# מבוא למערכות לומדות

# תרגיל 5

רן שחם - 203781000 - ransha

20 ביוני 2017

## 1 תת־אופטימליות של ID3

נניח ש־$\mathcal{X} = \{0,1\}^3, \mathcal{Y} = \{0,1\}$ ומדגם האימון נתון על ידי:

$$\begin{gathered}\left(\left(1,1,1\right),1\right)\\ \left(\left(1,0,0\right),1\right)\\ \left(\left(1,1,0\right),0\right)\\ \left(\left(0,0,1\right),0\right)\end{gathered}$$

### 1.1

נחשב את ה־information gain לכל אחד מהפיצ'רים. במקרה הנ"ל, $\mathcal{Y} = \{0,1\}$ ולכן אפשר לרשום את $H$ (מהגדרת information gain) באופן הבא:

$$H\left(S\right) = -p_0 \log p_0 - p_1 \log p_1 = -p \log p - \left(1-p\right) \log\left(1-p\right)$$

כאשר $p$ הוא אחוז הדוגמאות המתויגות 1 ב־$S$. מתקיים:

1.

$$\begin{aligned}Gain\left(S;x_1\right) &= H\left(S\right) - \left(\frac{|S_1|}{|S|}H\left(S_1\right) + \frac{|S_0|}{|S|}H\left(S_0\right)\right)\\ &= \underbrace{H\left(S\right)}_{p=\frac{1}{2}} - \left(\frac{3}{4}\underbrace{H\left(S_1\right)}_{p=\frac{2}{3}} + \frac{1}{4}\underbrace{H\left(S_0\right)}_{p=0}\right) \approx 0.22\end{aligned}$$

כאשר $S_1$ היא קבוצת הדוגמאות עבורה $x_1 = 1$, ויש 3 כאלה; ובאופן דומה $|S_0| = 1$. החישוב הנ"ל מתבסס על:

$$\begin{aligned}H\left(S\right) &= -p\log p - \left(1-p\right)\log\left(1-p\right) = -\log\frac{1}{2} \approx 0.69\\ H\left(S_1\right) &= -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} \approx 0.63\\ H\left(S_0\right) &= -0\log 0 - 1\log 1 = 0\end{aligned}$$

(כי $\lim_{x\to 0} x\log x = 0$)

2. באופן דומה (אם כי בכתיב קצר יותר):

$$Gain\left(S;x_2\right)=H\left(\frac{1}{2}\right)-\left(\frac{2}{4}H\left(\frac{1}{2}\right)+\frac{2}{4}H\left(\frac{1}{2}\right)\right)=H\left(\frac{1}{2}\right)-H\left(\frac{1}{2}\right)=0$$

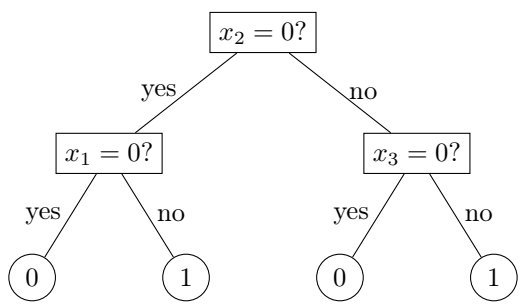
3. זהה לחישוב ב־2, כלומר $Gain\left(S;x_3\right)=0$

ולכן ID3 יקח את $x_1=0?$ להיות השורש של עץ ההכרעה. אם כך, 3 הדוגמאות הראשונות יגיעו לתת עץ אחד. נזכור כי גובה העץ חסום ב־2, כלומר נותרה לנו שאילתה אחת על אחת מהקורדינטות $x_2,x_3$ ואחריה עלינו להגיע לעלה - כלומר לקבל החלטה (0 או 1). מכיוון שבאחד מתתי העץ של הקדקוד $x_1=0?$ יש 3 דוגמאות, לא משנה איזה שאלה נבחר יהיו לפחות 2 דוגמאות שיקבלו את אותה פרדיקציה, ובמקרה שלנו אחת מהן תהיה לא נכונה. אם נשאל $x_2=0?$ נקבל שהדוגמה $(1,1,1)$ והדוגמה $(1,1,0)$ יתויגו באותו תיוג, ומכיוון שתיוגם האמיתי שונה, האלגוריתם יטעה על אחת מהן. אם נשאל $x_3=0?$ נקבל ש־ $(1,1,0)$ ו־$(1,0,0)$ יתויגו באותו תיוג, וגם כאן האלגוריתם יטעה על אחת.

אם כך, על מדגם מגודל 4 אנחנו טועים לפחות בתיוג אחד, כלומר טעות האימון היא לכל הפחות $1/4$.

### 1.2

נתבונן בעץ ההכרעה הבא:




ברור שהוא מעומק 2 וכן שהוא משיג שגיאת אימון 0 על המדגם, ונסיק שאלגוריתם ID3 אינו אופטימלי (במובן $ERM$).

## 2 שכנים קרובים $k$

### 2.1

ערך $k$ הממזער את שגיאת האימון הוא 1. ברור ששגיאת האימון במקרה זה היא 0, כי כל נקודה היא השכן הקרוב ביותר שלה (מרחק 0 בכל נורמה, ובפרט באוקלידית). שגיאת האימון היא אי־שלילית, ולכן זו שגיאת אימון מינימלית על המדגם הנתון.

### 2.2

**הערה 2.1** ננתח את השגיאה עבור ערכי $k$ אי זוגיים כדי להימנע ממצב של "תיקו" (ובהתאם לכתוב בפורום)

- נשים לב ש־$k=1$ הוא בחירה גרועה עבור השגיאה בleave-one-out cross-validation. נתבונן במקבץ העליון (השיקולים למקבץ התחתון סימטריים): הדוגמה $(2,6)$ תתוייג $+$[1] באופן שגוי כי השכן הקרוב ביותר שלה הוא $((2,7),+)$. באופן דומה, $(2,7)$ תתוייג $-$ כי $((2,6),-),((3,7),-)$ הם שכניה הקרובים ביותר (ובוחרים אחד מהם). בצורה זו, כל הדוגמאות במקבץ העליון למעט השתיים הקיצוניות ($(1,5),(5,9)$) יתויגו לא נכונה, כלומר האלגוריתם ישגה על $\frac{10}{14}\approx 71.4\%$ מהמדגם.

[1] כאשר משמיטים אותה, מאמנים את האלגוריתם בלעדיה ובודקים את התיוג שלה

- $k=3$ - במקרה זה הדוגמאות שיתויגו לא נכונה הן:[2] $(2,7),(3,7),(3,8)$. בנק' אלה 2 מתוך 3 השכנים הקרובים מתוייגים בתיוג ההפוך לזה הנכון, ולכן האלגוריתם שוגה עליהן. אם כך, השגיאה עבור $k$ זה היא: $\frac{6}{14} \approx 42\%$
- $k=5$ - קל לראות שהדוגמאות היחידות שמתויגות לא נכונה הן בתת המקבץ הקטן יותר: $(2,7),(3,8)$, כי רב שכניהן הם $-$, וכן שהיות והן 2 נק' במקבץ זה, הן לא משפיעות על התיוג של ה$-$ים בסביבתן (כי לוקחים 5 שכנים, אז לפחות 3 במקבץ יהיו בתיוג הנכון). השגיאה במקרה זה היא רק על נק' אלה בכל אחד מהמקבצים, כלומר $\frac{4}{14} \approx 28.5\%$
- בכל מקבץ גדול יש 7 נק', ולכן ערכי $k$ גדולים יותר לא יתנו שגיאה נמוכה יותר מאשר $k=5$

**לכן** $k=5$ הוא הערך הממזער את השגיאה הנ"ל, והשגיאה שלו היא $4/14$.[3]

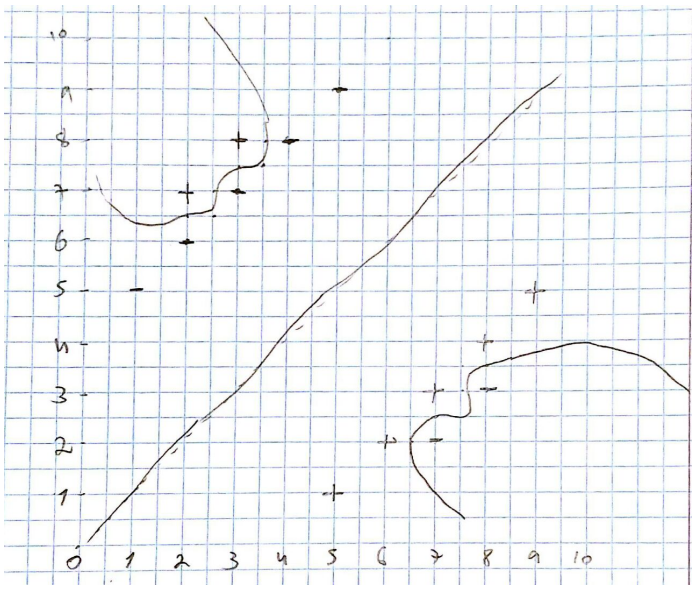
### 2.3

מקבצי הנקודות בעלות תיוג זהה מכילים מס' קטן של נקודות. כך למשל, עבור המקבץ העליון ביותר (המכיל שני $+$ים) כל ערך $k$ הגדול מ-4 יגרום ל"ניחוש" הדוגמאות באזור זה, כי תמיד יילקחו בחשבון גם 2 $-$ים. כלומר נאבד לחלוטין את המידע לגבי הימצאות מקבץ $+$ים למעלה (ומינוסים למטה).

ערכי $k$ נמוכים

### 2.4

הנה הניסיון הטוב ביותר שלי (הוא לא כ"כ טוב):




[2] שוב, כאשר מסתכלים על המקבץ העליון בלבד ומסיקים מסימטריה על המקבץ התחתון

[3] כאשר מחשבים שגיאה לפי $\frac{1}{m}\sum_{i=1}^{m}\ell^{0-1}\left(h_i,(\mathbf{x}_i,y_i)\right)$ ו-$h_i$ היא אימון האלגוריתם בלי הדוגמה ה-$i$ ($m=14$)

# 3 חלק תכנותי

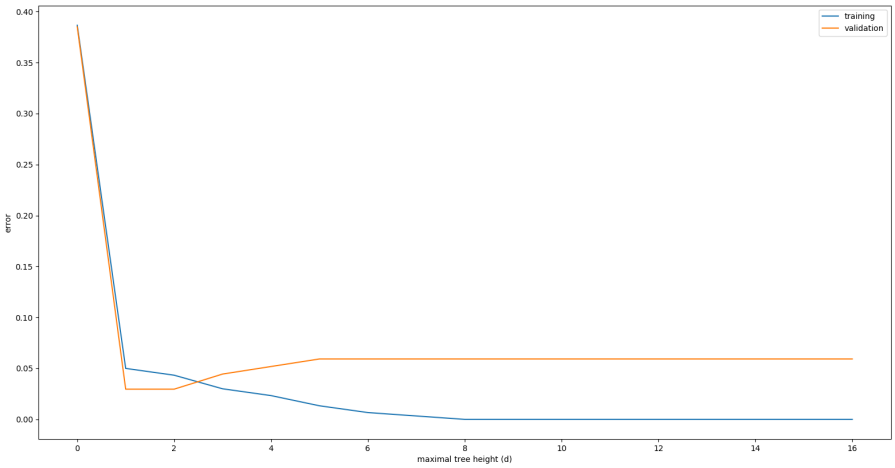
**הערה 3.1** נא להריץ את הפקודה `pip install -r requirements.txt` על מנת לוודא שכל החבילות מותקנות ובגרסה המתאימה. זה נבדק באקווריום, אבל אני לא בטוח שלא התקנתי ב־user שלי חבילות שאינן כלולות כברירת מחדל (pandas, anytree, etc.).

### 3.1

**הערה 3.2** בחרתי לממש את ה־Information gain

### 3.2

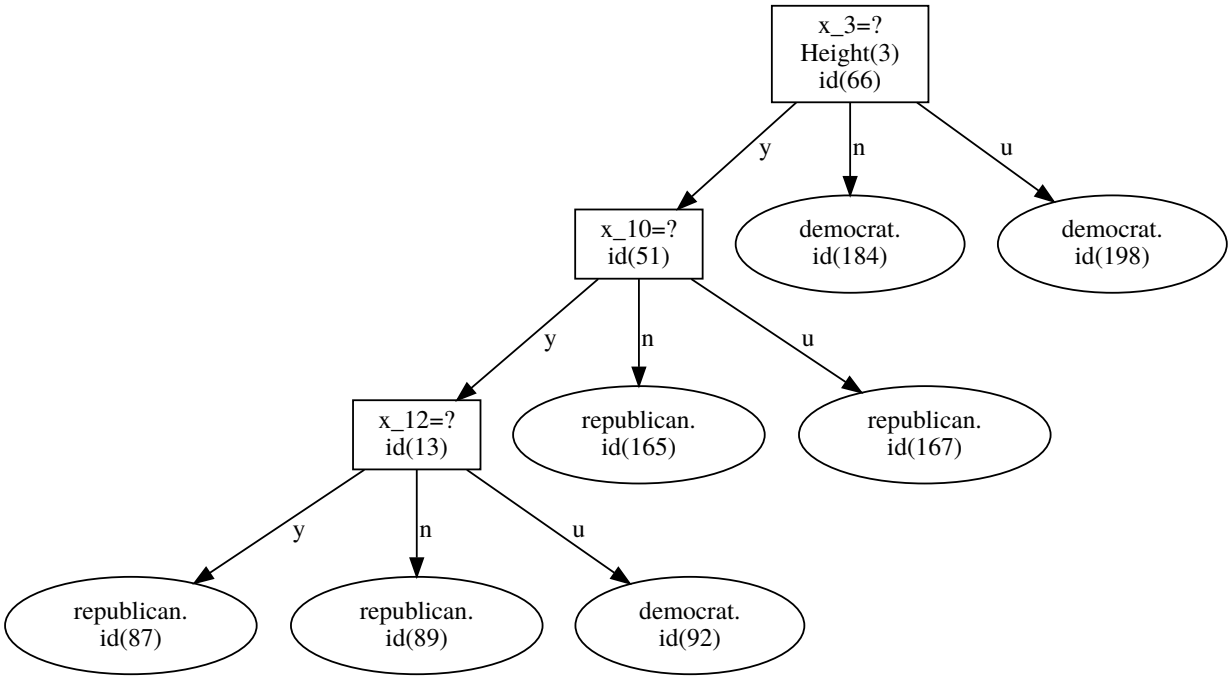
העצים שהתקבלו מופיעים בנספח (A). גרף השגיאות מתואר להלן:




איור 1: שגיאת האימון מול שגיאת הולידציה - שגיאות האימון קטנה מהר ל־0, כלומר האלגוריתם מתייג נכונה את כל דוגמאות האימון, אך מבצע overfitting לדוגמאות הולידציה עם עצים מגובה $> 2$. מהגרף הזה אפשר להסיק שהגובה שימזער את שגיאת ההכללה הוא 1 או 2, שכן במקרים אלה שגיאת הולידציה היא הנמוכה ביותר.

### 3.3

העץ שהתקבל על ידי גיזום העץ המקסימלי מהסעיף הקודם הוא הבא:



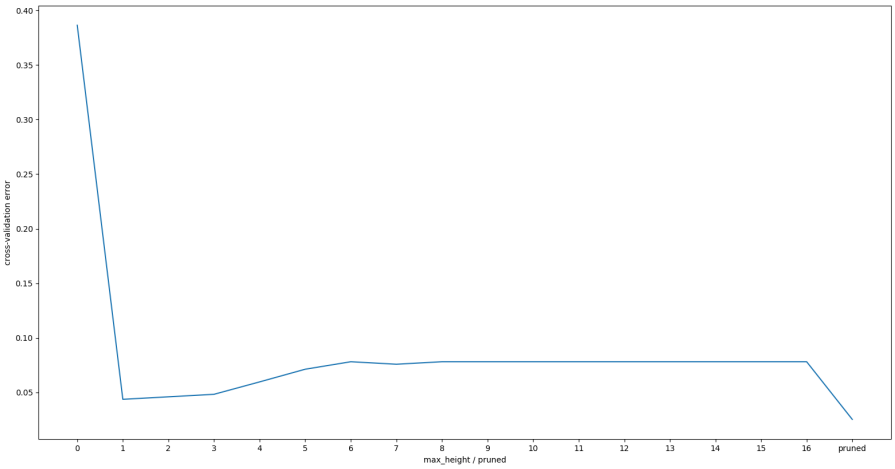

```
question 3
----------
generalization error of the un-pruned tree:
0.0592592592593
generalization error of the pruned tree:
0.02777777777778
```

איור 2: העץ שהתקבל מהרצת האלגוריתם הגנרי לגיזום (pruning), וערכי שגיאת ההכללה (כפי שמתואר בהוראות התרגיל - ההפרש בין שגיאת הולידציה לאימון)

אפשר לראות שהוא קטן משמעותית מהעץ הממזער את טעות האימון.

### 3.4

להלן גרף השגיאות הממוצעות בתהליך הקרוס־ולידציה:

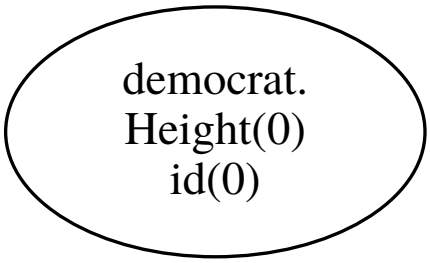



איור 3: הגרף דומה לגרף שגיאת הולידציה מסעיף 1, אם כי נראה שגם עצים מגובה 3 מקבלים במקרה זה שגיאה נמוכה יחסית.

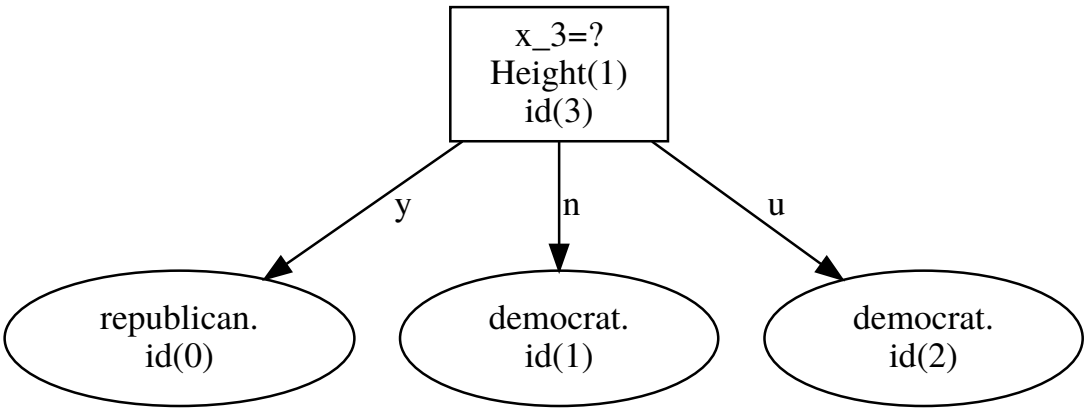
נראה שארכיטקטורת העץ הגזום הוא הבחירה הטובה ביותר, על סמך תהליך ה־cross-validation. עם זאת, בהיעדר דוגמאות "מבחן", קשה לתת הערכה טובה לשגיאת העץ הטוב ביותר (שגיאת estimation). זאת מכיוון שבתהליך זה כל דוגמה מהמדגם הנתון שימשה לצורך אימון האלגוריתם ובחינתו, ולכן בחינה נוספת של ביצועי האלגוריתם על אותו המדגם תסבול במידה כזו או אחרת (שאני לא בטוח שלמדנו לשלוט בה) מ־overfitting.

כמוכן, גם בתהליך זה קיבלנו שגובה העץ המומלץ הוא נמוך (1 או 2, על אף שבמקרה זה אפשר לקחת גם את 3 בחשבון), בהתאם לאינטואיציה מאחורי MDL לפיה היפותזה פשוטה (בעלת תיאור קצר; במקרה שלנו, עץ נמוך) היא טובה יותר.

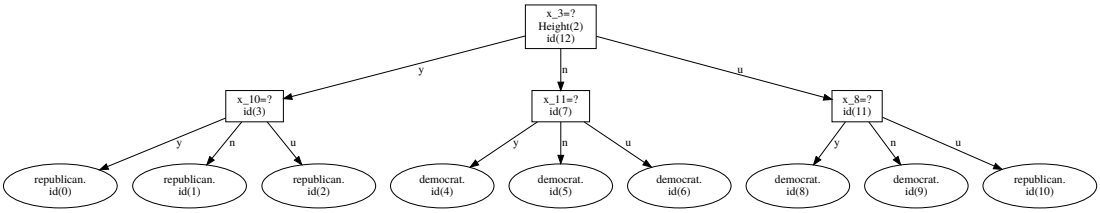
# א' עצים



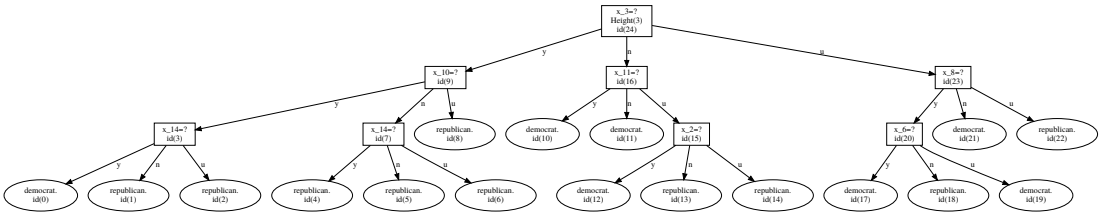

איור 4: העץ שהתקבל מהרצת ID3 עם גובה מקסימלי 0



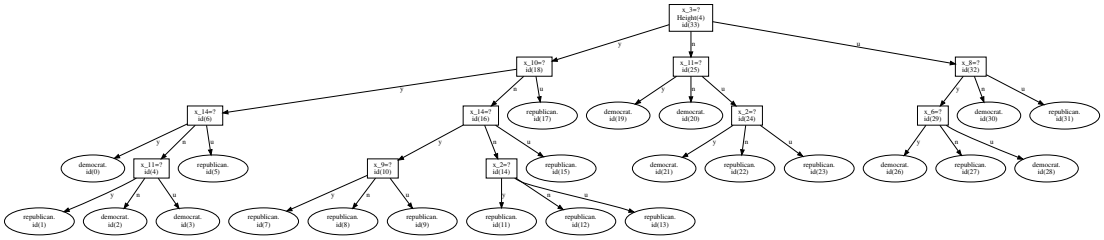

איור 5: העץ שהתקבל מהרצת ID3 עם גובה מקסימלי 1



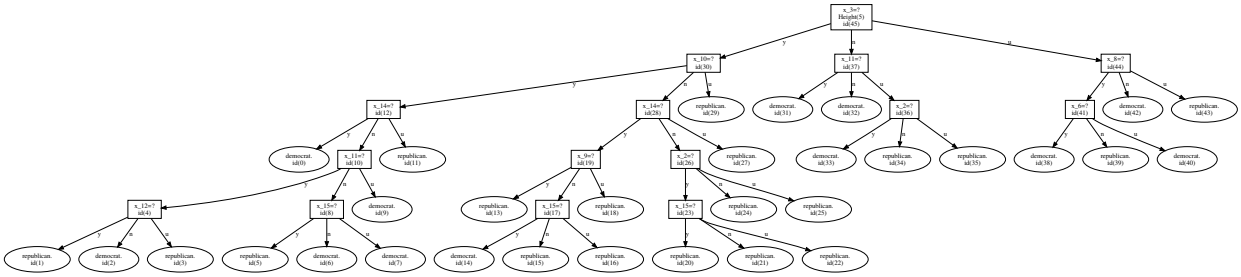

איור 6: העץ שהתקבל מהרצת ID3 עם גובה מקסימלי 2



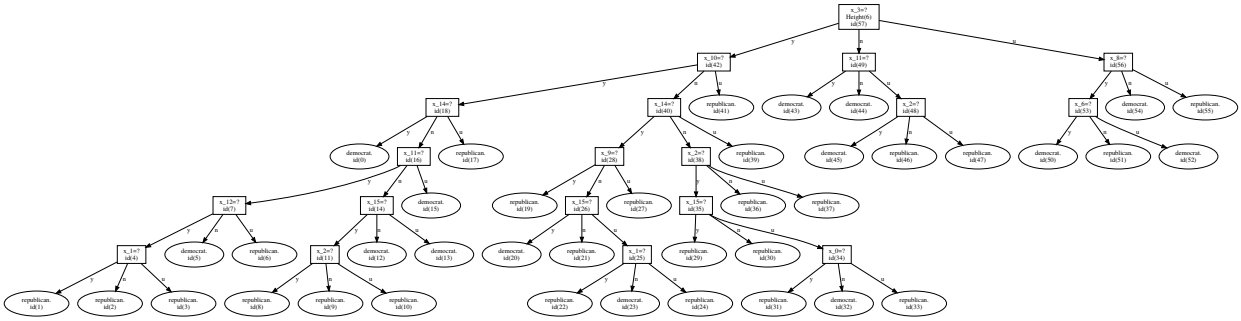

איור 7: העץ שהתקבל מהרצת ID3 עם גובה מקסימלי 3



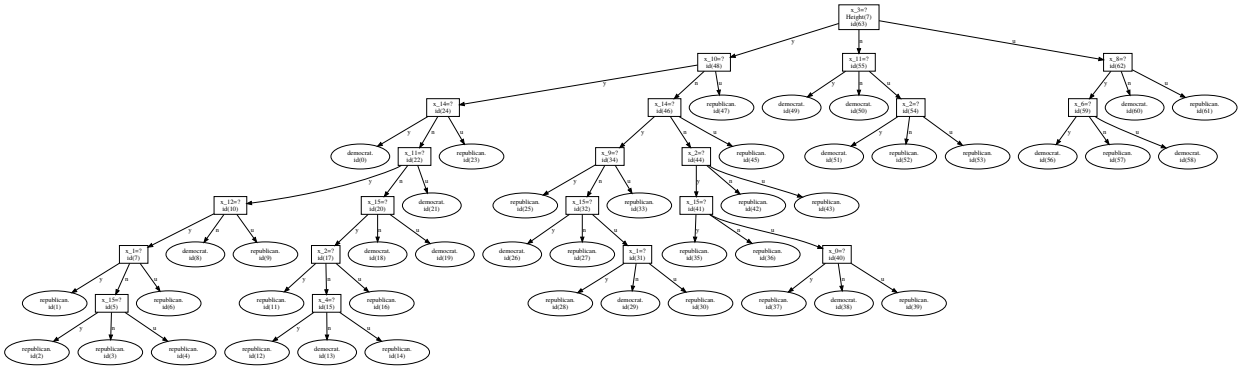

איור 8: העץ שהתקבל מהרצת ID3 עם גובה מקסימלי 4



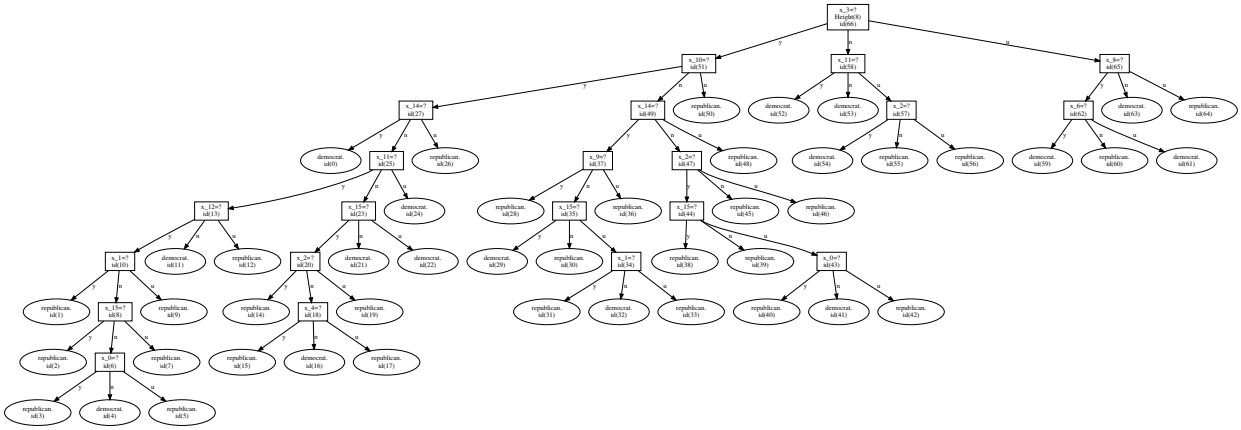

איור 9: העץ שהתקבל מהרצת ID3 עם גובה מקסימלי 5




איור 10: העץ שהתקבל מהרצת ID3 עם גובה מקסימלי 6




איור 11: העץ שהתקבל מהרצת ID3 עם גובה מקסימלי 7




איור 12: העצים שהתקבלו מהרצת ID3 עם גובה מקסימלי 16־8 ־ כולם יצאו זהים (ונתנו שגיאת אימון 0)