**מבוא לבינה מלאכותית**

**236501**

תרגיל בית 3

מגישים:

תום אגמי ת.ז: 302485628

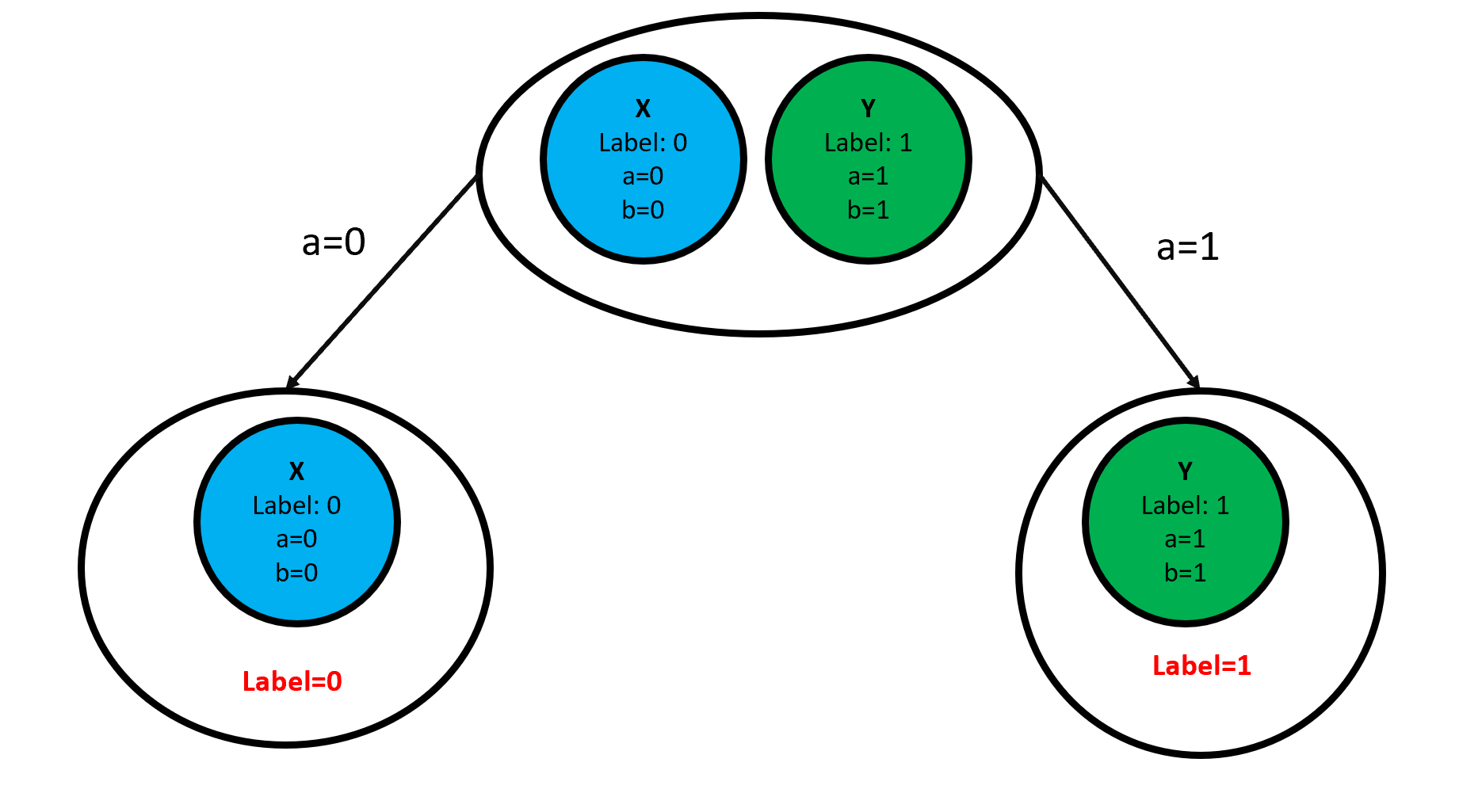
שיר רורברג ת.ז: 204380703

**תרגיל 1**

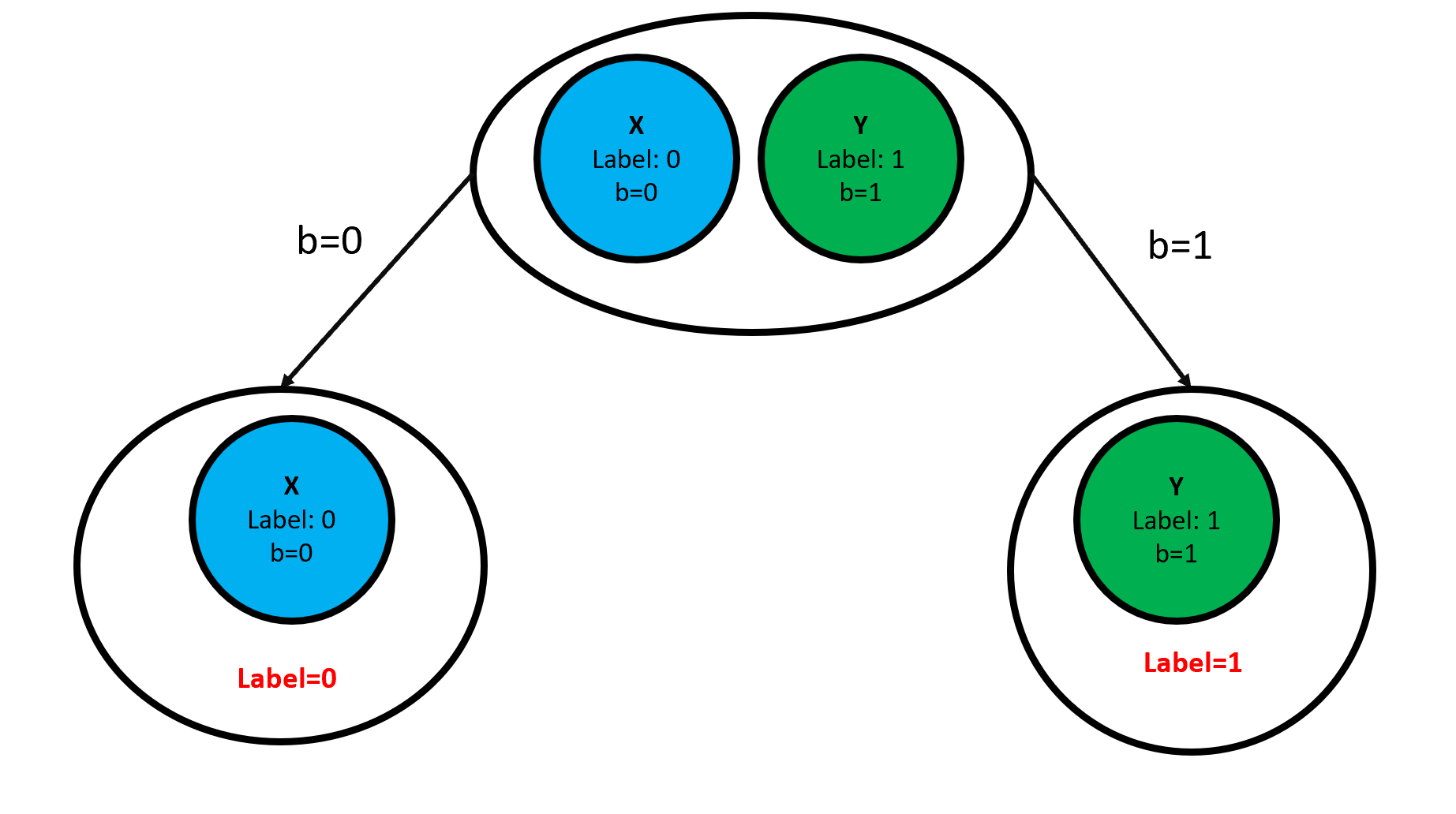
1. הטענה לא נכונה. נפריך על ידי דוגמה נגדית:

נגדיר:

עץ החלטה T אפשרי עבור הבעיה:



עץ ההחלטה עבור הבעיה:



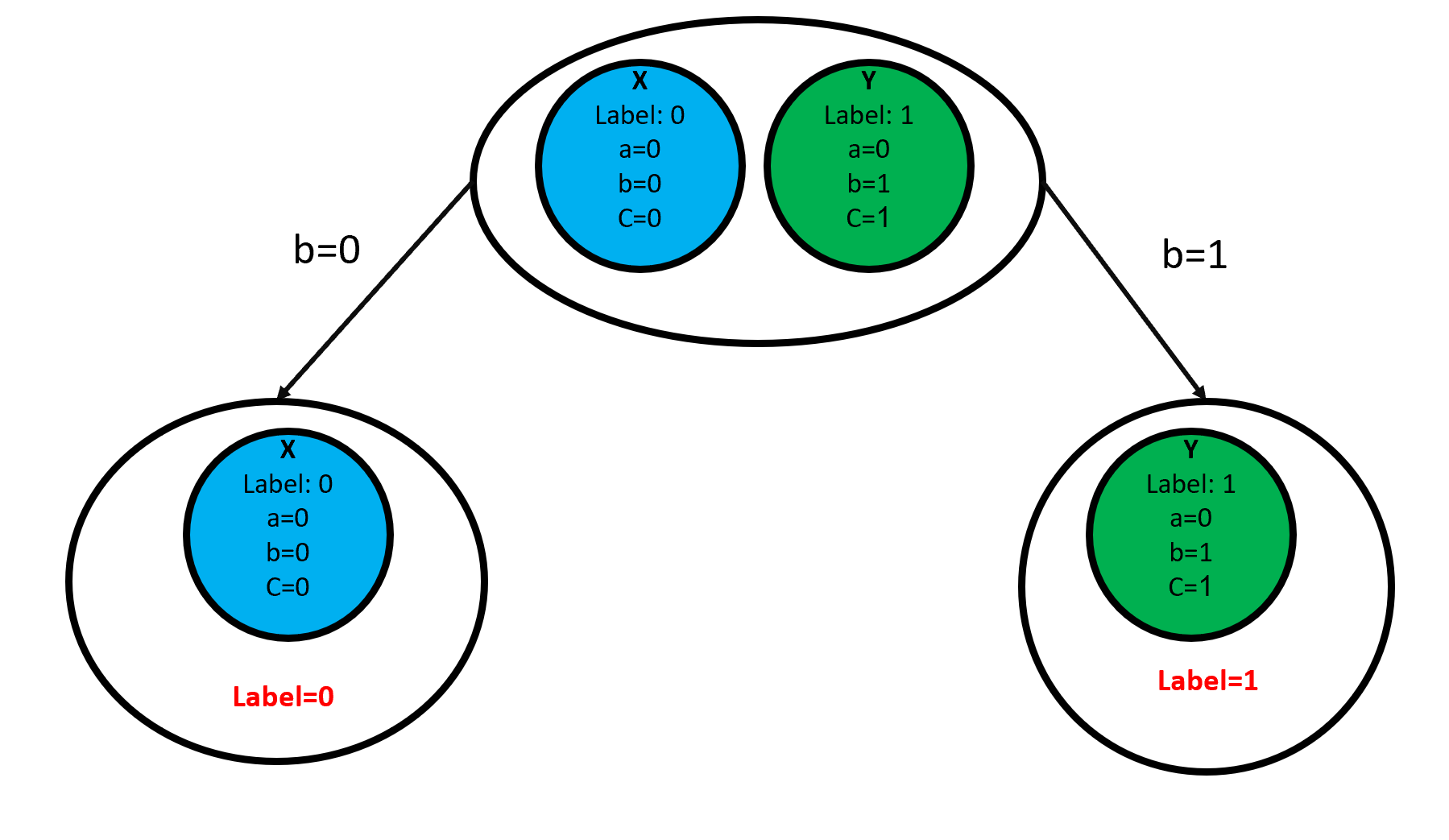
עבור דוגמת מבחן z המקיימת:

עץ T יסווג דוגמה זאת ב-0 ועץ יסווג ב-1.

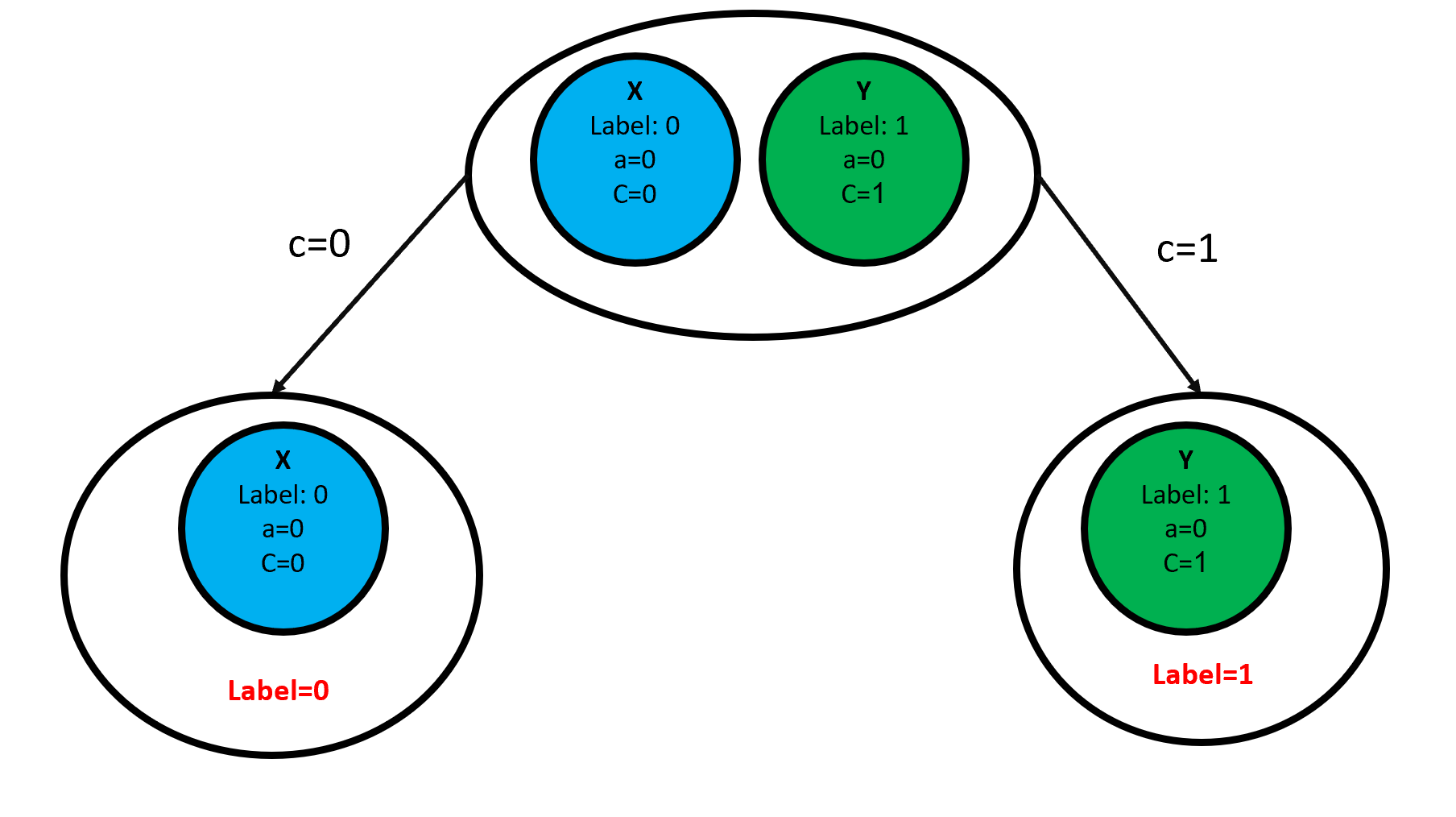
1. הטענה לא נכונה. נפריך על ידי דוגמה נגדית:

נגדיר:

*עץ ההחלטה T המתקבל:*

**

*עץ ההחלטה המתקבל:*

**

דוגמת המבחן:

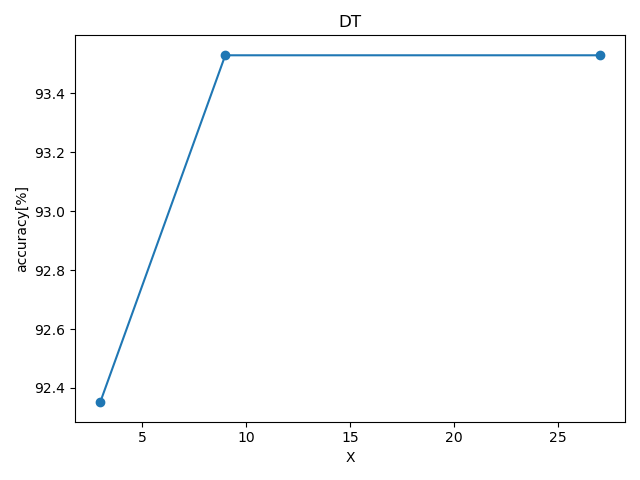
תסווג ב-1 עם עץ וב-0 בעץ , בסתירה לטענה.

**תרגיל 2**

הדיוק של קבוצת המבחן שהתקבל: 92.3529%

**תרגיל 3**

הגרף שהתקבל:

****

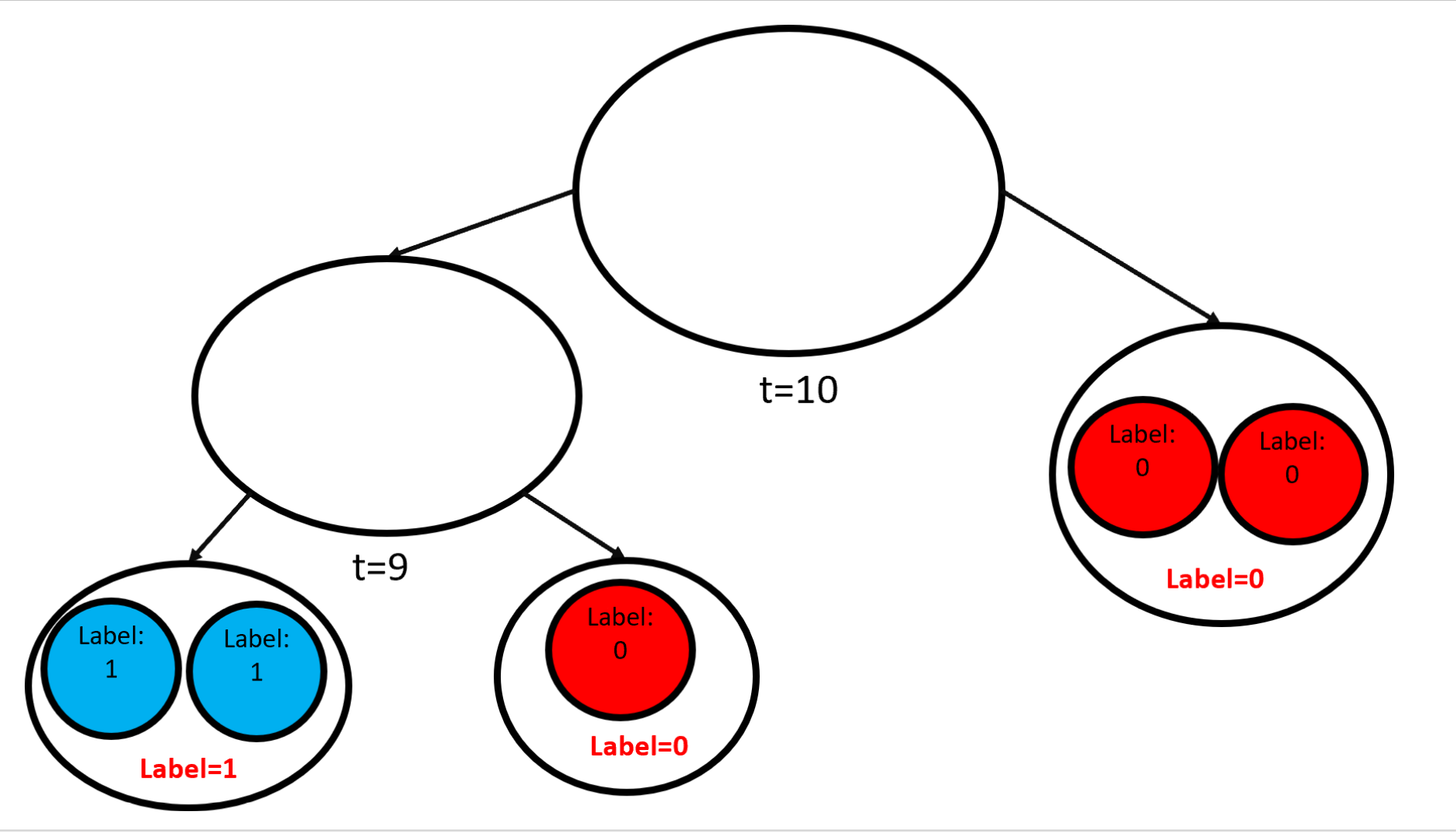
ניתן לראות שדיוק הסיווג עולה עד ערך 9. ניתן להסביר זאת על ידי כך שעבור x=3 וללא גיזום יש התאמת יתר לתוצאות המבחן. יש רעש בתוצאות ולכן עץ עקבי (ללא גיזום) ועץ עם x=3 לא מביא לדיוק המיטבי בקבוצת המבחן. עבור ומעלה, אנו מפחיתים את התאמת היתר על ידי מיצוע של הערכים כאשר יש 9 דוגמאות ופחות. עבור -ים גדולים מ-50 הדיוק יורד (מהרצות נוספות שביצענו), לכן עבור ערכים גדולים מערך זה גזמנו את העץ יתר על המידה וקיבלנו עץ קטן מדי בשביל לייצג כראוי את המבנה הבסיסי של הנתונים (תת התאמה).

**תרגיל 4**

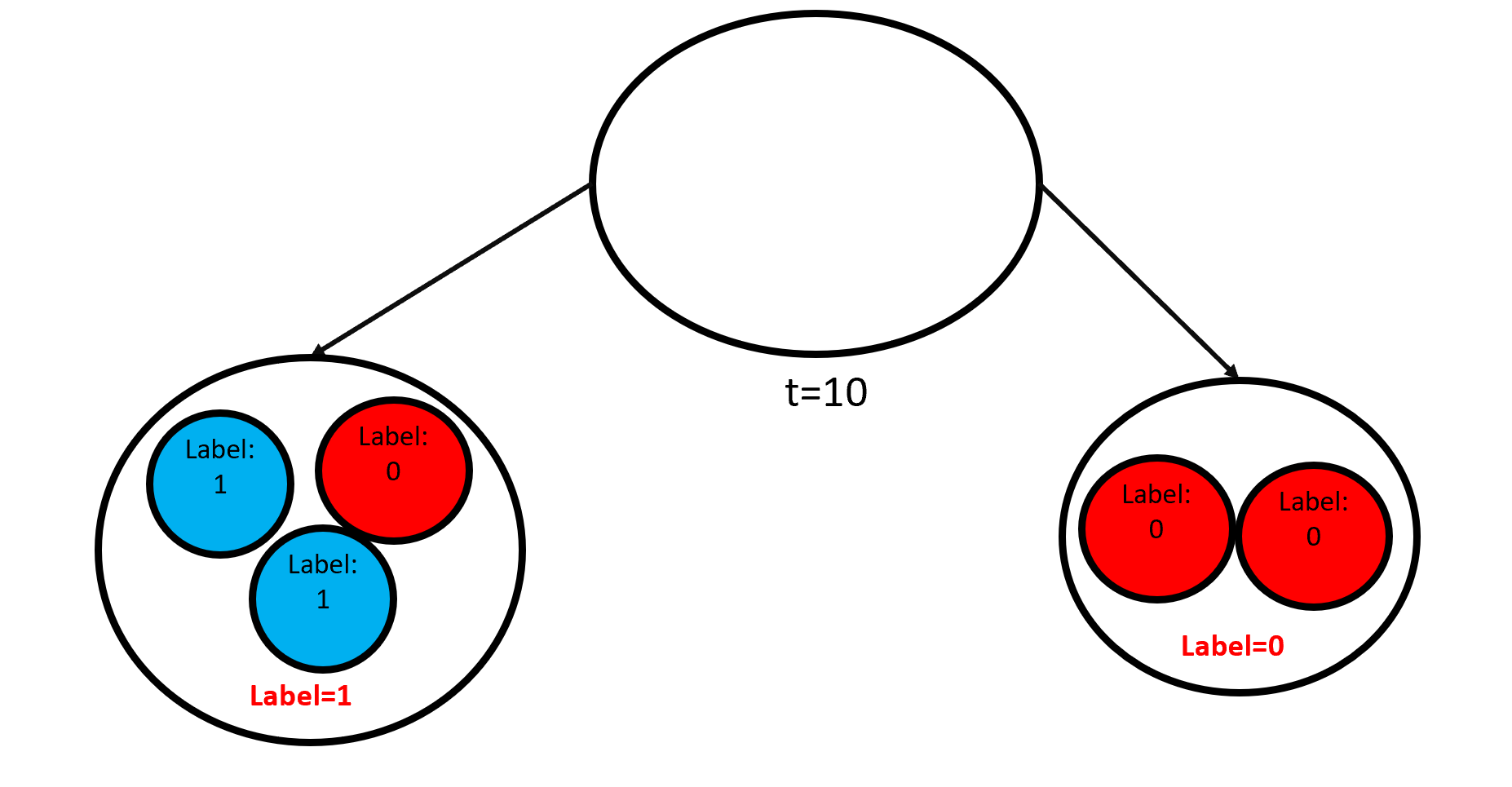
הטענה לא נכונה. נציג דוגמה נגדית.

מאגר האימון מכיל דוגמאות עם תכונה אחת רציפה.

העץ T:



העץ המתקבל:



עבור דוגמה עם , העץ יסווג ב-1 (יילך שמאלה בצומת הראשון ויגיע לעלה השמאלי).

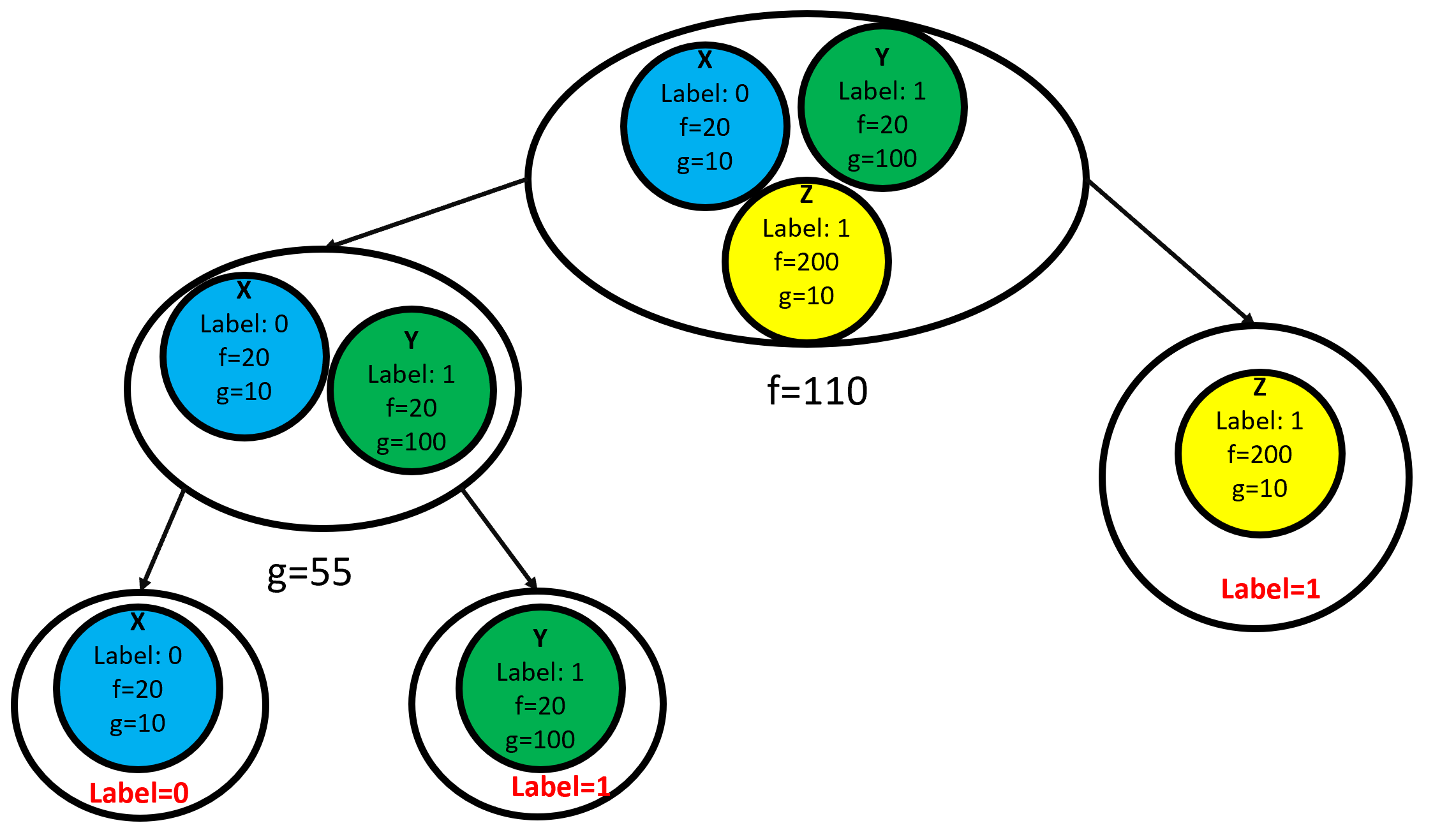
עבור , בצומת הפיצול הראשון לא נמצא בסביבת של 10 ולכן יילך שמאלה. בצומת הפיצול השני גם לא נמצא בסביבת של 9 ולכן יילך ימינה ויסווג את הדוגמה ב-0.

עבור , 9.5 נמצא בסביבת גם של הצומת הראשון וגם של הצומת השני ולכן תתקבל החלטה לפי כל העלים, מכיוון שרוב העלים הם 0, הדוגמא תתויג כ-0.

לכן לכל ערך של הדוגמה תסווג כ-0 בעץ T ותסווג כ-1 בעץ בסתירה לטענה.

**תרגיל 5**

הטענה לא נכונה. דוגמה נגדית:



עבור דוגמה זו ו- העץ יהיה זהה לעץ , מפני שאין דוגמאות מבחן שנופלות בתוך סביבת של ערך הסף.

אבל עבור דוגמת מבחן שעבורה: יתקיים:

* עם כלל ההחלטה הרגיל בעץ המתואר למעלה (שזהה לעץ ), המסלול עבור הדוגמה יגיע לעלה השמאלי ביותר ולכן יהיה מסווג ב-
* עם כלל החלטה שני הפיצולים יוצאים בתוך סביבת של דוגמת המבחן ולכן המסלול יגיע לכל העלים בעץ ויקבל החלטה לפי הסיווג של רוב העלים. מכיוון שברוב הדוגמאות בעלים , זה גם יהיה הסיווג שהאלגוריתם ייתן.

לכן העץ עם כלל ההחלטה הרגיל והעץ עם כלל החלטה יסווגו שונה דוגמה זו, בסתירה לטענה.

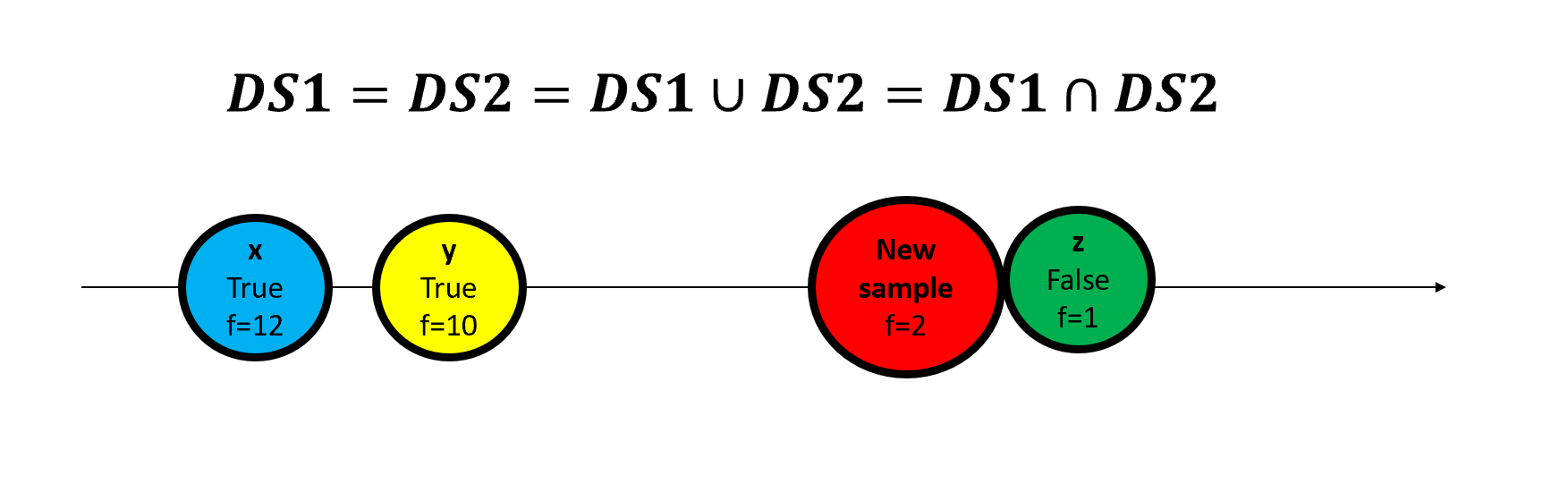
**תרגיל 6**

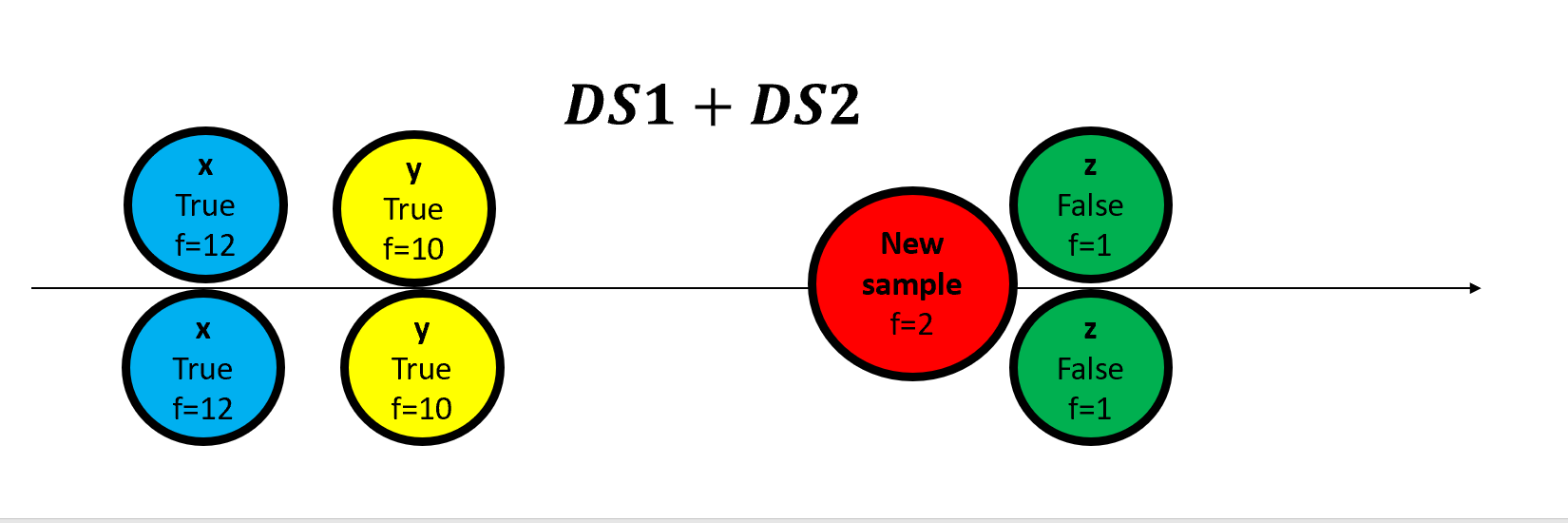
הדיוק שהתקבל עבור קבוצת המבחן: 93.5%

**תרגיל 7**

הטענה לא נכונה. נפריך על ידי דוגמה נגדית.

נדגים באמצעות האיור הבא:





באיור הקבוצות זהות ומכילות את הדוגמאות (ירוק, צהוב וכחול) הדוגמאות בעלות תכונה אחת שמסומנת ב-. דוגמת מבחן חדשה מסומנת באדום ובעלת . מכיוון שהאלגוריתם הוא , הוא יבחר לסווג את הדוגמה האדומה לפי תווית הרוב של 3 הדוגמאות הקרובות אליה, כלומר לפי מכיוון שרוב הדוגמאות () עם תווית True, דוגמת המבחן תסווג ב- Trueמכיוון ש- זהות, גם זהות ולכן גם הן יסווגו את דוגמת המבחן ב-True. כלומר כל תנאי השאלה מתקיימים.

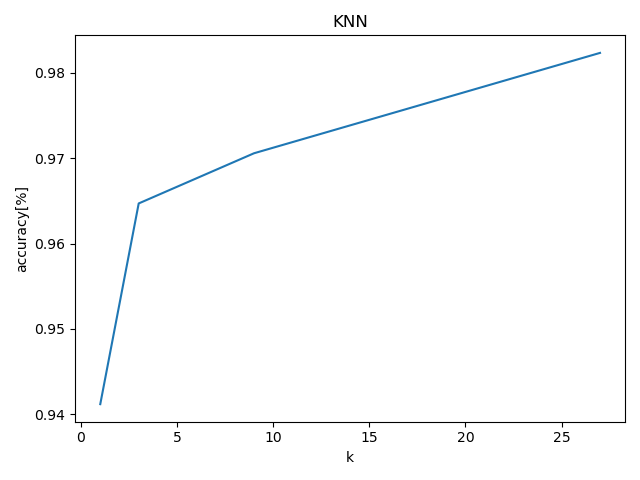
לעומת זאת, בקבוצה , הדוגמאות משוכפלות ולכן 3 הדגימות הקרובות ביותר לדוגמת המבחן הן פעמיים z ופעם אחת y, כלומר התווית של רוב הדגימות (של z) היא False ולכן דוגמת המבחן תסווג כ-False, בסתירה לטענה.

**תרגיל 8**

הדיוק שהתקבל עבור קבוצת המבחן: 97%

**תרגיל 9**

הגרף שהתקבל:



ניתן לראות שהדיוק עולה ככל ש-k עולה. ניתן להסיק מכך שעבור ערכים קטנים מדי של k מתקבלת התאמת יתר לנתונים. עבור ערכים גדולים יותר של k ממצעים על פני מספר גדול יותר של ערכים קרובים ובכך מפחיתים את הרעש של הדגימות. היינו מצפים שהחל מערך כלשהו של k נקבל ירידה בדיוק עקב מיצוע על פני מספר גדול מדי של דגימות אימון. (עבור ערך שגדול ממספר דגימות האימון, כל דוגמה חדשה תסווג באותו אופן לפי הסיווג הנפוץ ביותר של כל דגימות האימון. במקרה זה, עבור דוגמאות מבחן עם התפלגות אחידה על התגיות נצפה לקבל דיוק בסביבה של 50%).

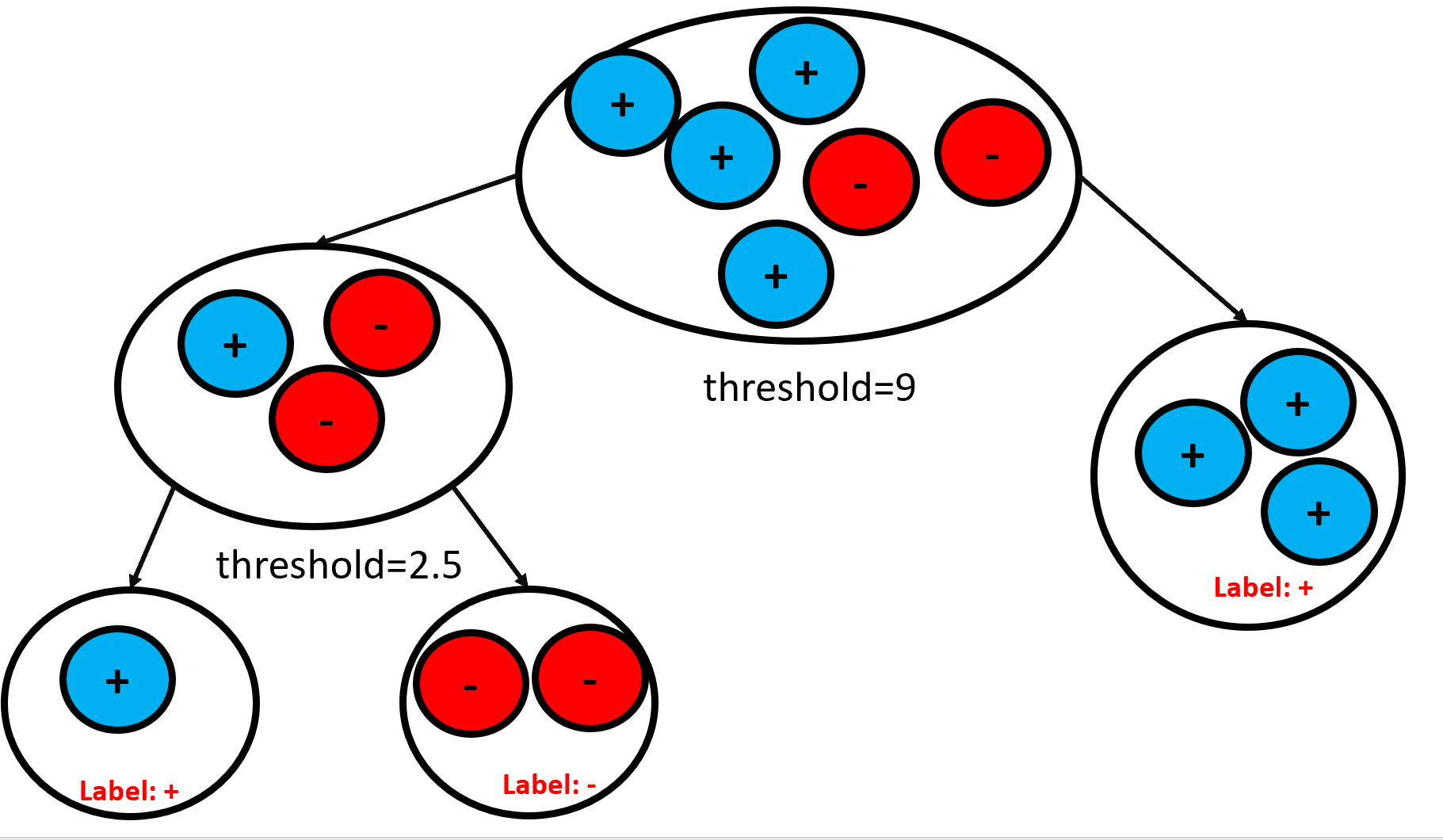
**תרגיל 10**

נראה שעבור שני המסווגים יסווגו את כל דוגמאות האימון בצורה זהה.

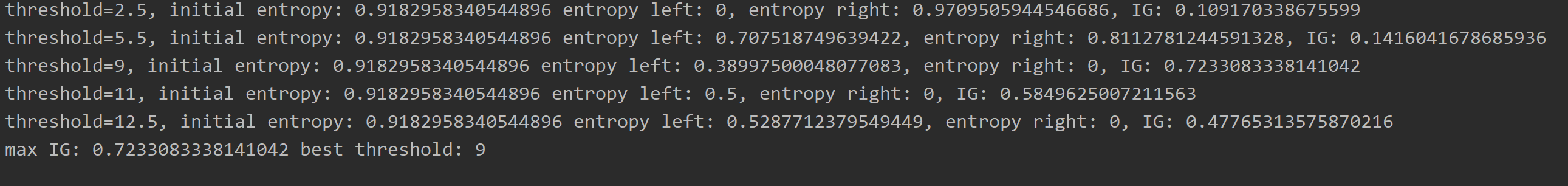
מסווג ה- יסווג את כל הדוגמאות עבורן ב- ואת כל הדוגמאות עבורן ב-

כלומר יסווג את דוגמאות האימון שעבורן ב- ואת שאר דוגמאות האימון (עבורן ) ב-

עץ ההחלטה עבור קבוצת האימון הנתונה:



כאשר הפיצול הראשון נקבע לפי IG המקסימלי. הערכים שהתקבלו:



הערך המקסימלי התקבל עבור ערך סף של 9 ועבור הפיצול השני ה-IG הגבוה ביותר מתקבל עבור הפרדה מוחלטת ל- ו-, כלומר ערך סף של 2.5.

עבור ערך של :

הדוגמא עם לא תהיה בסביבת של הפיצול הראשון, אך תהיה בסביבת של הפיצול השני ולכן תסווג לפי שני העלים השמאליים, כלומר ב- כנדרש.

הדוגמא עם לא תהיה בסביבת של הפיצול הראשון, אך תהיה בסביבת של הפיצול השני ולכן תסווג גם לפי שני העלים השמאליים, כלומר ב- כנדרש.

הדוגמא של תהיה בסביבת של הפיצול הראשון ולכן תגיע לעלה הימני ולא תהיה בסביבה של הפיצול השני ולכן תגיע גם לעלה האמצעי. כלומר תסווג לפי שני מינוסים ושלושה פלוסים – כלומר ב- כנדרש.

הדוגמא של תהיה בסביבת של הפיצול הראשון ולכן תגיע לעלה הימני ולא תהיה בסביבה של הפיצול השני ולכן תגיע גם לעלה האמצעי. כלומר תסווג לפי שני מינוסים ושלושה פלוסים – כלומר ב- כנדרש.

הדוגמאות של ו- לא יהיו בסביבת של הפיצול הראשון ולכן יגיעו רק לעלה הימני כלומר ויסווגו לפי שלושה פלוסים – כלומר ב- כנדרש.

**תרגיל 11**

הדיוק שהתקבל עבור קבוצת המבחן: 94.1%