SHENKAR — שנקר
ENGINEERING. DESIGN. ART. הנדסה. עיצוב. אמנות.

תכנות מתקדם - 3101803

מרצים: מר פרץ אור, מר גוטמן דוד

- משך הבחינה: שלוש שעות (180 דקות).
- חומר פתוח.
- מועד א' .
- הבחינה מכילה 2 חלקים אשר יש לענות על כל השאלות. יש לצרף את הפתרון לתיבת ההגשה בפורמט py או ipynb בלבד.
- **לאחר 3 שעות, תיבת ההגשה תיסגר באופן אוטומטי. יש לנהל את זמן הבחינה היטב ולהגיש בזמן הקצוב. בחינות אשר יאחרו את המועד לא ייבדקו.**

בהצלחה!

(35 נק׳) חלק א׳

1. (5 נק׳) - כתבו את הפונקציה get_matrix אשר מקבלת מספר שלם וחיובי n ומחזירה מטריצה בגודל n x n כאשר איבריה הם מספרים אקראיים בין 1 ל10.
2. (3 נק׳) - הפעילו את הפונקציה מסעיף (1) וצרו 2 מטריצות: אחת בגודל 5x5 ואחת בגודל 7x7.
3. (15 נק׳) - כתבו את הפונקציה sort_matrix_diagonal אשר מקבלת מטריצה וממיינת את האלכסון הראשי של המטריצה בלבד.

למשל,

4,3,1	→	1,3,1
1,3,4		1,3,4
1,2,1		1,2,4

<u>הערה</u>: אין להעתיק לרשימה צדדית לצורך מיון.

ללא תלות בסעיפים הקודמים:

(12 נק׳) - *<u>הגדרה</u>: בהינתן רשימה A, הזוג (i,j) נקרא "היפוך" אם* $A[i] > A[j]$.

למשל, עבור הרשימה [1,9,6,4,5] מספר ההיפוכים הוא 5 וערכם הוא:

(9, 6), (9, 4), (9, 5), (6, 4), (6, 5)

כתבו את הפונקציה count inversion אשר מקבלת רשימה של מספרים שלמים ומחזירה את רשימת ההיפוכים.

(65 נק׳) חלק ב׳

לבחינה זו, מצורף קובץ בשם cancer.csv אשר מכיל מידע על צילומי ריאות וחזה של מטופלים החשודים לסרטן. עמודת diagnosis מכילה 2 ערכים: M - גידול ממאיר, B - גידול שפיר.

(30 נק׳) - תחקור ראשוני, שאילתות וגרפים

1. (2 נק׳) - קראו את קובץ הcsv לDataFrame.
2. (2 נק׳) - מהי כמות המטופלים אשר מופיעים בקובץ?
3. (4 נק׳) - מהו ממוצע הradius_mean עבור כל סוג diagnosis?
4. (4 נק׳) - עבור diagnosis=M, מהו הממוצע הגדול יותר - area_mean או area_se?
5. (4 נק׳) - כמה מטופלים נמצאים מתחת לממוצע של perimeter_mean?
6. (4 נק׳) - עבור כל קבוצת diagnosis, הציגו את כמות הנבדקים אשר ממוצע compactness_mean שלהם הוא מתחת לממוצע הכללי של compactness_mean.
7. הציגו 3 גרפים (מסוגים שונים) לבחירתכם אשר יענו על השאלות הנ״ל.

(20 נק׳) - עיבוד נתונים

1. (3 נק׳) - **במידה וישנם** ערכים חסרים בטבלה:
   a. ערך נומרי (מספרי) - השתמשו בפעולת ה-Imputation והשלימו את הנתונים באמצעות הממוצע עבור אותה עמודה.
   b. ערך קטגוריאלי - מחקו את הרשומות בהן הערך חסר.
2. (5 נק׳) - עבור כל עמודה נומרית, נסמן m - ממוצע העמודה, s - סטיית תקן של העמודה. בצעו נרמול לנתונים, כך שעבור כל ערך a בטבלה, הערך החדש יחושב:

$$\frac{a-m}{s}$$

3. (3 נק׳) - צרו שני העתקים של אוסף הנתונים. אחד ישמש ללמידה לא-מונחית והשני ללמידה מונחית.

4. (6 נק׳)
   a. עבור הלמידה המונחית - פצלו את אוסף הנתונים לdata, labels כאשר data הינו אוסף הנתונים ללא עמודת diagnosis, והמשתנה labels הינו עמודת diagnosis.
   b. עבור הלמידה הלא-מונחית - מחקו את עמודת diagnosis.

(15 נק׳) - מודל למידה

1. (5 נק׳) - הריצו KMeans עם הערכים $K = 2\ to\ 15$, כאשר בכל ריצה שמרו ברשימה את SSE (סכום הטעויות בריבוע). השתמשו ב-Elbow Method ובחרו K משוערך. עבור הK שבחרתם, הציגו את מדד הסילואט ו describe עבור האשכולות שהתקבלו.
2. (5 נק׳) - צרו מופע של Random Forest המשתמש ב100 עצי החלטה שונים, ובצעו אימון על הקבוצה המתאימה. הציגו את מדד הדיוק ומטריצת הבלבול (Confusion Matrix).
3. (5 נק׳) - רשמו מסקנה **<u>אחת</u>** עבור כל מודל.