תרגיל 4 - Python

1. (5 נקודות) טען את קובץ נתוני האימון train_Loan.csv למשתנה מסוג Dataframe
2. (5 נקודות) הדפס את התפלגות הערכים (Frequency Distribution) של המשתנים השונים בקובץ האימון.
3. (5 נקודות) הדפס את הטיפוס (Type) של כל אחד מהשתנים בקובץ האימון.
4. (5 נקודות) בצע תיקון Missing Values עבור כל המשתנים שיש להם ערכים חסרים.
5. (5 נקודות) בצע דיסקרטיזציה (מכל סוג) למשתנה LoanAmount. ציין את ה-bins השונים.
6. (5 נקודות) בצע Outlier Detection למשתנה LoanAmount, על פי 3 סטיות תקן בהתפלגות המשתנה. עליך להשאיר בקובץ האימון רשומות ללא ערכים חריגים.
7. (5 נקודות) הוסף עמודה חדשה בשם Normalized_Income לקובץ אימון אשר מיוצגת ע"י הנוסחה: 0.5*sqrt('Applicant_Income').
8. (5 נקודות) צור Dummy Variable עבור המשתנה Education. הוסף את תוצאות משתנה ה- Dummy Variable ל- Dataset המכיל את נתוני הטבלה המקורית.
9. (5 נקודות) הפוך משתנים קטגוריים למשתנים נומרים על מנת לאפשר שימוש בספרייה SK-Learn.
10. (5 נקודות) ייצא את טבלת נתוני האימון המעודכנת לקובץ בשם train_Loan_updated.csv
כמו כן, חשב כמה רשומות יש בקובץ האימוץ המעודכן.
11. (15 נקודות) הפעל עץ החלטה מסוג Decision Tree Classifier על קובץ האימון על מנת לחזות את Loan_Status על סמך כל הפיצ'רים האחרים, ואמן את המודל תוך שימוש ב-10-fold Cross Validation.
12. (10 נקודות) חשב את דיוק האימון של המודל (Accuracy) ואת ה- Cross Validation Score
13. (10 נקודות) צייר את עץ ההחלטה באמצעות הספרייה GraphViz
14. (5 נקודות) האם עץ החלטה הוא דוגמה ללמידה מונחית או למידה ללא הנחיה? הסבר את תשובתך
15. (10 נקודות) בהתאם לתשובתך בסעיף הקודם, בחר אלגוריתם נוסף, מאותו תחום למידה (מונחית או ללא הנחיה), ובנה מודל על פיו. הסבר מדוע בחרת במודל זה (מדוע הוא מתאים לנתוני הקלט).

הוראות הגשה

א. אי עמידה בכל אחת מההוראות יגרור הורדת ציון או פסילת העבודה.
ב. הגשת העבודה ב**זוגות** בלבד. רק אחד מבני הזוג יגיש את המטלה!
ג. שפת תכנות – Python גרסה 3.7 ומעלה, סביבת פיתוח – PyCharm גרסה 2019.1 ומעלה.
ד. יש להגיש את העבודה **לתיקיית ההגשה הרלוונטית באתר הקורס (Moodle).**
באחריותכם האישית לבדוק לפני הגשה כי כל הקבצים נפתחים כראוי.
ה. **יש להגיש קובץ ZIP** - שם הקובץ יהיה מורכב משני מספרי תעודות הזהות של המגישים באופן הבא: ID1_ID2.zip
הקובץ יכיל את הקבצים הבאים:
- קובץ קוד אחד ב- Python הממומש את כל הסעיפים שלעיל ברצף. יש לרשום הערות בתוך הקוד כך שניתן יהיה להבין באילו שורות מתחיל ומסתיים המימוש של כל אחד מהסעיפים שלעיל.
- קובץ word המכיל צילום מסך של תוצאות ההרצה ב- Python של כל אחד מהסעיפים.
- קובץ בשם train_Loan_updated.csv המכיל את טבלת נתוני האימון המעודכנת.

ו. אין לשתף קטעי קוד ואין להעתיק פתרונות!
ז. בנוסף, זוהי עבודה תכנותית ולפיכך יהיה משקל לכך בבדיקה. כלומר: יש לדאוג לקוד מסודר, הערות בקוד, לשמות משתנים בעלי משמעות וכדומה. יש לחלק את הקוד לפונקציות (במידת האפשר ולפי הצורך).
ח. שאלות בנוגע לתרגיל יש לשאול **אך ורק** בפורום השאלות הרלוונטי המופיע ב-moodle (ולא במייל - שאלות במייל לא יענו).

בהצלחה !!