תרגיל כתה: clustering

בתרגיל זה תתנסו ב-clustering בפייתון.

1. הטעינו את הקובץ points.csv למשתנה בסביבת העבודה שלכם. קובץ זה מכיל קואורדינטות של נקודות במרחב (עמודות x, y, z) ואת הסיווג שלהם (העמודה target). כמה שורות יש בקובץ? אילו עמודות מכילה הטבלה?
2. ציירו scatter plots של כל הזוגות האפשריים של מימדים (X ו-Y, X ו-Z, ו-Y ו-Z) וצבעו את הנקודות בהתאם לסיווג שלהן. באילו מימדים ההפרדה נראית הכי טובה?
3. חלקו את הנקודות לקבוצת training (70% מהנקודות) ולקבוצת testing (30% מהנקודות).
4. הריצו hierarchical clustering על ה-dataset כולו והציגו את התוצאה. בחרו את הגובה המתאים ל-5 clusters אותו מצאתם בסעיף 2 והציגו את התוצאה על התרשים. חפשו כיצד להציג את העץ שנוצר ואת גובה החיתוך (מופיע גם בשקפים).
5. הציגו scatter plot עם שני המימדים שמצאתם ב-(2) וצבעו את הנקודות בהתאם לסיווג שקיבלתם. האם הייתם מגדירים את התוצאה כטובה?
6. קראו את השקפים על k-means מהמצגת וחפשו מידע נוסף על האלגוריתם, למשל כאן: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

   מצאו כיצד להריץ k-means בפייתון.
7. הריצו k-means clustering על המידע והציגו את התוצאה על scatter plot עם שני המימדים האינפורמטיבים ביותר. האם התוצאה נראית נכונה?
8. הריצו פעם נוספת hierarchical clustering ו-k-means, הפעם תוך שימוש בשני המימדים האינפורמטיבים ביותר בלבד. הציגו scatter plots עם שני המימדים בהם השתמשתם וצבעו את הנקודות בהתאם לתוצאת ה-k-means. האם התוצאה נראית טוב יותר?
9. חשבו את ה-Rand Index עבור התוצאה של k-means עם כל המשתנים (לעומת הסיווג האמיתי) ו-k-means עם שני המשתנים אינפורמטיבים ביותר. השתמשו ב-adjusted_rand_index מ-sklearn. הדפיסו את התוצאה.

בונוס:

10. ננסה לבנות מסווג knn . נסו, על קבוצת האימון, ערכי k שונים בין 1 ל-30. צרו מסווג על קבוצת האימון ובידקו את הביצועים של המסווג על קבוצת ה-testing. צרו גרף שבו תראו את הדיוק (accuracy) עבור כל אחד מהערכים כמו שראינו בכתה, ומצאו את ערך ה-k שעבורו הדיוק הוא הגבוה ביותר.
11. עבור הערך שבחרתם, צרו confusion matrix ובדקו איזה מהקבוצות "התערבבו". הדפיסו את המטריצה ו-classification report.