**פרק 1 : מטריקות.**

קרא את המאמר המצורף:

20 Popular Machine Learning Metrics. Part 1: Classification & Regression Evaluation Metrics

ולאחר מכן פתור את התרגיל הבא:

בתיקיה זו יש קובץ בשם “CANCER\_TABLE” , בקובץ זה יש 1000 רשומות.  
כל רשומה מייצגת ביופסיה שנלקחה מנבדקת. העמודה הראשונה בטבלה מייצגת את הקוטר של הביופסיה בס"מ, והעמודה השנייה מייצגת האם הנבדקת חלתה בסרטן בתקופה של עד שנה לאחר מועד ביצוע הביופסיה.  
חברת "פלא-סיינס" הציע מודל שלטענתה עובד נפלא: "אם קוטר הגידול מעל 7 ס"מ אז הנבדק יחלה בסרטן בשנה הקרובה ואחרת לא". בתרגיל זה ננתח ביצועי מודל זה.

1. חשב והצג confusion matrix עבור המודל.
2. חשב FP, TP, FN, TN.  
   מה המשמעות העסקית של מדדים אלו עבור בעיה זו?  
   נסח אותם בצורה כזו שגם לקוח שאיננו מבין בתחום יוכל להבין את משמעותם.
3. חשב TPR, FPR עבור המודל וספק משמעות עסקית.
4. חשב את הדיוק (accuracy), הנכונות (precision) והכיסוי (recall) של המודל.
5. חשב את מדד F1 של המודל.

לאחר ניתוח המודל שביצעת, חברת "פלא-סיינס" הציעה שדרוג למודל: במקום מודל מסווג, החברה תספק מודל שידרג את הביופסיות לפי סבירותן לייצג חולה סרטן (בשנה הקרובה). ביופסיה עם דירוג גבוה יותר מייצגת סיכוי גבוה יותר שהיא נלקחה מאדם שיחלה בסרטן בשנה הקרובה.

1. למה מודל כזה יכול להיות טוב? שווק אותו ללקוח.
2. שרטט עקומת ROC ללא שימוש ב-scikit-learn או כל פונקציה מובנית אחרת של פייתון שעושה זאת.
3. חשב את ה-AUC.  
   האם המודל טוב מבחינת קריטריון זה?  
   תן משמעות עסקית לגודל שחישבת שתהיה ברת הבנה ללקוח.
4. הפוך את מודל הדירוג למודל סיווג באמצעות עקומת ה-ROC ששרטטת.  
   הסבר היטב את תשובתך.
5. חשב confusion matrix עבור המודל החדש.
6. חשב accuracy, precision, recall, TP, FP, TN, FN, TPR, FPR עבור המודל החדש.
7. השווה המדדים שחישבת בסעיף הקודם לאלו של המודל הראשוני.  
   על איזה מודל היית ממליץ?
8. חזור על סעיף 7 כאשר אתה עושה שימוש בפונקציות של ספריית scikit-learn תוך כתיבת קוד מינימלי מצידך.