סמסטר ב' תשפ"ה | למידת מכונה 364-1-1811

**פרויקט חיזוי סנטימנט - חלק ב'**

לאחר שבחלק א' התמודדתם עם דילמות מעולם מדע הנתונים, בהקשר לחקר הנתונים ויצירת dataset, בחלק זה תשתמשו בסט הנתונים שהכנתם לצורך אימון ובחינה של מודלי למידת מכונה שונים, השוואה ביניהם ושיפורם.

דגשים:

* **קובץ הדו"ח, קוד ה python שכתבתם וקובץ הטסט** יוגשו לתיבת ההגשה במודל עד לתאריך ה-**26.06.2025 בשעה 59:23.** מספיקה הגשה של אחד מבני הזוג.
* **אורך הדו"ח לא יעלה על 10 עמודים** (לא כולל שער ותוכן עניינים), בגודל כתב 12, פונט Arial ורווח של שורה וחצי. פלטים מרכזיים וגרפים יש לצרף **בגוף הדו"ח** **ולא בנספחים**.
* אנא שמרו על תמציתיות והתמקדו בתובנות המרכזיות. בחלק זה ישנן שאלות פתוחות, **ענו עליהן בקצרה (שורה - שתיים).**
* בכתיבת הדו"ח יש להשתמש **במונחים מקצועיים** בהקשרים הרלוונטיים.
* יש להגיש קובץ Word המאפשר השארת הערות על גביו - **אין להגיש קובץ PDF**.

שאלותבנוגע לפרויקט, יש לפרסם בפורום הייעודי שייפתח במודל כדי שחברכם לקורס יוכלו ללמוד מהן***.***

# ***לפני שמתחילים (5 נק')***

* **חלקו את סט הנתונים לסט אימון וולידציה** בהתאם לשיטת הוולידציה שבחרתם בחלק א'. כדי לשמור על הדירות (reproducibility), קבעו random state כלשהו בעת החלוקה.
* מהי **זליגת נתונים** (Data leakage)? הסבירו את הצעדים שביצעתם כדי להימנע מכך

שימו לב: מכוון שמשתנה המטרה לא מאוזן, **לא נבחן את המודלים על פי קריטריון Accuracy, כל המודלים ימדדו לפי AUC-ROC**

# ***Model Training (55 נק')***

# ***Decision Trees 1.1 (15 נק')***

1. אימון וכוונון היפר-פרמטרים

* בנו עץ החלטה ובצעו תהליך של כוונון היפר-פרמטרים על מנת למצוא את הקונפיגורציה האופטימלית של המודל. ציינו אילו פרמטרים כיוונתם, מה משמעותם ואיזה טווח ערכים ניסיתם עבורם.
* מה **ביצועי המודל** (AUC-ROC) המתקבלים על סט האימון וסט הוולידציה עבור המודל הטוב ביותר? מה ניתן להסיק מתוצאות אלה?

1. פירוש והבנת המודל

* אחד היתרונות של מודל עץ החלטה הוא יכולת **פירוש המודל** (interpretability), הסבירו את חשיבות תכונה זאת בהקשר חיזוי הסנטימנט וכיצד באה לידי ביטוי בעץ החלטה.
* הציגו גרף של השכבות העליונות של העץ שהתקבל
* התבוננו בגרף שייצרתם ונסו להפיק ממנו תובנות לגבי דפוסים בנתונים ואופן קבלת ההחלטות של המודל.
* חשבו את חשיבות הפיצ'רים (importances feature) של המודל. איזה מסקנות בנוגע למשימת הלימוד ניתן להסיק? האם קיימת הלימה בין תכונות אלו לבין הצמתים הראשיים בעץ?

# ***Artificial Neural Networks 1.2 (17 נק')***

1. אמנו מודל MLP עם ערכי ברירת המחדל של ספרייה שתבחרו לצורך האימון.

* תארו את **מבנה המודל** והסבירו את משמעות המרכיבים השונים שלו (מספר נוירונים בשכבת הכניסה, מספר שכבות חבויות ומספר נוירונים חבויים בכל שכבה).
* מהם הביצועים המתקבלים על סט האימון וסט הוולידציה? מה ניתן להסיק מתוצאות אלה?

1. בחרו מספר **היפר-פרמטרים** שברצונכם לכוון (כגון: מספר שכבות חבויות, מספר נוירונים בכל שכבה, learning rate, גודל batch, פונקציית אקטיבציה וכו').

* עבור כל היפר פרמטר שבחרתם: הסבירו מה תפקידו ברשת, מה המשמעות של הגדלה או הקטנה של ערכו, ואילו טווחי ערכים בחנתם ולמה.
* הציגו בטבלה את ערכי ההיפר-פרמטרים שנבחנו כפונקציה של אחוז הדיוק על סט האימון והוולידציה. מה ניתן להסיק ממצאים אלו?
* מהם אחוזי הדיוק המתקבלים על סט האימון וסט הוולידציה באמצעות המודל המדויק ביותר? אם קיים שיפור, האם ניתן לקשור אותו לבחירת ערכי ההיפר-פרמטרים? פרטו.

1. הציגו גרף של הביצועים והLoss כפונקציה של **מספר ה epoch**, עבור סט האימון והוולידציה. האם המודל מתכנס? ציינו באיזה מספר תבחרו ומדוע.

**הערה**: בהרצאה לא למדנו על כל הפרמטרים שניתן לכוון במודל MLP. מומלץ להיעזר בדוקומנטציה של הספרייה שבחרתם כדי להבין איזה פרמטרים ניתנים לכוונון.

# ***SVM 1.3 (8 נק')***

אמנו מודל SVM ובצעו Hyperparameter Tuning למציאת הקונפיגורציה המיטבית - **השתמשו בKERNEL לינארי**

* הציגו את אחוזי הדיוק של המודל הנבחר על סט האימון והוולידציה
* הציגו את משוואת הישר המפריד, מה ניתן להבין ממשואת הישר בנוגע לחשיבות כל פיצ'ר לצורך הסיווג? האם תואם מסקנות קודמות שלכם?

# ***Clustering 1.4 (15 נק')***

1. במה שונה שיטת ניתוח אשכולות (Clustering) לעומת המודלים הקודמים שהרצתם? התייחסו להבדל בסט הנתונים ולמטרת משימת הלימוד

* בחרו סט של פיצ'רים לצורך ביצוע Clustering, והריצו עליו את **אלגוריתם K-medoids עם ערכי K שונים**. ציינו באיזה סט פיצ'רים ברחתם, ובאיזו מטריקת למדידת מרחק ברחתם. הסבירו את בחירותכם.

1. השוו את תוצאות האישכול עבור ערכי K שונים שבחנתם.

* ציינו באילו מדדים השתמשתם כדי להשוות בין האשכולות ומדוע.
* מהו מספר המחלקות בו תבחרו?
* האם יש קשר בין הקבוצות השונות למשתנה המטרה?

# ***Evaluation (10 נק')***

1. בחירת מודל

* השוו בין ביצועי שלושת מודלי הסיווג (SVM, DT, MLP) בטבלה מסכמת
* על איזה מהמודלים תמליצו? הסבירו מדוע (ניתן לערב שיקולים שאינם מדדי ביצוע)
* האם יש מגבלות למודל שבחרתם? תארו מצב בו חסרון זה בא לידי ביטוי.

1. ניתוח ויזואלי של ביצועי המודל

* צרו עקומת ROC על המודל שבחרתם (כולל השוואה למסווג הנאיבי), וסמנו את נקודת הסף (threshold) שנבחרה.
* צרו מטריצת מבוכה. האם קיימת הטיה כלפי אחת מהמחלקות? כיצד ניתן להסביר זאת בהתבסס על הנתונים?

1. סיכום

* התבוננו בחשיבות הפיצ'רים שביצעתם עבור כל מודל. האם קיימת הטיה כלפי אחת מהמחלקות? כיצד ניתן להסביר זאת בהתבסס על הנתונים?
* האם נתקלתם במצב של overfitting בעת יצירת אחד המודלים? כיצד זיהיתם זאת? (אם לא מצאתם – כיצד ניתן לזהות שמדובר בתופעה), תארו שיטה אחת להתמודדות.

# ***Improvements (15 נק')***

1. בסעיף זה תידרשו להציע וליישם דרכי שיפור למודל הטוב ביותר שבניתם, תוך התבססות על ידע נוסף מעבר לחומר שנלמד בקורס. שיפורים אלו עשויים לתרום לשיפור הביצועים ולחיזוי מדויק יותר בסעיף הבא. הניקוד על סעיף זה יינתן בהתאם לרמת החדשנות, המורכבות והיישום המוצלח של ההצעות שתבחרו לממש.
2. הציעו **לפחות 2 רעיונות לשיפור המודלים שיצרתם** – רעיון אחד של שיפור מתחום עיבוד הנתונים (חלק א' של הפרויקט), ורעיון אחד של שיפור בתחום יצירת המודל (חלק ב'). תוכלו להציע רעיונות ואלגוריתמים שלא נלמדו בקורס (ראו נספח מקורות נוספים).
3. **ממשו את הרעיונות שהצעתם**, האם שיפרתם את דיוק הוולידציה? אם לא, שערו מדוע.
4. אם שיפרתם את המודל הטוב ביותר שלכם ציינו את דיוק הוולידציה שהתקבל באמצעותו

# ***Test (15 נק')***

שלב זה מדמה **חיזוי על נתוני העולם האמיתי**, כלומר נתונים שאין ברשותכם בעת פיתוחהמודל. על חלק זה תקבלו ציון בהתאם לביצועי הקלסיפיקציה שלכם.

הורידו מהMoodle את הקובץ test.csv, העבירו את הנתונים את תהליך העיבוד (בדומה לנתונים ביצירת הdataset), ובצעו עליו חיזוים עם המודל הטוב ביותר שלכם (המשופר, גם אם השתמשתם בחומר שלא נלמד בקורס).

העלו לMoodle קובץ csv עם החיזויים של המודל שלכם. קראו לקובץ בשם הקבוצה שלכם (לדוג' group16.csv). הקובץ יכיל את הId של ההודעה ואת הסיווג בלבד.

דוגמה:

|  |  |
| --- | --- |
| sentiment | id |
| positive | 09c85963-40b5-478b-9fb9-6c252d23bd46 |
| … | … |

**שימו לב שאתם מעבירים את נתוני הtest את אותו תהליך של עיבוד מקדים כמו שהעברתם את הנתונים שעליהם בניתם את המודל, עם יכולת הוצאת פיצ'רים של כל שורה בנפרד.**

בהצלחה !

***נספח - מקורות נוספים:***

ככל שארגז הכלים של מדען הנתונים רחב ומגוון יותר, כך עולה היכולת שלו להתמודד עם בעיות מורכבות ומגוונות בתחומים שונים. אנו מעודדים אתכם להרחיב את הידע שלכם גם מעבר לחומר שנלמד בקורס, ולהיחשף למקורות נוספים שיכולים להעשיר את העבודה שלכם ולשפר את איכות הפתרונות.

להלן מספר מקורות מומלצים ללמידה עצמאית, העמקה וחיפוש פתרונות:

* **Documentation**:

תיעוד (דוקומנטציה) של ספריות שאתם משתמשים בהן, כגון:[scikit-learn](https://scikit-learn.org/stable/index.html), [keras](https://keras.io/), [tensorflow](https://www.tensorflow.org/resources/learn-ml) ועוד. פעמים רבות תמצאו שם פונקציות, פרמטרים וטכניקות שיכולות לחסוך לכם זמן ולשפר את המודל.

* **קהילות אונליין ושאלות-תשובות:**
* קבוצת פייסבוק: מומלץ לחפש פתרונות קיימים או לפרסם שאלות ולקבל עזרה מקהילת המומחים. דוגמה: [Machine & Deep learning Israel](https://www.facebook.com/groups/MDLI1/?locale=he_IL)
* Stack Overflow - שאלות על תקלות בקוד, פתרונות נפוצים וביצועים.
* Reddit - דיונים, המלצות על כלים ומאמרים. [MachineLearning](https://www.reddit.com/r/MachineLearning/) [DataScience](https://www.reddit.com/r/datascience/?rdt=40809)
* Kaggle Discussions - קהילת Data Science פעילה עם דוגמאות, קוד פתוח וניתוחים. [קישור](https://www.kaggle.com/discussions?sort=hotness)
* **מאמרים אקדמיים וחומר מתקדם**
* Google Scholar – חיפוש מאמרים אקדמיים לפי נושא, אלגוריתם או גישה מסוימת.