ממן 11

# שאלה 1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | ת.ז. שחקן | מגדר | גיל | ילדים | סוג רכב | סיכון לתאונה |
| 1 | 11111 | ז | 38 | כ | משפחתי | נמוך |
| 2 | 22222 | נ | 24 | ל | ספורט | גבוה |
| 3 | 33333 | ז | 29 | ל | ספורט | ?? |
| 4 | 44444 | נ | 36 | כ | משפחתי | נמוך |
| 5 | 55555 | ז | 38 | כ | ספורט | גבוה |
| 6 | 66666 | נ | 28 | כ | משפחתי | גבוה |
| 7 | 77777 | ז | 39 | כ | 99999 | גבוה |
| 8 | 88888 | ז | 32 | כ | משפחתי | נמוך |
| 9 | 99999 | ז | 48 | כ | ספורט | גבוה |
| 10 | 10111 | ז | ?? | ל | משפחתי | נמוך |
| 11 | 11222 | נ | 37 | כ | משפחתי | נמוך |
| 12 | 11AB3 | ז | 26 | כ | משפחתי | גבוה |
| 13 | 13444 | נ | 31 | 0 | ספורט | גבוה |
| 14 | 14555 | נ | 22 | ל | משפחתי | גבוה |

**שורה 3 עמודה "סיכון לתאונה"** - ערך חסר בעמודת הסיווג, ולכן נתעלם מכל הרשומה

**שורה 7 עמודה "סוג רכב"** – מידע רועש מכיוון שסוג רכב זו עמודה קטגוריאלית והערך הנתון הוא נומינלי. נבחר בשכיח של כל הרשומות בעלות אותו סיווג. השכיח הוא "ספורט" ולכן נתקן את ערך התא ל"ספורט"

**שורה 10 עמודה "גיל"** – ערך חסר בעמודה נומינלית. נחשב את ממוצע הגילאים של השורות בעלות סיווג "נמוך". כתוצאה מכך נקבל כי הערך החסר הוא 35.75

**שורה 12 עמודה "ת.ז. שחקן"** – מידע רועש. נתעלם מהעמודה מכיוון שהעמודה מתארת מידע ייחודי שאין לו השפעה על הסיווג

**שורה 13 עמודה "ילדים"** – מידע רועש בעמודה קטגוריאלית. השכיח עבור סיווג "גבוה" הוא "כ" ולכן נשנה את ערך התא ל-"כ"

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **ת.ז. שחקן** | **מגדר** | **גיל** | **ילדים** | **סוג רכב** | **סיכון לתאונה** |
| 1 | 11111 | ז | 38 | כ | משפחתי | נמוך |
| 2 | 22222 | נ | 24 | ל | ספורט | גבוה |
| 3 | 44444 | נ | 36 | כ | משפחתי | נמוך |
| 4 | 55555 | ז | 38 | כ | ספורט | גבוה |
| 5 | 66666 | נ | 28 | כ | משפחתי | גבוה |
| 6 | 77777 | ז | 39 | כ | ספורט | גבוה |
| 7 | 88888 | ז | 32 | כ | משפחתי | נמוך |
| 8 | 99999 | ז | 48 | כ | ספורט | גבוה |
| 9 | 10111 | ז | 35.75 | ל | משפחתי | נמוך |
| 10 | 11222 | נ | 37 | כ | משפחתי | נמוך |
| 11 | 11AB3 | ז | 26 | כ | משפחתי | גבוה |
| 12 | 13444 | נ | 31 | כ | ספורט | גבוה |
| 13 | 14555 | נ | 22 | ל | משפחתי | גבוה |

עמודת הגיל היא בעלת ערכים נומינליים וקיימים ערכים "חריגים" (לדוגמא: גיל 48). לכן במקרה זה נבצע תהליך דיסקרטיזציה על העמודה לפי שיטת Equal-depth.

נחשב את כמות הBin-ים על פי הנוסחה כך ש הוא מספר הערכים התקניים.   
לאחר ניקוי המידע הרועש נשארו 13 רשומות תקינות.   
לכן נקבע כי וכמות הBin-ים היא , נעגל כלפי מטה ונקבל כי כמות הBin-ים תהיה 3 וכמות הרשומות בכל Bin יהיה 4-5.  
לכן, הרשומות יתחלקו לBin-ים הבאים:  
1. Age <=28 (0-28)  
2. 28 < Age <= 36 (28-36)  
3. 36 < Age (36+)

נעדכן את הטבלה ונקבל

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **ת.ז. שחקן** | **מגדר** | **גיל** | **ילדים** | **סוג רכב** | **סיכון לתאונה** |
| 1 | 11111 | ז | 36+ | כ | משפחתי | נמוך |
| 2 | 22222 | נ | 0-28 | ל | ספורט | גבוה |
| 3 | 44444 | נ | 28-36 | כ | משפחתי | נמוך |
| 4 | 55555 | ז | 36+ | כ | ספורט | גבוה |
| 5 | 66666 | נ | 0-28 | כ | משפחתי | גבוה |
| 6 | 77777 | ז | 36+ | כ | ספורט | גבוה |
| 7 | 88888 | ז | 28-36 | כ | משפחתי | נמוך |
| 8 | 99999 | ז | 36+ | כ | ספורט | גבוה |
| 9 | 10111 | ז | 28-36 | ל | משפחתי | נמוך |
| 10 | 11222 | נ | 36+ | כ | משפחתי | נמוך |
| 11 | 11AB3 | ז | 0-28 | כ | משפחתי | גבוה |
| 12 | 13444 | נ | 28-36 | כ | ספורט | גבוה |
| 13 | 14555 | נ | 0-28 | ל | משפחתי | גבוה |

# שאלה 2

## סעיף א

נבנה עץ החלטה על פי מדד אנתרופיה (Information Gain).

נחשב את האנתרופיה עבור התכונות בטבלה

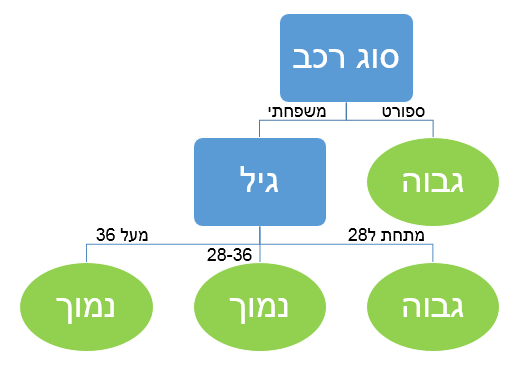
*קיבלנו כי התכונה בעלת האנתרופיה המינימלית היא "סוג רכב" ולכן נבחר בה לשורש עץ ההחלטה.*

*כעת, נסתכל על תת העץ שבו כל סוגי הרכב הם "משפחתי" ונחשב את האנתרופיה עבור התכונות שנותרו בו.*

*עבור התכונה "גיל" קיבלנו את האנתרופיה המינימלית ביותר האפשרית. לכן, נקבע כי הפיצול הבא בתת העץ "רכב משפחתי" יהיה על פי תכונת הגיל.*

*נסתכל על תת העץ שבו כל סוגי הרכב הם "ספורט". נשים לב כי הסיווג של כל הרשומות בתת העץ הוא "גבוה". לכן, לא נדרש לפצל את תת העץ.*

*לסיכום, העץ שהתקבל הוא*



## סעיף ב

ניתן להסיר את התכונות הבאות:

מס שחקן – תכונה בעלת נתון ייחודי עבור כל רשומה ולכן אין לה השפעה על סיווג הרשומה.

ילדים, מגדר – בפיתוח עץ ההחלטה הגענו לצמתים שבהם כל הרשומות שייכות לאותו הסיווג. זהו תנאי העצירה של האלגוריתם, ולכן אין שימוש בתכונה זו בעץ ההחלטה.

יפה