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)  
3) הסיכוי של בן אדם לחלות ב hypothyroid (תפקוד לא תקין של בלוטת התריס) – [קישור לדאטא](https://www.kaggle.com/datasets/yasserhessein/thyroid-disease-data-set)  
4) הסיכוי של בן אדם לקבל שבץ – [קישור לדאטא](https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-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.

**הדאטא:**

1. **סכרת:**זה קובץ הדאטא הכי גדול שמצאנו (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, הכל במספרים (F,T למשל).

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.