מבוא לבינה מלאכותית

תרגיל בית 3

מגישים:

אביב כספי 311136691

יקיר יהודה 205710528

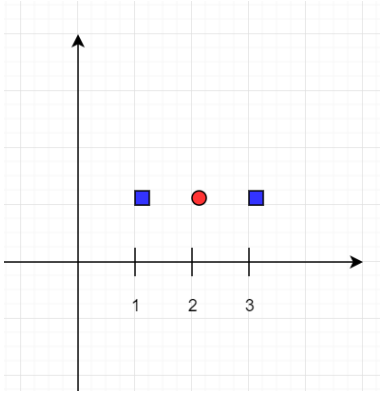
חלק א'

1. א. עץ החלטה מפריד בין דוגמאות על ידי קווים ישירים המקבילים לצירים, לכן נדרוש m=0 וn ממשי כלשהו. כך נקבל כי y=n כלומר קו ישר המקביל לציר x, ואז בפיצול יחיד עם כלל פיצול זה (y>n) נוכל לפצל את כל הדאטה שלנו.

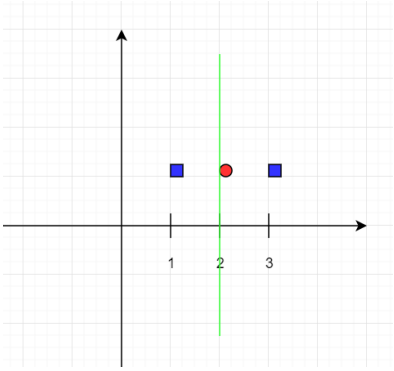
ב. עץ החלטה מפריד בכל איטרציה על ידי מציאת קו ישר המקביל לאחד הצירים, המפריד בין הדוגמאות, כאשר העקום המתקבל, בסיום בניית העץ, יהיה איחוד של כל הקווים שמצאנו. כלומר מרחב ההיפותזות מכיל עקומים המורכבים מישרים מקבילים לצירים.
כאשר כל פיצול מוסיף לנו קו נוסף , לכן כל מקרה בו הדוגמאות לא ניתנות להפרדה על ידי קו ישר מקביל לצירים, נצטרך לבצע יותר מפיצול אחד.

ג. נשנה את כלל הפיצול שיכיל יחס בין הפיצ'רים של כל אובייקט. לדוגמא אם ניתן להפריד בין הדוגמאות על ידי ישר מהצורה y=mx+n , כלל הפיצול שלנו יהיה y<mx+n , כאשר אם התשובה היא כן, נסווג בצורה אחת, ואם התשובה היא לא נסווג בצורה שנייה. כך למעשה אנחנו מתייחסים ליותר מפיצ'ר אחד בכל כלל, ומאפשרים שמרחב ההיפותזות שלנו, יכיל גם קווי הפרדה שלא מקבילים לצירים.

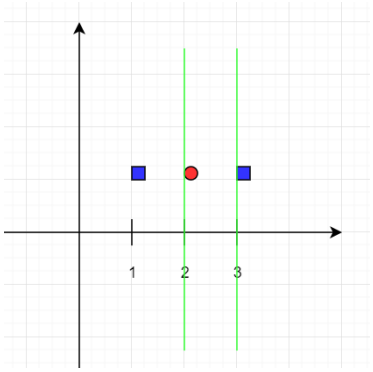
ד. כן, כאשר התכונות רציפות, ייתכן כי בין שתי דוגמאות חיוביות נמצאת דוגמא שלילית ונפריד על ידי שימוש באותה תכונה פעמיים .
נסתכל על הדוגמא הבאה :



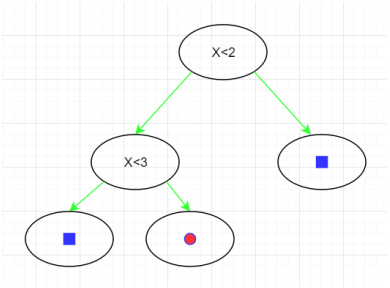
נשים לב כי לפי תכונה y לא ניתן להפריד בין הדוגמאות, אך לפי דוגמא x ניתן להפריד דוגמת ריבוע אחת מבין הדוגמאות האחרות (information gain) זהה להפרדה עבור x>2 ו x<2). כלומר הכלל פיצול הראשון יהיה לדוגמא x>2.



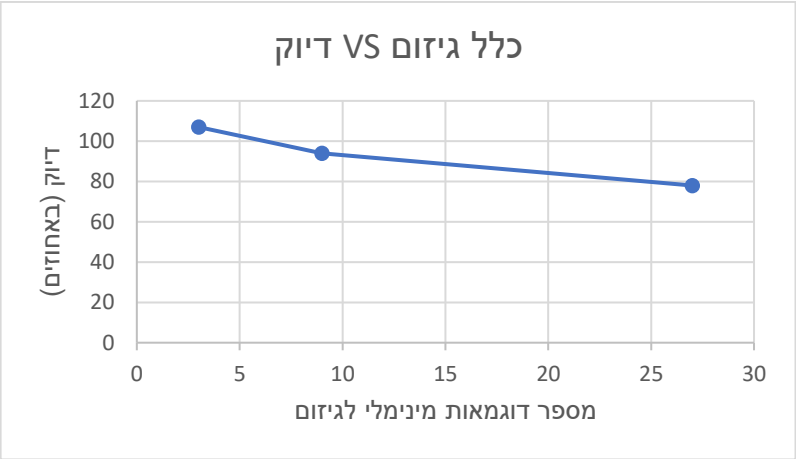
כעת נרצה לפצל שוב כי לא הגענו לצמתים הומוגניים. ושוב נראה כי לא ניתן להפריד לפי תכונה y אלא רק לפי x , כלומר יבחר כלל פיצול נוסף מהצורה x<3.



כלומר, הגענו להפרדה מלאה של הדוגמאות, כאשר נבדוק ברצף שני כללים בנוגע לתכונה x.

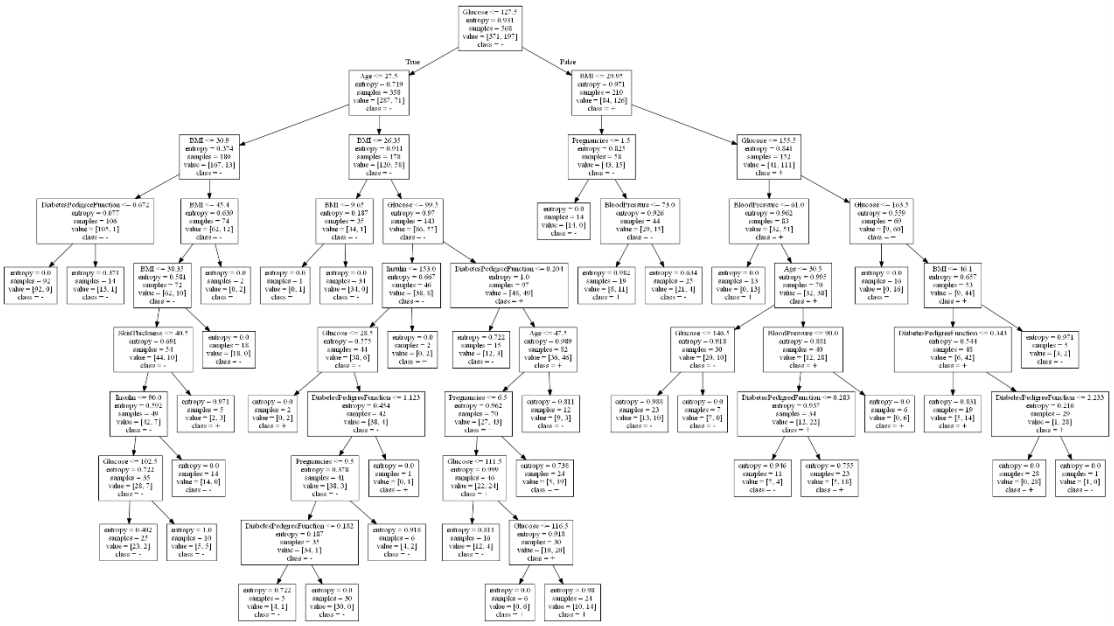


3. הגרף שקיבלנו בסעיף זה:



ניתן לראות בגרף כי ככל שנעלה את מספר הדוגמאות המינימלי לפיצול, נקבל ירידה בדיוק המסווג שלנו.
לדעתנו ירידה זו בביצועים נובעת מכך שהדאטה שלנו אינו מאוזן, כלומר ישנם יותר דוגמאות שליליות מאשר חיוביות, וזה גורם לכך שכאשר נגזום את העץ עם מספר מינימלי גדול יותר, ניצור עלים בהם יש יותר דוגמאות. בגלל החוסר איזון בדאטה, נקבל כי יש סיכוי גדול יותר כי בעלה יהיה יותר דוגמאות שליליות מאשר חיוביות ולכן נסווג יותר דוגמאות כשליליות ולכן נפגע ביכולת שלנו להכללה ובהתאם בדיוק המסווג.

4. מבנה העץ עבור x=27



ניתן לראות כי אחרי הגיזום יש יותר עלים שמסווגים במינוס כאשר הם לא הומוגניים.

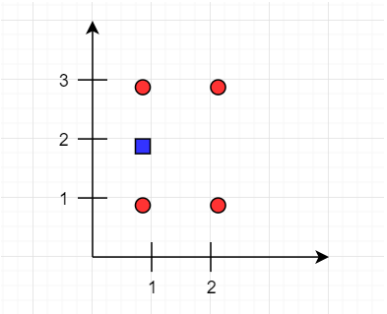
חלק ב'

5.

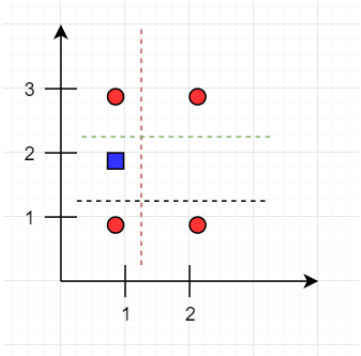
א. נעריך כי ההסתברות שהמסווג שלנו יסווג את X כשלילית בהסתברות קרובה ל-p.
A עץ ID3 לא גזום, כלומר הוא ממשיך לפצל בזמן האימון, עד אשר כל העלים הומוגניים.
בגלל שנתון כי יש הרבה יותר דוגמאות שליליות מחיוביות, נוכל להסיק כי יהיו הרבה יותר מסלולים לעלים שליליים מאשר חיוביים. כאשר נוכל להסיק כי ההסתברות לסווג את X כשלילית הינה בערך p בגלל שהדוגמאות בלתי תלויות (כלומר, יש הסתברות זהה להגיע לכל עלה בעץ).

ב. כאשר יש הרבה יותר דוגמאות "-" מאשר "+" , אם נגזום את העץ, ייתכן כי נגיע למצב בו בכל העלים יש מספר גדול יותר של דוגמאות "-" מאשר "+", לכן עץ ההחלטה שלנו יסווג כל דוגמא כשלילית בהסתברות גבוהה מאד (גדולה מ-p). כמו שראינו בסעיף 4, כאשר גוזמים עץ עם דאטה לא מאוזן, רוב העלים הלא הומוגניים בעץ יכריע לכיוון הסיווג עם מספר הדוגמאות הגדול יותר, כאשר ככל שהיחס בין הדוגמאות גדל כך גם הסיכוי לסיווג המחלקה הגדולה.

6. נפריך את הטענה בעזרת דוגמא נגדית :



נסתכל על הגרף הנ"ל המתאר את דוגמאות האימון שלנו לפני איזון הדאטה, כאשר עיגולים הינם דוגמאות שליליות וריבועים דוגמאות חיוביות.
כאשר נבנה את עץ ID3 נעבור על כל כלל הפיצול האפשריים ונחשב לכל אחד מהכללים את הIG שלו, ונבחר את הכלל פיצול בעל IG מקסימלי.
אם נסתכל על הדוגמא הנ"ל נשים לב כי יתכנו הפיצולים הבאים:

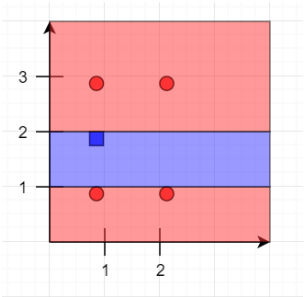


כאשר אם נשים לב, כל כלל פיצול (קו מקווקו) מפצל את הדאטה שלנו בצורה זהה לחלוטין, לקבוצה אחת עם 2 דוגמאות שליליות, וקבוצה שנייה עם 2 דוגמאות שליליות ואחת חיובית. כלומר, הIG של כל הפיצולים הללו הינו זהה ולכן האלגוריתם יבחר את אחד הפיצולים הנ"ל בצורה אקראית.

בשביל הדוגמא שלנו נניח כי האלגוריתם בחר בפיצול הירוק כלומר כלל הפיצול הינו y>2.

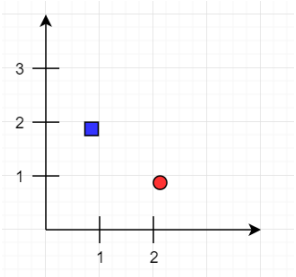
כעת שוב האלגוריתם יבחר כלל פיצול חדש אשר ממקסם את IG , נשים לב כי הפיצול השחור מפצל בצורה מושלמת את הדאטה שנותר ולכן הIG שלו יהיה מקסימלי, כלומר הפיצול הבא יהיה y>1.

נסתכל על איזור ההחלטה של כל סיווג לפי כללי הפיצול שנבחרו:

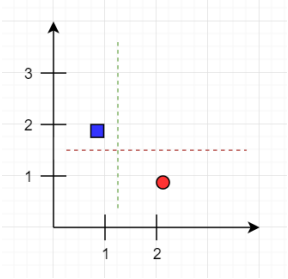


כלומר אם נקבל דוגמת מבחן עם התכונות (2,2) , עץ ההחלטה שלנו יסווג אותה כחיובית.

כעת נאזן את הדאטה שלו:



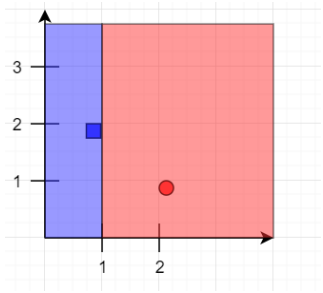
כעת נותרנו עם דוגמא שלילית יחידה, והאלגוריתם שלנו יבחר כלל פיצול שממקסם את IG.



ישנם שני כללי פיצול אפשריים המסומנים בקווים המקווקוים, שניהם מפצלים את הדאטה בצורה מושלמת ולכן ה IG שלהם זהה ומקסימלי.

נניח כי נבחר כלל הפיצול הירוק, כלומר כלל הפיצול שלנו הוא x>1.

נסתכל על איזור ההחלטה של כל סיווג לפי כלל הפיצול שנבחר:



כעת אם ניקח את דוגמת המבחן הקודמת עם התכונות (2,2) , עץ ההחלטה שלנו יסווג אותה כשלילית.
כלומר הראנו כי דוגמת מבחן כלשהי סווגה על ידי העץ ID3 הלא גזום עם קבוצת אימון מאוזנת בצורה שלילית, אך אותו עץ ID3 עם קבוצת האימון הלא מאוזנת מסווג את הדוגמא כחיובית.

7. א. כמו שראינו בסעיף 5 , כאשר הדאטה שלנו לא מאוזן, ישנה הסתברות גבוהה יותר שדוגמא כלשהי תסווג בסיווג אשר מופיע יותר בדוגמאות האימון כאשר אנחנו גוזמים את העץ. כלומר בדוגמא שלנו, יש יותר דוגמאות שליליות מאשר חיוביות, לכן בעץ גזום ההסתברות שנסווג דוגמא כשלילית גבוהה יותר (ההסתברות שבעלים יהיו יותר דוגמאות שליליות מאשר חיוביות). כלומר בהתאם ישנו סיכוי גדול יותר כי נגדיל את מספר הFN (ונקטין את FP) ולכן השגיאה תגדל.

ב. תוצאות ההרצה שקיבלנו :

```
dt1 not trimmed
Error_w = 129
dt1 trimmed
Error_w = 145
```

ניתן לראות כי כמו שציפינו, השגיאה עבור עץ גזום הינה יותר גדולה מאשר השגיאה עבור עץ לא גזום.

8. א. כמו שראינו בסעיפים הקודמים, ההסתברות לסווג דוגמא כשלילית כאשר הדאטה שלנו לא מאוזן הינה גדולה מאד ולכן נסווג הרבה דוגמאות כשליליות ומכך יעלה ערך השגיאה $Error_w$. כאשר נאזן את הדאטה שלנו כך שיהיו מספר דוגמאות שווה לכל סיווג, נעלה את ההסתברות (כאשר הדוגמאות בלתי תלויות אחת בשנייה) לסווג את הדוגמא כחיוביות, כלומר ערך השגיאה יקטן בהתאם.

ב. קיבלנו בסעיף זה עם דאטה מאוזן ועץ לא גזום שגיאה של 129, לעומת העץ DT1 לא גזום עם דאטה לא מאוזן שקיבלנו שגיאה של 123.
נשים לב כי זה סותר את התשובה שנתנו בסעיף הקודם, ולאחר בדיקה של התוצאות, שמנו לב כי הסיבה לכך שיש עליה בשגיאה כאשר מאזנים את הדאטה נובעת בגלל ירידה בדיוק המסווג. במקרה שלנו עבור דאטה לא מאוזן קיבלנו את השגיאות הבאות $FN = 24, FP = 27$ ועבר הדאטה המאוזן קיבלנו : $FN = 23, FP = 37$
כלומר, ההנחה שלנו כי מספר ה*FN* ירד אכן הייתה נכונה (למרות שקיבלנו ירידה של 1 בלבד) , אך מצד שני פגענו בביצועי המסווג בגלל הורדת דאטה שתרם לסיווג ולכן העלנו בהרבה את *FP* ומכאן קיבלנו כי למרות ירידה ב*FN* הייתה עלייה בשגיאה.

9. א. על מנת שיהיה כדאי להגדיר p=1 , נרצה כי השגיאה שלנו אחרי ההגדרה תהיה קטנה יותר.
נסמן TP,FP,TN,FN את הערכים של המודל שלנו לפני הגדרת p וב- TP',FP',TN',FN' את הערכים אחרי הגדרת p .
נרצה שיתקיים :
4FN + FP >4FN' + FP'
נשים לב כי כאשר p=1 , נסווג כל דוגמא כחיובית, כלומר מתקיים :

FN'=0, TN'=0, FP'=TN+FP

כאשר השגיאה החיובית שלנו כרגע היא כל הדוגמאות השליליות הקיימות.

כלומר : $4FN + FP > 4FN' + FP' = 0 + TN + FP$

$4FN > TN$

במקרה זה יהיה לנו עדיף להגדיר p=1.

ב. נרצה למצוא את התוחלת של השגיאה שלנו.

נשים לב כי שגיאת האימון של עץ DT1 מבלי להשתמש בp הינה 0, מפני שממשיכים לפצל את הצמתים עד אשר מגיעים לעלים הומוגניים, כלומר לכל דוגמא מקבוצת האימון נסווג אותה בצורה נכונה (עץ DT1 הינו עץ עקבי).

כלומר, לאחר שימוש בשיטה המוגדרת, יתכן כי נשנה רק דוגמאות שמסווגות כשליליות לחיוביות. כלומר, לא יתכן כי נסווג דוגמת אימון חיובית כשלילית ולכן תמיד יתקיים FN=0.

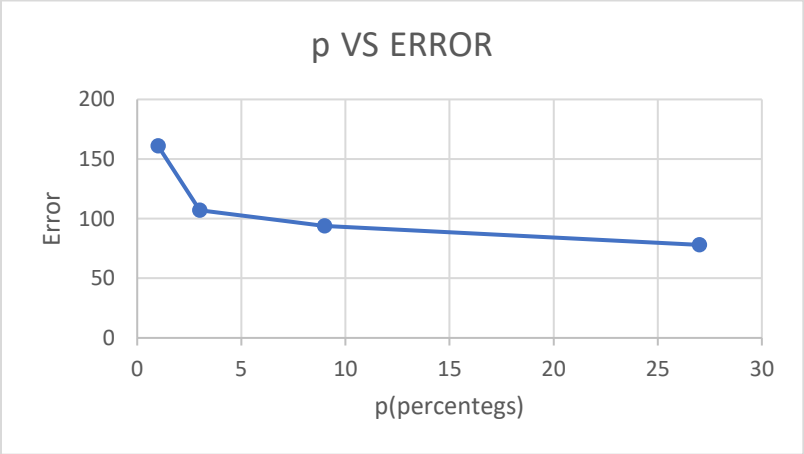
ניתן לראות כי FP מתפלג בצורה בינומית, כאשר מספר הניסויים שלנו הוא כגודל קבוצת הדוגמאות השליליות - |F| , ונגדיר הצלחה כשינוי של סיווג שלילי לחיובי, בהסתברות p.

$$E(Error_w) = 4 * E(FN) + E(FP) = E(FP)$$

לפי תוחלת של התפלגות בינומית נקבל כי :

$$E(FP) = p|F|$$

ג. הגרף שקיבלנו בסעיף זה:



כאשר p=0 אנחנו עם עץ DT1 המקורי.

ניתן לראות בגרף כי ככל שp גדל כך גדלה השגיאה שלנו, לאחר בדיקת נתוני השגיאות, שמנו לב כי השגיאות FN שלנו יורדות בצורה משמעותית כאשר מגדילים את p מ-24 עבור p=0 ל 17 עבור p=0.2 והשגיאות FP שלנו עולות בצורות משמעותית , מ-27 עבור p=0 ל-66 עובר p=0.2 . כלומר באמת מתקיים כי השגיאה FN שלנו יורדת אך ישנה פגיעה רצינית בדיוק המסווג, אשר גורמת לעלייה בשגיאה הכוללת. זה כמובן נובע מפני שאנחנו בצורה ידנית משנים סיווג שקיבלנו על ידי המסווג שלנו, ללא ידע נוסף אלא בצורה אקראית.

10. על מנת למזער את ערך השגיאה, נרצה כי המסווג שלנו, יסווג יותר דוגמאות בצורה חיובית מאשר שלילית וכך בעצם נקטין את ה*FN* וייתכן כי נגדיל את *FP* לכן ננסה לשמור על יחס כך ש*FP* יגדל פחות מפי 4 מאשר ה*FN* יורד (על מנת שסך השגיאה תקטן $Error_w = 4FN + FP$).

על מנת לסווג יותר דוגמאות כחיוביות, תחילה נגדיר את $\alpha = 4$ כך בעצם נחזק בעלים את הדוגמאות החיוביות, וכל דוגמא חיובית תהיה בעלת משקל כפול פי 4 . בנוסף לכך, נגדיר את $\delta = \frac{4}{5}$ על מנת לשמור על היחס של *1:4* בכמות האינפורמציה (*IG*) של דוגמא חיובית ביחס לשלילית, כך

למעשה נבחר כללי פיצול, אשר ישמרו על יחס זה בין הדוגמאות בצמתים פנימיות בעץ.

13. כעת נתון לנו כי מספר הדוגמאות השליליות גדול בהרבה ממספר הדוגמאות החיוביות בקבוצת האימון.
לכן, נוכל להסיק כי ככל ש *k* יגדל, כך גם ההסתברות שדוגמת מבחן כלשהי תסווג כשלילית תגדל בהתאם. כאשר עבור *k* ששווה לפי 2 מגודל הקבוצה החיובית, נקבל כי ההסתברות לסיווג דוגמת מבחן כשלילית היא 1. (מספר הדוגמאות החיוביות הקרובות יהיה במקסימום גודל הקבוצה החיובית, לכן תמיד יהיה יותר דוגמאות שליליות קרובות)
כאשר *k* קטן (שווה ל1), בגלל אי תלות בין הדוגמאות, נוכל להסיק כי ישנה הסתברות של *p* שדוגמת מבחן *x* כלשהי, תהיה קרובה יותר לדוגמת אימון שלילית, ולכן תסווג כשלילית בהסתברות שווה ל*p*.

כאשר *k* קטן אך גדול מ1, ההסתברות שדוגמת מבחן תסווג כשלילית גדולה מ- *p* (מפני שההסתברות שדוגמת המבחן תהיה קרובה ליותר מדוגמת מבחן אחת חיובית הינה קטנה יותר מאשר *q*).

14. קיבלנו : $Error_w = 154$

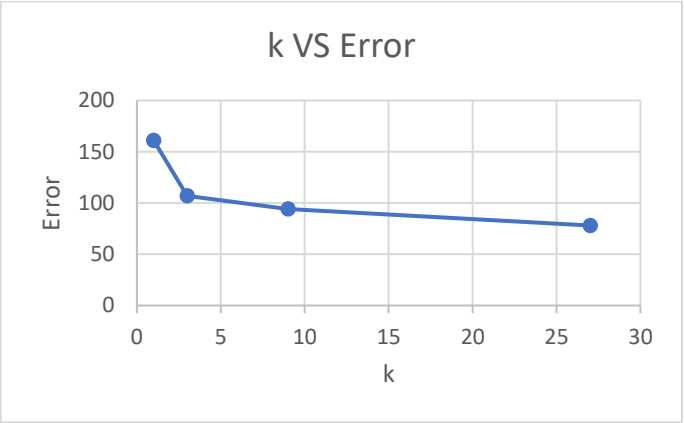
15. א. התכונות האופטימליות שמצאנו הן:

[1, 2, 5, 6, 7] עם דיוק של 82 אחוז

ב. התכונות שמצאנו עם אלגוריתם *SFS* הן:

[1, 7] עם דיוק של 78 אחוז

17. הגרף שקיבלנו בסעיף זה :



ניתן לראות בגרף כי ככל ש*k* עולה כך יורדת השגיאה.
נזכור כי בסעיף זה החשבנו כל דוגמא חיובית כ-4 דוגמאות חיוביות.
כלומר, כאשר נבדוק את *k* השכנים הקרובים ביותר, מספיק שיהיו לנו פי 4 פחות דוגמאות חיוביות מאשר שליליות על מנת שנסווג את הדוגמא כחיובית.
בגלל הגדרה זו, ובגלל שהיחס בין הדוגמאות השליליות לחיוביות בדאטה שלנו הוא הינו קטן מ-4.
נקבל כי ככל שנגדיל את *k*, מספר הדוגמאות החיוביות שנהיה קרובים אליהם יגדל ב-1 יותר מהר מאשר מספר הדוגמאות השליליות יגדל ב-4. כלומר, ככל ש *k* גדול יותר נסווג יותר ויותר דוגמאות כחיוביות. לכן, *FN* שלנו יקטן ו*FP* יעלה.
אם נסתכל על תוצאות ההרצה, נראה כי אכן *FN* הגיע ל-1 עבור *k* גדול, לעומת 32 עבור *k* קטן.
ובהתאם, *FP* הגיע ל 74 עבור *k* גדול, לעומת 33 עבור *k* קטן.