ניתוח מידע מדעי - תרגיל 2

כתבו קובץ פיית'ון בשם **ex2.py**. בתוכו ממשו את הפונקציה **analyse_s2_stats_ex2**. הפונקציה עובדת על הקובץ **S2.csv** (ראו באתר הקורס תחת חומר למידה מסוג: תרגיל כיתה). ממשו בפונקציה את הפעולות הבאות (כמובן, ניתן לממש בקובץ פונקציות נוספות ולקרוא להן):

1. טענו הקובץ לאובייקט DataFrame בשם df המכיל את כל העמודות.
   הערה: בשלב זה מומלץ לוודא שכל הקובץ נקרא במלואו, כלומר בדקו שמספר השורות באובייקט ה df שיצרתן תואם לציפייה.
2. דווחו את מספר הערכים החסרים (כלומר, מספר השורות עבורן הערך הוא 0, במקרה הזה) ואת ה missing rate (כלומר המספר שחישבנו לחלק למספר השורות בטבלה המקורית), עבור העמודות הבאות.
   - PWM_ref
   - MES_ref
   - NNSplice_ref
   - HSF_ref
   - GeneSplicer_ref
   - GENSCAN_ref
   - NetGene2_ref
   - SplicePredictor_ref

   להלן טבלת התוצאות שעליכן לקבל (הדפיסו את ה missing rate עם דיוק של 3 ספרות אחרי הנקודה, כמו בטבלה). תוכלו להשתמש ב

Table 2. Missing rates of the prediction scores for eight *in silico* tools

Tool	No. of missing	Missing rate
PWM	77	0.026
MES	82	0.028
NNSplice	68	0.023
HSF	66	0.022
GeneSplicer	563	0.190
GENSCAN	2466	0.833
NetGene2	1887	0.638
SplicePredictor	2252	0.761

3. מצאו את הדגימות שהנן missing בלפחות אחת מארבע השיטות שהניבו missing rate שקטן מ 0.05. שמרו את האינדקסים של השורות שאינן missing בקובץ טקסט שנקרא non_missing_top_4.txt. במילים אחרות, הקובץ מכיל את מספרי השורות שלא מכילות 0 בכל העמודות: PWM_ref, MES_ref, NNSplice_ref, HSF_ref.

4. חלקו את הדאטה של טבלה S2 לשתי קבוצות: training - המהווה 90% מהדאטה ו test המהווה את 10% הנותרים. בחרו את השורות כך שתחילת הקובץ ילך ל train וסוף הקובץ ל test. במילים אחרות, אם למשל בקובץ היו 100 שורות, 90 השורות הראשונות היו ה train והיתר ב test. השתמשו ב assert על מנת לוודא שהחלוקה שבחרתן מכסה את כל הדאטה הנתון.השתמשו רק בדגימות שאינן missing (לפי הגדרת סעיף 3).

עבור כל אחת מהקבוצות (כל הדאטה, רק ה train ורק ה test) דווחו איזה חלק מתוך הקבוצה הוא positive ואיזה חלק הוא negative. דווחו את הכמות הן בספירות והן באחוז יחסי.

5. חלקו את הדאטה של טבלה S2 לשתי קבוצות: training - המהווה 90% מהדאטה, ו test המהווה את 10% הנותרים. דאגו לכך שיחס הדגימות החיוביות (positive) לעומת השליליות (negative) יהיה דומה בכל הקבוצות. השתמשו ב assert על מנת לוודא שהחלוקה שבחרתן מכסה את כל הדאטה הנתון. השתמשו רק בדגימות שאינן missing.

   הפעם דאגו לכך שיחס הדגימות החיוביות (positive) לעומת השליליות (negative) יהיה דומה בכל הקבוצות. השתמשו ב assert על מנת לוודא שהחלוקה שבחרתן מכסה את כל הדאטה הנתון.

   עבור כל אחת מהקבוצות (כל הדאטה, רק ה train ורק ה test) דווחו איזה חלק מתוך הקבוצה הוא positive ואיזה חלק הוא negative. דווחו את הכמות הן בספירות והן באחוז יחסי.

6. צרו את העמודה **PWM_ratio** המוגדרת בתור היחס בין העמודות PWM_ref חלקי PWN_alt.

   א. נגדיר בעזרת העמודה החדשה classifier המסווג דגימות כדלקמן: אם הערך גדול מ-1, הוא מסווג את המוטציה כ positive ואחרת, כ negative. דווחו מה ה-confusion matrix עבור קבוצת האימון (training set) וקבוצת המבחן (test set), עבור מסווג זה עבור הקבוצות מסעיפים 4 ו-5.

   ב. מצאו את הסף שמניב את ה TPR הגבוה ביותר עבור FPR שלא עולה על 0.1 עבור ה training set. דווחו את ה-confusion matrix עבור קבוצת האימון (training set) וקבוצת המבחן (test set), עבור מסווג זה עבור הקבוצות שנוצרו מסעיפים 4 ו-5.

7. בצעו ten-fold cross validation עבור סף של 1 (כמו בסעיף 6א') דווחו מה ה confusion matrix **הממוצעת** עבור 10 קבוצות האימון (training set) ו-10 קבוצות המבחן (test set), עבור מסווג זה.

8. חזרו על סעיפים 6+7 גם עבור השיטות: MES, NNSplice, HSF.

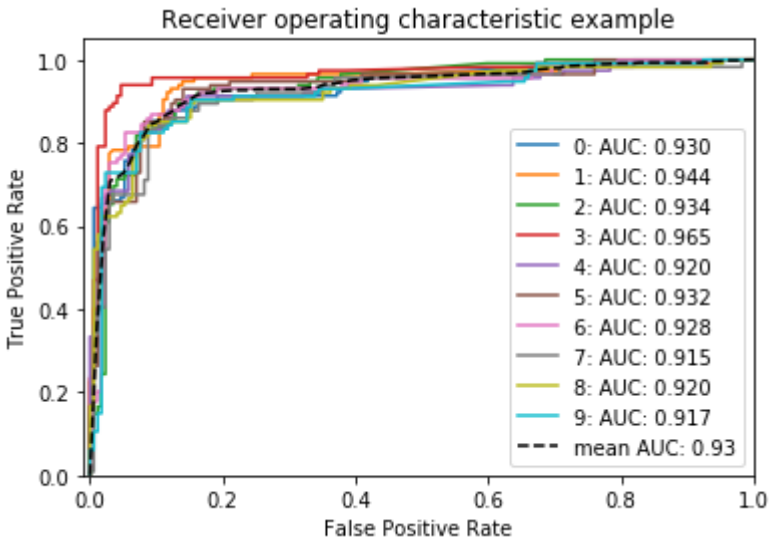
9. עבור השיטה PWM חשבו AUC וציירו את ה- ROC curve עבור קבוצה אחת של ה test.

10. עבור השיטה PWM חשבו AUC וציירו את ה- ROC curve בעזרת

10-fold cross validation

כלומר על אותו הגרף ציירו את ה ROC curve עבור ה test כל אחת מעשר האיטרציות. בנוסף ציירו את הגרף הממוצע מ-10 האיטרציות.

כלומר עליכן לקבל גרף דומה לגרף הזה:



11. עבור ארבע השיטות PWM, MES, NNSplice ו- HSF חשבו ROC curve וכן AUC על קבוצת test אחת. ציירו את ארבעת ה ROC curve על אותו הגרף. גם כאן יש להמנע משימוש בדגימות שהינן missing בלפחות אחת מהשיטות.

הסיקו מסקנה מהתוצאות: איזו שיטה היא הטובה ביותר לחיזוי האם מוטציה היא Positive או Negative?

הערות:

1. הגישו קובץ בשם ex2.py המכיל את קוד הפיית'ון שמימשתם (אם יש פונקציות עזר נוספות שמימשתן, כללו אותן בקובץ זה).
2. הוסיפו לפונקציה פונקציית main (כמו בתרגיל 1) הקוראת לפונקציה **analyse_s2_stats_ex2**. וודאו שהקוד רץ ללא שגיאות, מדפיס תוצאות ומייצר גרפים כנדרש.