תרגיל בית 4 - שפת Python

הוראות הגשה

א. אי עמידה בכל אחת מההוראות יגרור הורדת ציון או פסילת העבודה.

ב. הגשת העבודה ב**זוגות** בלבד. רק אחד מבני הזוג יגיש את המטלה!

ג. שפת תכנות – Python, סביבת פיתוח – JetBrains PyCharm.

ד. יש להגיש את העבודה **לתיקיית ההגשה הרלוונטית באתר הקורס (Moodle).**
אחריותכם האישית לבדוק לפני הגשה כי כל הקבצים נפתחים כראוי.

ה. יש להגיש קובץ zip - שם הקובץ יהיה מורכב משני מספרי תעודות הזהות של המגישים באופן הבא: ID1_ID2.zip
הקובץ יכיל את הקבצים הבאים:

- הפרויקט המלא: קבצי קוד + GUI, חשוב : ללא קבצי הנתונים.
- קובץ word המכיל תיאור של מבנה הפרויקט שיצרתם ותפקיד של כל מחלקה ושיטה בפרויקט. יש לציין בפינה השמאלית בכל עמוד את ת"ז ושמות הסטודנטים.

ו. אין לשתף קטעי קוד ואין להעתיק פתרונות!

ז. בנוסף, זוהי עבודה תכנותית ולפיכך יהיה משקל לכך בבדיקה. כלומר: יש לדאוג לקוד מסודר, הערות בקוד, לשמות משתנים בעלי משמעות וכדומה. יש לחלק את הקוד לפונקציות (במידת האפשר ולפי הצורך).

ח. אין להשתמש בספריות שאינן מובנות ב Python, Anaconda for python או הספריה plotly. יש להשתמש בספריות שלמדנו במעבדה. **לא תתבצע התקנה של סיפריות חיצוניות אחרות!**

ט. את חלון ה GUI אין צורך לבנות מאפס, אפשר להשתמש בקובץ calculator.py שהוצג במעבדה (introduction python) ולשנות את ה class הנתון (להוסיף/ לשנות שדות וכפתורים על פי הנדרש).

י. שאלות בנוגע לתרגיל יש לשאול **אך ורק** בפורום השאלות הרלוונטי המופיע ב-moodle (ולא במייל - שאלות במייל לא יענו).

I. הגדרות התרגיל

בתרגיל זה עליכם לממש את אלגוריתם Naïve Bayes תוך שימוש ב- m-estimator (m=3).
לצורך המימוש, ניתן להשתמש בפונקציות מכל הספריות שלמדנו במעבדה.
קובץ הנתונים מכיל גם תכונות רציפות וגם תכונות קטגוריאליות. בנוסף, ייתכנו רשומות עם ערכים חסרים, בהם יש לטפל כחלק מתהליך ניקוי הנתונים.

תיאור הקבצים שלרשותכם:

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 התכנית 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).
   - תכונת המטרה תקרא תמיד "class", ותהיה אחרונה בקובץ ובכל רשומה.
   - 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.
3. **train** – קובץ המכיל רשומות שישמשו לבניית המסווג. לשם פשטות, הקובץ מסוג CSV. כל רשומה מופיעה בשורה נפרדת.
4. **test** – קובץ המכיל רשומות שאותן תצטרכו לסווג. לשם פשטות, הקובץ מסוג CSV. כל רשומה מופיעה בשורה נפרדת. שימו לב שבקובץ זה מופיע הסיווג האמיתי של כל רשומה, אך **אין לכם כל צורך להשתמש בו, אלא ביתר התכונות בלבד**.

II. משימות התרגיל

- ממשק משתמש פשוט שיוצג עם הרצת התכנית. הממשק יכיל את התהליכים הבאים:

1. הזנת ה-path לתיקייה בה נמצאים נתוני התרגיל (יש לממש אפשרות זו בעזרת browser).

על הממשק להכיל תיבת טקסט אחת בלבד, אליה יוכנס הנתיב לקבצים.

בתיקייה זו יופיעו כל הקבצים הדרושים במהלך ריצת התכנית (קבצים 2-4 המתוארים בסעיף הקודם). אין להניח שכל הקבצים נמצאים בתקייה ויש לטפל במקרים שונים של הימצאות הקבצים. בנוסף, קובץ הפלט שעליכם לייצר ייוצר באותה תיקייה. **הטקסט אשר יופיע על הכפתור יהיה "Browse".**

2. תיבת טקסט בה ניתן להזין את כמות ה- Bins (אינטרוולים) שאליהם יחולקו הערכים הרציפים כחלק מתהליך הדיסקרטיזציה. אין להניח כי מספר ה-Bins יהיה תקין ויש לטפל במקרים האפשריים השונים. **שם תיבת הטקסט יהיה "Discretization Bins".**
3. לחצן לטעינת ה-train ובניית המודל. **הטקסט אשר יופיע על הכפתור יהיה "Build".**
4. לחצן לטעינת ה-test וסיווג הרשומות בו. **הטקסט אשר יופיע על הכפתור יהיה "Classify".**

- טעינת ה-train ובניית המודל (תהליך 3):

1. עם לחיצה על הלחצן המתאים התכנית תקרא את הקובץ Structure.txt אשר מופיע בתיקייה שצוינה ע"י המשתמש. במהלך קריאת הקובץ, יש להסיק את מבנה המודל שאותו עתיד המסווג להגדיר.
2. לאחר טעינת מבנה המודל, תטען התכנית את הקובץ train.csv ותפריד עבור כל instance בין התכונות השונות בו. התכנית תשלח מבנה (לבחירתכם) אשר יכיל את כל תכולת הקובץ למחלקה המייצגת את המסווג. בשלב זה תתבצע בניית המסווג (הגדרת כל הערכים הנחוצים לשם סיווג רשומות חדשות בעתיד).

2.2.1 לאחר קריאת הקובץ, יתבצע ניקוי הנתונים. כחלק מתהליך זה:

א. יש **<u>להשלים ערכים חסרים</u>** :

- עבור ערכים נומריים: ערך הממוצע של כל הרשומות השייכות לאותו class.
- עבור ערכים קטגוריאליים: הערך השכיח ביותר (Mode).
- ניתן להניח כי אין ערכים חסרים בתכונת ה-class.

ב. יש **<u>לבצע דיסקרטיזציה</u>** לכל משתנה נומרי (רציף), לשם פשטות, יש להשתמש ב- Equal-width Partitioning. ניתן להחליף את הערכים לתגית (label) מספרית פשוטה ("1", "2", וכך הלאה).

3. בשלב זה יופיע dialog אשר יכיל את הודעה ***"Building classifier using train-set is done!"*** המתריאה על סיום בניית המסווג ויאפשר למשתמש ללחוץ על "OK" להמשך.

- לחיצה על הלחצן לסיווג (תהליך 4) תוביל ראשית לקריאת הקובץ test.csv שמופיע בתיקייה שציין המשתמש, כפי שתואר עבור קובץ הלימוד קודם לכן.

1. הרשומות שנטענו מתוך הקובץ יישלחו (במבנה לבחירתכם) למסווג. המסווג יעבור על כל רשומה ויסווג אותה בעזרת אלגוריתם Naïve Bayes, **תוך שימוש ב- m-estimator** (m=3). עבור כל רשומה, יש להדפיס לקובץ ששמו "output.txt" את התוצאות במבנה הבא (הקובץ ימוקם בתיקייה שהגדיר המשתמש לצד קבצי הקלט הנתונים):
   א. מספר הרשומה (מספור הרשומות בקובץ ה-test מתחיל מ-1) .
   ב. *רווח*.
   ג. ערך תכונת המטרה, כפי שסווגה ע"י המסווג שבניתם.

בין רשומה מסווגת אחת לאחרת יש לעבור שורה. דוגמה לתכולה תקינה של הקובץ:

1 yes

2 no

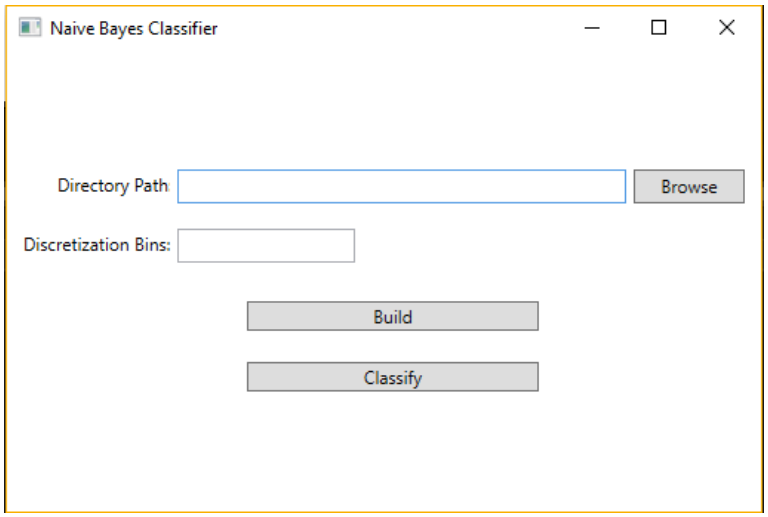
3 no

…

2. יש להציג dialog נוסף שיעדכן שהסיווג הסתיים. לחיצה על "OK" תסיים את ריצת התכנית ותסגור את החלון הראשי.

הערות חשובות:

- הטקסט על הכפתורים לא ניתן לשינוי, וחשוב שיהיה זהה למוגדר לעיל.
- ניתן להניח שהפעולות יבוצעו בסדר הנכון – כלומר, בניית מודל ורק לאחר מכן סיווג הרשומות.
- כותרת **כל** החלונות (כולל הדיאלוגים שפורטו לעיל) צריכה להיות:
  Naïve Bayes Classifier
- מצורף תצלום חלון ה-GUI הנדרש לתרגיל זה. מומלץ להשתמש בממשק פשוט ביותר של tkinter (from tkinter import *).



- **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 **קטגורי/נומינאלי** ותכיל ערכים אלפאנומריים).
  - ההנחה היחידה בעבודה היא כי סדר הפעולות הנדרשות יתבצע בסדר הנכון, כלומר קודם בניית המודל ("Build") ורק אז סווג ה-test-set ("Classify").
- אין להשתמש בקוד קיים של האלגוריתם שעלול להימצא באינטרנט. אין להשתמש בסיפריות מוכנות של Python המממשות את האלגוריתם. שימוש במקורות חיצוניים כאלה ואחרים ייחשבו כהעתקה. **הקפידו על כך**!
- על התכנית לדעת להתמודד עם שגיאות כמו למשל קובץ נתונים ריק, קבצים חסרים בנתיב התיקייה המוזן או מספר לא תקין את bins עבור תהליך הדיסקרטיזציה. במקרה של נתון לא תקין, יש להציג הודעת שגיאה מתאימה (המעידה על סוג השגיאה) ולא לאפשר לחיצה על כפתור ה-"Build". כלומר, יש לוודא תקינות קלט מיד לאחר הזנת הנתונים בכל שדה בטופס. רק אם כל ערכי השדות מלאים ותקינים, אז כפתור ה- BUILD יהיה זמין ללחיצה.

בהצלחה!