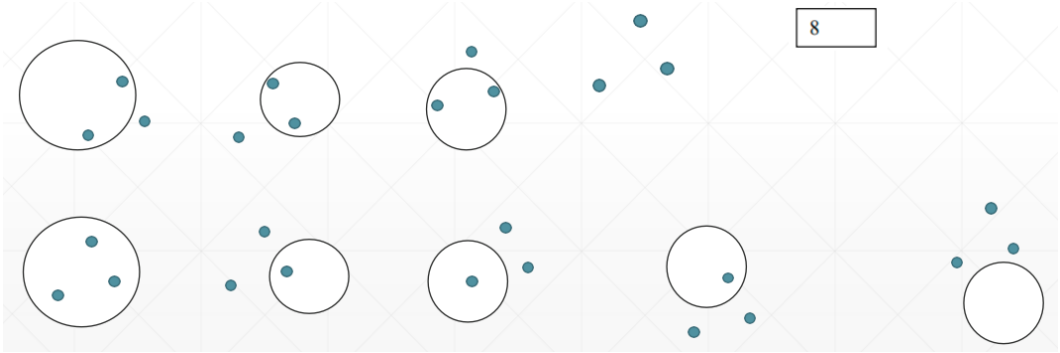
# מבוא ללמידה ממוכנת: שיעורי בית 3 –

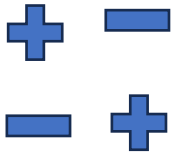
**מגישים:**
נעם לביא - 325674513
ליאון גורין – 214511214

**שאלה 1:**

<u>Section a</u>: ניעזר בציור מתוך התרגול לקראת המבחן כדי להראות שה $VCdim(\mathcal{H}) \geq 3$ (נקודות במעגל יסווגו עם +, אחרת עם -):




נראה דוגמה המפריכה את הטענה $VCdim(\mathcal{H}) = 4$:



לא ניתן להכיל את 2 הדגימות החיוביות במעגל ללא עוד דגימה שלילית שתיכנס ותסווג לא נכון.

לכן $VCdim(\mathcal{H}) = 3$.

<u>Section b</u>: ניעזר בנוסחה:

$$m \in O\left(\frac{1}{\epsilon^2}(VCDim(\mathcal{H}) + \ln\left(\frac{1}{\delta}\right)\right)$$

נציב את הנתונים בנוסחה:

$$m \in O\left(\frac{1}{\left(\frac{\varepsilon}{2}\right)^2}\left(3 + \ln\left(\frac{1}{e^{-2}}\right)\right)\right) = O\left(\frac{4}{\varepsilon^2}(3+2)\right) = O\left(\frac{20}{\varepsilon^2}\right)$$

כלומר אנחנו צריכים סדר גודל של $\frac{20}{\varepsilon^2}$ דגימות כדי ש $\mathcal{H}$ יהיה AGNOSTIC PAC Learnable.

Section c: תחת ההנחת ה Realizability ועם הנתון ש $C_1 = 1$ נוכל להשתמש בנוסחה:

$$m \geq C_1 \cdot \frac{1}{\epsilon}\left(VCDim(\mathcal{H}) + \ln\left(\frac{1}{\delta}\right)\right)$$

נציב את הנתונים בנוסחה:

$$m \geq 1 \cdot \frac{1}{\frac{\varepsilon}{2}}\left(3 + \ln\left(\frac{1}{e^{-2}}\right)\right) = \frac{2}{\varepsilon}(3+2) = \frac{10}{\varepsilon}$$

ולכן צריך לפחות: $\frac{10}{\varepsilon}$ דגימות.

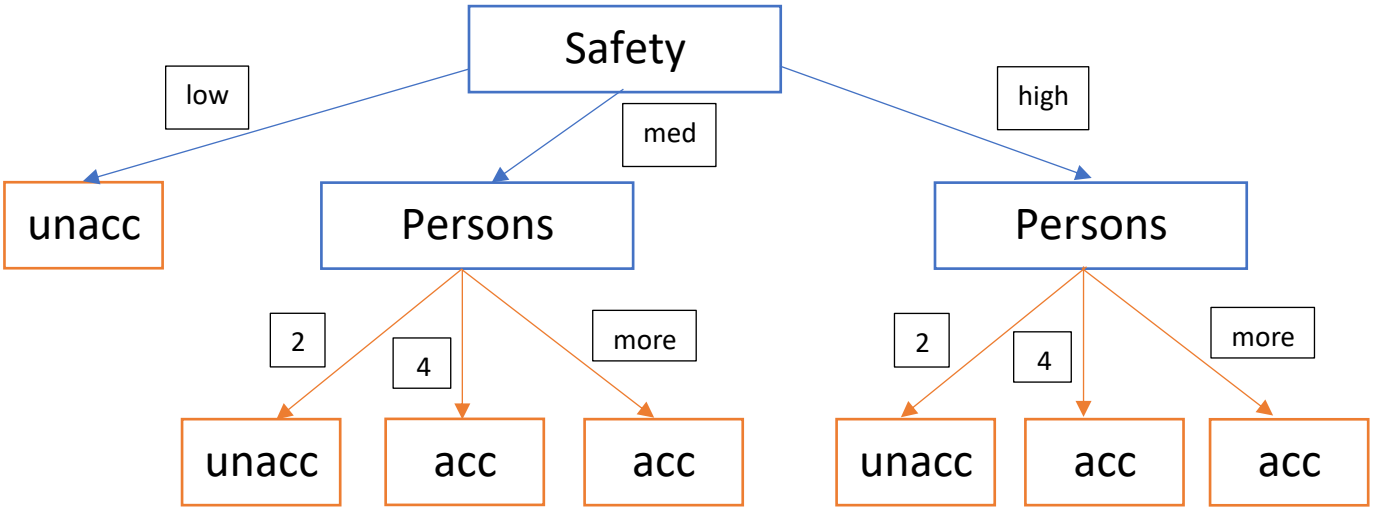
**שאלה 2:**

Section b: קיבלנו שאחוז ההצלחה הוא 100 כי עץ החלטה תמיד יעשה overfit לפי ה data הנתון לו במהלך ה training אם לא נגביל לו את הגובה.

Section c:

כאשר maxdepth=2 קיבלנו שאחוז ההצלחה באימון היה: 77.13% וטסט: 80.35%.
במקרה זה קיבלנו תוצאה של underfitting כי סיבוכיות המודל הייתה קטנה מדי – שגיאות אימון גבוהה.

כאשר ה maxdepth=5 אחוז ההצלחה באימון היה: 96.89% ובטסט: 92.49% לעומת
במקרה זה קיבלנו תוצאה של overfitting, שגיאת האימון הייתה נמוכה אבל עבור data שעדיין לא ראינו נקבל שגיאה גבוהה יותר.



**שאלה 4:**

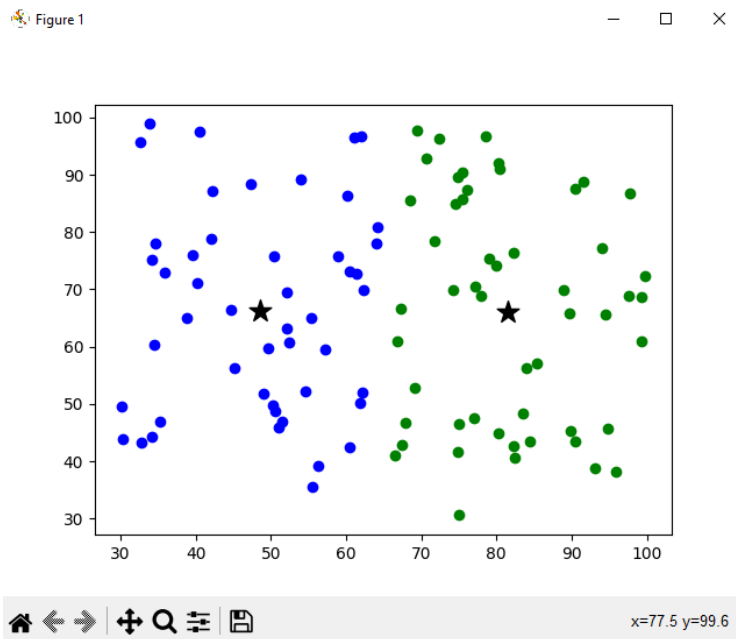
<u>Section a</u>: נרצה למצוא את הנגזרת של הביטוי כדי לראות מה הערך שמוביל למינימום עבור $\mu_i$:

$$\frac{\partial}{\partial \mu_i} J_{SSE\,(i)} = -2 \sum_{x \in D_i} (x - \mu_i) = 0$$

$$\sum_{x \in D_i} x = \sum_{x \in D_i} \mu_i = n_i \mu_i$$

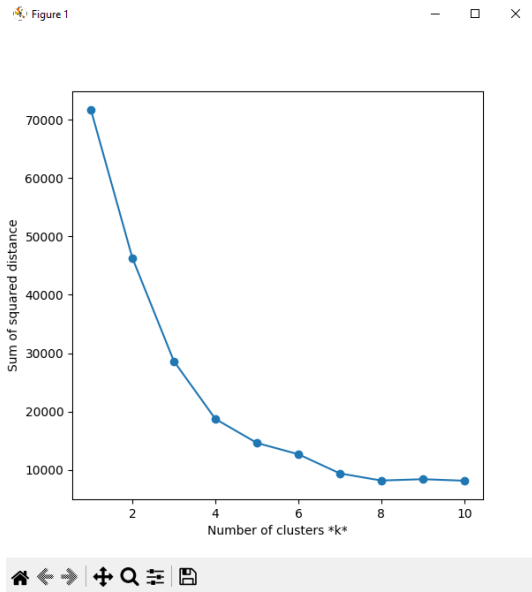
$$\mu_i = \frac{1}{n_i} \sum_{x \in D_i} x\,, \qquad Q.E.D$$

<u>Section b</u>:



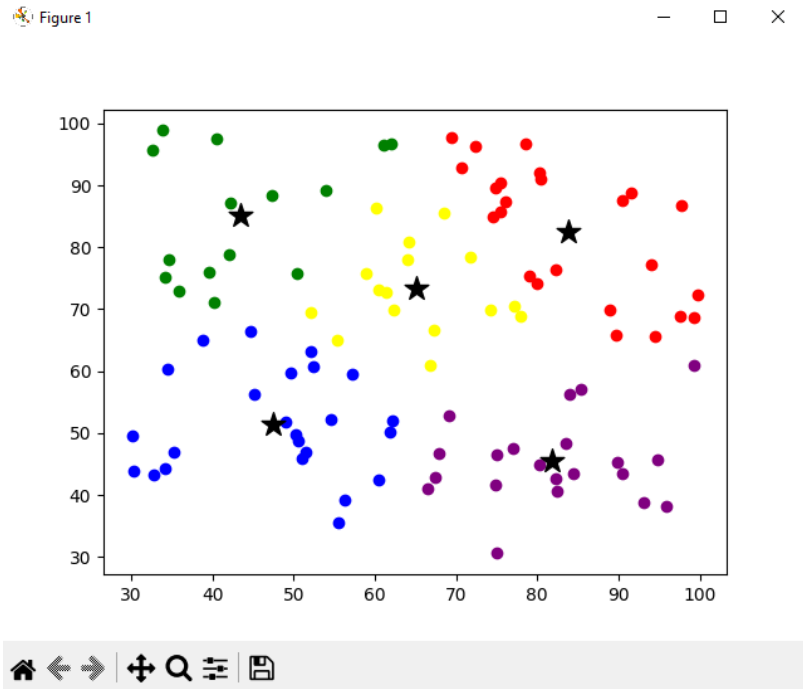
:Section c

**– Elbow plotting**

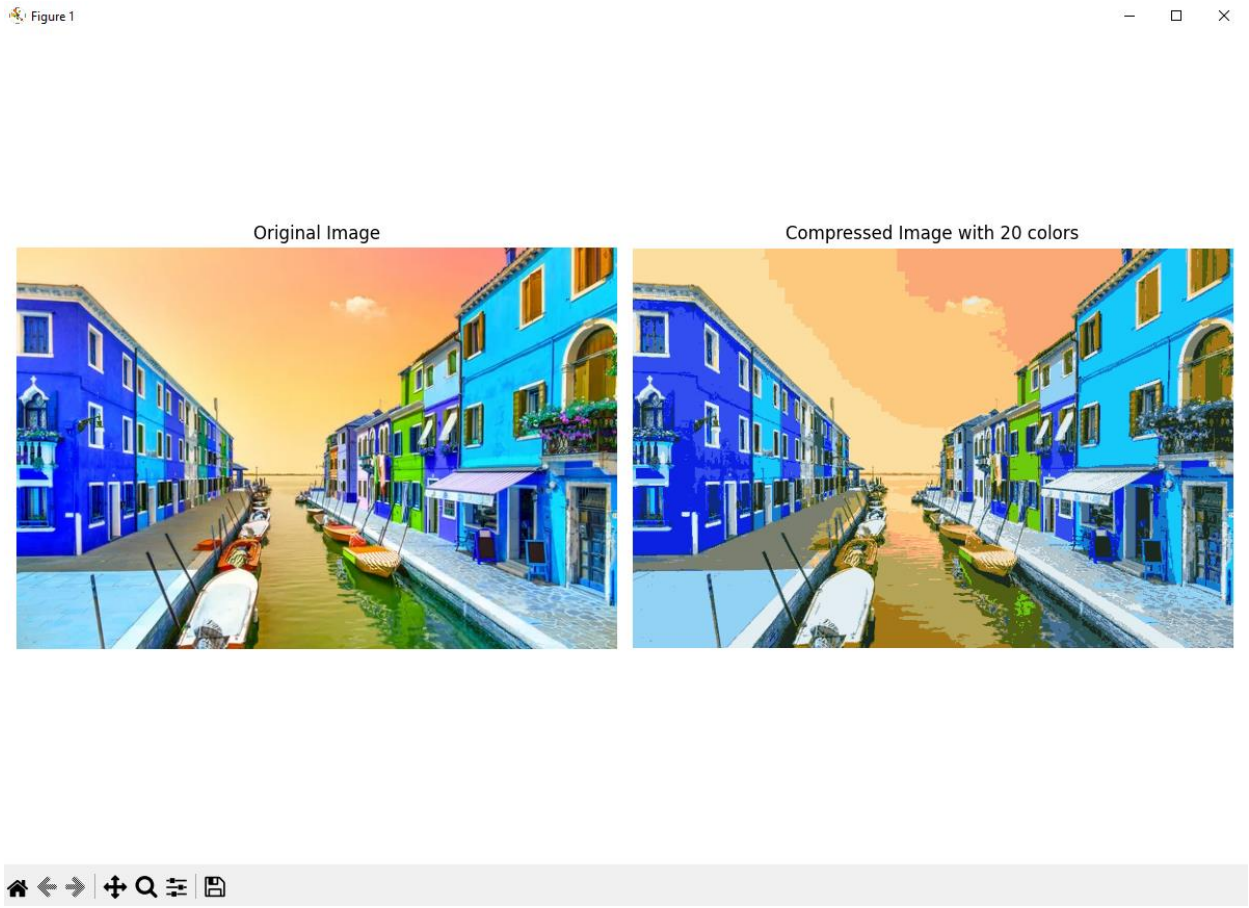



לפי הגרף, בחרנו ב k=5 כי ישר אחריו השיפור מתחיל להשתטח והשיפור נהיה מזערי.

**סיווג הנקודות לפי k=5**



:Section d



# תרגיל בית 3 - למידה ממוכנת

**<u>תרגיל 3</u>**

נניח שיש לנו אוסף של משתנים מקריים $\{Y_i\}_{i=1}^n$ עם ממוצע $\mu$ ושונות $\sigma^2$.
במקרה שלנו, משתנה $Y_i$ מייצג את החיזוי שנוצר על ידי המסווג $i$.

**<u>סעיף א'</u>**

**חשבו את הממוצע והשונות, $Y = \frac{1}{n}\Sigma Y_i$.**

תוחלת:

$$E[Y] = E[\frac{1}{n} \cdot \sum Y_i]$$

$$= \frac{1}{n} \cdot E[\sum Y_i]$$

$$= \frac{1}{n} \cdot \sum E[Y_i]$$

$$= \frac{1}{n} \cdot n \cdot \mu = \boldsymbol{\mu}$$

שונות:

$$Var(Y) = Var[\frac{1}{n} \cdot \sum Y_i]$$

$$= \frac{1}{n^2} \cdot Var[\sum Y_i]$$

מכיוון ש$Y_i$ בלתי תלויים:

$$= \frac{1}{n^2} \cdot \sum Var[Y_i]$$

$$= \frac{1}{n^2} \cdot n \cdot \sigma^2 = \boldsymbol{\frac{\sigma^2}{n}}$$

התוצאה מראה שהחיזוי של $Y$ הוא יותר טוב מחיזוי $Y_i$כלשהו, מכיוון שממוצע החיזויים שלנו יישאר $\mu$ כפי שציפינו.

כלומר, בממוצע, $Y$ "יסכים" עם המסווגים היחידים. מכך נובע שבממוצע החיזוי יהיה מתאים לערך האמיתי.

והשונות מראה, שככל שמספר המסווגים היחידים ($n$) יגדל, השונות של החיזויים שלנו $Y$ תשאף לאפס.

השונות מייצגת את ההתפזרות של החיזויים. כאשר השונות שלהם קטנה, זה מראה שהחיזויים יותר מרוכזים סביב הערך האמיתי שלהם.

בנוסף, החיזוי הממוצע יהיה פחות מושפע מ$outliers$ ורעש.

החיזויים הרבה פחות מפוזרים ולכן זה מראה רמה גבוהה יותר של ביטחון בחיזוי.

**<u>סעיף ב'</u>**

התוצאות הן כאלה:

Decision Tree Training accuracy is 0.8147

Decision Tree Test accuracy is 0.838

Random Forest Training accuracy is 0.808

Random Forest Test accuracy is 0.843

למודל מסוג Random forest יש שגיאת אימון גדולה יותר מאשר עץ בודד יחיד.

עם זאת, למודל מסוג Random forest יש שגיאת בדיקה נמוכה יותר בקצת. אין הבדל גדול בין השגיאת טסט של שני המודלים.

התוצאות האלו מראות לנו שRandom forest מצליח להכליל יותר מאשר עץ בודד יחיד.

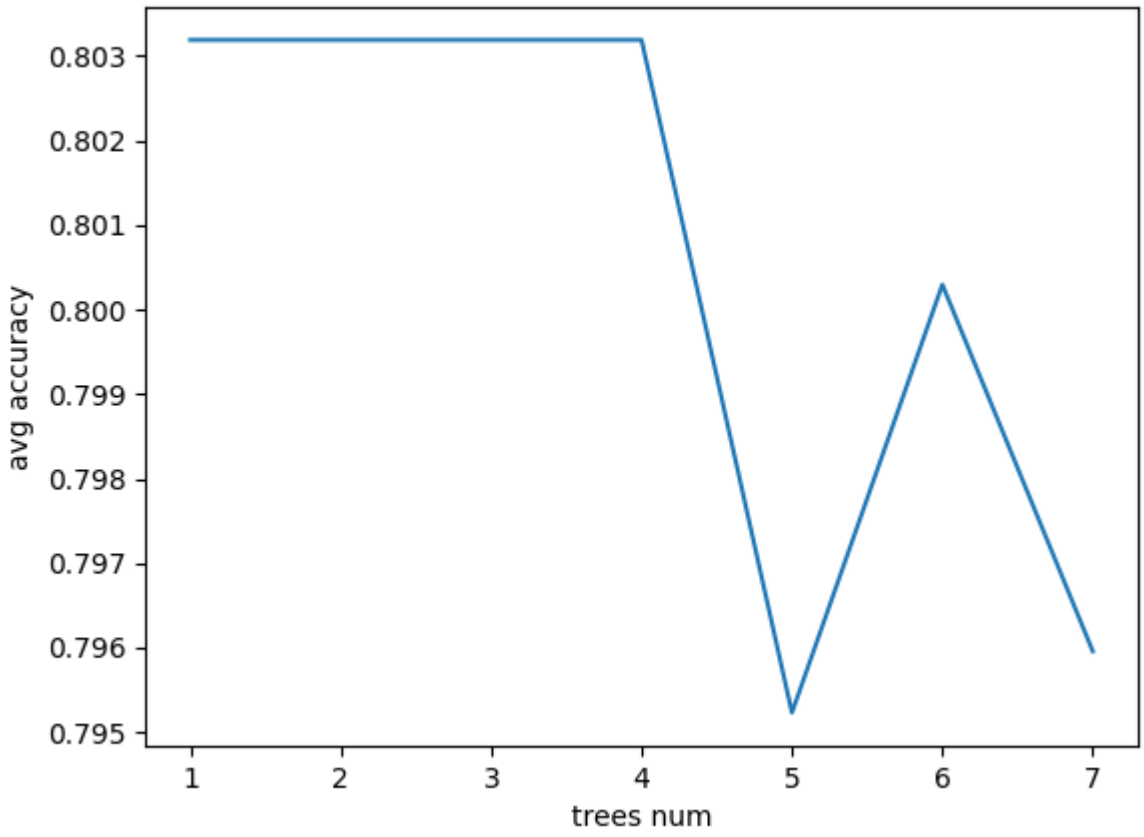
לעץ יחיד יש אחוז הצלחה גבוה ביותר באימון, מה שמראה שהוא יותר מתאים לtraining data, ומתקרב לoverfitting.

עץ יחיד לא מבצע הכללה, בהינתן הדאטה של האימון הוא יעשה את הכי טוב שאפשר, מה שלאחר מכן כשנבדוק על דאטה שלא ראינו לפני, נקבל פתאום תוצאות פחות טובות. דבר זה נקרא overfitting. לעומת זאת, בRandom forest שגיאת האימון גדולה יותר, אך לאחר מכן אחוז הדיוק עבור דאטה שעוד לא ראינו (Test set) הוא טוב יותר. ולכן נוכל להגיד שהוא יותר טוב בהכללה (generalization) ולכן יקרה לו פחות overfitting.

לגבי קורולציה, Random forest עמידים יותר עבור פיצ'רים שיש ביניהם קורולציה, בזכות היכולת שלהם לקחת חלק רנדומלי של פיצ'רים עבור כל עץ. זה מאפשר להם לטפל בקורולציות ופיצ'רים מתאימים בצורה יעילה יותר. עם זאת, אי אפשר לזהות זאת ישר מהתוצאות הללו.

**סעיף ג'**

כך נראה הגרף שקיבלנו:




על פי הגרף, מספר הestimators הטוב ביותר מבחינת דיוק הוא 4־1.