**Microsoft Security Bulletin Analysis**

*דוח הכנת הנתונים מוצג על ידי בר כהן וסהר חיים יעקב.* ​  
​ דוח זה עוסק בשלבי הכנת הנתונים כחלק מהליך ניתוח הנתונים בפרויקט. מטרתו היא להעריך את איכות הנתונים שנאספו, לזהות בעיות אפשריות כגון ערכים חסרים או שגויים, ולבצע עיבודים מקדימים לניתוח מעמיק. התהליך כולל בחירת נתונים רלוונטיים, ניקוי ואינטגרציה, יצירת מאפיינים חדשים, התאמת פורמט הנתונים, וביצוע ניתוח נתונים ראשוני (EDA) כדי לזהות דפוסים ומגמות. דוח זה מספק תיעוד מפורט של שלבי ההכנה שבוצעו, במטרה להבטיח שימוש אופטימלי בנתונים להמשך העבודה.

***שם מרצה : מר אבי זכאי.***

***שם מנחה : מר חנן לב.***

***מגישים :***

***בר כהן 208110254***

***סהר יעקב 314741851***



תוכן עניינים

[בחירת הנתונים 3](#_Toc192246301)

[תכונות ה – Dataset 3](#_Toc192246302)

[ניקוי נתונים 7](#_Toc192246303)

[שלבי ניקוי הנתונים 7](#_Toc192246304)

[בניית נתונים חדשים 12](#_Toc192246305)

[Orange Data Mining 13](#_Toc192246306)

[שילוב נתונים 17](#_Toc192246307)

[אתר ההורדות של מייקרוסופט 19](#_Toc192246308)

[עיצוב נתונים 19](#_Toc192246309)

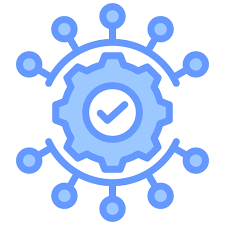
[הסבר על המודלים השונים 23](#_Toc192246310)

[הסברי מדדים 23](#_Toc192246311)

[מסקנות מדדים 24](#_Toc192246312)

[ניתוח נתונים (EDA) 26](#_Toc192246313)

[טבלאות תדירות 26](#_Toc192246314)

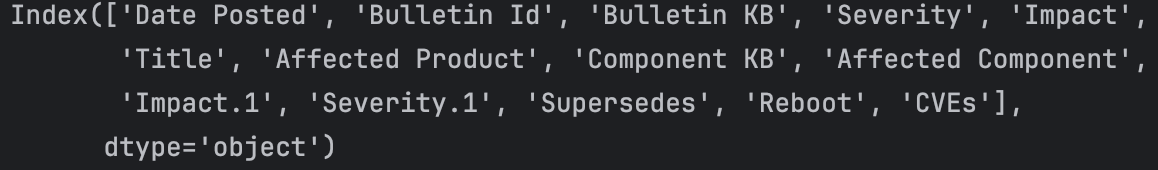


# בחירת הנתונים

בסעיף זה אנו בוחנים את מידת ההשפעה של כל פיצ׳ר על משתנה המטרה Severity, אשר משמש לחיזוי מידת הפגיעה והנזק באירועי האבטחה.

כדי לבחור את הפיצ׳רים הרלוונטיים ביותר, ננתח את מטריצת הקורלציה, ונזהה את התכונות בעלות הקשר החזק ביותר למשתנה המטרה. פיצ׳רים שערכם קרוב ל-0 מצביעים על חוסר קשר סטטיסטי ל -Severity, ולכן הם פחות רלוונטיים למודל החיזוי שלנו. בנוסף, פיצ׳רים עם קורלציה חיובית או שלילית מובהקת כלומר, ערכים גבוהים או נמוכים משמעותית מאפס עשויים להיות משמעותיים.

התכונות איתם אנו מתמודדים:



איור 1 : רשימת התכונות של DataSet

## תכונות ה – Dataset

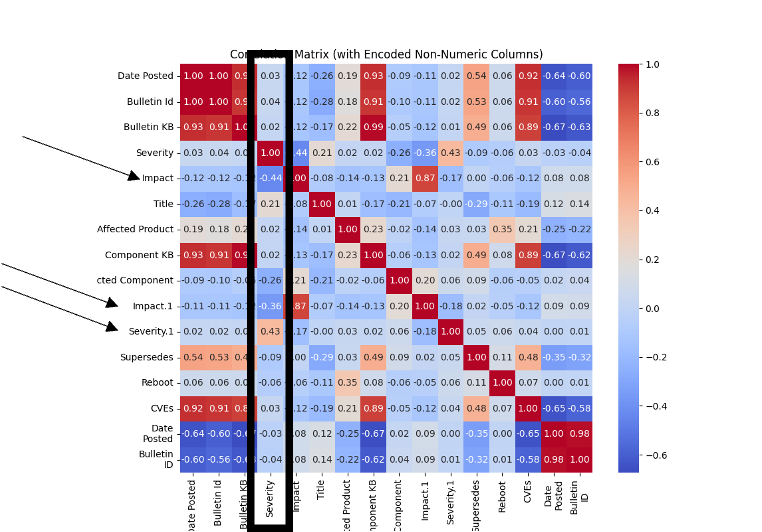
* Date Posted - תאריך הפרסום של הפגיעה מופרד ב- [',' '/'], נפריד את התאריך ליום, חודש ושנה.  
   התאריך הוא חסר משמעות אך ניתן להוציא משמעות מתוך החודש מכיוון שאם בחודשים מסוימים דווחו על פגיעות זה יתרום למודל, מאשר יום ושנה שגם הם פחות רלוונטיים.
* Bulletin Id - מספר מזהה, יכול להיות רלוונטי.
* Bulletin KB - המדריך שפורסם, נשתמש בו כפיצ׳ר אך סיכויי ההשפעה שלו נמוכים.
* **Severity** - מידת הפגיעה הסופית, משתנה המטרה.
* Impact - השפעת הבעיה על המערכת, ניתן להשתמש בו כפיצ׳ר והוא רלוונטי למודל.
* Title - כותרת המתארת את עדכון האבטחה, נוכל להשתמש בו כפיצ׳ר, עם זאת מידת ההשפעה שלו כנראה תהיה נמוכה.
* Affected Product - המוצר או הרכיב שהושפע בערבות תיאור הבעיה, רלוונטי מאוד.
* Component KB - מדריך שמסביר על הרכיב המושפע, אפשרי להשתמש בו כפיצ׳ר.
* Affected Component - רכיב המערכת המושפע, רלוונטי מאוד!
* Impact.1 - מידת השפעה משנית ניתן להשתמש.
* Severity.1 - מידת פגיעה ראשונית, ניתן להשתמש והיא חיובית להבנה ראשונית.
* Supersedes - מציין האם יש בעיה שעדכון תיקן, פיצ׳ר שיכול להיות רלוונטי.
* Reboot - מציין האם צריך לבצע הפעלה מחדש לאחר העדכון, רלוונטי למודל.
* CVEs - מספק אמה זיהה את בעיית האבטחה, רלוונטי.

קבוצת הנתונים הנוכחית איכותית מאוד, וכוללת מספר רב של משתנים רלוונטיים.  
 ישנם מספר משתנים אשר עשויים להשפיע ישירות על משתנה המטרה, כגון Impact, Component,, Severity.1 ועוד.   
 בנוסף, קיימים פיצ'רים שיכולים להוסיף מידע מועיל, אך השפעתם עשויה להיות נמוכה יותר, כגון Component KB.

כמות הנתונים הרחבה יכולה לסייע למודל להפיק תחזיות מדויקות, יחד עם הנדסת תכונות נכונה וניקוי נתונים יסודי.  
 האיכות הגבוהה של הנתונים והיקפם הרחב מאפשרים יצירת מודל חיזוי מדויק, במיוחד כאשר משלבים הנדסת תכונות וניקוי נתונים קפדני.

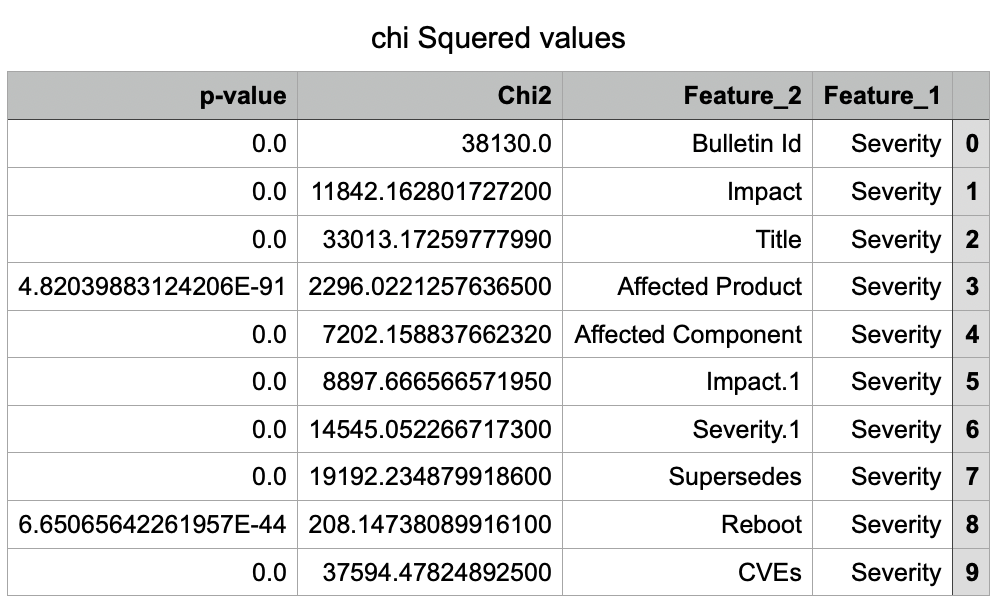
בשלב זה של המחקר, ביצענו ניתוח מעמיק של התכונות הקיימות בנתונים, במטרה לזהות אילו פיצ'רים עשויים להשפיע בצורה המשמעותית ביותר על המודל. באמצעות בחינה יסודית של הנתונים ושימוש במטריצת קורלציה, הצלחנו לזהות קשרים בין משתני הקלט למשתנה המטרה, שלוש התכונות הכי יעילות מבחינת המטריצה קורלציה הם – **impact**, **impact.1**,**severity1**, **title** – רמת פגיעות נוספת (ראשונית) , מידת השפעה ראשונית ומשנית ותיאור עדכון האבטחה.

ניתוח זה אפשר לנו להבין אילו פיצ'רים רלוונטיים ביותר לחיזוי רמת הפגיעות, וכיצד הם תורמים למודל. כמו כן, גילינו כי ישנם פיצ'רים מסוימים שהשפעתם נמוכה יחסית, אך עדיין עשויים להוסיף מידע מועיל. שלב זה ביחס לשלבים קודמים, מהווה התקדמות משמעותית בתהליך פיתוח המודל, שכן הוא מספק תובנות קריטיות לגבי מבנה הנתונים ומאפשר לבצע בחירות מושכלות להמשך האופטימיזציה של המודל.



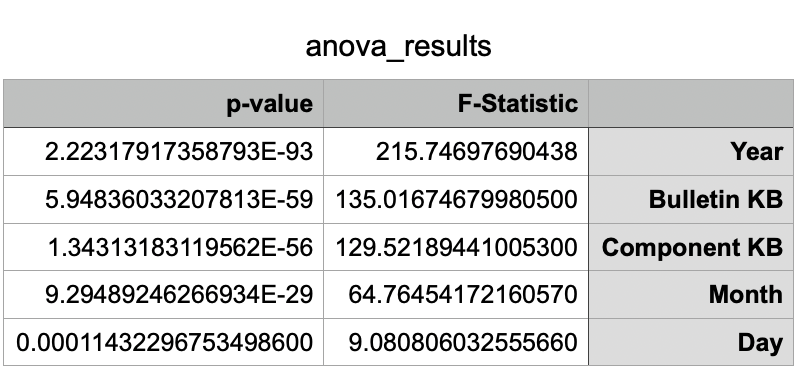
איור 2 : מטריצת הקורלציה המלאה

בנוסף, נוכל לבצע מבחן חי בריבוע (Chi-Square Test) כדי לבדוק קשרים בין משתנים קטגוריאליים לבין משתנה המטרה. מבחן זה מאפשר להעריך האם קיימת תלות בין המשתנים ולזהות אילו מהם משפיעים על המשתנה המוסבר. ערך חי בריבוע גבוה מצביע על קשר חזק יותר בין המשתנים, בעוד שערך p קטן מעיד על כך שהקשר בין המשתנים אינו מקרי. ככל שערך ה-p קטן יותר, כך גוברת ההסתברות לקיומו של קשר מובהק בין המשתנים. נשתמש במבחן זה לכל משתנה קטגורי מול משתנה המטרה, כדי להבין את הקשרים ולשפר את בחירת התכונות למודל.



איור 3 : מבחן chi^2 עבור ערכים קטגוריאלים

נבחין כי אלה הם עשרת המשתנים הקטגוריאליים שנתנו את הערכים הטובים ביותר.  
יתר על כן נשתמש במבחן נוסף שבוחן קשרים בין משתנה מטרה המטרה קטגוריאלי לבין משתנים מספריים. נשתמש ב- **Anova test** :



איור 4 : מבחן F עבור ערכים נומרים

אם ערך ה-F גבוה מאוד, ניתן להסיק כי יש הבדל מובהק בין הקבוצות, כלומר שקיים קשר משמעותי בין המשתנה המסביר למשתנה המטרה. יחד עם זאת, יש להתחשב גם בערך ה-p, אשר מעיד על רמת המובהקות הסטטיסטית של הקשר – ערך קטן יותר מצביע על כך שההבדלים אינם מקריים.

כחלק מתהליך בחירת התכונות (Feature Selection), נוכל להתמקד במשתנים בעלי השפעה משמעותית על המודל, תוך סינון תכונות פחות רלוונטיות שעשויות להוסיף רעש למודל ולפגוע בביצועים.

סיכום -

בשלב זה של המחקר, אנו מתמקדים בבחירת הפיצ'רים הרלוונטיים ביותר לחיזוי משתנה המטרה Severity, המתאר את מידת הפגיעה של עדכוני האבטחה. תהליך זה הינו קריטי להצלחת המודל, שכן הוא מאפשר לנו להתמקד במשתנים התורמים ביותר ולסנן מידע לא רלוונטי שעשוי להוסיף רעש.

בכדי לבחור את הפיצ'רים המשמעותיים ביותר, ניתחנו את מטריצת הקורלציה וזיהינו את התכונות בעלות הקשרים החזקים ביותר עם משתנה המטרה. מטריצת הקורלציה מאפשרת לנו להבין את הפיצ'רים, כאשר:   
פיצ'רים עם ערכים הקרובים ל-0 מצביעים על קשר חלש או היעדר קשר סטטיסטי ל-Severity, ולכן הם פחות רלוונטיים למודל.  
פיצ'רים עם קורלציה גבוהה (חיובית או שלילית) נחשבים לבעלי פוטנציאל גבוה להשפעה על מידת הפגיעות ויכולים להוות מדדים חשובים בחיזוי.

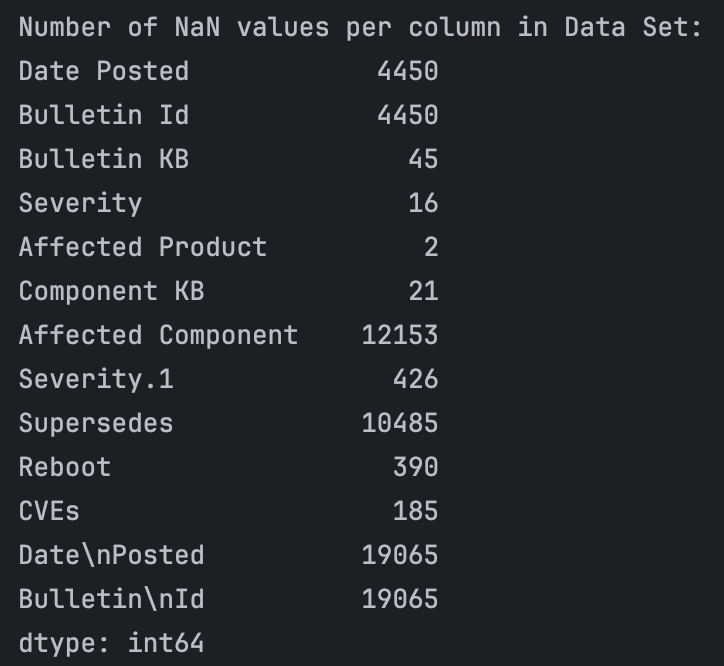
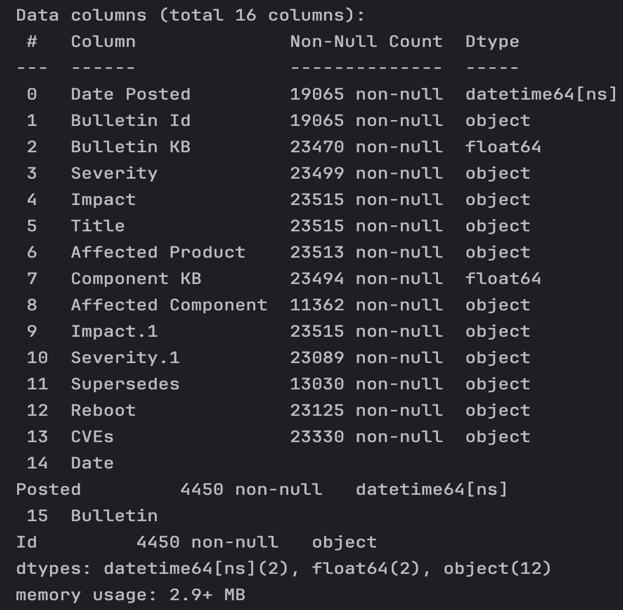
בנוסף , כדי להעריך את השפעתם של משתנים קטגוריאליים על משתנה המטרה, ביצענו מבחן חי בריבוע (Chi-Square Test). מבחן זה מאפשר לקבוע האם קיימת תלות בין משתנים קטגוריאליים שונים לבין Severity.  
מבחן זה סייע לנו לזהות עשרת המשתנים הקטגוריאליים שנתנו את הערכים הטובים ביותר, ובכך לצמצם את מספר התכונות שניתן לכלול במודל החיזוי.

ניתוח משתנים מספריים - מבחן ANOVA (F-Test)  
כדי להעריך את הקשר בין משתנה מטרה קטגוריאלי למשתנים מספריים, ביצענו מבחן ANOVA. מטרת מבחן זה היא לבדוק האם קיים הבדל מובהק בין קבוצות המשתנים המספריים ביחס למשתנה המטרה.

תהליך זה מהווה התקדמות משמעותית בפיתוח המודל, התהליך מספק לנו תובנות חשובות על מבנה הנתונים ומאפשר לבצע בחירות חיוניות להמשך האופטימיזציה של המודל. בעזרת הנדסת תכונות מדויקת, נוכל לשפר את ביצועי המודל ולספק תחזיות מדויקות יותר לגבי רמת הפגיעות של עדכוני האבטחה.

# ניקוי נתונים

בסעיף זה אנו נחקור את הנתונים עצמם ונבדוק האם ואיפה ישנם נתונים חסרים, בתכונות ובמשתנה המטרה. הנתונים שלנו מכילים נתונים חסרים, דבר שיפגע באמינות ודיוק המודל, עם זאת ננסה למזער ולמלא את הנתונים החסרים בצורה מועילה ועם כמה שפחות נזק למודל.  
בנתונים שלנו יש נתונים חסרים הן במטרה והן בתכונות. נטפל בערכים חסרים, כאשר שלבי העבודה הם: מילוי, הפרדה, הרצת מודל ראשוני, חיזוי , איחוד.

****כהקדמה אנו נבחן את הנתונים שלנו ונבדוק איפה קיימים ערכים חסרים והיכן :

איור 5 : הצגת כמויות non nulls בנתונים

איור 6 : הצגת כמויות הnulls בנתונים

בנוסף קיימים 219 רשומות כפולות אותם נרצה להסיר מתוך הדאטה , כדי למנוע הטיות ושיפור איכות הנתונים.

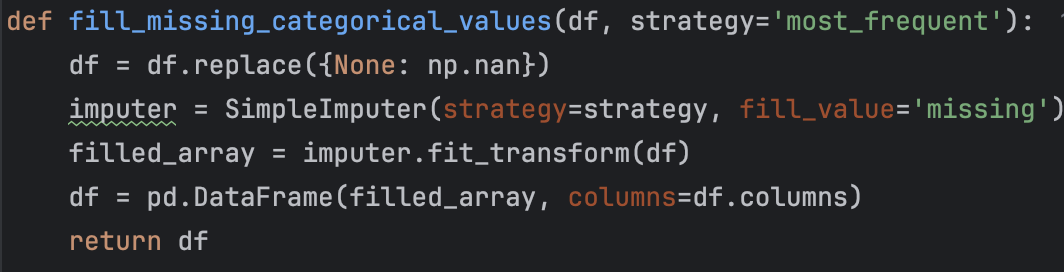
## שלבי ניקוי הנתונים

**שלב 1 :**

**מילוי** - בשלב עיבוד הנתונים, חשוב להבטיח שאין ערכים חסרים שעלולים לפגוע בניתוח או בביצועי המודל. לכן, נבצע תהליך מילוי Impute עבור כל הערכים החסרים בעזרת שיטות שונות. ראשית, נחליף את כל הערכים מסוג None בערך np.NAN כדי להבטיח שכל הספריות שמבצעות חישובים על הנתונים יזהו אותם כערכים חסרים. לאחר מכן, נתמודד עם הערכים החסרים באמצעות simple imputer .  
עבור משתנים שסוגם מוגדר כקטגוריאליים ונומריים, נשתמש באסטרטגיית **הערך השכיח**, כך שכל ערך חסר יוחלף בערך שמופיע הכי הרבה פעמים באותה עמודה. שיטות אלו נבחרו מכיוון שמילוי עם הערך השכיח עבור משתנים קטגוריאליים מונע הכנסת ערכים חדשים שעלולים לשבש את ההתפלגות הקיימת, בעוד ששימוש בממוצע עבור משתנים מספריים מסייע להפחתת השפעות קיצוניות ומשמר את מגמות הנתונים.

איור 7 : פונקצית מילוי ערכי null

**שלב 2 :**

**הפרדה** - בשלב עיבוד הנתונים, נבצע הפרדה של הנתונים בהתאם להימצאות ערכים חסרים במשתנה המטרה. הפרדה זו מדגישה את שיפור איכות הנתונים ומאפשרת גישה ממוקדת לטיפול בערכים החסרים, תוך שמירה על דיוק ואמינות הנתונים התקינים. לשם כך, נחלק את הנתונים לשני Data Frames נפרדים:

מערך נתונים חסר – יכיל אך ורק רשומות שבהן המשתנה המטרה חסר. קבוצה זו תעבור תהליכי עיבוד מתאימים, כגון השלמת ערכים חסרים באמצעות חיזוי מבוסס מודלים, שימוש בממוצעים או בשיטות סטטיסטיות אחרות, או לחלופין הסרה של רשומות אם החוסרים משמעותיים מדי.

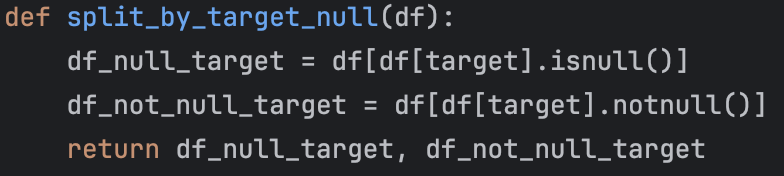
מערך נתונים מלא – יכלול רשומות שבהן המשתנה המטרה וכל שאר התכונות מלאים. נתונים אלה ישמשו ישירות לאימון המודל, ללא צורך בהתערבות נוספת.

**למה להפריד?**

כדי להשלים את אופן טיפול מדויק ומלא של הנתונים נצטרך לבצע הפרדה ראשונית לניתוח מיטבי שיועיל לרמת הדיוק של המודל הסופי.

טיפול נפרד בנתונים החסרים מאפשר לנו לבנות מודלים ספציפיים להשלמת ערכים, מה שעשוי לשפר את דיוק התחזיות. לדוגמה, המודל יכול להתמקד בהבנת הדפוסים שבהם חסרים ערכים ולממש שיטות להשלמה באופן יותר מדויק. ניתוח נפרד של הנתונים החסרים עשוי לחשוף דפוסים או מגמות ייחודיות שלא היו ניכרות בניתוח הכולל. ייתכן שנמצא קשר בין החסרים לבין תכונות אחרות, מה שיכול לעזור במתן הסברים לתופעות הנתונים. בנוסף, הפרדה מאפשרת לנו לנסות גישות שונות על כל קבוצת נתונים בנפרד ולבחור את השיטה המיטבית לכל אחת. לדוגמה, אם אנחנו מטפלים בנתונים חסרים בצורה אחרת מאלה המלאים, נוכל לבדוק שיטות שונות ולהשוות את הביצועים.

הפרדה זו מאפשרת לנו למקד את תהליך הטיפול בערכים החסרים ללא פגיעה בנתונים התקינים, ובכך לשמר את אמינות המודל. בנוסף,

****

איור 8 : חלוקת הנתונים לשני dataframes שונים

איור 9 : שם המסמך

**שלב 3 :**

**הרצת מודל ראשוני** - לאחר עיבוד הנתונים והשלמת הערכים החסרים בתכונות, נבצע הרצת מודל ראשוני. שלב זה כולל שני תהליכים עיקריים: עיבוד הנתונים המלאים והשלמת הערכים החסרים עבור הנתונים שבהם המשתנה המטרה חסר. נבחר במודל עץ החלטה (**Decision Tree**) . עץ החלטה הוא מודל פשוט ויעיל לחיזוי, ויתרונו בכך שהוא מסוגל ללמוד דפוסים לא לינאריים בצורה ברורה וניתנת להבנה. המודל ייבנה על בסיס הנתונים המלאים ויכלול את התכונות הידועות לחיזוי הערכים החסרים של המשתנה המטרה, בנוסף הוא יכול להשתמש בערכים קטגוריאליים לאחר קידוד בצורה מהירה.

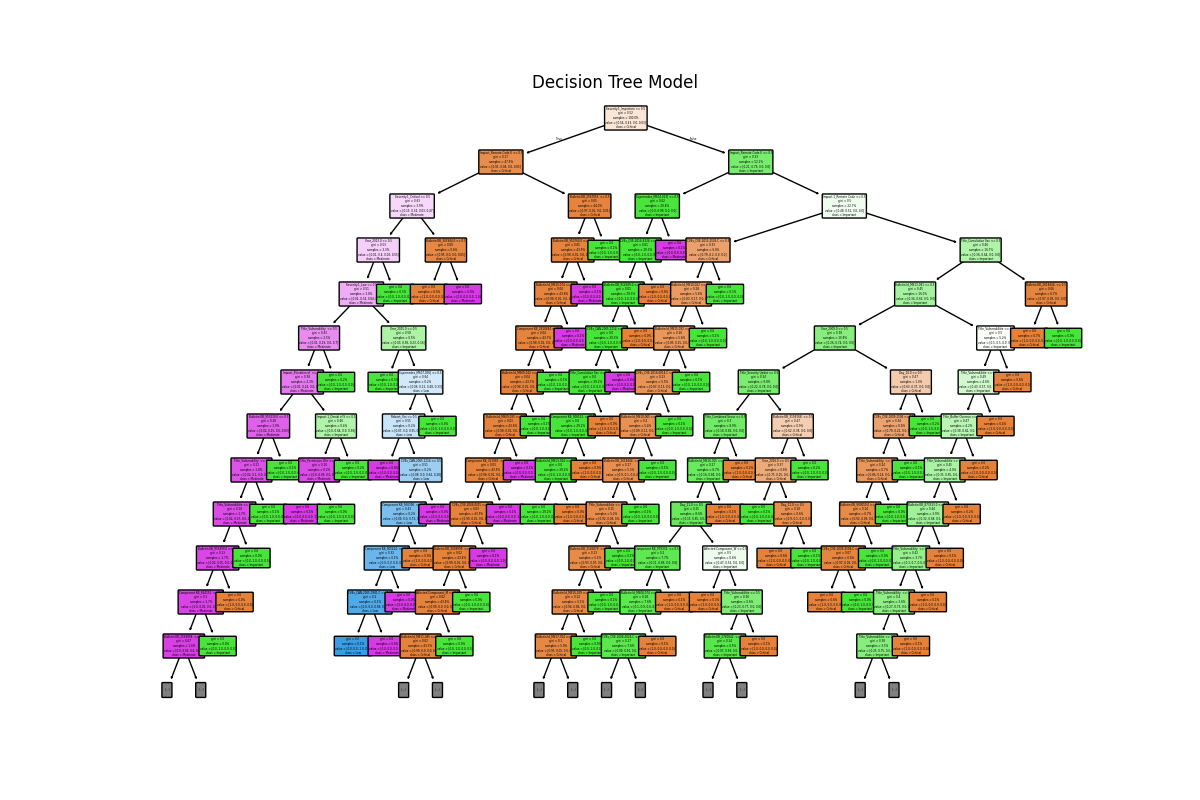
**איך יתבצע התהליך?**

**אימון המודל**: המודל מאומן על נתונים מלאים, בהם כל המשתנים (כולל המשתנה המטרה) מכילים ערכים מלאים.

**חיזוי הערכים החסרים**: לאחר שהמודל מאומן, אנו מיישמים אותו על ה-Data Frame שבו המשתנה המטרה חסר. המודל מנצל את הדפוסים שנלמדו מהנתונים המלאים כדי לחזות את הערכים החסרים של המשתנה המטרה.

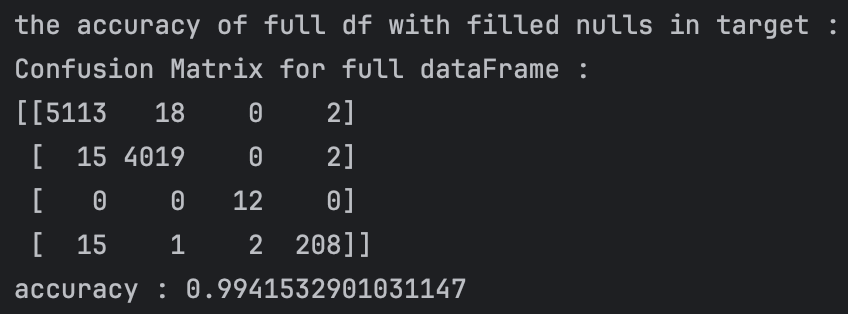
תהליך זה מאפשר לנו לבצע השלמה מבוססת-מודל בצורה שמשקפת את המגמות הקיימות בנתונים.

לסיכום, מטרת הרצת המודל הראשוני עם עץ החלטה לשפר את איכות המודל הסופי.



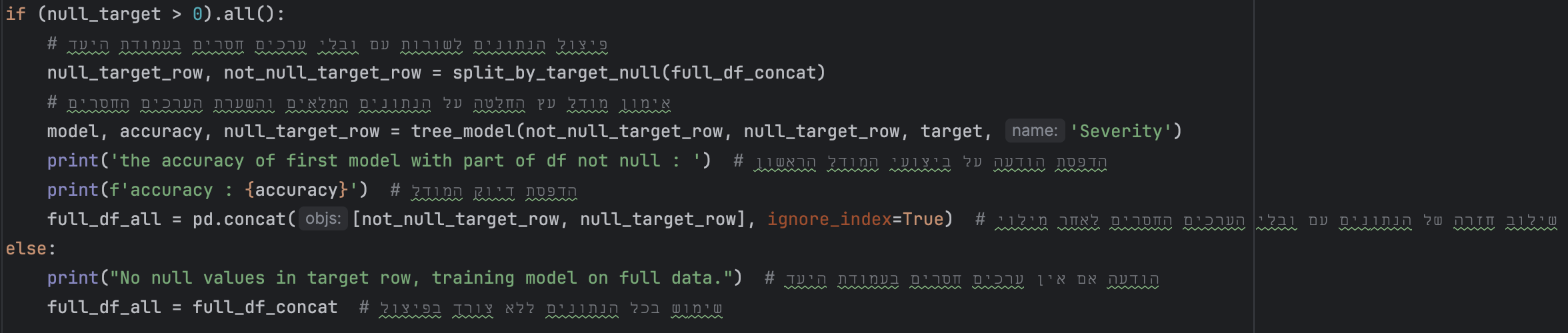
איור 10 : הרצת מודל ראשוני על הנתונים בעלי משתנה מטרה מלא

איור 11 : דיוק המודל ומטריצה בלבול

****

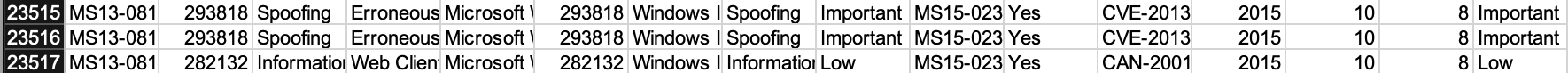
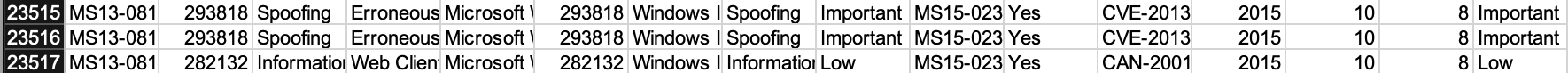
**שלב 4 :**

**חיזוי** - לאחר הרצת המודל הראשוני, נשתמש בו כדי לבצע **חיזוי** לערכים החסרים במשתנה המטרה. המודל, שאומן על ה -**Data Frame** יחזה את ערכי המטרה עבור הרשומות שבהן הם חסרים. כתוצאה מכך, נקבל **Data Frame** , שבו לכל רשומה יש ערך מטרה משוער, מה שיאפשר לנו להמשיך בניתוחים ובשלבי עיבוד המידע הבאים.



איור 12 : חיזוי הנתונים החסרים וחיבור הנתונים

איור 13 : הצגה מתוך הנתונים המלאים עם 23,517 נתונים



ניתן לראות שלאחר החיזוי וחיבור הנתונים שמרנו על כמות נתונים מלאה להרצת המודל הסופי , תהליך זה יוביל להעלאת הדיוק של המודל ותרומה למטרת הפרויקט.

**שלב 5 :**

**איחוד** - לאחר חיזוי הערכים החסרים במשתנה המטרה, נבצע **איחוד** של שני ה-**Data Frames**: האחד עם הנתונים המקוריים והמלאים, והשני עם הנתונים שהשלמנו באמצעות המודל. פעולה זו מאפשרת לנו לשמור על **הגודל המקורי של הדאטה**, ובמקביל להגדיל את כמות הנתונים הזמינים למודל. כתוצאה מכך, נוכל לשפר את **רמת הדיוק של המודל** בשל ניצול מרבי של כל הנתונים הקיימים.



איור 14 : איחוד ושמירה של הנתונים המלאים

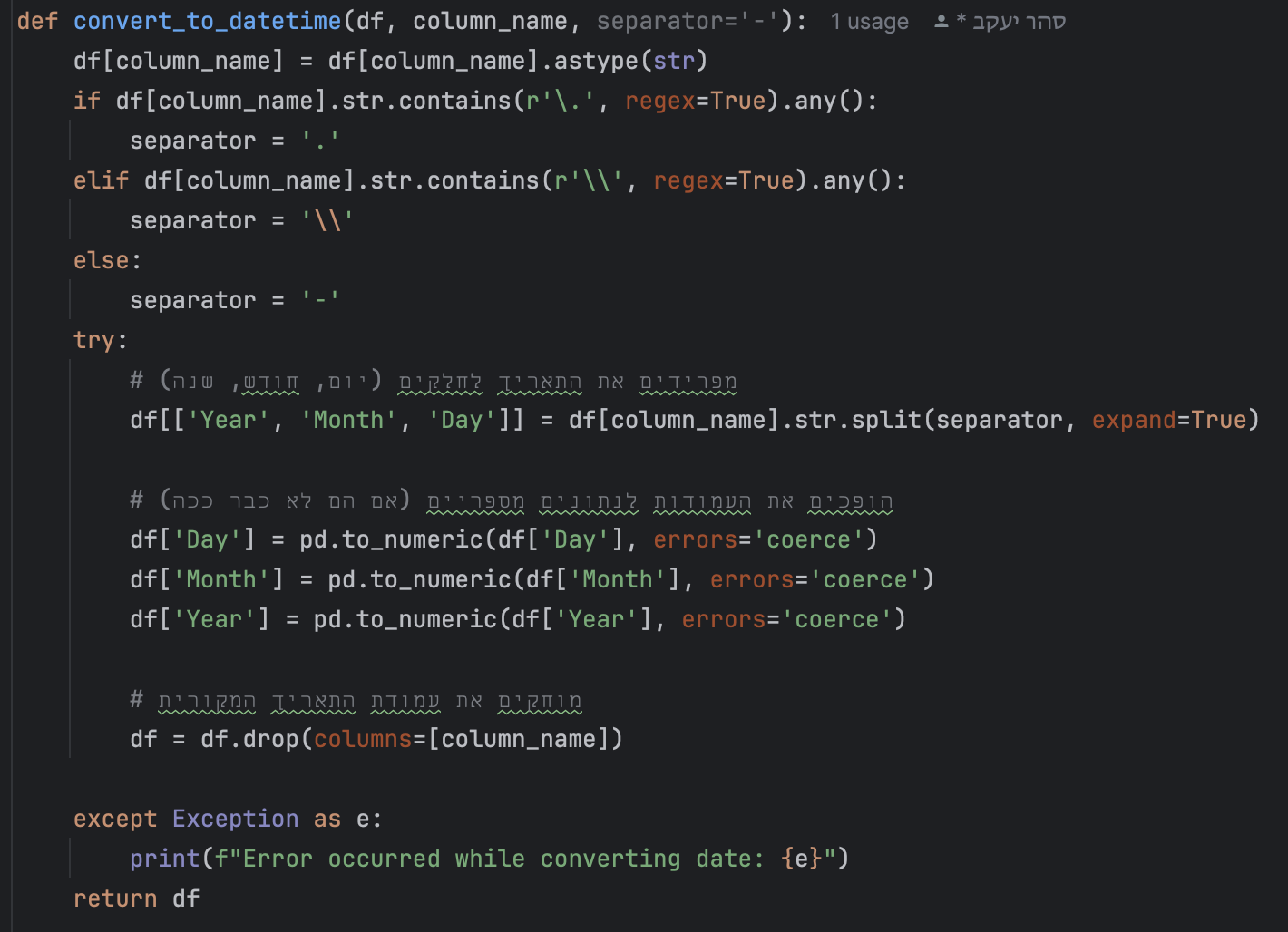
**הסבר על אופן השמירה –**

יצרנו תיקייה בשם output data באמצעות ספריית os בפיתון, שמסייעת ביצירת תיקיות ושינויי מערכת הפעלה.

לאחר מכן, שמרנו את הקובץ בפורמט excel שמאפשר עיצוב ברור של הטבלה. זאת כדי לשמור על סדר וארגון הקבצים.

בנוסף בעמודת התאריך אנו נתקלנו בחוסר עקביות בקידוד, כאשר מספר רב של תצפיות תאריך מופרדות באמצעות '/' ואחרים באמצעות '.' דבר שהקשה על הפרדת התאריך.

כדי להתגבר על חוסר עקביות בתאריך נזהה את המפריד בתאריך ונפעיל:



איור 15 : המרה של תאריך ל3 עמודות בנתונים

לבסוף, נשמור את ה-**Data Frame** המעודכן עם הערכים החזויים, כך שנוכל להשתמש בו עבור **מגוון המודלים הסופיים** שלנו. שמירת הפלט תאפשר לנו לבצע **השוואה בין ביצועי המודלים** השונים ולבחון את ההשפעה של שלבי העיבוד על התוצאות. תהליך זה **קריטי** לשיפור הדיוק ולבחירת המודל האופטימלי, שכן הוא מבטיח שנוכל לנתח, להשוות ולכוונן את האלגוריתמים בצורה מבוססת נתונים.

סיכום -

בפרק זה ביצענו חקר נתונים מקיף וזיהינו כי קיימים ערכים חסרים הן בתכונות והן במשתנה המטרה, מצב שעלול להשפיע לרעה על ביצועי המודל. כדי לטפל בכך, השתמשנו בשיטות מילוי ערכים חסרים. לאחר מכן, חילקנו את הנתונים לשתי קבוצות: רשומות עם ערכי מטרה חסרים ורשומות עם ערכים מלאים, כדי שנוכל לבצע השלמות מבוססות-מודל. בשלב הבא, אימנו מודל עץ החלטה על הנתונים המלאים וניצלנו אותו כדי לחזות את הערכים החסרים במטרה. לאחר שהתקבלו התוצאות החזויות, ביצענו איחוד של הנתונים ושחזרנו את המסגרת המקורית, אך עם מידע מלא יותר. נוסף לכך, זיהינו חוסר עקביות בקידוד התאריכים – חלקם הופרדו על ידי '/' ואחרים על ידי '.', ולכן ביצענו המרה תקנית כדי לאחד את הפורמט. בסוף התהליך, שמרנו את הנתונים המעובדים לשימוש עתידי במודלים נוספים ולביצוע השוואות בין שיטות ניבוי שונות. השלבים שננקטו בתהליך זה הבטיחו שימוש יעיל במידע הזמין, שיפור הדיוק של המודל וניהול נתונים איכותי יותר.

# בניית נתונים חדשים

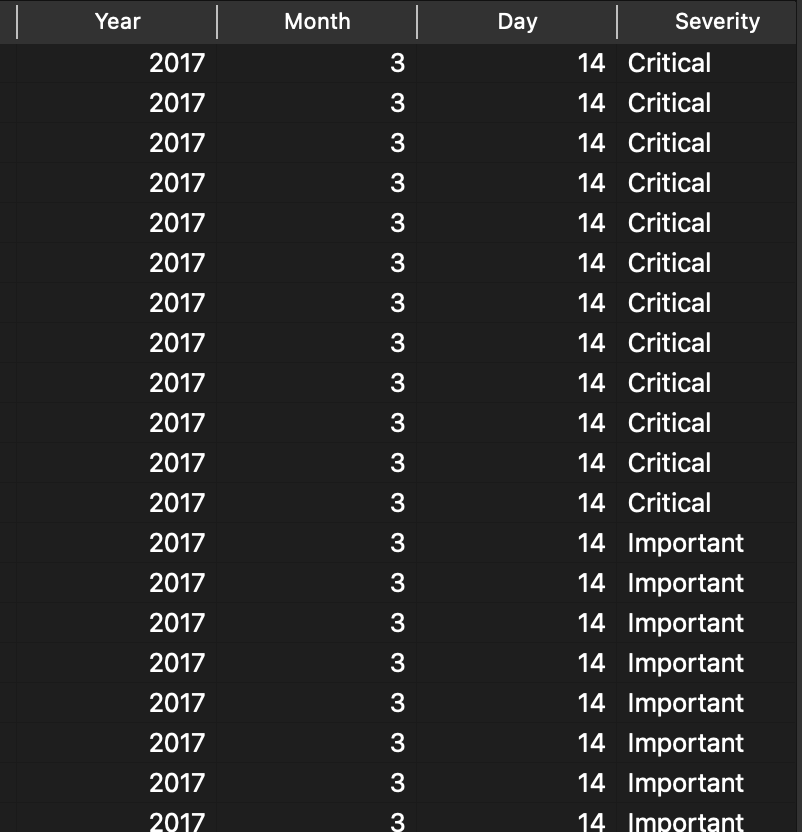
יצירת נתונים חדשים והכנתם לניתוח

בשלב זה, אנו נתאר את הנתונים החדשים ואת תהליך יצירת הרשומות והעמודות החדשות.

**גזירת נתונים** - בשלב הראשון, נבצע גזירת נתונים, כלומר הפקת עמודות חדשות מתוך עמודות קיימות. לדוגמה:

תאריכים: נתונים הכוללים תאריך ישופצו כך שנפריד אותם ליום, חודש ושנה. פעולה זו מאפשרת ניתוח מעמיק יותר של הנתונים על בסיס זמן, כגון זיהוי מגמות עונתיות או שבועיות.

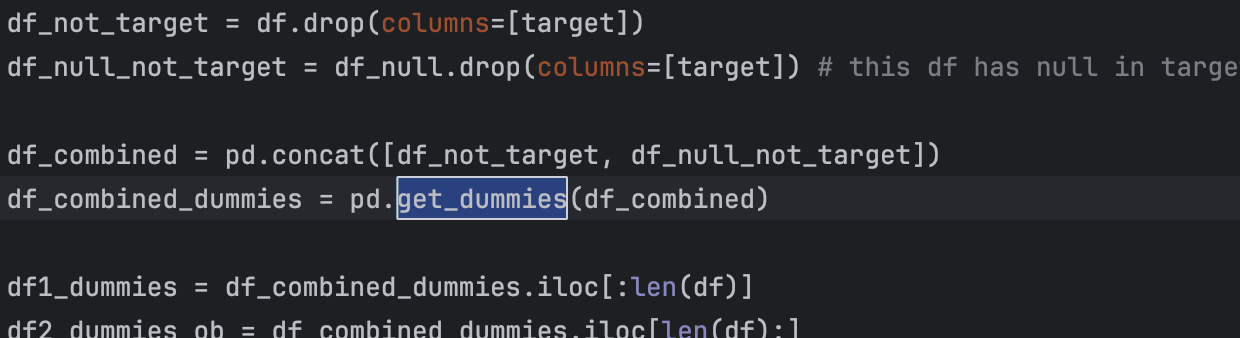
יצירת רשומות חדשות - בנוסף, נבצע יצירת רשומות חדשות, כלומר הוספת עמודות חדשות לנתונים שלנו, בהתאם לצורך. זה יכול לכלול הוספת מידע חדש שנאסף או נוצר במהלך הניתוח.



איור 16 : יצירת עמודות יום, חודש ושנה מתוך עמודת התאריך

**קידוד משתנים דאמי:**

כדי להכין את הנתונים לשלב המודל הסופי, עלינו להתחשב במשתנים הקטגוריאליים ולהמיר אותם לייצוג מספרי מתאים. לשם כך, נבצע קידוד משתני דאמי (Dummy Variables), שייצור עמודה בינארית לכל פרמטר בעמודת התכונה. כך נקבל N עמודות חדשות בהתאם למספר הערכים הייחודיים בכל משתנה קטגוריאלי. פעולה זו חשובה כדי לאפשר למודל להבין טוב יותר את ההשפעה של כל תכונה ולשפר את ביצועי החיזוי.

ההערכה מדגישה את החשיבות של כל שלב בתהליך ומספקת הסבר מעמיק יותר על הפעולות הנדרשות להכנת הנתונים לניתוח ולחיזוי.

איור 17 : המרה של משתנים קטגוריאלים באמצעות דמי

Orange Data Mining

**הסבר על תוכנה Orange Data Mining – ממשק גרפי**

**רקע על התוכנה –**

**Orange Data Mining** נועדה לשפר את ניתוח הנתונים וההבנה של אלגוריתמים ללמידת מכונה בצורה נגישה, אינטואיטיבית ויעילה. התוכנה הושקה כחלק ממחקר בתחום אינטליגנציה מלאכותית וניתוח נתונים, על ידי צוות חוקרים . היא נועדה לספק כלי ניתוח נתונים מתקדמים לגורמים מקצועיים כמו מדעני נתונים, חוקרים, אנשי אקדמיה וסטודנטים, וגם לאפשר למשתמשים פשוטים עם מעט ניסיון בתכנות להפיק תובנות משמעותיות מנתונים.

איור 18 : לוגו התוכנה Orange Data Mining

הפלטפורמה כוללת יותר מ-100 גדג'טים , שהם רכיבים גרפיים שמבצעים פעולות כמו טעינת נתונים, ניתוחם, והפעלת אלגוריתמים למידת מכונה שונים. Orange תומכת במגוון רחב של פורמטים של קבצים, כולל CSV, Excel, JSON, SQL, ומסדי נתונים אחרים, מה שמאפשר לה להתמודד עם סוגי נתונים מגוונים.

היא תומכת גם בהפקת ויזואליזציות אינטראקטיביות ומתקדמות, שמאפשרות להציג את התוצאות בצורה ברורה ונגישה. היא כוללת גרפים, דשבורדים, תרשימים ועוד, אשר מסייעים להבין את התוצאות ולהפיק תובנות מהנתונים.

הפלטפורמה מציעה גם יכולות עיבוד נתונים מתקדמות, כולל טיפול בנתונים חסרים, סינון נתונים, ניתוח נתונים קטגוריאליים, הפחתת ממדים (Dimensionality Reduction), ניתוח סדרות זמן, ועוד. בנוסף, היא מציעה כלים לבחירת מודל, חיפוש אחר המודל המתאים ביותר לנתונים, והערכת המודל באמצעות שיטות כמו cross-validation, confusion matrix, precision, recall ועוד.

**ממשק הגרפי -**

ממשק גרפי הוא אחד היתרונות, הוא מאפשר יצירה של זרימות עבודה תוך גרירה ושחרור של רכיבים. ממשק זה מאפשר ליצור ולבצע תהליכים מורכבים של ניתוח נתונים בצורה ויזואלית, כך שאין צורך בכתיבת קוד. התוכנה גם תומכת בקוד **Python**, כך שניתן לשלב אלגוריתמים מותאמים אישית בתוך הממשק, או להריץ קוד מתקדם כשיש צורך בכך.

התוכנה פותחה בשפת **Python** ומבוססת עליה, כך שהקוד שלה כתוב בעיקר ב-**Python**. כל רכיב בתוכנה הוא מודול **Python** שניתן להרחיב ולהתאים אישית. יתרה מכך, התוכנה מציעה API של **Python** שמאפשר לעבוד עם הנתונים שלה בצורה תכנותית ומתקדמת, ומספקת את הגמישות של עבודה עם קוד לצד היתרון של עבודה ויזואלית עם ממשק גרפי.

**מה Orange יכולה לעשות? -**

התוכנה מספקת כלים מגוונים בתחומים של ניתוח נתונים, חיזוי, ולמידת מכונה:

טעינת נתונים: ניתן לטעון נתונים ממגוון מקורות, כמו קבצי CSV, Excel, מסדי נתונים SQL, ועוד.

עריכת נתונים: כולל אפשרויות לניקוי נתונים, המרת סוגי משתנים (כגון המרה בין משתנים קטגוריאליים למספריים), טיפול בערכים חסרים ע״י impute ומילוי ערכים ע״י ממוצע או השכיח, ושינוי טיפוסים של עמודות.

המרת משתנים קטגוריאליים למשתני דמי: התוכנה תומכת בהמרת משתנים קטגוריאליים למשתנים דמי (One-Hot Encoding), תהליך מותאם לצורך הפעלת מודל רצוי.

עיבוד נתונים: מתבצע עיבוד נתונים מתקדם כמו הפעלת מודלים לשיפור נתונים, טיפול בערכים חסרים, וביצוע הפחתת ממדים (PCA).

למידת מכונה: התוכנה תומכת במודלים מתקדמים של למידת מכונה, כולל חיזוי, סיווג, קיבוץ ובמגוון רחב של סוגי מרחקים כמו אוקלידי וקוסינוס.

ויזואליזציות: יש בתוכנה מגוון רחב של כלים להפקת ויזואליזציות של נתונים ותחזיות, כולל גרפים, תרשימים, ודשבורדים אינטראקטיביים.  
בנוסף המערכת תומכת במגוון רכיבים שניתנים להתקנה וניתן לבצע בהם ויזואליזציות של מודלים שונים כמו מודל ארימה ועוד.

הערכה ותחזית: לאחר בניית המודל, ניתן להעריך את ביצועי המודל על ידי מדדים שונים כמו דיוק, Confusion Matrix, AUC (Area Under Curve), F1 Score ועוד.

**אלגוריתמים במערכת Orange -**

התוכנה מציעה מגוון רחב של מודלים של למידת מכונה, בהם ניתן להשתמש לביצוע חיזוי, סיווג, קיבוץ ועוד:

סיווג (Classification):

* Decision Trees
* Random Forest
* Naive Bayes
* k-Nearest Neighbors (k-NN)
* Logistic Regression
* Neural Network

קיבוץ (Clustering):

* K-Means
* Hierarchical Clustering

רגרסיה (Regression):

* Linear Regression
* Ridge/Lasso Regression

חיזוי ותחזית (Time Series Forecasting):

* Time Series Forecasting

הפחתת ממדים (Dimensionality Reduction):

* PCA (Principal Component Analysis)
* t-SNE

Orange Data Mining מציעה כלים למידת מכונה, כולל אלגוריתמים לסיווג, קיבוץ, רגרסיה, חיזוי והפחתת ממדים, כגון Decision Trees, K-Means, Linear Regression ו-PCA, המאפשרים ניתוח ויצירת מודלים חיזוי מדויקים.

**הערכת מודלים ב-Orange -**

לאחר אימון המודל, ניתן להעריך את ביצועי המודל על ידי מדדים שונים. Orange מאפשרת הערכה של המודל במספר דרכים:

Cross-validation: מאפשרת לבדוק את ביצועי המודל בצורה לא מוטה, על ידי חלוקה אוטומטית של הנתונים ל-k קבוצות.

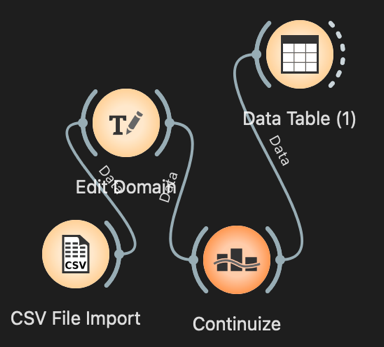
Test & Train Split: חלוקה של הנתונים לסטים נפרדים של אימון ובדיקה.

Confusion Matrix: להערכת ביצועים של מודלים לסיווג, כולל מדדים כמו דיוק, זיכרון, Precision, Recall, ו-F1 Score.

ROC Curve & AUC: להצגת ביצועי המודל על ידי גרף המתאר את יחס ה-FPR וה-TPR (True Positive Rate) המתקבלים בכל סף של המודל.

באמצעות אלו, ניתן להעריך את איכות המודל ולבצע אופטימיזציה אם יש צורך ניתן להוסיף קוד מותאם אישית של פיתון.

תוכנת Orange Data Mining היא פלטפורמה גמישה וחזקה שמאפשרת ביצוע ניתוחי נתונים מתקדמים, יצירת מודלים של למידת מכונה והפקת ויזואליזציות של נתונים בצורה אינטואיטיבית. עם התמחות בלמידת מכונה, הערכת ביצועים, וטיפול בנתונים, Orange מציעה את הכלים הנדרשים לחוקרים, מדעני נתונים ומשתמשים מתחילים כאחד.



איור 19 : תמונה זו מראה את תהליך encoder של התוכנה בנוסף לשינוי סוגי הנתונים והצגתם בטבלה.

טעינת הנתונים 🡨 שינוי סוג הנתונים 🡨 encoder 🡨 הצגה בטבלה את טבלת הencoder.

איור 19 : תהליך dummies בoranger data mining

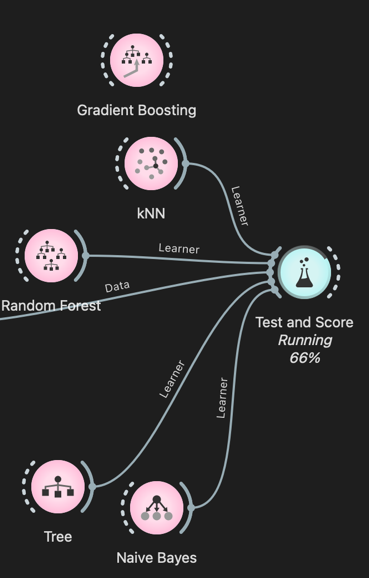
איור 20 מציג את תהליך יצירת המודלים השונים לצד שלב הערכתם. התהליך מתחיל מחיבור הנתונים אשר עברו עיבוד מוקדם בשלבים הקודמים, ולאחר מכן הם מוזנים למודלים שונים לצורך למידת מכונה. כל מודל מקבל את הנתונים שנוקו, הומרו והתאימו לפורמט אופטימלי ללמידה, ובונה מהם דפוסים על פי האלגוריתם שבו נעשה שימוש.

בשלב הבא, מתבצעת הערכת המודלים באמצעות רכיב Test and Score Running, אשר אחראי על בדיקת ביצועי המודלים שהוכשרו. רכיב זה מחלק את הנתונים לסטים אימון ובדיקה, מריץ את המודלים על הנתונים ומחשב מדדי ביצועים שונים. המדדים כוללים בין היתר:

F1-Score – מדד מאוזן הבוחן את איכות המודל תוך שקלול Precision ו-Recall.

Precision – אחוז התחזיות החיוביות שנמצאו נכונות מתוך כלל התחזיות החיוביות.

Recall – אחוז הדגימות החיוביות שהמודל הצליח לזהות מתוך כלל הדגימות החיוביות.



איור 20 : תהליך הערכת מודלים בorange data mining

לאחר חישוב, אנו נוכל לייצר את מטריצת הבלבול עבור כל אחד מהמודלים ולראות את מידות הנכונות של כל מודל.

בתמונה זו חישבנו מודלים שונים ובניהם , עץ החלטה , נאיב בייס ועוד.

סיכום -

בשלב זה של הפרויקט, אנו יוצרים נתונים חדשים ומכינים אותם לניתוח. תחילה, אנו גוזרים נתונים מעמודות קיימות, כמו פירוק תאריכים או חישוב כמות מכירות. לאחר מכן, אנו מאחדים את כל מסמכי האקסל למסד נתונים אחיד. כדי להיערך למודל, אנו ממירים משתנים קטגוריאליים לערכים מספריים באמצעות קידוד משתני דאמי. לבסוף, אנו בוחרים את המודל בעל הביצועים הטובים ביותר לחיזוי מדויק ואמין.

# שילוב נתונים

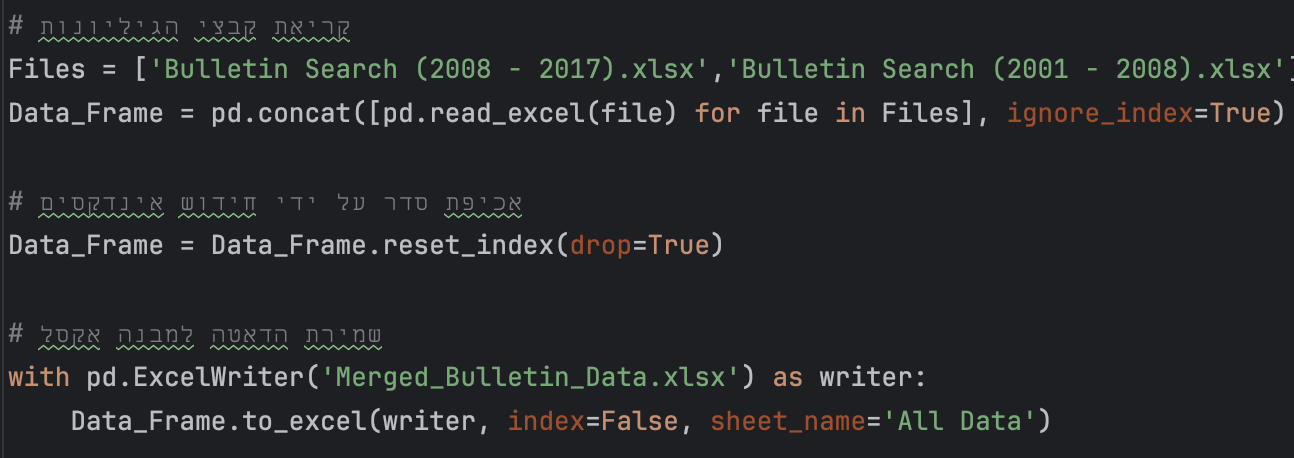
בשלב זה אנו נתאר שלב שילוב הנתונים , תהליך מיזוג וצירוף הנתונים שלנו מתוך אתר ההורדות של מיקרוסופט. הנתונים מכילים מגוון מסמכים שמתוארים באתר ההורדות :

CVRF Information.docx – מידע על פורמט דיווח CVRF.  
BulletinSearch2001-2008.xlsx – קובץ Excel עם מידע על תיקוני אבטחה בין השנים 2001-2008.  
MSRC-CVRF.zip – קובץ ZIP הכולל תיקוני אבטחה בפורמט CVRF החל מיוני 2012.  
BulletinSearch.xlsx – קובץ Excel עם מידע על תיקוני אבטחה מחודש נובמבר 2008 ועד היום.

**מה הם קובצי cvrf –**

קובץ CVRF (Common Vulnerability Reporting Framework) הוא פורמט מבנה נתונים סטנדרטי שנועד לשימוש במתן דיווחים אודות פגיעויות אבטחה במערכות מידע.  
בין הרכיבים העיקריים בקובץ CVRF:

Document Title: כותרת המסמך.ֿ  
Document Publisher: פרטי המפרסם, כולל פרטי יצירת קשר.  
Product Tree: רשימה של המוצרים הפגיעים.  
Vulnerability Information: פרטים על הפגיעות, כולל תיאור, CVE (מספר מזהה של פגיעות אבטחה), ורמות סיכון.  
Threats: תיאור של האיומים שנגרמים מהפגיעות.  
Fixes or Mitigations: פתרונות או אמצעי מניעה להפחתת הסיכון מהפגיעות.  
Document Tracking: מידע על גרסאות ושחרורים של המידע.

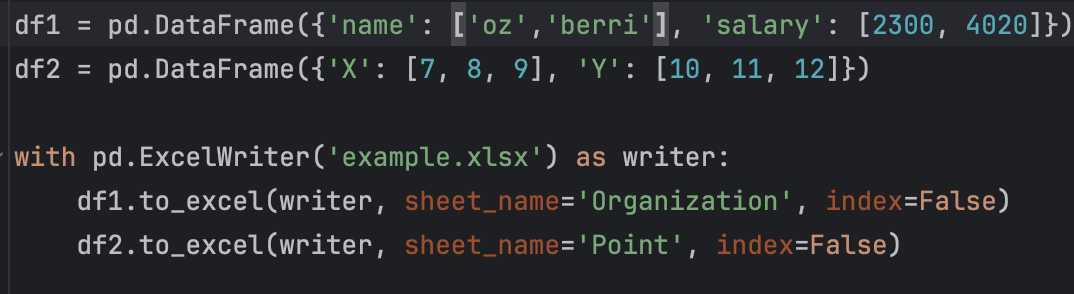
המסמכים כוללים שני מסמכי אקסל מופרדים שאיתם נעבוד, אותם אנו נצטרך לצרף באמצעות append וליצור את המסמך שאיתו נעבוד.

איור 22 : שם הקובץ המאוחד

איור 21 : שילוב קבצי הxlsx לקובץ אחד



**הסבר על אופן השמירה**השמירה כאן היא שונה מהדרך הרגילה, ההבדל הוא יצרת הקובץ וביצוע שינויי כתיבה בפורמט Excel, לעומת שמירה רגילה כמו ב- csv. זה מאפשר להוסיף נתונים מרובים לאותו הקובץ במספר גיליונות (sheets), במקרה זה גיליון בשם 'All Data'.  
ההבדל הוא שהקובץ נשמר באופן מובנה ונתמך ביכולות נוספות כמו גיליונות מרובים.

בדרך זו ניתן לשמור מספר נתונים בגיליונות נפרדים בקלות , לדוגמא :

איור 21 : שמירה של 2 גיליונות שונים לקובץ אחד



איור 22 : שמות הגליונות

כאן יצרנו קובץ אקסל בשם example שמכיל 2 גיליונות בשם : organization, point.

## אתר ההורדות של מייקרוסופט

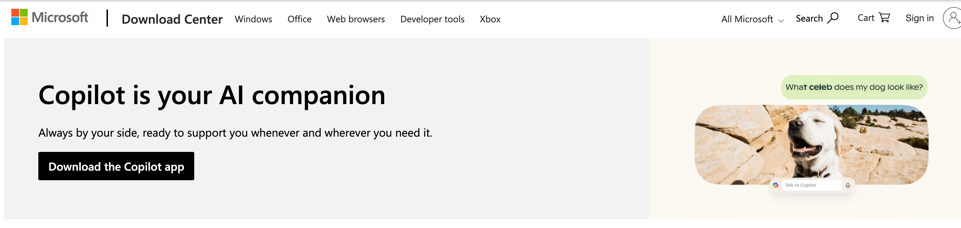
מרכז ההורדות של Microsoft הוא מקור רשמי להורדת תוכנות, עדכונים וכלים שונים של Microsoft, כגון Windows, Office, Visual Studio ועוד. באמצעותו, ניתן להוריד גרסאות עדכניות של תוכנות, חבילות שירות ועדכוני אבטחה.

האתר מתעדכן באופן קבוע, עם הוספת גרסאות חדשות, עדכונים ותיקונים בהתאם לצורך ולמוצרים החדשים של Microsoft. תדירות העדכונים משתנה בהתאם למוצר ולגרסה, אך Microsoft מקפידה לספק עדכונים שוטפים כדי להבטיח את אבטחת וביצועי המוצרים שלה.

חלק ממטרותיו של אתר ההורדות הם : הפצת תוכנות , שיפור אבטחה ועוד .   
הפצת תוכנה רשמית: להציע למשתמשים גישה ישירה להורדה של מערכות הפעלה, חבילות תוכנה, כלים לפיתוח, אפליקציות, ועוד, כולם באישור ובתמיכה של Microsoft.

עדכונים ושדרוגים: מספק עדכונים שוטפים לכל התוכנות המיוצרות על ידי Microsoft, כולל תיקוני אבטחה, שדרוגים לגרסאות חדשות ותיקוני ביצועים.  
שיפור אבטחה: מהבטיח שמשתמשי מיקרוסופט מקבלים את הגרסאות העדכניות ביותר של תוכנות, דבר שמסייע בהגנה על המידע ובשמירה על ביצועי המערכת.  
הדרכות וכלים לפיתוח: להעניק גישה למדריכים, כלים ו-API.

האתר מציע קטגוריות שונות להורדות, וכולל ביניהם:  
Windows: הורדות עבור מערכות ההפעלה Windows, כולל עדכונים וכלי עזר.  
Office: הורדות עבור חבילת Microsoft Office, כולל עדכונים וערכות שפה.  
Developer Tools: כלים למפתחים, כגון Visual Studio.  
**Security** **Updates: עדכוני אבטחה למוצרים שונים של Microsoft.**



איור 23 : אתר ההורדות של מיקרוסופט

[אתר ההורדות של מיקרוסופט](https://www.microsoft.com/en-us/download)

# עיצוב נתונים

השלב של בחירת המודל הסופי בתהליך ניתוח נתונים הוא קריטי להצלחת המודל וליכולת להפיק חיזויים מדויקים ואמינים מהנתונים. כדי להשיג תוצאות מיטביות. בשלב זה נעבוד בשיטתיות על מנת להבטיח מודל חזק ואמין , עם זאת חשוב לציין כי לכל מודל יש חסרונות משלו וננסה לצמצם את הפערים של המודל הסופי שנבחר.

**בחירת המודל הסופי**

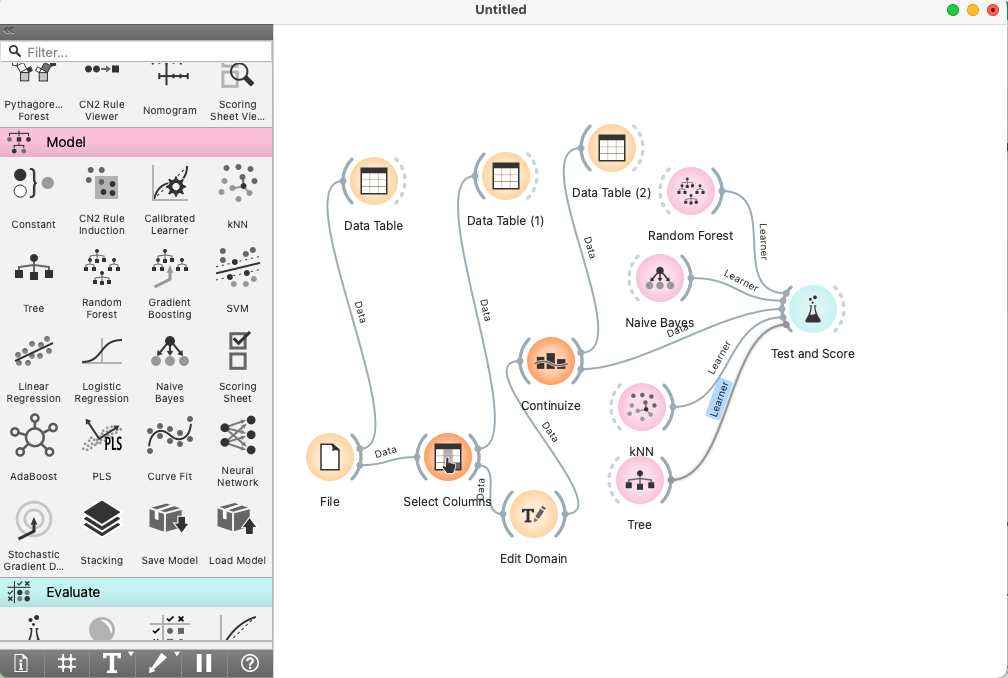
בסופו של התהליך, נבחר את המודל בעל הביצועים הטובים ביותר, שיספק את החיזוי המדויק והאמין ביותר עבור הנתונים שלנו. התהליך כולל השוואה של מודלים שונים ובחירת המודל שמספק את התוצאות הטובות ביותר בהתאם למטרות הניתוח.

**איך נבחר את המודל הסופי?**

נבחר את המודל הסופי על ידי הרצת מגוון מודלים באמצעות תוכנת **Data Mining**, לאחר שביצענו הכנה מקיפה של הנתונים, כולל יצירת משתני דמי ובחירת השורות הרלוונטיות לניתוח. לאחר הפעלת המודלים, נבחן את ביצועיהם בהתאם לקריטריונים שנקבעו, ונבחר את המודל שמציג את הביצועים הטובים ביותר והחיזויים המדויקים ביותר עבור הנתונים שלנו.

שיטת העבודה לבחירת המודל הטוב ביותר הוא הרצת מגוון מודלים שונים והערכת ביצועים שלהם.

הבחירה במודל הסופי היא שלב מרכזי בתהליך, נבצע אותה בצורה מסודרת על מנת להבטיח שהמודל יפיק את התוצאות המדויקות והאמינות ביותר. הרצת מגוון מודלים, בדיקת ביצועיהם ושיפורם במידת הצורך, ולאחר מכן בחירת המודל שמספק את הביצועים הטובים ביותר, מאפשרת להפיק את התועלת המקסימלית מהנתונים.



איור 24 : תהליך מלא בorange data mining להערכת מודלים

לאחר מילוי הנתונים המקיף שהצגנו בשלבים הקודמים, כעת נרצה למצוא את המודל היעיל והטוב ביותר מתוך המודלים הקיימים , ביניהם רשת נוירונים , עץ החלטה , נאיב בייס ועוד, התמונה מתארת איך אנו יכולים ליצור באמצעות תוכנת Data mining את התהליך החל מהעלאת הנתונים ועוד מדידת ביצועי המודל. התמונה היא הממשק המלא של Orange data mining, נפרט את שלבי התהליך של התוכנה:

* 1. File – העלאת קובץ האקסל לאחר עיבוד ומילוי הערכים החסרים.
  2. Select columns – בחירת העמודות הרצויות ובחירת משתנה המטרה Severity,

הפיצ׳רים הם : Affected Product , Affected component , Title , Supersedes , CVEs , Reboot , Impact , Impact.1 , Component KB , Bulletin KB , Severity.1 , Year , Month.

* 1. Edit Domain – שינוי סוג הפיצ׳רים, בנתונים שלנו ישנם ערכים שאומנם מיוצגים כמספרים נצטרך להתייחס אליהם כמשתנים קטגוריאליים.
  2. Continuize – קידוד הערכים לפי משתנה דמי.
  3. Test and Score Running – הערכת המודלים לפי הנתוני דמי.
  4. Test and Score – חישוב מגוון מדדים על מודלים שונים.

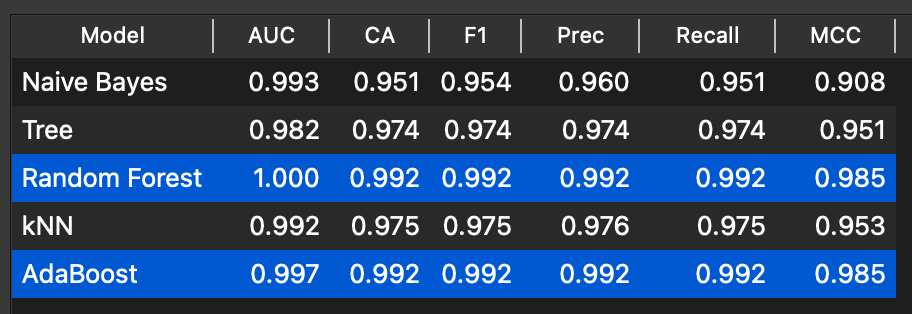
שלבי ביניים:

Data table – הצגה של המשתנים כטבלה של data frame , למשל data table (2) מציג את משתני הדמי של התכונות.

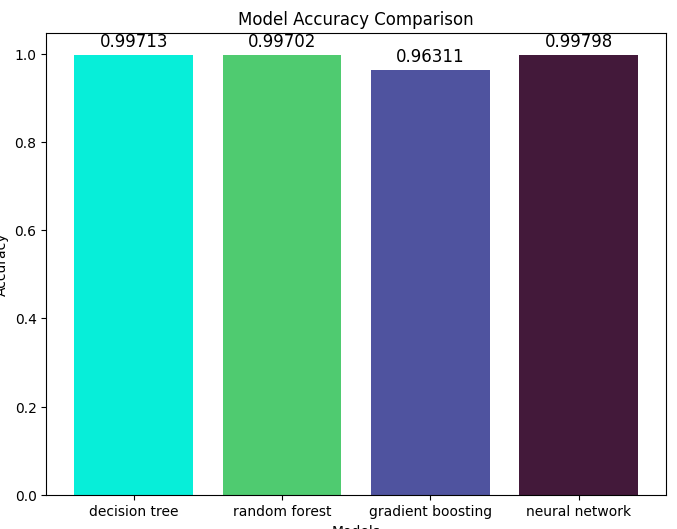
Models – הרצת מודלים שונים לחישוב המדדים שלהם לבחירת מיטבית של המודל הטוב ביותר.

לאחר חישוב המדדים של המודלים שהרצנו קיבלנו שהמודל המועדף ביותר עליו נתמקד:

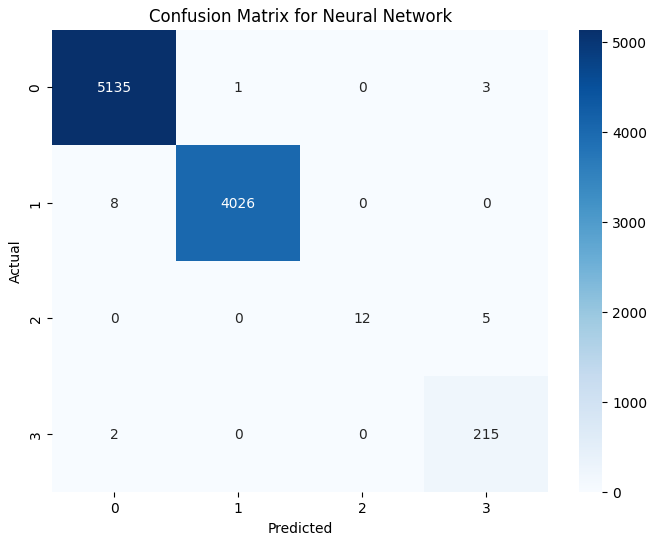
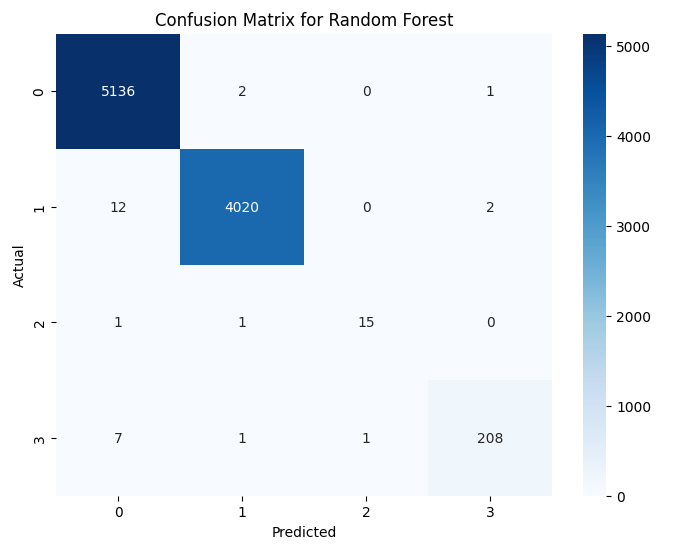
איור 25 : מגוון המדדים של המודלים



בנוסף מדדי דיוק של מודלים שונים :

****

איור 26 : השוואת דיוק המודלים

מודל רשת הנוירונים עם כ-50 שכבות מוסתרות ופונקצית הפעלה של relu מציג יכולות טובות .

איור 27 : מטריצת בלבול לRandom forest

איור 28 : מטריצת בלבול לרשת נוירונים

מטריצות הבלבול הללו מציגות את יעילות שני המודלים , משמאל עץ החלטה ומימין מודל Random Forest.

מודל רשת הנוירונים מציג יכולת טובה עם חיזוי של 9389ֿֿ תחזיות נכונות , לעומת מודל Random forest שמציג 9379 תחזיות נכונות.

## הסבר על המודלים השונים

Random Forest – מבוסס על שילוב של מספר עצי החלטה, כאשר כל עץ מאומן על דגימה אקראית מהנתונים. מפחית overfitting ומשפר דיוק על ידי הצבעה משותפת של כל העצים.

Naïve Baye – מודל הסתברותי שמבוסס על משפט בייס ומניח כי כל המשתנים בתכונות הם בלתי תלויים. מהיר ופשוט אך מניח הנחות שעלולות לא תמיד להתקיים.|  
  
AdaBoost – מודל Boosting המחזק מודלים חלשים (לרוב עצי החלטה קטנים) על ידי שינוי משקלים בדוגמאות הקשות לחיזוי. משפר דיוק אך רגיש לרעש וערכים חריגים.  
לאחר בנייה של מודלים ״מוחלשים״, כל המודלים שנבנו משוקללים בהתאם לביצועים שלהם, כך שלמודלים מדויקים יותר יש השפעה רבה יותר על התחזית הסופית. בצורה זו, AdaBoost משפר את הדיוק על ידי התמקדות בדוגמאות הקשות.

KNN – מסווג דגימה לפי הקטגוריה של ה-K השכנים הקרובים ביותר על בסיס מרחק. קל ליישום אך דורש חישובים רבים לכן עלול להיות איטי.

Decision Tree – מודל היררכי שמחלק את הנתונים לפי תנאים על התכונות עד להגעה לתוצאה סופית. קל להבנה ופרשנות, אך עלול לסבול מ-overfitting ללא הגבלת עומק.

GBM – מודל Boosting מתקדם שבונה סדרת של עצים כך שכל עץ חדש מתקן את הטעויות של העצים הקודמים. נותן ביצועים טובים אך דורש זמן חישוב ארוך ורגיש לפרמטרים.

Neural Network – מודל מבוסס על שכבות של נוירונים המחוברים ביניהם. מתאים למשימות מורכבות כמו עיבוד. יכול להשיג תוצאות מעולות אך דורש חישובים גדולים ומשך אימון ארוך ביחס למודלים קודמים.

במקרה של משתנים קטגוריאליים עם משתני דמי, **Random Forest** ו-**Gradient Boosting** הם לרוב הבחירות הטובות ביותר, שכן הם יודעים להתמודד היטב עם נתונים כאלה ולספק חיזוי יציב ומדויק.  
 **Naïve Bayes** מתאים בעיקר לבעיות שבהן המשתנים אכן בלתי תלויים, אך פחות מומלץ אם יש תלות ביניהם.   
**AdaBoost** יכול לשפר את דיוק החיזוי על ידי מתן דגש לדוגמאות שקשות לסיווג, אך רגיש לרעש.  
 **KNN** פשוט ליישום אך פחות יעיל כאשר יש משתנים רבים. **Decision Tree** קל לפרשנות אך עלול לסבול מ - overfitting.  
 **Neural Networks** מתאימות יותר למשימות מורכבות כמו תמונה וטקסט.

## הסברי מדדים

**1. AUC) Area Under the Curve)**

AUC מודד את השטח מתחת לעקומת ה-ROC, ומייצג את יכולת המודל להבדיל בין קטגוריות שונות. ערך AUC הקרוב ל-1 מציין יכולת הבחנה מצוינת, בעוד שערך AUC הקרוב ל-0.5 מציין יכולת הבחנה מקרית, ורוב המודלים נחשבים לאי-מועילים במקרה זה.

**2. CA) Classification Accuracy)**

CA הוא אחוז התחזיות הנכונות מתוך כלל התחזיות שהמודל ביצע. זהו מדד פשוט, אך לא תמיד משקף את הביצועים האמיתיים, במיוחד כשיש חוסר איזון בין הקטגוריות. לדוגמה, במצבים של חוסר איזון חריף, מדד זה יכול להטעות ולהציג דיוק גבוה למרות ביצועים גרועים במצבים מסוימים.

**3. F1-Score**

F1-Score הוא ממוצע הרמוני בין הדיוק (Precision) לרגישות (Recall). הוא משלב את שני המדדים הללו למספר יחיד. F1-Score שימושי במיוחד כשיש חוסר איזון בין הקטגוריות, שכן הוא מוודא שהמודל לא רק מדויק, אלא גם מזהה היטב את כל הקטגוריות. ערך F1-Score גבוה מציין שילוב טוב של Precision ו-Recall.

**4. Precision (דיוק)**

Precision מודד את אחוז התחזיות החיוביות שהמודל סימן כחיוביות, שמאמתות כאמיתיות. אם המודל עושה הרבה טעויות מסוג חיוביים כוזבים (False Positives), Precision יהיה נמוך. מדד זה חשוב במיוחד במקרים שבהם חיוביים כוזבים יכולים להיות מזיקים או לא רצויים, כמו במערכת רפואית שבה המודל מסמן הרבה חולים כלא חולים.

**5. Recall (רגישות)**

Recall מודד את אחוז המקרים האמיתיים החיוביים שסומנו בצורה נכונה כחיוביים. אם המודל מפספס הרבה חיוביים אמיתיים (False Negatives), ה-Recall יהיה נמוך. זהו מדד קריטי במצבים שבהם חשוב לזהות את כל המקרים החיוביים, אפילו במחיר של טעויות חיוביות כוזבות, כמו במערכות לזיהוי מחלות.

**6. MCC) Matthews Correlation Coefficient)**

MCC הוא מדד שמספק תמונה מאוזנת של ביצועי המודל, גם במקרים של חוסר איזון בין הקטגוריות. הוא לוקח בחשבון את כל ארבעת הערכים של מטריצת הבלבול: True Positives (TP), False Positives (FP), True Negatives (TN), ו-False Negatives (FN). ערכים קרובים ל-1 מעידים על מודל מצוין, ערכים קרובים ל-0 מעידים על מודל גרוע, וערכים שליליים מצביעים על ביצועים גרועים במיוחד.

## מסקנות מדדים

**AUC -** ערך גבוה מצביע על כך שהמודל מסוגל להפריד היטב בין הקטגוריות, כלומר יש לו יכולת טובה לזהות מקרים חיוביים ושליליים ללא תלות בערך הסף שנקבע.

**CA -** ערך גבוה מציין שהמודל מסווג אחוז גדול יותר מהדוגמאות נכון, כלומר מתוך כלל הדוגמאות שסווגו, רובן קיבלו את התווית המתאימה.

**- F1-Score** מדד מאוזן בין דיוק (Precision) ושליפה (Recall), כך שערך גבוה מעיד על כך שהמודל מצליח לזהות נכון חיוביים אמיתיים תוך צמצום השגיאות בזיהוי חיובי כוזב ושלילי כוזב.

**Precision -** ערך גבוה מצביע על כך שכאשר המודל מסווג דוגמה כחיובית, יש סבירות גבוהה יותר שהיא באמת חיובית, כלומר המודל מייצר פחות חיוביים כוזבים (False Positives).

**Recall -** ערך גבוה מעיד על כך שהמודל מזהה את רוב המקרים החיוביים בפועל, כלומר הוא מחמיץ פחות חיוביים אמיתיים (False Negatives).

**MCC -** ערך גבוה מציין שהמודל שומר על איזון טוב בין ארבעת סוגי הסיווגים (True Positives, True Negatives, False Positives, False Negatives), ולכן הוא מדד אמין להערכת איכות המודל, במיוחד אם הנתונים לא מאוזנים.

**סיכום -**

בשלב זה של הפרויקט התמקדנו בבחירת המודל הסופי, זאת לאחר שביצענו הכנה מקיפה של הנתונים והרצת מספר מודלים שונים על הנתונים המעובדים. ההשוואה בין המודלים בוצעה על בסיס מדדים שונים, שנבחרו כדי להעריך את איכות החיזוי והביצועים של כל מודל.  
לאחר חישוב המדדים והשוואת תוצאות המודלים, נמצא כי **Random Forest** הוא המודל היעיל ביותר, שכן הוא השיג את התוצאות הגבוהות ביותר במדדים המרכזיים. מודל זה הציג דיוק גבוה, רגישות גבוהה, איזון טוב בין חיוביים ושליליים, והפחית את מספר השגיאות בזיהוי.

בסיכום, לאחר תהליך בחירה מדויק ומעמיק, החלטנו להתמקד במודל **Random Forest** או **רשת נוירונים ו -**  **Ada Boost** כמועמדים הסופיים לחיזוי הנתונים, בשל הביצועים האופטימליים שלו בכל המדדים שנבחנו ומהירות ההרצה של האלגוריתם שלו- אותם נבחן לעומק בפרק הבא.

מסקנות –  
המודלים המועדפים במקרה זה הם Random Forest ו-Gradient Boosting, שכן הם מציגים חיזוי יציב ומדויק במיוחד עבור נתונים עם משתני דמי. בנוסף רשת נוירונים עם relu וכ-50 שכבות מוסתרות מציגה מדדים די טובים.

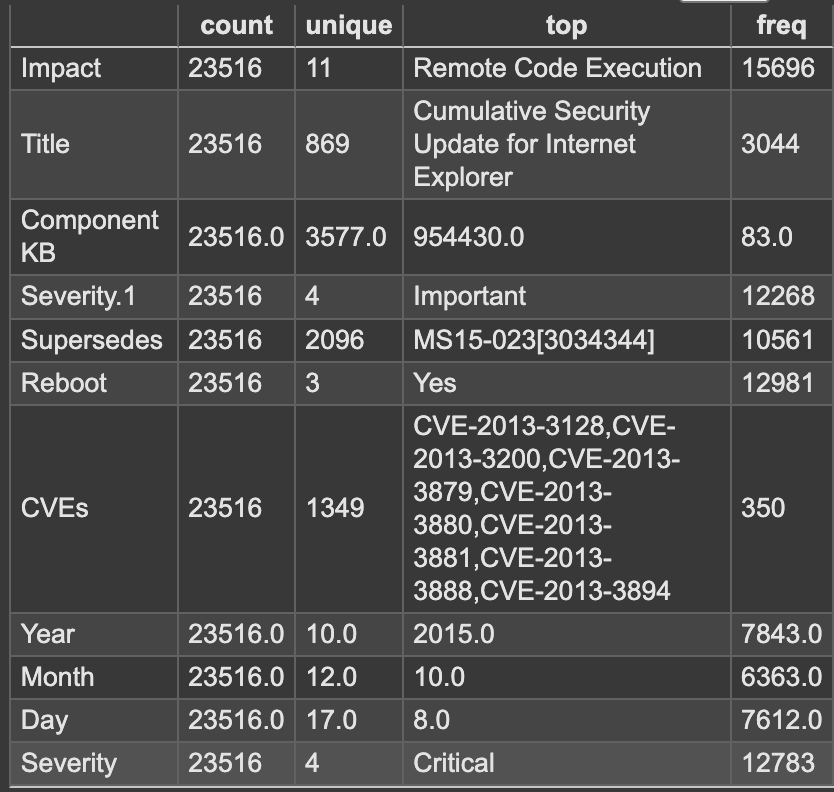
מודלים אחרים כמו Naïve Bayes ו-KNN פחות מתאימים, בשל ההנחות הסטטיסטיות המגבילות או הצורך בעיבוד חישובי כבד.

# 

# ניתוח נתונים (EDA)

בסעיף זה, נבחן את קבוצת הנתונים הסופית באמצעות ניתוח נתונים. הניתוח יסייע לנו להבין את המאפיינים העיקריים של הנתונים ולזהות קשרים פוטנציאליים בין התכונות למטרה.  
הנתונים הסופיים שנבחרו עבור ייעול המודל הם : Date Posted, Impact, Title, Severity.1, Supersedes, Reboot, CVEs, Affected component, Component KB.

משתנה המטרה של המודל הוא : Severity.  
אנו נתמקד במודלים של **Random Forest** , **רשת נוירונים ו -**  **Ada Boost**.



## 

איור 29 : טבלת תדירות וערך מוסף

## טבלאות תדירות

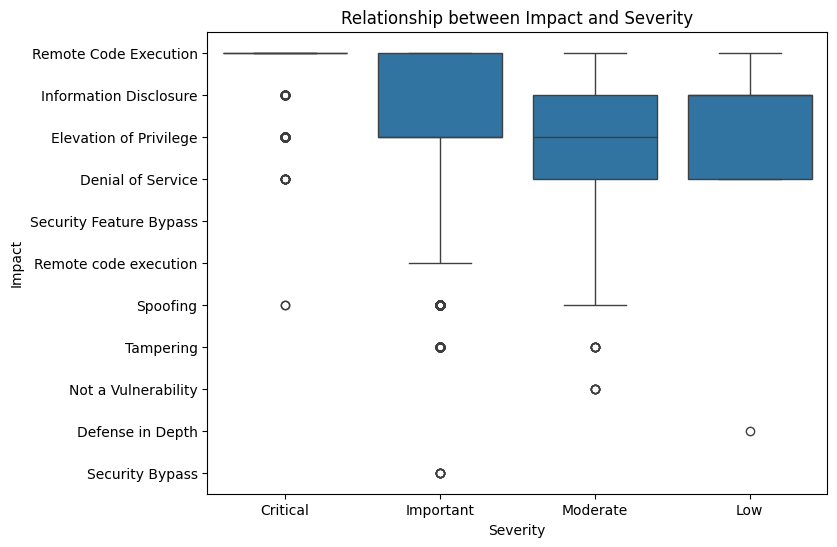
לאחר בחירה קפדנית של הנתונים נציג ערכים על המשתנים שבחרנו למודל :

ניתן להבחין כי הערך השכיח ביותר במידת הפגיעה הראשונית והסופית שונות זו מזו.

**Severity** - רוב העדכונים מתייגים את הבעיות כ"חמורות" (Critical), מה שמצביע על חשיבות גבוהה לטיפול מהיר בבעיות אלו.

**צורך באתחול מחדש (Reboot)** - רוב העדכונים דורשים אתחול מחדש , מה שמצביע על כך שהעדכונים משפיעים על המערכת באופן משמעותי.

**Supersedes** - ישנם 2,096 ערכים ייחודיים בעמודה, מה שמצביע על כך שישנם עדכונים רבים שמחליפים עדכונים קודמים.



איור 30 : דרך Box מתאר את ההשפעה ומידת הנזק

**הסבר הגרף :**

הגרף מציג את היחס בין ההשפעה (impact – ציר y ) לחומרת הפגיעה (Severity – ציר x).

שטח הקופסא של התרשים מבטא את התפלגות 0.5 הנתונים המרכזיים ושטח קופסא קטן מעיד על שונות נמוכה, בנוסף שהקו האנכי בתוך התפלגות 0.5 מהנתונים הוא החציון שמחלק את הנתונים באופן שווה. זרועות הקופסא מייצגות את מרבית הנתונים ללא השפעה על ערכים חריגים ומספקים ערך מוסף על התפלגות הנתונים.  
הערכים הקיצוניים מובאים לידי ביטוי על ידי הנקודות המופרדות מתחת או מעל הדיאגרמה.

תרשים הקופסאות מספק כלי חזותי חזק להבנת התפלגות הנתונים, והזרועות והערכים הקיצוניים מספקים מידע חשוב על הווריאציה והיוצאים מן הכלל בנתונים.

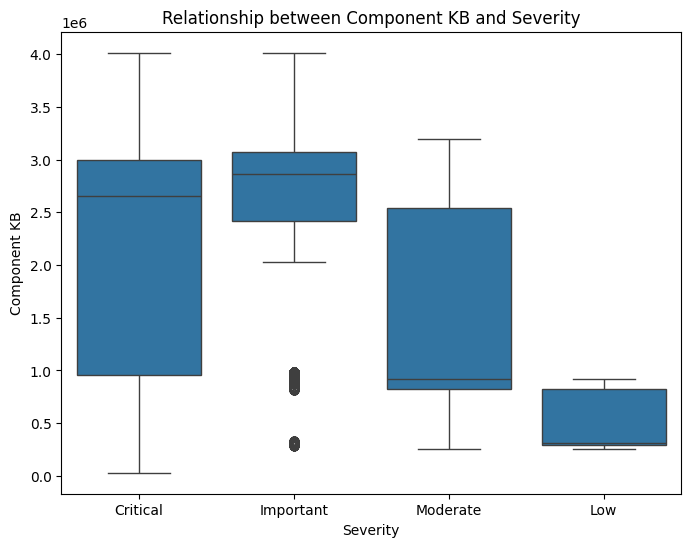
**הקשר בין המשתנים :**

הצגת הקשר בין השפעה לחומרה ע״י הגרף מראה את הקשרים בין סוגי הפגיעות השונות (כמו הרצת קוד מרוחקת, חשיפת מידע, העלאת הרשאות) לבין חומרת הפגיעה. כל סוג פגיעה מוצג בשורה אחת, והקופסאות מראות את התפלגות החומרה של כל סוג פגיעה.

**מסקנות עיקריות :**

הגרף מציג את הקשר בין סוגי פגיעות אבטחה לחומרתן. פגיעות כמו **הרצת קוד מרחוק**  מסווגות כחמורות ביותר באופן עקבי, בעוד שפגיעות כמו **Spoofing** וגם **Tampering** מדורגות לרוב כחומרה נמוכה. שטח הקופסאות הקטן מצביע על שונות נמוכה בתוך כל קטגוריה, כלומר רוב הפגיעות מסווגות באופן יציב ללא חריגות משמעותיות.

תרשים זה מאפשר להבין איזה סוגי פגיעות נחשבים חמורים יותר וכיצד הן מתפלגות בהקשר של חומרתן.ֿ



איור 31 : גרףBox מתאר את הרכיב שהושפע למידת הנזק

ניתן להסיק מספר מסקנות מתוך הגרף שמציג יחס בין מידע על רכיב המושפע ורמת הפגיעות :

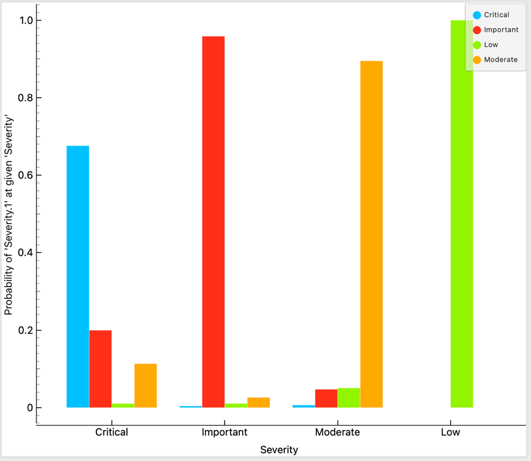
**חומרה קריטית (Critical)** - התפלגות רחבה מאוד, עם טווח גדול של ערכים מהמינימום למקסימום. החציון נמצא באמצע הטווח, מה שמצביע על התפלגות סימטרית יחסית.

ישנם מספר ערכים מופרדים מתחת לקופסה, ניתן להסיק על עדכונים קטנים יותר מאוד בחומרה זו.

**חומרה חשובה (Important) -** החציון נמצא באזור העליון של הקופסה, מצביע על נטייה לערכים גבוהים יותר.ישנם מספר ערכים מופרדים מתחת לקופסה, אך פחות מאשר בחומרה קריטית.

**חומרה מתונה (Moderate) -** החציון נמצא באמצע הטווח, מה שמצביע על התפלגות סימטרית יחסית.ישנם מספר ערכים מופרדים מתחת לקופסה.

**חומרה נמוכה (Low) -** התפלגות צרה מאוד, מה שמצביע על ווריאציה מועטה ( הגיוון נתונים נמוך ) . החציון נמצא באמצע הטווח, והקופסה קטנה יחסית, מה שמצביע על עדכונים קטנים יותר באופן כללי.



איור 32 : גרף המייצג את יחס בין סוג החומרה הראשוני והסופי

גרף העמודות מציג ההסתברות של סוגי החומרה הראשוניים והתפלגותם על סוגי החומרה הסופיים הגרף מאפשר להבין את התפלגות החומרה של הפגיעות ואת ההסתברות של כל רמת חומרה. ציר y מייצג את מייצג הסתברות ראשונית של כל חומרה.

ציר x מחולק לפגיעות ראשוניות וסופיות.

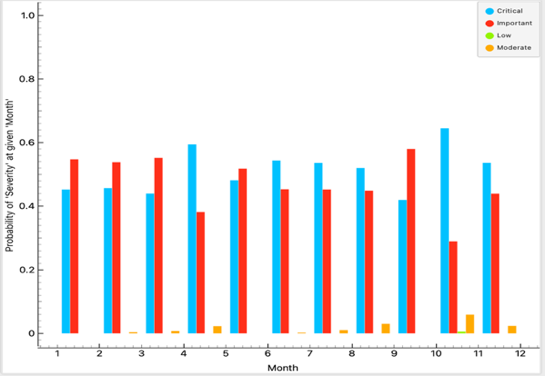
**מסקנות :**

הגרף מראה שרמות החומרה important וModerate הן הנפוצות ביותר בין הפגיעות, כאשר לשתיהן יש הסתברות גבוהה יחסית.

רמת החומרה Critical מופיעה פחות, אך עדיין בהסתברות משמעותית.

רמת החומרה low היא הנפוצה ביותר בקרב הפגיעות המדווחות, כפי שניתן לראות מהעמודה הירוקה הגבוהה.

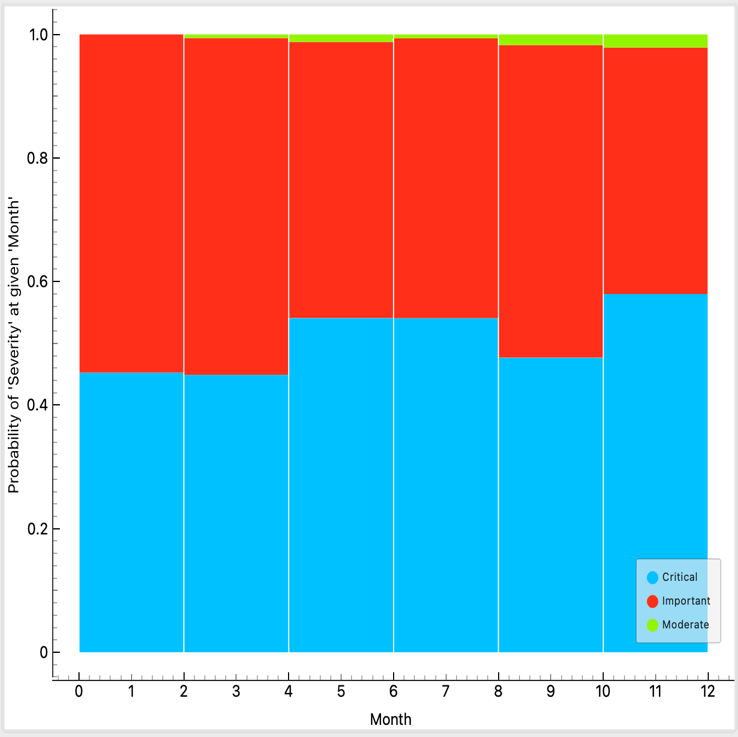
הגרף מספק תמונה חזותית של התפלגות החומרה של הפגיעות, ויכול לסייע בקבלת החלטות בנוגע לאבטחה.



איור 33 : גרף המציג את מידת הנזק לפי חודשים

גרף עמודות מציג את ההסתברות של רמות חומרה שונות של פגיעות או אירועים לאורך תקופה של 12 חודשים. הגרף מאפשר לראות איך התפלגות החומרה של הפגיעות משתנה לאורך הזמן.

גרף העמודות המוערם נותן ערך מוסף ומציג את החלק מהשלום של חומרת הפגיעה לפי החודשים.



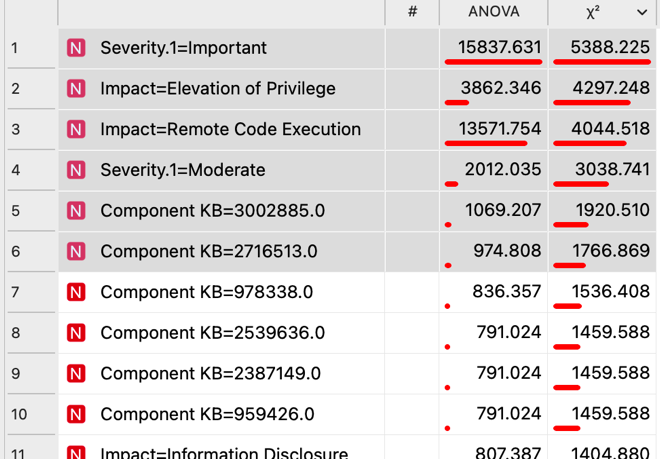
**מסקנות :**

ניתן לזהות עוצמות תקיפה שונות לאורך זמן, דרך הגרף אפשר להבחין ולהבין אם האסטרטגיות הנוכחיות יעילות או זקוקות לשינוי. בנוסף ניתן לראות שלאורך כל החודשים ישנם תקיפות ברמות גבוהות, חודשים עם הסתברות גבוהה של פגיעות קריטיות או חשובות עשויים להצביע על נקודות ביקורת שבהן יש צורך בהתערבות או בחיזוק אמצעי ההגנה.

מירב התקיפות הקריטיות מתרחשות בחודשים 4, 6, 7, 10, 11 ניתן להסיק מכך שנוכל למקד את מירב ההגנה מפני התקפות קשות בחודשים האלו ויכול לסייע בתכנון משאבים.

יתר על כן דרך הגרף המוערם ניתן להסיק כי החודשים שצוינו מתרחשים בהם מרבית התקיפות הקריטיות. **החברה מיקרוסופט מתמודדות עם מתקפות בחומרות גבוהות לאורך כל השנה.**

איור 34 : גרף מוערם של מידת הנזק על פני שנה



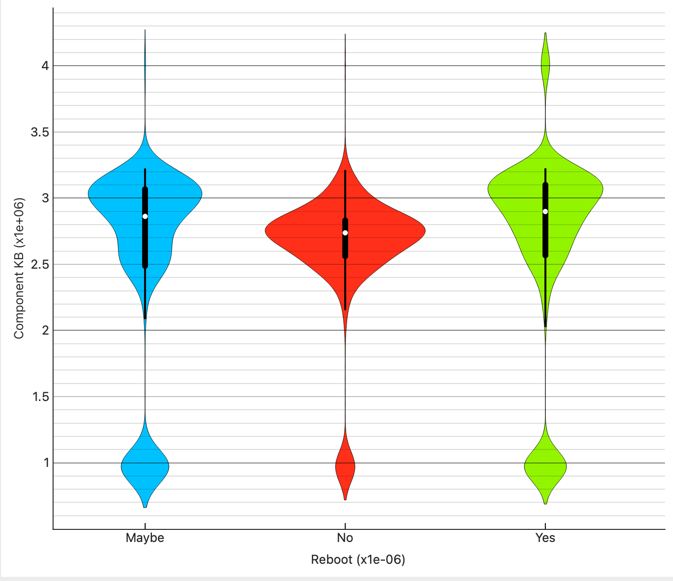
ANOVA: ANOVA מייצגת ניתוח שונות. זהו מבחן סטטיסטי המשמש להשוואת הממוצעים של שתי קבוצות או יותר. בדרך כלל, ערך גבוה יותר של ANOVA מצביע על הבדל משמעותי יותר בין הקבוצות שנבדקו.

X^2: מייצג סטטיסטיקה של Chi-squared. מבחן Chi-squared משמש בדרך כלל כדי לקבוע אם יש קשר משמעותי בין שני משתנים קטגוריים. כמו ANOVA, ערך גבוה יותר עשוי לרמז על קשר משמעותי יותר.

שורות: כל שורה מייצגת סוג מסוים של פגיעות או בעיה, יחד עם הערכים הסטטיסטיים שלה עבור ANOVA ו-X^2.

איור 35 : תצוגה של מבחנים סטטיסטים על משתנה המטרה

לסיכום, הטבלה מספקת השוואה כמותית של קטגוריות באמצעות מבחנים סטטיסטיים (ANOVA ו-Chi-squared). הערכים הללו יכולים לעזור בתיעדוף ולהבין את המשמעות של בעיות שונות



גרף הכינור מציג את התפלגות רכיב KB בקטגוריות שונות של אתחול מחדש: 'אולי', 'לא' ו'כן'.

הגרף מספק יותר פרטים על התפלגות הנתונים, במיוחד צפיפות ההסתברות של הנתונים בנקודות שונות.

איור 36 : גרף כינור המייצג את התפלגות הרכיב על פני טגוריות אתחול מחדש

**ציר ה-x:** ציר ה-x מייצג את הקטגוריה 'אתחול מחדש', המחולקת לשלוש קבוצות: 'אולי', 'לא' ו'כן'.

**ציר ה-y:** ציר ה-y מייצג את הערך של רכיב KB (בסולם של 1e+06).

**צורות הכינור:** לכל קטגוריה של אתחול מחדש, יש צורת כינור המייצגת את התפלגות רכיב KB עבור קטגוריה זו**. הרוחב** של צורת הכינור מציין את צפיפות הנתונים בערך זה, כאשר חלקים רחבים יותר מצביעים על סבירות גבוהה יותר של מציאת נקודת נתונים בערך זה.

נקודות על צורת הכינור מייצגות מדדים סטטיסטים שונים ,הנקודה הלבנה בתוך כל צורת כינור מייצגת את החציון, סרגל האמצעי השחור מייצג כנראה את הטווח הבין-רבעוני, הוא מראה את התפשטות האמצעית של 50% מהנתונים עבור כל קטגוריה של Reboot.

**מסקנות :**

עבור Reboot הקטגוריה 'Maybe' יש התפלגות רחבה יותר לרכיב KB, מה שמצביע על מגוון גדול יותר של ערכי KB בקרב דוגמאות אלה.

החציון של רכיב KB, שמיוצג על ידי הנקודות הלבנות, נראה דומה באופן יחסי בכל שלוש הקטגוריות.

התפלגויות שונות מעידות על כך שההחלטה לאתחל מחדש עשויה להיות קשורה בערכי רכיבי KB שונים, אך יש צורך בניתוח סטטיסטי נוסף כדי לקבוע אם הבדלים אלה משמעותיים.

הגרף הבא מייצא את המילים הנפוצות ביותר של הרכיבים שהושפעו במהפך התקיפות.

איור 37 : גרף ענן שמראה את השכיחות של הרכיבים שהושפעו

בין הרכיבים נמצא את windows , מערכת ההפעלה ואתרי אינטרנט שונים .