

מבוא לניתוח נתונים - 094202
מבחן מועד א' – סמסטר אביב תשפ"א
טור א'

טיוטת פתרון לדוגמה

תאריך הבחינה: 15.07.2021

מרצים: עפרה עמיר, אורי פלונסקי

מתרגלים: רפאל שללה, זהר גלעד, אלכסנדר טואיסוב

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 בonus. 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 מילים מעבר לשורות המיועדות. מחברות הטייטה 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. בפרט, חובה להחזיר בסיום הבחינה את הטופס וכל מחברת בה השתמשתם.

בהצלחה!

שאלה 1 (15 נק')

הנהלת הטכניון מעוניינת לבדוק האם במסגרת הלמידה ההיברידית היו הבדלים בהערכות המרצים (סקר ההוראה) שניתנו על ידי סטודנטים שנכחו באופן פיזי בהרצאות והערכות המרצים שניתנו על ידי סטודנטים שצפו בהרצאות באופן סינכרוני בזום (ניתן להתעלם מסטודנטים שלא נכחו בהרצאה באופן סינכרוני). הניחו שלטכניון יש גישה לסקרי ההוראה ולזהות הסטודנטים שמילאו אותם (זהות הסטודנטים מותממת, כלומר עברה אנונימיזציה). בנוסף, נאספו מכל הקורסים הלוגים (logs) של הזום (כלומר שמות הסטודנטים שהתחברו לשיעור בזום באופן סינכרוני) והוצאו מהמודל (Moodle) הרישומים של הסטודנטים להגעה פיזית לכיתה (הנתונים מהמודל גם עברו התממה ויכולים להיות מוצלבים עם נתוני סקר ההוראה).

א. נסחו את השערת האפס וההשערה האלטרנטיבית (3 נק')

תשובה: השערת האפס: ממוצע הערכות הסטודנטים שנכחו באופן פיזי שווה לממוצע

הערכות הסטודנטים שצפו בהרצאות באופן סינכרוני בזום

השערה אלטרנטיבית: ממוצע הערכות הסטודנטים שנכחו באופן פיזי שונה מממוצע הערכות

הסטודנטים שצפו בהרצאות באופן סינכרוני בזום

ב. ציינו מהו הפרמטר שהייתם בודקים ומה יהיה סטטיסטי המבחן. (4 נק')

תשובה: פרמטר: ההפרש בין ממוצע דירוגי הסטודנטים שנכחו פיזית לבין ממוצע דירוג

הסטודנטים שצפו בזום בקרב אוכלוסיית כלל הסטודנטים בטכניון

סטטיסטי המבחן: ההפרש בין ממוצע דירוגי הסטודנטים שנכחו פיזית לבין ממוצע דירוג

הסטודנטים שצפו בזום במדגם (סטודנטים שמילאו את הסקר)

ג. תארו הטיית מדידה שעשויה להיות בנתונים שנאספו, וציינו את ההשלכות האפשריות שלה

על ניתוח הנתונים. (4 נק')

תשובה: דוגמה להטית מדידה: סטודנטים שנרשמו להגעה פיזית לכיתה ובסוף לא הגיעו

אלא צפו בזום. השלכות אפשריות – אומדן לא נכון של דירוגים של סטודנטים בקבוצות

השונות שיפגע במבחנים הסטטיסטיים.

ד. הניחו שבתום הניתוח נמצא כי ישנו הבדל מובהק בין ציוני סקר ההוראה שניתנו על ידי סטודנטים שנכחו באופן פיזי לבין סטודנטים שנכחו בזום, כך שסטודנטים שנכחו באופן פיזי דירגו בממוצע את המרצים בציוני הוראה גבוהים יותר מסטודנטים שנכחו באופן מקוון. על פי נתונים אלה, הנהלת הטכניון קבעה שביחס ללמידה מקוונת, למידה בקמפוס גורמת לשביעות רצון גבוהה יותר של הסטודנטים מסגל ההוראה ומהקורס. האם לדעתכם ניתן להסיק זאת (בהנחה שהניתוח בוצע באופן סטטיסטי נכון)? אם כן, הסבירו מדוע. אם לטענתכם לא ניתן להסיק זאת, הציעו הסבר חלופי. (4 נק')

תשובה: לא, מכיוון שקורלציה אינה גוררת סיבתיות. יכולים להיות מספר הסברים חלופיים – למשל, סטודנטים שאוהבים את שיטת ההרצאה של מרצה מסוים בוחרים להגיע לכיתה וסטודנטים שלא אוהבים את שיטת ההוראה של המרצה מתחברים בזום. לכן, באופן שיטתי ייראה כאילו ההוראה בכיתה גורמת לשביעות רצון גבוהה יותר בעוד שההפך הוא הנכון (שביעות רצון גבוהה מהוראה גורמת להגעה פיזית לכיתה).

שאלה 2 (5 נק')

חוקרים מצאו כי רמת האושר **יורדת** עם העלייה בגיל. בנוסף, נמצא כי בממוצע לאנשים מבוגרים יותר יש רמת אושר **גבוהה** יותר מאשר לאנשים צעירים. ניתן להסביר פרדוקס זה בכך שהנתונים קטומים מימין (right-censored data). הסבירו למה הכוונה ב-3 משפטים לכל היותר (תוך התייחסות לקשר בין אושר וגיל).

תשובה: נתונים קטומים מימין משמעותם שהערך הנמדד בנקודת זמן מסוימת נותן לנו רק חסם תחתון לערך האמיתי. בהקשר של אושר וגיל, ככל הנראה אנשים מאושרים יותר חיים זמן ארוך יותר ולכן למרות שרמת האושר יורדת עם הגיל, בעת המדידה לאנשים מבוגרים שנותרו בחיים (נתוני הגיל שלהם קטומים מימין) יש רמת אושר גבוהה יותר מאשר לאנשים צעירים (שרובם ככולם, מאושרים או לא, עדיין בחיים).

שאלה 3 (8 נק')

הגרף הבא לקוח ממאמר אשר חקר את הנהיגה של רופאים מהתמחויות שונות, בהסתמך על בסיס נתונים של דו"חות על נהיגה מהירה בין השנים 2004-2017. "מהירות קיצונית" הוגדרה במאמר כנהיגה במהירות הגבוהה ב-20 מייל לשעה או יותר מהמהירות המותרת בכביש בו ניתן הדו"ח. הגרף הבא מראה עבור כל תחום התמחות רפואית, את אחוז הדו"חות שניתנו על נהיגה קיצונית (מתוך כלל הדו"חות במדגם שניתנו עבר רופאים מתחום ההתמחות). למשל, מבין הדו"חות שניתנו לפסיכיאטרים (psychiatrists), 34% היו במקרים של עבירת מהירות "קיצונית".



א. ציינו בעיה מהותית הקיימת בגרף עצמו (ולא בנתונים המוצגים) והסבירו מדוע זו בעיה.

(צבעים דומים מדי אינה בעיה של הגרף אלא אילוז' של הדפסת המבחן) (3 נק')

תשובה: בעיה מרכזית – התחלת הציר לפני 0%, מה שמעוות את הפרופורציות בגרף

העמודות. (בנוסף, גם הציר של המכוניות מקשה לקרוא בבירור את היחסים ועומד בניגוד

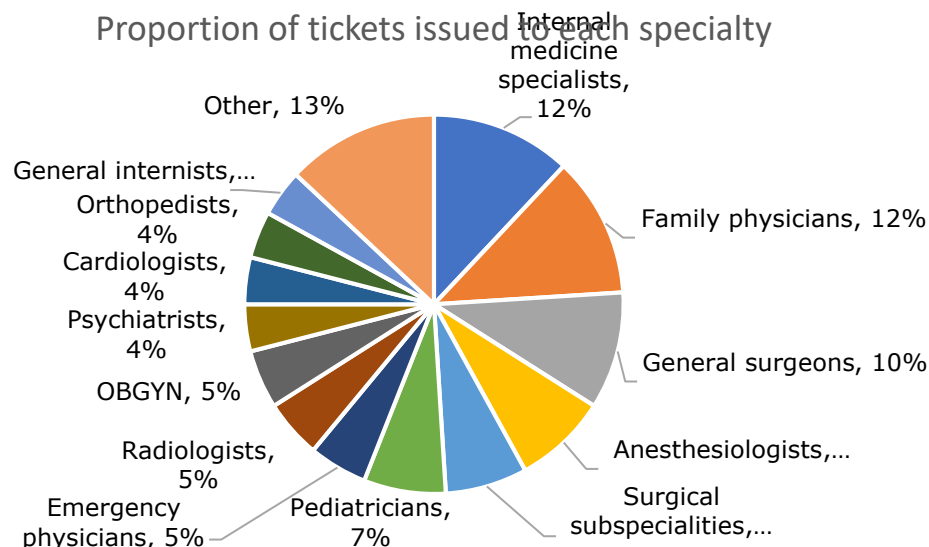
לעיקרון ה-proportional ink).

ב. הציעו דרך לתקן את הבעיה (הדגימו בשרטוט, לא צריך לכלול את כל הנתונים, רק הדגמה

שבין מה הרעיון) (2 נק') [מקום לאיור]

תשובה: ציור הגרף כך שהמספרים מתחילים מ-0 (ובעמודות רגילות, בלי מכוניות)

ג. הגרף הבא לקוח גם הוא מאותו המחקר, ומראה את אחוז דו"חות המהירות שחולקו לרופאים מההתמחויות השונות מתוך כלל הדו"חות שחולקו. מטרת הגרף היא להציג את הסיכוי של רופא/ה מכל אחת מההתמחויות לקבל דו"ח על עבירת מהירות ביחס לרופא/ה מההתמחויות אחרות. ציינו בעיה **בנתונים המוצגים בגרף** (ולא בויזואליזציה עצמה) אשר מקשה עלינו ללמוד על מידת הנהיגה המהירה של רופא/ה מהתמחות אחת לעומת רופא/ה מהתמחות אחרת. הציעו דרך לתקן את הנתונים המוצגים כדי להתגבר על בעיה זו. (3 נק')



תשובה: הבעיה המרכזית היא שלא ידוע כמה רופאים יש בכל התמחות, כלומר אין

התייחסות לשיעור הבסיס של כל קבוצת רופאים, ולכן אין משמעות לגרף. כדי לתקן זאת, צריך לנרמל את הנתונים בהתאם. למשל, כרגע נראה שרופאי משפחה אחראים לנתח גדול מדו"חות המהירות, אבל אם יש יותר רופאי משפחה מרדיולוגיים (למשל), זה לא מפתיע.

שאלה 4 (14 נק')

אתם עובדים באתר חדשותי גדול. ברשותכם נתוני גלישה באתר המכילים עבור כל כניסה לכתבה:

- מזהה הכתבה [מספר זיהוי ייחודי לכל כתבה]
- קטגוריית הכתבה [חדשות/ספורט/...]
- כתובת IP [כתובת הרשת של הגולש]
- תאריך כניסה
- שעת כניסה
- משך קריאה
- מנוי פרימיום [כן/לא]

- מזהה ייחודי למשתמש [רק עבור מנוי פרימיום]

לדוגמה שתי שורות מתוך הטבלה:

IP_address	article_id	article_category	date_enter_article	time_enter_article	time_on_article_seconds	premium_user	user_id
103.22.5.6	431587	sports	05/07/2021	10:03:22	72	True	reader123
212.13.8.12	552413	culture	05/07/2021	10:04:55	720	False	NaN

צוות ניתוח הנתונים מעוניין לאמן מסווג שיחזה אם קורא שאינו מנוי פרימיום (מזוהה על ידי כתובת IP) בחודש מסוים יירשם למינוי פרימיום בחודש שלאחר מכן.

א. הציעו 2 משתנים חדשים שתייצרו על בסיס הנתונים הגולמיים עבור כל משתמש, בהם תיעזרו בבעיית הסיווג. תארו בבירור איך יחושב המשתנה (לא צריך קוד) ומדוע אתם חושבים שהוא עשוי להיות שימושי. (6 נק')

תשובה: הרבה תשובות אפשריות, למשל:

- משך קריאה כולל של כתבות בחודש. יחושב על ידי חישוב סכום העמודה של

time_on_article_seconds לאורך כל חודש לכל כתובת IP

- כמות הכתבות הנקראות בחודש. יחושב על ידי מספר הערכים הייחודיים של עמודת

article_id בכל חודש

שני המשתנים הללו עשויים לסייע מכיוון שסביר לחשוב שמשתמשים שמתעמקים יותר בכתבות, כמו גם משתמשים אשר קוראים יותר כתבות באתר באופן כללי, יהיו בסבירות גבוהה יותר לעשות מנוי פרימיום.

ב. הניחו שלאחד מהמשתנים שייצרתם יש קורלציה של 0.8- עם המשתנה של מנוי פרימיום (הניחו ש-True מיוצג על ידי הערך 1 ו-False על ידי 0). למשתנה נוסף יש קורלציה של 0.1 עם המשתנה של מנוי פרימיום. באיזה מהם הייתם מעדיפים להשתמש לצורך בעיית הסיווג? הסבירו מדוע במשפט אחד. (3 נק')

תשובה: המשתנה עם הערך מינוס 0.8, כיוון שהוא בעל קשר חזק יותר למשתנה המטרה.

ג. צוות ניתוח הנתונים של האתר הציע לחלק את הנתונים שנאספו בין החודשים ינואר-יוני 2021 באקראי, כך ש-80% ישמשו לאימון ו-20% לבדיקה. ציינו בעיה בחלוקה זו והציעו חלוקה חלופית לסט בדיקה וסט אימון. (5 נק')

תשובה: הבעיה היא data leakage פוטנציאלי מנתונים עתידיים של משתמש אל העבר.

פתרון אפשרי הוא לחלק את סט האימון והבדיקה לפי משתמשים שונים.

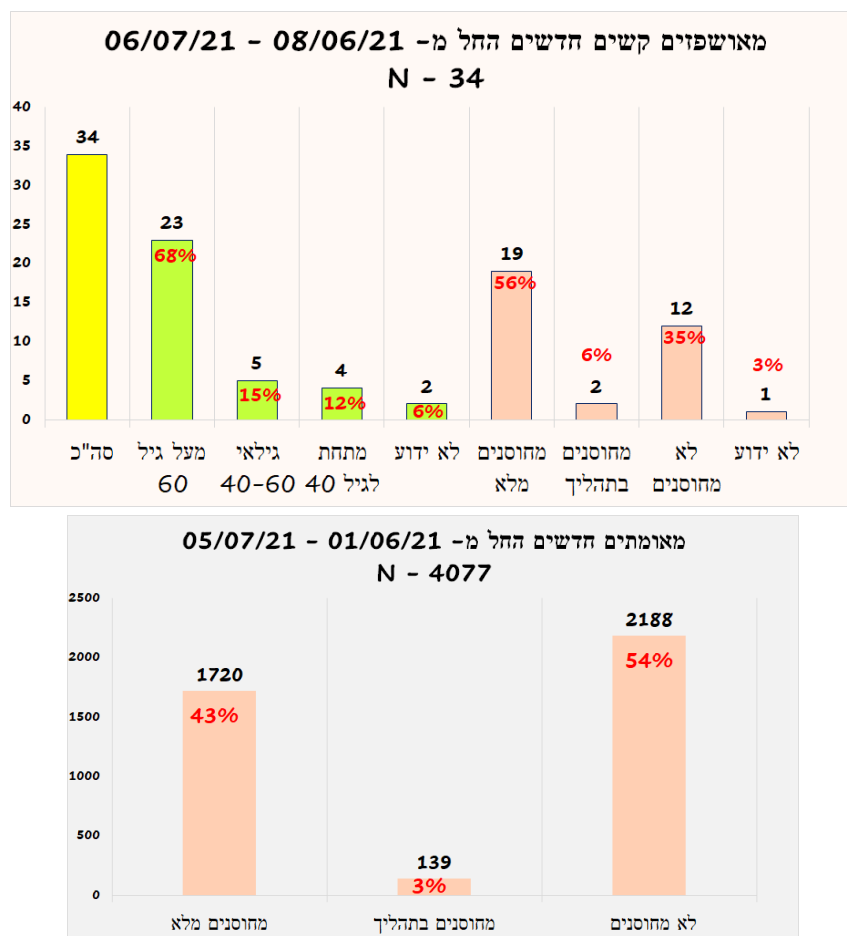
שאלה 5 (4 נק')

תמר חישה רוח סמך ברמת ביטחון 95% לממוצע באוכלוסייה בהתבסס על מדגם אקראי וגדול. רוח הסמך שהתקבל הוא [62.4, 71.9]. לאחר מכן אילנית קיבלה גישה לנתונים על האוכלוסייה כולה ומצאה כי הממוצע באוכלוסייה הוא 73.5. הסבירו כיצד זה יתכן (בהנחה שתמר חישה את רוח הסמך נכון ואילנית חישה את הממוצע נכון)

תשובה: רוח הסמך (שתמר חישה) הינו ברמת ביטחון 95% ולעיתים הוא אינו כולל את הערך האמיתי של הפרמטר (שאילנית חישה)

שאלה 6 (8 נק')

מואיז פרסם את הגרפים הבאים:



א. בתגובה, אפרת כתבה כי מאחר שיש יותר מאושפזים קשים "מחוסנים מלא" מאשר "מחוסנים בתהליך", הגרף מראה שעדיף להיות "מחוסן בתהליך" מאשר "מחוסן מלא". הסבירו מדוע לא ניתן להסיק מסקנה כזו מהגרף והציעו הסבר חלופי לנתונים. (4 נק')

תשובה: זו דוגמה להזנחת שיעור הבסיס. בעת המדידה ישנם באוכלוסייה הרבה יותר מחוסנים מלא מאשר מחוסנים בתהליך ולכן בפרט, ישנם גם הרבה יותר מאושפזים המחוסנים מלא מאשר מאושפזים שעדיין בתהליך חיסון.

ב. יוסף כתב שמאחר ששיעור המחוסנים מלא מתוך כלל המאומתים החדשים נמוך יותר משיעור המחוסנים מלא מתוך כלל המאושפזים הקשים, הנתונים מראים שהחיסון אינו אפקטיבי במניעת מחלה קשה. הסבירו מדוע לא ניתן להסיק מסקנה זו מהגרפים והציעו הסבר חלופי. (4 נק')

תשובה: הגיל הממוצע של המחוסנים מלא (ובכלל התפלגות הגיל שלהם) גבוה מהגיל הממוצע של הלא מחוסנים (ומהתפלגות הגיל שלהם). מאחר שמבוגרים נוטים לפתח מחלה קשה יותר, הנתונים לא מהווים עדות לכך שהחיסון לא אפקטיבי במניעת מחלה קשה.

שאלה 7 (6 נק')

בסט נתונים מסוים יש 100 תצפיות, מתוכן ל-75 יש תגית "כחול" ול-25 תגית "לבן". אליהו חילק את הנתונים לסט אימון עם 80 תצפיות ולסט מבחן עם 20 תצפיות. הוא הריץ אלגוריתם kNN על סט האימון ומצא שה-k הטוב ביותר עבור הנתונים שלו בסט האימון הוא 11. בסט האימון, דיוק האלגוריתם של אליהו הוא $93.4\% = 75/80$.

כאשר אליהו בחן את דיוק האלגוריתם על סט המבחן הוא מצא להפתעתו שהדיוק שקיבל בסט המבחן הוא בדיוק 0 (אפס). כלומר, אף תצפית מסט המבחן לא סווגה נכון. הסבירו ב-4 משפטים לכל היותר כיצד זה יתכן, מה סביר שקרה לאליהו ומה הייתם ממליצים לו לעשות.

תשובה: כנראה שסט האימון היה ממוין לפי הצבע והחלוקה שביצע אליהו היתה לפי הסדר של התצפיות בנתונים כך שלקח את כל התצפיות הכחולות ורק 5 מהלבנות. כאשר $K=11$ ויש רק 5 תצפיות לבנות אז כל סיווג של תצפית יהיה כחול ולכן הדיוק של האלגוריתם על סט האימון יהיה $75/80$ אבל הדיוק על סט המבחן יהיה 0. היינו ממליצים לעשות לנתונים shuffle ולהריץ מחדש.

שאלה 8 (13 נק')

הפלט של הפקודה `my_df.describe()` עבור סט נתונים מסוים הוא:

	variable1	variable2	variable3	target
count	3882.000000	3882.000000	3.882000e+03	3882.000000
mean	51.134982	109.883823	3.872725e+07	0.494333
std	69.851621	22.721818	4.792372e+07	0.500032
min	0.000000	34.000000	2.180000e+02	0.000000
25%	7.000000	95.000000	1.000000e+07	0.000000
50%	28.000000	106.000000	2.400000e+07	0.000000
75%	65.750000	120.000000	5.000000e+07	1.000000
max	761.000000	330.000000	7.000000e+08	1.000000

א. מהו ערך הרבעון השלישי של המשתנה variable2? (2 נק')

תשובה: 120

[השאלה ממשיכה בעמוד הבא]

מורן הריצה את הקוד הבא על מנת לחשב רווח סמך ברמת בטחון 95% לממוצע של המשתנה

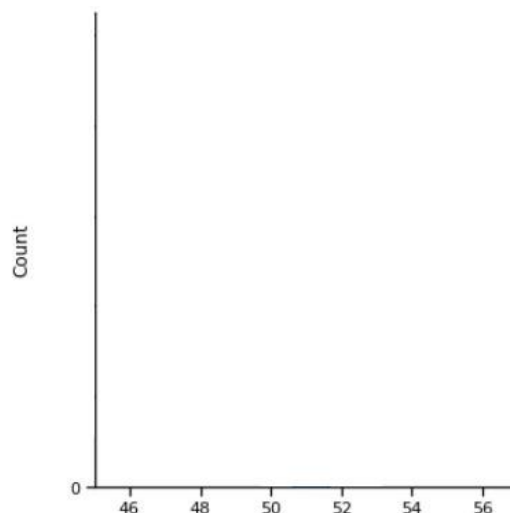
variable1 באוכלוסייה:

```
1 def bootstrap_mean(original_sample, column_name, num_replications):
2     '''This function returns an array of bootstrapped sample averages:
3     original_sample: df containing the original sample
4     column_name: name of column containing the variable of interest
5     num_replications: number of bootstrap samples'''
6     original_sample_size = original_sample.shape[0]
7     original_sample_var_of_interest = original_sample[[column_name]]
8     bstrap_means = np.empty(num_replications)
9     for i in range(num_replications):
10         bootstrap_sample = original_sample_var_of_interest.sample(original_sample_size, replace=False)
11         resampled_mean = bootstrap_sample.mean()
12         bstrap_means[i] = resampled_mean
13
14     return bstrap_means
15
16 means_bootstrapped = bootstrap_mean(my_df, 'variable1', 5000)
17 ax=sns.displot(means_bootstrapped)
18 ax.set(xlim=(45,57))
```

ב. השלימו את ההיסטוגרמה שתתקבל כפלט (כלומר הפלט שיודפס למסך מקוד זה), כולל

השלמת הערכים על ציר Y. רמז: ישנה טעות בקוד של מורן. (4 נק')

תשובה: גרף בעל עמודה יחידה הנמצאת ב- $x=51.135$ ומגיעה בציר Y ל-5000.



ג. מהו רווח הסמך שתקבל מורן כאשר תחשב אותו על סמך הקוד הנתון לעיל (כלומר הקוד

שבו יש טעות)? (3 נק')

תשובה: [51.135, 51.135]

ד. איתי בונה מסווג שינבא את ערכו של המשתנה target באמצעות המשתנים variable1

variable2, variable3. לשם כך חילק את הנתונים לסט אימון וסט מבחן וביצע את הליך ה-

cross-validation בצורה נכונה. מהנתונים לעיל, ציינו בעיה אחת שאיתי עלול להתקל בה

ללא עיבוד נוסף של הנתונים לפני הליך ה-cross-validation. (4 נק')

תשובה: על איתי לבצע scaling לנתונים שכן variable3 נמצא בסקלה אחרת לחלוטין (עשרות ומאות מיליונים) מאשר המשתנים האחרים.

שאלה 9 (12 נק')

ארגון למען ילדים בסיכון החליט לייצר אלגוריתם שינבא על בסיס נתונים מדו"חות של עובדי רווחה האם הילד או הילדה בסיכון. לאחר שעובדי הארגון אימנו את המודל, הם בחנו אותו על סט מבחן של 1000 תצפיות. להלן הנתונים שהם חישבו עבור האלגוריתם:

- דיוק (accuracy) 0.89
- רגישות (sensitivity) 0.2

א. השלימו את מטריצת הבלבול (8 נק):

סה"כ	תגית אמיתית: לא בסיכון	תגית אמיתית: בסיכון	
50	30	20	ניבוי המודל: בסיכון
950	870	80	ניבוי המודל: לא בסיכון
1000	900	100	סה"כ

ב. האם הייתם ממליצים לארגון להשתמש במודל זה כדי לנבא האם ילדים חדשים (שאינ עליהם נתונים) נמצאים בסיכון? הסבירו ב-3 משפטים לכל היותר. (4 נק')

תשובה: לא היינו ממליצים מפני שלמודל יש רגישות נמוכה מאוד וגם סגוליות נמוכה כך שהמודל גרוע יחסית בתפיסת ילדים בסיכון וגם גרוע יחסית באבחון האם ילד הוא בסיכון (ניתן בהחלט לקבל תשובה אחרת אם יש נימוק הגיוני למשל על מודל אלטרנטיבי שהוא גרוע יותר)

שאלה 10 (15 נק')

- ענו נכון או לא נכון. נמקו תשובתכם במשפט אחד.
- א. ההנחה העומדת בבסיס שיטת bootstrap היא שההתפלגות של המדגם מייצגת בקירוב טוב את ההתפלגות של הסטטיסטי תחת השערת האפס. נכון / לא נכון.
- נימוק (במשפט אחד): ההנחה העומדת בבסיס השיטה היא שההתפלגות של המדגם מייצגת בקירוב טוב את ההתפלגות של האוכלוסייה.
- ב. אם הדיוק (accuracy) של מסווג א' גבוה מהדיוק (accuracy) של מסווג ב', ה-true positive rate (TPR) של מסווג א' בהכרח גבוה מה-true positive rate (TPR) של מסווג ב'. נכון / לא נכון.
- נימוק (במשפט אחד): כנימוק, ניתן להראות דוגמה נגדית או להראות לפי הנוסחאות של המדדים האלה.
- ג. כשמפעילים על סט נתונים מסוים 5-fold cross validation, משתמשים בכל אחת מהתצפיות לאימון המודלים הנבחרים בדיוק 4 פעמים. נכון / לא נכון.
- נימוק (במשפט אחד): ב-K-fold CV מאמנים את המודל K פעמים, אחד לכל פולד, כך שכל תצפית משמשת לאימון K-1 פעמים ולמבחן פעם אחת.
- ד. בגרף קו שמייצג שינוי, התחלת ציר ה-Y מאפס לא הכרחית ואף עלולה להטעות. נכון / לא נכון.
- נימוק (במשפט אחד): בגרף קו (שלא כמו בגרף עמודות) אין הכרח להתחיל את ציר ה-Y מאפס והתחלה מאפס אף עלולה לעוות את תפיסת השינוי של הצופה בגרף.
- ה. בידינו נתונים על כל ציוני המבחן במבוא לניתוח נתונים בטכניון בסמסטר אביב תשפ"א. דרך אחת סבירה להעריך את הממוצע במבחן היא לקחת דגימת בוטסטראפ מתוך נתונים אלו ולחשב לממוצע רווח סמך ברמת ביטחון 95%. נכון / לא נכון.
- נימוק (במשפט אחד): בידינו נתונים על כלל האוכלוסייה ולכן אין להשתמש בשום שיטה לאמידת הממוצע (ובפרט לא בוטסטראפ) אלא לחשב אותו ישירות.

שאלת בונוס (2 נק')

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

הדובר בתמונה עושה טעות של הזנחת שיעור בסיס (של אחוז הימניים באוכלוסייה) כאשר הוא מדבר על אחוז הטעויות של הזנחת שיעור בסיס

