**פרק 1 : מטריקות.**

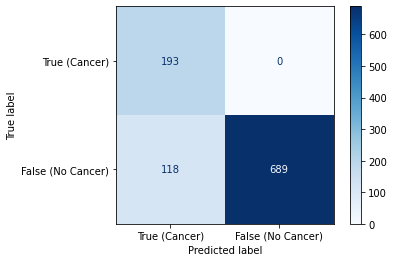
קרא את המאמר המצורף:

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.



1. חשב FP, TP, FN, TN.

TN – כל הפרדיקציות השליליות עבור תוצאות שליליות – 689

FN – כל הפרדיקיות השליליות עבוא תוצאות חיוביות – 0

TP – כל הפרדיקציות החיוביות עבור תוצאות חיוביות – 193

FP – כל הפרדיקיות החיוביות עבור תוצאות שליליות – 118  
מה המשמעות העסקית של מדדים אלו עבור בעיה זו? נסח אותם בצורה כזו שגם לקוח שאיננו מבין בתחום יוכל להבין את משמעותם.

המשמעות העסקית היא שמודל שמסווג אנשים חולים ולא חולים חשוב שלא יהיו לו FN כדי לא לפספס חולים, גם נרצה ש ה FP יהיה כמה שיותר נמוך כדי לא לגרום לחולים טיפולים שהם לא צריכים. וכמו בכל מודל נרצה כמה שיותר תשובות נכונות...  
חשב TPR, FPR עבור המודל וספק משמעות עסקית.

TPR - TP / (TP + FN)= 193/(193 + 0) = 1

FPR - FP / (FP + TN) = 118/(118+689) = 0.15

1. חשב את הדיוק (accuracy), הנכונות (precision) והכיסוי (recall) של המודל.

Accuracy –( (193 + 689)/1000)\*100 = 88.2%

Precision – TP/(TP + FP)

Precision-cancer = 193/(193+118) = 62%

Precision-non-cancer = 689/(689+0) = 100%

Recall – TP/(TP+FN)

Recall-cancer = 193/(193 + 0) = 100%

Recall-non-cancer = 689/(689+118) = 85%

1. חשב את מדד F1 של המודל.

F1-score-cancer= 2\*Precision\*Recall/(Precision+Recall) = 2\*0.6\*1/(0.6+1) = 76.5%

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.

מודל כזה יכול להיות טוב בגלל שהוא ייצג את הסיכוי ביחס לכל הדגימות ולא רק ביחס לדגימה אחת וthreshold קבוע.

1. שרטט עקומת ROC ללא שימוש ב-scikit-learn או כל פונקציה מובנית אחרת של פייתון שעושה זאת.
2. חשב את ה-AUC.

Auc = 0.5  
האם המודל טוב מבחינת קריטריון זה?

תן משמעות עסקית לגודל שחישבת שתהיה ברת הבנה ללקוח.

לא, עם ערך auc כזה המודל שווה ערך להחלטה רנדומלית לגבי הדאטה.

1. הפוך את מודל הדירוג למודל סיווג באמצעות עקומת ה-ROC ששרטטת.  
   הסבר היטב את תשובתך.
2. חשב confusion matrix עבור המודל החדש.
3. חשב accuracy, precision, recall, TP, FP, TN, FN, TPR, FPR עבור המודל החדש.
4. השווה המדדים שחישבת בסעיף הקודם לאלו של המודל הראשוני.  
   על איזה מודל היית ממליץ? המודל החדש
5. חזור על סעיף 7 כאשר אתה עושה שימוש בפונקציות של ספריית scikit-learn תוך כתיבת קוד מינימלי מצידך.

סעיפים 8-11 נמצאים במחברת