

מטלה 1 (מטלה אישית)

מבוא ללמידת מכונה

תאריך: 27/11/24

תאריך הגשה: עד 12/12/24

אתם מתבקשים לבנות 3 מסווגים מבוססי עצים שלמדנו.

התרגיל בנוי מ-3 חלקים:

- א. (60 נקודות) כתיבת קוד שיבצע את המטלה.
- ב. (35 נקודות) הרצת ניסויים על מנת לאפסם כל אחד מהמסווגים.
- ג. (5 נקודות) להכין עמוד **GitHub** לפרויקט ולהעלות את המטלה לשם.

פרטים של כל חלק:

חלק א:

1. יש להוריד את הקובץ המצורף **ex1.ipynb** למחשב שלכם.
2. יש לשנות שם הקובץ כך שבשם הקובץ יופיע שמכם הפרטי והמשפחה. למשל, אילו הייתי סטודנט שמגיש, שם הקובץ שלי היה **igor_nor_ex1.ipynb**
3. יש להעלות את הקובץ ל-**Google Colab** שלכם.
4. בקובץ נמצא תיאור של הבעיה ושלבים לפתרון.
5. חובה לעקוב אחרי ההוראות במדויק ולהוסיף קוד במקומות המסומנים בלבד. קוד שלא יתאים מבחינת התוכן או הפורמט המפורט יפגע מאוד בציון.
6. שימו לב כי בנוסף להוראות, הקובץ מכיל מספר לינקים שעשויים לעזור לפתח את המטלה.
7. שימו לב כי שלב 2 (המפורט בקובץ) הוא במשקל נמוך של 5 נקודות עבור מי שרוצה להשקיע מאמץ בהכרות של הצגת נתונים בממד גבוה.
8. שימו לב כי תוצאות של כל מסווג יש לבחון בעזרת 5 שיטות:
f-score, recall, precision, confusion_matrix, accuracy
9. בסיום ביצוע הקוד יש להוריד את הקובץ עליו עבדתם מ-**Google Colab** שלכם. הקובץ הזה הוא החלק הראשון במטלה אותו צריך להגיש.

חלק ב:

1. עליכם לגלות פרמטרים עבורם כל אחד מהמסווגים משיג את התוצאה הטובה ביותר.
2. לכל מסווג יש להריץ לפחות 5 ניסויים בהם הפרמטרים שונים.
3. הפרמטרים השונים אפשר לראות בהגדרה של כל מסווג.
4. בסה"כ תבצעו לפחות 15 ניסויים.
5. כל ניסוי חייב להיות מתועד בקובץ נפרד ששמו: **hyperparams.pdf** בתוספת שם פרטי ומשפחה. במקרה שלי זה יהיה: **igor_nor_hyperparams.pdf**
6. פורמט התיעוד: לכל ניסוי שורה נפרדת בה פירוט הפרמטרים שונים מברירות המחדל והתוצאה שהתקבלה. את גרסה הטובה ביותר שמצאתם יש להשאיר בקוד הסופי.
7. הקובץ שלכם שמתעד את הניסויים הוא הקובץ השני שצריך להגיש.

חלק ג:

- יש להעלות את המטלה ל-GitHub.
- יש ליצור חשבון באתר (אם עדיין אין) ובו ליצור עמוד למטלה.
- הסבר טוב אפשר למצוא כאן: <https://docs.github.com/en/repositories/creating-and-managing-repositories/quickstart-for-repositories>
- שימו לב כי בתרגיל אין צורך לנהל פרויקט אלא רק לדאוג להעלות את הקבצים.
- את הלינק לעמוד ב-GitHub יש להוסיף בשורה הראשונה בקובץ הקוד.

הערות נוספות:

- המלצה חמה לא להעתיק – בין הסיבות, שימוש בתוכנה לבדיקת מקוריות.
- במטלה תשתמשו בנתונים על זיהוי סרטן מנתונים.
- שימו לב כמטלה לא משתמשים בתמונות, אלא רק בנתונים מספריים.
- פרטים על הנתונים אפשר למצוא כאן: https://scikit-learn.org/1.5/modules/generated/sklearn.datasets.load_breast_cancer.html

בהצלחה!