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.
- **עבור כל שאלה, יש לרשום קוד יעיל ככל הניתן (לא יינתן ניקוד מלא לקוד אשר מבצע את הדרישה אך בצורה לא יעילה).**

בהצלחה!

(35 נק׳) חלק א׳

השירות המטאורולוגי מאחסן מדי יום את הטמפרטורה הממוצעת אשר נמדדה בתל אביב במהלך חודש מאי. לאחר איסוף המידע, הבחינו החזאים כי הצטבר מידע בפורמט הבא:

[24,25,…,27,28,21,22,...,23]

כלומר, מתחילת החודש ועד לנקודה מסוימת הטמפרטורה עולה מדי יום (בדוגמה הנ״ל, זהו היום בו נמדדו 28 מעלות). לאחר מכן, הטמפרטורה יורדת והחל מנקודה זו עולה מדי יום. במילים אחרות,

<u>מתקבלת רשימה אשר מורכבת מ 2 רשימות ממוינות בזו אחר זו.</u>

1. (16 נק׳) - כתבו את הפונקציה search אשר מקבלת את הרשימה הנ״ל וטמפרטורה ומחזירה את האינדקס בו נמצאת הטמפרטורה (במידה והטמפרטורה מופיעה מספר פעמים, יש להחזיר את המופע הראשון). במידה והטמפרטורה לא נמצאה, יש להחזיר 1-.
2. (3 נק׳) - הפעילו את הפונקציה על רשימה בגודל 10 אשר מקיימת את הדרישות הנ״ל. (ניתן לבחור מספרים לבחירתכם)
3. (12 נק׳) - נסמן את הפרמטר m להיות ממוצע הרשימה. כתבו את הפונקציה get_diff אשר מחזירה את החישוב הבא:

$$\sum_{i=1}^{n}(x_i - m)^2$$

כאשר x_i הוא האיבר ה-i ברשימה.

למשל, עבור הרשימה [1,2,3] הממוצע הינו 2 ולכן הערך שיחושב:

$$(1-2)^2 + (2-2)^2 + (3-2)^2 = 1 + 0 + 1 = 2$$

4. (4 נק׳) – הפעילו את הפונקציה get_diff על הרשימה אשר הגדרתם בסעיף (2).

(65 נק׳) חלק ב׳

לבחינה זו, מצורף קובץ בשם "salary.xlsx" המכיל מידע על משכורות עובדים. מבנה הקובץ:

age - גיל

marital-status - סטטוס נישואים / משפחה

relationship - בעל / אישה / ילד / ...

gender - מגדר

hours-per-week - שעות עבודה שבועיות

country - מדינה

salary - האם המשכורת מעל 50 אלף או פחות

(30 נק׳) - תחקור ראשוני, שאילתות וגרפים

בסעיפים 1-7 יש להדפיס תשובה מלאה. למשל, "**The number of users**..."

1. (2 נק׳) - קראו את קובץ האקסל (xlsx) לDataFrame.
2. (2 נק׳) - מהו הגיל הממוצע של נשים אשר מופיעות בקובץ?
3. (4 נק׳) - באיזו מדינה יש הכי הרבה גברים מעל גיל 30?
4. (4 נק׳) - מהו סוג הrelationship עבורו קיימים יותר נשים מגברים?
5. (4 נק׳) - עבור כל סטטוס נישואים (marital-status), הציגו את כמות האנשים אשר מרוויחים מעל ל50 אלף ש״ח.
6. (4 נק׳) - עבור כל מדינה (country), הציגו את ממוצע שעות העבודה השבועיות של גברים לעומת נשים.
7. (4 נק׳) - עבור עובדים אשר מרוויחים פחות מ50 אלף ש״ח, מהו הrelationship הנפוץ ביותר?
8. הציגו 2 גרפים (<u>מסוגים שונים</u>) לבחירתכם:
   a. (3 נק׳) - גרף 1 - יענה על התשובה לשאלה מספר 4, יציג את היחס בין הגברים לנשים.
   b. (3 נק׳) - גרף 2 - יענה על התשובה לשאלה מספר 6.

(20 נק׳) - עיבוד נתונים

1. (3 נק׳) - **במידה וישנם** ערכים חסרים בטבלה, מחקו את השורות בהם ערכים אלו מופיעים.
2. (6 נק׳) - עבור **כל** עמודה נומרית,
   a. מצאו את ממוצע העמודה - m.
   b. מצאו את סטיית התקן של העמודה - s.
   c. בצעו נרמול לנתונים, כך שעבור כל ערך a בטבלה, הערך החדש יחושב:
$$\frac{a - m}{s}$$
3. (2 נק׳) - מחקו את עמודות salary.
4. (5 נק׳) - המירו את עמודת gender כך שגברים יסומנו במספר 0 ואילו נשים במספר 1. לאחר מכן, המירו את שאר העמודות הקטגוריאליות לקידוד one-hot-vector.
5. (2 נק׳) - צרו העתק של אוסף הנתונים ומחקו ממנו את עמודות age, hours-per-week.
   הערה: לאחר סעיף זה, יש ברשותכם את אוסף הנתונים המנורמל ועוד אוסף נתונים מקודד one-hot-vector ללא עמודות age, hours-per-week.

(15 נק׳) - מודל למידה

1. (6 נק׳) - הריצו KMeans עם הערכים $K = 2\ to\ 7$, כאשר בכל ריצה יש לאמן 2 מודלים:
   a. הראשון עבור אוסף הנתונים.
   b. השני עבור ההעתק שיצרתם בחלק הקודם סעיף (5).
   יש לשמור ברשימה את SSE (סכום הטעויות בריבוע) עבור כל k וכל מודל בנפרד. השתמשו ב-Elbow Method ובחרו K משוערך כל אחת מן הבעיות.
2. (5 נק׳) - הציגו את מדד הסילואט עבור ה-k שבחרתם לכל אחת מהבעיות. הסבירו במשפט <u>אחד</u> את משמעותו.
3. (4 נק׳) - מה ניתן להגיד על האופן בו התבצעה החלוקה עבור hours-per-week? האם הקבוצות שהתקבלו מייצגות חלוקה טובה עבור hours-per-week? הציגו תשובה זו בקוד + הסבר מילולי.