המכללה האקדמית להנדסה ע"ש סמי שמעון

המחלקה למדעי המחשב

קורס: מבוא לכריית הנתונים – תשפ"ג

מטלה 3 – **classification**

תאריך הגשה: 11/06/2023

מטרת המטלה

התנסות הסטודנטים בעץ החלטה, random forest, xgboost .

תוכן המטלה

בקורס כריית נתונים נלמד לחזות תוצאות המסווגות למספר מוגבל של אפשרויות, ערכי הסיווג יכולים להיות כל דבר, בעיות סיווג מחפשות למעשה רק איזשהו "קו מפריד" בין קבוצה של תצפיות לקבוצה אחרת. אנחנו מצפים לקבל החלטה כלומר קטגוריה מסוימת אליה התצפית שייכת ולא ערך.

במשימה זו הנכם מתבקשים להשתמש בסט מידע (dataset) שמצורף במודל.

<u>שלב א': עץ החלטה (40 נקודות)</u>

יבואן רכב החליט לנהל תוכנית מכירות חדשה ולשלוח סוכני מכירות לגבי מספר סוגי רכבים אך ורק ללקוחות עם פוטנציאל גבוה לרכוש אותם. היבואן פנה לחברת מערכות מידע. בחברה החליטו לנתח את הנתונים של רכישות קודמות ולבנות עץ החלטה לכל סוג של רכב על מנת לסווג את הלקוחות לפי רמת פוטנציאל הקנייה (יקנה / לא יקנה).

הנתונים לגבי קנייה של דגם רכב ספציפי מכילים: רמת הכנסה של הלקוח, האם הלקוח נשוי, מספר השנים שהרכב הנוכחי בבעלות הלקוח והאם הרכב נקנה ע"י הלקוח.

להלן טבלה מרכזת בעניין:

ID	Income	Married	Old Car Age	Bought?
1	High	No	1-7	Yes
2	High	No	8+	No
3	High	No	1-7	Yes
4	High	Yes	8+	Yes
5	Low	Yes	1-7	No
6	Low	No	8+	No

7	Low	Yes	8+	Yes
8	High	No	1-7	No
9	High	Yes	1-7	Yes
10	Low	Yes	8+	No
11	Low	Yes	1-7	No
12	Low	No	8+	No
13	High	Yes	8+	Yes
14	Low	Yes	1-7	No
15	High	No	8+	Yes

השתמשו בנתונים אלו על מנת לבנות ולבדוק מודל ID3.

1. יש לחלק את הנתונים ל-training set (10 תצפיות הראשונות) ול-test set (5 תצפיות האחרונות).
2. בנו מודל ID3 על ה-training set (כאשר מספר מינימלי של תצפיות בעלה הוא 2).
3. בדקו את המודל על ה-test set – מהו הaccuracy שקיבלתם?
4. לפי המודל שקיבלתם בסעיף 2, למי מהלקוחות החדשים הבאים כדאי לשלוח סוכן?

ID	Income	Married	Old Car Age	Bought?
16	High	No	1-7	?
17	High	Yes	8+	?
18	Low	Yes	1-7	?

שלב ב': **RandomForest** ו-**XGBoost** (60 נקודות)

עבור שלב זה ושלב הבא השתמשו בדאטאסט המצורף המייצג נתונים לגבי רכבים ונסו לחזות את העמודה class

העמודות הן:
```
['buying', 'maint', 'doors', 'persons', 'lug_boot', 'safety', 'class']
```
המכילות ערכים:

buying: vhigh, high, med, low.
maint: vhigh, high, med, low.
doors: 2, 3, 4, 5 more.
persons: 2, 4, more.
lug_boot: small, med, big.
safety: low, med, high.

שימו לב vhigh מייצג very high , med מיצג medium.

1. הציגו את המידע בעזרת הפונקציות של pandas והן info, head, tail, shape והדפיסו עבור כל עמודה את שמה ואת הערכים שמופיעים בה וכמותם (השתמשו ב-value_counts) , בדקו האם יש ערכים חסרים.
2. צרו X ו-y מהנתונים (מכל הרשומות בדאטאסט) ככה שX יכיל את כל העמודות מלבד class ו-y יכיל את העמודה class. השתמשו בlabel encoder על מנת לקודד את ערכי העמודות.
3. פצלו את הנתונים לאימון ובדיקה בעזרת הפונקציה train_test_split כך שהבדיקה תכיל 33% מהנתונים בצורה רנדומלית עם הערך 42.
4. אמנו מודל עץ החלטה על נתוני האימון ובצעו חיזוי עבור נתוני הבדיקה.
5. הציגו את העץ שהתקבל.
6. הציגו את הaccuracy ואת Confusion matrix.
7. אמנו מודל Random Forest Classifier עם 100 estimators על נתוני האימון ובצעו חיזוי עבור נתוני הבדיקה.
8. הציגו את הaccuracy ואת Confusion matrix.
9. שמרו את המודל בעזרת החבילה pickle בקובץ בשם RandomForestClassifier.pkl.
10. אמנו מודל XGBClassifier על נתוני האימון ובצעו חיזוי עבור נתוני הבדיקה.
11. הציגו את הaccuracy ואת Confusion matrix.
12. שמרו את המודל בעזרת החבילה pickle בקובץ בשם XGBClassifier.pkl.

בהצלחה

הוראות הגשה

1. עליכם להגיש קובץ zip הכולל את מספרי תעודות הזהות של שני המגישים עם קו תחתון מפריד ביניהם בצורה הבאה: ID1_ID2. zip.
2. תשובה סופית, גם אם היא נכונה, ללא הצגת דרך החישוב המלאה, לא תזכה בניקוד.
3. את הקובץ pdf שיצרתם עליכם להעלות לתיבת ההגשה.
4. הגשה בזוגות בלבד.
5. תאריך ההגשה המעודכן ומיקום ההגשה יופיעו באתר הקורס.