עבודה 2- ניתוח מאגרי מידע

בעיה 3 (עץ החלטות):

א. על מנת לבנות את עץ ההחלטות נחשב Entropy עבור כל אופציה בקטגוריה ונחשב עבור כל קטגוריה את ה Information Gain. נבחר את השאלה לפי הקטגוריה עם הרווח מידע הגבוה ביותר:

תחילה נחשב Entropy עבור כל הדאטא סט, כאשר 5 משחקים לא משחקים ו-10 כן (15 סה"כ):

$$entropy = -p_+ * \log_2 p_+ - p_- * \log_2 p_- = -\frac{10}{15} * \log_2 \frac{10}{15} - \frac{5}{15} * \log_2 \frac{5}{15} = 0.918$$

על מנת להבין איזה שאלה כדאי לשאול ראשונה, נחשב לכל קטגוריה entropy לפי הנוסחא למעלה, והנוסחא בה נשתמש לחישוב ה Information Gain:

$$information\ gain = entropy(s) - \sum_v \frac{|s_v|}{|s|} * entropy(s_v)$$

עבור outlook:

outlook, sunny: $P_+ = 2$, $P_- = 3$ } Entropy = 0.971

outlook, rainy: $P_+ = 4$, $P_- = 2$ } Entropy = 0.918

outlook, overcast: $P_+ = 4$, $P_- = 0$ } Entropy = 0

$$\text{Information Gain (outlook)} = \text{Entropy}(s) - \sum \frac{|s_v|}{|s|} \cdot \text{Entropy}(s_v)$$

$$= 0.918 - \left[\frac{5}{15} \cdot 0.971 + \frac{6}{15} \cdot 0.918 + \frac{4}{15} \cdot 0\right] = 0.227$$

עבור temp:

temp, hot: $P_+ = 2$, $P_- = 2$ } Entropy = 1

temp, mild: $P_+ = 4$, $P_- = 2$ } Entropy = 0.918

temp, cool: $P_+ = 4$, $P_- = 1$ } Entropy = 0.722

$$\text{Info gain (temp)} = 0.918 - \left[\frac{4}{15} \cdot 1 + \frac{6}{15} \cdot 0.918 + \frac{5}{15} \cdot 0.722\right] = 0.043$$

עבור humidity:

humidity, high: $P_+ = 3$, $P_- = 4$ } Entropy = 0.985

humidity, normal: $P_+ = 7$, $P_- = 1$ } Entropy = 0.544

$$\text{info gain} = 0.918 - \left[\frac{7}{15} \cdot 0.985 + \frac{8}{15} \cdot 0.544\right] = 0.168$$

עבור Windy:

Windy, weak: $P_+ = 6$, $P_- = 1$ } Entropy = 0.592

Windy, strong: $P_+ = 4$, $P_- = 4$ } Entropy = 1

$$\text{info gain} = 0.918 - \left[\frac{7}{15} \cdot 0.592 + \frac{8}{15} \cdot 1\right] = 0.108$$

סיכום info gain עבור כלל הדאטא (לשאלה הראשונה):

Outlook = 0.227

Temp = 0.043

Humidity = 0.168

Windy = 0.108

כיוון ש- **outlook** בעל הרווח הכי גבוה הוא יהיה השאלה הראשונה.

עבור השאלה השניה:

<u>Outlook = overcast:</u>

כל המשחקים שוחקו ולכן entropy = 0, לכן לא נשאל שאלה נוספת וישר נשחק.

<u>Outlook = sunny:</u>

Entropy(s) = 0.971 (חושב מקודם)

קיימים 5 משחקים: 3-לא, 2-כן

עבור temp:

sunny -> temp, hot: $p+ = 0$, $p- = 2$ } Entropy = 0

sunny -> temp, mild: $p+ = 1$, $p- = 1$ } Entropy = 1

sunny -> temp, cool: $p+ = 1$, $p- = 0$ } Entropy = 0

$$\text{info gain (temp)} = 0.971 - \left[\frac{2}{5}\cdot 0 + \frac{2}{5}\cdot 1 + \frac{1}{5}\cdot 0\right] = 0.571$$

עבור humidity:

sunny -> humidity, high: $p+ = 0$, $p- = 3$ } Entropy = 0

sunny -> humidity, normal: $p+ = 2$, $p- = 0$ } Entropy = 0

$$\text{info gain (humidity)} = 0.971 - 0 = 0.971$$

עבור Windy:

sunny -> Windy, weak: $p+ = 1$, $p- = 1$ } Entropy = 1

sunny -> Windy, strong: $p+ = 1$, $p- = 2$ } Entropy = 0.918

$$\text{info gain (windy)} = 0.971 - \left[\frac{2}{5}\cdot 1 + \frac{3}{5}\cdot 0.918\right] = 0.020$$

סיכום info gain עבור sunny:

Temp = 0.571

Humidity = 0.971

Windy = 0.02

הרווח הכי גבוה הוא מ- **humidity** ולכן נבחר בו לשאלה השניה.

:Outlook = rainy

Entropy = 0.918 (חושב קודם)

קיימים 6 משחקים: 4-כן, 2-לא

עבור temp:

rainy -> temp, hot: $P+ = 0$, $P- = 0$ } Entropy = 0

rainy -> temp, mild: $P+ = 2$, $P- = 1$ } Entropy = 0.918

rainy -> temp, cool: $P+ = 2$, $P- = 1$ } Entropy = 0.918

$$\text{info gain (temp)} = 0.918 - \left[\frac{3}{6}\cdot 0.918 + \frac{3}{6}\cdot 0.918\right] = 0$$

עבור humidity:

rainy -> humidity, high: $P+ = 1$, $P- = 1$ } Entropy = 1

rainy -> humidity, normal: $P+ = 3$, $P- = 1$ } Entropy = 0.811

$$\text{info gain (humidity)} = 0.918 - \left[\frac{2}{6}\cdot 1 + \frac{4}{6}\cdot 0.811\right] = 0.044$$

עבור Windy:

rainy -> Windy, weak: $P+ = 3$, $P- = 0$ } Entropy = 0

rainy -> Windy, strong: $P+ = 1$, $P- = 2$ } Entropy = 0.918

$$\text{info gain (windy)} = 0.918 - \left[\frac{3}{6}\cdot 0 + \frac{3}{6}\cdot 0.918\right] = 0.459$$

סיכום info gain עבור sunny:

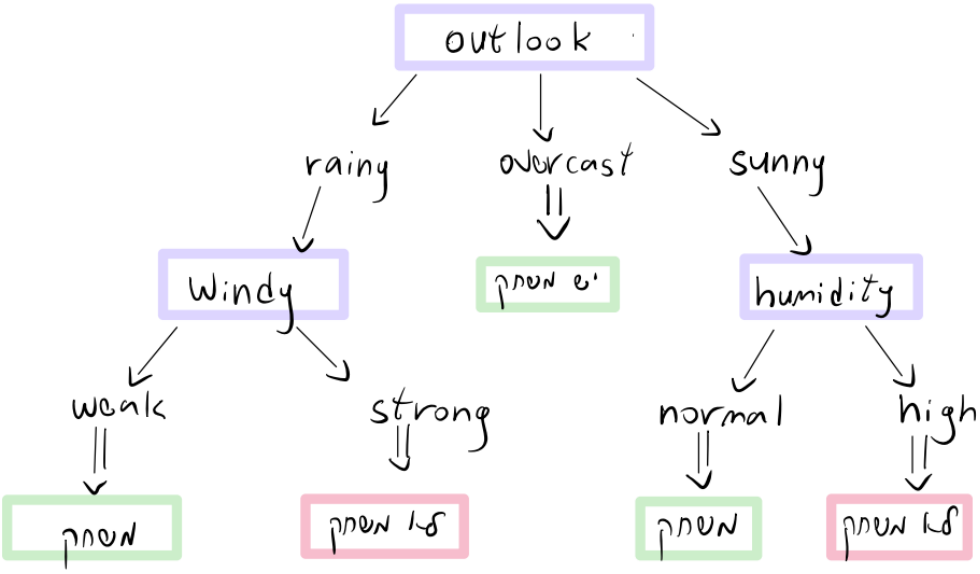
Temp = 0

Humidity = 0.044

Windy = 0.459

הרווח הכי גבוה הוא מ- **windy** ולכן נבחר בו לשאלה השניה.

מקרא: שאלות, תשובה חיובית למשחק, תשובה שלילית למשחק



ב. בעץ החלטות זה השגיאה עומדת על 1/15 טעויות כלומר: 6.667% סיכוי לטעות.

ג. לפי עץ ההחלטות שלי, הימים בהם הייתי משחקת הם 1,2 ו-3 (משחקת ב- D16, D17, D18 ולא הייתי משחקת בD19) כלומר משחקת בימים בהם מעונן או גשום ורוח חלשה.