מבוא למדעי הנתונים – תשפ״א

מטלה מסכמת

כללי:

- מטלה זו מאפשרת התנסות מעשית ותרגול של עקרונות ושיטות שנלמדו במהלך הסמסטר.
- סביבת העבודה: python3 בלבד. ניתן להשתמש בחבילות המובנות לצורך כתיבת המטלה. שימוש ב-jupyter notebook או בקבצי py – לבחירתכם.
- המטלה מורכבת משני חלקים:
  - חלק א׳: חיזויי קליקים על לינק או פרסומת - Click-Through Rate Prediction. קובץ הנתונים מכיל מידע על משתמשים (אפליקציה שבה השתמשו, סוג המכשיר, וכד׳) והאם הקליקו או לא.
  - חלק ב׳: משימת clustering.
- ציון המטלה ייקבע לפי:
  - 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.
- שאלות על המטלה יש לשאול רק בפורום המתאים ב-Moodle.
- מייל עוזרת ההוראה: reutregev@mail.tau.ac.il

הגשה:

- הגשה בזוגות בלבד. יש להירשם לקבוצות במודל.
- יש להגיש קובץ zip בשם <group_number> הכולל קבצי jupyter notebook, קבצי py וקובץ חיזויי המודל (חלק א׳ של המטלה).

  על הקבצים לכלול הסברים ותיעוד של הקוד. בבדיקה יינתן דגש על כתיבת קוד מסודר וקריא!

הוראות:

יש למלא את כל הסעיפים. ניתן להוסיף ניתוחים נוספים וויזואליזציות כדי להעשיר את העבודה.

חלק א׳:

בחלק זה תשתמשו בקבצים: ctr_dataset_train, ctr_dataset_test. הקובץ Readme מכיל מידע קבצי הדאטה.

1. **טעינת הנתונים**:
הורידו את קובץ הנתונים, שמרו אותו כ-pandas dataframe, והציגו את השורות הראשונות של הטבלה.
2. **Data exploration**:
   - הראו סטטיסטיקות שונות של הדאטה.
   - הציגו את מטריצת הקורלציות של הדאטה.
   החליטו האם יש צורך להסיר עמודות מהדאטה. אם כן, הסירו אותן והסבירו.
   - בדקו האם הדאטה מאוזן. במידה ולא, טפלו בכך באמצעות אחת מהשיטות: SMOTE, ADASYN, או שתיהן ביחד. הסבירו את השיטה שבחרתם. בחנו האם מימוש השיטה משפר את ביצועי המודל.
3. **Missing values**:
   - בדקו האם קיימים ערכים חסרים בדאטה והציגו נתונים סיכומיים על כך.
   - עבור עמודות שמכילות ערכים חסרים, החליטו כיצד לטפל בהם והסבירו את החלטתכם.
4. **Feature engineering**:
הוסיפו לפחות 4 עמודות נוספות על בסיס העמודות הקיימות והסבירו מדוע בחרתם להוסיף עמודות אלו.
5. **Data normalization**:
בצעו נורמליזציה של הנתונים לפי מידת הצורך (בהתאם למודלים שאתם מתכננים להריץ), באיזו שיטה שתבחרו.
6. **Training**:
   - חלקו את הדאטה ל-train, validation, test.
   - אמנו שלושה מודלים והציגו עבור כל אחד מהם את המדדים הבאים:
   recall, precision, AUC, accuracy.
   חוו את דעתכם על המדדים שהתקבלו.
   - בצעו hyperparameters tuning לכל אחד מהמודלים. הסבירו בקצרה מהם הפרמטרים שאותם בחרתם לכייל.
   - בחרו את המודל בעל המדדים הגבוהים ביותר שישמש אתכם לשלבים 7-8.

7. **Explainable AI**:

השתמשו ב-SHAP על מנת לפרש את חיזויי המודל.

- הסבירו מהו shap value.
- Global interpretability – הציגו את ה-summary_plot, הסבירו את המשמעות של הגרף ומה ניתן להסיק ממנו.
- Local interpretability – הגרילו שלוש שורות רנדומליות מהדאטה והשתמשו ב-shap_plot כדי להסביר את תוצאות המודל עבור כל אחת מהן.

8. **Inference**:

בשלב זה תבחנו את המודל שבניתם בשלב הקודם על test-set חיצוני (קובץ ctr_dataset_test).
עליכם לבצע את שלבי ה-pre-processing שביצעתם בשלב ה-train גם על ה-test-set. לאחר מכן, השתמשו במודל כדי לחזות האם בוצעה הקלקה או לא.
הפלט של שלב זה הוא קובץ בשם output_<group_number>.txt המכיל את חיזויי המודל עבור כל שורה (לפי סדר השורות), כאשר 1=clicked, 0=not clicked. לדוגמה:

```
1
0
1
1
1
0
```

9. **בונוס**:

בעולם ה-Online advertising עושים שימוש במודלים מסוג זה כדי להציג למשתמשים פרסומות או לינקים שקיימת סבירות גבוהה יותר שהם יקליקו עליהם. מה לדעתכם הבעיה שעלולה לצוץ בעת שימוש במודל ML כחלק מתהליך אימון מתמשך (Continuous learning)? כתבו בקצרה פתרון אפשרי לבעיה.

חלק ב׳:

בחלק זה תשתמשו בקובץ clustering_data.

1. הריצו KMeans ו-DBSCAN על הדאטה.
2. מצאו את הפרמטרים האופטימליים למודלים. הסבירו והראו כיצד מצאתם אותם, הוסיפו גרפים מתאימים.
3. בחנו את ביצועי המודלים בעזרת מדדים המתאימים לבעיות clustering והשוו בין שתי השיטות שהצגתם.
4. הציגו את אופן החלוקה של כל מודל לקאלסטרים (בגרפים נפרדים).

בהצלחה!