



# Microsoft Security Bulletin Analysis

דוח הבנת הנטוונים מוצג על ידי בר כהן וסחר חיים יעקב.

מטרת דוח "הבנת הנטוונים" היא לנתח, להעריך ולהציג את יכולות הנטוונים שנאספו, לזהות בעיות אפשריות כמו ערבים חסריים או לא מדויקים, ולהבין את המבנה, הדפוסים והਮוגמות שבнетוונים.



שם מרצה : מר אבי זכאי.

שם מנהה : מר חנן לב.

מגייסים :

בר כהן 208110254  
סחר חיים יעקב 314741851



## תוכן עניינים

- 1 - ..... **דוח הבנת נתוניים**
- 3 - ..... **איסוף נתונים**
- 4 - ..... **מקורות הנתוניים**
- 8 - ..... **בדיקות נתונים ראשונית**
- 13 - ..... **תיאור נתונים**
- 15 - ..... **כמויות הנתוניים**
- 17 - ..... **סוגי ערבים**
- 19 - ..... **ערבות קידוד**
- 24 - ..... **חקירתי נתונים**
- 28 - ..... **aicoot הנתוניים**





## איסוף נתונים

איסוף נתונים הוא שלב קריטי בתהליך של ניתוח נתונים וביוזע פרויקטי Data Science. אינטואיטיביות הנתונים, האמינות שלהם קובעת את התובנות וההחלפות שיתקבלו בהמשך. נתונים מדויקים ומלאים מאפשרים ייצור תמונה נconaה ומעמיקה של בעיות וצרכים, ומספקים את הבסיס לתובנות משמעותיות. תהליך זה לא מתייחס רק למידעväם, אלא גם למידעväם אינטואיטיבי, כמו נתונים טקסטואליים או תיאורים. כל פרט כזה יכול להוסיף ערך למודלים המבוססים על למידת מכונה או תיאורים. כמו כן, ניתן לשלב תמונה מלאה ומעמיקה יותר.

הנתונים משמשים לא רק להנחות את מקבלי החלטות, אלא גם לבניית מודלים ואלגוריתמים חכמים. במודלים של למידת מכונה, הנתונים הם הכוח המניע מאחורי הלמידה ולהסיק מסקנות. נתונים אינטואיטיביים מאפשרים להצליח ולספק תוצאות מדויקות. באותו הזמן, אינטואיטיביות הנתונים היא קריטית לשיפור ביצועים, שכן נתונים לא מדויקים או חסרים עשויים להוביל לתוצאות מוטעות. תהליכי ניקוי הנתונים, כולל את הסרת ערכים חסרים או כפולים, הוא קריטי להבטחת אמינותם של הנתונים וליצירת מודלים שימושיים וтоוצואות אמינות ומדויקות.

באמצעות הנתונים, ניתן לשפר תהליכי ולבצע אופטימיזציה מתמשכת, להעריך את הצלחה של פעולות ולבצע שיפורים. איסוף נתונים מאפשר גם לעמود בדרישות רגולציה ולודא תאימות לחוקים, במיוחד בנוגע לפרטיות המידע. התמודדות עם אתגרים טכנולוגיים, כמו חיבור למאגרי נתונים שונים או ניתוח נתונים בזמן אמת, יכולה להציג טכנולוגיות מתקדמות ופתרונות ענן ניהול נתונים גדולות. בסופה של דבר, איסוף נתונים אינטואיטיביים משפר את יכולת התמודד עם אתגרים ומספק כלים חשובים להצלחת הפרויקט.





## מקורות הנתונים

הנתונים שהורדו הם חלק מפרויקט "חיזוי אירועי אבטחה של מיקרוסופט" (Microsoft Security Prediction Events), מספק תצפיות אמיטיות על אירועי סייבר.

מאגר מרכז ההודאות של חברת מיקרוסופט הוא מאגר ציבורי גדול, ומכיל נתונים חשובים שנitin להשתמש בהם לצורכי מחקר בתחום אבטחת המידע. הנתונים מספקים מידע על סוגי התקפות, דפוסים של איומים, ואופי המתקפות שנצפו במערכות מידע, הערכת חומרות ועוד.

### מטרת הנתונים:

מטרת הנתונים היא לאתגר את קהילת מדעי הנתונים לפתח דרכים לחזות את רמת המורכבות של אירועי סייבר עתידיים. הנתונים נועדו לשפר את המודלים לחיזוי תקירות סייבר ולעוזר בפיתוח מערכות תגובה שמשיעות למועד אבטחת מידע להגביל לאירועים בצורה אוטומטית או מונחית.

### מאפיינים עיקריים:

הנתונים המפורטים כוללים מידע מפורט על פגיעות אבטחת מידע (CVE), תאריכי פרסום, דרגות חומרה והשפעה, וכן מידע על המוצרים והמערכות המושפעים. בנוסף, הנתונים מספקים עדכנים על תיקוני אבטחה ושיפורים טכניים במערכות ובמוצרים של Microsoft. מטרת הנתונים היא לתרום לחיזוי תקיפות סייבר, לשפר את ההגנה מפניהם, ולסייע במקב אחר פגיעות אבטחה תוך מתן הנחיות ברורות ליישום תיקונים.

### זמן יצירת הנתונים וטווח זמני:

הנתונים נאספו בין השנים 2008 ל-2017, כאשר פורסמו ב-2017. תאריכי העדכנים מוצגים בטבלה בפורמט של ערכיים מופרדים בנקודה ובמקף, ומיצגים את הזמן המדויק שבו זוהתה הפגיעה או פורסם עדכון האבטחה.



### **מטרות השימוש בנתונים:**

לפתח ולבחון מודלים של מידת מכונה **לחיזוי רמותTKRIOT סיבר ופריצות**. לשמש כבסיס למערכות **הכוונה ותגובה מונחות** עבור SOCs. לאתגר את קהילת המחקר בתחום אבטחת המידע לפתח פתרונות שיכולים לסייע בתגובה מהירה **זיהוי פגיעות החומרה** של המערכת וצריך בעדכוניים נוספים. לספק מודד סטנדרטי לשיפור מערכת **התגובה המונחת** (GR) לפי רמה ולבחון את יכולות ההסתמודות של מערכות אלו בرمות שונות.

### **פלטי המהוזלה:**

גרסת הנתונים שפורסמה היא גרסה 1.0, ותאריך הפרסום שלה הוא **14 במרץ 2017**. הנתונים זמינים להורדה תחת רישיון **CDLA-Permissive-2.0**. עם דגש על אבטחת מידע מוקפת: מזהים אישיים עוברים גיבוב ומוחלפים למזהים אקרים. המידע נועד לתמוך בקהילת המחקר בתחום אבטחת המידע, תוך קידום פיתוח מערכות חיזוי ותגובה אוטומטיות לאירועי הסיבר.

### **מרכיבי הקבצים:**

1. **2MB BulletinSearch.xlsx** - מכיל מידע על עדכוני אבטחה, תוכנות מושפעות, דרישות הפעלה מחדש, זיהויי פגיעות (CVE). המידע בקובץ זה מכסה עדכוניים מנובמבר 2008 ועד למועד הפרסום האחרון שהוא ב 14/03/2017 ( 19,066 תצלפיות ).
2. **505.7KB Bulletin Search 2001-2008.xlsx** - כולל נתונים עבר משנים 2001-2008 ( 4,451 תצלפיות ).
3. **1.8MB MSRC-CVRF.zip** - אוסף של עדכוני אבטחה בפורמט CVRF החל מיוני 2012.
4. **19.2KB CVRF Information.docx** - מסמך עם מידע נוסף על פורמט CVRF.

**קישור לדאטה - קישור להורדת הנתונים**



**סיכום:** הנתונים שנמסרו ממיקרוסופט נתונים הזדמנות ייחודית לפתח ולבחנו מודלים לחיזוי ומענה לرمות אירועי סייבר, עדכוני מערכת וחומרה תוך שימוש נתונים מהמציאות, ויכולים לשמש בסיס לפיתוח מערכות אוטומטיות או מונחות בתחום אבטחת המידע.

#### **סיכום מבנה הדעתהסט: (סיכום טפני)**

ראיות (Evidence) : הרמה הבסיסית ביותר בנתונים. כל ראייה תומכת בהתראה ויכולת לכלול פרטי מידע כמו אתחול , תוכינה מחליפה , 2 סוגים השפעות , חומרת פגיעה ועוד ...  
התרעות (Alert) : ארגזיה (תהליך של חיבור או סיכון נתונים מספר מקורות או ערכיים, במטרה להפיק תוצאה כוללת או תובנה מדמיינית) של מספר ראיות שמצביעות על איום או תקלת אבטחה פוטנציאלית.

#### **תכונות עיקריות:**

טריאז' (קביעת סדר העדיפויות) של תקלות : הדעתהסט בניו כדי לחזות את דרגת החומרה של התקלות.  
למידת מכונה : הדעתהסט יכול להוות יעד לחיזוי משתנה מטרה של רמת החומרה או חומרה .

#### **שימושים פוטנציאליים:**

הදעתהסט מאפשר לפתח מערכות תגובה מונחות שיכולות לסייע לצוותי ה- SOC (Security Operations Centers) בקבלת החלטות מבוססות נתונים, ולהציג את תהליכי האוטומציה בمعנה לאירועי סייבר. באמצעות המידע שבדאטה, ניתן לפתח מודלים לחיזוי רמות חומרת התקלה והתקפות, ולספק המלצות לפעולות תיקון ושיקום. המודלים הללו יכולים לשפר את יכולת ההתרמודדות עם תקלות, ולהתאים את פעולות השיקום לפי הסיכון והחומרה של כל אירוע אבטחה, כך שתהליך הזיהוי והtagובה יהיה מדויק ומהיר יותר.



### מצדי הערכה:

הדאטהסט כולל כ – 23,517 תצפיות וכ – 14 מאפיינים, והוא לא מחולק לשט אימון ושט בדיקה. המדד העיקרי להערכתה של מחקרים שמתבצעים עם הדאטהסט הוא F1 score, הכולל מדרדים נוספים כמו precision ו-recall, שנמצאים בשימוש לצורך הערכת הביצועים של המודלים המתמודדים עם הזיהוי והחיזוי של אירועי סייבר.

### רישוי:

רישוי : הדאטהסט זמין תחת הסכם רישוי Community Data License Agreement – Permissive – Version 2.0 (CDLA-Permissive-2.0) ומאפשר שימוש חופשי למטרות מחקר ופיתוח, תוך שמירה על פרטיות המידע.

### סיכום:

הדאטהסט כולל תצפיות ומחלקה לראיות, התרעות ותקלות, שמייעות בזיהוי תקלות ואירועים אבטחתיים. הוא מאפשר חיזוי דרגות חומרה ואוטומציה בתגובה לאיירוע סייבר, ומוצע לשימוש במודלים של למידת מכונה. המדד העיקרי להערכת המודלים הוא F1 score, הכולל מדרדים נוספים כמו precision ו-recall. הדאטהסט זמין לשימוש תחת רישוי Community Data License Agreement – Permissive – Version 2.0.





## בדיקות נתוניים ראשונית

נתמקד בתוכנות שיכולות לספק תובנות שימושיות עבור זיהוי מתקפות סייבר או שיפור האבטחה. התוכנות המבטיחות ביותר יהיו אלה שמספקות מידע ישיר ומדויק על המתקפה, המכשירים המעורבים, והפעולות שהתרחשו.

### תכונות לlionetti:

#### Bulletin KB .1

- **למה זה חשוב?** זהו המדריך שפורסם ב-Knowledge Base (KB)-בו נמצאות פרטי הפתרון לבעה. מדריך זה חשוב לצורך הבנת התהליכים הפתרוניים שננקטו.

#### Severity .2

- **למה זה חשוב?** דרגת חמורת הבעה, שנوعדה לציין עד כמה היא משפיעה על המערכת. דרגת חמורה זו עוזרת לקבוע את סדר העדיפויות לתקן.

#### Impact .3

- **למה זה חשוב?** זהו תיאור השפעת הבעה על המערכת. מידע זה חיוני להערכת גודל הבעה ולהבנה של מה בדיקונ פגעה

#### Affected Product .4

- **למה זה חשוב?** המוצר המושפע מתיאור הבעה. מידע זה חשוב כדי לזהות איזה רכיב במערכת נדרש לעדכון או לתקן.

#### Affected Component .5

- **למה זה חשוב?** רכיב המערכת המושפע. מידע זה מאפשר לך להבין מה בדיקון במערכת דורש תשומת לב ושדרוג.

#### Impact2 .6

- **למה זה חשוב?** השפעה נוספת של הבעה, שעשויה לשפוך אור על תסמים או בעיות נוספות שיכולות להתרחש בעקבות הבעה הראשית.

#### Severity.1 .7

- **למה זה חשוב?** דרגת חמורה נוספת, שיכולה לעזור להבין את ההשפעה של הבעה על המערכת בסקללה נוספת או אחרת.



## Supersedes .8

### • **למה זה חשוב ?**

בעיה שנכassa. זה מציין אם בעיה ישנה כווסתה בעדכון הנוכחי, מה שעזר למנוע כפיפות ולסייע בשיפור חווית המשתמש.

## Reboot .9

### • **למה זה חשוב ?**

מצין אם נדרש לבצע פעולה מחדש של המערכת לאחר העדכון. מידע זה חשוב למי שביצע את העדכון על מנת למנוע תקלות.

## CVEs .10

### • **למה זה חשוב ?**

מספק את מזהה ה CVE (Common Vulnerabilities and Exposures) -שימושים לצורך זיהוי בעיות אבטחה מסוימות. המידע הזה חשוב להבנה אם הבעיה קשורה לפרצות אבטחה ידועות.

### תבונות לא רלוונטיות (או לפחות עלולות להיות לא רלוונטיות):

## Date Posted .1

### • **הסביר :**

זהו תאריך פרסום העדכון, שיכול לשמש כمدד בזמן שבו נחשפה הבעיה. מידע זה חשוב כדי לזהות עדכונים בזמן אמיתי ולהבין את סדר העדיפויות של תיקון בעיות.

## Bulletin Id .2

### • **הסביר :**

מדובר בمزאה ייחודי לכל תצפית. מזהה זה מאפשר לך לאתר ולזהות עדכונים באופן מדויק, ולהתחקות אחריו שינויים או תיקונים שנעשו.

## Component KB .3

### • **הסביר :**

מדובר במדריך ידע שմסביר על הרכיב המושפע. מדריך זה מסייע למשתמשים להבין את הדרך בה יש לטפל בעיה ברמת הרכיב.

## Title .4

### • **הסביר :**

מדובר בכותרת המתארת את סדרת עדכוני האבטחה (Security Updates) לתוכנות או רכיבים שונים.

- ישנו מספיק נתונים כדי להסיק מסקנות הניתנות להכללה ולביצוע תחזיות מדויקות, אך יש לקחת בחשבון את האיכות והמגון של הנתונים, ואת הצורך בהערכת מקיפה של הנתונים לפני ביצוע התחזיות.



## סיכום全文:

### 1. כמויות נתונים גדולות:

יש כמויות נתונים מספקת כדי להבחן בדפוסים ולהסיק מסקנות על בסיס היסטורי. כמויות נתונים גדולות מספקת עושר של מידע שיכול לעזור למודלים ללמוד ולהקליל בצורה טובה יותר, וכך לבצע תחזיות מדויקות יותר.

### 2. גיוון נתונים:

גיוון הנתונים תורם לזיהוי דפוסים מורכבים ולהבנת תלויות בין תכונות, תוך שיפור אמינות המסקנות והיכולת להתמודד עם שונות רכיבים. הוא מאפשר למודלים חוזיים להיות מדויקים יותר, מסייע באימוט ניתוחים וمبטיח תגובה רחבה לאירועים אבטחת מודיעין מגוונים. בנוסף, גיוון הנתונים מספק בסיס לזיהוי ותיקון בעיות בעלות השפעה רחבה, תוך שיפור הגנת המערכת והתאמתו לתרחישים רבים, כולל תרחישי קצה.

### 3. דרישות ניקוי נתונים:

לפני ביצוע תחזיות הדיקוק, נבצע ניקוי נתונים או נתמודד עם מילוי הערכים על מנת להימנע משגיאות או בעיות של ערכים חסרים, עיוותים או חוסר עקביות נתונים. ברגע שהנתונים נקיים, המודלים יהיו מוכנים לפתח תחזיות אמינות.

### 4. יכולת לבצע תחזיות מדויקות:

אם הנתונים נלמדים בצורה נכון, ובבסיס המידע מספק, אפשר להשتمש בהם לביצוע **תחזיות מדויקות** על מתקפות סייבר וחוזקה שלחמת בעתיד. עם מודלים מתקדמים של **למידת מכונה** ונתונים אינטלקטואליים, ניתן לחזות סוג מתקפות, עיתויים ותגובה במהירות גבוהה.

### 5. דרישות משאבים:

נתונים בגודל כזהו מושכים **משאבים** **היישובים** נכון, כמו חישוב על גבי ענן או שימוש בטכנולוגיות כמו **Apache Spark** (פלטפורמת עיבוד נתונים מבוירת בזמן אמת). זאת על מנת לנצל את העיבוד והאנליזה בצורה היעילה ביותר.

**לסיכום**, הנתונים מספקים כדי להסיק מסקנות הניתנות להכללה ולבצע תחזיות מדויקות, בתנאי שהם יעמדו ויתנקו נכון. כל תוכנה צריכה להיבחן במתפקות ובמערכות אפשר להפיק תובנות מדויקות ואמינות לצורך זיהוי מתקפות סייבר עתידיות.

בנוסף אין תשובה חד משמעות אם יש יותר מדי תכונות. כל תוכנה צריכה להיבחן לפי חשיבותה ותרומתה למודול ולתוצאה הצפואה. בהינתן כמות גדולה של תכונות, נבצע **צמצום** **תכונות** באמצעות שיטות כמו **PCA**, **Feature Importance**, או כלים אחרים שמתאים למטרות שלנו, וכך נמנע בעיות של **overfitting** ונוכל לשפר את ביצועי המודל.

מעבר לכך מאחר וכל הנתונים מגיעים ממוקור אחד (Microsoft), **לא נדרש מיזוג** **של נתונים** **מקורות** **שוניים**, מה שmpshet את התהילה. במצב זה, הבעיה המרכזית אינה מיזוג נתונים, אלא **בעיקר הכנה וניקוי נתונים**. חשוב לוודא שכל הנתונים



מסודרים בצורה איחודית ונכונה, ולודא שאין בעיות כמו **ערבים חסרים, נתוניים שגויים או כפליות**.

ניתן להסיק ש : **הנתוניים מגיעים ממוקור אחד, ולכן אין צורך במיזוג של מקורות נתונים שונים**, אלא יש להתרכז בהכנה נכונה של הנתוניים על מנת לוודא שהמודל לא יתקל בבעיות של אי-coresight נתונים או בתהליכי אימון שגוי.

כדי להתמודד עם **ערבים חסרים** בנתוניים, יש לגשת לשאלת בצורה מסודרת ולקבוע את השיטה המתאימה לכל תconaה בהתאם לאופייה. כאשר עובדים עם נתונים שגיעים ממוקור אחד, כמו Microsoft, דרך כלל יש צורך בבדיקה ערבים חסרים בתכונות השונות, ולאחר מכן לבחור את הדרך המתאימה לטפל בהם.

בין השיטות הנפוצות לטיפול בערבים חסרים :

Bulletin KB  
Component KB

- **השלמת ערבים מספריים**: תכוונות שמכילות ערבים מספריים כמו , כמו **Bulletkn KB** או **Component KB** , ניתן להשלימו **בממוחע או חציו** , על מנת לשמור על יציבות הנתוניים.
- **השלמת ערבים קטגוריים** : עבור תכוונות קטגוריות, כמו **Impact** או **Severity** , ניתן להשלימו **בערך הנפוץ ביותר** (המוד) על מנת לשמור על עקבות הנתוניים.
- **השלמת ערבים בתוכנות זמן** : בתכוונות שמכילות תאריכים, כמו **Date Posted** , ניתן להשלימו עם הערך הנפוץ ביותר או עם הערך האחרון בתוצאות.

במקרים בהם ערבים חסרים מופיעים בתכוונות קרייטיות, אפשר לבחור **בהתו** **שורות עם ערבים חסרים** אם הן לא מופיעות באופן משמעתי על התוצאות. במקרים חמורים יותר, ניתן להשתמש בשיטות מתקדמות להשלמת נתונים כמו **Multiple Imputation** או **KNN Imputation**.

#### כעת נציג את התכוונות הקיימות בDATAEST עם חלום נယוב:

- מאגר המידע שמייקروسופט פרסמה לאחרונה כולל נתונים מקיפים על עדכוני אבטחה (Microsoft Security Bulletins), המתוירים באמצעות 14 מאפיינים שונים. את אופן ניתוח התכוונות ביצעו עיי שימוש ב – AI, כלומר נתנו למערכת הבינה המלאכותית לקרוא את הנתוניים, לאחר מכן ביקשו שתפרט לנו אודוט השורה הראשונה (שורת שמות העמודות), ככלומר התכוונת. לאחר שפירטה את התכוונת, ביקשו שתכתוב זאת ישירות לטבלת אקסל ייועודית .



- תאריך פרסום העדכון . Date Posted

- מזהה ה - Bulletin Id

- קישור לעדכון עם מזהה . Bulletin KB

.(Critical, Important, Severity

, Remote Code Execution, Information - Impact

.(Disclosure

- כוורתה העדכון . Title

- המוצר המושפע . Affected Product

- מספר קישור רכיב העדכון . Component KB

.(Adobe Flash Player (כגון , Affected Component

.(Remote Code Execution (למשל , Impact

.(Critical, Severity

.(MS17-005 . Supersedes

. (Yes/Maybe - האם נדרש להפעיל מחדש המערכת לאחר העדכון Reboot

A	B	C
Feature Name	Definition	Description
Date Posted	The date when the bulletin was published.	Indicates when the security update or bulletin was officially released.
Bulletin Id	The unique identifier for the security bulletin.	Helps in referencing and categorizing security updates.
Bulletin KB	Knowledge Base (KB) number associated with the bulletin.	Links to detailed technical documentation about the update.
Severity	The criticality level of the update.	Defines the importance of applying the update, such as Critical, Important, etc.
Impact	The type of vulnerability the update addresses.	Specifies whether the vulnerability impacts Remote Code Execution, Denial of Service, etc.
Title	The title of the security bulletin or update.	Provides a brief description of the update or its purpose.
Affected Product	The product impacted by the vulnerability.	Lists the operating systems, applications, or services requiring the update.
Component KB	The Knowledge Base (KB) number for the affected component.	Specifies the technical documentation for the impacted component.
Affected Component	The specific component affected by the vulnerability.	Details the part of the product or service that is vulnerable.
Impact.1	Secondary impact description for the vulnerability.	Further clarifies the vulnerability's effect, if needed.
Severity.1	Secondary severity level for the update.	Provides an additional severity classification, if applicable.
Supersedes	The bulletin or update that this one replaces.	Indicates updates that are deprecated or no longer applicable.
Reboot	Whether a reboot is required after applying the update.	Indicates if the system needs to restart to complete the update installation.
CVEs	Common Vulnerabilities and Exposures identifiers.	Lists standardized IDs for vulnerabilities addressed by the update.
שם תכונה	הגדרה	תאור
אחור רכיב	האחור שvu פוטס העילן.	מצין מה עדכון אבחנה או עלן שחזור אונליין רשמי.
מחהה עליון	המחהה הייחודי של עליון האבטחה.	עדר בהפעיה או זיאוג עדכוני אבטחה.
KB עלין	מספר מאגר הדיע (KB) המשויך לעליון.	קישורם ליעדר טכני מפורט על העדכון.
חווארה	רמת הקיטוחות של העדכון.	מגדיר את החישובות של יישום העדכון, כגון קרייטי, חשב וכו'.
פיעעה	סוג הפיענוח שבב העדכון מטל.	מצין מה פעולה מסוימת על בעזע קוד מרחוק, מינעת שירות וכו'.
כונתרת	הគורת של עליון האבטחה והעדכון.	מספק תיאור קצר על העדכון או מטרתו.
מוצר מושפע	המוצר שהושפע מהഫגעות.	מפורט את התוצאות ההפועלן, היישומים או השירותים הדורשים עדכון.
רכיב KB	מספר מאגר הדיע (KB) עבור הרכיב המושפע.	מצין את התיאור הטכני עבור הרכיב המושפע.
ביב מושפע	הביב הופכי המשוער מהפגיעה.	פריטח תחולן של המזוז או השירות ההפוך.
1. השפעה	תיאור השפעה משנית עבור הפגיעה.	מחבר עד יותר את השפעת הפגיעה, במידת הצורך.
1. חומרה	רמת חומרה משנית עבור העדכון.	מספק סיווג חומרה נסוף, אם ולולו.
מחלף	עלון או העדכון אחדך זהה מחלף.	מצין עדכון שיצאו משימוש או שאים ישים עוד.
לאתול	האם נדרש אטול מחדש הולה העדכון.	מצין אם המערכת צריכה לשובל מחדש כדי להשלים את הקונטן העדכון.
CVEs	כתוב פונציונון וושופר וטואן	מכוון מודדים נזירים ובורגונדיים תומכו במתחולט על ידי גודלו.

## תיאור נתוניים

תיאור הנתוניים (Data Description) הוא אחד שלביהם הكريיטיים ביותר בתהליך העבודה עם נתונים, במיוחד כאשר מדובר בפרויקטים גדולים הכוללים כמות גדולה של מידע (Big Data). שלב זה מספק הבנה עמוקה של המבנה, האיכות, והמאפיינים של הנתוניים שעימם אנו עוסדים. תיאור הנתוניים מסייע לזהות מאפיינים חשובים כמו סוגים משתנים, ערכים חריגים (Outliers), ערכים חסרים (Missing Values), ומגוון כלויות שעשוות להשפיע על התהליך הניתוה.

תיאור נכון ומדויק של הנתוניים מאפשר למנוע טעויות בניתוחים מאווררים יותר. מפחית סיכון לפרשנות שגوية של המידע, וטומך בקבלת החלטות מדעיות ואמינות. בפרויקטים המשלבים ניתוחים סטטיסטיים או למידת מכונה, תיאור הנתוניים הוא תנאי הכרחי לבחירת המודלים המתאימים ולביצוע שיפורים באיכות התוצאות.

### از מה בעצם הקשיים שאנו מתחווים איתם?

1. **נתוניים לא מסוורים או חסרים:** נתונים לעיתים קרובות אינם מסוורים או כוללים ערכים חסרים ובלתי תקינים, מה שמקשה על ביצוע ניתוחים מדויקים או בניית מודלים אפקטיביים.

2. **תלות בכלים ובמשאבים:** שימוש במערכות כמו Google Colab או Python תלוי בזיכרון זיכרון (RAM) מוגבל ובמערכות אחרות המותאמות לעיבוד נתונים. במקרה של חיבור לא יציב או זיכרון מוגבל, התהליך עשוי להיעצר או לגרום לאובדן נתונים.

3. **זמן עיבוד ארוך:** ניתוח קבצים גדולים דורש זמן עיבוד ממושך, במיוחד אם אין שימוש בטכניקות אופטימיזציה מתקדמות כמו חלוקת הנתוניים או עבודה מקבילה.

### از מה הפתרונות לכך?

1. **נתוניים לא מסוורים או חסרים:** נתונים לעיתים קרובות אינם מסוורים או כוללים ערכים חסרים ובלתי תקינים, מה שמקשה על ביצוע ניתוחים מדויקים או בניית מודלים אפקטיביים.

2. **תלות בכלים ובמשאבים:** שימוש במערכות כמו Google Colab או Python תלוי בזיכרון זיכרון (RAM) מוגבל ובמערכות אחרות המותאמות לעיבוד נתונים. במקרה של חיבור לא יציב או זיכרון מוגבל, התהליך עשוי להיעצר או לגרום לאובדן נתונים.

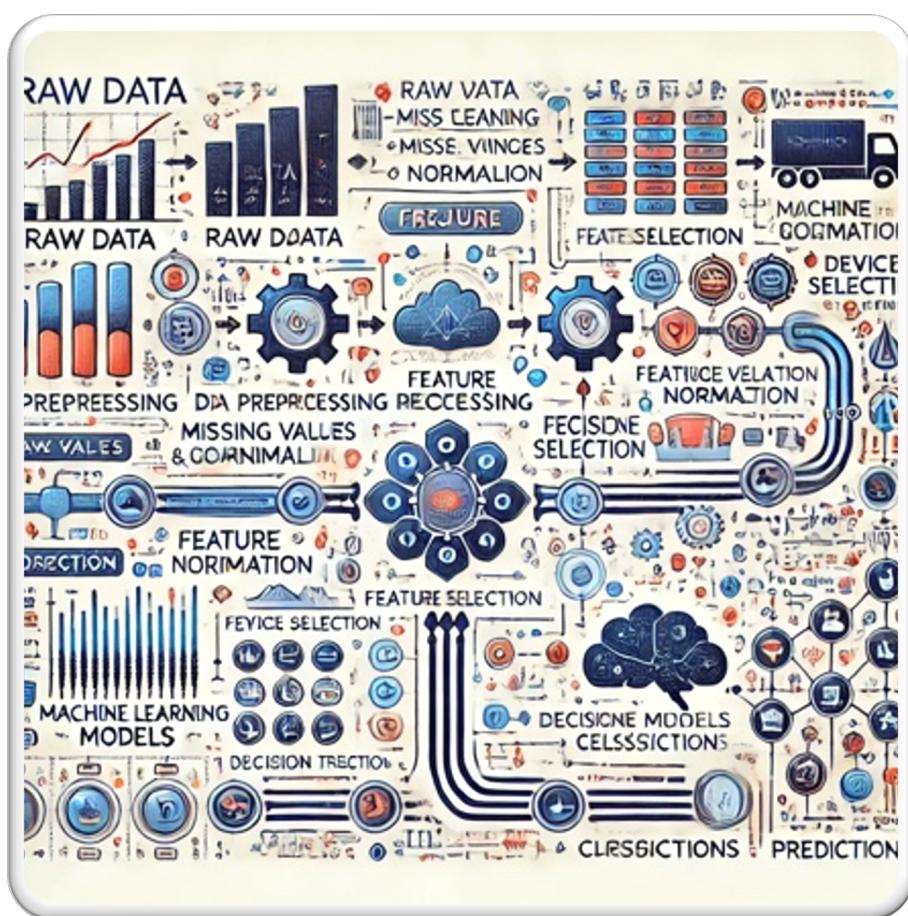
3. **זמן עיבוד ארוך:** ניתוח קבצים גדולים דורש זמן עיבוד ממושך, במיוחד אם אין שימוש בטכניקות אופטימיזציה מתקדמות כמו חלוקת הנתוניים או עבודה מקבילה.



## از מה אנחנו הולכים לבצע בעצם?

בתחילת העבודה שלנו, אנו מתחילהים באיסוף הנתונים והבאתם לסביבת הנכונה לעיבוד (לדוגמא, הعلاה ל - Colab) לאחר מכן, אנו מבצעים שלב תיאור ראשון של הנתונים על מנת להבין את מבנה המידע ואיכותו. תיאור זה כולל ניתוח סטטיסטי בסיסי (כמו ממוצעים וסטיות תקן), זיהוי ערכים חריגים ומילוי נתונים חסרים.

במקרים שבהם הנתונים גדולים מדי לעיבוד ישיר, **אנו מפעילים אותם לנתחים קטנים יותר**, מבצעים אופטימיזציה לעובדה, ושומרים את החלקים במיקום נגיש (כגון, Google Drive) במקביל, אנו משתמשים בכלים מתקדמיים ויעילים כדי לזרז את תהליך הניתוח ולהבטיח שנוכל להפיק תובנות משמעותיות בזמן סביר.





## כמויות הנתונים

בפרויקט שלנו, אנו עובדים עם מערך נתונים המורכב מכ-14 תכונות שונות, המתארות מידע מגוון אודוט אירופים ותהליכיים בארגון. הנתונים כוללים מזהים ייחודיים עבור כל אירוע, תיאורים של קטגוריות שונות של אירועים, ועוד. כל שורה מייצגת אירוע סייבר או פרופיל חסוד הקשור לעניות האירועים שקרו במערכת בזמן הנתון.

בנוסף, הנתונים כוללים כ-23,517 תכיפות, מה שמצויב על כמות גדולה של מידע. ניתוח נתונים כזה מחייב טכניקות עבודה מתקדמות כדי להתמודד עם המספרים והמידע המגוון. הנתונים כוללים אף תכונות שעברו קידוד/סיווג כדי להקל על ניתוחם וסודיותם.

כמויות הנתונים מאפשרת לנו להשוות בין תכיפות שונות, לzechות חריגות ולבדק איך שינויים במערכת משפיעים על התנהגות האירועים. עם כמות נתונים גבוהה, אנחנו יכולים לראות את התמונה המלאה ולהבין יותר טוב את הקשרים וההשפעות בין משתנים שונים, ווזר לzechות מגמות.

כדי להבין את הנתונים טוב יותר ולהזות תוצאות בצורה מדויקת, אפשר להשתמש בטכניקות כמו למידת מכונה. הטכניקות האלה עוזרות לנו לzechות קשרים שלא תמיד רואים בעבודת ניתוח רגילה. זה גם מאפשר לנו לzechות בעיות בזמן אמיתי ולפועל מהר ככל שינויים או חריגות במערכת, מה שסייע לנו להיות יותר גמישים ויעילים.

ఈ הנתונים רבים, אנחנו מקבלים תמונה ברורה ומלאה של המצב, שמאפשרת לנו להבין את המערכת בצורה טובה יותר. עם יותר נתונים, אנחנו יכולים לzechות דפוסים ושינויים שלא היו נראים עס כמות קטנה של נתונים. בנוסף, נתונים רבים מאפשרים לנו למצוא קשרים חדשים בין המשתנים, דבר שימושי לשיפור הדיק של החיזויים והבנה הכללית. בסופו של דבר, כמות גדולה של נתונים עוזרת לנו לשפר את הביצועים, ליעל את התהליכיים ולהפיק תוצאות מדויקות יותר.

יתר על כן, איקות הנתונים חשובה מאוד להצלחת הניטוחים. יש כמה בעיות שיכולות להשפיע על איקות הנתונים, כמו עריכים חסרים, טוויות הקלדה ונתונים שלא תואמים אחד לשני. הרבה נתונים יכולים להיות חסרים וזה עלול לגרום לטעוויות בתוצאו.

בעובדה עם נתונים, יש לנקה בחשבון את בעיות האיקות שיכולות להיווצר, כמו עריכים חסרים או תכיפות עם מידע חסר, אשר דורשות טיפול מוקדם לפני עיבוד הנתונים (Pre processing). בתחילת חקר הנתונים ויצירת מודלים מתקדמים, נבצע אופטימיזציה לטיפול בנתונים חסרים ובטיה שמירה על יציבות המערכת והיכולת לבצע תחזיות מדויקות ככל הנitin.



כמויות גבואה של נתונים יכול להוביל למספר מסקנות חשובות :

1. דיקט גבואה – עם כמויות גבואה של נתונים מודלים יכולים ללמידה בצורה טוביה יותר, לモזער את ערך השגיאה ולספק תחזיות מדויקות יותר.
2. יכולת זיהוי קשרים (גרף קורלציה) – כמויות גבואה של נתונים מאפשרת זיהוי קשרים מורכבים ומידית השפעתם אחד על השני, דבר שימושי לזיהוי מגמות טוב יותר וחזקוי מדויק יותר.
3. הבנה מעמיקה – כאשר יש כמויות גבואה של נתונים נקלט תמונה עמוקה ומאפשר הבנת משתנים וקשריהם.

- כתה נראה את כמויות הנתונים בפועל בהרצה מהירה בפייטון -

```
# ביצוע קידום נתונים טריניטריים
Data_Frame = pd.read_excel("Merged_Bulletin_Data.xlsx")

# Data Set
צגון המציג את ה
print("Data Set Shape:", Data_Frame.shape)
```

```
=====
Data Set Shape: (23515, 14)
=====
```

- לפני קריאת הגילוונות המאוחדים ביצענו קוד המציג את כל הגילוונות ע"פ אינדקסים קבועים, לאחר מכן ווידאנו את הדבר בExcel.

```
# DataFrame-
צגון סטטיסטי הנתונים של כל טבלה
print("DataTypes of Data Set:")
print(Data_Frame.info(), "\n")
```

```
DataTypes of Data Set:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23515 entries, 0 to 23514
Data columns (total 14 columns):
 #   Column           Non-Null Count Dtype
 ---  -- 
 0   Date Posted     23515 non-null  datetime64[ns]
 1   Bulletin Id    23515 non-null  object
 2   Bulletin KB    23470 non-null  float64
 3   Severity        23499 non-null  object
 4   Impact          23515 non-null  object
 5   Title           23515 non-null  object
 6   Affected Product 23513 non-null  object
 7   Component KB   23494 non-null  float64
 8   Affected Component 11362 non-null  object
 9   Impact.1        23515 non-null  object
 10  Severity.1     23089 non-null  object
 11  Supersedes      13030 non-null  object
 12  Reboot          23125 non-null  object
 13  CVEs            23330 non-null  object
dtypes: datetime64[ns](1), float64(2), object(11)
memory usage: 2.5+ MB
```

- נשים לב לסוגי הנתונים כי קיימים לנו הרבה Object, כלומר הדבר מעיד על דברים, או שיש ערבוב של DATA או שהDATA פשוט אינו מוגדר נכון.



## סוגי ערכאים

במהלך העבודה עם הנתונים, נתקלו במספר סוגים ערכאים שבהם יש להתמודד בדרכים שונות כדי לשמר על איות הנתונים ולהבטיח ביצועים אופטימליים במודלים שנבנה עליהם. להלן הסוגים העיקריים של הערכאים שפגשו:

### 1. ערכאים מספריים

הערכאים המספריים משמשים לתיאור אובייקטים כמותיים ויכולים להיות שלמים, חסרים או מכילים ערכים קיצוניים. علينا להשלים באמצעות חיזון, מומצע או שכיחות כדי לשמר על דיקוק המודל. כמו כן, יש לטפל בערכאים קיצוניים ולודא שאין הטענות לא סימטרית בנתונים.

### 2. ערכאים קטגוריאליים

הערכאים הקטגוריאליים הם כמו סוגים שונים של אירועים או תכנים בתציפות, ערכאים אלו מייצגים קטגוריות ולא מספרים. חשוב שבוצע קידוד נכון עבור ערכאים קטגוריאים אלו, כמו בשיטות Label Encoding או One-Hot Encoding, כדי שנוכל לאפשר למודלים להשתמש בהם כערכאים מספריים ולקראם אותם באופן תקין, או להשתמש בספריית python יייעודית.

### 3. ערכאים חסרים

ערכאים חסרים יכולים להופיע בכל אחד מסוגי הערכאים. בחלק מההmarkerים, הערכאים החסרים עשויים להיות תוצאה של חוסר נתונים בזמן יצירת הדוח או תקלות אחרות. טיפול בערכאים חסרים חשוב במיוחד, כי נתונים חסרים עשויים לעוזת את תוצאות המודלים. علينا להתמודד עם ערכאים חסרים באמצעות אלגוריתמים להשלמה כמו חיזון, ממוצע או שיטות מבוססות מודלים כמו KNN.

### 4. ערכאים של תאריכים

תאריכים יכולים לעזור לנו לזהות מגמות, לתקן תחזיות, ולבוד אם زمنי אירועים. נזדאת שההתאריכים נכתבים בפורמט הנכון, כדי שנוכל לנצל את המידע בצורה הטובה ביותר. לעיתים יש צורך לעבד את הערכאים הללו, כמו חיבור זמני אירועים או יצירת משתנים חדשים מתוך התאריך.

לסיום, סוג הערכאים לעיל דורשים טיפול שונה בהתאם למאפייניהם וצרכיהם השונים. נדרש לבצע התאמות כמו קידוד, השלמה של ערכאים חסרים, טיפול בערכאים קיצוניים ועיבוד תאריכים, כדי להcin את הנתונים בצורה שתאפשר יצירת מודלים חזקים ומדויקים.



### כעת נציג כמה בעיות שנתקלנו בהם -

```
# Data Set-ב中有NaN
# 針對各欄位中有多少個NaN值
nan_columns_DataSet = Data_Frame.isna().sum()

# 打印出各欄位中有多少個NaN值
print("Number of NaN values per column in Data Set:")
print(nan_columns_DataSet[nan_columns_DataSet > 0], "\n")

# 打印出所有欄位中有NaN值的總數量
print(f"The total of all columns with 'NaN' values in all the Data Set is: {nan_columns_DataSet[nan_columns_DataSet > 0].count()}\n")
```

```
Number of NaN values per column in Data Set:
Bulletin KB          45
Severity             16
Affected Product     2
Component KB          21
Affected Component   12153
Severity.1           426
Supersedes            10485
Reboot                390
CVEs                  185
dtype: int64
```

The total of all columns with 'NaN' values in all the Data Set is: 9

- נשים לב כי קיימות אצלונו 9 עמודות המכילות ערך ריק "NaN"

```
# בדיקת דוחות דווקאים בדאטן סט
duplicates_DataSet = Data_Frame.duplicated().sum()
print(f"Number of duplicate rows in Data Set: {duplicates_DataSet}", "\n")
```

```
=====
Number of duplicate rows in Data Set: 318
=====
```

- נשים לב כי קיימים לנו בדאטה סט כ – 318 תכפיות דומות (כפילות)

## עלכות קידוד

במהלך תהליך עיבוד הנתונים, אחד שלביהם החשובים הוא קידוד הערכים בתוכנות קטגוריות כך שהם יהיו ניתנים לעיבוד על ידי אלגוריתמים של למידה מוכונה. תהליך זה קרוי **קידוד נתונים** לאור העובדה שלמכונה קשה להבין ערכים טקסטואליים או קטגוריים, علينا להמיר את הערכים הללו לפורמט מספרי או מתמטי.

אחת הטכניקות הנפוצות ביותר לעיבוד קידוד היא **קידוד תווית** (Label Encoding), בה כל ערך קטגוריאלי מומר לערך מספרי, כך שהאלגוריתם יכול לעבוד איתו במדויק. לדוגמה, במקרים להשתמש בערכים טקסטואליים כמו "Public" ו- "Private", נוכל להמיר אותם ל-1 ו-2, או לכל מספר מתאים אחר.

בשלב זה, נבחן את כל התכונות בסיסד הנתונים שלנו וננתח האם עברה קידוד או לא? איזה ערכים קיבלו? האם יש לה ערכים חסריים? או מהם הערכים הנפוצים?

### נבדוק את התכונות :

#### Date Posted

• תכונה זו היא תאריך, והיא לא עברה קידוד. היא מיוצגת כפורמט תאריך datetime. ייתכנו ערכים חסריים אם הנתונים לא הוזנו כראוי. ערכים נפוצים יהיו תאריכים מתפסטים בטוח נתון.

#### Bulletin Id

• תכונה זו היא מזזה ייחודי לכל גלילון מידע (ID) ולא עברה קידוד. היא מיוצגת כמספר או כטקסט, תלוי במבנה הנתונים. ייתכן שיש ערכים חסריים אם המזזה לא הוזן. הערכים הנפוצים יהיו סדרה של מזחים שונים שנפוצים בקובץ.

#### Bulletin KB

• תכונה זו מצינית מזחה עבור מידע מותוך (KB) Knowledge Base, ולא עברה קידוד. מדובר במזחה טקסטואלי (למשל קוד גישה למידע). ייתכן ויש ערכים חסריים במרקירים של מידע לא הוזן. הערכים הנפוצים יהיו סדרות טקסט שמייצגות את מזחי ה - KB.

#### Severity

• תכונה זו מצינית את דרגת החומרה של הבעיה ולא עברה קידוד. ייתכנו ערכים כמו "Low", "Medium", "High" : קיימים ערכים חסריים אם לא הוזן דירוג. הערכים הנפוצים יהיו לפי התפלגות של דרגות החומרה.

#### Impact

• תכונה זו מצינית את השפעת הבעיה על המערכת, והיא לא עברה קידוד. ניתן



להניח שזו תכונה עם ערכיים כמו כמו "Critical", "Major", "Minor", "Chronic" אך ייתכן שיש ערכיים חסרים אם לא הוזנה השפעה. הערכיים הנפוצים יהיו המילים המתארות את עוצמת ההשפעה.

### Title

- תכונה זו היא טקסט חופשי (cotext הבעה), ולא עברה קידוד. היא מכילה טקסט חופשי ללא הגבלה מיידית לערכיים. קיימים ערכיים חסרים אם לא הוזןcotext. הערכיים הנפוצים יהיו כוורות שמצוות עם סוגים הבעיות.

### Affected Product

- תכונה זו מצינית את המוצר שנפגע, ולא עברה קידוד. ייתכן שמדובר בכמה מוצרים שונים (כמו, "Product A", "Product B", ...) קיימים ערכיים חסרים אם המוצר לא הוזן. הערכיים הנפוצים יהיו שמות מוצרים שימושיים באופן תדירים.

### Component KB

- תכונה זו מצינית את מזהה KB של הרכיב, ולא עברה קידוד. ייתכנו ערכיים כמו קוד מזהה עבור רכיב מתוך Knowledge Base. קיימים ערכיים חסרים אם לא הוזן מזהה עבור רכיב. הערכיים הנפוצים יהיו מזהה KB עבור רכיבים.

### Affected Component

- תכונה זו מצינית את הרכיב שנפגע, ולא עברה קידוד. מדובר בטקסט או בקוד שמייצג רכיב ספציפי (כמו, "Component A", ...) ערכיים חסרים עשויים להופיע במקרים של מידע חסר. הערכיים הנפוצים יהיו שמות רכיבים.

### Impact.1

- תכונה זו מצינית את השפעת הבעיה, כפי שהוזכרה בתכונה Impact, אך היא כנראה גרסה נוספת שלה, ולא עברה קידוד. ייתכנו ערכיים כמו "Critical", "Major", "Minor". הערכיים הנפוצים יהיו הערכיים שמתארים את עוצמת ההשפעה.

### Severity.1

- תכונה זו מצינית את דרגת החומרה, כפי שהוזכרה בתכונה Severity, אך היא כנראה גרסה נוספת שלה, ולא עברה קידוד. ייתכנו ערכיים כמו "Low", "Medium", "High". הערכיים הנפוצים יהיו דרגות חומרה תואמות.

### Supersedes

- תכונה זו מצינית אם יש עדכו או גרסה חדשה שמחילה את הבעיה הנוכחית. היא לא עברה קידוד, וויתכנו ערכיים כמו "Yes" או "No". קיימים ערכיים חסרים אם לא הוזן מידע על הגרסה המחליפה. הערכיים הנפוצים יהיו "Yes" ו- "No".

## Reboot

- תכונה זו מצינית אם יש צורך בהפעלה מחדש של המערכת לאחר התקלה. היא לא עברה קידוד, ויתכנו ערכים כמו "No", "Yes". קיימים ערכים חסריים אם לא הוזן מידע לגבי הפעלת המערכת מחדש. הערכים הנפוצים יהיו "Yes" ו- "No".

## CVEs

- תכונה זו מצינית את מזהה ה CVE (Common Vulnerabilities and Exposures), לא עברה קידוד. יתכנו ערכים כמו "CVE-2021-1234" שמייצגים מזהים של בעיות אבטחה. קיימים ערכים חסריים אם לא הוזן מידע על ה CVE, הערכים הנפוצים יהיו מזהה CVE שנפוצים בתחום האבטחה.

הסבר על שיטות הקידוד :

### - Label Encoding .1

**מה זה אומר ?**

Label Encoding היא שיטת קידוד הממירה ערכים קטגוריים (טקסטואליים) לערכים מספריים, כך שכל קטgorיה מקבלת מספר ייחודי.

לדוגמה :

"Public" -> 1

"Private" -> 2

### - One-Hot Encoding .2

**מה זה אומר ?**

שיטה שבה כל קטgorיה בודדת ממופha לעמודה נפרדת. כל עמודה מכילה ערך ביןארי (0 או 1), המציין אם רשומה מסוימת שייכת לקטgorיה זו.

לדוגמה :

קטgorיות : ["User", "Admin", "Suspicious"]

User -> [1, 0, 0]

Admin -> [0, 1, 0]

Suspicious -> [0, 0, 1]

### - Hash Encoding (Hashing) .3

**מה זה אומר ?**

כדי למפות Hash Function (Hash Function) משתמש בפונקציית הגיבוב Hash Encoding. השיטה מועילה במיוחד במקרים עם קטgorיות רבות.



לדוגמה :

"Category A" -> Hash("Category A") = 82435

"Category B" -> Hash("Category B") = 39284

#### - Binary Encoding .4

**מה זה אומר ?**

Binary Encoding ממיר ערכים קטגוריים למספרים בינאריים, ואז מפנה אותם למספר ביטים בעמודות נפרדות.

לדוגמה :

קטגוריות : ["A", "B", "C"]

"A" -> 1 -> [0, 1]

"B" -> 2 -> [1, 0]

"C" -> 3 -> [1, 1]

#### - Frequency Encoding .5

**מה זה אומר ?**

שיטת שבה כל קטgorיה מקבלת ערך מספרי המבוסס על שכיחות הופעתה בתווים.

לדוגמה :

קטגוריות : ["A", "B", "C"]

50 > מופיעה 50 פעמים "A"

30 > מופיעה 30 פעמים "B"

20 > מופיעה 20 פעמים "C"

#### - Ordinal Encoding .6

**מה זה אומר ?**

Ordinal Encoding מתאים לKatgoriyot שיש להן יחס סדר. מפנה את הערכים למספרים על בסיס סדרם.

לדוגמה :

"Low" -> 1



"Medium" -> 2

"High" -> 3

## 7. (החלפת ערכים חסרים) - Replacement Encoding

**מה זה אומר ?**

שיטת להשלמת ערכים חסרים באמצעות ערכים חלופיים כמו ממוצע, חציון או ערך שכיח.

לדוגמה :

תמונה עם ערך חסר "Age" -> משלימים עם ממוצע הגילאים.

## 8. (קידוד מותאם אישית) - Custom Encoding

**מה זה אומר ?**

שיטת שבה משתמש במצב קידוד מותאם אישית המבוסס על ידע תחום או דרישות ספציפיות.

לדוגמה :

"Yes" -> 1, "No" -> 0

"Male" -> 1, "Female" -> 2





## חקירת נתונים

חקירת נתונים (Data Exploration) היא שלב קרייטי בתהליך ניתוח הנתונים, שבו מתבצעת בחינה עמוקה של הנתונים הגולמיים שנאספו. בשלב זה, המטרה היא להבין את מאפייני המידע, לזהות תובנות ראשוניות, ולהכין את הנתונים לעיבוד מתקדם. חקירה זו היא הבסיס לציצרתמודלים חכמים ומדוקים, והיא משפיעה משמעותית על הצלחת הפרויקט כולו.

### הבנייה מבנה הנתונים:

חקירת הנתונים מספקת תמונה ברורה של מבנה המידע – סוגי התכונות, טווחי הערכים, וnochותם של ערכים חסרים. ידע זה מאפשר לזהות אילו עמודות חשובות לפרוייקט ואילו דרישות עיבוד נוספת.

### זיהוי בעיות בנתונים:

במהלך החקירה ניתן לאתר ערכים חריגים, ערכים שגויים, וכפיליות. טיפול מוקדם בעיות אלו יכול לשפר את איכות המודלים ולמנוע טעויות ניתוח בהמשך.

### גיבוש אסטרטגיות לעיבוד הנתונים :

חקירת הנתונים עוזרת לקבוע אילו טכניקות עיבוד יש לify, כמו קידוד משתנים קטגוריאליים, השלמת ערכים חסרים, או שינוי קנה מידת תכונות כמותיות.

### זיהוי תבניות וקשרים:

באמצעות ניתוח סטטיסטי בסיסי וקורלציית ניתן לחושף קשרים בין משתנים ולקבל כיוון ראשוני לפתרון הבעיה העסקית. לדוגמה, אם קיימת תכונה חזקה שמנבאת את היעד, ניתן להתמקד בה בשלב המודלים.

### כליים וטכניקות:

חקירת הנתונים משלבת כלים מתקדמים להדמיה ולניתוח, כמו בפייתון (בשימוש הסि�פריות pandas, matplotlib – seaborn) או ב – SQL, גרפים כמו היסטוגרמות, גרפי פיזור וטבלאות סיכום סטטיסטיות הם כלים מרכזיים המספקים תובנות מהירות ו互動יביות.

### לסיכום:

שלב חקירת הנתונים אינו רק שלב טכני, אלא שלב אסטרטגי, שמחבר בין הנתונים הגולמיים להבנה عمוקה של הבעיה העסקית. ביצוע נכון של חקירה זו משפר את האיכות הכלולית של תהליכי מדעי הנתונים, מצמצם טעויות אפשריות, וחוסך זמן ומשאבים בשלבים מתקדמים יותר.



- בעת חקירת הנתונים, אנו מתחילהים בגיבוש השערות המבוססות על התבוננות ראשונית והבנת התחום.

#### דוגמאות לשערות (לגביה הקשרים):

"האם יש קשר בין כמות האירועים הקשורים לאבטחת סייבר לבין פרמטר הזמן (יום/שבה')?"

אם נזהה דפוס עונתי (למשל, יותר תקלות בחודשים), נוכל לגבות מודל לתחזית עתידית.

"האם יש תכונה מסוימת המשפיעה באופן ישיר על חומרת האירועים?"  
ניתן לבדוק אם תכונה מסוימת משפיעה באופן ישיר על חומרת האירועים, כמו "סוג האירוע (Impact/Severity)."

#### דוגמאות לשערות (לגביה הנתונים עצם):

"האם כדאי קודם לאפיין את הנתונים (למשל להכין טבלת אקסל על כל נתון)"?

"אם כדאי לבצע נרמול של תאריך ושעה כדי לוזות דפוסים עונתיים או מגמות לאורך זמן?"

"האם علينا להמיר את הערכים הקטגוריאליים (כגון Bulletin Id, Affected Product) לקלידוד בינארי או One-Hot Encoding לשיפור ביצועי המודל?"

"האם כדאי להתמקד באירועים עם חומרת "Critical" בלבד לצורך בניה מודל תחזית, או לכלול גם את היתר ?"

"האם יש צורך בשינוי מחדש של חומרת האירועים (Severity) למספר רמות (למשל, נמוך, בינוני, גבוה) כדי להקל על בניית המודל?"

- במהלך חקירת הנתונים, חלק מההתכונות מתבלטות בשל פוטנציאל השפעתן או יכולת להפיק מהן תובנה.

#### דוגמאות לתוכנות מבטיחות -

##### **תכונה "Severity"**

- גילינו שהתכנית מצינית את רמת החומרה של האירועים.
- התכונה יכולה לשמש לסיווג תקלות לפי חומרתן ולביצוע אנליזות על פי סוג החומרה.



## "Impact2" ו- "Impact"

- תוכנות אלו מtarות את סוג האיום וההשפעה של האירוע, כמו "Remote Security Update" או "Code Execution".
- ניתן לבצע ניתוחים לפי קשרים בין סוג האיום לモוצר המושפע או לחומרת האירוע.

## "Bulletin KB"

- מס'KB של העדכון (Knowledge Base).
- מאפשר זיהוי פרטי העדכון וקשר אותו למסמכים ותרגומים נוספים של מיקרוסופט.

## "Affected Product"

- המוצר המושפע מהעדכון.
- מאפשר לדעת אילו תוכנות או מערכות חייבות לעדכון, כדי להימנע מפגיעות.

## "Component KB"

- מס'KB של הרכיב המושפע.
- נותן מידע נוסף על רכיב מסוים במערכת או במוצר המושפע, כך שניתן להتمיקד בהיבט ספציפי יותר בעדכון

## "Supersedes"

- האם העדכון מחליף עדכון קודם.
- מאפשר לעקב אחריו עדכניםים שמחליפים או עדכניםים שהיו עדכנים קודמים להם, מה שעוזר בניהול תהליכי העדכון.

## "Reboot"

- האם נדרש אחזור לאחר התקנת העדכון.
- חשוב למנהל המערכת כדי להבין את הצעדים שצריך לבצע אחרי התקנה."

## "CVEs"

- שימושי לזיהוי פגיעות קודמות וקשר לעדכניםים ולהומרה או תוכנה שנפגעו מהם.



## קשרים אפשריים בין תוכנות

”**Affected Bulletin Id**” לביון **Product**.

- קשרים אלו עשויים לסייע בהבנת התקלות המשותפות במספר מוצרים או קטגוריות אבטחה.

## תוכנות עם ערכים חסרים

- התגלו ערכים חסרים רבים.
- נתונים אלו דורשים התייחסות מיוחדת, כמו השלמה (imputation) או טיפול בערכים החסרים בדרכים אחרות.

## שימושים פוטנציאליים:

מאפיינים אלו יאפשרו לנו לבצע סיווגים מדויקים יותר של התקלות, לאמן מודלים**Machine Learning** ולוזהות דפוסי התנהגות חריגים.

בנוסף, החקירה יכולה לעזור את תהליך בחירת התוכנות ולהשபיע על עיצוב המודל הסופי.

במהלך הממחקר ניתוח הנתונים, במיוחד כאשר אנו בודקים את הבעיות), אך אנו עשויים לשנות את ההשערות הראשונות שלנו.

למשל, אם היינו מניחים כי ההתקפלות של סוגים אירוניים האבטחה אחידה בין כל קבצי האימון והטסט, ייתכן ומהמחקר שלנו היה מוצא פערים בהתקפלות התקלות לפי רמות חומרה או לפי סוג האיוומים. הדבר יכול לשנות את ההשערה הראשונית ולגרום לנו להתרכז יותר במודלים שמתמודדים עם התקלות חמורות או התקלות הקשורות למוצרים מסוימים. בנוסף, אם נבצע ניתוח נוסף על נתוני ה'חומרה', זה עשוי לשנות את החשיבה שלנו לגבי איך לאמן את המודל, למשל תוך שימוש בשלושתנים שמצביעים על סוג האיוומים וה להשפעות החומריות.

לסיכום, ניתן לומר שהמחוקרים שביצעו לא שינו את ההשערות הראשונות שלנו לגבי הדאטה, אך היינו מוכנים אליהם

• באופן כללי, הממחקר לא בהכרח שינה את המטרות הראשונות שלנו, אלא עשו רק לעדכו את הגישות וההשערות בהן אנחנו משתמשים.  
לסיכום, אם הנתונים יגלו לנו תובנות חדשות לגבי המיקוד בתקופות פנימיות או חיצונית, זה ידרש שיפור בהתאם של המודל שלנו לאותן התקופות. ככלומר, ככל שתיגלה לנו יותר תובנות בעבר תוכנות חדשות או קיימות נדע לבצע אופטימיזציה למודל.





## aichot ha-natoniim

נתונים מהווים את עמוד השדרה של כל פרויקט מדעי נתונים, איקות הנתונים, הדיק שלם והמכנות לעיבוד המסורתיים קритיים המשפיעים ישירות על הצלחת המודל שלנו. אנו עוסקים בחקר, עיבוד וניתוח של מאגר נתונים כולל תוכנות מגוונות המשקפות את הסביבה הטכנולוגית במקויסוף.

מה כבר עשינו?

1. **חקר ראשוני של הנתוניים** – הבנו את מבנה הנתוניים, את ההקשרים בין התכונות, ואת האתגרים כמו ערכים חסרים, חריגים וחוסר איזון בתנאים.

2. **שיפור איקות הנתוניים** – ניקוי נתונים, טיפול בערכים שגויים, ושיתוב טכניות כמו חציון, ממווץ או שכיח בצורה מבודדת לצורך שלמות הדאטה.

3. **זיהוי תכונות רלוונטיות** – הتمקדמו בתכונות בעלות פוטנציאל חיזי גבוה.

מה עוד נותר לעשות?

1. **העשרה ושיפור התכונות** -

הנדסת תכונות (Feature Engineering): נבצע יצירה של משתנים חדשים שעשוים להיות שימושיים למודל, למשל אינדיקציה לרמת חומרה של התקלות, סיוגים מותאמים של פעולות חריגות ועוד.

שיפור איקות הנתוניים החסרים: המשך שימוש בשיטות סטטיסטיות ומודלים להשלמה חכמה.

2. **טיפול בחוסר איזון בתנאים** -

במקרים בהם יש חוסר סימטריה, נבחן טכניות כמו Undersampling oversampling למניעת הטיות במודל.

3. **אימות והערכת המידע** -

נבחן את המהימנות והעקביות של הנתונים לאורך זמן, במיוחד כאשר יש קשרים בין תכונות מסוימות לתוצאות החיזוי.

4. **הרחבת מקורות המידע** -

שילוב מקורות חיצוניים או פנימיים נוספים לצורך הצלבת מידע ושיפור דיקן המודל.



## למה זה חשוב?

המטרה שלנו היא לבנות מודל חכם, מדויק ויעיל, אך מודל כזה טוב בדיקות כמו הנתונים שעליהם הוא מתבסס. נתונים באיכות גבוהה מאפשרים לנו :

- להפחית רעש וטעויות בתהיליך החיזוי.
- להגדיל את הדיקות והמהימנות של התהווות.

לזהות **דפוסים סטטיסטיים** ולהפיק תובנות עסקיות בעלות ערך למיקיروسופט.

אנחנו מתמקדים לא רק בשיפור איכות הנתונים המקוריים, אלא גם ביצירת תהליכי שיבתיחו כי בעתיד נוכל לעבוד עם נתוניםקיימים, מאורגנים ומועילים יותר. התהיליך הזה כולל עבודה מתמשכת של מחקר, בדיקות ושיפור מתמיד.

**במבט קדימה**, ככל שנעמיק בחקר הנתונים ונתמודד עם אתגרים חדשים, נוכל לפתח מודל שיספק חיזוי מתקפות ברמה גבוהה, יתרום ליעילות ולבטיחות התשתיות, ויאפשר למיקרוסופט להישאר צעד אחד קדימה בתחום אבטחת הסיביר.

הסבר מושגיים -

### Oversampling .1

**הגדרה** : טכנית להגדלת מספר הדוגימות של הקטגוריה הנדירה (minority class) בסט הנתונים על מנת לאזן את שכיחות הקטגוריות השונות.

**מטרה** : להתגבר על בעיית חוסר איזון בנתונים כך שהמודל לא יטה לטובות הקטגוריה הדומיננטית.(majority class).

### Undersampling .2

**הגדרה** : טכנית להקטנת מספר הדוגימות של הקטגוריה השכיחה (majority class) בסט הנתונים כדי להתאים את הכמות לקטגוריה הנדירה (minority class).

**מטרה** : למנוע חוסר איזון על ידי הפחיתת הדוגימות של הקטגוריה הדומיננטית, מה שעזר למנוע הטוות בסיווג.

שתי הטכניקות משמשות כפתרונות לבעיית חוסר איזון בנתונים בעיות סיווג (Classification).



## חלק מהבניות שאנו מתמודדים איתם -

נתונים חסרים: ערכים חסרים בעמודות קריטיות שיכולים להשפיע על איכות המודל.

תלות שגوية בין תכונות: קשרים לא נכונים בין תכונות עשויים להוביל למסקנות שגויות.

הבדלים בין קטגוריות או תיאורים: תיאורים לא אחידים או חפיפות בין קטגוריות.

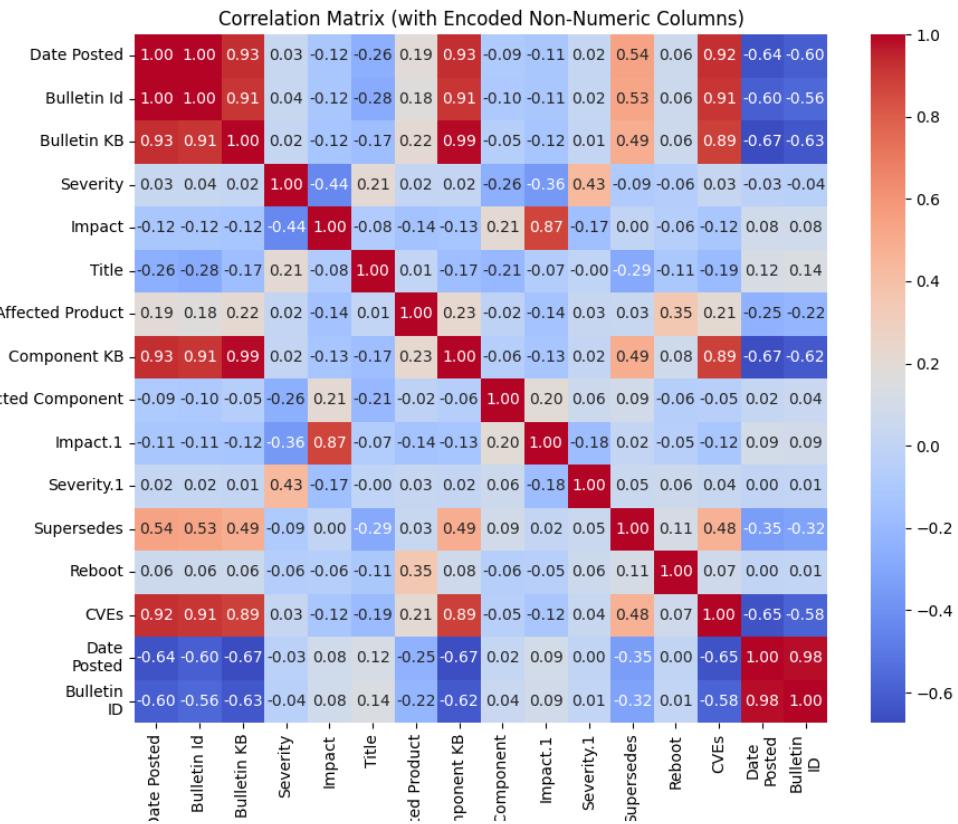
נתונים קטגוריאליים לא מסווגים: קטגוריות לא ברורות מקשות על ניתוח הנתונים.

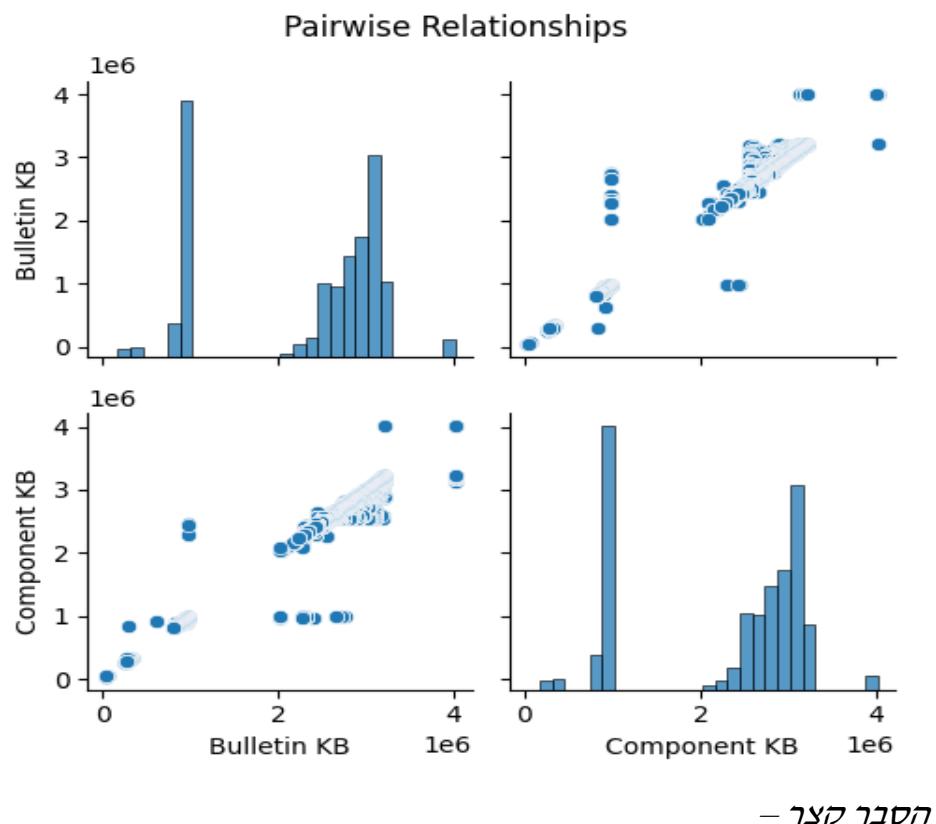
נתונים חופפים בין עמודות: חפיפות בין תכונות כמו "Severity" ו-"1". שМОBILETY לבלבול.

ערכים שגויים בעמודות תאריך: פורמטים שגויים בתאריך עשויים למונע ניתוח נכון של זמנים.

תיאורים לא ברורים: כוורות לא ברורות או כלליות עשויות להקשות על הזיהוי מהיר של הבעיה.

## לסיכום, נציג כמה גורפים מעוניינים המבאים בעיתיות וגם קצר סקרנות לחקר נתוני הפרויקט -

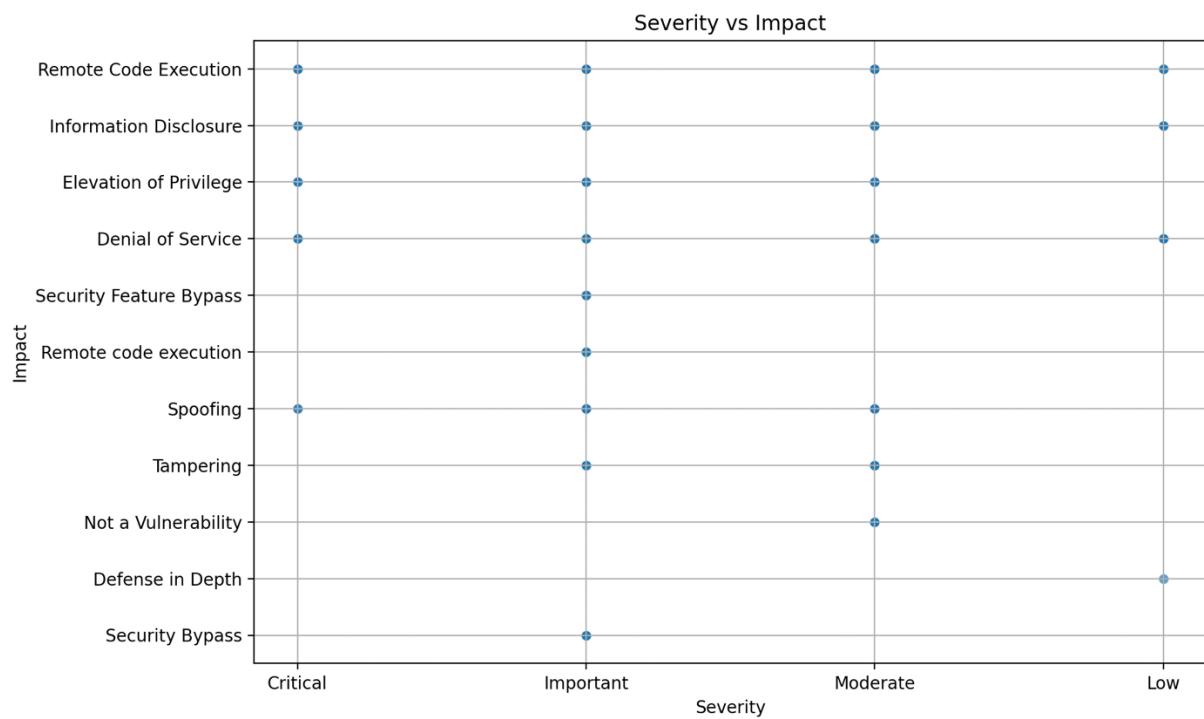




**מזהה ה KB:** **Bulletin KB** - של הבוליטין, שמתאר את העדכון שהופץ על ידי מיקרוסופט. תוכנה זו משמשת לזיהוי העדכון הספציפי שנינו כדי לפתור בעיה מסוימת.

**מזהה ה KB:** **Component KB** - של הרכיב המושפע מהעדכון, אשר מתאר את החלק הספציפי במערכת או ב מוצר שדורש תיקון או עדכון.

כאשר יש עדכון עבור "Bulletin KB" ישנה סבירות גבוהה שגם אותו רכיב ישתמש ב "Component KB"-דומה או משוויך.



- ניתן לראות את מידת ההשפעה בין פגיעה החומרה, כפי שרואים החומרה מוגדרת מ – Critical ועד Low –. ל墈נו את כל אפשרויות שיכולות להופיע בעמודות ההשפעה, והציבנו אותן על הגרף. כפי שניתן לראות יש לנו חמישה סיווגים תחת ההשפעה המוגדרים בקריטריום למערכות Microsoft.