תכנות מתקדם - מטלה 5

נושא: Supervised learning and Unsupervised learning

הנחיות:

1. נא לכתוב קוד מסודר, תוך שימוש בשמות משתנים בעלי משמעות והערות היכן שנדרש.
2. כל שאלה צריכה להיות מוגשת בנפרד
3. בתחילת כל מענה, נא להוסיף כהערה את שמות המגישים והאם התייעצתם נעזרתם בסטודנטים נוספים.

הגשה:

1. יש להגיש את העבודות בזוגות (עפ״י הקבוצות שנרשמתם).
2. יש להגיש מחברת בודדת (Jupyter notebook) עם המטלה.
3. שם הקובץ המוגש צריך להיות: student1_id_student2_id.ipynb

בהצלחה!

חלק א' - תחקור ראשוני

לתרגיל זה מצורף קובץ csv המכיל נתונים על תסמינים בקרב חולי סוכרת. הנתונים מכילים את העמודות הבאות:

Pregnancies - מספר ההריונות בעבר. במידה והמספר הוא 0, מדובר בגבר.

Glucose - רמת גלוקוז בדם

BloodPressure - לחץ דם

SkinThickness - עובי העור בס"מ

Insulin - רמת אינסולין בדם

BMI

DiabetesLevel - רמת סוכרת

AgeCategory - קטגוריית גיל אליה משוייך המטופל

Outcome - האם זוהתה סוכרת אצל המטופל, 1 זוהתה ו-0 לא זוהתה

1. הורד את קובץ הנתונים הנ"ל, וקרא את הנתונים לתוך DataFrame.
2. הצג את התוצאה של פונקציית describe והסבר בקצרה (2-3 משפטים) מה ניתן להסיק מטבלה זו. למשל, מה הם הערכים הנמוכים והגבוהים של כל עמודה, ממוצע וכו'.
3. מהו לחץ הדם הגבוה ביותר שנמדד בקרב גברים?
4. אצל איזו קבוצה ממוצע עובי העור גדול יותר, אצל גברים או נשים?
5. הצג התפלגות של **לחץ הדם**, כאשר גברים ונשים מופיעים בצבעים שונים. מה ניתן להגיד על כל אחת מההתפלגויות? (hist, bar)
6. מהי קטגוריית הגיל עבורה ממוצע רמת הגלוקוז בדם (של המטופלים השייכים לקבוצה זו) הוא הגבוה ביותר?
7. מהו ממוצע האינסולין בדם עבור נשים אשר עברו מעל ל-8 הריונות?
8. הצג גרף הממחיש את הקשר בין לחץ הדם לרמת הגלוקוז בדם.

חלק ב' - עיבוד נתונים + Clustering

1. **<u>במידה וישנם</u>** ערכים חסרים בטבלה, השתמש בפעולת ה-Imputation והשלם את הנתונים באמצעות הממוצע עבור אותה עמודה.
2. השתמש בקידוד one-hot-vector והמר את ה**משתנים קטגוריאליים**.
3. בצע נרמול לנתונים בטבלה (ניתן ע"י MinMaxScaler שראינו בתרגול, **<u>או</u>** ע"י הורדת הממוצע וחילוק בסטיית התקן). שמור העתק של אוסף הנתונים המקוריים (לאחר סעיף 2, ולפני סעיף 3).
4. הריצו KMeans עם הערכים $K = 2\ to\ 15$, כאשר בכל ריצה - יש לאמן 2 מודלים שונים: אחד עבור הנתונים המקוריים ואחד עבור הנתונים המנורמלים. עבור כל אחד, שמרו ברשימה את SSE (סכום הטעויות בריבוע).
5. השתמשו ב-Elbow Method ובחרו K משוערך עבור כל אחת מהבעיות. בנוסף, הצג את מדד הסילואט עבור כל מודל.
6. עבור אוסף הנתונים המנורמל, הציגו סקירה קצרה עבור כל אחד מהקלאסטרים שנוצרו. למשל, אם נוצרו 2 קלאסטרים, ניתן להציג עבור כל אחד נתונים סטטיסטים (באמצעות הפונקציה describe). **<u>מומלץ</u>** - השתמשו בויזואליזציה להמחיש את ההבדלים.

חלק ג' - Classification

עבור חלק זה, עמודת "outcome" היא הסיווג באוסף הנתונים שלנו. כלומר, כל שורה היא אוסף נתונים על נבדק בודד, כאשר עמודת outcome היא 1 אם לאותו נבדק יש סוכרת, אחרת 0.

1. הצג את היחס בין כמות הנבדקים אשר חולי סוכרת לבין השאר.
2. כתבו את הפונקציה split_train_test **<u>המקבלת מספר בין 0 ל-1</u>** המייצג את גודל קבוצת המבחן ומחזירה חלוקה של הנתונים בהתאם.

   הערה: פונקציה זו שימושית כאשר נרצה לבדוק מספר אפשרויות לגודל קבוצת המבחן.
3. הפעילו את הפונקציה הנ"ל וחלקו את אוסף הנתונים לסה"כ 75% אימון ו-25% מבחן.

4. צרו מופע של Decision Tree ובצעו אימון על הקבוצה המתאימה.
5. בדקו ניבוי (prediction) על קבוצת המבחן, והציגו את מדד ה-accuracy ואת מטריצת הבלבול. הוסיפו תיאור קצר מה ניתן להבין ממטריצת הבלבול.
6. צרו מופע של Random Forest המשתמש ב-50 עצי החלטה שונים, ובצעו אימון על הקבוצה המתאימה.
7. חזרו על סעיף מספר 5, עבור התוצאה של 6.
8. הדפיסו את חשיבות התכונות בנתונים. מי היא החשובה ביותר? מה הכי פחות חשובה?