VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

I LABORATORINIS DARBAS

DIABETO PROGNOZAVIMAS

MAKSIM ČIŽOV, ANTON CIFIROV

Docentė

Dr. Rūta Levulienė

TURINYS

ĮVADAS IR DUOMENYS

Laboratoriniam darbui atlikti pasirinkome viešai prieinamą duomenų rinkinį „Diabetes“, turinti 768 stebinius ir 8 skaitinius rodiklius (neštumų sk., gliukozės lygis, kraujospūdis ir kiti), bei dvejetainį atributą „Outcome“, nurodanti, ar pacientas serga diabetu. Pasirinkus šį atributą kaip atsako kintamąjį, analizei atlikti nusprendėme pasinaudoti binarinio atsako regresijos modeliu. Modeliui sukurti naudojome R, Python ir SAS programavimo kalbas.

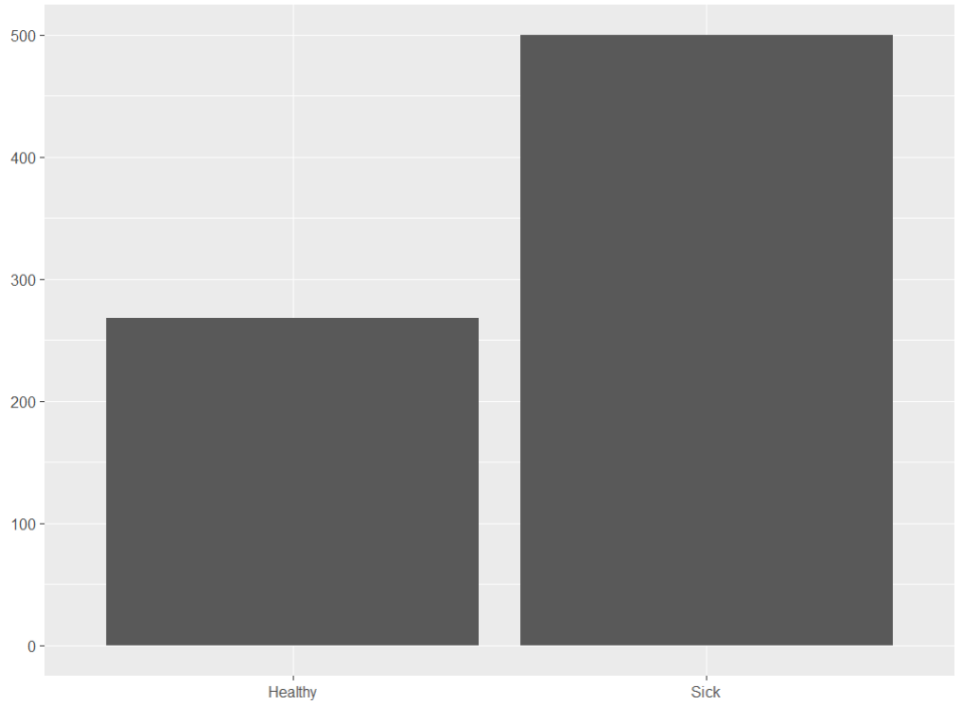
Tikslas - sukruti logistinį binarinio atsako regresijos modelį, kuris iš turimu požymių klasifikuotų, ar pacientas turi diabetą.

Uždaviniai:

1. Atlikti pradinę duomenų analizę
2. Išfiltruoti netinkamus duomenis
3. Atrinkti tinkamus regresorius
4. Sudaryti binarinio atsako modelį ir įvertinti rezultatus

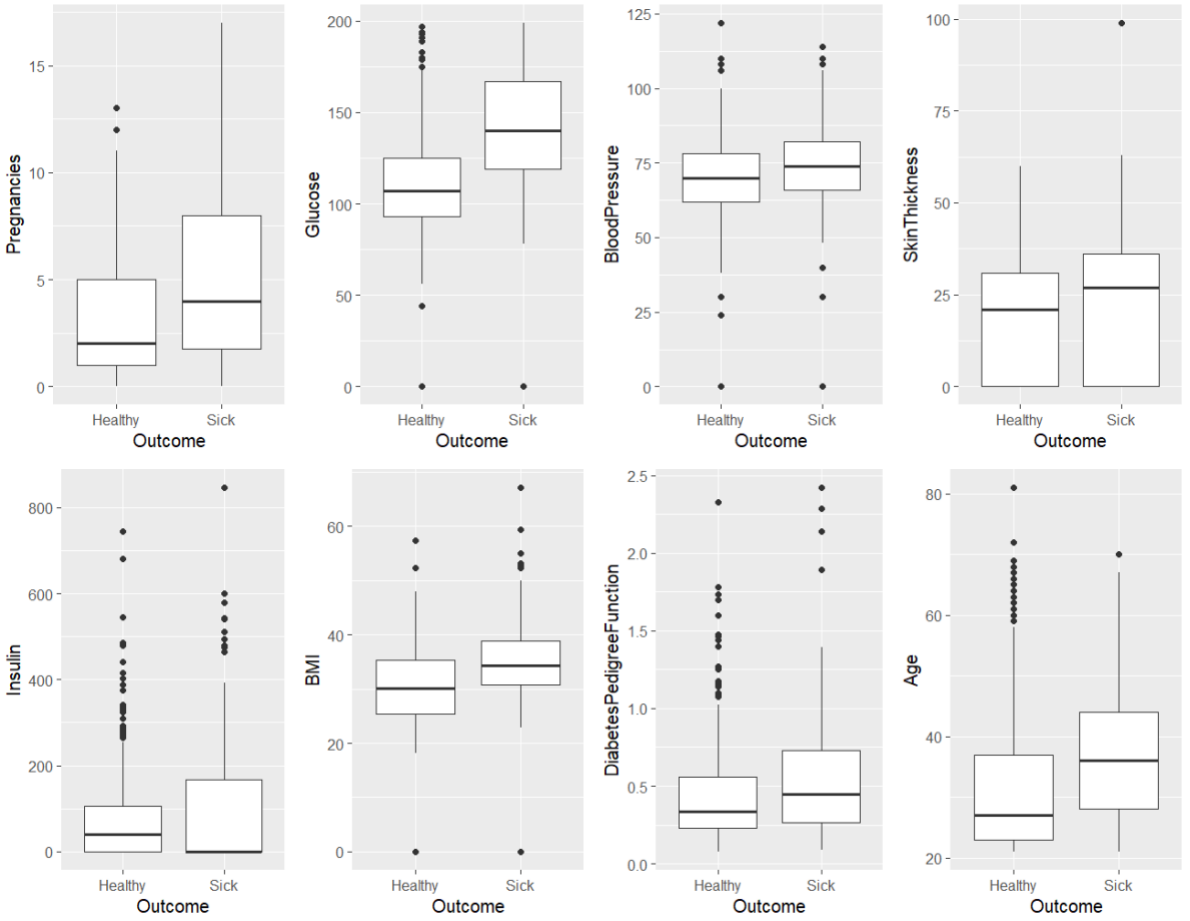
SPRENDIMAS SU R

Įkėlus duomenis į „R“, patikrinome ar įvykių (ne įvykių) dalis yra ne mažiau 20% nuo visų duomenų. Iš stulpelinės diagramos matome, jog taip ir yra.



Pav 1 Duomenų kintamojo Outcome stulpeline diagrama.

Įsitikinus, jog duomenys tinka, išrinkome iš jų 20% stebinių atsitiktinę imtį testavimui, o likusius duomenys palikome modeliui treniruotis. Nubrėžėme 8 stačiakampes diagramas kiekvienam kintamajam – patikrinti, ar matomas grupių atskyrimas tyrimo kintamųjų atžvilgiu. Kadangi mūsų atsako kintamasis „Outcome“ įgyja skaitines reikšmes iš aibės {Healthy, Sick}, transformavome jį į kategorinį kintamąjį, pavertus jį faktoriumi.



Pav 2 Palyginamosios stačiakampės diagramos

Iš stačiakampių diagramų matome, jog didžiausias skirtumas tarp grupių pasiskirstymų yra su kintamaisiais „Pregnancies“, „Glucose“, „BloodPressure“, „BMI“, bei „Age“. Kintamieji „Age“, „DiabetesPredigreeFunction“, bei „Insulin“ turėjo daug išimčių, todėl nuo jų atsisakėme. Po grafinės analizės, sukūrėme pradinį logistinės regresijos modelį, bei patikrinome regresorių reikšmingumą. Kadangi analizė buvo atliekama medicinos srityje, reikšmingumo lygmeniu pasirinkome $\alpha = 0.01$. Iš Pav 3 matome, jog regresorių „Age“, „BloodPressure“, „SkinThikness“ ir „Insulin“ p-reikšmės buvo didesnės už reikšmingumo lygmenį α, todėl jas atmetame. Taigi, pasinaudojus informaciją iš stačiakampių diagramų, bei regresorių p-reikšmėmių, modeliui palikome tik kintamuosius „Pregnancies“, „Glucose“ ir „BMI“.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6972  -0.7341  -0.4111   0.7579   2.9283

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -8.5295878  0.7971213 -10.700  < 2e-16 ***
Pregnancies              0.1337869  0.0366612   3.649 0.000263 ***
Glucose                  0.0349482  0.0041702   8.381  < 2e-16 ***
Age                      0.0051747  0.0106973   0.484 0.628571
BloodPressure           -0.0115066  0.0058512  -1.967 0.049235 *
SkinThickness           -0.0041024  0.0078461  -0.523 0.601071
Insulin                 -0.0007167  0.0010830  -0.662 0.508108
BMI                      0.0997861  0.0169186   5.898 3.68e-09 ***
DiabetesPedigreeFunction 1.0289815  0.3388817   3.036 0.002394 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 800.9  on 614  degrees of freedom
Residual deviance: 579.9  on 606  degrees of freedom
```

Pav 3 Tikėtinumų santykio kriterijai pirminiam modeliui

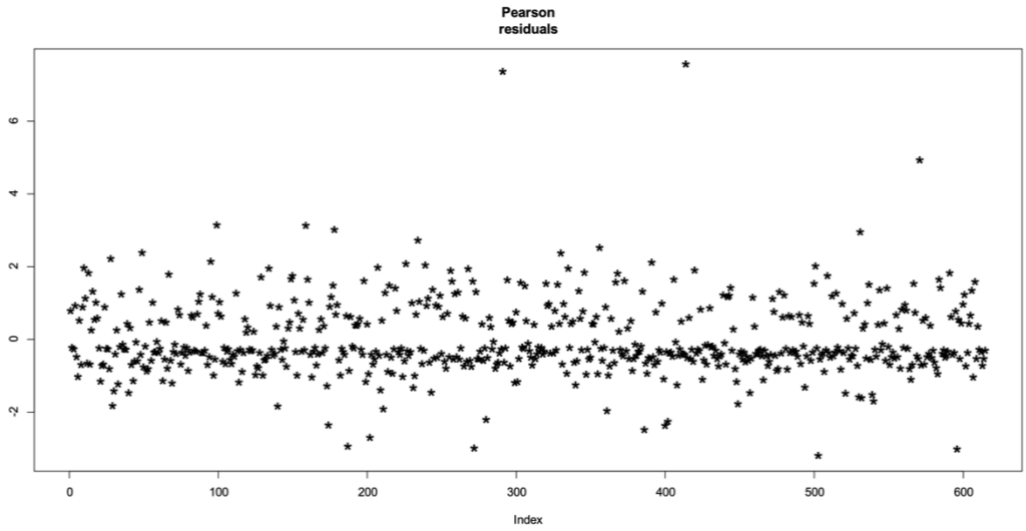
Pašalinus netinkamus kintamuosius, sukūrėme naują modelį (žr. Pav 4), turintį didesnį reikšmingumą.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1956  -0.7237  -0.4213   0.7648   2.8490

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.379294   0.723007 -11.590  < 2e-16 ***
Pregnancies  0.130774   0.030527   4.284 1.84e-05 ***
Glucose      0.033843   0.003662   9.243  < 2e-16 ***
BMI          0.091120   0.015468   5.891 3.85e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 800.90  on 614  degrees of freedom
Residual deviance: 594.49  on 611  degrees of freedom
```

Pav 4 Tikėtinumų santykio kriterijai tarpiniam modeliui

Toliau atlikome Pearson‘o liekanų analizę išimtims surasti. Tam buvo nubrėžtas Pearson‘o liekanų grafikas, iš kurio matome, jog apart bendro triukšmo aplink nulį, yra 6 reikšmės, kurių liekanos gavosi didesnės už 3. Kruopščiau pažiūrėjus į tas išimtis, pastebėjome,

jog keturiuose iš šešių stebinių kintamųjų „Glucose“, „BMI“ ir „BloodPressure“ reikšmės buvo nulinės, kas akivaizdžiai yra neįmanoma. Todėl nusprendėme pašalinti visus stebinius su nulinėmis „Glucose“, „BMI“ ir „BloodPressure“ kintamųjų reikšmėmis iš visų duomenų. Tokių stebinių buvo 44 iš 768, kas neturėtų pabloginti rezultatų.



Pav 5 Tarpinio modelio Pearsono liekanų grafikas

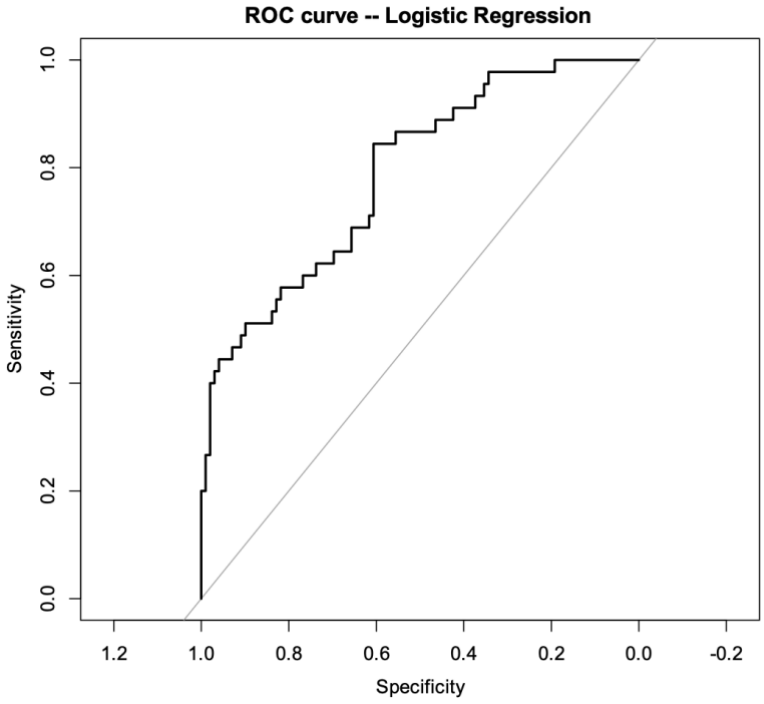
Taigi, išrinkus tinkamus regresorius, bei panaikinus netinkamus stebinius, sukūrėme galutinį binarinio atsako regresijos modelį. Šį modelį patikrinome testiniais duomenimis ir nubrėžėme klasifikavimo lentelę, iš kurios aišku, jog modelio bendras teisingumas yra lygus 77.08 %. Taip pat matome jog “False Negative” klaidos tikimybė yra ganėtinai didelė - 48%, kas yra blogai ypač mūsų atvejui, kadangi “False Negative” reikštų sergantį žmogų suklasifikuoti kaip sveiką.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2396  -0.7100  -0.4049   0.7255   2.2496

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.929394   0.784732 -11.379  < 2e-16 ***
Pregnancies  0.130265   0.032010   4.070 4.71e-05 ***
Glucose      0.036890   0.003869   9.535  < 2e-16 ***
BMI          0.093624   0.016564   5.652 1.58e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 752.27  on 579  degrees of freedom
Residual deviance: 547.65  on 576  degrees of freedom
```

```
> conf_m1

model_pred1 Healthy Sick
    Healthy      88   22
    Sick         11   23
>  test_acc1
[1] 0.7708333
> fn_prob1
[1] 0.4888889
> fp_prob1
[1] 0.1111111
>
```

Pav 6 *Tikėtinumų santykio kriterijai ir klasifikavimo lentelė galutiniam modeliui slenkstis = 0.5*

Parinkus optimalų slenkstį naudojant *coords()* R funkcija gavome galutinį logistinės regresijos modelį.




```
> opt_slenkstis$roc_curve.thresholds
[1] 0.2541632
> conf_m2

model_pred2 Healthy Sick
    Healthy      65   15
    Sick         34   30
> test_acc2
[1] 0.6597222
> fn_prob2
[1] 0.3333333
> fp_prob2
[1] 0.3434343
```

Pav 7 ROC kreivė ir klasifikavimo lentelė galutiniam modeliui, kai slenkstis = 0.2541

Taigi, bendras modelio tikslumas sumažėjo iki 66%, tačiau tikimybe klasifikuoti sergantį žmogų kaip sveiką, sumažėjo iki 33.3%, taip pat tikimybe sveiką žmogų klasifikuoti kaip sergantį padidėjo iki ~34.3%.

SPRENDIMAS SU SAS

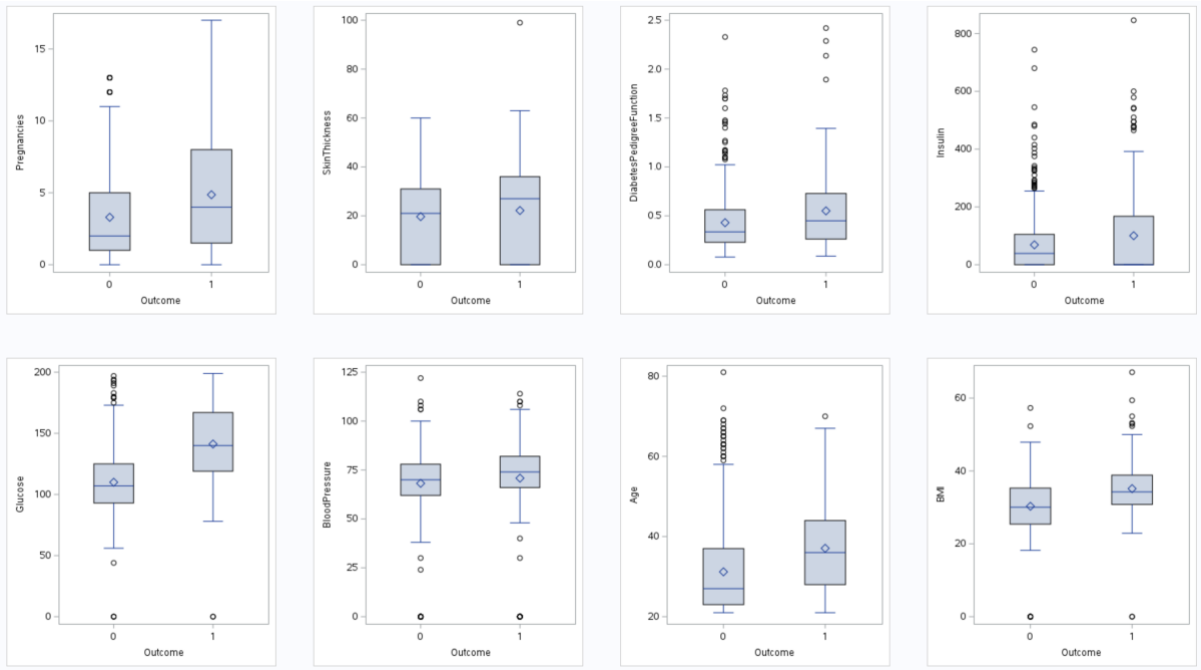
Įkėlus duomenys į SAS studio aplanką, nuskaitėme jas ir peržiūrėjome dažnių lentele. Tokiu būdu patikrinome ar įvykio (ne įvykio) dalis sudaro bent 20 procentų visų duomenų.

The FREQ Procedure

Outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	500	65.10	500	65.10
1	268	34.90	768	100.00

Pav 8 Duomenų kintamojo Outcome dažnių lentelė.

Patikrinus įvykio (ne įvykio) santykį, nubraižėme stulpelines diagramas kiekvienam busimo modelio kintamajam. Remiantis gautais grafikais, vizualiai patikrinome mūsų duomenys, ar matomas grupių atskyrimas tyrimo kintamųjų atžvilgiu.



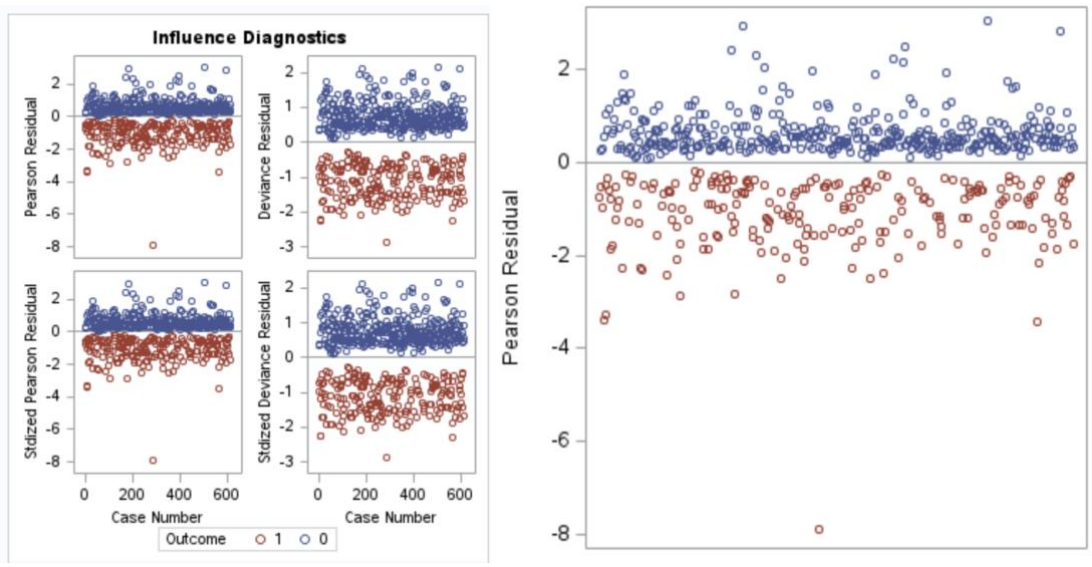
Pav 9 Palyginamosios stačiakampės diagramos

Padalinome duomenys į apmokymo ir testavimo dalys atitinkamai 80% ir 20%. Sukūrėme pirmini modelį į kurį įtraukėme visus galimus mūsų lenteles kintamuosius: „Pregnancies“, „Glucose“, „BloodPressure“, „SkinThickness“, „Insulin“, „BMI“, „DiabetesPedigreeFunction“ ir „Age“. Sukūrus pirminį modelį, apskaičiavome SAS’o LOGISTIC koeficientų įverčius:

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-7.8146	0.7728	102.2465	<.0001
Pregnancies	1	0.1137	0.0347	10.7409	0.0010
Glucose	1	0.0354	0.00409	75.1460	<.0001
BloodPressure	1	-0.0126	0.00565	4.9803	0.0256
SkinThickness	1	-0.00078	0.00751	0.0107	0.9176
Insulin	1	-0.00122	0.000957	1.6364	0.2008
BMI	1	0.0812	0.0163	24.7215	<.0001
DiabetesPedigreeFunc	1	0.6618	0.3288	4.0515	0.0441
Age	1	0.0113	0.0102	1.2204	0.2693

Pav 10 Didžiausio tikėtinumo įverčiai pirminiam modeliui

Medicinos srityje dažnai naudojamas mažesnis statistinio reikšmingumo lygmuo alpha = 0.01, todėl ir mes pasirinkome tokią lygmens reikšmę. Kadangi kintamųjų BloodPressure SkinThickness Insulin DiabetesPedigreeFunction Age p-reikšmės buvo didesnės už mūsų pasirinktos alpha = 0.01, jas atmetame. Be to, remiantis R‘o analize, pašalinome iš duomenų stebėjimus, kur BloodPressure, Glucose ar BMI kintamieji yra lygus nuliui, nes akivaizdu kad taip būti negali. Atlikome Pearsono liekanų analize



Pav 11 Tarpinio modelio Pearsono liekanų grafikas

Po kintamųjų atrinkimo, bei duomenų filtravimo, gavome galutinį modelį, kurio tikslumą tikrinome iš didžiausio tikėtinumo lentelės, bei klasifikavimo lentelės.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-7.6644	0.6881	124.0677	<.0001
Pregnancies	1	0.1275	0.0289	19.5034	<.0001
Glucose	1	0.0341	0.00365	87.2100	<.0001
BMI	1	0.0706	0.0147	23.1004	<.0001

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	805.243	623.403
SC	809.665	641.089
-2 Log L	803.243	615.403

R-Square	0.2632	Max-rescaled R-Square	0.3610

Paskaičiavome, jog bendras modelio tikslumas = 83%. Iš Pav 13 matome, jog "False positive" tikimybė = 34%, "False negative" tikimybe = 9.4%. Kadangi „False Positive" tikimybė gavosi aukšta, bandėme parinkti optimalų atmetimo slenkstį.

The FREQ Procedure

Table of F_Outcome by I_Outcome

F_Outcome(From: Outcome)	I_Outcome(Into: Outcome) 0	1	Total
0	96 62.75	10 6.54	106 69.28
1	16 10.46	31 20.26	47 30.72
Total	112 73.20	41 26.80	153 100.00

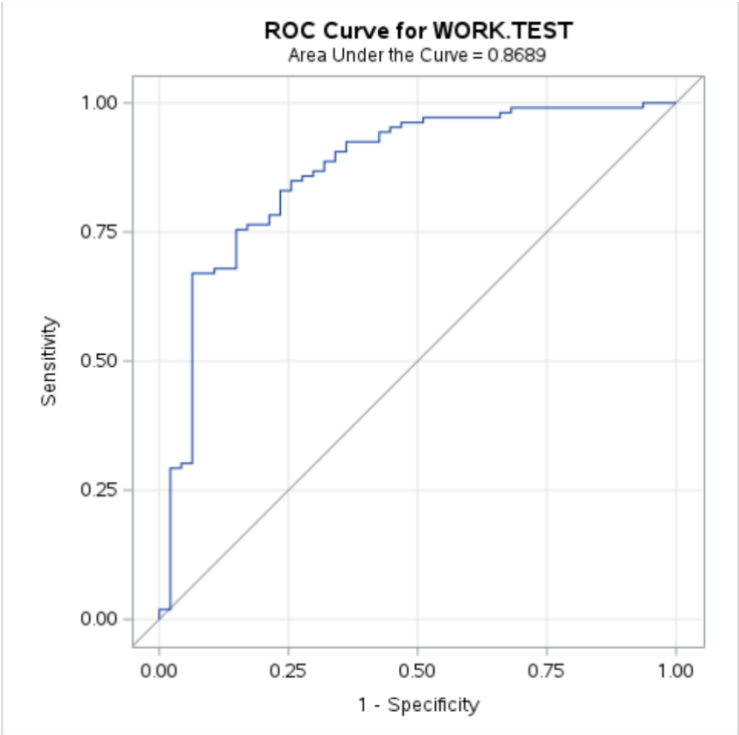
Pav 12 Klasifikavimo lentelė

	Positive	Tikimybes
1	16	0.3404255319
2	31	0.6595744681

	Negative	Tikimybes
1	96	0.9056603774
2	10	0.0943396226

Pav 13 Klasifikavimo lentelė

Optimaliajam slenksčiui parinkti, nubrėžėme ROC grafiką. Plotas po ROC kreivė gavosi lygus 0.8689, kas yra pakankamai gerai. Optimalų slenkstį suradome naudojant Youden indeksą ir gavome reikšme = 0.350993.



Pav 14 ROC grafikas

Pasirinkus optimalųjį slenkstį, vėl atlikome tikslumo analizę. Kaip matome, „False Positive“ tikimybė žymiai sumažėjo - iki 23.4%, tačiau tuo pat metu truputi sumažėjo bendras tikslumas iki 79.7%, o „False Negative“ tikimybė užaugo iki 18.9%.

The FREQ Procedure

Table of F_Outcome by I_Outcome_n			
	I_Outcome_n		
F_Outcome(From: Outcome)	0	1	Total
0	86 56.21	20 13.07	106 69.28
1	11 7.19	36 23.53	47 30.72
Total	97 63.40	56 36.60	153 100.00

Pav 15 Klasifikavimo lentelė galutiniam modeliui

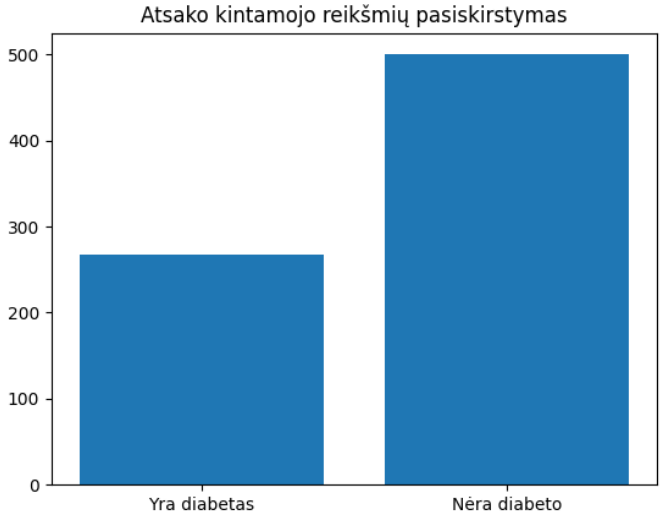
	Positive	Tikimybes
1	11	0.2340425532
2	36	0.7659574468

	Negative	Tikimybes
1	86	0.8113207547
2	20	0.1886792453

Pav 16 Klasifikavimo lentelė galutiniam modeliui

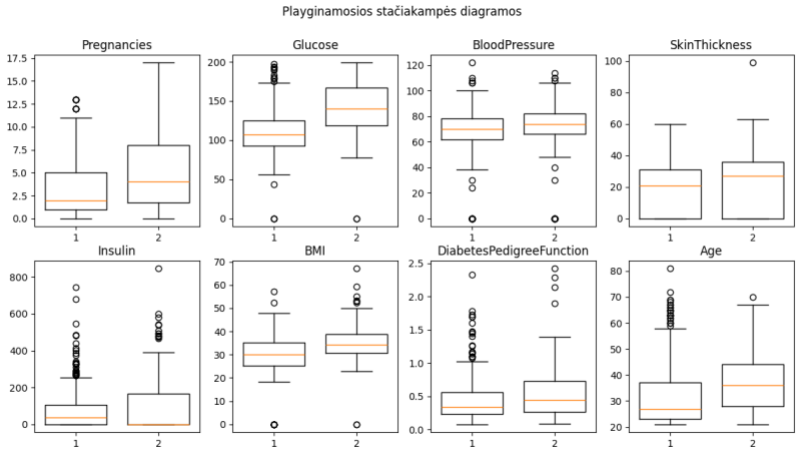
SPRENDIMAS SU PYTHON

Nuskaičius duomenis į „Python“, pirmiausiai patikrinome, ar atsako kintamojo teigiamos reikšmes sudaro bent 20% dalį iš visų duomenų. Kaip matome iš Pav 17 , 34.5% atsako kintamojo reikšmių yra teigiamos. Todėl galime rinktis šį kintamąjį kaip atsako.



Pav 17 Duomenų kintamojo Outcome stulpelinė diagrama.

Įsitikinus, jog duomenis yra tinkami, iš jų išrinkome 20% atsitiktinę imtį modelio testavimui. Taip pat nubrėžėme stačiakampes diagramas, kurios sutapo su „R“ diagramomis, tad išvados gavosi tokios pat: kintamieji „Age“, „DiabetesPredigreeFunction“, bei „Insulin“ turėjo daug išimčių, todėl nuo jų atsisakėme.



Pav 18 Palyginamosios stačiakampės diagramos

Sukūrėme binarinio atsako modelį su visais turimais kintamaisiais ir patikrinome regresorių reikšmingumą. Reikšmingumo lygmeniu vėl pasirinkome $\alpha = 0.01$. Iš Pav 19

matome, jog šįkart regresoriai su p-reikšmėmis didesnėmis už α liko tokie patys : „Age“, „BloodPressure“, „SkinThikness“ ir „Insulin“, tad jas vėl atmetėme. Taigi, pasinaudojus informaciją iš stačiakampių diagramų, bei regresorių p-reikšmėmis, modeliui palikome kintamuosius „Pregnancies“, „Glucose“ ir „BMI“.

Generalized Linear Model Regression Results

Dep. Variable:	Outcome	No. Observations:	615
Model:	GLM	Df Residuals:	606
Model Family:	Binomial	Df Model:	8
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-296.51
Date:	Fri, 01 Mar 2024	Deviance:	593.02
Time:	16:30:22	Pearson chi2:	665.
No. Iterations:	5	Pseudo R-squ. (CS):	0.2854
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-7.9415	0.787	-10.089	0.000	-9.484	-6.399
Pregnancies	0.1281	0.034	3.726	0.000	0.061	0.195
Glucose	0.0338	0.004	8.366	0.000	0.026	0.042
Age	0.0133	0.010	1.294	0.196	-0.007	0.033
BloodPressure	-0.0165	0.006	-2.899	0.004	-0.028	-0.005
SkinThickness	0.0009	0.008	0.118	0.906	-0.014	0.016
Insulin	-0.0009	0.001	-0.868	0.386	-0.003	0.001
BMI	0.0895	0.017	5.350	0.000	0.057	0.122
DiabetesPedigreeFunction	0.8367	0.319	2.622	0.009	0.211	1.462

Pav 19 Tikėtinumų santykio kriterijai pirminiam modeliui

Antro modelio regresorių p-reikšmės gavosi $< \alpha = 0.01$, todėl tesėmė analizę ieškodami išimčių

Generalized Linear Model Regression Results

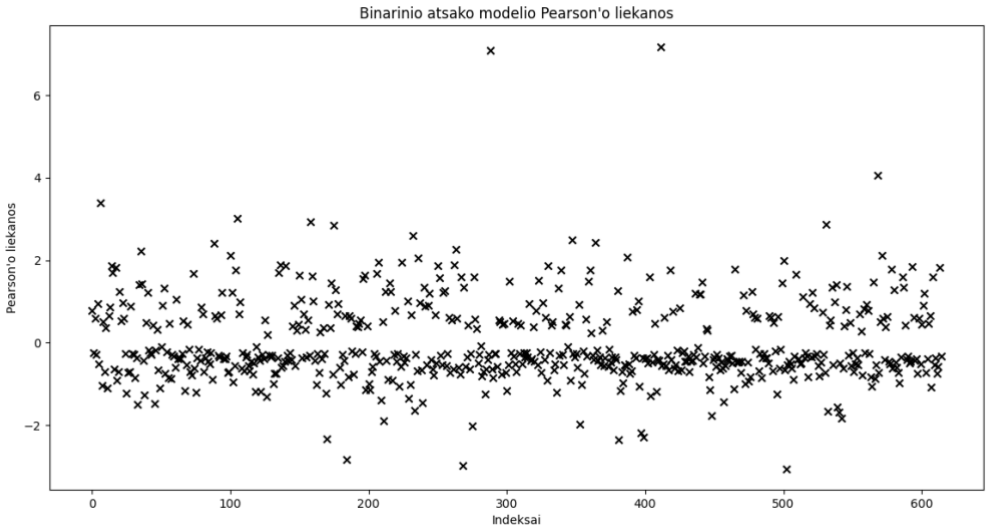
Dep. Variable:	Outcome	No. Observations:	615
Model:	GLM	Df Residuals:	611
Model Family:	Binomial	Df Model:	3
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-305.27
Date:	Fri, 01 Mar 2024	Deviance:	610.54
Time:	16:30:22	Pearson chi2:	660.
No. Iterations:	5	Pseudo R-squ. (CS):	0.2648
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-7.9533	0.711	-11.188	0.000	-9.347	-6.560
Pregnancies	0.1359	0.029	4.699	0.000	0.079	0.193
Glucose	0.0329	0.004	9.153	0.000	0.026	0.040
BMI	0.0819	0.015	5.410	0.000	0.052	0.112

Pav 20 Tikėtinumų santykio kriterijai tarpiniam modeliui

Išimtis ieškojome pasinaudojus Pearson‘o liekanų analizę. Nubrėžėme Pearson‘o liekanų grafiką, iš kurio pastebime, jog šįkart gavome tik 4 reikšmes, kurių liekanos buvo didesnės už

3. Kadangi jau pastebėjome, jog tarp išimčių yra daug nulinių reikšmių, jas ištrynėme. Ištrintu stebėjimų vėl gavosi 44.



Pav 21 Tarpinio modelio Pearsono liekanų grafikas

Taigi, sukūrėme galutinį binarinio atsako regresijos modelį. Šio modelio regresorių p-reikšmės gavosi tinkamos, todėl toliau tikrinome modelio tikslumą.

Generalized Linear Model Regression Results

Dep. Variable:	Outcome	No. Observations:	580
Model:	GLM	Df Residuals:	576
Model Family:	Binomial	Df Model:	3
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-281.24
Date:	Fri, 01 Mar 2024	Deviance:	562.49
Time:	16:30:22	Pearson chi2:	547.
No. Iterations:	5	Pseudo R-squ. (CS):	0.2744
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-8.3595	0.766	-10.918	0.000	-9.860	-6.859
Pregnancies	0.1388	0.030	4.589	0.000	0.080	0.198
Glucose	0.0354	0.004	9.338	0.000	0.028	0.043
BMI	0.0819	0.016	5.065	0.000	0.050	0.114

Pav 22 Tikėtinumų santykio kriterijai galutiniam modeliui

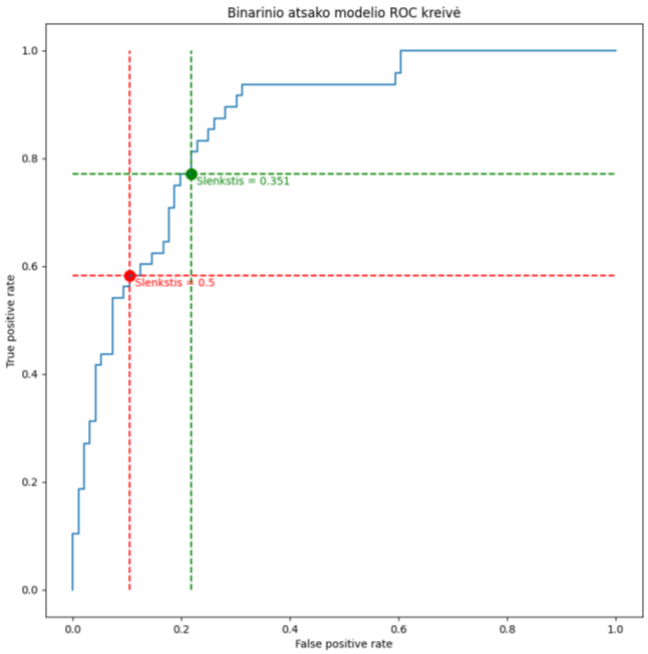
Modelio tikslumą vertinome patikrinus modelį su testiniais duomenimis, bei pasirinkus atmetimo slenkstį = 0.5 . Nubrėžėme klasifikavimo lentelę, iš kurios gavome bendrą

teisingumą lygu 79.17%, o „False Negative“ ir „False Positive“ tikimybės gavosi lyhios atitinkamai 41.7% ir 10.4%. Kadangi „False Negative“ tikimybė gavosi aukšta, bandėme parinkti optimalų atmetimo slenkstį.

```
Klasifikavimo lentele
      0    1
FALSE 86   20
TRUE  10   28

Klasifikavimo lentele
          0      1
FALSE  0.895833 0.416667
TRUE   0.104167 0.583333

Patikimuno lygis =  0.7916666666666666
False Positive rate =  0.10416666666666667
False Negative rate =  0.4166666666666667
```

Pav 23 Klasifikavimo lentelė

Optimaliajam slenksčiui parinkti nubrėžėme ROC grafiką, iš kurio matome, jog plotas po kreivė yra lygus 0.86610, kas yra ganėtinai gerai, o optimalus slenkstis šiuo atveju = 0.3507



Pav 24 ROC grafikas

Pasirinkus optimalųjį slenkstį, vėl atlikome tikslumo analizę. Kaip matome, „False Negative“ tikimybė žymiai sumažėjo - beveik dukart, iki 22.9%, tačiau tuo pat metu truputi sumažėjo bendras tikslumas iki 77.8%, o „False Postitive“ tikimybė užaugo iki 21.9%.

```
Klasifikavimo lentele
      0    1
FALSE 75   11
TRUE  21   37

Klasifikavimo lentele
          0      1
FALSE  0.78125 0.229167
TRUE   0.21875 0.770833

Patikimuno lygis =  0.7777777777777778
False Positive rate =  0.21875
False Negative rate =  0.22916666666666666
```

Pav 25 Galutinio modelio klasifikavimo lentelė

IŠVADOS

Sukurėme logistinės regresijos modelius SAS, Python ir R programavimo kalbomis, padalinome duomenys į apmokymo ir testavimo imtys, išmetėme netinkamus stebinius. Pasinaudojus Pearsono liekanomis, pašalinome išskirtis, bei regresorius, kurių koeficientų statistinis reikšmingumas buvo mažesnis už norimą. Atlikdami modelio tikslumo analizę su testiniais duomenimis, parinkome optimalų slenkstį. Naudojant testinius duomenys patikrinome mūsų modelio bendrą tikslumą, bei "False positive" ir "False negative" tikimybes. Modelių, sukurtų skirtingomis kalbomis, rezultatai gavosi skirtingi, dėl to, nes testavimo ir apmokymo imtys buvo sugeneruotos atsitiktinai iš visų duomenų, o kiekviena programavimo kalba tai padarė skirtingai. Vienas iš galimu būdų išspręsti šią problemą būtų tikslumo analizei naudoti cross-validation ar bootstrap algoritmus.