Wyjaśnialne Uczenie Maszynowe - Praca domowa 1

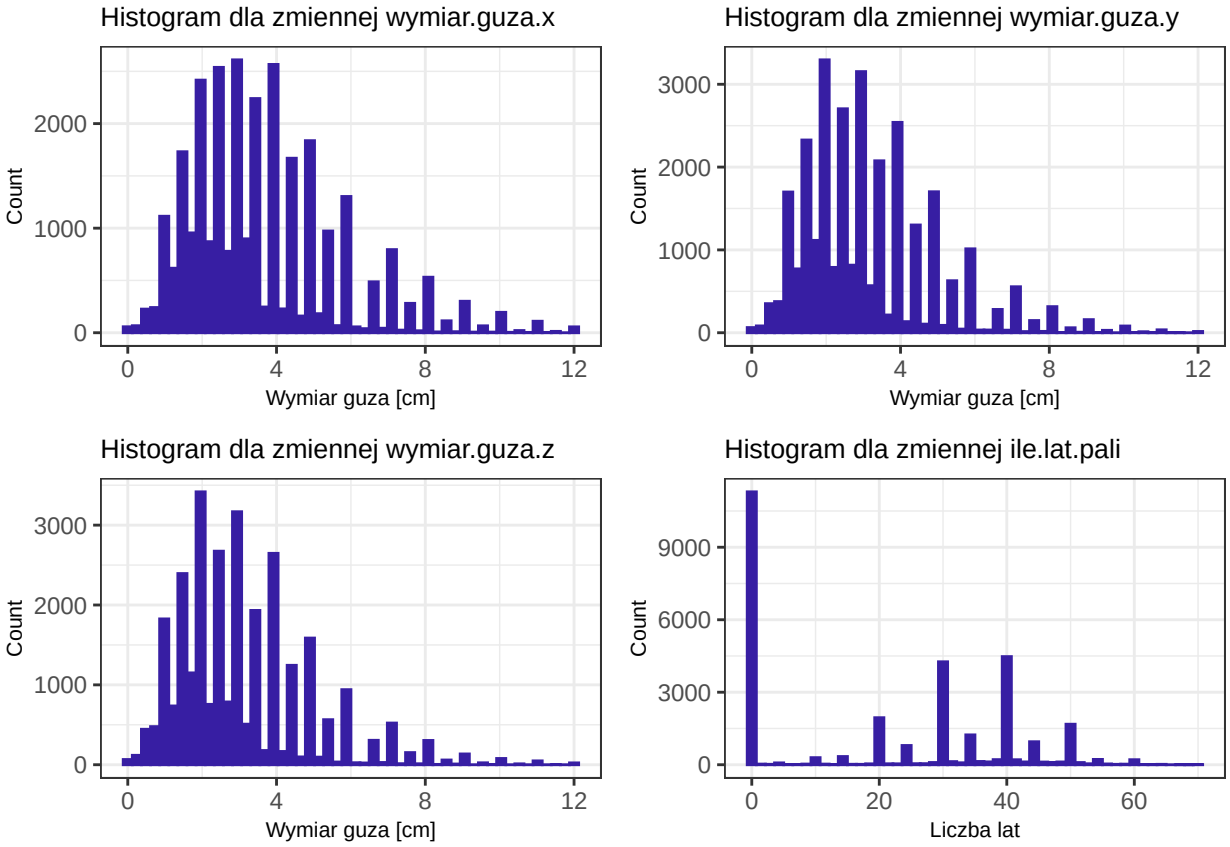
Anna Kozak

Zadanie

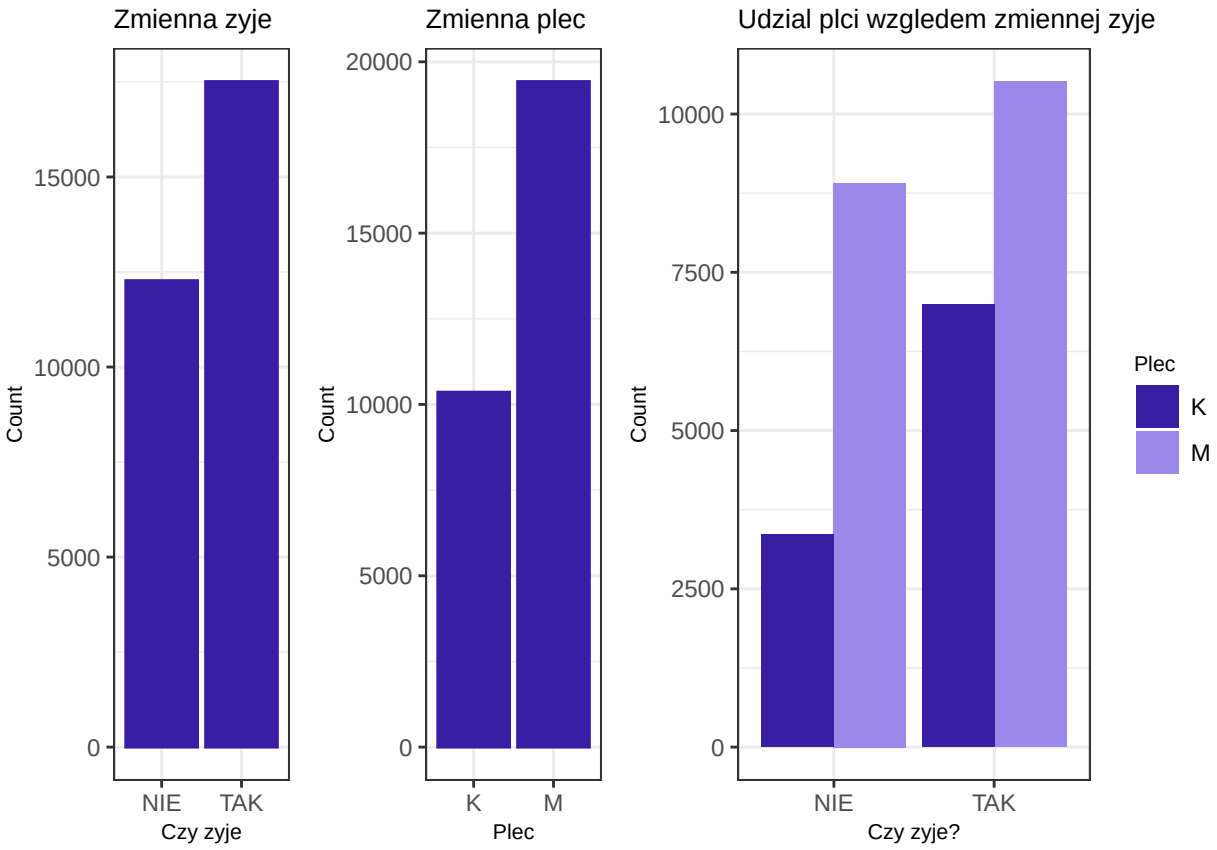
Zbudować model predykcyjny dla wybranego problemu.

Wybrany zbiór danych oraz kilka słów o nim

Do wykonania pracy domowej wybrano zbiór danych odpowiadający szansom przeżycia po operacji nowotworu płuca w polskiej populacji. Jest to zbiór danych z 30 tyś. obserwacji oraz 14 zmiennymi. Dane oczyszczono z braków danych oraz z wartości wyraźnie odstających. Do modelowania wykorzystano ostatecznie 12 zmiennych, m.in. informację o płci, wymiarze guza, stadium choroby oraz wieku pacjenta w roku rozpoznania i czasie oczekiwania na operację (w latach).



Zmienna objaśniania (*zyje*) przyjmuje dwie katergie: TAK i NIE. Poniżej wykresy ukazujące podział zmiennych *zyje* oraz *plec*. Klasy dzielą się w stosunku 0.41:0.59.



Model predykcyjny

Na podstawie tak przygotowanych danych zbudowano dwa modele:

- regresję logistyczną (glm)
- las losowy (random forest).

Jeden z modeli to regresja, czyli model interpretowalny, natomiast drugi z nich to las losowy czyli przykład modelu "czarnej szkrzynki" (*black box*).

Podział zbioru na treningowy oraz testowy w proporcji 2:1.

```
inds <- partition(data$zyje, c(train = 0.65, test = 0.35), seed = 21)
train <- data[inds$train, ]
test <- data[inds$test, ]
```

Model *glm*

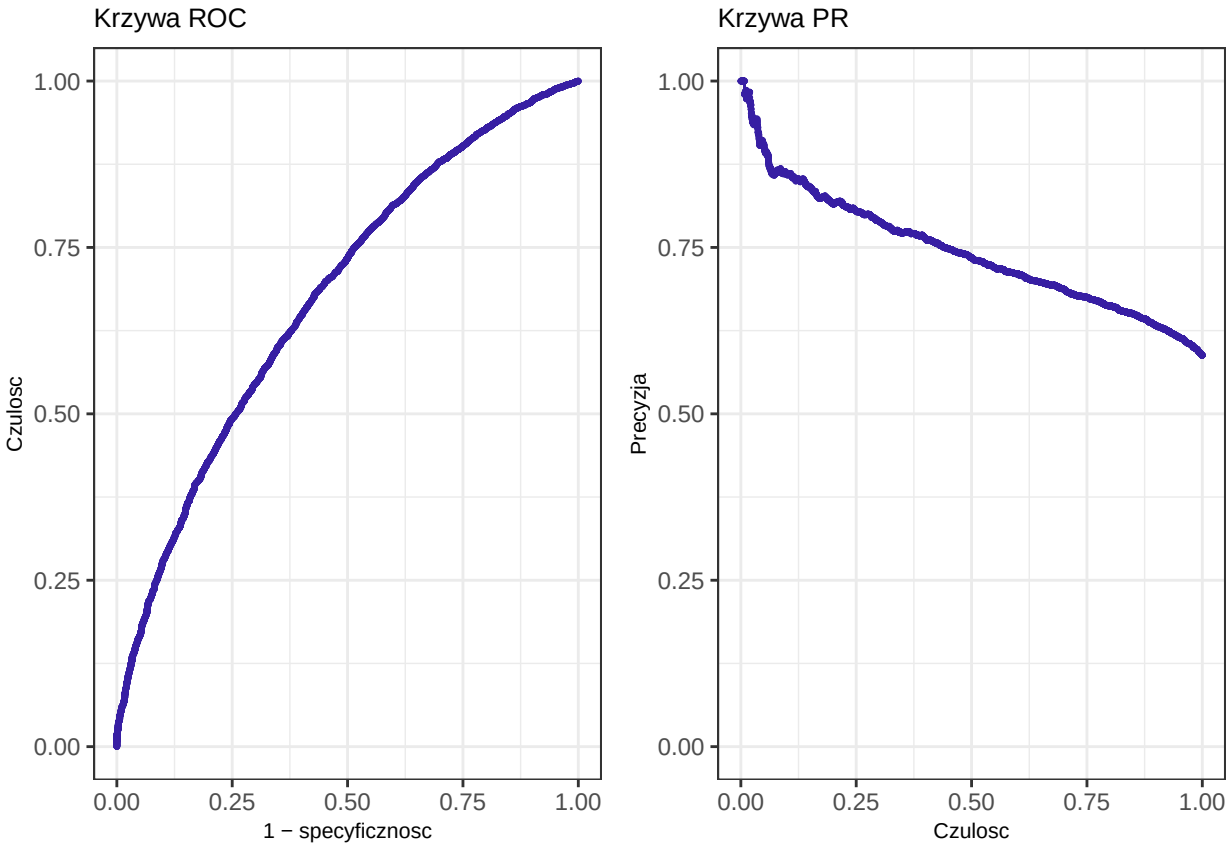
Budowa modelu regresji logistycznej. Do oceny predykcji wykorzystano miarę AUC (Area Under ROC Curve) oraz AURPC (Area under the PR curve)

```r
#model
model <- glm(zyje == "TAK" ~., data = train)
#predykcja
pred_model_test <- predict(model, newdata = test, type = "response")

score_0 <- pred_model_test[test$zyje == "TAK"]
score_1 <- pred_model_test[test$zyje == "NIE"]

# ROC Curve
roc <- roc.curve(scores.class0 = score_0, scores.class1 = score_1, curve = TRUE)
roc$auc
```

```
## [1] 0.6762198
```

```r
# PR Curve
pr <- pr.curve(scores.class0 = score_0, scores.class1 = score_1, curve = TRUE)
pr$auc.integral
```

```
## [1] 0.7434212
```



Model *randomForest*

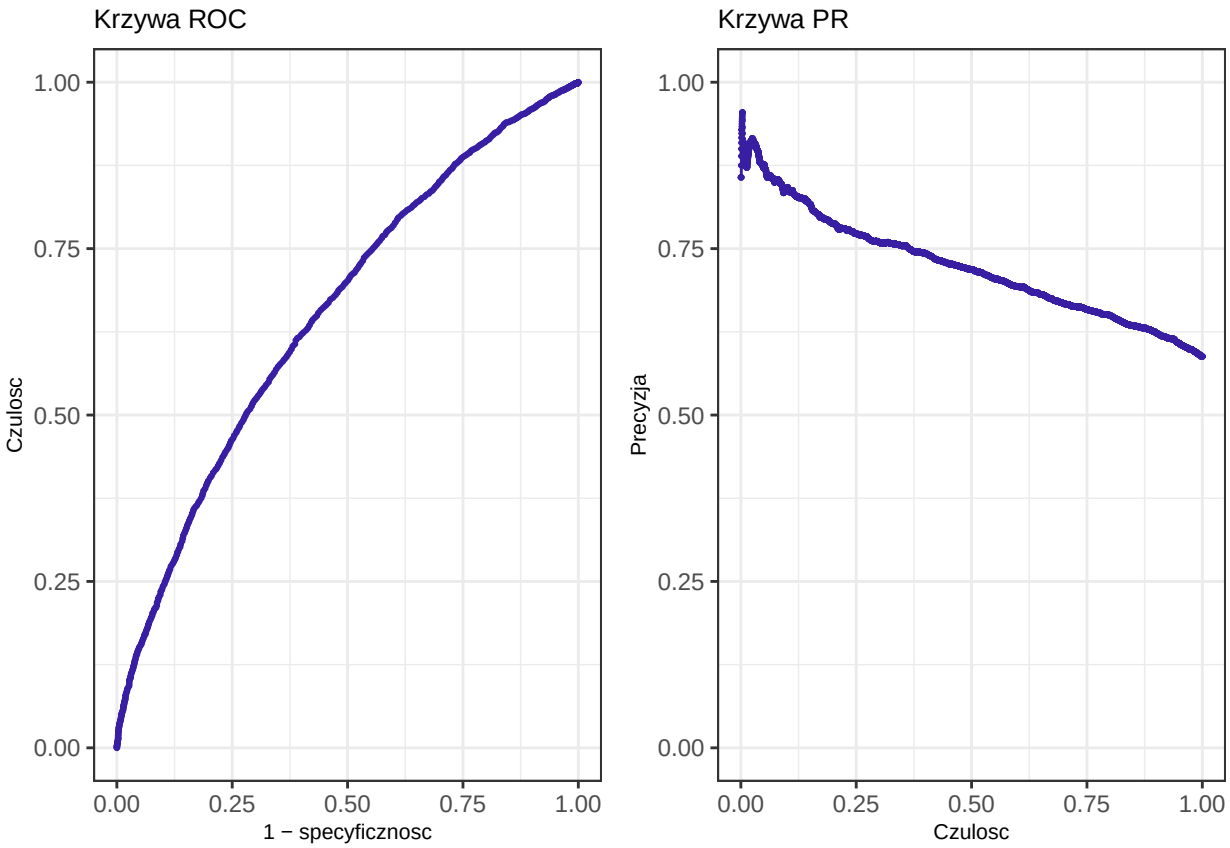
Budowa modelu lasu losowego z pakietu `randomForest`. Do oceny predykcji wykorzystano miarę AUC (Area Under ROC Curve) oraz AURPC (Area under the PR curve)

```r
#model
model <- randomForest(zyje ~., data = train)
#predykcja
pred_model_test <- predict(model, newdata = test, type = "prob")[,2]

score_0 <- pred_model_test[test$zyje == "TAK"]
score_1 <- pred_model_test[test$zyje == "NIE"]

# ROC Curve
roc <- roc.curve(scores.class0 = score_0, scores.class1 = score_1, curve = TRUE)
roc$auc
```

```
## [1] 0.653354
```

```r
# PR Curve
pr <- pr.curve(scores.class0 = score_0, scores.class1 = score_1, curve = TRUE)
pr$auc.integral
```

```
## [1] 0.7228618
```



Podsumowanie

Na zbiorze danych o nowotworze płuc przeprowadzono analizę danych, następnie na podstawie wiedzy eksperckiej oczyszczono dane z błędnych wartości. Podzielono dane, zbudowano dwa modele, interpretowalny re-

gresji logistycznej oraz model lasu losowego. Uzyskane wyniki zmierzono miarą AUC oraz AUPRC, w obu przypadkach model regresji logustycznej miał niewielką przewagę.