תרגיל בית 3 – מבוא ללמידה

הנחיות כלליות:

- תאריך ההגשה: 27/01/2021 ב23:59
- את המטלה יש להגיש בזוגות בלבד.
- שאלות בנוגע לתרגיל יש לשלוח למוחמד דהאמשה: muhammad.dah@campus.technion.ac.il עם הכותרת AI_HW3, עדיף יותר להשתמש בפיאצה של הקורס. לפני ששואלים שאלה, יש לבדוק אם היא נשאלה בעבר. קישור: piazza.com/technion.ac.il/winter2022/236501
- בחלק מהסעיפים קיימת הגבלת שורות לפתרון שלכם. אם אתם חושבים שבסעיף מסוים הגבלה זו לא ריאלית אנא שילחו נימוק על כך במייל ואם הנימוק יתקבל נשנה את ההגבלה לכלל הסטודנטים.
- יתכן שיהיו שינויים במהלך התרגיל. כל העדכונים מחייבים.
- ציון בית המקסימלי הינו 105, אך ניתן לקבל עד 100, כלומר הציון הסופי הוא $\min(\text{grade}, 100)$.
- בחלק ד' תתבקשו להציע אלגוריתם לבחירת מאפיינים, כתוצאה מכך יש בחלק זה דרגת חופש, הציון בשאלות אלו יינתן בהתאם לטיב הפתרון, הן מבחינת יצירתיות והן מבחינת תוצאות אמפיריות. יש לתאר בצורה ברורה, פורמלית ומדויקת את הפתרון.
- התשובות לסעיפים בהם מופיע הסימון ✍ צריכים להופיע בדוח.
- לחלק הרטוב מסופק שלד של הקוד (שאתם נדרשים להשלים החלקים שמסומנים **TODO**)
- מותר להשתמש בספריות sklearn, pandas ,numpy, random, matplotlib, argparse, abc, typing, all the built in packages in python אך כמובן שאין להשתמש באלגוריתמי הלמידה, או בכל אלגוריתם או מבנה נתונים אחר המהווה חלק מאלגוריתם למידה אותו תתבקשו לממש.

מטרות התרגיל:

❖ הבנת ההשפעה של תכונות ושל דוגמאות על הביצועים של אלגוריתמי למידה.
❖ הבנת תהליך כיוונון פרמטרים והערכת ביצועים של אלגוריתמי למידה.
❖ תרגיל זה עוסק בבעיית סיווג בינארית. לאורך התרגיל נתנסה בבניית סוגים שונים של מסווגים בסיסיים (decision Tree, KNN) והרחבות שלהם. נעסוק בלמידה של **עצי החלטה ובחירת מאפיינים**. במהלך התרגיל תתבקשו להריץ מספר ניסויים ולנתח את תוצאותיהם. אנא בצעו **ניתוח מעמיק ומפורט** של התוצאות וצרפו אותו לדו"ח כפי שיוסבר בהמשך התרגיל.

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

רקע - DATA SETS:

בתרגיל זה נעסוק בשני $data-sets$, 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. לשימושכם ניתן להיעזר בפונקציות:
load_data_set, create_train_validation_split, create_train_validation_split, get_dataset_split
בקובץ utils.py אשר טוענות/מחלקת את הדאטא למערכים np.array בצורה נוחה (קראו את תיעוד הפונקציות).

1) **<u>עבור החלק של ID3:</u>** דאטה 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 מיני תכונות רפואיות של אותו אדם (התכונות קצת מורכבות ואינכם צריכים להתייחס למשמעות שלהן כלל).

2) **<u>עבור החלק של KNN ובחירת מאפיינים:</u>** מטרת data-set זה היא לחזות באופן אבחנתי אם למטופל יש סוכרת או לא, על סמך מדידות אבחנתיות מסוימות הכלולות במסד הנתונים. להלן המפרט הרשמי, התכונות הבאות סופקו כדי לעזור לנו לחזות אם אדם חולה סוכרת או לא:
<u>Pregnancies</u>: מספר הפעמים בהריון.
<u>Glucose</u>: ריכוז גלוקוז בפלזמה במשך שעתיים (בבדיקת סבילות לגלוקוז דרך הפה).
<u>BloodPressure</u>: לחץ דם דיאסטולי (mm Hg).
<u>SkinThickness</u>: עובי קפל העור התלת ראשי (mm).
<u>Insulin</u>: אינסולין בסרום שעתיים (mu U/ml).
<u>BMI</u>: מדד מסת הגוף (weight in kg/(height in m)2).
<u>DiabetesPedigreeFunction</u>: תפקוד אילן יוחסין של סוכרת (פונקציה המציינת את הסבירות לסוכרת על סמך היסטוריה משפחתית).
<u>Age</u>: גיל (שנים).

<u>Outcome</u>: תוצאה בינארית (0 אם ללא סוכרת, 1 אם חולה סוכרת).

חלק א׳ - היכרות עם הקוד

תיקיות $ID3-dataset$ / $KNN-dataset$:

- תיקיות אלו מכילות את קבצי הנתונים עבור $ID3$ ו- KNN, בהתאמה. הדאטה חולק עבורכם לשתי קבוצות: קבוצת אימון train.csv וקבוצת מבחן test.csv ככלל

קובץ $utils.py$:

- קובץ זה מכיל פונקציות עזר שימושיות לאורך התרגיל, כמו טעינה של $dataset$ וחישוב הדיוק.
- עליכם לממש את הפונקציות $accuracy, l2\_dist$. קראו את תיעוד הפונקציות ואת ההערות הנמצאות תחת התיאור **TODO**.
- מסופק לכם $unit\_test.py$ בתור קובץ טסט בסיסי שיכול לעזור לכם לבדוק את המימוש.

קובץ $DecisionTree.py$:

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

קבצים $ID3\_experiments.py$ / $ID3.py$:

- קובץ זה מכיל את המחלקה של $ID3$, עיינו בהערות ותיעוד המתודות.
- קובץ הרצת הניסויים של ID3, בקובץ זה יש את הניסויים (שנסביר עליהם בהמשך):
  cross_validation_experiment, basic_experiment

קבצים $KNN\_experiments.py$ / $KNN.py$ / KNN_CV.pyc:

- קובץ זה מכיל את המחלקה המומשת של KNN, עיינו בהערות ותיעוד המתודות.
- קובץ הרצת הניסויים של KNN, בקובץ זה נממש גם את חלק ד׳ (בחירת המאפיינים).
- קובץ KNN_CV.pyc מריץ cross validation עבור KNN (הסבר מפורט בחלק ד׳)

✍️ חלק ב׳ – חלק היבש (25 נק׳)

1. (7 נק׳) הוכח/הפרך: בהינתן דאטה כלשהו עם תכונות רציפות ותיוגים בינאריים המחולק לקבוצת אימון ומבחן, הפעלה של פונקציית נירמול Minmax, על הדאטה אינה משפיעה על דיוק של מסווג ID3 הנלמד על קבוצת האימון והנבחן על קבוצת המבחן.
[אורך התשובה מוגבל ל20שורות]

2. (12 נק׳) נגדיר דאטה סט $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ שבו n דוגמאות מתויגות עם סיווג בינארי $y_i \in \{0,1\}$. כל דוגמה היא וקטור תכונות המורכב משתי תכונות רציפות $x_i = (v_{1,i}, v_{2,i})$.
הניחו כי קיים מסווג מטרה $f(x): R^2 \to \{0,1\}$ שאותו אנו מעוניינים ללמוד (הוא אינו ידוע לנו) וכן שהדוגמאות ב-D עקביות עם מסווג המטרה (כלומר שאין דוגמאות רועשות ב- D).
בסעיפים הבאים, עבור KNN, הניחו פונק׳ מרחק אוקלידי.
כמו כן, הניחו שאם קיימות נקודות במרחב כך שעבורן יש מספר דוגמאות במרחק זהה, קודם מתחשבים בדוגמאות עם ערך v_1 מקסימלי ובמקרה של שוויון בערך של v_1, מתחשבים קודם בדוגמאות עם ערך v_2 מקסימלי. הניחו כי אין דוגמאות זהות לחלוטין (כלומר גם עם ערך v_1 זהה וגם עם ערך v_2 זהה).
בכל סעיף, **הציגו מקרה המקיים את התנאים המוצגים בסעיף, הסבירו במילים, וצרפו תיאור גרפי (ציור) המתאר את המקרה (הכולל לפחות תיאור מסווג המטרה והדוגמאות שבחרתם)** . סמנו דוגמאות חיוביות בסימן '+' (פלוס) ודוגמאות שליליות בסימן '-' (מינוס). בכל אחת מתתי הסעיפים הבאים אסור להציג מסווג מטרה טריוויאלי, דהיינו שמסווג כל הדוגמאות כחיוביים או כל הדוגמאות כשליליים.

[3 שורות לכל סעיף, אין הגבלה על הגרפים, מלל ופתרון שאינו מוגדר היטב כמתבקש לא יקבל ניקוד]

סעיף(א)(3נק׳)
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.

סעיף(ב)(3נק׳)
הציגו מסווג מטרה $f(x): R^2 \to \{0,1\}$ 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), אך למידת עץ ID3 תניב מסווג אשר עבורו קיימת לפחות דוגמת מבחן אפשרית אחת עליה הוא יטעה.

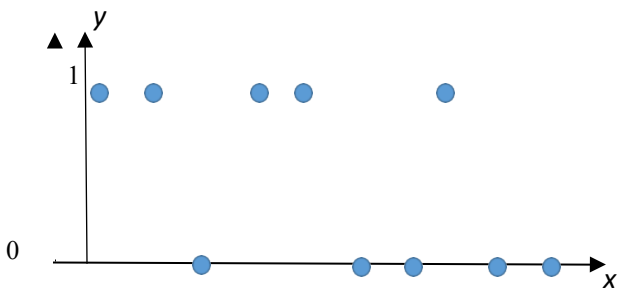
סעיף(ג)(3נק׳)
הציגו מסווג מטרה $f(x): R^2 \to \{0,1\}$ וקבוצת אימון בעלת לכל היותר 10 דוגמאות כך שלמידת מסווג KNN עבור ערך K מסוים תניב מסווג אשר עבורו קיימת לפחות דוגמת מבחן אפשרית אחת עליה הוא יטעה, וגם למידת עץ ID3 תניב מסווג אשר עבורו קיימת לפחות דוגמת מבחן אחת אפשרית עליה הוא יטעה.

סעיף(ד)(3נק׳)
הציגו מסווג מטרה $f(x): R^2 \to \{0,1\}$ 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).

3. (6 נק') נניח שאתם משתמשים ב-Majority Classifier, עבור קבוצת האימון הבאה שמכילה 10 דוגמאות כאשר לכל דוגמה יש תכונה אחת בעלת ערך ממשי x, ותווית בינארית y, עם ערך '0' או '1'. נגדיר במסווג Majority Classifier אשר מסווג את התווית לפי הרוב בקבוצת האימון, <u>ללא קשר לערך הקלט</u>. במקרה של תיקו, הסיווג הוא המחלקה '1'.

<u>למשל:</u> עבור קבוצת האימון $\{(1,'1'),(2,'0'),(3,'1')\}$ הסיווג של Majority Classifier הוא:
$Predict(x=5)='1'$ כי רוב התוויות בקבוצת האימון הוא '1'
$Predict(x=2)='1'$ למרות שהנקודה נמצאת בקבוצת האימון

ענה על השאלות הבאות בהתבסס על קבוצת האימון המתוארת בגרף הבא:



סעיף (א) (3 נק')
מהו ערך הדיוק של המסווג Majority Classifier על קבוצת האימון?

סעיף (ב) (3 נק')
מהו ערך הדיוק המתקבל ע"י הרצת 2-fold Cross-Validation ?
הניחו ש-5 הנקודות השמאליות ביותר (כלומר, 5 הנקודות עם ערכי ה-X הקטנים ביותר) נמצאות ב fold אחד ו-5 הנקודות הימנית ביותר נמצאות בfold השני.

חלק ג' – חלק רטוב ID3 (50 נק')

4. (5 נק') השלימו את הקובץ utils.py ע"י מימוש הפונקציות $accuracy, l2\_dist$. קראו את תיעוד הפונקציות ואת ההערות הנמצאות תחת התיאור TODO.
(הריצו את הטסטים המתאימים בקובץ $unit\_test.py$ לוודא שהמימוש שלכם נכון)

5. (25 נק') **אלגוריתם ID3:**

a. ממשו את אלגוריתם $ID3$ כפי שנלמד בהרצאה.
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.
המימוש צריך להופיע בקובץ בשם $ID3.py$ (השלימו את הקוד החסר אחרי שעיינתם והפנמתם את הקובץ $DecisionTree.py$ ואת המחלקות שהוא מכיל).

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

6. (20 נק') **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.** הסבירו מה החשיבות של הגיזום באופן כללי ואיזה תופעה הוא מנסה למנוע?
[אורך התשובה מוגבל ל 3 שורות]

b. ממשו את הגיזום המוקדם כפי שהוגדר בהרצאה. הפרמטר M מציין את מספר המינימלי בעלה לקבלת החלטה. על המימוש של הגיזום המוקדם להיות גם כן בתוך המחלקה ID3 שנמצאת בקובץ $ID3.py$)

c. בצעו כיוונון לפרמטר M על קבוצת האימון:

i. בחרו לפחות חמישה ערכים שונים לפרמטר M.

ii. עבור כל ערך, חשבו את הדיוק של האלגוריתם על ידי $K-\text{fold cross validation}$ על קבוצת האימון בלבד. כדי לבצע את חלוקת קבוצת האימון ל- K קבוצות יש להשתמש בפונקציה sklearn.model_selection.KFold עם הפרמטרים $n\_split = 5$, $shuffle = True$ ו- $random\_state$ שווה למספר תעודת הזהות שלכם. (כל כיוונון פרמטרים בתרגיל יעשה בצורה דומה).

✍️ iii. 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 $utils.py$).

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

מימוש כיוונון הפרמטר נמצא בפונקציה בשם $cross\_validation\_experiment$ (קראו את התיעוד והשלימו אותה) בקובץ $ID3\_experiments.py$.

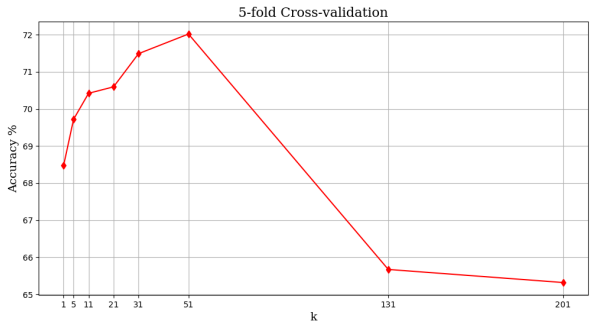
d. השתמשו באלגוריתם ID3 עם הגיזום המוקדם כדי ללמוד מסווג מתוך **כל** קבוצת האימון ולבצע חיזוי על קבוצת המבחן. השתמשו בערך ה- M האופטימלי שמצאתם בסעיף c. (ממשו $best\_m\_test$ שנמצאת ב $ID3\_experiments.py$ והריצו את החלק המתאים ב $main$)
ציינו בדו"ח את הדיוק שקיבלתם. האם הגיזום שיפר את הביצועים ביחס להרצה ללא גיזום בשאלה 5?

חלק ד' – חלק רטוב - בחירת מאפיינים (30 נק')

נרצה לשפר אלגוריתם הלמידה k-nearest neighbors אשר המימוש שלו מסופק לכם בקובץ $KNN.py$, בעיקר במחלקה של $KNNClassifier$ שתי פונקציות שהם fit(X, Y) שמבצעת אימון, ו- predict(X) שמבצעת את הפרדיקציה.

בנוסף, מסופק לכם קובץ KNN_CV.pyc שמכיל bytecode – הוראות שונות של ה-interpreter- אך כאמור זה אינו קובץ בשפת מכונה , הסבר: בשלב הראשון, קוד פייתון מהודר לשפת ביניים נמוכה, מבוססת מחסנית (Bytecode). תוצאת ההידור הזה נשמרת בקבצים עם הסיומת pyc. בעת הרצה, סביבת זמן הריצה (המפרש) מריצה את קוד הביניים. קובץ זה מריץ cross validation אשר בהרצתו נקבל את ה K הטוב ביותר למסווג.

7. (5 נק') הריצו הפונקציה run_cross_validation שנמצאת בקובץ $KNN\_experiments.py$, הערכים השונים לפרמטר K הם $[1, 5, 11, 21, 31, 51, 131, 201]$. ודאו שהתוצאה המתקבלת וגרף הדיוק של המסווג הם :




```
 K value   | Validation Accuracy
    1      | 68.48%
    5      | 69.72%
   11      | 70.42%
   21      | 70.59%
   31      | 71.49%
   51      | 72.02%
  131      | 65.67%
  201      | 65.32%
==========================
  Best K   | Validation Accuracy
   51      | 72.02%
Test Accuracy: 75.50%
```

ניתן לראות שערך הפרמטר הטוב ביותר, לפי מדד הדיוק, עם מרחק אוקלידי, הוא $K = 51$, נשתמש בערך זה לאורך המשך התרגיל.

✍️ הסבירו בקצרה איך עובד אלגוריתם KNN, ציינו 2 חסרונות ו- 2 יתרונות של האלגוריתם.
[אורך התשובה מוגבל ל 5 שורות]

בחירת מאפיינים - (25 נק')

בחלק זה נתמקד בבעיית בחירת תת קבוצה של מאפיינים מתוך קבוצת המאפיינים המלאה.
בהינתן קבוצת המאפיינים המלאה בגודל D: $S = \{x_1, \dots, x_D\}$ נרצה לבחור מתוך מאפיינים אלו רק תת קבוצה של מאפיינים בגודל b $(1 \leq b \leq D)$ אשר ישמשו אותנו בבניית המסווג.

הסיבות שבגינן נרצה לבצע בחירת תת קבוצה של מאפיינים הן מגוונות, להלן חלק מהסיבות:

- מחיקת מאפיינים מיותרים, כלומר כאלו אשר אינם קשורים ואינם תורמים מידע עבור הפרדה בין המחלקות השונות, או לחלופין מאפיינים אשר ניתן לבטא אותם ע"י מאפיינים אחרים בקבוצה ולכן לא תורמים לנו מידע רלוונטי ויכולים אף לפגוע בביצועי המסווג.
- מתן תובנות על בעיית הסיווג: בחירת המאפיינים תיתן לנו מידע לגבי הרלוונטיות של המאפיינים לתהליך הסיווג.

הקטנת הסיבוכיות החישובית.

8. ✍ (5 נק') נגדיר את כל תתי הקבוצות של S בתור $P(S) = \{X | X \subseteq S\}$, כלומר, כל תתי הקבוצות האפשריות בגודל קטן או שווה ל-D.

a. מהו מספר כל תתי הקבוצות של הקבוצה S?
b. מהו מספר כל תתי הקבוצות של הקבוצה S בגודל b $(1 \leq b \leq D)$?

9. ✍ (20 נק') בשאלה 8 ראינו כי מספר תת הקבוצות האפשריות של מאפיינים הוא לא פולינומיאלי בגודל הקבוצה. באופן כללי מציאת קבוצת המאפיינים האופטימלית עבור מושג מטרה היא בעיה NP קשה.
נרצה למצוא דרך (אלגוריתם) לבחור תת קבוצה של מאפיינים טובה בצורה חכמה. אשר לוקח את קבוצת האימון x, y ומבצע $preprocessing$ לבחירת המאפיינים החשובים.

ממשו את get_top_b_features אשר לוקחת את הפרמטרים x, y, b, k ומחזירת קבוצה של אינדיקסים (ממוינת) של המאפיינים שבהם נרצה לבחור. קראו את התיעוד.
אלגוריתם הלמידה יאומן מחדש על קבוצת האימון החדשה x', שהיא סינון הקבוצה x לקבוצת המאפיינים שקיבלנו מ get_top_b_features.

<u>הערה:</u> אתם נדרשים להציע אלגוריתם לבחירת מאפיינים, מכאן בחלק זה קיימת דרגת חופש.
הציון בשאלה זו יינתן בהתאם לטיב הפתרון, הן מבחינת יצירתיות והן מבחינת תוצאות אמפיריות.
יש לתאר בצורה ברורה, פורמלית ומדויקת את הפתרון.
<u>כיוון למחשבה:</u> נסו להשתמש בכלים שצברנו עד כה בקורס (יתקבלו כמובן גם פתרונות אחרים).

[מימוש $brute\ force$ אשר עובר על כל תת הקבוצות בגודל b לא ייתקבל]

<u>a.</u> על איזו קבוצה (קבוצת אימון ולידציה או מבחן), יש לבדוק את הביצועים? הסבר.
<u>b.</u> מה גודל תת הקבוצה של המאפיינים האופטימלי שקיבלתם, האם קיבלתם שיפור בביצועים בעזרת בחירת מאפיינים?
<u>c.</u> הסבירו במפורש את האלגוריתם שביצעתם.

הוראות הגשה

✓הגשת התרגיל תתבצע אלקטרונית בזוגות בלבד.
✓הקוד שלכם ייבדק (גם) באופן אוטומטי ולכן יש להקפיד על הפורמט המבוקש. הגשה שלא עומדת בפורמט תקבל קנס.
✓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.PDF המכיל את תשובותיכם לשאלות היבשות.
- ✓ כל קבצי הקוד שנדרשתם לממש בתרגיל:
  - קובץ – utils.py
  - בחלק של עצי החלטה – ID3.py, ID3_experiments.py
  - בחלק של שכנים קרובים ביותר ובחירת מאפיינים – KNN_experiments.py
  - כל קוד עזר שמימשתם בתרגיל.

בהצלחה! 😊