בע"ה

למידת מכונה ולמידה עמוקה בדיאגנוסטיקה קלינית - **פרוייקט גמר**

אוראל כרמיאל 318912862

הקוד: https://colab.research.google.com/drive/1DUvNDU6I5C43L1kyOQC9MLQFhiqafr-Y?usp=sharing

הורדתי וחילצתי את תיקיית הנתונים, מחקתי שורות לא רלוונטיות, עריכתי כותרות והמרתי את סוג הדאטה מאובייקט לעשרוני. לבחירת מאפיינים, השתמשתי בסטיית תקן כדי לבחור גנים עם השונות הגדולה ביותר והסרתי גנים עם מתאם גבוה, השארתי 424 גנים. בניתי שני מודלים מסווגים, SVM ו-RF, עם פרמטרי ברירת מחדל. לאחר מכן השתמשתי באימות צולב כדי למצוא את הפרמטרים הטובים ביותר עבור כל מודל (פירוט עליהם ניתן למצוא בהערות על הקוד). בשל כמות הנתונים הקטנה, נעלתי את הפונקציה train_test_split על זרע ספציפי לשמירה על עקביות. לבסוף, הערכתי את הביצועים של שני המודלים באמצעות מטריצת הבלבול, מדדירגישות ודיוק, F1 ועקומת ROC.

ככלל שני המודלים הגיעו לתוצאות די טובות בכל המדדים. כל אחד מהם מצביע על רגישות מסויימת במודל ובמיוחד נבחן את הRecall כדי לא לפספס את בעלי סיכוי לחלות.
בכל המדדים הSVM השיג תוצאות טובות יותר מהRF אך לא באופן משמעותי. יער רנדומלי הוא שיטת אנסמבל שמבטיחה חיזוי כמעט מדוייק, אך SVM מתמודד טוב יותר מאשר RF כאשר מספר התכונות גבוה או כאשר מספר הדוגמות קטן כמו בדאטה שלנו.

לסיכום ולאחר שקראתי קצת באינטרנט אני חושבת שמודל הSVM הוא המוצלח ביותר עבור מחקר כמו שלנו ובאופן ראשוני אפשר להעזר בו. אם כי אני מסתייגת וטוענת שהמדגם הנוכחי קטן ולא הייתי סומכת על החיזוי שלו באופן מלא. לו הייתי חוקרת בתחום בהחלט הייתי רוצה להמשיך למחקר עומק נוסף עם מדגם גדול יותר. כחלק ממחקר זה הייתי גם משתמשת ביתרונו של היער הרנדומלי לחקור את הגנים שנבחרו להוות נקודת פיצול בעץ, כנראה הם משמעותיים לחיזוי האפשרות לחלות בהתקף לב.

לא נדרש (כתבתי תוך כדי עבודה):

פרוטוקול בניית המודלים:

<u>עיבוד מקדים:</u>
הורדתי את תקיית הנתונים מהמודל וחילצתי אותה.
העלתי את הקובץ המחולץ לדרייב.
מחקתי את 59 השורות הראשונות- הן לא היו חלק מהדאטה עצמה.
ערכתי את כותרות השורות והעמודות כך שיהיו אינפורמטיביות.
שיחלפתי את הטבלה באופן שהדוגמות יהיו בשורות והגנים בעמודות, מקובל שהמאפיינים הן העמודות,
בנוסף הזזתי את עמודת החיזוי לסוף כמקובל.
בשל העובדה שכמות המאפיינים די גדולה בחרתי להשמיט את מעט (בסך הכל 0.08% מכלל הגנים) הנתונים החסרים במקום להשלים אותם.
כאשר קלטתי את הטבלה הנתונים בה היו מסוג אובייקט, המרתי את כולם למספר עשרוני מלבד נתוני החיזוי

בע"ה

<u>בחירת מאפיינים:</u>
כשיטה ראשית לבחירת מאפיינים בחרתי להשתמש בסטיית תקן.
בחרתי בשיטה זו מכמה סיבות,
היא פשוטה להבנה וחישוב ואינה כוללת אלגוריתמיקה מורכבת,
היא שומרת על הערכים המקוריים כלומר לא מפעילה מוניפולציות כלשהן על הנתונים אלא שומרת אותם בערכים הטעיים שלהם וכך יהיה ניתן בהמשך להעמיק את המחקר אודות גנים מסויימים (מה שיכול להוות חיסרון בשיטות כמו PCA). בנוסף כאשר אנו עובדים עם מידע ביולוגי ובפרט ביטוי גנים נרצה באמת לתפוס את הגנים בעלי השונות הגדולה בין הנבדקים כדי ללמוד על ההבדל בין קבוצות הנבדקים השונות.
לאחר סינון משמעותי של כמות הגנים על ידי בחירת שונות גבוהה החלטתי גם להסיר גנים בעלי קורלציה גבוהה, בעת בניית המודל אין משמעות לשני גנים בעלי אותה התנהגות וניתן לבחור רק את אחד מהם.
נותרתי עם 424 גנים.

יש שיאמרו שביחס לנתוני המקור נשארתי עם מעט מידע, אך אני סבורה כי גם בהקשר הביולוגי וגם בהקשר החישובי זוהי כמות סבירה.
מבחינה ביולוגית, אנו מתמקדים במערכת הדם אותה מרכיבים מספר מסלולים ביולוגים ספציפיים, ודאי כי 400 הגנים שנבחרו בצורה מושכלת יכולים ללמד אותנו באופן ממוקד על הסיכוי לחלות בהתקף לב
מבחינה חישובית, בניסוי הספציפי הזה יש בידינו מעט מאוד דגימות מה שעלולו להביא להתאמת יתר בקלות, המודל עשוי למצוא מתאמים מזויפים בין המאפיינים למשתנה היעד, שאינם מכלילים היטב לנתונים חדשים אפשריים. לכן נרצה להקטין את כמות הגנים שלנו במידת האפשר. יתרון נוסף יהיה זמני ריצה יותר מהירים בעבודה עם מערך נתונים קטן יותר.

<u>בניית המודל:</u>
ראשית אקדים ואומר שמלבד זוג המודלים המופיעים בקוד הרצתי בטיוטה עוד כמה סוגי מסווגים כדי לקבל תחושה אילו מודלים יתאימו ביותר לניסוי זה, עקב הדרישה ל2 מודלים הסרתי אותם מהקובץ הסופי.
נשים לב שמשתנה המטרה הינו קטגורי ולכן נשתמש במודלים מסווגים ולא רגרסורים.
פיצלתי את הנתונים לקבוצת אימון וקבוצת מבחן.
בחרתי את המסווגים SVM - מכונת וקטורים תומכים וRF - יער אקראי.
עבור כל מודל אימנתי מודל עם היפרפרמטרים בסיסיים (ברירת מחדל)
וחישבתי ציון דיוק על מנת לקבל הבנה כללית על איכות המודל
לאחר מכן עבור כל מודל בניתי מילון היפרפרמטרים מתאים ובעזרת Cross-validation מצאתי את ההיפרפרמטרים הטובים ביותר לנתונים אלו.
הערה חשובה: כמו שציינתי בשלב בחירת המאפיינים היות וכמות הדגימות קטנה שלב הלמידה עלול לסבול מכך ולהחזיר תוצאות לא אמינות. אכן כאשר מריצים את הקוד שכתבתי יכולות להקבל תוצאות שונות לחלוטין אחריי כל הרצה (במידה הדאטה הייתה גדולה יותר משמעותית היו הריצות מתכנות לאותה תוצאה). ב"עולם האמיתי" ישנן כמה דרכים לפתור בעיה זו; הוספת משתתפים לניסוי, הכפלת שורות באופן רנדומלי או מבוקר על מנת ליצור דאטה מנופחת יותר ועוד. במטלה זו היה עלינו להתמקד בעיקר בבניית המודלים עצמם ולכן לאחר כמה ניסויים בחרתי לנעול את הפונקציה train_test_split האחראית על חלוקת הנתונים על זרע ספציפי (random_state=42) על מנת להיות עקבית לאורך ההרצות והערכת הביצועים (יתכן שעבור זרע שונה יתקבלו ביצועים שונים ובעקבות כך מסקנה שונה).

<u>הערכת ביצועים:</u>
את כל המדדים תמיד בדקתי עבור שני המודלים והדפסתי אותם זה לצד זה.
בתחילה בחרתי במטריצת בלבול, מכיוון שהיא גם מודפסת באופן חזותי היא מצליחה להסביר באופן אינטואיטיבי את רמת ההצלחה של המודלים.
לאחר מכן חישבתי את המדדים הקלאסים: Recall, Accuracy, Precision, F1 Score, ROC AUC
ולבסוף הדפסתי את עקומת ROC באופן ויזואלי.