**הצעת מחקר – חיזוי שבץ מוחי באמצעות למידת מכונה על סמך מאגר רפואי**

**מגישים:**  
לירון כהן – 312324247

עומר אפטר - 206766677

**המאגר:**  
הפרויקט מבוסס על מאגר הנתונים **Stroke Prediction Dataset** מאתר**Kaggle** .

המאגר כולל מידע רפואי, דמוגרפי והתנהגותי על כ־5,110 מטופלים, ומתעד האם כל מטופל עבר שבץ מוחי (stroke = 1) או לא.  
המאגר כולל משתנים כגון: גיל, מין, לחץ דם, מחלת לב, רמת גלוקוז בדם, BMI, מצב עישון, מצב משפחתי, אזור מגורים וסוג עבודה.

במטרה לחזות את הסבירות לאירוע מוחי (stroke).

המשימה היא **לסווג מטופלים לפי הסיכון לשבץ** באמצעות אלגוריתמים שונים של למידת מכונה, תוך ניתוח הנתונים, השוואת דיוק המודלים, ובדיקת טכניקות נוספות כמו הפחתת ממדים.

**שאלות המחקר:**

1. **אילו משתנים רפואיים ודמוגרפיים משפיעים ביותר על הסיכון לשבץ מוחי, וכיצד ניתן לזהות אותם מתוך הנתונים?**
2. **השוואת ביצועים בין מודלים שונים של למידת מכונה:**

מהם הביצועים (דיוק) של אלגוריתמים שונים בלמידת מכונה עבור סיווג מטופלים שעברו או לא עברו שבץ מוחי?

כיצד משתנה רמת הדיוק של אלגוריתמים שונים כגון Logistic Regression, KNN, Decision Tree, Adaboost ו־ SVM כאשר הם מופעלים על הנתונים במטרה לחזות שבץ מוחי?

1. **כיצד ניתוח חשיבות תכונות (Feature Importance) יכול לשפר את הבנת המודל ואת יעילות הסיווג?**

כמה מתוך המטופלים עם מחלת לב או יתר לחץ דם קיבלו שבץ מוחי בפועל?

1. **כיצד מתפלגים מקרי השבץ בקבוצות שונות באוכלוסייה, למשל לפי לחץ דם, מחלת לב או מצב עישון?**
2. **באיזו מידה ניתן להשתמש בהפחתת ממד (PCA) כדי לבצע סיווג יעיל ומהיר מבלי לפגוע בדיוק?**

**הכלים והשיטות:**

* **עיבוד מקדים (Preprocessing):**
  + **טעינת הנתונים**

קריאת קובץ stroke\_dataset.csv באמצעות pandas.

* + **ניקוי נתונים**

טיפול בערכים חסרים (במיוחד ב־BMI).

הסרת מזהה (id) שאינו תורם ללמידה.

* + **קידוד משתנים קטגוריים**
  + המרה של משתנים קטגוריים כמו gender, work\_type, smoking\_status למספרים באמצעות replace(). המרה של משתנים קטגוריים ל־Dummy Variables
  + **נירמול** (Standardization)

נירמול של משתנים המספריים באמצעות StandardScaler

* + **בחירת תכונות משמעותיות**

שימוש ב־ExtraTreesClassifier לזיהוי המשתנים המשפיעים ביותר על סיכון לשבץ (Feature Importance).

* **מודלים לבחינה:**
  + **Logistic Regression**
  + **K-Nearest Neighbors (KNN)**
  + **Decision Tree**
  + **AdaBoost**
  + **Support Vector Machine (SVM)**
* **ניתוח חשיבות משתנים:**
  + שימוש ב־ExtraTreesClassifier כדי לזהות משתנים קריטיים
* **הפחתת ממדים:**
  + PCA (Principal Component Analysis) לשימור 95% מהשונות
  + השוואת ביצועים לפני ואחרי PCA

**הערכת ביצועים:**

* נעשה שימוש במדד דיוק (Accuracy) ככלי מרכזי להערכת ביצועי המודלים.
* הדיוק חושב עבור כל אחד מהאלגוריתמים: Logistic Regression, KNN, Decision Tree, AdaBoost ו־SVM.
* התוצאות הוצגו בצורה כמותית (גרף השוואה בין האלגוריתמים) וכן נכתבה השוואה מילולית לניתוח חוזקות וחולשות של כל מודל. תוצאות הדיוק הוצגו בגרף השוואתי ברור, הממחיש את ההבדלים בין המודלים.
* בוצע ניתוח של חשיבות משתנים (Feature Importance) באמצעות ExtraTreesClassifier.
* בוצע השוואה בין ביצועים לפני ואחרי הפחתת ממד (PCA), במטרה לבחון האם ניתן לשפר את מהירות המודלים תוך שמירה על רמת דיוק סבירה.

**סיכום:**

מטרת המחקר היא להשוות בין מודלים שונים של למידת מכונה על בסיס ביצועיהם במאגר ,Stroke Prediction לזהות את הגורמים המשפיעים ביותר על הסיכון לשבץ מוחי, ולבחון שיפורים פוטנציאליים באמצעות ניתוח תכונות, הפחתת ממדים (PCA), ועיבוד מקדים לנתונים.  
תוצאות המחקר יספקו תובנות לגבי המודלים האופטימליים לסיווג מטופלים בסיכון לשבץ מוחי, ויתרמו להבנה טובה יותר של קשרים בין משתנים רפואיים לבין הופעת שבץ.