# מטלת מנחה (ממ"ן) 11

במטלה זו נתייחס לטבלת נתוני האימון הבאה –

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| מס רשומה | גיל נבדק | משקל | תזונה | ספורט | מצב סוציוקונומי |
| 1 | 58 | גבוה | רגילה | ל | נמוך |
| 2 | 60 | גבוה | רגילה | Y | גבוה |
| 3 | 32 | גבוה | צמחונית | ל | נמוך |
| 4 | 58 | תקין | רגילה | כ | גבוה |
| 5 | 663 | נמוך | צמחונית | כ | גבוה |
| 6 | 39 | נמוך | רגילה | ל | נמוך |
| 7 | 70 | נמוך | צמחונית | כ | גבוה |
| 8 | 27 | 72 | צמחונית | ל | נמוך |
| 9 | 45 | נמוך | רגילה | ל | גבוה |
| 10 | 64 | תקין | רגילה | כ | גבוה |
| 11 | 48 | תקין | רגילה | כ | גבוה |
| 12 | 62 | תקין | צמחונית | כ | גבוה |
| 13 | 50 | גבוה | רגילה | ל | גבוה |
| 14 | 24 | תקין | צמחונית | ???? | נמוך |

**שאלה 1 – הכנת הנתונים**

נבצע תהליך עיבוד מקדים (Preprocessing) של נתוני האימון הר"מ. שלב העיבוד המקדים של הכנת הנתונים כולל שלל אספקטים/צעדים, בין היתר – ניקוי (טיוב), אינטגרציה, רדוקציה וטרנספורמציה של הנתונים.

במקרה הנידון, נתוני האימון מרוכזים בסט נתונים יחיד אז אין צורך באינטגרציה. כמו כן, מספר הרשומות והיקף המאפיינים יחסית מצומצם, כך שנוכל בנתיים לוותר גם על רדוקציה של הנתונים. לפיכך נתמקד בעיקר בטיוב (ניקוי) הנתונים, כדוגמת טיפול בערכים חסרים ונתונים "רועשים" הכוללים שגיאות או אנומליות/סטיות חריגות, וטרנספורמציה ע"פ הצורך (למשל דיסקרטיזציה, אם כי ניתן גם לראות אותה גם בתור רדוקציה).

**טיוב הנתונים (Data cleaning)**

ניכר שאין רשומות כפולות/חוסר עקביות. לפיכך ראשית נטפל בערכים החסרים ונתונים חריגים ש-"חשודים" כשגויים, המסומנים להלן באדום –

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| מספר רשומה | גיל נבדק | משקל | תזונה | ספורט | מצב סוציוקונומי |
| 1 | 58 | גבוה | רגילה | ל | נמוך |
| 2 | 60 | גבוה | רגילה | Y | גבוה |
| 3 | 32 | גבוה | צמחונית | ל | נמוך |
| 4 | 58 | תקין | רגילה | כ | גבוה |
| 5 | 663 | נמוך | צמחונית | כ | גבוה |
| 6 | 39 | נמוך | רגילה | ל | נמוך |
| 7 | 70 | נמוך | צמחונית | כ | גבוה |
| 8 | 27 | 72 | צמחונית | ל | נמוך |
| 9 | 45 | נמוך | רגילה | ל | גבוה |
| 10 | 64 | תקין | רגילה | כ | גבוה |
| 11 | 48 | תקין | רגילה | כ | גבוה |
| 12 | 62 | תקין | צמחונית | כ | גבוה |
| 13 | 50 | גבוה | רגילה | ל | גבוה |
| 14 | 24 | תקין | צמחונית | ???? | נמוך |

כאסטרטגיית טיפול כללית בסט נתונים מצוצמם זה, נעדיף שלא להסיר רשומות פגומות/חסרות ולהשלים/להחליק אותן ע"פ ערך משוער, מאחר שלאור היקף הרשומות המצומצם כל רשומה הינה בעל השפעה ניכרת על הסיווג והסרתה יכולה להיות קריטית.

* רשומה מספר 2 - מאפיין ספורט: ניכר שתחום הערכים החוקי הוא בוליאני/בינרי - {'כ', 'ל'}, כנגד 'כן' ו-'לא' בהתאמה, לכן הערך 'Y' חשוד כערך לא חוקי (invalid). נקבע במקומו את הערך השכיח של מאפיין זה מתוך יתר הרשומות שבהן המצב הסוציוקונומי מסווג כגבוה: 6/8 מסווגות ב-'כ' בספורט, ורק 2/8 מסווגים כ-'לא'. ממצא זה גם מאשש את ההשערה לפיה יש לפרש את הערך 'Y' עם ''Yes . לפיכך, נתקן את ערך המאפיין ברשומה זו ל-'כ'.
* רשומה מספר 5 – מאפיין גיל: ניכר שהגיל נמדד בשנים, לכן הערך '663' אינו חוקי. עבור סט נתונים קטן זה, מאחר שקיים רק אחד שגוי, נוכל לטפל בו ידנית. ננקוט בהנחה הסבירה שמדובר בטעות אנוש בהקלדה, מה שגורר הערך המתאים הינו 66 או 63 (תלוי איזו ספרה נוספת נלחצה בטעות). נקבע ערך ממוצע (Mean) של רשומות עם ערכי גיל בטווח זה (63-66) ונקבל 64.33, נעגל ל-64, מאחר שתחום הערכים הוא ערכים שלמים.
* רשומה מספר 8 – משקל: ניכר שמאפיין המשקל הוא נומינלי/קטגוריאלי, ולא נומרי, לכן הערך 72, (שכנראה מייצג 72 ק"ג) אינו חוקי. לא נוכל למפות אותו באופן מיידי לאחת ממחלקות התחום (נמוך/תקין/גבוה) מאחר שסיווג כזה תלוי לא רק במשקל, אלא גם בגובה (למשל ע"פ מדד BMI) ונתון זה אינו בנמצא. לפיכך, נתייחס לערך זה כאל ערך חסר ונקבע עבורו את הערך השכיח ביותר (mode value) מבין יתר הרשומות - 'תקין' – 5/13 לעומת 4/13 של 'נמוך'/'גבוה' כ"א.
* רשומה מספר 14 – ספורט: הערך חסר, נשלים אותו עם הערך השכיח עבור הרשומות שמסווגות גם כן למצב סוציואקונומי נמוך, ונבחין כי כל הרשומות הללו מסווגות כ-'ל' (לא עושות ספורט), לכן נקבע גם עבור רשומה זו, שמסווגת למצב סוציואקונומי נמוך ערך זה.

**להלן בסיס הנתונים המתוקן לאחר ההשלמות/עדכונים הר"מ -**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| מספר רשומה | גיל נבדק | משקל | תזונה | ספורט | מצב סוציוקונומי |
| 1 | 58 | גבוה | רגילה | ל | נמוך |
| 2 | 60 | גבוה | רגילה | כ ~~Y~~ | גבוה |
| 3 | 32 | גבוה | צמחונית | ל | נמוך |
| 4 | 58 | תקין | רגילה | כ | גבוה |
| 5 | 64 ~~663~~ | נמוך | צמחונית | כ | גבוה |
| 6 | 39 | נמוך | רגילה | ל | נמוך |
| 7 | 70 | נמוך | צמחונית | כ | גבוה |
| 8 | 27 | תקין ~~72~~ | צמחונית | ל | נמוך |
| 9 | 45 | נמוך | רגילה | ל | גבוה |
| 10 | 64 | תקין | רגילה | כ | גבוה |
| 11 | 48 | תקין | רגילה | כ | גבוה |
| 12 | 62 | תקין | צמחונית | כ | גבוה |
| 13 | 50 | גבוה | רגילה | ל | גבוה |
| 14 | 24 | תקין | צמחונית | ל  ~~????~~ | נמוך |

כשלב אחרון בהכנת נתונים נבצע טרנספורמציית "החלקה" (smoothing) של מאפיין עמודת הגיל, באמצעות דיסקרטיזציה של הערכים השונים לטווחי ערכים. פעילות זו תביא לרדוקציה בהיקף טווח הערכים האפשרים ותייעל את שלב הסיווג בהמשך.

נחלק את טווח הגילאים בטכניקת 'equal-frequency(depth) binning', כלומר נחלק את הערכים השונים לקבוצות טווחים כך שכל קבוצה תכיל מספר רשומות זהה במידת האפשר, מה שיוביל לפיצול מאוזן יותר מבחינת שוויון הגובה/עומק בשלב הסיווג, בהתאם לשכיחות הופעת הערכים בנתוני האימון.

לאחר מיון הגילאים מתקלבל חלוקה שוות עומק (פחות או יותר) כדלהלן –

Bin 1: {24, 27, 32, 39}

Bin 2: {45, 48, 50, 58, 58}

Bin 3: {60, 62, 64, 64, 70}

לפיכך נבצע חלוקה לטווחים כדלהלן: (0,39], [45,59] ו-[60,∞).

נשייך כל גיל בטבלה המעודכנת מעלה לטווח אליו הוא משתייך. ונקבל לבסוף את בסיס הנתונים המטוייב הבא:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| מס רשומה | גיל נבדק | משקל | תזונה | ספורט | מצב סוציוקונומי |
| 1 | [45,59] | גבוה | רגילה | ל | נמוך |
| 2 | [60,∞) | גבוה | רגילה | כ | גבוה |
| 3 | (0,39] | גבוה | צמחונית | ל | נמוך |
| 4 | [45,59] | תקין | רגילה | כ | גבוה |
| 5 | [60,∞) | נמוך | צמחונית | כ | גבוה |
| 6 | (0,39] | נמוך | רגילה | ל | נמוך |
| 7 | [60,∞) | נמוך | צמחונית | כ | גבוה |
| 8 | (0,39] | תקין | צמחונית | ל | נמוך |
| 9 | [45,59] | נמוך | רגילה | ל | גבוה |
| 10 | [60,∞) | תקין | רגילה | כ | גבוה |
| 11 | [45,59] | תקין | רגילה | כ | גבוה |
| 12 | [60,∞) | תקין | צמחונית | כ | גבוה |
| 13 | [45,59] | גבוה | רגילה | ל | גבוה |
| 14 | (0,39] | תקין | צמחונית | ל | נמוך |

**שאלה 2 –עץ החלטה**

סעיף א – בניית עץ ההחלטה

נשתמש כעת בסט נתוני אימון המטוייב לבניית עץ החלטה שישמש כמודל לחיזוי מצב סוציואקונומי. נשתמש במדד Information Gain (המבוסס על אנטרופיה), בדומה לנעשה באלגוריתם הבסיסי ID3.

ראשית נחשב על סמך נתוני האימון המטוייבים, מהי רמת האינפורמציה הנדרשת לסיווג רשומה כלשהי כשייכת למצב סוציוקונומי גבוה/נמוך, ללא מידע מוקדם על ערכי מאפיינים אחרים ברשומה. נסמן את המאפיין 'מצב סוציואקונומי' ב-.

ההסתברות עבור רשומה כלשהי בנתוני האימון להיות מסווגת למצב סוציואקונומי גבוה היא כמספר היחסי של אותן רשומות מכלל סט נתוני האימון:

לפיכך הסתברות המאורע המשלים להיות מסווג למצב סוציוקונומי נמוך הינה  . מכאן, נציב בנוסחה לחישוב האנטרופיה של 'מצב סוציואקונומי' ע"ס נתוני האימון ונקבל –



כעת נחשב ונבדוק באיזו מאפיין כדאי להשתמש בתור שורש הפיצול של העץ. נעשה זאת ע"י חישוב האנטרופיה המותנית של המצב הסוציוקונומי בהינתן מידע על כל אחד ואחד מהמאפיינים האחרים הרלוונטיים (למשל מאפיין מספר הרשומה משמש כמזהה בלבד ואינו רלוונטי לסיווג), ולאחר מכן נבדוק ידיעת ערכו של איזה מאפיין תניב רווח אינפורמטיבי מרבי. נחשב איפה:

נתחיל במאפיין ' **גיל נבדק'**. נבדוק מהי ההתפלגות של סיווג המצב הסוציואקונומי (גבוה/נמוך) עבור כל אחד מערכי הטווח האפשריים של הגיל ונחשב את האנטרופיה המותנית בהתאם –



לפיכך המידע הנוסף שעוד נצטרך בכדי לחזות/לסווג את המצב הסוציו-אקונומי של רשומה בהינתן הגיל שלה, הינו , ומכאן ש**הרווח האינפורמטיבי** המושג ע"י פיצול לפי מאפיין **גיל** נתון ע"י –



נבצע חישוב דומה עבור כל יתר המאפיינים הרלוונטיים. עבור **משקל** נקבל -



ומכאן **הרווח האינפורמטיבי** עבור מאפיין **משקל** הינו-



עבור **תזונה** נקבל -



ומכאן **הרווח האינפורמטיבי** עבור מאפיין **תזונה** הינו-



עבור **ספורט** נקבל -



ומכאן **הרווח האינפורמטיבי** עבור מאפיין **ספורט** הינו-



כעת נשווה בין הרווח האינפורמטיבי המושג ע"י ידיעת כל אחת מהתכונות ונסיק כי הרווח האינפורמטיבי המקסימאלי מושג עבור מאפיין **גיל**.

לפיכך נפצל את שורש עץ ההחלטה לפי מאפיין זה, ונסדר מחדש את הרשומות בכל תת-קבוצה בהתאם למקטע גיל אליו הן שייכות. להלן העץ לאחר הפיצול (עמוד הבא) –



כפי שניתן לראות מהעץ, הקביעה של המצב-סוציואקונומי עבור רשומות בטווח גילאים עד 39 ו-+60 היא חח"ע לפי מודל זה, ולכן בענפים המתאימים יש עלה טרמינלי עם הסיווג המתאים.

עבור טווח הגילאים [45-59] , עדיין יש חוסר ודאות מסויים, בכדי לחדד את הדיוק נמשיך בפיצול ענף זה בעץ. מאפיין תזונה לא יתרום בפיצול שכן יש לו ערך קבוע,

אי לכך נשווה את מידת הפיצול בין 'משקל' לבין 'ספורט':

אנטורפיה מעודכנת עבור 'מצב סוציואקונומי' לאחר הפיצול -



נחשב את הרווח האינפורמטיבי שנשיג מפיצול לפי **משקל** –



ומכאן **הרווח האינפורמטיבי** עבור מאפיין **משקל** בהינתן טווח הגיליאים  הינו-



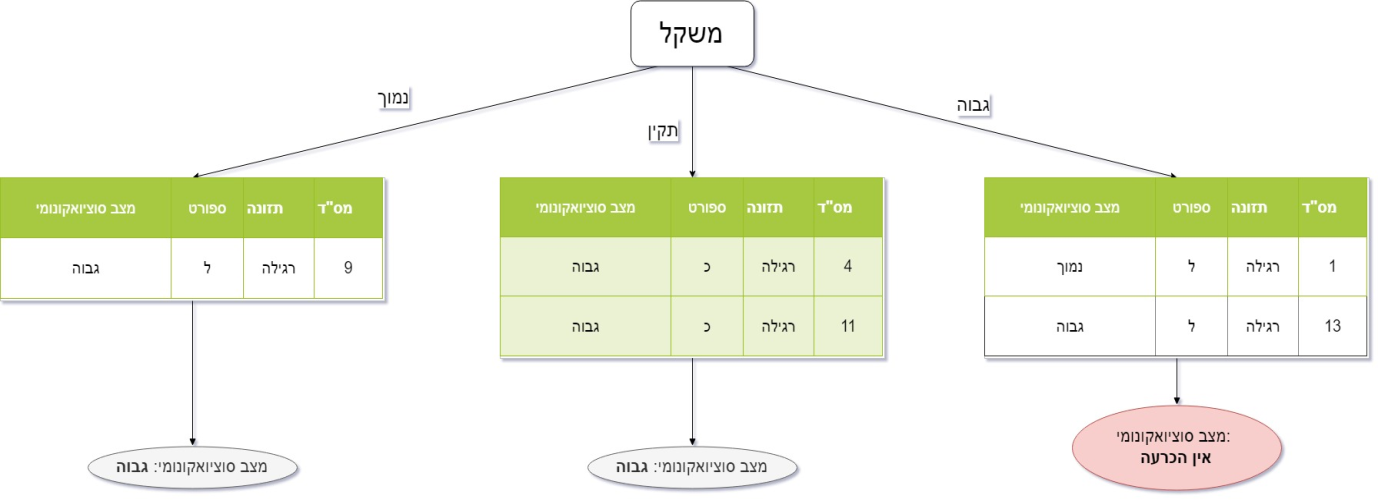
באופן דומה נחשב את הרווח האינפורמטיבי שנשיג מפיצול לפי מאפיין **ספורט** –



ומכאן **שהרווח האינפורמטיבי** עבור מאפיין **ספורט** בהינתן טווח הגיליאים  הינו-



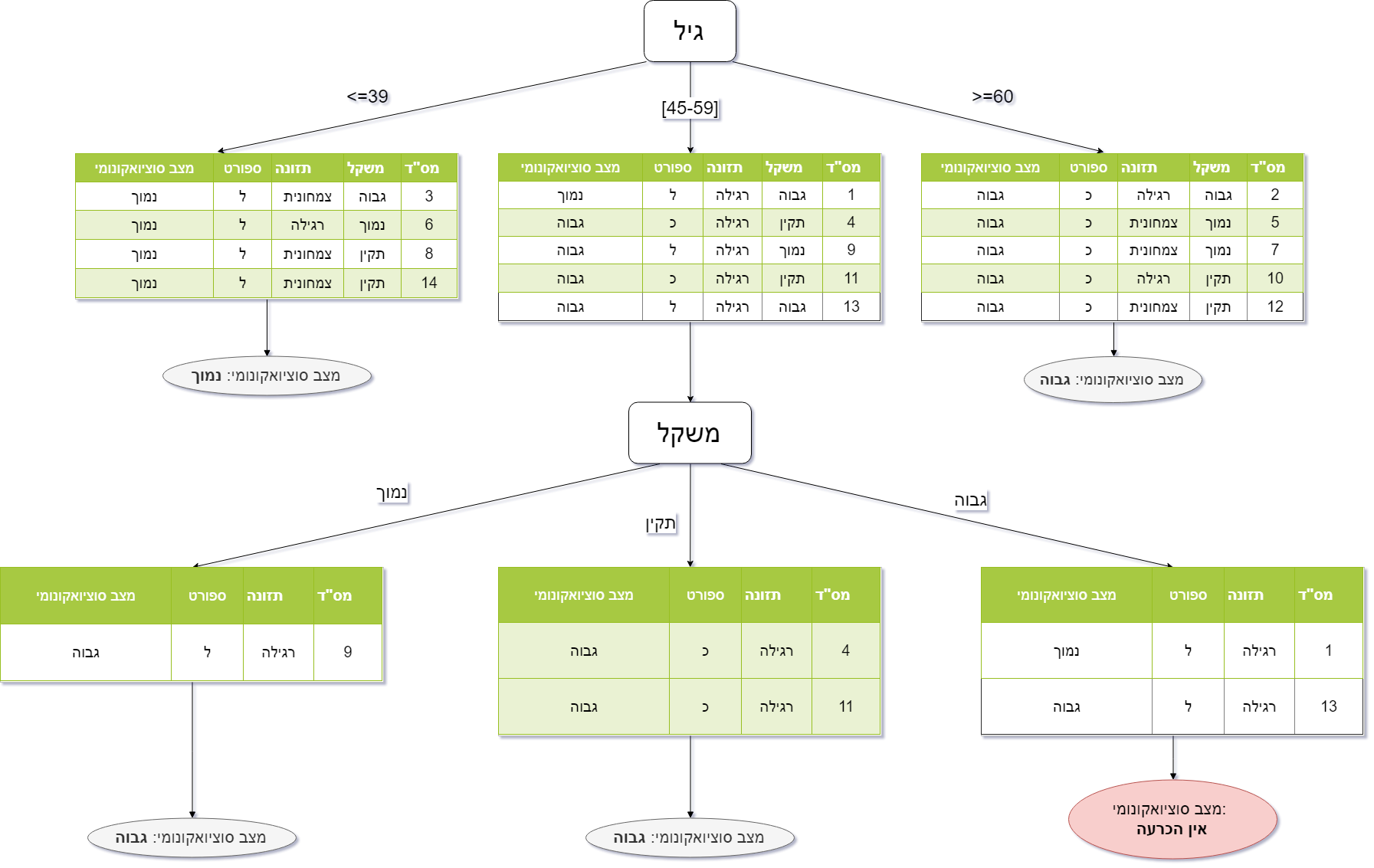
ניכר שנשיג רווח גבוה יותר מפיצול לפי משקל. נפצל ונקבל את תת-העץ, ממנו נוכל לקבוע את המצב הסוציו אקונומי עבור משקל נמוך ותקין, אך לא נוכל להכריע עבור תתי-הרשומות עם משקל גבוה, שכן יש בדיוק 2 רשומות שנבדלות בתכונה זו וזהות בכל יתר התכונות -



ע"כ זהו מבנה העץ הסופי המתקבל לפי מודל זה. בעמוד הבא נתון העץ המלא:



העץ המלא -



סעיף ב – תכונות יתירות ( characteristics)

ניתן להסיר את התכונות שלא היו שימושיות לסיווג/חיזוי, תכונות שלא השתמשנו בהן בעץ החלטה. השתמשנו רק מאפיינים של גיל ומשקל, לכן עבור מודל זה, כל יתר התכונות יתירות וניתן להסיר אותן, שכן הן אינן תורמות לתהליך:

* מספר סידורי של הרשומה;
* תזונה;
* ספורט;