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-5 ליוני. הגשה באיחור עד ה-12 ליוני (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)
- GGPlot (plotnine), pygal ,bokeh , Matplotlib, Seaborn, pyplot,
- string, re, math, statistics
- מותר השימוש במודולים רלוונטים לנושאים מתקדמים, או אם אתם עושים מטלה בנושא עיבוד תמונה או ניתוח טקסט, <u>אם התקבל אישור על כך בפורום המטלה</u>
- אסור להשתמש בשום מודול (ספריות/חבילות תוכנה) נוסף מלבד אלו המוזכרים לעיל, <u>אלא אם כן ישנה סיבה מיוחדת לכך והתקבל אישור מיוחד בפורום המטלה</u>
- אסור להשתמש בשום קובץ חיצוני, <u>אלא אם כן ישנה סיבה מיוחדת לכך והתקבל אישור מיוחד בפורום המטלה</u>

הקבצים המצורפים למטלה:

קבצי data

- עבור כל dataset מופיעים קבצי csv עבור trainset ועבור test-set
- מכיוון שישנה אפשרות לעשות מטלות גם בנושאים של עיבוד תמונה ועיבוד טקסט, אז אפשר לבחור גם datasets בנושאים אלו מ- Kaggle, בקישורים הבאים:
  - עבור בעיות סיווג או רגרסיה בראיה ממוחשבת ועיבוד תמונה: https://www.kaggle.com/datasets?tags=13207-Computer+Vision
  - עבור בעיות סיווג או רגרסיה ניתוח טקסט ועיבוד שפה טבעית (NLP): https://www.kaggle.com/datasets?tags=13204-NLP

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, בזוגות או בשלשות. ניתן יהיה לצרף סטודנט רביעי מילואמניק (ואז יש לידע את המרצה), או באישור מיוחד מהמרצים (ולרי או משה, במייל המופיע במודל)
- יש להירשם באקסל המשותף את שמות המשתתפים והמייל 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 או לרשום עיבוד תמונה/ניתוח טקסט, הכוללים:
  - עבור למידת רגרסיה: Diabetes, House-pricing
  - עבור למידת סיווג: Titanic, Wine, Breast cancer Wisconsin (diagnostic)
  - עבור עיבוד תמונה/ניתוח טקסט, עליכם להוסיף אם מדובר בבעיית סיווג או רגרסיה וקישור אם ידוע כבר בשלב זה.

אופן ההגשה

כל משתתף ירשום בהגשה את 2 (או 3) 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, בו ניתן לצפות בסרטו). **על הסרטון להיות קצר באורך של כ 3-5 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.
3. הקישור ל- dataset במקרה שבחרתם במטלה בנושא עיבוד תמונה וניתוח טקסט

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 – הקדמה (בתחילת המטלה) – 10 נקודות

- **פרטי הסטודנט** – בתחילת המטלה, יהיה עליכם לרשום את השם הפרטי והאות הראשונה של שם המשפחה ובנוסף 4 ספרות אחרונות של ת.ז.
  - בהצגה בסרטון, יש להציג בהתחלה את שמות המשתתפים בברור
- **פרומפטים ב AI LLM או צ'ט בוטים, עזרים נוספים** יש להקדיש תא בו תכתבו את ה-prompt בו השתמשתם ב- AI chatbot, קישורים נוספים בהם נעזרתם ומה היתה המטרה של השימוש בהם – הדבר מותר, אך כמובן שעליכם להראות הבנה
  - **אנחנו מצפים שגם תהיה לכך התייחסות בעל פה.**
- **הסבר על בעית הלמידה וה-dataset** – נדרש סיכום קצר של הבעיה וה- dataset בתחילת קובץ ההגשה באורך של פסקה. עליכם להסביר בצורה קצת יותר מפורטת על כך בע"פ בסרטון.

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 וויזואליזציות על הנתונים
  - יש להציג 4 תוצרים - לפחות 3 וויזואליזציות (אפשר גם להשתמש גם בטבלה אחת במקום ויזואליזציה).
  - על הויזואליזציות, לשרת שלבים שונים ב- flow, כמו ניתוח מאפיינים, ניתוח תוצאות, הדגמת feature engineering, קשרים מעניינים וכדו'
  - יש להסביר בקצרה גם כן את מטרת הויזואליזציה

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

- **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 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
  - את הניסויים הללו יש לבצע על הסעיפים המוסברים להלן ב- *התנסות ב - Feature engineering* וב- *התנסות במודלים 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 בלעדיה, שימוש בפרמטרים שונים של feature engineering.
    - שימו לב, על ה- feature engineering להיות כחלק מה- cross validation
- **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.
- **בונוסים של עד20 נקודות**
  - התנסות עם Feature engineering מורכב יותר, כמו feature selection למשל, התנסות עם אלגוריתמי למידה חדשים אותם לא למדנו, שימוש במטריקות מתקדמות של הערכת איכות המודל, טיפול בבעיות מתקדמות כמו imbalanced data, יצירה סנטטית (מלאכותית) של דוגמאות, וגם בחירה של עבודות בתחום עיבוד תמונה או ניתוח טקסט
    - מצופה הסבר מתאים ויסודי יותר בכל חלקים אלה, הן בגוף המטלה ובמיוחד בהצגתה.

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 התאמנתם *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).