תרגיל 7- Classification

הוראות הגשה:

1. בתרגיל הבא יש לענות על השאלות באמצעות שימוש בקוד **פייתון ושימוש ב-Scikit-Learn**.
2. יש להגיש את העבודה **בזוגות בלבד**.
3. שם הקובץ יהיה מספרי הזהות של המגישים בצורה הבאה: זהות1_זהות2 במחברת הפתרון, יש לציין את מספר השאלה עליה עניתם עבור כל חלק בפתרון

חלק KNN 1:

ענו על השאלות הבאות באמצעות הנתונים על מחירי יהלומים:

1. חלקו את המידע ל train וtest. באמצעות אלגוריתם KNN (K=3) בנו מודל שחוזה clarity של היהלום באמצעות נתוני carat, depth, price, table, y, x. לאחר מכן ,חשבו את מדדי הaccuracy וf1_score עבור המודל שיצרתם .

2. באמצעות seaborn ציירו גרף של ביצועי המודל. כלומר, יש לצייר גרף שבהם ציר ה-X הוא ערך ה-K של המודלים ואילו ערכי ה-Y הם מדדי הaccuracy ו-f1_score שמתאימים לכל מודל. (מי שלא מצליח לשלב באותו גרף ניתן בשני גרפים שונים- כל עוד הסקאלות תואמות)

3. חזרו על בניית המודל בדומה לסעיף ה1, רק הפעם בנו את המודל על ידי הוספת מידע מעמודות clarity וcolor. כלומר, בנו את המודלים בתוספת שתי העמודות הנוספות. האם הaccuracy של המודלים השתפרו?
רמז: יש להשתמש בLabelEncoder בסעיף זה

4. בדומה לסעיף 4, בנו מודלי KNN (K=5) החוזים את הcut של היהלום, רק שהפעם על מנת לבנות את המודלים, השתמשו בגדלים שונים של נתונים:
5%,10%,50%,75%,90%,
כלומר, יש לבנות את המודל רק על ידי שימוש ב5% מהמידע ,10% מהמידע , וכו׳ צרו גרף שבו ציר ה-X הוא גודל הtrainset באחוזים ואילו ציר ה-Y הוא הaccuracy של המסווג.

הערה חשובה: חשוב להשתמש כ-test באותם נתונים בדיוק עבור כל המודלים

חלק 2 decision tree:

ענו על השאלות הבאות באמצעות הנתונים על השחקנים בFIFA23:
אנחנו נרצה לחזות מהו התפקיד של השחקן לפי הנתונים שלו

בחלק זה נבצע את התהליך של דאטה סיינס.

1. Preprocessing - בשלב זה אנו מסדרים ומנקים את הנתונים עבור אימון המודל. הפעילו שיקולים כיצד לנקות ולסדר את הנתונים.
להלן כמה קווים מנחים:
   - בדקו חוסרים בדאטה – חשבו על דרכים כיצד להתמודד עם החוסרים
   - התמודדו עם עמודות קטגוריאליות (לא ניתן להסיר את העמודות הללו)
   - עמודות מזהות רשומות (לדוגמה מספר ID)
   - פישוט עבור המודל – החליפו תפקידים של חלק מהשחקנים לצורה גנארית.
לדוגמא: 'LWB' : 'LW', 'RWB': 'RW', 'CF': 'ST', 'CAM': 'CM', 'CDM': 'CM'
   - עבור העמודה של BestPosition- בצעו מיפוי לדאטה, כך שיהיה ברור מה סדר השחקנים במגרש. לדוגמא : שוער יקבל את הספרה 0, מגן שמאלי את הספרה 1

2. בצעו חלוקה של הדאטה לtrain-test לפי 0.2, קבעו את החלוקה (seed) באמצעות התעודות זהות שלכם.

3. בנו 3 עצי החלטה בעומקים שונים (3-7) ועבור כל עץ הציגו את 4 המדדים עליהם למדנו מצאו מהו העץ שנותן לנו את הדיוק הטוב ביותר (הclass הוא BestPosition)

4. ציירו את העץ בעל הדיוק הגבוה.

5. השתמשו בfeature_importance:
א. בצעו את סעיף 3 מחדש רק בשימוש בtop-10 פיצ'רים בעלי התרומה הרבה למודל.
הסבירו למה לדעתכם אותם פיצ'רים תרמו הכי הרבה למודל
ב. קבעו trash-hold לבחירתכם עבור ה-score, ובצעו את סעיף 3 רק בשימוש באותם פיצ'רים שעברו את הtrash-hold. הסבירו לפי מה בחרתם את ה trash-hold.

6. השתמשו ב(your_id) random.seed והגרילו 5 שחקנים מהטווח 1-1000 המייצגים את האינדקס שלהם בבסיס הנתונים הכולל , בדקו עליהם את הפרדיקציה על סמך המודל הכי טוב מסעיף 5, והסבירו את התוצאות שקיבלתם לפי דעתכם.
(רמז: נסו להבין מהו path שהוא עבר)

7. **בונוס** - חקרו על GridSearch
ובנו את המודל הטוב ביותר שניתן באמצעות הפרמטרים הבאים:

```
'max_depth':[2,3,4,5,6,7,8,9,10,11,12,13],
'criterion':['gini','entropy'],
'max_features': ['auto', 'sqrt', 'log2'],
```

```
'class_weight' : ['balanced'],
'splitter':[ 'best', 'random']
```

החזירו את הפרמטרים הכי טובים שקיבלתם