|  |  |  |
| --- | --- | --- |
| **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 המתייחסות לנתוני האימון הרצ"ב. ענו במפורט על **ארבע** השאלות.

נתונה טבלת נתוני אימון:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **מס'**  **ת.ז.** | **גיל** | **מגדר** | **תואר** | **שפת תכנות** | **ידע ב-WEKA** | **ניסיון**  **בעבודה** | **ניסיון בהוראה** | **העסקה כמנחה בקורס כריית מידע** |
| 12345 | 25 | נ | מדעי המחשב | Java | כן | לא | לא | לא |
| 67891 | 30 | נ | מדעי המחשב | C++ | לא | כן | כן | כן |
| 71294 | 28 | ז | ניהול מערכות ויישומים | Java | ?? | כן | לא | כן |
| 34568 | 35 | נ | מדעי המחשב | C++ | כן | כן | לא | לא |
| 24680 | 29 | ז | ניהול מערכות ויישומים | Java | לא | לא | כן | X |
| 12578 | 38 | נ | מדעי המחשב | C++ | כן | כן | כן | כן |
| 25804 | 33 | ז | ניהול מערכות ויישומים | Java | כן | ?? | כן | כן |
| 16923 | 40 | ז | מדעי המחשב | C++ | כן | כן | לא | לא |
| 34890 | 24 | נ | מדעי המחשב | Java | לא | לא | לא | לא |

### שאלה 1 (20 נקודות) – הכנת הנתונים (data preprocessing)

### ציינו והדגימו את שלבי הכנת הנתונים לביצוע כריית מידע.

### שאלה 2 (30 נקודות) – עצי החלטה (decision tree)

1. בנו עץ החלטה חלקי, הכולל את רמת השורש ורמה אחת נוספת בלבד, עבור נתוני האימון שבטבלה לחיזוי העסקת מועמד כמנחה בקורס כריית מידע. בתשובתכם הדגימו את שלבי בחירת התכונה המפצלת בעץ.

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

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

**שאלה 3 (25 נקודות) – אלגוריתם Naïve Bayes**

סווגו את המופע הבא <'27001 ', 31, נ, מדעי המחשב, C++, כן, לא, כן> תוך שימוש בנאייב בייס.

בתשובתכם הדגימו את ההסתברויות המותנות אשר ערכתם בהן שימוש.

### שאלה 4 (25 נקודות) – ניתוח אשכולות (clustering)

### בצעו אשכול היררכי אגלומרטיבי לנתוני האימון .

### בתשובתכם יש לכלול:

* אופן הגדרת המרחקים
* קריטריון לעצירת בניית הדנדרוגרם

**שימו לב,**

בתשובתכם הסופית יש להדגים את כל שלבי בניית הדנדרוגרם .

**המשך הבחינה בעמוד הבא**

## חלק ב (10 נקודות)

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

### שאלה 5 (10 נקודות)

במהלך קורס כריית מידע טען אחד הסטודנטים: "באלגוריתם אפריורי בכל מחזור יוצרים מועמדים וגוזמים מועמדים אשר לא מקיימים את תכונת האפריורי".

חוו דעתכם.

**בהצלחה!**