תרגיל בית 3 –מבוא ללמידה

עברו על כלל ההנחיות לפני תחילת התרגיל.

הנחיות כלליות:

- תאריך ההגשה: עד לסוף מועדי א' - **17/05/2024** ב23:59
- את המטלה יש להגיש **בזוגות בלבד.**
- יש להגיש מטלות מוקלדות בלבד. פתרונות בכתב יד לא ייבדקו.
- ניתן לשלוח שאלות בנוגע לתרגיל בפיאצה בלבד.
- המתרגל האחראי על תרגיל זה: **דניאל אלגריסי.**
- בקשות דחיה מוצדקות (מילואים, אשפוז וכו') יש לשלוח למתרגל האחראי (**ספיר טובול**) בלבד.
- במהלך התרגיל ייתכן שנעלה עדכונים, למסמך הנ"ל – תפורסם הודעה בהתאם.
- העדכונים הינם מחייבים, ועליכם להתעדכן עד מועד הגשת התרגיל.
- שימו לב, העתקות תטופלנה בחומרה.
- התשובות לסעיפים בהם מופיע הסימון ✍️ צריכים להופיע בדוח.
- לחלק הרטוב מסופק שלד של הקוד.
- אנחנו קשובים לפניות שלכם במהלך התרגיל ומעדכנים את המסמך הזה בהתאם. גרסאות עדכניות של המסמך יועלו לאתר. הבהרות ועדכונים שנוספים אחרי הפרסום הראשוני יסומנו כאן בצהוב. ייתכן שתפורסמנה גרסאות רבות – אל תיבהלו מכך. השינויים בכל גרסה יכולים להיות קטנים.

לצורך הנוחות:

הבהרות ועדכונים גרסה ראשונה סומנו ככה.

הבהרות ועדכונים גרסה שניה סומנו ככה.

הבהרות ועדכונים גרסה שלישית סומנו ככה.

שימו לב שאתם משתמשים רק בספריות הפייתון המאושרות בתרגיל (מצוינות בתחילת כל חלק רטוב)
לא יתקבל קוד עם ספריות נוספות

מומלץ לחזור על שקפי ההרצאות והתרגולים הרלוונטיים לפני תחילת העבודה על התרגיל.

חלק ב' - מבוא ללמידה (56 נק')

✍️חלק א' – חלק היבש (28 נק')

kNN – נעים להכיר

בחלק זה תכירו אלגוריתם למידה בשם kNN, או בשמו המלא k-Nearest Neighbors, כאשר ה־k הוא למעשה פרמטר!
יהי סט אימון עם n־דוגמות, $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$, כאשר $\forall i: x_i \in \mathbb{R}^d, y_i \in \mathcal{Y}$.
כלומר הדוגמות הינן וקטורים d־ממדיים והתגיות הינן מדומיין כלשהו, הבעיה היא בעיית קלסיפיקציה (סיווג).
אם לא נאמר אחרת, הקלסיפיקציה תהיה בינארית, כלומר $\mathcal{Y} = \{-, +\}$.
עבור כל דוגמה בסט האימון, ניתן להסתכל על הכניסה ה־i בווקטור כעל הfeature ה־i של הדוגמה, קרי כל דוגמה x_i מיוצגת על ידי d־ערכים: $f_1(x_i), f_2(x_i), \dots, f_d(x_i)$.
תהליך ה"אימון" של האלגוריתם הוא טריוויאלי – פשוט שומרים את סט האימון במלואו.
תהליך הסיווג הוא גם פשוט למדי – כאשר רוצים לסווג דוגמה מסט המבחן מסתכלים על k השכנים הקרובים ביותר שלה במישור הd־ממדי מבין הדוגמות בסט האימון, ומסווגים את הדוגמה על פי הסיווג הנפוץ ביותר בקרב k השכנים.
על מנת להימנע משוויון בין הסיווגים, נניח בדרך כלל כי k־אי זוגי, או שנגדיר היטב שובר שוויון.
אם לא נאמר אחרת, במקרה של שוויון בקלסיפיקציה בינארית, נסווג את הדוגמה כחיובית $+$.

שאלות הבנה

א. (3 נק') כאמור, בתהליך הסיווג אנו בוחרים עבור הדוגמה את הסיווג הנפוץ ביותר של k השכנים הקרובים ביותר, אולם עלינו להגדיר את פונקציית המרחק עבור קביעת סט שכנים זה.
שתי פונקציות מרחק נפוצות הינן מרחק אוקלידי ומרחק מנהטן.
(1 עבור איזה ערכים של d, k נקבל שאין תלות בבחירה בין פונקציות המרחק הנתונות ~~בבחירה פונקציית המרחק~~? (נמקו)

עבור k=1 וכל d. כאשר יש פיצ'ר אחד, מתקיים כי מרחק מנהטן שווה למרחק האוקלידי (המישור חד מימדי). נוכיח זאת:
יהיו שתי נקודות במרחב- a = (x_1), b = (x_2)
מתקיים: $manhattan\ distance = \sqrt{(x1 - x2)^2} = |x1 - x2| = euclidean\ distance$

נראה באינדוקציה כי אין תלות בבחירה בין שתי פונקציות המרחק עבור כל k:
בסיס- k=1: הסיווג יהיה לפי השכן הקרוב ביותר. השכן הקרוב ביותר לפי מרחק מנהטן הוא גם השכן הקרוב ביותר לפי מרחק אוקלידי, לכן הסיווג יהיה זהה כאשר נשתמש בכל אחת מפונקציות המרחק.
צעד- נניח שנכון עבור k, ונראה שנכון עבור k+1: מתקיים כי הסיווג לפי k שכנים זהה לפי כל אחת מפונקציות המרחק, ונבחרו אותם k שכנים קרובים ביותר. נסיר k שכנים אלה מהגרף, וכעת נרצה למצוא את הדוגמה הקרובה ביותר ולהוסיף אותה לקבוצת השכנים. דוגמה זו תהיה זהה לפי מרחק

מנהטן ולפי מרחק אוקלידי, כלומר לפי שתי הפונקציות קבוצת k+1 השכנים תהיה זהה לכן הסיווג יהיה זהה.

2) עבור בעיית קלסיפיקציה בינארית תנו דוגמה <u>פשוטה</u> לערכי d, k, סט אימון ודוגמת מבחן בה השימוש בכל אחת מפונקציות המרחק הנ"ל משנה את סיווג דוגמה המבחן.

עבור d=2, k=1:
דוגמת המבחן תהיה t = ((0,0), y)
סט האימון יהיה: x_1 = ((0,4), +), x_2 = ((3,2), -)

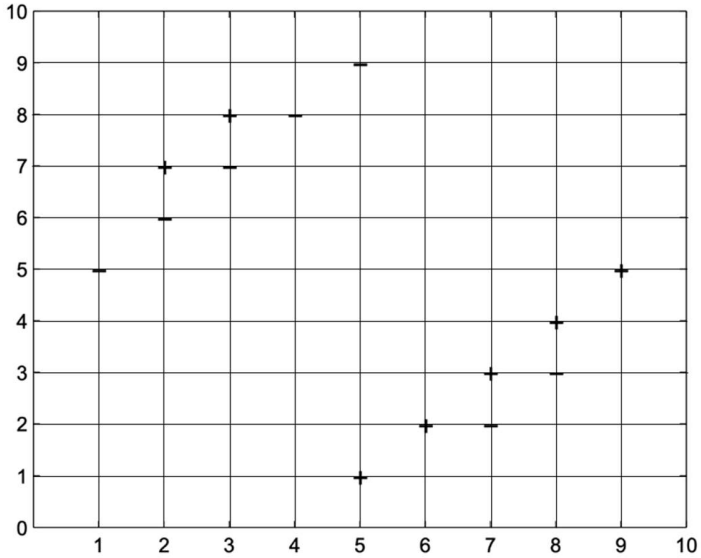
לפי מרחק מנהטן:
$d_{MD}(t, x_1)$ = |0-0|+|4-0| = 4
$d_{MD}(t, x_2)$ = |0-3|+|0-2| = 5
כלומר x_1 הוא השכן הקרוב ביותר, לכן y = +.

לפי מרחק אוקלידי:

$$d(t, x1) = \sqrt{(0-0)^2 + (0-4)^2} = \sqrt{16} = 4$$
$$d(t, x2) = \sqrt{(0-3)^2 + (0-2)^2} = \sqrt{13} = 3.61$$

כלומר x_2 הוא השכן הקרוב ביותר, לכן y = -.

מעתה, אלא אם כן צוין אחרת, נשתמש במרחק אוקלידי
נתונה קבוצת האימון הבאה, כאשר $d = 2$:



3) (1 נק') איזה ערך של k עלינו לבחור על מנת לקבל את הדיוק המרבי **על קבוצת האימון**? מה יהיה ערך זה? (הדוגמא לא יכולה להיות שכנה של עצמה)

עבור k=5 נקבל את הדיוק המירבי, שהוא 10/14. עבור k קטן יותר תהיה השפעה של דוגמאות חריגות (הדוגמאות שבפינות הימנית תחתונה והשמאלית עליונה), ועבור k גדול תהיה השפעה של דוגמאות רחוקות מידי.

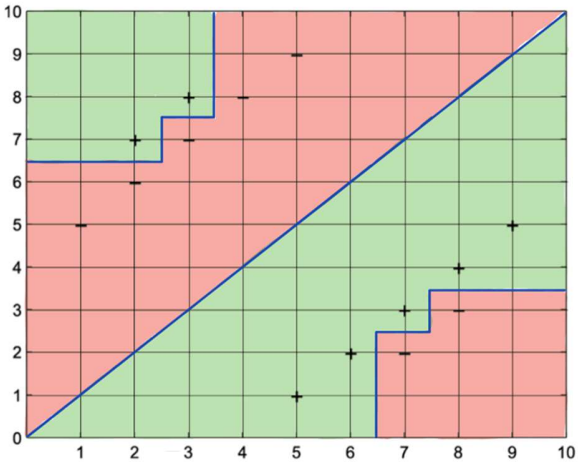
4) (1 נק') עבור איזה ערך של k נקבל מסווג $majority$ של קבוצת האימון? קרי כל דוגמת מבחן תקבל את הסיווג הנפוץ של כלל קבוצת האימון?

עבור ערך k בגודל קבוצת האימון, כלומר בגודל 14. במקרה זה, k השכנים הקרובים יהיו כל קבוצת האימון, לכן הסיווג שייבחר יהיה הסיווג הנפוץ של כלל קבוצת האימון.

5) (2 נק') נמקו מדוע שימוש בערכי k גדולים או קטנים מדי יכול להיות גרוע עבור קבוצת הדגימות הנ"ל.

ערך k קטן מידי יכול לגרום ל-overfitting, שייגרם מדוגמאות רועשות בקבוצת האימון. למשל, במקרה בו הפלוסים בצד שמאל למעלה הם רעשים, נקודות באזור שלהן יסווגו גם באופן שגוי.
ערך k גדול מידי עשוי לגרום ל-underfitting, כלומר הסיווג יהיה מושפע משכנים רחוקים מידי ולכן יתקבל שגוי. עבור ערכי k גדולים מגודל קבוצת האימון נקבל מסווג majority.

6) (2 נק') שרטט את גבול ההחלטה של 1-nearest neighbor עבור הגרף.



השוואה בין מודלי למידה – יש לנמק בקצרה את הפתרונות

1) (3 נק') הציגו מסווג מטרה $f(x): R^2 \to \{0,1\}$ וקבוצת אימון בעלת לכל היותר 10 דוגמאות כך שלמידת עץ ID3 תניב מסווג אשר עונה נכון עבור כל דוגמת מבחן אפשרית (כלומר יתקבל מסווג המטרה), אך למידת KNN תניב מסווג שעבורו קיימת לפחות דוגמת מבחן אחת עליה הוא יטעה, לכל ערך K שייבחר.

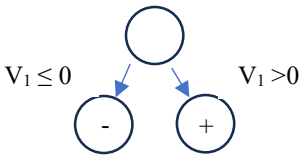
נגדיר את מסווג המטרה:

$$f(v1, v2) = \begin{cases} + & v1 > 0 \\ - & v2 \leq 0 \end{cases}$$

נגדיר את קבוצת האימון:

	V_1	V_2	
X_1	2	2	+
X_2	-2	-2	-
X_3	-3	-4	-

עץ ההחלטה ב-ID3 יהיה:



כלומר התקבל בדיוק מסוג המטרה.

עבור מסווג KNN, ודוגמת המבחן (1, -4), שתי הנקודות הקרובות ביותר הן x_2, x_3, כלומר עבור k=1, k=2 יסווג בתור -, וכאשר k>2 יתקבל מסווג majority לכן יסווג גם בתור -, למרות שהדוגמה אמורה להיות מסווגת בתור +, כלומר KNN יטעה לכל k.

2) (3 נק') הציגו מסווג מטרה $f(x): R^2 \rightarrow \{0,1\}$ וקבוצת אימון בעלת לכל היותר 10 דוגמאות כך שלמידת מסווג KNN עבור ערך K מסוים תניב מסווג אשר עונה נכון עבור כל דוגמת מבחן אפשרית (בלומר יתקבל מסווג המטרה), אך למידת עץ ID3 תניב מסווג אשר עבורו קיימת לפחות דוגמת מבחן אפשרית אחת עליה הוא יטעה.

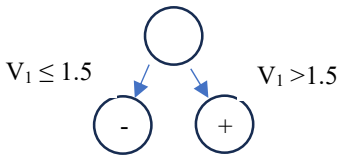
נגדיר את מסווג המטרה:

$$f(v1, v2) = \begin{cases} + & v1 \geq v2 \\ - & else \end{cases}$$

נגדיר את קבוצת האימון:

	V_1	V_2	
X_1	2	1	+
X_2	1	2	-

עץ ההחלטה ב-ID3 יהיה:



לכן ID3 וטעה עבור הדוגמה (2,3) משום שיסווג אותה בתור + למרות שלפי מסווג המטרה היא -.
עבור KNN עם k=1, כאשר $v_1 \geq v_2$ הנקודה תהיה קרובה יותר ל-x_1, ואחרת תהיה קרובה יותר ל-x_2 (גבול ההחלטה הוא הישר y = x, ושתי הנקודות x_1,x_2 במרחק זהה ממנו), לכן כל דוגמת מבחן תסווג נכון.

3) (3 נק') הציגו מסווג מטרה $f(x): R^2 \rightarrow \{0,1\}$ וקבוצת אימון בעלת לכל היותר 10 דוגמאות כך שלמידת מסווג KNN עבור ערך K מסוים תניב מסווג אשר עבורו קיימת לפחות דוגמת מבחן אפשרית אחת עליה הוא יטעה, וגם למידת עץ ID3 תניב מסווג אשר עבורו קיימת לפחות דוגמת מבחן אחת אפשרית עליה הוא יטעה.

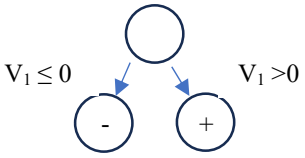
נגדיר את מסווג המטרה:

$$f(v1, v2) = \begin{cases} + & v1 * v2 > 0 \\ - & else \end{cases}$$

נגדיר את קבוצת האימון:

	V_1	V_2	
X_1	1	1	+
X_2	-1	1	-

עץ ההחלטה ב-ID3 יהיה:



כלומר עבור הדוגמה (-1,-1) ID3 יסווג בתור – למרות שצריך להיות +.
גם עבור KNN עם k=1, עבור אותה דוגמה השכן הקרוב יהיה x_2 ולכן גם כאן יסווג בתור -.

4) (3 נק') הציגו מסווג מטרה $f(x): R^2 \rightarrow \{0,1\}$ וקבוצת אימון בעלת לכל היותר 10 דוגמאות כך שלמידת מסווג KNN עבור ערך K מסוים תניב מסווג אשר עונה נכון עבור כל דוגמת מבחן אפשרית (כלומר יתקבל מסווג המטרה), וגם למידת עץ ID3 תניב מסווג עונה נכון עבור כל דוגמת מבחן אפשרית (כלומר יתקבל מסווג המטרה).

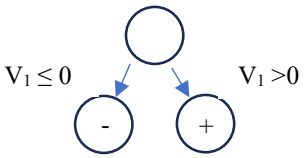
נגדיר את מסווג המטרה:

$$f(v1, v2) = \begin{cases} + & v1 > 0 \\ - & v2 \leq 0 \end{cases}$$

נגדיר את קבוצת האימון:

	V_1	V_2	
X_1	1	1	+
X_2	-1	1	-

עץ ההחלטה ב-ID3 יהיה:



כלומר התקבל בדיוק מסוג המטרה.

גם עבור KNN עם k=1, כאשר $v_1 \geq 0$ יהיה קרוב יותר ל-x_1 ויסווג בתור +, אחרת כאשר $v_1 < 0$ יהיה קרוב יותר ל-x_2 ויסווג בתור -. כלומר כל דוגמה תסווג נכון גם על פי KNN.

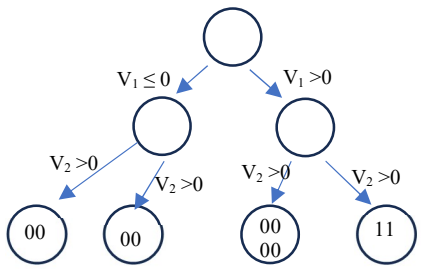
מתפצלים ונהנים

(7 נק') בידוע, בעת סיווג של דוגמת מבחן על ידי עץ החלטה, בכל צומת בעץ אנו מחליטים לאיזה צומת בן להעביר את דוגמת המבחן על ידי ערך סף v שמושווה לfeature של הדוגמה. לפעמים ערך הסף קרוב מאוד לערך הfeature של דוגמת המבחן. היינו רוצים להתחשב בערכים "קרובים" לערך הסף בעת סיווג דוגמת מבחן, ולא לחרוץ את גורלה של הדוגמה לתת־עץ אחד בלבד; לצורך כך נציג את האלגוריתם הבא:

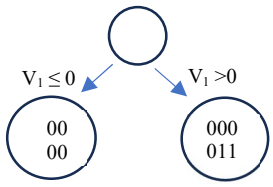
יהיו עץ החלטה T, דוגמת מבחן $x \in \mathbb{R}^d$, ווקטור $\varepsilon \in \mathbb{R}^d$ המקיים $\forall i \in [1,d]: \varepsilon_i > 0$.
כלל אפסילון־החלטה שונה מכלל ההחלטה הרגיל שנלמד בכיתה באופן הבא:
נניח שמגיעים לצומת בעץ המפצל לפי ערכי התכונה i, עם ערך הסף v_i.
אם מתקיים $|x_i - v_i| \le \varepsilon_i$ אזי ממשיכים **בשני** המסלולים היוצאים מצומת זה, ואחרת ממשיכי לבן המתאים בדומה לכלל ההחלטה הרגיל. לבסוף, מסווגים את הדוגמה x בהתאם לסיווג הנפוץ ביותר של הדוגמאות הנמצאות בכל העלים אליהם הגענו במהלך הסיור על העץ (במקרה של שוויון – הסיווג ייקבע להיות $True$).

יהא T עץ החלטה לא גזום, ויהא T' העץ המתקבל מ־T באמצעות גיזום מאוחר שבו הוסרה הרמה התחתונה של T (כלומר כל הדוגמות השייכות לזוג עלים אחים הועברו לצומת האב שלהם).
הוכיחו\הפריכו: **בהכרח** קיים ווקטור ε כך שהעץ T עם כלל אפסילון־החלטה והעץ T' עם כלל ההחלטה הרגיל יסווגו כל דוגמת מבחן ב$\mathbb{R}^d$ בצורה זהה.

הטענה לא נכונה.
יהיה עץ החלטה T (0 מסמן דוגמאות שסווגו בתור -, 1 מסמן דוגמאות שסווגו בתור +):



העץ T' יהיה:



נניח בשלילה שקיים וקטור $\varepsilon = [\varepsilon 1, \varepsilon 2]$ כך שהעץ T והעץ T' יסווגו כל דוגמת מבחן באופן זהה.
תהיה הדוגמה $(3 * \varepsilon 1, 3 * \varepsilon 2)$.
בעץ T נלך ימינה בצומת הראשון ואז שוב ימינה ונקבל סיווג +.
בעץ T' נלך ימינה בצומת הראשון ונקבל סיווג -.
כלומר קיבלנו דוגמה בה הסיווג לפי כל עץ יהיה שונה, בסתירה להנחה.

חלק ב' - היכרות עם הקוד

רקע

חלק זה הוא רק עבור היכרות הקוד, עבורו עליו במלואו וודאו כי הינכם מבינים את הקוד.
בחלק של הלמידה, נעזר ב $dataset$, הדאטה חולק עבורכם לשתי קבוצות: קבוצת אימון train.csv וקבוצת מבחן test.csv.
בכלל, קבוצת האימון תשמש אותנו לבניית המסווגים, וקבוצת המבחן תשמש להערכת ביצועיהם.

בקובץ utils.py תוכלו למצוא את הפונקציות הבאות לשימושכם:
load_data_set, create_train_validation_split, get_dataset_split
אשר טוענות/מחלקת את הדאטה בקבצי ה־csv למערכי np.array (קראו את תיעוד הפונקציות).

הדאטה של ID3 עבור התרגיל מכיל מדדים שנאספו מצילומים שנועדו להבחין בין גידול שפיר לגידול ממאיר. כל דוגמה מכילה 30 מדדים כאלה, ותווית בינארית **diagnosis** הקובעת את סוג הגידול (0=שפיר, 1=ממאיר). כל התכונות (מדדים) רציפות . העמודה הראשונה מציינת האם האדם חולה (M) או בריא (B). שאר העמודות מציינות כל תכונות רפואיות שונות של אותו אדם (התכונות מורכבות ואינכם צריכים להתייחס למשמעות שלהן כלל).

תיקיית $ID3 - dataset$:
- תיקיה זו אלו מכילה את קבצי הנתונים עבור $ID3$.

קובץ $utils.py$:
- קובץ זה מכיל פונקציות עזר שימושיות לאורך התרגיל, כמו טעינה של $dataset$ וחישוב הדיוק.
- בחלק הבא יהיה עליכם לממש את הפונקציה $accuracy$. קראו את תיעוד הפונקציות ואת ההערות הנמצאות תחת התיאור **TODO**.

קובץ $unit\_test.py$:
- קובץ בדיקה בסיסי שיכול לעזור לכם לבדוק את המימוש.

קובץ $DecisionTree.py$:
- קובץ זה מכיל 3 מחלקות שימושית לבניית עץ $ID3$ שלנו.
  - המחלקה $Question$: מחלקה זו מממשת הסתעפות של צומת בעץ. היא שומרת את התכונה ואת הערך שלפיהם מפצלים את הדאטה שלנו.
  - המחלקה $DecisionNode$: מחלקה זו מממשת צומת בעץ ההחלטה.
    הצומת מכיל שאלה $Question$ ואת שני הבנים $true\_branch, false\_branch$ כאשר $true\_branch$ הוא הענף בחלק של הדאטה שעונה $True$ על שאלת הצומת (הפונקציה $match$ של ה$Question$ מחזירה $True$).
    ו־ $false\_branch$ הוא הענף בחלק של הדאטה שעונה $False$ על שאלת הצומת (הפונקציה $match$ של ה$Question$ מחזירה $False$).
  - המחלקה $Leaf$: מחלקה זו מממשת צומת שהוא עלה בעץ ההחלטה. העלה מכיל לכל אחד מהמחלקות בדאטה את מספר הדוגמאות בעלה עבור כל מחלקה (למשל: $\{'B' : 5, 'M' : 6\}$).

קובץ $ID3.py$:
- קובץ זה מכיל את המחלקה של $ID3$ שתצטרכו לממש חלקים ממנה, עיינו בהערות ותיעוד המתודות.

קובץ $ID3\_experiments.py$:
- קובץ הרצת הניסויים של ID3, הקובץ מכיל את הניסויים הבאים, שיוסברו בהמשך:

$cross\_validation\_experiment, basic\_experiment$

חלק ג' – חלק רטוב ID3 (28 נק')

עבור חלק זה מותר לכם להשתמש בספריות הבאות:

All the built in packages in python, sklearn, pandas ,numpy, random, matplotlib, argparse, abc, typing.

<u>אך כמובן שאין להשתמש באלגוריתמי הלמידה, או בכל אלגוריתם או מבנה נתונים אחר המהווה חלק מאלגוריתם למידה אותו תתבקשו לממש.</u>

1. (3 נק') השלימו את הקובץ utils.py ע"י מימוש הפונקציה $accuracy$.
קראו את תיעוד הפונקציה ואת ההערות הנמצאות תחת התיאור TODO.
(הריצו את הטסטים המתאימים בקובץ $unit\_test.py$ לוודא שהמימוש שלכם נכון).
שימו לב! בתיעוד ישנן הגבלות על הקוד עצמו, אי־עמידה בהגבלות אלו תגרור הורדת נקודות.
בנוסף, שנו את ערך הID בתחילת הקובץ מ־123456789 למספר תעודת הזהות של אחד מהמגישים.

2. (10 נק') אלגוריתם ID3:

a. השלימו את הקובץ ID3.py ובכך ממשו את אלגוריתם $ID3$ כפי שנלמד בהרצאה. TODO
שימו לב שכל התכונות רציפות. אתם מתבקשים להשתמש בשיטה של חלוקה דינמית המתוארת בהרצאה. כאשר בוחנים ערך סף לפיצול של תכונה רציפה, דוגמאות עם ערך השווה לערך הסף משתייכות לקבוצה עם הערכים הגדולים מערך הסף. במקרה שיש כמה תכונות אופטימליות בצומת מסוים בחרו את התכונה בעלת <u>האינדקס המקסימלי</u>.
כלל המימוש הנ"ל צריך להופיע בקובץ בשם $ID3.py$, באזורים המוקצים לכך.
(השלימו את הקוד החסר אחרי שעיינתם והפנמתם את הקובץ $DecisionTree.py$ ואת המחלקות שהוא מכיל).

b. ממשו את $basic\_experiment$ שנמצאת ב $ID3\_experiments.py$ TODO
והריצו את החלק המתאים ב $main$ ציינו בדו"ח את הדיוק שקיבלתם. ✍️

```
Test Accuracy: 96.12%
```

3. גיזום מוקדם.
פיצול צומת מתקיים כל עוד יש בו יותר דוגמאות מחסם המינימום m, כלומר בתהליך בניית העץ מבוצע "גיזום מוקדם" כפי שלמדתם בהרצאות. שימו לב כי פירוש הדבר הינו שהעצים הנלמדים אינם בהכרח עקביים עם הדוגמאות .לאחר סיום הלמידה (של עץ יחיד), הסיווג של אובייקט חדש באמצעות העץ שנלמד מתבצע לפי רוב הדוגמאות בעלה המתאים.

a. ✍️ (2 נק') הסבירו מה החשיבות של הגיזום באופן כללי ואיזה תופעה הוא מנסה למנוע?

Pruning in decision trees is crucial for preventing **overfitting**, where the model learns the training data too well, including its noise and anomalies, leading to poor generalization to new data.

b. (3 נק') <u>עדכנו</u> את המימוש בקובץ $ID3.py$ כך שיבצע גיזום מוקדם כפי שהוגדר בהרצאה.
הפרמטר $min\_for\_pruning$ מציין את המספר המינימלי בעלה לקבלת החלטה, קרי יבוצע גיזום מוקדם אם ורק אם מספר הדוגמות בצומת קטן שווה לפרמטר הנ"ל. TODO

c. (8 נק') **שימו לב, זהו סעיף יבש ואין צורך להגיש את הקוד שכתבתם עבורו.**

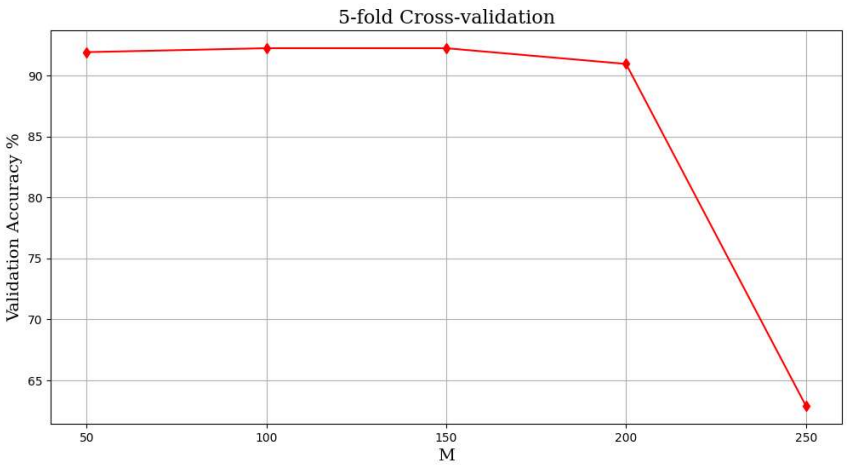
בצעו כיוונון לפרמטר M על קבוצת האימון:

1. בחרו לפחות חמישה ערכים שונים לפרמטר M.
2. עבור כל ערך, חשבו את הדיוק של האלגוריתם על ידי $\mathrm{K-fold\ cross\ validation}$ על קבוצת האימון בלבד.

כדי לבצע את חלוקת קבוצת האימון ל- K קבוצות יש להשתמש בפונקציה sklearn.model_selection.KFold עם הפרמטרים $\mathrm{shuffle = True, n\_split = 5}$ ו־random_state אשר שווה למספר תעודת הזהות של אחד מהשותפים.

✍️ i. השתמשו בתוצאות שקיבלתם כדי ליצור גרף המציג את השפעת הפרמטר M על הדיוק. צרפו את הגרף בדו"ח. (לשימושכם הפונקציה $util\_plot\_graph$ בתוך הקובץ $utils.py$).

```
 M value    | Validation Accuracy
    50      | 91.94%
   100      | 92.26%
   150      | 92.26%
   200      | 90.97%
   250      | 62.90%
============================
  Best M    | Validation Accuracy
   100      | 92.26%
best_m = 100
```



✍️ ii. הסבירו את הגרף שקיבלתם. לאיזה גיזום קיבלתם התוצאה הטובה ביותר ומהי תוצאה זו?

In Y axis we can see the accuracy and in X axis we can see the M values. As we can see from the graph itself or from the results, we get the **optimal accuracy for M=100 and 150 with 92.26% accuracy**. For higher values, we can notice the over-pruning effect that takes place since model has become too constrained to effectively learn from the training data.

✍️ **.d** (2 נק') השתמשו באלגוריתם ID3 עם הגיזום המוקדם כדי ללמוד מסווג מתוך **כל** קבוצת האימון ולבצע חיזוי על קבוצת המבחן.
השתמשו בערך ה־M האופטימלי שמצאתם בסעיף c. (ממשו $best\_m\_test$ שנמצאת ב $ID3\_experiments.py$ והריצו את החלק המתאים ב $main$). ציינו בדו"ח את הדיוק שקיבלתם. האם הגיזום שיפר את הביצועים ביחס להרצה ללא גיזום?

```
Test Accuracy: 97.09%
```

As we can see we got a 97.09% accuracy for m=100 which is higher than the 96.12% accuracy we got initially without pruning. Which means we slightly improved the performances.

הוראות הגשה

- ✓ הגשת התרגיל תתבצע אלקטרונית בזוגות בלבד.
- ✓ הקוד שלכם ייבדק (גם) באופן אוטומטי ולכן יש להקפיד על הפורמט המבוקש. הגשה שלא עומדת בפורמט לא תיבדק (ציון 0).
- ✓ המצאת נתונים לצורך בניית הגרפים אסורה ומהווה עבירת משמעת.
- ✓ הקפידו על קוד קריא ומתועד. התשובות בדוח צריכות להופיע לפי הסדר.
- ✓ יש להגיש קובץ zip יחיד בשם AI3_<id1>_<id2>.zip (ללא סוגריים משולשים) שמכיל:

- קובץ בשם AI_HW3_LEARNING.PDF המכיל את תשובותיכם לשאלות היבשות.
- קבצי הקוד שנדרשתם לממש בתרגיל **ואף קובץ אחר:**
  - קובץ – utils.py
  - בחלק של עצי החלטה – ID3_experiments.py, ID3.py

אין להכיל תיקיות בקובץ ההגשה, הגשה שלא עומדת בפורמט לא תיבדק.