תרגיל בית 3 - **Data science**

מועד הגשה: 24.5 בשעה 23:55

הנחיות כלליות להגשת תרגילי הבית **:**

- הגשת התרגילים תתבצע באמצעות אתר הקורס בלבד.
- יש להגיש את התרגיל לתיבת ההגשה המתאימה, בהתאם לפורמט הבא: **HW#_ID1_ID**2 , כאשר התו # מציין את מספר התרגיל, **ID**1מציין את תעודת הזהות של הסטודנט הראשון, **ID**2 מציין את תעודת הזהות של הסטודנט השני.

  לדוגמה:**HW1_000000000_111111111.py** זהו קובץ בודד שפותר את תרגיל בית 1.
- בנוסף, יש לכתוב כהערה (בתוך הקוד) בראש כל קובץ את השמות המלאים (שם פרטי ושם משפחה) בלועזית ואת תעודת הזהות של כל אחד מהשותפים, כאשר הפרטים של כל אחד רשומים בשורה נפרדת.
- ההגשה מתבצעת באמצעות קבוצות ההגשה בלבד.
- תרגיל אשר יוגש באיחור לא ייבדק, למעט אישורים מיוחדים כגון: מילואים, אשפוז וכו' המתקבלים לפני מועד ההגשה הרשמי.
- יש לכתוב הסברים ומענה מילולי על השאלות בגוף הקובץ בעזרת הערות ובאנגלית בלבד.
- יש להפריד בין פתרונות של כל שאלה בצורה הבאה. לפני המענה לשאלה מספר 1 כתבו בהערה **Q**1 וכן הלאה. במידה והשאלה מכילה מספר סעיפים (נניח 2 סעיפים) יש לכתוב בהערה **Q1a** לפני פתרון סעיף **a** בשאלה 1, ו- **Q1b** לפני פתרון סעיף **b** בשאלה 1 וכן הלאה.

בהצלחה!

תרגיל בית 3

Decision Trees

מצורף הקובץ votersdata.csv המכיל נתונים על בוחרים בארה״ב , כולל עמודת תוצאת הבחירה (רפובליקני/ דמוקרטי).

1. תחילה הגדירו RSEED עם הערך 123 והשתמשו בו בפיצול הנתונים ובאתחול המודל.
2. הכרת הדאטה: נסו להכיר את הקשרים בין המשתנים השונים לבין עמודת המטרה vote.
   2. א. שרטטו stacked bar plots עבור המשתנים הקטגוריים.
   2. ב. שרטטו multivariate boxplot עבור הנומריים.
3. תקנו את הנתונים במידת הצורך. זה כולל טיפול בערכים חסרים, ערכים לא תקינים, נרמול והמרות נדרשות.
4. חלקו את הדאטה באופן רנדומי ל-70% training set ו-30% test set בשימוש ה RSEED שהגדרתם:

```
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.33,random_state=RSEED)
```

5. בנו מודל עץ בעזרת train set לחיזוי ערך המשתנה vote וענו על השאלות הבאות:

זכרו להציב RSEED : DecisionTreeClassifier(random_state=RSEED)

Model Evaluation

6. בנו Confusion matrix עבור חיזוי על ה- test set בעזרת המודל שבניתם בשאלה הקודמת. הניחו כי ״דמוקרטי״ = Positive וחשבו את המדדים הבאים:
   א. Accuracy
   ב. Precision
   ג. Recall
7. השוו את המדדים עבור ה train set. האם מתקיימת תופעת ה-overfitting במודל החיזוי שבניתם? נמקו.
8. צרו מודל חדש משופר על סמך המסקנות מ 7. הגבילו את גובה העץ ל5 ואת כמות הרשומות לחיתוך ל40 וענו:
   א. מהו עומק העץ ?
   ב. כמה עלים יש בעץ ?
   ג. מהו פיצ'ר החלוקה הטוב ביותר בעץ ?

ד. האם יש פיצ'רים שלא נכללו במודל ? מהם?

ה. האם התצפית ה 68 (בדאטה סט המקורי) סווגה נכונה במודל ? נמקו.

9. בצעו שוב חיזוי על ה train set וה test sets ובנו מטריצות חדשות.

★★10. מאור הריץ מודל משופר ויצאו לו המדדים הבאים:

Test set result:

Accuracy: 0.7946428571428571

Precision: 0.7142857142857143

recall: 0.9433962264150944

Train set results:

Accuracy: 0.7961538461538461

Precision: 0.7261904761904762

recall: 0.9457364341085271

מה אפשר להסיק מהתוצאות הנל לגבי ביצועי המודל על הדאטה ?

Decision tree - Multiclass

שנו את עמודת המטרה להיות "status" ובנו עץ החלטה לזיהוי הסטטוס.

בנו confusion matrix והדפיסו את מדד ה **accuracy**לאחר בדיקה על הtest set.

כתבו את התוצאה גם בקוד כהערה וענו:

10. האם המודל יחזה טוב את הסטטוס המשפחתי, מדוע?

11. חשבו את מדד ה **precision** עבור הקטגוריה single.