תרגיל בית 3 –מבוא ללמידה

**עברו על כלל ההנחיות לפני תחילת התרגיל.**

# **הנחיות כלליות:**

* תאריך ההגשה: עד לסוף מועדי א' - 17/05/2024ב23:59
* 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 [Writing Hand on Apple ](https://emojipedia.org/apple/ios-14.6/writing-hand/) 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׳)**

[Writing Hand on Apple ](https://emojipedia.org/apple/ios-14.6/writing-hand/) **חלק א׳ – חלק היבש (28 נק')**

**k****NN – 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 ־דוגמות, , כאשר .  
כלומר הדוגמות הינן וקטורים ־ממדיים והתגיות הינן מדומיין כלשהו, הבעיה היא בעיית קלסיפיקציה (סיווג).  
אם לא נאמר אחרת, הקלסיפיקציה תהיה בינארית, כלומר .  
עבור כל דוגמה בסט האימון, ניתן להסתכל על הכניסה ה־ בווקטור כעל הfeature ה־ של הדוגמה, קרי כל דוגמה מיוצגת על ידי ־ערכים: .  
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 השכנים הקרובים ביותר שלה במישור ה־ממדי מבין הדוגמות בסט האימון, ומסווגים את הדוגמה על פי הסיווג הנפוץ ביותר בקרב השכנים.  
על מנת להימנע משוויון בין הסיווגים, נניח בדרך כלל כי ־אי זוגי, או שנגדיר היטב שובר שוויון.  
אם לא נאמר אחרת, במקרה של שוויון בקלסיפיקציה בינארית, נסווג את הדוגמה כחיובית .

שאלות הבנה

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

עבור d=1 נקבל כי מרחק מנהטן שווה למרחק אוקלידי, לכל k, ולכן אין תלות בין תוצאות הסיווג לבין בחירת פונקציית המרחק.  
בנוסף, עבור K שגדול או שווה לגודל כל קבוצת האימון, נקבל שלכל d, K השכנים הקרובים הם כל קבוצת האימון, בין אם המרחק הוא מרחק אוקלידי ובין אם מנהטן.

1. עבור בעיית קלסיפיקציה בינארית תנו דוגמה פשוטה לערכי , סט אימון ודוגמת מבחן בה השימוש בכל אחת מפונקציות המרחק הנ"ל משנה את סיווג דוגמה המבחן.נבחר d=2, k=1.  
   סט אימון: דוגמת מבחן:   
   עבור מרחק אוקלידי נקבל:

לכן דוגמת המבחן תקבל את הסיווג של של שהוא -.

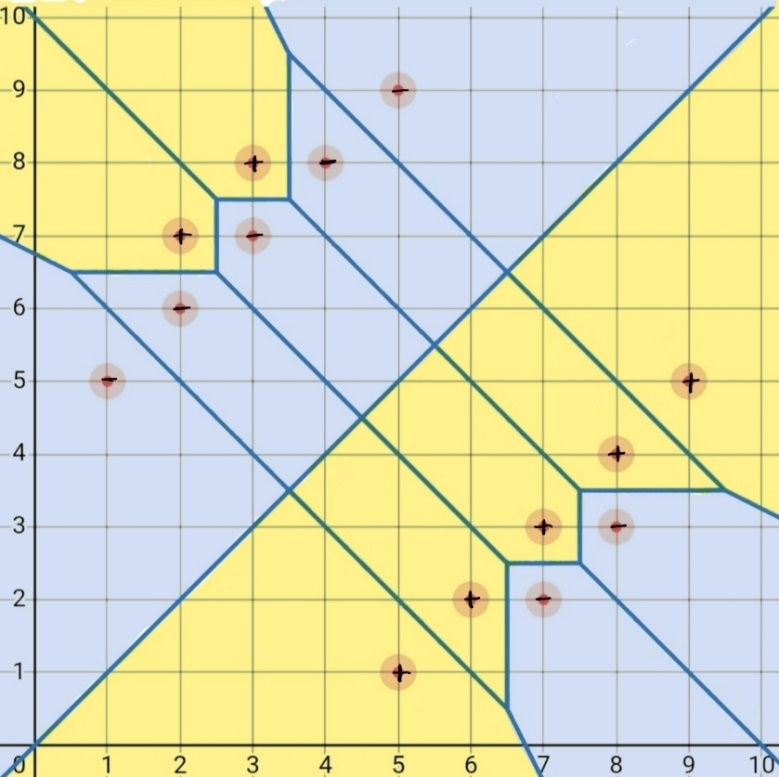
עבור מרחק מנהטן נקבל:

לכן דוגמת המבחן תקבל את הסיווג של של שהוא +.

מעתה, אלא אם כן צוין אחרת, נשתמש במרחק אוקלידי.

נתונה קבוצת האימון הבאה, כאשר :



1. (1 נק') איזה ערך של 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 שכל איבר בקבוצת האימון, מתסכל על 5 האיברים הקרובים ביותר אליו,  
   במקרה זה נקבל שכל איבר יקבל את הסיווג של רוב הקבוצה בחלק שלו (יש חלק עליון וחלק תחתון),  
   במקרה זה נקבל דיוק של 10 מתוך 14.
2. (1 נק') עבור איזה ערך של נקבל מסווג של קבוצת האימון? קרי כל דוגמת מבחן תקבל את הסיווג הנפוץ של כלל קבוצת האימון?  
   עבור k שהינו כגודל קבוצת האימון, כל דוגמת מבחן תיקח בחשבון את הסיווג של הרוב מתוך כל קבוצת האימון שלו, כלומר פשוט הרוב.
3. (2 נק') נמקו מדוע שימוש בערכי גדולים או קטנים מדי יכול להיות גרוע עבור קבוצת הדגימות הנ״ל.  
   ראשית נבחין שיש לנו 4 דוגמאות 'חריגות', שנתייחס אליהן כרועשות. אלו שתי הדוגמאות השונות בכל צד.  
   עבור ערכי K קטנים מידי, ההשפעה של הרעש עלולה להיות גדולה יותר, ולכן בדוגמא הנל היא עלולה להוביל ל overfitting.  
   לדוגמא עבור דוגמת מבחן שנמצאת מעל הנקודות החיוביות בצד שרובו שלילי, עבור k=2 תקבל ערך חיובי, בהתאם לערך הרועש, במקום הערך של רוב הקבוצה.  
   שימוש בערכי K גדולים מידי עלול להוביל גם הוא לoverfiring, שכן יכול להיות מצב שבו נתחשב יותר מידי ברוב הנקודות הרחוקות יותר מהקבוצה שלנו, במקום להתמקד ברוב 'באזור שלנו'.   
   לדוגמא דוגמת מבחן שנמצאת בקבוצה העליונה בקרבת הנקודות השליליות, עבור K גדול מידי תתחשב בכל הפלוסים שנמצאים מטה ותקבל סיווג חיובי במקום שלילי.
4. (2 נק') שרטט את גבול ההחלטה של 1-nearest neighbour - עבור הגרף.  
   נשרטט את דיאגרמת Voronoi כמו שראינו בהרצאה:  
   

**השוואה בין מודלי למידה – יש לנמק בקצרה את הפתרונות**

1. (3 נק') הציגו מסווג מטרה 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 המסווג הבא:

נגיד תכונה הוא תוכן הוקטור במקום

קבוצת אימון:

אלגוריתם ID3 עם התכונות יפצל כל צומת פעם לפי ערך הביט הראשון ופעם לפי ערך הביט השני (או הפוך, לא משנה), ויסווג כל דוגמת מבחן בהתאם למסווג המטרה.

נתבונן על דוגמת המבחן .  
אלגוריתם ID3 היה נותן לה תיוג של 1, בהתאם למסווג המטרה.  
אולם, לכל K, היינו מקבלים כי רוב השכנים הקרובים ביותר לדוגמת המבחן הם בעלי תיוג של 0, ולכן גם x היה מקבל תיוג של 0.

1. (3 נק') הציגו מסווג מטרה 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), אך למידת עץ ID3 תניב מסווג אשר עבורו קיימת לפחות דוגמת מבחן אפשרית אחת עליה הוא יטעה.
2. (3 נק') הציגו מסווג מטרה 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.
3. (3 נק') הציגו מסווג מטרה 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).

**מתפצלים ונהנים**

(7 נק') כידוע, 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 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 , דוגמת מבחן , ווקטור המקיים .  
כלל אפסילון־החלטה שונה מכלל ההחלטה הרגיל שנלמד בכיתה באופן הבא:   
נניח שמגיעים לצומת בעץ המפצל לפי ערכי התכונה , עם ערך הסף .  
אם מתקיים  אזי ממשיכים **בשני** המסלולים היוצאים מצומת זה, ואחרת ממשיכי לבן המתאים בדומה לכלל ההחלטה הרגיל. לבסוף, מסווגים את הדוגמה *בהתאם לסיווג הנפוץ ביותר של הדוגמאות הנמצאות בכל העלים אליהם הגענו במהלך הסיור על העץ (במקרה של שוויון – הסיווג ייקבע להיות ).*

יהא עץ החלטה לא גזום, ויהא העץ המתקבל מ־ באמצעות גיזום מאוחר שבו הוסרה הרמה התחתונה של (כלומר כל הדוגמות השייכות לזוג עלים אחים הועברו לצומת האב שלהם).  
הוכיחו\הפריכו: **בהכרח** קיים ווקטור כך שהעץ עם כלל אפסילון־החלטה והעץ עם כלל ההחלטה הרגיל יסווגו כל דוגמת מבחן ב בצורה זהה.

**הפרכה:**יהיה .

ויהי T עץ החלטה המסווג ראשית לפי התכונה: ושנית לפי . ירך הסף הינו .  
נשרטט את T לאחר אימון על קבוצה בגודל 4, ונסמן ב+,- את הסיווג הסופי של כל דוגמא, וליד כל עלה את הסיווג הכולל של העלה (לפי הרוב):

**+**

**-**

**-**

T' המתקבל מT הינו:

**+**

**-**

כעת, נניח בשלילה שקיים וקטור המקיים *, ועבורו לכל דוגמת מבחן x, הסיווג של x תחת T עם כלל החלטה , זהה לסיווג של x תחת T' עם כלל החלטה רגיל.*

*תהי דוגמת מבחן .  
עבור העץ T מתקיים:   
ולכן בצומת ההחלטה הראשון נבחר להתפצל לענף של T, ובצומת ההחלטה השני נבחר להתפצל לפי F, ונקבל סיווג של "–" עבור דוגמת המבחן.*

*עבור העץ T' מתקיים:*

**חלק ב׳ - היכרות עם הקוד  
רקע**

חלק זה הוא רק עבור היכרות הקוד, עבורו עליו במלואו ווודאו כי הינכם מבינים את הקוד.  
בחלק של הלמידה, נעזר ב 𝑑𝑎𝑡𝑎𝑠𝑒𝑡, הדאטה חולק עבורכם לשתי קבוצות: קבוצת אימון train.csv וקבוצת מבחן test.csv.   
ככלל, 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 :

* תיקיה זו אלו מכילה את קבצי הנתונים עבור .

קובץ :

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

קובץ :

* קובץ בדיקה בסיסי שיכול לעזור לכם לבדוק את המימוש.

קובץ :

* קובץ זה מכיל 3 מחלקות שימושית לבניית עץ שלנו.
  + המחלקה : מחלקה זו מממשת הסתעפות של צומת בעץ. היא שומרת את התכונה ואת הערך שלפיהם מפצלים את הדאטה שלנו.
  + המחלקה : מחלקה זו מממשת צומת בעץ ההחלטה.   
    הצומת מכיל שאלה ואת שני הבנים כאשר הוא הענף בחלק של הדאטה שעונה על שאלת הצומת   
    (הפונקציה של ה מחזירה ).  
    ו־ הוא הענף בחלק של הדאטה שעונה על שאלת הצומת   
    (הפונקציה של ה מחזירה ).
  + המחלקה : מחלקה זו מממשת צומת שהוא עלה בעץ ההחלטה. העלה מכיל לכל אחד מהמחלקות בדאטה את מספר הדוגמאות בעלה עבור כל מחלקה (למשל: ).

קובץ :

* קובץ זה מכיל את המחלקה של שתצטרכו לממש חלקים ממנה, עיינו בהערות ותיעוד המתודות.

קובץ :

* קובץ הרצת הניסויים של ID3, הקובץ מכיל את הניסויים הבאים, שיוסברו בהמשך:

**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.

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

1. (3 נק') השלימו את הקובץ utils.py ע"י מימוש הפונקציה .   
   קראו את תיעוד הפונקציה ואת ההערות הנמצאות תחת התיאור TODO.  
   (הריצו את הטסטים המתאימים בקובץ *לוודא שהמימוש שלכם נכון).*  
   שימו לב! בתיעוד ישנן הגבלות על הקוד עצמו, אי־עמידה בהגבלות אלו תגרור הורדת נקודות.  
   בנוסף, שנו את ערך ה בתחילת הקובץ מ־ למספר תעודת הזהות של אחד מהמגישים.
2. (10 נק') **אלגוריתם ID3:**
   1. השלימו את הקובץ ID3.py ובכך ממשו את אלגוריתם 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 האינדקס המקסימלי.   
      כלל המימוש הנ"ל צריך להופיע בקובץ בשם , באזורים המוקצים לכך.  
      (השלימו את הקוד החסר אחרי שעיינתם והפנמתם את הקובץ ואת המחלקות שהוא מכיל).
   2. ממשו את שנמצאת ב TODO

[Writing Hand on Apple ](https://emojipedia.org/apple/ios-14.6/writing-hand/)*והריצו את החלק המתאים ב* ציינו בדו"ח את הדיוק שקיבלתם.

1. **גיזום מוקדם.**

פיצול צומת מתקיים כל עוד יש בו יותר דוגמאות מחסם המינימום ,𝑚כלומר 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.

* 1. (2 נק') [Writing Hand on Apple ](https://emojipedia.org/apple/ios-14.6/writing-hand/)הסבירו מה החשיבות של הגיזום באופן כללי ואיזה תופעה הוא מנסה למנוע?

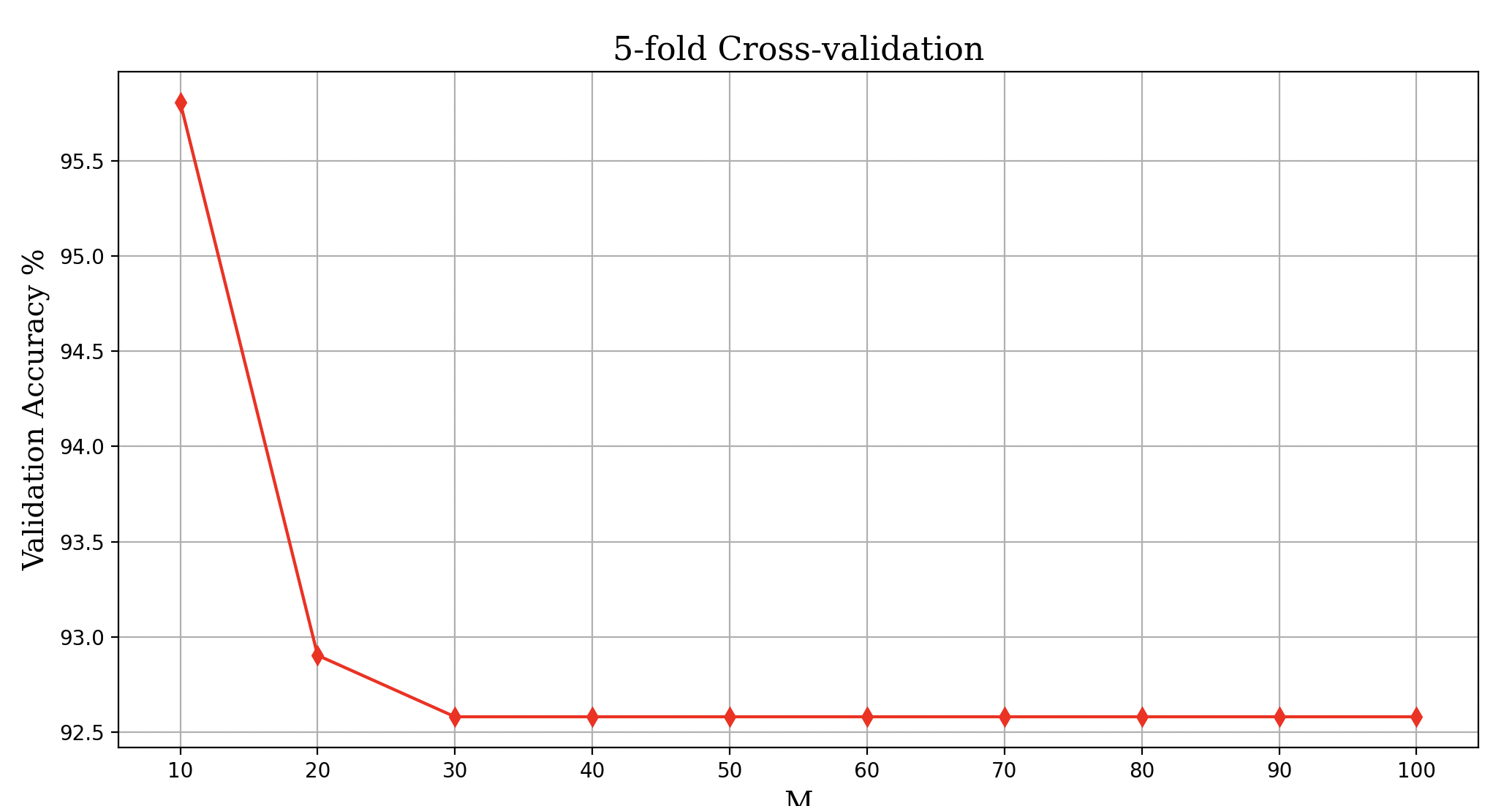
הגיזום חשוב כי הוא מנסה למנוע או להחליש את בעיית הoverfitting, ובכך משפר את הסיווג

* 1. (3 נק') **עדכנו** את המימוש בקובץ כך שיבצע גיזום מוקדם כפי שהוגדר בהרצאה.   
     הפרמטר מציין את המספר המינימלי בעלה לקבלת החלטה, קרי יבוצע גיזום מוקדם אם ורק אם מספר הדוגמות בצומת קטן שווה לפרמטר הנ"ל. TODO
  2. (8 נק') **שימו לב, זהו סעיף יבש ואין צורך להגיש את הקוד שכתבתם עבורו.**בצעו כיוונון לפרמטר על קבוצת האימון:

1. בחרו לפחות חמישה ערכים שונים לפרמטר .

2. עבור כל ערך, חשבו את הדיוק של האלגוריתם על ידי  על קבוצת האימון בלבד.   
כדי לבצע את חלוקת קבוצת האימון ל- קבוצות יש להשתמש בפונקציה [sklearn.model\_selection.KFold](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html" \l "sklearn.model_selection.KFold" \o "sklearn.model_selection)עם הפרמטרים ,   
ו־ אשר שווה למספר תעודת הזהות של אחד מהשותפים.

* + 1. [Writing Hand on Apple ](https://emojipedia.org/apple/ios-14.6/writing-hand/)השתמשו בתוצאות שקיבלתם כדי ליצור גרף המציג את השפעת הפרמטר על הדיוק.   
       צרפו את הגרף בדו״ח. (לשימושכם הפונקציה בתוך הקובץ ).



* + 1. [Writing Hand on Apple ](https://emojipedia.org/apple/ios-14.6/writing-hand/) הסבירו את הגרף שקיבלתם. לאיזה גיזום קיבלתם התוצאה הטובה ביותר ומהי תוצאה זו?

עבור גיזום של 10, נקבל את רמת הדיוק הכי גבוהה (שהיא לא 0). רמת הדיוק המתקבלת היא 95.81%

(טל אני לא יודע אם חרטטתי פה או לא, ואולי שווה לבדוק עוד m)

* 1. [Writing Hand on Apple ](https://emojipedia.org/apple/ios-14.6/writing-hand/)(2 נק׳) השתמשו באלגוריתם ID3 עם הגיזום המוקדם כדי ללמוד מסווג מתוך **כל** קבוצת האימון ולבצע חיזוי על קבוצת המבחן.   
     השתמשו בערך ה־ האופטימלי שמצאתם בסעיף c. (ממשו שנמצאת ב  *והריצו את החלק המתאים ב ).* ציינו בדו"ח את הדיוק שקיבלתם. האם הגיזום שיפר את הביצועים ביחס להרצה ללא גיזום?

עבור הרצה עם m = 10, קיבלנו רמת דיוק של 95.15%, והיא אותה רמת דיוק שקיבלנו ללא גיזום. לכן אין שינוי בביצועים.

**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.py, ID3\_experiments.py

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