

# Microsoft Security Bulletin Analysis

דו"ח הערכה מוצג על ידי בר כהן וסחר חיים יעקב.

דו"ח זה מסכם את שלב ההערכת פרויקט, ובו חונן האם התוצאות שהתקבלו בתהליכי המידול עומדות בייעדים שהוגדרו בתחלת הדרך – הן מבחינת האנליטיות והן מבחינת העסקית. הדוח כולל ניתוח של התאמה המודלים למטרות הפרויקט, הציגת הממצאים המרכזיים, והסקת מסקנות באשר לאפקטיביות הכללית של הפרויקט. בנוסף, נבחן תהליכי העבודה עצמו, תוך זיהוי נקודות חזקה והצעות לשיפור עתידי.

## Evaluation

Presented by Bar Cohen  
Sahar haim Yaakov

This report summarizes the evaluation phase of the project, examining whether the results meet the goals defined at the outset. It includes an analysis of the models' alignment with the objectives, the effectiveness of solution, and a review of the work proce itself.



שם מרצה : מיר אביה זלאי.  
שם מנהה : מיר חנן לב.  
מגייסים :  
בר כהן 208110254  
סחר יעקב 314741851



### חובן עניינים

1. **העלאת התוצאות:** ..... - 3 -
2. **תהליכי הסquia:** ..... - 6 -

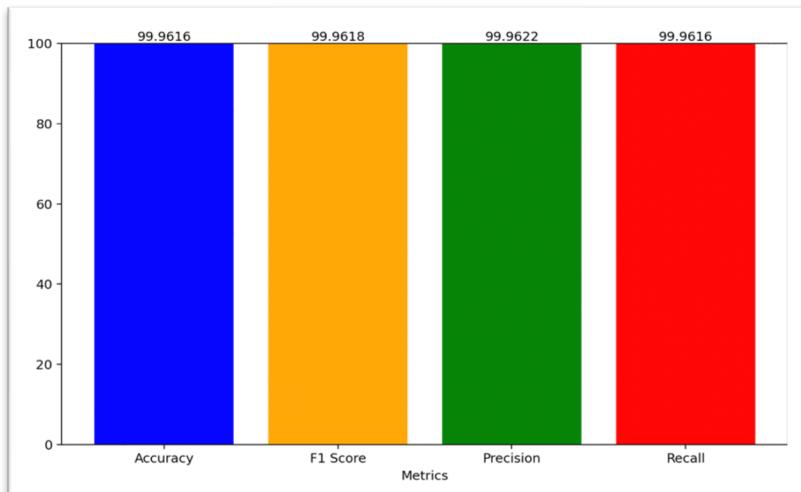


## 1. הערכת התוצאות:

בשלב זה של הערכת תוצאות המודל, נבחן את המטרה של הפרויקט מבחן קriticרוני הצלחה. המטרה של הפרויקט היא חיזוי מדויק של חומרת תקיפה עבור הארגון, באופן שיאפשר תגובה לרמת האיום, הפרדה של תקיפות חזקות יותר והערכות בהתאם. לאחר שהציגו את דרכם ההתחמಡות שלנו, חקרו את הפתרונות האפשריים, ביצעו ניסויים מורביים עם מגוון מודלים תוך התאמת של פרמטרים לכל אחד מהם. התוצאות שהתקבלו מראים כי ניתן להגיע לרמת דיקijk בגובה מואוד בחיזוי חומרת התקיפה, ובפרט כאשר נעשה שימוש במודלים כמו CatBoost.

בנוסף בחנו גם ממדים נוספים כמו F1, Precision ו- Recall, וכן התייחסנו ליכולת של המודלים להבחין בין רמות חומרה שונות – קלות, בינוניות וקריטיות – כדי עסקי חשוב ביותר עבור הארגון.

### התפלגות עבור המודל שנבחר :



איור 1 : השוואת הדיקijk בין ארבעת המודלים המובילים

התוצאות שהתקבלו ממודל CatBoost וממודלים נוספים שבחנו במהלך הפרויקט מציגות רמה גבוהה מאוד של ביצועים, בהבוסס על דיקijk ורגישות. הנתונים מראים בצורה כמעט מושלמת, אפקטיבית ויעילה שהמודלים מצליחים לחזות את ערך המטרה - חומרת התקיפה.

במהלך ניתוח המודלים זיהינו כ-2-3 תכונות שהשפיעו על תוצאה החיזוי היא גבוהה ביחס לאחרות. תכונות אלו בלטו במודל ותרמו להסביר ולשיפור יכולת הניבוי שלו. בנוסף המודל חשוב אילו ייחידות בארגון ניתנים לפירצה ולתקיפה וזוקקים לשיפור ותחזקה גבוהה יותר.

בחלק מתהליך העבודה, הרצינו גרשאות שונות של מודלים תוך שינוי פרמטרים כמו עומק העצים, קצב הלמידה, כמות האיטרציות, ובחנו כיצד השינויים האלו משפיעים על איכות החיזוי. ניסויים אלו הובילו לתובנות חשובות ושיפורים מדדיים ביצועים. דירגנו כל מודל בהתאם לביצועיו, כך שיכלנו לזהות בצורה ברורה את המודל המוביל.

המודל יהווה יתרון עבור הארגון מכיוון שהוא מאפשר לארגון תיעוד וטיפול בתקיפות חמורות יותר, דבר שיוביל להקצותות להן משאבים בצורה יותר, ולבנות תוכנית תגובה מבוססת סיכון, ובכך ישנו שיפור ממשי בניהול האיום ובהגנה על נכסיו הארגוני, מה שתומך בקבלת החלטות אסטרטגיות מדויקות יותר.



### איזה נדרג את המודלים?

המודלים ידורגו לפי היכולת שלהם לעמוד במידדים העסקיים, תוך התמקדות בכמה פרמטרים -

**יעילות חיצונית:** המודל שיצילח להוכיח את רמות הפגיעה ורמת החומרה בצורה המדוייקת ביותר -  
ידורג גובה יותר.

**פרמטרי המודל:** נרצה להימנע ממודלים שיכולים לסוג תקיפות "קלות" בקלות, כדי להימנע מהצמדות יתר - קבענו פרמטרים קבועים קטנים אך נשמר על דיוק גבוהה.

הקריטריונים העיקריים להצלחה היו דיוק בזיהוי תקיפות - קבענו דיוק של מעל 90% למודל אמין, אפקטיבי, בעל זמן תגובה מהיר, בעל הפחיתה התוראות שגויות - יכולת הפחתה של שיעור התהווות השגויות שנחשו ללא מזיקות או להפץ, ובעל חיסכון במשאבים.

השתמשנו במודלים שונים לצורכי השוואה ובבחינה של איזה מהם הכיiesel ומתאים לעיד העסקי שלנו ומוטבים בחרנו את המודל שנתן את הביצועים הטובים ביותר.

דירוג המודלים מתבצע ע"פ רמת דיוקם לזיהוי ההתקפות בקצב הבדיקה וכך יתרמו לעידים העסקיים בהמשך.

### אלו הם 7 המודלים החזקים ביותר שתורמים למטרת העסקית:

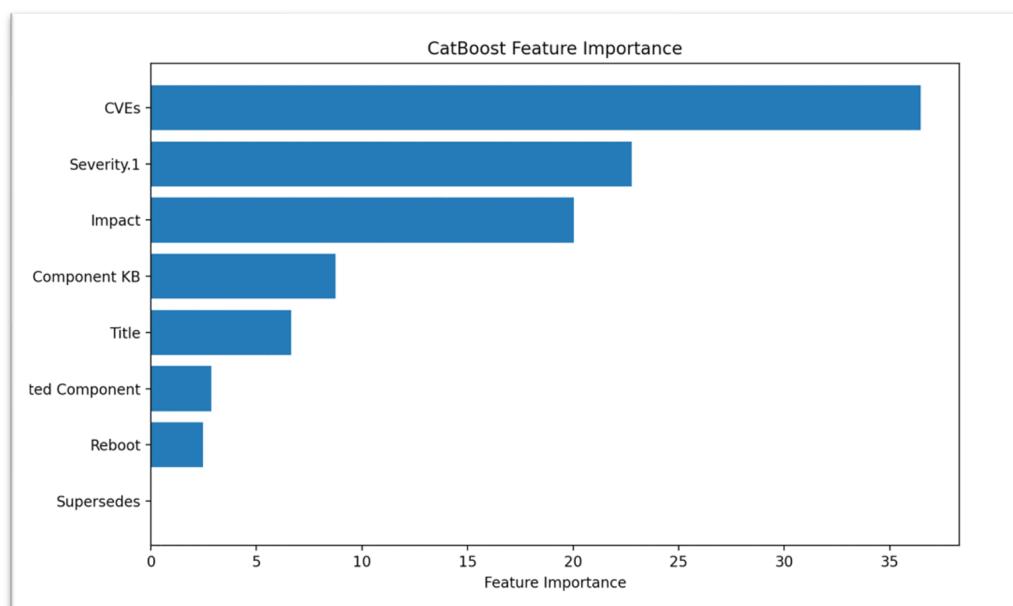
#### **נבעוד לדילוג :**

.0.999646	- סט בדיקה : 48.11%	deep : 5, עומק : 5, דיוק : 47.51%	<b>CatBoost</b>	1
.0.999642	- סט בדיקה : 47.51%	deep : 6, עומק : 6, דיוק : 47.21%	<b>CatBoost</b>	2
.0.99964	- סט בדיקה : 47.21%	deep : 6, עומק : 6, דיוק : 42.219%	<b>CatBoost</b>	3
.0.999597	- סט בדיקה : 42.219%	deep : 5, עומק : 5, דיוק : 40%	<b>CatBoost</b>	4
.0.999574	- סט בדיקה : 40%	deep : 6, עומק : 6, דיוק : 12000, חזרות : 40	<b>CatBoost</b>	5
.0.992980	- סט בדיקה : 40%	deep : 12000, דיוק : TabNet	<b>TabNet</b>	6
.0.980020	- סט בדיקה : 40%	deep : 40, דיוק : XGBoost	<b>XGBoost</b>	7

איור 2 : 7 המודלים המובילים עם פרמטרים הטובים ביותר.

### עם זאת במהלך המודלים השונים לمعנה על המטרות התפתחו שאלות כגון :

- האם ניתן לחזות את חומרת התקיפה לפי השפעת הבעה על המערכת?
- מה מידת ההשפעה של מידת החומרה ראשונית על החיזוי המשני, האם עיליה?
- עד כמה מספר זיהוי (ID) של בעית האבטחה הוא רלוונטי לחומרת התקיפה?
- האם כל התקיפה (חזקה או חלש) מצריכה פעולה מחדש לאחר העדכון?
- מה הקריטריונים שמאפיינים התקיפה חזקה?
- האם ניתן לחזות את הרכב המושפע או המוצר שהושפע (מאותה התקיפה) ע"פ התוכנות?



איור 3 : גרפּ המציג את התוכנה המשפיעה ביותר

### **לסיכום,**

בשלב זה הערכנו את ביצועי המודלים לחיזוי חומרת תקיפות הסייבר, במטרה לאפשר תגובה מותאמת לסייעו ותיעודו משאבים שיובילו לעד העסקי. מודלים כמו **CatBoost** הרואו דיקוב גביה מאד, במיוחד בזיהוי תקיפות חמורות וזהו מספר מאפיינים קריטיים שתורמים לחיזוי. דירוג המודלים נעשה ע"פ דיקוק זמני התגובה .

בנוסף הועלו מס' שאלות להמשך חקירה כגון : מהי השפעת החומרה הראשונית? מהי חשיבות מזיהוי תקיפות? והאם חקירה עתידית של משתנה אחר יכולה להועיל?.

שאלות אלו מהוות בסיס לפוליה עתידית במסגרת תהליכי שיפור המודל, בדגש על שיטות Feature Engineering מתקדמיות, ניתוח סדרות זמן, ושיילוב מידע הקשור נסף, שיכול לתרום לייצירת מערכת ניבוי חכמה, מהירה ומדויקת יותר.



## 2. תהליכי הסקירה:

בסעיף זה נסקור את תהליכי הלמידה שבערנו במהלך העבודה על הפרויקט ועל חברות מיקרוסופט. נבחן את החלטות והפעולות שביצענו – מה נעשה בצורה טובה, ומה ניתן היה לבצע טוב יותר. מטרת הסקירה היא להפיק לקחים שימושתיים אשר יסייעו לנו בפרויקטים עתידיים. נציג את התוצאות שבhem ניתן היה להשתפר – החל משלב בחירת הנושא ועד להצגת הדוחות לאורך הדרך. כתע' מעבור דוח – דוח וננתח את השלבים שביצענו וננסה להסיק מסקנות לפרויקטים הבאים.

### • דוח בחירת המשא

בחירת הנושא היא שלב מרכזי וקריטי בפרויקט בתחום מדעי הנתונים, שכן מדובר בהחלטה אסטרטגית שמשפיעה על כל שלבי הפרויקט – החל מהשגת הנתונים, דרך ניתוחם, ועד להסקת המסקנות. בפרויקט זה, בחרנו לעסוק בחיזוי רמת החומרה של תקיפות סייבר, בהתבסס על נתונים אבטחה שנאספו מארגון Microsoft.

#### מה הצליח?

- רלוונטיות - בחרנו נושא אקטואלי, רלוונטי ומעורר עניין.
- שילוב - הנושא משלב עולם תוכן עסקית וטכנולוגית – אבטחת מידע, ניתוח נתונים, מודלים חכמים.
- אמינות - הדעתה שנבחר מגיע ממקור אמין ומכליל מעל 20 אלף תצפיות.

#### מה ניתן לשפר?

- ניתן היה לבחון נושאים נוספים או לשלב כמה תת-נושאים וחיזוי מספר מטרות (כגון: חיזוי חומרה של תקיפה גם בזמן תגובה).
- כדאי היה לבדוק מראש אילו מדרדים קיימים לצורך הערכת הצלחה עסקית, כדי להתאים את מטרות הנитוח.

### • דוח עסקי

הדו"ח עסקי מكيف ומסביר את הארגון ממנו לכוחה הבעייתי שבנתונים.

#### מה הצליח?

- הסבר על הארגון בצורה מפורטת ומקיפה, התעמקות במגוון רחב של נושאים של הארגון ומאפייניהם, הסבר כולל על הפרטיהם של הארגון הגדול.
- חיברנו את הניתוח ל蹶ה עסקית ברור: תיעוד תקיפות, חיסכון בזמן/כסף, שיפור תגובה ארגונית.

#### מה ניתן לשפר?

- ארגון מיקרוסופט הוא גוף גדול ומרכזי, אשר מושך אליו קהילות ענק של מדענים נתונים. חיפוש ממוקד ועמוק יותר אחר מקורות נתונים היה עשוי להוביל למציאת DATA אינטליית ורלוונטי אף יותר לפרויקט.
- מיקרוסופט פועלת בתחוםים מגוונים ורבים. בדיעבד, הינו יכולים לצמצם את המיקוד ולהתמקד ביחסים מסוימים בתוך הארגון – לדוגמה, מחלקה Azure, מחלקה Windows, מחלקה Microsoft 365, או מחלקה Defender (שבבעלות מיקרוסופט). מיקוד כזה היה מאפשר להעמק את ההבנה ולהפיך תובנות מוקדמות ומדויקות אף יותר.



## • דוח הבנת נתונים

דוח זה עוסק בהבנת מאפייני מערך הנתונים ששימש אותנו בפרויקט. הוא כולל ניתוח של איות הנתונים, תיאור של השדות והמשתנים השונים, הסבר על המשמעות שלהם, וכן סקירה של קשרים ותבניות שנמצאו בין משתנים. מטרת הדוח היא לספק תמונה ברורה ו邏輯ית של הנתונים לפני תחילת שלבי הניתוח והבנית המודלים.

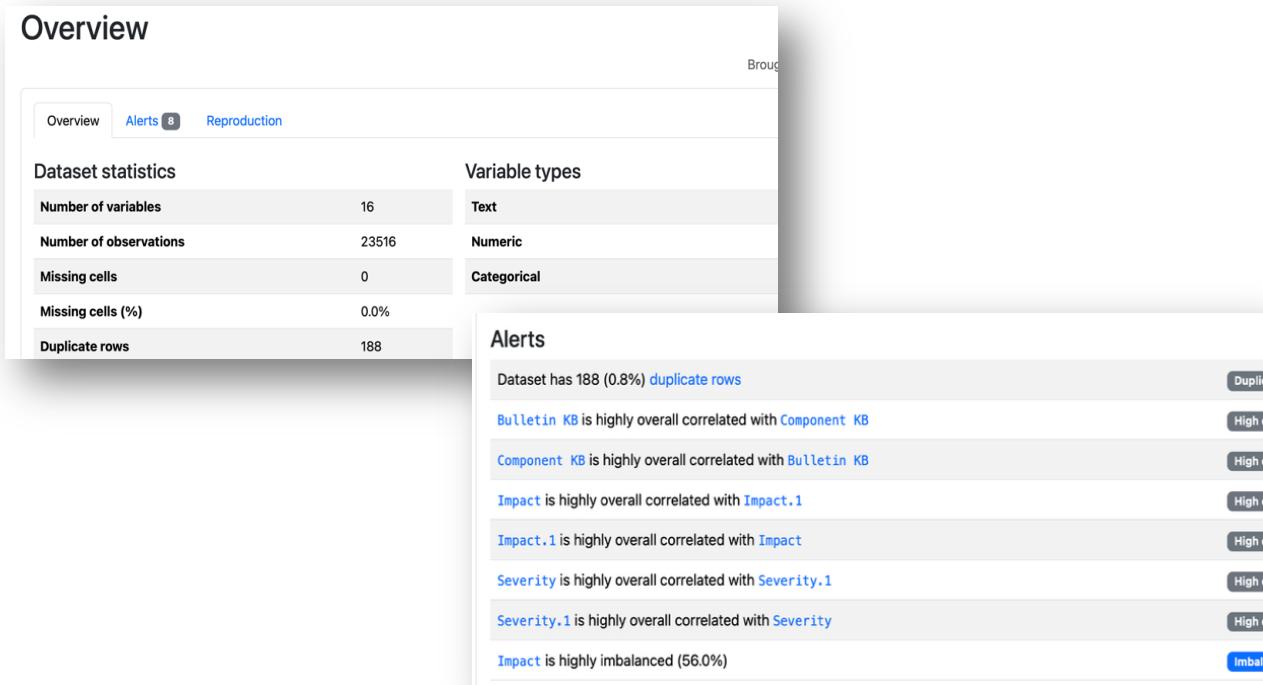
### מה הצלח?

- זיהנו את סוגי הפיצרים וקשרים שונים למשתנה המטרה.
- ביצעו ויזואлизציות וגרפים שונים שמסבירים את הנתונים.

### מה ניתן לשפר?

- יכולנו להשתמש בספריות אוטומטיות לחקר נתונים כמו - **import ProfileReport** שזוהי ספרייה שיצרת דף HTML מותאם עבור ה-**Data Frame**.

## Overview



### הסכלה

הו דוח המועד לניתוח והבנה של פרופיל הנתונים המיובאים לארגון או למערכת. דוח זה מספק תובנות לגבי הנתונים שנכנסים למערכת, כולל מאפיינים, איות, וקשרים בין הנתונים השונים.



## • דוח הכנות הנתוניים

דוח זה מתאר את שלבי ההכנה והעיבוד שבוצעו על מערך הנתוניים במטרה להתאיםו ללמידה מכונה. תהליכי ההכנה כללו טיפול בערכים חסרים (NA) תיקון שגיאות, סינון משתנים לא רלוונטיים, המרת טיפוסי נתונים, והנדסת תכונות במידת הצורך. שלבים אלה בוצעו על מנת לשפר את איכות הדadata ולהבטיח שהמודל יתבסס על מידע נקי, עיקבי ומשמעותי.

### מה הצליח?

- ביצעו ניקוי כולל עם שלבים ברורים ומוגדרים.
- בדקנו קשרים חשובים למשתנה מטרה.
- בחרנו את העמודות הרלוונטיות

### מה ניתן לשפר?

- תהליכי ניקוי הנתוניים היה יכול להתבצע בדרךים נוספות, כגון שימוש בשיטות חכמות יותר למילוי ערכים חסרים (כמו היישוב ממוצעים לפי קבוצות, שימוש במודלים לחיזוי ערכים חסרים, או טכניקות איטרפטטיביות), במקום הסירה או מילוי גורף.
- אופן ההתמודדות עם רשותות חלקיות – לדוגמה, רשותות הכלולות בערכים חסרים – היה יכול להיבחן לעומק: במקרים מסוימים ייתכן שהיה עדיף להשאיר את הרשותות ולملא ערכים חסרים, ולא להסירן, על מנת לשמור כמה שיותר מה מידע המקורי.

## • דוח המידיל

דוח זה מתאר את תהליכי הבנייה וההערכה של אלגוריתמים שנעודו לחזות את תוכנות המטרה שנבחרה במסגרת הפרויקט. הוא כולל סקירה של האלגוריתמים שנבחרו, הסבר על אופן הבחירה בהם, תיאור של שלבי האימון והבדיקה של המודלים, וכן ניתוח ביצועים באמצעות מדדים מתאימים (כגון: דיקוק, AUC, F1, ועוד). מטרת הדוח היא להציג כיצד תהליך המידול בוצע בפועל, מהן התוצאות שהתקבלו, ומה ניתן להסיק מיהן לגביייעיל ואמין.

### מה הצליח?

- בפרויקט זה בחנו מספר מודלים מתקדמים, בהם : CatBoost, LightGBM, XGBoost ואחרים.
- השווינו בין המודלים באמצעות מדדי ביצועים שונים, וביצעו ניסויים עם פרמטרים מגוונים על מנת לשפר את תוצאות החיזוי. התהlikן כלל הערכת ביצועים, כוונון היפר-פרמטרים ידני, והשוואה שיטתיות בין האלגוריתמים.

### מה ניתן לשפר?

- ניתן היה לשלב כלים אוטומטיים מתקדמים כגון Optuna – ספרייה לאופטימיזציה פרמטרים המבצעת חיפוש חכם ומהיר אחר שילובי הפרמטרים הטובים ביותר למודלים, ובכך לייעל את תהליכי הכוון ולהשוך זמן.
- כמו כן, ניתן היה לשלב שיטות למידת Ensemble נספנות (כגון או Blending - שילוב תוצאות של כמה מודלים בעזרת מודל נוסף "עלון") (meta-model). כדי לשפר את הביצועים הכלליים של המודל.
- בנוסף, שילוב של כלי AutoML – אשר מבצעים אוטומטית לכל שלבי בניית המודל, מבחרת האלגוריתם ועד לכונן פרמטרים – עשוי היה להוביל לפתרונות מדויקים ויעילים יותר, תוך חיסכון משמעותי בזמן ובמשאבים.



### כליים של AutoML לבחינה עתידית:

במהלך המשך העבודה והעמקת היכולות בתחום חיזוי ואופטימיזציה של מודלים, ניתן לשקלות שימוש בכלים מתאימים של AutoML שמבצעים אוטומציה לתהליכי המלא של בניית מודלים. להלן מספר כלים בולטים:

#### ○ H2O AutoML

תומך במשימות של רגרסיה, סיוג וסדרות זמן. משלב מודלים מתאימים כגון Random Forest ו-XGBoost Ensemble (Ensembling) שמאפשר שיפור ביצועים על ידי שילוב מספר מודלים.

#### ○ Auto-sklearn

مبוסס על ספריית scikit-learn, ותומך ברגression ו-classification. משתמש באלגוריתמים מוכרים כגון Random Forest ו-SVM, ומבצע חיפוש חכם אחר שילוב המודלים והפרמטרים האופטימליים.

#### ○ PyCaret

מספקת ממשק פשוט ונוח לישום משימות של סיוג, רגרסיה וניתוח שפה טבעית (NLP). עשויה שימוש באלגוריתמים כמו XGBoost ו-LightGBM, ומבצעת את כל תהליכי ההשווואה והכוונון של מודלים בפקוודות בודדות.

#### ○ FLAML (Fast and Lightweight AutoML)

מתמקד בהפחיתה עלויות חישוב וזמן ריצה. תומכת ברגression, סיוג וסדרות זמן, וمبוססת בעיקר על מודלים כגון LightGBM ו-XGBoost. אידיאלית לפרויקטים שדרושים פתרונות יעילים בזמן קצר.

#### ○ תהליכי optuna ופלט מלא:

```
● Running Optuna optimization for CatBoost...
[I 2025-05-10 00:41:38,369] Trial 0 finished with value: 0.9917915245018049 and parameters: {'iterations': 509, 'depth': 10, 'learning_rate': 0.038358359899}
[I 2025-05-10 00:41:41,729] Trial 1 finished with value: 0.9287939474855363 and parameters: {'iterations': 307, 'depth': 4, 'learning_rate': 0.023534960147}
[I 2025-05-10 00:41:57,972] Trial 2 finished with value: 0.986648865153538 and parameters: {'iterations': 549, 'depth': 7, 'learning_rate': 0.0380849241816}
[I 2025-05-10 00:42:10,725] Trial 3 finished with value: 0.9973297738387077 and parameters: {'iterations': 426, 'depth': 7, 'learning_rate': 0.192519739505}
[I 2025-05-10 00:43:29,652] Trial 4 finished with value: 0.9957968649557435 and parameters: {'iterations': 651, 'depth': 10, 'learning_rate': 0.04199170882}
[I 2025-05-10 00:43:36,474] Trial 5 finished with value: 0.9886762597042971 and parameters: {'iterations': 501, 'depth': 5, 'learning_rate': 0.067027825222}
[I 2025-05-10 00:43:40,645] Trial 6 finished with value: 0.9449636552440291 and parameters: {'iterations': 251, 'depth': 5, 'learning_rate': 0.040883993396}
```

אייר 4 : תהליכי ההרצה של Optuna

```
def objective(trial):
    params = {
        'iterations': trial.suggest_int("iterations", 300, 800),
        'depth': trial.suggest_int("depth", 6, 14),
        'learning_rate': trial.suggest_float("learning_rate", 0.01, 0.3, log=True),
        'l2_leaf_reg': trial.suggest_float("l2_leaf_reg", 1.0, 10.0, log=True),
        'border_count': trial.suggest_int("border_count", 64, 128),
        'verbose': 0,
        'random_seed': 42
    }
    model = CatBoostClassifier(**params)
    score = cross_val_score(model, X_train, y_train, cv=3, scoring='accuracy').mean()
    return score
```

אייר 5 : הגדרת טווח הפרמטרים לחיפוש



בתמונה ניתן לראות את Optuna בשלבי הריצה שלה, כל שורה אדומה היא הרצה של מודל CatBoost או מודל לבחירה עם טווח הפרמטרים שניתנו למודל, לדוגמא בשורה הראשונה מקבל מדד value והפרמטרים של המודל Optuna. הריצה מספר מודלים בהב� לקלט המשמש או אם אין שינוי מהותי בערך value והמודל לא משפר את עצמו, Optuna מסתיים ורואה את התוצאות הרצויות של המודל בעל value הטוב ביותר. עבור דאטה רב, Optuna מרים מספר רב של מודלים עם ביג דאטה.

**ב-”0 Trial”,** המודל הוגדר עם 509 עצים, עומק מקסימלי של 10, שיעור למידה של 0.038, רגולציה של 0.135 וגבול של 148. הערך של פונקציית המטרה עבור ניסיון זה היה 0.9917915245018049. כל ניסיון מסתיים עם>Status "finished", המציין שהניסיון הושלם בהצלחה.

תהליך האופטימיזציה מנסה למצוא את הצירוף הטוב ביותר של פרמטרים כדי לשפר את הביצועים של המודל, כאשר כל ניסיון מציג את הערכים השונים של הפרמטרים ואת התוצאה המתבקשת. זהו תהליך אטרקטיבי שמטרתו למצוא את המודל האופטימלי עבור המשימה הנANTAה.

### כעת ננתן את שלבי הריצה של Optuna:

#### מה קורה בכל ריצה?

בכל ריצה Optuna בחרת ערכים להיפר-פרמטרים כמו מספר העצים, עומק העצים, גודל הדגימה ועוד. המודל מאומן עם הערכים שנבחרו, מבוצעת הרצה של הביצועים ונשמרת תוצאה. על סמך תוצאות אלו, Optuna לומדת אילו שילובים משתלימים יותר וمعدכנת את הבחירה שלה לבניית הבאים – כך נבנית אופטימיזציה חכמה.

#### מה קורה עם ביג דאטה?

כאשר מדובר ב大数据 ביג דאטה, משך הזמן של כל הריצה מודרך ממשמעותית. לכן, מומלץ לבצע במדדי ביצוע כמו - Cross Validation בוסף, יותר שידרשו יותר ניסויים כדי להגיע למודל האופטימלי. במצבים כאלה חשוב לנצל תשתיות כמו GPU או ריבוי ליביות, ולהפעיל מגננו שmpsיק הריצת מודלים שלא מראים שיפור מובהק.

#### איך Optuna יודעת متى לעצור?

ניתן להגדיר מראש מספר ניסויים מקסימלי (n\_trials) או הגבלת זמן כוללת (timeout). מעבר לכך, Optuna עשויה להפסיק את תהליך האופטימיזציה מוקדם יותר, אם היא מזהה שאין שיפור ממשמעותי בתוצאה – לדוגמה, אם 20 הריצות אחרונות לא הביאו לשיפור במדד היעד.

בסיום התהליך, Optuna מדפסת את המודל עם התוצאה הטובה ביותר, יחד עם ערכי ההיפר-פרמטרים שהניבו אותה. ניתן לשמור את המודל, לציר גրפים של ביצועים לאורך הזמן, ולהשוות בין מודלים שונים.

תהליך Optuna **איינו מטאים לביג דאטה** והוא יכול לקחת מספר שעות לאמן על ביג דאטה, עם זאת נריץ על הדאטה המקורי שכולל מעל 20,000 רשומות. עבור 15 הריצות שיפור קיבלנו את הנtau�ים הבאים, accuracy = 0.9975 :

	precision	recall	f1-score	support	accuracy
Critical	0.998348	0.997799	0.998073	1817.000000	0.997571
Important	0.997843	0.997843	0.997843	1391.000000	0.997571
Low	0.600000	1.000000	0.750000	3.000000	0.997571
Moderate	1.000000	0.987805	0.993865	82.000000	0.997571
accuracy	0.997571	0.997571	0.997571	0.997571	0.997571
macro avg	0.899048	0.995862	0.934945	3293.000000	0.997571
weighted avg	0.997813	0.997571	0.997645	3293.000000	0.997571
{'iterations': 758, 'depth': 8, 'learning_rate': 0.2894324972736559, 'l2_leaf_reg': 0.4447181713345161, 'border_count': 172}					

איור 6 : מדריך Optuna



### הסבר פלט פרמטרים :

במהלך תהליך הטיבוב (Hyperparameter Tuning) של מודל Random Forest, התקבלו הפרמטרים הבאים אשר סייפקו את הביצועים הטובים ביותר :

**iterations :** מספר האיטרציות (עצים) שנבנו במהלך האימון של המודל. ערכים גבוהים יותר מצינים מודל מורכב יותר.

**depth :** העומק המקסימלי של כל עץ במודל. ערכים גבוהים יותר מצינים מרכיבים יותר.

**learning\_rate :** קצב הלמידה של המודל, המשפיע על כמה כל עץ חדש תורם לתיקון השגיאות של העצים הקודמים. ערכים נמוכים יותר מציגים תיקון איטי יותר, אך יכולים להוביל לביצועים טובים יותר.

**l2\_leaf\_reg :** פרמטר רגולרייזציה L2 עבור העלים בעצים. ערכים גבוהים יותר מציגים רגולרייזציה חזקה יותר, מה שיכל למנוע התאמה יתר (overfitting).

**border\_count :** מספר הגבולות (split points) המשמשים לחלוקת התוכנות בעצים. ערכים גבוהים יותר מציגים חלוקות מורכבות יותר.

מתי להשתמש	זמן חישוב	סיכון ל-Overfitting	רמת דיק פוטנציאלית	border_count
כאשר רצים הרצה מהירה או עבר נתונים פשוטים מאוד. טוב לניסויים וראשוניים.	מהיר	נמוך מאוד	נמוכה	32–16
איזון טוב לbijoux מהירים עם מעט אובדן מידע. מתאים לפרויקטים עם משאבים מוגבלים.	מהיר	נמוך	בינונית	64
ברירת מחדל טובה – לרובה מציע איזון טוב בין דיק וזמן.	בינו	בינוני	גבוהה	128
לשיפור מקסימלי של דיק, בעיקר במקרים סופיים או תחרויות. יש לבדוק אם המודל לא לומד רעש.	אחר	גבוהה	גבוהה מאוד	255–254
עלול לפגוע ביצועים; לא מומלץ לשימוש בפועל. CatBoost מוגבל בד"כ ל-255.	לא נתמך או לא יציב	גבוהה מאוד	<255 (לא מומלץ)	

איור 7 : תוכנת Optuna והערכת border count של ערכים שלו

עם קידוד נכון של הנתונים אפשר להפעיל את תהליך Optuna על מספר מודלים שונים ביניהם Random Forest, lightGBM, XGBoost. ניתן להניח כי שימוש בתהליכי Optuna יראה תוצאות אופטימליות יותר, הפחיתה רעש והפחיתה התאמת יתר הן תוצאות ישירות של השימוש ב-Optuna לאופטימיזציה של היפר-פרמטרים. הפחיתה רעש יכולה להתרחש כאשר אנו מתאים את הפרמטרים לצורה חכמה, במקום הסתמך על ניסויים אקראיים או הערכות ידניות של פרמטרים. חישוב חכם אחורי ערכים אופטימליים בעזרת Optuna ממצמצם את האפשרות להגיע לפרמטרים לא מתאים או לא אופטימליים שימושיים לרעש בתוצאות, בכך שהוא מקטין את השפעת הסטיית-תקן של תוצאות המודל (variance). בנוסף, הפחיתה התאמת יתר מתבצעת כיוון ש-Optuna מבצע אופטימיזציה



תוך שימוש במדדי ביצועים. ולרוב אנחנו משתמשים גם בנתונים נפרדים כדי לבדוק את הביצועים על נתונים מבחן.

מבצע אופטימיזציה תוך כדי שימוש ב-validation על כל פרמטר, כך שנמנע מהמודל ללמידה יותר מדי על הנתונים (overfitting) ולהתאים את עצמו בדיקות יתר על המידה לנתונים ספציפיים של קבוצת האימון, ובכך מצמצם את הסיכון של התאמת יתר.