מטלה בלמידת מכונה – הפעלת flow של למידה מונחית – מסמך הסבר

שאלות על המטלה

שאלות אישיות, נא לפנות במייל למרצים, שאלות בתוכן נא להשתמש בפורום היעודי במודל.

על המטלה

מטרת המטלה לתרגל הפעלה של flow של למידת מכונה מההתחלה ועד הסוף.

- החלק המרכזי של העבודה כולל ניסויים שבהם תבחנו את הרכיבים המיטביים עבור בעיית למידת המכונה בה אתם עוסקים ועבור ה- dataset הספציפי, תוך שימוש ב- cross validation.

מסלולי הבחירה של המטלה

- מסלול רגיל – במסלול זה הדגש הוא על מימוש חלקים עיקריים ב-flow ברמת התשתית והבנה עמוקה שלהם.
- מסלול אפליקטיבי למעוניינים בלבד– במסלול זה הדגש הוא על אפליקציות של למידת מכונה, כמו ניתוח טקסט ו-NLP וההתעסקות המיוחדת בהם.

בסעיפים הבאים נתייחס לנקודות הבאות:

פרטים טכניים הנוגעים למטלה

הקבצים המצורפים למטלה

קבצי data

- מסלול רגיל – עליכם לבחור את אחד מה- datasets המופיעים במודל, להבין את הבעיה ואת ה-data ולזהות את סוג הבעיה (ולבחור את מדד האיכות המתואר בהמשך) בהתאם. עבור כל dataset מופיעים קבצי csv עבור trainset ועבור test-set (כמובן שאין לבצע פיצול נוסף של ה- trainset לצורך test-set).
  - ה- datasets המופיעים במודל: Diabetes, Wine, Breast cancer Wisconsin (diagnostic)
    - **עבור מגיש/ה יחיד/ה <u>בלבד</u>** – ניתן גם לעבוד עם ה- datasets של House-pricing ו- Titanic
- מסלול אפליקטיבי – תוכלו לבחור אפליקציה של עיבוד תמונה או עיבוד טקסט או סייבר.
  - מי שבוחר באפליקציה של סייבר יש dataset במודל בנושא packet classification כאנומליות או לא אנומליות.
  - עבור בעיות סיווג או רגרסיה לאתגר בראיה ממוחשבת ועיבוד תמונה עליכם לבחור dataset בנושאים אלו מ- Kaggle, בקישורים הבאים:
    - בראיה ממוחשבת: https://www.kaggle.com/datasets?tags=13207-Computer+Vision
    - בניתוח טקסט (ו- NLP): https://www.kaggle.com/datasets?tags=13204-NLP

מחברת הגשה ריקה להגשת התרגיל

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

הרשמה בטבלת המטלה (קישור מופיע במודל)

- ניתן להגיש את העבודה בין ביחידים או בקבוצות של עד 5 סטודנטים.
- יש להירשם באקסל המשותף את הפרטים הבאים:
  - **Assignment_type** (סוג המטלה) – יש למלא implementation (עבור מטלת רגיל), או application (עבור מטלה אפליקטיבית, כמו ניתוח טקסט וכדו')
  - **Supervised_learning_type** (סוג בעיית הלמידה המונחית) – יש למלא classification או regression
  - **Dataset_Name** – יש למלא את שם ה- dataset, כפי שמופיע במודל למטלה רגילה או כפי שמופיע ב-Kaggle
  - **dataset_URL** – יש למלא את הקישור ל- dataset ב- Kaggle (רלוונטי רק בסוג מטלה אפליקטיבית).
  - **Video_URL** – יש למלא את הקישור לסרטון ההסבר של המטלה (יש לוודא הרשאות צפיה, לכל מי שיש לו קישור). הקישור ל- YouTube או לכל מקום אחר בו ניתן לצפות בסרטון (ללא הרדה של הסרטון)
  - **Repository_URL** – יש למלא את הקישור ל-repository בו העליתם את המטלה, אשר ניתן לצפות במטלה (יש לוודא הרשאות צפיה, לכל מי שיש לו קישור), ללא הורדה שלה וניתן לצפות גם בתוצאות ההרצה (ההרצה של המטלה היא על אחריותכם).
  - **פרטי הסטודנטים -** בכל קבוצה יכולים להיות עד שלושה סטודנטים. יש לרשום את פרטי הסטודנטים:
    - **contact_emails**: פרטי הקשר של הסטודנטים, רק אם הם שונים מהמייל במודל. יש להפריד ע"י פסיק (,) במידה ויש יותר ממייל אחד.

- **student_name_x** – יש לכתוב את השם בעברית כפי שמופיע מודל (בעמודות student_name_1, student_name_2, student_name_3, student_name_4 ו- student_name_5)
- **email_student_x** – יש לכתוב את המייל כפי שמופיע במודל (בעמודות email_student_1, email_student_2, email_student_3, email_student_4 ו- email_student_5)

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, pyplot , bokeh, pygal (plotnine), GGPlot,
- string, re, math, statistics
- מותר השימוש במודולים רלוונטיים 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>

תוצרי ההגשה

התוצרים ישלחו בקישורים (כפי שמתואר בהמשך)

- הקישורים צריכים להיות זהים הן בהגשה אצל כל הסטודנטים בקבוצה והן בקובץ ההרשמה
- **<u>יש לבדוק את תקינות כל אחד מהקישורים מבחינת הגישה ומבחינת עדכניות התוכן</u>**
  - וודאו שישנה גישה לכל אחד מהקישורים, עבור כל מי שקיבל את הקישור.

הסרטון

- מירב הערכה של המטלה תתקיים מול הסרטון שתכינו.
- **על הסרטון להיות קצר באורך של כ 4-5 דקות (לא יותר)**
- בהצגת המטלה, עליכם להציג את הדברים, בהנחה שמי שצופה בסרטון אינו יודע את החומר.
- בתחילת הסרטון, עליכם להציג את הסטודנטים בקבוצה.
- עליכם להציג את החלקים השונים במטלה, כשאתם מראים את הקוד ואת הפלט ומלווים אותו בהסברים, שמראים הבנה של מה שעשיתם ושל התוצאות.
- אנחנו מצפים מכם לשתף את כל חברי הקבוצה בסרטון, במידה שווה ככל האפשר.
- המיקוד בסרטון הוא של הקוד והתוצרים ואין צורך להראות את חברי הצוות.
- אי הגשת סרטון, או סרטון לא זמין או לא תקין, תגרור קנס משמעותי בציון.
- **<u>סיכום</u> חובה** לשלוח קישור לסרטון ברור, מקיף וכולל שמראה וסוקר את העבודה שלכם ואת התוצרים באופן מקיף וברור ומראה **שהבנתם אותם** והם פועלים כראוי.

מחברת הקוד

- **התוכן הכלול במטלה מוסבר בסעיף *פרטי המטלה* בהמשך**.
- וודאו שישנה גישה לסרטון לכל מי שקיבל את הקישור.
- על המחברת לכלול גם את התוצרים והויזואליזציות, ללא הורדה והרצה במחשב של מי שיבדוק אותה.
- על המחברת להיות מלווה בהערות והסברים קצרים שמסבירים את העבודה והתוצרים.
- שימו לב – ההערכה במטלה היא בעיקרה על תהליך הניסויים.
- אי הגשת מחברת הפתרון, או מחברת הפתרון לא זמינה או לא תקינה, תגרור קנס משמעותי בציון.
- **<u>סיכום</u> חובה** לשלוח **<u>קישור</u>** מחברת מטלה ברורה, מקיף וכולל שמראה וסוקר את העבודה שלכם ואת התוצרים באופן מקיף וברור

תאריך הגשת המטלה

את המטלה יש להגיש עד יום ראשון בערב ה- 14.9ל ההגשה באיחור עד יום חמישי בערב, ה- 18.9 ל (קנס סימלי של חצי נקודה ליום על הגשה באיחור).

אופן ההגשה

- נבסס את ההגשה על אקסל ההרשמה והנתונים הבאים הינם לצורך גיבוי.

כל סטודנט/ית בקבוצה ירשום בהגשה את 2 (או 3) הקישורים הבאים (עם הפרדה של רווח בינהם).

- ההגשה באופן המתואר להן, הינה לצורך גיבוי. המטלה תיבדק רק פעם אחת לכולם.
- **יש לבדוק את תקינות כל אחד מהקישורים מבחינת הגישה ומבחינת עדכניות התוכן**
- דוגמא להגשה במסלול רגיל (הקישורים צריכים להיות זהים אצל כל הסטודנטים בקבוצה):

https://youtu.be/kqtD5dpn9C8 https://git.new/FF6Dp2F

- דוגמא להגשה במסלול אפליקטיבי (הקישורים צריכים להיות זהים אצל כל הסטודנטים בקבוצה):

https://youtu.be/kqtD5dpn9C8 https://git.new/FF6Dp2F https://tinyurl.com/bdezpa8x

1. הגשת חובה – **קישור לסרטון** (תצטרכו להעלות את הסרטון ל- YouTube, או למקום אחר ברשת, בו ניתן לצפות בסרטו)., בו אתם מציגים ומסבירים את עבודתכם ואת התוצאות.
2. הגשת חובה – **קישור למחברת הקוד שיפתח בדף ה-GitHub / Colab / Azure** של אחד המשתתפים ויכיל קובץ ה-jupyter notebook (Assignment_supervised_learning_flow.ipynb), עם המימוש, ההרצה וההערות.
3. **במסלול האפליקטיבי** - יש להוסיף את הקישור ל- dataset במקרה שבחרתם במטלה (כמו בדוגמא השניה לעיל).

מה עליכם לבצע במטלה?

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 – פרטי הסטודנטים, פרומפטים, על ה- dataset (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 הייתה המטרה של השימוש בהם – הדבר מותר, אך כמובן שעליכם להראות הבנה
  - **אנחנו מצפים שגם תהיה לכך התייחסות בעל פה.**
- **הסבר על בעיית 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 – טעינה, EDA – 10 נקודות

- **טעינה** (2 נקודות) - על המטלה לכלול טעינת ה- trainset וה- test-set
  - 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 ויזואליזציות על הנתונים
  - יש להציג 4 תוצרים - לפחות 3 ויזואליזציות (אפשר גם להשתמש גם בטבלה אחת במקום ויזואליזציה).

- למגיש/ה יחיד – מספיקים 2 תוצרים.
- 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 הוויזואליזציה

חלק 3 – הניסויים (70 נקודות)

מדד האיכות

- בבעיות רגרסיה – האיכות תוערך לפי מדד r^2.
- בבעיות סיווג multi-class (או עם 2 מחלקות, אך ללא מחלקה מרכזית) – macro-average-f1
- בבעיית סיווג בינארית (עם מחלקה מרכזית אחת בלבד) – f1 (רק על המחלקה המרכזית)
- **השימוש במדד האיכות**
  - לצורך בחירת הפרמוטציה המוצלחת ביותר ב- cross validation (ממוצע המדד הגבוה ביותר ב-5 ה-folds)
  - לצורך שיערוך האיכות על ה- test-set
- **Feature engineering** –
  - מטריקות שיש להתנסות בהם, הינם מטריקות אותם למדנו, כמו סילום, feature selection וכדומה.
  - 1-2 מגישים – התנסות בלפחות מטריקה אחת של feature engineering – יש להתנסות לפחות בקונפיגורציה אחת או ללא אותה קונפיגורציה (למשל סילום עם standardization או ללא סילום כלל).
  - 3-5 מגישים - צריכים להתנסות בלפחות 2 מטריקות, אך עם 2 קונפיגורציות (למשל סילום עם standardization ו- minmax ו- feature selection, על ידי info-gain או לפי שכיחות) לכל מטריקה, או 3 מטריקות עם קונפיגורציה אחת.
- **התנסות במודלים וב- hyper parameters**
  - יש להתנסות באלגוריתמי למידה שונים עם hyper parameters שונים.
  - מגיש יחיד – התנסות בלפחות אלגוריתם למידה אחד, אך עם לפחות 2 hyperparameters, עם 2 ערכים לכל hyperparameter.
  - 2-3 מגישים – התנסות בלפחות 2 אלגוריתמי למידה עם לפחות 2 hyperparameters, עם 2 ערכים לכל hyperparameter.
  - 4-5 מגישים – התנסות בלפחות 3 אלגוריתמי למידה עם לפחות 2 hyperparameters, עם 2 ערכים לכל hyperparameter.
  - יש כמובן להראות הבנה גם אם מדובר באלגוריתם/ hyperparameter אותו לא למדנו.

ניהול הניסויים עם grid-search k-fold cross-validation

- **grid-search** – עליכם לבצע grid-search, בו תבחנו את כל הפרמוטציות. כלומר, מכפלה קרטזית של הפרמוטציות של feature engineering, המודלים השונים שנוצרים ע"י אלגוריתמי הלמידה אותם בחרתם עם ה-hyperparameters השונים.
- **K-fold cross validation** – 5-fold cross validation שיעטוף את כל החלקים בניסויים. יש לחלק את ה-trainset ל-5 חלקים שווים ובכל פעם משתמשים ב-4 חלקים לאימון ו-1 ל-validation, תבצעו כל פרמוטציה מה- grid search 5 פעמים (1 לכל fold) ותבחנו את המדד על ה- חלק הנוסף (validation). עליכם לחשב מדד ממוצע המדד ל-5 ה- folds לכל פרמוטציה.
- הצגת כל הפרמוטציות והתוצאות (הממוצעות) של המדד ב- dataframe
- ציון הפרמוטציה הטובה ביותר.
- בשלב האימון (בהמשך) עליכם לאמן את כל ה- trainset עם הפרמוטציה המוצלחת ביותר.

חלק *3 – המלצות למסלול אפליקטיבי

כללי – בסעיף זה מתואר המסלול האפליקטיבי. במסלול זה, מצופה ללמוד קצת מעבר למה שלמדנו בקורס, כדי לדעת איך להתעסק עם סוג הבעיה וב-data מקצה לקצה.

הערה חשובה – המסלול האפליקטיבי, דורש יותר עבודה והוא מומלץ בעיקר להתפתחות אישית ולמי שרוצים להציג תיק עבודות.

התאמות מיוחדות של ה- data– התאמות המיוחדות של ה- data הדרושות לבעיה הספציפית, ל- dataset ולסוג האפליקציה, טיפולים שונים ב- imbalanced data שמבוססות בעיקר על under-sampling, over-sampling, וייצור של דוגמאות סינטטיות.

Feature Engineering מיוחד – מצופה לבצע עיבוד מוקדם ייעודי ומיוחד הדרוש לבעיה הספציפית, ל- dataset ולסוג האפליקציה, למשל שימוש בדמיון ווקטורי של word embeddings לצורך NLP, "SYN/ACK/FIN" rates ב- cyber security או Data Augmentation לראיה ממוחשבת.

בדיקת איכות מיוחדת – מצופה לבצע בדיקת איכות ספציפית ומיוחדת נוספת הדרושה לבעיה הספציפית, ל- dataset ולסוג האפליקציה, בנוסף למדדים המתוארים לעיל (או במקומם, במידה ואין להם מקום בבעיה הספציפית), כמו למשל Intersection over Union (IoU) לבעיות localization בבעיות ראיה ממוחשבת, BLEU Score לאפליקציות תרגום ב-NLP או מדדים מבוססים Tactics, Techniques, and Procedures (TTPs) ב-cyber security .

Explainability - הבנת התוצאות והנגשתם והנגשה של המאפיינים – הבנה וניתוח של הנתונים החשובים, בעזרת טכניקות המוטמעות במידע שמוחזק כתוצאה מהלמידה, שמוחזר עם המודל עם החיזוי, או בטכניקות שונות שמנתחות את החיזויים והמאפיינים לאחר החיזוי, כמו SHAP.

הגשות כתלות בכמות הסטודנטים בקבוצה

- **1-2 מגישים** – יבצע לפחות אחד מהרכיבים המתוארים לעיל: התאמות מיוחדות של ה-data, Feature Engineering מיוחד, בדיקות איכות מיוחדות או Explainability - הבנת התוצאות והנגשתם.
- **3-5 מגישים** – יבצע לפחות אחד מהרכיבים המתוארים לעיל: התאמות מיוחדות של ה- data, Feature Engineering מיוחד, בדיקות איכות מיוחדות או Explainability - הבנת התוצאות והנגשתם.

הרצת ה- flow – כמתואר לעיל, בסעיף *3.א. הניסויים – כללי*

- ***רק לצורך הניקוד ממליצים להגיש flow בסיסי במסגרת הזמנים***

חלק 4 – אימון - הפעלת ה- flow לפי הפרמטרים השונים (10 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 (10 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).