ת.ז. _____

מבחן מועד ב' במבוא לניתוח נתונים בפייתון (096202) סמסטר אביב 2019

מרצה: עפרה עמיר, מתרגלים: רפי שללה, שרון הירש, בודק: תום בר

1. לרשותכם **שעתיים וחצי** לפתור את הבחינה.
2. הבחינה היא עם **חומר סגור**. אסור להשתמש בכל חומר עזר.
3. הבחינה כוללת **10 עמודים** ובהם **11 שאלות**. יש לענות על כל השאלות (השאלה האחרונה היא שאלת בונוס). הניקוד של כל שאלה מופיע לצידה.
4. **את התשובות יש לכתוב רק על גבי טופס הבחינה.**
5. על טפסי שאלות ותשובות הבחינה יש לכתוב רק מספר תעודת זהות (ולא שם). חובה לכתוב מספר ת.ז. על כל דפי הטופס והמחברת.
6. אין להפריד את דפי טופס הבחינה.
7. חובה להחזיר בסיום הבחינה את טופס שאלות הבחינה. אין צורך להחזיר מחברות טיוטא.

בהצלחה!

שאלה	ניקוד
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
בונוס	

ת.ז. _____

שאלה 1 (12 נק')

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

1. נניח שהמדגם שלנו כולל את הדגימות הבאות {2, 5, 6, 6}. יותר סביר שבדגימת בוטסטראפ תתקבל הדגימה {5, 6, 6, 6} מאשר הדגימה {2, 5, 2, 2}. נכון/לא נכון. הסבר:

_____

_____

1. לבחירה של מרכזי הקלאסטרים הראשונים באלגוריתם k-means אין השפעה על הקלאסטרים הסופיים שייווצרו. נכון/לא נכון. הסבר:

_____

_____

2. שיטת בוטסטראפ לא תהיה טובה בשביל לקבל רווח סמך לערך המקסימלי של משתנה מסוים באוכלוסיה .נכון/לא נכון. הסבר:

_____

_____

3. למסווגים שונים יכול להיות מדד דיוק כללי (accuracy) זהה אבל מדד true positive שונה. נכון/לא נכון. הסבר:

_____

_____

שאלה 2 (5 נק')

הרצתם KNN על דוגמאות האימון (training set) וקיבלתם דיוק של 95%. על דוגמאות הבדיקה (test set) קיבלתם דיוק של 80%. כיצד ניתן להסביר את ההבדל בין הביצועים של המסווג על הדוגמאות השונות?

_____

_____

_____

_____

_____

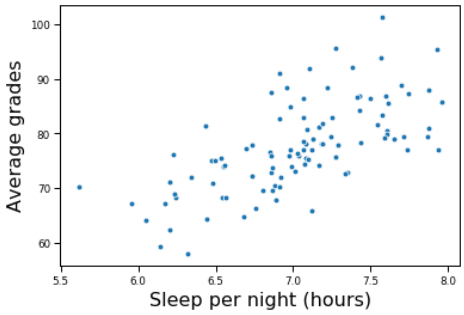
_____

ת.ז. _____

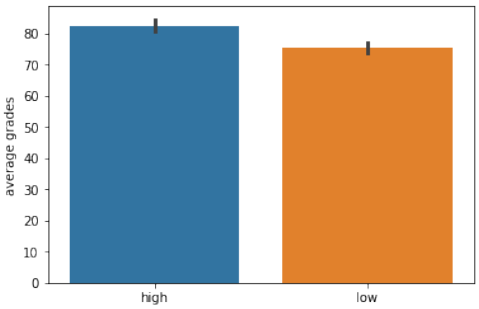
שאלה 3 (18 נק')

במסגרת מחקר, נאספו הנתונים הבאים על קבוצת סטודנטים בטכניון: פקולטה, גיל, שעות שינה בלילה, שעות צפייה בטלוויזיה ביום, מגורים לבד או עם שותפים, ממוצע ציונים. נתוני המחקר נאספו באופן הבא: נשלח אימייל לכלל הסטודנטים (10,000) עם קישור לסקר, ומתוכם ענו על הסקר 536 סטודנטים.

(א) (4 נק') לפניכם גרף אשר מראה את מספר שעות השינה בלילה עבור כל סטודנט (ציר X) וממוצע הציונים (ציר Y). מה ניתן להסיק על פי הגרף לגבי הקשר בין שעות שינה לציונים? האם הייתם ממליצים לסטודנטים להגדיל את כמות שעות השינה שלהם? (התעלמו מהשפעות של הטיות אפשריות והתייחסו רק לגרף)



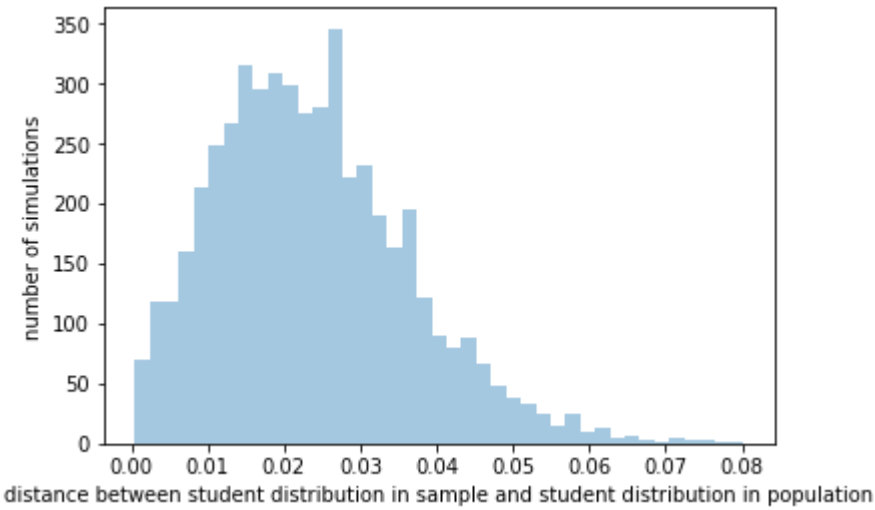
(ב) (7 נק') החוקרים חילקו את שעות הצפייה בטלוויזיה לשתי קטגוריות – "גבוה" (מעל שעתיים ביום) ו"נמוך" (מתחת לשעתיים). להלן גרף של ממוצעי הציונים של סטודנטים בכל אחת מהקטגוריות. ההבדל בין הציונים יצא מובהק סטטיסטית.



מקום לאיור של גרף:

תנו דוגמה למקרה בו הקשר בין צפייה בטלוויזיה לציונים הוא תוצאה של פרדוקס סימפסון. ציירו (במקום המיועד למעלה) דוגמה לגרף שיוכל להראות את התופעה במקרה זה (אין צורך לדייק במספרים, רק העיקרון חשוב).

(ג) (7 נק') החוקרים רצו לבדוק האם הפקולטות השונות מיוצגות במדגם באופן שתואם את ייצוג המחלקות בכלל אוכלוסיית הסטודנטים בטכניון. בטכניון יש 3 פקולטות - הנדסת תעשייה וניהול (5000 סטודנטים), מדעי המחשב (3000 סטודנטים), והנדסת חשמל (2000 סטודנטים). לצורך הבדיקה, החוקרים הריצו 5000 סימולציות, כאשר בכל אחת הם דגמו 536 סטודנטים על פי ההתפלגות של 50% תעשייה וניהול, 30% מדעי המחשב ו-20% הנדסת חשמל. עבור כל סימולציה, הם חישבו את המרחק (total variation distance) בין התפלגות הסטודנטים בין הפקולטות שיצאה בסימולציה, לבין ההתפלגות האמיתית באוכלוסייה. התקבלה ההיסטוגרמה הבאה:



במדגם של החוקרים היו 252 סטודנטים מתעשייה וניהול (47%), 172 ממדעי המחשב (32%) ו-112 סטודנטים מהנדסת חשמל (21%). האם החוקרים יכולים לדחות ברמת מובהקות של 0.05 את ההשערה שהתפלגות הסטודנטים בין הפקולטות במדגם תואמת את ההתפלגות בין הפקולטות בכלל אוכלוסיית הסטודנטים? הסבירו בבירור את תשובתכם והצדיקו אותה באופן מספרי.

ת.ז. _____

שאלה 4 (16 נק')

חוקרים רוצים לבדוק האם הסיכוי שסטודנטים מהטכניון שלקחו קורס מבוא לניתוח נתונים יעבדו כמדעני נתונים גבוהה מאשר הסיכוי שסטודנטים מהטכניון שלא לקחו את הקורס יעבדו כמדעני נתונים.

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

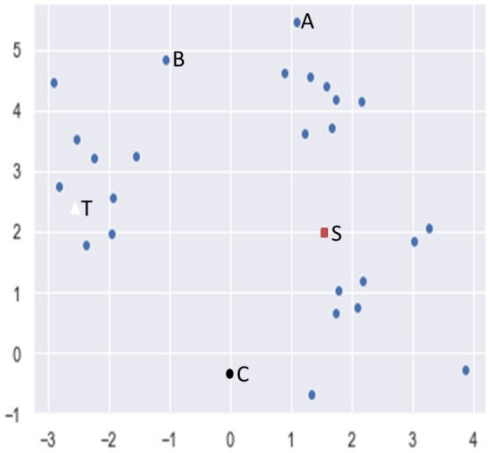
(ב) (6 נק') הציעו דרך לאסוף נתונים כדי לבחון את השערת המחקר. נסחו בבירור מה האוכלוסיה הכללית שמעניינת אתכם, אילו אנשים תכללו במדגם שלכם, ובאיזו דרך תאספו את הנתונים.

ת.ז. _____

(ג) (6 נק') ציינו הטיה אחת מסוג הטיית בחירה (selection bias) והטיה אחת מסוג הטיית מדידה (measurement bias) שעלולות להיות בנתונים שתאספרו וכיצד הן ישפיעו על הניתוח שתבצעו.

שאלה 5 (12 נק')

ברצונכם לבצע קלאסטרינג על הנקודות המסומנות בכחול על הגרף, באמצעות k-means. מרכזי הקלאסטרים אותחלו להיות העיגול שלימינו האות C (x=0,y=-0.2), הריבוע שלימינו האות S (x=2,y=1.5) והמשולש שלימינו האות T (x=-2.5,y=2.3) ומצוירים גם על הגרף. בכל השאלות הבאות, אתם יכולים להניח שהמרחקים בין היחידות בציר X וY זהים (כלומר מספיק להעריך מרחקים בעין, אין צורך לחשב).



(א) (2נק') בסוף איטרציה אחת של האלגוריתם, כמה נקודות ישויכו לקלאסטר שמרכזו בעיגול C? הסבירו.

ת.ז. _____

(ב) (2נק') בסוף איטרציה אחת של האלגוריתם, לאיזה קלאסטר תשויך הנקודה A (x=1.05, y=5.8)?

_____

(ג) (2נק') בסוף איטרציה אחת של האלגוריתם, לאיזה קלאסטר תשויך הנקודה B (x=-1.05, y=4.9)?

_____

(ד) (3 נק') לאן יזוז מרכז הקלאסטר C בסוף האיטרציה הנוכחית של האלגוריתם? הסבירו.

_____

_____

_____

(ה) (3 נק') אם היינו ממשיכים להריץ את האלגוריתם איטרציה נוספת, האם מספר הנקודות המשויכות לקלאסטר שמרכזו הראשוני היה בריבוע S יעלה או ירד? הסבירו.

_____

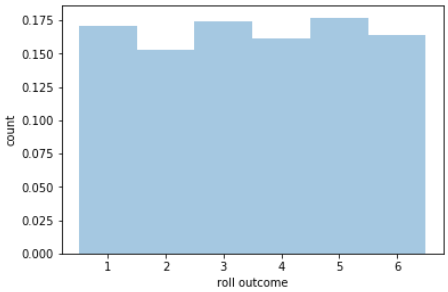
_____

_____

_____

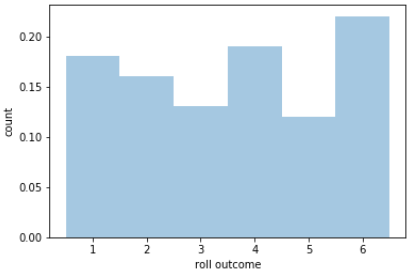
שאלה 6 (6 נק')

סטודנטים התבקשו להריץ 1000 סימולציות של הטלת קוביה הוגנת. כך נראית התפלגות לדוגמה על בסיס 1000 סימולציות:

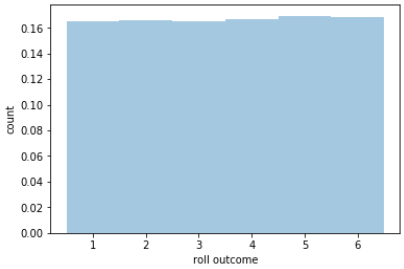


לדני היתה טעות בקוד והוא הריץ בטעות 100 סימולציות. ליעל היתה טעות בקוד והיא הריצה בטעות 10000 סימולציות. איזה מהגרפים הבאים קיבל דני ואיזה מהגרפים קיבלה יעל? הסבירו את תשובתכם.

גרף א':



גרף ב':



_____

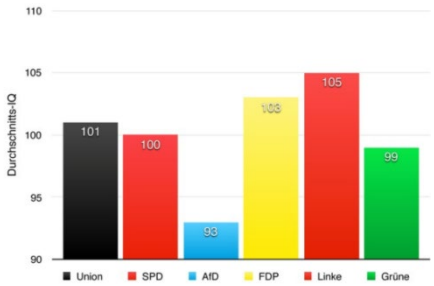
_____

_____

ת.ז. _____

שאלה 7 (8 נק')

הגרף הבא מציג את ממוצע ה-IQ (ציר Y) עבור קבוצות שונות של אנשים בגרמניה (ציר Y).



(א) ציינו בעיה הקיימת בגרף ושעלולה ליצור רושם מוטעה לגבי הנתונים עבור מי שמסתכל על הגרף.

_____

_____

_____

_____

(ב) הציעו דרך לתקן את הבעיה. הראו איור סכמתי של התיקון המוצע.

[מקום לאיור]

שאלה 8 (6 נק')

הסבירו את המושג "מדגם נוחות" (convenience sample). הניחו שברצונכם לבצע סקר כדי לחזות את תוצאות הבחירות. תנו דוגמה לדרך בחירת נשאלים שתהווה מדגם נוחות, והסבירו מה הבעיה בשימוש במדגם כזה.

_____

_____

_____

_____

_____

_____

ת.ז. _____

שאלה 9 (5 נק')

נתונים המשפטים הבאים:

- חם מאוד בקיץ בישראל
- אסור לגעת בתנור חם
- בקיץ אוכלים גלידה

רשמו עבור כל משפט מה יהיה הייצוג bag of words שלו. הניחו שלא מורידים stopwords.

_____

_____

_____

_____

_____

שאלה 10 (12 נק')

בארה"ב נעשה שימוש במערכת COMPAS, אשר חוזה את הסיכון שנאשם במשפט יפשע שוב אם ישוחרר מהכלא. המערכת מסייעת לשופטים לקבל החלטות בנושא שחרור בערבות. עלו טענות על כך שבמערכת יש הטיה כנגד נאשמים ממוצא אפרו-אמריקאי. נתונות מטריצות הבלבול (confusion matrix) שהתקבלו עבור נאשמים ממוצא אפרו-אמריקאי ועבור נאשמים לבנים ומדד הדיוק הכללי שהתקבל עבור כל אוכלוסיה.

מטריצת הבלבול עבור נאשמים לבנים:

דיוק (accuracy) כללי: 0.78

חיזוי האלגוריתם: לא יפשע שוב	חיזוי האלגוריתם: יפשע שוב	
215	49	סיווג אמיתי: פשע שוב
852	34	סיווג אמיתי: לא פשע שוב

מטריצת הבלבול עבור נאשמים שחורים:

דיוק (accuracy) כללי: 0.85

חיזוי האלגוריתם: לא יפשע שוב	חיזוי האלגוריתם: יפשע שוב	
15	81	סיווג אמיתי: פשע שוב
642	108	סיווג אמיתי: לא פשע שוב

ת.ז. _____

(א) הביעו את עמדתכם לגבי הטענה על אפליה כנגד נאשמים ממוצא אפרו-אמריקאי בהתבסס על תוצאות הביצועים של אלגוריתם הסיווג. הסבירו את טענתכם בהתבסס על נתוני הביצועים של המסווג.

(ב) כדי להימנע מסכנה של אפלייה, הוצע להוציא את המשתנה "גזע" מטבלת הנתונים, ולהריץ מחדש את הסיווג. האם לדעתכם צעד זה ימנע סכנה של אפלייה? נמקו את עמדתכם.

שאלת בונוס (2 נק'):

הסבירו מה מצחיק בקומיקס הבא:

Null hypothesis: השערת האפס

Conclusively: באופן ודאי