למידת מכונה – מטלה 2 – הפעלת flow של למידה מונחית – מסמך הסבר

פרטים טכניים הנוגעים למטלה

תאריך הגשת המטלה

את המטלה יש להגיש עד יום ראשון בערב ה-16 ליוני. הגשה באיחור עד ה-23 ליוני (קנס חצי נקודה ליום על הגשה באיחור).

החומרים בהם יהיה מותר להשתמש

מותר להשתמש בכל חומר אותו למדנו הכולל

- python בסיסי
- המודולים (ספריות/חבילות תוכנה): NumPy, Pandas, Scikit-learn (sklearn)

החומרים בהם אסור להשתמש

- אסור להשתמש בשום מודול (ספריות/חבילות תוכנה) נוסף מלבד אלו המוזכרים לעיל
- אסור להשתמש בשום קובץ חיצוני.

הקבצים המצורפים למטלה:

קבצי data

- עבור כל dataset מופיעים קבצי csv עבור trainset ועבור test-set

מחברת הגשה ריקה להגשת התרגיל

- **שם הקובץ**: Assignment2_supervised_learning_flow.ipynb - המחברת שתריצו בה את הקוד, ההסברים, הניסויים והתוצאות. **המחברת אינה מכילה כל קוד** (זה יהיה תפקידכם :-))

אופן ההרשמה

- ניתן להגיש את העבודה בקבוצות של בין 4 – 6 סטודנטים.
- יש להירשם באקסל המשותף את שמות המשתתפים ומס' ת.ז. של כל משתתף (לפי מה שמופיע במודל).
- שימו לב, שכחלק מהבחירה, יש להירשם בשורה המתאימה ל dataset -אותו אתם בוחרים ולבעיית הלמידה אותה אתם בוחרים הכוללים:
  - עבור למידת רגרסיה: Diabetes, House-pricing
  - עבור למידת סיווג: Titanic, Wine, Breast cancer Wisconsin (diagnostic)

אופן ההגשה

כל משתתף ירשום בהגשה את 2 הקישורים הבאים (עם הפרדה של רווח בינהם). שימו לב, בקשת ההגשה מכל סטודנט, היא לצורך גיבוי. המטלה תיבדק רק פעם אחת:

1. הגשת חובה – **קישור לסרטון** (תצטרכו להעלות את הסרטון ל- YouTube, או למקום אחר ברשת)**. על הסרטון להיות קצר באורך של כ 2-3 דקות (לא יותר)**, בו אתם מציגים ומסבירים את עבודתכם ואת התוצאות.
2. הגשת חובה – **קישור לפרויקט שיפתח בדף ה-GitHub / Google Colab/ Azure** של אחד המשתתפים.

דף ה- GitHub / Google Colab/ Azure - יכיל את Assignment2_supervised_learning_flow.ipynb , קובץ ה-jupyter notebook, המכיל את כל הקוד של המטלה, על השלבים השונים, ואת הניסויים אשר עשיתם. יש ללוות את הקוד שלכם בהערות הסבר בגוף הקוד.

יש לבדוק את תקינות הקישורים לפני ההגשה (גם מבחינת גישה פתוחה לכולם וגם מבחינת התוכן העדכני).

פרטי המטלה:

על המטלה להפעיל flow של למידה מונחית (למידת סיווג או למידת רגרסיה, לפי בחירתכם).

- יש להסביר את כל השלבים אותם אתם עושים בסרטון, כאשר אתם מציגים את הקוד אותו תעלו לפרויקט ה-GitHub
- הניקוד יכלול גם הסבר ברור, שמראה שהבנתם מה שעשיתם

חלק 1 - פרטי הסטודנטים

עליכם למלא בכל שורה את השם הפרטי ו4 ספרות אחרונות של ת.ז. של כל סטודנט בקבוצה

חלק 2 – הניסויים (70 נקודות + אפשרות של עד 10 נקודות בונוס)

- על המטלה לכלול טעינת ה- trainset וה- testset (2 נקודות)
  - שימו לב – אין לחלק את ה-datasets הללו שוב ל- train ו- test.
  - עליכם להציג את 5 השורות הראשונות של כל dataset
- EDA – הצגת סטטיסטיקות וויזואליזציות על הנתונים (8 נקודות)
  - יש להציג לפחות 2 טבלאות ו2 וויזואליזציות.
- Feature engineering (20 נקודות)
  - עליכם להתנסות לפחות בסוג אחד של מטריקה של Feature engineering אותם למדנו, יש לזכור, שכל שלב של feature engineering אותם אתם מפעילים יש ללמוד מה-train ולהפעיל על ה-train ועל ה-test. מבחינת ההתנסות, תוכלו לבדוק שילוב של כמה מטריקות של feature engineering, סוגים שונים שלהם, איתם או בלעדיהם.
  - Feature engineering מורכב (יותר מהבסיס הנ"ל) יכול לתת עד 5 נקודות בונוס.
- אימון (20 נקודות)
  - יש לבדוק לפחות 2 אלגוריתמי למידה, מתוכם, לפחות אחד אותו למדנו. יש להתנסות לפחות עם 2 hyper parameters עבור כל אלגוריתם למידה.
  - יש להסביר קצת יותר אם מדובר באלגוריתם/ Hyper parameter אותו לא למדנו.
  - התנסות מורכבת של אימון ו- hyperparameters (יותר מהבסיס הנ"ל) יכולה לתת עד 5 נקודות בונוס.
- בחירת פרמוטציה המיטבית 5-fold-cross-validation בשיטת grid search (20 נקודות)
  - בחירת פרמוטציה של ה- Feature engineering, מודל הלמידה ו- hyper parameter המיטביים על 5-fold-cross-validation בשיטת grid search.
  - את התוצאות יש לבחור לפי r^2 בבעיות רגרסיה ולפי macro-average-f1 בבעיות סיווג בהם יש יותר ממחלקה אחת חשובה או עם f1 רגיל בבעיות סיווג בהם יש רק מחלקה חשובה אחת
  - שימו לב – עליכם להתנסות בכל האלמנטים הנ"ל ולהראות את התוצאות שנתנו כל אחד של מהאפשרויות, עם דגש, על האפשרות שנתנה את התוצאות הטובות ביותר על ממוצע ה- 5-fold-cross-validation
  - יש להראות טבלה מסכמת (dataframe) של השוואת התוצאות

חלק 3 – הפעלת ה- flow לפי הפרמטרים השונים (15 נקודות)

- לאחר בחירת הקומבינציה המוצלחת ביותר, עליכם לאמן את כל ה- train עם קומבינציה זו.

חלק 4 – הפעלה על ה- test set ושערוך המודל (15 נקודות)

- ליישם את ה- feature engineering הנבחר על ה-test ולחזות את דוגמאות ה- test (במודל הנבחר)
- יש להראות את הסיווגים הראשונים על ה- test ולהראות את איכות המודל (לפי התיאור לעיל ב cross validation).