**מטלה מסכמת**

סמסטר ב' תשפ"א

1242324801 – גילוי ידע ורשתות נוירונים

ד"ר אילן ששון

**כללי**

* מטרת המטלה: התנסות מעשית ותרגול שיטות בגילוי ידע ורשתות נוירונים שנלמדו במהלך הקורס.
* סביבת העבודה: python **בלבד**.
* ניתן להשתמש בחבילות המובנות לצורך כתיבת המטלה.
* ניתן להשתמש ב-IDE לבחירתכם (דוגמת jupyter notebook או pycharm).
* הרכב המטלה: המטלה מורכבת מ-2 חלקים:
  1. **חלק א'** – חיזוי רכישה של מוצר פרמיום. קובץ הנתונים מכיל מידע על משתמשים באתר והאם חזוי שהם ירכשו בעתיד מוצר פרמיום או שלא.
  2. **חלק ב'** – משימת clustering.
* ציון המטלה: ייקבע לפי הפרמטרים הבאים:
  + הערכת אופן ביצוע המטלה ואיכותה.
  + הערכת ביצועי המודל על test-set חיצוני, מול הביצועים של קבוצות אחרות בכיתה (בסגנון תחרות Kaggle) - רלוונטי לחלק א' של המטלה.
* חלק ההוראות מכיל סדר פעולות מפורט אותו יש להשלים. כל פעולה נוספת או החלטה צריכה להיות מלווה בהסבר.
* דף הסברים: מצורף לקבצים.
* שאלות לגבי המטלה: ניתן לשאול שאלות הקשורות במטלה **אך ורק** בפורום הייעודיב-Moodle.
* עוזר ההוראה בקורס: גיא קרן ([guykern@mail.tau.ac.il](mailto:guykern@mail.tau.ac.il)).
* הגשת המטלה: תתבצע **בזוגות בלבד** (יש להירשם לקבוצות במודל).
* אופן הגשת המטלה: יוגש קובץ zip בשם <group\_number> הכולל את הסקריפט כפי שנרשם ב-IDE בו בחרתם להשתמש וקובץ הכולל את הפלט שהתקבל ממודל החיזוי (חלק א' של המטלה).

**לתשומת ליבכם, על הקבצים לכלול הסברים ועל הקוד להיות מתועד, מסודר וקריא!**

**הוראות**

הערות כללית:

* + - יש למלא אחר כל הסעיפים.
    - ניתן להוסיף ניתוחים נוספים וויזואליזציות כדי להעשיר את העבודה.

**חלק א'**

בחלק זה תשתמשו בקבצים הבאים:

* ctr\_dataset\_train
* ctr\_dataset\_test

1. **טעינת הנתונים**

* הורידו את קובץ הנתונים ושמרו אותו כ-pandas dataframe.
* הציגו את השורות הראשונות של ה-dataframe ששמרתם.
* עבור משתנה המטרה, הקטגוריה 0 מייצגת את "לא ירכוש" ו-1 את הקטגוריה "ירכוש" כ-1.

1. **Data exploration**
   * הציגו סיכומים סטטיסטים שונים אודות הנתונים (הנמצאים ב-dataframe). על הסיכומים הסטטיסטים לכלול לפחות 5 מדדים סטטיסטיים שונים.
   * הציגו את מטריצת המתאם (*Correlation matrix*) של הנתונים. אם החלטתם שלא לכלול חלק מהמשתנים במטריצת המתאם, ציינו מהם והסבירו מדוע.
   * החליטו האם יש צורך להסיר משתנים מה-dataframe. אם כן, ציינו מהם המשתנים שבחרתם להסיר והסבירו מדוע.
   * בדקו האם הקטגוריות של משתנה המטרה מאוזנות (*Class imbalance*). אם הקטגוריות אינן מאוזנות, אתם רשאים לבחור בשיטות הדגימה SMOTE או ADASYN. הסבירו בקצרה את ההיגיון העומד בבסיס השיטה בה בחרתם, והאם מימושה עשוי לשפר את ביצועי המודל.
2. **Missing values**
   * בדקו האם קיימים ערכים חסרים בנתונים באמצעות סיכומים (*Summaries*).
   * רשמו כיצד בחרתם להתמודד עם הערכים החסרים בנתונים והסבירו מדוע.
3. **Feature engineering**

הוסיפו לפחות 5 משתנים חדשים על בסיס המשתנים הקיימים.

1. **Data normalization**

אתם רשאים לנרמל את כל הנתונים או את חלקם. אם בחרתם לעשות כן, נמקו את בחירתכם.

1. **Training**
   * חלקו את הדאטה ל-3 חלקים:
2. Train
3. Validation
4. Test
   * אמנו לפחות 2 מודלים כאשר לפחות אחד מהמודלים הוא מודל של רשתות נוירונים והציגו עבור כל אחד מהם את המדדים הבאים:
5. Precision
6. Recall
7. Accuracy
8. AUC
   * בצעו כיול היפר-פרמטרים (*Hyperparameter tuning*) לכל אחד מהמודלים.
   * ציינו מיהם הפרמטרים שאותם בחרתם לכייל והסבירו בקצרה על מה הם משפיעים באלגוריתם.
   * בחרו את המודל שישמש אתכם בשלבים הבאים והסבירו מדוע חברתם במודל זה.
9. **Explainable AI**

* השתמשו ב-SHAP על מנת לפרש את פלט החיזוי המתקבל על-ידי המודל שבחרתם.
* הסבירו מהו shap value.
  + **Global interpretability** – הציגו את ה-summary\_plot, והסבירו את המשמעות של הגרף.
  + **Local interpretability** – הגרילו 5 שורות באקראי מתוך הנתונים והשתמשו ב- shap\_plot כדי להסביר את תוצאות המודל עבור כל אחת מהן.

1. **Inference**

* עליכם לבחון את המודל שבניתם בשלב הקודם על אוסף נתוני test חיצוני (ctr\_dataset\_test).
* בצעו את שלבי העיבוד המקדים (*Preprocessing*) שביצעתם בשלב אימון המודל (אוסף נתוני train) גם על אוסף נתוני ה-test.
* הריצו את המודל וחזו את הערך של משתנה המטרה על אוסף נתוני ה-test.

**פלט של חלק א'**

* הפלט של שלב זה יוגש כקובץ בשם output\_<group\_number>.txt המכיל את הערכים שחזה המודל עבור כל שורה (לפי סדר השורות).

**חלק ב'**

בחלק זה תשתמשו בקובץ הבא:

* Clustering.csv

1. שמרו את הקובץ כ-dataframe.
2. הריצו KMeans ו-DBSCAN על הקובץ.
3. מצאו את הפרמטרים האופטימליים עבור כל אחד מהמודלים, והסבירו כיצד עשיתם זאת. הוסיפו גרפים מתאימים.
4. בחנו את ביצועי המודלים בעזרת מדדים המתאימים לבעיות *clustering*, בחרו שני מדדים והשוו ביניהם.
5. הציגו את אופן החלוקה של כל מודל לאשכולות (*Clusters*), כל אחד בגרף נפרד.

**בהצלחה!**