TECHNION Israel Institute of Technology | FACULTY OF DATA AND DECISION SCIENCES MAKING SENSE OF IT

ת"ז : _____

מבוא לניתוח נתונים; - 094202

מבחן מועד א' – סמסטר אביב תשפ"ג

טור א' - פתרון

תאריך הבחינה : 12.07.2023

מרצה : אורי פלונסקי

מתרגלים : ספיר גרשוב, טל איפרגן, ניב ברדס, טל קרייצר

הוראות :

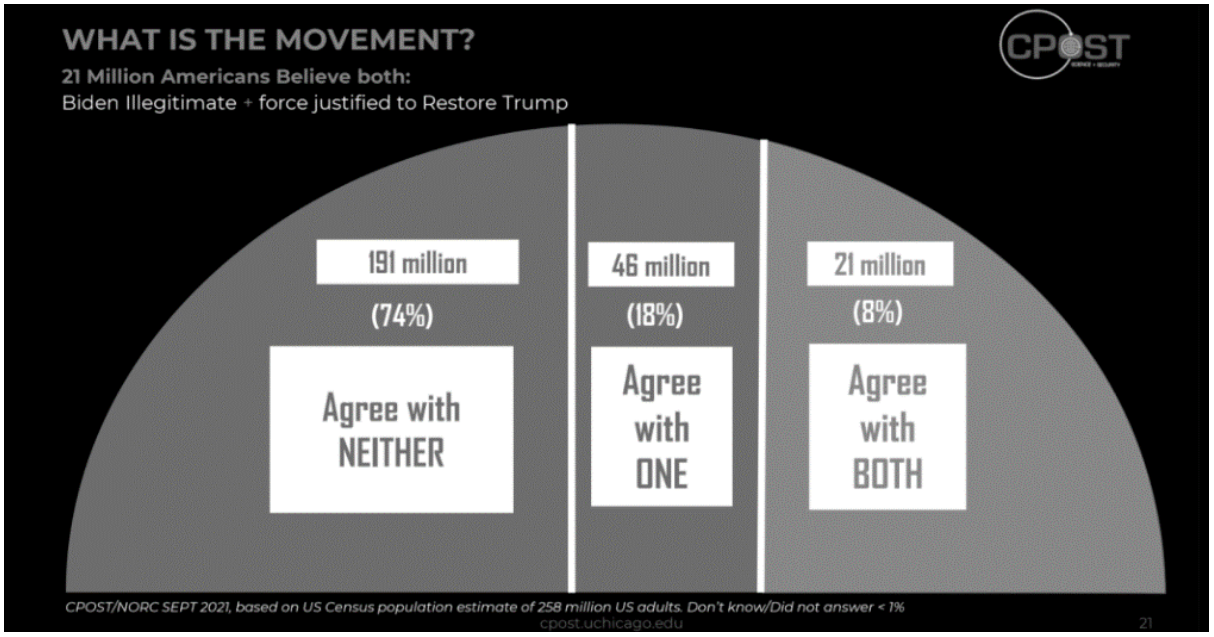
1. לרשותכם **שלוש שעות** לפתור את הבחינה.
2. מותר להשתמש במחשבון פשוט בלבד. אין להשתמש בכל חומר עזר אחר.
3. הבחינה כוללת 13 עמודים (כולל עמוד זה) ובהם 10 שאלות ושאלת בונוס. בדקו בתחילת הבחינה שיש ברשותכם את כל העמודים.
4. ניקוד כל שאלה (ולרוב כל סעיף) מצוין לידה. סך כל **הניקוד האפשרי בבחינה הוא 102** נקודות.
5. **את התשובות יש לכתוב רק על גבי טופס הבחינה, ובמקומות המיועדים לכך בלבד.** אין להוסיף מלל מעבר לשורות המיועדות. מחברות הטיוטה יושמדו לאחר סיום הבחינה וממילא לא יבדקו.
6. חובה לכתוב מספר תעודת זהות על כל אחד מדפי טופס הבחינה.
7. אסור להפריד את דפי טופס הבחינה.
8. בשום שלב, במהלך הבחינה או לאחר סיומה, אסור להוציא מחדר הבחינה שום דבר שלא הבאתם אתכם לחדר הבחינה. חובה להחזיר בסיום הבחינה את הטופס וכל מחברת בה השתמשתם.

בהצלחה!

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

שאלה 1 (9 נק')

נתונה הוויזואליזציה הבאה :



א. פרטו בנוגע לשתי בעיות בוויזואליזציה המטעות או מבלבלות את הקורא. (אין להתייחס לצבעים של הוויזואליזציה). (6 נק')

- הפרה של עקרון הדיו הפרופורציונלי : 26% הם יותר מחצי מהגרף (אפשר לכתוב גם על היחס בין 8% ל-18% לעומת היחס בשטחים שלהם)

- השימוש בגרף (חצי) עוגה (pie chart) באופן כללי אינו מומלץ ובפרט שגוי כאשר יש חשיבות ליחסים בין הקבוצות השונות, ובמקרה הזה יש חשיבות.

- אפשר לקבל גם הערה על כך שהשימוש בחצי עוגה מרמז על כך שהנתונים מייצגים חצי משלם כלשהו.

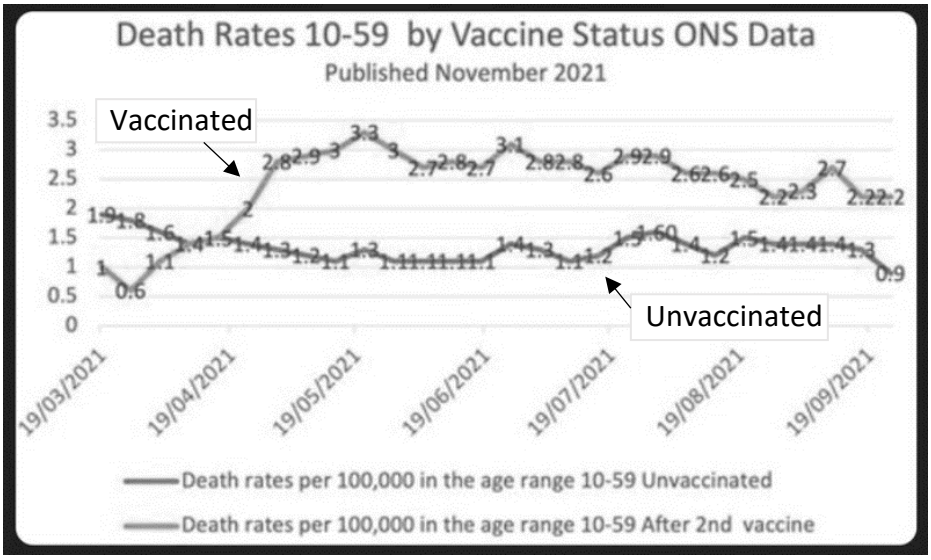
- אפשר לקבל גם הערה על כך שלא באמת יתכן שנשאלו 258 מיליון אמריקאים בוגרים. אלו כנראה הערכות על סמך אחוז מענה בסקר, ואקסטרפולציה כזו היא די בעייתית.

ב. ציירו מחדש את הוויזואליזציה כך שלא תהיה מטעה (אין צורך לדייק, רק שנבין מה הרעיון) (3 נק') [מקום לאיור]

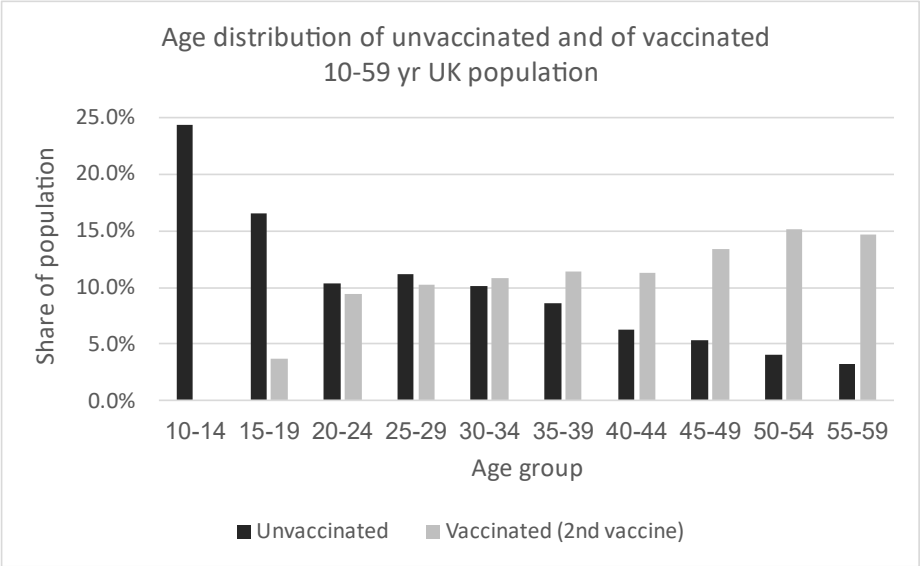
גרף עמודות עם ציר Y שמתחיל מאפס ומראה את האחוזים בלבד עבור כל אחת משלוש הקטגוריות. (אין להשתמש בגרף עוגה משום צורה.)

שאלה 2 (7 נק')

הגרף הבא, שפורסם בחודש נובמבר 2021, מתאר את שיעורי התמותה לכל 100,000 איש בגילים 10-59 בבריטניה, לאורך מספר חודשים, עבור שתי קבוצות אוכלוסייה: מחוסנים פעמיים לקורונה ולא מחוסנים כלל.



בנוסף, פורסם הגרף הבא, אשר מתאר את התפלגות הגיל של המחוסנים פעמיים והתפלגות הגיל של הלא מחוסנים באוכלוסיית בריטניה בגיל 10-59, נכון לנובמבר 2021:



אלכס הציג את הגרף העליון (בלבד) וטען "איני יודע איך ניתן להסביר את הנתונים בלי להניח שהחיסונים מגדילים את שיעור המוות". הסבירו, בהתבסס על הגרף התחתון, מדוע טענה זו אינה נובעת בהכרח מהנתונים. השתמשו במושגים שנלמדו בקורס: איזו תופעה סטטיסטית מיוצגת בגרפים אלו? תנו לגרף העליון הסבר חלופי להסברו של אלכס. איזה נתונים עליכם יהיה לאסוף כדי לבחון אם ההסבר החלופי שלכם עקבי עם הנתונים על חיסוני קורונה ושיעורי התמותה?

התופעה היא (ככל הנראה) הפרדוקס של סימפסון. הסבר חלופי הוא שלקבוצות הגיל המבוגרות יותר מבין האנשים בגילים 10-59 (נניח מעל 40) יש שיעורי מוות הרבה יותר גבוהים מלקבוצות הגיל הנמוכות יותר (ילדים ונוער), באופן טבעי. במצב כזה, מאחר וקבוצת המחוסנים כוללת יחסית הרבה אנשים מקבוצות הגיל המבוגרות וקבוצת הלא מחוסנים כוללת הרבה אנשים מקבוצות הגיל הצעירות, נקבל

שמחוסנים מתים בשיעור גבוה יותר מלא מחוסנים כאשר מסתכלים על כל הקבוצות יחד, אבל בכל תת קבוצה נראה שללא מחוסנים יש שיעורי מוות גבוהים יותר ממחוסנים פעמיים. כדי לבחון את ההסבר, יש לאסוף נתונים על שיעורי המוות של תתי קבוצות גיל מבין גילי 59-10 (במנותק מסטטוס חיסון), למשל לפי החלוקה לתתי קבוצות שבגרף השני.

שאלה 3 (17 נק')

חברת "המכשירים" מקיימת תכניות הכשרה ופיתוח מיומנויות לעובדיה על בסיס קבוע. לאחרונה, הוצע לחברה ליישם תכנית חדשה להכשרה ליצירתיות. מאחר ויש לחברת "המכשירים" תכנית קיימת להכשרה ליצירתיות, הנהלת החברה רוצה לבחון את ההבדל בין התוצאות של שתי התכניות. לשם כך, הנהלת החברה החליטה לערוך ניסוי על מדגם מקרי גדול של עובדיה שטרם קיבלו תכנית הכשרתה ליצירתיות. תכנית הניסוי היא שהעובדים במדגם יחולקו לשתי קבוצות בגודל דומה, כך שקבוצה א' תקבל את תכנית ההכשרה החדשה וקבוצה ב' תקבל את תכנית ההכשרה הקיימת. בתום ההכשרות, בכוונת ההנהלה למדוד את היצירתיות של העובדים משתי הקבוצות ואז לבצע מבחן בדיקת השערות הבוחן אם, בממוצע, ציון היצירתיות של עובדים שקיבלו את התכנית החדשה שונה מציון היצירתיות של עובדים שקיבלו את התכנית הקיימת.

א. כיצד הייתם מציעים להנהלת החברה לחלק את העובדים במדגם לשתי הקבוצות? הסבירו הצעתכם ב-2-1 משפטים. (3 נק')

כדאי לחלק את העובדים בצורה אקראית בין הקבוצות. חלוקה כזו תאפשר להשיג קבוצות טיפול וביקורת הדומות זו לזו (בממוצע), חוץ מתכנית ההכשרה ו(אולי) להסיק סיבתיות.

ב. מהי השערת האפס ומהי ההשערה האלטרנטיבית? (3 נק')

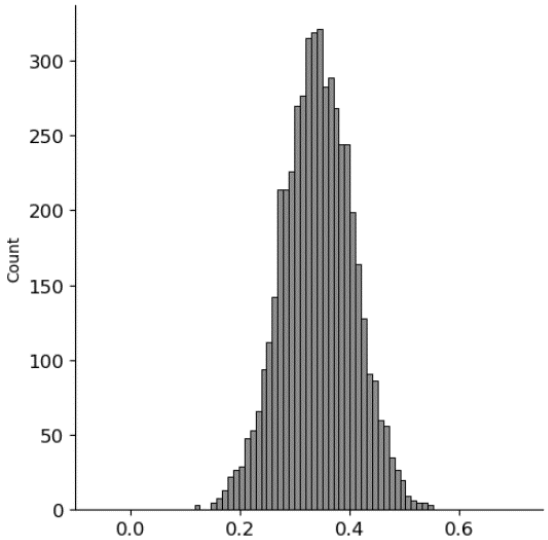
השערת האפס : ציון היצירתיות של עובדי החברה שמקבלים את התכנית החדשה זהה לציון היצירתיות של עובדי החברה שמקבלים את התכנית הקיימת.
השערה אלטרנטיבית : ציון היצירתיות של עובדי החברה שמקבלים את התכנית החדשה שונה מציון היצירתיות של עובדים שמקבלים את התכנית הקיימת.

ג. מהו הפרמטר שעבורו מתבצעת בדיקת ההשערות ומהו האומד לפרמטר זה? (2 נק')

פרמטר : ההפרש בין ממוצע ציון יצירתיות של עובדי החברה המקבלים את התכנית החדשה וממוצע ציון היצירתיות של עובדי החברה המקבלים את התכנית הקיימת.
אומד : ההפרש בין ממוצע ציוני היצירתיות של עובדים במדגם שקיבלו את התכנית החדשה ועובדים במדגם שקיבלו את התכנית הקיימת.

[השאלה ממשיכה בעמוד הבא]

ד. כדי לבצע את מבחן בדיקת ההשערות, ההנהלה החליטה למצוא רווח סמך לפרמטר מסעיף ג', באמצעות שיטת הבוטסטראפ. להלן היסטוגרמה של הערכים שהתקבלו עבור האמד לפרמטר כתוצאה מאלפי רפליקציות של תהליך הבוטסטראפ.



לפי ההיסטוגרמה, האם ניתן לדחות את השערת האפס ברמת מובהקות של 0.01? נמקו. (3 נק')

כן, ניתן לדחות בכל רמת מובהקות סבירה, ובפרט בר"מ 0.01, שכן באף אחת מאלפי הרפליקציות לא נמצא הפרש ממוצעים השווה לאפס ובפרט ר"ס ברמת ביטחון 99% לא יכלול את 0.

ה. לאחר סיום הניסוי התברר כי כל המפגשים בתכנית ההכשרה החדשה התקיימו בשעות הבוקר, בעוד שכל המפגשים בתכנית ההכשרה הקיימת התקיימו בשעות אחר הצהריים. האם נתון נוסף זה משנה את <u>מסקנת מבחן בדיקת ההשערות</u> מסעיף ד'? הסבירו. (3 נק')

נתון זה לא משנה את מסקנת המבחן. מבחן בדיקת השערות לא מניח למשל שהקבוצות דומות בהכל חוץ מהטיפול. ניתן להפעיל מבחן על כל שתי קבוצות ברות השוואה.

ו. האם הנתון הנוסף מסעיף ה' משנה את <u>הפירוש שצריכה לתת הנהלת החברה לתוצאות הניסוי</u>? הסבירו. (3 נק')

כן. אם החלוקה המקורית לקבוצות היתה אקראית ולא היו הבדלים נוספים בין הקבוצות, אז ניתן היה להסיק סיבתיות: התכנית החדשה גורמת לציון יצירתיות ממוצע שונה מהתכנית הקיימת. מאחר ויש כאן ערפלן (confounder), לא ניתן כעת להסיק סיבתיות. בפרט, יתכן שההבדל בין הקבוצות נובע מהשעות השונות בהן התקיימו התכניות.

שאלה 4 (7 נק')

העץ הנדיר סטטיס-טרי יכול לגדול לאחד משישה גבהים אפשריים : 55, 66, 77, 88, 99, או 110 ס"מ.

הטבלה הבאה מתארת את התפלגות אוכלוסיית העץ בישראל :

גובה העץ (ס"מ)	מספר העצים בישראל
55	27
66	48
77	60
88	32
99	14
110	1

א. נלקח מדגם מקרי פשוט (simple random sample) של עצים מאוכלוסיית העץ הנ"ל בישראל. לכל אחד מהבאים ציינו אם הוא יכול להיות סט הגבהים של העצים שעלו במדגם. אם תשובתכם היא לא, נמקו במשפט אחד (אין צורך לנמק בסעיפים בהם תשובתכם היא שכן). (3 נק')

1. {55, 66, 77, 88, 99, 110} : **כן** / לא נימוק (אם לא) : _____

_____

2. {99, 99, 99, 99, 99, 99} : **כן** / לא נימוק (אם לא) : _____

_____

3. {66, 66, 88, 88, 110, 110} : כן / **לא** נימוק (אם לא) : באוכלוסייה יש רק עץ אחד בגובה 110 ס"מ. לא יתכן שמדגם מקרי פשוט מאוכלוסייה זו יכלול שני עצים בגובה 110

ב. נניח שנלקח מדגם מקרי פשוט של עצים מאוכלוסיית העץ הנ"ל בישראל וסט הגבהים של העצים שעלו במדגם הוא : {66, 77, 77, 88, 99, 110}.
לכל אחד מהבאים ציינו אם הוא יכול להיות סט הגבהים של עצים במדגם בוטסטראפ שנלקח מהמדגם המקרי הפשוט המקורי. אם תשובתכם היא לא, נמקו במשפט אחד (אין צורך לנמק בסעיפים בהם תשובתכם היא שכן). (4 נק')

1. {55, 66, 77, 88, 99, 110} : כן / **לא** נימוק (אם לא) : מדגם בוטסטראפ לא יכול לכלול תצפיות מקטגוריה שלא קיימת במדגם המקורי (כמו עץ בגובה 55)

2. {99, 99, 99, 99, 99, 99} : **כן** / לא נימוק (אם לא) : _____

_____

3. {66, 66, 88, 88, 110, 110} : **כן** / לא נימוק (אם לא) : _____

_____

4. {66, 77, 77, 77, 88} : כן / לא נימוק (אם לא) : מדגם בוטסטראפ חייב להיות באותו גודל כמו המדגם המקורי.

שאלה 5 (11 נק')

שרון ודני מנתחים נתונים של היסטוריית מזג האוויר בניו יורק בעשרים השנים האחרונות. להלן ארבע השורות הראשונות מקובץ הנתונים ותיאור הפיצ'רים :

Date	Temperature	Humidity	Precipitation	Cloudiness
13/03/2003	49	45	0.1	1
14/03/2003	55	65	1.2	1
15/03/2003	71	62	0	0
16/03/2003	67	52	0	0

Date – תאריך (בפורמט DD/MM/YYYY)

Temperature - טמפרטורה יומית ממוצעת במעלות פרנהייט (טווח : [8, 101], חציון : 55)

Humidity - לחות יומית ממוצעת באחוזים (טווח : [0, 100], חציון : 63)

Precipitation - כמות משקעים יומית כוללת באינצ'ים (טווח : [0, 7.1], חציון : 0)

Cloudiness - (target class) משתנה בינארי : יום מעונן (1) או לא (0)

הם מעוניינים לבנות מסווג שינבא את מחלקת המטרה : האם היתה או לא היתה עננות ביום מסוים, ולשם כך מבקשים לחלק את הנתונים לסט אימון ולסט מבחן.

בתחילה, שרון מבצעת את החלוקה באמצעות הקוד הבא :

```
train_size = int(0.8*df.shape[0])
train_set = df.iloc[:train_size]
test_set = df.iloc[train_size:]
```

דני טוען שעליהם קודם כל לערבל (shuffle) את סדר התצפיות בטבלת הנתונים ואז להריץ את אותו הקוד. שרון לא מסכימה איתו.

א. הסבירו את ההיגיון מאחורי הטענה של כל אחד מהם : מתי מתאים להשתמש בכל אחת מהגישות, ומדוע? (5 נק')

כאשר המטרה היא לנבא עננות ביום עתידי, בהסתמך על מזג האוויר ההיסטורי (למשל על סמך הימים הקודמים, או לפי חודש מסוים בשנה), לא סביר לערבל את הנתונים כי אז נקבל מודל שאומן על נתונים מהעתיד כדי לנבא את העבר. עם זאת, אם המטרה היא רק לבנות מודל שינבא עננות ביום מסוים על סמך הפיצ'רים הנומריים, בלי להתייחס לפיצ'ר התאריך (כאילו מזג אוויר ביום אחד הוא בלתי תלוי במזג האוויר ביום אחר) אז ערבול הנתונים נכון.

[השאלה ממשיכה בעמוד הבא]

ב. לאחר שקיבלו החלטה כיצד לחלק את הנתונים, הם אימנו מודל kNN באמצעות שלושת הפיצ'רים הנומריים (טמפרטורה, לחות וכמות משקעים). בנוסף, כדי לבחון אם אחד הפיצ'רים מיותר, הם אימנו שלושה מודלים נוספים, כל אחד באמצעות שניים מאותם שלושה פיצ'רים (בכל אחד משלושת המודלים לא נעשה שימוש בפיצ'ר נומרי אחר). הם הופתעו לגלות שלמודל שלא משתמש בפיצ'ר Precipitation (כמות משקעים) יש דיוק דומה מאוד למודל שמשתמש בכל שלושת הפיצ'רים. תנו הסבר לתוצאה המפתיעה הזו והציעו דרך שתאפשר למודל להשתמש בפיצ'ר זה באופן אפקטיבי. (4 נק')

ניתן לראות שפיצ'ר המשקעים נמצא בסקאלה שונה לחלוטין מהפיצ'רים האחרים ולכן ללא scaling שיביא את הפיצ'רים לאותה הסקאלה הוא יתרום מעט מאוד למרחקים בין התצפיות (השכנים הקרובים ביותר). Scaling יפתור את הבעיה.
בנוסף לשימוש ב-scaling, רצוי גם להפוך את הפיצ'ר הזה לבינארי, כך שאם המשקעים גדולים מאפס, הוא יקבל את הערך 1 ואחרת אפס, שהרי המידע הרלוונטי שנמצא בפיצ'ר הוא רק האם היו בכלל משקעים. עם זאת, ללא scaling לא תהיה לכך השפעה.

ג. כעת דני ושרון מעוניינים לנסות למדל את מבנה הנתונים ולהבין אם ניתן למצוא דפוסים שיראו שיש קבוצות דומות של סוגי מזג אוויר דומים. באיזו שיטה מאלו שנלמדו בקורס כדאי להם להשתמש במקרה זה? (2 נק')

K-Means clustering

שאלה 6 (4 נק')

בטבלת הנתונים df הכוללת 1500 שורות, ישנו משתנה בשם var. הריצו את הקוד הבא על הטבלה:

```
num_repetitions = 1000
secret = []
for i in range(200):
    something = df.sample(n=100, replace=True)['var']
    secret.append(something.mean())

```

א. כמה מדגמים של df נדגמים סך הכל בקוד? (1 נק') _____200_____

ב. מהו גודל כל אחד מהמדגמים שמייצר הקוד? (1 נק') _____100_____

ג. מה מכיל המשתנה secret בסיום הרצת הקוד? (2 נק')

200 ממוצעים של מדגמים בגודל 100 שנלקחו עם החזרה מהמשתנה var בטבלת הנתונים df.

[נשים לב שלכתוב שאלו דגימות בוטסטראפ תהיה טעות בגלל גודל המדגם]

שאלה 7 (12 נק')

מדען נתונים בסטארט-אפ מפתח מודל למידת מכונה מפוקח (supervised learning) כדי לחזות אם אדם הוא נשא לגן (gene) נדיר שרק 2% מהאוכלוסייה נושאת. המודל משתמש במאפיינים שונים כגון גיל, מין, מוצא, לחץ דם, רמות כולסטרול ומדדי בריאות אחרים. לנשאות לגן ישנן השלכות בריאותיות ולכן העלות של החמצת נשאים (אמיתיים) לגן גבוהה בהרבה מהעלות של ביצוע בדיקות גנטיות נוספות למי שחשוד כנשא.

למודל שפיתח המדען יש את הנתונים הבאים: המודל מזהה את הגן (מחזיר תשובה חיובית) בקרב מי שהוא באמת נשא של הגן בהצלחה של 90%. עם זאת, המודל מחזיר תשובה חיובית גם אצל מי שלא נושא את הגן, ב-6% מהמקרים.

א. מנהלת צוות הפיתוח הציגה את נתוני המודל הנ"ל בפני ההנהלה וטענה שהנתונים האלה מראים שאין באמת צורך לערוך בדיקות נוספות לאדם שהמודל מנבא שנושא את הגן: בהינתן שהמודל מנבא שאדם נושא את הגן, הסבירות שהוא באמת נשא שלו גבוהה מאוד. איזו טעות מפגינה מנהלת צוות הפיתוח בטענתה זו (השתמשו במושג שנלמד בקורס)? (2 נק')

_____ הזנחת שיעור בסיס _____

ב. נניח שהמודל אומן על 1000 תצפיות שנלקחו מהאוכלוסייה. השלימו את מטריצת הבלבול (confusion matrix) שהתקבלה (אין צורך להראות חישובים): (4 נק')

	סיווג אמת: נשא	סיווג אמת: לא נשא
מודל מנבא: נשא	18	59 [הערה: במטריצת בלבול ערכים חייבים להיות שלמים]
מודל מנבא: לא נשא	2	921

ג. חשבו את המדדים Accuracy (דיוק), Precision, Recall של המודל. יש להציג את החישובים במונחים של TP, TN, FP, FN ולבסוף לתת תשובה מספרית. (3 נק')

Acc = (TP+TN)/(TP+TN+FP+FN) = (18+921)/1000 = 93.9%

Precision = TP/(TP+FP) = 18/(18+59) = 0.233

Recall = TP/(TP+FN) = 18/20 = 0.9

ד. איזה מהמדדים שחישבתם בסעיף ג' הוא החשוב ביותר, בהינתן הידוע לגבי העלויות של טעויות? נמקו ב-2 משפטים לכל היותר (3 נק')

Recall חשוב יותר מפני שנתון שחשוב שלא להחמיץ נשאים.

שאלה 8 (7 נק')

תמר היא מדענית נתונים בחברת כרטיסי אשראי שעשתה בשנה האחרונה שלושה קמפיינים גדולים לגיוס לקוחות: ביוני 2022, בנובמבר 2022 ובמאי 2023. קמפיינים לגיוס לקוחות נבחנים (בין היתר) על פי הזמן הממוצע שחולף בין תאריך ההצטרפות של לקוח/ה ועד לתאריך שבו הוא או היא עשו שימוש מצטבר של לפחות 10,000 שקלים בכרטיס. תמר רוצה להשוות בין אפקטיביות שלושת הקמפיינים על סמך מדד זה. לשם ההשוואה, היא משתמשת בנתונים על לקוחות שעשו שימוש מצטבר של לפחות 10,000 שקלים בכרטיס. הנתונים כוללים, לכל לקוח/ה, באיזה משלושת הקמפיינים הוא או היא גויסו ואת התאריך שבו השלים/ה שימוש מצטבר של 10,000 שקלים.

איזו הטייה הקיימת בנתונים אלה יכולה להשפיע על תוצאות המודל? הסבירו את מקור ההטייה ומהן ההשלכות האפשריות שלה. האם תוכלו לשער איזה משלושת הקמפיינים התגלה כאפקטיבי ביותר תחת המדד שבודקת תמר (ותחת ההנחה שבניתוח שלה היא לא מתחשבת בהטייה הקיימת בנתונים)?

נתונים קטומים מימין: כאשר מסתכלים רק על לקוחות שכבר השלימו רכישות מצטברות של 10000 שקלים, חסר לנו מידע על לקוחות שעוד ישלימו רכישה כזו בעתיד. לכן נתוני "זמן ממוצע בין גיוס להשלמת רכישות ב-10000 ₪" יהיו מוטים. בפרט, אם לקוח נמצא במאגר הנתונים וגויס במאי 2023, אז הזמן שחלף בין הגיוס שלו לבין תאריך השלמת השימוש הוא בהכרח קצר, בעוד שלקוח שגויס ביוני 2022 ונמצא במאגר, בהחלט יתכן שלקח לו זמן רב להשלים שימוש של 10000 שקלים. כלומר, בעוד שלא כל הלקוחות שגויסו במאי 2023 משלימים מהר רכישות ב-10000 שקל, אלו מביניהם שבנקודת הזמן הנוכחית כבר השלימו רכישות ב-10000 שקל, עשו זאת מהר. מכאן, סביר שהקמפיין האחרון הוא זה שמתגלה תחת המדד הזה כאפקטיבי ביותר.

בנוסף ניתן לכתוב על פרדוקס הבדיקה: ההסתברות של לקוח להיכלל בנתונים של תמר לגבי הזמן עד להשלמת רכישות ב-10,000 ₪ תלויה בזמן שחלף עד להשלמת רכישות ב-10,000 ₪, כך שככל שהזמן ארוך יותר כך יותר סביר שהלקוח נכלל בנתונים. הניתוח לגבי אורך זמן ממוצע לרכישה נשאר כנ"ל.

שאלה 9 (11 נק')

רוני וליטל עובדים כמנתחי נתונים בחברת פיתוח אפליקציות. בילי, מנהל הצוות, ביקש מכל אחד מהם, בנפרד ובאופן בלתי תלוי, לבנות רווח סמך לזמן השימוש הממוצע של משתמשים באפליקציה תחת פורמט ויזואלי חדש שהושק לאחרונה. כל אחד מהם לקח מדגם מקרי גדול של זמני שימוש תחת הפורמט החדש ובנה רווח סמך ברמת ביטחון של 90% לזמן השימוש הממוצע. רוני קיבל רווח סמך של [119, 171], וליטל קיבלה רווח סמך של [103, 152].

א. הסבירו בקצרה כיצד יתכן שהם קיבלו רווחי סמך שונים למרות שבחנו את אותו הפרמטר. (3 נק')

רווחי הסמך שבנו מתבססים על מדגמים מקריים שונים (בנוסף, אנחנו למדנו רק איך לבנות רווח סמך באמצעות בוטסטראפ שמוסיף עוד דרגת אקראיות שיכולה לייצר רווחי סמך שונים)

ב. בילי מנהל הצוות אמר שהוא מעוניין לבדוק, ברמת מובהקות של 0.01, את ההשערה שזמן השימוש הממוצע באפליקציה תחת הפורמט החדש שונה מזמן השימוש הממוצע בה תחת הפורמט הקודם. ידוע כי בפורמט הקודם זמן השימוש הממוצע היה 107 שניות. מי מהם, אם בכלל, יכול להסיק מסקנה רלוונטית בלי לבצע ניתוחים נוספים? הסבירו. (4 נק')

כדי לבדוק את ההשערה, צריכים להשתמש ברווח סמך ברמת ביטחון של 99% : אם הוא לא יכלול את הערך 107, אז נדחה את השערת האפס בר"מ 0.01. רווחי סמך ברמת ביטחון גבוהה יותר הם רחבים יותר ולכן רווח הסמך של ליטל, שכבר כשהוא ברמת ביטחון של 90% כולל את 107, בוודאות יכלול את הערך 107 כשיורחב לרמה של 99%. מכאן שליטל יכולה להסיק שאין לדחות את השערת האפס. לא ניתן לדעת מה יסיק רוני בלי לבצע את הניתוח במלואו.

ג. בילי החליט לבדוק בצורה אחרת אם יש שינוי בזמני השימוש באפליקציה : לכל משתמש/ת, הוא בדק אם זמן השימוש הממוצע שלו/ה גדל לאחר שינוי הפורמט. הנתונים הראו כי לרוב גדול מאוד של משתמשים זמן השימוש אכן גדל. בילי הסיק שהשינוי בפורמט הויזואלי של האפליקציה גרם להגדלה של זמן השימוש בה. חוו דעתכם על מסקנה זו והציעו הסבר חלופי לנתונים (4 נק').

בילי הסיק מסקנה סיבתית, אך לכל היותר הוא מצא קורלציה. הסבר חלופי לדוגמה הוא שיחד עם שינוי הפורמט הויזואלי השתנו דברים נוספים באפליקציה, או שפשוט משתמשים נוטים להשתמש לאורך זמן רב יותר באפליקציה ככל שהיא מותקנת להם יותר זמן בטלפון. [אפשר לקבל הסברים סבירים אחרים.]

שאלה 10 (15 נק')

בכל סעיף, הקיפו : נכון או לא נכון. נמקו תשובתכם במשפט אחד. (3 נק' כל סעיף)

א. באלגוריתם kNN, השימוש במספר שכנים k גדול יותר מוביל לבעייה הנקראת קללת המימד (Curse of dimensionality). נכון / **לא נכון**

נימוק (במשפט אחד) : קללת המימד קשורה למספר הפיצ'רים ולא למספר השכנים

ב. שיטת המרפק (elbow method) למציאת מספר האשכולות (clusters) ב-K-Means תמיד מוצאת את המספר האופטימלי של אשכולות. נכון / **לא נכון**

נימוק (במשפט אחד) : שיטת המרפק היא שיטה יוריסטית שלא ניתן לומר שתמיד מוצאת מספר אופטימלי של אשכולות.

ג. אלון ביצע בדיקת השערות וקיבל p-value של 0.03. לכן, הוא ידחה את השערת האפס ברמת מובהקות 0.01 אבל לא ברמת מובהקות 0.05. נכון / **לא נכון**

נימוק (במשפט אחד) : ההפך. הוא ידחה את השערת האפס בר"מ 0.05 אך לא בר"מ 0.01

ד. כשמשתמשים בשיטת K-fold cross validation, כל תצפית בנתונים משויכת בדיוק ל-fold אחד.
נכון / לא נכון

נימוק (במשפט אחד) : בשיטה זו מחלקים את הנתונים ל-K קבוצות mutually exclusive וכל תצפית משויכת בדיוק לפולד אחד. [ניתן לקבל תשובה "לא נכון" אם הנימוק הוא שישנן תצפיות בנתונים שמראש אינן נכללות בשיטה כי הן חלק מסט המבחן]

ה. אם למסווג מסוים יש ערך F1 של 1, אז הדיוק (accuracy) שלו על הנתונים הוא 100%.
נכון / לא נכון

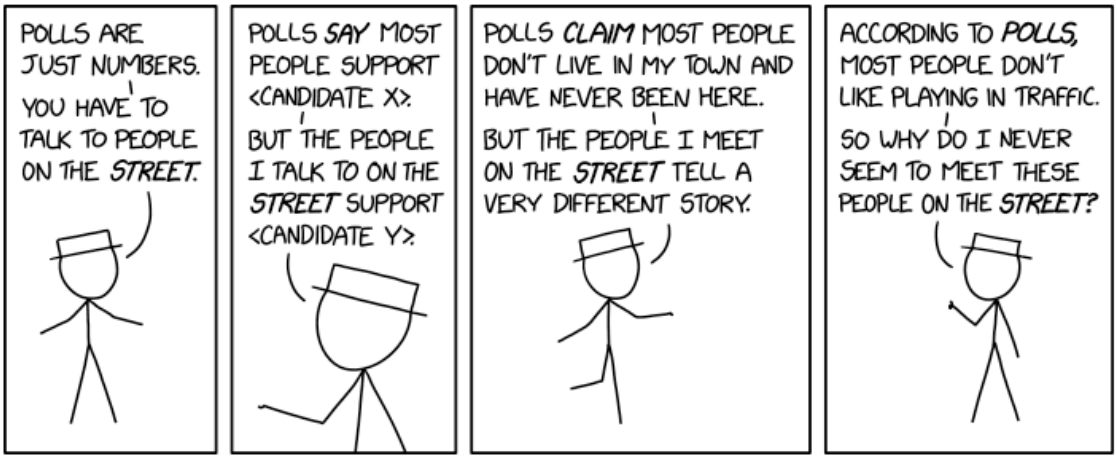
נימוק (במשפט אחד) : אם F1=1 אז precision=recall=1 ומכאן לפי הנוסחאות FP=FN=0.

כלומר המודל לא עושה אף טעות והדיוק שלו 100%.

שאלת בונוס (2 נק')

הסבירו בקצרה למה קריקטורה זו (אמורה להיות) מצחיקה. השתמשו במושגים שנלמדו בקורס.

[Polls = סקרים ; Candidate = מועמד ; Traffic = תנועה (בכביש)]



(מקור : xkcd)

הבחור בכובע טוען שהנתונים שלו "מהאנשים שהוא מדבר איתם ברחוב" משקפים טוב יותר את המציאות מאשר סקרים, אבל בבירור המדגם שלו הוא מדגם נוחות שסובל מהטיית בחירה – כפי שמתבלט למשל בבועית האחרונה שבה הוא אומר ש"האנשים שהוא פוגש ברחוב" הם אף פעם לא אלה שלא אוהבים להיות ברחוב.