שאלה 1.

נעתיק את הדאטה ונסמן את הערכים הבעייתיים.

COPD	Inhaler use	Phlegm	Wheezing	Dyspnea on exertion	Marital status	Smoke	Age range	Subject num
No	Yes	No	Yes	Yes	Marriage	No	51-60	**1**
Yes	Yes	Yes	Yes	Yes	Marriage	??	70+	**2**
No	No	Yes	Yes	No	Single	No	31-40	**3**
Yes	Yes	Yes	Yes	Yes	XXX	Yes	70+	**4**
No	No	Yes	No	No	Single	Yes	51-60	**5**
??	No	Yes	Yes	No	Marriage	Yes	51-60	**6**
Yes	No	No	Yes	Yes	Marriage	Yes	61-70	**7**
No	No	No	No	Yes	Marriage	Yes	51-60	**8**
No	??	No	Yes	No	Marriage	No	61-70	**9**
No	No	Yes	No	No	Marriage	No	41-50	**10**
Yes	Yes	No	Yes	Yes	Marriage	Yes	61-70	**11**
No	No	Yes	??	No	Marriage	Yes	30-	**12**

יש חמישה נבדקים עם נתונים בעייתיים.

- נבדק מס' 2: חסרה התכונה 'מעשן' – כיוון שבטווח הגילאים של מעל גיל 70 אין נבדק עם חרחורים, ליחה, קוצר נשימה במאמץ, ומחלת ריאות חסימתית כרונית ביחד, לכן נשלים את הערך ל'מעשן'.
- נבדק מס' 4: חסרה התכונה 'סטאטוס משפחתי' – כיוון שמהנתונים אין אפשרות לשער האם הנבדק אכן היה נשוי או לא (אין אפילו קורלציה בין מאפיין זה למאפיינים אחרים), ובנוסף לכך שנתון זה לכאורה לא אמור להשפיע על התכונה שאנו רוצים לחזות (COPD), לכן נשלים את הערך החסר ל-null.
- נבדק מס' 6: חסרה התכונה 'COPD' (מחלת ריאות חסימתית כרונית) – כיוון שזו התכונה שאנו רוצים לחזות נמחק את הרשומה.
- נבדק מס' 9: חסרה התכונה 'שימוש במשאף' – יש הגיון בלהשלים את הערך ל-'לא משתמש במשאף', מחמת שלנבדק אין קוצר נשימה במאמץ, ליחה ומחלת ריאות, והנבדק איננו מעשן. אך כיוון שטווח הגילאים של נבדק זה גבוה (כלומר מבוגר ביחס לנבדקים האחרים), ובנוסף הוא סובל מחרחורים, השלמה של הערך ל-'לא משתמש במשאף' עלולה להוות הטיה לא תקינה של הנתונים. לכן נשלים את הערך החסר ל-null.
- נבדק מס' 12: חסרה התכונה 'חרחורים' – מאותם שיקולים שקיימים אצל נבדק מס' 9 (אין מובהקות להשלים לצד זה או אחר), בנוסף לכך שאין לנו נתונים על נבדקים בטווח גילאים של הנבדק (צעיר מ-30), לכן נשלים את הערך החסר ל-null.

טבלת הנתונים לאחר התיקון:

COPD	Inhaler use	Phlegm	Wheezing	Dyspnea on exertion	Marital status	Smoke	Age range	Subject num
No	Yes	No	Yes	Yes	Marriage	No	51-60	**1**
Yes	Yes	Yes	Yes	Yes	Marriage	**Yes**	70+	**2**
No	No	Yes	Yes	No	Single	No	31-40	**3**
Yes	Yes	Yes	Yes	Yes	**null**	Yes	70+	**4**
No	No	Yes	No	No	Single	Yes	51-60	**5**
~~??~~	~~No~~	~~Yes~~	~~Yes~~	~~No~~	~~Marriage~~	~~Yes~~	~~51-60~~	~~6~~
Yes	No	No	Yes	Yes	Marriage	Yes	61-70	**7**
No	No	No	No	Yes	Marriage	Yes	51-60	**8**
No	**null**	No	Yes	No	Marriage	No	61-70	**9**
No	No	Yes	No	No	Marriage	No	41-50	**10**
Yes	Yes	No	Yes	Yes	Marriage	Yes	61-70	**11**
No	No	Yes	**null**	No	Marriage	Yes	30-	**12**

<u>נבצע רדוקציה:</u> מחמת שאנו מעוניינים לחקור את מאפיין ה-COPD , מעוניינים ונראה נשים לב שמאפיין המצב המשפחתי (לכאורה) לא רלוונטי, לכן נמחק את העמודה מהנתונים.

אין צורך לבצע דיסקרטיזציה כיוון שהנתונים כבר נתונים עם ערכים דיסקרטיים.

טבלת הנתונים המטויבת:

COPD	Inhaler use	Phlegm	Wheezing	Dyspnea on exertion	Smoke	Age range	Subject num
No	Yes	No	Yes	Yes	No	51-60	**1**
Yes	Yes	Yes	Yes	Yes	Yes	70+	**2**
No	No	Yes	Yes	No	No	31-40	**3**
Yes	Yes	Yes	Yes	Yes	Yes	70+	**4**
No	No	Yes	No	No	Yes	51-60	**5**
Yes	No	No	Yes	Yes	Yes	61-70	**7**
No	No	No	No	Yes	Yes	51-60	**8**
No	null	No	Yes	No	No	61-70	**9**
No	No	Yes	No	No	No	41-50	**10**
Yes	Yes	No	Yes	Yes	Yes	61-70	**11**
No	No	Yes	null	No	Yes	30-	**12**

<u>שאלה 2.</u>

נשתמש כעת בסט נתוני אימון המטוייב לבניית עץ החלטה שישמש כמודל לחיזוי COPD – מחלת ריאות חסימתית כרונית. נשתמש במדד Information Gain (מבוסס אנטרופיה).

$$Info(inhaler) = \frac{4}{10}I(3,1) + \frac{6}{10}I(5,1) = 0.71$$
$$Info(phlegm) = \frac{5}{11}I(2,3) + \frac{6}{11}I(4,2) = 0.94$$
$$Info(wheezing) = \frac{7}{10}I(4,3) + \frac{3}{10}I(3,0) = 0.69$$
$$Info(dysphea) = \frac{6}{11}I(2,4) + \frac{5}{11}I(5,0) = 0.5$$
$$Info(smoke) = \frac{7}{11}I(3,4) + \frac{4}{11}I(4,0) = 0.63$$
$$Info(age) = 3\left(\frac{1}{11}I(1,0)\right) + \frac{3}{11}I(3,0) + \frac{3}{11}I(2,1) + \frac{2}{11}I(2,0) = 0.25$$

נקבל שהרווח האינפורמטיבי (IG) הגבוה ביותר מתקבל מהמאפיין 'גיל', ולכן נבחר בו כתכונה מפצלת ראשונה בעץ החלטות.

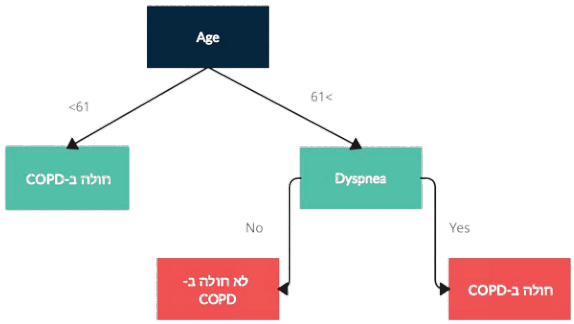
נחלק את העץ ל-2 מסלולים; נבדקים מתחת לגיל 60 ונבדקים מעל גיל 60.

א. מתחת לגיל 60: נקבל עלה של 'לא חלה ב-COPD'.

ב. מעל גיל 60: במסלול זה נקבל את הטבלה הבאה. נשים לב שהרווח האינפורמטיבי לפיצול ע"פ קוצר נשימה הוא הגבוה ביותר, מחמת שבמקרה זה נקבל חלוקה מלאה (הנבדק סובל מקוצר נשימה במאמץ אם ורק אם הוא מעשן). לכן נקבע שזה יהיה מאפיין הפיצול השני.

COPD	Inhaler use	Phlegm	Wheezing	Dyspnea on exertion	Smoke	Age range	Subject num
Yes	Yes	Yes	Yes	Yes	Yes	70+	**2**
Yes	Yes	Yes	Yes	Yes	Yes	70+	**4**
Yes	No	No	Yes	Yes	Yes	61-70	**7**
No	null	No	Yes	No	No	61-70	**9**
Yes	Yes	No	Yes	Yes	Yes	61-70	**11**

לסיכום נקבל את העץ:



<u>סעיף ב':</u>

ניתן להוריד את התכונות: חרחורים, ליחה, שימוש במשאף, ומצב משפחתי. 'מצב משפחתי' הוסבר לעיל, וכל שאר התכונות ראינו שאין שימוש בהן בעץ ההחלטה מחמת שהגענו לתנאי העצירה של האלגוריתם.