<u>מבוא לניתוח נתונים - 094202</u>

<u>מבחן מועד ב' – סמסטר אביב תשפ"א</u>

טור א'

תאריך הבחינה: 17.10.2021

מרצים: עפרה עמיר, אורי פלונסקי

מתרגלים: רפאל שללה, זהר גלעד, אלכס טואיסוב

הוראות:

1. לרשותכם **שלוש שעות** לפתור את הבחינה.
2. מותר להשתמש במחשבון <u>פשוט</u> בלבד. אין להשתמש בכל חומר עזר אחר.
3. הבחינה כוללת 14 עמודים (כולל עמוד זה) ובהם 11 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 מילים 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. בפרט, חובה להחזיר בסיום הבחינה את הטופס וכל מחברת בה השתמשתם.
9. יש לכתוב בעט כחול או שחור בלבד. כתיבה בעפרון מבטלת את זכות הערעור על הבחינה.

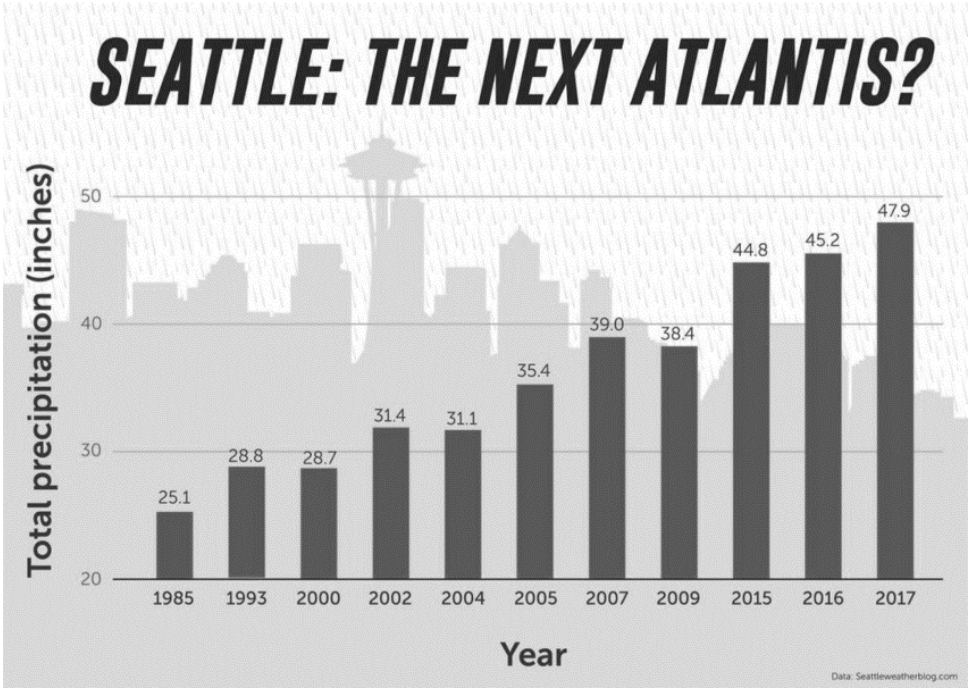
בהצלחה!

שאלה	ניקוד
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
בונוס	
סה"כ	

ת"ז: _____

<u>שאלה 1 (6 נק')</u>

נתונה הויזאוליזציה הבאה (precipitation = משקעים):



א. פרטו בנוגע לשתי בעיות לפחות בויזואליזציה המטעות או מבלבלות את הקורא. (4 נק')

_____

_____

_____

_____

_____

ב. ציירו מחדש את הויזואליזציה כך שלא תהיה מטעה (אין צורך לדייק, רק שנבין מה הרעיון) (2 נק') [מקום לאיור]

ת"ז: _____

שאלה 2 (13 נק')

תושבי שכונת נווה שאנן התלוננו על גידול במספר חזירי הבר הפוקדים את השכונה. עיריית חיפה שוקלת לקשור את הפחים על מנת למנוע מחזירי הבר לאכול מהם, ובכך להפחית את נוכחותם בשכונות. על מנת לבדוק אם קשירת הפחים אפקטיבית, העירייה החליטה לערוך ניסוי ולקשור את הפחים רק בחלק מהרחובות בשכונה. לרשות העירייה עומד תקציב המאפשר לקשור את הפחים ב-10 רחובות. העירייה בחרה לערוך את הניסוי בעשרת הרחובות שמהן הגיעו הכי הרבה תלונות על חזירים וקשרה ברחובות אלו את הפחים. הפחים בשאר הרחובות בשכונה לא נקשרו. הניסוי נערך לאורך חודש יוני. בסוף החודש, אספו את מספר התלונות שהתקבלו על חזירי בר ברחובות השונים. ברשות העירייה גם מספר התלונות שהתקבלו מתושבים בכל רחוב בחודש שלפני כן (חודש מאי). עיריית חיפה ביקשה את עזרתכם בניתוח הנתונים.

א. רשמו באיזה מדד (על בסיס הנתונים שנאספו) הייתם משתמשים כדי לבחון את יעילות קשירת הפחים. כלומר, מה הפרמטר שתרצו להעריך. (2 נק')

_____

_____

_____

ב. רשמו את השערת האפס וההשערה האלטרנטיבית וציינו מה יהיה סטטיסטי המבחן. (4 נק')

_____

_____

_____

_____

_____

ג. האם לדעתכם הבחירה בקבוצת הניסוי (10 הרחובות בהם נקשרו הפחים) טובה? אם כן, הסבירו מדוע. אם לא, הציעו דרך אחרת להחליט באיזה רחובות לקשור את הפחים. (4 נק')

_____

_____

_____

_____

_____

[השאלה ממשיכה בעמוד הבא]

ת"ז: _____

ד. בתקופות הסגרים התושבים התלוננו יותר מהרגיל על גידול במספר חזירי הבר. נניח כי לא היה באמת שינוי בביקורי חזירי הבר בשכונה. הציעו הסבר סטטיסטי לתחושת התושבים כי נוכחות חזירי הבר עלתה. הסבירו ב-3 משפטים לכל היותר. (3 נק')

_____

_____

_____

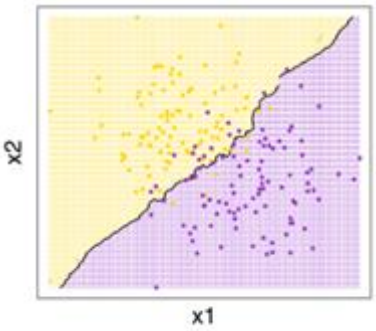
_____

_____

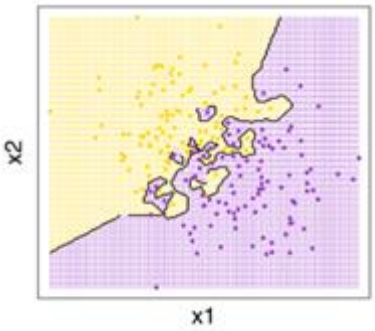
<u>שאלה 3 (4 נק')</u>

להלן שלושה גרפים של תוצאות הסיווג של אלגוריתם kNN על אותם נתונים, כאשר הסיווג הוא בין שתי מחלקות. הנקודות הצהובות מייצגות דוגמאות ממחלקה א' והנקודות הסגולות מייצגות דוגמאות ממחלקה ב'. הצבעים ברקע מייצגים את משטח ההחלטה (decision boundary) בין המחלקות.
ציינו באיזה מהמקרים (A/B/C) השתמשו בערך גבוה יותר של k ובאיזה השתמשו בערך נמוך יותר? הסבירו את תשובתכם ב-3 משפטים לכל היותר.

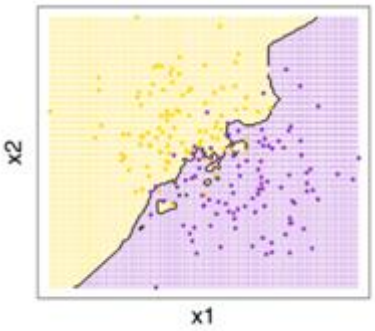
A



B



C



_____

_____

_____

_____

ת"ז: _____

<u>שאלה 4 (19 נק')</u>

מרק ושריל מנסים לזהות פרופילים מזויפים בפייסבוק. לצורך כך הם אספו נתונים על פרופילים רגילים ועל פרופילים שידוע שהם מזויפים, ואימנו מסווג שמטרתו לחזות אם פרופיל מסוים הוא מזויף. בסט המבחן היו 100 פרופילים, מתוכם 20 מזויפים ו-80 רגילים.

המסווג של מרק סיווג 15 מתוך 20 הפרופילים המזויפים כמזויפים ו-60 מתוך ה-80 הנותרים כלא מזויפים.

המסווג של שריל סיווג 10 מתוך 20 הפרופילים המזויפים כמזויפים, ו-70 מתוך ה-80 הנותרים כלא מזויפים.

א. צרו לכל אחד מהמסווגים את מטריצת הבלבול שלו. (8 נק') [מקום למטריצות]

ב. עבור כל אחד מהמסווגים, חשבו את ה-precision ואת ה-recall שלו. (4 נק')

_____

_____

_____

_____

ג. בהנחה שהמודל משמש על מנת לזהות פרופילים חשודים כמזויפים שלאחר מכן נבדקים ידנית על ידי בן אדם כדי לזהות האם הם אכן מזויפים, באיזה מסווג הייתם בוחרים? מדוע? (2 נק')

_____

_____

_____

_____

[השאלה ממשיכה בעמוד הבא]

ת"ז: _____

ד. גיליתם שמרק לא הפריד בין סט האימון לסט הבדיקה. כלומר, המודל שלו כולו אומן על כל הנתונים שנאספו. מה הבעיה שעלולה להיות עם תוצאות המסווג במקרה זה? הסבירו במשפט אחד. (3 נק')

_____

_____

_____

ה. בעת אימון המסווגים הראשוניים, היו חסרים למרק ושריל נתונים על אחד המשתנים בהם רצו להשתמש. כעת השיגו אותם ואימנו מחדש את המסווגים. האם דיוק המסווגים בהכרח ישתפר בזכות הוספת הנתונים החסרים? הסבירו. (2 נק')

_____

_____

_____

<u>שאלה 5 (5 נק')</u>

להלן טבלה של אחוז בדיקות הקורונה החיוביות של הנכנסים דרך כל אחד משלושת מעברי הגבול הפתוחים של מדינת ישראל וכן עבור כלל הנכנסים לישראל דרך מעברים אלו, עבור שני תאריכים:

	אחוז בדיקות חיוביות	
מעבר גבול	1.7.2021	8.7.2021
נתב"ג	1.4%	1.7%
סיני	0.4%	0.7%
ירדן	0.2%	0.9%
כלל הנכנסים לישראל	**1.3%**	**1.2%**

כלומר, בעוד שאחוז הבדיקות החיוביות עלה בכל אחד ממעברי הגבול בין שני התאריכים, אחוז הבדיקות החיוביות הכללי ירד. הסבירו כיצד תופעה זו הגיונית, בהתבסס על הנלמד בקורס.

_____

_____

_____

_____

_____

ת"ז: _____

שאלה 6 (5 נק')

מירב המנהלת חנות מסחר אלקטרוני מעוניינת לבנות מודל שינבא, לכל לקוח חדש, האם הוא צפוי לבצע קנייה נוספת בתוך חודש מיום הקנייה הראשונה. ברשותה 400 תצפיות. כל תצפית מייצגת נתונים של לקוח חדש יחיד הכוללים, בין היתר, את גיל הלקוח ומידע האם הלקוח ביצע רכישה נוספת בתוך חודש מיום הקנייה הראשונה. עם זאת, הנתונים על גיל הלקוח חסרים עבור 180 מהתצפיות (ניתן להניח שאין נתונים חסרים אחרים). כדי לפתור בעיה זו, מירב מחליטה לבצע השלמת נתונים (imputation) למשתנה "גיל" כך שבמקום הנתון החסר יוכנס הגיל החציוני של 220 הלקוחות שעבורם קיים נתון זה. לאחר השלמת הנתונים, היא מחלקת את הנתונים לסט אימון וסט מבחן בצורה רגילה, בונה את אלגוריתם הניבוי כפי שנלמד בקורס, ומקבלת דיוק של 90% על סט המבחן.

הסבירו מהי הבעיה בצורה שבה הכינה מירב את הנתונים לפני בניית האלגוריתם ומדוע עולה חשש שהדיוק שלו על נתונים חדשים יהיה נמוך משמעותית מ-90%.

שימו לב: עצם השימוש בחציון להשלמת נתונים חסרים אינו בהכרח בעייתי לכשעצמו.

_____

_____

_____

_____

שאלה 7 (13 נק')

ברשותך מדגם מקרי של ספרים מתוך מאגר נתונים של חנות ספרים הכולל מידע על למעלה מ-300,000 ספרים. להלן תיאור הנתונים במדגם:

```
books.describe(include='all')
```

	title	author	binding	num_pages	rating	rating_count
count	37545	37545	37545	37545.000000	37545.000000	37545.000000
unique	35225	21025	2	NaN	NaN	NaN
top	Selected Poems	Agatha Christie	Paperback	NaN	NaN	NaN
freq	14	64	26358	NaN	NaN	NaN
mean	NaN	NaN	NaN	335.430284	3.996210	10501.922946
std	NaN	NaN	NaN	202.309382	0.333102	19381.937880
min	NaN	NaN	NaN	0.000000	0.000000	0.000000
25%	NaN	NaN	NaN	222.000000	3.800000	509.000000
50%	NaN	NaN	NaN	309.000000	4.000000	2775.000000
75%	NaN	NaN	NaN	400.000000	4.200000	10280.000000
max	NaN	NaN	NaN	3420.000000	5.000000	123372.000000

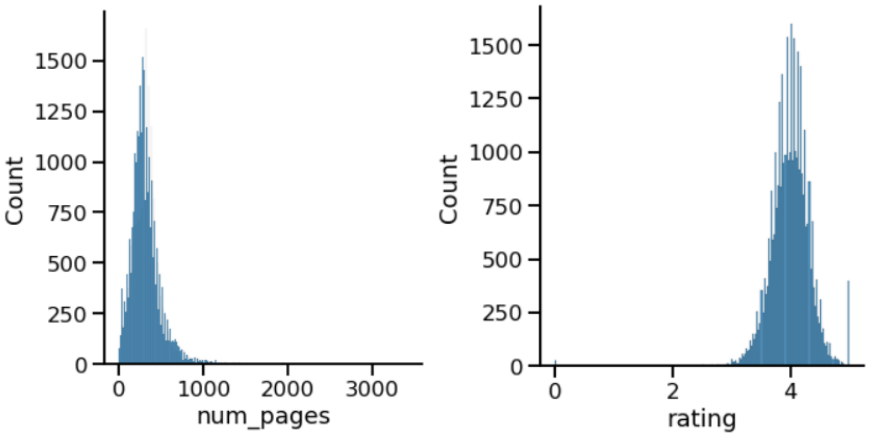
[השאלה ממשיכה בעמוד הבא]

ת"ז: _____

א. אם נחליט למחוק את כל הנתונים על הספרים שיש בהם יותר מ-1000 עמודים, האם חציון מספר העמודים (num_pages) שבנתונים שיישארו יגדל, יקטן, לא ישתנה או שלא ניתן לדעת? (1 נק')

_____

ב. נתונות שתי ההיסטוגרמות הבאות של שניים מהמשתנים:



נכון או לא נכון: מנתונים אלו ניתן להסיק שיש קורלציה שלילית בין מספר העמודים בספר (num_pages) לבין הציון (rating) שלו. נמקו במשפט אחד. (2 נק')

_____

_____

ג. מוכרת הספרים מעוניינת לדעת האם יש הבדל בין הציון (rating) של ספרים בכריכה קשה וספרים בכריכה רכה, כלומר לפי הערכים במשתנה binding. רשמו את השערת האפס ואת ההשערה האלטרנטיבית. (2 נק')

_____

_____

_____

ד. כדי לבחון את ההשערה מסעיף ג', מוכרת הספרים מחליטה להשתמש בהליך בוטסטראפ עם 5000 חזרות (replications) כדי למצוא רווח סמך ברמת בטחון 95%. מהן ההנחות המרכזיות שעליה להניח כדי שלתוצאה שתתקבל תהיה משמעות? (1-2 משפטים) (2 נק')

_____

_____

_____

[השאלה ממשיכה בעמוד הבא]

ת"ז: _____

ה. הסבירו (במילים ובלי לכתוב קוד) מה הקוד של מוכרת הספרים אמור לעשות בכל חזרה (replication) של הליך הבוטסטראפ. (2 נק')

_____

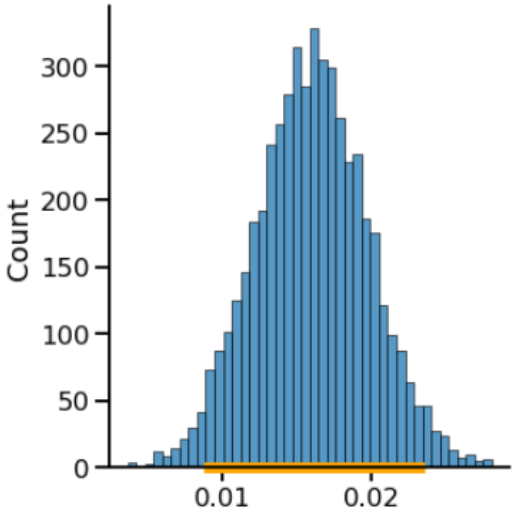
_____

_____

_____

ו. להלן הפלט שקיבלה מוכרת הספרים. בהנחה שהיא ביצעה את התהליך בצורה נכונה, מה מייצג ציר X? (2 נק')

```
The 95% boostsrap confidence interval is [0.008681716800374772, 0.02360655494546693]
```



_____

_____

ז. לפי הפלט מסעיף ו', מהי מסקנתה של מוכרת הספרים? (2 נק')

_____

_____

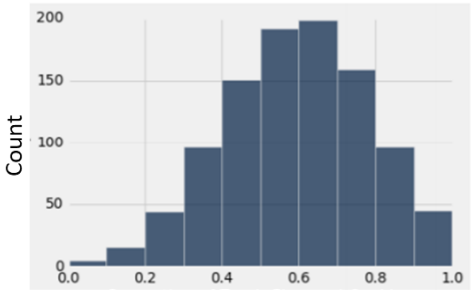
_____

ת"ז: _____

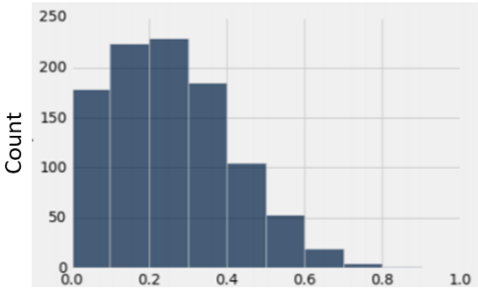
<u>שאלה 8 (9 נק')</u>

המפעל של עוגיפלצת מייצר קופסאות עוגיות אשר בהן 20% מהעוגיות הן בטעם אשפה (trash-flavored) ו-80% מהעוגיות הן בטעם שוקולד. בקופסת העוגיות שלכם יש 10 עוגיות, מתוכן 4 עוגיות בטעם אשפה, אבל אינכם זוכרים אם קניתם את הקופסה מהמפעל של עוגיפלצת. החלטתם לבצע מבחן סטטיסטי כדי להחליט אם הקופסה הגיעה מהמפעל של עוגיפלצת. **דחיתם את השערת האפס ברמת מובהקות של 0.05**. אחת מההיסטוגרמות הבאות מייצגת את ההתפלגות של **סטטיסטי המבחן** תחת השערת האפס.

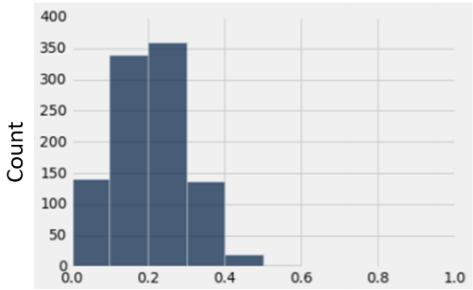
היסטוגרמה א':



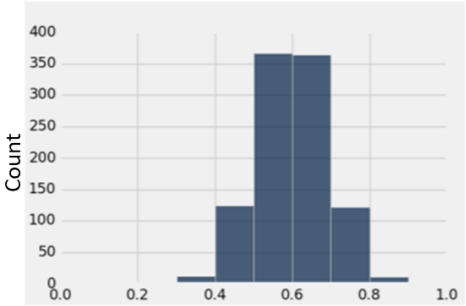
היסטוגרמה ב':



היסטוגרמה ג':



היסטוגרמה ד':



לכל אחת מההיסטוגרמות **שלא** מייצגת את ההתפלגות של סטטיסטי המבחן, הסבירו מדוע היא <u>אינה יכולה</u> לייצג את ההתפלגות. ציינו בבירור את שם ההיסטוגרמה אליה אתם מתייחסים

_____

_____

_____

_____

_____

_____

_____

ת"ז: _____

שאלה 9 (5 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 מנויי 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

א. לצורך הבנת קהל הלקוחות של האתר, צוות ניתוח הנתונים של האתר החליט לבצע קלאסטרינג (חלוקה לאשכולות). הם אינם בטוחים כיצד לקודד את קטגוריית הכתבה לערכים מספריים לפני ביצוע K-Means. כיצד הייתם ממליצים לבצע את הקידוד ומדוע? הניחו שיש 4 קטגוריות - חדשות, ספורט, תרבות, כלכלה. הראו את הקידוד בו הייתם משתמשים. (3 נק')

_____

_____

_____

_____

ב. צוות ניתוח הנתונים הריץ את אלגוריתם K-Means עם K=3. הם הריצו את האלגוריתם (עם אותו K) חמש פעמים, ובכל פעם קיבלו חלוקה מעט שונה. כיצד הייתם ממליצים להם לבחור באיזה חלוקה לאשכולות להשתמש? (2 נק')

_____

_____

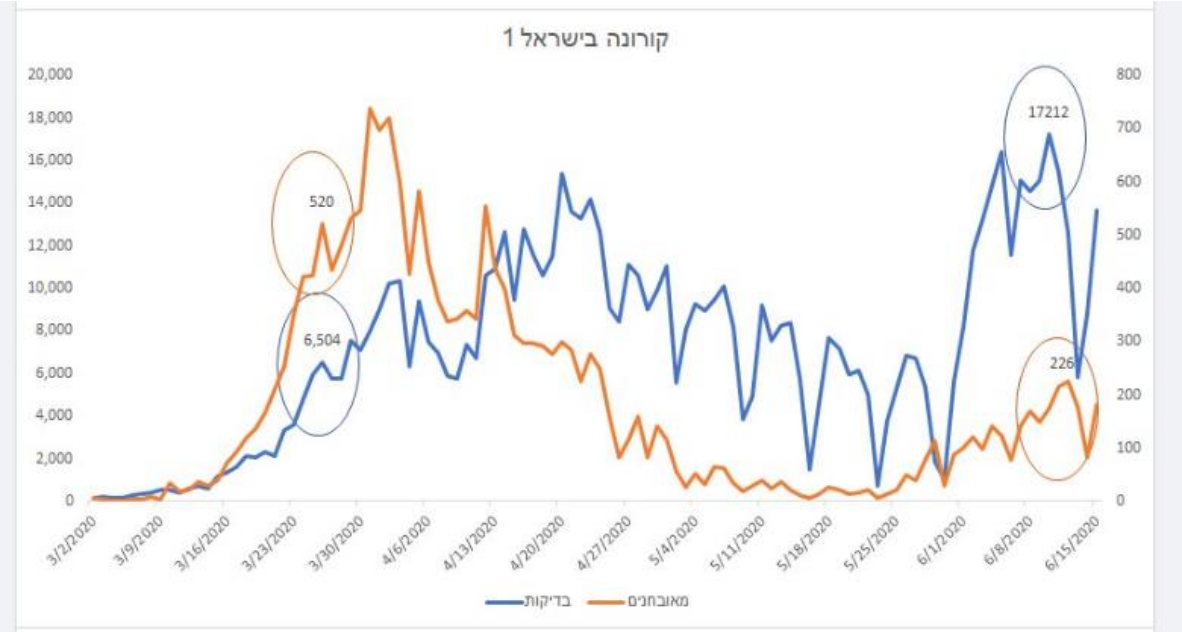
_____

_____

ת"ז: _____

<u>שאלה 10 (6 נק')</u>

הגרף הבא מראה את מספר בדיקות הקורונה שבוצעו (כחול, ציר Y שמאלי) ומספר החיוביים שנמצאו (כתום, ציר Y ימני) בין מרץ ליוני 2020. למשל, לקראת סוף מרץ, בתאריך המסומן בעיגולים משמאל, התבצעו 6504 בדיקות ומתוכן 520 יצאו חיוביות.



לגבי כל אחד מההיגדים הבאים. הסבירו האם ניתן להסיקו מהגרף (ללא שימוש בנתונים נוספים).

הסבירו את תשובתכם לגבי כל היגד במשפט אחד:

א. קיים קשר חיובי בין מספר הבדיקות שבוצעו לבין מספר המאובחנים.

ניתן להסיק / לא ניתן להסיק.

הסבר:_____

_____

ב. הגידול בכמות המאובחנים נובע מגידול בכמות הבדיקות.

ניתן להסיק / לא ניתן להסיק.

הסבר:_____

_____

ת"ז: _____

<u>שאלה 11 (15 נק')</u>

ענו נכון או לא נכון. נמקו תשובתכם במשפט אחד.

א. לאורך כל תקופת הקורונה, נאספים נתונים על כמות נבדקים, על אחוז האנשים שדיווחו שיש להם סימפטומים מבין הנבדקים ועל אחוז האנשים שדיווחו שיש להם סימפטומים מבין הנבדקים החיוביים. ישנה טענה כי בגל הראשון של הקורונה, כאשר היה מחסור בבדיקות, אנשים טענו שיש להם סימפטומים (גם כשלא היו להם) על מנת לקבל בדיקת קורונה. תופעה זו מהווה הטיית מדידה (measurement bias). נכון / לא נכון.

נימוק (במשפט אחד): _____

_____

ב. לכל מסווג שתמיד מחזיר את הערך "0" בעת סיווג של משתנה בינארי שערכיו "0" או "1" צפוי להיות דיוק של כ-50%. נכון / לא נכון.

נימוק (במשפט אחד): _____

_____

ג. חישבנו רווח סמך ברמת ביטחון של 95% עבור ממוצע מספר הודעות הווטסאפ שסטודנטים בטכניון מקבלים ביום מתוך מדגם גדול ואקראי של סטודנטים בטכניון. רווח הסמך חושב באמצעות בוטסטראפ עם 1000 סימולציות. קיבלנו רווח סמך של [25,90]. לפי רווח הסמך, ניתן לדעת כי בכ-950 מתוך 1000 הסימולציות, הערך הממוצע של מספר הודעות הווטסאפ היומי שחושב עבור המדגם מהסימולציה היה בטווח [25,90]. נכון / לא נכון.

נימוק (במשפט אחד): _____

_____

ד. חוקרים מצאו כי בקרב אנשים שעושים אמבטיה חמה פעמיים בשבוע יש 26% פחות מקרי שבץ ו-28% פחות מקרים של מחלות לב מאשר אצל אלו שלא עושים אמבטיה חמה פעמיים בשבוע. מכך ניתן להסיק שאמבטיות חמות מגנות מפני מחלות לב ושבץ. נכון / לא נכון.

נימוק (במשפט אחד): _____

_____

ה. ככל שנגדיל את כמות הפיצ'רים (מאפיינים) בהם נשתמש בסיווג, הדיוק של המסווג ישתפר (או לפחות לא ירד). נכון / לא נכון.

נימוק (במשפט אחד): _____

_____

ת"ז: _____

<u>שאלת בונוס (2 נק')</u>

הסבירו בקצרה מדוע meme זה (אמור להיות) מצחיק. השתמשו במושגים שנלמדו בקורס:

_____

_____

_____

_____

_____

_____

_____