מיני פרויקט:
ניתוח נתוני עתק לשם אבטחת המרחב המקוון

202-1-4591

Time-Series Analysis
מנחים: פרופ' דני הנלדר, מר עמיר רובין

מגישים: יובל גבע, נוי משט, ערן סלומון
אוניברסיטת בן גוריון בנגב, סמסטר ב', תש"פ 2020, 19/07/2020

הקדמה

מערך הנתונים ש-Microsoft סיפקה עבור הפרוייקט בנושא time-series analysis כולל מידע על הורדות קבצים שנשלחו כ-webmail attachment ב-14 הימים הראשונים של שנת 2017.
המטרה היא לבדוק את דפוסי ההורדות של קבצים נקיים לעומת קבצים זדוניים על ידי סדרות זמן ומאפיינים נוספים שנמצאים בדאטה שקיבלנו.
חילקנו את מערך הנתונים ל-11 ימים ראשונים, קבוצת ה-train, מהם יצרנו מאגר של מאפיינים, ששימשו אותנו על מנת לאמן מודל של machine learning שיוכל לתייג האם קובץ הוא נקי או זדוני.
ב-3 הימים הנותרים, קבוצת ה-test, השתמשנו כדי לבדוק את המודל עם המאפיינים שיצרנו.

נגדיר מהו time series (סדרת זמן): זהו וקטור המייצג ערכים כתלות בזמן.
לכל קובץ ממערך הנתונים יצרנו סדרת זמן המייצגת את מספר ההורדות של הקובץ על מכונות שונות, כלומר אינן כוללות הורדות חוזרות של הקובץ על אותה המכונה בטווח יומי עבור קבוצת ה-train.

ניתוח ראשוני של ה-data

בחרנו לייצג בסדרות הזמן את המופע הראשון של הורדה לכל מכונה ייחודית בכל יום, משום שזוהי אינדיקציה טובה יותר על כמות האנשים שמשתמשים בקובץ ועל ההתפשטות שלו במרחב המקוון.
זאת אומרת שעבור הורדת קובץ על מכונה מסויימת, המכונה מוגדרת "ייחודית" עבור אותו היום אם זהו מופע ההורדה הראשון על המכונה ביום זה.

מידע כללי על הדאטה

	קבוצת ה-Train													קבוצת ה-Test			
	יום 1	יום 2	יום 3	יום 4	יום 5	יום 6	יום 7	יום 8	יום 9	יום 10	יום 11	יום 12	יום 13	יום 14	סה"כ
מספר רשומות	42,244	111,133	142,948	127,616	152,317	158,867	76,073	54,324	100,892	117,495	96,664	109,110	114,948	110,788	**1,515,419**
רשומות מתוייגות malicious	7,793	25,894	62,875	79,207	57,824	40,279	22,896	20,468	42,741	49,901	40,794	65,928	55,797	31,437	**603,834**
מספר הקבצים (עפ"י Sha1 ID)	22,461	22,443	53,463	32,027	59,836	78,249	33,894	21,209	37,722	37,843	23,264	21,534	31,336	42,605	**517,886**
מספר מכונות ייחודיות	20,611	20,595	37,144	23,418	45,324	55,529	31,186	23,371	36,138	49,800	44,102	33,172	46,276	61,242	**527,908**

(רשומה- הורדה).

תחילה, הגדרנו סף התייחסות לקבצים- אם הם בעלי יותר מ-5 הורדות על מכונות שונות.
זאת אומרת שקובץ מוגדר כנקי אם הוא לא תוייג כזדוני במהלך כל טווח הבדיקה, וגם יש לו יותר מ-5 הורדות על מכונות שונות. קובץ מוגדר כזדוני אם תוייג כזדוני באחד המופעים שלו, ויש לו יותר מ-5 הורדות על מכונות שונות.
סף זה נקבע כדי לסנן קבצים שירדו באופן חד פעמי על מספר קטן מאד של מכונות שונות, כך שלא ניתן ללמוד על דפוס הורדתם, והם עלולים להטות את הנתונים ולשבש את המסקנות.
לאחר שקבענו את הסף לעיל, זהו הדאטה שנותר לנו אשר עונה על ההגדרה:

השוואה בין Train ל-Test

	Train	Test
ימים	11	3
מספר קבצים זדוניים	264	75
מספר קבצים נקיים	2256	1061
סה"כ	2520	1136

התחלנו עם חלוקה של טווחי הזמן לפי שעות ולפי ימים. מכיוון ששעות ביום משתנות בין אזורי זמן לא ניתן להסיק מסקנות לגבי הורדות בשעות מסוימות של היום, אלא רק לגבי ההפרשים בקצב ההורדות בטווחי זמן של שעה. בהמשך החלטנו להמשיך עם סדרות זמן בהפרשים של ימים בלבד (מסיבות טכניות של זמני הרצת הקוד).

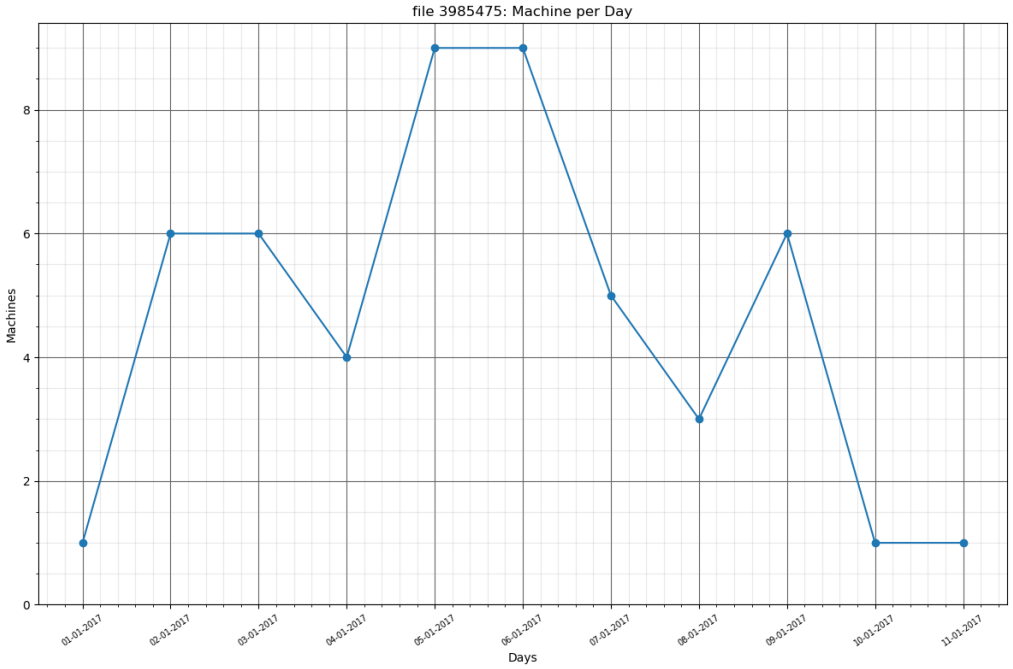
בנוסף, נגדיר שני מאפיינים הנוגעים לקפיצות במספר ההורדות-
המאפיין **Peaks** מציין את מס׳ הימים בסדרת הזמן בעלי ״פיק״-עלייה מתונה במספר ההורדות ביחס לכמות ההורדות בסביבת אותו היום.
המאפיין **Sharp Peaks** מציין את מס׳ הימים בסדרת הזמן בעלי ״פיק״ חד-עלייה חדה במספר ההורדות ביחס לכמות ההורדות בסביבת אותו היום.

סדרות זמן

גרפים מייצגים של סדרות זמן עבור קבצים נקיים:

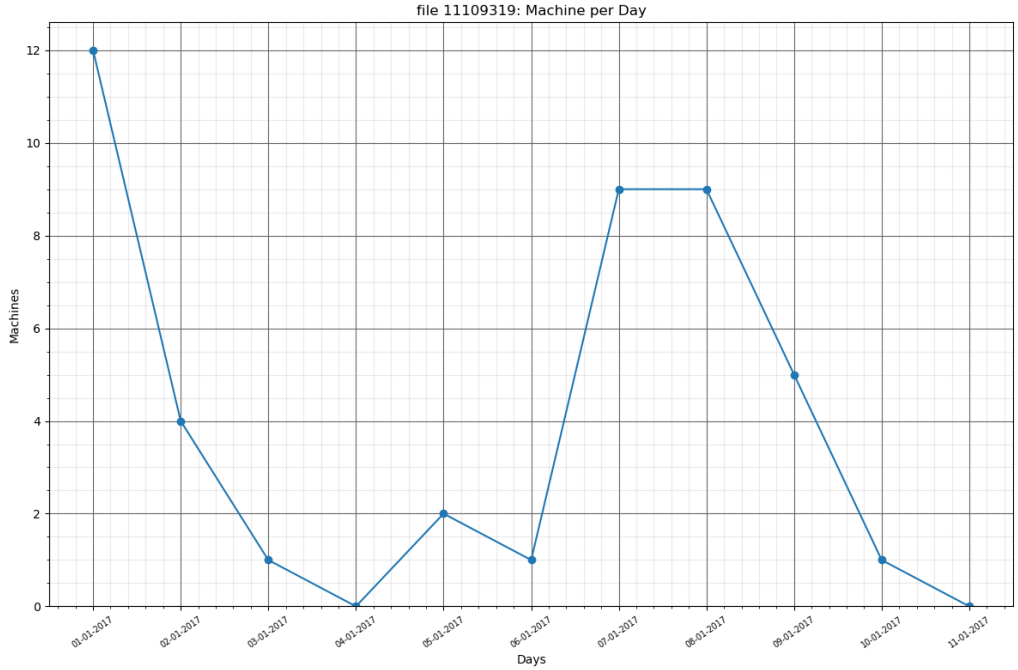
עבור סדרות זמן של קבצים נקיים אנו מצפים שכמות ההורדות לא תשתנה בצורה דרסטית מיום ליום, ושדפוס ההורדות יישאר יציב יחסית.

בפועל, קיבלנו תבנית מייצגת של כמות הורדות קבועה, עם לכל היותר קפיצה אחת מתונה בהורדה וללא קפיצות קיצוניות.

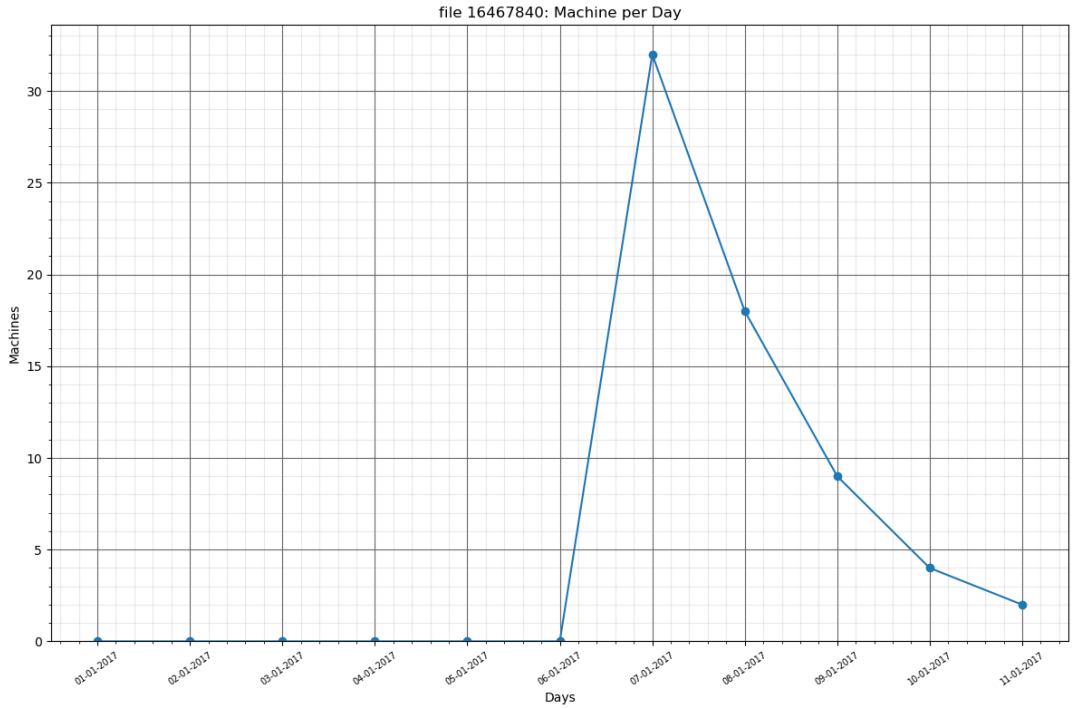


בדוגמה המתוארת לעיל ניתן לראות שכמות ההורדות נשארת יחסית קבועה, עם שינויים קלים, במהלך הימים.

<u>גרפים מייצגים של סדרות זמן עבור קבצים זדוניים:</u>
עבור סדרות זמן של קבצים זדוניים נצפה ל"פיק" חד בכמות ההורדות, ולאחריו כמה ימים עם מספר מועט של הורדות.
בפועל, קיבלנו קבצים שבחלקם היו כמות הורדות מתונה ובחלקם היו קפיצות חדות.



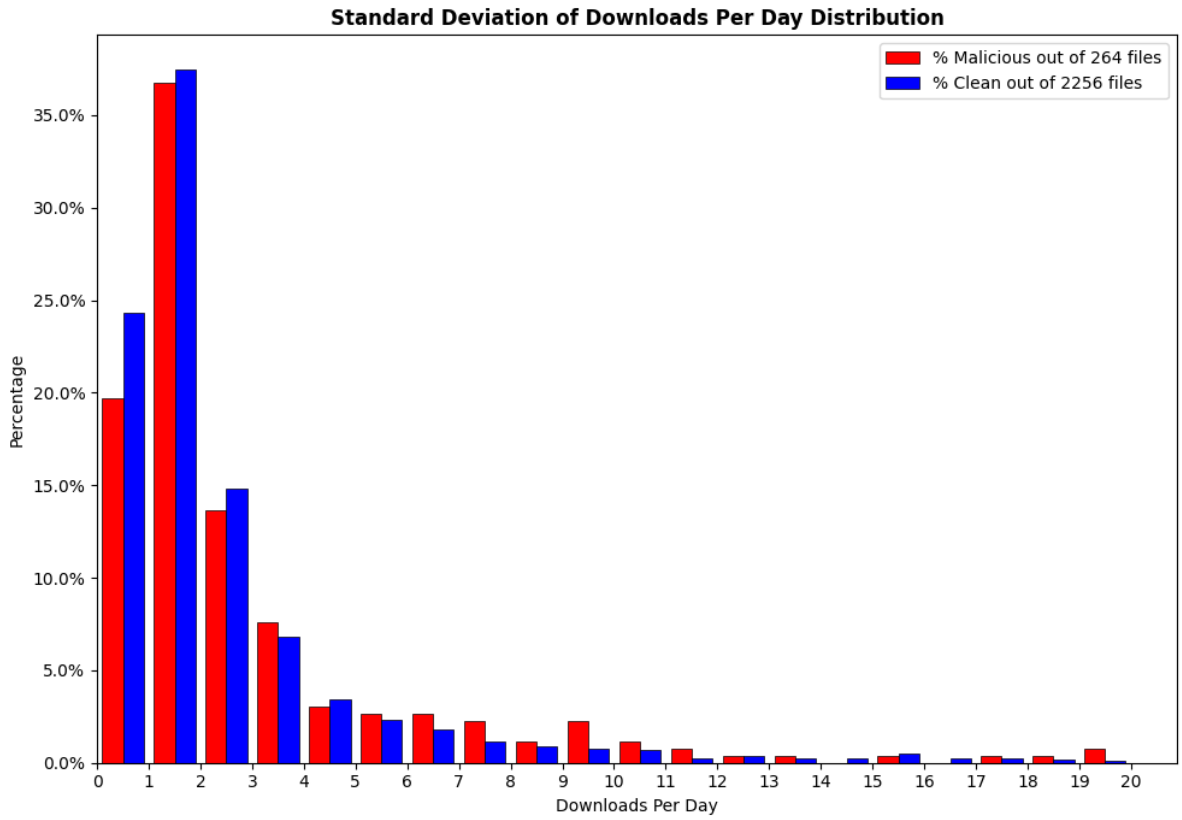
בדוגמה זו ניתן לראות שכמות ההורדות לא משתנה בצורה משמעותית במהלך הימים.



לעומת זאת, עבור קובץ זה רואים כי יש קפיצה חדה בכמות ההורדות בין ה-6-7/1/2017, ולאחריה ירידה מתונה.
הדפוס הראשון היה נפוץ יותר במרבית הקבצים.
בניגוד לציפייה לראות בקבצים הזדוניים דפוס שונה וקיצוני יותר, קיבלנו כי גם עבור הזדוניים וגם עבור הנקיים רוב תבניות ההורדה דומות ומתונות.

השערה אחת שלנו היא שטווח הבדיקה שלנו קטן מדי בשביל לזהות דפוס של קפיצות גבוהות בהורדות לאורך זמן, השערה נוספת היא שמספר הקבצים הזדוניים קטן ועבור דאטה גדול יותר עם מספר קבצים זדוניים רב יותר ניתן לקבל אינפורמציה יותר כוללת.

<u>סטיית תקן:</u>
שיערנו שעבור הקבצים הזדוניים סטיית התקן של כמות ההורדות היומית תהיה גדולה יותר מזו של הקבצים הנקיים, עקב האופי של דפוסי ההורדות שציפינו שיאפיין את הקבצים הזדוניים.



בניגוד לצפוי, התפלגות סטיות התקן גם עבור הקבצים הנקיים וגם עבור הקבצים הזדוניים היא דומה, כך שלא ניתן להסיק מסטיית התקן דפוס התנהגות רלוונטי.

חישוב מרחקים בין סדרות זמן
על מנת להשוות בין סדרות הזמן יש לחשב את המרחק בין כל שתי סדרות זמן. ככל שהמרחק שהתקבל קטן יותר, ניתן לומר שהסדרות דומות יותר. מידע זה ישמש אותנו להשוואה בין מרחקים של קבצים משני הסוגים.
מרחק זה ניתן לחישוב בדרכים שונות, אנחנו בחרנו בשתי שיטות:

מרחק אוקלידי:
בהינתן שתי סדרות זמן $(t_1,..t_n),(s_1,..,s_n)$ נקבל את המרחק האוקלידי ביניהן ע״י החישוב הבא: $\sqrt{(t_1-s_1)^2+\cdots+(t_n-s_n)^2}$.
בנוסף, ביצענו חישוב מקדים של שינוי מרכז המסה על מנת שכל סדרה תתחיל מהמספר הראשון שאינו 0, ע״י ביצוע circular shift- העברת רישא של 0 לסוף הסדרה. חישוב זה נועד לכך שניתן יהיה לזהות דפוסים של עלייה/ירידה בקצב מספר ההורדות גם בזמנים שונים.

אלגוריתם Dynamic Time Wrapping (DTW):
אלגוריתם תכנון דינאמי אשר בהינתן שתי סדרות מחפש את ההתאמה הטובה ביותר ביניהן ע״י הנוסחה הבאה:

$$DTW(m,n)=\begin{cases}0, & m=0 \text{ or } n=0\\ d(m,n)+\min\begin{pmatrix}DTW(m-1,n)\\ DTW(m,n-1)\\ DTW(m-1,n-1)\end{pmatrix}, & otherwise\end{cases}$$

בשיטה זו לא היה צורך בחישוב מקדים מכיוון שהאלגוריתם מבצע התאמה בין שתי סדרות זמן גם אם הן במהירויות שונות, כלומר הוא מתאים בין הדפוסים בקצב ההורדות ואז משווה אותן.

לכל קובץ ממערך הנתונים חישבנו את המרחק בין סדרת הזמן שלו לבין כל שאר סדרות הזמן של הקבצים האחרים ע״י שתי השיטות.

זמני ריצה
זמן ריצה מעשי:
מרחק אוקלידי- הריצה החלה ב- 14:44:23 והסתיימה ב- 15:11:19, זמן הריצה של החישוב בשיטה זו הוא בערך חצי שעה.
DTW- הריצה החלה ב-15:11:19 והסתיימה ב-16:09:02, זמן הריצה של החישוב בשיטה זו הוא בערך שעה.
זמן הריצה המעשי של DTW הינו כמעט פי 2 מזמן הריצה של החישוב בשיטה האוקלידית.

ניתוח זמן ריצה תיאורטי:
נסמן ב-n את מספר הקבצים עליהם מחשבים את המרחקים. נסמן ב-m את אורך סדרת הזמן.
עבור שתי השיטות, נעבור על כל רשימת הקבצים ונחשב את המרחקים מכל קובץ לכל הקבצים האחרים, זהו חישוב שמתבצע ב-$O(n^2)$.

מרחק אוקלידי- חישוב מקדים: שינוי center of mass עובר על n הקבצים ומבצע על כל אחד חישוב ב-$O(m)$, מכאן שזמן הריצה שלו הינו $O(mn)$.
בתוך הלולאה המקוננת חישוב כל מרחק בין זוג סדרות מתבצע ע״י סדרה של פעולות בזמן קבוע, כלומר $O(2m*1)=O(m)$.
בסך הכל, חישוב המרחקים האוקלידיים בין כל קובץ לכל שאר הקבצים מתבצע ב- $O(mn+n^2m)=O(n^2m)$.
DTW- עבור שתי סדרות באורך m סיבוכיות חישוב המרחק ביניהן לפי האלגוריתם הוא $O(m^2)$.
לכן בסך הכל חישוב המרחקים על פי אלגוריתם זה מתבצע ב-$O(n^2m^2)$.

השוואה בין השיטות לחישוב מרחק

שיטת חישוב המרחקים	ממוצע	חציון	סטיית תקן	מרחק מקסימלי	מרחק מינימלי
Euclidian	28.3	9.4	128.2	2622.8	0
DTW	17126	57	304573	11600390	0

על פי ערכי הממוצע והמקסימום המוצגים בטבלה ניתן לראות כי המרחקים שהאלגוריתם DTW מחשב הם בסדר גודל ובטווח הרבה יותר רחב מאשר המרחקים שמחושבים בשיטה האוקלידית.
עוד ניתן לראות כי סטיית התקן, הממוצע והחציון של המרחקים ב-DTW גדולים בהרבה מאשר של המרחקים בשיטה האוקלידית, ומכאן ניתן להבין שהתפלגות המרחקים ב-DTW גם רחבה הרבה יותר מאשר בחישוב האוקלידי.

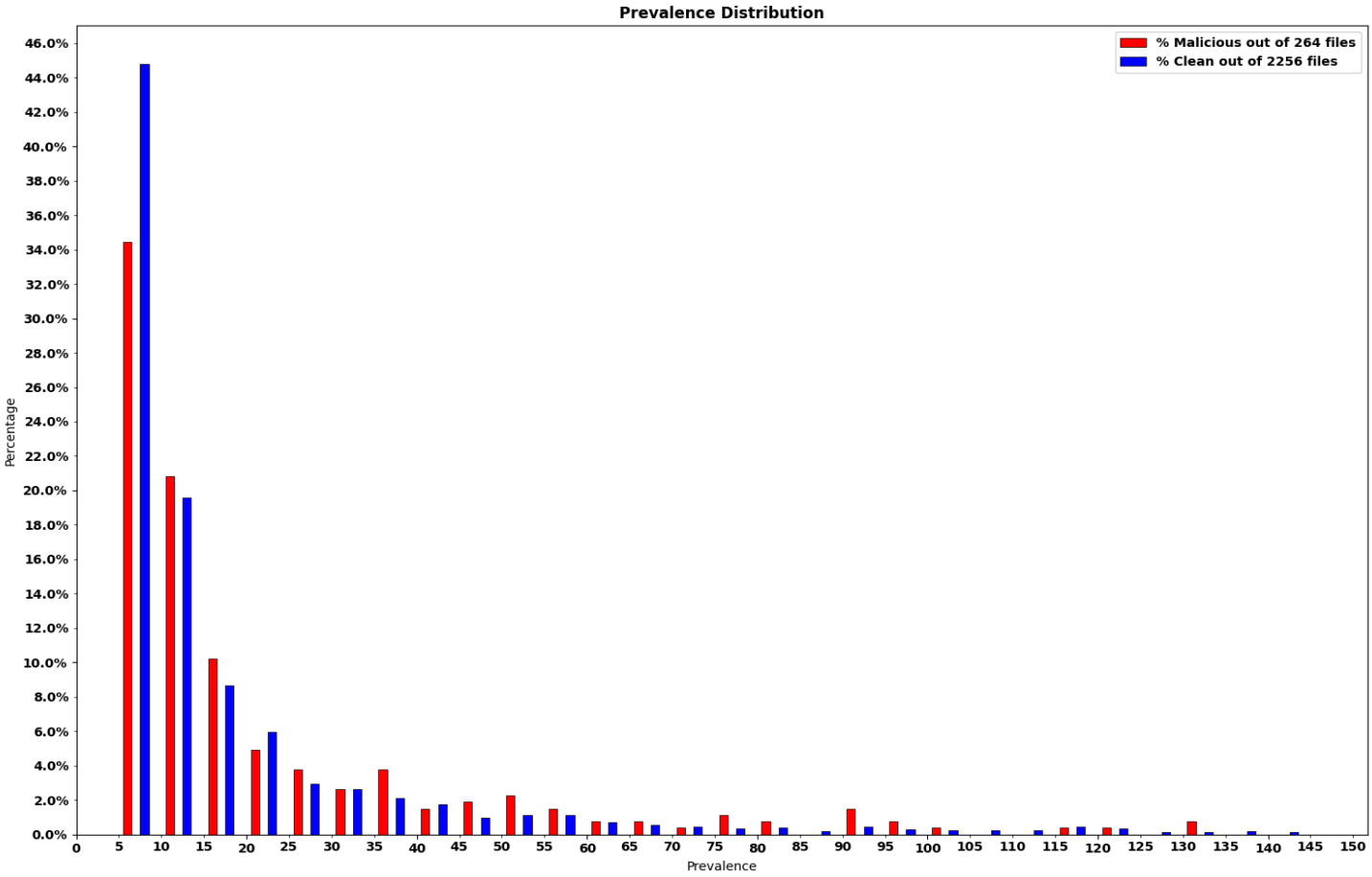
מעצם היותו אלגוריתם דינאמי, DTW מתאים בין 2 סדרות זמן בצורה יותר קפדנית, והשיטה רגישה יותר לשינויים. זו הסיבה שטווח והתפלגות המרחקים רחבה יותר, ושתוצאות האלגוריתם אמורות להיות מדויקות יותר מאשר התוצאות בשיטה האוקלידית.
למרות זאת, גם בשיטה האוקלידית עם שינוי מרכז המסה קיבלנו תוצאות שניתן להפיק מהן מידע שימושי ולכן החלטנו להמשיך עם שתי השיטות.

ניתוח המאפיינים (Features)

לכל קובץ בחרנו מספר מאפיינים, על פיהם נאמן את המודל להבחין בין קבצים נקיים לבין קבצים זדוניים. לכל קובץ הוצאנו מאפיינים משלושה סוגים:

1. מאפיינים גלובליים-

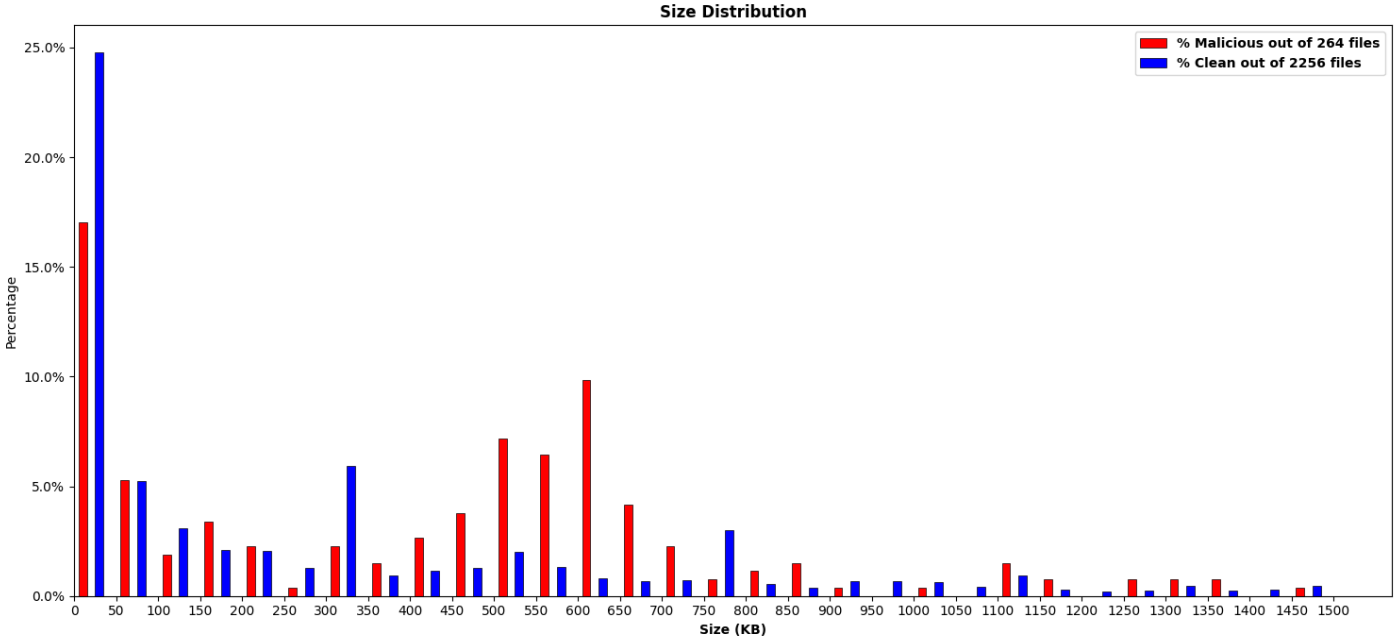
מאפיין **Prevalence** -תדירות הורדת הקובץ, סך ההורדות על מכונות שונות בכל יום.



ניתן לראות שמבחינת **שכיחות הקובץ**, קשה יותר לעשות הפרדה בין התפלגות קבצים נקיים לזדוניים, כאשר עבור שני סוגי הקבצים השכיחות שקטנה מ-5 שווה ל-0 בגלל הסינון הראשוני של הורדות מעל ל-5 מכונות ייחודיות.

הערה- בגרפי ההיסטוגרמה כל bin (טווח מספרים על ציר ה-x) כולל את הערך השמאלי ואינו כולל את הערך הימני.

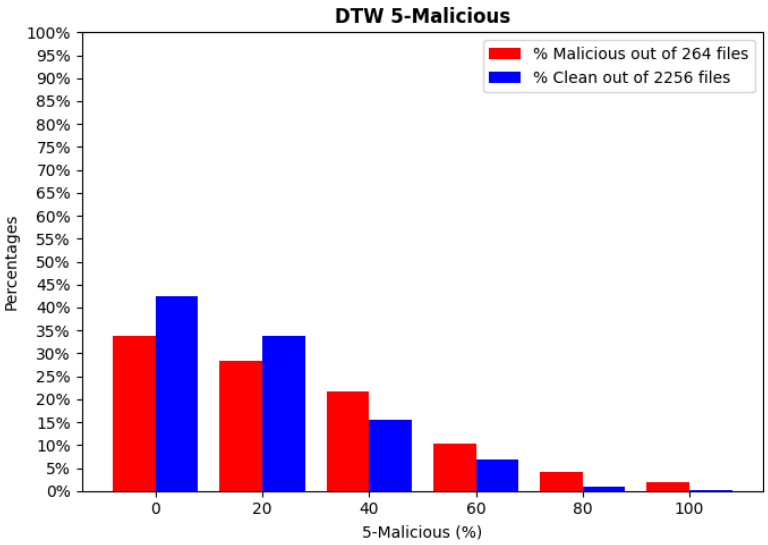
מאפיין **Size**- גודל הקובץ ב-KB.

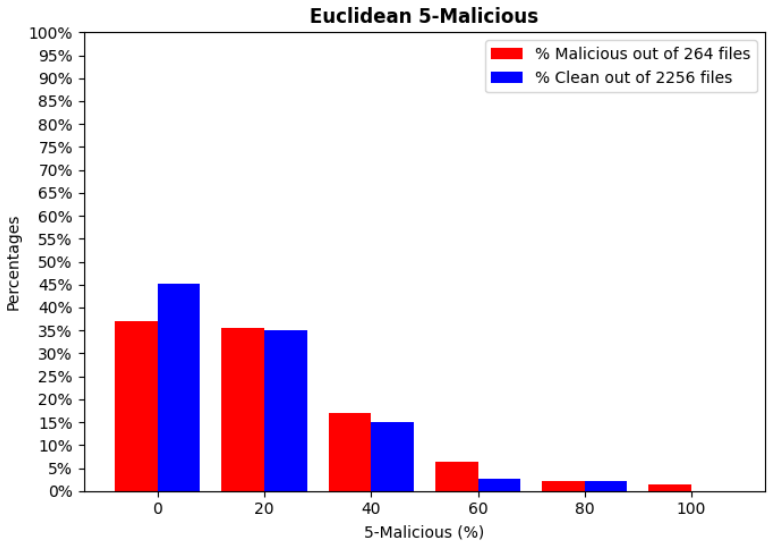


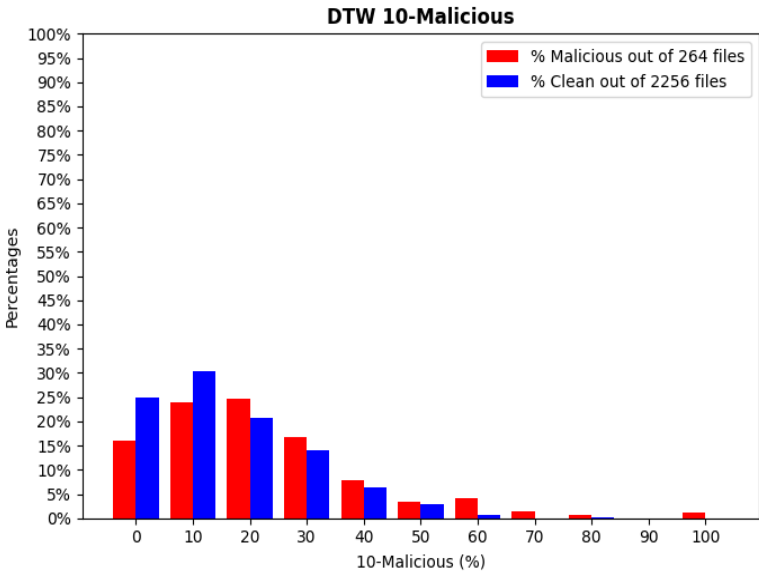
עבור התפלגות **גודל הקובץ**, יש שינוי בין סוגי הקבצים. בעוד שכרבע מהקבצים הנקיים בגודל קטן (0-50 KB), יתר הקבצים הנקיים מתפלגים בצורה דומה על שאר גדלי הקבצים. מבחינת קבצים זדוניים כשישית מהקבצים בגודל קטן, ובנוסף כ-40% מהקבצים שוקלים 350-750 KB.

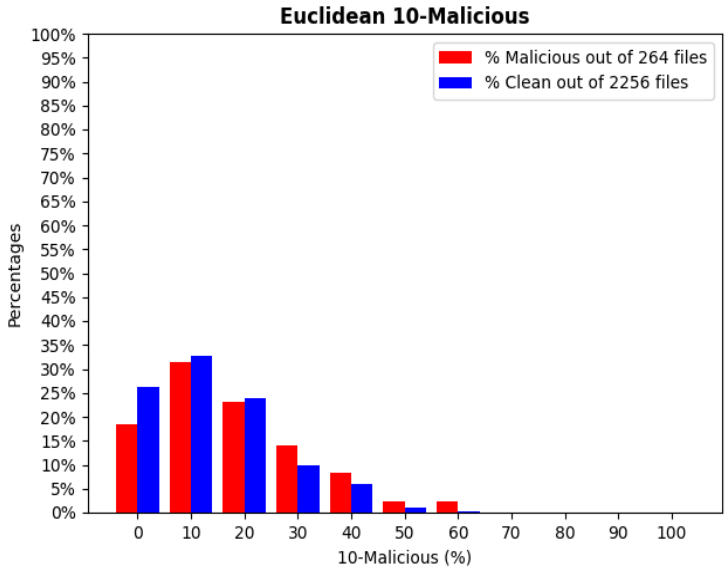
2. <u>מאפיינים על סמך המרחקים שחישבנו-</u> עבור כל אחת משתי השיטות לחישוב מרחקים מצאנו לכל קובץ את k=5,10,15 הקבצים עם המרחק הכי קטן ממנו.

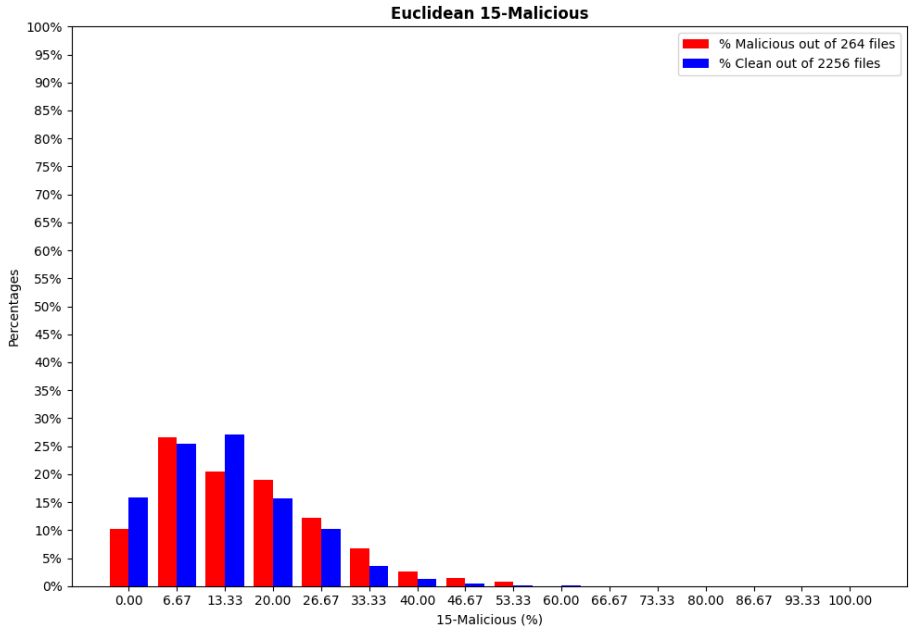
המאפיין **DTW/Euclidean k-Malicious** מייצג את אחוז הקבצים ה<u>זדוניים</u> מתוך k הקבצים הקרובים ביותר לכל קובץ בכל שיטה.

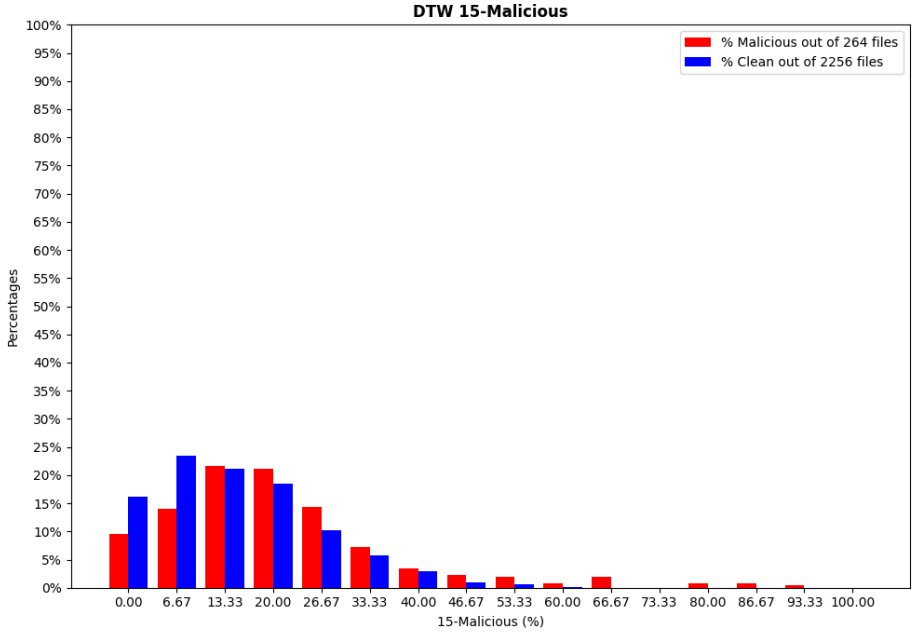








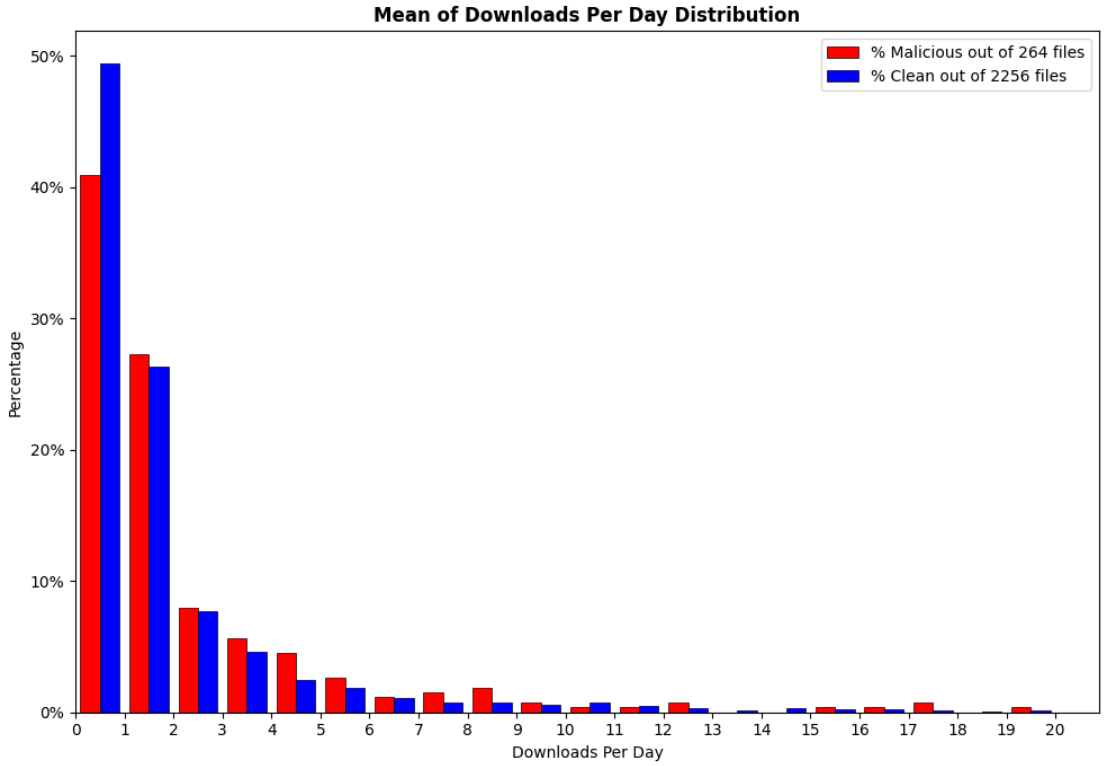




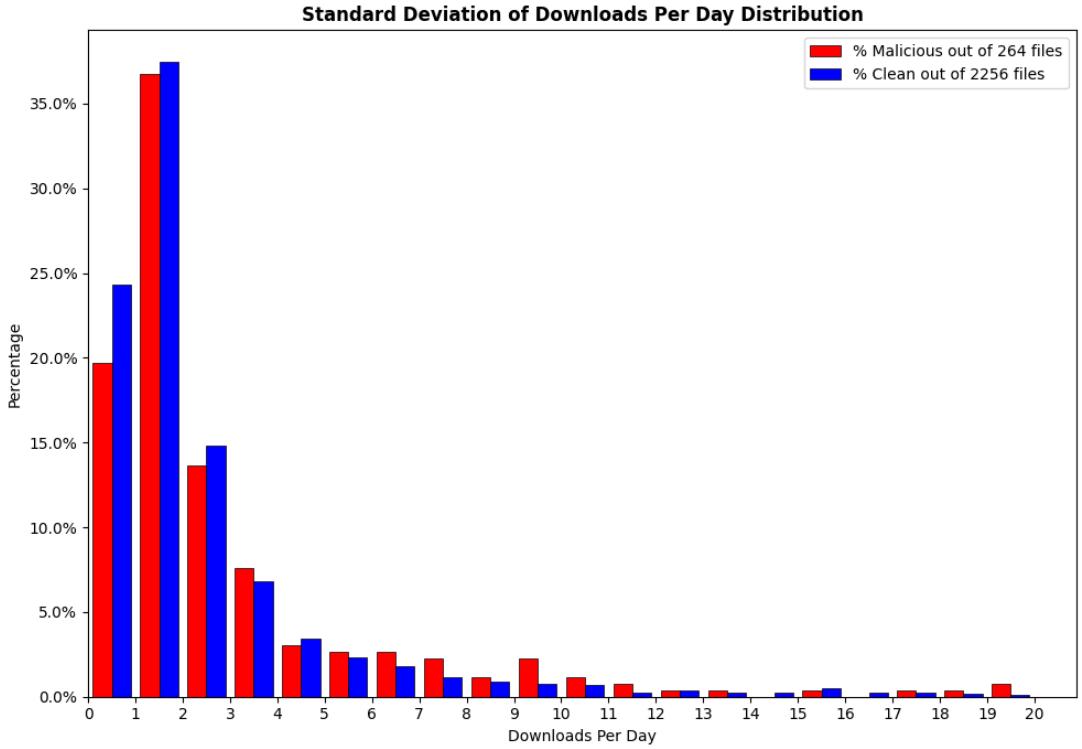
ניתן לראות שעבור כל קובץ, אחוז הקבצים הזדוניים מתוך ה-k הכי קרובים אליו נשאר דיי דומה בין שתי השיטות, דבר שאולי מצביע על דמיון רב בין קבצים מסוימים אשר מתבטא בשתיהן, לעומת קבצים "גבוליים" שהיחס בין מרחקם אינו קוהרנטי בין שתי השיטות.
בנוסף, עבור k=15 ניתן לראות שבחישוב האוקלידי לא היו כלל קבצים שמתוך 15 הקבצים הכי דומים להם 66% ומעלה היו זדוניים, ואילו בחישוב ע״י DTW רק עבור קבצים זדוניים קיבלנו נוכחות של 66% ומעלה של קבצים זדוניים מתוך ה-15 הקרובים ביותר לכל קובץ.

3. <u>מאפיינים על סמך סדרות הזמן-</u>

המאפיין **Day Count Mean** מציין את ממוצע ההורדות היומי של הקבצים. ניתן לראות כי את כמחצית מהקבצים הורידו 0-1 פעמים בממוצע ביום בתקופת הזמן אותה בדקנו. כרבע מהקבצים ירדו על 1-2 מכונות, ושאר הקבצים ירדו על 2 מכונות או יותר.

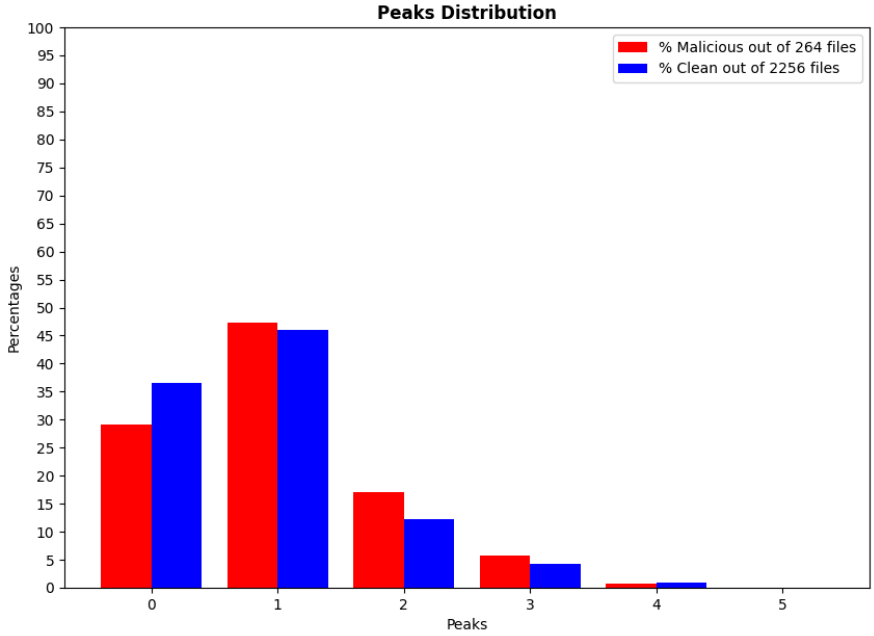


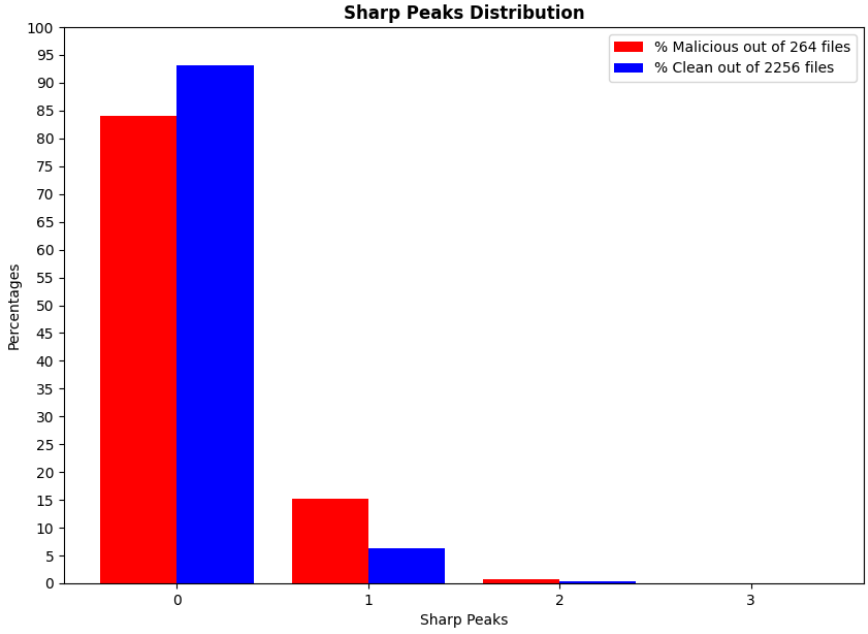
המאפיין **Day Count STD** מציין את סטיית התקן של מספר ההורדות ליום, כלומר זהו ההפרש בין ממוצע ההורדות ליום למספר ההורדות הכולל באותו יום בתקופת ימי ה-Train. גם כאן לא ניכר הבדל משמעותי בין שני סוגי הקבצים.



כפי שציינו, המאפיינים **Peaks** ו-**Sharp Peaks** מייצגים קפיצות במספר ההורדות.
המאפיין **Peaks** מציין את מספר הימים בסדרת הזמן בהם מספר ההורדות עלה על 3, ויש עלייה ביחס למספר ההורדות בסביבה שלו.
כדי לאפיין את ה-Peaks השתמשנו באלגוריתם למציאת מקסימום יחסי בדאטה.

המאפיין **Sharp Peaks** מציין את מספר הימים בסדרת הזמן בהם ההפרש בין מספר ההורדות באותו יום להורדות בסביבתו עולה על 15.
כדי לאפיין את ה-Sharp Peaks השתמשנו באלגוריתם למציאת נישאות טופוגרפית (Topographic prominence) שהיא המידה שבה בולטת פסגה מעל סביבתה. תכונה זו משמשת למיון של הרים ופסגות כ"ראשיים" או "משניים". ככל שפסגה בולטת יותר מעל סביבתה, כך היא נחשבת יותר כפסגה "ראשית" (ויקיפדיה).

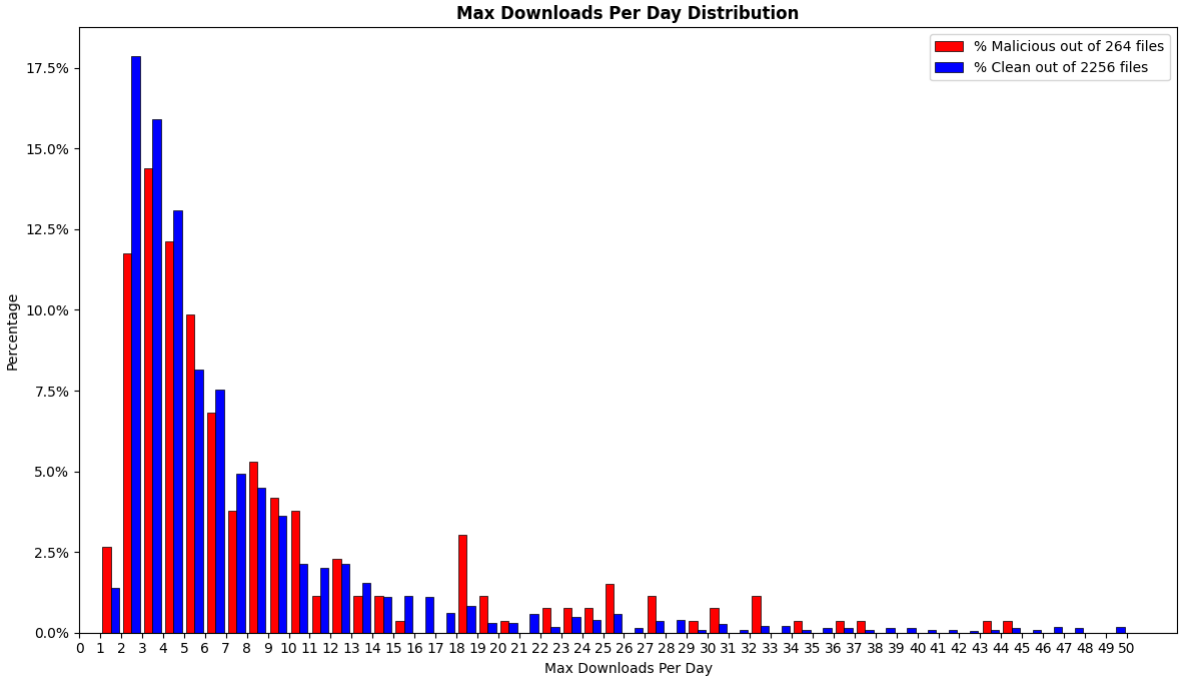




בחרנו את ערכי הסף בהתאמה לטווחי ההורדות היומיים, ונציין ששינוי קל בערכי הסף הללו מביאים לתוצאות דומות.

ניתן לראות כי רוב הקבצים מכילים לכל היותר "פיק" מתון אחד. לרוב הקבצים הנקיים אין "פיקים" חדים, דבר התואם לתבנית ההורדות המתונה של קבצים נקיים, ו-1-0 עליות חדות לקבצים זדוניים.

המאפיין **Max Day Count** מציג את התפלגות מספר ההורדות המקסימלי בכל סדרת זמן עבור כל קובץ. ניתן לראות שההתפלגות דומה למדיי בין קבצים נקיים לזדוניים, כך שרוב הקבצים ירדו לכל היותר על 2 מכונות שונות ביום. במבט רחב יותר, ההתפלגות מתרכזת בעיקר באיזור 2-20 הורדות, כך שניתן להבין שרוב הקבצים, גם הנקיים וגם הזדוניים, ירדו לכל היותר על 2-20 מכונות ביום.



המאפיין **Min Day Count** נבחן אך נמצא כלא יעיל משום שההתפלגות הייתה מאד צרה ולרוב המוחלט של הקבצים הערך המינימלי היה 0, לכן הורדנו אותו מרשימת המאפיינים.

למרות שנראה כי אין הבדלים גדולים בין התפלגות מאפיינים ביחס לקבצים נקיים וביחס לזדוניים, אנו צופים כי למאפיינים הקשורים ל"פיקים" תהיה החשיבות הגבוהה ביותר. פיצ'רים אלו מייצגים את תבניות ההורדה של קבצים נקיים וזדוניים שלפי ההיגיון אמורות להיות שונות.
בנוסף לכך, גם למאפיינים הקשורים לאחוז הזדוניים מתוך k הקבצים הקרובים ביותר לכל קובץ נשער כי הם יתנו לנו מידע משמעותי. אנו מניחים כי קבצים מסוג מסוים יהיו בעלי תבנית דומה ולכן יהיו קרובים ביותר לקבצים מאותו הסוג.

בנוסף, ביחס לשיטות חישוב המרחקים, אנו צופים כי מעצם היותו אלגוריתם דינאמי, DTW ייתן תוצאה מדויקת יותר מבחינת אמינות הפיצ'רים ולכן יניב תוצאה גבוהה יותר במדד ה-AUC ביחס למרחק אוקלידי.

Machine Learning

זהו השלב בו המאפיינים שאספנו באים לידי ביטוי כאמצעים לאימון מכונה לומדת, כך שתוכל לאבחן קבצים זדוניים על סמך המטה-דאטה שלהם ועל סמך מופעים אחרים שלהם במרחב המקוון.
בחרנו לאמן שני סוגים של מודלים- Logistic Regression ו-TreeClassifier כאשר הרצנו כל אחד מהם פעמיים בשינוי של אחד מהיפר-הפרמטרים שלו.
על כל אחד מהמודלים בשינוי היפר-הפרמטר שלו הפעלנו 5-Fold Cross Validation, כאשר הדאטה בו השתמשנו הוא קבוצת ה-Train. בכל איטרציה מתוך החמש החזיר כל מודל את ערך ה-AUC עבור היפר הפרמטרים שנתנו לו. AUC זהו השטח מתחת לעקומה, כאשר ככל שערכו קרוב יותר ל-1 המודל טוב יותר. הציון של כל מודל נקבע על ידי ממוצע של הציונים מחמש הריצות.

תוצאות 5-Fold Cross Validation

		DTW		Euclidean	
מודל	היפר-פרמטר	תוצאות ה-AUC	ממוצע תוצאות ה-AUC	תוצאות ה-AUC	ממוצע תוצאות ה-AUC
Logistic Regression 1	penalty=l2 solver='liblinear' max iter=300	[0.89484127 0.89484127 0.89484127 0.89484127 0.8968254]	0.8952	[0.89484127 0.89484127 0.89484127 0.89484127 0.8968254]	0.8952
Logistic Regression 2	penalty=l2 solver='liblinear' max iter=500	[0.89484127 0.89484127 0.89484127 0.89484127 0.8968254]	0.8952	[0.89484127 0.89484127 0.89484127 0.89484127 0.8968254]	0.8952
TreeClassifier 1	random state=0 max depth=10 min samples leaf=1	[0.88293651 0.89087302 0.86706349 0.89087302 0.89484127]	0.8853	[0.875 0.89087302 0.87301587 0.87896825 0.87301587]	0.8782
TreeClassifier 2	random state=0 max depth=3 min samples leaf=1	[0.8968254 0.89880952 0.89087302 0.89880952 0.9047619]	0.8980	[0.8968254 0.9047619 0.89285714 0.90674603 0.8968254]	0.8996

מהטבלה ניתן לראות שהציון הגבוה ביותר התקבל עבור מודל של עץ בעל עומק ששווה ל-3.

לאחר הרצה של הTEST על המודל הנבחר נוכל לראות את חשיבותו של כל מאפיין.

מקדם (Coefficient) של כל פיצ'ר מייצג את המתאם בינו לבין סיווג הקובץ כזדוני ע"י המודל, כאשר ככל שהמקדם קרוב יותר ל1, יש קשר חזק יותר בין המשתנים.
מספר חיובי – מעיד על יחס ישר בין המשתנים, מספר שלילי - יחס הפוך.
נבחין כי אצלנו כל המקדמים חיוביים לכן יש יחס ישר בין ערך המאפיין לזדוניותו, ז"א ככל שהמקדם גדול יותר כך ערך גבוה יותר של המאפיין מעיד כי הקובץ יסווג כזדוני יותר.

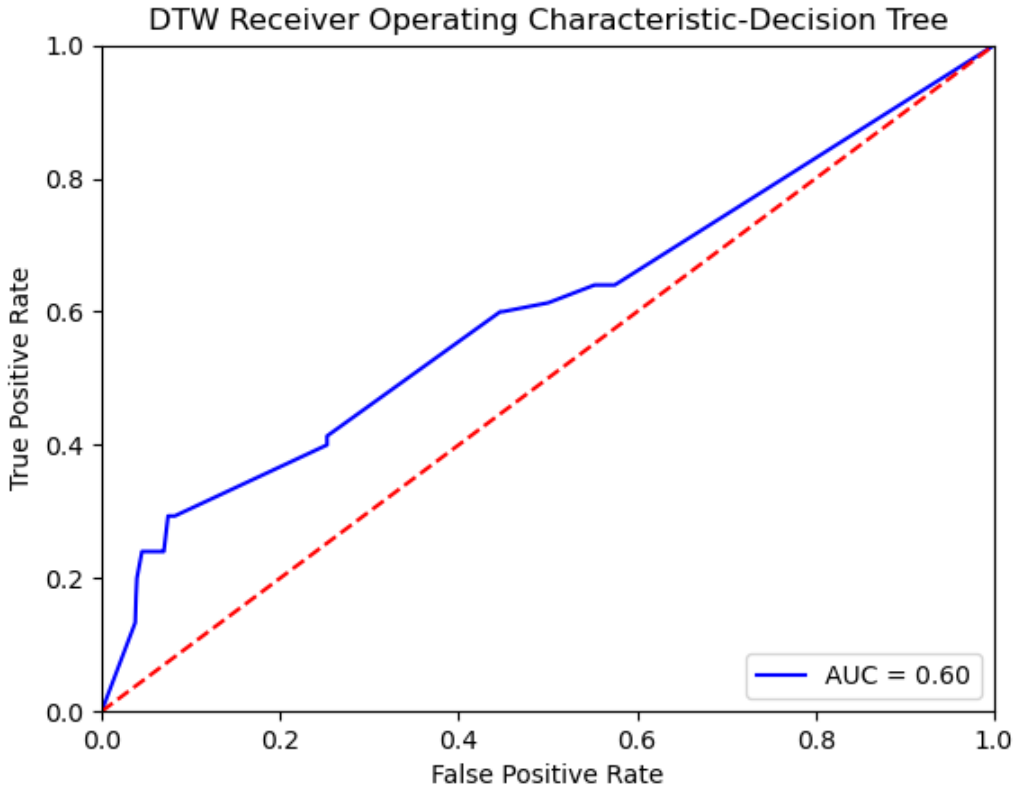
נסמן:
p – ההסתברות שקובץ מסווג כזדוני.
c – coeff.
נקבל את הנוסחה הבאה: $\log_2\left(\frac{p}{1-p}\right) = c$, והנוסחה הנגזרת ממנה: $2^c = 2^{\log_2\frac{p}{1-p}} = \frac{p}{1-p}$

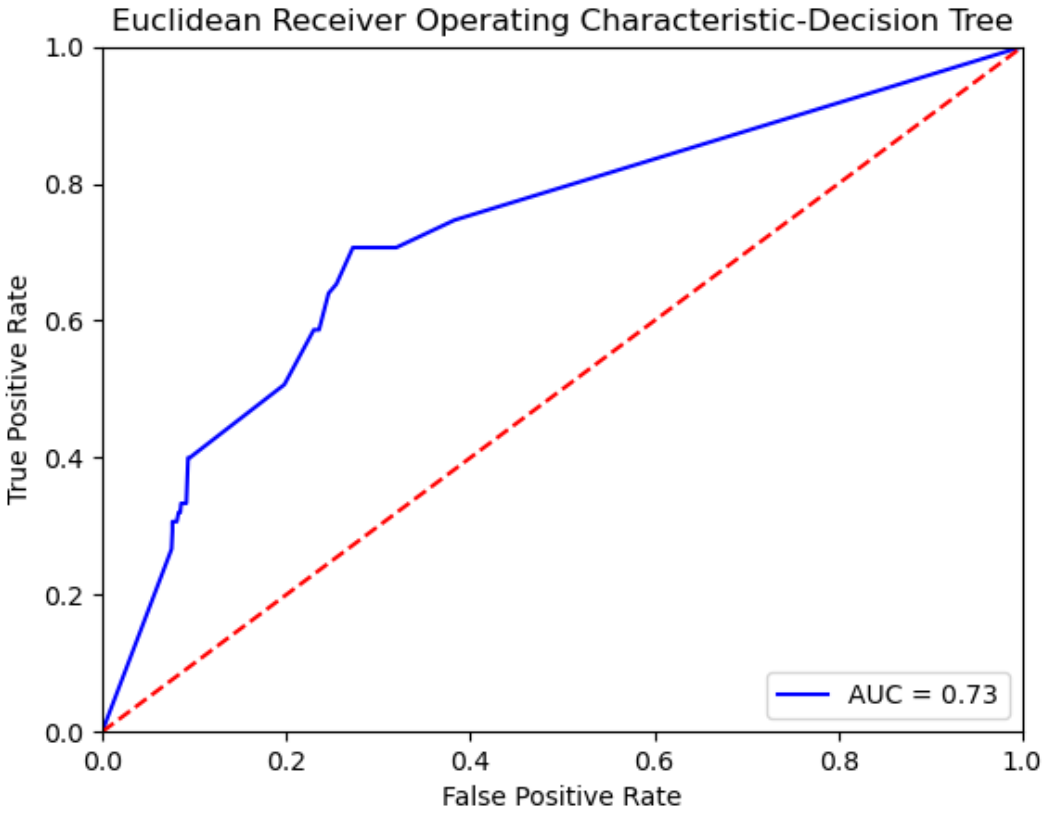
לפיכך, ככל ש- 2^c גדול יותר מ-1 כך p גדל ושואף ל-1, זאת אומרת שיש קשר חזק יותר בין ערך המאפיין עבור כל קובץ לבין סיווגו של הקובץ כזדוני, כך שחשיבותו בתהליך הסיווג גבוהה יותר.
וההפך, ככל ש- 2^c גדול יותר מ-1 כך p גדל ושואף ל-0, יש קשר חלש יותר ומשקלו בסיווג לזדוני יהיה נמוך יותר.

Decision Tree Classifier: Feature Importance

	DTW		Euclidean	
מאפיין	**מקדם**	**2 בחזקת המקדם**	**מקדם**	**2 בחזקת המקדם**
ממוצע הורדות ליום (Day Count Mean)	0.05782196	1.040893135	0.079813233	1.056881210990 6793
סטיית תקן הורדות ביום (Day count STD)	0.11986302	1.086631685	0.124117817	1.08984111
מספר הורדות מקסימלי ביום (Max Day Count)	0.088227192	1.063063069	0.141231545	1.10284615
גודל הקובץ (size)	0.484950138	1.399537494	0.490043769	1.404487485
אחוז הקבצים הזדוניים מתוך 15 הקבצים הקרובים ביותר (15-Malicious(%))	0.107041929	1.077017677	0.082445605	1.058811379
אחוז הקבצים הזדוניים מתוך 10 הקבצים הקרובים ביותר (10-Malicious(%))	0.037787041	1.026538005	0.033993254	1.023842113
אחוז הקבצים הזדוניים מתוך 5 הקבצים הקרובים ביותר (5-Malicious(%))	0.027598358	1.019313869	0.019348826	1.013501923
תדירות הורדת הקובץ-סך הורדות על מכונות שונות (Prevalence)	0.060643304	1.042930704	0.017130288	1.011944584
מספר ימים עם קפיצה בכמות ההורדות (Peaks)	0.01606706	1.011199083	0.007873346	1.005472306
מספר ימים עם קפיצה חדה בכמות ההורדות (Sharp peaks)	0	1	0.004002317	1.002778046

תוצאות המודל עבור כל שיטת חישוב מרחקים:





מסקנות:
ניתן לראות שבניגוד להנחה כי DTW יספק תוצאות מדויקות יותר, הסתבר לנו ששיטת המרחק האוקלידי הניבה תוצאות טובות יותר במדד ה-AUC.

בנוסף, בניגוד לציפייה כי המאפיינים הקשורים לחישוב ה"פיקים" יהיו בעל משקל משמעותי בסיווג הקובץ, מאפיינים אלו התגלו כבעלי החשיבות הנמוכה ביותר. הדבר מחזק את מסקנתנו בחלק הראשון, כי אולי על מנת לראות הבדלים בדפוסי ההורדות יש צורך בהגדלת כמות הקבצים ו/או את טווח הזמנים הנמדד.

המאפיין המשפיע ביותר הוא גודל הקובץ, גם בחישוב ע"י השיטה האוקלידית וגם ע"י DTW.
כבר בגרף התפלגות הגודל ניתן היה להבחין בהבדלים בין קבצים זדוניים לנקיים, למרות שהבדלים אלו לא נראו גדולים. התברר כי לעומת מאפיינים אחרים בהם ההבדלים היו מינוריים אף יותר, הבדלים אלו אכן היו משמעותיים והובילו לכך שזהו המאפיין הטוב ביותר לסיווג.

לאחר מכן נמצאים במדד החשיבות מאפיינים הקשורים למספר הורדות.
למאפיינים הקשורים למרחקים ישנה חשיבות נמוכה יותר משציפינו, אך מבין כלל המאפיינים, מאפיין המרחק עבור k=15 התגלה כבעל חשיבות מעט גבוהה יותר מזאת של k=5,10 גם ב-DTW וגם באוקלידי, כך שהוא יכול להיות מעט יותר רלוונטי עבור המודל.

כבר בשלב הגרפים ניתן היה לחשוד כי, בניגוד לציפייה, הפיצ'רים הקשורים למספר ה"פיקים", כלומר מאפיינים הקשורים לתבניות הורדה- אינם משפיעים כפי שחשבנו על סיווג הקבצים.
בנוסף, שלב ה-ML הפריך את ההנחה כי DTW יהיה מדויק יותר מהחישוב האוקלידי, שכן קיבל ציון פחות טוב ממנו במדד ה-AUC.
בנוסף לכך, עבור רוב המאפיינים שבדקנו, משקלם בשלב ה-ML קרוב יותר ל-0, מה שמעיד שהפיצ'רים לא יעילים במתן תמונה טובה בהינתן מסד הנתונים איתו עבדנו.
להפתעתנו, דווקא גודל הקובץ, שבהנחה הראשונית לא שיערנו שיהיה בעל משמעות רבה בסיווג הקובץ, קיבל משקל השואף ל- 0.5, ועלה בפער על משקלי הפיצ'רים האחרים. זאת אומרת שהשפעתו בסיווג הקובץ לזדוני ונקי הייתה המשמעותית ביותר.

לסיכום, בהינתן כמות הנתונים שהייתה בידינו, המודלים בהם השתמשנו סיפקו תמונה שונה מאוד מזו שציפינו לה, גילינו כי הפיצ'רים להם ייחסנו חשיבות גבוהה בשלב ההשערה בתור גורם משפיע על המודלים, התגלו כפיצ'רים עם חשיבות אפסית, בעוד שפיצ'רים שלא ייחסנו להם חשיבות גדולה בשלב ההשערה הם אלו שקיבלו את החשיבות הגדולה ביותר, אך גם פה, בצורה חלקית, שלא יכלה לספק לנו תמונה חזקה וחד משמעית לגבי סיווגם של הקבצים.
לצערנו, משקלי הפיצ'רים שהתקבלו לא סיפקו את התוצאה הרצויה כלל. ייתכן שאם היו לנו מאפיינים אחרים אשר היו מביאים תמונה אמינה יותר, היינו יכולים לחזות את סיווג קבוצת ה-test שלנו בצורה טובה יותר. היינו יכולים להסתמך עליהם עד כדי ויתור על פיצ'רים אחרים שמשקלם היה נמוך ונכללו בפועל, וזאת על מנת לספק תוצאה מדויקת יותר.