מבחן בית למשרת DS – מכבי שירותי בריאות

שם: איתן בוקר

1. פתיח

נהניתי מאוד מהעבודה על פרויקט זה, שהיה עבורי גם מאתגר וגם מעשיר במיוחד. תחום הבריאות והעולם הרפואי באופן כללי מסקרן ומושך אותי, וזו הייתה עבורי הזדמנות משמעותית להעמיק בו. מכיוון שאין לי רקע רפואי קודם, מצאתי לנכון להתחיל את התהליך בלימוד והעמקה בספרות המקצועית, כדי להבין מהם הפיצ'רים הקליניים המרכזיים הקשורים לרעלת הריון ולסיכון במהלך ההיריון. רק לאחר שהצלחתי למפות ולהבין את המשתנים המשמעותיים – התחלתי את מבחן הבית.

2. ספרות מקצועית

בסעיף זה אציג את עיקרי הדברים שמהווים אינדיקטור לזיהוי רעלת היריון בשלב מוקדם בהיריון. סקר הספרות בוצע מ-guidelines המהווה מסקנות מאוסף של מאמרים שונים בתחום.

1. WHO recommendations for Prevention and treatment of pre-eclampsia and eclampsia

רעלת הריון היא אחת ההפרעות המשמעותיות בלחץ דם במהלך ההריון, ומובילה מרכזית לתחלואה ולתמותה אימהית ופרינטלית.
גורמי הסיכון כוללים: השמנה, יתר לחץ דם כרוני, סוכרת, הריון ראשון, הריון בגיל ההתבגרות, וכן הריון מרובה עוברים.

נהוג לסווג רעלת הריון לשתי דרגות חומרה – קלה וחמורה. רעלת הריון תיחשב חמורה כאשר מתקיים אחד מהתנאים הבאים:

- יתר לחץ דם - כאשר לחץ הדם הדיאסטולי נותר גבוה מ־90 מ"מ כספית.
- פרוטאינוריה משמעותית - יותר מ־0.3 גרם חלבון בשתן שנאסף במשך 24 שעות.
- פגיעה באיברים חיוניים של האם (כגון כבד, כליות, מערכת עצבים מרכזית).

עיכוב בלידה עלול להוביל להחמרת המצב, לאי-ספיקה שלייתית ולפגיעה מערכתית באם – מצב הקשור לעלייה בסיכון לתמותה גם של האם וגם של היילוד.

פגיעה באיברים כתוצאה מרעלת הריון עשויה להתבטא קלינית בתסמונות חמורות, כגון:

- אקלמפסיה– פרכוסים כלליים שאינם מוסברים מגורמים אחרים (כגון אפילפסיה).
- תסמונת HELPP המתבטאת בהמוליזה, עלייה באנזימי כבד וירידה במספר הטסיות.

מצבים אלו מהווים אינדיקציה להחמרה מערכתית ומעלים את הסיכון לסיבוכים קשים ואף למוות.

2. Hypertension in Pregnancy

להלן מאפיינים חמורים של רעלת הריון (מספיק קיום אחד מהבאים):

- לחץ דם סיסטולי של 160 מ"מ כספית או יותר, או לחץ דם דיאסטולי של 110 מ"מ כספית או יותר, בשתי מדידות בהפרש של לפחות 4 שעות, כאשר המטופלת נמצאת במנוחה (אלא אם כבר התחילה טיפול תרופתי נגד יתר לחץ דם לפני כן).
- טרומבוציטופניה – ספירת טסיות נמוכה מ־100,000 למיקרוליטר.
- תפקוד כבד לקוי, כפי שמודגם בעלייה לא תקינה באנזימי כבד (פי שניים מהנורמה), או כאב מתמשך וחמור בבטן ימנית עליונה / אפיגסטריום שאינו מגיב לטיפול תרופתי ולא מוסבר על ידי אבחנות אחרות – או שניהם יחד.
- אי ספיקת כליות מתקדמת – ריכוז קריאטינין בדם הגבוה מ־1.1 מ"ג/ד"ל, או הכפלה של ריכוז הקריאטינין יחסית לרמה הבסיסית – בהיעדר מחלת כליות אחרת.
- בצקת ריאתית
- הפרעות מוחיות או ראייתיות חדשות – הופעה פתאומית של תסמינים נוירולוגיים או ראייתיים.

BOX E-1. Severe Features of Preeclampsia (Any of these findings) ⇦

- Systolic blood pressure of 160 mm Hg or higher, or diastolic blood pressure of 110 mm Hg or higher on two occasions at least 4 hours apart while the patient is on bed rest (unless antihypertensive therapy is initiated before this time)
- Thrombocytopenia (platelet count less than 100,000/microliter)
- Impaired liver function as indicated by abnormally elevated blood concentrations of liver enzymes (to twice normal concentration), severe persistent right upper quadrant or epigastric pain unrespon-sive to medication and not accounted for by alternative diagnoses, or both
- Progressive renal insufficiency (serum creatinine concentration greater than 1.1 mg/dL or a doubling of the serum creatinine concentration in the absence of other renal disease)
- Pulmonary edema
- New-onset cerebral or visual disturbances

קריטריונים אבחנתיים לרעלת הריון (Preeclampsia)

<u>לחץ דם</u>:

- סיסטולי ≥140 מ"מ כספית או דיאסטולי ≥90 מ"מ כספית, בשתי מדידות בהפרש של לפחות 4 שעות, לאחר שבוע 20 להריון אצל אישה עם לחץ דם תקין קודם לכן

או

- סיסטולי ≥160 מ"מ כספית או דיאסטולי ≥110 מ"מ כספית, כאשר ניתן לאשר ביתר מהירות (בפרקי זמן של דקות) לצורך התחלת טיפול תרופתי דחוף

<u>פרוטאינוריה</u> (נוכחות חלבון בשתן):

- ≥300מ"ג חלבון באיסוף שתן של 24 שעות

או

- יחס חלבון/קריאטינין ≥0.3

או

- תוצאה של +1 בבדיקת strip

TABLE E-1. Diagnostic Criteria for Preeclampsia ⇦

Blood pressure	• Greater than or equal to 140 mm Hg systolic or greater than or equal to 90 mm Hg diastolic on two occasions at least 4 hours apart after 20 weeks of gestation in a woman with a previously normal blood pressure • Greater than or equal to 160 mm Hg systolic or greater than or equal to 110 mm Hg diastolic, hypertension can be confirmed within a short interval (minutes) to facilitate timely antihypertensive therapy
and	
Proteinuria	• Greater than or equal to 300 mg per 24-hour urine collection (or this amount extrapolated from a timed collection) or • Protein/creatinine ratio greater than or equal to 0.3* • Dipstick reading of 1+ (used only if other quantitative methods not available)
Or in the absence of proteinuria, new-onset hypertension with the new onset of any of the following:	
Thrombocytopenia	• Platelet count less than 100,000/microliter
Renal insufficiency	• Serum creatinine concentrations greater than 1.1 mg/dL or a doubling of the serum creatinine concentration in the absence of other renal disease
Impaired liver function	• Elevated blood concentrations of liver transaminases to twice normal concentration
Pulmonary edema	
Cerebral or visual symptoms	

* Each measured as mg/dL.

גורמי סיכון:

- הריון ראשון
- רעלת הריון בעבר
- יתר לחץ דם כרוני או מחלת כליות כרונית, או שניהם יחד
- היסטוריה של טרומבופיליה (נטייה לקרישיות יתר בדם)
- הריון מרובה עוברים
- הפריה חוץ-גופית
- היסטוריה משפחתית של רעלת הריון
- סוכרת סוג 1 או סוג 2
- השמנת יתר
- זאבת מערכתית
- גיל אם מתקדם – מעל גיל 40

BOX 3-1. Risk Factors for Preeclampsia ⇦

- Primiparity
- Previous preeclamptic pregnancy
- Chronic hypertension or chronic renal disease or both
- History of thrombophilia
- Multifetal pregnancy
- In vitro fertilization
- Family history of preeclampsia
- Type I diabetes mellitus or type II diabetes mellitus
- Obesity
- Systemic lupus erythematosus
- Advanced maternal age (older than 40 years)

לסיכום ממצאי הסקירה הספרותית, זיהוי מוקדם של רעלת הריון מתבסס על שילוב של גורמי סיכון אישיים, סימנים קליניים ומדדים ביולוגיים.

הפיצ'רים המרכזיים שנמצאו כרלוונטיים ביותר הם:

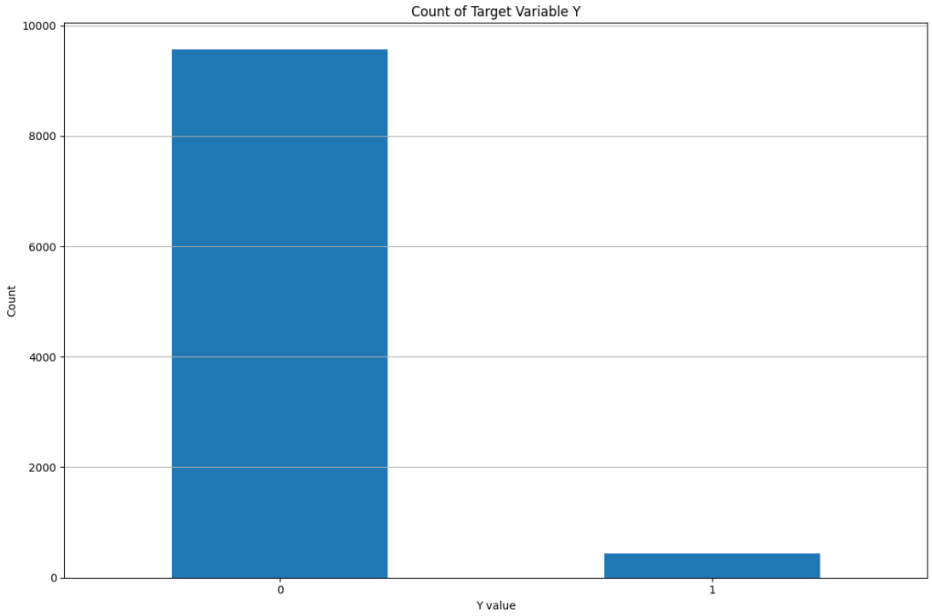
- **לחץ דם**: מדדים סיסטוליים ודיאסטוליים הגבוהים מהנורמה לאחר שבוע 20 להריון, ובמיוחד ערכים מעל 140/90 מ"מ כספית (או 160/110 במקרים חמורים).
- **פרוטאינוריה**: הפרשה של **חלבון בשתן** מעל 300 מ"ג ב-24 שעות, יחס חלבון/קריאטינין ≥0.3, או תוצאה חיובית בבדיקת strip.
- **תפקודי כבד וכליה**: עלייה משמעותית **באנזימי כבד**, **רמות קריאטינין** מוגברות או הכפלת רמת הבסיס.
- **ספירת טסיות** נמוכה מ־100,000 למיקרוליטר.
- **תסמינים נוירולוגיים וראייתיים חדשים**
- **בצקת ריאתית.**
- **גורמי סיכון אישיים ורפואיים**: הריון ראשון, רעלת הריון בעבר, יתר לחץ דם כרוני, סוכרת (סוג 1 או 2), השמנה, זאבת, טרומבופיליה, הריון מרובה עוברים, הפריה חוץ-גופית, גיל אם מעל 40 והיסטוריה משפחתית של רעלת הריון.

נתונים אלה, כאשר נאספים ונמדדים כבר עד שבוע 15 להריון, יכולים לשמש תשתית לפיתוח אלגוריתם חיזוי שיזהה נשים בסיכון גבוה ויאפשר התערבות מונעת מוקדמת.

3. Data Exploration (חלק I)

במחברת הjupyter מופיעים גם גרפים תומכים לצורכי תחקור, אך כאן אביא רק את <u>הממצאים המרכזיים</u> שעלו בתהליך ה-EDA.

3.1 משתנה מטרה Y



ניתן לראות כי משתנה המטרה אינו מאוזן:

1 – מעט מאוד רשומות שייכות למחלקה זו 4.4% (מצב מסוכן / רעלת היריון)

0 – רוב הרשומות שייכות למחלקה זו 95.6% (ללא סיבוך / מצב לא מסוכן)

יש לקחת בחשבון מצב זה כאשר בוחרים מודל למידה (/מתן משקל גבוה יותר) ושימוש במטריקות המתאימות.

באוכלוסייה, לפי WHO אחוז הנשים שיש להן רעלת היריון עומד על 10% וזה קרוב להתפלגות שלנו.

“Hypertensive disorders of pregnancy affect about 10% of all pregnant women around the world”

3.2 ערכים חסרים

- קיימים 76 פיצ'רים עם ערכים חסרים, מתוכם כ-53 פיצ'רים עם מעל ל50% ערכים חסרים בפיצ'רים מסוג measures ו last_diag.
- לאחר בדיקה, כ-9 פיצ'רים הם ללא השפעה על משתנה המטרה, בעוד ששאר הפיצ'רים כן משפיעים על יכולת הניבוי של משתנה המטרה.

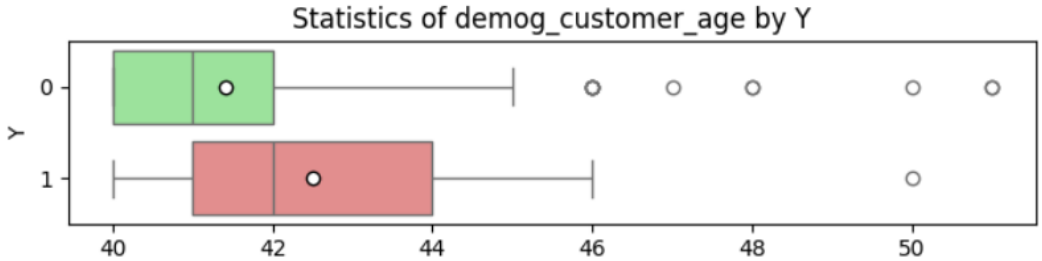
feature	percent_Y_missing	percent_Y_non_missing
24_diag_53_days_since_last_diag	0.042952	0.250000
4_diag_98_days_since_last_diag	0.043043	0.200000
24_diag_83_days_since_last_diag	0.042113	0.153061
4_diag_118_days_since_last_diag	0.042738	0.101266
4_diag_124_days_since_last_diag	0.041352	0.097859
4_diag_130_days_since_last_diag	0.043038	0.093750
4_diag_114_days_since_last_diag	0.043147	0.090909
24_diag_81_days_since_last_diag	0.039337	0.089378
24_diag_71_days_since_last_diag	0.042837	0.078431

- מתוך הערכים הקיימים בפיצ'רים אלו, אחוז ה-1-ים במשתנה המטרה Y נע בין 1.5%-25% ולכן לא נסיר אותם בהמשך.
- מבחינה קלינית, קיימת הצדקה לערכי Null בעמודות מסוג last_diag. מדובר במקרים בהם לא בוצעה אבחנה רלוונטית באותו מועד, ולכן חוסר המידע הוא צפוי ואינו מעיד על בעיה בנתונים.
- באופן דומה, בעמודות מסוג measures (מדידות לחץ דם), הערכים החסרים נובעים מכך שלא בוצעה מדידה בפועל, ולכן ערך ה־Null מייצג מצב רפואי תקף ולא טעות נתונים.
- ערכים שיש בהם null ולא משפיעים על משתנה המטרה יוסרו (סה"כ 9 פיצ'רים).

3.3 בדיקת הקשר בין המשתנים בלתי תלויים ומשתנה המטרה

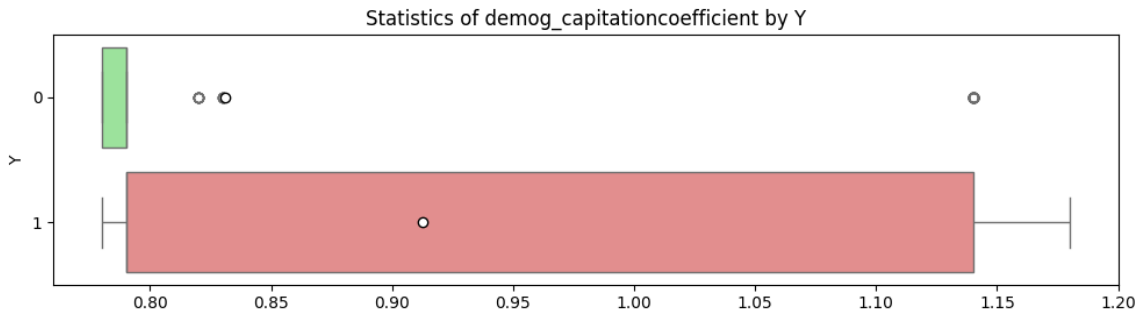
3.3.1 משתנים דמוגרפיים

- עפ"י סקר הספרות מצופה שבגיל מבוגר יותר הסיכוי לרעלת היריון יגדל (+40).
- <u>משתנה גיל</u>: כאשר מסתכלים על כל הsample, ההתפלגויות מאוד דומות אך כאשר מסתכלים על תת האוכלוסייה של +40, כ-13.6% אחוז מהנשים הן בעלות סיכוי לחלות ברעלת היריון.



יתרה מזאת, נשים שאובחנו הן מבוגרות יותר בממוצע מהנשים שלא אובחנו עם פיזור רחב יותר.

- <u>משתנה קפיטציה</u>: בנוסף, לפי סקר הספרות, נשים בעלי סוציו אקונומי נמוך יצרכו יותר שירותים רפואיים ויהיו בעלי סיכוי גבוה יותר לחלות ברעלת היריון. אם נסתכל על תת האוכלוסייה בגיל +40:



מקדם הקפיטציה גבוה יותר, עם טווח רחב הרבה יותר (כ-0.85 עד 1.15), וממוצע ~0.92.

- משמעות קלינית: נשים בגיל 40+ שפיתחו סיבוך, הן בעלות מקדם קפיטציה גבוה יותר בממוצע — כלומר, יש להן פרופיל רפואי מורכב יותר. תוצאה זו עשויה להעיד על קשר בין פרופיל רפואי מורכב לבין סיכון להתפתחות סיבוכים במהלך ההריון.

3.3.2 משתני עישון

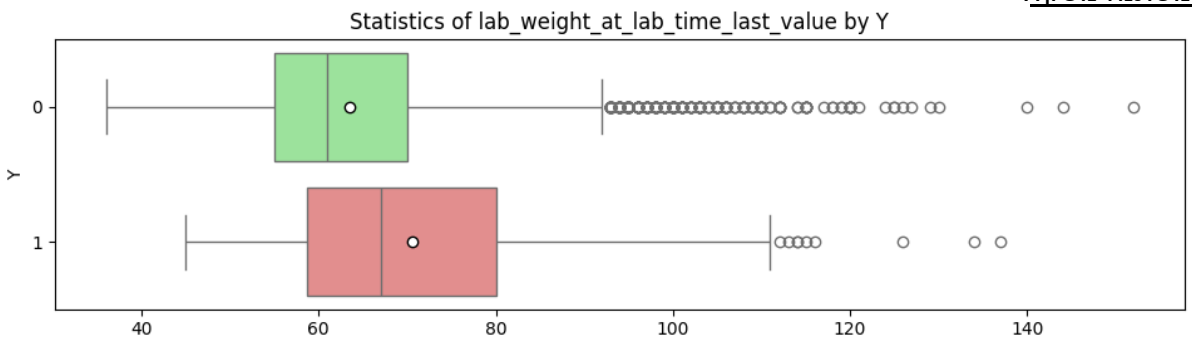
- <u>משתנה מספר שנות עישון</u>: יש כאן ערכים לא הגיונים לדוגמא 107 שנות עישון (ערך מקסימלי 122 שנות עישון.
- <u>משתנה האם מעשנת</u>: מקבל שלושה ערכים: 0,1,2.

0 – לא מעשנת. 1 ו-2 אני מניח שזה כן מעשנת אך זה לא ערך בינארי כפי שציפיתי

- יצרתי משתנה חדש is_smoker – אם הערך הוא 1,2 אז פיצ'ר יקבל 1 אחרת 0. בדקתי עם mosaic plot וראיתי שאין קשר בין משתנה המטרה לis_smoker.
- <u>משתנה מעשן כבד</u>: רק סאמפל אחד שערכו שווה ל2, משפיע על משתנה המטרה.

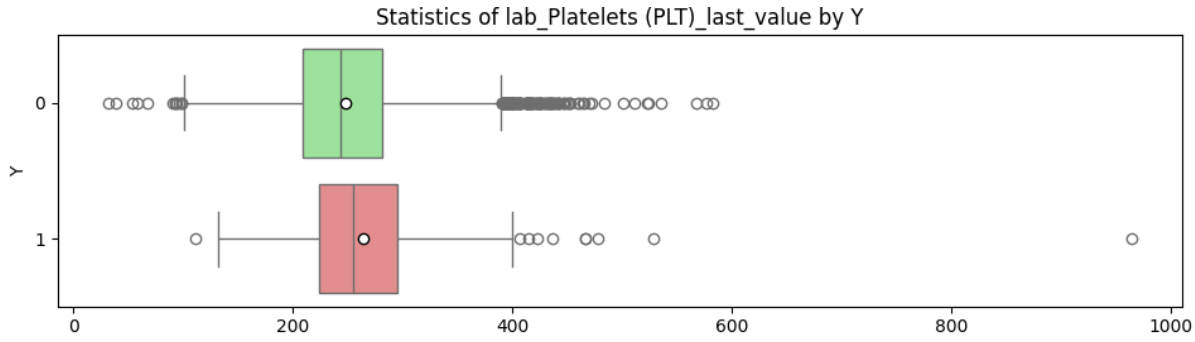
3.3.3 משתני מעבדה

- מסקירת הספרות, חלבון בשתן, ספירת טסיות, היריון מרובה עוברים והשמנת יתר (BMI) אלו גורמים המשפיעים על רעלת היריון.
- <u>משתנה משקל</u>:



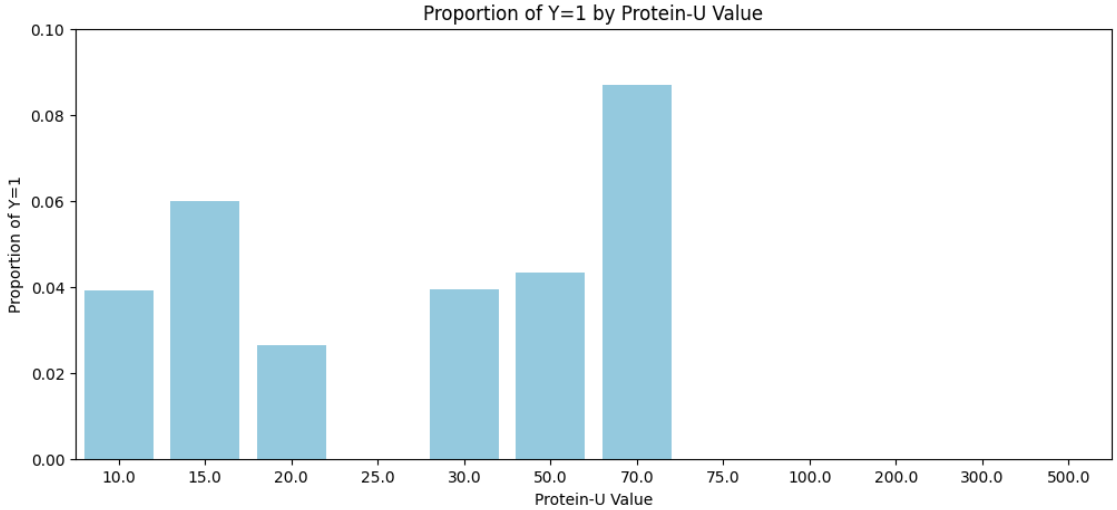
נשים שפיתחו סיבוך במהלך ההריון מציגות משקל גבוה יותר בממוצע ובחציון כבר בשלב מוקדם של ההריון (בזמן בדיקת המעבדה). ממצא זה תואם את הידוע בספרות, לפיו השמנת יתר מהווה גורם סיכון מובהק לרעלת הריון.

- <u>משתנה טסיות:</u>



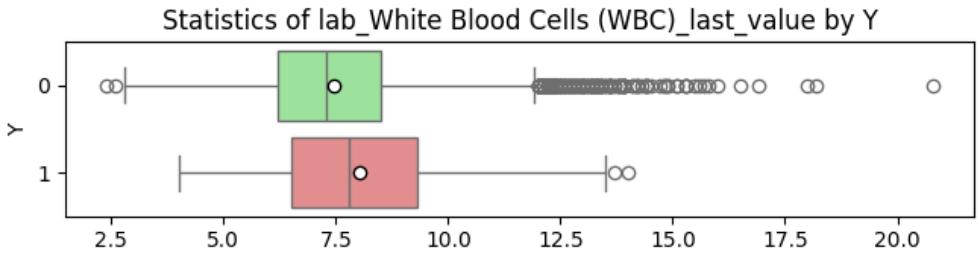
בנתונים הקיימים אין הבדל מהותי בין הקבוצות, לא בערך החציוני ולא בהתפלגות הכללית. ייתכן כי רוב מקרי הסיבוך בדאטה זה אינם מייצגים רעלת חמורה, או שהמדידה האחרונה אינה משקפת את השלב הקריטי בהתפתחות הסיבוך.

- <u>משתנה חלבון בשתן:</u>



לפי הספרות הרפואית ערכי חלבון בשתן של 300 מ"ג ומעלה מהווים קריטריון קליני לרעלת הריון, אך בדאטה הנוכחי לא נמצאו כלל נשים שפיתחו סיבוך בערכים אלו.
עם זאת, נצפה כי גם בערכים נמוכים יחסית (15–50 מ"ג) שיעור הנשים שפיתחו סיבוך עומד על כ־4%, ואילו בערך 70 מ"ג שיעור זה מטפס לכ־9%.
ממצא זה עשוי להעיד על כך שגם ערכים מתחת לסף הקליני המקובל, עשויים לשאת מידע מנבא מסוים.

- <u>משתנה מספר עוברים:</u> כלל הdataset עם עובר אחד בלבד.
- <u>משתנה תאי דם לבנים:</u>



רעלת היריון היא מצב דלקתי ומערכות החיסון מגיבות בהתאם. ניתן לראות כי הממוצע והחציון מעט גבוהים יותר לעומת הנשים הבריאות אך ממצא זה אינו נתמך באופן מובהק בספרות הרפואית.

- <u>קורלציות:</u> קיימת קורלציה בין המשתנים הבאים:

Feature 1	Feature 2	Correlation
lab_NT_abs_last_value	lab_NT_MoM_last_value	0.942412
lab_Neutrophils_2_last_value	lab_White Blood Cells (WBC)_last_value	0.930620
lab_Eosinophils_2_last_value	lab_Eosinophils_1_last_value	0.929864
lab_Hematocrit (HCT)_last_value	lab_Hemoglobin (HGB)_last_value	0.924765
lab_Mean Corpuscular Hemoglobin (MCH)_last_value	lab_Red Cell Count (RCC)_last_value	0.910560
lab_papp_a_MoM_last_value	lab_papp_a_abs_last_value	0.878515
lab_Neutrophils_1_last_value	lab_Lymphocytes_1_last_value	-0.954148

❖ lab_NT_abs ↔ lab_NT_MoM
שני המשתנים מודדים את אותה בדיקה (שקיפות עורפית) — פשוט באחד זה ערך רגיל ובשני ערך מתוקנן — לכן הם דומים מאוד.

❖ lab_Neutrophils_2 ↔ lab_WBC
נויטרופילים הם סוג עיקרי של תאי דם לבנים, אז כשיש מהם הרבה — גם סך תאי הדם הלבנים עולה.

❖ lab_Eosinophils_1 ↔ lab_Eosinophils_2
נגזר מאותו בדיקה ערכים כמעט זהים.

❖ lab_HCT ↔ lab_HGB
שניהם מודדים את מצב תאי הדם האדומים — כשאחד עולה, גם השני בדרך כלל עולה.

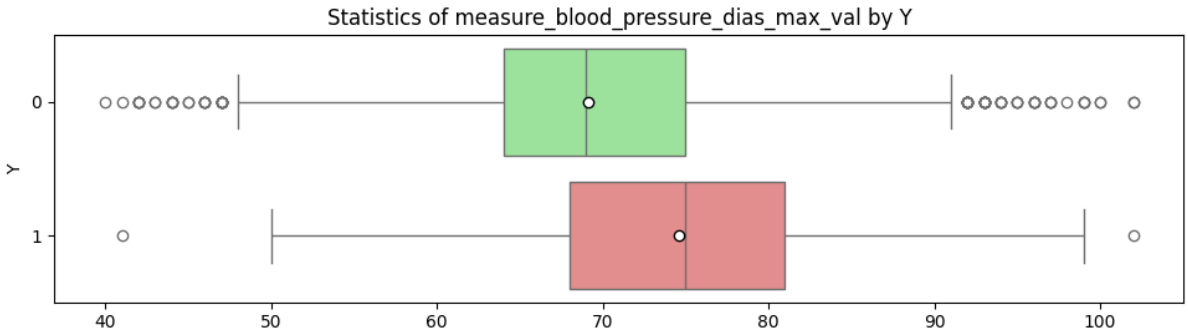
❖ lab_MCH ↔ lab_RCC
כשיש יותר תאי דם אדומים גם רמות ההמוגלובין בתא (MCH) נוטות להיות גבוהות.

❖ lab_papp_a_abs ↔ lab_papp_a_MoM
אותו בדיקה, רק אחד רגיל והשני מותאם לגיל ההיריון.

❖ lab_Neutrophils_1 ↔ lab_Lymphocytes_1
כשסוג אחד של תאי דם לבנים עולה (נויטרופילים), הסוג השני (לימפוציטים) יורד — לכן יש קשר הפוך חזק.

3.3.4 משתני מדידה

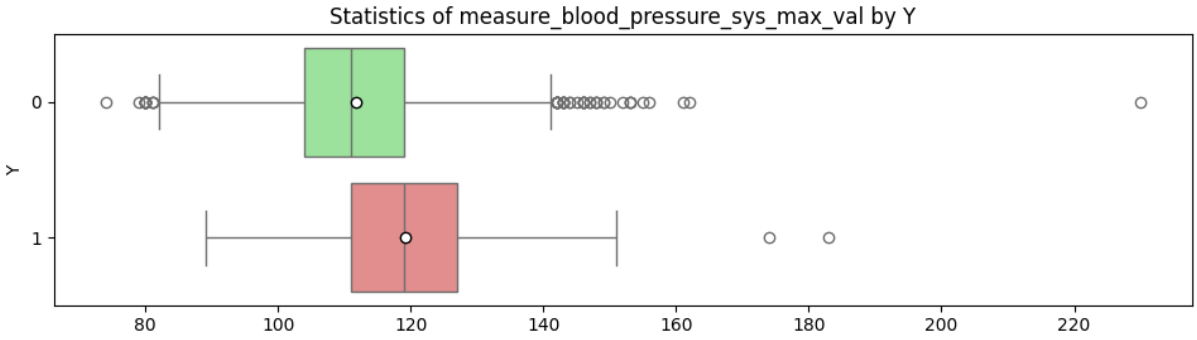
- מסקירת הספרות, משתני לחץ הדם הסיסטוליים והדיאסטוליים (כגון max_val, mean_val ו-last_val) הם מהפיצ'רים המרכזיים לזיהוי מוקדם של רעלת הריון, בהתאם לספרות הקלינית המדגישה ערכים גבוהים מ־140/90.
- <u>משתנה לחץ דם דיאסטולי מקסימלי</u>: אם נסתכל על ההתפלגות מקסימלית של לחץ דם דיאסטולי:



אצל נשים שפיתחו סיבוך, ניתן לראות גם הממוצע וגם החציון של לחץ הדם הדיאסטולי גבוהים יותר. התפלגות הערכים רחבה מעבר לערך 90.

משמעות קלינית: זה כן תומך בהשערה שלחץ דם דיאסטולי גבוה הוא סמן מוקדם לסיבוך, בדיוק כפי שצויין בסקר הספרות.

- <u>משתנה לחץ דם סיסטולי מקסימלי:</u> אם נסתכל על ההתפלגות מקסימלית של לחץ דם סיסטולי:



אצל נשים שפיתחו סיבוך, ניתן לראות גם הממוצע וגם החציון של לחץ הדם הדיאסטולי גבוהים יותר בהשוואה לנשים שלא פיתחו סיבוך. רוב הנשים שפיתחו סיבוך הן בעלת לחץ סיסטולי של 115-125 אך ישנן מטופלות עם ערך גדול מ140, דבר שמרמז על מצבים חמורים יותר

ממצאים אלו תואמים את הקריטריונים הקליניים המוכרים לרעלת הריון, שבהם לחץ דם סיסטולי מעל 140 מהווה סמן מובהק לסיכון.

- <u>קורלציות</u>: קיימת קורלציה גבוהה מאוד בין משתנים אלו (min/max/mean/stdev):

Feature 1	Feature 2	Correlation
measure_blood_pressure_sys_count	measure_blood_pressure_dias_count	1.000000
measure_blood_pressure_sys_mean_val	measure_blood_pressure_sys_min_val	0.921097
measure_blood_pressure_sys_mean_val	measure_blood_pressure_sys_max_val	0.918387
measure_blood_pressure_dias_mean_val	measure_blood_pressure_dias_min_val	0.916962
measure_blood_pressure_dias_mean_val	measure_blood_pressure_dias_max_val	0.916553
measure_blood_pressure_sys_last_val	measure_blood_pressure_sys_mean_val	0.912786
measure_blood_pressure_sys_min_max_percent	measure_blood_pressure_sys_stddev_val	0.909861
measure_blood_pressure_sys_first_val	measure_blood_pressure_sys_mean_val	0.909530
measure_blood_pressure_dias_last_val	measure_blood_pressure_dias_mean_val	0.908059
measure_blood_pressure_dias_first_val	measure_blood_pressure_dias_mean_val	0.907958
measure_blood_pressure_dias_min_max_percent	measure_blood_pressure_dias_stddev_val	0.900993
measure_blood_pressure_sys_first_val	measure_blood_pressure_sys_max_val	0.876991
measure_blood_pressure_dias_first_val	measure_blood_pressure_dias_max_val	0.871618
measure_blood_pressure_sys_last_val	measure_blood_pressure_sys_min_val	0.868881
measure_blood_pressure_dias_last_val	measure_blood_pressure_dias_min_val	0.862545
measure_blood_pressure_sys_last_val	measure_blood_pressure_sys_max_val	0.806012
measure_blood_pressure_sys_first_val	measure_blood_pressure_sys_min_val	0.804853
measure_blood_pressure_dias_last_val	measure_blood_pressure_dias_max_val	0.804358

נמצאו קורלציות גבוהות מאוד בין מדדים שונים של לחץ דם סיסטולי ודיאסטולי — כולל ערכי ממוצע, מינימום, מקסימום, ערך ראשון ואחרון.

הדבר טבעי, שכן כל המדדים מבוססים על אותן סדרות מדידות, ומשקפים מגמות דומות בפרופיל לחץ הדם של המטופלת.

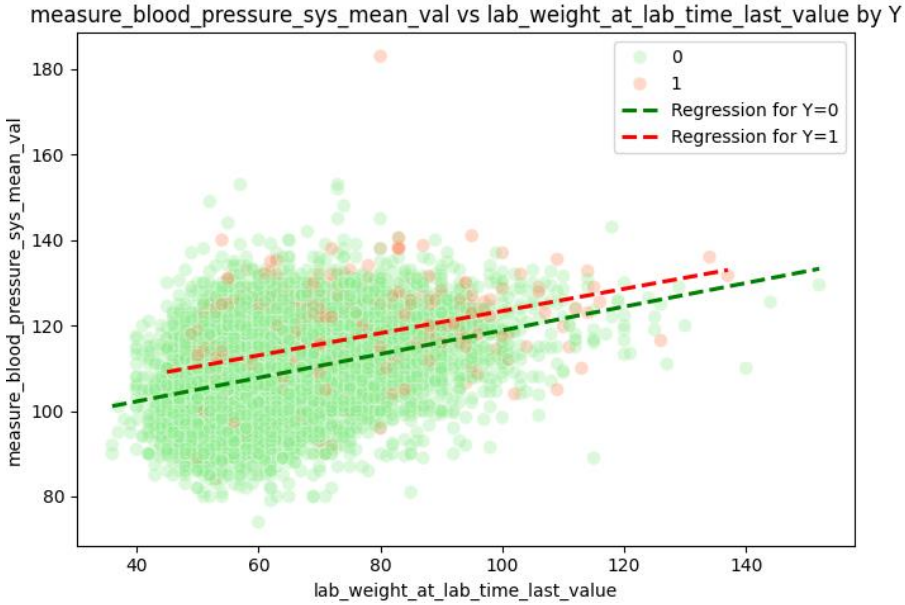
כתוצאה מכך, ייתכן שיש חפיפה במידע בין חלק מהפיצ'רים, ולכן ניתן לשקול להשאיר רק חלק מהם כדי למנוע כפילויות ולפשט את המודל.

נתייחס לזה ב-feature selection.

3.4 בדיקת הקשר בין המשתנים בלתי תלויים

בסעיף זה אני אנסה לענות על מספר שאלות קליניות שעוררו את סקרנותי.

3.4.1 האם נשים עם משקל גבוה יותר נוטות ללחץ דם גבוה יותר?

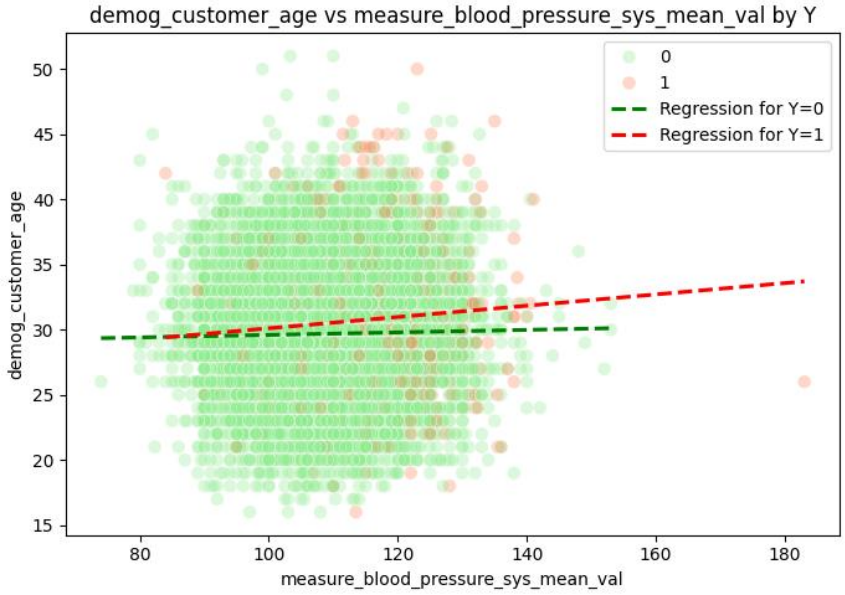


ניתן לראות מגמה שבה לחץ הדם הסיסטולי הממוצע עולה עם המשקל, בשתי הקבוצות.

עם זאת, בקרב נשים שפיתחו סיבוך, הלחץ גבוה יותר באופן עקבי לאורך כל טווח המשקל.

ממצא זה מחזק את ההשערה שלשילוב בין השמנת יתר ולחץ דם גבוה יש תרומה לסיכון לרעלת הריון.

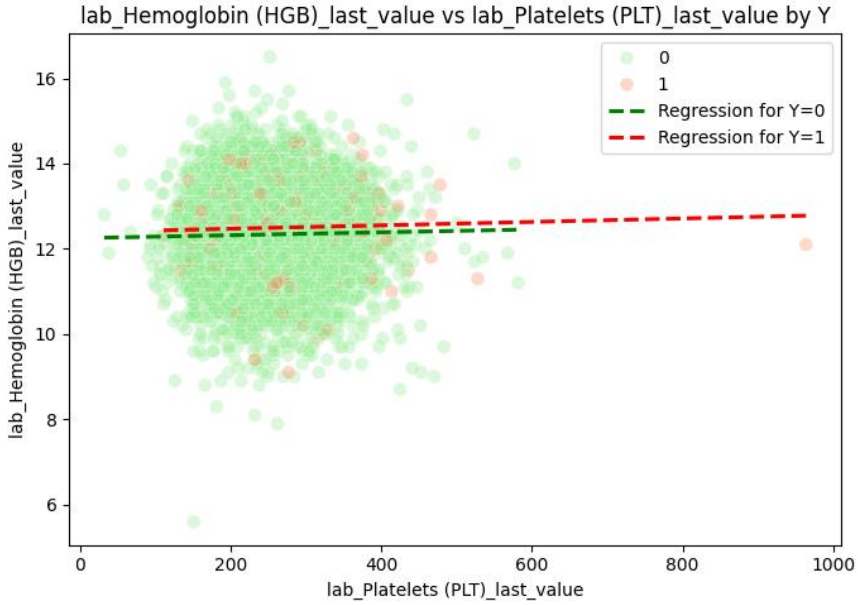
3.4.2 האם נשים מבוגרות נוטות ללחץ דם גבוה יותר?



בקרב נשים שלא פיתחו סיבוך, לא נמצא קשר משמעותי בין גיל ללחץ דם. לעומת זאת, בקרב נשים שפיתחו סיבוך, ניכרת מגמה של עלייה בלחץ הדם עם הגיל.

ממצא זה עקבי עם הספרות הרפואית - גיל מתקדם מהווה גורם סיכון, במיוחד כאשר הוא מלווה בלחץ דם גבוה.

3.4.3 האם בעיות בקרישה ו/או אנמיה עשויות להיות קשורות לסיבוכים הריוניים?

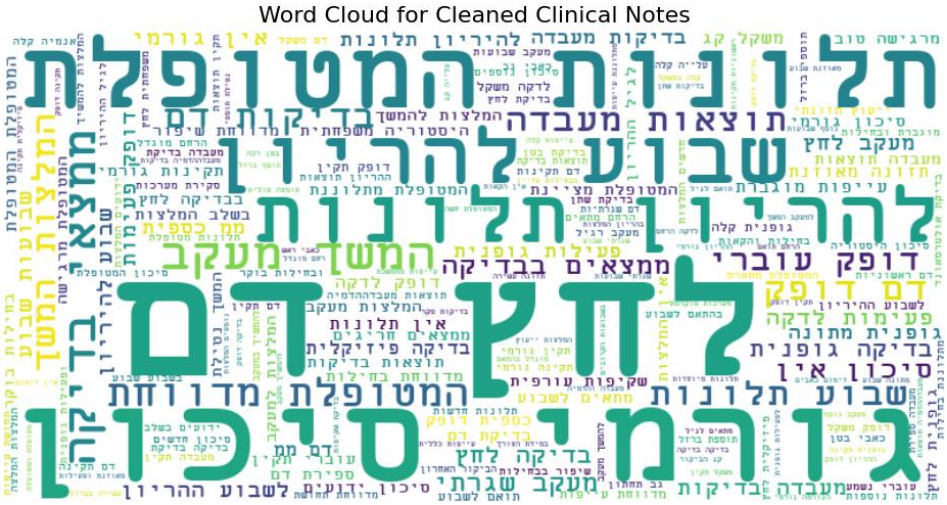


שני הקלאסים מרוכזים סביב אותם ערכים.

לא נמצא קשר משמעותי בין רמות טסיות לבין רמות המוגלובין.

3.5 משתנה טקסט – ניתוח טקסט חופשי

3.5.1 ניתוח כלל הטקסט



לצורך זיהוי תכנים חוזרים במסמכים הקליניים, בוצע ניקוי טקסט שכלל הסרת תווים שאינם עבריים והסרת מילות עצירה בעברית.

לאחר הניקוי, נבנה WordCloud אשר מציג את המילים הבולטות ביותר בכלל הרשומות. גודל המילים בגרף משקף את התדירות שלהן בטקסטים.

מהמילים הבולטות ניתן לזהות:

- מונחים קליניים מרכזיים כגון "לחץ דם", "גורמי סיכון", "תלונות המטופלת"
- מונחים פרוצדורליים כמו "ממצאי בדיקה", "תוצאות מעבדה."

ממצאים אלו משקפים היטב את אופי הנתונים: רישומי ביקור רפואיים המתארים מצבים קליניים, תסמינים, תוצאות בדיקות, והמלצות טיפול.

3.5.2 ניתוח טקסט (Y=1)

גם כאן בוצע ניקוי טקסט כפי שצויין בסעיף הקודם.

מהמילים הבולטות בקבוצה זו:

- לחץ דם – מופיעה באופן בולט במיוחד, בהתאם להקשר הקליני של רעלת הריון.
- גורמי סיכון, ממצאים חריגים, תלונות – מעידים על תיעוד קליני פעיל ומורכב יותר.
- תלונות המטופלת, בדיקות, המשך מעקב, כאב, עייפות מוגברת, המלצות – מצביעים על ריבוי דיווחים קליניים והתערבויות.

בהשוואה בין WordClouds של כלל הדאטה לזה של נשים שפיתחו סיבוך, ניכר שבקבוצה השנייה מופיעות בתדירות גבוהה יותר מילים הקשורות למעקב רפואי, תסמינים וגורמי סיכון. הדבר עשוי להעיד על עומס תסמינים, ניטור תכוף יותר או ריבוי אינטראקציות עם המערכת הרפואית / רופא.

3.6 משתני אבחנה (sum)

משתני האבחנה שייכים לשלב שלאחר הופעת המקרה (Y=1), ולכן לא ישולבו במודל. עם זאת, הם מספקים תובנות חשובות על חומרת המצב. נמצא כי רעלת הריון היא האבחנה המרכזית במקרים המאומתים.

preeclampsia_sum	128
pregnancy_hypertension_sum	111
essential_hypertension_sum	100
labs_sum	49
eclampsia_sum	17
secondary_hypertension_sum	0
hypertensive_heart_disease_sum	0
hypertensive_chronic_kidney_disease_sum	0
hypertensive_heart_and_chronic_kidney_disease_sum	0

לעיתים רעלת היריון מופיעה יחד עם אבחנות נוספות — מה שעשוי להעיד על החמרה קלינית.

essential_hypertension_sum,pregnancy_hypertension_sum	21
pregnancy_hypertension_sum,preeclampsia_sum	19
essential_hypertension_sum,preeclampsia_sum	12
essential_hypertension_sum,pregnancy_hypertension_sum,preeclampsia_sum	7
preeclampsia_sum,labs_sum	5
preeclampsia_sum,eclampsia_sum	5
essential_hypertension_sum,preeclampsia_sum,labs_sum	3
essential_hypertension_sum,pregnancy_hypertension_sum,preeclampsia_sum,labs_sum	3
essential_hypertension_sum,preeclampsia_sum,eclampsia_sum	3
pregnancy_hypertension_sum,preeclampsia_sum,eclampsia_sum	2
essential_hypertension_sum,pregnancy_hypertension_sum,preeclampsia_sum,eclampsia_sum	2
pregnancy_hypertension_sum,eclampsia_sum	1
pregnancy_hypertension_sum,preeclampsia_sum,labs_sum	1
essential_hypertension_sum,pregnancy_hypertension_sum,labs_sum	1
essential_hypertension_sum,preeclampsia_sum,eclampsia_sum,labs_sum	1
essential_hypertension_sum,eclampsia_sum	1
essential_hypertension_sum,pregnancy_hypertension_sum,eclampsia_sum	1

3.7 שונות

המשתנה int_date מכיל רק ערך אחד 01-01-1970 ולכן יוסר בהמשך.

4. Training Pipeline

(חלק II)

4.1 פיצול לסט אימון ובחינה

פיצול הדאטה לסט אימון וסט בחינה בוצע בשלב מוקדם, על מנת למנוע דליפת מידע (data leakage).
הפיצול נשמר בצורה stratified, כך שהתפלגות הקלאסים נשמרה באופן זהה בשני הסטים.

4.2 ניקוי נתונים

ניקוי הנתונים יתבסס עפ"י התוצאות מחלק I:

- הסרת 9 פיצ'רים אשר הראו שלא קיימת השפעה מובהקת שלהם על המשתנה המוסבר
- משתנה שנות עישון (smoking_smoking_years): זיהוי ערכים אנומליים ותיקונם – ערכים בלתי סבירים הומרו ל-NaN, על מנת למנוע הטיה ו"חוסר איזון" בתוך הפיצ'ר.
- הסרת משתנה int_date – בעל ערך אחד 01-01-1970.

4.3 Feature Engineering

בשלב זה בוצע תהליך "הנדסת פיצ'רים" שמטרתו לתמצת המידע האבחנתי ולהעשיר את הדאטה בתובנות רלוונטיות, תוך הפחתת סיבוכיות מיותרת. להלן הפעולות שבוצעו:

- diag_non_missing_count: מונה את מספר העמודות האבחנתיות שאינן ריקות (non-null) עבור כל מטופלת. מהווה מדד לעושר המידע האבחנתי שנאסף בפרק הזמן הרלוונטי.
- diag_non_missing_sum : סכום כלל הערכים בעמודות האבחנתיות. מדד לשכיחות מצטברת של אבחנות ב־4 או 24 החודשים האחרונים.
- total_diag_sum: סך כל מופעי האבחנות שנרשמו לפי ספירת עמודות *_num_of_diag אשר משקף את כמות הביקורים או הרישומים האבחנתיים בפועל.
- active_diag_types: מספר סוגי אבחנות שונים שהופיעו לפחות פעם אחת. ישמש כמדד למגוון האבחנות הקליני שנצפו עבור המטופלת.

בוצע ניתוח טקסט פשוט חוקים ו-regex:

- clinical_sentiment_score: מדד המבוסס על ניתוח מילות מפתח. ניתנה נקודה שלילית עבור כל מונח "בטוח" (כגון "בדיקה תקינה") ונקודה חיובית עבור כל מונח המצביע על סיכון (כגון "לחץ דם גבוה", "החמרה"). ככל שהציון גבוה יותר – כך סביר שהרופא ציין חשש או בעיה קלינית.
- high_blood_pressure, is_smoker_text – פיצ'רים בינאריים שנבנו בזיהוי ביטויים מתוך הטקסט החופשי. למשל: "מעשנת", "לחץ דם גבוה".

בוצע הסרת פיצ'רים שיכולים לגרום לdata leakage:

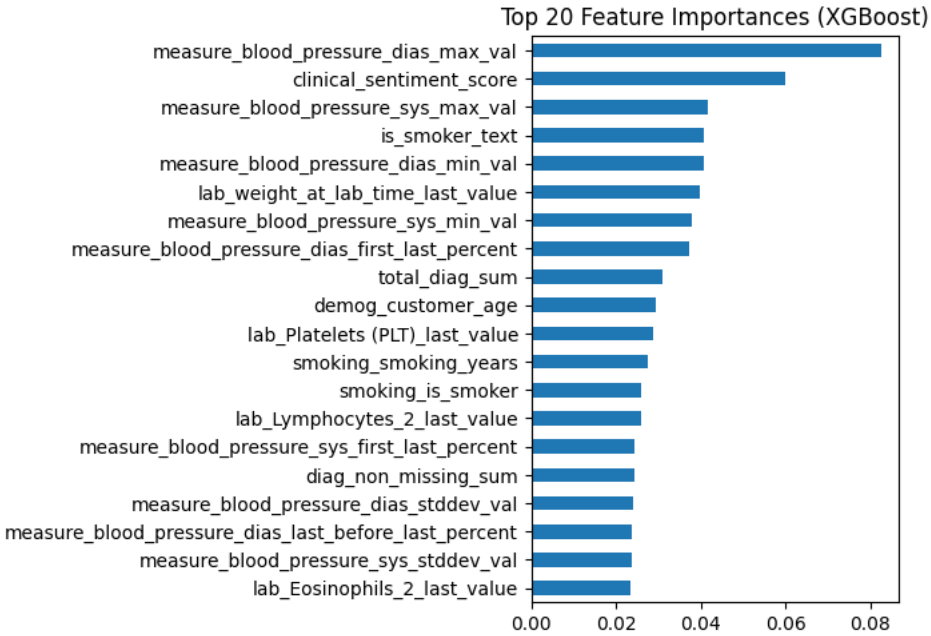
- בוצעה הסרה של משתנים שעלולים לחשוף את משתנה המטרה (לדוגמא הפיצ'רים של _match).

4.4 Feature Selection

בוצעו שלושה צעדים עיקריים לבחירת משתנים:

- הסרת משתנים עם שונות נמוכה: באמצעות variance Threshold הוסרו משתנים עם שונות נמטכה בתוך הפיצ'ר שלא תרמו מידע.
- הסרת משתנים עם קורלציה גבוהה לפי קבוצות (lab, measures, smoking): פיצ'רים בעלי מתאם גבוהה מעל 0.8 (בערך מוחלט) הוסרו.

- שקלול חשיבות פיצ'רים באמצעות xgboost: בתרשים למטה מוצגים 20 הפיצ'רים החשובים ביותר:



פיצ'רים הקשורים ללחץ דם סיסטולי ודיאסטולי הם פיצ'רים חשובים בחיזוי רעלת היריון.
יתר על כן, 2 פיצ'רים שחילצתי: clinical sentiment score (ציון רגשי של טקסטים קליניים) וis_smoker_text דורגו גם הם גבוה. הדבר מצביע על תרומתם האפשרית של נתונים לא מובניים ליכולת הניבוי של המודל.

- מגבלות: בשל נוכחות ערכים חסרים, לא היה ניתן להשתמש בשיטות מתקדמות יותר כגון mutual_information.

4.5 Model Training

לפני בחירת המודל אני רוצה לדעת בנקודה חשובה שקשורה להתפגלות הקלאסים.

הדאטה לא מאוזן כלומר שיעור הנשים אשר בסיכון לרעלת היריון נמוך משמעותית לעומת הנשים הבריאות. יכולתי לאזן את הדאטה באמצעות upsampling לקלאס המיעוט אך בחרתי לא לבצע זאת עקב הסיבות הבאות:

1) מדובר במידע רפואי רגיש: לא רציתי לייצר דגימות סינתטיות שעלולות לא לשקף תרחישים קליניים אמיתיים.
2) שיטות כמו SMOTE יוצרות דגימות חדשות על בסיס שכנים קרובים לפי מרחקים במרחב הפיצ'רים, מה שעלול להכניס רעש או הטיה — במיוחד בהקשרים רפואיים רגישים. לכן העדפתי לבחור מודל שיודע להתמודד עם דאטה לא מאוזן ועם ערכים חסרים.
3) העדפתי להשתמש במודל שיודע להתמודד עם דאטה לא מאוזן בצורה מובנית – מבלי לסנתז דגימות חדשות.

המודל הנבחר: xgoobst ובחרתי בו מהסיבות הבאות:

1) יש לו את היכולת לטפל בקלאס לא מאוזן ע"י היפר-פרמטר scale_pos_weight. המודל יודע להעניש שגיאות של הקלאס הנדיר ולהפחית את ההשפעה של חוסר האיזון
2) יש לו את היכולת להתמודד עם ערכים חסרים
3) קיימים שימושים קליניים אמיתיים במודל xgboost לדוגמא early detection of sepsis.

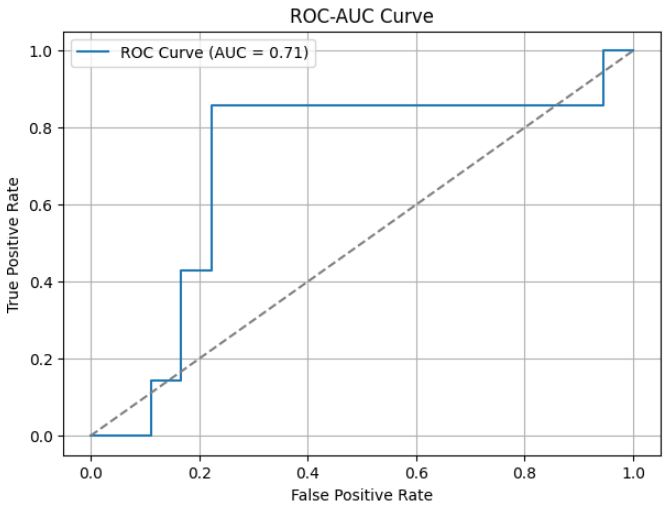
4.6 Model Evaluation

בשל מגבלת משאבים, לא ניתן לבצע בדיקות מעבדה לכל הנשים ההריוניות . לכן, נדרש מהמודל לזהות באופן מדויק ככל הניתן את הנשים בסיכון כאשר ניתן לשלוח לבדיקה רק אחוז קטן מהאוכלוסייה.

במקרה זה, הערכתי את ביצועי המודל על האחוזון העליון של ההסתברויות, אחוזון 99%.

להלן התוצאות:

- Recall 100% - המודל הצליח לזהות את כל שבעת המקרים האמיתיים של רעלת היריון שזה הישג חשוב במיוחד – מנענו מקרים מסכני חיים למטופלת ולעובר שלה.
- Precision 28% - מתוך 25 נשים שהמודל סיווג כחשודות, 18 לא חלו בפועל, שיעור ה FP גבוה יחסית.
- המודל מזהיר יתר על המידה אך לא מפספס אף מקרה מסוכן
- AUC 71% - קיימת הבחנה בינונית-טובה בין הקלאסים. זה מגובה בכך שבחלק הראשון של העקומה, העקומה עולה בחדות – המודל מצליח לזהות נכון חלק מהמקרים בthresholds שמרניים (TPR גבוה ו FPR נמוך)



- בהקשרים רפואיים, Recall גבוה נחשב קריטי גם בריבוי של FP. כמובן שיש מקום לשיפור ה Precision וארחיב על כך בסעיף הבא.

4.7 מסקנות והמלצות

- המודל הנוכחי הצליח לזהות את כל הנשים שאובחנו בפועל (Recall = 1.0), גם תחת מגבלת תקציב. יחד עם זאת, שיעור ה־Precision נמוך מאוד, דבר המעיד על ריבוי false positives – **וזו נקודת חולשה שדורשת שיפור.**
- עקב אילוצי זמן, הפוטנציאל של הנתונים הטקסטואליים לא מומש במלואו. הפיצ'ר clinical_sentiment_score שימש כאינדיקציה כללית בלבד וראינו שהוא מדורג שני ברשימת הפיצ'רים הכי חשובים במודל, כלומר יש עוד המון מה לעשות!
- לדוגמא: שימוש במודלי LLM לצורך חילוץ פיצ'רים/ישויות רפואיות חשובות מהטקסט כגון ממצאים קליניים, בדיקות שתן, דופק, תלונות, המלצות רופא, היסטוריה רפואית וגורמי סיכון.
- ניתן לבחון בניית שני מודלים נפרדים:
  - מודל עצים עבור הנתונים המובנים.
  - מודל מתקדם עבור הנתונים הטקסטואלים הקליניים (לדוג' רשתות נוירונים) ואז נבצע ensemble באמצעות מיצוע ההסתברויות.
- הערה: שימוש ב-LLM בטקסטים קליניים ידרוש רגולציה ומשנה זהירות בהיבט של פרטיות ואבטחת מידע.

- בכל תהליך שיבחר, מומלץ לבצע תהליך ולידציה מסודר למודל, במטרה לשפר ולטייב את תוצאות המודל.
- ✓ שילוב חכם בין נתונים מובנים לבין טקסטים קלינים עשויים לשפר את ביצועי המערכת ולהוריד את שיעור ה-FP ויאפשרו שימוש מדוייק יותר במשאבים הרפואיים.