Microsoft Security Bulletin Analysis

פרויקט "Bulletin Security Analysis" מוצג על ידי בר כהן וסהר חיים יעקב.

מטרת הפרויקט חיזוי אירועי סייבר במערכות "Microsoft" באמצעות מדעי הנתונים.

שם מרצה : מר אבי זכאי.
שם מנחה : מר חנן לב.
מגישים :
בר כהן 208110254
סהר יעקב 314741851

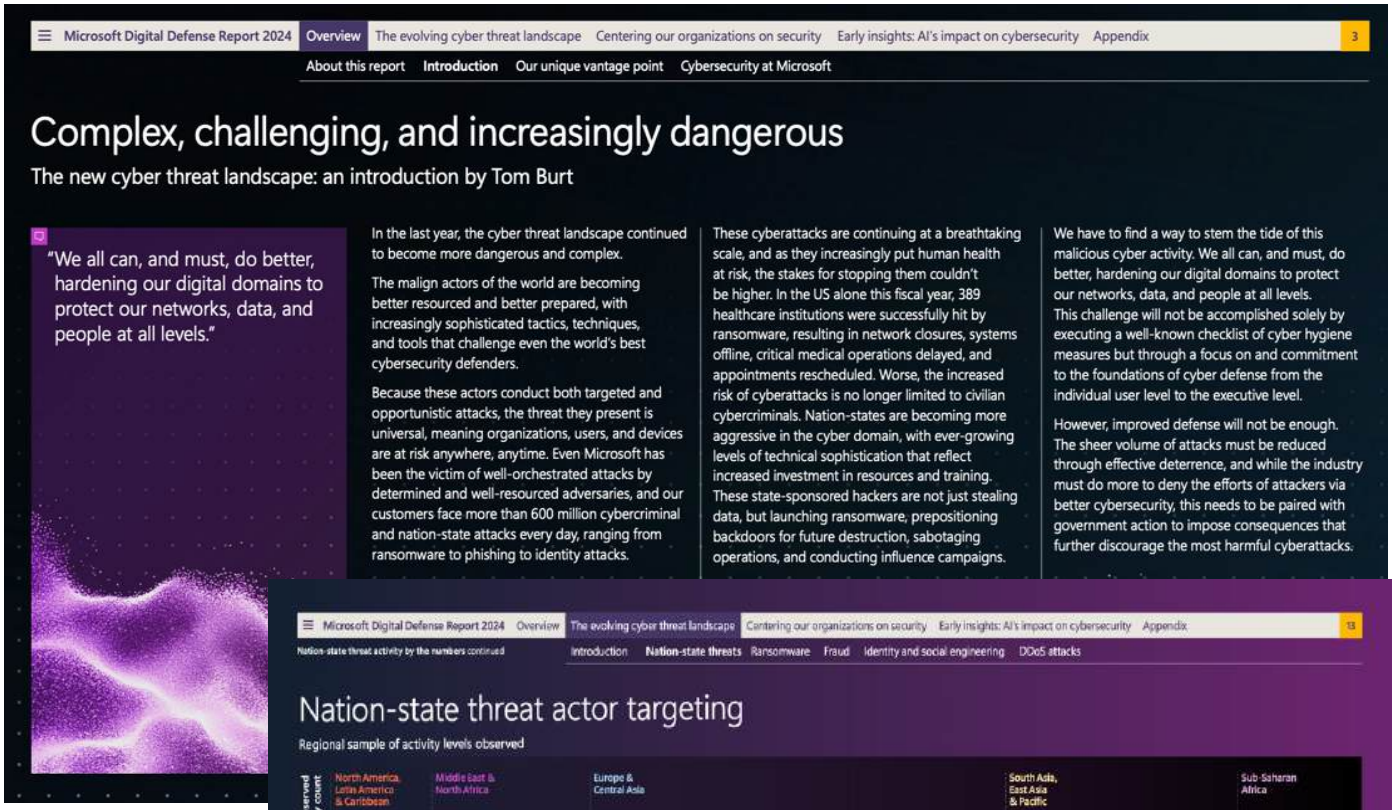
תוכן עניינים

רקע

חברת מיקרוסופט היא אחת החברות המובילות בעולם בעלת היסטוריה עשירה ותרומה רבה לפיתוח תשתיות טכנולוגיות שמסייעת לארגונים רבים . נוסדה על ידי ביל גייטס ופאול אלן ב1975 , מאז החברה מצליחה לצמוח בקצב אדיר בתחומי הטכנולוגיה והמחשוב האישי , אופיס ופלטפורמות ענן . במהלך השנים החברה מתמודדת עם מספר רב של תקיפות , ולכן משקיעה משאבים לפיתוח אמצעים טכנולוגיים כך שיאפשרו לה ביטחון ברמת האבטחה של הלקוחות.

להלן טבלה המציגה את המערכות הגנה השונות הקיימות במיקרוסופט ומטרותיה :

שם המערכת	מטרה
Azure Active Directory	ניהול הרשאות בארגון ואימות רב גורמי
Azure DDoS Protection	הגנת מתקפות בענן
Azure Firewall	חומת אש לענן לניהול ובקרה
Azure Information Protection	ניהול והגנה באמצעות הצפנה ונהלים
Azure Security center	ניהול אבטחת ענן עם ניטור ותיאם של התקפות אבטחה
Azure Sentinel	מערכת SIEM לניהול איומים תוך שימוש ב AIולמידת מכונה
Microsoft Defender	זיהוי איומים וחיזוי תקיפות
Microsoft Defender For EndPoint	הגנה על תחנות עבודה ושרתים
Microsoft Defender For identity	הגנה על חשבונות משתמשים ויישומים ארגוניים
Microsoft Defender For Office 365	הגנה על דוא"ל ואופיס ממתקפות פישינג כמו סוס טרויאני
Microsoft 365 Defender	פתרון לניהול על סביבת מיקרוסופט שכולל קבצים ישומים ודוא"ל.
Microsoft EndPoint Manager	ניהול מכשירים ותחנות עבודה בארגון
Microsoft Intune	פתרון מכשירים ניידים והגנה על מכשירים מקומיים
Windows Defender ATP	הגנה על תחנות עבודה ושרתים ביצוע ניטור ותגובה מיידית

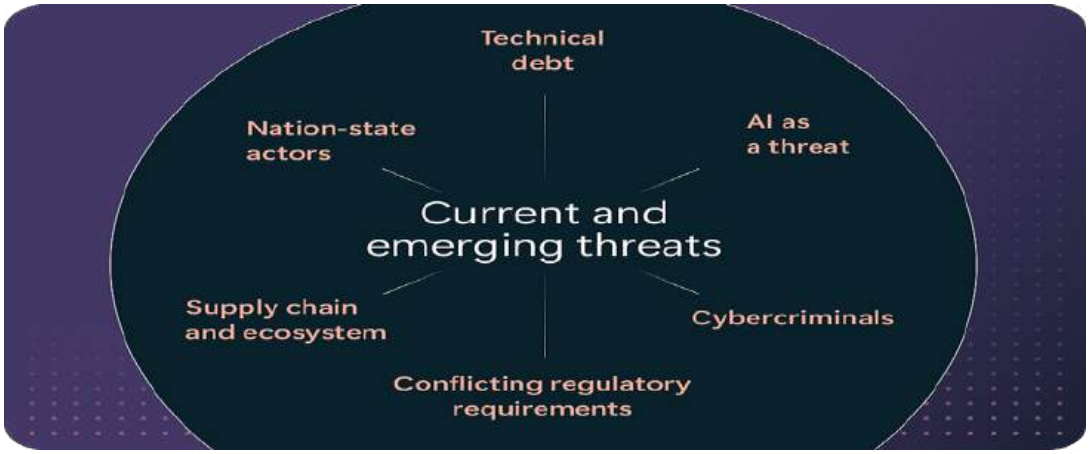


Microsoft Digital Defense Report 2024

בשנת 2024 התפרסם דוח על מצבה של הארגון בתחום הסייבר , מתמודדת עם כ-600 מיליון מתקפות , עם זאת מיקרוסופט בעלת ייתרון ייחודי על פני מתקפות שונות : צבירת נתוני אבטחה מחברות שונות , בשל מספר הלקוחות העצום ויצירת מגמות מרכזיות , מעקב אחר קבוצות איום , הקצאת מספר מהנדסים גבוה מאוד המתמקד בשדרוג ופיתוח רב גורמי כעדיפות ראשונה.

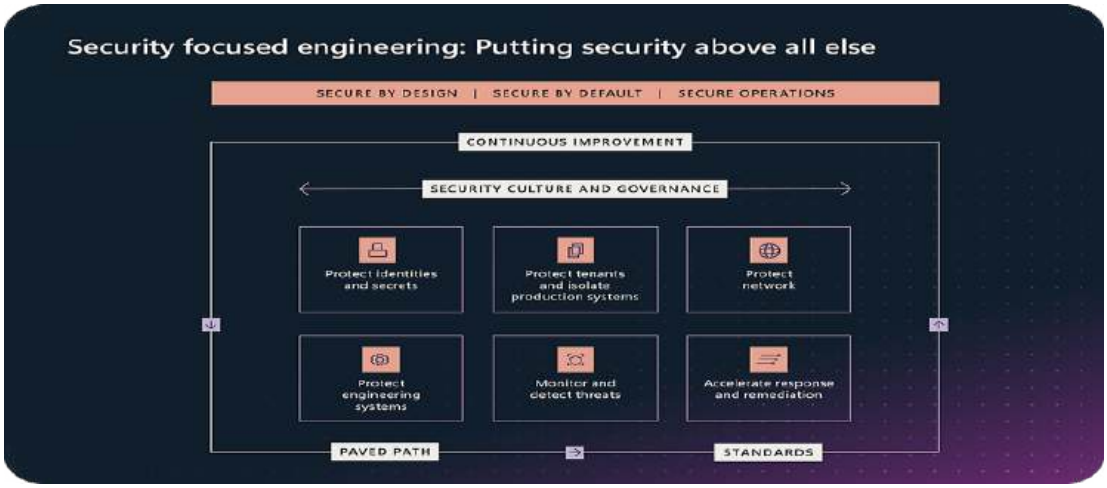
דוח הגנה דיגיטלית של מיקרוסופט 2024

דוח הגנה דיגיטלית של מיקרוסופט 2023



Source : Microsoft Digital Defense p.7

התמונה מפרטת את האיומים הנוכחיים של 2024 , איומי , AI סייבר ודרישות סותרות.



Source : Microsoft Digital Defense p.57

Source : Microsoft Digital Defense p.16

התמונה מציגה את חסמי האבטחה של מיקרוסופט כקריטריון ראשי.
גורמי הסיכון על חברת האבטחה לצורך רווח ומודיעין על החברה , מטרות האיומים משתנות בהדרגה , מתוך סקר שבוצע ניתן להבחין כי תחום ה , IT-פיתוח, ניהול ותמיכה של פתרונות טכנולוגיים המיועדים לשיפור ביצועי הארגון ולקוחותיו היא המטרה החזקה ביותר לאיומים.

הבנת נתיבי הגנה לגבי איומים על ידי זיהוי נתיב המתקפה הוא חשוב מאוד ומהווה סבירות גבוהה לצמצום מתקפות. הניתוח מציין ספירות נכסים , נתוני פגיעות ושטחי מתקפה חיצוניים.
השפעתה של בינה מלאכותית מחוללת שינוי בנוף איומים והגנה, שיפור אמצעי הגנה באמצעות הבינה המלאכותית מבססת תקנים ומסגרת עבודה עבור אבטחה.
שימושים בבינה מלאכותית משרתות גם כן את התוקפים , מיקרוסופט לוקחת חלק פעיל עם מערך הסייבר לפיתוח תועלת מהבינה המלאכותית בעלי עקרונות.
מתוך סקר שבוצע נמצא כי סין ורוסיה משתמשות כדי לייצר ולהפיץ תוכן מותאם להשגת יעדים כמו עיצוב דעת קהל, הפצת דיסאינפורמציה, ותמיכה באג'נדות פוליטיות באמצעות טקסט ברמה נמוכה\בינונית , תמונה ווידאו או שמע ברמה גבוה.

גם איראן והפרוקסים שלה מבצעות אך בסבירות נמוכה יותר.
רוסיה (סרט תיעודי מפוברק אודות איילון מאסק) , סין (תעמולה שנוצרה ע"י Taizi Flood) איראן (סרטון איום שנוצר לקראת מבצע צבאי איראני).
תקיפות סייבר שונות שמאופיינות כמניע של ביטחון לאומי , תקיפת איראן על ידי ישראל נעשית בעיקר באמצעות מתקפות סייבר בכדי לשבש את תוכנית הגרעין ויכולותיה הצבאיות. תקיפות כאלה כוללות את וירוס Stuxnet שפותח בשיתוף פעולה עם ארה"ב ונגרם נזק למתקני הגרעין האיראניים. בנוסף, ישראל מבצעת מתקפות סייבר נוספות על תשתיות אנרגיה ונמלים באיראן. מטרתן העיקרית היא למנוע פיתוח נשק גרעיני ולפגוע בארגונים הנתמכים על ידי איראן, וחלק מהמתקפות נעשות בעזרת טכנולוגיות מיקרוסופט. מניעים אלו חושפים את החשיבות של מערכות טכנולוגיה כמו מיקרוסופט.
מיקרוסופט הייתה מעורבת לאחרונה בזיהוי ובתגובות לאיומי סייבר, כולל אלה שמגיעים מקבוצת Lazarus ,שהיא קבוצת האקרים הנתמכת על ידי צפון קוריאה.
Lazarus ניצלו את הפגיעות במערכות הענן של מיקרוסופט, כגון Azure ו Office 365, כדי לחדור לרשתות ארגוניות ולגנוב מידע רגיש.

HAFNIUM היא קבוצת אקרים המיוחסת בדרך כלל לממשלת סין. העומדת מאחורי מתקפות רבות שניצלו את הפגיעות במערכות Microsoft Exchange (שרת דואר אלקטרוני ולוח שנה שפותח על ידי מיקרוסופט), והשתמשו בהן לחדור לרשתות של ארגונים גדולים ולקבל גישה למידע רגיש.

מיקרוסופט היא יעד מרכזי לפריצה ותקיפה, במטרה לגנוב נתונים ופרטי משתמשים.

שתי תקיפות משמעותיות שבוצעו על חברת מיקרוסופט -

1. התקפות על תשתיות ענן של (2020) Azure ו - Office 365 : Lazarus Group, הידועה כקבוצת תקיפות סייבר הנתמכת על ידי צפון קוריאה, ידועה בהתקפותיה נגד מערכות טכנולוגיה, הכוללות מערכות ענן. הקבוצה השתמשה (ועודנה משתמשת) בטכניקות מתוחכמות כדי לנצל פרצות באבטחה ולהשיג גישה בלתי מורשית. עם זאת, אין תיעוד ספציפי מהתוצאות שהוצגו לגבי תקיפה ישירה על Azure ו-Office 365 על ידי Lazarus Group.

https://www.microsoft.com/en-us/security/blog/2023/10/18/multiple-north-korean-threat-actors-exploiting-the-teamcity-cve-2023-42793-/vulnerability

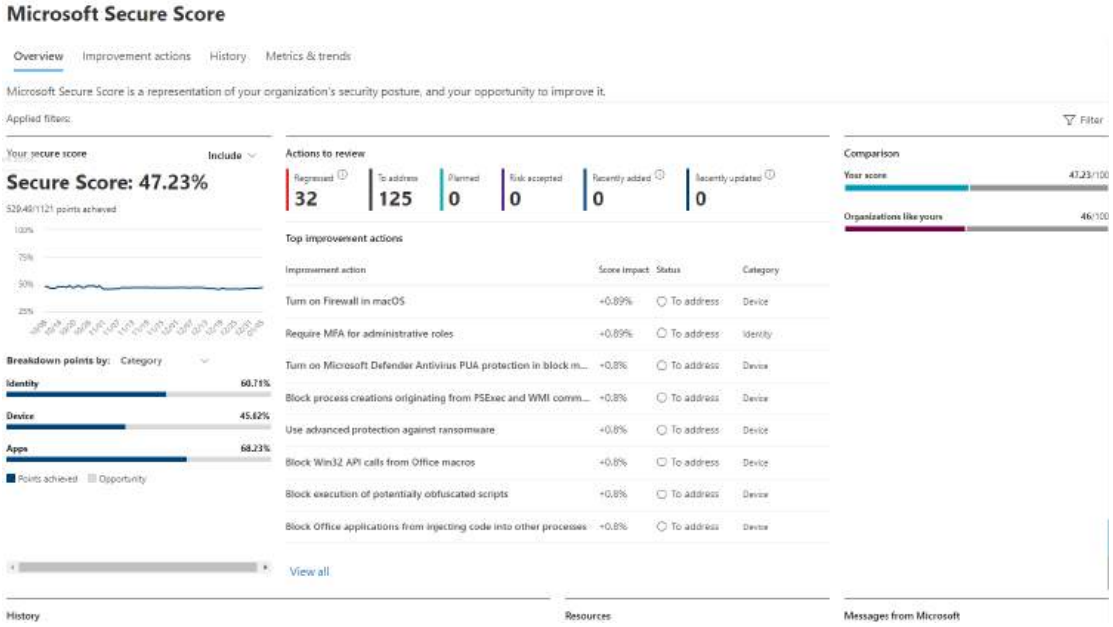
https://www.sisainfosec.com/weekly-threat-watch/lazarus-group-/exploits-windows-driver-zero-day-vulnerability

2. התקפות על (2021) Microsoft Exchange : קבוצות הפריצה של HAFNIUM, המיוחסת לסין, ניצלו חולשות במערכות Microsoft Exchange כדי לחדור לשרתי לקוחות ברחבי העולם. התקפות אלו התאפיינו בגניבת מידע רגיש ובהפצת קוד זדוני.
אופי התקיפה : מפעילי HAFNIUM התקינו מעטפות אינטרנט (web shells) על השרתים שנפרצו, אשר שימשו לניהול שליטה מתמשכת על השרתים.

https://en.wikipedia.org/wiki/2021_Microsoft_Exchange_Server_data_breach

דרכי התמודדות של החברה -

1. מיקרוסופט הגיבה לתקיפה על תשתיות הענן של חברת Lazarus ונקטה באמצעים שונים , שימוש במערכת למדידת רמת האבטחה והגברת ההגנה של המשתמשים , הפעלת הגדרות ברירת מחדל וחסימה של פרוטוקולים ישנים באמצעות אימות רב גורמי.
2. בעקבות חדירה לשרתי מייל ולוח שנה של משתמשים וגניבת מידע החברה תיארה צעדים לניהול חולשות וחיזקה את המערכת בגורמים שונים שיצליחו להגן על המשתמשים כמו : הפצת עדכוני מערכת לניהול חולשות , כלים לאיתור 'מעטפות אינטרנט' (קבצים זדוניים שהותקנו בשרתים נגועים המאפשרות שליטה מרחוק והרצת פקודות , וכולל סקריפטים שונים בשפת PHP או PERL לניצול חולשות אבטחה מערכתיות) שהותקנו בשרתים נגועים , שיתוף פעולה עם גורמים שונים בממשלה וספקי אבטחה עולמיים לסיוע ללקוחות שנפגעו.



https://learn.microsoft.com/en-us/defender-xdr/microsoft-secure-score

בתמונה זו ניתן לראות איך מיקרוסופט מודדת עם האיום על ידי Secure Score ומספקת המלצות שונות להגנה על הארגון מפני איומים - מדידת תנוחת אבטחת זהות, תוכנית שיפורי אבטחה, סקירת השיפורים, שימוש בציון מאובטח.
מדידת הציון נקבעת על פי מספר קריטריונים - השוואה נתונים מול גורמים אחרים בעלי תכונות דומות , ניקוד פעולה לשיפור המערכת בין 1-10 , ניקוד על סמך עדכונים ומשימות אבטחה שגרתיות , הוא משקף את מצב האבטחה הנוכחי לאורך זמן . מטרת הציון היא שיפור ההגנה של הארגון וזיהוי חולשות.

ישנה חשיבות רבה לאימוץ פתרונות הגנה מתקדמים, כמו AI לאיתור איומים, תחזוקה עדכנית של מערכות ותהליכים, ותמיכה במודעות כללית לנושא אבטחת המידע.

מהו בעצם וירוס Stuxnet?

וירוס מחשב מתוחכם שנועד לתקוף את מערכות הבקרה התעשייתיות ובעיקר את מתקני הגרעין האיראני , הוא התגלה ב2010 והיה בין הווירוסים הראשונים שנועדו לפגיעה במערכות פיזיות ולא במחשבים. הווירוס תוכנן בצורה שיוכל להחדיר את עצמו למערכות בקרה תעשייתיות המשתמשות בתוכנות שונות , לשנות את פעולת הציוד הפיזי ולקלקל אותם בהדרגה . נחשב לווירוס המסוכן אי פעם.
אופן פעולה - החדרה למערכות בקרה תעשייתיות ע"י זיהוי פתח פרצה , ושינוי מהירות סיבוב הצנטריפוגות של העשרת האורניום , וכך גרם להם להתקלקל בהדרגה , בנוסף הצלחת הווירוס תוך הסתרת פעילותו מפני מערכות ניטור ובקרה.
השפעה - גרם נזק משמעותי למתקני הגרעין האיראנים והאטת קצת תוכנית הגרעין.
ע"פ –
https://www.tandfonline.com/doi/full/10.1080/18335330.2012.653198

מה זה מעטפות אינטרנט?

מעטפות אינטרנט (Web Shells) הן כלי תוכנה או סקריפטים המאפשרים לתוקפים לקבל גישה מרחוק לשרת שנפרץ, ולעיתים אף שליטה מלאה עליו. מעטפות אלו פועלות על גבי שרת אינטרנט ומאפשרות לתוקף לבצע מס פעולות שונות, כגון:

1. הרצת פקודות מרחוק: הפעלת פקודות על השרת, כאילו התוקף נמצא במערכת עצמה.
2. העלאת והורדת קבצים: הוספת קבצים זדוניים לשרת או גניבת מידע ממנו.
3. גישה לנתונים רגישים: איסוף נתוני משתמשים, סיסמאות, או מסדי נתונים.
4. שמירה על שליטה מתמשכת: יצירת "דלת אחורית" שמאפשרת לתוקף לחזור ולשלוט בשרת בכל עת.

* מעטפות אינטרנט נכתבות לעיתים קרובות בשפות כמו PHP, ASP.NET, או Python, ונראות כמו קבצים לגיטימיים במערכת, מה שמקשה על זיהוין.

קביעת יעדים עסקיים

אבטחת סייבר ורמות חומרתם הם אתגרים מכוננים של זמננו. ארגונים בכל סדר גודל בכל תעשייה ברחבי העולם מרגישים את הדחיפות והלחץ של הגנות מפני התקפות מתוחכמות. בעוד שבינה מלאכותית משנה את אבטחת הסייבר וממעיטה את רמת חומרתם , כלומר בעזרת הבינה המלאכותית אקרים עלולים ללמוד התקפות מתוחכמות ולמצוא בעזרת חומרים רגישים עליהם החברה לא מודעת. השימוש בה לצורכי הגנה דורש כמויות אדירות של נתונים מגוונים. במיקרוסופט, יותר מ-10,000 מומחי אבטחה אשר מנתחים יותר מ-65 טריליון אותות בכל יום בעזרת AI, וצוותי Microsoft Threat Intelligence (צוותי מודיעין של מיקרוסופט) עוקבים אחר מאות איומים, איומים המשתייכים לקבוצות שחקנים ברחבי העולם. המערכת האקולוגית של מיקרוסופט כוללת יותר מ-15,000 שותפי אבטחה עם פתרונות מיוחדים, בעוד שהקהילה הגלובלית של חוקרי ובוחני אבטחה תורמים ומציפים בעיות ואתגרי אבטחה.

הארגון מצפה להרוויח ממדעי הנתונים את היכולת לחזות תקריות אבטחת סייבר ולדעת לסווג את רמת חומרתן ואת רמת החומרה של עדכוני תוכנה משמעותיות ולשפר את היעילות והביצועים של מערכות תגובה מודרכות באמצעות עדכונים . באמצעות ניתוח נתוני אבטחת סייבר אמיתיים או ומידע חיוני על פגיעות ותיקונים במערכות הגנה וניתוח מכוון, הארגון שואף לשפר את יכולות ההתגוננות מפני איומים על ידי פיתוח מודלים חכמים המנבאים את פעולות התגובה הנכונות והיעילות ביותר, תוך הבטחת פעולה נכונה שברוב המקרים (99%) מבלי לגרום להשבתה בטעות של נכסים ארגוניים קריטיים.

פירוט:

65 טריליון : אותות מסונתזים מדי יום.

10,000+ : מומחים לאבטחה ולסיכונים.

4,000 : מתקפות זיהוי שנחסמות בכל שנייה.

300+ : שחקני איום מעוקבים.

100,000+ : תחומים הוסרו.

15,000+ : שותפים במערכת האקולוגית לביטחון.

135 מיליון : מכשירים מנוהלים.

750 מיליון אותות בכל שנייה : מיקרוסופט מנתחת כמויות אדירות של מידע כדי לזהות ולמנוע התקפות סייבר.

10,000+ אנשי מקצוע : כולל מהנדסים, חוקרים, מדעני נתונים, ציידי איומים, אנליסטים גיאופוליטיים, חוקרים ומגיבים בקו הראשון ברחבי העולם.

300+ שחקני איום : כולל 160 מדינות-אומות ,ו- 50 קבוצות תוכנות כופר (בעלי תכונות סחיטה) ומאות אחרים.

100,000+ תחומים הוסרו : שכללו תחומים של ל-Cybercriminals, ו-כללו יותר מ-600 שחקנים שפעלו בשם מדינות-אומות.

15,000+ שותפים במערכת האקולוגית : שותפים אלו מספקים פתרונות שונים שמשפרים את חוסן הסייבר של הלקוחות.

135 מיליון מכשירים מנוהלים: מיקרוסופט מספקת פתרונות אבטחה למספר עצום של מכשירים, ומספקת תובנות ותמיכה.

סיכום:

1. בעיות:

הבנת נתונים: יש קושי בהבנה ובניתוח של כמויות גדולות של נתונים שנאספו.

קבלת החלטות: קיים צורך בשיפור תהליכי קבלת החלטות על בסיס נתונים.

תחרות: התחרות בשוק מחייבת שימוש בטכנולוגיות מתקדמות כדי לייעל תהליכים ולשפר את השירותים.

אבטחת מידע: עם הגידול בכמויות הנתונים, יש צורך בהגנה על המידע ובשמירה על פרטיות הלקוחות.

2. מטרות:

שיפור ביצועים: לנצל את מדעי הנתונים כדי לשפר את הביצועים העסקיים ולהגביר את היעילות.

חיזוי מגמות: להשתמש בניתוח נתונים כדי לחזות מגמות עתידיות בשוק ובצרכים של הלקוחות.

אופטימיזציה של תהליכים: לייעל תהליכים פנימיים בעזרת ניתוחים מתקדמים.

שיפור חוויית לקוח: להבין את צרכי הלקוחות על מנת לשפר את חוויית השירות.

3. משאבים (מה בעצם קיים?):

נתונים: נתונים פנימיים (מכירות, לקוחות) ונתונים חיצוניים (שוק, מתחרים).

כלים טכנולוגיים: תוכנות לניתוח נתונים, מערכות (Business Intelligence-BI) ופלטפורמות ל-Machine Learning.

צוות מומחים: מדעני נתונים, אנליסטים, ומומחים בתחום הטכנולוגי שיכולים ליישם את הפתרונות.

תקציב: משאבים כספיים המוקצים לפיתוח טכנולוגיות ולגיוס עובדים.

4. מידע רקע על המצב העסקי הנוכחי:

תחרות גוברת: השוק מתפתח במהירות, עם חברות חדשות המציעות פתרונות דומים.

שינוי בצרכים: לקוחות מחפשים פתרונות מותאמים אישית ומהירים יותר.

התקדמות טכנולוגית: טכנולוגיות חדשות כמו AI ו-Big Data הופכות לזמינות ונגישות יותר (ChatGPT, BlackBox, Claude ועוד המון תוכנות מבוססות AI אחרות שאפילו מייעדות לתחומים ספציפיים כגון: יצירת מצגות, ניתוח נתונים, יצירת תמונות, ,ווידיאו ופוטושופ מתקדמים ,תכנות קבלת מסקנות והחלטות וכ'ו...).

5. יעדים עסקיים ספציפיים:

הגדלת הכנסות: באמצעות ניתוח נתונים, החברה שואפת לזהות הזדמנויות חדשות למכירה (בדיקת קשרים לצורך הזדמנויות).

חיסכון בעלויות: זיהוי תהליכים לא יעילים ושיפורם באמצעות אוטומציה.

פיתוח מוצרים חדשים: הבנת הצרכים של הלקוחות כדי לפתח מוצרים חדשים המותאמים לשוק.

---

1 **בעיות**	הבעיות בפרויקט כוללות את הצורך להתמודד עם איומי סייבר מתקדמים כמו מתקפות על ענן, דליפות מידע ופרצות אבטחה. חברת מיקרוסופט שואפת להפחית מתקפות סייבר ופריצות למערכות, שמסכנות הן את הלקוחות והן את החברה עצמה, לדוגמה פגיעות וחדירות ל - Windows. כל אלו דורשים פתרונות מתקדמים לעדכון אבטחה, זיהוי והתרעה על מתקפות בזמן אמת.
2 **מטרות**	מטרת הפרויקט היא לשפר את הביצועים העסקיים באמצעות ניתוח נתונים מתקדם, לחזות מגמות עתידיות ולייעל תהליכים פנימיים. בנוסף, יש לשפר את חוויית הלקוח ולהגביר את אבטחת המידע באמצעות חיזוי והתרעה על מתקפות סייבר ופריצות למערכות.
3 **משאבים**	לצורך פיתוח הפרויקט בצורה אופטימלית נדרשים משאבים מרכזיים כמו זמן, טכנולוגיה וכלים לניתוח נתונים. כלים כגון עץ החלטה ומודלים מתקדמים אחרים מסייעים בהבנה והבחנה בתבניות חשודות מתוך הנתונים. כמו כן, יש להקצות זמן מסודר לכל שלב בפרויקט כדי להבטיח מיצוי מלא של כל משאב. בסופו של דבר, המטרה היא להבטיח שכל המשאבים הנדרשים זמינים ומנוהלים בצורה אפקטיבית ויעילה.

רקע עסקי

מיקרוסופט היא חברה בינלאומית המתמקדת בפיתוח, רישוי ותמיכה במגוון מוצרים, שירותים, ומכשירים. החברה מציעה מערכות הפעלה, פתרונות תוכנה עסקיים, כלים לניהול נתונים, מכשירי חומרה כמו קונסולות Xbox, ו-Surface ושירותי ענן מתקדמים דרך Azure. מיקרוסופט גם עוסקת בפיתוח טכנולוגיות בינה מלאכותית ובמתן פתרונות אבטחת סייבר המגנים על מערכות ארגוניות ודיגיטליות.

תחומי פעילות מרכזיים:

1. **תוכנה ושירותים** : כולל מערכת ההפעלה Windows, Office, Dynamics 365, וכלים מבוססי ענן.
2. **מכשירים** : Surface, Xbox ומוצרים נוספים המותאמים לצרכי עבודה ובידור.
3. **טכנולוגיות ענן** Azure : מספקת פתרונות לעיבוד, ניתוח, ואחסון בענן.
4. **אבטחת מידע** : פתרונות AI מתקדמים המיועדים לארגונים לשמירה על נכסים דיגיטליים.

ב-2023, מיקרוסופט שירתה מעל מיליון ארגונים ברחבי העולם, באמצעות פתרונות מתקדמים מבוססי AI שנועדו להגן על נכסים דיגיטליים ולשפר את הביצועים הארגוניים. החברה מדווחת על השקעות נרחבות בתחומי בינה מלאכותית וחדשנות טכנולוגית, כולל שימוש בטכנולוגיות ענן ופתרונות אבטחה בעידן הסייבר המתקדם בנוסף להתפתחות המהירה שמובילה לגידול של פועלי איום שמעלה את האתגרים של מרכזי התפעול האבטחתיים בארגונים. תוכניות כמו **Microsoft AI Cloud Partner Program** מעניקות תמיכה ומשאבים לשותפים עסקיים, ומדגישות את חזון החברה להטמיע AI כבסיס לחדשנות בעתיד.

" All up, more than 1 million organizations now count on our comprehensive, AI-powered solutions to protect their digital estates, and our security business surpassed $20 billion in annual revenue, as we help protect customers across clouds and endpoint platforms. "

הסברים:

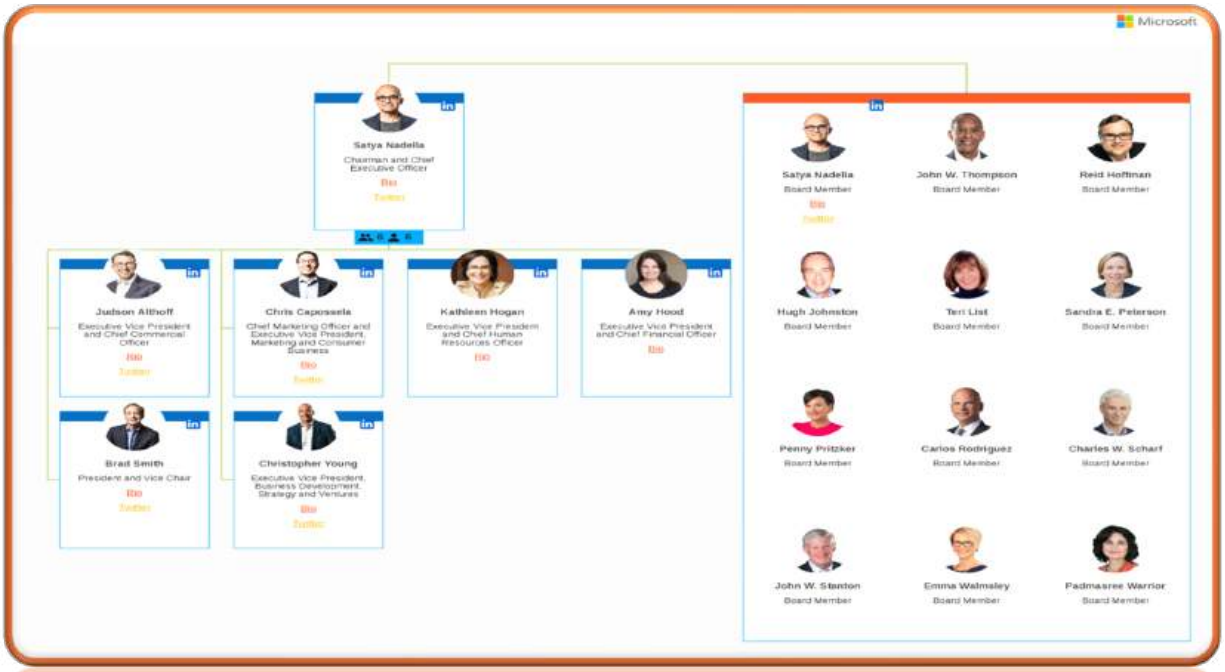
Dynamics 365	ניהול קשרי לקוחות ומשאבים ארגוניים
Surface	סדרת מחשבים ניידים וגם טבלטים מבית מיקרוסופט לצרכים אישיים
Azure	פלטפורמת ענן מבית מיקרוסופט המספקת שירותי מחשוב, אחסון, ניתוח נתונים, AI, ולמידת מכונה.

משימה 1 - קביעת מבנה ארגוני

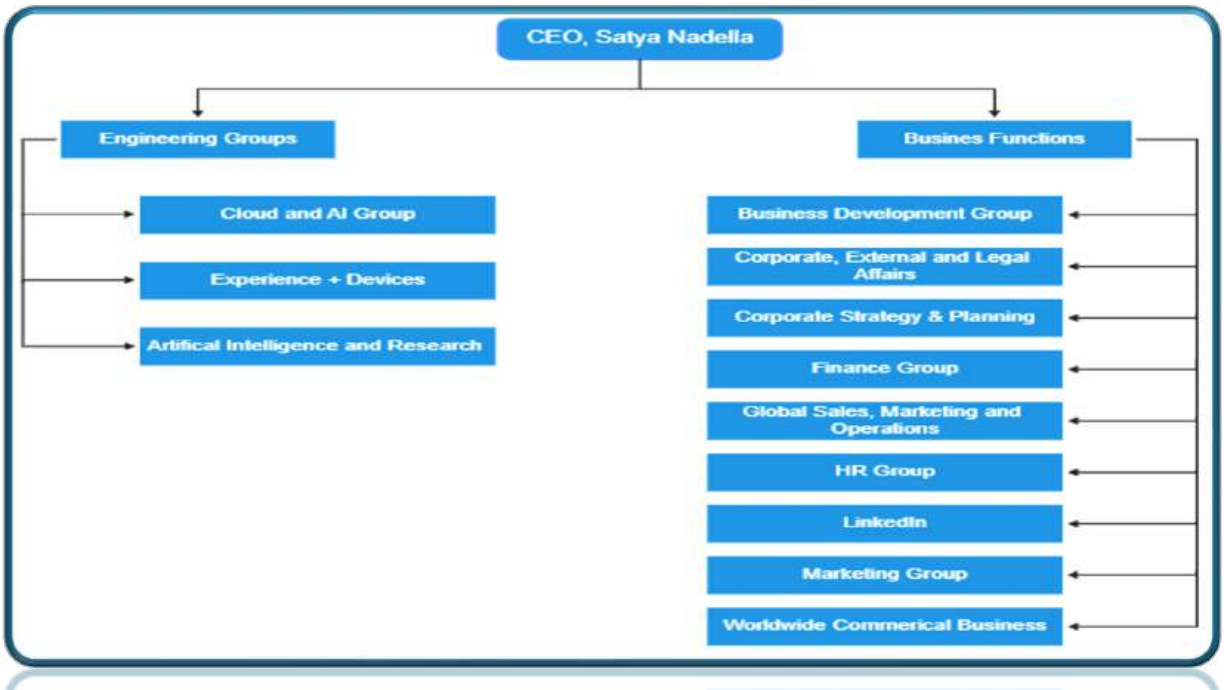
מיקרוסופט פועלת במבנה ארגוני המבוסס חטיבות מוצר (Product-Type Divisional Structure), שבו כל חטיבה מתמקדת במוצר או בשירות המסוים:

1. Windows & Devices: חטיבה המתמקדת במערכות ההפעלה Windows ומכשירי חומרה כמו Surface.
2. Azure & AI: חטיבה המתמקדת בשירותי מחשוב ענן ובינה מלאכותית.
3. Office Product: אחראית על Microsoft Office וכלי פרודוקטיביות נוספים.
4. Gaming: חטיבה המתמקדת ב - Xbox ובתעשיית המשחקים.
5. LinkedIn: פועלת כיחידה עצמאית תחת מיקרוסופט.
6. Security & Compliance: אחראית על פתרונות אבטחת מידע וציות לרגולציות.

אנשי המפתח בארגון:



יחידות עסקיות :



היחידות העסקיות שעלולות להיות מושפעות מפרויקטים במדעי הנתונים הם כולם, וכל יחידה בתחומה עקב המידע והידע הקיים בכל יחידה:

- **היחידות הכלולות בחטיבת ה – Engineering Groups:**

- (קבוצת ענן ובינה מלאכותית), Cloud & AI Group - ביחידה זו עלולים לעדכן ולהוסיף מנגנוני אבטחה לצורך שמירת מידע הענן ולצורך שמירת הבינה המלאכותית שעלולים להשתמש בה לרעה.

- (ניסיון + התקנים), Experience & Devices - בעלי תפקידים אלו ידעו כיצד ליישם מנגנוני אבטחה נוספים בעזרת ניסיונם החומרי והתיאורטי ובכך להקל ולזרז על תהליך העדכון והחידוש הן בקודים/אלגוריתמים והן בחומרה (למשל הגדלת כוח העיבוד, הגדלת נפח האחסון וכו...).

- (בינה מלאכותית ומחקר), Artificial Intelligence and Research - אנשי המחקר בתחום זה יוכלו לבצע ולנתח אלגוריתמים נוספים לצורך חיזוק האלגוריתמים הקיימים, כלומר יכולים להסתמך ולחקור את הנתונים שחקרו בפרויקטים, ולהבין כיצד הגיעו למסקנות ולהחלטות עצם היותם אנשי ידע ומדע, שכן מבינים את מהות פרויקטים אלו.

- **<u>היחידות הכלולות בחטיבת ה – Business Development Group:</u>**

- (קבוצת פיתוח עסקי), Business Development Group - לקבוצה זו עלולים לשלוח מידע על היכן כדאי לפתח את תשתיות העסק והיכן לא, כלומר מניתוח נתונים רבים ניתן לדעת ולחזות על מקומות לא כדאיים ואף מסוכנים שכן העסק שואף לרווח אופטימלי במינימום סיכון.

- (ענייני חברה, חיצוניים ומשפטיים), Corporate, External and Legal Affairs - ניתוח נתונים יוכל להוביל את ענייני החברה לשיקולים פוליטיים-עסקיים ואף במקרים מסוימים גם ביטחוניים, לצורך העניין, בפרויקט זה נעסוק בחיזוי אירועי סייבר על חברת מיקרוסופט, שלבסוף יהיו מסקנות והחלטות עצם ניתוח הנתונים, כלומר לאחר ניתוח הדאטה נדע איזה מסקנה ואיזה החלטה היא הנכונה לארגון (ע"פ ה Dataset שהוצאנו על הארגון)

- (אסטרטגיה ותכנון חברה), Corporate Strategy & Planning - ניתוח נתונים יכול להמליץ ואף לציין על מקומות עסקיים כדאיים, לצורך העניין ניתוח נתונים על לקוחות, מדינות וכו... יכול לראות בברור את הכדאיות העסקית הנכונה לארגון. שיקול אסטרטגי נכון יכול להתקבל על סמך אותם נתונים שנותחו, לדוג' אם נבחרה מדינה בה החברה מעוניינת לקום, אך ע"פ ניתוח שוק וניתוח דאטה, אותו אזור מסוכן לה, אז יהיה אפשר למנוע שיקול אסטרטגי מוטעה.

- (קבוצת הכספים), Finance Group - לקבוצת הכספים עלולה להיכנס פרויקט חדש עליו יצטרכו להקצות תקציבים, כלומר, פרויקט אבטחה נוסף / פרויקט מוצר נוסף שנותח על סמך נתונים כבעל משמעות ובעל הזדמנות עסקית, כדאי עד מאוד לפיתוחו שכן אותה חברה תוכל להניב רווח גדול ממנו.

- (מכירות גלובלי, שיווק ופעילות), Global Sales, Marketing and Operations - לניתוח שווקים, לקוחות, משתמשים וחברות השפעה רבה על החברה, אנליסטים ומדעני נתונים יכולים להסיק מסקנות ולקבל החלטות נכונות על אותם הפונקציות. חקר הידע והמדע על תחומים אלו יכולים לחזות אילו צעדים יהיו נכונים ואילו לא, כלומר אם ניקח דאטה סט

על כל פונקציה שונה ונבדוק האם קיים קשר בניהם נוכל לנבא האם כדאי להשקיע או לא כדאי, האם קיים קשר בין שתי הפונקציות או לא קיים, וכך נוכל לקחת החלטה מושכלת.

- (קבוצת משאבי אנוש), HR Group - לקבוצת משאבי האנוש תהיה השפעה על אופן קבלת ומיון האנשים לארגון, כלומר אם פרויקט מסוים עסק בניתוח אנשים בארגון או בנזילת אינפורמציה מהארגון (אפרופו חיזוי תרחישי סייבר), יוכל הפרויקט לשמש אותם בקבלת ההחלטות הנוגעות לקבלת עובדים חדשים לארגון, לדוג', האם קיים קשר בין מדינה שבה מתבצעת הרבה מתקפות סייבר לעובדי החברה?.

- (הרשת החברתית), LinkedIn - היחידה תוכל לבצע סינון ומחיקת משתמשים זדוניים באופן מהיר ויעיל מחיזויי אירועי תקיפה, כלומר אם יש xמשתמשים ולכמה מהם יש קשר לאירועי תקיפה בסייבר יוכלו אותם מוקדי אבטחה להיזהר ואף להסיר חשבונות אלו.

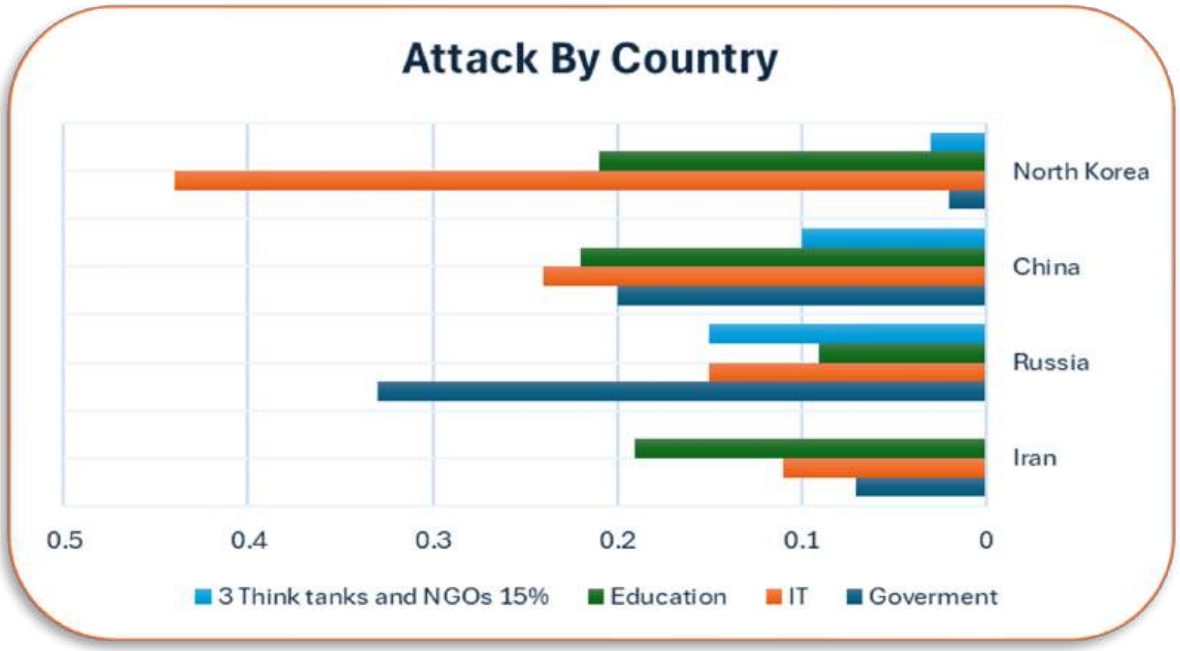
- (קבוצת שיווק), Marketing Group - משווקים יוכלו לשווק בבטחה במקומות שבהן יש כדאיות כלכלית ואהדה לאותה החברה. כיום אירועי שיווק יקרים מאוד, וכהשקעה ראשונית הארגון ירצה לראות את המקסימום רווחיות מהם, ניתוח דאטה על מדינות, ערים, משתמשים, דירוגים, ומנויים (שלא רק משתמשים אלא גם קונים), יוכל להוות פונקציה חשובה בעת יישום השיווק לאותו המקום, כלומר מקבלת החלטה מושכלת תהיה אפשרות לקבל את יעדי רווח השיווק שכן אליו ניבא אותו חיזוי.

- (עסקים מסחריים ברחבי העולם), Worldwide Commercial Business - כחברה בין לאומית, הרצון הוא לבצע כמה שיותר עסקים ולהגדיל את רווחי ההון העצמי שלה באופן מושכל. כלומר, ביצוע עסקים על הגלובוס הוא חשוב לכל חברה בין לאומית אך יש צורך בניתוח וקריאת הנתונים שכן כל התרחבות של כל חברה היא השקעה עצומה. פרויקטי דאטה יכולים לסייע במתן קריאת הנתונים והשוק לאותם המדינות אליהם החברה רוצה להתרחב. בניתוחי נתונים אנו יכולים לנבא את אחוזי ההצלחה פר מדינה וכמובן שע"פ דאטה סט איכותי, בסיום הרצת האלגוריתם על

הנתונים, נוכל לקבל מסקנה והחלטה מושכלת האם התרחבות זו נכונה או לא וכמה סיכויי הצלחה יש לה.

משימה 2 - תיאור אזור בעיה

הבעיה המרכזית היא בתחום **אבטחת הסייבר**. חברת מיקרוסופט מתמודדת עם כמות הולכת וגדלה של איומי סייבר מורכבים, כולל ניסיונות פריצה, התקפות כופר וגניבת מידע. בעיות אלו משפיעים על אמינות השירותים שלה, פגיעות בלקוחות וגורם להפסדים כספיים רבים.



ניתן לחזות ולהתמודד באופן פרואקטיבי עם התקפות סייבר באמצעות שימוש בטכנולוגיות של מדעי הנתונים ובינה מלאכותית.

האתגר כולל:

- זיהוי דפוסים חשודים בהתנהגות משתמשים.
- גילוי מוקדם של פעולות זדוניות ברשת.
- קבלת החלטות בזמן אמת על בסיס ניתוח נתונים רחב היקף.

- מיקרוסופט משתמשת במדעי הנתונים כבר כיום, אך היא שואפת לשפר את היכולת לנבא איומים עתידיים בצורה מדויקת יותר. הנתונים קיימים במערכות כמו Azure Sentinel ומבוססים על פעילות מיליוני משתמשים.

- הפרויקט דורש שילוב של מומחים מתחומים כמו אבטחת מידע, מדעי נתונים ותפעול מערכות.

- מניע הפרויקט הוא הגנה על לקוחות ושירותים קריטיים, חיזוק אמון הלקוחות בחברה, והפחתת עלויות הנובעות מהתקפות סייבר.

- חברת מיקרוסופט מכירה ומודעת לתרחישים סביב מדעי הנתונים, וכתוצאה מכך מעסיקה אלפי אנליסטים ואלפי מדעני נתונים לצורך חקר הידע והמידע. דבר שיכול לעזור לקידום והבנה של הפרויקט, שכן ימשוך תשומת לב רבה בקרב מהנדסי הנתונים ומדעני הנתונים.

הסבר מהו Azure Sentinel?

Azure Sentinel הוא פתרון ה SIEM (Security Information and Event Management) ותזמורת אבטחה, אוטומציה ותגובה (Security Orchestration, Automation and Response, SOAR) של מיקרוסופט. עם Azure Sentinel עסקים יכולים לאסוף, לנתח ולהגיב את/על הנתונים ממספר המקורות ולתת לארגונים הבנה מלאה על סביבת האבטחה שלהם.

משימה 3 - תיאור הפתרון הנוכחי

תיאור הפתרונות המשמשים כעת לטיפול בבעיה העסקית :

1. מערכות Security Information and Event Management (SIEM) : מיקרוסופט עושה שימוש במערכות SIEM כמו Azure Sentinel, המאפשרות איסוף, ניתוח, וניטור נתוני אבטחת סייבר ממקורות מגוונים בזמן אמת.

2. פתרונות Endpoint Detection and Response (EDR) : מערכות כמו Microsoft Defender מספקות הגנה מתקדמת על תחנות קצה תוך זיהוי ותגובה מהירים לאיומים.

3. בינה מלאכותית ומודלים לחיזוי : מיקרוסופט מטמיעה AI ו- Machine Learning על מנת לחזות התקפות על סמך דפוסי נתונים היסטוריים ואנומליות (וחריגים).

4. מרכזי תגובה לאירועים Security Operations Center (SOC) : מיקרוסופט מפעילה צוותי SOC המספקים ניטור רציף ותגובה מיידית לאירועי סייבר.

יתרונות וחסרונות של הפתרונות הנוכחיים:

יתרונות:

- **תפיסה פרואקטיבית:** מערכות AI וחיזוי מאפשרות מניעה מוקדמת של התקפות.
- **יכולת סקיילביליות (**היכולת של מערכת לעמוד בעומס שהולך וגדל**) :** הכלים של מיקרוסופט מתמודדים עם כמויות נתונים אדירות ומתאימים לארגונים גדולים.
- **שילוביות:** הפתרונות עובדים בסנכרון עם שאר המוצרים של מיקרוסופט, כגון Azure ו- Office 365.

חסרונות:

- **תלות במערכות מורכבות:** יישום וניהול של פתרונות מתקדמים דורש מומחיות טכנית גבוהה.

- **עלויות גבוהות:** רכישת רישיונות, תחזוקה, וצוותי SOC כרוכים בעלויות משמעותיות.
- **זמן תגובה:** בעוד AI משפר את הזיהוי, ייתכנו עיכובים במקרים שבהם המודלים מתמודדים עם תוקפים המשתמשים בטכניקות חדשות. כלומר במרכזי SOC מערכות מבוססות AI עוזרות לזהות איומים על בסיס דפוסים קיימים או התנהגויות חריגות, ובכך משפרות את המהירות והדיוק בזיהוי מתקפות סייבר. עם זאת, ישנם מקרים שבהם תוקפים משתמשים בטכניקות חדשות שטרם נלמדו או שולבו במודלי ה-AI.

רמת קבלה בארגון:

מיקרוסופט מתמקדת באבטחת סייבר כמרכיב אסטרטגי מרכזי בפעילותה, ומשלבת פתרונות טכנולוגיים מתקדמים כמו Azure Sentinel ו- Microsoft Defenderבארגונים. עם זאת, כדי למקסם את היעילות, נדרש גם חינוך והסברה למשתמשים בארגון, כדי להתמודד עם סיכונים הנובעים מטעויות אנושיות ומתקפות כמו "דיוג".

<u>הסבר מהו "דיוג" (Phishing)?</u>

"דיוג" (Phishing) היא שיטת תקיפה במרחב הסייבר שמטרתה להערים על משתמשים ולגרום להם לחשוף מידע רגיש, כגון סיסמאות, פרטי כרטיסי אשראי או פרטי גישה לחשבונות. התוקפים מתחזים לגורם אמין ומוכר, כמו בנק, ספק שירותים או ארגון, כדי לזכות באמון הקורבן.

<u>סיכום:</u>

חברת מייקרוסופט בעלת מספר מרובה של משאבים חומריים ואנושיים כאחד על מנת להשיג את מטרותיה. מייקרוסופט שמה דגש חזק המתמקדת בגיוס ופיתוח אנשי מפתח בתפקידים שונים להבטחת ביצוע יעיל. התשתית הטכנולוגית בעלת מגוון גבוהה של מרכזי נתונים ושירותם בייחוד Azure. ניתן להצביע על אתגר בולט במיוחד, מיקרוסופט מתמודדת עם אתגרים מתמשכים הקשורים להגנת נתונים ואיומי אבטחת סייבר. בין היתר מטרותיה של החברה הן חדשנות , שירותי ענן , חווית לקוח , הגנת נתונים , פיתוח עובדים ועוד .

---

1 מבנה הארגוני של החברה בנוי כחטיבות , המתעסקות בתחומים שונים.
החטיבות העיקריות הן צריכה , ענן , AI ופיתוח .

2 יש זיהוי של בעיית הפריצות וסייבר על חברת מייקרוסופט, מתקפות כל מרכזי נתונים.
המגמה לפריצות ותקפות שונות לא מרפה ומהוות איום , פוגעת במוניטין ועוד.

מייקרוסופט מציעה הגנה באימות והצפנה , צוות ייעודי למלחמה בסייבר , ניתוח איומים והעלאת מודעות. פתרונות אלה בעלי יתרון של הגברת הגנה , מניעת כניסה זרה , נטרול רשתות פשע סייבר , תגובה לאיומים ועוד. עם האלגוריתם חיזוי ניתן יהיה לחזות כמעט במאת האחוזים את התקיפה הבאה ולהיערך אליה באופן המיטבי ביותר.

יעדים עסקיים וקריטריונים להצלחה

יעדים עסקיים וקריטריונים להצלחה הם הבסיס להצלחת כל פרויקט, במיוחד כשמדובר במדעי הנתונים. היעדים העסקיים מגדירים את המטרה המרכזית של הארגון, כמו שיפור יעילות, הגדלת הכנסות, או צמצום סיכונים. מטרות אלה מהוות את הבסיס שעל פיו יפותחו הפתרונות הטכנולוגיים והמודלים המדעיים בפרויקט. מנגד, קריטריונים להצלחה הם המדדים המאפשרים לבדוק אם הושגה המטרה, באמצעות נתונים כמותיים ואיכותיים ברורים.

יעדים עסקיים במדעי הנתונים באים לידי ביטוי בחיבור בין הצרכים הארגוניים לטכנולוגיה. לדוגמה, במערכת לזיהוי מתקפות סייבר, היעד העסקי עשוי להיות צמצום עלויות הנזק והתייעלות צוותי האבטחה. קריטריוני ההצלחה, לעומת זאת, יגדירו במדויק מה נחשב הצלחה טכנית, כמו דיוק של מעל 90% בזיהוי מתקפות, זמן תגובה מהיר, או חיסכון כלכלי משמעותי כתוצאה מהשימוש במערכת.

השפעתם על הארגון היא רחבת היקף. מצד אחד, היעדים העסקיים מספקים כיוון ברור, המניע את הצוותים הטכנולוגיים להתרכז בפתרונות הרלוונטיים ביותר. מצד שני, הקריטריונים להצלחה מבטיחים שכל מאמץ מדעי הנתונים ייבחן על פי ערך אמיתי ותועלת כלכלית, ולא רק על פי חדשנות טכנולוגית. כאשר מדעי הנתונים פועלים מתוך ראייה עסקית ברורה, הם הופכים לכלי אפקטיבי שמשפר את תהליכי קבלת ההחלטות, מייעל את העבודה, וממקסם את הערך הארגוני.

בסופו של דבר, חיבור נכון בין יעדים עסקיים לקריטריונים להצלחה יוצר מערכת יעילה יותר. זה מאפשר לארגונים להתאים את המודלים המדעיים למציאות העסקית, להקטין סיכונים ולמנף את הפתרונות כדי ליצור יתרון תחרותי בשוק.

בנוסף, עדכוני אבטחה מהווים מרכיב קריטי בהשגת היעדים העסקיים והקריטריונים להצלחה בפרויקטים של מדעי הנתונים, במיוחד כאשר מדובר בהגנה על מערכות נתונים ותקשורת. עדכוני אבטחה תכופים וממוקדים מסייעים לשמור על המערכות מוגנות מפני איומים חדשים, ומבטיחים שהפתרונות הטכנולוגיים לא יהיו פגיעים לניצול על ידי תוקפים. בכך, הם תורמים לצמצום סיכונים, שיפור יעילות צוותי האבטחה, והבטחת שלמות הנתונים – כל אלו מהווים מדדים חשובים להצלחה, ומסייעים לארגון להישאר תחרותי בשוק.

יעדים עסקיים

היעד העסקי של הפרויקט הוא פיתוח מערכת חיזוי מתקדמת שתספק ניתוח בזמן אמת לזיהוי ומניעת התקפות סייבר ורמותיה על מערכות מיקרוסופט, תוך שיפור ביצועים עסקיים, שמירה על חוויית משתמש ושילוב אינטגרטיבי עם פתרונות קיימים. ניתוח ניתוח נתונים יאפשר לחזות את רמת הפגיעות ואת רמת החומרה שנפגעה לפי מאפיינים כמו המדדים הבאים: תאריך פרסום, מזהה בליטין (מזהה של הודעת האבטחה), דרגת חומרה, השפעה, כותרת עדכון האבטחה, המוצר המושפע, הרכיב שנפגע , רכיב המושפע, אתחול ו-CVEs (זיהוי פגיעות ידועות למעקב אחר בעיות ידועות).

<u>תיאור הבעיה</u>

מיקרוסופט מתמודדת עם איומים סייבריים מגוונים המנסים לפגוע במערכותיה ובמשתמשיה. המטרה היא לחזות ולזהות בזמן אמת אירועי סייבר באמצעות ניתוח נתונים קיימים מהתשתיות שלה, במטרה למנוע התקפות ולעצור את ההתפשטות שלהן.

<u>שאלות עסקיות מכוונות</u>

- אילו סוגי איומים סייבריים פוגעים לעיתים קרובות ביותר במערכות מיקרוסופט?
- האם קיימים דפוסי התנהגות שמובילים לזיהוי מוקדם של התקפות?
- מהו פרופיל התוקפים?
- אילו יחידות בארגון נמצאות בסיכון הגבוה ביותר ומהן נקודות החולשה?
- כיצד ניתן לשפר את רמת ההגנה מבלי לפגוע בביצועי המערכות או בחוויית המשתמש?

<u>דרישות עסקיות שונות</u>

- שיפור דיוק חיזוי להתקפות סייבר ב-15% תוך שמירה על זמני תגובה מהירים.
- פיתוח מודל המאפשר ניטור בזמן אמת ללא האטת ביצועי המערכות.
- מניעת פגיעה באמון המשתמשים על ידי הגנה חזקה ובלתי נראית.

- שילוב מערכת ההגנה החדשה עם פתרונות קיימים של Microsoft Defender.
- הגברת הגנה על מערכות חומרה והפחתת פגיעות באמצעות עדכונים אוטומטיים ומדויקים, כך שהמשתמשים לא ירגישו בהשפעה של עדכונים אלה.

<u>יתרונות צפויים</u>

- הפחתת הנזק הפיננסי: חסכון בעלויות הנגרמות כתוצאה מפריצות (כגון שיקום מערכות וקנסות).
- שיפור אמון לקוחות: צמצום פרצות המובילות לפגיעה בפרטיות המשתמשים.
- ייעול המשאבים הארגוניים: ניתוח נתונים אוטומטי המפחית את התלות באנליסטים אנושיים.
- שימור לקוחות: שיפור תדמית המותג כבטוח ויעיל, מה שימנע נטישת לקוחות
- שיפור ביצועים: עדכוני תוכנה משפרים את ביצועי המערכת, ומונעים בעיות טכניות כמו איטיות או תקלות.
- מניעת תקלות עתידיות: מערכות מעודכנות מונעות בעיות מבוססות ממערכות ישנות.

קריטריונים להצלחה

קריטריונים להצלחה:

1. **מטרה (קריטריונים אובייקטיביים):**
   - שיפור דיוק המודל לזיהוי אירועי סייבר לפחות ב-15%, המשמעות היא להבטיח שהמודל שמנתח את אירועי הסייבר ינבא נכון יותר אירועים קריטיים. זה נמדד באמצעות מדדים סטנדרטיים כמו דיוק, שלמות (recall), ודיוק ממוצע (F1 Score). לדוגמה, במודל שמזהה מתקפות, דיוק גבוה אומר שפחות התקפות מוחמצות.
   - הפחתת כמות ההתראות השגויות (false positives) ב-25%, כמות התראות שגויות פוגעת ביעילות של צוותי אבטחה. אם הפרויקט יכול להקטין את התופעה הזו ב-25%, המשמעות היא יעילות גבוהה יותר וזמן עבודה טוב יותר.
   - קיצור זמן העיבוד של נתונים למקסימום 5 דקות לאירוע.

2. **קריטריונים סובייקטיביים:**
   - זיהוי דפוסים בלתי צפויים באירועי סייבר המשפרים את יכולת ההגנה של הארגון, המדע מבוסס על היכולת לגלות מידע לא צפוי, כמו דפוסים חדשים שמחמיאים או משפרים את האבטחה.
   - שביעות רצון גבוהה מצוותי אבטחת המידע והמנכ"ל על השימושיות והערך העסקי של הכלים שפותחו, בשונה ממדדים טכניים, שביעות רצון נמדדת על ידי משוב מבעלי עניין שמשתמשים בכלים שנבנו.
   - היכולת לשלב את המערכת החדשה בתהליכים קיימים ללא שינויים מהותיים.

רשימת משימות:

1. **תיעוד קריטריונים:**
   - לוודא שהמטרות העסקיות מוגדרות וברורות.
   - לקשר כל מטרה לקריטריון מדיד להצלחה.

2. **<u>תיאום עם בעלי עניין:</u>**
   - להיפגש עם בעלי עניין (למשל מנהלי IT, הנהלה) ולאשר את המדדים הסובייקטיביים.
   - לרשום הערות בנוגע לציפיות ולדרישותיהם.

3. **<u>בדיקות מדידה:</u>**
   - לערוך בדיקות כדי לוודא שהמדדים ניתנים למדידה בצורה אפקטיבית.

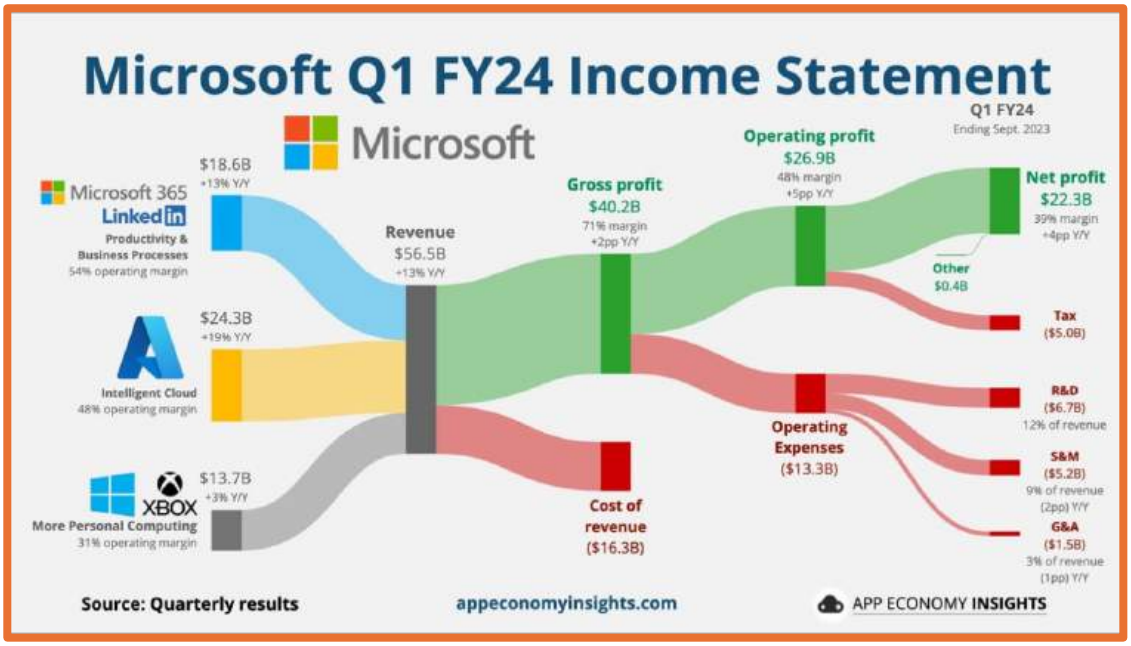
4. **<u>מעקב שוטף:</u>**
   - להקים מערכת דיווח שתנתח באופן קבוע את הצלחת הפרויקט בהתאם לקריטריונים שנקבעו.

<u>סיכום :</u>

תיעוד קריטריונים : להגדיר מטרות בצורה מסודרת ולאפשר לעובדים להבין מה נחשב להצלחה.

תיאום עם בעלי עניין : מתן אפשרות למנהלים ובעלי החלטות להשתתף בהגדרת הציפיות והמדדים.

מעקב ובדיקות : שימוש בכלים מדידים כדי להעריך את הצלחת הפרויקט לאורך זמן.



Source: media.licdn.com

הערכת מצב

סוג נתונים זמינים:

הנתונים המתקבלים מקורם במרכז ההורדות של מיקרוסופט, שם זמינים קבצי Excelתחת הכותרת ,"Microsoft Security Bulletin Data" המפורסמים להורדה לציבור. נתונים אלו מספקים מידע מפורט על פגיעות אבטחה שזוהו במוצרים שונים של מיקרוסופט, כולל פרטי תיקונים, גרסאות המושפעות, והשלכות של בעיות אבטחה. המידע שבקבצים מהווה מקור חשוב למעקב אחרי עדכוני אבטחה ולטיפול בפגיעויות מוכרות ,(CVEs) תוך הצגת השפעות בעיות האבטחה על רכיבים שונים במערכת. נתונים אלו משמשים ככלי חשוב עבור מנהלי אבטחה, המאפשר להם לעקוב אחרי איומים פוטנציאליים וליישם עדכונים שוטפים על פי הצורך.

תיאור הנתונים:

הנתונים בקובץ מציגים מידע מקיף על עדכוני אבטחה עבור רכיבי תוכנה במערכות ההפעלה שלMicrosoft . כל שורה בקובץ מכילה את פרטי העדכון, כולל מזהה עדכון (Bulletin Id), קוד בסיס הידע של העדכון (KB), דרגת חומרת הפגיעות (Severity), השפעת הפגיעות (Impact), מערכת ההפעלה המושפעת (Affected Product), ורכיב התוכנה או החומרה שנפגעו (Affected Component). נוסף על כך, הנתונים כוללים פרטים על דרישה להפעיל מחדש את המערכת (Reboot) וקישורים לזיהוי פגיעויות ספציפיות (CVE). הנתונים מכסים מגוון רחב של מערכות הפעלה, כגון : Windows 7, 8.1, 10 וגרסאות שונות של Windows Server, כאשר רוב העדכונים מתייחסים לפגיעויות קריטיות כמו "Remote Code Execution" (ביצוע קוד מרחוק), אשר מאפשרות לתוקפים לגשת למערכת בצורה לא מורשית.

החשיבות של ניתוח הנתונים:

ניתוח הנתונים מספק הבנה מעמיקה על דפוסי העדכונים והפגיעויות במערכות ההפעלה של Microsoft. ניתוח כזה חשוב עבור מספר סיבות עיקריות :

1. **מניעת תקיפות:** פגיעויות קריטיות, כגון ביצוע קוד מרחוק (Remote Code Execution), עשויות להוות סיכון חמור למערכות ולארגונים. ניתוח הנתונים מאפשר לזהות את העדכונים הקריטיים ביותר ולוודא שהם מיושמים במועד הנכון, כדי למנוע תקיפות שעלולות לנצל את הפגיעויות.

2. **שיפור מדיניות האבטחה:** ניתוח מעמיק של הפגיעויות מאפשר לארגונים לשפר את מדיניות האבטחה שלהם ולמנוע סיכונים עתידיים. זהו כלי חשוב בשמירה על רמות אבטחה גבוהות לאורך זמן.
3. **זיהוי מגמות:** ניתוח כמותי של הנתונים יכול לחשוף מגמות בעדכוני האבטחה, למשל, אם יש עלייה במספר הפגיעויות בתחום מסוים, מה שמסייע למקד את המשאבים והמאמצים בניהול הסיכונים.
4. **תיעוד וציות:** עבור ארגונים גדולים, ניתוח ויישום העדכונים באופן שיטתי מהווה חלק מהדרישות הרגולטוריות בתחום אבטחת המידע. המידע יכול לשמש כתיעוד חיוני עבור בקרות פנימיות ודו"חות ניהוליים.

ניתוח זה מספק למומחי אבטחת מידע תמונה ברורה ומפורטת של מצב האבטחה במערכותיהם, ועוזר להבטיח סביבה בטוחה, מוגנת ומעודכנת.

מטרת הנתונים:

מטרת הנתונים המפורסמים היא לאתגר את קהילת מדעי הנתונים לפתח טכניקות לחיזוי אירועי סייבר עתידיים. הנתונים נועדו לשפר את המודלים לחיזוי תקריות סייבר ולסייע בפיתוח מערכות תגובה מונחות עבור מרכזי אבטחת מידע (SOCs), המאפשרות סינון ותגובה אוטומטיים לאירועים.

מאפיינים עיקריים:

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

זמן יצירת הנתונים וטווח זמנים:

הנתונים נאספו בין השנים 2008 ל-2017, כאשר פורסמו ב - 2017. תאריכי העדכונים מוצגים בטבלה בפורמט של ערכים מופרדים בנקודה ובמקף, ומייצגים את הזמן המדויק שבו זוהתה הפגיעות או פורסם עדכון האבטחה.

מטרות השימוש בנתונים:

לפתח ולבחון מודלים של למידת מכונה **לחיזוי רמות תקריות סייבר 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 וצורך בעדכונים נוספים. לספק מדד סטנדרטי לשיפור מערכת **התגובה המונחית** (GR) לפי רמה ולבחון את יכולות ההתמודדות של מערכות אלו ברמות שונות.

פרטי המהדורה:

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 מוקפדת: מזהים אישיים עוברים גיבוב ומוחלפים למזהים אקראיים. המידע נועד לתמוך בקהילת המחקר בתחום אבטחת המידע, תוך קידום פיתוח מערכות חיזוי ותגובה אוטומטיות לאירועי הסייבר.

מרכיבי הקבצים:

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

קישור למחקר - קישור להורדת הנתונים

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

צוות הפרויקט:

הערכת מיומנות הצוות לקיום והקדשת זמן ומשאבים ליצירת מודל אולטימטיבי מתאים, בניית המודל תהיה על סמך איסוף , ניקוי (Cleaning Data) , הפרדה ומידול נתונים. צוות הפרויקט מורכב מ – 2 סטודנטים למערכות מידע המתמחים במדעי הנתונים בשנתם האחרונה ומנחה מתחום השיווק דיגיטלי, אוטומציית שיווק וניתוח נתוני שיווק ואינטליגנציית שיווק.

גורמי הסיכון המעורבים:

גורמי סיכון יכולים לשבש את יכולת הניהול של הנתונים בצורה טובה ולהפריע בעת ביצוע המשימה , זיהוי והברה מוקדמת של סיכונים אלו יכולה ואף תשמש בעתיד את הצוות למציאת נקודות חולשה ותיקנם האפשרי.

בין הסיכונים האפשריים ניתן למצוא:

- איכות ירודה של נתונים ומציאת נתונים חסרים.
- מורכבות הנתונים.
- דיוק מודל פחותה.
- מטרות מעורפלות (לא ברורות) בין חברי הצוות.
- ביג דאטה(?) (עלול לקחת זמן לפירוקו ותלוי עד כמה הצוות מנוסה לזאת).
- חוסר ידע (עלולים להיתקל בחוסר ידע, כגון: ספריות חדשות, מבנה קוד חדש וכ"ו).
- עמודת תאריכים לא מסודרת , מופרדת לפי מקף וחלקם נקודה.

תוכניות מגירה:

תוכניות חלופיות או תוכניות מגירה שיכולות לשמש כפתרון, עבור כל סיכון יהווה אבן דרך בהתגברות על נקודות החולשה ותאפשר ביצועים טובים יותר בעתיד.
להלן תוכניות המגירה עבור כל סיכון:

- תהליך בקרת איכות הנתונים הגבוה וטיהור הנתונים.
- שימוש בכלים מתקדמים והתמודדות עם ערכים חסרים.
- ניתן להשתמש במודלים חלופיים וחזרה עקבית על תהליך האלגוריתם להעלאת הדיוק.
- וידוא דרישות ברורות והבנה מדויקת של המטרה של הפרויקט .
- שימוש בכלים מתקדמים לצורך פירוק הביג דאטה באופן מהיר ויעיל (כדוגמות שפות התכנות: Python, R, Excel PowerBI, וכ'ו).

את חוסר הידע ניתן להשלים דרך תוכנות ה – AI המתקדמות כיום בשוק, בנוסף במקור הנתונים קיימים דיונים על אופן ניתוח הדאטה ובו כל משתמש/אורח משתף את הידע שלו ואף מעלה קוד כדוגמא.

סיכום מבנה הדאטהסט: (סיכום טכני)

ראיות (Evidence): הרמה הבסיסית ביותר בנתונים. כל ראיה תומכת בהתראה ויכולה לכלול פרטי מידע כמו אתחול , תכונה מחליפה , 2 סוגי השפעות , חומרת פגיעה ועוד...
התרעות (Alert): אגרגציה (תהליך של חיבור או סיכום נתונים ממספר מקורות או ערכים, במטרה להפיק תוצאה כוללת או תובנה מדומיינת) של מספר ראיות שמצביעות על איום או תקלת אבטחה פוטנציאלית.
תקלות (Incident): הרמה הגבוהה ביותר, המאגדת אחת או יותר התרעות ליצירת תמונה מקיפה של איום או תקלה ורמת החומרה .

תכונות עיקריות:

טריאז' (קביעת סדר העדיפויות) של תקלות: הדאטהסט בנוי כדי לחזות את דרגת החומרה של התקלות.

למידת מכונה: הדאטהסט יכול להוות יעד לחיזוי משתנה מטרה של רמת החומרה או חומרה .

שימושים פוטנציאליים:

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

מדדי הערכה:

הדאטהסט כולל כ – 23,517 אלף תצפיות וכ – 14 מאפיינים, והוא לא מחולק לסט אימון וסט בדיקה. המדד העיקרי להערכה של מחקרים שמתבצעים עם הדאטהסט הוא ה-macro F1 score, כולל מדדים 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 ומחולק לראיות, התרעות ותקלות, שמסייעות בזיהוי תקלות ואיומים אבטחתיים. הוא מאפשר חיזוי דרגות חומרה ואוטומציה בתגובה לאירועי סייבר, ומיועד לשימוש במודלים של למידת מכונה. המדד העיקרי להערכת המודלים הוא macro F1 score, עם מדדים נוספים כמו precision ו-recall.
הדאטהסט זמין לשימוש תחת רישוי Community Data License Agreement – Permissive – Version 2.0.

מלאי משאבים

משימה 1 - משאבי חומרה

1. **תמיכה בחומרה:**

מאחר והנתונים נלקחו ממאגר ,ההורדות של מיקרוסופט , יש צורך בחומרה חזקה כדי לעבד את הנתונים ובחירה נכונה של אמצעי למידת מכונה, כולל:

- **מעבד (CPU) :** לפחות 8 ליבות לתמיכה במודלים מורכבים (יחידת העיבוד המרכזית של המחשב, המכונה גם "מעבד").
- **זיכרון (RAM) :** מינימום 32 GB, עם העדפה ל-64 GB, כדי להתמודד עם עומסי עיבוד (זיכרון הגישה האקראית של המחשב – הזיכרון המיידי).
- **כונן אחסון SSD :** בנפח 512 GB לפחות, כדי לטעון ולקרוא נתונים במהירות (כונן אחסון מתקדם המבוסס על זיכרון פלאש).
- **מעבד גרפי (GPU) :** מודלים מתקדמים דורשים GPU עם לפחות 8 GBזיכרון (מעבד גרפי, רכיב חומרה שאחראי על עיבוד הגרפיקה, וויזואליזציה של תמונות, וידאו וגרפיקה תלת-ממדית במחשב).

אם התשתית המקומית אינה מספיקה, ניתן לשקול שימוש בענן Azure, AWSאוGoogle Cloud .

משימה 2 - זיהוי מקורות נתונים ומאגרי ידע

1. **מקורות נתונים זמינים:**
   - הדאטהסט - מכיל נתוני אבטחת סייבר , חומרה ועדכונים אמיתיים עם מבנה היררכי.
   - הנתונים זמינים בפורמט סטנדרטי בטופסי Excel
2. **אחסון וגישה:**
   - הנתונים מאוחסנים וזמינים להורדה.
   - חשוב להעלות אותם למחסן נתונים כגון Azure Data Lake (אחסון נתונים מבית מיקרוסופט, המיועד לאחסון נתונים גדולים Big)

(Data) או כלי מקומי עם אפשרויות גישה חיה כגון : SQL או NoSQL.

3. **רכישת נתונים חיצוניים:**
   - הדאטהסט מכיל נתונים שיכולים לשמש לשיפור המודלים לחיזוי רמות חומרת התקלה ועדכוני אבטחה. הקריטריונים המרכזיים כוללים מידע על תאריך פרסום, מזהה התראה, חומרה והשפעה של התקלה, סוג המוצר המושפע, ופרטי) CVEs פגיעות נפוצות). נתונים אלו יכולים להעשיר את המודלים ולסייע במיפוי ומיון תקלות על פי קריטריונים של חומרה והשפעה, מה שיאפשר פיתוח מערכות תגובה מונחות ומודלים לחיזוי פעולות שיקום ותיקון.
4. **בעיות אבטחה:**
   - אין בעיות אבטחה מובנות מכיוון שהנתונים עברו תהליך אנונימיזציה והחלפת מזהים.
   - יש לוודא שהגישה למידע מבוקרת ומאובטחת, במיוחד בשימוש בענן.

<u>משימה 3 - זיהוי משאבי כוח אדם</u>

1. **מומחי עסקים ונתונים:**
   - מומחים בתחום אבטחת המידע (SOC) יכולים לספק הקשר עסקי למידע.
   - מדעני נתונים אחראים על עיבוד הנתונים ואימון המודלים.
   -
2. **מנהלי מסד נתונים וצוות תמיכה:**
   - נדרשים מנהלי DB עם ידע ב - SQL ובכלי אחסון בענן, לדוגמה Azure Synapse (מחסן נתונים מבוסס ענן, המאפשר אחסון וניהול נתונים בעוצמה גבוהה ובעלות נמוכה יחסית).
   - צוות תמיכה טכני שיכול לתחזק את חומרת העיבוד המקומית או לקשר בין השרתים בענן.

דרישות הנחות ואילוצים

משימה 1 - קביעת דרישות

הגבלות אבטחה ומשפטיות:

הגבלות על הנתונים:

הנתונים פורסמו תחת רישיון ההפצה איתם עובדים מיקרוסופט (https://cdla.dev/permissive-2-0/), המאפשר שימוש למחקר ופיתוח, אך יש להבטיח שמירה על תנאי הרישיון, כולל:

מניעת חשיפה של מידע שעלול לזהות ארגונים או יחידים.

עמידה בחוקי פרטיות כמו **GDPR** באירופה או חוקים מקומיים.

השימוש בטכניקות כמו פיסו-אנונימיזציה והחלפת מזהים אקראיים מבטיח הגנה מסוימת, אך מומלץ לבצע בדיקות כדי לוודא שאין סיכוני זיהוי מחדש.

הגבלות על התוצאות:

תוצאות המחקר עשויות לדרוש אישור נוסף אם מתוכננת הצגה פומבית או שיתוף עם גורמים חיצוניים.

תאימות לדרישות תזמון הפרויקט:

שלבי הפרויקט מותאמים למסגרת הזמן שנקבעה.

איסוף נתונים והכנתם – (Data Cleaning) שלב זה עשוי לקחת זמן רב אם איכות הנתונים נמוכה.

אימון מודלים – (Model Training) במיוחד עם מודלים גדולים, עשוי לקחת ימים או שבועות תלוי במשאבי החישוב.

הערכת המודל (Evaluation) ותיקונים – יש לקחת בחשבון זמן לביצוע שיפורים.

על כל שלב להיות מתועד ולוודא התאמה ללוח הזמנים הכולל.

דרישות לפריסת תוצאות:

פרסום באינטרנט:

התוצאות צריכות להיות מוצגות לציבור. יש להבטיח שהפורמטים יהיו תואמים לכלים כמו **Power BI** או **Tableau**.

קריאת ציונים במסד נתונים:

נדרש לאחסן או לשלוף נתוני תחזיות ממסד נתונים, יש לוודא תמיכה בפורמט הנתונים במסד הנתונים הקיים, כמו **SQL Server** או **Azure Data Lake**.

גישה מאובטחת:

יש להבטיח שאופן הפריסה עונה על דרישות אבטחת המידע של הארגון, כמו שימוש ב - **Azure Key Vault** לניהול גישה.

<u>משימה 2 - הבהרת הנחות</u>

גורמים כלכליים שעשויים להשפיע על הפרויקט:

עלויות ייעוץ:

אם מערבים יועצים חיצוניים, יש לוודא שתעריפי הייעוץ מתאימים לתקציב הפרויקט.

לדוגמה, אם נדרש ייעוץ מתמחים בלמידת מכונה או ניתוח נתונים, יש להעריך עלויות ולבדוק אפשרות לשלב מומחיות פנימית של הארגון.

מוצרים תחרותיים:

ייתכן שמיקרוסופט מתחרה במוצרים או פתרונות קיימים של חברות אחרות בתחום הSOC - , זה עשוי לדרוש התאמות בתכנון התוצאה, כדי לבדל את הפתרון ולשפר את ההיתרון העסקי.

שימוש במשאבים:

חשוב להבין האם יש צורך בשדרוגים טכנולוגיים (לדוגמה, כוח חישוב גבוה יותר בעזרת הענן) שיכולים להוביל לעלויות נוספות.

הנחות לגבי איכות הנתונים:

הנתונים עשירים ומובנים היטב, אך יש לבדוק:

שלמות נתונים : האם כל התקלות, החומרות ששומשו והראיות מספקות תמונה מלאה, או שיש חסרים הדורשים טיפול?

עקביות : לבדוק האם הנתונים תואמים בפורמט, תיוגים (Labels) וזמנים.

טיפול בבעיות : נדרש לבנות תהליך לניקוי נתונים, כמו טיפול בחסרים או סילוק רשומות בלתי תקינות.

ציפיות נותן החסות/צוות הנהלה:

הבנת המודל עצמו:

נותן החסות מעוניין להבין כיצד המודל פועל, יש לספק דוח מפורט – מבנה המודל (Features, Hyperparameters).

דוגמאות של תחזיות מוצלחות וכישלונות (False Positives/Negatives).

תוצאות בלבד:

אם המיקוד הוא בתוצאות, נדרש להציג דוחות ברורים וגרפיים, למשל בדשבורד אינטראקטיבי ,ב Power-BI או ב Excel כך שיציג:

דירוגי תקלות (Severity Ratings) .

תחזיות פעולות שיקום מוצלחות.

מגמות ושיפורים ב – SOC.

בכל אופן, מאחר ומדובר בחיזוי רמות תקיפות סייבר וחומרה , רצוי להראות גם את המודל וגם את התוצאות.

<u>משימה 3 - אימות אילוצים</u>

גישה לנתונים:

יש את כל הסיסמאות הנדרשות לגישה לנתונים.

צריך לוודא כי לכל חברי הצוות קיימים הנתונים, כלומר אם הצליחו להוריד אותם (יש מקרים בהם קיימת הגבלת הורדה לפי בקשת המפרסם).

מגבלות משפטיות על השימוש בנתונים:

אימות המגבלות המשפטיות על השימוש בנתונים.

הנתונים מופצים תחת רישיון שמאפשר שימוש למחקר ופיתוח. עם זאת, חשוב לוודא:

שהשימוש עומד בתנאי הרישיון, כולל מניעת זיהוי מחדש של נתונים.

שאין סתירה עם חוקי פרטיות מקומיים, כמו **GDPR** באירופה או **חוק הגנת הפרטיות** בישראל.

אם הפרויקט מערב שיתוף תוצאות, נדרש לבדוק תנאי פרסום המותרים על ידי מיקרוסופט.

מגבלות פיננסיות ותקציב:

המגבלות הפיננסיות מכוסות בתקציב הפרויקט.

עלויות צפויות :

שימוש בשירותי ענן כמו **Azure Synapse** או **Azure Machine Learning Studio**.

כוח חישוב (GPU/CPU) לצורך אימון מודלים.

כלים תומכים כמו Power BI להצגת תוצאות.

סיכונים ומקריות

סיכון 1: תזמון

בעיה אפשרית:

הפרויקט עשוי להימשך זמן רב מהצפוי בשל עיכובים באיסוף הנתונים, ניתוחם, או בעיות טכניות במודלים. (לדוג' : ב – Big Data ניכר עומס הזמן לניתוח הנתונים לעד הפקת מסקנות והחלטות, כלומר מאחר ונתון מאגר נתונים עצום (למעלה מכמה מיליוני נתונים) ייקח זמן לעד ניתוחו.

השפעה:

עיכובים עלולים להוביל להחמצת תאריכים קריטיים ולפגיעה בלוחות הזמנים העסקיים של נותן החסות.

תוכנית מגירה:

- הגדרת לוחות זמנים ריאליים מראש עם מרווחי ביטחון.
- חלוקה לשלבי עבודה עצמאיים כדי שניתן יהיה לספק תוצרים חלקיים.
- תגבור צוותי עבודה בשעות נוספות או גיוס יועצים זמניים בשעת הצורך.

סיכון 2: פיננסי

בעיה אפשרית:

נותן החסות עשוי להיתקל בקשיים תקציביים, מה שעלול לגרום להפסקת מימון או צמצום היקף הפרויקט.

השפעה:

חוסר במימון יכול לעכב או להפסיק את העבודה לפני השלמתה.

תוכנית מגירה:

- זיהוי של שלבי מפתח בפרויקט שניתן להשלים בעלות נמוכה אך בעלי ערך גבוה.
- תכנון אפשרות לצמצם את היקף הפרויקט תוך שמירה על הערך המרכזי.
- בדיקת מימון אלטרנטיבי, כמו גיוס תקציב נוסף דרך קרנות או שותפים.

סיכון 3: נתונים

בעיה אפשרית:

הנתונים עשויים להיות באיכות נמוכה (למשל עם - חוסרים, שגיאות, או רעשים) או בעלי כיסוי חלקי שאינו מייצג את הבעיה העסקית (כלומר שהנתונים לא כוללים את כל המידע הנדרש להבנת הבעיה העסקית או לפתרונה בצורה המדויקת.)

השפעה:

נתונים באיכות ירודה Missing Data, Underrepresentation, Outdated Data, Irrelevant data (Biased Data) עלולים להוביל לתוצאות לא מדויקות או לחוסר יכולת לבנות מודל מתאים.

תוכנית מגירה:

- שימוש בטכניקות ניקוי נתונים ושחזור ערכים חסרים.
- גיוס מקורות נתונים נוספים או השלמה ידנית לפי הצורך.
- תיעוד מגבלות הנתונים והשפעתם על הפרויקט.

סיכון 4: תוצאות

בעיה אפשרית:

התוצאות הראשוניות עלולות להיות פחות דרמטיות (מעניינת) או בעלות ערך נמוך מהמצופה, מה שיגרום לאכזבה מצד נותן החסות.

השפעה:

זה עשוי לערער את התמיכה בפרויקט ולפגוע במוניטין הצוות.

תוכנית מגירה:

- הסבר מפורט מראש על מגבלות הפרויקט והמודלים.
- עבודה אינטראקטיבית עם נותן החסות כדי למקד את המטרות לאורך הדרך.
- שיפור המודל על ידי הוספת נתונים, תיקון הנחות, או שימוש בטכניקות מתקדמות יותר.

מה זה אומר נתונים חסרים/חלקיים ?

נביא כמה דוגמאות והסברים על מנת שנבין יותר טוב את הדברים.

1. נתונים חסרים (Missing Data)

- **מה זה אומר ?**חלק מהשדות או הערכים החיוניים אינם מתועדים במערכת.

- **השפעה :**ייתכן שחלק מהשאלות העסקיות לא יקבלו מענה, כי אין מספיק מידע כדי לבצע ניתוח מלא.
- **דוגמה :**ניתוח תקלות ייצור, מנגד חוסר נתונים על זמני ההשבתה של המכונות, הניתוח שלך יהיה חלקי בלבד.

2. חוסר בכיסוי (Underrepresentation)

- **מה זה אומר ?**הנתונים הזמינים אינם מייצגים את כלל האוכלוסייה או התהליך שנחקר.
- **השפעה :**תוצאות המודל או התובנות עלולות להיות מוטות ולא אמינות.
- **דוגמה :**בדיקת בעיות באיכות מוצר, מנגד הנתונים נאספו רק מקו ייצור אחד, המסקנות לא בהכרח תקפות לגבי כל קווי הייצור.

3. נתונים מיושנים (Outdated Data)

- **מה זה אומר ?**הנתונים לא עדכניים ואינם משקפים את המציאות הנוכחית.
- **השפעה :**ייתכן שהמסקנות שלך לא יהיו רלוונטיות להחלטות עסקיות עדכניות.
- **דוגמה :** התבססות על נתונים ישנים , התובנות לא יהיו מועילות.

4. נתונים לא רלוונטיים

- **מה זה אומר ?**הנתונים אינם קשורים ישירות לבעיה.
- **השפעה :**זמן רב מושקע.
- **דוגמה :**אם מטרתך לחקור תקלות במכונות, אבל יש לך רק נתונים על משמרות העובדים, קשה יהיה להסיק מסקנות לגבי ציוד.

5. הטיה בנתונים (Biased Data)

- **מה זה אומר ?**הנתונים שנאספו מייצגים דפוס מסוים בלבד ולא את כלל המקרים.
- **השפעה :**הפתרונות שיגזרו מהנתונים עלולים להעדיף קבוצה מסוימת או להתעלם מקבוצות אחרות.
- **דוגמה :**אם הנתונים מגיעים רק ממפעל אחד מתוך רשת של מפעלים, ייתכן שהם לא משקפים את האתגרים של כלל הרשת.

כיצד מטפלים בכל השגיאות והחוסרים הללו?

1. **איסוף נתונים נוסף:** ננסה לקבל גישה למקורות מידע נוספים שמייצגים טוב יותר את הבעיה העסקית.
2. **תיעוד מגבלות הנתונים:** נתעד בדיוק איפה יש חוסרים או בעיות, כדי שנוכל להתחשב בכך בפרשנות התוצאות.
3. **בחירת מדדים חלופיים:** אם חסר מידע מסוים, ננסה למצוא מדדים קרובים שיכולים לעזור לנו להעריך את הבעיה, כלומר מדדים קרובים הם משתנים או נתונים שניתן להשתמש בהם באופן עקיף כדי להעריך את המידע החסר או להבין את התופעה העסקית בצורה עקיפה.
4. **שימוש בטכניקות לניהול נתונים חסרים:** למשל, השלמת נתונים חסרים (imputation) או שימוש במודלים שמתגברים על חוסר מידע.

טרמינולוגיה

בפרק זה נעסוק במילון מונחים על מנת שנוכל להבטיח שצוותים עסקיים ומדעני נתונים יוכלו להבין את מטרת הפרויקט, ביצועו ואופן פעילותיו.

כעת נציג את המונחים מתחום הסטטיסטיקה ,מתחום למידת המכונה ומתחום הלמידה העמוקה שעם חלקם נעשה שימוש:

1. **סטיית תקן (Standard Deviation)** – מדד לפיזור הנתונים סביב ממוצע.
2. **ממוצע (Mean)** – סכום כל הערכים מחולק במספר הערכים.
3. **שונות (Variance)** – ריבוע של סטיית התקן, מדד נוסף לפיזור הנתונים (רעש).
4. **רגרסיה לינארית (Linear Regression)** – שיטה למידת מכונה לניבוי ערך רציף על בסיס קשר לינארי בין משתנים.
5. **קשר רגרסיה (Regression Coefficient)** – מדד המצביע על עוצמת הקשר בין משתנים ברגרסיה לינארית.
6. **בדיקת השערות (Hypothesis Testing)** – תהליך סטטיסטי לקביעת תקפות השערות על בסיס נתונים (H1, H0).
7. **p-value** – הוא מדד בסטטיסטיקה המשמש בבדיקות השערות כדי להעריך את המובהקות הסטטיסטית של התוצאות.
8. **מבחן t-test**– מבחן סטטיסטי לבדוק אם יש הבדל משמעותי בין שתי קבוצות.
9. **היסטוגרמה (Histogram)** – תרשים המשמש להצגת התפלגות הנתונים.
10. **הפרשנות השגויה (Bias)** – הטיה במודל או בממצאים.
11. **אלגוריתם (Algorithm)** – סדרת שלבים שמבוצעים כדי לפתור בעיה מסוימת או לבצע משימה.
12. **מודל (Model)** – ייצוג של מערכת או תהליך שיכול להיות בשימוש לפתרון בעיות, חיזוי או הבנה של נתונים.
13. **למידת מכונה (Machine Learning)** – תחום בלמידה המאפשר למערכות ללמוד מנתונים ולבצע תחזיות או החלטות ללא התערבות ישירה.
14. **למידה מפוקחת (Supervised Learning)** – שיטה בלמידת מכונה שבה המודל מאומן על ידי נתונים עם תוויות (labels).
15. **למידה לא מפוקחת (Unsupervised Learning)** – שיטה בלמידת מכונה שבה המודל לומד לזהות מבנים או קבוצות בנתונים ללא תוויות.

16 **מודל רגרסיה (Regression Model)** – מודל שמטרתו לחזות ערך רציף.

17 **מודל סיווג (Classification Model)** – מודל שמטרתו לסווג נתונים לקטגוריות.

18 **דיוק (Accuracy)** – מדד שמראה את שיעור ההצלחה של המודל בניבוי נכון של התוויות.

19 **Precision (דיוק)** – מדד המראה את שיעור הניבויים הנכונים מתוך כלל הניבויים החיוביים.

20 **Recall (היזכרות)** – מדד המראה את שיעור המקרים החיוביים שנמצאו מתוך כלל המקרים החיוביים האמיתיים.

21 **F1-Score** – מדד שמשלב את ה-Precision וה-Recall, כדי לספק תמונה כוללת של הביצועים.

22 **Overfitting (התרגלות יתר)** – כאשר המודל מתאים את עצמו יותר מדי לנתוני האימון ומאבד את יכולתו כלפי נתונים חדשים.

23 **Underfitting (התאמה לקויה)** – כאשר המודל אינו מצליח ללמוד את הקשרים בין הנתונים בצורה מספקת.

24 **Cross-Validation (אימות צולב)** – טכניקת הערכה שבה הנתונים מחולקים לסטים שונים לאימון ובדיקה, כדי להבטיח יציבות המודל.

25 **תכונה (Feature)** – מאפיין או משתנה בנתונים המהווה חלק מהקלט של המודל.

26 **הפחתת ממדיות (Dimensionality Reduction)** – טכניקות כמו PCA המפחיתות את מספר המאפיינים בנתונים תוך שמירה על המידע החשוב.

27 **קבוצת אימון (Training Set)** – הקבוצה שבה המודל מתאמן ולומד.

28 **קבוצת בדיקה (Test Set)** – הקבוצה שבה המודל נבדק על מנת להעריך את ביצועיו.

29 **Normal Distribution (התפלגות נורמלית)** – התפלגות סטטיסטית בה רוב הנתונים מתרכזים סביב הממוצע.

30 **Correlation (קורלציה)** – מדד המראה את עוצמת וכיוון הקשר בין שני משתנים.

31 **Outlier (נתון חריג)** – נתון השונה באופן משמעותי מהנתונים האחרים בקבוצה.

32 **Gradient Descent (ירידת גרדיאנט)** – אלגוריתם אופטימיזציה למציאת ערכים אופטימליים של פרמטרים במודל.

33 **Loss Function (פונקציית אובדן)** – פונקציה שמודדת את הטעות של המודל במידת ההתאמה לנתונים.

34 **Over-sampling and Under-sampling (דגימה חוזרת ודגימה מצומצמת)** – טכניקות לניהול מחסור או יתר של נתונים בקבוצות אימון.

כעת נציג את האלגוריתמים מתחום למידת המכונה ומתחום הלמידה העמוקה שעם חלקם נעבוד:

1. אלגוריתמים של למידת מכונה מפוקחת (Supervised Learning)

- **רגרסיה לינארית (Linear Regression) -** שיטה לניבוי ערך רציף.
- **לוגיסטית רגרסיה (Logistic Regression) -** שיטה לסיווג בינארי.
- **עץ החלטה (Decision Tree) -** מודל החלטה שמבצע חלוקה של הנתונים בהתבסס על תנאים.
- **יער אקראי (Random Forest) -** קבוצת עצי החלטה שמבצעים תחזיות משולבות.
- **K-Nearest Neighbors (KNN's)** - שיטה לסיווג או ניבוי, בהתבסס על הקרבה למשתנים אחרים.
- **Naive Bayes** - מודל מבוסס הסתברות לסיווג.
- **Support Vector Machine (SVM)** - מודל סיווג המבוסס על מציאת הגבול האופטימלי בין קבוצות.
- **XGBoost / LightGBM / CatBoost -** אלגוריתמים מתקדמים של עץ החלטה המבצעים תחזיות מדויקות יותר.
- **רשתות נוירונים (Neural Networks)** - מודלים המספקים פתרונות לשורות נתונים מורכבות, כמו סיווג תמונות וניתוח טקסט.

2. אלגוריתמים של למידת מכונה לא מפוקחת (Unsupervised Learning)

- **אלגוריתם K-Means -** טכניקת חלוקה לקבוצות (Clustering).
- **אלגוריתם Hierarchical Clustering -** שיטה נוספת לחלוקה לקבוצות המבוססת על היררכיה.
- **אלגוריתם DBSCAN -** שיטה לחלוקה לקבוצות שמבוססת על צפיפות.
- **Principal Component Analysis (PCA) -** טכניקת הפחתת ממדיות (Dimensionality Reduction).

- **t-SNE** - טכניקת הפחתת ממדיות שמיועדת להצגת נתונים בצורה ויזואלית.
- **Isolation Forest** - אלגוריתם לאיתור נתונים חריגים (Outlier Detection).

3. אלגוריתמים של למידת חיזוק (Reinforcement Learning)

- **Q-Learning** - אלגוריתם למידת חיזוק שבו הלמידה מתבצעת על ידי למידת ערכים של פעולות.
- **Deep Q-Networks (DQN)** - הרחבה של Q-Learning באמצעות **רשתות נוירונים** עמוקות.
- **Policy Gradient Methods** - שיטות שבהן המודל לומד את המדיניות המיטבית ישירות.
- **Actor-Critic Methods** - שיטות המשלבות רשתות נוירונים לשם למידת מדיניות וערכים בו זמנית.

4. אלגוריתמים של למידה עמוקה (Deep Learning)

- **רשתות נוירונים מלאות (Fully Connected Networks)** - רשתות בהן כל נוירון מחובר לכולם בשכבה הקודמת.
- **Convolutional Neural Networks (CNNs)** - רשתות נוירונים שמיועדות במיוחד לבעיות של עיבוד תמונה.
- **Recurrent Neural Networks (RNNs)** - רשתות שמיועדות לבעיות של סדרות-זמן (Time Series) כמו ניתוח טקסט או קול.
- **Long Short-Term Memory (LSTM)** - גרסה מתקדמת של RNN, המתמודדת עם בעיות של זיכרון לטווח ארוך.
- **Transformer Models** - מודלים מתקדמים לשפה, כמו BERT ו-GPT המיועדים לעיבוד טקסט.

כעת נציג את הספריות שעם חלקם נעבוד:

1. ספריות לעיבוד נתונים וסטטיסטיקה:

- **Pandas -** ספריה לעיבוד נתונים וניתוחם, נוחה מאוד לעבודה עם DataFrames.
- **NumPy** - ספריה מתמטית המספקת פונקציות מתקדמות לעבודה עם מערכים.
- **SciPy** - ספריה המרחיבה את NumPy עם פונקציות מתקדמות לסטטיסטיקה, אינטגרציה, פתרון משוואות ועוד.
- **Statsmodels -** ספריה לסטטיסטיקה מתקדמת הכוללת מודלים לניתוח רגרסיה, מבחני סטטיסטיקה וניתוחים רבים נוספים.
- **Seaborn** - ספריה ליצירת גרפים ויזואליים עם יכולות מתקדמות להצגת נתונים בצורה אסתטית.
- **Matplotlib -** ספריה ליצירת גרפים ותרשימים פשוטים וגלויים.
- **Sympy-Plotting -** ספריה להצגת גרפים תלת ממדיים וקשר בין 3 משתנים.

2. ספריות למידת מכונה:

- **scikit-learn -** אחת הספריות הפופולריות ביותר לביצוע משימות של למידת מכונה, כולל רגרסיה, סיווג, קיבוץ ועוד.
- **XGBoost -** ספריה לאימון מודלים מתקדמים של עץ החלטה וגרדיאנט (Gradient Boosting).
- **LightGBM -** גרסה מהירה יותר של XGBoost, יעילה במיוחד בעבודה עם כמויות נתונים גדולות.
- **CatBoost -** ספריה נוספת למידול מבוסס עץ שמיועדת להתמודד עם נתונים קטגוריאליים ביעילות.
- **TPOT -** ספריית אוטומציה של למידת מכונה (AutoML) המיישמת חיפוש פרמטרים אוטומטי ומספקת מודלים אופטימליים.
- **H2O.ai -** כלי ללמידת מכונה שמספק גם פתרונות AutoML וגם כלים לביצוע ניתוחים מתקדמים.

3. ספריות ללמידה עמוקה:

- **TensorFlow -** ספריה עוצמתית של Google לבניית מודלים של למידה עמוקה, פופולרית מאוד בשימוש בתעשייה ובאקדמיה.
- **Keras -** ספריה עם ממשק פשוט לבניית רשתות נוירונים, כיום היא משולבת עם.TensorFlow
- **PyTorch -** ספריה של Facebook המציעה גמישות רבה וקלות בשימוש לבניית מודלים של למידה עמוקה.
- **Theano -** ספריה ישנה יותר (כיום לא מתוחזקת) שעדיין נמצאת בשימוש במקרים מסוימים לבניית רשתות נוירונים.
- **MXNet -** ספריה של Apache, מתאימה לבניית מודלים גדולים מאוד ותומכת במודלים מבוזרים.
- **Fastai -** ספריה מבוססת על PyTorch המפשטת את תהליך הבנייה של מודלים מתקדמים בלמידה עמוקה.

4. ספריות לניתוח תמונה וראייה ממוחשבת:

- **OpenCV -** ספריה מאוד פופולרית ויעילה לביצוע עיבוד תמונה וראייה ממוחשבת.
- **Pillow -** ספריה לעיבוד תמונה, המהווה גרסה משודרגת של הספריה המפורסמת.PIL
- **scikit-image -** ספריה לניתוח ועיבוד תמונה מבית.scikit-learn
- **ImageAI -** ספריה לבניית מודלים של ראייה ממוחשבת עם פשטות גבוהה.

5. ספריות לעיבוד שפה טבעית (NLP):

- **NLTK (Natural Language Toolkit) -** ספריה מעשית לעיבוד שפה טבעית, כוללת כלים רבים להערכת טקסטים, ניתוחים וכו.'
- **spaCy -** ספריה לעיבוד שפה טבעית עם דגש על מהירות ודיוק.
- **TextBlob -** ספריה פשוטה המיועדת לניתוח טקסטים ולהפקת משמעויות בסיסיות.
- **Transformers (by Hugging Face) -** ספריה מתקדמת המכילה מודלים לשפה, כגון BERT ו - GPT, שמאפשרת אימון מודלים מתקדמים ויישום פתרונות NLP.

6. ספריות לניהול פרויקטים וניתוח נתונים:

- **Dask -** ספריה המאפשרת עיבוד נתונים גדולים על ידי חלוקה לחלקים קטנים שמאוחדים לתוצאה סופית.
- **Vaex -** ספריה מעולה לניתוח נתונים גדולים (Big Data) בצורה אינטראקטיבית ומהירה.
- **PySpark -** חיבור של PySpark לעבודה עם Apache Spark, עוזר לעבודה עם נתונים בקנה מידה גדול.

כעת נציג את התכונות הקיימות בדאטהסט שעם חלקם נעבוד:

- אז כידוע מדובר בדאטה שמיקרוסופט פרסמו (לפני מס' שנים, ב 2017), בו נמצאים כ – 14 מאפיינים/תכונות ו – 23,517 תצפיות. את אופן ניתוח התכונות ביצענו ע"י שימוש ב – AI, כלומר נתנו למערכת הבינה המלאכותית לקרוא את הנתונים, לאחר מכן ביקשנו שתפרט לנו אודות השורה הראשונה, כלומר שורת התכונות. לאחר שפירטה את התכונות, ביקשנו שתכתוב זאת ישירות לטבלת האקסל שאותה ייעדנו לטבלת ניתוח מאפיינים/תכונות הנתונים.
- נשים לב ללוגים ולתכונות השונות הקיימות בטלה.

	A	B	C	D
1				
2	Feature Name	Definition	Description	
3	Date Posted	The date when the bulletin was published.	Indicates when the security update or bulletin was officially released.	
4	Bulletin Id	The unique identifier for the security bulletin.	Helps in referencing and categorizing security updates.	
5	Bulletin KB	Knowledge Base (KB) number associated with the bulletin.	Links to detailed technical documentation about the update.	
6	Severity	The criticality level of the update.	Defines the importance of applying the update, such as Critical, Important, etc.	
7	Impact	The type of vulnerability the update addresses.	Specifies whether the vulnerability impacts Remote Code Execution, Denial of Service, etc.	
8	Title	The title of the security bulletin or update.	Provides a brief description of the update or its purpose.	
9	Affected Product	The product impacted by the vulnerability.	Lists the operating systems, applications, or services requiring the update.	
10	Component KB	The Knowledge Base (KB) number for the affected component.	Specifies the technical documentation for the impacted component.	
11	Affected Component	The specific component affected by the vulnerability.	Details the part of the product or service that is vulnerable.	
12	Impact.1	Secondary impact description for the vulnerability.	Further clarifies the vulnerability's effect, if needed.	
13	Severity.1	Secondary severity level for the update.	Provides an additional severity classification, if applicable.	
14	Supersedes	The bulletin or update that this one replaces.	Indicates updates that are deprecated or no longer applicable.	
15	Reboot	Whether a reboot is required after applying the update.	Indicates if the system needs to restart to complete the update installation.	
16	CVEs	Common Vulnerabilities and Exposures identifiers.	Lists standardized IDs for vulnerabilities addressed by the update.	
17				
18	שם תכונה	הגדרה	תאור	
19	תאריך פרסום	התאריך שבו פורסם העלון.	מציין מתי עדכון האבטחה או העלון שוחררו באופן רשמי.	
20	מזהה עלון	המזהה הייחודי של עלון האבטחה.	עוזר בהפניה וסיווג עדכוני אבטחה.	
21	עלון KB	מספר מאגר הידע (KB) המשויך לעלון.	קישורים לתיעוד טכני מפורט על העדכון.	
22	חומרה	רמת הקריטיות של העדכון.	מגדיר את החשיבות של יישום העדכון, כגון קריטי, חשוב וכו'.	
23	פְּגִיעָה	סוג הפגיעות שבה העדכון מטפל.	מציין אם הפגיעות משפיעה על ביצוע קוד מרחוק, מניעת שירות וכו'.	
24	כּוֹתֶרֶת	הכותרת של עלון האבטחה או העדכון.	מספק תיאור קצר של העדכון או מטרתו.	
25	מוצר מושפע	המוצר שהושפע מהפגיעות.	מפרט את מערכות ההפעלה, היישומים או השירותים הדורשים עדכון.	
26	רכיב KB	מספר מאגר הידע (KB) עבור הרכיב המושפע.	מציין את התיעוד הטכני עבור הרכיב המושפע.	
27	רכיב מושפע	הרכיב הספציפי המושפע מהפגיעות.	פירוט החלק של המוצר או השירות הפגיע.	
28	השפעה.1	תיאור השפעה משנית עבור הפגיעות.	מבהיר עוד יותר את השפעת הפגיעות, במידת הצורך.	
29	חומרה.1	רמת חומרה משנית עבור העדכון.	מספק סיווג חומרה נוסף, אם רלוונטי.	
30	מחליף	העלון או העדכון שהאחד הזה מחליף.	מציין עדכונים שיצאו משימוש או שאינם ישימים עוד.	
31	לְאַתחֵל	האם נדרש אתחול מחדש לאחר החלת העדכון.	מציין אם המערכת צריכה להפעיל מחדש כדי להשלים את התקנת העדכון.	
32	CVEs	מזהי פגיעויות וחשיפה נפוצות.	מפרט מזהים סטנדרטיים עבור נקודות תורפה המטופלות על ידי העדכון.	
33				

עלויות ותועלות

עלויות משוערות:

1. איסוף נתונים וכל מידע חיצוני בו נעשה שימוש -

- מקורות נתונים : הכוונה כאן היא לנתונים שנכנסים למערכת (במקרה הזה, למערכות של מיקרוסופט כמו Azure Sentinel, הנתונים הללו יכולים לכלול התראות סייבר, פעילות רשת, נתונים ממערכות אבטחה כגון Microsoft Defender ועוד. כל שימוש בנתונים אלו עלול להיות כרוך בעלויות אחסון ועיבוד גבוהות.
  - לדוגמה, עלויות השימוש ב Azure Sentinel- כוללות אחסון נתונים ועיבוד נתונים לצורך ניתוחים בזמן אמת.
  - עלות משוערת $2,500 – $1,500 לחודש ,עלות זו תלויה בנפח הנתונים שנאגר במערכת ובסוג הנתונים.
- ניקוי והעשרת נתונים : התהליך כולל את הצעד שבו הנתונים שייכנסו למערכת צריכים להיות מאורגנים וממוינים בצורה שתאפשר להם להיות שימושיים. מדובר בתהליך של "ניקוי" נתונים (מחיקת המידע שאינו רלוונטי) ו"עשרת" הנתונים (הוספת פרטים או עיבוד לשם ביצוע חיזויים מדויקים יותר.
  - עלות משוערת $10,000 – $5,000 לפרויקט ,מדובר בעלויות של אנשי מקצוע כמו מדעני נתונים שמתכננים את ניקוי והעשרת הנתונים.

2. פריסת תוצאות -

- מודלים בענן : הכוונה כאן היא להטמיע את המודלים (מודלים של חיזוי מתקפות סייבר באמצעות למידת מכונה) על תשתית ענן כמו Azure, פריסה זו כוללת את הפעולה של הפעלת המודלים בסביבה חיה בה ייבדקו ויתבצעו חיזויים בזמן אמת.
  - עלות משוערת $3,000 – $2,000 לפריסה ראשונית ,כאן מדובר בעלויות של פריסת המודלים במערכות ענן כמו Azure Machine Learning.
- אינטגרציה עם מערכות קיימות : אחרי שפרסנו את המודלים, יש צורך לשלב אותם עם מערכות קיימות כמו Microsoft Defender כדי לקבל תוצאות באופן אוטומטי וישיר. זה כולל את פעולת החיבור בין המודלים לבין הפלטפורמות הקיימות בארגון.

- עלות משוערת $5,000 – $3,000 ,אינטגרציה כזו דורשת זמן עבודה של צוותים טכנולוגיים והבנה עמוקה של המערכות השונות.

3. עלויות תפעול -

- תחזוקה ועדכונים : כדי לשמור על המודלים ולטפל בשינויים בזמן אמת, יש צורך בעדכונים שוטפים של המודלים, זאת משום שהסכנות בתחום אבטחת המידע משתנות כל הזמן.
  - עלות משוערת $2,000 – $1,000 לחודש ,זוהי עלות התפעול השוטף של המודלים - כולל עדכונים ושיפורים.
- משאבי ענן : עלות שימוש ההמשך בתשתית הענן לצורך עיבוד הנתונים וסטוראז' (אחסון) למטרות ניתוח. לדוגמה, במערכת Azure יש צורך בשירותים שמאפשרים להפעיל את המודלים ולנהל את נתוני האבטחה בזמן אמת.
  - עלות משוערת $1,000 – $500 לחודש ,זהו סכום חיבור המערכת לענן.

יתרונות:

1. המטרה העיקרית -

המטרה העיקרית של הדאטה המבוסס על הנתונים שבקישור שציינת היא לשפר את יכולת חיזוי והתגובה לאירועי סייבר. באמצעות הנתונים, ניתן לפתח מודלים של למידת מכונה שמאפשרים זיהוי מתקפות וסיווג של אירועים בזמן אמת, תוך צמצום התראות שגויות והגדלת הדיוק בזיהוי. השגת הצלחה בתהליך זה תוכל להפחית נזקים פוטנציאליים לארגונים על ידי זיהוי מוקדם של מתקפות, מה שיביא לחיסכון ניכר בהוצאות שקשורות למתקפות סייבר.

בנוסף, הפחתת כמות ההתראות השגויות תייעל את העבודה של צוותי SOC (Security Operations Center), ותשפר את היכולת שלהם להגיב בצורה מהירה ויעילה יותר לאירועים.

2. תובנות נוספות שנוצרו מחקירת נתונים -

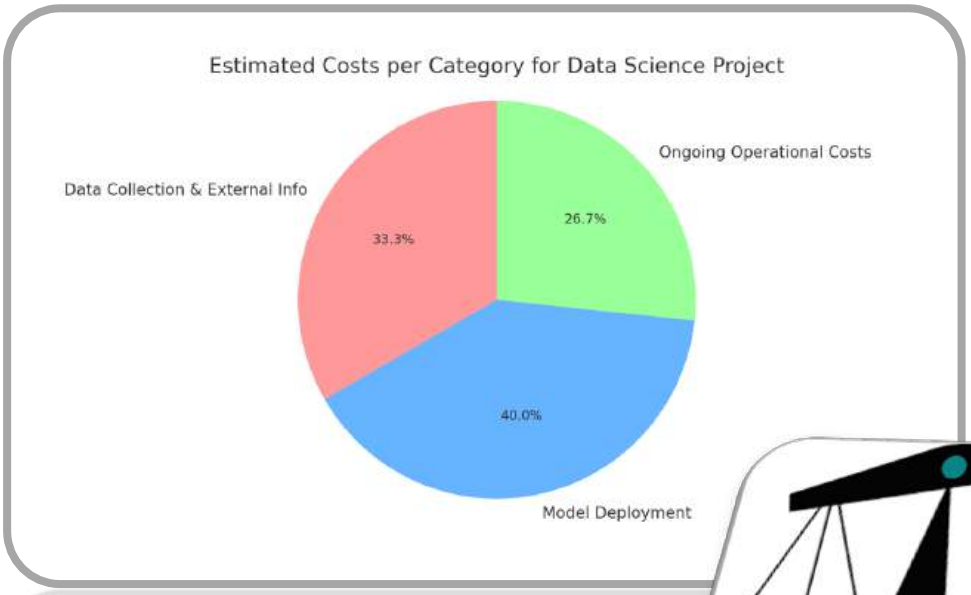
- החקירה והניתוח של הנתונים מאפשרת לארגון לגלות מגמות נסתרות ופרצות אבטחה שלא היו נחשפות באופן אחר. תובנות אלה יכולות לשפר את כל מערך האבטחה.

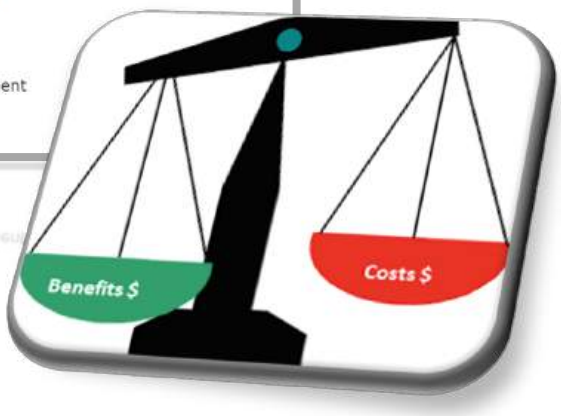
3. יתרונות אפשריים מהבנה טובה יותר של הנתונים -

- חיזוק יכולות זיהוי בעיות ופרצות אפשריות, תוך יצירת מערכת נתונים חזקה שתשמש לפיתוח עתידי.

סיכום:

השקעה בפרויקט זה עשויה להוביל לחיסכון משמעותי בהוצאות שנוגעות למתקפות סייבר ועדכוני אבטחה , ויכול גם לחזק את היכולות הארגוניות והטכנולוגיות של החברה בטווח הארוך.





יעדי מדעי הנתונים וקריטריונים להצלחה

החיבור בין **עסקים** ו**טכנולוגיה** הוא קריטי להצלחה של פרויקטי מדעי הנתונים, כי מדובר לא רק בהבנת הנתונים עצמם, אלא גם בהבנה של איך הם יכולים לשרת את המטרות העסקיות.

<u>שיתוף הפעולה בין אנשי העסקים ומדעני הנתונים -</u> כדי להצליח בפרויקטים כאלה, אנשי **העסקים** צריכים להציג את הצרכים והדרישות שלהם (כמו זיהוי מתקפות בזמן אמת או שיפור יעילות צוותי האבטחה), בעוד ש**מדעני הנתונים** צריכים להשתמש בטכניקות מתקדמות של **למידת מכונה** ו**בינה מלאכותית** כדי לפתח מודלים שיכולים לעזור לפתור את הבעיות הללו.

- לדוגמה, מדען נתונים יכול ליצור מודל חיזוי שמבוסס על **Azure** ו-**Microsoft- Sentinel** כדי לחזות מתקפות סייבר.
- הצוות העסקי יכול להנחות את אנשי הטכנולוגיה על הנתונים החשובים ביותר וכיצד ניתן לתעדף את המודל כדי שיהיה שימושי למטרותיהם.

<u>הפיכת תובנות לפעולות עסקיות -</u> המטרה הסופית היא שהמודלים שפותחו יאפשרו לעסק לפעול בצורה אפקטיבית יותר. לדוגמה, אם המודל יכול לחזות רמות פריצה או חומרה ועדכוני אבטחה, הצוות העסקי יוכל לפעול במהירות ולמנוע נזק , או לשפר את השירותים ללקוחות על ידי אופטימיזציה של תהליכים ובנוסף לעזור במדד הפיננסי של ההשקעה עבור כל עדכון.

<u>איך מדעי הנתונים יכולים לשפר את המטרה העסקית?</u> מדעי הנתונים לא רק עוזרים למצוא פתרונות אלא גם מאפשרים **שיפור מתמיד** -

- **זיהוי מגמות** : באמצעות נתונים, ניתן לזהות דפוסים ומגמות שלא היו נראים לעין.
- **חיזוי בעיות עתידיות** : מדעי הנתונים מאפשרים לחזות תקלות .

- **ייעול תהליכים** : כלי למידת מכונה יכולים לשפר את אופן קבלת ההחלטות, להפחית טעויות ולייעל את האבטחה בצורה ממוקדת ויעילה יותר.

<u>איך משפרים את שיתוף הפעולה בין טכנולוגיה ועסקים?</u>

1. **שיח פתוח והבנה הדדית** : על אנשי העסקים להסביר את הצרכים והבעיות האמיתיות שהם רוצים לפתור, ועם זאת אנשי הטכנולוגיה צריכים להסביר את הפתרונות האפשריים בטכנולוגיות של מדעי הנתונים.
2. **שימוש בטכנולוגיות מותאמות** : יש לבחור בטכנולוגיות שמותאמות לארגון ולבעיה העסקית. מדעי הנתונים אינם פתרון כללי, אלא דורשים התאמה ספציפית לצרכים של כל ארגון.
3. **הדרכה והכשרה** : חינוך והדרכה לשני הגורמים, יכול לעזור להבין את המגבלות והיתרונות של כל תחום, יכולים לשפר את שיתוף הפעולה ולהביא לתוצאות מדויקות יותר.

<u>סיכום:</u>

השילוב של **עסקים** ו**טכנולוגיה** הוא המפתח להצלחת פרויקטי מדעי הנתונים, מכיוון שהוא מבטיח שהתוצאה הסופית לא רק תהיה טכנולוגית עבור אותו הארגון, אלא גם תתאים למטרות העסקיות הספציפיות ולקבלת ההחלטות המדויקות והנכונות יותר לאותו הארגון.

יעדי מדעי נתונים

1. תיאור סוג הבעיה במדעי הנתונים

כאשר מדובר בבעיה בתחום הפריצות עדכוני אבטחה ורמות חורמה ,ישנם סוגים שונים של בעיות במדעי הנתונים שיכולות לעזור לנו :

- ***אשכולות (Clustering):***
  - **בעיה** : זיהוי מתקפות דומות על פי דפוסים (למשל, מתקפות ממקור אחד או מתקפות שמבוססות על שיטות תקיפה דומות).
  - **לדוגמה** : אשכול אירועים שבהם כל המתקפות נעשו תוך שימוש בווירוסים או תוכנות זדוניות עם דפוסים דומים (למשל, מתקפות "Ransomware"או"Botnet") .
- ***רגרסיה (Regression):***
  - **בעיה** : חיזוי של **סיכון מתקפות** או **השלכות נזקים עתידיים** בעקבות מתקפה פוטנציאלית.
  - **לדוגמה** : חיזוי של **הפסדים כספיים** כתוצאה מהתקפה על פי נתונים היסטוריים של מתקפות קודמות (למשל, חישוב הסיכון הכלכלי של מתקפת "פישינג" על בסיס המידע שנגנב).
- ***סיווג (Classification):***
  - **בעיה** : סיווג **אירועים כמתקפה או לא**.
  - **לדוגמה** : סיווג אירועי סייבר כהתקפות או כפעולות תקינות, כמו למשל זיהוי אם ההתראה היא מתקפת דואר זבל ("spam") או לא.

2. תיעוד יעדים טכניים עם יחידות זמן ספציפיות

היעדים הטכניים בפרויקט בתחום **הסייבר** יכולים להיראות כך :

- ***יעד טכני :*** פיתוח מודל **למידת מכונה** לזיהוי מתקפות **"DDoS"** בזמן אמת.
  - **תוצאות רצויות** :
    - תחזית המתקפות היא למשך **שלושה החודשים** הבאים.
    - דיוק של המודל בכ - **95%** בזיהוי מתקפות "DDoS".
- ***יחידות זמן ספציפיות :***
  - **שלב 1 (חודש 1)** : איסוף נתונים היסטוריים על מתקפות "DDoS".

- ***יחידות זמן ספציפיות:***
  - **שלב 1 (חודש 1)** : איסוף נתונים היסטוריים על מתקפות "DDoS".
  - **שלב 2 (חודש 3-2)** : יצירת המודל, ביצוע אופטימיזציה ותחזיות בזמן אמת.

3. סיפוק מספרים בפועל עבור התוצאות הרצויות

1. **הקטנת התראות שגויות** : הפחתת התראות שגויות ב-40%, דבר שיביא להורדת העומס על צוותי האבטחה וייעול תהליך הזיהוי והתגובה לאירועים חשובים, כך שצוותי האבטחה יתמקדו רק בהתרעות הרלוונטיות.
2. **זיהוי מתקפות מוקדם יותר** : באמצעות ניתוח הנתונים, ניתן לזהות בעיות במערכות בזמן אמת, מה שיביא להפחתת נזקי מתקפות סייבר פוטנציאליות, וחיסכון כלכלי של מאות אלפי דולרים בשנה.
3. **שיפור בניהול תקלות והתקפות** : ניתוח מדויק של חומרת התקלה או ההשפעה שלה על המערכת יאפשר התמקדות בתקלות קריטיות יותר, תוך מתן עדיפות לפתרונות המתאימים ביותר.

סיכום: הבעיה העסקית כאן היא להקטין את הסיכון וההפסדים כתוצאה ממתקפות סייבר ורמותיהם ועדכוני אבטחה, תוך שימוש בבעיות במדעי הנתונים כמו ב**אשכולות** ,**רגרסיה** ו**סיווג**.

כל יעד טכני מוצע יהיה קשור בזמן ובתוצאות כמותיות (למשל, דיוק המודל, הפחתת התראות שגויות, והחיסכון הכלכלי) כדי להבטיח פתרון טכנולוגי בר-השגה שיאפשר למנוע מתקפות ולשפר את האבטחה בארגון.

מה זה? Ransomware - רנסומוור (Ransomware) הוא סוג של תוכנה זדונית (malware)שמשתלטת על מחשב או מערכת, ומבצעת הצפנה (encryption) של הקבצים בתוך המחשב או השרת. ברגע שהתוקפים מצליחים להדביק את המערכת בתוכנה כזו, הם דורשים תשלום-כופר כדי להחזיר את הגישה לקבצים המוצפנים.

מה זה? Botnet - (**Botnet** רשת בוטים) היא רשת של מחשבים או מכשירים מדביקים שמשתמשים בהם התוקפים כדי לבצע התקפות רשת או לשלוט בהם באופן מרוחק. כל מחשב או מכשיר ברשת נקרא "**bot**", והוא פועל על פי פקודות שמגיעות מהתוקפים.

בדרך כלל, בוטנט משמש להתקפות "**DDoS**" שיתואר בהמשך, או להפצת תוכנות זדוניות אחרות.

מה זה? DDoS - **DDoS** היא מתקפת מניעת שירות מבוזרת. במתקפה כזו, התוקפים שולחים כמויות עצומות של בקשות או תנועה (traffic) אל השרת או לאתר האינטרנט המסוים, במטרה להעמיס עליו ולמנוע ממנו להעניק שירותים למשתמשים אחרים.

במקום שתוקף אחד ינסה להיכנס לאתר או לשרת, התוקפים משתמשים **בבוטנט** כדי לשלוח מתקפה מבוזרת מכמה מקורות במקביל, מה שמקשה על מניעת ההתקפה.

קריטריונים להצלחה במדעי הנתונים

כדי להבטיח הצלחה בפרויקט, נגדיר קריטריונים מבוססים על יעדים ברורים ותוצאות מדידות. להלן התוכנית המפורטת -

1. שיטות להערכת מודל

- ***דיוק (Accuracy):*** אחוז התחזיות הנכונות מתוך כלל התחזיות.
  - רלוונטי במיוחד לבעיות סיווג (Classification), כמו זיהוי מתקפות סייבר.
- ***זמן תגובה (Latency):*** משך הזמן שלוקח למודל לנתח את האירוע ולהגיב לו .
- ***דיוק בהתראה מוקדמת (Precision & Recall):***
  - **Precision** : כמה מתוך ההתראות שהמודל זיהה כמתקפה באמת היו מתקפות.
  - **Recall** : כמה מתוך המתקפות בפועל המודל זיהה.
  - משמש להערכת איזון בין **False Positives** ל - **False Negatives**.
- ***F1-Score:*** מדד משוקלל בין Precision ו – Recall, כאשר חשובה ההדגשה על איזון בין דיוק וזיהוי מלא.

2. אמות מידה להערכת הצלחה

- ***דיוק המודל (Accuracy):*** המודל ייחשב מוצלח אם יגיע לדיוק של **90% ומעלה** בזיהוי מתקפות סייבר.
- ***הפחתת התראות שגויות (False Positives):*** המודל יפחית את שיעור ההתראות השגויות ב **- 40% לפחות**.
- ***זמן תגובה*** : המודל יפעל בזמן אמת, עם זמן תגובה שאינו עולה על **500 מילי-שניות** לאירוע.
- ***זיהוי מתקפות חוזרות*** : זיהוי של לפחות **95% מהמתקפות החוזרות** על סמך דפוסים דומים.

3. מדידות סובייקטיביות ובורר הצלחה

- ***מדדים סובייקטיביים:***
  - שביעות רצון מצוותי האבטחה: הצוותים יצטרכו להעיד שההתראות הרלוונטיות עוזרות להם להגיב מהר יותר לאירועים ,ומשפרות את עבודתם.
  - חוויית המשתמש: עד כמה הצוותים מצליחים להבין ולהשתמש בתובנות שמייצר המודל.
- ***בורר הצלחה:***
  - הפרויקט ייחשב מוצלח אם צוותי האבטחה ימדדו שיפור של **30%** **ביעילות** העבודה היומית שלהם בעקבות השימוש במודל.

4. פריסה מוצלחת כחלק מההצלחה

- ***האם הפריסה מוצלחת?***
  - הצלחת המודל נמדדת גם בפריסתו במערכות קיימות, כמו **Azure Sentinel**או **Microsoft Defender**.
  - יש לוודא שהמודל משתלב בצורה חלקה עם תשתיות קיימות ומספק תובנות בזמן אמת.
- ***שלבי תכנון לפריסה:***
  - **חודש 1**: הכנת סביבות בדיקה עבור המודל.
  - **חודש 2**: בדיקת הפריסה בסביבת הייצור (בסביבה בו המוצר רץ עם נתונים חיים) על קבוצת נתונים קטנה.
  - **חודש 3**: השקת המודל לסביבת הייצור המלאה (פריסה מלאה במוצרים או במערכות שהוא **נועד לשרת**).

סיכום:

- **שיטות הערכה:** Precision, Recall, Accuracy, Latency.
- **אמות מידה:** דיוק > 90%, הפחתת False Positives ב - 40%, זיהוי מתקפות חוזרות > 95%.
- **מדידות סובייקטיביות:** שביעות רצון הצוותים ושיפור פרודוקטיביות ב - 30%.
- **פריסה:** הצלחה תכלול שילוב מלא במערכות קיימות ותכנון לפריסה מתוזמנת.

תוכנית הפרויקט

בעידן הדיגיטלי, שבו איומי הסייבר מתעצמים ומאתגרים את מערכות האבטחה ועדכוני תוכנה ורמותיהן של הארגונים ברחבי העולם, תוכנית פרויקט במדעי הנתונים מקבלת משמעות אסטרטגית מיוחדת. בתחום זה, תוכנית הפרויקט אינה רק מסמך ארגוני – היא מצפן שמכוון את המאמצים לזהות, לנתח ולחזות איומים פוטנציאליים. תוכנית זו מספקת מסגרת ברורה שמאגדת מטרות, לוחות זמנים, משאבים וסיכונים, ומבטיחה שכל המעורבים – מהמובילים העסקיים ועד לצוותי הנתונים – עובדים בתיאום מלא להשגת תוצאות קריטיות. בתחום רגיש כמו סייבר, התוכנית מסייעת לא רק בשיפור האבטחה, אלא גם במניעת נזקים משמעותיים למערכות המידע, תוך דגש על שקיפות, יעילות וצמצום סיכונים. עבודה מסודרת ושיטתית זו תאפשר לארגון להתמודד עם אתגרי העתיד בעולם הסייבר המשתנה במהירות.

<u>מטרות הפרויקט:</u>

הגדרת הבעיה העסקית והתוצאות הרצויות:

הפרויקט מתמקד בזיהוי איומי סייבר ומתן פתרונות להגברת האבטחה הארגונית. המטרה היא לאפשר זיהוי מוקדם של תקלות ולספק כלי חיזוי המבוססים על נתוני העבר ותובנות בזמן אמת. לדוגמה, דיוק חיזוי של $\geq$ 90% בטווח של שלושה חודשים ישפר משמעותית את יכולת המניעה והתגובה.

מטרות טכניות ברורות:

היעדים כוללים יצירת מודלים סטטיסטיים ומבוססי למידת מכונה שיאפשרו חיזוי תקלות, איתור אנומליות ובניית תשתית להמלצות פרואקטיביות.

<u>משאבים ולוחות זמנים:</u>

הערכות זמן לכל שלב:

- ניקוי נתונים (3 שבועות)
- ניתוח ראשוני (2 שבועות)
- בניית מודלים (4 שבועות)
- אימות (2 שבועות)

- פריסה (3 שבועות)

הסך הכולל הוא 14 שבועות עם גמישות לשינויים בהתאם לתקלות או אתגרים שיצוצו.

משאבים נדרשים:

- כוח אדם: צוות הכולל מנתחי נתונים, מומחי סייבר, מהנדסי נתונים, ומנהלי פרויקט.
- תשתיות: שרתים לביצוע חישובים כבדים, כלים לניתוח נתונים כמו Pythonאו R וגיבוי מתמיד של מאגרי מידע.

<u>תהליך העבודה:</u>

שלבים עיקריים:

- איתור נתונים: זיהוי מקורות מידע מתאימים (דוחות תקלות, יומני סייבר).
- יצירת מודלים: שימוש באלגוריתמים כמו סיווג, רגרסיה ואשכולות.
- אימות: השוואת ביצועי המודלים באמצעות מדדי ביצוע כמו MSE ו F1. Score -
- פריסה: הטמעת המודל במערכות קיימות לצורך בדיקות והטמעה בארגון.

דגשים על איטרציות:

תהליך שיפור המודלים יכלול התאמות חוזרות במקרים של דיוק נמוך או כשלים בזיהוי תקלות חדשות.

<u>ניהול סיכונים:</u>

זיהוי סיכונים:

- חוסר בנתונים איכותיים או מלאים.
- ביצועי מודלים נמוכים בשל מורכבות או רעשי נתונים.

הצעת פתרונות:

- שימוש בטכניקות השלמת נתונים והפחתת רעש.
- בדיקות חלופיות ואינטגרציה של מקורות נתונים נוספים.
- גיבוי מתמיד ושימוש במודלים רגרסיביים כפיילוט.

נקודות החלטה:

אבני דרך בתהליך:

- סיום שלב ניקוי הנתונים.
- השלמת בניית מודלים ראשוניים.
- קבלת אישור על דיוק חיזוי מול דרישות הפרויקט.

שלבי אישור קריטיים:

- בדיקת איכות הנתונים לפני המשך לעיבוד.
- בדיקות ייעודיות למודל לפני הפעלתו בסביבה הארגונית.
- קבלת אישור לפריסה מההנהלה הבכירה.

אופן הפעולה וההתנהלות:

הפרויקט שלנו מורכב ממשימות מוגדרות, כאשר כל משימה מהווה חלק אינטגרלי בתהליך ומשקפת את עיקרי הנושא המרכזי. אחד השלבים המרכזיים והמורכבים ביותר היה איתור הנתונים ובחינתם, תוך הדגשת רלוונטיות הנושא לפרויקט הגמר. הנושא שנבחר – איומי סייבר – משפיע על חברות רבות, במיוחד על מיקרוסופט, אשר מתמודדת עם מגמת עלייה בכמות האיומים. זיהוי מוקדם של האיומים יכול לספק ביטחון משמעותי למנויי החברה, להפחית סיכונים, ולהגביר את תחושת הבטיחות.

התהליך כלל עבודה אינטנסיבית ובחינת מספר רב של מקורות נתונים. חשיבות הנושא חיזקה את ההשקעה שלנו בחקירה מעמיקה שלו, תוך התייעצות ותשומת לב למציאת נתונים מדויקים ועדכניים. הניתוח כלל הבנה מעמיקה של סוגי התקיפות השונות, אופן התרחשותן, והשפעתן על הארגון.

במהלך העבודה למדנו רבות על תחום איומי הסייבר ואימוץ טכניקות מתקדמות בלמידת מכונה ו .AI-המטרה הסופית היא פיתוח פתרון פרואקטיבי לצמצום הפגיעות והגברת הבטיחות, תוך שימוש במודלים מבוססי נתונים. כל שלב בתהליך הקנה לנו ידע נוסף והבנה טובה יותר על הדרכים להתמודד עם איומים ולהתמקד בפרמטרים המרכזיים מתוך מאגר הנתונים הנרחב של מיקרוסופט.

הערכת זמנים היא נדבך מרכזי בפרויקט רחב היקף מסוג זה. ניהול מוקפד של הזמן עבור כל שלב בפרויקט, לצד עמידה בלוחות הזמנים שנקבעו, מספק אינדיקציה חיובית ותורם לשביעות רצון המעורבים. תהליך חלוקת הזמנים בפרויקט נעשה באמצעות תאריכים מוגדרים והערכת דרישות לכל שלב, תוך התמקדות בכמות

הזמן הנדרשת לכל משימה. הגדרה ברורה של הדרישות סייעה לנו לבצע הערכה מדויקת ולהקצות את המשאבים הנדרשים בצורה מיטבית.
תכנון מוקפד והקפדה על ניצול זמן מיטבי העניקו תחושת מעורבות והובילו לעמידה מוצלחת ביעדים. הערכה נכונה של משאבים וזמן, לצד דרישות ברורות, הוכיחו את חשיבותם בשיפור המוטיבציה והמחויבות לפרויקט. הבנת החשיבות של ניצול אופטימלי של הזמן מדגישה את תפקידה המרכזי של תכנון נכון בניהול פרויקטים ומבטיחה עמידה בלוחות זמנים תוך מיצוי מקסימלי של הפוטנציאל.

המאמץ בפרויקט נמשך לאורך כל שלביו, החל מבחירת הנושא ועד לסיומו המוצלח. כל שלב בפרויקט דרש משאבים שונים, כולל זמן, ידע טכנולוגי, ומיומנויות תפעוליות. הפרויקט לא הסתמך רק על הידע הקיים שלנו, אלא גם דרש העמקה וחקירה מעמיקה בשטחי התקיפה של חברת מייקרוסופט ותקיפות סייבר כלליות שהתרחשו, כדי להבין את מאפייני האיומים והסכנות הקיימות בתחום.
לאחר פיתוח האלגוריתם, התבצעו פעולות של הטמעתו במערכות שונות ועריכת בדיקות קריטיות להתאמת תפקודו בכל חברה, במטרה להבטיח את התקינות והביצועים האופטימליים שלו. בהסתמך על נתונים שנלקחו מ ,Microsoft-
האלגוריתם הזה צפוי לתרום רבות לשיפור תחום האבטחה והסייבר. יישום האלגוריתם בחברות גדולות ובארגונים ממשלתיים יהפוך אותו לכלי חיוני עבור גופים המעוניינים לחזק את מערכות האבטחה שלהן ולמנוע התקפות עתידיות.
נקודות החלטה הן שלבים קריטיים בפרויקט, בהם הצוות נדרש להתמודד עם קונפליקטים ולבצע החלטות על מנת להתקדם לעבר המטרה. במהלך הפרויקט היו מספר נקודות בהן נדרשנו לקבל החלטות משמעותיות. השלב הראשון היה בחירת הנושא, תהליך לא פשוט שדרש לבחור נושא שיתאר בצורה מדויקת את מטרת המערכת שברצוננו לפתח – מערכת שתסייע בהגנה על חברה. בנוסף, רלוונטיות הנושא הניעה אותנו לחפש נתונים יומיים שיכולים לתרום להצלחת חברות.
במהלך חיפושי הנושא, החלטנו להסתמך על דאטה מ - Microsoft הכולל מיליוני רשומות של תקיפות סייבר. מדובר בנתונים המכילים קריטריונים רבים ולעיתים ערכים חסרים, דבר שמאלץ אותנו לקבל החלטות קריטיות בנוגע לקטגוריות שייבחרו לאלגוריתם. עלינו להתמודד עם סיכון הקשור להשלמת נתונים חסרים, אשר ידרוש הרצת האלגוריתם מספר פעמים עם שילובים שונים של פיצ'רים, עד שנגיע לתוצאות אופטימליות.

בקשות בדיקה הן תהליך חשוב לבדיקת דיוק ותפקוד המערכת, ומבצעות תהליך של

הערכת הביצועים והמדדים החשובים. עם קבלת הנתונים, שמנו לב לחוסרים ואי נתונים מושלמים, לשם כך, נבצע אנליזות בעזרת כלים כמו numpy ו- pandas ונבחר את הפיצ'רים המתאימים באמצעות כלים כמו Scikit-learn, Boruta ו- XGBoostהפיתוח יכלול מודלים מתקדמים כגון Random Forest ונבחן את ביצועיהם בעזרת מדדים כגון דיוק (Accuracy) ורגישות (Recall).

בנוסף, נשתמש ב - SimPy כדי להעריך את ביצועי המודל דרך סימולציות, וב - Flaskלשם בדיקות אינטגרציה עם מערכות קיימות. לאחר סיום הבדיקות, תבוצע פריסה של המערכת בסביבה חיה באמצעות Docker, אשר יאפשר גמישות ויכולת התאמה לצרכים משתנים, ובכך יגביר את תפקוד מערכות האבטחה ויביא לתוצאות טובות יותר.

הערכה ראשונית של כלים וטכניקות

לשם ביצוע הערכה ראשונית, נבחרים כלים פופולריים ומובילים בתחום מדעי הנתונים.

שלב 1: סקירה של כלים בולטים

Python .1

- **מה זה?**

Python הוא אחד הכלים הפופולריים ביותר במדעי הנתונים. יש לו ספריות רבות שמאפשרות ניתוח נתונים, חישובים מתקדמים, ביצוע למידת מכונה ו-ויזואליזציה.

- **ספריות עיקריות**:
  - **Pandas** - לניהול נתונים וניתוחם.
  - **NumPy** - לחישובים מתמטיים ומדעיים.
  - **Scikit-learn** - לביצוע אלגוריתמים של למידת מכונה.
  - **Matplotlib / Seaborn** - ליצירת ויזואליזציות על גבי גרפים.
  - **TensorFlow / PyTorch** - לפיתוח מודלים של למידה עמוקה.
- **יתרונות**:
  - קל ללמוד ולהשתמש.
  - קהילה רחבה ותמיכה רבה.
  - תמיכה מצוינת ללמידת מכונה וללמידה עמוקה.

R .2

- **מה זה?**

R הוא שפה ותוכנה שמתמקדת בניתוח סטטיסטי, במיוחד עבור נתונים כמותיים. הוא מצוין לעבודות שמתמקדות בניתוחים סטטיסטיים מורכבים.

- **ספריות עיקריות**:
  - **dplyr** - לניהול וניתוח נתונים.
  - **ggplot2** - ליצירת ויזואליזציות מתקדמות על גבי גרפים.
  - **caret** - לביצוע למידת מכונה.
- **יתרונות**:
  - חזק מאוד בניתוחים סטטיסטיים.

- מצוין עבור ניתוח נתונים חזותיים וגרפיים.
- מתעדכן כל הזמן עם שיפורים חדשים.

3. SQL

- **מה זה?**

SQL (Structured Query Language) הוא כלי עיקרי לשאילת נתונים מתוך מאגרי מידע. הוא חשוב בשלב איסוף הנתונים, ויש לו תפקיד מרכזי בפרויקטים מבוססי נתונים גדולים.

- **יתרונות**:
  - נוח לגישה ולעיבוד נתונים במאגרי מידע רגישים.
  - תומך בעבודה עם נתונים שנמצאים בתוך מסדי הנתונים.
  - אפשר להשתמש ב - SQL יחד עם כלים אחרים כמו Python ו - R.

4. Apache Spark

- **מה זה?**

Apache Spark הוא כלי עוצמתי לעיבוד נתונים גדולים. הוא מצוין עבור נתונים גדולים ומאפשר עיבוד מקבילי על פני אוספים גדולים של נתונים.

- **יתרונות**:
  - מיועד לעבודה עם **Big Data**.
  - תומך בלמידת מכונה ובטכניקות אחרות על נתונים גדולים.
  - מספק ביצועים גבוהים מאוד.

<u>שלב 2: בחירת טכניקות מדעיות</u>

לאחר שבחרנו את הכלים, נבחר את הטכניקות המתאימות ביותר לבעיה העסקית שלנו. הפרויקט מתמקד במתקפות סייבר, אז נדבר על טכניקות שיכולות להועיל:

1. למידת מכונה (Machine Learning)

- **מה זה?**

טכניקות למידת מכונה מאפשרות למודלים ללמוד ולחזות תוצאות על בסיס נתונים. למשל, תחום **למידת מכונה מפוקחת** (Supervised Learning) יכול לשמש לסיווג מתקפות סייבר.

- **טכניקות רלוונטיות**:
  - **סיווג (Classification):** לסווג אירועים כמתקפה או לא.
  - **למידת חיזוק (Reinforcement Learning):** למודלים שיוצרים תגובות בזמן אמת כמו במתקפות DDoS. (**למידת חיזוק** הוא סוג של אלגוריתם שמטרתו **ללמוד ולבצע סדרת פעולות נכונות** כדי להשיג את התוצאה הטובה ביותר, בהתאם למטרות שנקבעו. האלגוריתם פועל באמצעות ניסוי וטעיה, תוך התאמת הפעולות שלו למקסימום **תגמול מצטבר**).

2. למידת עמוקה (Deep Learning)

- **מה זה?**

מדובר בטכניקות מתקדמות יותר של למידת מכונה, שבהן המודלים עושים שימוש ברשתות נוירונים. מתאים במיוחד כשיש כמות עצומה של נתונים (כמו נתוני רשת או נתונים ממערכות אבטחה).

- **טכניקות רלוונטיות**:
  - **רשתות נוירונים עמוקות (Deep Neural Networks):** מודלים אלה יכולים לזהות דפוסים מורכבים במתקפות סייבר.
  - **למידת מכונה לא מפוקחת (Unsupervised Learning):** לגילוי דפוסים חדשים במתקפות.

3. אנליזת זמן אמת (Real-time Analytics)

- **מה זה?**

מדובר בניתוח נתונים בזמן אמת כדי לזהות איומים באופן מיידי, ולקבל התראות על מתקפות בהן נתקלות המערכות.

- **כלים רלוונטיים**:
  - **Apache Kafka:** לאיסוף וניתוח נתונים בזמן אמת.
  - **Apache Flink:** לפיתוח פתרונות של ניתוח בזמן אמת.

4. נתונים גדולים (Big Data Analytics)

- **מה זה?**

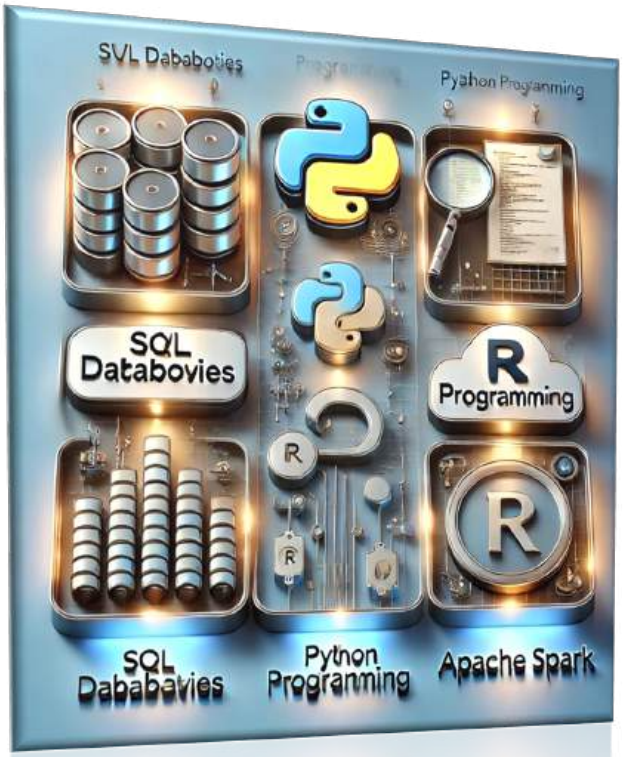
עבודה עם נתונים בקנה מידה גדול שדורשים טכניקות ופתרונות עיבוד נתונים מתקדמים.

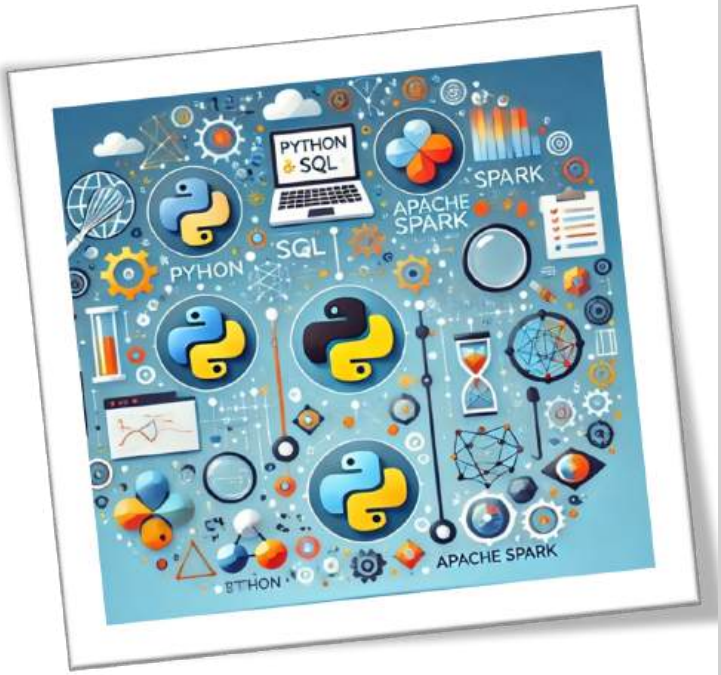
- **כלים רלוונטיים**:
  - **Apache Spark**: לעיבוד נתונים במהירות ובכמות גדולה.

<u>שלב 3: הכלים והטכניקות שבהם נשתמש לפרויקט</u>

בהתאם לבעיה העסקית שלנו (זיהוי מתקפות סייבר), הכלים והטכניקות שיכולים להצליח לעזור הן:

1. **<u>Python</u>**: עם ספריות כמו **Scikit-learn**, **TensorFlow**, ו- **Matplotlib** ללמידת מכונה ו- לויזואליזציה.
2. **<u>Apache Spark</u>**: עבור עיבוד נתונים גדולים.
3. **<u>SQL</u>**: כדי לשלוף נתונים ממאגרי מידע לצורך עיבוד וניתוח.
4. **<u>למידת מכונה</u>** **<u>(Classification)</u>**: כדי לזהות מתקפות, ולסווג אירועים כמתקפות או לא.
5. **<u>למידה עמוקה</u>** **<u>(Deep Learning)</u>**: במקרה של כמות גדולה מאוד של נתונים.





הערות מהמנחה חנן לב:

הפרויקט שלנו נבחר על נתונים אחרים והוחלף לאחר דיון מעמיק עם ראש החוג והמנחה , הפרויקט שלנו עוסק בנתונים על מערך האבטחה של מיקרוסופט , הדאטה החדש הוא שונה מאוד מהדאטה הקודם שכן הדאטה הקודם חוקר במדויק מה הם הקרטריונים והרמות פגיעה בארגון לפי , מדינה , אימיילים, לוגים ,סוגי משתמשים , מזהים שונים ועוד, בעוד הדאטה החדש שלנו עוסק ברמות חומרה ,פגיעות ,כמה סוגי לוגים ועדכוני אבטחה.
במסגרת הערות לקחנו לתשומת לבנו את שינוי הדאטה של הפרויקט והתאמתו מחדש לנושא הפרויקט, שכן ניתן לבצע את אותו נושא רק ב"מיקרו". הנתונים אינם לקוחים מאתר שגוי , הם לקוחים ישירות מממיקרוסופט . כלומר הפעם לא לקחנו את הדאטה מאתר צד שלישי כמו "Kaggel", שם פורסם הדאטה הקודם.

מעבר לזה, סעיף של אילו כלים נשתמש לבנייה לא מחייב ואין התחייבות בשלב זה . מסמך זה הוא מסמך ראשוני של הפרויקט , אקדמאי ומקצועי המנחה לא נתן לנו דרישה על סוג הגופן וגודל הגופן , לכן נעשה בו שימוש לפי רצוננו , אנחנו התאמנו את הגודל לתוכן.

לסיכום, הדאטה הקודם התעסק בכ – 14 מיליון תצפיות מעניינות מאוד ובכ – 46 מאפיינים עליהם היה "מולבש" הדאטה. מאחר הדאטה הקודם נלקח מאתר העוסק בפתרון מדעי הנתונים (כלומר ישנם פתרונות ממדענים באתר) – לא אושר הדאטה הזה ולכן הפעם בחרנו בדאטה ישירות ממיקרוסופט אותו חיפשנו ובחרנו בקפידה רבה. עברנו מביג דאטה לסמול דאטה, מ – 14 מיליון תצפיות וכ – 46 תכונות ל כ – 20 אלף + תצפיות וכ – 14 תכונות. דבר זה הוא מהותי כיוון שהכלים שבהם נשתמש יהיו שונים ומותאמים בהתאם להרצה ,ניתוח ומידול הדאטה.

מסקנה, הנושא נשאר אותו נושא רק הנתונים התחלפו.

מראי מקום

https://www.organimi.com/organizational-structures/microsoft
https://www.microsoft.com/en-us/security/security-insider/intelligence-reports/microsoft-digital-defense-report-2024
https://www.microsoft.com/en-us/security/security-insider/microsoft-digital-defense-report-2023
prediction
https://www.microsoft.com/investor/reports/ar23/index.html
https://www.microsoft.com/en-us/research/group/m365-defender-research
https://learn.microsoft.com/en-us/dynamics365/guidance/organizational-strategy/define-organizational-strategy
https://ctraining.co.il/%D7%A7%D7%95%D7%A8%D7%A1-azure-%D7%9C%D7%9E%D7%A2%D7%A8%D7%9B%D7%AA-%D7%A0%D7%99%D7%94%D7%95%D7%9C-%D7%90%D7%99%D7%A8%D7%95%D7%A2%D7%99-%D7%90%D7%91%D7%98%D7%97%D7%AA-%D7%9E%D7%99%D7%93%D7%A2/
https://www.microsoft.com/he-il/
https://www.python.org/
/https://www.r-project.org
https://www.w3schools.com/sql/sql_intro.asp
/https://spark.apache.org
https://www.microsoft.com/en-us/download/details.aspx?id=36982
https://www.microsoft.com/en-us/search/explore?q=download+details+for+Web+browsers
https://www.microsoft.com/he-il/download