ב"ה

# הנחיות למטלה מסכמת pipeline-דאטה סיינאס בסייבר

## 1. בחירת דאטאסאט

- בחרו דאטאסאט מ־ Kaggle שנראה לכם מעניין.
- תוודאו שהוא כולל מספיק תכונות(features) , ושאפשר לבצע עליו ניתוח מגוון.
- נסחו בקצרה למה בחרתם דווקא את הדאטאסאט הזה.

## 2. מערכת קבצים(System stage)

- תארו את קבצי הדאטה: שמות, גודל, סוגי קבצים.
- אם רלוונטי – האם הגיע מאתר? מה הפרוטוקול/פורמט?
- האם יש גרסאות שונות? שמרו על ניהול גרסאות.

## 3. Meta Data

- סוגי הנתונים (מספרי/טקסטואלי/תאריכים).
- האם יש ערכים חסרים?
- האם יש ערכים מיוחדים כמו למשל "-999", "unknown" וכו'.

## 4. סטטיסטיקות

- תארו את ההתפלגויות של עמודות נבחרות.
- חישבו סטטיסטיקות מרכזיות (mean, std, min/max) וכו'.
- בדקו מתאמים בין משתנים.
- האם יש שכפולים/ערכים ייחודיים בלבד?
- האם כדאי לבצע הפחתת מימדים (PCA) או דומה?

## 5. איתור חריגות(Abnormality detection)

- חריגות על תכונה אחת(outliers)
- חריגות על פני כמה תכונות.
- האם יש הסבר סיבתי/תחומי לחריגות?

## 6. Clustering

- בצעו clustering (כמו KMeans ).
- האם אפשר לתת משמעות לקבוצות?
- האם יש נקודות שלא שייכות לשום קלאסטר?

## 7. ניתוח מגזרים/סגמנטים

- אילו תכונות מאפיינות כל סגמנט?
- האם יש מרכיבים זמניים(temporal) ?

- האם יש ידע מוקדם מהתחום שמחזק את המסקנות?

8. **עיבוד שפה טבעית (אם יש טקסט)**
   - האם ניתן לבצע:
     - ניתוח סנטימנט?
     - חלוקת נושאים?
     - זיהוי סגנון כתיבה?

9. **גרפים**
   - צרו גרפים משמעותיים (למשל: גרף מידע, מבנה ארגוני, זרימת מידע).
   - אם רלוונטי – מיפוי נקודות תורפה (כמו מחשבים רגישים).

10. **מודלים**
    - בנו מודלים מתאימים לבעיה.
    - האם הם מסבירים את הנתונים?
    - האם יש התאמה טובה(Goodness of fit)
    - האם המודל ניתן להסבר(Explainable)

11. **דיווח**
    - הציגו ממצאים בצורה ברורה.
    - גרפים/טבלאות תומכים.
    - מה לדעתכם צריך לחקור הלאה?

12. **שיפור(Improve)**
    - אילו צעדים הייתם מבצעים כדי לשפר את המערכת?
    - לדוגמה:
      - שדרוג מודל.
      - חידוד פילוח.
      - הצעות לפדרציה או להפרדת רכיבים.

---

**תוצר סופי:**

- דוח כתוב (PDF או Jupiter notebook)
- מצגת מסכמת של כל השלבים.
- קוד עובד (רצוי פייתון).
- לינק לדאטאסאט שנבחר.