[redacted]

<u>שאלה 2</u>

הוכחה:

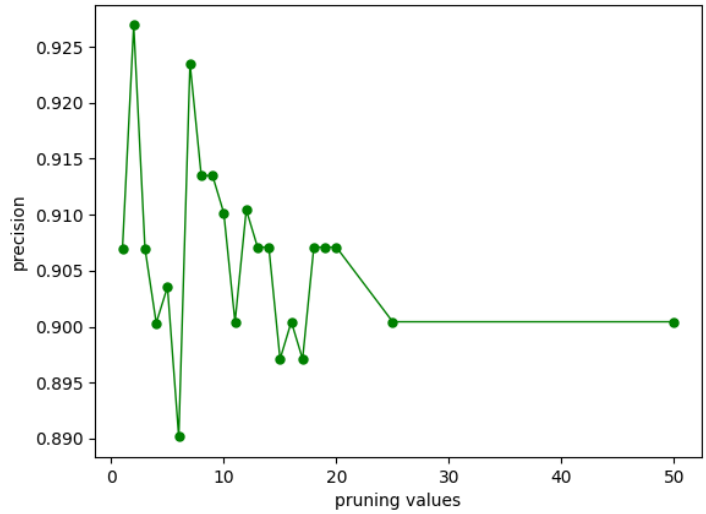
תחילה נגדיר מספר סימונים: בהינתן קבוצת train, נסמן ב-$v_{i,j}$ את הערך של התכונה ה-j של דגימה ה-i בקבוצת ה-train. נסמן ב-c_iאת הסיווג שניתן לדגימה ה-iעל ידי מסווג המטרה שלנו.נסמן ב-min_j, max_j את הערכים המינימליים והמקסימליים של התכונה ה-j בקבוצת ה-train.כאשר נבצע נרמול על קבוצת ה-train נקבל את הקבוצה הבאה: $\{(c_i, \frac{v_{i,j} - min_j}{max_j - min_j}) \mid i \text{ is sample in train}, \text{and } j \text{ is feature in train}\}$

כעת נניח כי בעץ המקורי קיבלנו שתכונה f כלשהי עם ערך חלוקה v מקבלת ערך ה-IG המקסימלי. לכן נקבל חלוקה של הצומת לשני בנים לפי f , v . החלוקה לשני בנים תיהיה זהה גם בעץ המנורמל בגלל שמההנחה שלנו אם בה"כ מתקיים עבור f , v ש- $x_{i,f} < v$ולכן מתקיים:$\frac{x_{i,f} - min_f}{max_f - min_f} < \frac{v - min_f}{max_f - min_f}$,ולכן נקבל שה-$IG$ יהיה מקסימלי גם בעץ המנורמל, ובכך החלוקה של צומת מסויים לשני בנים בעץ החדש (מבחינת הדגימות שיפוצלו) תיהיה זהה לחלוקה בעץ לפני הנירמול.מכאן נוכל להסיק ישירות שהעץ המנורמל יהיה זהה לעץ המקורי, בכל הפיצולים שלו לבנים ולפי אותו $feature$ חלוקה בכל פיצול לבנים. בכך נקבל גם שהדיוק של המסווג על קבוצת ה-train יהיה זהה לפני ואחרי הנירמול כי מדובר בפיצולים זהים בשני העצים. בהינתן איבר מקבוצת ה-test, נרצה להראות ששני העצים יתנו סיווג זהה. בכך נוכח שהדיוק על קבוצת ה-test ההוא זהה בשני העצים. יהי $(c, \frac{v_f - min_f}{max_f - min_f}) \in test$. בהינתן צומת החלטה בעץ המקורי שמפצלת לפי תכונה f וערך v. כעת נניח כי עבור הדגימה $(c, \frac{v_f - min_f}{max_f - min_f})$ בעץ המקורי לפני הנרמול, בצומת המדובר היה מתקיים: $v_f < v$. במקרה זה גם מתקיים: $\frac{v_f - min_f}{max_f - min_f} < \frac{v - min_f}{max_f - min_f}$. כלומר עבור כל צומת פיצול בעץ המקורי, אותה הצומת בעץ המנורמל תבצע את אותה הבחירה בצומת הבן להמשיך אליה כמו בעץ המקורי. בגלל שהבחירה שלנו של צומת בעץ הייתה שרירותית. למעשה הראנו שעבור כל צומת שנבחר בשני העצים (הראנו שהעצים זהים מבחינת צמתים), תתקבל בדיוק אותה ההחלטה עבור דגימה כלשהי מקבוצת ה-test. לכן נקבל את אותו הסיווג על קבוצת ה-test, ובכך נקבל את אותו הדיוק בשני העצים על קבוצת ה-test.

■

שאלה 3

3. 1.



3. 2.

חשיבות הגיזום היא הגבלת התאמת יתר בכך נתעלם מדגימות שגורמות לרעש. הגיזום נעשה ע״י הסתכלות על כל קבוצת מדגם, ואם יש בה לכל היותר m דגימות יוצאות דופן משאר הקבוצה, נהפוך את הקבוצה לעלה המתאים לסיווג תת הקבוצה הגדולה. בכך אנו מעלים את שגיאת האימון אבל מקטינים את שגיאת המבחן.

3. 3.

עבור $m \geq 25$ לא היה שינוי בדיוק כיוון שהגיזום על הקבוצות היה גס מידי ולכן מבצע הכללה גסה על כל הדגימות מה שפוגע בדיוק. עבור $m \leq 25$ ניתן לראות שהגיזום עבור m קטן יותר נתן רמות דיוק שונות, כלומר מדוייקות יותר ביחס לגיזום. אך עם זאת ניתן לראות התכנסות עבור m גדול יותר מה שאומר שמתבצעת פחות התאמת יתר. נציין כי הגרף מבוסס גם על סדר החלוקה ל-k קבוצות לפי k fold cross validation. אך המגמתיות של הגרף נשמרת.

3. 4.

הגיזום הטוב ביותר הוא עבור m = 2 , והוא טוב יותר מריצה ללא גיזום:

Without pruning, precision: 0.9069398907103826

m=2, precision: 0.9269398907103825

4. 1.

הלוס הממוצע שלנו עבור גיזום עם m=2 הוא 17.0. מטה מצורפת טבלה המפרטת עבור כל איטרציה של K-Fold כמה טעויות של FN ו-FP היו:

K FOLD ITERATION	1	2	3	4	5
False Negative	2	3	1	1	2
False Positive	3	3	0	6	1
Average loss: 17.0					

4. 2.

אם כל הדוגמאות שלנו מסווגים כאנשים חולים, אנו נקבל בקבוצת ה-train רק אנשים אנשים חולים. אז האלגוריתם שלנו יבנה עץ, שתמיד יסווג כל אדם כחולה. במצב זה לא יתכן מצב שנקבל FALSE NEGATIVE מכיוון שעל כל דגימה שהעץ יצטרך לבדוק הוא תמיד יחזיר שהאדם חולה. הממוצע על פני 5 ריצות **k-fold** הוא 30.8, נשים לב שבקובץ המידע הנתון לנו, מתקיים ש 154 אנשים בריאים. ולמעשה בהרצת k-fold כל האנשים הבריאים חולקו לקבוצות כך שכל אחד מהם נכח פעם אחת בלבד בקבוצת ה-test. כל אחד מהם יתרום FP אחד לחישוב ה-loss ולכן נוכל לצפות את תוצאות הריצה ($\frac{154}{5} = 30.8$). בהשוואה לתוצאה הקודמת, מספר ה- False Positive עלה במקרה זה לעומת הסעיף הקודם כי כל מי שבריא סווג כחולה.

K-fold ITERATION	1	2	3	4	5
False Negative	0	0	0	0	0
False Positive	26	31	32	35	30
Average loss: 30.8					

4. 3. ביצענו מחקר ובדיקות על אפשרויות שונות שיכולות לעזור לבעיה שלנו וחלטנו שני מקרים שנרצה לחקור:

<u>מקרה א:</u> כאשר צומת הוא הומוגני מבחינת התיוג של הדגימות בו, אנו נגדיר את הצומת כעלה. זה למעשה תנאי מאוד נוקשה, נבדוק אם ניתן להחליש מעט תנאי זה. בכך נקבל עץ קטן יותר, שפחות נוטה להתאמת יתר.

<u>מקרה ב:</u> ננסה למצוא hyper-parameter, נבחר בתור פרמטר את עומק העץ ונראה כיצד מגבלה על עומק העץ משפיעה לנו על הפונקציה של ה-loss.

<u>מקרה א:</u>

לפי אלגוריתם 3ID, כאשר אנו מגיעים לצומת בעץ אשר בו כל הדגימות אחידות מבחינת התיוג שלהן אנו ניצור עלה אשר יחזיר את התיוג המתאים לקבוצת הדגימות. אך יש מקרים שבהם בכל זאת ניתן לבצע הכללה על צומת מסויים בעץ, זאת בהתאם יחס סדר הגודל בין התיוגים החולים לבריאים. נציע את השיפור הבא:

בנוסף לבדיקה של אחידות הדגימות נבצע:

$$if: \ sick\ subjects\ /healthy\ subjects\ \geq 10\ and\ healthy\ subjects\ \leq 7\ ;\ then\ classify\ =\ SICK$$
$$if: \ healthy\ subjects/sick\ subjects\ \geq 40\ and\ healthy\ subjects\ \leq 2\ ;\ then\ classify\ =\ HEALTHY$$

פירוש התנאי הוא שנבצע הכללה של צומת כאשר:

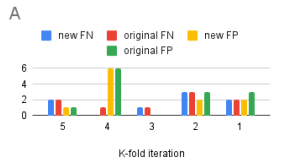
- מתקיים שכמות החולים גדולה לפחות פי 10 מהבריאים. כאשר יש לכל היותר 7 בריאים.
- מתקיים שכמות הבריאים גדולה לפחות פי 40 מהחולים. כאשר יש לכל היותר 2 חולים.

היתרון בשיפור הן שהעץ יבצע הכללה על קבוצת ה-train שלנו כלומר ננסה להתעלם כמה שיותר מ-"רעש" בדגימות, וגם נבצע כמה שפחות התאמת יתר לעץ הנתון.
המספרים הקבועים שנבחרו להשוואה מתאימים מצד אחד לסדר גודל של קבוצת ה-train שקיבלנו. ומצד שני למחירים של פונקצית ה-loss. להלן טבלה המתארת את מספר הניסויים שביצענו על הקבועים השונים:
בגרפים המצורפים מצויינים גם הנתונים של ה-FN ו-FP לפני השיפור (original) וגם לאחריו (new).

A:**Average loss: 15.0**

$sick\ subjects\ /healthy\ subjects\ \geq 10\ and\ healthy\ subjects\ \leq 7$

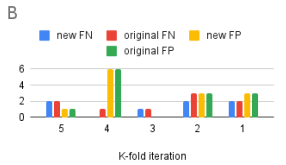
$healthy\ subjects/sick\ subjects\ \geq 40\ and\ healthy\ subjects\ \leq 2$



B:**Average loss: 13.8**

$sick\ subjects\ /healthy\ subjects\ \geq 10\ and\ healthy\ subjects\ \leq 7$

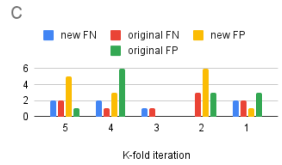
$healthy\ subjects/sick\ subjects\ \geq 60\ and\ healthy\ subjects\ \leq 3$



C:**Average loss: 14.2**

$sick\ subjects\ /healthy\ subjects\ \geq 10\ and\ healthy\ subjects\ \leq 8$

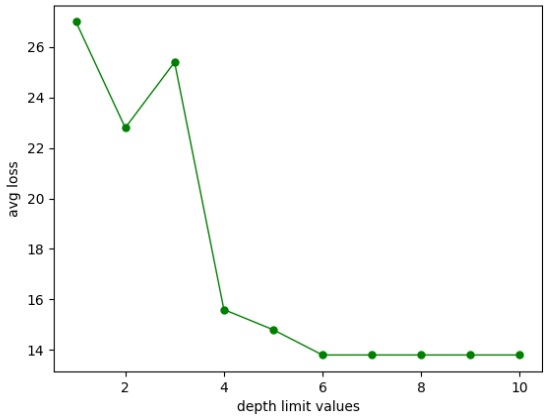
$healthy\ subjects/sick\ subjects\ \geq 60\ and\ healthy\ subjects\ \leq 3$



כל שלושת הניסיונות הנ"ל שיפרו את ה-loss הממוצע שקיבלנו בסעיף 1. לכן נבחר את את הקבועים שממזערים את ה-loss . לכן נבחר את הקבועים שנמצאים בניסוי B. בהשוואה למקרה שלא בוצע השיפור הצלחנו להוריד את כמות ה-FN ולכן שיפרנו את ה-loss כי לאלו משקל גדול יותר.

<u>חלק ב':</u>

מטרה נוספת שלנו היא למצוא פרמטרים שאינם קשורים באופן ישיר ל-DATA הנתון לנו, אשר ביכולתם לשפר את הביצועים של העץ שניצור. נבחר בתור hyper-parameter מגבלה על עומק העץ. לפרמטר זה יש פוטנציאל להכליל לנו את העץ ולמנוע התאמת יתר (זאת בנוסף לשיפור שביצענו בחלק א'):



בגרף ניתן לראות את העומק הממוצע (הממוצע הוא על כל ריצה של k-fold) כפונקציה של הגבלה על עומק העץ. ככל שמגבלת העומק היא קטנה יותר באופן כללי קיבלנו loss גבוהה יותר. כלומר האלגוריתם סיפק לנו מסווג שפחות התחשב ב-train וניסה יותר להכליל. נשים לב שהחל מעומק 6 אנו מקבלים בדיוק את ה-loss שקיבלנו בחלק א' עם השיפור B. זה נובע מהעובדה שקבוצת ה-train לא מאוד גדולה, ולמעשה כבר אנו מבצעים גיזומים בעץ. בנוסף על כך ניתן לראות שהגבלות העומק 4 ו-5 גם כן משפרות את ה-loss הממוצע (ערכי ה-loss בהתאמה: 15.6 ו - 14.8) בהתאם לכך אנו יכולים להניח שאם נגביל את העומק של העץ שלנו ל-5 אז נרוויח שני דברים:

1. נפחית את התאמת היתר של המסווג שלנו לקבוצת ה-train.
2. עדיין נשפר את ה-loss הממוצע שלנו בהשוואה ל-3ID הרגיל.

לכן אנו נכניס גם מגבלת עומק על העץ שלנו (עומק 5).