**בחינת גמר לדוגמה**

|  |  |  |
| --- | --- | --- |
| **מבנה הבחינה:** | לבחינה שני חלקים. יש לענות על כל השאלות. | |
|  | **חלק א:** | ענו על **ארבעת** 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 .

**שימו לב,**

**הבחינה לדוגמה שמופיעה להלן תוכל לשמש לכם כלי עזר נוסף ללימוד, ולעזרה בהכנה למבחן.**

**אך אין כאן שום התחייבות מצידנו שהבחינות בסמסטר זה תהיינה זהות במבנה, באופי וכו' לבחינה שהפצנו.**

**הבחינה כמו המטלות משמשת כלי ללימוד, ומבטיחה הכנה טובה למבחן**

## חלק א (90 נקודות)

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

**שאלה 1 (20 נקודות) – חוקי הקשר (association rules)**

נתון סל קניות הכולל 100 תנועות ו- 20 פריטים. התמיכה (support) לפריט a הוא 25%, התמיכה לפריט b הוא 90% והתמיכה לקבוצת הפריטים (item set) {a,b} היא 20% . בהנחה שהתמיכה היא 10% וסף הבטחון confidence thresholds)) הוא 60% :

1. חשבו את הבטחון (confidence) עבור חוק ההקשר {a}🡪{b}
2. בהמשך לסעיף א', האם החוק הוא בעל עניין (interesting) .

### שאלה 2 (25 נקודות) – ניתוח אשכולות (clustering)

### **בצעו אשכול לעשר הנקודות הבאות תוך שימוש באלגוריתם Agglomerative Hierarchical Clustering** .

### {(1,2), (4,8), (3,9), (7,3),(4,3),(2,4),(5,2),(3,5),(2,5), (6,6)}

בתשובתכם הניחו:

1. פונקצית המרחק בין פריטים – מנהטן
2. פונקצית מרחק בין clusters – minimum distance

שימו לב,

בתשובתכם הדגימו את כל השלבים והגדירו את האשכולות הסופיים .

### שאלה 3 (25 נקודות) – עצי החלטה (decision tree)

נתונה טבלת נתוני אימון:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Income** | **Car Type** | **Gender** | **Customer ID** |
| C0 | Low | Family | M | 1 |
| C0 | Medium | Sports | M | 2 |
| C0 | Medium | Sports | M | 3 |
| C0 | High | Sports | M | 4 |
| C0 | Very High | Sports | M | 5 |
| C0 | Very High | Sports | M | 6 |
| C0 | Low | Sports | F | 7 |
| C0 | Low | Sports | F | 8 |
| C0 | Medium | Sports | F | 9 |
| C0 | High | Luxury | F | 10 |
| C1 | High | Family | M | 11 |
| C1 | Very High | Family | M | 12 |
| C1 | Medium | Family | M | 13 |
| C1 | Very High | Luxury | M | 14 |
| C1 | Low | Luxury | F | 15 |
| C1 | Low | Luxury | F | 16 |
| C1 | Medium | Luxury | F | 17 |
| C1 | Medium | Luxury | F | 18 |
| C1 | Medium | Luxury | F | 19 |
| C1 | High | Luxury | F | 20 |

1. בנו עץ החלטה עבור נתוני האימון שבטבלה לחיזוי סוג הרכב.

בתשובתכם הדגימו את שלבי בחירת התכונה המפצלת בעץ.

**הערה:** בתשובתכם יש לכלול לפחות חישוב של מדד אחד כדוגמת אנטרופיה,

Gain ratio , מדד גיני.

1. איזה מבין התכונה/תכונות ניתן להסיר ומדוע? באיזה שלב ניתן להסיר את התכונות הללו?

### שאלה 4 (20 נקודות)- סיווג וחיזוי (classification and prediction)

נתון סט נתונים בו ה Concept הוא XOR של N משתנים בינאריים לא תלויים.

הסיכוי ל True הוא חצי בכולם.

1. חשבו את מספר הענפים הנדרשים לייצוג ה Concept בעץ החלטה?
2. בהמשך לסעיף א' , מהי סיבוכיות האלגוריתם ?

## חלק ב (10 נקודות)

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

**שאלה 5 (10 נקודות)**

במהלך קורס כריית מידע טען אחד הסטודנטים "בחברה בה אני עובד הנתונים מעודכנים במחסן נתונים. כמו כן, יש לנו חבילת תוכנה לכריית מידע. לדעתי משך זמן ביצוע כרית המידע יהיה כיום אחד בלבד, מאחר ויש להריץ את התוכנה עם הנתונים הקיימים במחסן הנתונים".

חוו דעתכם.

**בהצלחה!**