המכללה האקדמית להנדסה ע"ש סמי שמעון

המחלקה למדעי המחשב

קורס: מבוא לכריית הנתונים – תשפ"ג

מטלה 2 – **Data Preprocessing**

תאריך הגשה: 03/05/2023

מטרת המטלה

התנסות הסטודנטים בחישוב אנטרופיה ועצי החלטה, עיבוד מידע וניתוח חקירתי של המידע (EDA - Exploratory data analysis) .

תוכן המטלה

בקורס כריית נתונים נלמד להוציא תובנות ותבניות שכיחות ומעניינות ממידע נתון. עם זאת, נרצה למצוא מאפיינים עיקריים בסט מידע (dataset) שברשותנו ולנתח אותם. נבצע על הנתונים שלנו עיבוד מקדים בעזרת מניפולציות או השמטה של נתונים לפני השימוש בסט מידע (dataset) על מנת לשפר את הביצועים. חלק זה מהווה שלב חשוב בתהליך כרית הנתונים.

במשימה זו הנכם מתבקשים להשתמש בסט מידע (dataset) שמצורף במודל.

ניתן להשתמש בחבילות/ספריות הבאות:

```
import pandas as pd
import numpy as np
import math
import scipy.stats as stats
from sklearn.preprocessing import MinMaxScaler
```

במידה וישנה חבילה/ספריה נוספת שתרצו להשתמש חובה לשאול בפורום המטלה.

<u>שלב א': חישוב אנטרופיה (50 נקודות)</u>

נתון המידע הבא:

x, z – משתנים מקריים בדידים. המשתנה x יכול לקבל ערכים a,b,c,d והמשתנה z יכול לקבל ערכים i, j, k. בתוך הטבלה, הערך מציין את מספר המופעים בבסיס הנתונים של הערכים המתאימים. יש להתייחס ל-2 ספרות לאחר הנקודה.

data	z=i	z=j	z=k	Total
x=a	7	8	17	
x=b	6	3	11	
x=c	14	1	9	

x=d	4	15	2	
Total				

א. יש לחשב את האנטרופיה של המשתנים x ו-z.
ב. מהי האנטרופיה המקסימלית עבור כל אחד מהמשתנים x ו-z?
ג. יש לחשב אנטרופיה מותנית H(z|x) וגם H(x|z).
ד. יש לחשב Mutual Information (MI) עבור x ו-z. הוכיחו (ע"י חישוב) ש-MI סימטרי.

<u>שלב ג': עיבוד מקדים (50 נקודות)</u>

קובץ ה- CSV בשם wine.csv שמצורף במודל מכיל את העמודות הבאות:

1. alcohol
2. malic acid
3. ash
4. Alkalinity of ash
5. magnesium
6. Total phenols
7. flavanoids
8. Nonflavanoid phenols
9. Proanthocyanins
10. Color intensity
11. hue
12. od280/od315_of_diluted_wines
13. proline
14. target

בשלב הזה עליכם לבצע עיבוד מקדים לנתונים:

- טענו את קובץ הcsv.
- השלימו נתונים חסרים לפי הממוצע בעמודה.
- מצאו את 3 העמודות המספריות (רמז: היעזרו בdescribe) ביניהן יש את ההבדל הגדול ביותר בין הערך המינימלי בעמודה לערך המקסימלי בעמודה:
  1. העמודה הראשונה זו עם ההבדל הגדול ביותר, בצעו נרמול ע"י min max
  2. עבור העמודה עם ההבדל הגדול השני בגודלו, בצעו סטנדרטיזציה ע"י z-score
  3. העמודה השלישית חלקו ל-3 פחים עם עומק שווה ובצעו החלקה ע"פ bin boundaries

הוראות הגשה

1. עליכם להגיש קובץ pdf הכולל את מספרי תעודות הזהות של שני המגישים עם קו תחתון מפריד ביניהם בצורה הבאה: ID1_ID2. pdf.
2. תשובה סופית, גם אם היא נכונה, ללא הצגת דרך החישוב המלאה, לא תזכה בניקוד.
3. את הקובץ pdf שיצרתם עליכם להעלות לתיבת ההגשה.
4. הגשה <u>בזוגות בלבד</u>.
5. תאריך ההגשה המעודכן ומיקום ההגשה יופיעו באתר הקורס.