לימוד מכונה 364-1-1811

תרגיל מספר 2

הנחיות הגשה: דו"ח התרגיל השני, החיזויים הסופיים וקוד ה-Python שכתבתם יוגשו לתיבת ההגשה במודל עד לתאריך ה- 31/01/2024 בשעה 23:59.

מטרת התרגיל: בתרגיל נשתמש בנתונים שסקרנו והכנו בתרגיל הראשון לצורך אימון ובחינה של מערכות לומדות והשוואה ביניהן. התרגיל יעסוק ויציג מודלים של מערכות לומדות אליהם יש להתייחס כמפורט למטה.

כל ההמלצות שניתנו בתרגיל מס' 1 לגבי השימוש בתוכנה וביצוע העבודה והדיווח רלוונטיות.

דגשים לדו"ח: אורך הדו"ח **לא יעלה על 12 עמודים** (לא כולל עמודים נלווים כמו שער, תוכן עניינים, נספחים וסעיף בונוס), בגודל כתב 12, פונט Arial, רווח שורה וחצי. הדו"ח יוגש כקובץ word. חריגה ממספר עמודים זה תגרור הורדת נקודות. **יש לשמור על תמציתיות ולהתמקד בתובנות המרכזיות שלכם בכל סעיף.**

אין להכניס פלטים **מרכזיים** בנספחים, אלא בגוף הדו"ח.

הסעיפים בכתום - לא חובה. מי שיחליט לענות עליהם בשום מקרה לא יאבד על זה נקודות.

שאלות בנוגע לתרגיל, יש לפרסם **בפורום הייעודי (לחלק ב') במודל.**

דגשים והכוונה

- מחקר data science הוא אמפירי במהותו, כלומר אין פתרון אנליטי מוגדר מראש שמושג על ידי פתירת משוואות עם תשובה סופית ברורה. בידי החוקרים והחוקרות יש ארגז כלים ושיטות שבעזרתם הם מתמודדים עם הבעיות בהן הם נתקלים. ככל שארגז הכלים נרחב יותר, כך יוכלו להתמודד עם מגוון רחב יותר של בעיות בעולמות תוכן שונים. אנחנו מעודדים פתרונות יצירתיים לבעיות שתתקלו בהם במהלך העבודה ועליהם יינתנו עד 5 נקודות בונוס לחלק ב' (עד ציון 100 ובהתאם למידת ההשקעה בדוח והפגנת הידע). לכן חשוב לנו לחשוף אתכם למגוון כלים מוכרים שניתן להשתמש בהם:

1. קבוצת machine and deep learning Israel בפייסבוק שם ניתן לערוך חיפוש ולקבל מגוון פתרונות של אנשים או לחלופין לעלות פוסט ולבקש עזרה בפתרון הבעיה.
2. מאמרים מחקריים שאליהם ניתן לגשת בעזרת Google scholar.
3. פורומים רלוונטים כמו stackoverflow (פתרונות קוד). בלוגים כמו medium, towardsdatascience (הסברים על קונספטים של ML).
4. צ'אט GPT ומודלי שפה שונים: שימו לב! הכלים האלה יכולים להיות מועילים במצבים רבים, אבל באותה מידה יכולים גם להפיל אתכם. אל תתנו להם להחליף

את החשיבה העצמאית שלכם, וקחו תמיד בעירבון מוגבל הסברים וקוד שמגיעים מהם. הסברים שייראו "מסונתזים" או לא קשורים, קוד חסר הגיון, או סימנים שיעידו על כך ש**אתם** לא מבינים עד הסוף מה עשיתם – יובילו להורדת נקודות.

- ניתן בכל שלב בעבודה לשנות החלטות קודמות. לדוגמה, אם החלטתם בחלק א' להסיר משתנה מסוים וכעת אתם רוצים לבדוק האם הוספה שלו משפרת את ביצועי המודל/ים. כמובן, יש לציין זאת בדו"ח.
- מומלץ להסתכל בדוקומנטציה של החבילות שאתם עובדים איתם, כדי להבין מה האפשרויות הקיימות (החבילות המרכזיות מכילות את מרבית הפונקציות שתצטרכו בעבודה שלכם, תחפשו טוב ותוכלו לחסוך זמן בכתיבת קוד נוסף ומיותר).
- ניתן להשתמש בחבילות שלא נלמדו במעבדות (גם מימושים שונים של אותם המודלים).
- לעיתים שני מודלים דורשים "עיבוד מקדים" שונה. לדוגמה: משתנים קטגוריאליים בעצי החלטה. עבור מודל מסוג DecisionTreeClassifier של sklearn יש להמיר את המשתנים הקטגוריאליים למשתני דמה (למשל, קידוד של 1/0, ניתן להשתמש בחבילה שבקישור, או בפונקציית ()get_dummies של חבילת pandas בקישור).
- עבור כל מודל, תתבקשו לבצע כוונון של היפר-פרמטרים כדי לבחור בקונפיגורציה הטובה ביותר עבורו. תוכלו לבצע זאת בכל שיטה לבחירתכם (ניתן להיעזר בלינק הבא). ניתן לבחור שיטה שונה עבור כל מודל. יש להסביר איזו שיטה בחרתם ואת מאפייניה. למשל, אם בחרתם להשתמש ב-k-fold, יש לציין את ערכו של k מדוע בחרתם בערך זה.
- מהו המדד אותו בחרתם למקסם במהלך האימון ולמה? התייחסו לאיזון/חוסר איזון של סט הנתונים בהסבר הבחירה. כדי לבחור מדד מתאים ניתן להיעזר בקישור הבא. במידה ובחרתם למקסם מדד שאינו הדיוק (accuracy) במהלך כיוונון הפרמטרים ניתן להגדיר את המדד הרצוי ע"י הגדרת הארגומנט 'scoring' ב-grid/random search. העזרו בקישור הבא.
- כאשר אתם מתבקשים לבצע "כוונון פרמטרים" – עליכם להחליט: 1) לאילו היפר-פרמטרים בחרתם לבצע אופטימיזציה. 2) איזה ערכים ברצונכם לבדוק. 3) באיזו שיטה אתם דוגמים ערכים (למשל, grid search). יש להסביר את המוטיבציה לכל החלטה.
- בחירת מאפיינים – עבור כל מודל ניתן לשלב שיטות לבחירת מאפיינים שנלמדו בקורס (איכותניות ו/או כמותניות) במטרה לשפר את ביצועי המודלים.
- **הגשה** – יש להגיש סה"כ שלושה קבצים: 1) דו"ח PDF. 2) קובץ קוד. 3) קובץ חיזויים (xlsx).

הכנת הנתונים לאימון ובחינה (3 נק')

חלקו את קובץ Xy_train לסט אימון, וסט "בחינה" (ולידציה). סט הבחינה המדובר יישמר בצד, ועליו תוכלו לבדוק את ביצועי המודלים שתבנו לאחר האימון. כתבו מה השיקולים שלכם לקביעת גודל סט הבחינה, ואופן החלוקה שבחרתם (אקראי/לפי חוקיות מסוימת...).
שימו לב: החלוקה הזו נועדה לתת **לכם** פידבק ובטחון בביצועי המודלים על נתונים עתידיים. אין קשר לקובץ X_test בשלב זה, שישמש כמבחן האמיתי לביצועי המודלים.

Decision Trees (24 נק')

1. הכינו את הנתונים כך שיתאימו למודל עץ ההחלטה. תארו בקצרה את הפעולות שביצעתם ומשמעויותיהן (3 נק').

2. בנו עץ החלטה **מלא** באמצעות סט האימון. מהם אחוזי הדיוק (או מדד הביצוע שבחרתם) המתקבלים על סט האימון וסט הבחינה? מה ניתן להסיק מתוצאות האלה? האם עץ מלא תמיד יביא לתוצאה שכזו על סט האימון? (3 נק').

3. בצעו תהליך של "כוונון פרמטרים" (Hyperparameter-tuning) למציאת הקונפיגורציה המיטבית עבור מודל זה (8 נק').
   - עבור כל היפר-פרמטר: מה הייתה המוטיבציה בבחירה לכוונן אותו? מה המשמעות על העץ כתוצאה מהגדלת/הקטנת ערכו של כל אחד מהם?
   - הציגו (גרפים ו/או טבלה) את ערכי מדד ההצלחה שבחרתם, כפונקציה של ערכי היפר-פרמטרים מסויימים שנבחנו. מה ניתן להסיק מממצאים אלו?

4. אמנו עץ החלטה עם הקונפיגורציה הטובה ביותר שהתקבלה בסעיף 2 וענו על הסעיפים הבאים (10 נק'):
   - מהם אחוזי הדיוק (או מדד הביצוע שבחרתם) המתקבלים על סט האימון וסט הבחינה? מה ניתן להסיק מתוצאות אלה? איך ניתן להסביר את ההבדל מתוצאות העץ המלא (אם קיימים)?
   - הציגו גרף של העץ שהתקבל (אם גדול מידי – הגבילו את העומק עד שיהיה ניתן להצגה ברורה).
   - ע"י התבוננות במבנה העץ, אילו תובנות הוא מספק על הבעיה ועל החשיבות השונה של מאפייני הבעיה?
   - השתמשו בפונקציית חשיבות המשתנים של המודל (במודל DecisionTreeClassifier: _feature_importances). הסבירו את

משמעות הפלט של פונקציה זו. האם התוצאות מתיישבות עם המסקנות מהסעיף הקודם?

- **לא חובה:** בחרו רשומה לדוגמה מסט הבחינה, העבירו אותה באופן ידני (על פי המופיע בגרף העץ) דרך עץ ההחלטה ודווחו מה הסיווג שהתקבל. האם הסיווג תואם את המציאות?
- **לא חובה:** נקודות למחשבה. האם האימפורטנס שקיבלתם תואמים את השערותיכם לגבי משתנים חשובים / מיותרים מחלק א'? האם קורלציה ברורה של משתנים רציפים עם משתנה המטרה אכן העידה על חשיבות גבוהה בעץ ההחלטה?

Neural Networks (22 נק')

1. הכינו את הנתונים להכנסה לרשת נוירונים, והסבירו בקצרה את הפעולות שביצעתם ומשמעותן (4 נק').

2. הריצו (אמנו ובחנו) את הרשת בערכי ברירת המחדל. הסבירו את משמעות הקונפיגורציה שנלמדה ע"י המודל (מספר נוירונים בשכבת הכניסה, מספר שכבות חבויות, מספר נוירונים חבויים בכל שכבה ופונקציית אקטיבציה). מהו מדד הביצוע שבחרתם המתקבל על סט האימון וסט הבחינה? מה ניתן להסיק מתוצאות האלה? (5 נק').

3. בצעו תהליך של "כיוונון פרמטרים" למציאת הקונפיגורציה המיטבית עבור מודל זה. (בהרצאה לא למדנו על כל הפרמטרים שניתן לכוון במודל MLP. כאן ניתן להרחיב ולקרוא באינטרנט על פרמטרים נוספים אותם תרצו "לכוון". מומלץ להיעזר בדוקומנטציה של המודל באתר של חבילת התוכנה כדי להבין איזה פרמטרים ניתנים לכיוונון. הערה נוספת: אימון של רשת נוירונים סיבוכי יותר משל עץ החלטה. אם תכוונו יותר מדי היפר-פרמטרים עם יותר מדי ערכים אתם יכולים להיתקע עם חיפוש שיקח ימים שלמים. נסו להבין איפה ניתן לצמצם את מרחב החיפוש והימנעו מהרצה של מעבר ללילה אחד, לטובת הבריאות הנפשית שלכם). (7 נק').
   - עבור כל היפר-פרמטר: מה הייתה המוטיבציה בבחירה לכוונן אותו? מה המשמעות על הרשת כתוצאה מהגדלתו/הקטנתו?
   - הציגו (גרפים ו/או טבלה) את ערכי מדד ההצלחה שבחרתם, כפונקציה של ערכי היפר-פרמטרים שנבחנו. מה ניתן להסיק?

4. הריצו (אמנו) רשת באמצעות הקונפיגורציה הנבחרת מהסעיף הקודם (6 נק').
   - תארו את הקונפיגורציה הטובה ביותר שנבחרה בתהליך כיוונון ההיפר-פרמטרים והסבירו את משמעותה (מספר נוירונים בשכבת הכניסה, מספר שכבות חבויות, מספר נוירונים חבויים בכל שכבה, פונקציית אקטיבציה..).
   - מהם מדדי הביצוע המתקבלים על סט האימון וסט הבחינה? מה ניתן להסיק מתוצאות אלה?
   - **לא חובה:** הציגו מטריצת מבוכה עבור שתי המחלקות ותארו את התוצאות בקצרה.

Unsupervised Learning - Clustering (10 נק')

1. הריצו מודל K-Means (על סט האימון) עם ערכי ברירת המחדל ("המאשכל הבסיסי"), פרט להגדרת מספר האשכולות, אותו תגדירו כך שיתאים לבעיה עמה אתם מתמודדים ועבורה ידוע לכם מספר המחלקות.
2. דונו בתוצאות. מה טיב ההתאמה בין האשכולות למחלקות? כיצד שויכו תצפיות לאשכולות? לוו את ההסברים בגרפים מתאימים שיציגו את הנתונים (או חלק מהם) במרחב. ניתן להשתמש בשיטות להורדת ממד כגון PCA.
3. **לא חובה:** במנותק ממשתנה המטרה של סיפור המקרה, בו יש מספר ידוע של קבוצות, אמנו כעת מחדש שמונה מודלים של אשכול עם ערכי K משתנים (כלומר, עם מספר שונה של אשכולות בכל פעם). מהו ה-K בו תבחרו? הסבירו את אופן בחירתכם. האם הערך מתקשר לסיפור המקרה? אם לא, מה יכולה להיות הסיבה לכך?
4. **לא חובה:** עבור מס' האשכולות הנבחר מסעיף 3, בחנו שיטת אשכול נוספת (שנלמדה או לא נלמדה בקורס) והשוו בין סכמת האשכול שהתקבלה ממודל זה וסכמת האשכול שהתקבלה בסעיף 3. מה היתרונות/חסרונות של מודל זה לעומת K-Means?

אימון מודל נוסף לבחירתכם (10 נק')

בחרו מודל סיווג נוסף מכל סוג שתרצו (חפשו באופן עצמאי), אמנו וכוונו את הפרמטרים שלו. פרטו בקצרה על מנגנון הסיווג שלו, על הפרמטרים שבחנתם ובחרתם, והציגו את ביצועיו.

השוואה בין מודלים (6 נק')

1. הסבירו מדוע למשימת הסיווג תעדיפו DT או NN לעומת K-Means? (3 נק')
2. השוו את ביצועי שלושת המודלים (הMLP הנבחר, הDT הנבחר, והמודל הנוסף שאימנתם). מהן מסקנותיכם? מי מהם הייתם בוחרים? מדוע? (3 נק')

המודל הנבחר (7 נק')

1. הציגו את המודל שבחרתם לטובת ההגשה לתחרות, ופרטו מה ערכי ההיפר-פרמטרים שנבחרו עבורו (3 נק').
   - ניתן לבחור בכמה מודלים ולשלב בין החיזויים שהתקבלו מהם ליצירת החיזויים הסופיים. במידה ובחרתם לעשות זאת ה**סבירו בפירוט** מהם המודלים שבחרתם, למה בחרתם בהם, וכיצד שילבתם בין החיזויים שלהם.
   - במידת הצורך, אתם יכולים לבצע שיפורים נוספים בהכנת הנתונים ואימון המודל (למשל יצירת משתנים חדשים, בחירת משתנים, כיוונון נוסף של היפר-פרמטרים וכו').
2. בצעו ניתוח של תוצאות המודל באמצעות מטריצת מבוכה (confusion matrix) על סט הבחינה. האם התוצאה תקינה או שישנה הטיה? הסיקו מסקנות, אם ניתן, על החולשות / חוזקות של המודל (4 נק').

<u>חיזוי סופי (11 נק' + 2 נק' בונוס)</u>

בצעו חיזוי באמצעות המודל שבחרתם, על קובץ הנתונים "X_test".

אל תשכחו לבצע השלמת ערכים בהתאם לשיטות שהפעלתם על סט האימון. מחיקת רשומות לא אפשרית במקרה הזה (לא נוכל לחזות עבורן ערכים), ולכן אם אין לכם ברירה ניתן להשלים ערכים בשיטה נאיבית.

את החיזויים שיתקבלו העלו למודל כקובץ אקסל, על פי הפורמט שמופיע בקובץ הדוגמה "y_test_example.xlsx" הנמצא במודל (הערכים בקובץ שבמודל הם אקראיים ומשמשים לדוגמה בלבד). בדקו שהקובץ שאתם מגישים הוא בעל מבנה זהה ובעל אותו מספר רשומות הקיימות בקובץ X_test.xlsx המתאים לכם, ושמו לפי הפורמט: "dataset_Gi_ytest.xlsx" כאשר i הוא מספר הקבוצה שלכם ו-dataset הוא בסיס הנתונים עליו עבדתם (diabetes/CI).

לדוגמה: שם הקובץ עבור קבוצה 9 שעבדה על בסיס הנתונים car insurance יהיה:

carinsurance_G9_ytest.xlsx

שימו לב שאתם לא משנים את הסדר של הרשומות בקובץ X_test כדי למנוע בעיה בהשוואה לLABELS האמיתיים בסוף.

<u>חלוקת הנקודות:</u>

4 נק' - על הגשת קובץ החיזויים באופן תקני.

7 נק' - בהתאם למיקומכם ביחס לשאר הצוותים בקורס (עם אותו בסיס נתונים).

2 נק' בונוס - לקבוצה אשר תשיג את הביצועים הגבוהים ביותר (בכל אחד מבסיסי הנתונים).

<u>איכות הדו"ח והקוד (7 נק')</u> – נק' אלו יינתנו בהתאם להערכה כללית בנוגע לאיכות הדו"ח ובהתאם לרמת השימוש בPython.

<u>בונוס כללי (5 נק')</u>

בונוס של עד 5 נקודות לחלק ב' יינתן לקבוצות שיגלו ויציגו תובנות מעניינות ולא טריוויאליות לאורך הדרך, יעשו שימוש **<u>בצורה מקיפה ואיכותית</u>** בפונקציות/הרחבות של המודלים (לדוגמא- דרכים מתקדמות לשיפור המודלים) ו/או במודלים שלא נדרשים בעבודה, או שיציגו תוצאות ומסקנות באמצעים ויזואליים באופן יוצא דופן.

בהצלחה!