|  |  |  |
| --- | --- | --- |
| **מבנה הבחינה:** | לבחינה שני חלקים. יש לענות על כל השאלות. | |
|  | **חלק א:** | ענו על **ארבעה** השאלות שבחלק זה.  (90 נקודות לחלק זה). |
|  | **חלק ב:** | ענו על **השאלה** שבחלק זה.  (10 נקודות לחלק זה). |

משקל כל שאלה מפורט בגוף השאלון.

אפשר להשתמש בהנחות במהלך הפתרון, במידה ותזדקקו לכך. פרטו הנחותיכם!

חומר עזר: כל חומר עזר מותר בשימוש.

## חלק א (90 נקודות)

בחלק זה **ארבע** שאלות. ענו במפורט על **ארבע** השאלות.

### שאלה 1 (30 נקודות) – עצי החלטה (decision tree)

נתונה טבלת נתוני אימון:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| מס' סידורי | מזג האויר | טמפרטורה | לחות | רוח | טיול |
| 1 | שמש | גבוהה | גבוהה | ל | ל |
| 2 | שמש | גבוהה | גבוהה | כ | ל |
| 3 | מעונן חלקית | גבוהה | גבוהה | ל | כ |
| 4 | גשום | בינונית | גבוהה | ל | כ |
| 5 | גשום | נמוכה | נורמלית | ל | כ |
| 6 | גשום | נמוכה | נורמלית | כ | ל |
| 7 | מעונן חלקית | נמוכה | נורמלית | כ | כ |
| 8 | שמש | בינונית | גבוהה | ל | ל |
| 9 | שמש | נמוכה | נורמלית | ל | כ |
| 10 | גשום | בינונית | נורמלית | ל | כ |
| 11 | שמש | בינונית | נורמלית | כ | כ |
| 12 | מעונן חלקית | בינונית | גבוהה | כ | כ |
| 13 | מעונן חלקית | גבוהה | נורמלית | ל | כ |
| 14 | גשום | בינונית | גבוהה | כ | ל |

**המשך הבחינה בעמוד הבא**

1. בנו עץ החלטה חלקי, הכולל את רמת השורש ורמה אחת נוספת בלבד, עבור נתוני האימון שבטבלה לחיזוי החלטה לצאת לטיול אופניים. בתשובתכם הדגימו את שלבי בחירת התכונה המפצלת בעץ.

**הערה:** יש לכלול חישוב של **אחד** המדדים כדוגמת אנטרופיה, Gain ratio, מדד גיני.

1. איזה מבין התכונות ניתן להסיר ומדוע? אם אין תכונה הניתנת להסרה, יש לציין זאת מפורשות.

### שאלה 2 (25 נקודות) – חוקי הקשר (association rules)

נתונה הטבלה :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| טרנזקציה | אופניים | קסדה | שרשרת | כפפות | משאבה | מגנים לברכיים |
| T1 | 1 | 1 | 0 | 1 | 0 | 1 |
| T2 | 1 | 1 | 1 | 1 | 1 | 0 |
| T3 | 1 | 1 | 1 | 0 | 1 | 0 |
| T4 | 1 | 1 | 0 | 1 | 0 | 0 |

הגדירו לפחות 5 חוקי ההקשר המקיימים :

Min\_support=60%

Min\_confidence=80%

### שאלה 3 (10 נקודות) – ניתוח אשכולות (clustering)

### נתון בסיס נתונים המכיל נתונים מסוגים שונים כדוגמת נתונים רציפים, בינאריים, נומינליים ועוד.

הגדירו והדגימו כיצד יש לייצג את הנתונים על מנת לבצע אשכול.

**שאלה 4 (25 נקודות) – למידה בייאסינית**

### באיור הבא נתונה רשת בייסיאנית המתארת את העובדות הבאות:

מחלת לב (תסומן H) מתפתחת כתוצאה מעישון (יסומן S) או/ו תזונה עתירת שומנים (יסומן D).

מחלת לב גורמת לעליה בלחץ הדם (יסומן B) או/ו לשינויים בתרשים א.ק.ג (יסומן E ).

תרשים א.ק.ג. מבטא את פעילותו החשמלית של הלב.

B S

H

E D

בהנחה ש: P(S)=0.3 P(D)=0.4

נגדיר:





1. חשבו את ההסתברות של נבדק לחלות במחלת לב – P(H) .
2. חשבו את ההסתברות שנבדק מעשן יהיה חולה לב- P(H|S) .

**חלק ב (10 נקודות)**

בחלק זה שאלה **אחת**. ענו עליה במפורט.

### שאלה 5 (10 נקודות)

במהלך קורס כריית מידע טען אחד הסטודנטים: " פתרתי את ממ"ן 21 והמודל שהתקבל הוא בעל התאמת יתר (overfitting). לכן, אנסה להגדיל את קבוצת נתוני האימון וכך שגיאת האימון (training error) תקטן ואילו שגיאת המבחן (test error) תגדל."

חוו דעתכם.

**בהצלחה!**