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. **Hyperthyroid:**

בדאטא הזה יש 3700 דוגמאות. הרוב המוחלט של התכונות הוא עם מלל במקום מספרים, לכן נמיר באופן דומה לדאטא של המחלת לב. בדאטא הזה ישנם גם סימני שאלה עבור ערכים שלא נמדדו. החלטנו להעיף את כל השורות שאחת התכונות שלהן היא סימן שאלה, ויצא שהעפנו 1000 דוגמאות.   
בנוסף העפנו 6 תכונות שהשונות שלהן היא 0.

תחום הגילאים הוא 1 עד 94, לכן נוכל לבדוק Hypothyroid עם מחלת לב או סכרת בשביל בן אדם מסויים.

1. **Stroke:**

בדאטא הזה יש 5000 דוגמאות, הורדנו 200 מהם בגלל שורות שבהן חסר מידע בתכונה של BMI. העפנו את התכונה של ה ID של הבן אדם כי היא לא מעניינת.

תחום הגילאים הוא 0.08 ועד 82. לכן נוכל לבדוק גם Stroke, גם Hypothyroid וגם סכרת או מחלת לב לבן אדם בתחום הגליאים המתאים