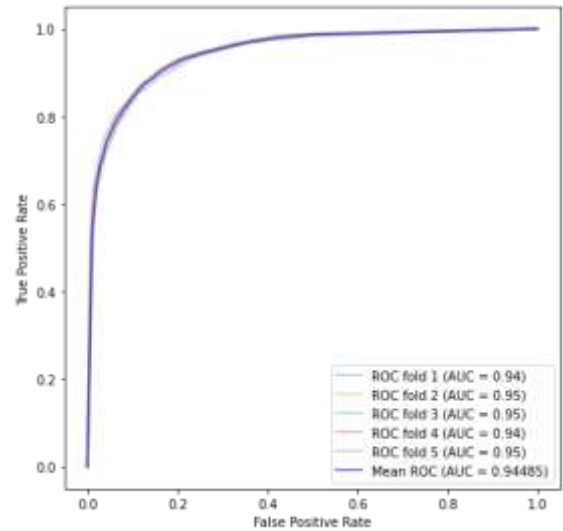
ניתן לראות כי ההפרש בין ה- roc – $mean$ של סט האימון ושל סט הולידציה קטן מאוד, דבר המעיד על אי קיום $overfitting$. למרות זאת, התוצאות של מודל זה נמוכות באופן יחסי לשאר המודלים ולכן בחרנו לא להשתמש במודל זה על סט המבחן.

Decision tree

ולידציה



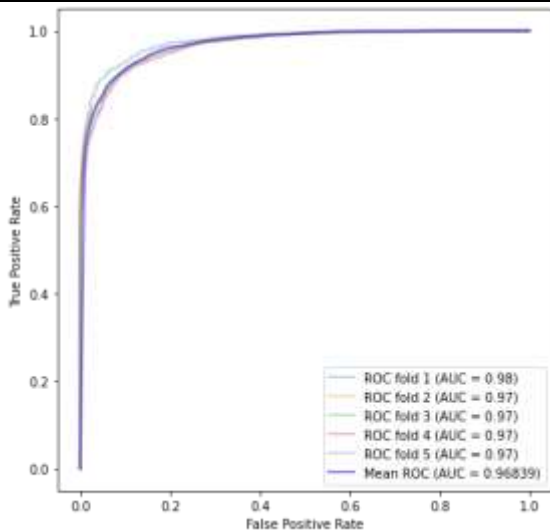
אימון



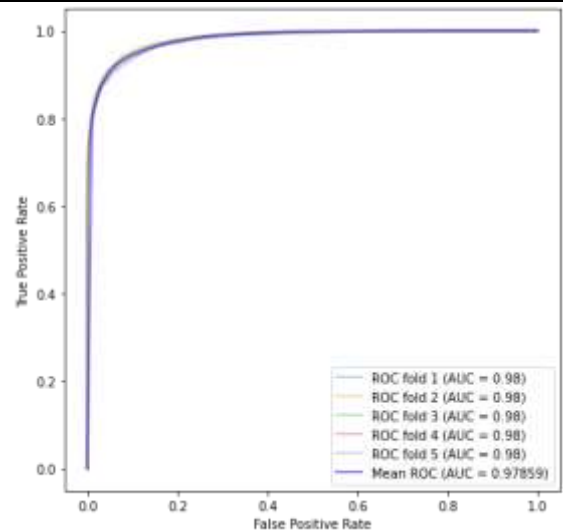
ניתן לראות כי ההפרש בין ה- $roc - mean$ של סט האימון ושל סט הולידציה די גדול, דבר היכול להעיד על $overfitting$ פוטנציאלי. חשוב לציין כי בחנו את המקרה ושינינו היפר-פרמטרים על מנת להתמודד עם בעיה זו, ולבסוף החלטנו לא להשתמש במודל זה על סט המבחן.

Random forest

ולידציה



אימון

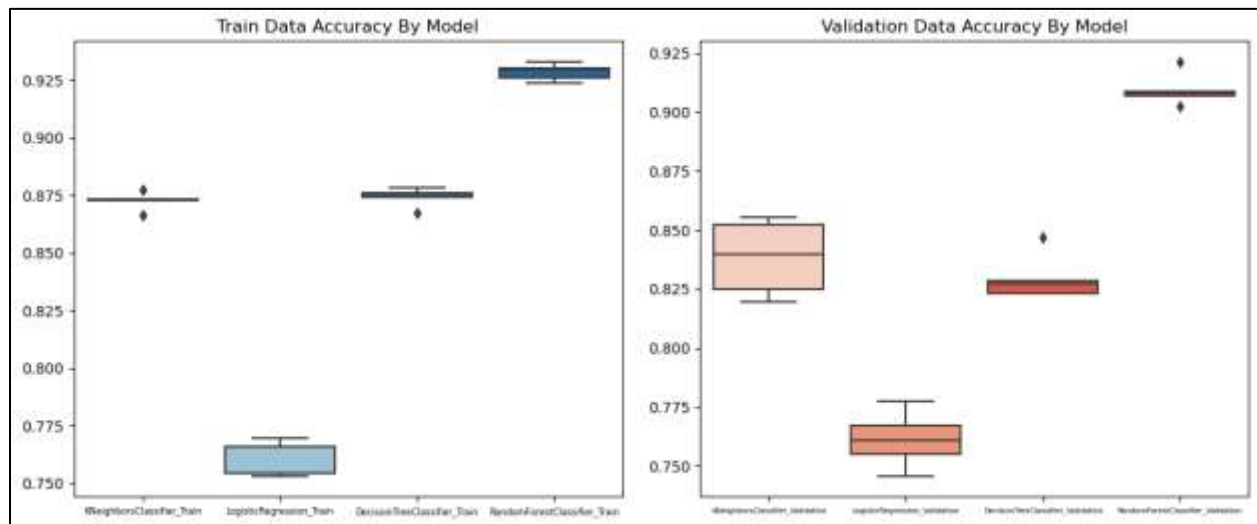


ניתן לראות כי ההפרש בין ה- $roc - mean$ של סט האימון ושל סט הולידציה קטן מאוד, דבר המעיד על אי קיום $overfitting$. בשל עובדה זו ולאור הביצועים הטובים של מודל זה, בחרנו בו בסופו של דבר למודל הסופי.

נספח 17 – השוואת מודלים

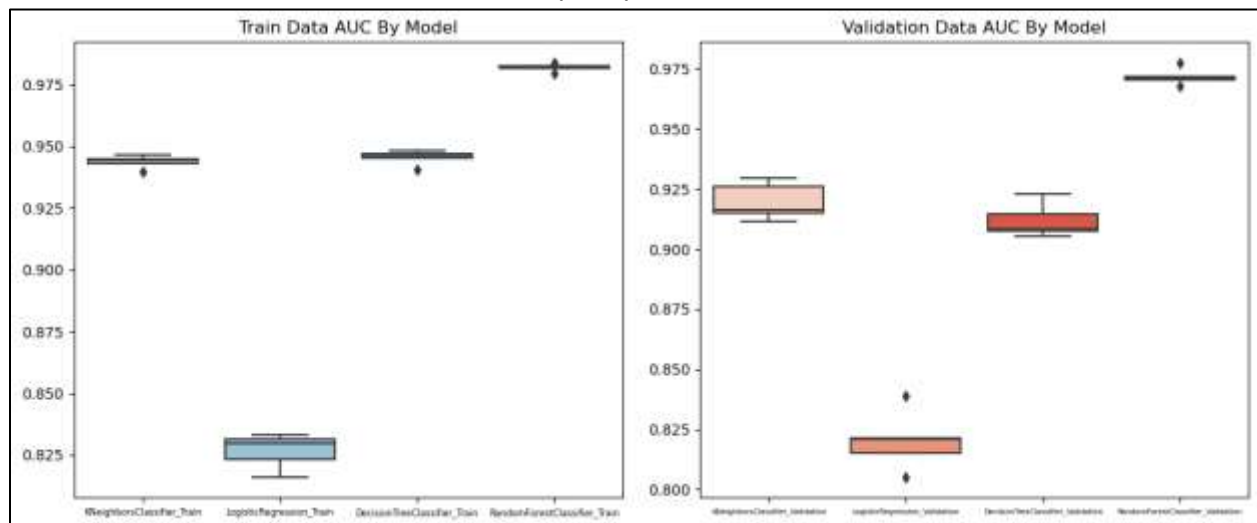
<i>random forest</i>	<i>decision tree</i>	<i>logistic reg</i>	<i>KNN</i>	
<i>n_estimators</i> = 140 <i>max_depth</i> = 30 <i>min_samples_split</i> = 2	<i>criterion</i> = <i>entropy</i> <i>max_depth</i> = 20 <i>min_samples_split</i> = 105	<i>c</i> = 0.1 <i>penalty</i> = <i>l2</i> <i>solver</i> = <i>newton - cg</i>	<i>n_neighbors</i> = 26 <i>weights</i> = <i>distance</i>	היפר פרמטרים נבחרים
0.928375	0.873479	0.76114	0.86268	<i>Accuracy</i> סט אימון
0.90883	0.82925	0.75850	0.83708	<i>Accuracy</i> סט ולידציה
0.019541	0.044229	0.00264	0.02560	פערי ביצועים <i>Accuracy</i>
0.982092	0.945555	0.82640	0.94497	<i>AUC</i> סט אימון
0.97150	0.91223	0.81957	0.92519	<i>AUC</i> סט ולידציה
0.01059	0.033318	0.00682	0.01977	פערי ביצועים <i>AUC</i>

נספח 18 – השוואת *Accuracy* בין אימון לולידציה של כל מודל

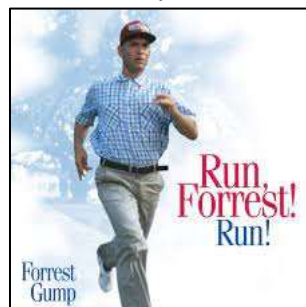


ניתן לראות כי מודל *random forest* – הוא בעל ערכי ה-*Accuracy* הגבוהים ביותר ועם הפרש קטן בין הביצועים על סט האימון לסט ולידציה

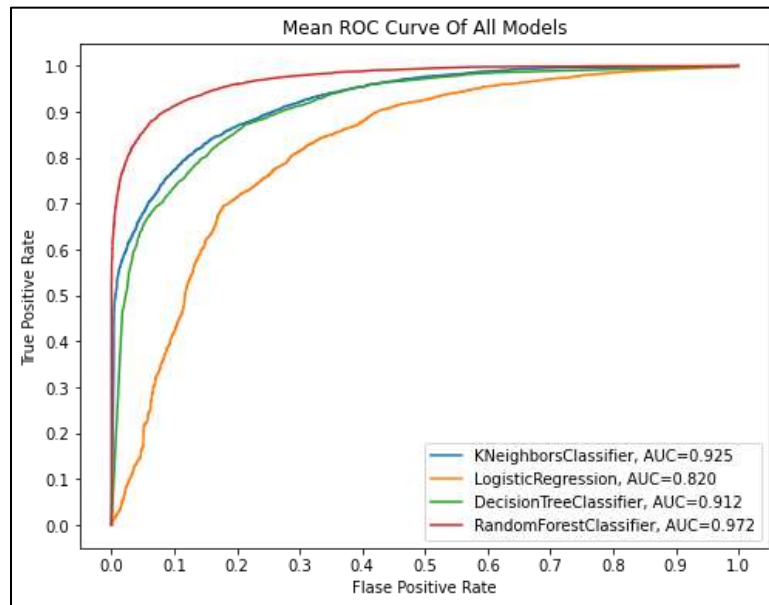
נספח 19 – השוואת *AUC* בין אימון לולידציה של כל מודל



ניתן לראות כי מודל *random forest* – הוא בעל ערכי ה-*AUC* הגבוהים ביותר ועם הפרש קטן בין הביצועים על סט האימון לסט ולידציה



נספח 20 – השוואת ROC ממוצע של כל מודל



ניתן לראות כי מודל *random forest* – הוא בעל ערך ה-*AUC* הממוצע הטוב ביותר ביחס לכל שאר המודלים

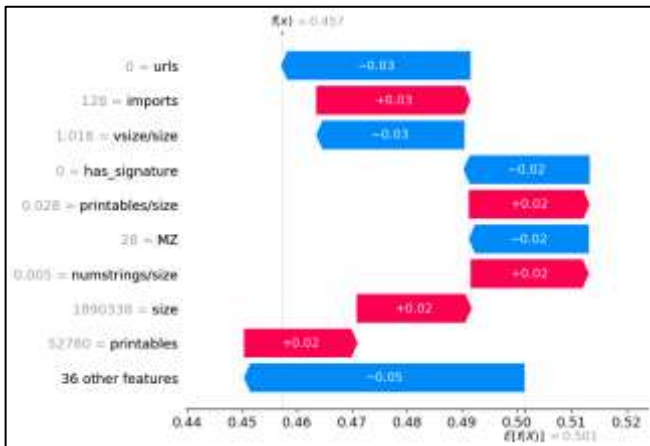
נספח 21 – כלים שלא למדנו – חבילת SHAP

ב-*waterfall – plot* ניתן לראות פירוט של האופן שבו כל פיצ'ר תורם לתחזית - בצד שמאל כל הפיצ'רים מדורגים בסדר יורד לפי מידת ההשפעה שיש להם על התחזית עבור הדוגמה הספציפית הזו. המספרים האפורים הם הערכים של הדוגמה הספציפית. בתחתית הגרף מופיע ערך הבסיס $E[f(x)]$ שהוא הערך הצפוי של הלייבל (=התוחלת של כל התחזיות). במקרה שלנו, $E[f(x)] = 0.501$, לכן, ניתן להסיק שהלייבלים של הטעויות מחולק באופן שווה לדדוני ולא דדוני. בתוך הגרף ערכי SHAP קובעים את התרומה (ערך ה-SHAP) והכיוון (משיכה ל-0 או 1) שבו כל פיצ'ר משפיע על התחזית.

- החיצים הכחולים "דוחפים" את תחזית הדוגמה לעבר חיזוי 0 – לא דדוני
 - החיצים האדומים "דוחפים" את תחזית הדוגמה לעבר חיזוי 1 - דדוני
- לבסוף $f(x)$ היא ההסתברות החזויה של המודל עבור דוגמה זו. ערך זה מחושב לפי ערך הבסיס בתוספת הסכום של כל ערכי ה-SHAP

גרף 1

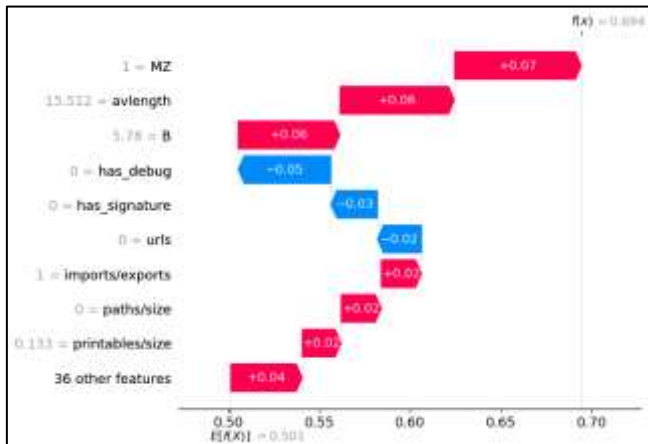
תזכורת - הגרף מנתח את הטעויות של המודל, כלומר, במקרה הזה החיזוי נטה ל-0- לא דדוני ($f(x) = 0.393$) והוא שגוי, התווית של הדוגמה היא למעשה 1 - דדוני. כך לדוגמה, הפיצ'ר *urls* הוא בעל ההשפעה הגדולה ביותר על החיזוי, אך למעשה זו השפעה "רעה" על המודל, יש לה תרומה רבה לטעות. לעומת זאת, לפיצ'ר *numstrings/size* הייתה השפעה גדולה ו"טובה" על התחזית (כמובן לא מספיק חזקה).



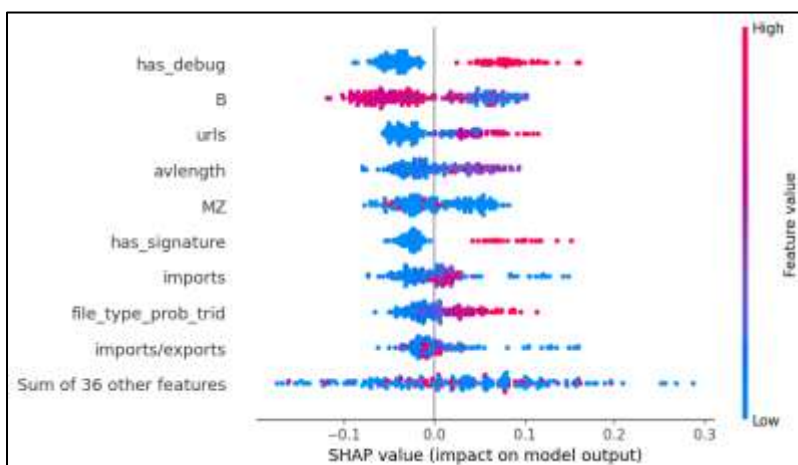
גרף 2

ננתח דוגמה אחרת ונראה את התרומה של כל פיצ'ר לתחזית שלה.

במקרה הזה החיזוי נטה ל-1 - דדוני ($f(x) = 0.743$) והוא שגוי. התווית של הדוגמה היא למעשה 0 – לא דדוני. כך למשל, לפיצ'רים *avlength*, *MZ* ו-*B* יש את ההשפעה הגדולה ביותר על החיזוי, אך למעשה זוהי השפעה "רעה" על המודל (הפיצ'רים מושכים לכיוון 1) ויש להם אחריות רבה בנוגע לטעות.



גרף 3



בצד שמאל כל הפיצ'רים מדורגים בסדר יורד לפי ערך ה- $SHAP$ (= הממוצעים שלהם עבור מערך הנתונים כולו) סרגל הצבע בצד ימין מציין את גודל הערכים הגולמיים של פיצ'ר עבור כל דוגמה בגרף. כאשר מסתכלים על התפלגות הצבע האופקית עבור כל פיצ'ר, ניתן ללמוד על הקשר בין הערך הגולמי של הפיצ'ר לבין ערך ה- $SHAP$ שלו.

לדוגמה, ניתן לראות שלערכים נמוכים יותר של $urls$ (בכחול) יש ערכי $SHAP$ שליליים (=דחיפה לכיוון 0) בעוד שלערכים גבוהים יותר של $urls$ (באדום) יש ערכי $SHAP$ חיוביים (=דחיפה לכיוון 1).

מעניין גם ללמוד מזה על פיצ'ר B שכן מדובר בפיצ'ר אנונימי וכפי שניתן לראות- יש לו חשיבות רבה. בנוסף, מהגרף הוא נראה שונה משאר התכונות - שכן לערך B גבוה יותר, יש השפעה שלילית וככל שערך B נמוך יותר, יש לו השפעה חיובית על חיזוי התוויית בניגוד לתכונות האחרות המתוארות בגרף.

בחינת אופן התפלגות ערכי ה- $SHAP$ מגלה כיצד פיצ'ר עשוי להשפיע על תחזיות המודל. לדוגמה, עבור $urls$ ניתן לראות אשכול צפוף של ערכים נמוכים יותר של כתובות אתרים (נקודות כחולות) עם ערכי $SHAP$ קטנים אך שליליים. ניתן לראות גם, מקרים של ערכים גבוהים יותר של כתובות אתרים (נקודות אדומות) מתרחבים יותר לכיוון ימין, כך שלערכים גבוהים יותר של $urls$ יש השפעה חיובית חזקה יותר על חיזוי התוויית מאשר ההשפעה השלילית של ערכים נמוכים יותר התחזית.