המסווגים

קראו בעיון את כל ההוראות לפני ביצוע העבודה

<u>הוראות כלליות:</u>

1. אי עמידה בכל אחת מההוראות יגרור הורדת ציון או פסילת העבודה.
2. הגשת העבודה ב**שלשות או זוגות** (גודל קבוצות אחרות באישור מראש בלבד).
3. שפת תכנות – Python 3.7/3.8 , סביבת פיתוח – מומלץ להשתמש ב PyCharm (שימוש בטכנולוגיה/שפה שונה באישור מראש בלבד, פייתון היא השפה המומלצת לתחום).
4. יש להגיש את העבודה **לתיקיית ההגשה הרלוונטית באתר הקורס (Moodle).** אחריותכם האישית לבדוק לפני הגשה כי **<u>כל הקבצים נפתחים כראוי</u>**.
5. יש להגיש קובץ zip - שם הקובץ יהיה מורכב משלוש מספרי תעודות הזהות של המגישים באופן הבא: ID_ID_ID.zip

   הקובץ יכיל את הקבצים הבאים:

   - הפרויקט המלא: קבצי קוד, חשוב : **ללא** קבצי הנתונים. יש לוודא ששמות הסטודנטים ותעודות הזהות רשומים בהערה בתחילת כל קובץ של קוד.
   - קובץ readme.txt
     i. המכיל את שמות הסטודנטים ותעודות הזהות.
     ii. פירוט התקנות של חבילות שהשתמשתם בהם.
     iii. תיאור סדר התלויות של הקבצים השונים בפרויקט שלכם.
     iv. כל הערה נוספת הרלוונטית לאופי ההרצה הטכני של התוכנה.
   - קובץ PDF (יש להיעזר ב Jupyter NoteBook) יש לרשום גם פה שם ותז.
     i. **<u>תיאור</u>** של מבנה הפרויקט שיצרתם ותפקיד של כל קובץ, מחלקה, שיטות ושדות בפרויקט.
        1. יש להסביר היכן בא לידי ביטוי המימוש האישי שלכם והיכן נעשה שימוש בכלים מן המוכן.
     ii. ניתוח של הנתונים
        1. EDA - Exploratory Data Analysis (קישור עם דוגמה מומלצת בנספח)

2. יש לסכם את התצפית והאבחנה שלכם על הנתונים בעזרת Jupyter NoteBook

iii. הרצת ניסויים - הצגת טבלת תוצאות דיוק ההרצה של סיווג ואישכול הנתונים המצורפים (עבור כל ההרצות יש להציג מדד דיוק אימון, דיוק בדיקה ואת דיוק חוק הרוב. בנוסף מדדי הערכה (Precision, Recall, F-measure, Accuracy, Confusion Matrix)

1. יש לבנות שני מודלים לסיווג של naive bayes מימוש שלכם ומימוש של ספריה קיימת
2. יש לבנות שני מודלים לסיווג של עצי החלטה מימוש שלכם ומימוש של ספריה קיימת
3. יש לבנות מודל סיווג של KNN
4. יש לבנות מודל אישכול בעזרת K-MEANS (שימו לב לדוגמה בנספח)
   a. יש לשים לב, בניית מודל אישכול אינו כולל את עמודת הסיווג.
   b. לאחר שנקבל את האשכולות ניתן לספור את הסיווגים בכל אחד מהאשכולות וועל פי הרוב להחליט על הסיווג של האשכול.
5. עבור הדיסקרטיזציה יש לבחור שלוש שיטות של דיסקרטיזציה, עבור כל הרצה (עומק שווה, רוחב שווה ומבוססת אנטרופיה) מימוש שלכם ומימוש של ספריה קיימת

.

iv. דיון והסבר על התוצאות כולל מבחני השערה, , גרפים וויזואליזציות מתאימות בעזרת Jupyter NoteBook

v. מסקנות.

6. בנוסף, זוהי עבודה תכנותית ולפיכך יהיה לכך משקל בבדיקה. כלומר: מצופה התייחסות לתכנון מושכל והנדסת התוכנה, יש לדאוג לתיעוד והערות בקוד, הסבר נרחב על פונקציות ממשק, חלוקה הגיונית לממשקים ומחלקות, פונקציות קצרות ועניינים וכדומה.
7. **תאריך הגשה סופי: 9/7, הצגת פרויקטים ב 10/7.**

הוראות:

בפרויקט זה עליכם לבנות מערכת סיווג ואישכול של נתונים על פי הנלמד בכיתה.
חלק מהמודולים בהם יש לעשות שימוש מומשו במעבדות השונות לאורך הסמסטר.
קובץ הנתונים מכיל גם תכונות רציפות וגם תכונות קטגוריאליות. בנוסף, ייתכנו רשומות עם ערכים חסרים, בהם יש לטפל כחלק מתהליך ניקוי הנתונים.

תיאור הקבצים לרשותכם (שימו לב הפרויקט עשוי להיבדק עם קבצי נתונים ומבנה שונים מהנתון):

1. **Dataset general info** – מידע כללי בנוגע לבסיס הנתונים ממנו לקוחים נתוני התרגיל. קובץ זה הינו לשימושכם בלבד ולא ישמש כנתון שעל תכניתכם לקרוא במהלך הריצה.
2. **Structure** – קובץ המתאר את התכונות המרכיבות כל רשומה בבסיס הנתונים (כולל ערך המטרה אשר מופיע אחרון ברשימה). הקובץ ישמש את התוכנית שלכם ללימוד מבנה בסיס הנתונים בו עליה לטפל (כפי שיודגש בהמשך, התכנית שלכם תיבדק בעזרת dataset שונה במבנהו מזה שנתון לכם בתרגיל).

   יש להקפיד על המבנה המתואר בקובץ ועל אופן הופעת התכונות (Features).

   1. תכונת המטרה תקרא תמיד "class", ותהיה אחרונה בקובץ ובכל רשומה.
   2. יש להשתמש בקובץ זה על מנת לחלץ את הערכים הייחודיים השונים אותם כל תכונה יכולה לקבל. בנוסף, ניתן לדעת על פי הקובץ מי מהתכונות היא נומרית או קטגוריאלית.

   - נומרית

```
@ATTRIBUTE [AttTitle] NUMERIC
```

   - קטגוריאלית

```
@ATTRIBUTE [AttTitle] {some comma separated categories}
```

דוגמא למבנה הקובץ שניתן לכם

```
@ATTRIBUTE age NUMERIC
@ATTRIBUTE job
{admin.,unknown,unemployed,management,housemaid,entrepreneur,student,blue-collar,self-
employed,retired,technician,services}
@ATTRIBUTE marital {married,divorced,single,widowed}
@ATTRIBUTE education {unknown,secondary,primary,tertiary}
@ATTRIBUTE default {yes,no}
@ATTRIBUTE balance NUMERIC
@ATTRIBUTE housing {yes,no}
@ATTRIBUTE loan {yes,no}
```

@ATTRIBUTE contact {unknown,telephone,cellular}
@ATTRIBUTE day NUMERIC
@ATTRIBUTE month {jan,feb,mar,apr,may,jun,jul,aug,sep,oct,nov,dec}
@ATTRIBUTE duration NUMERIC
@ATTRIBUTE campaign NUMERIC
@ATTRIBUTE previous NUMERIC
@ATTRIBUTE poutcome {unknown,other,failure,success}
@ATTRIBUTE class {yes,no}

3. **train** – קובץ המכיל רשומות שישמשו לבניית המסווג. לשם פשטות, הקובץ מסוג CSV. כל רשומה מופיעה בשורה נפרדת. (ניתן להניח שהקובץ יגיע עם כותרות בעמודות)

P	O	N	M	L	K	J	I	H	G	F	E	D	C	B	A	
class	poutcome	previous	campaign	duration	month	day	contact	loan	housing	balance	default	education	marital	job	age	1
no	unknown	0	1	261	may	5	unknown	no	yes	2143	no	tertiary	married	manageme	58	2
no	unknown	0	1	151	may	5	unknown	no	yes	29	no	secondary	single	technician	44	3
no	unknown	0	1	76	may	5	unknown	yes	yes	2	no	secondary	married	entreprene	33	4
no	unknown	0	1	92	may	5	unknown	no	yes	1506	no	unknown	married	blue-collar	47	5
no	unknown	0	1	198	may	5	unknown	no	no	1	no	unknown	single	unknown	33	6
no	unknown	0	1	139	may	5	unknown	no	yes	231	no	tertiary	married	manageme	35	7
no	unknown	0	1	217	may	5	unknown	yes	yes	447	no	tertiary	single	manageme	28	8
no	unknown	0	1	380	may	5	unknown	no	yes	2	yes	tertiary	divorced	entreprene	42	9
no	unknown	0	1	50	may	5	unknown	no	yes	121	no	primary	married	retired	58	10
no	unknown	0	1	55	may	5	unknown	no	yes	593	no	secondary	single	technician	43	11
no	unknown	0	1	222	may	5	unknown	no	yes	270	no	secondary	divorced	admin.	41	12
no	unknown	0	1	137	may	5	unknown	no	yes	390	no	secondary	single	admin.	29	13
no	unknown	0	1	517	may	5	unknown	no	yes	6	no	secondary	married	technician	53	14
no	unknown	0	1	71	may	5	unknown	no	yes	71	no	unknown	married	technician	58	15
no	unknown	0	1	174	may	5	unknown	no	yes	162	no	secondary	married	services	57	16

4. **test** – קובץ המכיל רשומות שאותן תצטרכו לסווג. לשם פשטות, הקובץ מסוג CSV. כל רשומה מופיעה בשורה נפרדת. שימו לב, בקובץ זה מופיע הסיווג האמיתי של כל רשומה, אך **אין לכם כל צורך להשתמש בו לצורך הסיווג, אלא ביתר התכונות בלבד**. השימוש בסיווג המקורי יהיה עבור מדידת הדיוק. (ניתן להניח שהקובץ יגיע עם כותרות בעמודות)

<u>תיאור המשימות שעליכם לממש במסגרת התרגיל</u>:

1. ממשק משתמש פשוט שיוצג עם הרצת התוכנית. הממשק צריך להיות נאמן לתיאור ולפונקציונאליות המתוארת. הממשק יכול להיות

1. GUI בונוס של 5 נק לציון העבודה הסופי (לא ניתן לקבל בפרויקט יותר מ 100)

2. תפריט טקסטואלי

3. ממשק CLI - שורת פקודה עם פרמטרים, ללא מגע יד אדם בכל התהליך, בונוס של 6 נק לציון העבודה הסופי (לא ניתן לקבל בפרויקט יותר מ 100)

הממשק יכיל את האופציות הבאות:

1.2. הזנת ה-path לתיקייה בה נמצאים נתוני התרגיל (למעט המחויבות על מבנה הקבצים - הפרויקט ייבדק עם אוסף נתונים שונה)

1.3. בחירת נתוני ניסוי ושלבי עיבוד מקדים

1. בכל ניסוי יתבצעו שלבי עיבוד ראשוניים, רובם מבוססים על קלט מהמשתמש:
   a. מחיקת שורות עם עמודת סיווג ללא ערך (יתבצע תמיד).
   b. יש להגדיר אופי השלמת ערכים חסרים - האם ההשלמה תעשה ביחס לערך סיווג או ביחס לכלל הנתונים. עבור ערכים רציפים יש להשלים ממוצע, ועבור ערכים בדידים יש להשלים שכיח.
   c. יש להגדיר האם בניסוי זה יש צורך בנרמול, אם כן יש להשתמש בשיטת נרמול לבחירתכם מתוך SKLEARN.
   d. יש לבחור האם תתקיים דיסקרטיזציה עבור נתונים רציפים בניסוי זה.
   a. בחירת סוג דיסקרטיזציה (יש לבחור בין ללא דיסקרטיזציה, עומק שווה, רוחב שווה או מבוססת

אנטרופיה) - דיסקרטיזציה זו תשמש עבור כל העמודות עם הנתונים הרציפים.

b. בחירה עבור כמות ה- Bins (אינטרוולים) שאליהם יחולקו הערכים הרציפים כחלק מתהליך דיסקרטיזציה. (נתון זה יקרא בתנאי ואכן יש דיסקרטיזציה)

e. בחירת אלגוריתם המודל

2. ביצוע תהליך העיבוד המקדים, יש לשמור את התוצאה בקובץ (הניקיון צריך להתבצע על שני קבצי הנתונים האימון והבדיקה):

[*FileName*]_clean.csv

1.4. בניית מודל ושמירתו כקובץ (יש להשתמש ב pickle או joblib)

1.5. הרצת מודל

1. הרצת המודל על קובץ האימון
2. הרצת המודל על קובץ הבדיקה

1.6. בדיקת דיוק - יש לבנות מטריצת טעות מלאה

1. השוואת נתוני הסיווג המקוריים לעומת הסיווג שהתבצע בפועל על קובץ האימון
2. השוואת נתוני הסיווג המקוריים לעומת הסיווג שהתבצע בפועל על קובץ הבדיקה
3. הנתונים יישמרו בקובץ במבנה לבחירתכם לצורך עיבוד נוסף וסיכום :
   a. נתוני העיבוד המקדים שנבחרו
   b. נתוני "חוק הרוב" עבור הניסוי
   c. תוצאות האימון והבדיקה

<u>הערות</u>

- לצורך חלון ה-GUI (בונוס). מומלץ להשתמש בממשק פשוט ביותר של Tkinter .
- ממשק CLI יאפשר עבודה ללא התערבות המשתמש לאחר ההפעלה - כל גרסה אחרת לא תתקבל כממשק מסוג זה.

- **התכנית שתצרו תיבדק בעזרת dataset שונה מזה שנתון לכם לטובת התרגיל.** עליכם לדאוג שהקוד שלכם יידע לעבד קבצי נתונים שונים (כמות תכונות שונה וכמות ערכים שונה לכל תכונה). על התכנית להתאים את עצמה למבנה הנתון בעזרת קובץ Structure נתון.
  - שימו לב – על התכנית לדעת להתמודד עם תכונות רציפות ונומינאליות (תכונת המטרה, אשר תהיה מסוג **קטיגורי/נומינאלי** ותכיל ערכים אלפאנומריים).
  - אין להניח תקינות כלשהי בפעולות או בסדר הפעולות - יש לבדוק ולהציג הודעות שגיאה עבור כל מקרה חריג.
- יש ורצוי להשתמש בספריות מוכנות של Python המממשות את האלגוריתמים בהם נעשה שימוש, רצוי להתחיל עם SKLEARN NUMPAY PANDAS
- **אין להעתיק** - מותר ורצוי להיעזר בגוגל או בחברי הכיתה **אבל**
  - אין להשתמש בקוד או חלקי קוד של קבוצות אחרות.
  - אין להשתמש בקוד קיים של האלגוריתם שעלול להימצא באינטרנט (ניתן להשתמש ב API מתוך IMPORT אבל לא בקטעי קוד) .
  - אין להשתמש ברפרנסים של עבודות משנים קודמות.

שימוש במקורות חיצוניים כאלה ואחרים ייחשבו כהעתקה ויובילו לוועדת משמעת. **<u>הקפידו על כך</u>**!

- על התכנית לדעת להתמודד עם שגיאות כמו למשל קובץ נתונים ריק, קבצים חסרים בנתיב התיקייה המוזן, או נתון לא תקין בתהליך העיבוד המקדים. במקרה של נתון לא תקין, יש להציג הודעת שגיאה מתאימה (המעידה על סוג השגיאה) ולא לאפשר הפעלה של בניית המודל.
- שאלות בנוגע לתרגיל יש לשאול בפורום השאלות הרלוונטי המופיע ב-moodle (ולא במייל - שאלות במייל לא יענו).

בהצלחה!

לינקים וחומרי עזר

איך לשמור את המודל שיצרנו :

https://machinelearningmastery.com/save-load-machine-learning-models-python-scikit-learn/

דוגמא להשוואת מסווגים שונים

https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html#sphx-glr-auto-examples-classification-plot-classifier-comparison-py

דוגמא לאשכול

https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_iris.html#sphx-glr-auto-examples-cluster-plot-cluster-iris-py

הסבר על EDA

https://www.kaggle.com/agrawaladitya/step-by-step-data-preprocessing-eda