Vilniaus Universitetas

Logistinė regresija

Laboratorinis darbas

Darbą atliko:

Vainius Gataveckas, Matas Gaulia, Dovydas Martinkus

Duomenų Mokslas

3 kursas 2 gr.

Vilnius, 2022

Naudoti metodai

Darbas atliktas naudojant R ir SAS.

Naudoti R paketai:

tidyverse

caret

MASS

cutpointr

yardstick

effects

Duomenys ir jų šaltiniai

Pimų tautybės moterų diagnostiniai matavimai skirti nustatyti ar pacientas sergama diabetu.

Duomenų šaltinis - Kaggle. Prieiga per internetą: https://www.kaggle.com/uciml/pima-indians-diabetes-database

„Pregnancies“ - neštumų kiekis.

„Glucose“ - gliukozės koncentracija plazmoje gliukozės tolerancijos testo metu.

„BloodPressure“ - diastolinis kraujo spaudimas.

„BMI“ – kūno masės indeksas.

„SkinThickness“ - tricepso odos plotis.

„Insulin“ - gliukozės tolerancijos testo rezultatas.

„DiabetesPedigreeFunction“ - diabeto tikėtinumas remiantis šeimos istorija.

„Age“ – amžius.

„Outcome“ – diabeto diagnozė (atsako kintamasis).

Tikslas ir uždaviniai

Tikslas: Rasti kokią įtaką tam tikri požymiai daro tikimybei sirgti diabetu ir prognozuoti diagnozę ar pacientas serga diabetu.

Uždaviniai:

Sudaryti binarinio atsako modelį.

Modelio tinkamumo analizė.

Paprastesnio (turinčio mažiau kovariančių) modelio suradimas.

Gautų modelio koeficientų interpretacija.

Slenkstinės reikšmės parinkimas.

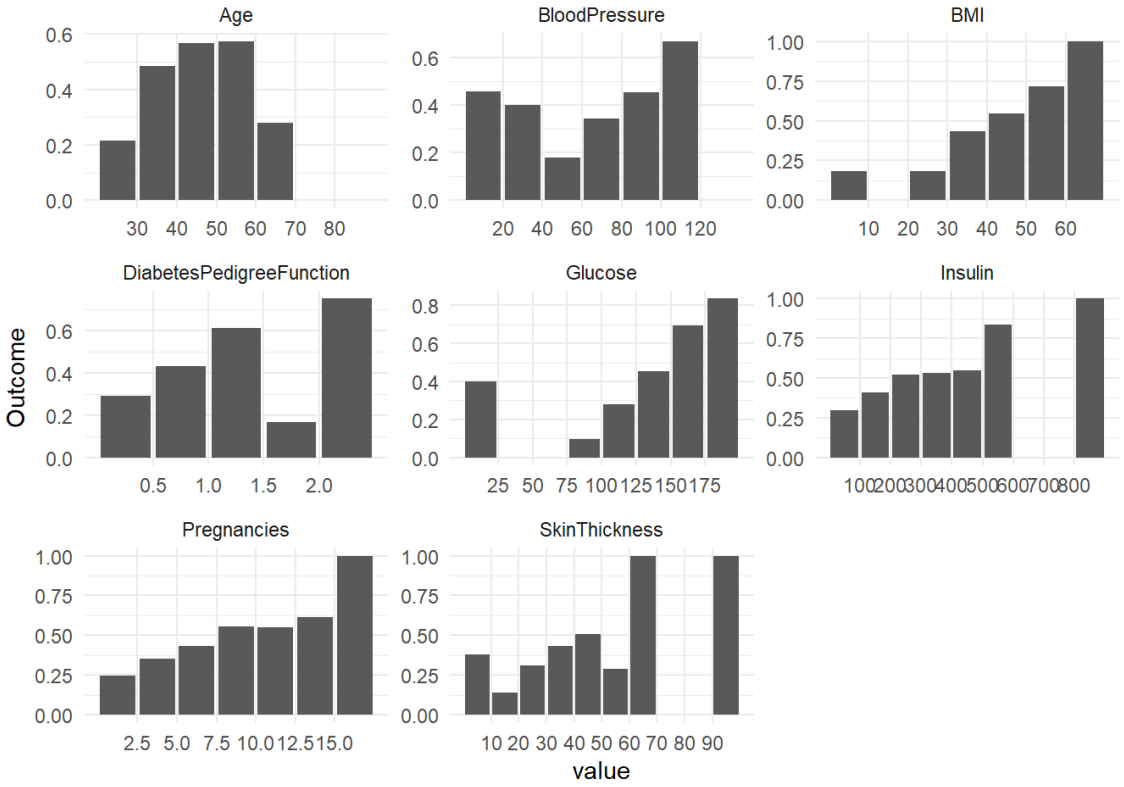
Modelio taikymas prognozėms.

Atliktos analizės aprašymas

1. Naudojant R

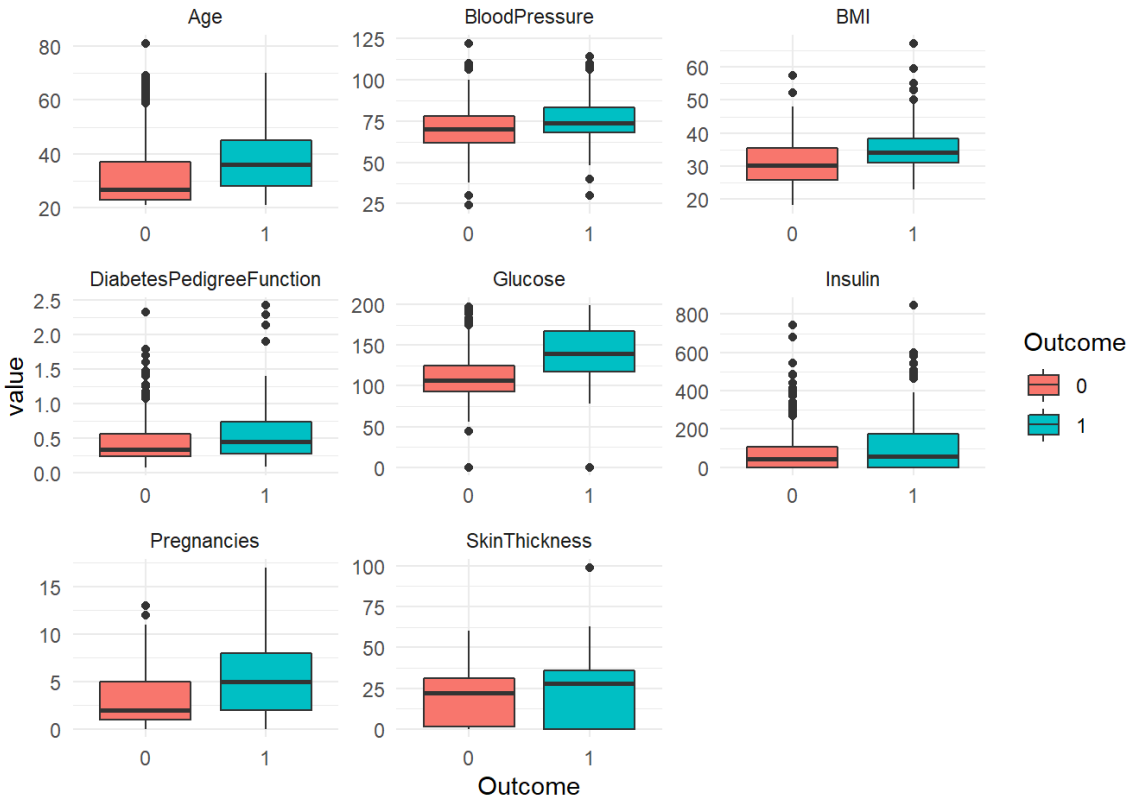
Duomenų aibę sudaro duomenys apie 500 diabetų nesergančių ir 268 sergančių pacientų. Pašalinus praleistas reikšmes doumenų aibėje lieka duomenys apie 478 nesergančius ir 251 sergančius pacientus. Atliekant tiriamąją duomenų analizę palygintas kovariančių pasiskirstymas abiejose grupėse naudojant stačiakampes diagramas, pavaizduotos empirinės sirgimu diabetu tikimybės pagal kiekvieną kovariantę.

```
library(tidyverse)

y <- read_csv("diabetes.csv")
y <- y %>% filter(BloodPressure != 0,BMI != 0)

table(y$Outcome)

##
##   0   1
## 478 251

# Empirinės tikimybės

y_plot <- y %>% pivot_longer(1:8)


y_plot %>% ggplot(aes(value, Outcome)) +
  stat_summary(fun = mean, geom = "bar") +
  facet_wrap(vars(name), scales = "free") +
  scale_x_binned(n.breaks = 8) +
  theme_minimal()
```




```
# stačiakampės diagramos
y <- y %>% mutate(Outcome = factor(Outcome))
y_plot <- y_plot %>% mutate(Outcome = factor(Outcome))
```

```
y_plot %>% ggplot(aes(Outcome, value, fill = Outcome)) +
  geom_boxplot() +
  facet_wrap(vars(name), scales = "free") +
  theme_minimal()
```




```
library(caret)
library(yardstick)

model <- glm(
  formula = Outcome ~ ., family = binomial(logit),
  data = y
)



summary(model)

##
## Call:
## glm(formula = Outcome ~ ., family = binomial(logit), data = y)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5812  -0.7207  -0.4158   0.7383   2.8530
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -8.8170408  0.8117630 -10.862  < 2e-16 ***
## Pregnancies               0.1150347  0.0330853   3.477 0.000507 ***
## Glucose                   0.0336410  0.0037572   8.954  < 2e-16 ***
## BloodPressure            -0.0100157  0.0086060  -1.164 0.244501
## SkinThickness             0.0007060  0.0070117   0.101 0.919797
## Insulin                  -0.0010220  0.0009144  -1.118 0.263709
## BMI                       0.0965139  0.0167561   5.760 8.42e-09 ***
## DiabetesPedigreeFunction  0.9995318  0.3061460   3.265 0.001095 **
```

```
## Age                         0.0179745  0.0097612   1.841 0.065558 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 938.74  on 728  degrees of freedom
## Residual deviance: 687.32  on 720  degrees of freedom
## AIC: 705.32
##
## Number of Fisher Scoring iterations: 5
```

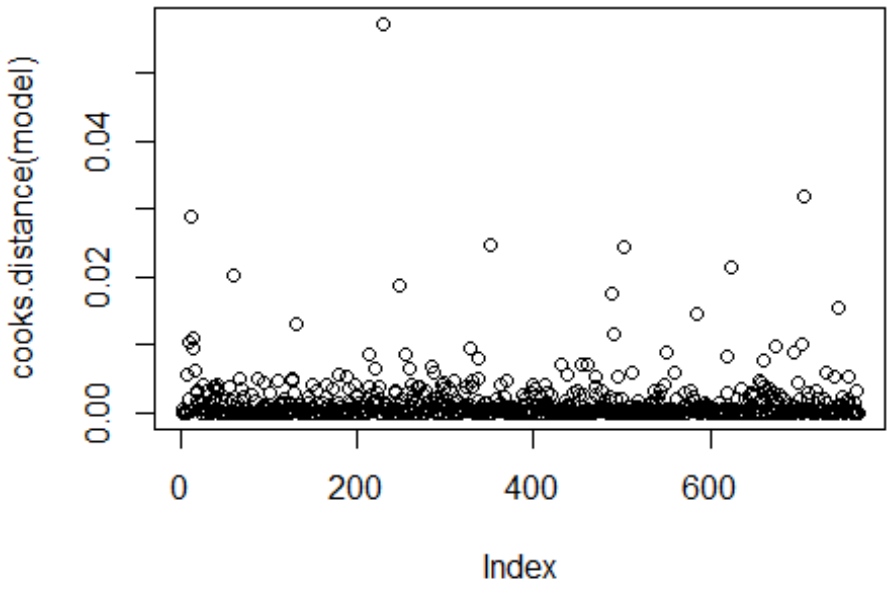
```
1-pchisq(model$null.deviance-model$deviance, model$df.null-model$df.residual) # globali nulinė hipotezė (tikėtinumo santykių testas likelihood ratio test)
```

```
## [1] 0
```

```
1-pchisq(model$deviance,model$df.residual) # residual goodness-of-fit testas
```

```
## [1] 0.8042267
```

```
# tikrinama, ar yra išskirtys
plot(cooks.distance(model))
```




```
confusionMatrix(factor(as.numeric(model$fitted.values > 0.5)), factor(y$Outcome),positive="1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 424 109
##          1  54 142
##
##                Accuracy : 0.7764
##                  95% CI : (0.7444, 0.8062)
##     No Information Rate : 0.6557
##     P-Value [Acc > NIR] : 8.155e-13
```

```
##
##                   Kappa : 0.4776
##
##  Mcnemar's Test P-Value : 2.341e-05
##
##             Sensitivity : 0.5657
##             Specificity : 0.8870
##          Pos Pred Value : 0.7245
##          Neg Pred Value : 0.7955
##              Prevalence : 0.3443
##          Detection Rate : 0.1948
##    Detection Prevalence : 0.2689
##       Balanced Accuracy : 0.7264
##
##        'Positive' Class : 1
##
```

```
# plotas po ROC
y_2 <- y %>% mutate(pred = model$fitted.values)
roc_auc(y_2, Outcome, pred, event_level = "second")
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.838
```

```
# multikolinearumo tikrinimas
signs <- (model$coefficients > 0)[-1]
name <- names(model$coefficients)[-1]

temp_model <- function(x) {
  x <- sym(x)
  temp_model <- glm(
    formula = Outcome ~ eval(x), family = binomial(logit),
    data = y)
  temp_model$coefficients[2] > 0
}


map(name,temp_model) == signs
```

```
##              Pregnancies                  Glucose            BloodPressure
##                     TRUE                     TRUE                    FALSE
##            SkinThickness                  Insulin                      BMI
##                     TRUE                    FALSE                     TRUE
## DiabetesPedigreeFunction                      Age
##                     TRUE                     TRUE
```

```
# pašalinamas kintamasis kurio koeficiento ženklas modelyje neatitinka jo įtakos
model <- glm(
  formula = Outcome ~ Pregnancies + Glucose + SkinThickness + BMI + DiabetesPedigreeFunction + Age, fam
ily = binomial(logit),
  data = y
)
```

Hipotezė apie reikšmingų koeficientų nebuvimą atmesta (p=0). Pradinio modelio su visomis kovariantėmis (naudojant logit jungties funkciją) tikslumas (angl. accuracy) 78%, plotas po ROC kreive 0.84.
Modelyje išskirčių nerasta (naudojant Kuko matą).
Tikrinant multikolinearumą rasta, kad kovariančių „BloodPressure“ ir „Insulin“ ženklai modelyje priešingi jų įtakai. Pasirinkta šias kovariantes pašalinti iš modelio.

```
#reikšmingų kovariančių atranka
model_2 <- MASS::stepAIC(model,direction = "both")
```

```
## Start:  AIC=703.63
## Outcome ~ Pregnancies + Glucose + SkinThickness + BMI + DiabetesPedigreeFunction +
##     Age
##
##                            Df Deviance    AIC
## - SkinThickness             1   689.73 701.73
## <none>                          689.63 703.63
## - Age                       1   692.48 704.48
## - DiabetesPedigreeFunction  1   700.49 712.49
## - Pregnancies               1   702.17 714.17
## - BMI                       1   725.28 737.28
## - Glucose                   1   795.99 807.99
##
## Step:  AIC=701.73
## Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction +
##     Age
##
##                            Df Deviance    AIC
## <none>                          689.73 701.73
## - Age                       1   692.73 702.73
## + SkinThickness             1   689.63 703.63
## - DiabetesPedigreeFunction  1   700.49 710.49
## - Pregnancies               1   702.33 712.33
## - BMI                       1   729.26 739.26
## - Glucose                   1   796.02 806.02
```

```
anova(model, model_2, test = "Chisq") # modelis statistiškai reikšmingai nesiskiria nuo modelio su visomis kovariantėmis
```

```
## Analysis of Deviance Table
##
## Model 1: Outcome ~ Pregnancies + Glucose + SkinThickness + BMI + DiabetesPedigreeFunction +
##     Age
## Model 2: Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction +
##     Age
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       722     689.63
## 2       723     689.73 -1 -0.10723   0.7433
```

```
model$aic
```

```
## [1] 703.6277
```

```
model_2$aic
```

```
## [1] 701.735
```

```
confusionMatrix(factor(as.numeric(model_2$fitted.values > 0.5)), factor(y$Outcome),positive="1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 425 109
##          1  53 142
##
##                Accuracy : 0.7778
##                  95% CI : (0.7458, 0.8075)
##     No Information Rate : 0.6557
##     P-Value [Acc > NIR] : 4.424e-13
##
##                   Kappa : 0.4803
##
##  Mcnemar's Test P-Value : 1.552e-05
##
##             Sensitivity : 0.5657
```

```
##             Specificity : 0.8891
##          Pos Pred Value : 0.7282
##          Neg Pred Value : 0.7959
##              Prevalence : 0.3443
##          Detection Rate : 0.1948
##    Detection Prevalence : 0.2675
##       Balanced Accuracy : 0.7274
##
##        'Positive' Class : 1
##
```

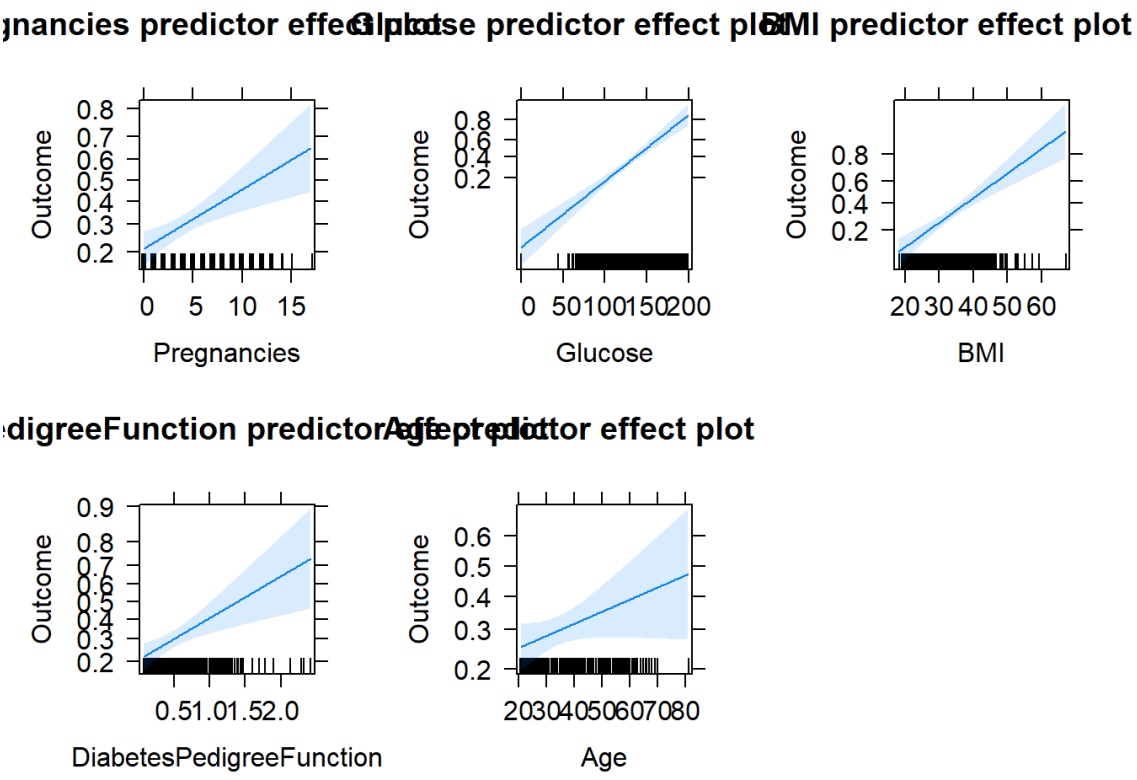
```
# koeficientų interpretacija
exp(coef(model_2))
```

```
##              (Intercept)              Pregnancies                  Glucose
##             0.0001142755             1.1218074285             1.0322329981
##                      BMI DiabetesPedigreeFunction                      Age
##             1.0935207918             2.6354624042             1.0164328523
```

```
exp(confint(model_2))
```

```
##                                 2.5 %       97.5 %
## (Intercept)                2.548831e-05 0.0004624005
## Pregnancies                1.052473e+00 1.1973554245
## Glucose                    1.025492e+00 1.0393887478
## BMI                        1.062607e+00 1.1266938678
## DiabetesPedigreeFunction 1.471138e+00 4.7896731129
## Age                        9.978306e-01 1.0353049391
```

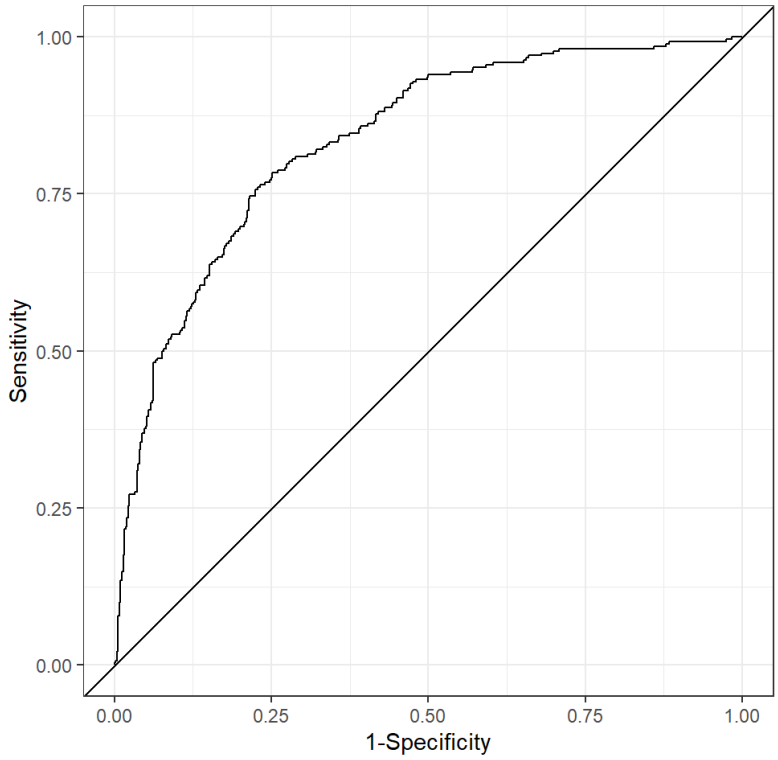
```
# modelio kovariačių efektai
library(effects)
plot(predictorEffects(model_2))
```



Naudojant pažingsninę regresiją remiantis AIC rasta, parinktas modelis be kovariantės „SkinThickness“ ir rasta, kad šis modelis statistiškai reikšmingai nesiskiria nuo modelio su visomis kovariantėmis (p=0.74). Modelio tikslumas 78%. Plotas po ROC kreive 0.84. Verta paminėti, kad šių metrikų reikšmės sutampa su prieš tai sudaryto sudėtingesnio modelio.

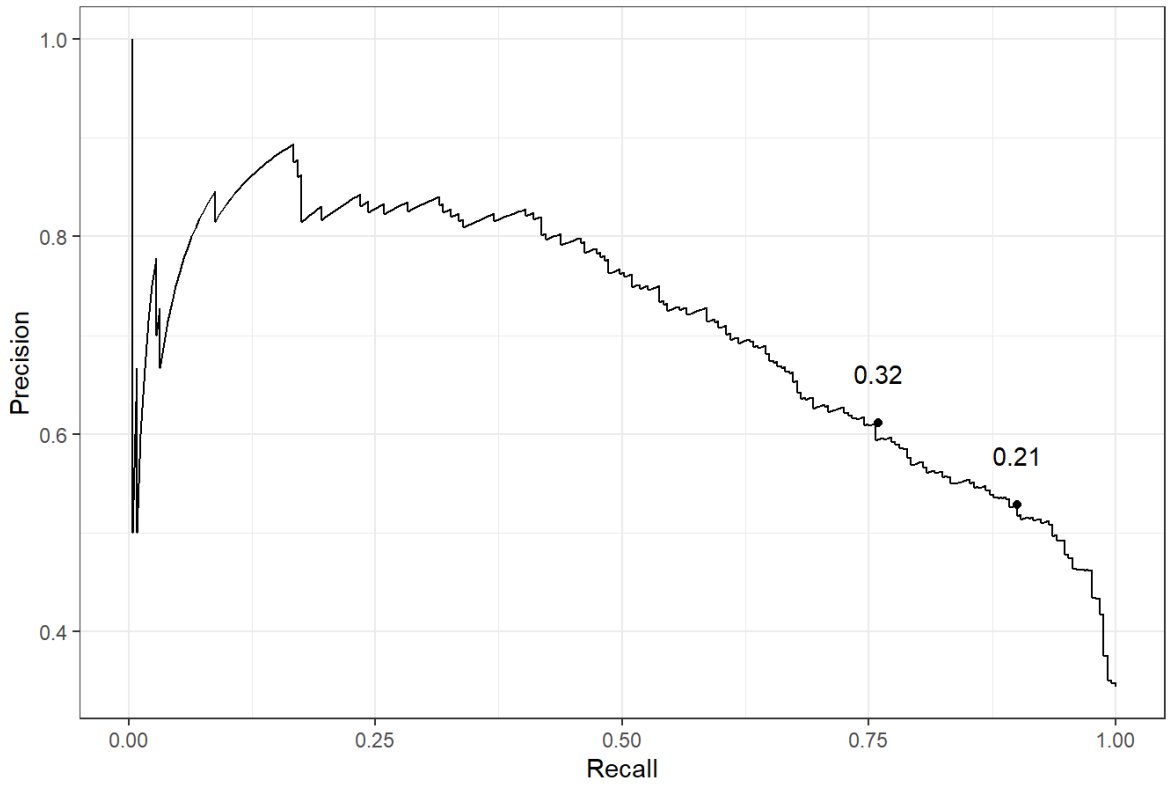
Modelio koeficientų interpretacija standartinė logit modeliui (pvz. Paciento kūno masės indeksui (angl. BMI) padidėjus 1, kitoms kovariantėms esant fiksuotoms, galimybė pacientui sirgti diabetu padidėja 1.09 kartų).

```
# ROC kreivė
library(cutpointr)

y_2 <- y %>% mutate(pred = model_2$fitted.values)

cp <- cutpointr(y_2, pred, Outcome,
  pos_class = "1", direction = ">=",
  method = maximize_metric, metric = youden
)


cp$roc_curve[[1]] %>%
  ggplot(aes(x = 1 - tnr, y = tpr)) +
  geom_path() +
  coord_equal() +
  geom_abline() +
  theme_bw() +
  xlab("Specificity") +
  ylab("1-Sensitivity")
```



Atsižvelgiant į didesnį nesergančių pacientų kiekį duomenyse (stulp. “Outcome” reikšmė 0) laikyta, kad ROC kreivė gali teigti per daug optimistišką informaciją apie modelio kokybę. Papildomai pavaizduotas modelio Precision-Recall grafikas. Atsižvelgiant į uždavinio specifiką (laikyta, kad neteisingai diagnozuotos neigiamos diagnozės (False Negative) kaina didesnė už neteisingai diagnozuotą teigiamą diabeto diagnozę (False Positive)) modeliui siekta parinkti kitą slenkstinę reikšmę (angl. cutoff value).

```r
roc_auc(y_2, Outcome, pred, event_level = "second")

## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.837

# dėl didelio TN skaičiaus ROC rezultatai gali būti per daug optimistiški, todėl papildomai naudojama P
R kreivė
cutoff <- cp$roc_curve[[1]] %>%
  filter(tpr > 0.9) %>%
  pull(m) %>%
  max()

labels <- filter(cp$roc_curve[[1]], (m %in% c(max(m), cutoff))) %>% round(2)


cp$roc_curve[[1]] %>%
  ggplot(aes(x = tpr, y = tp / (fp + tp))) +
  geom_point(data = labels) +
  geom_text(data = labels, aes(label = x.sorted), nudge_y = 0.05) +
  geom_path() +
  coord_equal() +
  theme_bw() +
  xlab("Recall") +
  ylab("Precision")
```

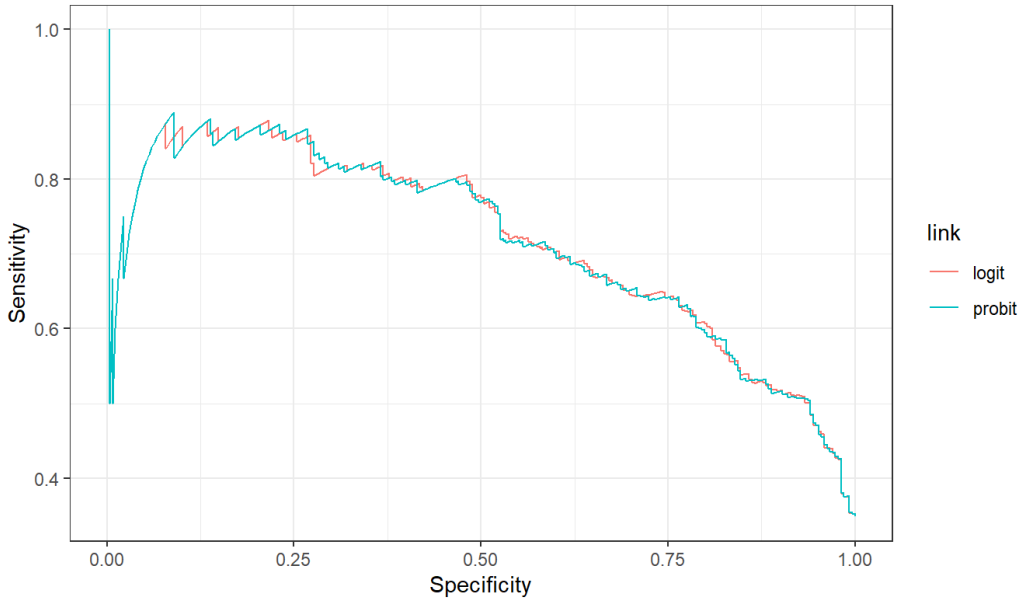



```r
# slenkstinių reikšmių parinkimas
# suskaičiuojamos optimalios slenkstinės reikšmės pagal Youden-J statistic ir pasirinkus ribą Sensitivi
ty > 0.9
  # (t.y. siekiant aptikti bent 90% sergančiųjų)
```

```r
# klasifikavimo lentelė su pasirinkta nauja slenkstine reikšme
confusionMatrix(factor(as.numeric(model_2$fitted.values > 0.21)), factor(y$Outcome),positive="1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 279  27
##          1 199 224
##
##                Accuracy : 0.69
##                  95% CI : (0.655, 0.7234)
##     No Information Rate : 0.6557
##     P-Value [Acc > NIR] : 0.02734
##
##                   Kappa : 0.4095
##
##  Mcnemar's Test P-Value : < 2e-16
##
##             Sensitivity : 0.8924
##             Specificity : 0.5837
##          Pos Pred Value : 0.5296
##          Neg Pred Value : 0.9118
##              Prevalence : 0.3443
##          Detection Rate : 0.3073
##    Detection Prevalence : 0.5802
##       Balanced Accuracy : 0.7381
##
##        'Positive' Class : 1
##
```

```r
# palyginimas su probit modeliu
model_3 <- glm(
  formula = Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction, family = binomial(probit)
,
  data = y
)


y_3 <- y %>% mutate(pred = model_3$fitted.values)


cp_2 <- cutpointr(y_3, pred, Outcome,
  method = maximize_metric, metric = F1_score
)


cp$roc_curve[[1]] %>%
  mutate(link = "logit") %>%
  rbind((cp_2$roc_curve[[1]] %>% mutate(link = "probit"))) %>%
  ggplot(aes(x = tpr, y = tp / (fp + tp), color = link)) +
  geom_path() +
  coord_equal() +
  theme_bw() +
  xlab("Specificity") +
  ylab("Sensitivity")
```




```
# skirtumų tarp modelių beveik nėra
```

Rasta slenkstinė reikšmė pagal Joudeno (Youden) indeksą - 0.32. Naudojant kriterijų, siekiantį teisingai aptikti bent 90% procentų sergančiųjų (Sensitivity > 0.9) - 0.21 (abi reikšmės pažymėtos Precision-Recall grafike). Naudojant PR kreives modelis palygintas su modeliu su tokiomis pačiomis kovariantėmis, tačiau naudojančiu probit junties funkciją. Reikšmingų skirtumų tarp modelių nerasta.

Rezultatai

Naudojant logistinę regresiją siekta rasti kokie požymiai susiję su didžiausia rizika sirgti diabetu, prognozuoti šios ligos diagnozę.

Tyrimo metu parinktas modelis su kovariantėmis „Pregnancies“, „Glucose“ „BMI“ ,“Age“ ir „DiabetesPedigreeFunction“. Modelio tikslumas (angl. accuracy) 0.78. Plotas po modelio ROC kreive = 0.84. Rasta, kad šių metrikų atžvilgiu modelius nesiskiria nuo sudėtingesnio modelio naudojančio visas 8 duomenyse esančias kovariantes.

Atsižvelgiant į užduoties specifiką, pasirinktos kitos modelio slenkstinės reikšmės: siekiant teisingai aptikti bent 90% teigiamų diabeto diagnozių pasirinkta slenkstinė riba 0.21.

Reikšmingų skirtumų tarp modelių su tokiomis pačiomis kovariantėmis, bet naudojančių atitinkamai logit ir probit jungties funkcijas nepastebėta.

2. Naudojant SAS

Atlikta analizė pakartotinai atlikta naudojant SAS.

```
PROC IMPORT DATAFILE='/home/u45871880/diabetes_cleaned.csv'
        DBMS=CSV
        OUT=data;
        GETNAMES=YES;
RUN;


%MACRO boxplot(column);
ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=WORK.DATA;
        vbox &column / category=Outcome;
        yaxis grid;
run;
%MEND;


%boxplot(Pregnancies);

%boxplot(Glucose);

%boxplot(BloodPressure);

%boxplot(SkinThickness)

%boxplot(Insulin);

%boxplot(Age);

%boxplot(DiabetesPedigreeFunction);

%boxplot(BMI);




* Modelis su visomis kovariantėmis;
PROC LOGISTIC DATA=data DESCENDING
        plots(only)=(roc(ID=cutpoint) effect(X=(Pregnancies Glucose BloodPressure SkinThickness
                                                                                Insulin Age
DiabetesPedigreeFunction BMI) CLBAND=YES ALPHA=0.05));
MODEL Outcome = Pregnancies Glucose BloodPressure SkinThickness
Insulin BMI DiabetesPedigreeFunction Age /
RSQUARE CTABLE PPROB=(0.1 TO 0.9 BY 0.1) EXPB LACKFIT scale=none clparm=wald
RUN;
```

Response Profile

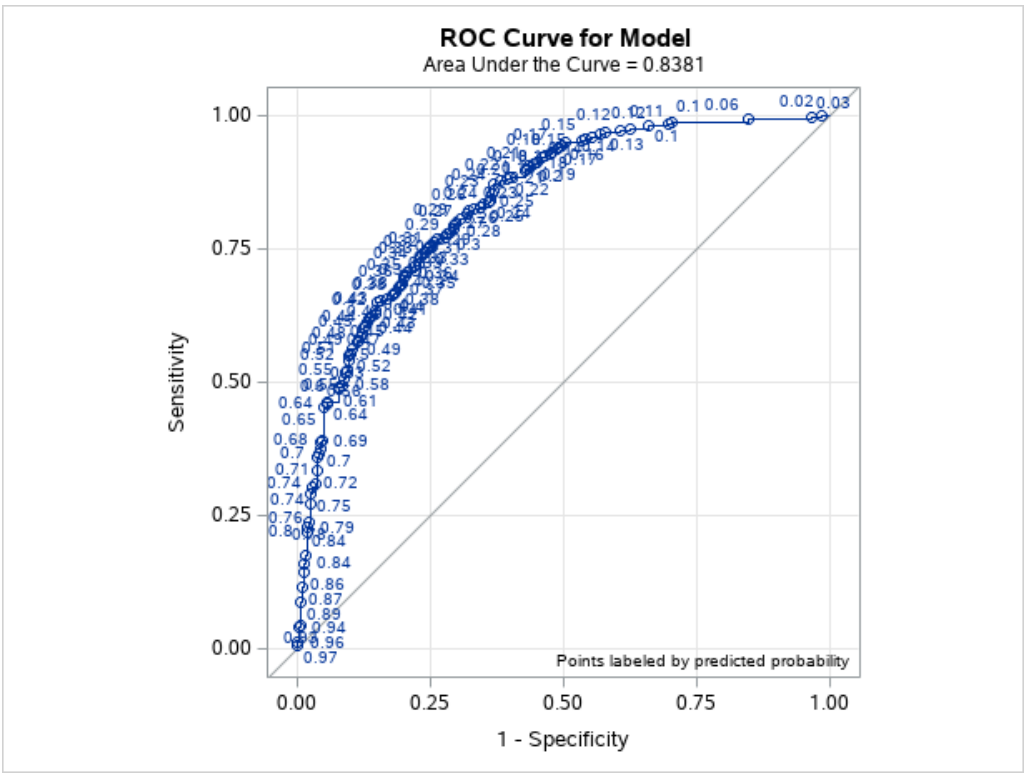
Ordered Value	Outcome	Total Frequency
1	1	251
2	0	478

Probability modeled is Outcome='1'.

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	251.4104	8	<.0001
Score	218.7755	8	<.0001
Wald	158.0839	8	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**	**Exp(Est)**
Intercept	1	-8.8170	0.8118	117.9740	<.0001	0.000
Pregnancies	1	0.1150	0.0331	12.0889	0.0005	1.122
Glucose	1	0.0336	0.00376	80.1674	<.0001	1.034
BloodPressure	1	-0.0100	0.00861	1.3544	0.2445	0.990
SkinThickness	1	0.000706	0.00701	0.0101	0.9198	1.001
Insulin	1	-0.00102	0.000914	1.2492	0.2637	0.999
BMI	1	0.0965	0.0168	33.1764	<.0001	1.101
DiabetesPedigreeFunc	1	0.9995	0.3061	10.6594	0.0011	2.717
Age	1	0.0180	0.00976	3.3909	0.0656	1.018

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
8.3548	8	0.3996



```
* Pažingsninė regresija kovariančių atrinkimui;
PROC LOGISTIC DATA=data DESCENDING plots(only)=(roc);
MODEL Outcome = Pregnancies Glucose BloodPressure SkinThickness
Insulin BMI DiabetesPedigreeFunction Age /
CTABLE PPROB=(0.1 TO 0.9 BY 0.1) EXPB
scale=none clparm=wald outroc=performance SELECTION=stepwise
RUN;
```

Summary of Stepwise Selection

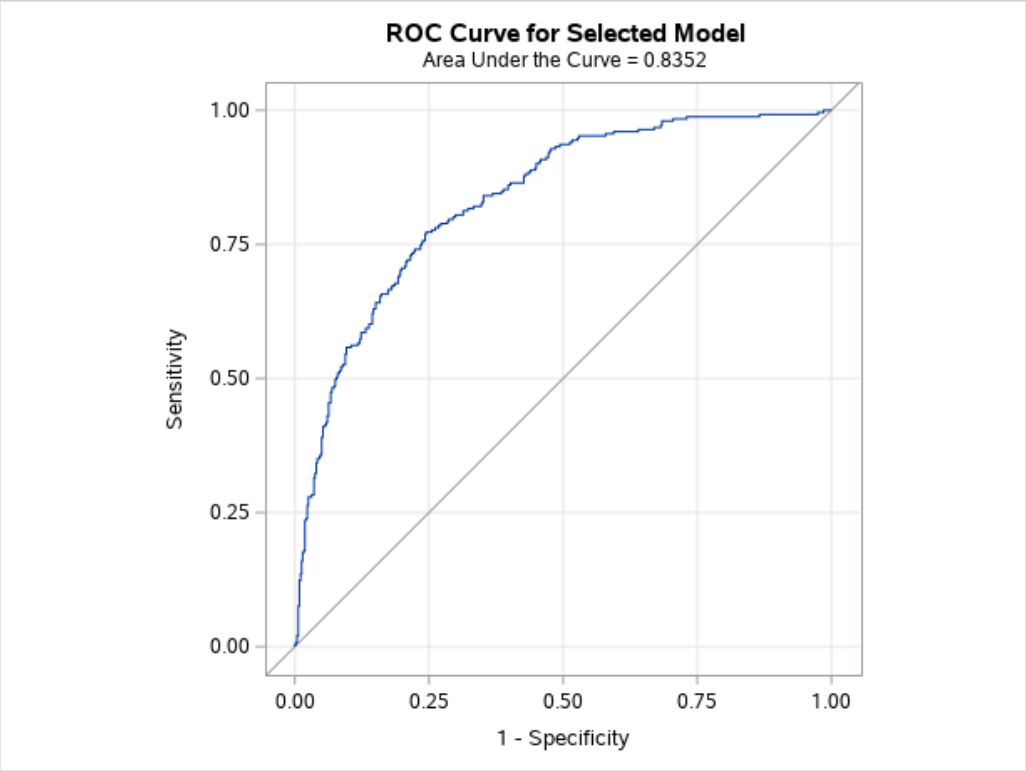
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	Glucose		1	1	153.7441		<.0001
2	BMI		1	2	38.3029		<.0001
3	Pregnancies		1	3	27.0672		<.0001
4	DiabetesPedigreeFunc		1	4	10.6849		0.0011

Analysis of Maximum Likelihood Estimates

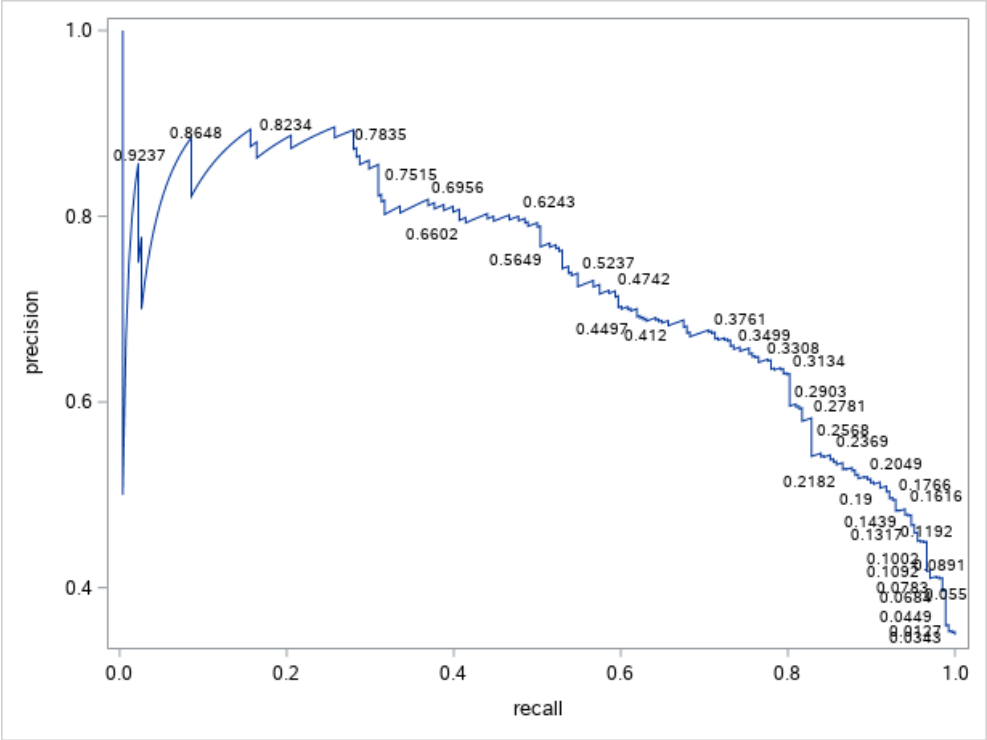
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept	1	-8.7195	0.7005	154.9483	<.0001	0.000
Pregnancies	1	0.1455	0.0280	26.9801	<.0001	1.157
Glucose	1	0.0329	0.00338	94.7882	<.0001	1.033
BMI	1	0.0874	0.0149	34.6227	<.0001	1.091
DiabetesPedigreeFunc	1	0.9726	0.2999	10.5158	0.0012	2.645

Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensi-tivity	Speci-ficity	Pos Pred	Neg Pred
0.100	247	136	342	4	52.5	98.4	28.5	41.9	97.1
0.200	222	263	215	29	66.5	88.4	55.0	50.8	90.1
0.300	196	347	131	55	74.5	78.1	72.6	59.9	86.3
0.400	165	390	88	86	76.1	65.7	81.6	65.2	81.9
0.500	141	422	56	110	77.2	56.2	88.3	71.6	79.3
0.600	119	444	34	132	77.2	47.4	92.9	77.8	77.1
0.700	89	454	24	162	74.5	35.5	95.0	78.8	73.7
0.800	57	468	10	194	72.0	22.7	97.9	85.1	70.7
0.900	18	474	4	233	67.5	7.2	99.2	81.8	67.0




```
* Atsižvelgiai į uždavinio specifiką
* Sukuriamas Precision-Recall grafikas alternativių slenksninių reikšmių parinkimui;
data precision_recall;
set performance;
precision = _POS_/(_POS_ + _FALPOS_);
recall = _POS_/(_POS_ + _FALNEG_);
F_stat = harmean(precision,recall);
if mod(_N_, 20) = 0 then _PROB_=_PROB_;
       else _PROB_ = .;
run;


Proc SQL;
create table precision_recall as
Select *
From precision_recall
having _step_ = max(_step_);
Quit;

proc sort data=precision_recall;
by recall;
run;


ods graphics / reset width=6.4in height=4.8in imagemap;
proc sgplot data=WORK.PRECISION_RECALL;
        SERIES X = recall Y = precision / DATALABEL=_PROB_;
run;
ods graphics / reset;
```



Dėl skirtingos pažingsninėje regresijoje naudojamos procedūros, naudojant SAS į modelį neįtraukta kovariantė „Age“. Kiti gauti rezultatai sutampa.