**פרויקט "הבייסיאני הנאיבי"**

מטרה:

עיצוב מונחה עצמים

תכנות מונחה עצמים

היכרות עם עבודה בקבצי נתונים

היכרות עם למידה מונחית וסיווג

**הוראות התרגיל:**

בתרגיל זה עליכם לממש את אלגוריתם הסיווג Naïve Bayes תוך שימוש ב- תיקון לפלס (Laplacian).

**נתונים לדוגמא:**

<https://archive.ics.uci.edu/ml/datasets/mushroom>

<https://www.kaggle.com/eswarchandt/phishing-website-detector?select=phishing.csv>

מומלץ להתחיל לפתח עם דוגמה קטנה כמו אוסף הנתונים שהוצג בכיתה.

שימו לב בכדי לענות על סעיף 4.1 על פי המקובל התחום יש לחלק את קובץ הנתונים לשני חלקים (ניתן אפילו באקסל):

1. 70% לבניית המודל
2. 30% לבדיקת המודל

**מהלך התוכנית:**

1. התוכנית תקבל קובץ נתונים בפורמט CSV
2. התוכנית תבנה מודל (אוסף הסטטיסטיקות המותנות הנדרשות לחישובים השונים)
3. לאחר בניית המודל בהצלחה תנתן הודעה למשתמש.
4. התוכנית תבקש קלט לסיווג באחת מ 2 אפשרויות:
   1. קובץ נתונים לבדיקת דיוק המודל – אוסף רשומות CSV שיש להם ערכי סיווג.
      1. עליכם להעביר לסיווג את הרשומות ללא ערך הסיווג
      2. עבור כל רשומה יש לבדוק האם ערך הסיווג תואם לערך המקורי או לא
      3. יש לספור את כמות הפעמים שהסיווג היה נכון
      4. בסיום סיווג כל רשומות הקובץ יש להציג אחוז דיוק כולל למודל
   2. קלט רשומה בודדת לסיווג ופלט תוצאת הסיווג

**תכנון מחלקות :**

1. עליכם לממש את המסווג כמחלקה עצמאית שאינה תלויה במחלקות אחרות בקוד.  
   עליכם להתייחס לנקודות הבאות:
   1. מהי הפונקציונליות הכללית הנדרשת לכל מסווג באשר הוא
   2. מהי הפונקציונליות היחודית לאלגוריתם
   3. איך לייצג את הנתונים
   4. איך לייצג את המודל
2. עליכם לתכנן מחלקה שתשתמש במודל וברשומה חדשה ללא סיווג, ותבצע את הסיווג של הרשומה.
   1. מומלץ שמחלקה זו תכיל את המסווג.
3. עליכם לתכנן ממשק משתמש כמחלקה.
   1. הממשק צריך להיות גנרי כך שאם יוחלט להחליפו בהמשך בממשק גרפי- זה יתבצע בקלות.
4. עליכם לתכנן מחלקה שתטפל בטעינת קובץ והמרתו למבנה מתאים עפי התכנון שלכם
5. עליכם לתכנן מחלקה שתאתחל ותכיל את האובייקטים השונים, מחלקה זו תנהל את ההתהליכים השונים (מחלקה זו תופעל מה main)

**הערות:**

למעט פונקציית ה main אין להשתמש בפונקציות כלליות אלא במחלקות מתודות ומאפיינים.

עליכם לתכנן **לפני תחילת הקידוד** את המחלקות השונות והקשרים ביניהם (בפרויקט זה אין חובה להשתמש בתורשה, ויש לתת דגש לקשרים של הכלה)

יש להקפיד על משתנים חבויים וגישה למשתנים אך ורק בעזרת מתודות ולוגיקה מתאימה.

ניתן להניח שהמידע בקובץ תקין ואין צורך לבדוק את תקינותו. (כמובן שיש לוודא שהקובץ נפתח ללא חריגות)

יש לנהל את הפרויקט בגיט

**רפרנסים:**

<https://machinelearningmastery.com/naive-bayes-tutorial-for-machine-learning/>

**מילות מפתח לחיפוש**

Classifier

Data mining

Machine learning

Naïve Bayes

Classification