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׳

לפני כשבוע, התקבלה פנייה במשטרת ישראל על פריצה שהתקיימה במוזיאון תל-אביב בין השעות 20:00 בערב ל8:00 בבוקר. משטרת ישראל הגיעה למוזיאון וראתה כי מחצית מהתמונות שבורות.
הצעד הראשון של המשטרה הוא לבקש את צילומי האבטחה (סרטון אשר אורכו 12 שעות).
לצורך תרגיל זה, צילומי האבטחה מאוחסנים ברשימה אשר כל אינדקס מציין True אם הפריצה התרחשה <u>לפני</u> הדקה הנוכחית, אחרת False. אורך הרשימה הינו 12x60=720 דקות, כאשר האינדקס 0 מייצג את השעה 20:00, אינדקס 1 את השעה 20:01, וכן הלאה.
למשל, עבור הרשימה

[False, False, True, True, …, True]

הפריצה התרחשה באינדקס 2, כלומר בשעה 20:02.

1. (3 נק׳) - צרו רשימה אקראית של ערכי True/False באורך 720.
2. (5 נק׳) - מיינו את הרשימה בעזרת מיון בועות / מיון בחירה.
3. (15 נק׳) - כתבו אלגוריתם <u>יעיל</u> ככל הניתן למציאת הזמן בו התרחשה הפריצה.
   (ניתן להדפיס את האינדקס, ולא שעה ספציפית)

ללא תלות בסעיפים הקודמים:

(12 נק׳) - *<u>הגדרה</u>: ״איבר-האמצע״ של רשימת מספרים הוא מספר אשר 50% משאר המספרים קטנים ממנו (ובהתאם, 50% מהמספרים גדולים ממנו).*

למשל, עבור הרשימה [5,1,4,2,6] הפונקציה תחזיר 4 מכיוון ש1,2 קטנים ממנו ו5,6 גדולים ממנו.
במקרה בו אורך הרשימה הינו זוגי (וקיימים שני איברים כאלו), ״איבר-האמצע״ הינו הראשון מבין השניים.

כתבו את הפונקציה find_middle אשר מקבלת רשימה של מספרים שלמים ומחזירה את ״איבר-האמצע״ של הרשימה.

(65 נק׳) חלק ב׳

לבחינה זו, מצורף קובץ בשם ״StudentsPerfomace.csv״ אשר מכיל מידע על ביצועי סטודנטים במקצועות מתמטיקה, קריאה וכתיבה. בנוסף, הקובץ מכיל מידע נוסף על כל סטודנט:

gender: מגדר

ethnicity: קבוצות A, B, C, D, E

education: רמת ההשכלה של הסטודנט

lunch: standard, free/reduced

pre-course: האם השלים מכינה

math, reading, writing: 0 - 100

(30 נק׳) - תחקור ראשוני, שאילתות וגרפים

1. (2 נק׳) - קראו את קובץ הcsv לDataFrame.
2. (2 נק׳) - מהי כמות הסטודנטים אשר מופיעים בקובץ?
3. (4 נק׳) - מהי כמות הגברים מול כמות הנשים?
4. (4 נק׳) - מהי כמות הנשים אשר השלימו מכינה?
5. (4 נק׳) - מהי כמות הסטודנטים אשר נכשלו לפחות במקצוע אחד וביצעו מכינה לפני הלימודים?
6. (4 נק׳) - עבור כל קבוצת Ethnicity, הציגו את כמות הגברים אשר נכשלו במתמטיקה ואוכלים ארוחת צהריים סטנדרטית.
7. (4 נק׳) - מהי רמת ההשכלה (education) אשר היחס בין מספר הנשים למספר הגברים הוא הנמוך ביותר?
8. הציגו 2 גרפים (מסוגים שונים) לבחירתכם:
   a. (3 נק׳) - גרף 1 - יענה על התשובה לשאלה מספר 3.
   b. (3 נק׳) - גרף 2 - יענה על התשובה לשאלה מספר 6.

(20 נק׳) - עיבוד נתונים

1. (3 נק׳) - **<u>במידה וישנם</u>** 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. (3 נק׳) - השתמשו בקידוד one-hot-vector והמירו את ה**משתנים קטגוריאליים**.
3. (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:

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

4. (3 נק׳) - צרו שני העתקים של אוסף הנתונים. אחד ישמש ללמידה לא-מונחית והשני ללמידה מונחית.
5. (6 נק׳) - עבור הלמידה המונחית
   a. צרו עמודת Target אשר שווה 1 אם הסטודנט עבר 2 קורסים לפחות (בציון של 60), אחרת 0.
   b. פצלו את אוסף הנתונים לdata, labels כאשר data הינו אוסף הנתונים ללא עמודת Target, והמשתנה labels הינו עמודת Target.

(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.