MATEMATIKOS IR INFORMATIKOS
FAKULTETAS

BINARIS ATSAKO MODELIS

Laboratorinis darbas

Atliko: Simona Gelžinytė,
Ugnė Kniukštaitė,
Laineda Morkytė,
Austėja Valeikaitė,
duomenų mokslas 3 k. 2gr.

Vilnius, 2023

TURINYS

1. ĮVADAS

1.1 Tyrimo tikslai

Parinkti tinkamą regresijos modelį.

1.2 Tyrimo uždaviniai

- Atlikti pirminę duomenų analizę;
- Patikrinti modelio prielaidas;
- Sukonstruoti modelį;
- Modelio tinkamumo įvertinimas;
- Apibendrinti gautus rezultatus, pateikti išvadas.

1.3 Duomenys ir programinė įranga

Pasirinktas duomenų rinkinys apie diabetą. Pateikti įvairūs medicininiai požymiai, kurie prognozuoja, ar pacientas serga diabetu, ar ne. Buvo tirtos tik moterys, jaunesnės nei 21 metų. Priklausomas kintamasis – rezultatas, rodantis ar pacientas serga diabetu (1 – serga, 0 – neserga) ir 8 kovariantės:

- Nėštumas – nėštumų skaičius;
- Gliukozė – gliukozės koncentracija plazmoje;
- Kraujo spaudimas – diastolinis kraujo spaudimas (mm Hg);
- Odos storis – tricepso odos raukšlės storis (mm);
- Insulinas – 2 valandų serumo insulinas (mU/ml);
- KMI – kūno masės indeksas;
- Diabeto kilmės funkcija;
- Amžius.

Iš viso yra 768 stebėjimų, praleistų reikšmių nėra. Tyrimo metu naudota programinė įranga: „SAS“, „R“ ir „Python“.

2. BINARINIO ATSAKO MODELIS

Parinkus binarinio atsako modelį pereiname visus modelio parinkimo etapus:

1. Pradinė analizė – vizualiai patikriname, ar nėra išskirčių. Jų nustatymui naudojame Kuko matą;
2. Prielaidų tikrinimas – paklaidų nepriklausomumas, multikolinearumas;
3. Reikšmingų kovariančių atranka;
4. Parametrų įvertinimas;
5. Modelio tinkamumo analizė;
6. Ryšių tarp kintamųjų interpretavimas.

2.1 Binarinio atsako modelis naudojant SAS

Prieš atliekant analizę, svarbu pabrėžti, jog duomenyse nebuvo pašalinti kintamieji, kurie lygūs 0.

Pirmiausia patikriname logistinio modelio prielaidas. Tikriname, ar nėra multikolinerumo. Tam naudojame VIF (dispersijos mažėjimo daugikliu) kriterijų.

1 lentelė. Parametrų įvertinių lentelė

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Pregnancies	Pregnancies	1	0.02264	0.00625	3.62	0.0003	3.26533
Glucose	Glucose	1	0.00355	0.00058347	6.08	<.0001	16.22928
BloodPressure	BloodPressure	1	-0.00472	0.00094534	-4.99	<.0001	14.10565
SkinThickness	SkinThickness	1	0.00041445	0.00137	0.30	0.7619	3.92282
Insulin	Insulin	1	0.00006917	0.00019083	0.36	0.7171	2.05316
BMI	BMI	1	0.00514	0.00229	2.25	0.0251	17.34631
DiabetesPedigreeFunction	DiabetesPedigreeFunction	1	0.09266	0.05683	1.63	0.1036	3.24311
Age	Age	1	-0.00136	0.00184	-0.74	0.4584	13.27403

Kaip galime matyti, daugumos regresorių VIF yra didesnis už 4, t. y. jie stipriai koreliuoja tarpusavyje. Regresoriai kaip nėštumas, odos storis, insulinas ir diabeto kilmės funkcija nekoreliuoja. Atlikus pažingsninę regresiją ir vėl patikrinus VIF galime pastebėti, jog tai – problemos neišsprendė.

2 lentelė. Parametrų įvertinių lentelė po pažingsninės regresijos

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Pregnancies	Pregnancies	1	0.01532	0.00540	2.84	0.0047	2.32267
Glucose	Glucose	1	0.00247	0.00050205	4.93	<.0001	11.45502
BMI	BMI	1	-0.00076418	0.00188	-0.41	0.6850	11.19550
DiabetesPedigreeFunction	DiabetesPedigreeFunction	1	0.08704	0.05724	1.52	0.1289	3.13669

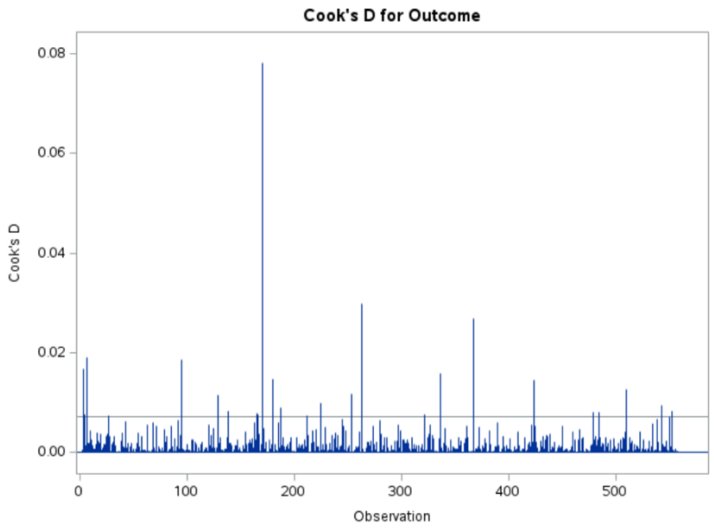
Toliau tikriname ar paklaidos yra nepriklausomos. Tam naudosime Durbino-Watsono statistiką.

3 lentelė. Durbin-Watson lentelė

Durbin-Watson Statistics			
Order	DW	Pr < DW	Pr > DW
1	2.1058	0.8932	0.1068

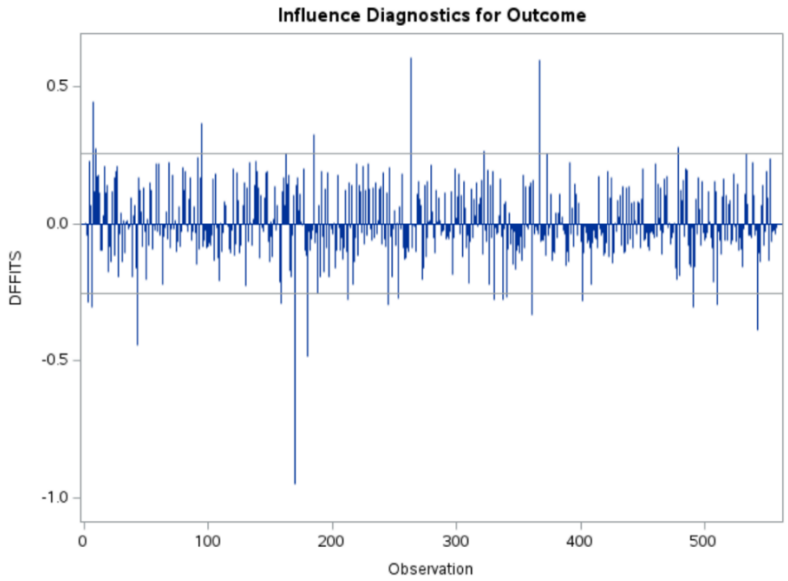
Kaip matome iš lentelės, p - reikšmė yra didesnė už reikšmingumo lygmenį 0.05, o tai reiškia, jog duomenyse nėra autokoreliacijos – paklaidos yra nepriklausomos.

Išskirčių analizei naudojame Kuko matą. Iš grafiko galime matyti, jog imtis išskirčių neturi (neviršija 1).



1 pav. Kuko matas

Pasitikrinimui, galime taikyti ir DFBetų statistiką.



2 pav. DFBetų statistika

Kaip matome iš grafiko, nei vienas stebėjimas neviršija 1, o tai dar kartą patvirtina, jog imtis neturi išskirčių.

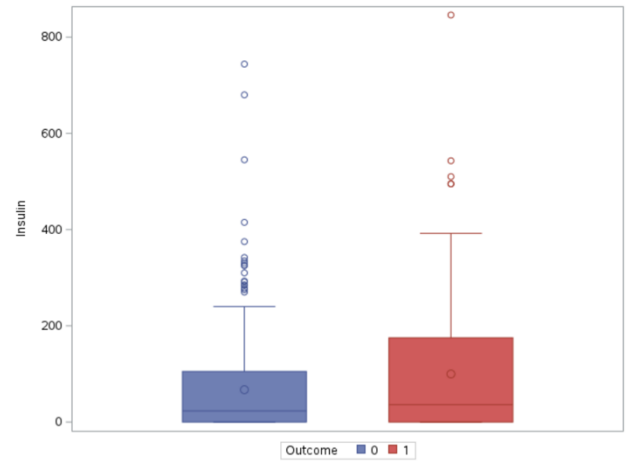
Pažiūrėkime kaip atrodo klasifikavimo lentelė.

4 lentelė. Klasifikavimo lentelė

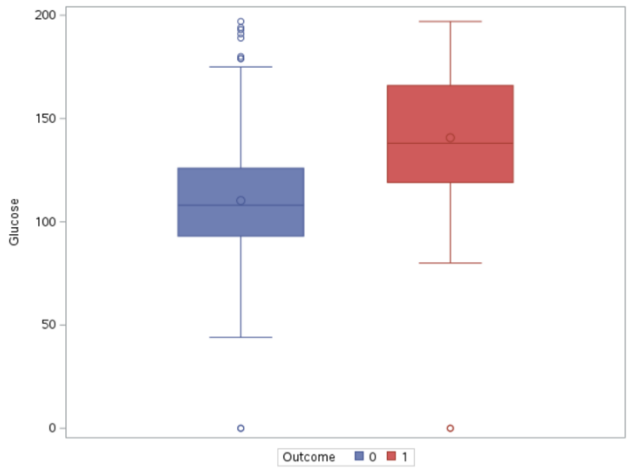
Outcome				
Outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	133	63.33	133	63.33
1	77	36.67	210	100.00

Iš lentelės matome, jog vyrauja abi y reikšmės. Stebimų įvykių yra daugiau nei 20%. Sergančiųjų yra 37 %, o nesergančių – 63 %.

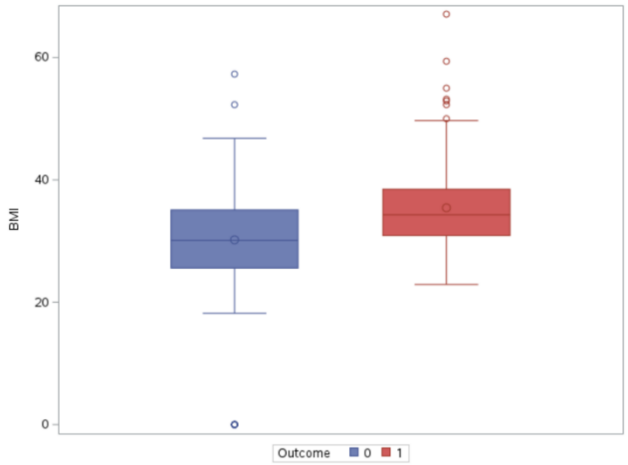
Pradinei analizei nubrėšime keletą grafikų pažiūrėsime kaip skiriasi rodikliai sergant ir nesergant diabetu.



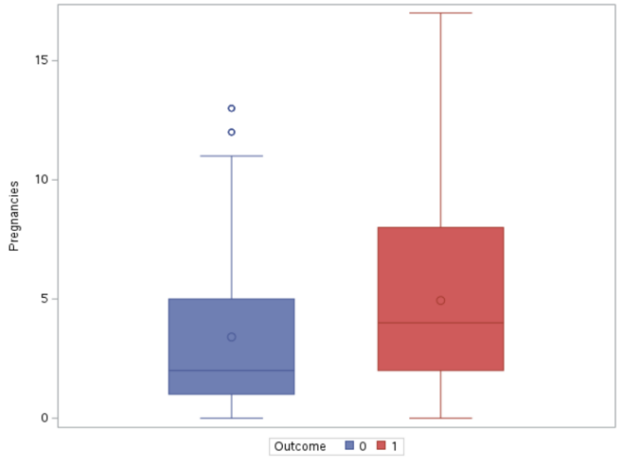
3 pav. Insulino stačiakampė diagrama



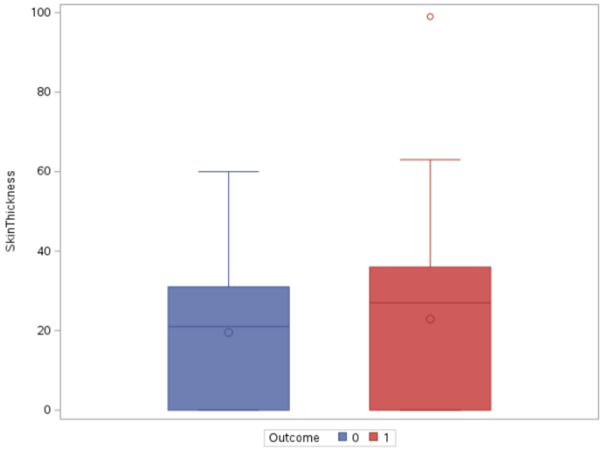
4 pav. Gliukozės stačiakampė diagrama



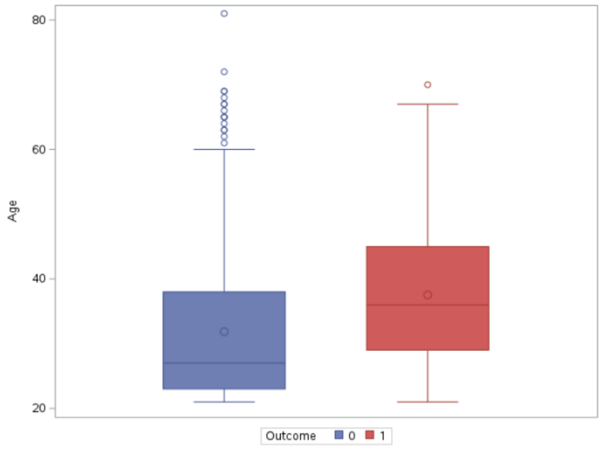
5 pav. KMI stačiakampė diagrama



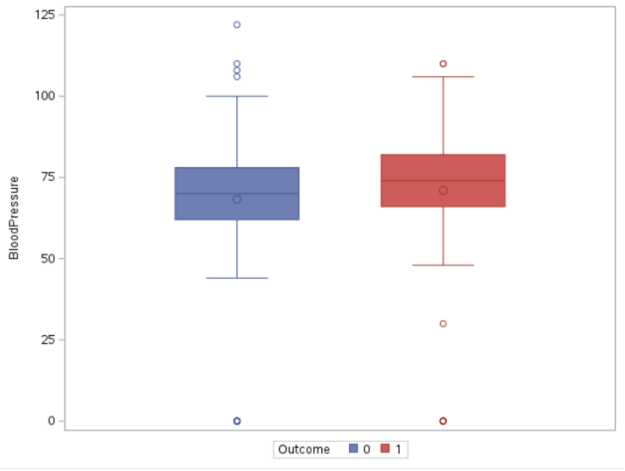
6 pav. Nėštumo stačiakampė diagrama



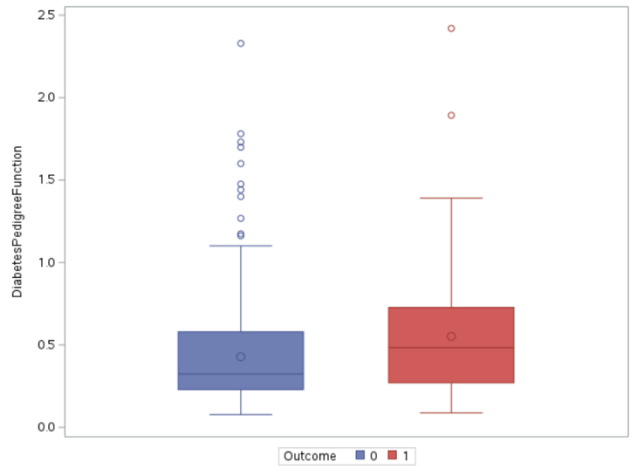
7 pav. Odos storio stačiakampė diagrama



8 pav. Amžiaus stačiakampė diagrama



9 pav. Kraujo spaudimo stačiakampė diagrama



10 pav. Diabeto atsiradimo funkcijos stačiakampė diagrama

Kaip galime pastebėti, sergant diabetu, beveik visada rodikliai didesni nei tų, kurių neserga.

Logit modelis

Pritaikius logistinę pažingsninę regresiją, galime patikrinti ar modelis statistiškai reikšmingas.

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	124.1546	1	<.0001
Score	112.5799	1	<.0001
Wald	92.4613	1	<.0001

5 lentelė. Hipotezės tikrinimas

Iš Chi-kvadrato reikšmės (tikėtinumų santykio kriterijaus) ir Valdo kriterijaus galime pamatyti, jog modelis - statistikai reikšmingas (abi reikšmės < 0.0001), o tai rodo, kad modelyje yra bent vienas reikšmingas regresorius. Tai taip pat galime patikrinti ir su Hosmerio-Lemešou kriterijumi.

6 lentelė. Hosmerio_Lemešou kriterijus

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
5.3229	8	0.7226

Kaip matome, jis taip pat parodo, jog duomenys gerai suderinti su modeliu (0.7226).

Kadangi jau buvo taikyta pažingsninė regresija, nereikšmingi regresoriai buvo išmesti, bet tuos kuriuos liko, galime patikrinti dar kartą su Valdo kriterijumi.

7 lentelė. Valdo kriterijus

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept	1	-8.9560	0.8163	120.3710	<.0001	0.000
Pregnancies	1	0.1262	0.0313	16.3034	<.0001	1.135
Glucose	1	0.0325	0.00389	69.6664	<.0001	1.033
BMI	1	0.0992	0.0171	33.6715	<.0001	1.104
DiabetesPedigreeFunc	1	0.9590	0.3472	7.6281	0.0057	2.609

Iš Valdo kriterijaus matome, jog yra likę keturi reikšmingi regresoriai: nėštumas, gliukozė, BMI ir diabeto kilmės funkcija.

Tada galime patikrinti determinacijos koeficientus (R^2 ir Nagelkerkės determinacijos koeficientą).

8 lentelė. Determinacijos koeficientai

R-Square	0.2881	Max-rescaled R-Square	0.3982

Abu koeficientus gavome gan mažus, o tai rodo, jog logistines regresijos modelis duomenims nelabai tinka, tačiau svarbu suprasti, jog šie rodmenys atlieka tik pagalbinį vaidmenį.

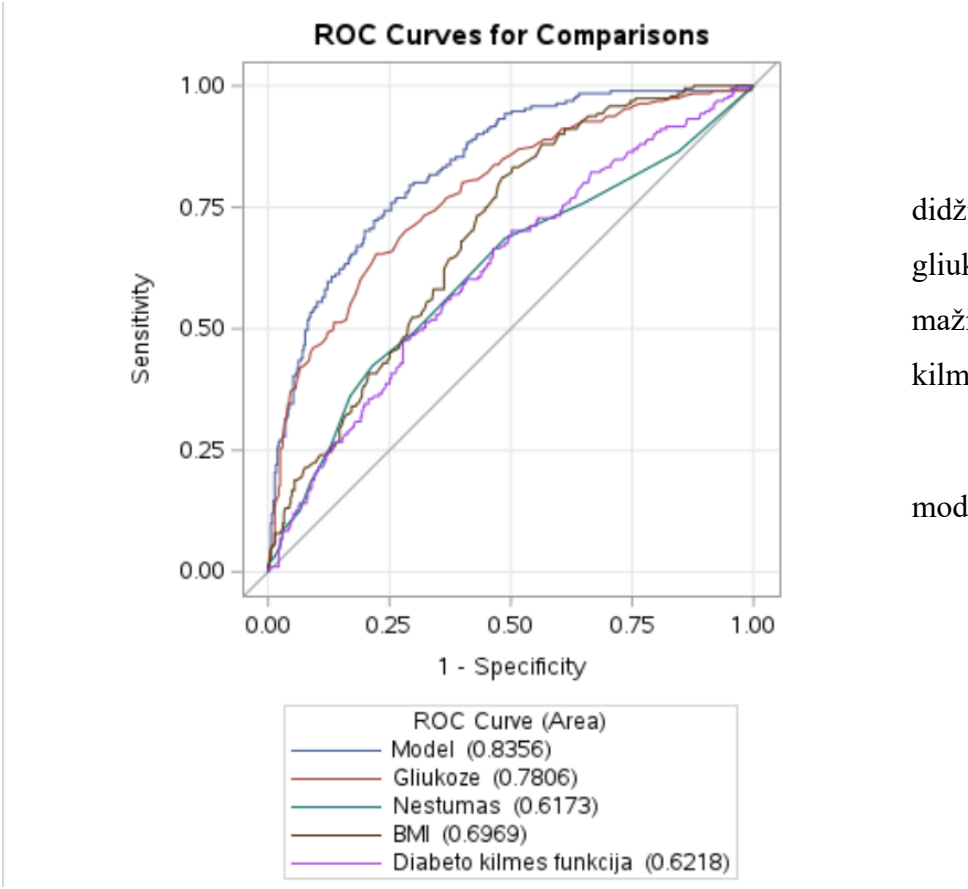
Porų suderinamumą galime patikrinti su Somerso D koeficientu.

9 lentelė. Somerso D koeficientas

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	83.6	Somers' D	0.671

Matome, jog suderintų porų yra 83.6%, o Somerso D=0.671, abu rodmenys, rodo pakankamai gera modelio tinkamumą duomenims.

Dabar galime palyginti ROC kreives.



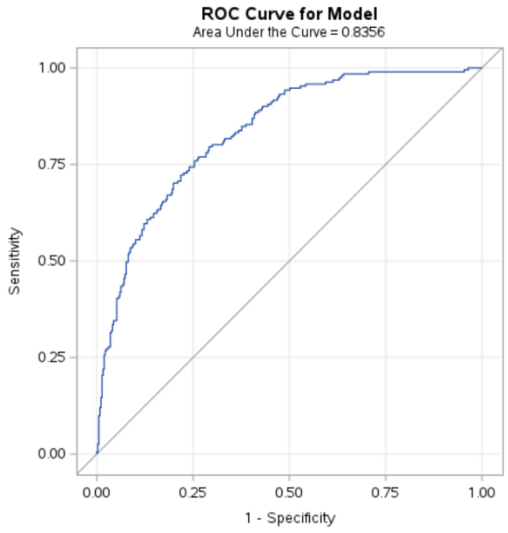
Kaip matome, didžiausią plotą turi gliukozės kreivė, o mažiausią – diabeto kilmės funkcija.

Sudarytas modelis atrodo taip:

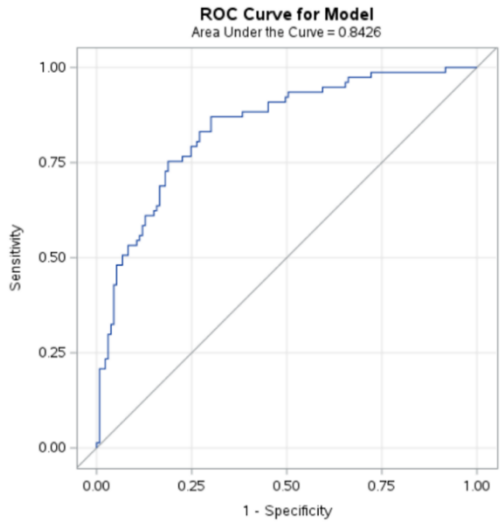
11 pav. Palyginamosios ROC kreivės

$$\ln\left(\frac{P(pacientas\ serga)}{P(pacientas\ neserga)}\right) = -8.9560 + 0.1262 * Nėštumas + 0.0325 * Gliukozė + 0.0992 * BMI + 0.9590 * Diabeto\ atsiradimo\ funkcija.$$

Gautos ROC kreivės mokymų ir testinėje aibėje.



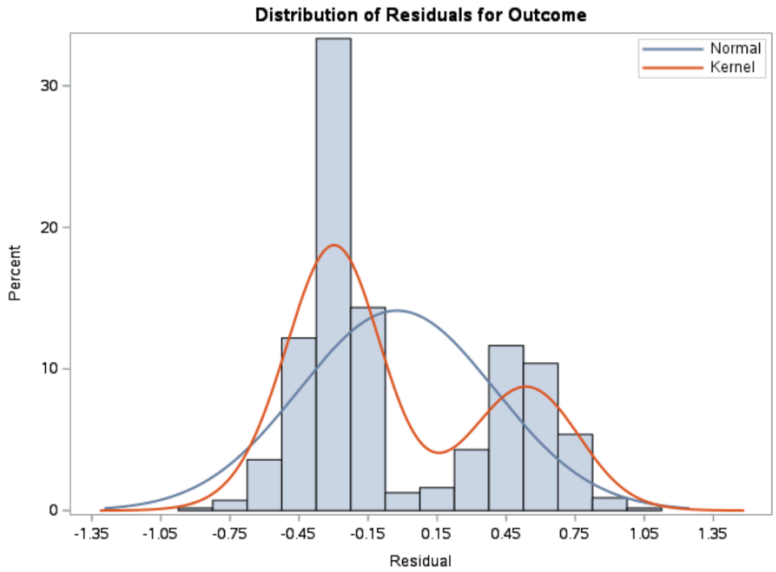
12 pav. Mokymosi aibės ROC kreivė



13 pav. Testavimo aibės ROC kreivė

Probit modelis

Dabar patikrinsime probit modelio prielaidas. Kadangi visos prielaidos identiškos kaip ir logit modelyje, tik prisideda viena papildoma – paklaidų normalumas, todėl tikrinsime tik šią vieną prielaidą.



14 pav. Histograma

Iš histogramos galime matyti, jog paklaidos nėra normalios, kad dar kartą įsitikinti, taikysime Shapiro-Wilk testą.

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.962788	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.107695	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.596869	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	8.645683	Pr > A-Sq	<0.0050

15 pav. Normalumo testo statistika

Šis testas taip pat patvirtina, jog paklaidos nėra normalios, o tai reiškia, kad analizei probit modelis nėra tinkamas.

2.2 Binarinio atsako modelis naudojant R

Atliekant analizę su R buvo ištrinti stebėjimai tokie, kaip: kraujo spaudimas lygus 0, gliukozės kiekis lygus 0, KMI lygus 0, insulino kiekis lygus 0 ir odos storis lygus 0. Liko 392 stebėjimai. Likę duomenys buvo padalinti į testavimo ir mokymo aibes 30 : 70 santykiu. Pradinis logistinės regresijos modelis atrodė taip:

$$\ln\left(\frac{P(pacientas\ serga)}{P(pacientas\ neserga)}\right) = \beta_0 + \beta_1 \times nėštumų\ skaičius + \beta_2 \times gliukozės\ koncentracija + \beta_3 \times kraujo\ spaudimas + \beta_4 \times insulino\ kiekis + \beta_5 \times odos\ storis + \beta_6 \times KMI + \beta_7 \times diabeto\ atsiradimo\ funkcija + \beta_8 \times amžius.$$

Pradėjome nuo prielaidų tikrinimo. Modelyje neturi būti multikolinearumo.

11 lentelė. Multikolinearumo tyrimas

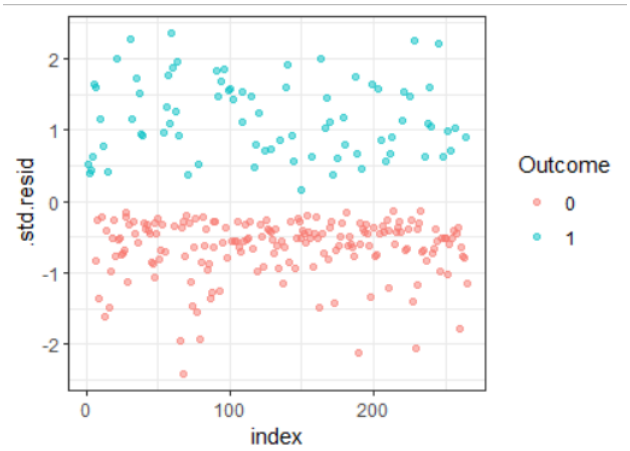
Pregnancies	Glucose	Blood pressure	Skin thickness	Insulin	BMI	Diabetes pedigree function	Age
1,974	1,367	1,199	1,667	1,397	1,836	1,017	2,124

Iš visų gautų koeficientų matome, jog nėra, nes neviršija 4.

Taip pat paklaidos turi būti nepriklausomos. Iš pradžių patikrinome hipotezę dėl autokoreliacijos.

$$\begin{cases} H_0: nėra\ autokoreliacijos\ tarp\ paklaidų, \\ H_1: yra\ autokoreliacija. \end{cases}$$

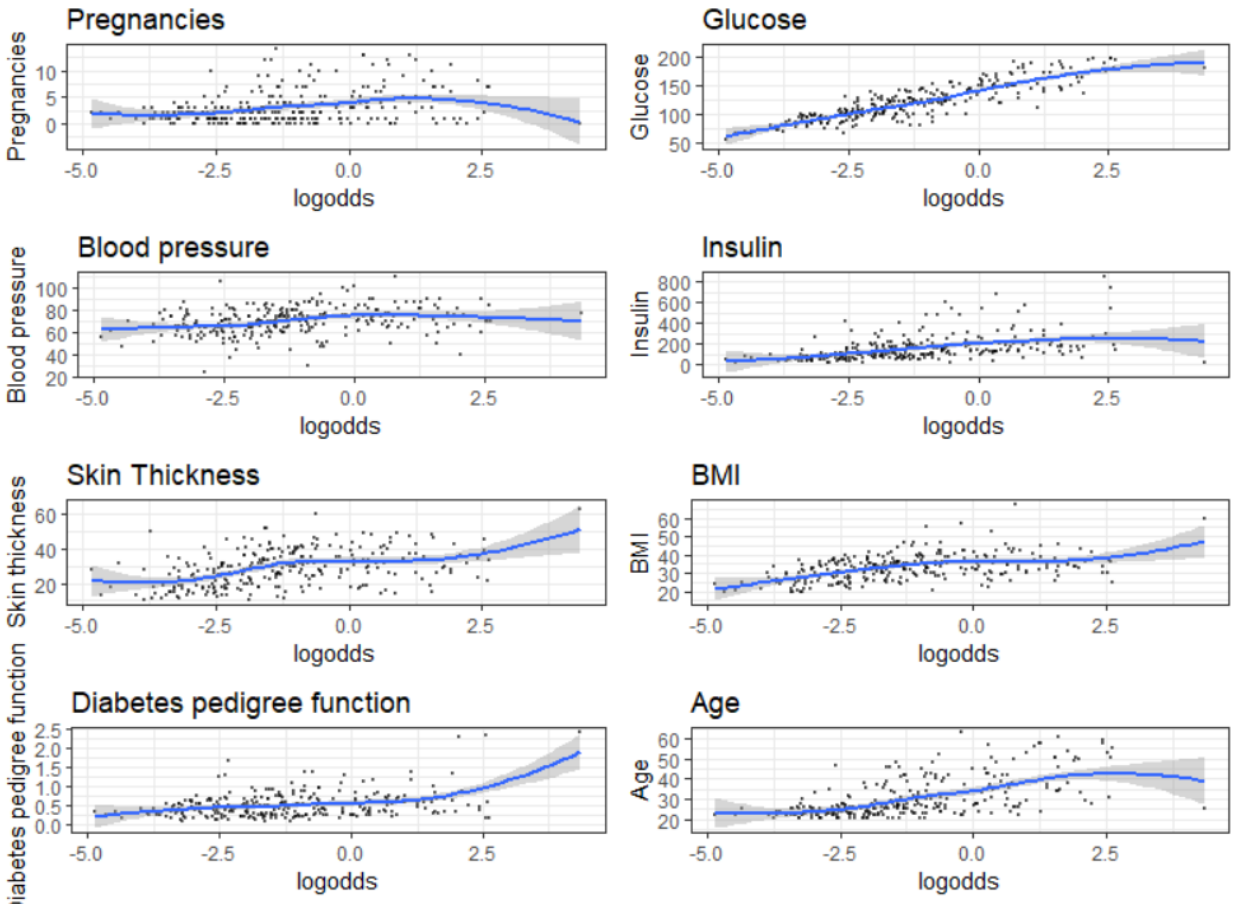
Hipotezę tikrinome naudodamiesi Durbin-Watson testu ir gavome, kad nulinės hipotezės negalime atmesti, nes p - reikšmė = 0,52 > 0,05 = reikšmingumo lygmuo.



16 pav. Standartizuotų liekanų paklaidų išsidėstymas

Taip pat iš pateikto paveiksliuko matome, jog standartizuotos paklaidos tarp sergančių ir nesergančių yra nepriklausomai pasiskirsčiusios.

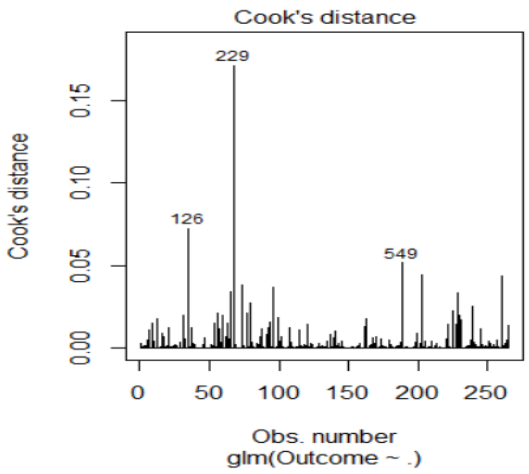
Toliau tikrinamas tiesiškumas tolydžiųjų kintamųjų atžvilgiu priklausomojo kintamojo.



17 pav. Tiesiškumo tyrimų panelė

Matome, kad tiesiškumo sąlyga beveik visomis kovariantėmis išpildyta. Diabeto atsiradimo funkcijai buvo taikytas logaritminis transformavimas, bet jis nepadėjo.

Taip pat patikriname ar duomenyse nėra išskirčių.



18 pav. Išskirčių tyrimas

Iš pateikto grafiko matome, kad išskirčių nėra, nes visos reikšmės mažesnės už 1.

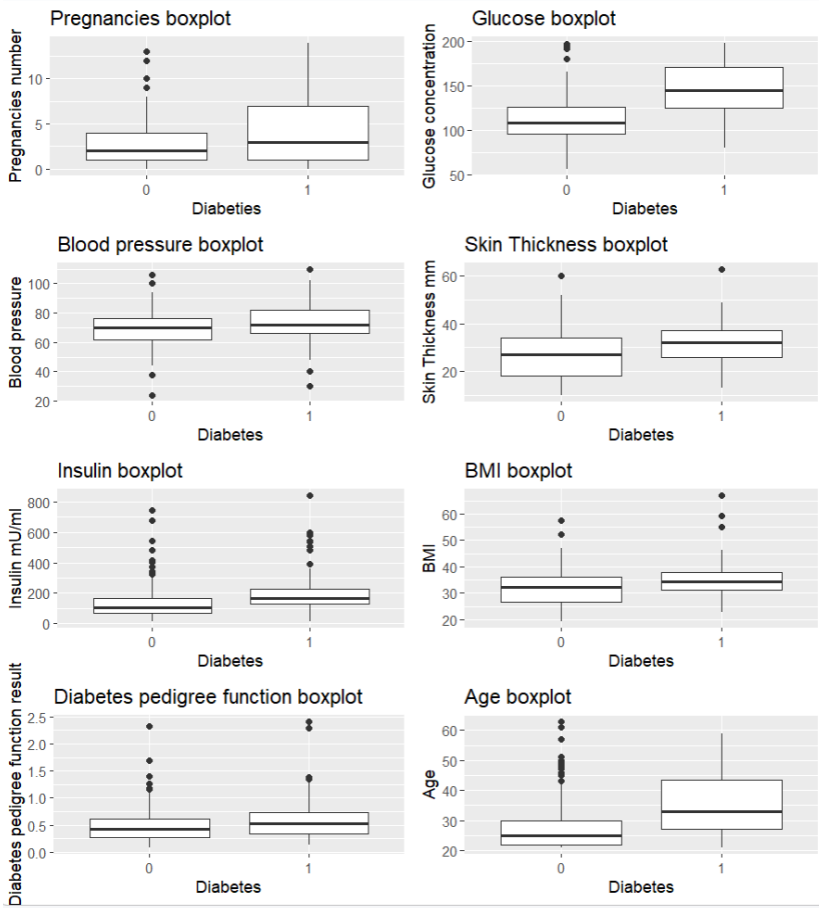
Dar reikia patikrinti, jog įvykio ar ne įvykio reikšmių būtų nemažiau nei 20 proc.

13 lentelė. Dažnių lentelė

Neserga	**Serga**
185	80

Nors ir matome iš lentelės, jog įvykių yra dvigubai mažiau nei ne įvykių, bet sąlyga yra išpildoma.

Toliau pareisime prie pirminės duomenų analizės – iš grafikų nuspręsime, ar sirgimas / nesirgimas diabetu atsiskiria pagal atskirus požymius.



19 pav. Stačiakampių diagramų panelė

Iš pateikto paveikslėlio matome, jog pagal gliukozę ir amžių geriausiai atsiskiria grupės serga / neserga cukriniu diabetu, o pagal kitas grupes beveik neatsiskiria.

Pereiname prie modelio konstravimo. Iš pradžių tikrinsime hipotezę:

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$$

$$H_1: bent\ viena\ lygybė\ neteisinga.$$

Hipotezę tikrinome pasinaudodami tikėtinumo santykio kriterijumi.

```
Analysis of Deviance Table

Model 1: Outcome ~ 1
Model 2: Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThickness +
Insulin + BMI + DiabetesPedigreeFunction + Age
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       264     324.60
2       256     229.01  8   95.588 < 2.2e-16 ***
```

Gavome p – reikšmę $< 2{,}2 \times 10^{-16}$, tai priimame alternatyvą, turime bent vieną reikšmingą kovariantę.

Toliau taikėme pažingsninę regresiją.

```
Start:  AIC=247.01
Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThickness +
    Insulin + BMI + DiabetesPedigreeFunction + Age
                          Df Deviance    AIC
- SkinThickness            1   229.03 245.03
- BloodPressure            1   229.32 245.32
- Pregnancies              1   229.50 245.50
- Insulin                  1   230.07 246.07
<none>                         229.01 247.01
- DiabetesPedigreeFunction  1   231.29 247.29
- BMI                      1   233.29 249.29
- Age                      1   239.48 255.48
- Glucose                  1   267.62 283.62
Step:  AIC=245.02
Outcome ~ Pregnancies + Glucose + BloodPressure + Insulin + BMI +
    DiabetesPedigreeFunction + Age
                          Df Deviance    AIC
- BloodPressure            1   229.34 243.34
- Pregnancies              1   229.52 243.52
- Insulin                  1   230.08 244.08
<none>                         229.03 245.03
- DiabetesPedigreeFunction  1   231.34 245.34
+ SkinThickness            1   229.01 247.01
- BMI                      1   236.04 250.04
- Age                      1   240.12 254.12
- Glucose                  1   267.68 281.68
Step:  AIC=243.34
Outcome ~ Pregnancies + Glucose + Insulin + BMI + DiabetesPedigreeFunction +
Age
                          Df Deviance    AIC
- Pregnancies              1   229.88 241.88
- Insulin                  1   230.32 242.32
<none>                         229.34 243.34
- DiabetesPedigreeFunction  1   231.86 243.86
+ BloodPressure            1   229.03 245.03
+ SkinThickness            1   229.32 245.32
- BMI                      1   236.15 248.15
- Age                      1   240.12 252.12
- Glucose                  1   267.72 279.72
Step:  AIC=241.88
Outcome ~ Glucose + Insulin + BMI + DiabetesPedigreeFunction + Age
                          Df Deviance    AIC
- Insulin                  1   230.81 240.81
<none>                         229.88 241.88
- DiabetesPedigreeFunction  1   232.66 242.66
+ Pregnancies              1   229.34 243.34
+ BloodPressure            1   229.52 243.52
+ SkinThickness            1   229.87 243.87
- BMI                      1   236.85 246.85
- Age                      1   244.18 254.18
- Glucose                  1   268.65 278.65
```

```
Step:  AIC=240.81
Outcome ~ Glucose + BMI + DiabetesPedigreeFunction + Age
                          Df Deviance    AIC
<none>                         230.81 240.81
- DiabetesPedigreeFunction  1   233.47 241.47
+ Insulin                   1   229.88 241.88
+ Pregnancies               1   230.32 242.32
+ BloodPressure             1   230.52 242.52
+ SkinThickness             1   230.80 242.80
- BMI                       1   237.00 245.00
- Age                       1   244.99 252.99
- Glucose                   1   275.17 283.17
Call:  glm(formula = Outcome ~ Glucose + BMI + DiabetesPedigreeFunction + Age,
family = binomial(logit), data = train)
Coefficients:
             (Intercept)                   Glucose                       BMI
DiabetesPedigreeFunction                       Age
                -9.86471                   0.03629                   0.05956
0.81849                   0.06126
Degrees of Freedom: 264 Total (i.e. Null);  260 Residual
Null Deviance:        324.6
Residual Deviance: 230.8      AIC: 240.8
```

Pažingsninė regresija paliko gliukozės kiekio, KMI rodiklio, diabeto atsiradimo funkcijos bei amžiaus kovariantes.

Dabar pasižiūrėsime gauto modelio santrauką.

```
Call:
glm(formula = Outcome ~ Glucose + BMI + DiabetesPedigreeFunction +
    Age, family = binomial(logit), data = train)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5743  -0.6501  -0.3739   0.6203   2.3330
Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)              -9.864711   1.324754  -7.446 9.59e-14 ***
Glucose                   0.036294   0.006096   5.954 2.62e-09 ***
BMI                       0.059561   0.024370   2.444 0.014525 *
DiabetesPedigreeFunction  0.818489   0.515816   1.587 0.112561
Age                       0.061263   0.016736   3.661 0.000252 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 324.60  on 264  degrees of freedom
Residual deviance: 230.81  on 260  degrees of freedom
AIC: 240.81
```

Matome, jog visos kovariantės yra statistiškai reikšmingos išskyrus diabeto atsiradimo funkciją, tačiau su šia kovariante turime mažesnį AIC rodiklį todėl paliksime kovariantę modelyje.

Taip pat matome, jog didėjant gliukozės kiekiui kraujyje, KMI rodikliui, diabeto atsiradimo funkcijos reikšmei bei amžiui didėja galimybė susirgti diabetu.

Kad galėtumėme interpretuoti detaliau kovariantes, kiekvienai kovariančių koeficiento reikšmei suskaičiavome eksponentę.

14 lentelė. Koeficiento eksponenčių reikšmės

Intercept	**Glucose**	**BMI**	**Age**
$5{,}19 \times 10^{-5}$	1,04	1,06	1,06

Iš pateiktos lentelės matome, jog didėjant gliukozės koncentracijai vienu vienetu 4 proc. labiau tikėtina, kad žmogus sirgs diabetu. Taip pat didinant KMI rodiklį vienu vienetu 6 proc. labiau tikėtina, kad žmogus sirgs cukriniu diabetu. Padidėjus amžiui 1 metais, 6 proc. labiau tikėtina, kad žmogus sirgs diabetu.

Kadangi diabeto atsiradimo funkcijos reikšmės yra iš intervalo $(0{,}05; 2{,}2)$, tai paskaičiuosime šiam regresoriui $exp(0{,}05 \times koeficientas\ diabeto\ atsiradimo\ funkcijos)$.

```
DiabetesPedigreeFunction
1.0417734
```

Tai padidinus šios funkcijos reikšmę per 0,05, 4 procentais bus labiau tikėtina, jog žmogus sirgs diabetu.

Apskaičiavome klasifikavimo lentelę mokymosi duomenims.

15 lentelė. Klasifikavimo lentelė mokymo duomenims

	1	**0**	
1	44	21	**Jautrumas**: 0,68
0	36	164	**Specifiškumas**: 0,55
	Tikslumas (precision): 0,55	**Negative predictive value**: 0,89	**Bendras tikslumas**: 0,78

Matome, jog bendras tikslumas yra 0,78. Taip pat gerai atpažįsta nesergančius asmenis, kai iš tikrųjų jie neserga, tačiau modelio tikslumas (precision) yra ganėtinai nedidelis 0,55, t.y. kiek iš visų priskirtų sergančių asmenų yra tikrai sergantys.

Optimaliausio slenksčio radimui pasinaudojome Youdeno indekso skaičiavimo formule:

$$J = jautrumas + specifiškumas - 1,$$

nes jautrumas ir specifiškumas vienodai svarbūs. Pritaikius Youdeno metodą, gavome, jog optimaliausias slenkstis yra 0,28.

Sudaryta klasifikavimo matrica testavimo duomenims, naudojant slenktį 0,28.

16 lentelė. Klasifikavimo matrica testavimo duomenims

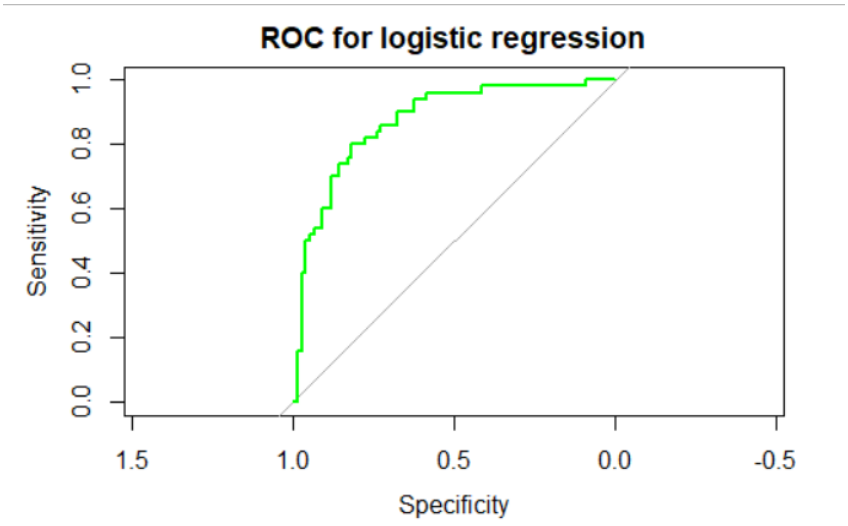
	1	**0**	
1	40	10	**Jautrumas**: 0,8
0	16	61	**Specifiškumas:** 0,79
	Tikslumas (precision): 0,71	**Negative predictive value**: 0,86	**Bendras tikslumas**: 0,80

Iš klasifikavimo matricos matome, jog pagrindinėje įstrižainėje yra teisingai suklasifikuoti pacientų skaičiai. Matome, jog gerai atpažįsta sergančius ir nesergančius asmenis. Bendras tikslumas yra šiek tiek geresnis nei mokymo duomenims, t.y. per 0,02 didesnis.

F_1 score gavome 0,75, kuris yra randamas pagal formulę:

$$F_\beta = \frac{(\beta + \beta^2) \times (precision * recall)}{(\beta^2 \times precision + recall)}, kur\ \beta\ gali\ būti\ 0{,}5, 1, 2.$$

Mes naudojome $\beta = 1$, nes jautrumas ir tikslumas vienodai svarbūs.



20 pav. ROC kreivė

Matome paveikslėlyje ROC kreivę sudarytam modeliui, plotas po kreive yra 0,8706.

Gautas modelis atrodo taip:

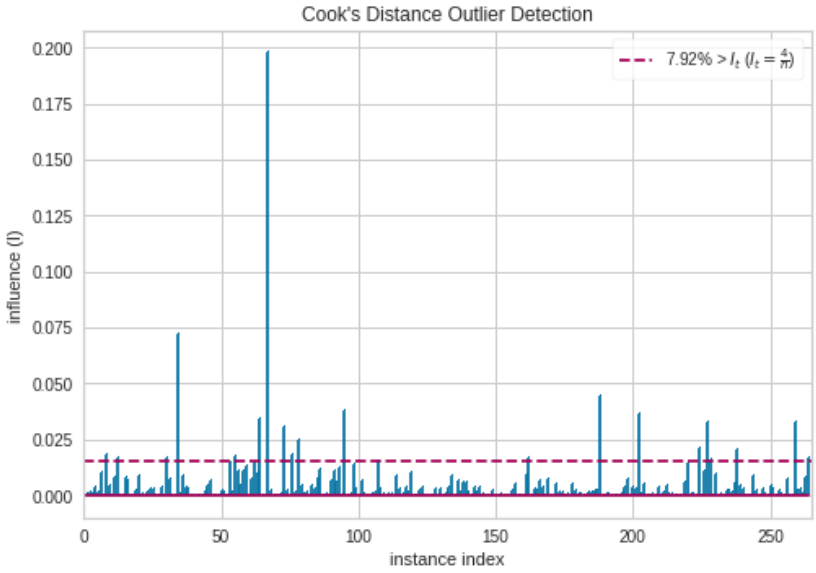
$$\ln\left(\frac{P(pacientas\ serga)}{P(pacientas\ neserga)}\right)$$
$$= -9{,}864711 + 0{,}036294 \times gliukozės\ koncentracija + 0{,}059561 \times KMI$$
$$+ 0{,}818489 \times diabeto\ atsiradimo\ funkcija + 0{,}061263 \times amžius.$$

Jo AIC 240.81, determinacijos pseudokoeficientas 0,289, kuris gaunamas:

$$R^2 = 1 - \frac{deviacija}{nulinė\ deviacija}.$$

2.3 Binarinio atsako modelis naudojant Python

Pirmiausia, vizualiai patikriname ar nėra išskirčių. Naudojame Kuko matą.



21 pav. Kuko mato grafikas

Kadangi Kuko matas yra mažesnis už 1, galime teigti, jog išskirčių duomenyse nėra. Toliau tikriname multikolinearumą.

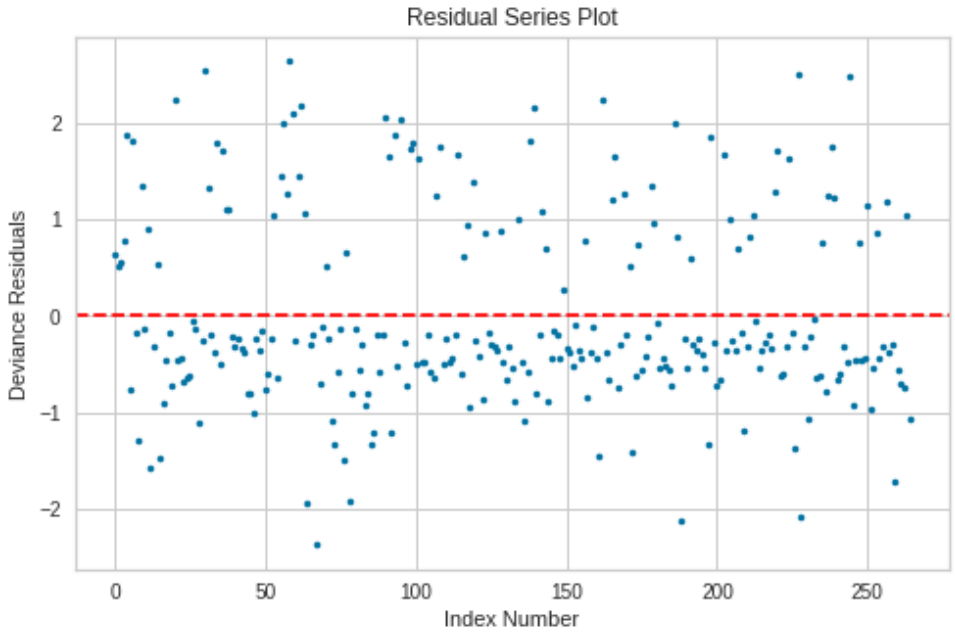
17 lentelė VIF kriterijus

Kovariantės	**VIF**
Nėštumas	1.96
Gliukozė	1.61

Kraujo spaudimas	1.25
Odos storis	1.9
Insulinas	1.52
BMI	1.99
Diabeto kilmės funkcija	1.07
Amžius	2.24

Iš lentelės matome, jog nei vienos kovariantės VIF kriterijus neviršija 4, todėl multikolinearumo problemos nėra.

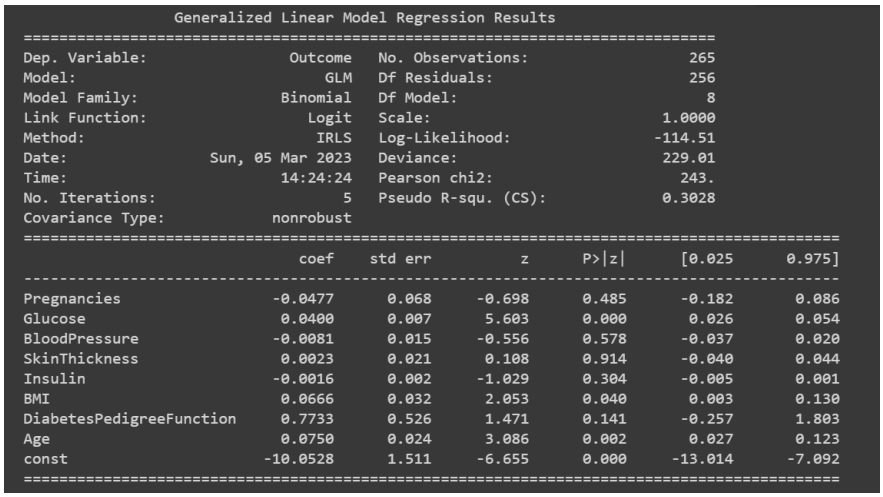
Kad patikrinti paklaidų nepriklausomumą nubrėžiame grafiką ir taikome Durbin-Watson testą.



22 pav. Paklaidų grafikas

Iš grafiko matome, jog paklaidos yra nepriklausomos. Pritaikius Durbin - Watson testą, gauname, jog paklaidos nėra autokoreliuotos, nes statistika yra 1.98 (jei reikšmė apytiksliai lygi 2, autokoreliacijos nėra).

Parinkus pradinį logit modelį, matome, jog nereikšmingos kovariantės yra: nėštumas, kraujo spaudimas, odos storis, insulinas ir diabeto kilmės funkcija. Taigi, toliau atliekame pažingsninę regresiją.



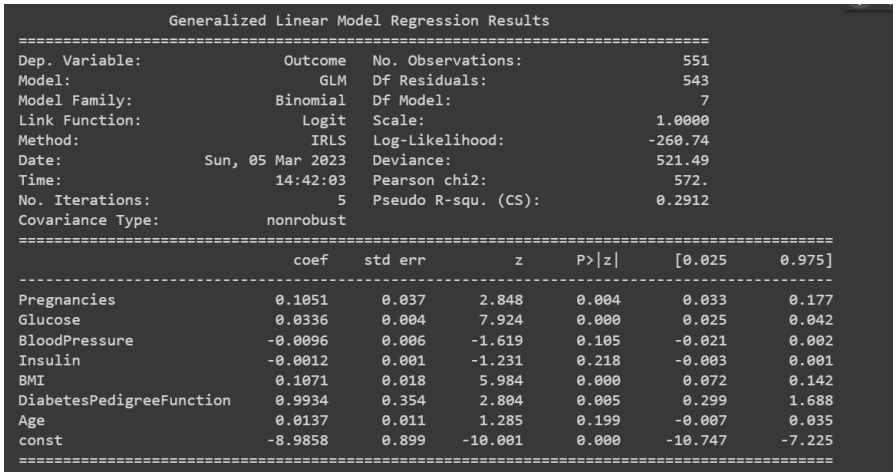
Generalized Linear Model Regression Results

Dep. Variable:	Outcome	No. Observations:	265
Model:	GLM	Df Residuals:	256
Model Family:	Binomial	Df Model:	8
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-114.51
Date:	Sun, 05 Mar 2023	Deviance:	229.01
Time:	14:24:24	Pearson chi2:	243.
No. Iterations:	5	Pseudo R-squ. (CS):	0.3028
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Pregnancies	-0.0477	0.068	-0.698	0.485	-0.182	0.086
Glucose	0.0400	0.007	5.603	0.000	0.026	0.054
BloodPressure	-0.0081	0.015	-0.556	0.578	-0.037	0.020
SkinThickness	0.0023	0.021	0.108	0.914	-0.040	0.044
Insulin	-0.0016	0.002	-1.029	0.304	-0.005	0.001
BMI	0.0666	0.032	2.053	0.040	0.003	0.130
DiabetesPedigreeFunction	0.7733	0.526	1.471	0.141	-0.257	1.803
Age	0.0750	0.024	3.086	0.002	0.027	0.123
const	-10.0528	1.511	-6.655	0.000	-13.014	-7.092

23 pav. Pirminis logit modelis

Šaliname odos storį. $AIC = 245.02$



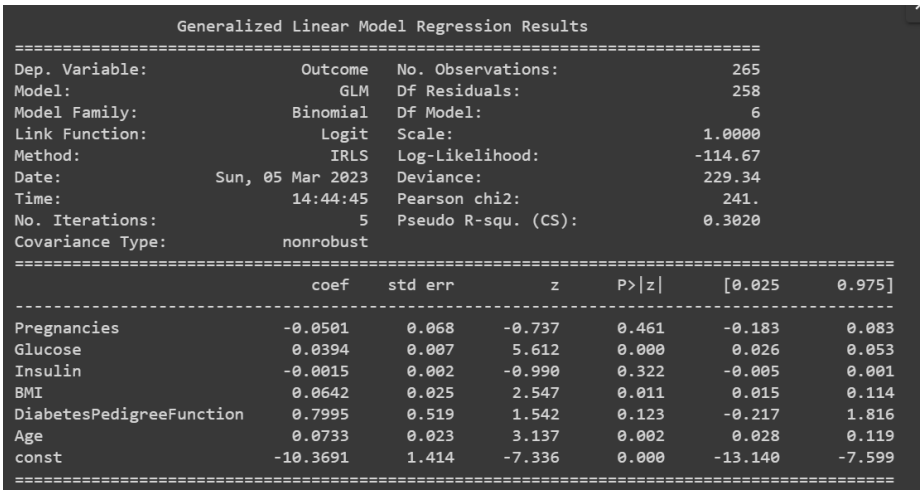
Generalized Linear Model Regression Results

Dep. Variable:	Outcome	No. Observations:	551
Model:	GLM	Df Residuals:	543
Model Family:	Binomial	Df Model:	7
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-260.74
Date:	Sun, 05 Mar 2023	Deviance:	521.49
Time:	14:42:03	Pearson chi2:	572.
No. Iterations:	5	Pseudo R-squ. (CS):	0.2912
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Pregnancies	0.1051	0.037	2.848	0.004	0.033	0.177
Glucose	0.0336	0.004	7.924	0.000	0.025	0.042
BloodPressure	-0.0096	0.006	-1.619	0.105	-0.021	0.002
Insulin	-0.0012	0.001	-1.231	0.218	-0.003	0.001
BMI	0.1071	0.018	5.984	0.000	0.072	0.142
DiabetesPedigreeFunction	0.9934	0.354	2.804	0.005	0.299	1.688
Age	0.0137	0.011	1.285	0.199	-0.007	0.035
const	-8.9858	0.899	-10.001	0.000	-10.747	-7.225

24 pav. Logit modelis pašalinus odos storį

Šaliname kraujo spaudimą. $AIC = 243.34$



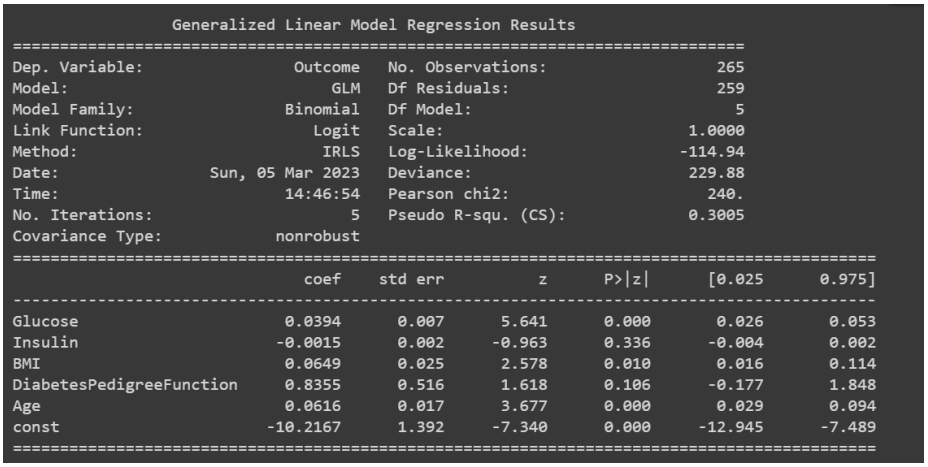
Generalized Linear Model Regression Results

Dep. Variable:	Outcome	No. Observations:	265
Model:	GLM	Df Residuals:	258
Model Family:	Binomial	Df Model:	6
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-114.67
Date:	Sun, 05 Mar 2023	Deviance:	229.34
Time:	14:44:45	Pearson chi2:	241.
No. Iterations:	5	Pseudo R-squ. (CS):	0.3020
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Pregnancies	-0.0501	0.068	-0.737	0.461	-0.183	0.083
Glucose	0.0394	0.007	5.612	0.000	0.026	0.053
Insulin	-0.0015	0.002	-0.990	0.322	-0.005	0.001
BMI	0.0642	0.025	2.547	0.011	0.015	0.114
DiabetesPedigreeFunction	0.7995	0.519	1.542	0.123	-0.217	1.816
Age	0.0733	0.023	3.137	0.002	0.028	0.119
const	-10.3691	1.414	-7.336	0.000	-13.140	-7.599

25 pav. Logit modelis pašalinus kraujo spaudimą

Šaliname nėštumą. $AIC = 241.88$



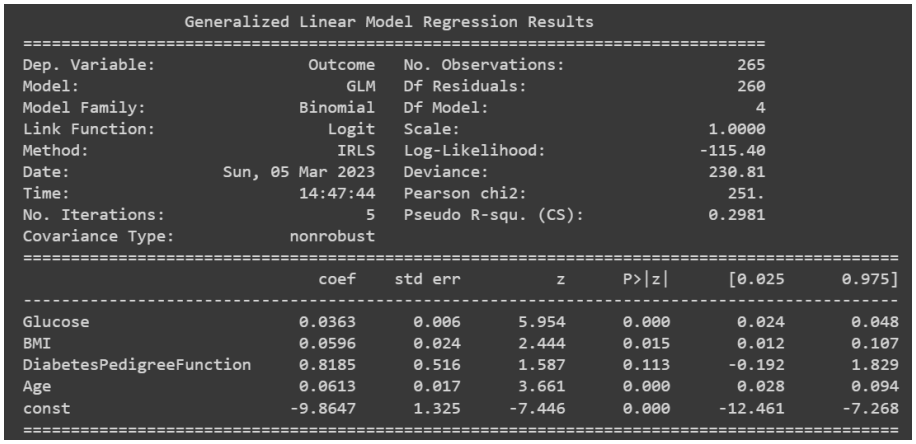
Generalized Linear Model Regression Results

Dep. Variable:	Outcome	No. Observations:	265
Model:	GLM	Df Residuals:	259
Model Family:	Binomial	Df Model:	5
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-114.94
Date:	Sun, 05 Mar 2023	Deviance:	229.88
Time:	14:46:54	Pearson chi2:	240.
No. Iterations:	5	Pseudo R-squ. (CS):	0.3005
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Glucose	0.0394	0.007	5.641	0.000	0.026	0.053
Insulin	-0.0015	0.002	-0.963	0.336	-0.004	0.002
BMI	0.0649	0.025	2.578	0.010	0.016	0.114
DiabetesPedigreeFunction	0.8355	0.516	1.618	0.106	-0.177	1.848
Age	0.0616	0.017	3.677	0.000	0.029	0.094
const	-10.2167	1.392	-7.340	0.000	-12.945	-7.489

26 pav. Logit modelis pašalinus nėštumą

Šaliname insuliną. $AIC = 240.80$



Generalized Linear Model Regression Results

Dep. Variable:	Outcome	No. Observations:	265
Model:	GLM	Df Residuals:	260
Model Family:	Binomial	Df Model:	4
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-115.40
Date:	Sun, 05 Mar 2023	Deviance:	230.81
Time:	14:47:44	Pearson chi2:	251.
No. Iterations:	5	Pseudo R-squ. (CS):	0.2981
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Glucose	0.0363	0.006	5.954	0.000	0.024	0.048
BMI	0.0596	0.024	2.444	0.015	0.012	0.107
DiabetesPedigreeFunction	0.8185	0.516	1.587	0.113	-0.192	1.829
Age	0.0613	0.017	3.661	0.000	0.028	0.094
const	-9.8647	1.325	-7.446	0.000	-12.461	-7.268

27 pav. Logit modelis pašalinus insuliną

Pašalinus diabeto kilmės funkciją AIC reikšmė padidėja (241.47), todėl šią kovariantę modelyje paliekame. Taigi, reikšmingos kovariantės yra: gliukozė, KMI, diabeto kilmės funkcija ir amžius. Toliau skaičiuojame kovariančių koeficientų eksponentes.

```
Intercept:  5.197748052553491e-05
Glucose:  1.0369668898990494
BMI:  1.06141189686234
DiabetesPedigreeFunction:  1.0417739695676995
Age:  1.063217831696686
```

28 pav. Kovariančių koeficientų eksponentės

Didėjant gliukozės koncentracijai ir diabeto atsiradimo funkcijai vienu vienetu 4 proc. labiau tikėtina, kad žmogus sirgs diabetu. Taip pat didinant KMI ir amžiaus rodiklį vienu vienetu 6 proc. labiau tikėtina, kad žmogus sirgs cukriniu diabetu.

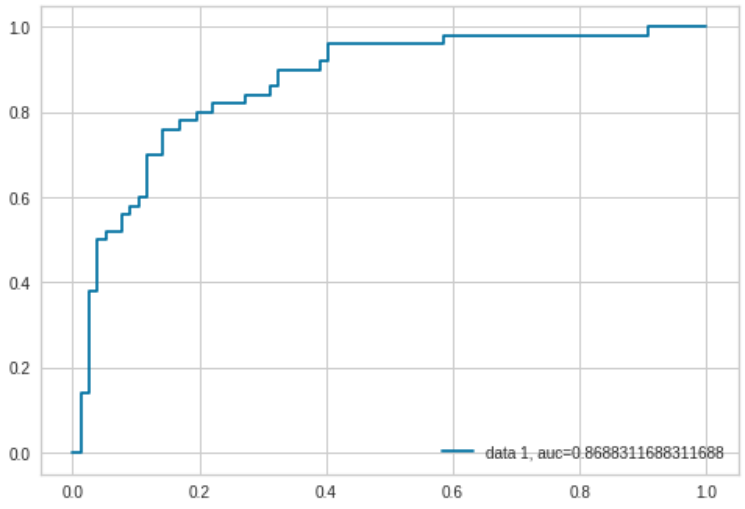
Galiausiai pasižiūrime klasifikavimo lentelę testiniams duomenims ir taip pat nubrėžiame ROC kreivę.

24 lentelė. Klasifikavimo lentelė

	1	**0**
1	69	8
0	20	30

```
Jautrumas:  0.8961038961038961
Specifiškumas:  0.6
Tikslumas:  0.7752808988764045
Neigiama prognozuojama reikšmė:  0.7894736842105263
Bendras tikslumas:  0.7795275590551181
```

29 pav. Kiti tikslumo rodikliai



30 pav. ROC kreivė

Iš tikslumo rodiklių, tokių kaip jautrumas, specifiškumas ir kt. ir ROC kreivės galime teigti, jog binarinio atsako modelis yra tinkamas šiems duomenims.

3. IŠVADOS

3.1 SAS išvados, kai duomenys nebuvo išfiltruoti

Atlikus pradinę analizę buvo pastebėta, jog moterų, sergančiu diabetu, visi rodikliai yra aukštesni, negu tų, kurios neserga diabetu, tačiau atlikus analizę su nepašalintomis nulinėmis reikšmėmis, buvo gauta, jog ne visi regresoriai yra reikšmingi ir gautas logistinės regresijos modelis atrodo taip:

$$\ln\left(\frac{P(pacientas\ serga)}{P(pacientas\ neserga)}\right) = -8.9560 + 0.1262 \times Nėštumas + 0.0325 \times Gliukozė + 0.0992 \times BMI + 0.9590 \times Diabeto\ atsiradimo\ funkcija\,.$$

Tai rodo, jog diabeto susirgimui daro įtaką nėštumų skaičius, gliukozės kiekis, BMI ir diabeto atsiradimo funkcija.

Gautas AUC lygus 0.8426.

Bandant atlikti analizę su probit modeliu, buvo pastebėta, jog duomenys neatitinka šio modelio prielaidų, todėl tolimesnės analizės nebuvo galima atlikti.

3.2 R ir Python išvados, kai duomenys buvo išfiltruoti

Pradinis nagrinėtas logistinės regresijos modelis atrodė taip:

$$Rezultatas\ serga\ ar\ neserga = \beta_0 + \beta_1 \times nėštumų\ skaičius + \beta_2 \times gliukozės\ koncentracija + \beta_3 \times kraujo\ spaudimas + \beta_4 \times insulino\ kiekis + \beta_5 \times odos\ storis + \beta_6 \times KMI + \beta_7 \times diabeto\ atsiradimo\ funkcija + \beta_8 \times amžius.$$

Pašalinus nereikšmingas kovariantes galutinis modelis atrodo taip:

$$\ln\left(\frac{P(pacientas\ serga)}{P(pacientas\ neserga)}\right) = -9{,}864711 + 0{,}036294 \times gliukozės\ koncentracija + 0{,}059561 \times KMI + 0{,}818489 \times diabeto\ atsiradimo\ funkcija + 0{,}061263 \times amžius.$$

Jo AIC 240.81, determinacijos pseudokoeficientas 0,289.

Didėjant gliukozės kiekiui kraujyje, KMI rodikliui, diabeto atsiradimo funkcijos reikšmei bei amžiui didėja galimybė susirgti diabetu. Didėjant gliukozės koncentracijai vienu vienetu 4 proc. labiau tikėtina, kad žmogus sirgs diabetu. Taip pat didinant KMI rodiklį vienu vienetu 6 proc. labiau tikėtina, kad žmogus sirgs cukriniu diabetu. Padidėjus amžiui 1 metais, 6 proc. labiau tikėtina, kad žmogus sirgs diabetu. Tai padidinus šios funkcijos reikšmę per 0,05, 4 procentais bus labiau tikėtina, jog žmogus sirgs diabetu.

Buvo rastas optimaliausias slenkstis, pasinaudojus Youden metodu: 0,28.

Bendras klasifikavimo tikslumas testavimo duomenims yra 0,8. F-1 score - 0,75, plotas po ROC kreive 0,8706. Iš visų rodiklių galime teigti, jog logistinės regresijos modelis tinka duomenims.