פרויקט בלמידה חישובית

מגישים:

אביב זבולוני – ת.ז. 211313333  
אליה אטלן – ת.ז. 318757200

**מבוא:**

פרויקט זה עוסק בחיזוי מחלות. האלגוריתם שלנו מנסה לחזות 4 מחלות שונות:

1) הסיכוי של בן אדם לחלות בסכרת – [קישור לדאטא](https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset)  
2) הסיכוי בן אדם לחלות במחלת לב כלשהי – [קישור לדאטא](https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction)  
2) הסיכוי של בן אדם לחלות בסרטן הריאות – [קישור לדאטא](https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link)  
1) הסיכוי של בן אדם לחלות בקורונה – [קישור לדאטא](https://www.kaggle.com/datasets/meirnizri/covid19-dataset)  
  
באמצעות השילוב של ארבעה קבוצות נתונים הנ"ל, האלגוריתם מביא למשתמש 4 פלטים, כל פלט מחזיק את הסיכוי עבור כל מחלה.  
  
אבל, כל דאטא מתאים לקבוצת גילאים אחרת, ולכן האלגוריתם לא יחזה את הסיכוי של בן אדם לחלות בכל הארבעה, אלא כל בן אדם יוכל לקבל חיזוי עבור המחלות שמתאימות לתחום הגילאים שלו.

בכל דאטא נשתמש ב 4 אלגוריתמים:

* One Vs One
* K Nearest Neighbors: KNN
* Gaussian Mixture: GMM
* Adaboost על עצי החלטה בעומק 1 (Stumps)

עבור כל אלגוריתם נמדוד:

* הדיוק של One Vs One באמצעות משחק עם הפרמטר C שמגדיר את סיבוכיות המודל, ועם הפרמטר degree שמגדיר את הדרגה של הפולינום של המודל שלנו (Linear Regression)
* ה K הכי אופטימלי עבור KNN, והשוואת הדיוק של KNN אל מול האלגוריתמים האחרים באמצעות השוואת ה F1 Score.
* היכולת של GMM למצוא את מספר המחלקות הנכון
* היכולת של Adaboost לשפר את הדיוק של המודלים הקיימים (נשתמש ב Stumps)

**הדאטא:**

1. **סכרת:**זה קובץ דאטא עצום (עם 250 אלף דוגמאות). בקובץ זה ישנם 3 מחלקות - 0 זה אומר שאין סכרת, 1 זה סיכון לחלות, ו 2 זה אומר שהבן אדם חולה בסכרת.  
   זה גם הקובץ הכי מסודר שמצאנו, מכיוון שאין שם נתונים בטקסט (F,T למשל), הכל במספרים.

הדאטא הזה מתאים לילדים בגילאים 1-13, ולכן המודל יכול לחזות את הסיכוי בלחלות בסכרת רק לקבוצת הגילאים הזאת.

לאחר בדיקת שונות על העמודות, לא מצאנו עמודה עם שונות שהיא 0, ולכן נשאיר את הדאטא כמו שהוא, ונעבוד על הגרסא המנורמלת

1. **מחלת לב:**

בדאטא הזה יש רק 913 דוגמאות, לכן התוצאות יהיו פחות מדוייקות מהדאטא של הסכרת. הפעם יש רק 2 קבוצות – האם לבן אדם יש מחלת לב כלשהי, או לא. אבל יש הרבה עמודות עם מלל, לכן נהפוך אותן למספרים – למשל עמודה שמכילה את הערכים {‘Flat’, ‘Up’, ‘Down’} תתורגם ל {0,1,2}עבור כל ערך בנפרד בטבלה.   
תחום הגילאים הוא 28-77.

1. **סרטן ריאות:**

בדאטא הזה יש 1000 דוגמאות. כל התכונות הם מספרים (למעט ה ID של הנבדק שאותו הסרנו). עמודת ה Y מתוארת על ידי Low, Medium, High וזה הסיכוי לחלות בסרטן ריאות. החלטנו להמיר את Low ל 1 (לא חולה) ואת Medium, High ל 1-.

תחום הגילאים הוא 14-73.

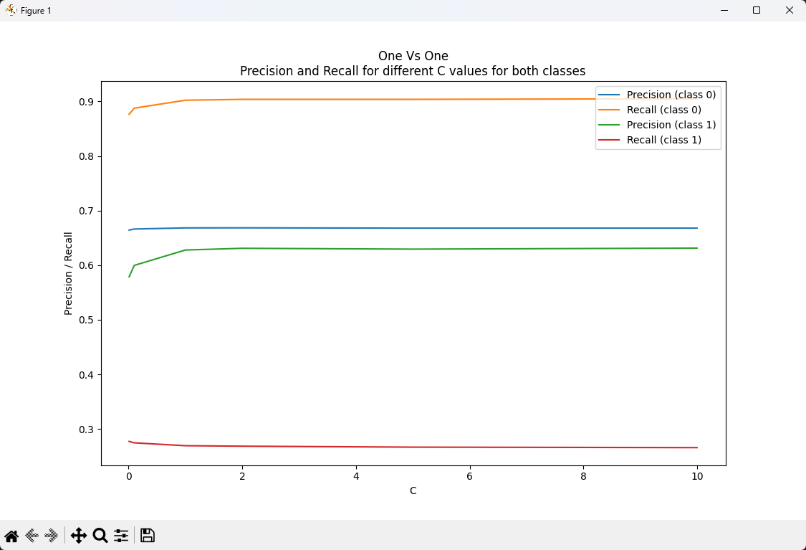
1. **קורונה:**

בדאטא הזה יש מליון(!) דוגמאות, מזערנו את הדאטא על ידי לקיחת 5000 שורות אקראיות.

הדאטא בא נקי וללא מלל ולכן נוכל להשתמש בו כמו שהוא (למעט השורה של התאריך שהסרנו).

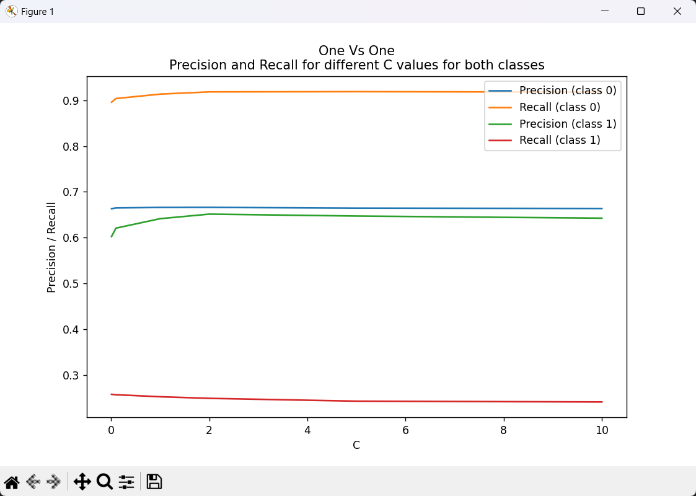
תחום הגילאים הוא 0 ועד 99. לכן נוכל לבדוק קורונה בשילוב עם כל המחלות האחרות לכל בן אדם.

**הרצת האלגוריתמים והסקת מסקנות - קורונה:**

ראשית נפריד את הדאטא לקבוצות של train, test.  
לאחר מכן נריץ One Vs One, ונחפש את הקבוע C שמביא את התוצאות הכי טובות. התוצאות הכי טובות נמדדות כ Precision, Recall. נציג אותם בגרף הבא:

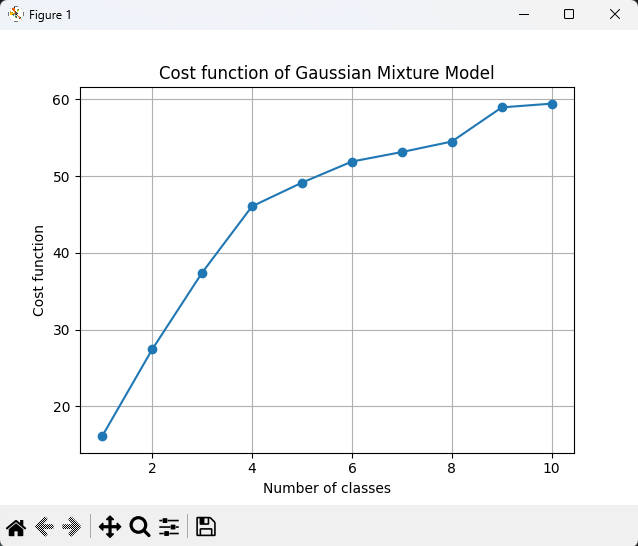
ניתן לראות כאן שה C האופטימלי הוא C = 1. כלומר סיבוכיות המודל שמתקבלת כתוצאה

מהדרישה:  היא אופטימלית עבור הדאטא הזה.

נשים לב שה recall הוא נמוך עבור מחלקת החולים אבל גבוה עבור מחלקת הבריאים.  
ננסה לשפר את הדרגה של התכונות ונקבל:

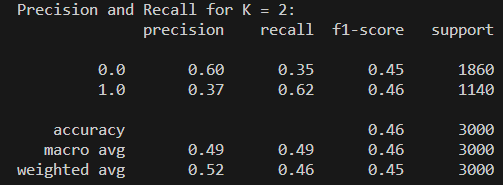
קיבלנו שיפור מזערי של ה precision של מחלקה 1.

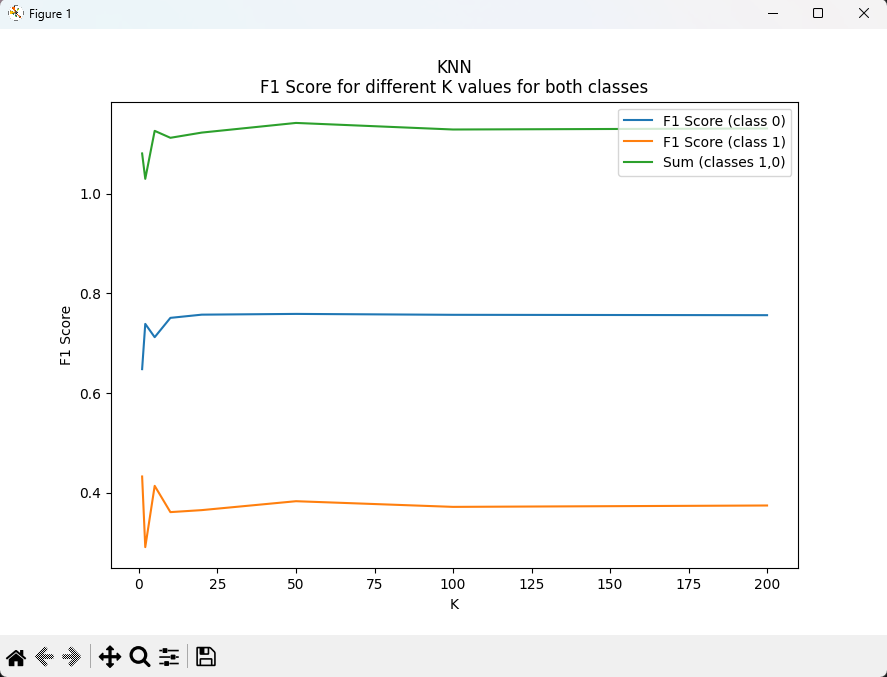
לכן במקרה הזה התכונות לא משפיעות הרבה על התוצאה.



נבדוק את אלגוריתם ה GMM:

נשים לב שהקפיצה הכי גדולה של פונקציית המחיר היא כאשר K = 2, ולכן מספר המחלקות האופטימלי הוא 2 (כפי שצפוי).

נבדוק את ה Recall, Precision של GMM עם K = 2:   
  
האלגוריתם הגיע לתוצאות פחות טובות גם עבור מחלקת הבריאים



נבדוק את אלגוריתם ה KNN:  
הפעם נשווה את ה F1 Score

נשים לב שה K הכי טוב פה הוא K = 50. הקו הירוק מיצג את הסכום של שני ה scores כדי שנוכל להסיק יותר בקלות את ה K הכי טוב.

בדומה לאלגוריתם ה One Vs One, החיזוי של מחלקת הבריאים הוא מדויק ודי טוב, אבל לא טוב אל מול מחלקת החולים.

נריץ Adaboost: