אוניברסיטת בן-גוריון בנגב
Ben-Gurion University of the Negev

Classification – 8 תרגיל

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. התרגיל יוגש כמחברת colab, כאשר לתיבת הגשה יש להגיש קובץ המכיל לינק למחברת עם

   הרשאות קריאה למייל **yaelhoc@post.bgu.ac.il** + **tamarin@post.bgu.ac.il**
4. שם הקובץ יהיה מספרי הזהות של המגישים בצורה הבאה: זהות1_זהות2

   במחברת הפתרון, יש לציין את מספר השאלה עליה עניתם עבור כל חלק בפתרון

ענו על השאלות הבאות באמצעות הנתונים על מחירי יהלומים:

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

2. חזרו על בניית המודל בסעיף 1 עבור ערכי k שונים. באמצעות plotly express ציירו גרפים של ביצועי המודלים עבור ערכי k שונים. כלומר, יש לצייר גרפים שבהם ציר ה-X הוא ערך ה-K של המודלים ואילו ערכי ה-Y הם מדדי הaccuracy ו-f1_score שמתאימים לכל מודל.

3. חלקו את המידע ל train וtest. לאחר מכן, באמצעות אלגוריתמי KNN (K=1,3,5) ו Decision Tree בנו מודלים שחוזים הcut של היהלום באמצעות עמודות carat, depth, price, table, x, and y. לפי מדד הaccuracy, מבין המודלים שיצרתם, איזה מודל חוזה את הcut טוב יותר?

4. חזרו על בניית המודלים בדומה לסעיף הקודם, 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

5. בדומה לסעיף 4, בנו מודלי KNN (K=3) החוזים את הcut של היהלום, רק שהפעם על מנת לבנות את המודלים, השתמשו בגדלים שונים של נתונים:
   5%,10%,20%, 50%,70%,80%,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