Ben-Gurion University of the Negev
Department of Industrial Engineering & Management

פרויקט סופי

בנושא: סיווג נשים אשר עתידות לחלות בסכרת ב-5 השנים הקרובות

תוכן

תקציר

מחלת הסכרת היא מחלה קשה לנשים, ועלותה בעולם יקרה מאוד. גילוי מקדים של המחלה עשוי לשמש את החולה ואת מערכת הבריאות. מטרת מחקר זה הינו מציאת מודל חיזוי ובחינת השפעת המשתנים השונים על משתנה המטרה. בעזרת סט נתונים של נשים אינדיאניות בחנו מודלים שונים של רגרסיה לוגיסטית, SVM, Random Forest, Naïve Bayes וכן KNN על מנת למצוא את מודל החיזוי שמביא לתוצאות הטובות ביותר. תחילה ביצענו רגרסיה לוגיסטית ומצאנו מודל מתאים באמצעות צעדים לאחור והמודלים השונים הוערכו ע"י מדד BIC. לאחר מכן, ביצענו התאמות למודל הרגרסיה הלוגיסטית באמצעות טרנספורמציות על המשתנים וביצענו פעם נוספת צעדים לאחור באמצעות מדד BIC. לאחר מכן, הרצנו מודלים שונים מעולם ה-Machine Learning המצוינים מעלה. לטובת כך, עשינו שימוש ב- Cross Validation (K-folds). עשינו שימוש בגישה זו גם על מודלי הרגרסיה על מנת שנוכל להשוות בין כל תוצאות המדדים בין המודלים השונים.

לבסוף קיימנו דיון על בחירת מודל חיזוי של Machine Learning אשר מחד הביא לתוצאות הטובות ביותר, ומאידך מהווה עבורנו "קופסה שחורה". זאת לעומת מודל הרגרסיה שהביא לתוצאות די דומות אך באמצעותו הצלחנו גם להצביע על המשתנים המובהקים ביותר במודל המשפיעים ביותר על משתנה המטרה.

המלצתנו היא לבצע חיזוי באמצעות מודל Random Forest, ובנוסף לציין את המסקנות הנובעות ממודלי הרגרסיה הלוגיסטית שהן שהמשתנים GLU, BMI תורמים רבות על הסיכוי לחלות בסכרת.

מבוא

סֻכָּרֶת (בלטינית: Diabetes mellitus) היא מחלה מטבולית המתאפיינת בריכוז גבוה של גלוקוז בדם (כאשר ריכוז של 180-200 מ"ג גלוקוז בדם נחשב למצב של היפרגליקמיה –עודף סוכר בדם). סכרת נגרמת מייצור לא מספק של אינסולין בגוף או כתוצאה מתגובה לא תקינה של תאי הגוף לאינסולין. 2 סוגים עיקריים של סכרת הינם סכרת מסוג 1, וסכרת מסוג 2. סכרת הריון היא סוג נוסף של סכרת בה נשים שאינן סובלות בדר"כ מהמחלה מפתחות רמות סוכר גבוהות בגוף במהלך תקופת ההריון.
זוהי מחלה קשה במיוחד לנשים שכן היא יכולה להשפיע הן על האם והן על ילדיהן. לנשים כאלו יש סיכוי גבוה יותר ללקות בהתקף לב, להפיל את ההריון או להוליד ילדים עם עם מומים מולדים.
בשל העלייה בשיעור הנשים הסובלות מתסמינים של סכרת, עולה הצורך בטיפול בסוגייה הדחופה והמשמעותית שהיא זיהוי הגורמים המשפיעים על הופעת סכרת אצל אנשים, ויותר מכך - אצל נשים.

עפ"י הערכות, העלות הכוללת של טיפולים באנשים חולי סכרת בשנת 2017 בארצות הברית היתה כ-327 מיליארד דולר. מדובר במגפה עולמית המשפיעה על בריאות החולים ועל הכלכלה.

מטרת המחקר

מטרת המחקר הינה סיווג נשים אשר עתידות לחלות בסכרת ב-5 השנים הקרובות. ברצוננו לבצע חיזוי ובנוסף להבין סיבתיות, קרי אילו משתנים משפיעים על משתנה המטרה, כלומר משפיעים על הסיכוי לחלות בסכרת.

שיטה ותוצאות ראשוניות

לטובת משימת החיזוי, בחרנו במספר מודלים של חיזוי: רגרסיה לוגיסטית (בחרנו ברגרסיה מסוג זה מאחר והמשתנה המוסבר הוא בינארי), Random Forest, SVM (Support Vector Machine), Naïve Bayes, K-Nearest Neighbors.

לאחר מכן, נבצע Evaluation ונשווה בין המודלים השונים באמצעות מס' מדדים, על מנת לבחור את המודל הטוב ביותר.

מבין המודלים הללו, מודל הרגרסיה הלוגיסטית הוא המודל היחיד אשר יוכל לשמש אותנו באופן ישיר גם בהסקת סיבתיות.

תיאור והסבר מפורט של המשתנים במחקר הסטטיסטי

המודל נשען על המשתנים הבלתי תלויים הבאים:

מספר	שם המשתנה	תיאור	סוג המשתנה	טווח ערכים	שם אינדקס
1	Pregnancies#	מספר הריונות של האישה	בדיד (אורדינלי)	[0,17]	PRE
2	Glucose	ריכוז הגלוקוז בדם במהלך שעתיים בבדיקה אורלית.	רציף	[4,199]	GLU
3	Blood Pressure	לחץ דם הנבדקת.	רציף	[24,122]	BP
4	BMI	מדד מסת גוף האישה.	רציף	[18.2,67.1]	BMI
5	Diabetes Pedigree Function	פונקציה אשר נותנת ציון לנראות של האישה לחלות בסכרת לפי היסטוריה משפחתית	רציף	[0,2.5]	DPF
6	Age	גיל האישה	רציף	[21,81]	AGE

טיוב הנתונים

ערכים חסרים

מצאנו כי בשלושה משתנים היו ערכים חסרים: 5 ערכים חסרים למשתנה Glucose, 11 ערכים חסרים למשתנה BMI, 35 ערכים חסרים למשתנה לחץ דם. מאחר ומדובר בכמות קטנה יחסית של רשומות עם ערכים חסרים ולא ידועה לנו מאיזו התפלגות הגיעו הנתונים, החלטנו למחוק את הרשומות עם הערכים החסרים. לאחר הסינון נותרנו עם 724 רשומות.

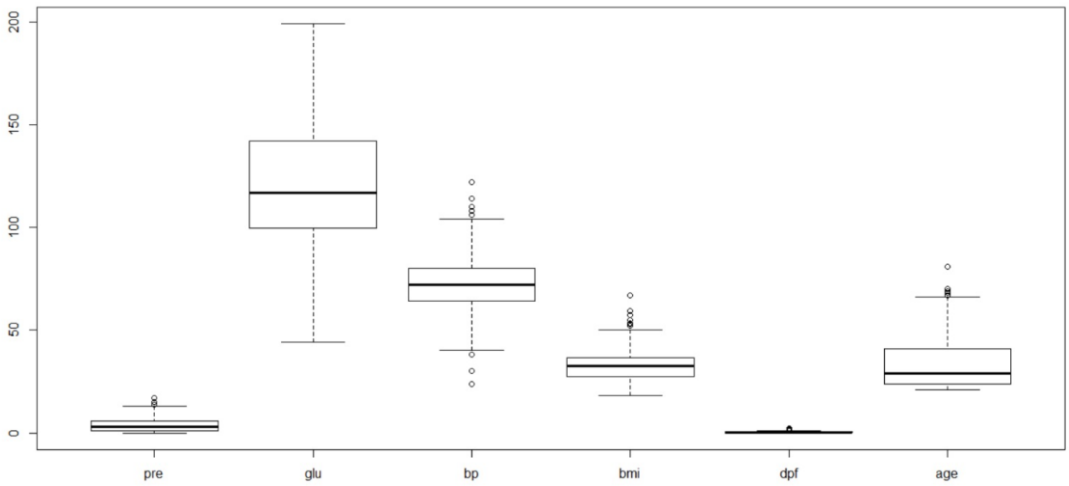
ערכים חריגים

על מנת להבין האם ישנם ערכים חריגים בנתונים, אנו מציגים תרשים-קופסה עבור כל משתנה. ניתן לראות בתרשים להלן כי קיימים ערכים חריגים, לפי האופן בו מוגדר תרשים-קופסה. ערכים חריגים מוגדרים כבאלו מעל הגבול העליון של התרשים שמוגדר Q3+1.5*IQR או מתחת לגבולו התחתון שמוגדר להיות Q1-1.5*IQR, כאשר IQR הינו גודל הקופסא (Q3-Q1). Qi הוא הרבעון ה-i.

עבור המשתנה **מס' ההריונות,** ניתן לראות כי קיימים ערכים חריגים וגבוהים. מהתבוננות בנתונים, נמצאו ערכים אשר בהתבסס על ידע כללי – נראים כחריגים ולא הגיוניים (כגון מס' הריונות ששווה ל-17), אשר הוחלט להסירם. מדובר בהסרה של 4 ערכים, כלומר הסרה של 4 רשומות בסה"כ. תשומת לב לכך שמדובר בבסיס נתונים המתבסס על אוכלוסייה שאיננה מוכרת לנו (pima

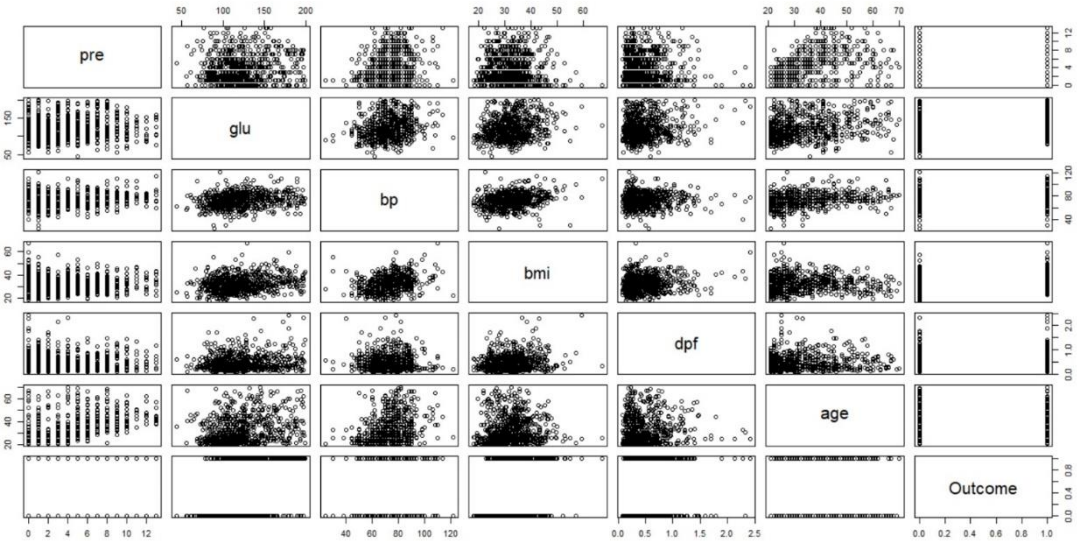
Indians), ולכן הוחלט שלא להסיר ערכים שיתכן והם חריגים ביחס לאוכלוסייה שלנו, אך אינם חריגים ביחס לנתונים (לפי תרשים הקופסה), מתוך ההבנה שיתכן כי הדבר נובע מהבדלים בין האוכלוסיות.
כמו כן, הוחלט להסיר רשומה אשר כוללת את הערך "80" עבור המשתנה **גיל**, מאחר ומדובר בערך חריג, גם ביחס לשאר הערכים החריגים, וכן כי מדובר ברשומה בודדת.

לאחר שלב טיוב הנתונים, בסיס הנתונים שלנו כולל **719 רשומות**, מתוכן 249 נשים המסווגות כחולות, המהווה כ-34.6% מכלל הנתונים.



מטריצת קורלציות

תחילה נציג את מטריצת הקורלציות בכדי לקבל תמונת מצב ראשונית על המשתנים:



	pre	glu	bp	bmi	dpf	age	Outcome
pre	1.000000000	0.1250400	0.219486170	0.001479778	-0.023748547	0.55804269	0.2085648
glu	0.125039985	1.0000000	0.226940083	0.222664286	0.137866914	0.26340677	0.4880508
bp	0.219486170	0.2269401	1.000000000	0.288467856	-0.001085197	0.32899471	0.1690805
bmi	0.001479778	0.2226643	0.288467856	1.000000000	0.154929855	0.02280682	0.2956637
dpf	-0.023748547	0.1378669	-0.001085197	0.154929855	1.000000000	0.02389278	0.1880753
age	0.558042689	0.2634068	0.328994712	0.022806818	0.023892784	1.00000000	0.2479447
Outcome	0.208564769	0.4880508	0.169080503	0.295663653	0.188075313	0.24794470	1.0000000

ניתן לראות כי הקורלציה הגבוהה ביותר עם משתנה המטרה (חולה בסכרת או לא) היא של המשתנה רמת הגלוקוז בדם – 0.488. המשתנה השני שלו הוא ה-BMI לו יש קורלציה של 0.295. ניתן לראות כי אין משתנה שלו יש קורלציה משמעותית עם משתנה המטרה. כמו כן, הקורלציה הגבוהה ביותר בין משתנים היא בין המשתנה מספר ההריונות לבין גיל האישה – 0.558. הדבר הגיוני שכן מספר ההריונות הוא מספר נצבר עם השנים. הקורלציה עדיין לא ממש גבוהה מהסיבה שמספר ההריונות נעצר בשלב מסוים (באיזור גיל 40-50).

המודלים בהם עשינו שימוש

רגרסיה לוגיסטית

השלב הבא היה ליצור מודל של רגרסיה לוגיסטית באמצעות הפונקציה GLM ב-R. מאחר ומשתנה המטרה הוא בינארי עשינו שימוש במשפחת "binomial".

המודל הכללי הוא:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 * PRE + \beta_2 * GLU + \beta_3 * BP + \beta_4 * BMI + \beta_5 * DPF + \beta_6 * AGE + \varepsilon$$

π – ההסתברות לחלות בסכרת ב-5 השנים לאחר המדידה (איסוף התצפית).

הנחות המודל:

המשתנה התלוי בינארי $Y_i \in \{0,1\} \quad \forall i$

התצפיות בלתי תלויות האחת בשנייה $cov(Y_i, Y_j) = 0 \quad \forall i,j$

אי קיום מולטיקולינאריות

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R_j^2} < 10$$

R_j^2 הוא מקדם המתאם המרובה בריבוע בין המשתנה המסביר ה-j לבין שאר המשתנים המסבירים.

לינאריות של לוג ההסתברויות של המשתנים הבלתי-תלויים במשתנה המטרה

$$\ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \alpha + \beta * X$$

מדגם יחסית גדול – כלל אצבע של כמות תצפיות לכל הפחות פי 10 מכמות המשתנים

$$n > 10 * Number\_Of\_Variables$$

השגיאות מתפלגות נורמלית $\varepsilon \sim N(0, \sigma^2)$

השערות המודל:

H_0: $\alpha = \beta_i = 0 \,, \forall i$

H_1: $else$

המודל שהתקבל:

$$\ln\left(\frac{\pi}{1-\pi}\right) = -8.962 + 0.117 * PRE + 0.035 * GLU - 0.008 * BP + 0.091 * BMI + 0.961 * DPF + 0.017 * AGE$$

```
Call:
glm(formula = Outcome ~ pre + glu + bp + bmi + dpf + age, family = "binomial",
    data = data1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8056  -0.7219  -0.4037   0.7203   2.3996

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.958438   0.820234 -10.922  < 2e-16 ***
pre          0.107591   0.034192   3.147  0.00165 **
glu          0.035141   0.003615   9.722  < 2e-16 ***
bp          -0.008934   0.008636  -1.035  0.30087
bmi          0.089354   0.015732   5.680 1.35e-08 ***
dpf          0.970200   0.306661   3.164  0.00156 **
age          0.019605   0.009988   1.963  0.04967 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 922.52  on 718  degrees of freedom
Residual deviance: 668.84  on 712  degrees of freedom
AIC: 682.84

Number of Fisher Scoring iterations: 5
```

מכיוון שראינו שאין קורלציה גבוהה באופן משמעותי (גדול מ-0.7) בין המשתנים המסבירים, איננו חושבים שיש בעיית מולטיקולינאריות בין משתנים (ייבדק סטטיסטית בהמשך).

בדיקת הנחות המודל והתמודדות עמן:

1. המשתנה התלוי בינארי

המשתנה התלוי בינארי מעצם הגדרת: 1 – חולה בסכרת, 0 – לא חולה בסכרת.

2. התצפיות בלתי תלויות האחת בשנייה

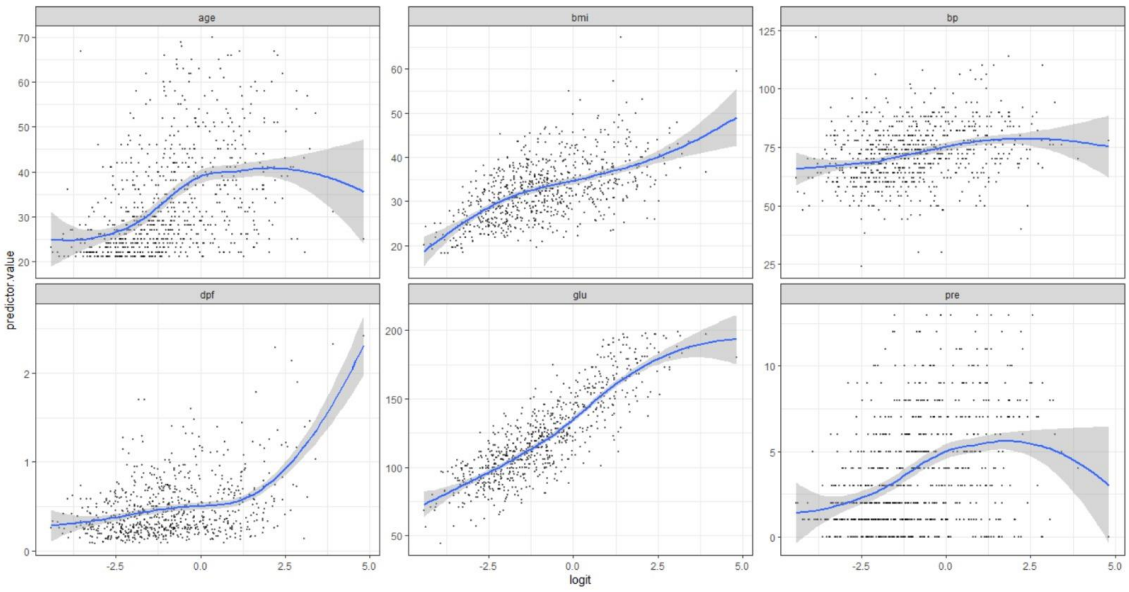
ניתן להניח שמאחר והתצפיות מייצגות נשים שונות, המדדים שלהם בלתי תלויים האחד בשני.

3. אי קיום מולטיקולינאריות

בפלט הבא ניתן לראות את ערכי VIF של הבטות של כלל המשתנים, אשר נמוכים מ-10, ולכן ניתן להסיק שאין בעיית מולטיקולינאריות בנתונים.

```
     pre      glu       bp      bmi      dpf      age
1.427313 1.049370 1.228112 1.134125 1.007664 1.544968
```

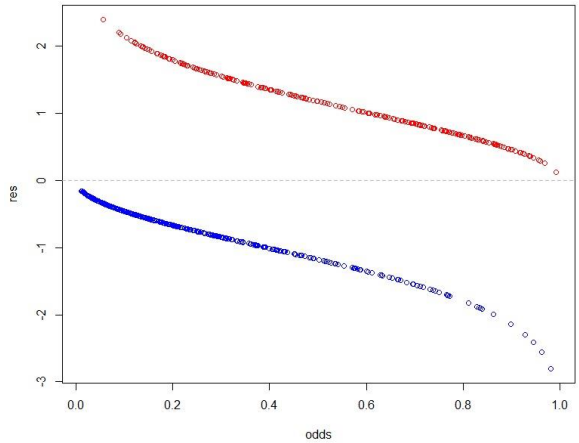
4. לינאריות של לוג ההסתברויות של המשתנים הבלתי-תלויים במשתנה המטרה



<u>ניתוח הפלט לעיל:</u> המשתנים bmi, glu, bp, dpf נראים לינאריים במשתנה המטרה.
עם זאת, המשתנים pre, age לא נראים לינאריים במשתנה המטרה, עבורן ניסינו מס' טרנספורמציות, אשר יפורטו בהמשך.

5. <u>מדגם יחסית גדול – כלל אצבע של כמות תצפיות לכל הפחות פי 10 מכמות המשתנים</u>
אכן מתקיים עבור סט הנתונים שלנו.
6. <u>השגיאות מתפלגות נורמלית</u>

<u>להלן תרשים שאריות:</u>



<u>ניתוח הפלט לעיל:</u> הסיבה שיש שני קווים של שאריות היא בגלל שאנו חוזים הסתברות של משתנה לקבל את הערך 0 או 1 (חולה בסכרת או לא). אם הערך שהוחלט הוא 0, אז החיזוי היה גבוה מהערך (כי חזינו הסתברות), לכן השאריות חייבות להיות שליליות (הנקודות הכחולות). בדיוק להפך אם ערך המטרה הוא 1 – השאריות חייבות להיות חיוביות (הנקודות האדומות). הצורה הלינארית של שני קווי השגיאות יכולה לאשש את הנחת הלינאריות של המודל.

תוצאות הניתוחים הסטטיסטים

נבצע מבחן LRT על מנת לבחון האם השערות המודל מתקיימות. במידה ונדחה את השערת האפס, אזי נאמר שלפחות אחת מההשערות לא מתקיימת.

נחשב את סטטיסטי המבחן:

$$\chi^2_{st} = Null\ Deviance - Residual\ Deviance = 922.52 - 668.84 = 253.686 > 12.591 = \chi^2_{cr} = \chi^2_{0.95,6}$$

כאשר ד"ח הן הפרש מספר הפרמטרים בין שני המודלים.

נדחה את השערת האפס, כלומר לפחות אחת מההשערות אינן מתקיימות. ה-P_value של המודל הינו 0, כלומר ממש קטן מ-0.05. כלומר, ניתן לומר שהמשתנים המסבירים אכן יכולים להסביר את $\ln(odds\ ration)$. אף על פי כן, נרצה לשפר את המודל שכן ניתן לראות שלפחות אחד מהמשתנים איננו מובהק, בהינתן שאר המשתנים. לכן, נבצע מודל רגרסיה לאחור עפ"י מדד BIC.

מודל רגרסיה לאחור עפ"י מדד BIC:

כפי שניתן לראות בפלט למטה, בהרצת המודל לאחור התחלנו עם המודל המלא שכולל את כלל המשתנים והחותך. בכל איטרציה, נבחן הוצאת משתנה מהמודל תוך שיפור (הקטנת) מדד BIC:

$$\mathrm{BIC} = \mathrm{n}\log\left(\frac{\mathrm{SSE}}{\mathrm{n}}\right) + \ln(\mathrm{n}) \times (\mathrm{p}+1)$$

נציין כי אמנם בפלטים הבאים רשום כי מדובר במשתנה AIC, אך אכן מדובר במשתנה BIC כפי שצוין.

באיטרציה הראשונה נבחר להוציא את המשתנה bp, שכן כפי שניתן לראות, במודל שלא כולל אותו התקבל ה-BIC הנמוך ביותר.

```
> step(mylogit, direction = "backward", k=log(nrow(data1)))
Start:  AIC=714.88
Outcome ~ pre + glu + bp + bmi + dpf + age

        Df Deviance    AIC
- bp     1   669.91 709.38
- age    1   672.67 712.14
<none>       668.84 714.88
- pre    1   678.91 718.38
- dpf    1   679.19 718.66
- bmi    1   703.67 743.14
- glu    1   788.03 827.49
```

באיטרציה השנייה, ניתן לראות כי מדד BIC הנמוך ביותר התקבל עבור המודל לאחר הוצאת המשתנה age, ולכן נבחר להוציא גם אותו מהמודל.

```
Step:  AIC=709.38
Outcome ~ pre + glu + bmi + dpf + age

        Df Deviance    AIC
- age    1   673.01 705.90
<none>       669.91 709.38
- pre    1   679.64 712.53
- dpf    1   680.57 713.46
- bmi    1   704.42 737.31
- glu    1   788.04 820.93
```

באיטרציה השלישית מדד ה-BIC הנמוך ביותר התקבל עבור none, כלומר עבור המודל הקיים, ללא הוצאת משתנה נוסף.

```
Step:  AIC=705.9
Outcome ~ pre + glu + bmi + dpf

       Df Deviance    AIC
<none>      673.01 705.90
- dpf   1   683.76 710.07
- pre   1   696.14 722.45
- bmi   1   706.18 732.49
- glu   1   806.07 832.39
```

בסופו של דבר, לאחר הוצאת שני המשתנים: bp, age התקבל המודל הבא:

$$\ln\left(\frac{\pi}{1-\pi}\right) = -8.941 + 0.137 * PRE + 0.036 * GLU + 0.083 * BMI + 0.982 * DPF$$

משמעות החותך במודל הוא ה-$log\ odds$ במצב שבו כל המשתנים שווים ל-0, אך מקרה זה איננו הגיוני במציאות (כך למשל, לא יתכן שערך ה-BMI של אדם הינו 0).

משמעות המקדם של כל משתנה הוא ההפרש בין ה-$log\ odds$ בין שני ערכים עוקבים של כל משתנה. במודל, ככל שמגדילים את ערך משתנה X – ההסתברות לחלות P(x) גדלה. כך למשל, עבור המשתנה גלוקוז – ככל שהערך של הגלוקוז גדל, ככה הסיכוי לחלות גדל וכו'. אמירה זו מתיישבת עם היכרותנו עם המשתנים בבעיה במציאות.

כפי שציינו תחת הנחת הלינאריות של לוג ההסתברויות של המשתנים הבלתי-תלויים במשתנה המטרה, ניתן לראות שחלק מהמשתנים אינם לינאריים במשתנה המטרה. על כן, נבצע טרנספורמציות על משתנים אלו.

רגרסיה לוגיסטית עם טרנספורמציות

לאחר הסתכלות על גרפי הלינאריות ליניאריות של לוג ההסתברויות של המשתנים הבלתי-תלויים במשתנה המטרה, ביצענו טרנספורמציה עבור המשתנים age, pre, dpf. הוחלט לבצע טרנספורמציה על המשתנים הללו שכן כפי שניתן לראות בתרשים של הלינאריות של לוג ההסתברויות של המשתנים הבלתי-תלויים במשתנה המטרה (ר' עמ' 7), בגרפים המשויכים אליהם הקווים אינם ליניאריים.

ניסינו מספר טרנספורמציות כגון העלאה בחזקת 2 ו-3, שורש, exp, sqrt, log.

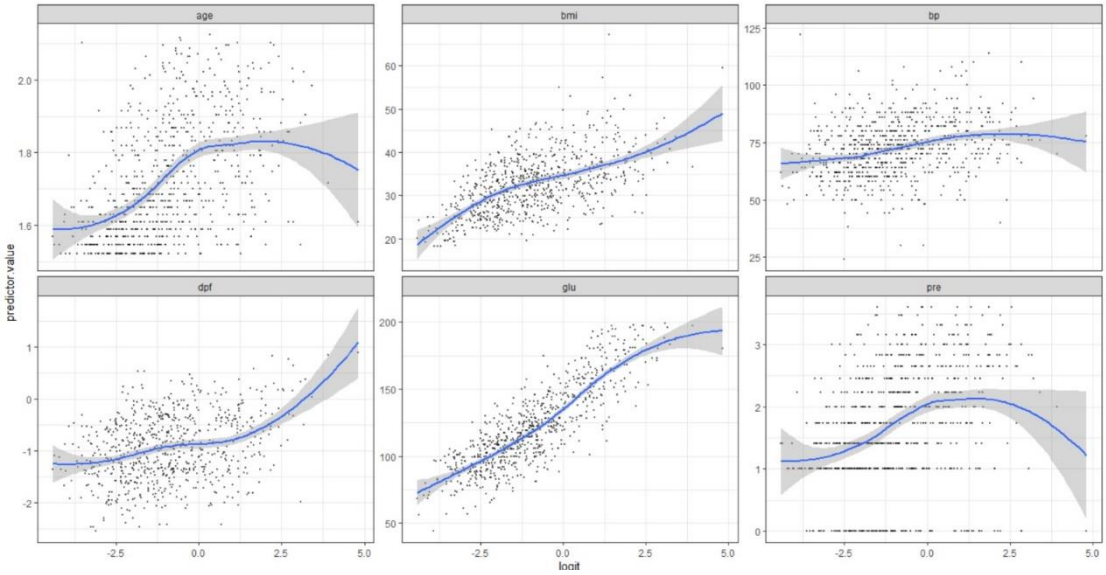
הטרנספורמציה שנתנה את המודל בעל ה-BIC הקטן ביותר היא:

$$PRE' = \sqrt{PRE}$$

$$DBF' = \log(DBF)$$

$$AGE = \log(AGE)$$

לאחר הטרנספורמציה הגרפים שקיבלנו נראים כך :



ניתן לראות כי משתנה הdpf הינו מתנהג בצורה יותר ליניארית. המשתנים pre ו-age עדיין לא נראים ליניאריים, אך דווקא נתנו תוצאות מוצלחות יותר מהמצב הקודם ומטרנספורמציות אחרות שניסינו.

לאחר ביצוע הטרנספורמציה הרצנו מודל רגרסיה לוגיסטית פעם נוספת וקיבלנו את התוצאות הבאות:

```
Call:
glm(formula = Outcome ~ sqrt(pre) + glu + bp + bmi + log(dpf) +
    log(age), family = "binomial", data = data1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5348  -0.7116  -0.3929   0.7123   2.4004

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.338355   1.339055  -8.467  < 2e-16 ***
sqrt(pre)     0.263855   0.120826   2.184 0.028980 *
glu           0.034664   0.003610   9.603  < 2e-16 ***
bp           -0.009936   0.008662  -1.147 0.251383
bmi           0.090742   0.015791   5.746 9.12e-09 ***
log(dpf)      0.568736   0.156293   3.639 0.000274 ***
log(age)      1.175850   0.379562   3.098 0.001949 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 922.52  on 718  degrees of freedom
Residual deviance: 663.03  on 712  degrees of freedom
AIC: 677.03

Number of Fisher Scoring iterations: 5
```

<u>מודל רגרסיה לאחור עפ"י מדד BIC:</u>

הרצנו פעם נוספת את מודל הרגרסיה בצעדים לאחור על מנת לנסות ולשפר את מדד ה-BIC כפי שביצענו במודל לפני ביצוע הטרנספורמציות:

באיטרציה הראשונה נבחר להוציא את המשתנה bp, שכן כפי שניתן לראות, במודל שלא כולל אותו התקבל ה-BIC הנמוך ביותר.

```
Start:  AIC=709.08
Outcome ~ sqrt(pre) + glu + bp + bmi + log(dpf) + log(age)

            Df Deviance    AIC
- bp         1   664.35 703.82
- sqrt(pre)  1   667.91 707.38
<none>           663.03 709.08
- log(age)   1   672.67 712.14
- log(dpf)   1   676.71 716.18
- bmi        1   698.88 738.35
- glu        1   779.04 818.51
```

באיטרציה השנייה, ניתן לראות כי מדד BIC הנמוך ביותר התקבל עבור המודל לאחר הוצאת המשתנה $\sqrt{pre}$, ולכן נבחר להוציא גם אותו מהמודל.

```
Step:  AIC=703.82
Outcome ~ sqrt(pre) + glu + bmi + log(dpf) + log(age)

            Df Deviance    AIC
- sqrt(pre)  1   669.24 702.12
<none>           664.35 703.82
- log(age)   1   672.76 705.65
- log(dpf)   1   678.41 711.30
- bmi        1   699.40 732.29
- glu        1   779.04 811.93
```

באיטרציה השלישית מדד ה-BIC הנמוך ביותר התקבל עבור none, כלומר עבור המודל הקיים, ללא הוצאת משתנה נוסף.

```
Step:  AIC=702.12
Outcome ~ glu + bmi + log(dpf) + log(age)

           Df Deviance    AIC
<none>          669.24 702.12
- log(dpf)  1   682.85 709.16
- log(age)  1   693.42 719.74
- bmi       1   702.36 728.67
- glu       1   782.48 808.79
```

ניתן לראות כי כרגע, מדד הBIC הנמדד עומד על 702.12 (לעומת 705.9 שקיבלנו לפני הטרנספורמציה), כלומר **באמצעות ביצוע הטרנספורמציות הצלחנו לשפר את המודל.**

<u>המודל שהתקבל</u>

המודל הסופי לאחר הטרנספורמציות:

$$\ln\left(\frac{\pi}{1-\pi}\right) = -12.282 + 0.033 * GLU + 0.082 * BMI + 0.564 * \log(dpb) + 1.48 * \log(age)$$

```
Call:
glm(formula = Outcome ~ glu + bmi + log(dpf) + log(age), family = "binomial",
    data = data1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6563  -0.7132  -0.3958   0.6952   2.4901

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.282948   1.277068  -9.618  < 2e-16 ***
glu           0.033893   0.003556   9.530  < 2e-16 ***
bmi           0.082218   0.014877   5.527 3.26e-08 ***
log(dpf)      0.564553   0.155520   3.630 0.000283 ***
log(age)      1.480436   0.304926   4.855 1.20e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 922.52  on 718  degrees of freedom
Residual deviance: 669.24  on 714  degrees of freedom
AIC: 679.24

Number of Fisher Scoring iterations: 5
```

בחירת מודל עבור כל אלגוריתם חיזוי

עבור כל אחד ממודלי החיזוי שנבחרו, חילקנו את סט הנתונים שלנו ל80% סט אימון, 20% סט בחינה. באמצעות סט האימון בחרנו את המודל המועדף (כיוונון היפר-פרמטרים) באמצעות שיטת K-folds, כאשר K=10. תוצאת שלב האימון באופן זה הינה מודל נבחר עבור כל אלגוריתם חיזוי. לבסוף, עבור כל אלגוריתם כזה – אימנו את סט האימון עם המודל שנבחר בשלב הקודם, ובחנו אותו על סט הבחינה. כך יכולנו לקבל אחוזי דיוק ומדדים נוספים לטובת השוואה בין המודלים.

את בחירת המודל בדרך זו ביצענו עבור שני המודלים המפורטים לעיל: רגרסיה לוגיסטית ורגרסיה לוגיסטית לאחר ביצוע טרנספורמציות. בנוסף אליהם, בחנו מודלים נוספים: Random Forest, KNN, SVM, NB.

רגרסיה לוגיסטית

באמור לעיל, מודל הרגרסיה הלוגיסטית שנבחר כולל את המשתנים: PRE, GLU, BMI, DPF וחותך. בכדי להשוות בין המודלים של שיטות החיזוי השונות הנ"ל לבין מודל הרגרסיה הלוגיסטית שנבחר, אימנו את סט האימון שלנו גם כן באמצעות שיטת K-folds, כאשר K=10. בכל איטרציה התקבלו ערכי $\hat{\beta}$ שונים עבור המודל.

מיצענו את ערכי ה-$\hat{\beta}$ השונים שהתקבלו על מנת להכריע בנוגע לערכי ה-$\hat{\beta}$ הסופיים של המודל.

	Intercept	PRE	GLU	BMI	DPF
[1,]	-9.835996	0.1354519	0.03960223	0.09165314	1.3172278
[2,]	-9.865790	0.1420013	0.03661386	0.10654819	1.0661784
[3,]	-9.140200	0.1497784	0.03441458	0.08878171	1.2270404
[4,]	-9.266118	0.1384734	0.03612652	0.08879582	1.1844559
[5,]	-9.106599	0.1101815	0.03854210	0.08207094	0.9043389
[6,]	-9.858444	0.1164664	0.03877775	0.10368996	0.8684864
[7,]	-9.762980	0.1414577	0.03725607	0.09807429	1.2131007
[8,]	-9.349791	0.1285459	0.03700715	0.09076467	0.9810906
[9,]	-9.757194	0.1355452	0.04262818	0.07979743	1.0639074
[10,]	-9.008641	0.1233593	0.03545787	0.08930963	0.9280532

ממוצעי ערכי ה-$\hat{\beta}$ הם:

```
[1] -9.49517521  0.13212610  0.03764263  0.09194858  1.07538798
```

כלומר המודל הסופי הוא:

$$\ln\left(\frac{\pi}{1-\pi}\right) = -9.4951 + 0.1321 * PRE + 0.0376 * GLU + 0.0919 * BMI + 1.0753 * DPF$$

לאחר הרצת המודל, התקבלו תוצאות המדדים: Accuracy – 0.7343, F1 Score – 0. 54476, TPR – 0. 4792, FPR – 0. 1368.

רגרסיה לוגיסטית – לאחר ביצוע הטרנספורמציות:

בדיוק באותו האופן שבוצע עבור מודל הרגרסיה הלוגיסטית, גם במקרה הזה אימנו את סט האימון שלנו גם כן באמצעות שיטת K-folds, כאשר K=10. בכל איטרציה התקבלו ערכי $\hat{\beta}$ שונים עבור המודל. מיצענו את ערכי ה-$\hat{\beta}$ השונים שהתקבלו על מנת להכריע בנוגע לערכי ה-$\hat{\beta}$ הסופיים של המודל.

	Intercept	GLU	BMI	Log(DPF)	Log(AGE)
[1,]	-13.30469	0.03752491	0.08939182	0.7580803	1.627141
[2,]	-14.03010	0.03458292	0.10417401	0.6572794	1.772380
[3,]	-12.83457	0.03289372	0.08531778	0.7846742	1.695541
[4,]	-12.66357	0.03406530	0.08798582	0.7170465	1.565314
[5,]	-13.09345	0.03679210	0.08428593	0.6193268	1.596595
[6,]	-13.97351	0.03659847	0.10347084	0.5832590	1.661168
[7,]	-13.77257	0.03493532	0.09746593	0.7184092	1.759388
[8,]	-13.52752	0.03487893	0.08837119	0.5959705	1.738925
[9,]	-13.80382	0.04068982	0.07687715	0.6715005	1.736684
[10,]	-12.49898	0.03353454	0.08580501	0.6107426	1.535158

ממוצעי ערכי ה-$\hat{\beta}$ הם:

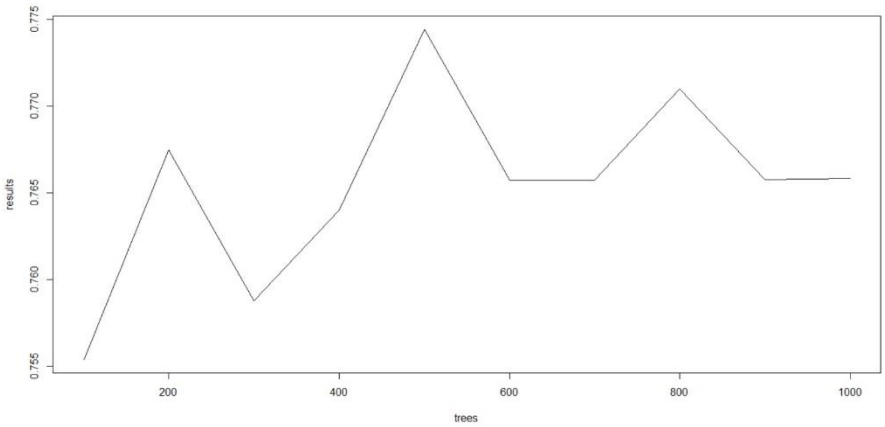
```
[1] -13.35027765   0.03564960   0.09031455   0.67162891   1.66882946
```

כלומר המודל הסופי הוא:

$$\ln\left(\frac{\pi}{1-\pi}\right) = -13.3502 + 0.0356 * GLU + 0.0903 * BMI + 0.6716 * \log(DPF) + 1.6688 * \log(AGE)$$

לאחר הרצת המודל, התקבלו תוצאות המדדים: Accuracy – 0.7413, F1 Score – 0.5843, TPR – 0.5417, FPR – 0.1579.

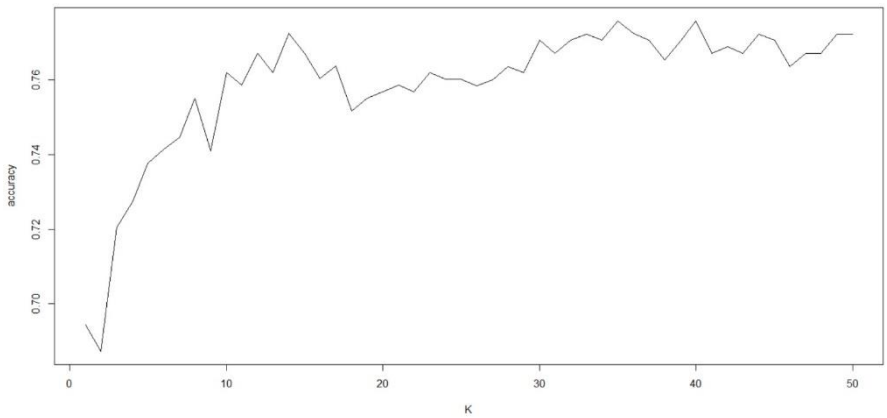
Random Forest



כפי שניתן לראות בתרשים, מודל Random Forest הנבחר כולל 500 עצים.

לאחר הרצת המודל, התקבלו תוצאות המדדים: Accuracy – 0.7692, F1 Score - 0.6526, TPR – 0.6078, FPR – 0.7692.

(KNN) K-Nearest Neighbors



כפי שניתן לראות בתרשים, במודל KNN הנבחר K=35.

לאחר הרצת המודל, התקבלו תוצאות המדדים: Accuracy – 0.7133, F1 Score – 0.506, TPR – 0.4374, FPR – 0.1473.

SVM

הוחלט לבחון מודלים בעלי Kernel שונה: Kernel לינארי (מסווג היוצר מפריד בצורת קו ישר), וכן Kernel='rbf', כלומר רדיאלי.

Linear SVM

מודל SVM לינארי הנבחר כולל את הפרמטר C=0.1.

לאחר הרצת המודל, התקבלו תוצאות המדדים: Accuracy – 0.7343, F1 Score – 0.5366, TPR – 0.4583, FPR – 0.1263.

Radial SVM

מודל SVM רדיאלי הנבחר כולל את הפרמטר C=0.1.

לאחר הרצת המודל, התקבלו תוצאות המדדים: Accuracy – 0.7273, F1 Score – 0.4935, TPR – 0.3958, FPR – 0.1053.

(NB) Naive Bayes

לאלגוריתם NB אין היפר פרמטרים אותם יש לכוונן, שכן הוא מניח הנחות חזקות עליהן הוא מתבסס.

לאחר הרצת המודל, התקבלו תוצאות המדדים: Accuracy – 0.7063, F1 Score – 0.533, TPR – 0.5, FPR – 0.1895.

תוצאות סופיות – השוואת תוצאות שלב ה-Evaluation

על מנת להשוות בין המודלים הנבחרים של כל אחד מאלגוריתמי החיזוי, נתבונן בטבלת ה-evaluation הבאה:

	Random Forest	KNN	Linear SVM	Radial SVM	NB	Logistic Regression	Logistic Regression - After Transformation
ACC	0.7692	0.7133	0.7343	0.7273	0.7063	0.7343	0.7413
F1 Score	0.6526	0.506	0.5366	0.4935	0.533	0.54476	0.5843
TPR	0.6078	0.4375	0.4583	0.3958	0.5	0.4792	0.5417
FPR	0.1413	0.1473	0.1263	0.1053	0.1895	0.1368	0.1579

בחרנו להציג, עבור כל אחד מהמודלים, את המדדים הבאים:

- Accuracy (ACC) – אחוזי הדיוק של המודל.
  מחושב כך:
  $$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$
- TPR (TRUE POSITIVE RATE) – אחוז הנשים המסווגות כחולות בסכרת ע"י המודל מתוך סך הנשים החולות בסכרת.
  מחושב כך:
  $$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} = 1 - FNR$$
- FPR (FALSE POSITIVE RATE) – אחוז הנשים המסווגות כחולות בסכרת ע"י המודל מתוך סך הנשים שאינן חולות בסכרת.
  מחושב כך:
  $$FPR = \frac{FP}{N} = \frac{FP}{FP + TN}$$
- F1 Score – ממוצע הרמוני של TPR, PPV.
  כאשר PPV הוא מדד המציין את אחוז הנשים החולות שאכן סווגו כך, מתוך סך הנשים המסווגות כחולות ע"י המודל.
  מחושב כך:
  $$F1 = 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$$

נשים לב שמאחר ומדובר במחלה מסכנת, טעות בסיווג אישה חולה (סיווגה כלא-חולה) הוא קריטי יותר מאשר טעות בסיווג אישה שאיננה חולה (שסווגה כחולה). לכן, המדדים הרלוונטיים ביותר עבורנו בבחירת המודל הינם F1, TPR.

דיון ומסקנות

ניתן לראות בטבלה כי עבור המדדים F1 Score, TPR (וכן Accuracy) המודל הנבחר הוא מודל Random Forest עם 500 עצים. מודל חיזוי כזה אמנם נותן פעמים רבות את התוצאות הטובות ביותר, אך מדובר ב"קופסה שחורה" עבור החוקר.

בהתאם לכך, נרצה להתייחס גם לשני המודלים של הרגרסיה הלוגיסטית (לפני ואחרי ביצוע הטרנספורמציות). שני המודלים הללו קיבלו תוצאות שאינן נמוכות בהרבה ממודל Random Forest, והיתרון שלהם על פניו הוא הסקת הסיבתיות של המשתנים למשתנה המטרה באופן ישיר. ככל הנראה, בהינתן אינסוף זמן וצעדים יכולנו להגיע בסופו של דבר לאותן התוצאות בביצוע הרגרסיה הלוגיסטית עם טרנספורמציות שונות על הדאטה ועם מודל החיזוי הנבחר.

בשני המודלים ניתן לראות כי המשתנה המובהק ביותר הוא משתנה GLU המציין את ריכוז הגלוקוז בדם במהלך שעתיים בבדיקה אורלית של נבדקת. כלומר, זהו המשתנה המשפיע ביותר על משתנה המטרה בהינתן שאר המשתנים במודל. אמירה זו מתיישבת עם הבנתנו את המציאות לאחר ביצוע מחקר בתחום שכן אחוז הגלוקוז בדם משפיע ישירות על הסיכוי לחלות בסכרת.

משתנה נוסף שיצא מובהק בשני המודלים הוא משתנה ה-BMI המציין את מדד מסת גוף האישה.

שני האומדים של המקדמים של המשתנים בשני מודלים הרגרסיה הם חיוביים, ולכן משפיעים באופן חיובי על משתנה המטרה. כלומר, ככל שהם גדלים – הסיכוי לחלות גדל.