מעבדה 5 - עץ החלטה

הוראות המעבדה:

בתרגיל זה עליכם לממש את האלגוריתם המסווג עץ החלטה ID3 .

קובץ הנתונים מכיל גם תכונות רציפות וגם תכונות קטגוריאליות. בנוסף, ייתכנו רשומות עם ערכים חסרים, בהם יש לטפל כחלק מתהליך ניקוי הנתונים.

שלבי עבודת התוכנית:

1. קלט לכמות בינים עבור דיסקרטיזציה.
2. קלט קובץ מבנה, קובץ אימון, וקובץ בדיקה.
3. טעינת קבצים
4. השלמת ערכים חסרים בשיטה לבחירתכם
5. דיסקרטיזציה לעמודות רציפות באופן זהה לקובץ האימון וקובץ הבדיקה.
6. אימון מודל בעזרת קובץ האימון
7. בדיקת המודל בעזרת קובץ הבדיקה
8. הצגת נתוני הסיווג של קובץ הבדיקה (כמה טעויות היו לנו)

תיאור הקלט (קבצים):

1. **Dataset general info** – מידע כללי בנוגע לבסיס הנתונים ממנו לקוחים נתוני התרגיל. קובץ זה הינו לשימושכם בלבד <u>ולא ישמש כנתון שעל תכניתכם לקרוא במהלך הריצה</u>.
2. **Structure** – קובץ המתאר את התכונות המרכיבות כל רשומה בבסיס הנתונים (כולל ערך המטרה אשר מופיע אחרון ברשימה). הקובץ ישמש את התוכנית שלכם להכרת מבנה בסיס הנתונים בו עליה לטפל .

   <u>יש להקפיד על המבנה</u> המתואר בקובץ ועל <u>סדר הופעת התכונות</u> (Features).

   1. תכונת המטרה תקרא תמיד "class", ותהיה אחרונה בקובץ ובכל רשומה.
   2. יש להשתמש בקובץ זה על מנת לחלץ את הערכים הייחודיים השונים אותם כל תכונה יכולה לקבל. בנוסף, ניתן לדעת על פי הקובץ מי מהתכונות היא נומרית או קטגוריאלית.
      - נומרית

@ATTRIBUTE [AttTitle] NUMERIC

      - קטגוריאלית

@ATTRIBUTE [AttTitle] {some comma separated categories}

דוגמא למבנה הקובץ שניתן לכם

@ATTRIBUTE age NUMERIC
@ATTRIBUTE job
{admin.,unknown,unemployed,management,housemaid,entrepreneur,student,blue-collar,self-employed,retired,technician,services}
@ATTRIBUTE marital {married,divorced,single,widowed}
@ATTRIBUTE education {unknown,secondary,primary,tertiary}
@ATTRIBUTE default {yes,no}
@ATTRIBUTE balance NUMERIC
@ATTRIBUTE housing {yes,no}
@ATTRIBUTE loan {yes,no}
@ATTRIBUTE contact {unknown,telephone,cellular}
@ATTRIBUTE day NUMERIC
@ATTRIBUTE month {jan,feb,mar,apr,may,jun,jul,aug,sep,oct,nov,dec}
@ATTRIBUTE duration NUMERIC
@ATTRIBUTE campaign NUMERIC
@ATTRIBUTE previous NUMERIC
@ATTRIBUTE poutcome {unknown,other,failure,success}
@ATTRIBUTE class {yes,no}

3. **train** – קובץ המכיל רשומות שישמשו לבניית המסווג. לשם פשטות, הקובץ מסוג CSV. כל רשומה מופיעה בשורה נפרדת. (ניתן להניח שהקובץ יינתן עם כותרות בעמודות)

P	O	N	M	L	K	J	I	H	G	F	E	D	C	B	A	
class	poutcome	previous	campaign	duration	month	day	contact	loan	housing	balance	default	education	marital	job	age	1
no	unknown	0	1	261	may	5	unknown	no	yes	2143	no	tertiary	married	manageme	58	2
no	unknown	0	1	151	may	5	unknown	no	yes	29	no	secondary	single	technician	44	3
no	unknown	0	1	76	may	5	unknown	yes	yes	2	no	secondary	married	entreprene	33	4
no	unknown	0	1	92	may	5	unknown	no	yes	1506	no	unknown	married	blue-collar	47	5
no	unknown	0	1	198	may	5	unknown	no	no	1	no	unknown	single	unknown	33	6
no	unknown	0	1	139	may	5	unknown	no	yes	231	no	tertiary	married	manageme	35	7
no	unknown	0	1	217	may	5	unknown	yes	yes	447	no	tertiary	single	manageme	28	8
no	unknown	0	1	380	may	5	unknown	no	yes	2	yes	tertiary	divorced	entreprene	42	9
no	unknown	0	1	50	may	5	unknown	no	yes	121	no	primary	married	retired	58	10
no	unknown	0	1	55	may	5	unknown	no	yes	593	no	secondary	single	technician	43	11
no	unknown	0	1	222	may	5	unknown	no	yes	270	no	secondary	divorced	admin.	41	12
no	unknown	0	1	137	may	5	unknown	no	yes	390	no	secondary	single	admin.	29	13
no	unknown	0	1	517	may	5	unknown	no	yes	6	no	secondary	married	technician	53	14
no	unknown	0	1	71	may	5	unknown	no	yes	71	no	unknown	married	technician	58	15
no	unknown	0	1	174	may	5	unknown	no	yes	162	no	secondary	married	services	57	16

4. **test** – קובץ המכיל רשומות שאותן תצטרכו לסווג. לשם פשטות, הקובץ מסוג CSV. כל רשומה מופיעה בשורה נפרדת. שימו לב, בקובץ זה מופיע הסיווג האמיתי של כל רשומה, אך **אין לכם כל צורך להשתמש בו לצורך הסיווג, אלא ביתר התכונות בלבד**.

השימוש בסיווג המקורי יהיה עבור מדידת הדיוק. (ניתן להניח שהקובץ יגיע עם כותרות בעמודות)

בהצלחה!