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.

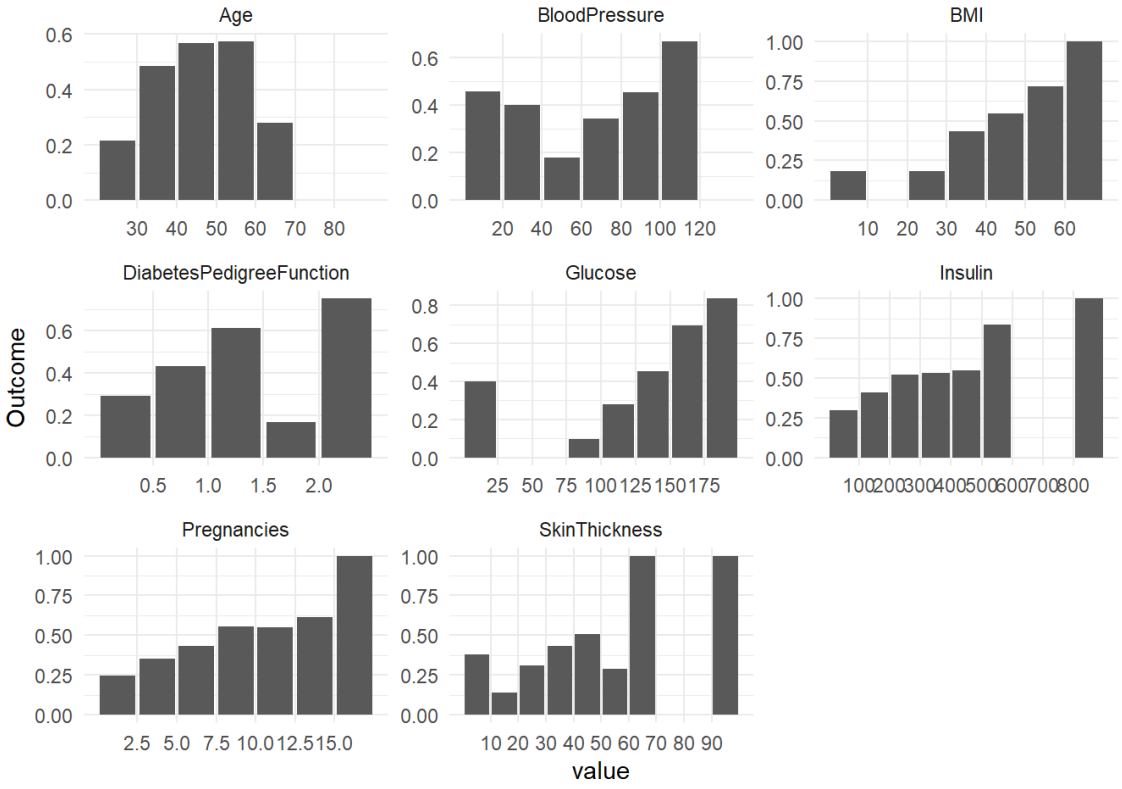
Modelio taikymas prognozėms.

Atliktos analizės aprašymas

1. Naudojant R

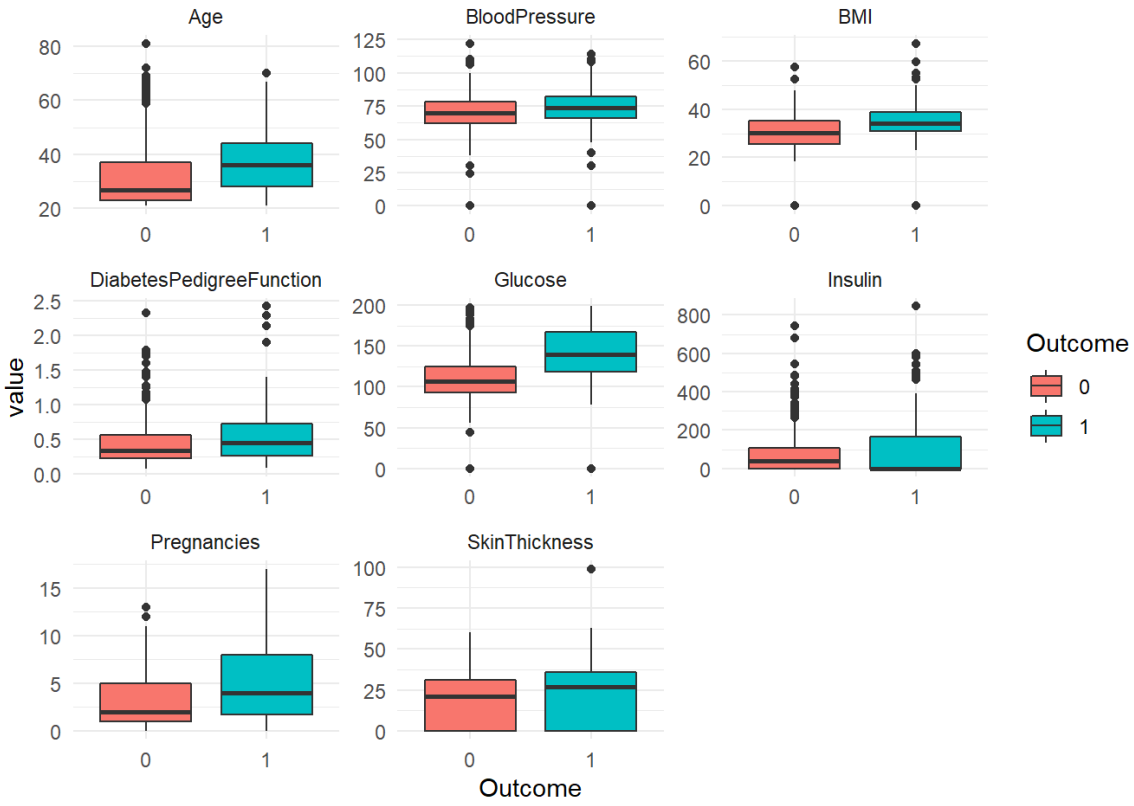
Duomenų aibę sudaro duomenys apie 500 diabetų nesergančių ir 268 sergančių pacientų. Atliekant tiriamąją duomenų analizę palygintas kovariančių pasiskirstymas abiejose grupėse naudojant stačiakampes diagramas, pavaizduotos empirinės sirgimu diabetu tikimybės pagal kiekvieną kovariantę.

```
library(tidyverse)

y <- read_csv("diabetes.csv")

table(y$Outcome)

##
##   0   1
## 500 268

# 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()
```

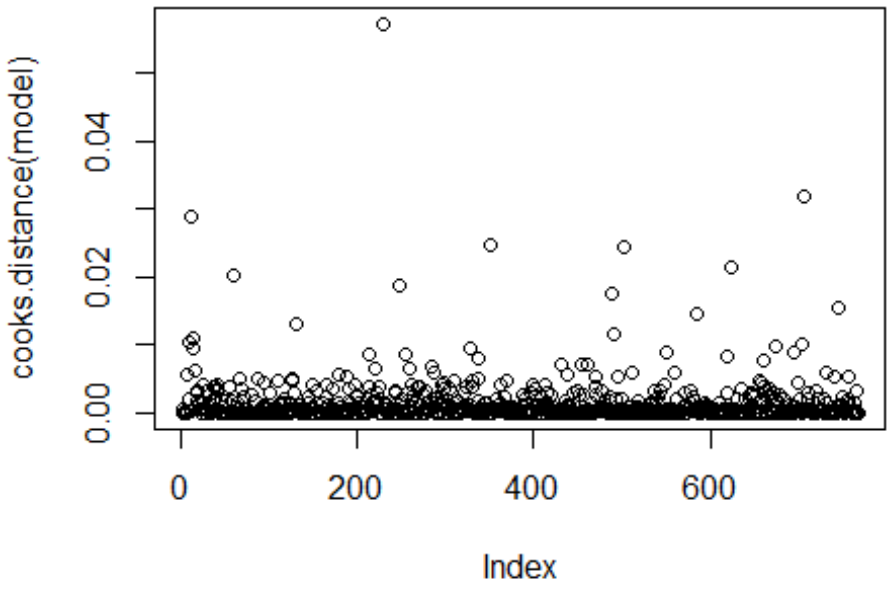



```
# stačiakampės diagramos
y <- y %>% mutate(Outcome = factor(Outcome))
y_plot <- y_plot %>% mutate(Outcome = factor(Outcome))
```

```
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
)

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] 0.8185965

# 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
##          0 445 112
##          1  55 156
##
##                Accuracy : 0.7826
##                  95% CI : (0.7517, 0.8112)
##     No Information Rate : 0.651
##     P-Value [Acc > NIR] : 1.373e-15
##
##                   Kappa : 0.4966
##
##  Mcnemar's Test P-Value : 1.468e-05
##
##             Sensitivity : 0.5821
##             Specificity : 0.8900
##          Pos Pred Value : 0.7393
##          Neg Pred Value : 0.7989
##              Prevalence : 0.3490
##          Detection Rate : 0.2031
##    Detection Prevalence : 0.2747
##       Balanced Accuracy : 0.7360
##
##        '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         0.839
```

```r
# 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
```

```r
# 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
)
```

Pradinio modelio su visomis kovariantėmis (naudojant logit jungties funkciją) tikslumas (angl. accuracy) 78%, plotas po ROC kreive 0.84.
Tikrinant multikolinearumą rasta, kad kovariančių „BloodPressure“ ir „Insulin“ ženklai modelyje priešingi jų įtakai. Šios kovariantės pašalintos iš modelio.

```r
# reikšminų kovariančių atranka
model_2 <- step(model,direction = "both")
```

```
## Start:  AIC=745.52
## Outcome ~ Pregnancies + Glucose + SkinThickness + BMI + DiabetesPedigreeFunction +
##     Age
##
##                            Df Deviance    AIC
## - SkinThickness             1   732.51 744.51
## - Age                       1   733.06 745.06
## <none>                          731.52 745.52
## - DiabetesPedigreeFunction  1   741.75 753.75
## - Pregnancies               1   746.24 758.24
## - BMI                       1   768.98 780.98
## - Glucose                   1   849.35 861.35
##
## Step:  AIC=744.51
## Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction +
##     Age
##
##                            Df Deviance    AIC
## - Age                       1   734.31 744.31
## <none>                          732.51 744.51
## + SkinThickness             1   731.52 745.52
## - DiabetesPedigreeFunction  1   742.10 752.10
## - Pregnancies               1   747.23 757.23
## - BMI                       1   770.61 780.61
## - Glucose                   1   850.23 860.23
##
## Step:  AIC=744.31
## Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction
```

```
##
##                            Df Deviance    AIC
## <none>                          734.31 744.31
## + Age                       1   732.51 744.51
## + SkinThickness             1   733.06 745.06
## - DiabetesPedigreeFunction  1   744.12 752.12
## - Pregnancies               1   762.87 770.87
## - BMI                       1   771.27 779.27
## - Glucose                   1   864.84 872.84
```

```
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
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       761     731.52
## 2       763     734.31 -2  -2.7831   0.2487
```

```
model$aic
```

```
## [1] 745.5228
```

```
model_2$aic
```

```
## [1] 744.3059
```

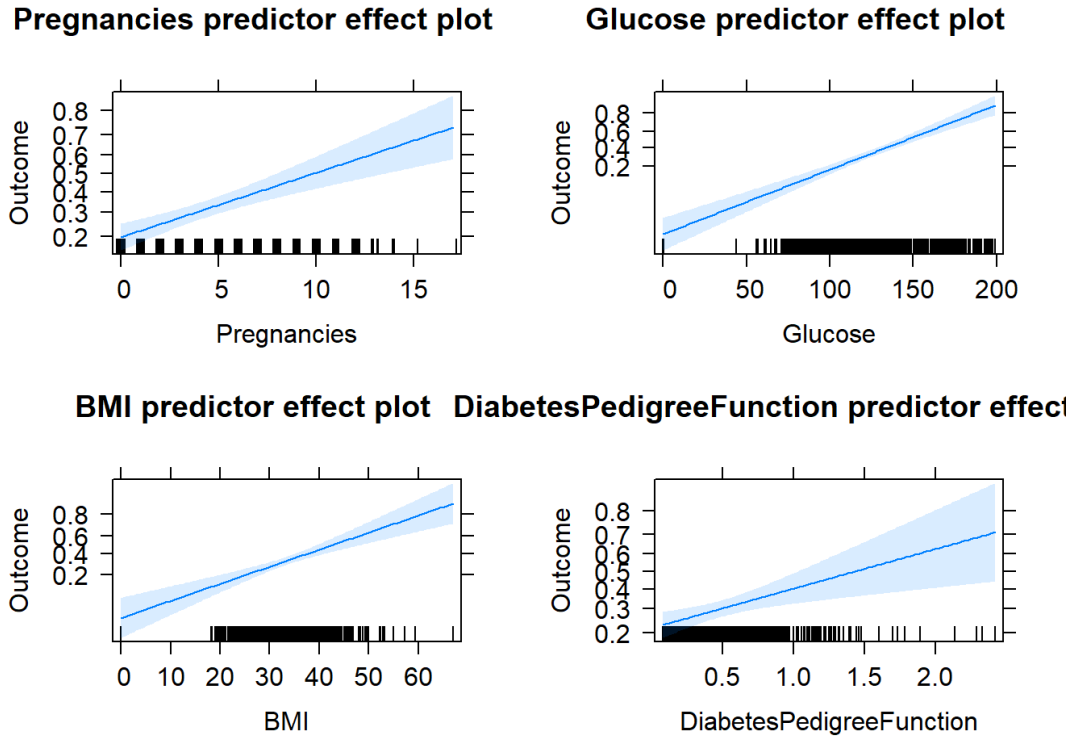
```
confusionMatrix(factor(as.numeric(model_2$fitted.values > 0.5)), factor(y$Outcome),positive="1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 442 117
##          1  58 151
##
##                Accuracy : 0.7721
##                  95% CI : (0.7408, 0.8014)
##     No Information Rate : 0.651
##     P-Value [Acc > NIR] : 2.172e-13
##
##                   Kappa : 0.4715
##
##  Mcnemar's Test P-Value : 1.163e-05
##
##             Sensitivity : 0.5634
##             Specificity : 0.8840
##          Pos Pred Value : 0.7225
##          Neg Pred Value : 0.7907
##              Prevalence : 0.3490
##          Detection Rate : 0.1966
##    Detection Prevalence : 0.2721
##       Balanced Accuracy : 0.7237
##
##        'Positive' Class : 1
##
```

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

```
##             (Intercept)             Pregnancies                 Glucose
##            0.0002213311            1.1524917181            1.0344049733
```

```
##                         BMI DiabetesPedigreeFunction
##              1.0812274690              2.4627867982

exp(confint(model_2))

##                                    2.5 %       97.5 %
## (Intercept)               5.819171e-05 0.0007665084
## Pregnancies               1.093442e+00 1.2162069846
## Glucose                   1.027820e+00 1.0414047033
## BMI                       1.053128e+00 1.1115861320
## DiabetesPedigreeFunction 1.397876e+00 4.3901539569

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

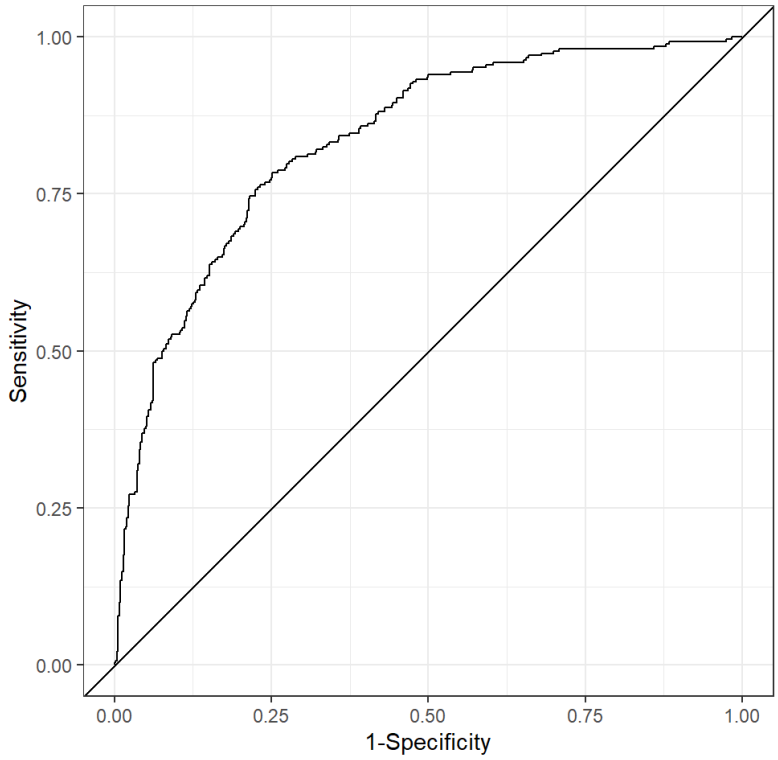


Naudojant pažingsninę regresiją rasta, kad modelis tik su kovariantėmis „Pregnancies“, „Glucose“, „BMI“ ir „DiabetesPedigreeFunction“ statistiškai reikšmingai nesiskiria nuo modelio su visomis kovariantėmis (p=0.25). Modelio tikslumas 77%. Plotas po ROC kreive 0.83.

Modelio koeficientų interpretacija standartinė logit modeliui (pvz. Paciento kūno masės indeksui (angl. BMI) padidėjus 1, kitoms kovariantėms esant fiksuotoms, tikimybė, kad pacientas serga diabetu padidėja 1.08 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]] %>%
```

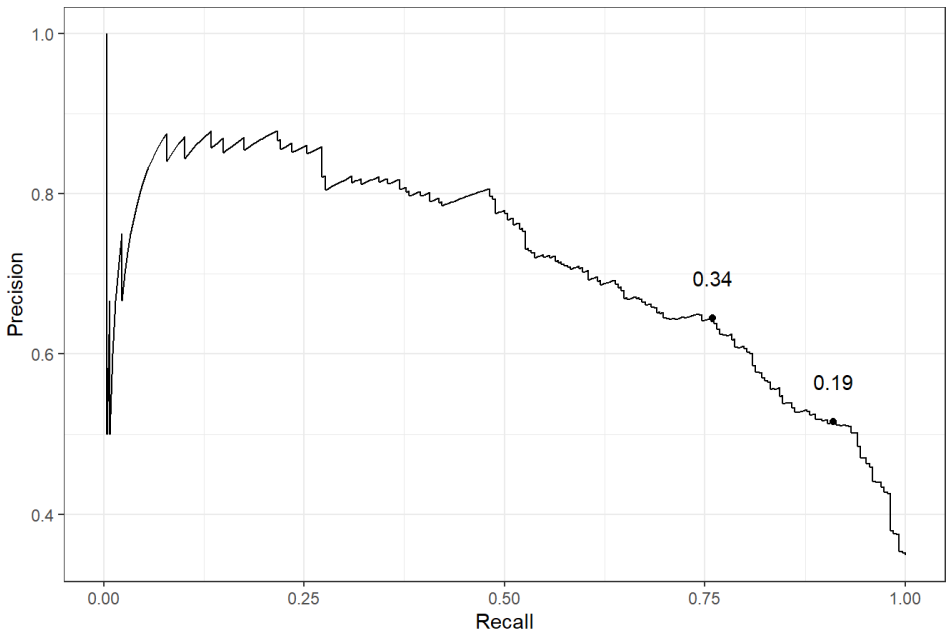
```
  ggplot(aes(x = 1 - tnr, y = tpr)) +
  geom_path() +
  coord_equal() +
  geom_abline() +
  theme_bw() +
  xlab("Specificity") +
  ylab("1-Sensitivity")
```



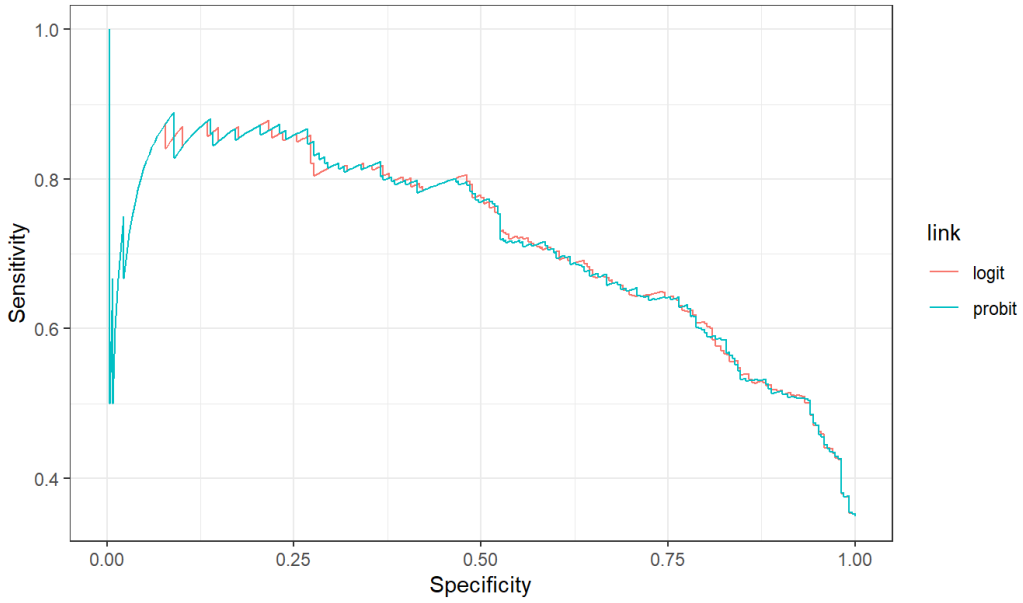
Atsižvelgiant į didesnį nesergančių pacientų kiekį duomenyse (stulp. “Outcome” reikšmė 0) laikyta, kad ROC kreivė gali teigti per daug optimistišką informaciją apie modelio kokybę. Papildomai pavaizduotas modelio Precision-Recall grafikas. Atsižvelgiant į uždavinio specifiką (laikyta, kad neteisingai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

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

## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.834

# 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")
```




```
# 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)

# 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 - 0.34. Naudojant kriterijų, siekiantį teisingai aptikti bent 90% procentų sergančiųjų (Sensitivity > 0.9) - 0.19 (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 rasta, kad modelis su kovariantėmis „Pregnancies“, „Glucose“ „BMI“ ir „DiabetesPedigreeFunction“ statistiškai reikšmingai nesiskiria nuo sudėtingesnio modelio su papildomis kovariantėmis ($p=0.25$). Modelio tikslumas (angl. accuracy) 0.77. Plotas po modelio ROC kreive = 0.83.

Atsižvelgiant į užduoties specifiką, pasirinktos kitos modelio slenkstinės reikšmės: siekiant teisingai aptikti bent 90% teigiamų diabeto diagnozių pasirinkta slenkstinė riba 0.19.

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

2. Naudojant SAS

```
PROC IMPORT DATAFILE='/home/u45871880/diabetes.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	268
2	0	500

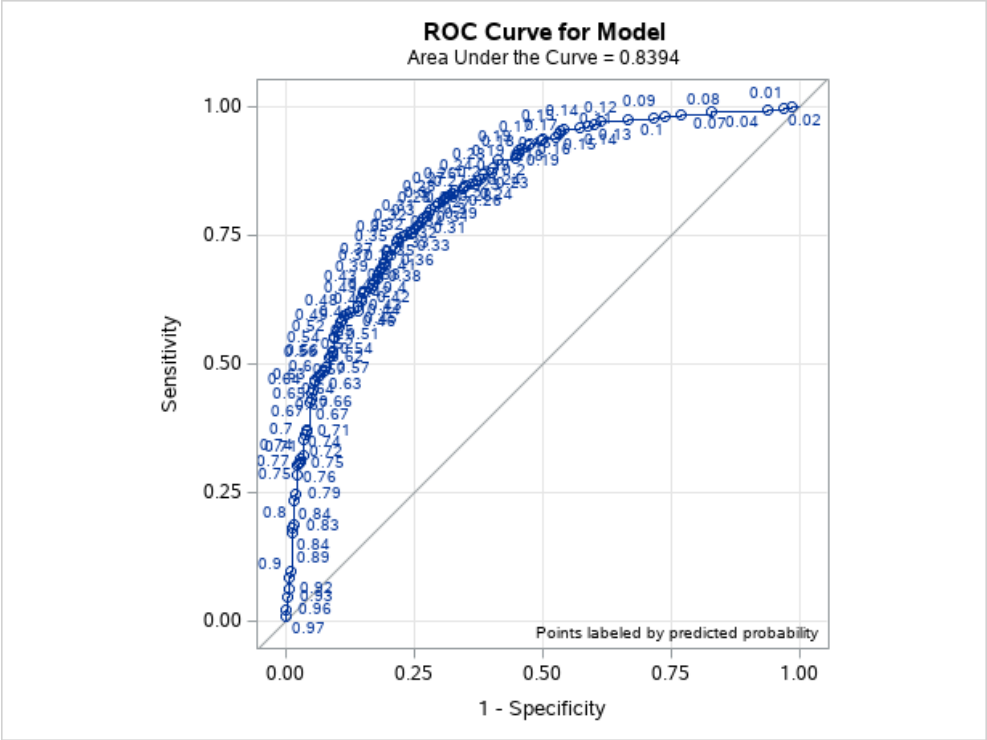
Probability modeled is Outcome='1'.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	270.0385	8	<.0001
Score	232.8984	8	<.0001
Wald	167.7255	8	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept	1	-8.4047	0.7166	137.5452	<.0001	0.000
Pregnancies	1	0.1232	0.0321	14.7466	0.0001	1.131
Glucose	1	0.0352	0.00371	89.8965	<.0001	1.036
BloodPressure	1	-0.0133	0.00523	6.4537	0.0111	0.987
SkinThickness	1	0.000619	0.00690	0.0080	0.9285	1.001
Insulin	1	-0.00119	0.000901	1.7485	0.1861	0.999
BMI	1	0.0897	0.0151	35.3467	<.0001	1.094
DiabetesPedigreeFunc	1	0.9452	0.2991	9.9828	0.0016	2.573
Age	1	0.0149	0.00933	2.5372	0.1112	1.015

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
8.3230	8	0.4026

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	Pos Pred	Neg Pred
0.100	261	146	354	7	53.0	97.4	29.2	42.4	95.4
0.200	240	277	223	28	67.3	89.6	55.4	51.8	90.8
0.300	209	356	144	59	73.6	78.0	71.2	59.2	85.8
0.400	177	408	92	91	76.2	66.0	81.6	65.8	81.8
0.500	154	443	57	114	77.7	57.5	88.6	73.0	79.5
0.600	130	457	43	138	76.4	48.5	91.4	75.1	76.8
0.700	97	475	25	171	74.5	36.2	95.0	79.5	73.5
0.800	62	489	11	206	71.7	23.1	97.8	84.9	70.4
0.900	22	495	5	246	67.3	8.2	99.0	81.5	66.8




```
* 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;
```

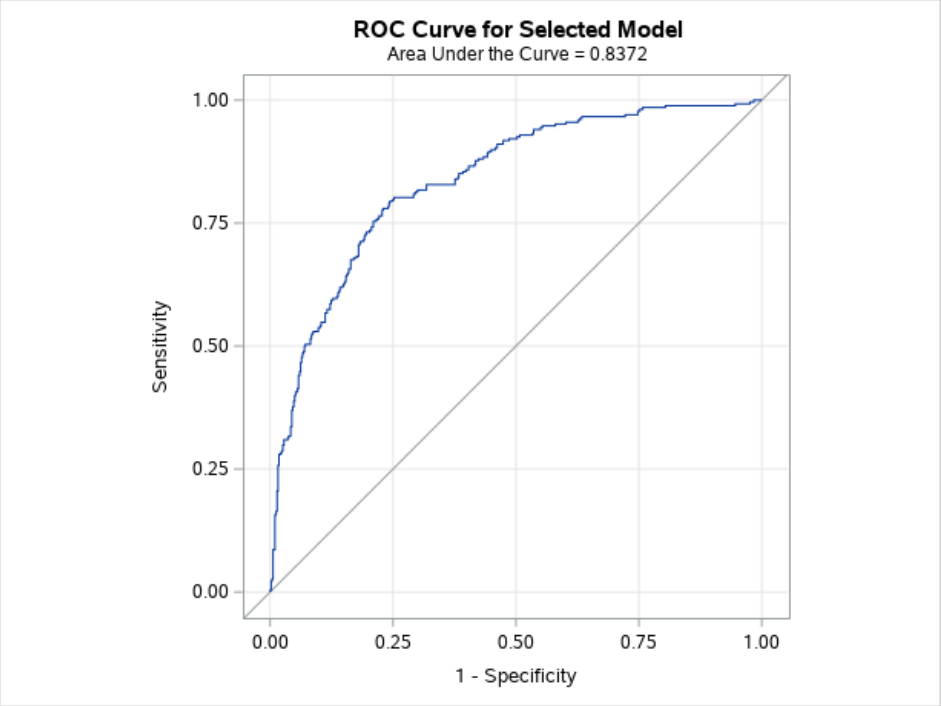
Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	Glucose		1	1	167.1922		<.0001
2	BMI		1	2	34.3033		<.0001
3	Pregnancies		1	3	27.3305		<.0001
4	DiabetesPedigreeFunc		1	4	9.6773		0.0019
5	BloodPressure		1	5	5.8123		0.0159

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept	1	-7.9549	0.6758	138.5505	<.0001	0.000
Pregnancies	1	0.1535	0.0278	30.4074	<.0001	1.166
Glucose	1	0.0347	0.00339	104.3051	<.0001	1.035
BloodPressure	1	-0.0120	0.00503	5.6969	0.0170	0.988
BMI	1	0.0848	0.0141	36.0703	<.0001	1.089
DiabetesPedigreeFunc	1	0.9106	0.2940	9.5919	0.0020	2.486

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	Pos Pred	Neg Pred
0.100	259	138	362	9	51.7	96.6	27.6	41.7	93.9
0.200	239	275	225	29	66.9	89.2	55.0	51.5	90.5
0.300	215	362	138	53	75.1	80.2	72.4	60.9	87.2
0.400	181	412	88	87	77.2	67.5	82.4	67.3	82.6
0.500	152	439	61	116	77.0	56.7	87.8	71.4	79.1
0.600	133	462	38	135	77.5	49.6	92.4	77.8	77.4
0.700	96	477	23	172	74.6	35.8	95.4	80.7	73.5
0.800	63	492	8	205	72.3	23.5	98.4	88.7	70.6
0.900	23	496	4	245	67.6	8.6	99.2	85.2	66.9




```
* 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;
```