

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

פרטים טכניים הנוגעים למטלה

הרשמה בטבלת המטלה

יש לרשום את פרטי הסטודנטים לבעיות ול-datasets באקסל המשותף במודל.

תאריך הגשת המטלה

את המטלה יש להגיש עד יום ראשון בערב ה-22 לספטמבר. הגשה באיחור עד ה-26 לספטמבר (קנס סימלי של חצי נקודה ליום על הגשה באיחור).

החומרים בהם יהיה מותר ואסור להשתמש

- מותר להשתמש ב-python בסיסי

- מותר להשתמש במודולים (ספריות/חבילות תוכנה): NumPy, Scipy, Pandas, Scikit-learn (sklearn), Matplotlib, Seaborn

- 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 ביחידים או בזוגות.
- יש להירשם באקסל המשותף את שמות המשתתפים ומס' ת.ז. של כל משתתף (לפי מה שמופיע במודל).
- שימו לב, שחלק 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 ביניהם). 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, 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 ו-4 ספרות אחרונות של ת.ז. של כל סטודנט בקבוצה

חלק 2 – הכנה – 10 נקודות

- **טעינה (2 נקודות)** - על המטלה לכלול טעינת ה- trainset וה- testset
 - שימו לב – אין לחלק את ה-datasets הללו שוב ל- train ו- test.
 - עליכם להציג את 5 השורות הראשונות של כל dataset
- **EDA (8 נקודות)** – הצגת סטטיסטיקות וויזואליזציות על הנתונים
 - יש להציג 4 תוצרים - לפחות 2 וויזואליזציות (אפשר גם להשתמש בטבלאות).

חלק 3 – הניסויים (60 נקודות + אפשרות של עד 10 נקודות בונים)

- **ניהול הניסויים עם cross validation (20 נקודות)** - בחירת פרמוטציה המיטבית 5-fold-cross-validation בשיטת grid search
 - 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 אחת חשובה
 - 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
 - את הניסויים הללו יש לבצע על הסעיפים המוסברים להלן ב- **התנסות ב- Feature engineering וב- hyper parameters**
- **התנסות ב- 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- validation. 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 בונים.
- **התנסות במודלים וב- hyper parameters (20 נקודות)**
 - יש לבדוק להתנסות לפחות עם 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 בונים.

חלק 4 – אימון - הפעלת ה-flow לפי הפרמטרים השונים (15 נקודות)

- לאחר בחירת הקומבינציה (של feature engineering, מודל ושל hyper parameters) המוצלחת ביותר (זו שנתנה את התוצאות הגבוהות ביותר, לפי הניסויים עם cross validation), עליכם לאמן **מחדש** (כלומר ביצוע feature engineering, אימון מודל ושל hyper parameter מחדש) את כל ה-train עם קומבינציה זו.

חלק 5 – חיזוי ובדיקת איכות - הפעלה על ה-test set ושערוך איכות המודל (15 נקודות)

- להשתמש ב-feature engineering, במודל וב-hyper parameters עליהם התאמתם מחלק 4 על ה-test ולחזות את כל דוגמאות ה-test
- יש להראות את תוצאות חיזוי 5 הסיווגים הראשונים על ה-test
- יש להראות את איכות המודל (לפי התיאור לעיל ב cross validation).