ת.ז. _____

מבחן מועד א' במבוא לניתוח נתונים בפייתון (094202) סמסטר אביב 2022

מרצה: עפרה עמיר, מתרגלים: רפי שללה, זהר גלעד, אלכס טואיסוב

לרשותכם שעתיים וחצי לפתור את הבחינה.

1. הבחינה היא עם חומר סגור. אסור להשתמש בכל חומר עזר.
2. הבחינה כוללת **10** עמודים ובהם **11** שאלות. יש לענות על כל השאלות (השאלה האחרונה היא שאלת בונוס). הניקוד של כל שאלה מופיע לצידה.
3. את התשובות יש לכתוב רק על גבי טופס הבחינה.
4. על טפסי שאלות ותשובות הבחינה יש לכתוב רק מספר תעודת זהות)ולא שם(. חובה לכתוב מספר ת.ז. על כל דפי הטופס והמחברת.
5. אין להפריד את דפי טופס הבחינה.
6. חובה להחזיר בסיום הבחינה את טופס שאלות הבחינה עם כל המחברות בהן השתמשתם. שימו לב: אי החזרה של טופס שאלות הבחינה, או המחברות תגרור כישלון מידי בבחינה.

בהצלחה!

ת.ז. _____

שאלה 1 (15 נק')

על כל אחד מההיגדים הבאים ציינו אם הוא נכון או לא נכון והסבירו. תשובה ללא הסבר לא תקבל ניקוד.

1. אם הprecision- של מסווג א' גבוה מהprecision- של מסווג ב', ה-recall של מסווג א' בהכרח נמוך מהrecall- של מסווג ב.' נכון/לא נכון. הסבר:

_____

_____

2. אם הממוצע וסטיית התקן של מדגם א' שווים לממוצע וסטיית התקן של מדגם ב', החציונים של שני המדגמים שווים בהכרח. נכון/לא נכון. הסבר:

_____

_____

3. אלגוריתם k-means לא תמיד ימצא את החלוקה האופטימלית לקלאסטרים על פי מדד ה-SSE. נכון/לא נכון. הסבר:

_____

_____

4. אם מנרמלים על ידי השיטה של max/min, הממוצע של המשתנה לאחר הנרמול יהיה 0. נכון/לא נכון. הסבר:

_____

_____

5. מוגדר כי y=0.5x. קורלציית Pearson בין המשתנים x ל-y הינה 1. נכון/לא נכון. הסבר:

_____

_____

שאלה 2 (5 נק')

בביתה בקורס מבוא לניתוח נתונים יושבים 100 סטודנטים. במסגרת הקורס, הוחלט לבצע ניסוי אשר יבדוק עבור מטבע מסוים האם המטבע הוגן. הניחו שכל אחד מהסטודנטים ביצע 50 הטלות של אותו המטבע (אותו המטבע עבר בין כל הסטודנטים). כל הסטודנטים בחנו את ההשערה שהמטבע הוגן, והשתמשו בp-value=0.05 בתור הסף להחלטה על דחיית השערת האפס. ניתן להניח שכולם ביצעו את ההטלות כמו שצריך ולא טעו בבחינת ההשערות. ידוע כי המטבע הוגן (כמובן שמידע זה לא ידוע לסטודנטים). כמה סטודנטים מתוך ה-100 אתם צופים שיסיקו שהמטבע אינו הוגן? הסבירו בקצרה (2-3 משפטים).

ת.ז. _____

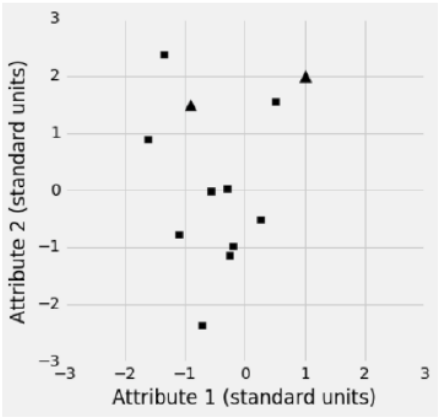
_____

_____

_____

שאלה 3 (5 נק')

ברשותכם מאגר נתונים שבו כל תצפית שייכת לאחת משתי מחלקות - משולשים או ריבועים. בנוסף, לכל תצפית שני פיצ'רים בהם משתמשים לסיווג (attribute 1 ו attribute 2). הגרף הבא מייצג את כל התצפיות בסט האימון שלכם אשר מכיל 12 תצפיות, כאשר המשתנים attribute 1 ו attribute 2 עברו נרמול לפי z-score.



בהתבסס על כל הנתונים והגרף, הניחו כי יש בסט הבדיקה תצפית אשר הערך של attribute 1 שלה קטן מהערך הממוצע של attribute 1 בסט האימון, והערך של attribute 2 של תצפית זו גם כן קטן מהערך הממוצע של attribute 2 בסט האימון. נניח שאתם מבצעים סיווג באמצעות k-nearest neighbor עם k=1. לאיזו מחלקה (משולשים או ריבועים) תסווג התצפית החדשה? הסבירו בקצרה (1-2 משפטים).

_____

_____

_____

שאלה 4 (5 נק')

יוסף גר בדירה עם שותף וחושד שהוא מוריד את הזבל יותר מהשותף שלו. על מנת לבדוק את החשד שלו, יוסף מנהל רישום של כל הפעמים שהוא הוריד את הזבל וכל הפעמים שהוא ראה את השותף שלו מוריד את הזבל. לאחר חודש הוא גילה שהוא הוריד את הזבל פי 2 פעמים מהשותף שלו. השותף של יוסף טוען שיש בעיה במדגם שהוא אסף. האם אתם מסכימים עם השותף של יוסף? הסבירו בקצרה (השתמשו במושגים שנלמדו בקורס).

ת.ז. _____

_____

_____

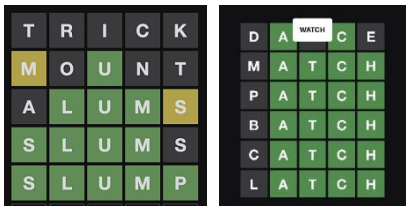
_____

_____

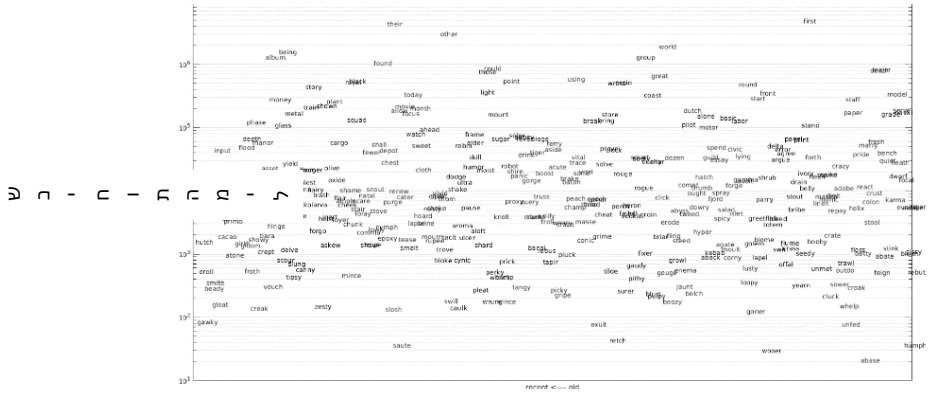
ת.ז. _____

שאלה 5 (20 נק')

המשחק Wordle הוא משחק בו צריך לנחש מילה תוך 6 ניסיונות. בכל ניחוש, אותיות אשר מופיעות במקום הנכון מסומנות בירוק, ואותיות אשר קיימות במילה אך מופיעות במיקום אחר מסומנות בצהוב. התמונה מימין מראה משחק לדוגמה בו המילה הנכונה (slump) נמצאה תוך 5 ניחושים, והתמונה משמאל מראה משחק בו המשתתף לא הצליח לנחש את המילה הנכונה (watch) תוך 6 ניחושים ולכן הפסיד.



יובל משחק במשחק כל יום, וחשד שהמילים במשחק נהיו קשות יותר לניחוש בחודש האחרון. על מנת לבחון את החשד שלו, יובל צייר את הגרף הבא, אשר מראה כל מילה מהמשחק כנקודה בגרף לפי הזמן בו פורסמה (ציר X, מילים חדשות יותר בצד שמאל, ישנות יותר בצד ימין) והשכיחות של המילה בשפה האנגלית (ציר y, מספרים גבוהים יותר מצביעים על שכיחות גבוהה יותר של המילה בשפה האנגלית).



תאריך הופעת המילה)תאריכים ישנים מימין ,חדשים משמאל(

(1) (3 נק') כיצד יובל יכול היה לקבוע את שכיחות המילים בשפה האנגלית? הציעו איזשהו מאגר נתונים בו הייתם משתמשים למטרה זו.

_____

_____

_____

(2) (3 נק') האם הגרף שצייר יובל תומך בחשד שלו שהמילים נהיות קשות יותר לניחוש (מילים פחות שכיחות בשפה האנגלית), או לא? הסבירו בקצרה.

ת.ז. _____

_____

_____

_____

(3) ברצונכם לעזור ליובל לבחון את ההשערה שלו באמצעות מבחן סטטיסטי של בדיקת השערות. יובל משתף איתכם את מאגר המילים שפורסמו ב- Wordleבעבר (עליו הגרף שלו התבסס).

a. (3 נק') הציעו מדגם/ים על בסיס מאגר הנתונים שבידיכם בו תוכלו להשתמש לצורך בדיקת ההשערה, וציינו מה יהיה הפרמטר שאותו תנסו לבחון. שימו לב, אין צורך לתאר את כל תהליך בדיקת ההשערות, רק לתאר את הנתונים בהם תשתמשו ואת הפרמטר שתבחנו.

_____

_____

_____

_____

b. (4 נק') ציינו את השערת האפס וההשערה האלטרנטיבית:

_____

_____

_____

_____

(4) ברצונכם לבדוק האם יובל באמת התקשה יותר בניחוש המילה בחודש האחרון מאשר בתקופה שלפני כן. לשם כך, קיבלתם מיובל גם את מספר הניחושים שלקח ליובל לנחש את המילה בכל פעם (מספר בין 1 ל6 או "לא הצליח" אם יובל לא מצא את המילה בששת הניחושים המותרים).

a. (3 נק') על איזה נתון סטטיסטי סיכומי מהמשחקים של יובל כדי לנסות להבין אם הוא באמת התקשה יותר לפתור את המשחק בחודש האחרון לעומת התקופה הקודמת. באיזה נתון תשתמשו?

_____

_____

b. (4 נק') מה הבעיה הקיימת בנתונים כתוצאה מכך שישנם מקרים בהם יובל לא הצליח לנחש את המילה תוך 6 ניסיונות? הסבירו בקצרה תוך שימוש במושג/ים מהקורס והסבירו כיצד תתמודדו עם הבעיה (כלומר, מה תעשו עם מקרים אלה)

_____

_____

ת.ז. _____

_____

_____

שאלה 6 (15 נק')

המרכז לקידום ההוראה בטכניון מעוניינים לחקור את הלמידה שנעשתה בקורסים במהלך הקורונה. לצורך כך, יש להם נתוני צפייה בהקלטות ממערכת הפנופטו ממנה ניתן להוריד דו"ח צפייה. הדו"ח מראה עבור כל אירוע צפייה בהרצאה את הנתונים המפורטים מטה (הטבלה מראה דוגמה לנתונים). בנוסף, בידי הטכניון הציונים בקורס. אנשי המרכז לקידום ההוראה מנסים לנבא כישלון של סטודנטים בקורסים לפי נתוני הצפייה.

Timestamp	CourseId	LectureId	StudentId	Number of minutes watched	Percent lecture watched
10/28/2021 10:35:34 AM	acfc004	4980c1a5	405edd0f	62	100
12/13/2021 6:43:41 PM	acfc004	4980c1a5	222f8299	31	50

מידע לגבי העמודות:

- Timestamp - זמן תחילת סשן הצפייה
- CourseId - מזהה ייחודי לכל קורס
- LectureId - מזהה ייחודי לכל הקלטה של הרצאה
- StudentId - מזהה ייחודי לכל סטודנט בקורס
- Number of minutes watched - מספר הדקות מתוך ההרצאה שנצפו במהלך הסשן
- Percent lecture watched - איזה אחוז מההרצאה נצפה

(1) (9 נק') הציעו 3 מאפיינים (פיצ'רים) שונים שתוכלו לייצר עבור הפעילות של סטודנט בקורס מסוים על פי הנתונים על מנת לחזות את הצלחת הסטודנט. בטבלה שתייצרו כל שורה תסכם את הפעילות של סטודנט בקורס מסוים. כלומר, הטבלה תכיל את העמודות CourseId, StudentId ועמודה נוספת לכל מאפיין שתייצרו. לכל מאפיין שאתם מציעים, הסבירו מדוע לדעתכם הוא עשוי להיות רלוונטי, והסבירו בבירור כיצד המאפיין יחושב מהנתונים הקיימים. יש להתבסס על המאפיינים הקיימים בטבלת הצפייה בלבד.

CourseId	StudentId	Your feature 1	Your feature 2	Your feature 3
acfc004	405edd0f			
acfc004	222f8299			

מאפיין 1:

_____

_____

ת.ז. _____

מאפיין 2:

_____

_____

מאפיין 3:

_____

_____

(2) (3 נק') לאחר קביעת המאפיינים שישמשו לחיזוי, משתמשים במרכז לקידום ההוראה ב-kNN לצורך סיווג. ידוע כי אחוז הנכשלים בקורסים נמוך באופן כללי. האם תייעצו למרכז לקידום ההוראה להשתמש בערך נמוך יחסית או גבוה יחסית של k? הסבירו בקצרה (2-3 משפטים).

_____

_____

_____

(3) (3 נק') באיזה מדד הערכה תייעצו למרכז לאיכות ההוראה להשתמש על מנת להעריך את המסווג? הסבירו בקצרה (2-3 משפטים).מדד הע

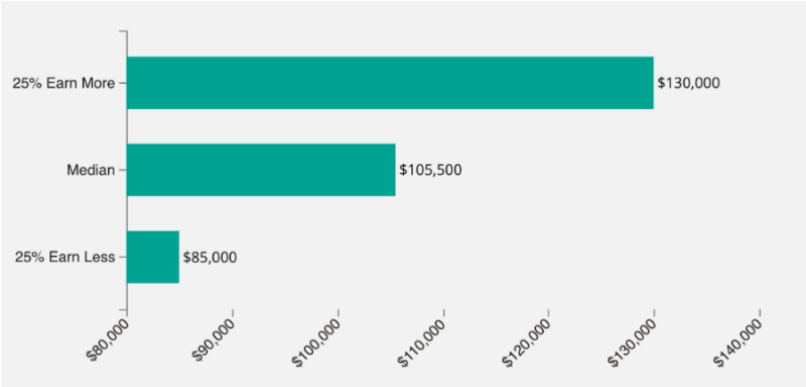
_____

_____

_____

שאלה 7 (8 נקודות)

הגרף הבא מציג נתונים לגבי משכורות במקצוע מסוים. כאשר החציון באותו המקצוע הוא 105,000, האחזון ה75 הוא 130,000 והאחוזון ה-25 הוא 85,000.



(1) (4 נק') ציינו שתי בעיות בויזואליזציה.

בעיה 1: _____

_____

ת.ז. _____

בעיה 2: _____

_____

(2) (4 נק') הציעו ויזואליזציה חלופית. ציינו את סוג הגרף המוצע והסבירו בקצרה מדוע הוא עדיף לדעתכם. ציירו סקיצה ברורה לויזואליזציה המוצעת (במלבן מתחת לשורות מטה).

_____

_____

_____

שאלה 8 (8 נקודות)

עמית הציע דרך למדוד מרחק בין שתי התפלגויות. נסמן את ההתפלגויות ב-X ו-Y. הקוד של עמית הינו:

```
import numpy as np

def compute_distance(x, y):
    if len(x) != len(y):
        return "distribution size doesn't match"
    distance = sum(abs(x-y))
    return distance

# example input
x_counts = np.array([1,2,3])
y_counts = np.array([1,1,4])
print(compute_distance(x_counts,y_counts))
```

(1) (2 נק') מה מטרתן של שורות 4-5 בקוד?

_____

_____

(2) (3 נק') ציינו בעיה הקיימת בנוסחה של עמית בשורה 6 (בעיה עקרונית, לא בעיית סינטקס של קוד), ותנו דוגמת קלט המדגימה את הבעיה. ציינו בבירור את הערכים של x ושל y, את הערך שהפונקציה של עמית תחזיר, ואת הערך המצופה.

ת.ז. _____

(3) (3 נק') הציעו תיקון לשורה 6. רשמו בבירור שורת קוד חלופית והסבירו במילים מה הנוסחה מחשבת וכיצד היא פותרת את הבעיה מהסעיף הקודם.

שאלה 9 (14 נקודות)

נניח שמנסים לחזות האם אנשים אוהבים קפה או לא. ידוע כי באוכלוסייה מסוימת הכוללת 1,000 אנשים, 90% מהאנשים אוהבים קפה ו-10% אינם אוהבים קפה. עבור אנשים שאוהבים קפה, למסווג יש דיוק (accuracy) של 80% (כלומר, מבין מי שאוהבים קפה, 80% יסווגו כאוהבי קפה ו-20% כלא). עבור אנשים שאינם אוהבים קפה, הדיוק (accuracy) הינו 98%.

(1) (5 נק') רשמו את מטריצת הבלבול של המסווג:

(2) (3 נק') אדם נדגם באקראי מהאוכלוסייה המתוארת למעלה. מה הסיכוי שהאדם יסווג נכון? (כלומר, יסווג כאוהב קפה אם הוא אכן אוהב קפה, או כלא אוהב קפה אם אינו אוהב קפה)

(3) (3 נק') בהנחה שנסמן את "אוהבי הקפה" בתור "positive", מה ערך ה-precision של המסווג? (אין צורך לחשב מספר עשרוני, אפשר להשאיר כשבר)

ת.ז. _____

(4) (3 נק') בהנחה שנסמן את "אוהבי הקפה" בתור "positive", איזה סוג טעויות יותר נפוץ עבור מסווג זה, False Positive או False Negative? הסבירו בקצרה.

_____

_____

_____

שאלה 10 (5 נקודות)

בחברת אמזון פיתחו מערכת אוטומטית לסינון קורות חיים, אך גילו כי המערכת מוטה כנגד נשים (כלומר, פחות סביר שאישה שתצליח בתפקיד תסווג כמתאימה לעומת הסיכוי שגבר שיצליח בתפקיד יסווג כמתאים). צוות משאבי האנוש הציעו להוריד מטבלת הנתונים את העמודה (מאפיין) של מגדר כדי להימנע מבעיה זו. האם לדעתכם זהו פתרון טוב לבעיה? הסבירו בקצרה (2-4 משפטים).

_____

_____

_____

_____

_____

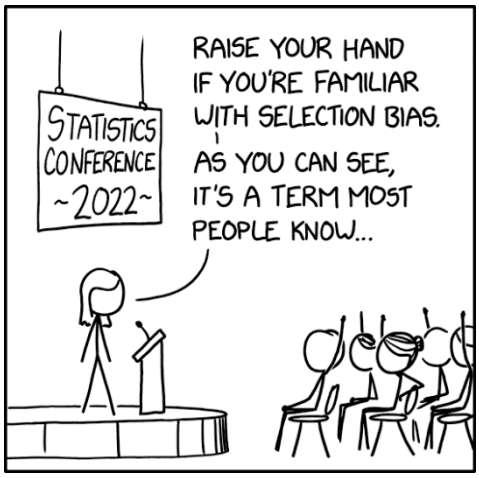
_____

ת.ז. _____

שאלת בונוס (2 נק'):

הסבירו מדוע הקומיקס הבא מצחיק:

תרגום הטקסט בקומיקס לעברית: "הרימו את היד אם אתם מכירים את המונח selection bias" ... "כפי שאתם יכולים לראות, זהו מושג שרוב האנשים מכירים..."



_____

_____

_____