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.

# Atliktos analizės aprašymas

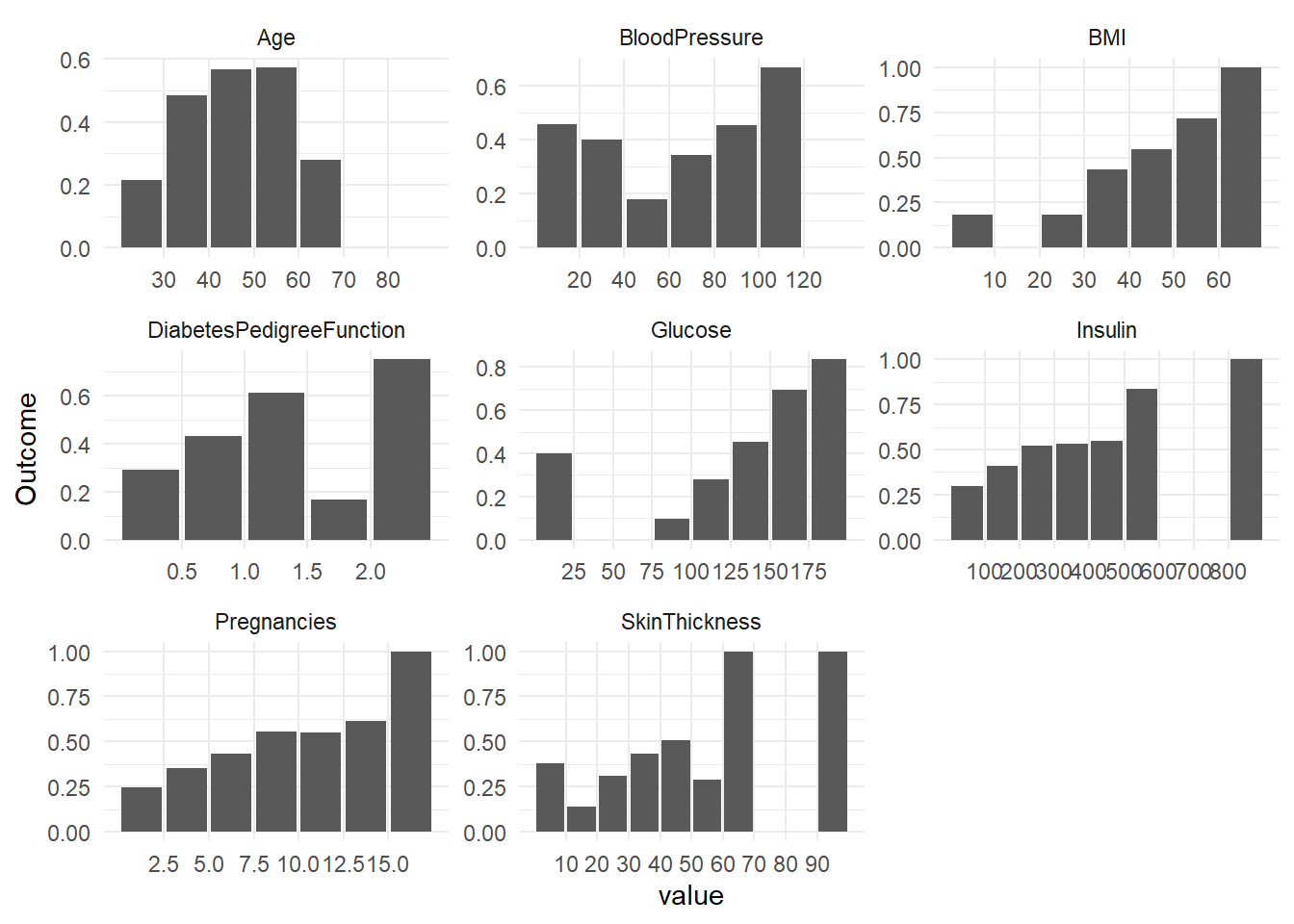
**1. Naudojant R**

Duomenų aibę sudaro duomenys apie 500 diabetų nesergančių ir 268 sergančių pacientų. Pašalinus praleistas reikšmes duomenų 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.

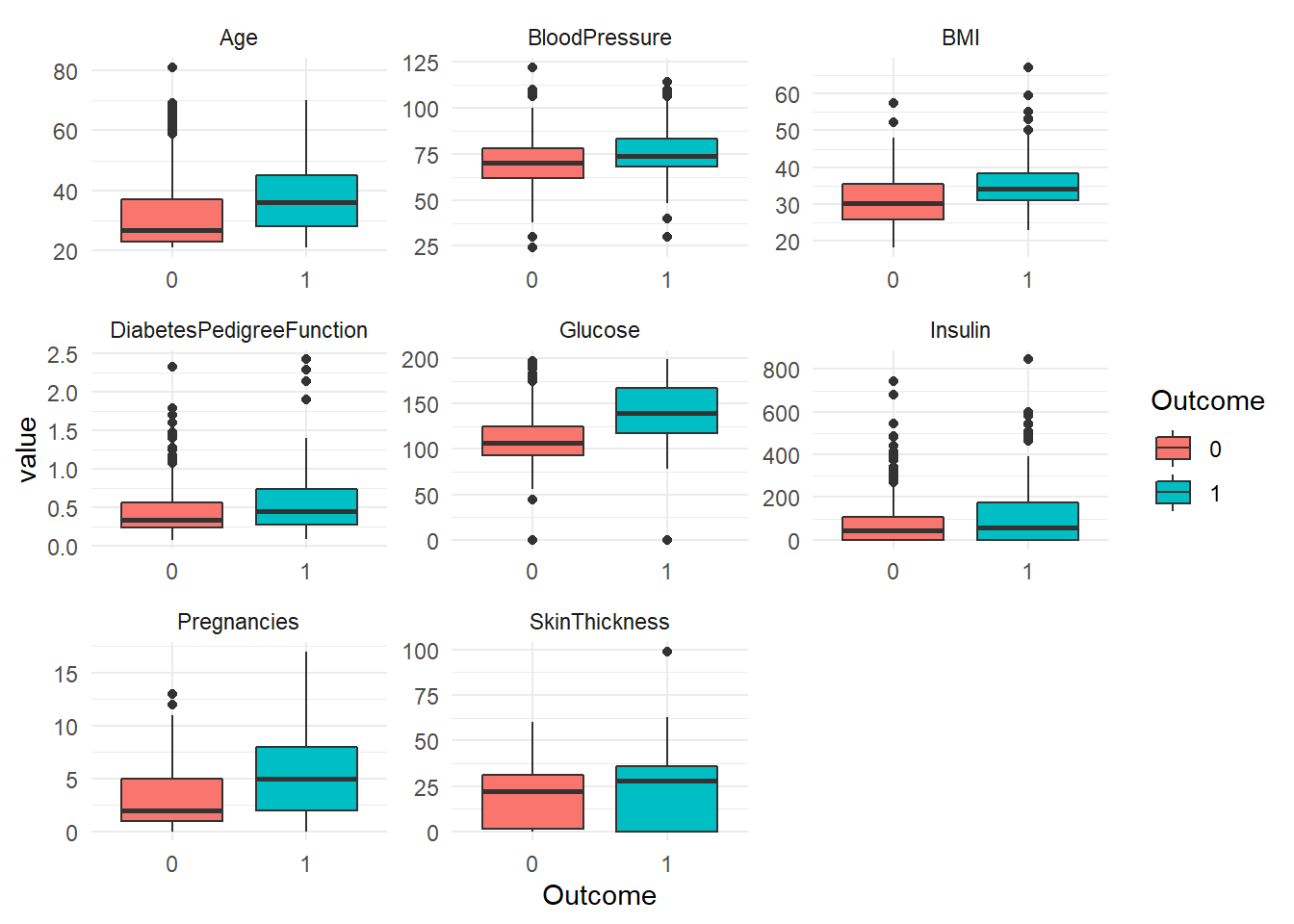
library(tidyverse)  
  
y <- read\_csv("diabetes.csv")  
y <- y %>% filter(BloodPressure != 0,BMI != 0)  
  
table(y$Outcome)

##   
## 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

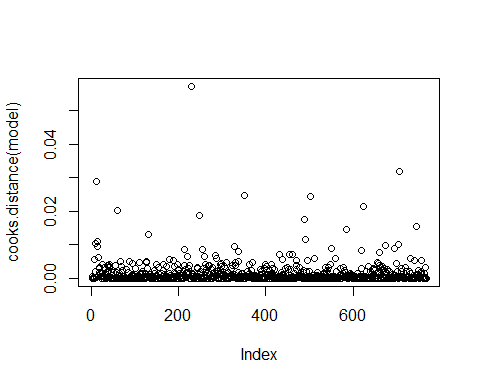
1-pchisq(model$null.deviance-model$deviance, model$df.null-model$df.residual) # globali nulinė hipotezė (tikėtinumo santykių testas likelihood ratio test)

## [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

Sudarytas logit modelis, naudojantis visas duomenyse esančias kovariantes. 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).

# reikšmingų kovariančių atranka  
model\_2 <- MASS::stepAIC(model,direction = "both")

## Start: AIC=705.32  
## Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThickness +   
## Insulin + BMI + DiabetesPedigreeFunction + Age  
##   
## Df Deviance AIC  
## - SkinThickness 1 687.33 703.33  
## - Insulin 1 688.57 704.57  
## - BloodPressure 1 688.68 704.68  
## <none> 687.32 705.32  
## - Age 1 690.71 706.71  
## - DiabetesPedigreeFunction 1 698.41 714.41  
## - Pregnancies 1 699.74 715.74  
## - BMI 1 723.25 739.25  
## - Glucose 1 788.23 804.23  
##   
## Step: AIC=703.33  
## Outcome ~ Pregnancies + Glucose + BloodPressure + Insulin + BMI +   
## DiabetesPedigreeFunction + Age  
##   
## Df Deviance AIC  
## - BloodPressure 1 688.69 702.69  
## - Insulin 1 688.71 702.71  
## <none> 687.33 703.33  
## - Age 1 690.71 704.71  
## + SkinThickness 1 687.32 705.32  
## - DiabetesPedigreeFunction 1 698.55 712.55  
## - Pregnancies 1 699.76 713.76  
## - BMI 1 728.24 742.24  
## - Glucose 1 789.65 803.65  
##   
## Step: AIC=702.69  
## Outcome ~ Pregnancies + Glucose + Insulin + BMI + DiabetesPedigreeFunction +   
## Age  
##   
## Df Deviance AIC  
## - Insulin 1 689.73 701.73  
## <none> 688.69 702.69  
## - Age 1 691.30 703.30  
## + BloodPressure 1 687.33 703.33  
## + SkinThickness 1 688.68 704.68  
## - DiabetesPedigreeFunction 1 700.08 712.08  
## - Pregnancies 1 700.82 712.82  
## - BMI 1 729.18 741.18  
## - Glucose 1 789.73 801.73  
##   
## Step: AIC=701.73  
## Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction +   
## Age  
##   
## Df Deviance AIC  
## <none> 689.73 701.73  
## + Insulin 1 688.69 702.69  
## + BloodPressure 1 688.71 702.71  
## - 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 + BloodPressure + SkinThickness +   
## Insulin + BMI + DiabetesPedigreeFunction + Age  
## Model 2: Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction +   
## Age  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 720 687.32   
## 2 723 689.73 -3 -2.4103 0.4917

model$aic

## [1] 705.3247

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   
##

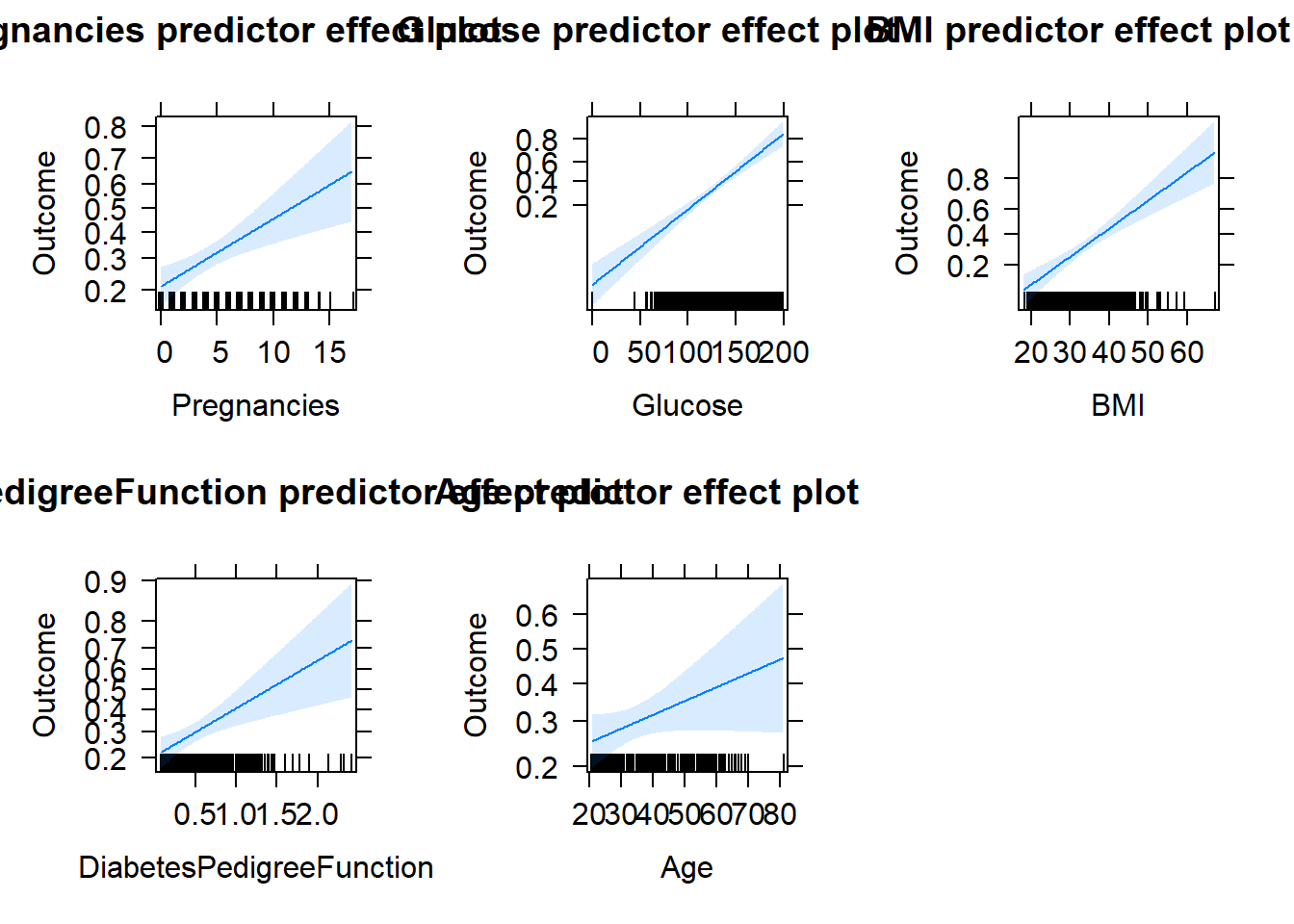
# koeficientų interpretacija  
exp(coef(model\_2))

## (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

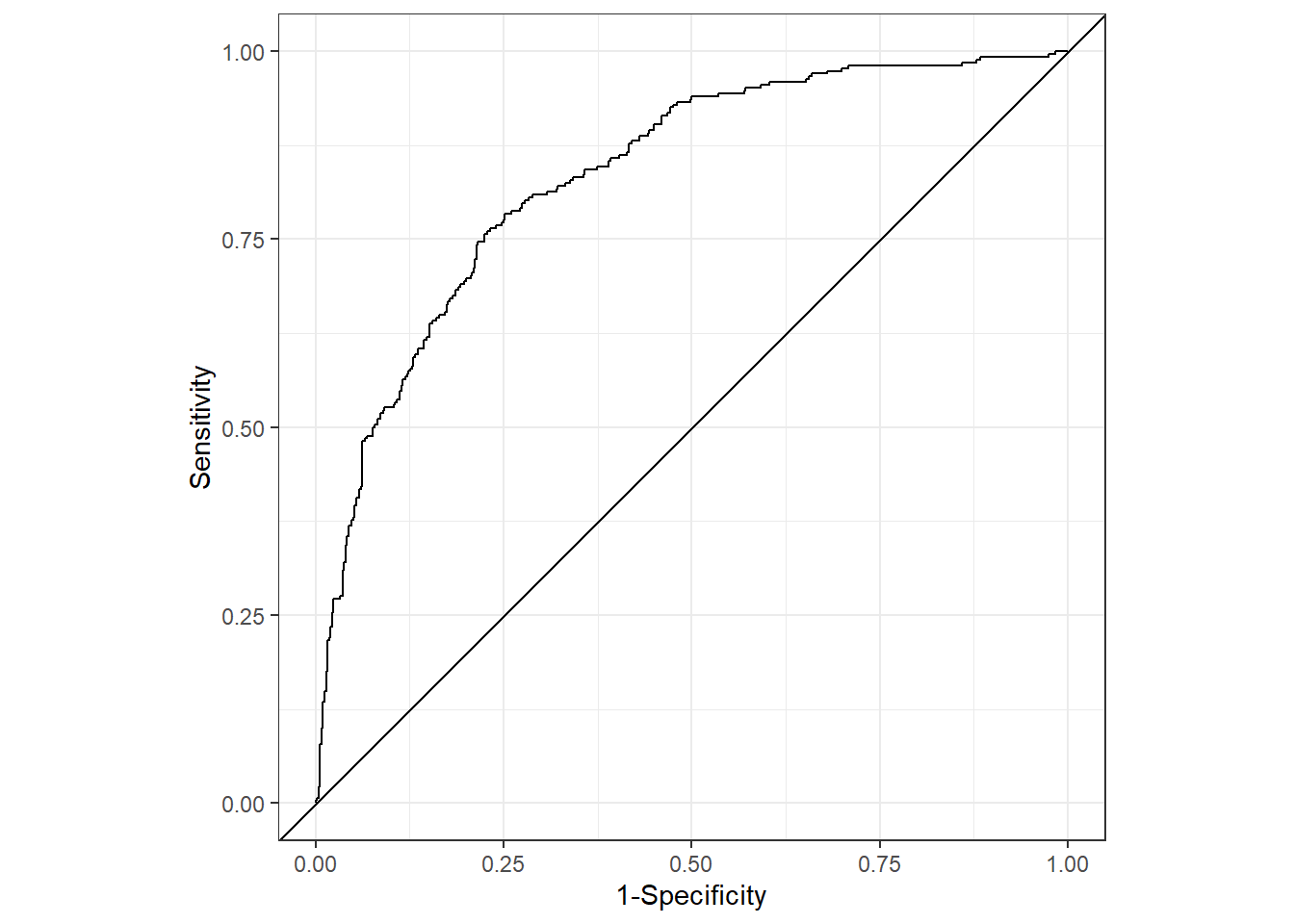
# modelio kovariačių efektai  
library(effects)  
plot(predictorEffects(model\_2))



Naudojant pažingsninę regresiją parinktas modelis be kovariančių „Insulin“, „BloodPressure“ ir „SkinThickness“. 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")

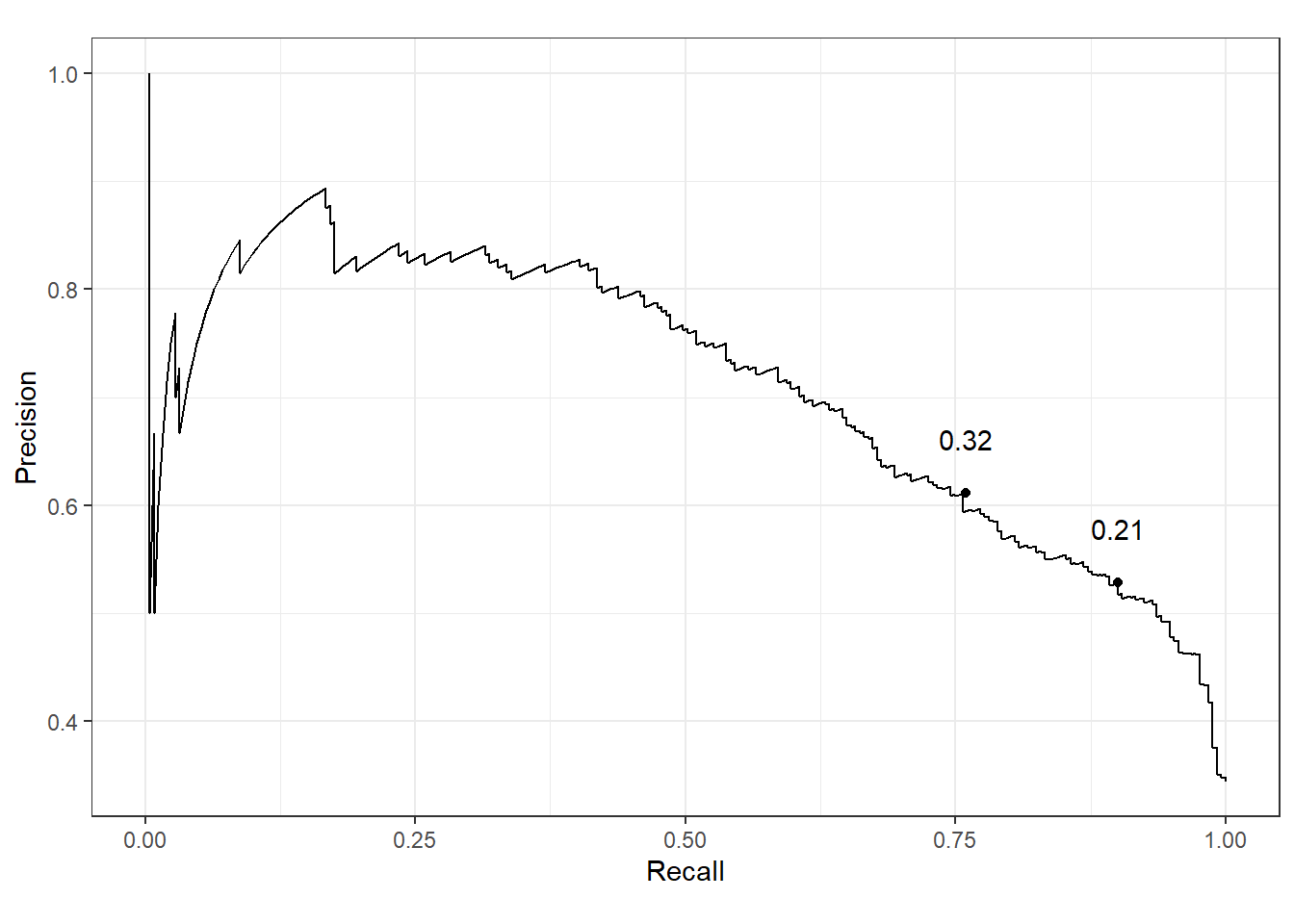


Atsižvelgiant į didesnį nesergančių pacientų kiekį duomenyse (stulp. “Outcome” reikšmė 0) laikyta, kad ROC kreivė gali teigti per daug optimistišką informaciją apie modelio kokybę. Papildomai pavaizduotas modelio Precision-Recall grafikas. Atsižvelgiant į uždavinio specifiką (laikyta, kad neteisingai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

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 PR 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 Sensitivity > 0.9  
 # (t.y. siekiant aptikti bent 90% sergančiųjų)

# 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   
##

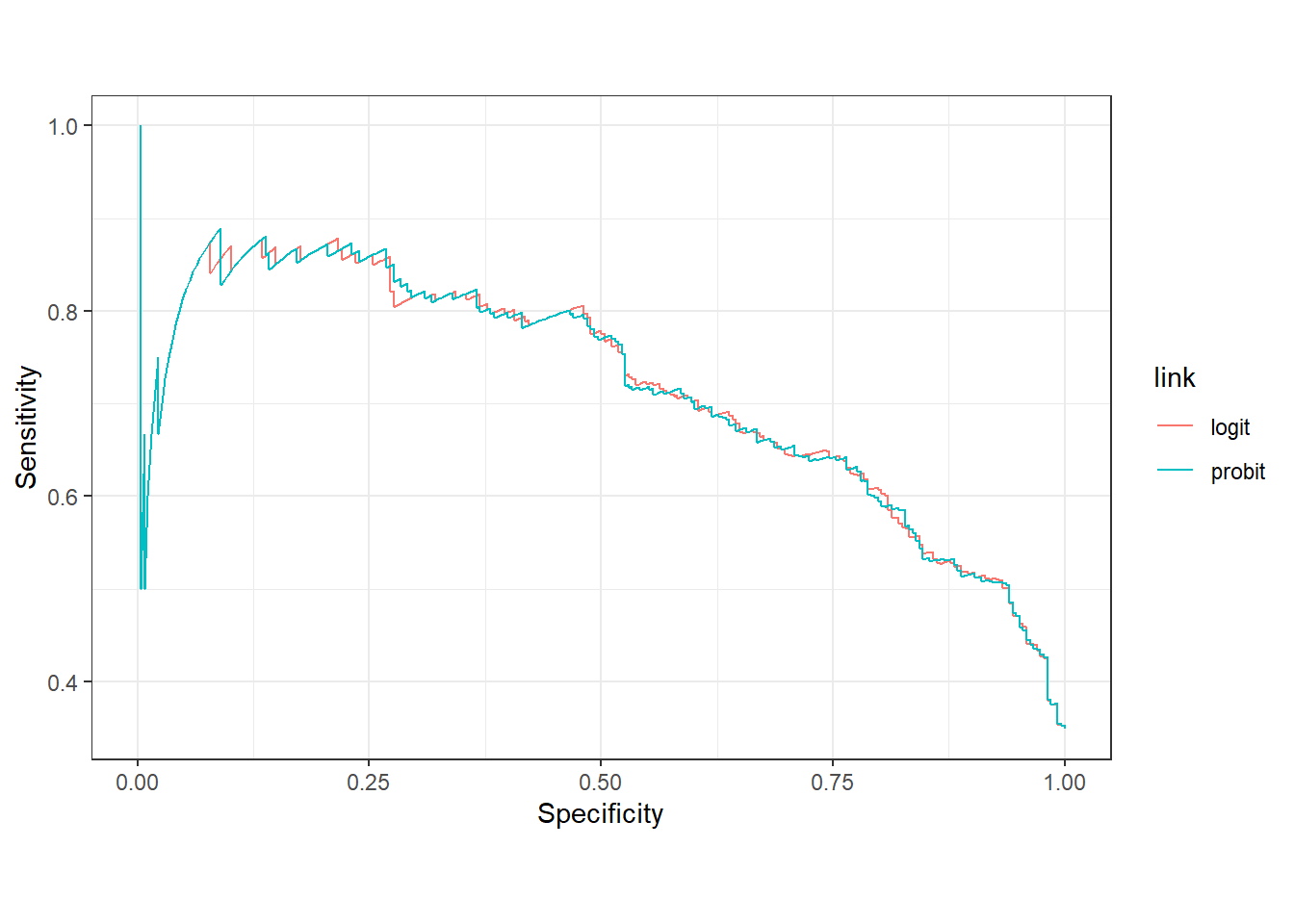
# palyginimas su probit modeliu  
model\_3 <- glm(  
 formula = Outcome ~ ., family = binomial(probit),  
 data = y  
)  
  
  
model\_3 <- MASS::stepAIC(model\_3,direction = "both")

## Start: AIC=706.29  
## Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThickness +   
## Insulin + BMI + DiabetesPedigreeFunction + Age  
##   
## Df Deviance AIC  
## - SkinThickness 1 688.38 704.38  
## - BloodPressure 1 689.66 705.66  
## - Insulin 1 689.81 705.81  
## <none> 688.29 706.29  
## - Age 1 692.59 708.59  
## - DiabetesPedigreeFunction 1 698.26 714.26  
## - Pregnancies 1 701.34 717.34  
## - BMI 1 726.64 742.64  
## - Glucose 1 786.31 802.31  
##   
## Step: AIC=704.38  
## Outcome ~ Pregnancies + Glucose + BloodPressure + Insulin + BMI +   
## DiabetesPedigreeFunction + Age  
##   
## Df Deviance AIC  
## - BloodPressure 1 689.75 703.75  
## - Insulin 1 689.85 703.85  
## <none> 688.38 704.38  
## + SkinThickness 1 688.29 706.29  
## - Age 1 692.61 706.61  
## - DiabetesPedigreeFunction 1 698.51 712.51  
## - Pregnancies 1 701.46 715.46  
## - BMI 1 732.97 746.97  
## - Glucose 1 787.59 801.59  
##   
## Step: AIC=703.75  
## Outcome ~ Pregnancies + Glucose + Insulin + BMI + DiabetesPedigreeFunction +   
## Age  
##   
## Df Deviance AIC  
## - Insulin 1 690.89 702.89  
## <none> 689.75 703.75  
## + BloodPressure 1 688.38 704.38  
## - Age 1 693.09 705.09  
## + SkinThickness 1 689.66 705.66  
## - DiabetesPedigreeFunction 1 699.97 711.97  
## - Pregnancies 1 702.51 714.51  
## - BMI 1 733.93 745.93  
## - Glucose 1 787.63 799.63  
##   
## Step: AIC=702.89  
## Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction +   
## Age  
##   
## Df Deviance AIC  
## <none> 690.89 702.89  
## + Insulin 1 689.75 703.75  
## + BloodPressure 1 689.85 703.85  
## - Age 1 694.73 704.73  
## + SkinThickness 1 690.87 704.87  
## - DiabetesPedigreeFunction 1 700.31 710.31  
## - Pregnancies 1 704.11 714.11  
## - BMI 1 733.98 743.98  
## - Glucose 1 793.58 803.58

names(model\_3$coefficients) == names(model\_2$coefficients)

## [1] TRUE TRUE TRUE TRUE TRUE TRUE

# atrinktos tos pačios kovariantės  
  
  
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("Recall") +  
 ylab("Precision")



# 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). Papildomai sudarytas modelis naudojantis probit jungties funkciją vietoje logit. Naudojant pažingsninę regresiją atrinktos tokios pačios kovariantės, kaip ir gautos logit modelyje. Modeliai tarpusavyje palyginti naudojant PR kreives. Reikšmingų skirtumų nerasta.

Rezultatai

Sudarytas logistinės regresijos binarinio atsako modelis diabeto diagnozei prognozuoti.

Tyrimo metu pasirinktas 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.

Papildomai sudarytas modelis naudojantis probit jungties funkciją, tačiau reikšmingų skirtumų tarp šio modelio ir modelio naudojančio logit jungties funkciją 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

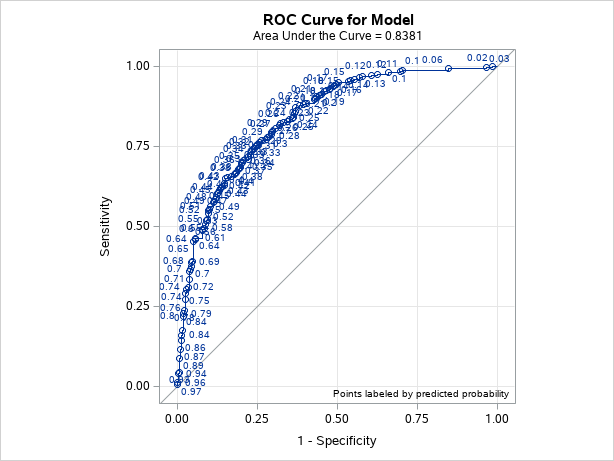
RUN;

| **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

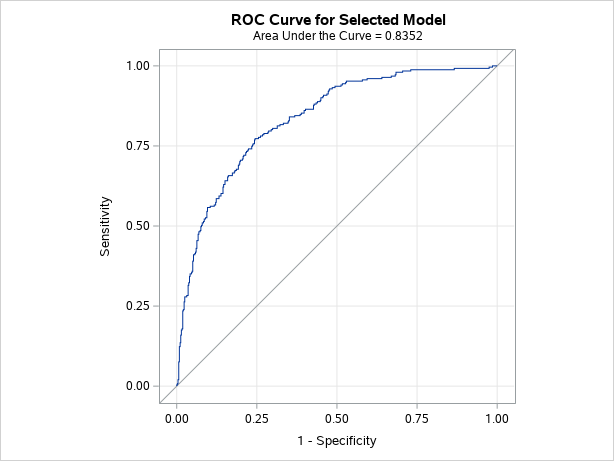
scale=none clparm=wald outroc=performance SELECTION=stepwise

RUN;

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Summary of Stepwise Selection** | | | | | | | | | --- | --- | --- | --- | --- | --- | --- | --- | | **Step** | **Effect** | | **DF** | **Number In** | **Score Chi-Square** | **Wald Chi-Square** | **Pr > ChiSq** | | **Entered** | **Removed** | | **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** | | **Incorrect** | | **Percentages** | | | | | | **Event** | **Non- Event** | **Event** | **Non- Event** | **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;

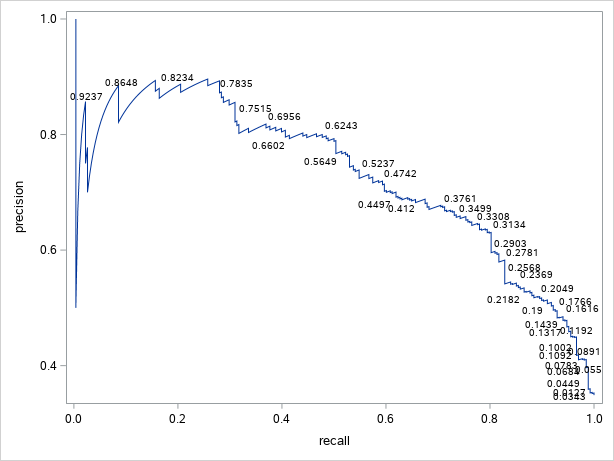
ods graphics / reset width=6.4in height=4.8in imagemap;

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 su analize atlikta naudojant R.