<https://colab.research.google.com/drive/1gL1YBxB3y2bT9K8sySth1oNnn_SwKHPe?usp=sharing>

**מטלה מספר 1**

**Data Preparation**

**קורס כריית נתונים**

**ספריות שהשתמשנו:**

Pandas, NumPy, Seaborn, matplotlib.pyplot, sklearn.preprocessing

* תחילה ביצענו פעולות פשוטות- כדי להבין את סדר הגודל של ה-Dataframe, כמה ערכים חסרים, איזה טיפוסים קיימים בו וכדומה.
* בחרנו לא להציג ויזואליציה למידע בשלב הזה עקב כך שה-Dataset גדול מידי והנתונים לא יהיו מתאימים לקריאה ולהבנה.
* לאחר מכן יצרנו Dataframe חדש שמכיל משתנים (עמודות) שיסייעו לנו ביצירת מודל למידת המכונה באופן היעיל ביותר.
* דבר ראשון, החלפנו את משתנה המטרה "ACTIVITY" לערכים בינאריים (כאשר Bitter=1, Non-Bitter=0).
* **תכונות שבחרנו לא להוסיף לטבלה החדשה הן:**
* כל התכונות מסוג “object” – השארנו את כל המספרים והבוליאנים.
* כל העמודות שיותר מ-50% מהערכים שלהן זהים.
* כל העמודות שיותר מ-5% מערכיהן הם null (מצאנו שכבר אין כאלה לאחר הסעיפים הקודמים).
* בעזרת מטריצת קורלציה, אחד מכל זוג של משתנים בעלי קורלציה גבוהה (התאמה גבוהה מ-0.8).
* בעזרת מדד IQR, כל המשתנים שמעל 10% מהערכים שלהם הם חריגים.
* סיננו שורות מהטבלה שיותר מ-20% מהמידע שלהן הוא outlier, באמצעות מדד z-score.
* כך קיבלנו ב-Dataframe החדש את המשתנים האידיאליים ביותר מתוך ה-Dataset הראשוני.
* הצגנו את ה-Dataframe באמצעותboxplots והיסטוגרמות, מכיוון שכל המשתנים שנותרנו איתם הם מספריים
* בסוף יצרנו טבלה המציגה סטטיסטיקה תאורטית לכל משתנה (טווח ערכים, ממוצע וסטיית תקן).

לא הצגנו תדירות משתנה בינארי, מאחר ולא נותרו לנו משתנים כאלו (מלבד משתנה המטרה ("ACTIVITY"()

* לבסוף, נרמלנו את הנתונים באמצעות standard scaler.