מבוא לבינה מלאכותית

תרגיל בית 3

מגיש: תובל גלוון

ת.ז.: 312419971

שאלה 1

1.2. תוצאת הדיוק של האלגוריתם id3 -

```
0.9469026548672567
```

שאלה 2

הוכחה:

בהינתן דאטה כלשהו עם תיוגים בינאריים ותכונות רציפות נגדיר:
E_{test}- דוגמאות קבוצת המבחן. E_{train}- דוגמאות קבוצת האימון. $Features$- קבוצת התכונות הרציפות.

פונקציית נרמול MinMax:

$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

כאשר $e_k, e_j \in E_{train}$. ועבור תכונה מסויימת $f \in Features$, מתקיים $f(e_k) = X_{min}, f(e_j) = X_{max}$ כלומר, זהו הערך המינימלי/מקסימלי עבור תכונה זו מכל דוגמאות האימון. ו-X זהו הערך של הדוגמא אותה אנו רוצים לנרמל עבור תכונה f.
נסמן בt_f את ערך הסף של התכונה f, ממוצע שני ערכים כלשהם אשר נבחר באלגוריתם. כאשר:

$$t_f = \frac{f(e_i) + f(e_{i+1})}{2} = \frac{x_i + x_{i+1}}{2}$$

תהא $e \in E_{test}/E_{train}$. נניח כי בעץ המתקבל ע"י האלגוריתם ID3 לפני הפעלת הנרמול, הערך עבור הדוגמא e הייתה מעל ערך הסף, כלומר $x = f(e) \geq t_f$. (*) בדומה עבור ערך קטן מהסף. $x < t_f$. (**)

נפעיל את הנרמול על הדוגמאות עבור תכונה f. ונקבל:

$$t_{f\,changed} = \frac{x_{i\,changed} + x_{i+1\,changed}}{2} = \frac{\frac{x_i - X_{min}}{X_{max} - X_{min}} + \frac{x_{i+1} - X_{min}}{X_{max} - X_{min}}}{2} = \frac{x_i + x_{i+1} - 2X_{min}}{2 \cdot (X_{max} - X_{min})}$$

$$x_{changed} = \frac{x - X_{min}}{X_{max} - X_{min}} \underbrace{\geq}_{(*)} \frac{\frac{x_i + x_{i+1}}{2} - X_{min}}{X_{max} - X_{min}} = \frac{x_i + x_{i+1} - 2X_{min}}{2 \cdot (X_{max} - X_{min})} = t_{f\,changed}$$

$$x_{changed} = \frac{x - X_{min}}{X_{max} - X_{min}} \underbrace{\leq}_{(**)} \frac{\frac{x_i + x_{i+1}}{2} - X_{min}}{X_{max} - X_{min}} = \frac{x_i + x_{i+1} - 2X_{min}}{2 \cdot (X_{max} - X_{min})} = t_{f\,changed}$$

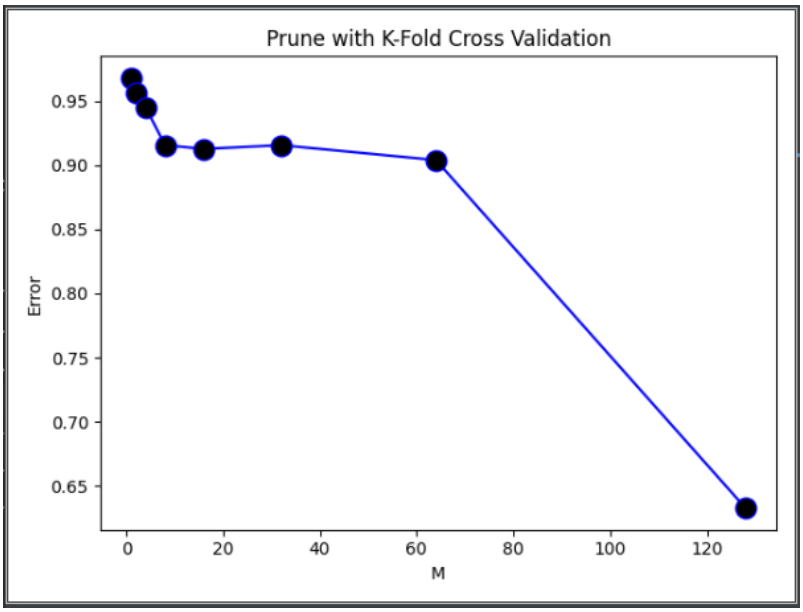
עבור כל תכונה נקבל כי אם ערך התכונה עבור דוגמא כלשהי היה מעל או מתחת לסף הוא יישאר כך לאחר הנרמול ולכן בכל שלב באלגוריתם נקבל פיצול זהה ולכן גם התכונה שתיבחר לפי הig תישאר זהה באימון והסיווג אשר נקבע על פי ערך הסף יישאר זהה עבור קבוצת המבחן.

שאלה 3

3.1. חשיבות הגיזום הוא הקטנת העץ על ידי הורדת ענפים שאינם תורמים לשיפור תוצאת האלגוריתם, ובכך לחסוך במשאבים ובזמן חיפוש. התופעה שאנו באים למנוע/להחליש היא תופעה של התאמת יתר. בכך שנגזום את העץ אנו נגדיל את שגיאת האימון, בתקווה להקטין את שגיאת המבחן, בכך שלא נתייחס לענפים שהתקבלו מדוגמאות רועשות.

3.3. i. $M = [1,2,4,8,16,32,64,128]$.

iii. הגרף שהתקבל-



iv. בציר x אנו רואים את הערך של המספר המינימלי בעלה, M. ובציר y אנו רואים את **הדיוק הממוצע** שחושב על ידי kfold אשר **שווה ל1 כאשר קבוצת המבחן לא טעתה כלל**. קיבלנו גרף עם שיפוע יורד. כלומר ככל שעשינו גיזום עבור ערך מינימלי של ילדים גדול יותר, למעשה ירד הדיוק הממוצע. כלומר שגינו ביותר ערכים בכך שגזמנו יותר את העץ. לכן ככל הנראה לא היה רעש רציני בדוגמאות שהעלה לנו את שגיאת המבחן. התוצאה הטובה ביותר היא עבור M=1 וערכה הוא: 0.9679454390451834. (על קבוצת האימון עם cv)

3.4. הדיוק שיצא עבור כל קבוצת המבחן עם הגיזום עבור M=1 הוא : 0.9469026548672567. אותו דיוק שהתקבל בשאלה 1 ללא גיזום. תוצאה זו הגיונית כיוון שיצא לנו שאנו נגזום רק אם מספר הדוגמאות קטן מ1 שזה שקול למצב ללא גיזום כיוון שבערך זה בהכרח הדוגמא עקבית ולכן לא נמשיך את העץ גם ללא גיזום.

שאלה 4

4.1. הloss שיצא עם הגיזום עבור M=1 הוא:

$$loss = 0.021238938053097345$$

זהו אותו ערך גם ללא גיזום. שוב, תוצאה זו הגיונית מאותה סיבה שהתקבל עץ זהה בשני המקרים (כי גזמנו רק אם מספר הילדים קטן מ-1 שזה שקול לזה שלא גזמנו כלל).

4.2. על מנת לגרום לאלגוריתם ללמוד מסווג טוב יותר נשתמש **בכיוון פרמטרים ובשיפור מדד האנטרופיה**.

כיוון פרמטרים :

ניתן לעץ פרמטרים שונים אשר יעזרו לו להחליט איך לגזום את העץ על מנת למקסם את ערך הloss. הרעיון הוא בכך שנוריד ענפים בעץ אשר גרמו לשגיאה בחיזוי נוריד את ערך הFP/FN בחישוב הloss ובכך הוא יקטן.

פרמטרים שבחרתי:

מספר מינימלי של דוגמאות בעלה- נשנה את מספר הבנים המינימלי בעלה.

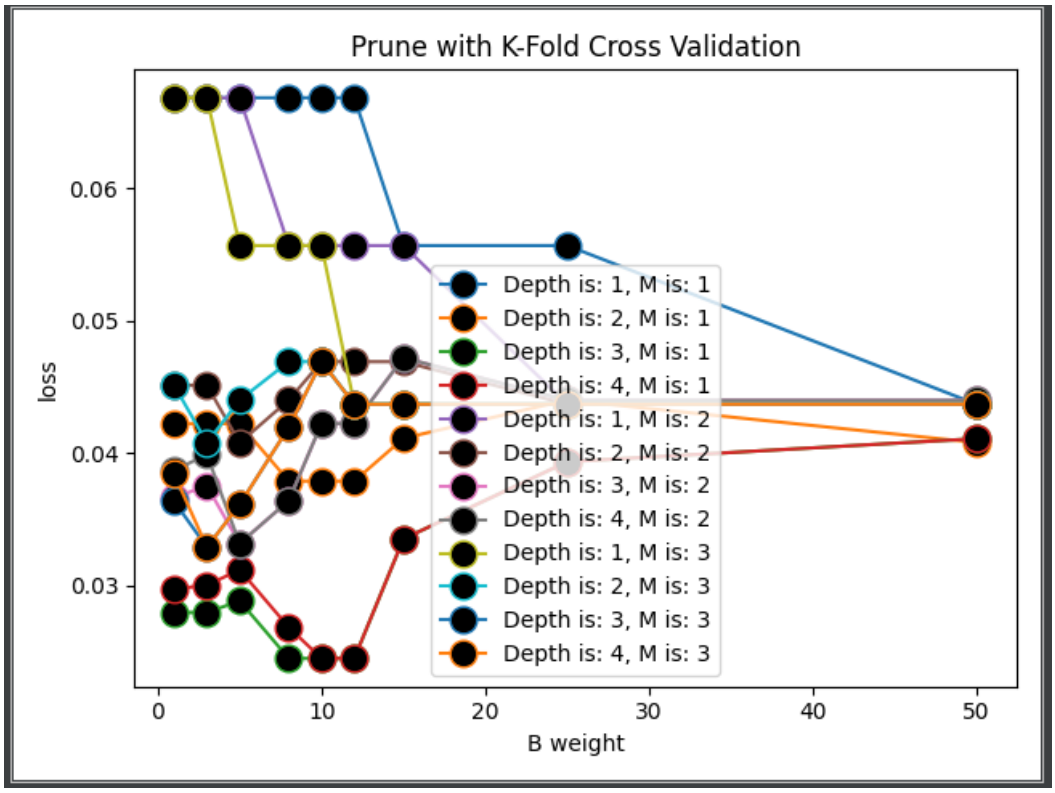
עומק העץ- נגביל את עומק העץ לערכים שונים.

ניתן משקל לזיהוי בריא- ניתן משקל גדול יותר לכל צומת שהרוב בה הוא בריא על מנת שיהיה סיכוי קטן יותר לקבלת false positive שזוהי הטעות החמורה יותר. ככל שהמשקל גדול יותר, נגזום קודם ענפים עם רוב בריא.

שיפור אנטרופיה:

בכך שניתן מדד טוב יותר לחולים מאשר בריאים נקטין את הסיכוי שנקבל FP, ובכך מדד הloss ישתפר.

4.3 לאחר **כיוון הפרמטרים** שתיארתי ב4.2 עם cross validation קיבלנו את הגרף הבא:



אנו רואים את המשקל לזיהוי בריא בציר x עבור ערכי עומק וגיזום ילדים מינימלי שונים. קיבלנו כי התוצאה הטובה ביותר היא עבור הערכים:

$$M = 1, depth = 3, B = 8$$

אנטרופיה:

שיניתי את מדד האנטרופיה כך שייתן משקל קטן פי 10 לחולה שסווג כבריא וכך נשאף למזער את FP מול FN בהתאם ליחס המשקלים בחישובם.

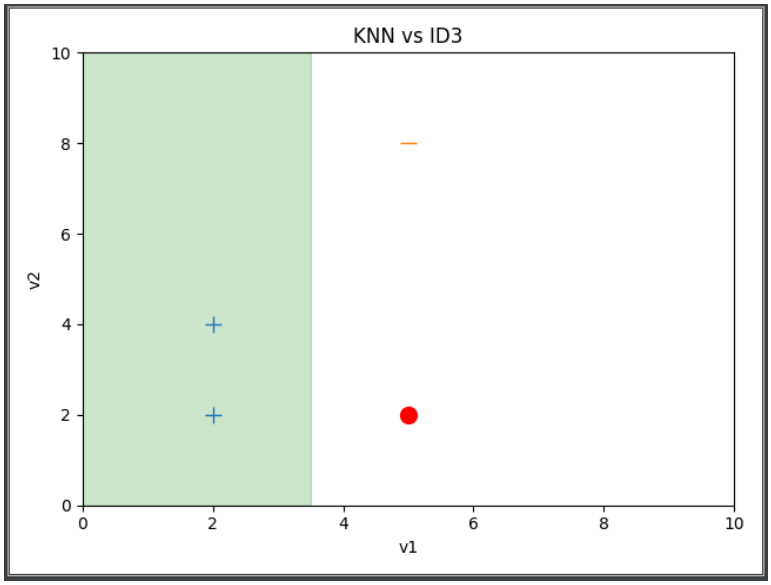
ערך הLosss החדש שקיבלתי לאחר השיפורים הוא:

```
0.001769911504424779
```

שאלה 5

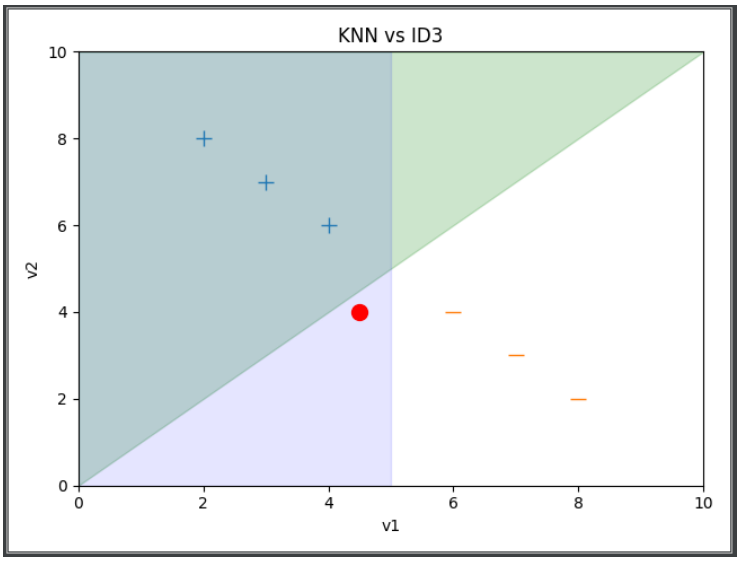
מסווג המטרה יסומן על ידי הרקע כאשר ירוק זהו סיווג חיובי. מסווג ID3 יסומן ברקע כחול בהיר כחיובי.

סעיף א'-



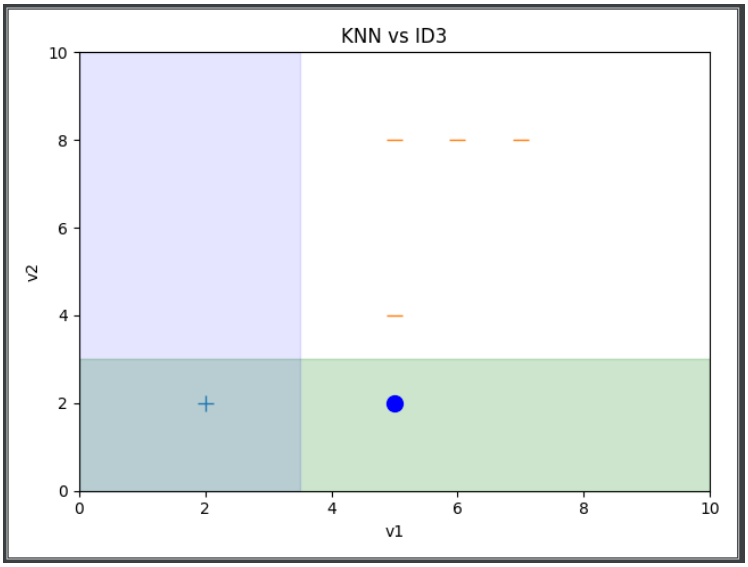
ID3 ייתן סיווג חיובי באזור הירוק ובלבן שלילי כי יפריד לפי v1. עבור הדוגמה השלילית בעיגול נקבל בID3 סיווג שלילי ובKNN נקבל סיווג חיובי כיוון שהוא בודק לפי מרחק מהדוגמאות ותמיד יהיו יותר דוגמאות חיוביות קרובות אליו לכל K.

סעיף ב'-



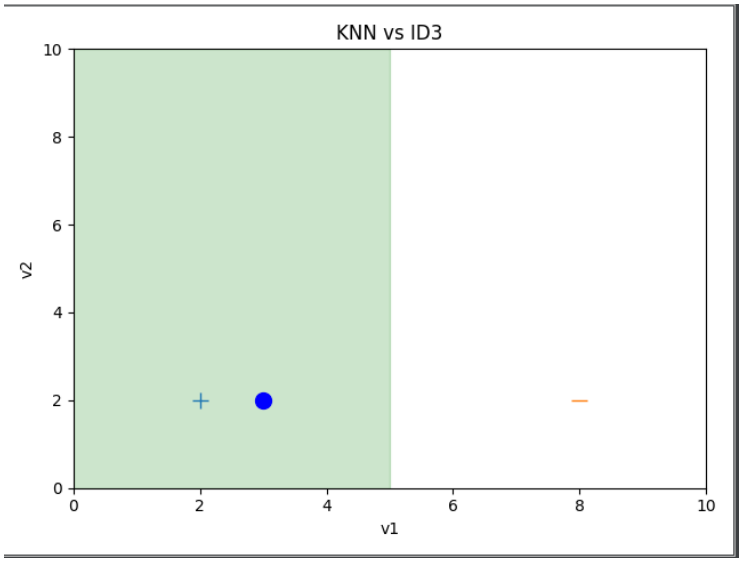
ID3 יפריד לפי v1 ולכן לא יתקבל מסווג המטרה והוא יסווג את הדוגמה השלילית כחיובית. עבור KNN עם K=1 נקבל את מסווג המטרה כיוון שיראה את הדוגמה הקרובה ביותר אליו, חיובי מעל האלכסון ושלילי מתחת.

סעיף ג'-



המסווג של ID3 מסומן בכחול. עבור הדוגמא החיובית בעיגול נקבל עם KNN וK=1 סיווג שלילי בגלל הדוגמה השלילית שקרובה אליו ונטעה. עבור ID3 גם נסווג אותו כשלילי כי הפרדנו לפי v1 במקום לפי v2.

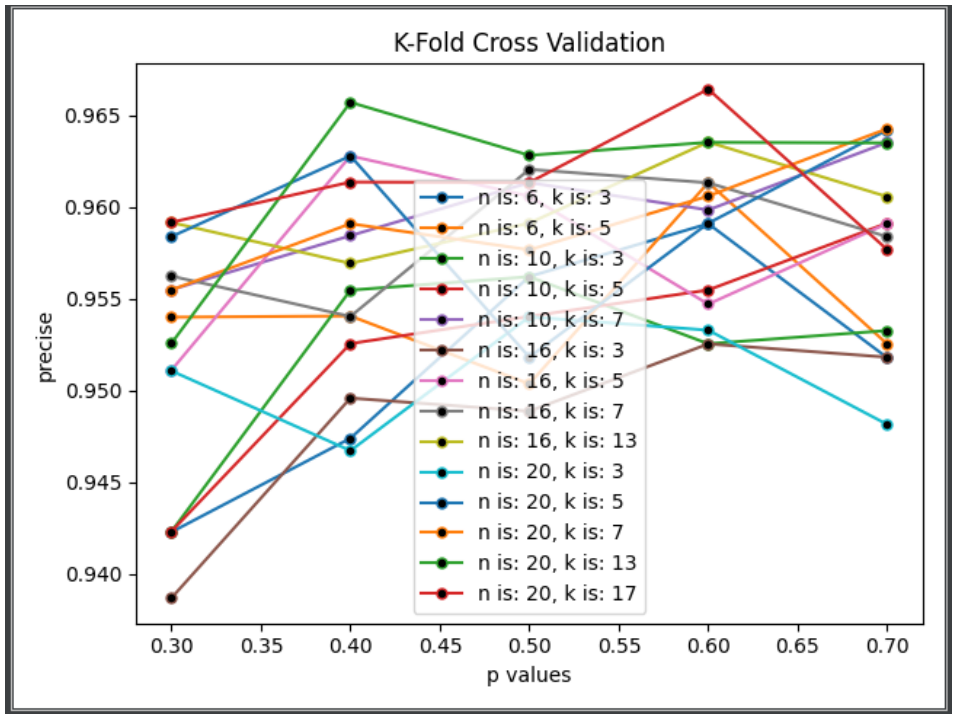
סעיף ד'-



עבור ID3 וגם KNN עם K=1 נקבל את מסווג המטרה כיוון שID3 יפריד לפי v1 וKNN יחפש את הדוגמה הקרובה ביותר.

שאלה 6

6.1. על מנת למצוא דיוק מקסימלי ביצעתי ניסויים על הפרמטרים p,K,N בעזרת cross validation על קבוצת האימון + קבוצת המבחן להגדלת הדאטה עליו אני מסיק מסקנות ועבור כל קבוצת ערכים וKfold מסוים ביצעתי שלושה ניסויים ולקחתי את השגיאה הממוצעת עליהם. (למניעת השגיאה שנובעת מrandom) תוצאות הניסוי שהתקבלו:



אנו רואים בציר x את ערכי p מול הדיוק בציר y. כל ערך של דיוק מבדקנו מול ערכי n וk שונים.

לאחר כיוון פרמטרים אנו רואים כי הדיוק המקסימלי מתקבל כאשר:

$$p = 0.4, N = 20, K = 13$$

הדיוק האופטימלי על קבוצת המבחן הוא:

```
accuracy is: 0.9823008849557522
```

הדיוק הממוצע על 5 הרצות של קבוצת המבחן הוא:

```
mean accuracy is: 0.9628318584070795
```

שאלה 7

7.1. נציע שיפור לאלגוריתם KNN: האלגוריתם אינו מתחשב במרחק היחסי מבין כל אחד מK העצים שהוא בוחר, לכן מתוך הנחה כי אם המרחק קרוב יותר התוצאה מדויקת יותר ניתן משקל יחסי למרחק ובכך נקבל את החלטת הרוב אשר קרוב לאותה דוגמא, זה מונע מאיתנו בעצם לבחור בדוגמאות רחוקות אשר פוגעות בדיוק האלגוריתם גם אם הן חלק מהוועדה. את השינוי נבצע במקום בחירת ה majority בפונקציית predict. את **פונקציית המשקל** נגדיר כך:

$$dist\_weight_i = \frac{1}{distance_i}, 1 \le i \le k$$

כך נקבל משקל גדול יותר למי שמרחקו קרוב.

בנוסף כדי לקבל השפעה גדולה יותר של המרחקים מול הרוב נוסיף פרמטר בונוס לפי קרבה.

על מנת לחשב את הסיווג הטוב השתמשתי בנוסחה הבאה: כאשר $dist\_weight$ ממוין מהרחוק ביותר לקרוב.

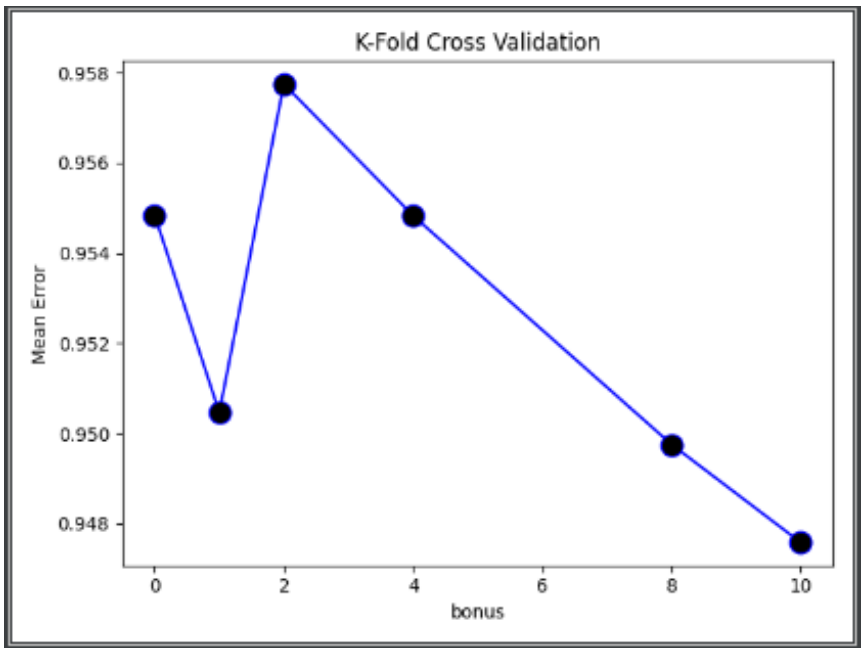
$$B_{weight} = \sum_{i=1}^{k} \mathbb{I}_{B\_predict} \cdot (dist\_weight_i \cdot i \cdot bonus + 1)$$

$$M_{weight} = \sum_{i=1}^{k} \mathbb{I}_{M\_predict} \cdot (dist\_weight_i \cdot i \cdot bonus + 1)$$

$$return\ B_{weight} \geq M_{weight} ?\ B : M$$

עבור בונוס 0 נקבל את השפעת הרוב בדיוק כמו בסעיף 6 ללא בונוס וככל שנגדיל את הבונוס נקבל השפעה גדולה יותר של עצים עם מרחק קרוב. בנוסף נשתמש ב**נרמול** מינימום מקסימום על מנת לקבל מרחק שאינו מושפע מסקלת הנתונים.

7.2. לאחר שינוי הסיווג לפי המרחק השתמשתי באותם פרמטרים עבור p, K, N משאלה 6.
על מנת לדעת מהו ערך הבונוס הנותן את השיפור המקסימלי ביצעתי כוון פרמטר באמצעות cross validation. תוצאות הניסוי עבור דגימה של 4 ניסויים על כל ערך ולקיחת ממוצע על הדיוק (סה"כ 20 ניסויים לכל ערך עם kfold) הם:



ניתן לראות כי עבור $bonus = 2$ אנו מקבלים דיוק גבוה יותר מ$bonus = 0$ אשר מייצג את החלטת הרוב ללא השיפור, כך ניתן לראות בבירור את השיפור המתקבל.

הדיוק האופטימלי על קבוצת המבחן הוא:

```
accuracy is: 0.9911504424778761
```

הדיוק הממוצע על 5 הרצות של קבוצת המבחן הוא:

```
mean accuracy is: 0.9805309734513272
```