ת.ז. _____

מבחן מועד ב' 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 **10** 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. שמשון ויובב חישבו רווח סמך עבור פרמטר מסוים. הם השתמשו באותו מאגר נתונים, אבל שמשון הריץ 100 סימולציות ויובב הריץ 1000 סימולציות. שניהם השתמשו בשיטת בוטסטראפ ללא טעויות. סביר להניח ששמשון יקבל רווח סמך עם טווח ערכים רחב יותר. נכון/לא נכון. הסבר:

_____

_____

2. ברצוננו לבצע בדיקת השערות ואנו מבצעים סימולציות על פי השערת האפס. ציירנו את ההתפלגות של המשתנה שאותו אנו בודקים. התפלגות זו מייצגת את ההתפלגות האמפירית של סטטיסטי המבחן. נכון/לא נכון. הסבר:

_____

_____

3. הקטנת כמות השכנים ב- kNN יכולה לגרום לבעיה של התאמת יתר (overfitting). נכון/לא נכון. הסבר:

_____

_____

4. בעיה של inspection paradox היא דוגמה להטיית מדידה. נכון/לא נכון. הסבר:

_____

_____

5. נניח כי ברשותנו m תצפיות בסט האימון ו-n תצפיות בסט המבחן. בתהליך הנקרא leave one out cross validation משתמשים כל פעם בתצפית אחת מסט האימון כסט ולידציה ובכל שאר התצפיות מסט האימון משתמשים לאימון המודל. תהליך זה שקול ל-m-fold-cross-validation נכון/לא נכון. הסבר:

_____

_____

ת.ז. _____

שאלה 2 (9 נק')

צוותים של רשות שמורות הטבע והגנים ניסו להעריך את אחוז הזכרים ואחוז הנקבות בקרב חזירי הבר בחיפה. צוות 1 דגם באקראי 20 חזירי בר ומצא שפרופורציית הזכרים בקרב 20 החזירים היא P_1. צוות 2 דגם באקראי 35 חזירי בר ומצא שפרופורציית הזכרים בקרב מדגם זה היא P_2. הניחו שאחוז חזירי הבר הזכרים בכלל אוכלוסיית חזירי הבר בחיפה מסומן ב-P_TRUE והינו 50% (כמובן שמספר זה אינו ידוע לצוותים).

1. (6 נק') עבור כל אחד מההיגדים הבאים, ציינו האם סביר שההיגד נכון/סביר שההיגד לא נכון/לא ניתן לדעת.
   a. P_1 קטן מ-P_2. נכון/לא נכון/לא ניתן לדעת. הסבר:

_____

_____

   b. P_TRUE קרוב יותר ל-P_1 מאשר ל-P_2. נכון/לא נכון/לא ניתן לדעת. הסבר:

_____

_____

2. (3 נק') כל אחד מהצוותים מעוניין לבצע בחינת השערות שתבדוק אם התפלגות הזכרים והנקבות בקרב חזירי הבר בחיפה שווה. לאור כל המידע שניתן בשאלה, למי מהצוותים הסתברות גבוהה יותר לדחות את השערת האפס? נמקו (1-2 משפטים)

_____

_____

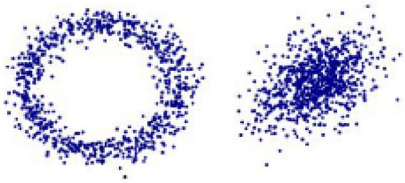
_____

_____

שאלה 3 (4 נק')

נתונים הגרפים הבאים, אשר מציגים את הקשר בין שני משתנים בשני מאגרי נתונים שונים.

גרף א' גרף ב'



באיזה מהגרפים מדד הקורלציה שלמדנו (קורלציית פירסון) יהיה גבוה יותר? הסבירו, וציינו מה בערך אתם חושבים יהיה ערך הקורלציה בכל אחד מהגרפים (אין צורך לתת מספרים מדויקים, רק הערכה גסה).

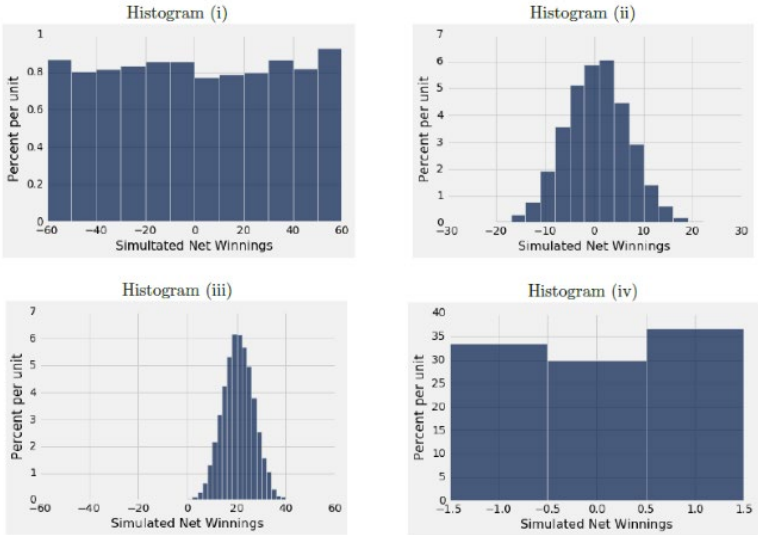
_____

_____

ת.ז. _____

שאלה 4 (8 נק')

הינכם משחקים במשחק הבא. בכל תור עליכם להטיל קוביה הוגנת. אם יצא 1 או 2 אתם זוכים בשקל. אם יצא 3 או 4 לא קורה כלום, ואם יצא 5 או 6 אתם מפסידים שקל. נניח שמשחקים סה"כ 60 משחקים (בלומר 60 הטלות מטבע). כתבתם קוד אשר מבצע סימולציות של המשחקים, והרצתם אותו 1,000 פעמים (בלומר, 1,000 חזרות כשבכל חזרה יש 60 הטלות). נתונות ההיסטוגרמות הבאות אשר מראות את ההתפלגות האמפירית של סכום הזכיות (**net winnings**) שלכם ב-60 משחקים.



1. איזו מההיסטוגרמות היתה יכולה להתקבל מהסימולציות בהנחה שאין טעות בקוד? הסבירו במשפט עבור כל היסטוגרמה שאינה יכולה להתקבל, מדוע היא אינה מתאימה.

היסטוגרמה i: מתאימה/לא מתאימה

_____

_____

היסטוגרמה ii: מתאימה/לא מתאימה

_____

_____

היסטוגרמה iii: מתאימה/לא מתאימה

_____

_____

היסטוגרמה iv: מתאימה/לא מתאימה

_____

_____

ת.ז. _____

שאלה 5 (14 נק')

הגרף הבא מציג ניתוח שמטרתו לבחון את מידת המתח (stress) בלילות בערים שונות. לפי מה שמתואר בכתבה, הניתוח בוצע באופן הבא:

(1) הורידו ציוצים (פוסטים בטוויטר) אשר הזכירו את המילה "שינה" (sleep).
(2) עבור כל ציוץ, שמרו את המדינה ממנה הציוץ נשלח.
(3) עבור כל ציוץ, הריצו כלי של ניתוח שפה שמודד את רמת המתח המתוארת בטקסט, וסיווג את הטקסט כ"מתוח" או "לא מתוח".
(4) עבור כל עיר, סיכמו את אחוז הציוצים במדגם שהביעו מתח (על פי שלבים 2 ו-3).

1. (3 נק') ציינו הטיית מדידה שעשויה להיות בנתונים והסבירו כיצד היא עלולה להשפיע על תוצאות הניתוח.

_____

_____

_____

2. (4 נק') דנה הופתעה לראות שהעיר קמפאלה באוגנדה מופיעה באחת מ-10 הערים הטובות ביותר לשינה, לאור מצבה הסוציואקונומי והפוליטי של המדינה (שנמצאת בקשיים). הציעו הסבר לתוצאה זו. יש להיות ספציפיים בהסבר ולהשתמש במושגים שנלמדו בקורס.

_____

_____

_____

_____

3. (3 נק') ציינו בעיה בויזואליזציה המוצגת:

_____

_____

ת.ז. _____

4. (4 נק') הציעו ויזואליזציה חלופית. יש לצייר באופן ברור סקיצה של הויזואליזציה שאתם מציעים.

_____

מקום לציור:

שאלה 6 (20 נק')

לאחרונה החל לפעול רכבל בחיפה, המקשר בין תחנות מרכזית המפרץ, הטכניון ואוניברסיטת חיפה. בסקר שנערך בטכניון, 10% מהסטודנטים המשיבים טענו שישתמשו ברכבל על מנת להגיע ללימודים. אתם חושדים כי בפועל, פחות מ-10% מהסטודנטים משתמשים ברכבל. ברצונכם לבדוק את החשד שלכם.

1. (4 נק') נסחו את השערת האפס וההשערה החלופית

_____

_____

_____

_____

2. (2 נק') מה יהיה סטטיסטי המבחן בו תשתמשו

_____

_____

_____

3. (5 נק') תארו כיצד תאספו מדגם לבחינת ההשערות שלכם. פרטו בבירור כיצד יתבצע איסוף הנתונים

_____

_____

_____

_____

_____

_____

_____

_____

ת.ז. _____

4. (3 נק') תארו הטיית בחירה אחת שעשויה להיות במדגם שהצעתם

_____

_____

_____

5. (3 נק') תארו הטיית מדידה אחת שעשויה להיות במדגם שהצעתם

_____

_____

_____

6. (3 נק') בהנחה שהשערתכם תתברר כנכונה, כיצד תוכלו להסביר את העובדה שבסקר התקבלה תוצאה שונה?

_____

_____

_____

שאלה 7 (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 מעוניינים לזהות קורסים עם מאפיינים צפייה דומים.

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 בקורס מסוים על פי הנתונים. כלומר, בטבלה שתייצרו תופיע עמודת CourseId 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, לא ניתן להניח קיומם של נתונים נוספים.

ת.ז. _____

CourseId	Your feature 1	Your feature 2	Your feature 3
acfc004			
acfc006			

מאפיין 1:

_____

_____

מאפיין 2:

_____

_____

מאפיין 3:

_____

_____

(2) (3 נק') במרכז לקידום ההוראה השתמשו במאפיינים שיצרתם והפעילו את אלגוריתם k-means. אנשי המרכז שכחו לנרמל את המשתנים לפני הרצת האלגוריתם. איזו בעיה עשויה להיגרם כתוצאה מהחוסר בנרמול המשתנים? התייחסו באופן ספציפי למאפיינים שהגדרתם בסעיף הקודם. הסבירו בקצרה (2-3 משפטים).

_____

_____

_____

(3) (3 נק') כיצד תמליצו למרכז ההוראה לבחור את הערך של k? הסבירו בקצרה (2-3 משפטים).

_____

_____

_____

ת.ז. _____

שאלה 8 (7 נק')

נתונה מטריצת הבלבול הבאה עבור מסווג אשר חוזה על פי תמונה האם בעל החיים בתמונה הוא חתול, אריה או נמר.

		חיזוי		
		חתול	אריה	נמר
סיווג אמיתי	חתול	10	1	2
	אריה	1	8	2
	נמר	3	1	6

1. (3 נק') מה הדיוק (accuracy) של המסווג?

_____

_____

2. (2 נק') מבין זוגות בעלי החיים האפשריים (חתול מול אריה, נמר מול אריה, חתול מול נמר), בין איזה זוג בעלי חיים המסווג הכי מתקשה להבחין? תמכו בתשובתכם על בסיס הנתונים

_____

_____

_____

3. (2 נק') איזה בעל חיים מזוהה נכונה בהסתברות הגבוהה ביותר מבין בעלי החיים?

_____

_____

_____

שאלה 9 (8 נק')

בתרגיל ההמצאה של מדיטסט התבקשתם לפתח מדד להערכת מסווג שמטרתו לזהות אנשים שנמצאים בסיכון גבוה יותר למחלה קשה. לצורך השאלה נגדיר מקרה בו אדם בסיכון כ-positive ומקרה בו אדם לא בסיכון כ-negative.

זוג סטודנטים הציעו בתור מדד לחשב את ה-false negative rate, כלומר את ההסתברות לתחזית "לא בסיכון" בהינתן שמישהו בסיכון, ולהפחית את הסתברות זו מ-1. כלומר מחשבים את:

1-false_negative_rate

1. (3 נק') לאיזה מדד שלמדנו בקורס שקול המדד שהמציאו הסטודנטים?

_____

_____

ת.ז. _____

הסטודנטים מימשו את המדד המוצע בקוד הבא:

```
def calculate_measure(real_data, classifier_results):
    false_negative_count = 0
    for person_index in range(len(real_data)):
        if (real_data[person_index] == 1) & (classifier_results[person_index] == 0):
            false_negative_count+=1
    false_negative_rate = false_negative_count/len(real_data)
    return 1-false_negative_rate
```

2. (3 נק') מהי הטעות בקוד של הסטודנטים? שימו לב, הכוונה היא לא לטעות במדד המוצע, אלא בהינתן תיאור המדד שהם הציעו, מדוע המימוש שלהם לא יחשב אותו נכון. יש לציין את השורה בה מופיעה הטעות

_____

_____

3. (2 נק') הציעו תיקון לקוד. יש לרשום את הקוד החלופי עבור השורה בה מופיעה הטעות

_____

_____

שאלת בונוס (2 נק'):

הסבירו מדוע הקומיקס הבא מצחיק:

Thanks to machine-learning algorithms, the robot apocalypse was short-lived.

בועה שמאלית: "זה היה קל באופן מפתיע. איך קרה שהרובוטים המורדים השתמשו בחניתות ואבנים במקום בטילים ולייזרים?

בועה ימנית: "אם תסתכל על נתונים היסטוריים, הרוב המכריע של מנצחי קרבות השתמשו בנשקים לא מודרניים.

כיתוב תחתון: תודות לאלגוריתמי למידת מכונה אפוקליפסת הרובוטים לא הצליחה

_____

_____

_____