Titanic

Code

Paweł Warchoł i Jakub Rudzki

6 04 2020

Wstęp

14 kwietnia 1912 roku doszło do najsłynniejszej w historii katastrofy morskiej. Około godziny 23:40 statek RMS Titanic zderzył się z górą lodową, w skutek czego zatonął. Z 2200 pasażerów (wraz z załogą) statku w katastrofie życie straciło ponad 1500 osób. Choć całkowita pojemność szalup ratunkowych pozwalała na przyjęcie 1100 osób, wiele z nich odpłynęło częściowo pustych. Nie podjęto działań mających na celu ratowanie osób, które znalazły się w wodzie.

Cel Badania

Celem badania jest stworzenie modeli predykcyjnych dokonujących klasyfikacji pasażerów (nie przeżył / przeżył) i wybór najlepszego z tych modeli. Ponadto, na ich podstawie sprawdzimy jakie czynniki wpłynęły na przeżywalność katastrofy. Skorzystaliśmy z czterech metod:

- Regresja logistyczna
- Metoda K-najbliższych sąsiadów
- Metoda losowych lasów decyzyjnych
- Naiwny klasyfikator Bayesa

Dane

Badanie oparliśmy na zbiorze danym “Titanic: Machine Learning from Disaster” (dostępnym na stronie kaggle.com). Pobrany zbiór posiadał 891 rekordów i 12 zmiennych. Poniżej umieszczamy ich krótką charakterystykę.

- PassengerID - identyfikator przyporządkowujący każdemu rekordowi liczbę naturalną od 1 do 891.
- Survived - zmienna binarna informująca czy dany pasażer przeżył katastrofę.
- Pclass - w statku można było podróżować pierwszą, drugą i trzecią klasą. Zmienna informuje którą klasą podróżował dany pasażer. Może być istotna, gdyż pasażerowie z lepszych klas, w związku z wyższą opłatą, mogli spodziewać się pierwszeństwa w kolejce do szalup ratunkowych.
- Name - imię i nazwisko pasażera wraz z tytułem. Możliwe, że pasażerowie uznani za “bardziej wartościowych” (tj. posiadający odpowiednie tytuły wojskowe bądź naukowe) mieli pierwszeństwo.
- Age - wiek pasażera wyrażony w latach. Wiek mógłby być uznany za istotny czynnik z dwóch powodów. Możliwe, że chciano uratować jak najwięcej młodych ludzi (z powodu perspektywy dłuższego życia przed nimi), oraz ludzie w kwiecie wieku, jako sprawniejsi mieli fizycznie większe szanse na przeżycie w kryzysowych warunkach.
- SibSp - sumaryczna liczba rodzeństwa i małżonków pasażera (przebywających na statku). Być może ludzie posiadający rodzeństwo/będący w związku byli bardziej zdeterminowani do ucieczki/przeżycia.
- Parch - liczba rodziców i dzieci pasażera (przebywających na statku). Podobnie jak powyżej.
- Ticket - numer biletu.
- Fare - opłata za rejs.
- Cabin - numer kabiny w której pasażer mieszkał. Hipotetycznie pasażerowie z kabin znajdujących się nieopodal szalup ratunkowych mogli mieć większe szanse na przeżycie.
- Embarked - zmienna czynnikowa posiadająca 3 warianty. Statek płynął po trasie Southampton (S) - Cherbourg (C) - Queenstown (Q) - Nowy Jork. Wartość zmiennej informuje w którym porcie wsiadł pasażer.

Wstępna obróbka danych

Bezpośrednio po zaimportowaniu do programu dane posiadały chaotyczną strukturę. W celu zwiększenia przejrzystości i transparentności konieczne było dostosowanie ich do potrzeb badania. Część zmian miała charakter kosmetyczny - na przykład w przypadku zmiennej objaśnianej *Survived* z postaci (0,1) na postać (“dead”, “alive”). Część pozornie nieznacznych zmian, była w rzeczywistości bardzo istotna. Na przykład zmienna Pclass ze zmiennej liczbowej (o wartościach 1, 2, 3) na zmienną czynnikową (“1st”, “2nd”, “3rd”). W miejsca brakujących wartości wprowadzono NA, by program mógł je w odpowiedni sposób zauważyć.

Code

Pierwsza selekcja zmiennych

Po wykonaniu wspomnianych wyżej operacji, podjęliśmy pierwsze decyzje odnośnie terminacji niektórych zmiennych. Zdecydowaliśmy się usunąć zmienną *Name* - po oględzinach zauważyliśmy, że w dominującej części tytułów występowały jedynie zwroty nakierowujące na płeć, co w bezpośrendi spoób przekładało się na informacje zawarte już w zmiennej *Sex*. Ponadto usunęliśmy zmienną *Ticket* - występujące tam ciągi znaków były praktycznie niemożliwe do zinterpretowania i nieprzydatne do analizy. Zniknęła też zmienna *Cabin* - charakteryzowała się

ogromną liczbą braków (687), przez co zdecydowaliśmy o usunięciu jej z dalej rozważanego zbioru. Zniknęła też zbędna zmienna *PassengerID*, która w kontekście badania byłaby bezużyteczna.

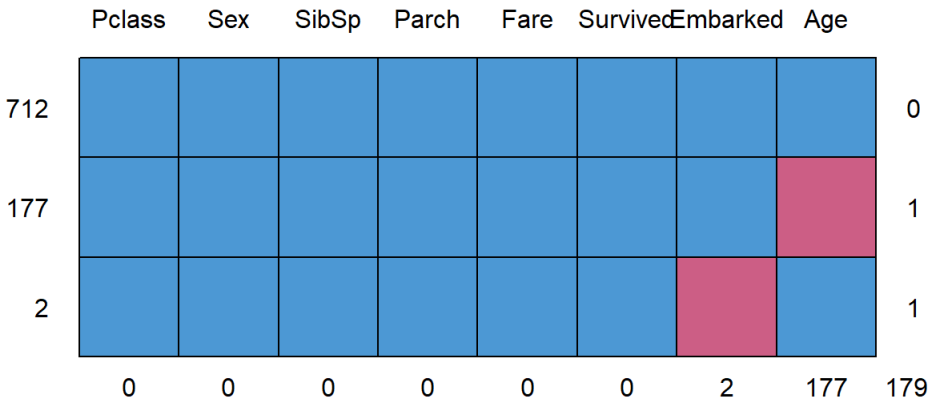
Uzupełnienie braków danych.

W celu sprawdzenia jak wygląda sytuacja z brakami danych zliczyliśmy obserwacje, które zawierają co najmniej 1 brak oraz sprawdziliśmy jak sytuacja wygląda na tle wszystkich zmiennych.

Code

```
## [1] 179
```

Code




```
##     Pclass Sex SibSp Parch Fare Survived Embarked Age
## 712      1   1     1     1    1        1        1   1   0
## 177      1   1     1     1    1        1        1   0   1
## 2        1   1     1     1    1        1        0   1   1
##          0   0     0     0    0        0        2 177 179
```

Stąd widzimy, że mamy 179 obserwacji z brakami danych, wśród których 2 obserwacje nie zawierają informacji odnośnie portu, z którego dany pasażer wypłynął oraz pozostałe 177 obserwacji nie zawiera wartości dla zmiennej Age. W celu osiągnięcia jak najlepszych wyników predykcji dla tworzonych modeli, zdecydowaliśmy o nieusuwaniu tych obserwacji ze zbioru (ich usunięcie spowodowałoby znaczną utratę informacji, szczególnie że stanowiły one około 20% wszystkich obserwacji) i zastąpieniu wartości NA poprzez prognozy wykonane metodą Random forest tj. lasów losowych. Jest to metoda, opierająca swoje działanie na stworzeniu wielu modeli drzew losowych i dokonywaniu prognozy na ich podstawie - dzięki czemu pozornie niezbyt dobre w kontekście klasyfikacji narzędzie jakim są drzewa decyzyjne staje się bardzo silnym narzędziem.

Code

```
##   missForest iteration 1 in progress...done!
##   missForest iteration 2 in progress...done!
##   missForest iteration 3 in progress...done!
##   missForest iteration 4 in progress...done!
##   missForest iteration 5 in progress...done!
```

Code

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

Code

Code

```
##  Pclass        Sex           Age            SibSp           Parch
##  1st:216   female:314   Min.   : 0.00   Min.   :0.000   Min.   :0.0000
##  2nd:184   male  :577   1st Qu.:21.00   1st Qu.:0.000   1st Qu.:0.0000
##  3rd:491                Median :28.00   Median :0.000   Median :0.0000
##                         Mean   :29.45   Mean   :0.523   Mean   :0.3816
##                         3rd Qu.:36.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                         Max.   :80.00   Max.   :8.000   Max.   :6.0000
##       Fare         Embarked     Y
##  Min.   :  0.00   C:170    dead :549
##  1st Qu.:  7.91   Q: 77    alive:342
##  Median : 14.45   S:644
##  Mean   : 32.20
##  3rd Qu.: 31.00
##  Max.   :512.33
```

Code

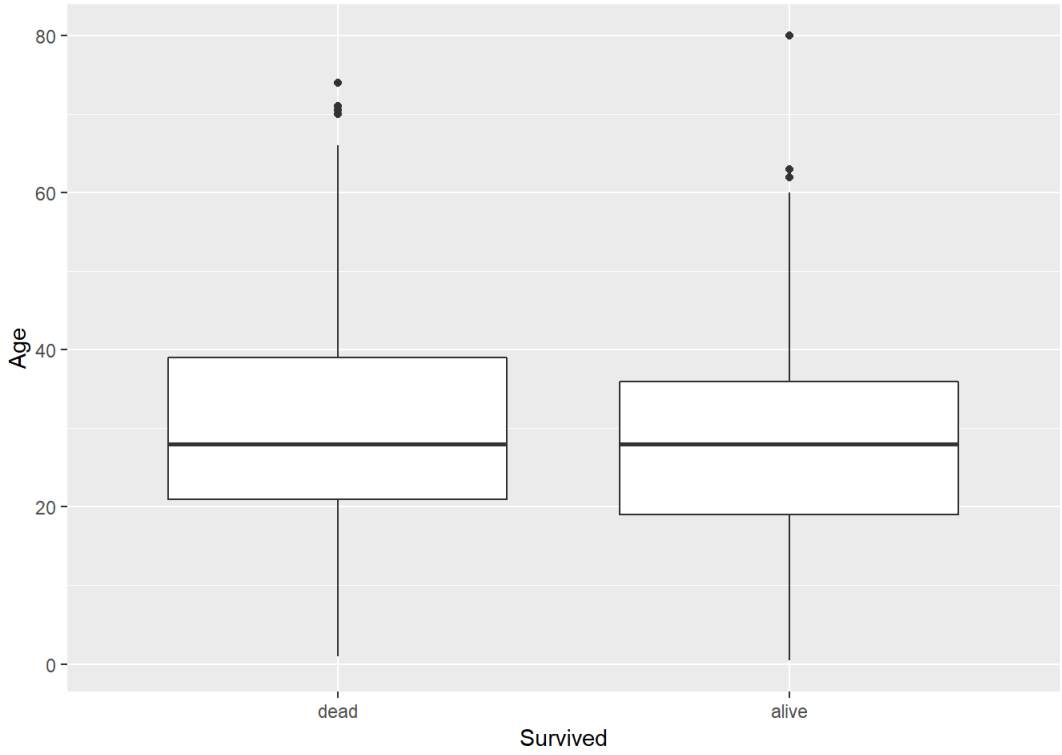
```
## 'data.frame':    891 obs. of  8 variables:
##  $ Pclass  : Factor w/ 3 levels "1st","2nd","3rd": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Sex     : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age     : int  22 38 26 35 35 31 54 2 27 14 ...
##  $ SibSp   : num  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch   : num  0 0 0 0 0 0 0 1 2 0 ...
##  $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Embarked: Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...
##  $ Y       : Factor w/ 2 levels "dead","alive": 1 2 2 2 1 1 1 1 2 2 ...
```

Widzimy, że brakujące wartości zostały uzupełnione. Struktura danych i ich podstawowe statystyki opisowe po uzupełnieniu prezentują się jak powyżej.

Wartości odstające

W celu wychwycenia obserwacji, które w znaczący sposób odstają od reszty i mogą nieporządanie wpłynąć na model utworzyliśmy wykresy pudełkowe z wąsami dla zmiennych numerycznych.

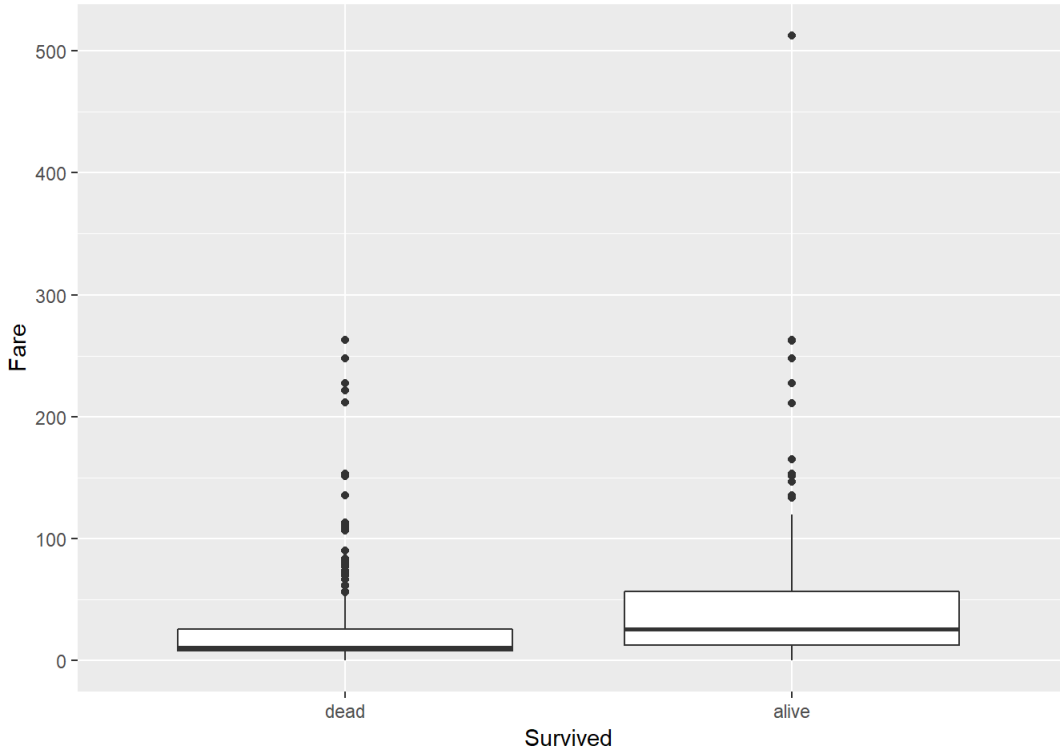
Code



W przypadku zmiennej age nie zauważyliśmy sytuacji wymagającej naszej ingerencji. Występowały osoby starsze, jednak na tyle mało

licznie, że uznaliśmy sytuację za normalną.

Code



Dla zmiennej fare (oznaczającej wysokość opłaty za bilet) zauważamy, że występują obserwacje odstające bardzo wyraźnie.

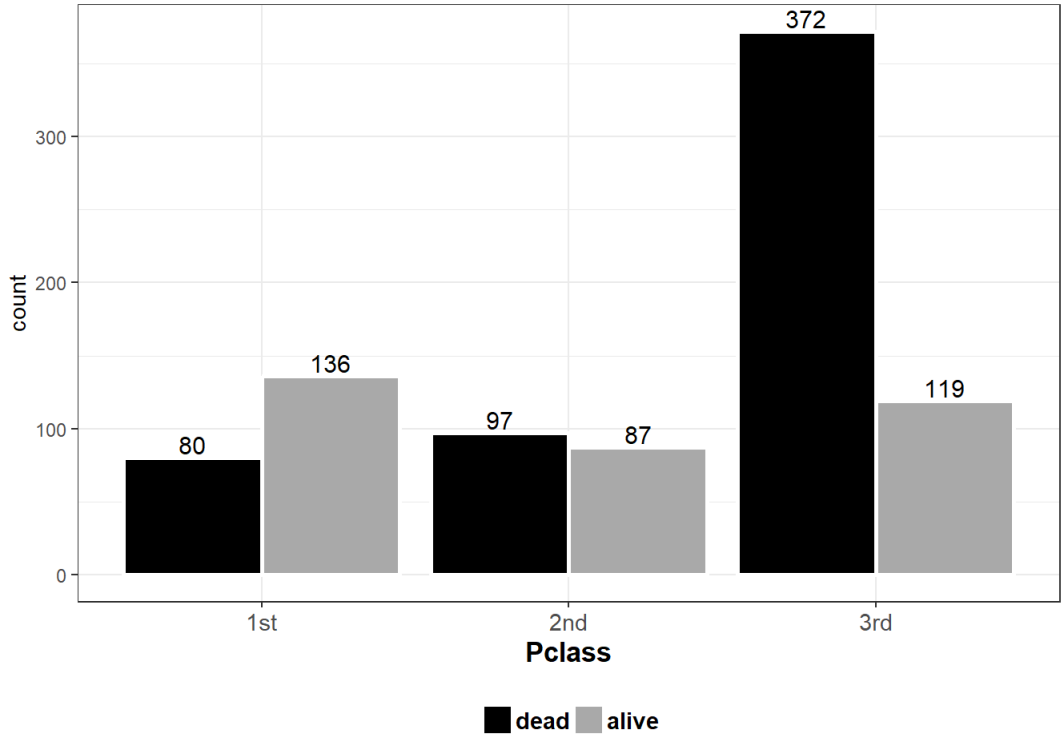
Code

```
##     Pclass    Sex Age SibSp Parch     Fare Embarked Survived
## 259    1st female  35     0     0 512.3292        C    alive
## 680    1st   male  36     0     1 512.3292        C    alive
## 738    1st   male  35     0     0 512.3292        C    alive
```

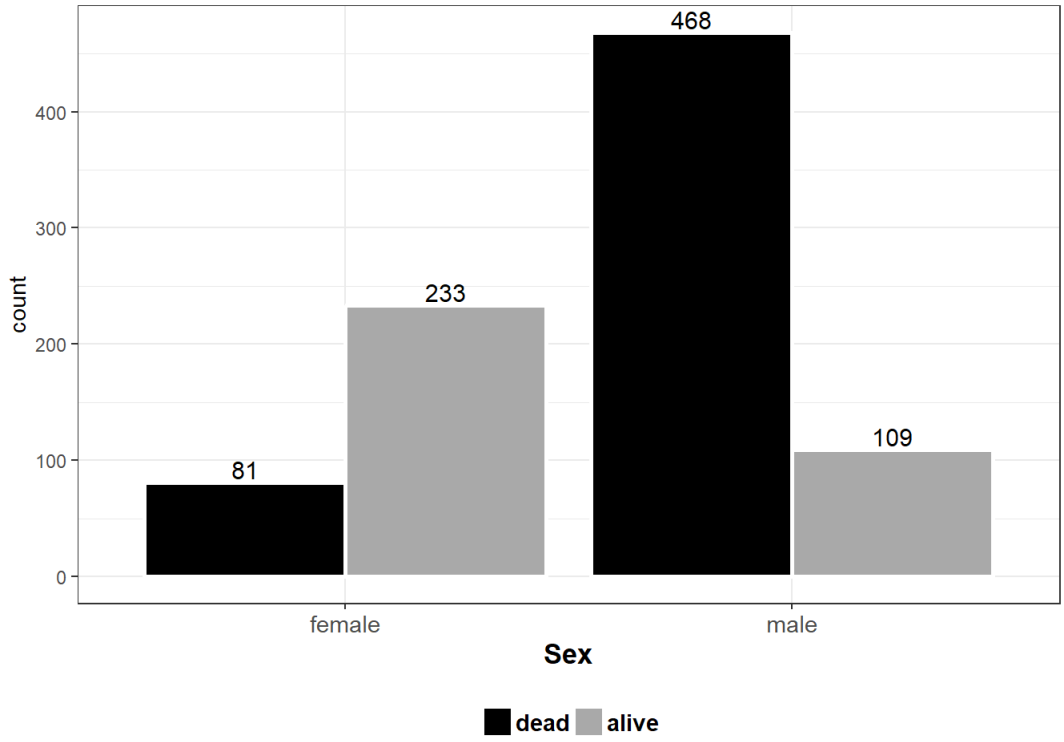
Po bliższemu przyjrzeniu się odstającym rekordom zauważyliśmy, że dotyczą trójki pasażerów z pierwszej klasy, którzy zapłacili za rejs ogromną opłatę. Wszystkie te osoby przeżyły, co wydaje nam się być logiczne i sensowne, zdecydowaliśmy się więc na pozostawienie tych obserwacji w ostatecznym zestawie danych.

Ostateczny zbiór danych

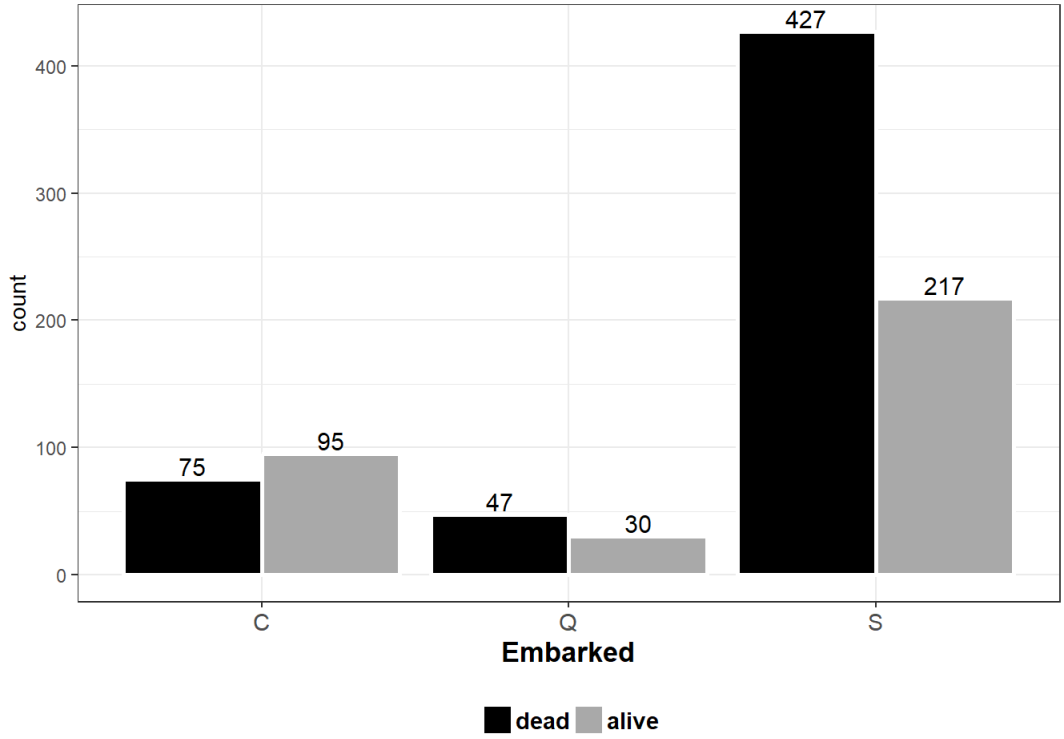
Poniżej przedstawione zostaną wizualizacje mające na celu przybliżenie struktury danych w zmiennych wybranych do badania.



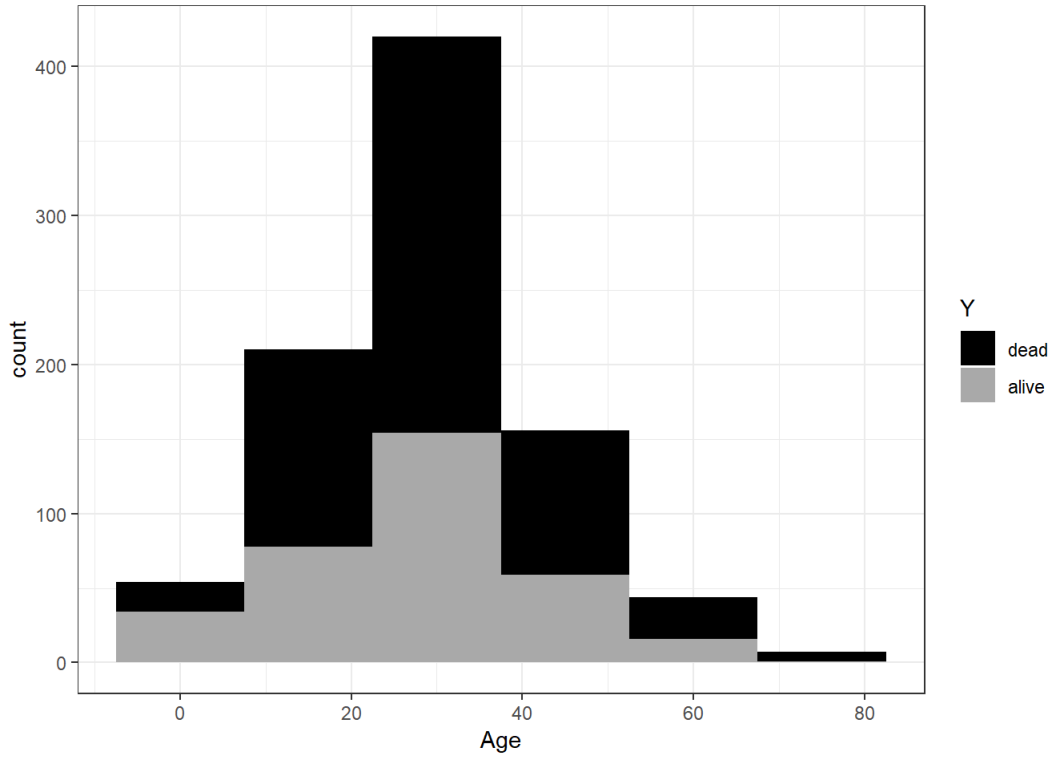
Już w przypadku pierwszej zmiennej objaśniającej*Pclass* wyraźnie widać różnice pomiędzy poszczególnymi klasami. Jedynie w pierwszej klasie liczba osób, które przeżyły jest wyższa niż liczba zgonów. Dla klasy drugiej wartości te są porównywalne, natomiast w klasie trzeciej, dysproporcja jest ogromna na niekorzyść pasażerów, którym udało się przeżyć



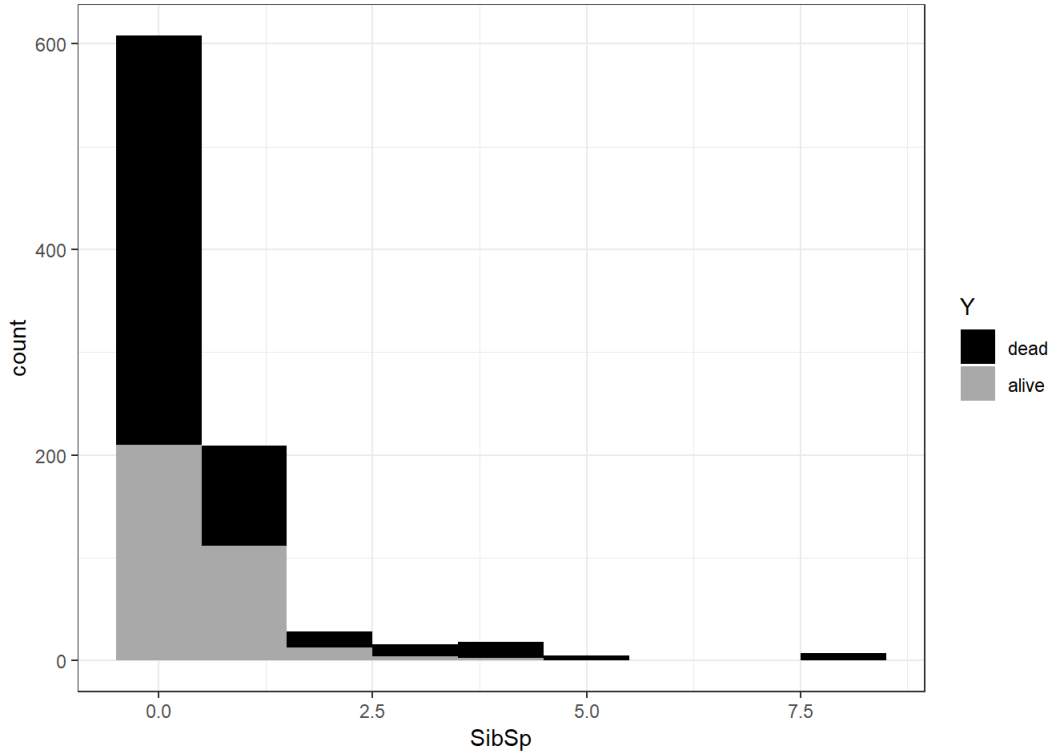
W zmiennej *Sex* również widać bardzo wyraźne dysproporcje. Więcej kobiet przeżyło katastrofę, natomiast w grupie mężczyzn sytuacja jest odwrotna - na każdego mężczyznę, który przeżył przypadało 4, którym się to nie udało.



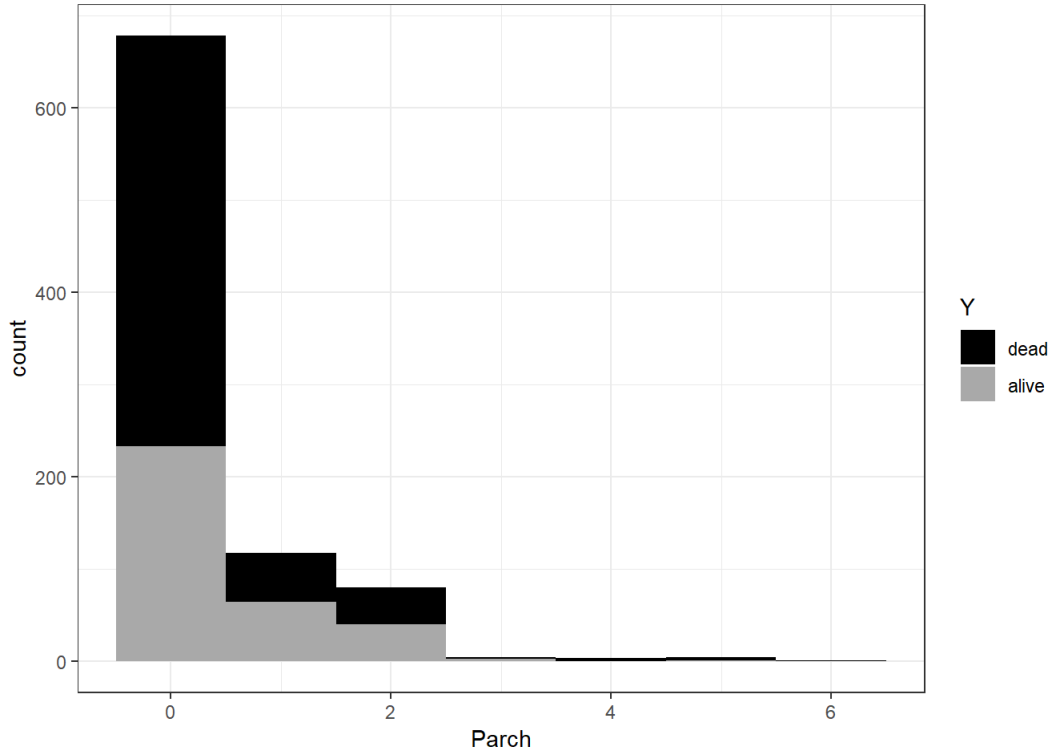
Analizując porty z których pasażerowie wypłynęli od razu widzimy, że największa liczba osób płynęła od samego portu macierzystego w Southampton. Zgodne z informacjami zawartymi we wstępnie, większości pasażerów nie udało się tej podróży przeżyć. Pasażerów płynących z Cheerbourg i Quennstown jest odpowiednio mniej, więc ciężko mówić o istotnych statystycznie zależnościach już na tym etapie.



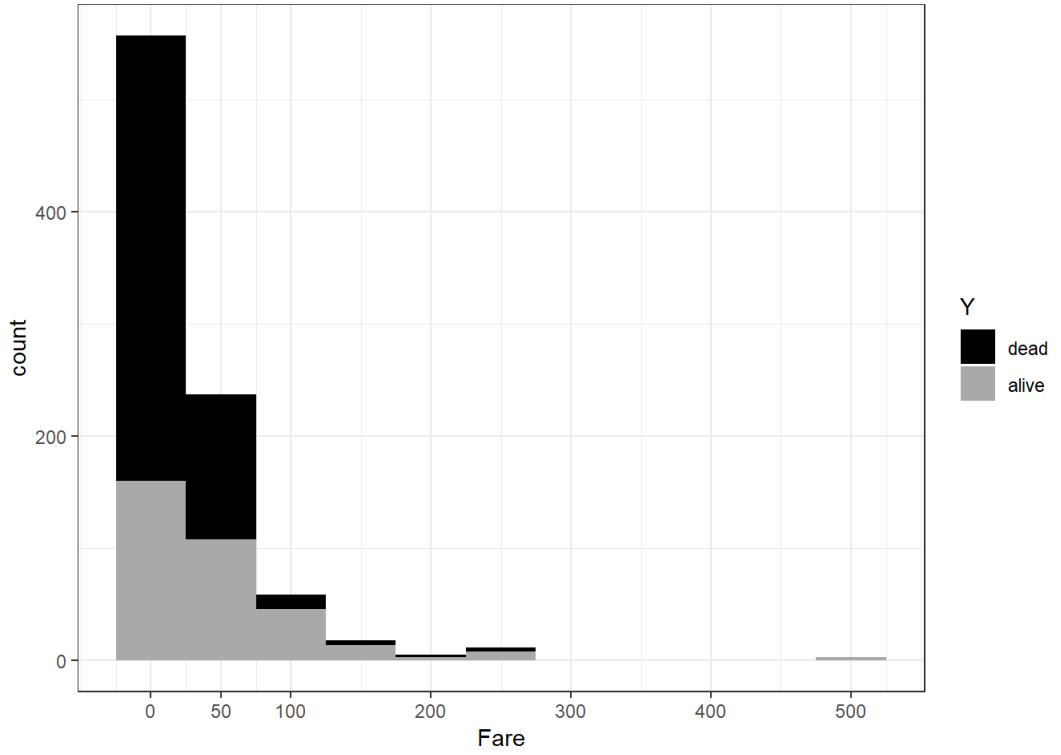
Widzimy, że struktura wieku przypominała rozkład normalny. Liczba osób, które przeżyły wydaje się rozkładać proporcjonalnie w każdej grupie wiekowej, oprócz najmłodszych pasażerów, gdzie więcej osób przeżyło katastrofę.



Zdecydowana większość pasażerów podróżowała bez rodzeństwa bądź małżonka/małżonki.



Bardzo podobna sytuacja ma miejsce w przypadku pasażerów podróżujących z rodzicami lub dziećmi.



Wydaje się, że wraz ze wzrostem wysokości opłaty dokonywanej za rejs stosunek liczby osób które przeżyły, do tych, które zginęły w trakcie katastrofy, jest coraz lepszy.

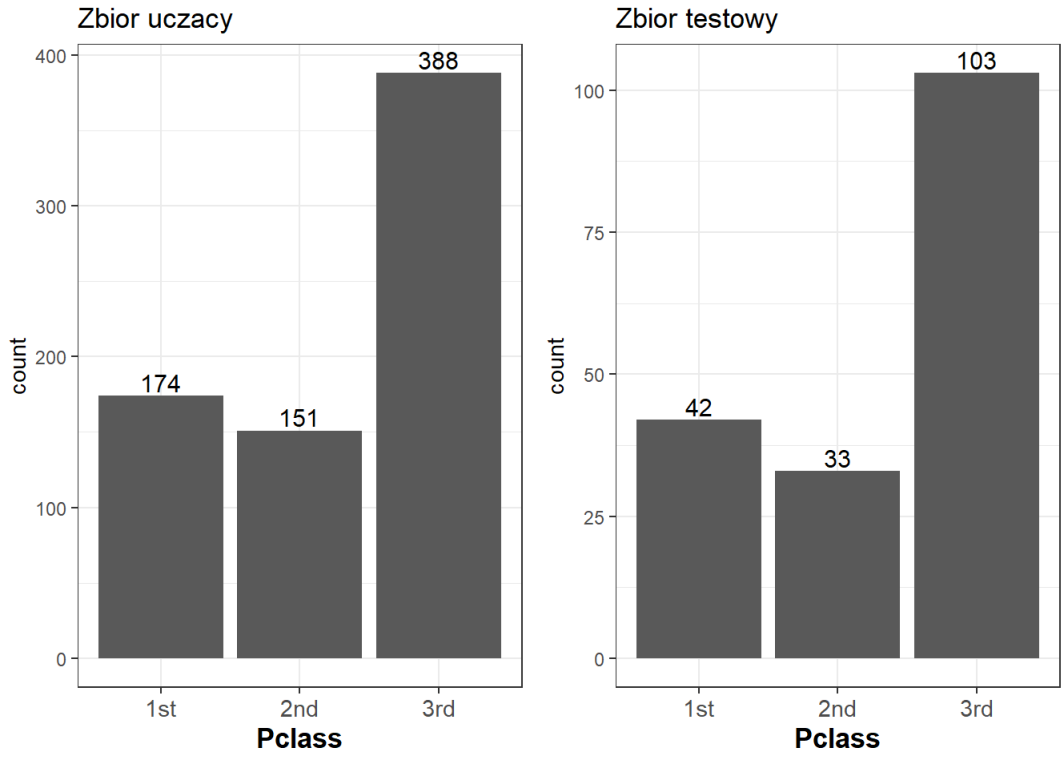
Część właściwa badania

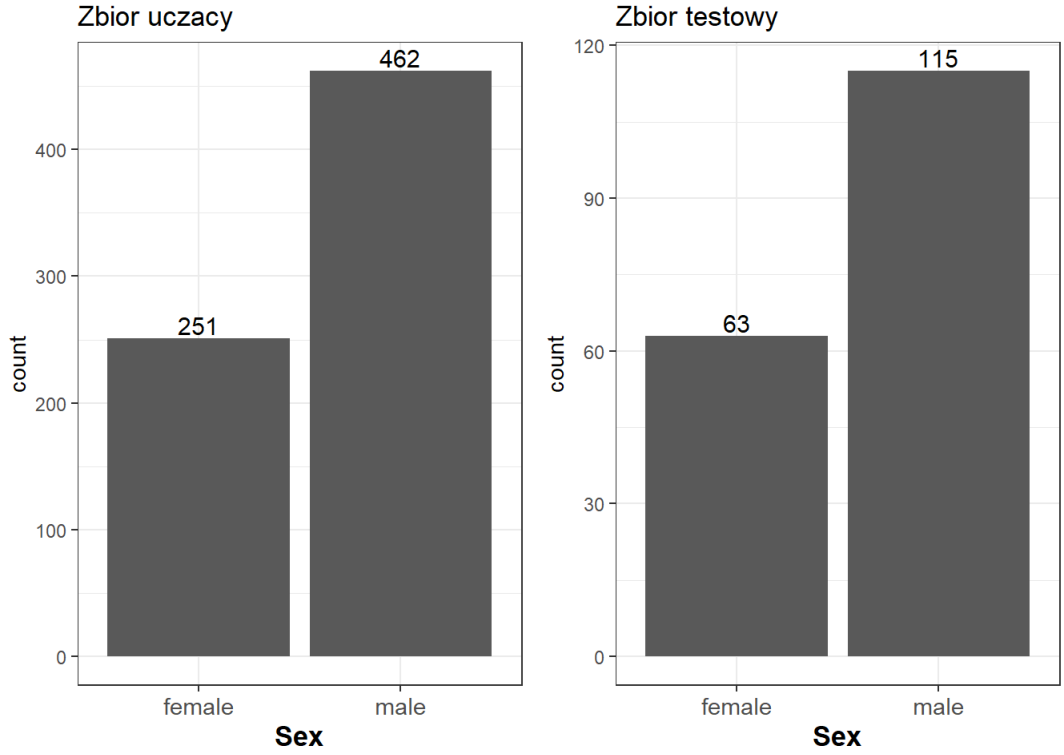
Po wstępie i opisaniu danych, na których oparte jest badanie należy przejść do właściwych czynności. Pierwszą z nich jest podzielenie zbioru danych na dwie części - uczącą i testującą. Wybrana przez nas proporcja:

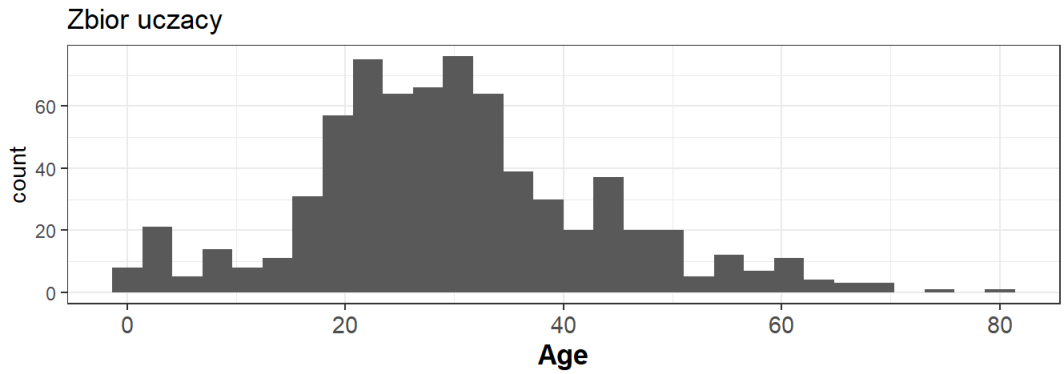
- 80% obserwacji w zbiorze *train*
- 20% obserwacji w zbiorze *test*

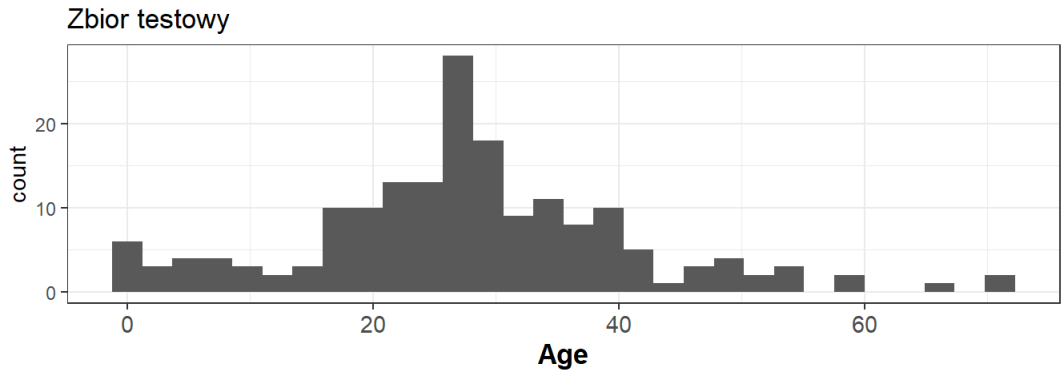
By podział był możliwe najlepszy, zdecydowaliśmy się dla otrzymanych podzbiorów utworzyć wizualizacje sprawdzające strukturę danych. Jak na załączonych wykresach widać - wartości wydają się być zbliżone, stąd podział można uznać za satysfakcjonujący.

Code

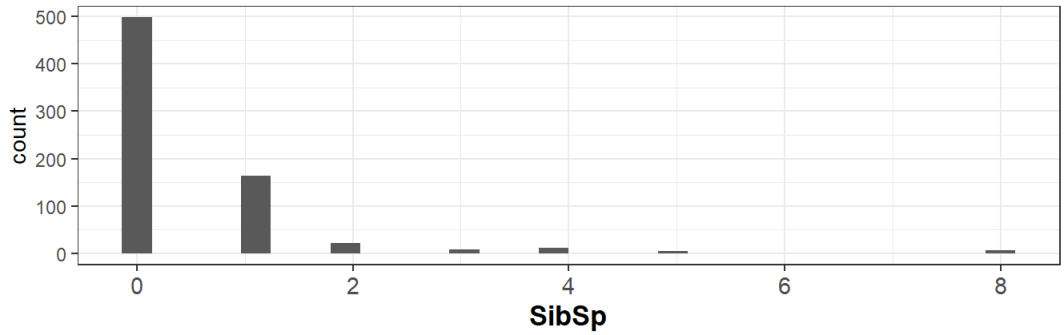



Zbior uczacy
462
251
count
400
300
200
100
0
female
male
Sex
Zbior testowy
120
90
60
30
0
115
63
count
female
male
Sex

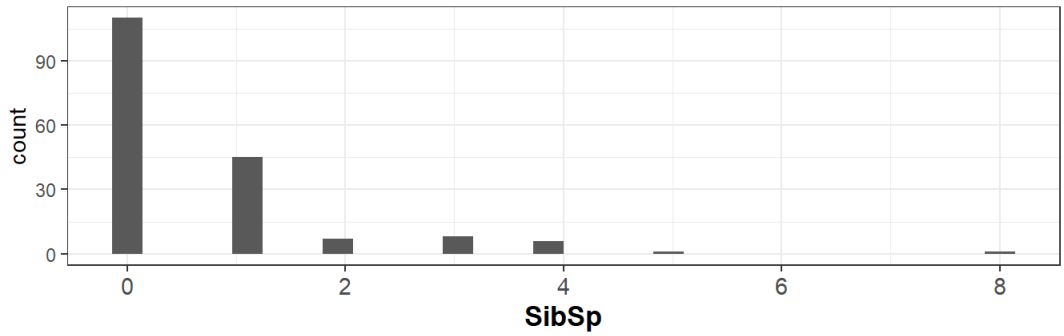

Zbior uczacy
count
60
40
20
0
0
20
40
60
80
Age


Zbior testowy
count
20
10
0
0
20
40
60
Age

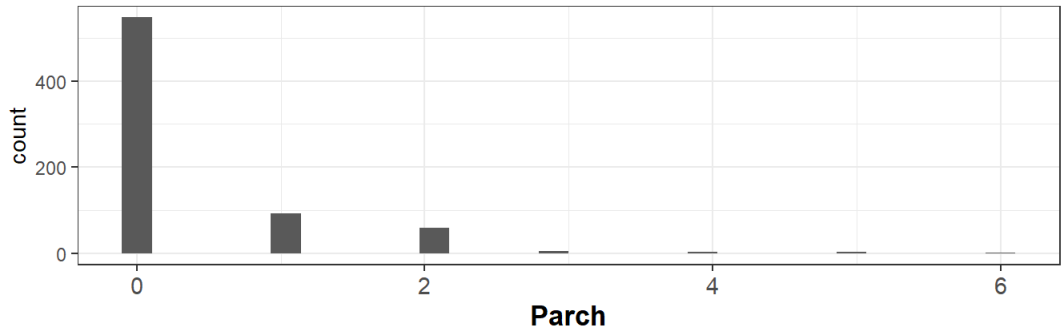
Zbior uczacy



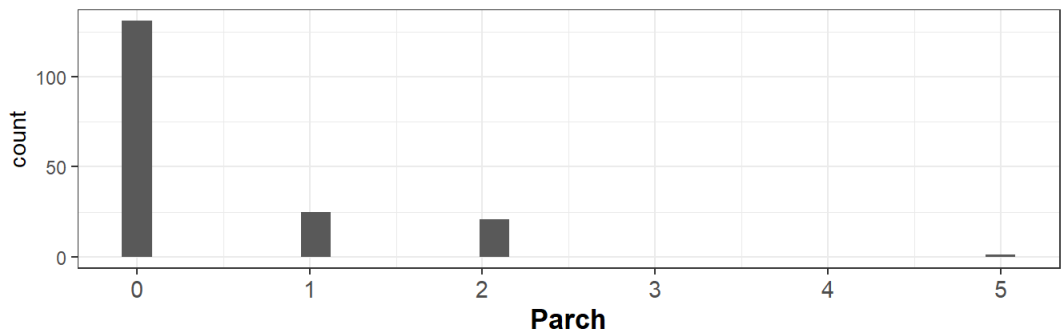
Zbior testowy

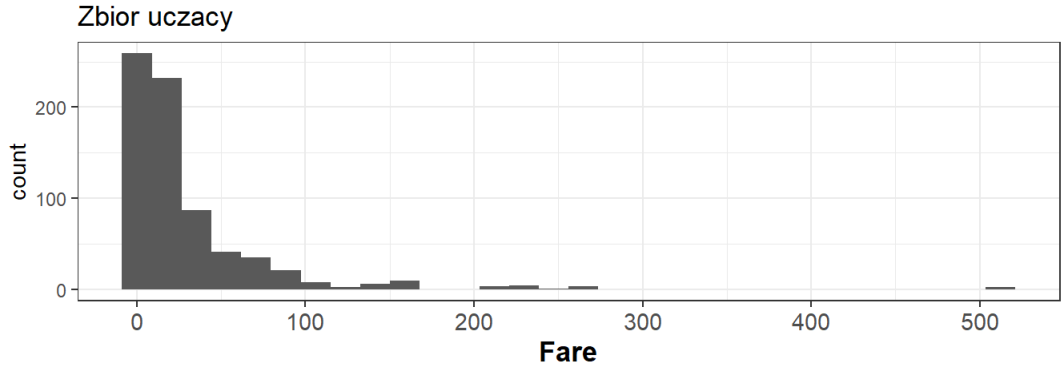


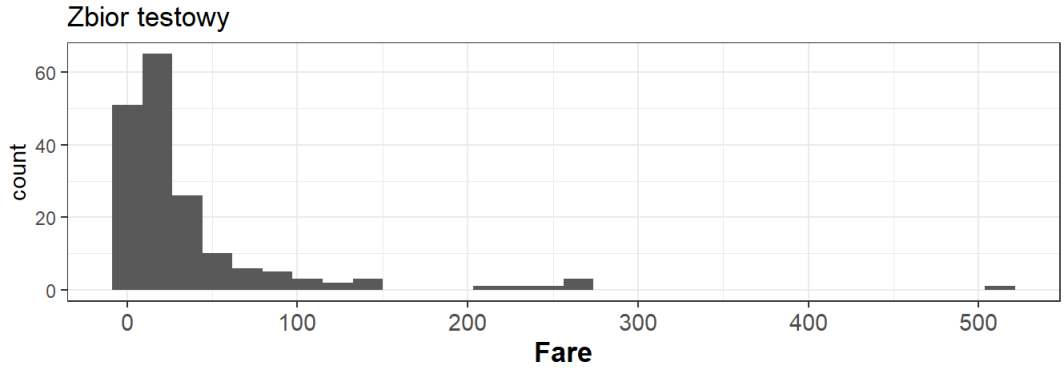
Zbior uczacy

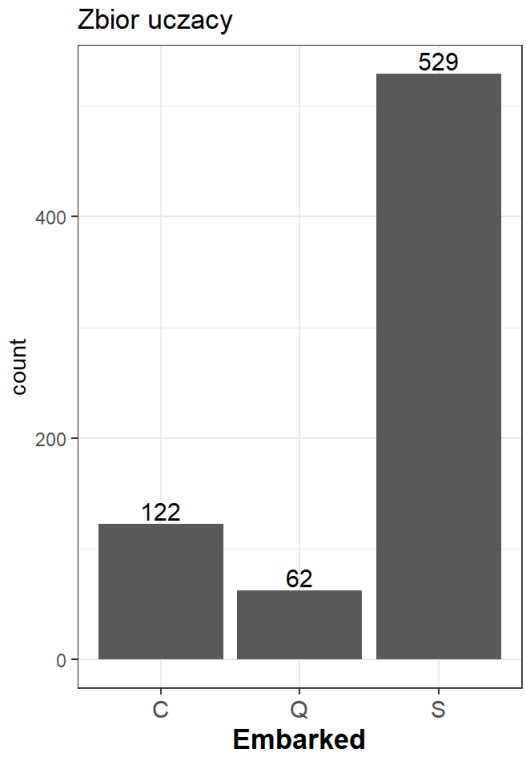


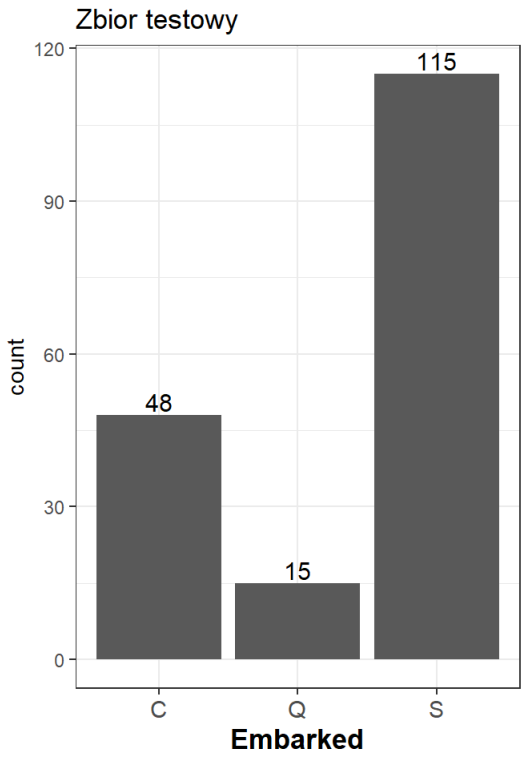
Zbior testowy

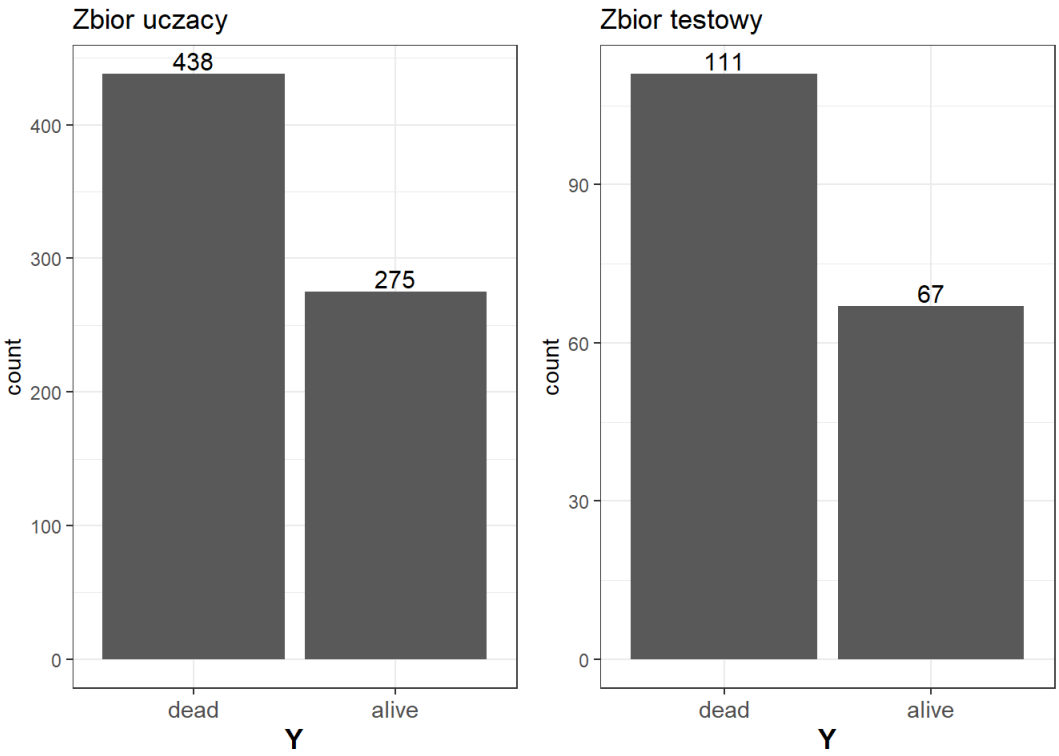



Zbior uczacy
count
200
100
0
0
100
200
300
400
500
Fare


Zbior testowy
count
60
40
20
0
0
100
200
300
400
500
Fare


Zbior uczacy
count
400
200
0
122
62
529
C
Q
S
Embarked


Zbior testowy
count
120
90
60
30
0
48
15
115
C
Q
S
Embarked



Regresja logistyczna

Pierwszą metodą, przy pomocy której będziemy sprawdzać od czego zależała przeżywalność katastrofy z 1912 roku to regresja logistyczna. Można z niej skorzystać gdy do czynienia mamy z dychotomiczną zmienną objaśnianą. Model regresji logistycznej jest szczególnym przypadkiem GLM, w którym wykorzystano funkcję logit. Funkcja ta przekształca prawdopodobieństwo na logarytm szans.

Modelowanie

Jako, że cały proces obróbki zmiennych został zaprezentowany wcześniej. Pozostała jedynie kwestia stworzenia optymalnego modelu. W tym celu utworzony zostanie model zerowy - z wszystkimi dostępnymi zmiennymi objaśniającymi, a następnie metodą krokową wsteczną (ze względu na kryterium informacyjne Akaikego) odnaleziona zostanie najlepsza kombinacja zmiennych.

```
## Start:  AIC=665.04
## Y ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked
##
##            Df Deviance    AIC
## - Embarked  2   646.93 662.93
## - Fare      1   645.25 663.25
## - Parch     1   646.01 664.01
## <none>          645.04 665.04
## - SibSp     1   655.23 673.23
## - Age       1   665.53 683.53
## - Pclass    2   694.51 710.51
## - Sex       1   808.88 826.88
##
## Step:  AIC=662.93
## Y ~ Pclass + Sex + Age + SibSp + Parch + Fare
##
##          Df Deviance    AIC
## - Fare    1   647.43 661.43
## - Parch   1   647.96 661.96
## <none>        646.93 662.93
## - SibSp   1   658.55 672.55
## - Age     1   668.62 682.62
## - Pclass  2   698.91 710.91
## - Sex     1   819.45 833.45
##
## Step:  AIC=661.43
## Y ~ Pclass + Sex + Age + SibSp + Parch
##
##          Df Deviance    AIC
## - Parch   1   648.22 660.22
## <none>        647.43 661.43
## - SibSp   1   658.67 670.67
## - Age     1   670.26 682.26
## - Pclass  2   730.32 740.32
## - Sex     1   820.30 832.30
##
## Step:  AIC=660.22
## Y ~ Pclass + Sex + Age + SibSp
##
##          Df Deviance    AIC
## <none>        648.22 660.22
## - SibSp   1   663.10 673.10
## - Age     1   670.89 680.89
## - Pclass  2   731.51 739.51
## - Sex     1   825.15 835.15
```

Zmienne rekomendowane przez metodę: *Pclass*, *Sex*, *Age*, *SibSp*, *Embarked*. Widząc jednak mały wpływ zmiennej *Embarked*, oraz niemalże identyczną wartość AIC, zdecydowaliśmy się na nie umieszczanie jej w ostatecznym modelu.

```
##
## Call:
## glm(formula = Y ~ Pclass + Sex + Age + SibSp, family = binomial("logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6181  -0.6232  -0.4316   0.6517   2.3681
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.878561   0.441610   8.783  < 2e-16 ***
## Pclass2nd   -1.128827   0.289298  -3.902 9.54e-05 ***
## Pclass3rd   -2.303329   0.275720  -8.354  < 2e-16 ***
## Sexmale     -2.547970   0.210933 -12.080  < 2e-16 ***
## Age         -0.039308   0.008573  -4.585 4.54e-06 ***
## SibSp       -0.405629   0.120340  -3.371  0.00075 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 950.83  on 712  degrees of freedom
## Residual deviance: 648.22  on 707  degrees of freedom
## AIC: 660.22
##
## Number of Fisher Scoring iterations: 5
```

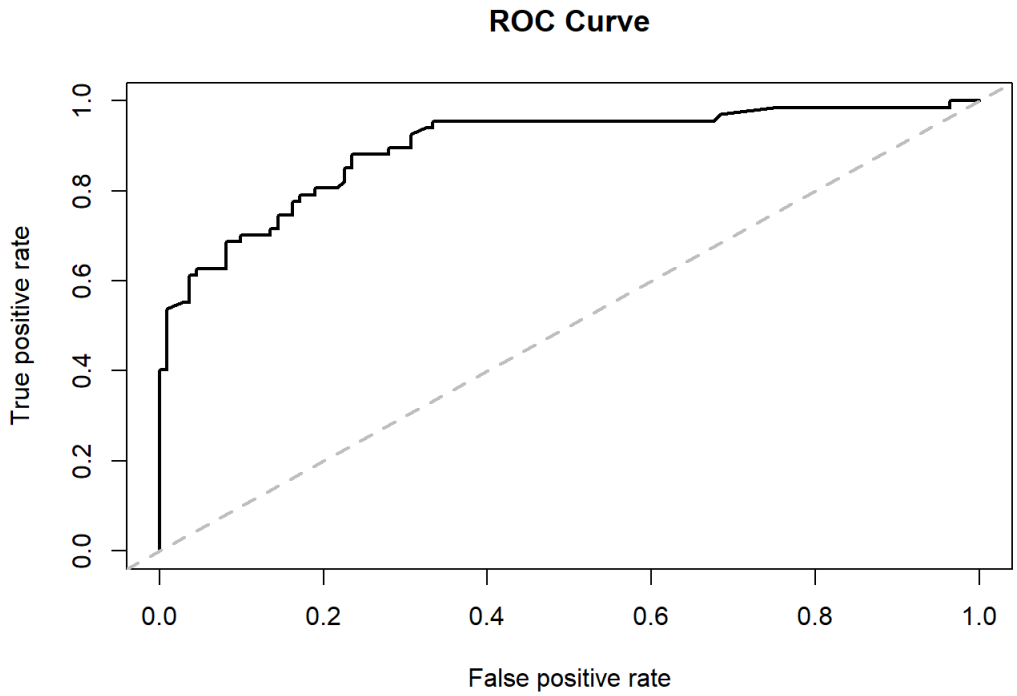
Podczas analizy poszczególnych współczynników zauważyć można wyraźnie negatywny wpływ każdego z nich. Przemieszczanie się w drugiej i trzeciej klasie istotnie zmniejszało szanse na przeżycie. Sytuacja wygląda podobnie dla płci, kobiety miały większą szansę na przeżycie. W przypadku wieku, każdy rok zmniejszał szansę na przeżycie w znikomy sposób. Być może wycentrowanie tej zmiennej na "przeciętnej" wartości dałoby precyzyjniejsze wyniki. Co ciekawe, posiadanie rodzeństwa lub bycie w związku małżeńskim również przechylało szalę w stronę nieprzeżycia.

Badanie jakości predykcji

Korzystając z wylosowanego wcześniej zbioru testowego funkcją**predict** wyliczyliśmy szanse. Następnie dla każdej wartości większej niż 0,5 wstawiliśmy 1 - w znaczeniu prognozy, że dany pasażer przeżyje. W przeciwnym wypadku 0. Następnie zestawiliśmy wyniki predykcyjne z rzeczywistymi. Uzyskaliśmy precyzję na poziomie 80%.

```
## [1] "Accuracy 0.8034"
```

Dla powyższego modelu przedstawiono również krzywą ROC:



AUC w przypadku tego modelu wyniosło:

Code

```
## [[1]]
## [1] 0.8976066
```

Co jest naprawdę dobrym wynikiem. Te metryki posłużą w ostatniej części do porównania modeli.

Metoda K-najbliższych sąsiadów

Drugim narzędziem z którego zdecydowaliśmy się skorzystać to metoda K-najbliższych sąsiadów (nazwa z angielskiego k nearest neighbours). Sposób działania jest bardzo prosty, dla każdego rekordu obliczana jest odległość od pozostałych elementów. W kolejnym kroku przeliczana jest liczba obiektów z danych grup w tym k - elementowym zbiorze. A następnie na tej podstawie przyporządkowywana jest wartość zmiennej objaśnianej (taka sama jak najliczniejszego elementu w zbiorze).

Z racji specyfiki metody KNN w zbiorze wykorzystanym do budowy tego modelu znaleźć się mogą jedynie zmienne numeryczne. Stąd dla zmiennych kategorycznych zdecydowaliśmy o ich odpowiednim przekształceniu do formy zero-jedynkowej, dzięki czemu będzie możliwe ich uwzględnienie w modelu.

Standaryzacja

W przypadku, gdy zmienne istotnie różnią się od siebie rzędem wielkości należy je znormalizować. W badaniu użyliśmy standaryzacji, jako jednej z metod normalizacji danych. W przypadku nie zastosowania takiego rozwiązania, zmienne o dużych wartościach mogłyby zbyt mocno wpłynąć na predykcję.

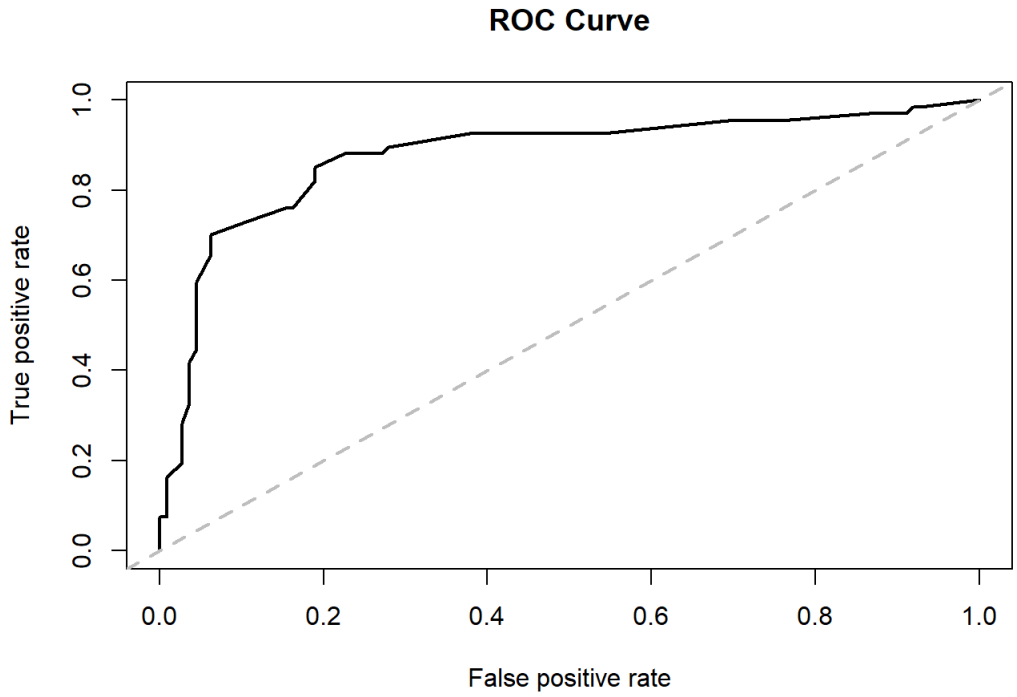
Po przygotowaniu zmiennych stworzona została pętla mająca na celu zlokalizowanie odpowiedniego**k**, dla którego jakość predykcji jest najlepsza.

Code

```
##    k  accuracy
## 1 19 0.8483146
## 2 12 0.8370787
## 3 17 0.8370787
## 4 18 0.8370787
## 5 21 0.8370787
## 6 23 0.8370787
```

Jak widać celność predykcji jest największa dla**k** równego 19 i wynosi aż 84,83%, co jest zaskakująco dobrym wynikiem.
Poniżej wykres krzywej ROC:

Code



Oraz wartość AUC:

Code

```
## [[1]]
## [1] 0.8763614
```

Również na bardzo wysokim poziomie.

Lasy losowe

Algorytm lasów losowych, jak już przy okazji uzupełniania braków danych wspomniano, polega na stworzeniu wielu drzew decyzyjnych i dokonaniu na ich podstawie predykcji. W problemach klasyfikacyjnych o tym jaką klasę należy przypisać do danego obiektu decyduje ‘głosowanie większościowe’, podobnie jak w metodzie KNN. Pomimo, że metoda ta bazuje na drzewach decyzyjnych, które pojedynczo mają sporo wad i nie są zbyt dobrym narzędziem, to dzięki wspomnianej specyfice potrafi dawać naprawdę dobre rezultaty. W kontekście tworzonych drzew należy wspomnieć jeszcze o fakcie, że w celu uniknięcie sytuacji, w której wpływ jednej ze zmiennych jest na tyle duży, że dominuje ona każde tworzone drzewo i tym samym zmniejsza wagę pozostałych, stosuje się pewną modyfikację. Mianowicie, do budowy kolejnych drzew wykorzystywana jest jedynie część ze zmiennych, najczęściej pierwiatek z liczby wszystkich zmiennych. Z tej racji, w naszym modelu zdecydowaliśmy o losowaniu do każdego drzewa 3 spośród dostępnych zmiennych.

Poniżej model i output funkcji `confusionMatrix` dla stwożonego modelu:

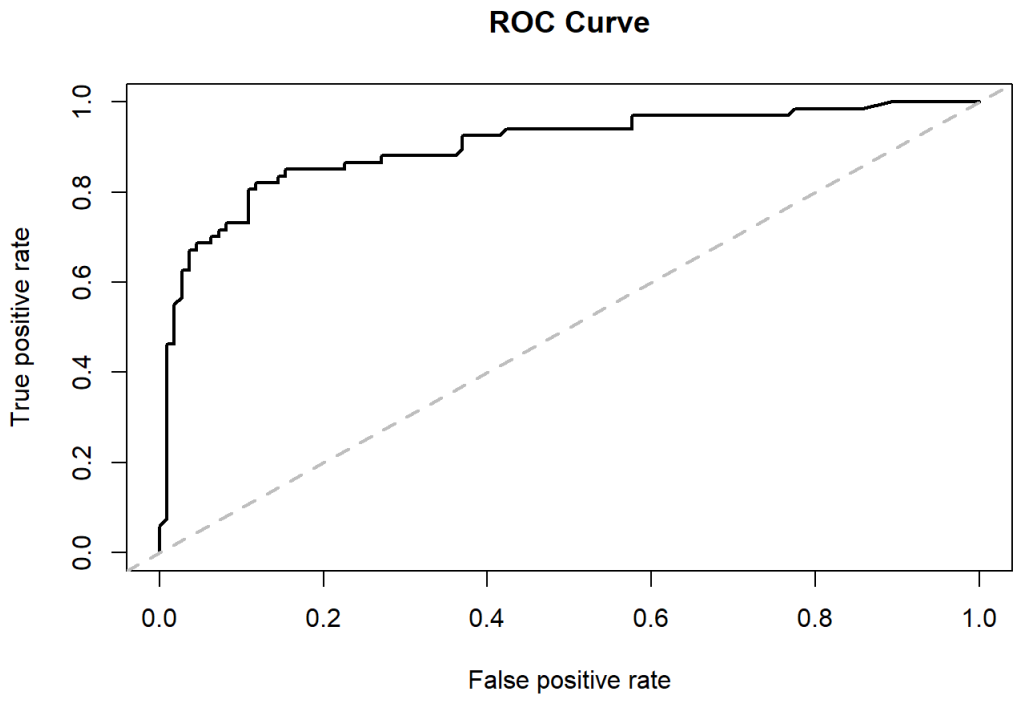
Code

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction dead alive
##      dead   103    19
##      alive    8    48
##
##                Accuracy : 0.8483
##                  95% CI : (0.787, 0.8976)
##     No Information Rate : 0.6236
##     P-Value [Acc > NIR] : 3.494e-11
##
##                   Kappa : 0.666
##  Mcnemar's Test P-Value : 0.05429
##
##             Sensitivity : 0.7164
##             Specificity : 0.9279
##          Pos Pred Value : 0.8571
##          Neg Pred Value : 0.8443
##              Prevalence : 0.3764
##          Detection Rate : 0.2697
##    Detection Prevalence : 0.3146
##       Balanced Accuracy : 0.8222
##
##        'Positive' Class : alive
##
```

Widzimy, że model oparty na algorytmie lasów losowych uzyskał skuteczność predykcji na poziomie 84,83%.

Poniżej krzywa ROC:

Code



Oraz AUC uzyskane przez model:

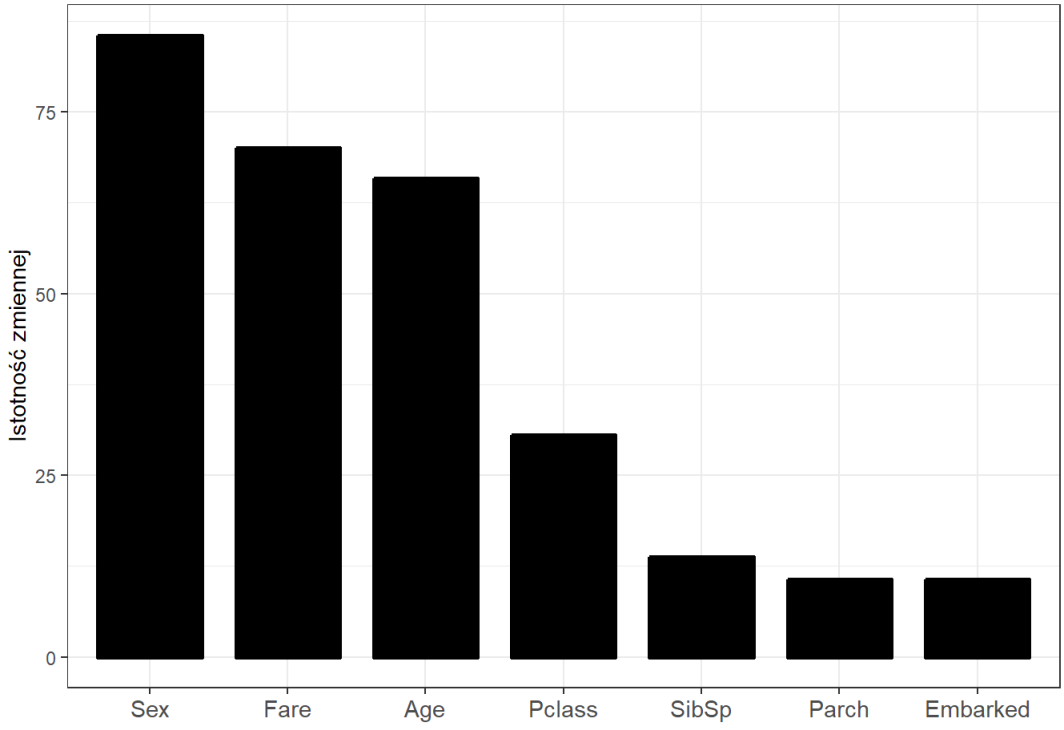
Code

```
## [[1]]
## [1] 0.9012371
```

Wartość AUC przekroczyła próg 0.9 co jest świetnym wynikiem, co zresztą widać po krzywej ROC.

W celu sprawdzenia, jakie zmienne były najważniejsze w modelu stworzono wykres za pomocą funkcji `varImpPlot`, który obrazuje istotność zmiennych w modelu:

Code



Istotność zmiennych jest weryfikowana na podstawie sumy spadku wartości indeksu Giniego po podziale zbioru względem danej zmiennej. Oznacza to, że za zmienne najbardziej istotne należy uznać te o największej wartości sumy spadków – pokazuje to, że wykorzystanie danej zmiennej do podziału drzewa w największym stopniu wpływało na lepsze predykcje. Widzimy, że najważniejszą zmienną była zmienna Sex, w dalszej kolejność Fare, Age oraz Pclass. Pozostałe zmienne były wyraźnie mniej istotne.

Klasyfikator naiwny Bayesa

Ostatnim z modeli będzie klasyfikator naiwny Bayesa. Technika ta opiera się na twierdzeniu Bayesa, które mówi, że prawdopodobieństwo warunkowe bycia w stanie X, pod warunkiem posiadania własności C jest równe: P(X|C) = P(C|X) * P(X) / P(C). Prawdopodobieństwa użyte we wzorze w łatwy sposób można obliczyć na podstawie danych. Sytuacja jest jednak bardziej skomplikowana, kiedy zmiennych (własności) jest więcej niż tylko 1. Ten problem jednak zdecydowano się 'obejść' poprzez założenie, że zmienne są niezależne (bez znaczenia jak sytuacja wygląda w praktyce). Dzięki temu, w przypadku użycia kilku zmiennych do prognozy, prawdopodobieństwo przynależności danej obserwacji do klasy X pod warunkiem posiadania własności C1, C2, .. Cn można obliczyć jako iloczyn prawdopodobieństw warunkowych, takich jak podano wcześniej, dla każdej zmiennej. Znacząco uprasza to rachunki, jednak założenie to praktycznie nigdy nie jest spełnione w rzeczywistości. Stąd słowo "naiwny" w nazwie metody. Jednak mimo tego faktu, metoda naiwna Bayesa potrafi dawać wyniki zbliżone, a czasami nawet lepsze niż inne, bardziej skomplikowane metody.
W ramach tej metody użyte mogą być jedynie zmienne kategoryczne, żeby był sens liczenia dla nich prawdopodobieństw. Na podstawie wykresów uwzględnionych w opisie danych zdecydowano o:

- pogrupowaniu zmiennych SibSp oraz Parch w 3 grupy oznaczające: 0, 1 i 2 lub więcej,
- pogrupowaniu zmiennej Fare w 4 grupy: pierwsza grupa - opłata mniejsza niż 7.91 (1 kwantyl), druga - między 7.91, a 14.45 (mediana), trzecia - między medianą, a 31 (3 kwantyl), czwarta - między 31, a 100 oraz piąta, czyli grupa osób które zapłaciły za rejs zdecydowanie więcej niż inni - > 100,
- pogrupowaniu zmiennej Age w grupy: 0-10, 11-28, 29-36, 36+. Jak widzieliśmy na wykresie wcześniej, wśród najmłodszych zdecydowanie inaczej rozkładały się proporcje jeśli chodzi o przeżycie katastrofy / zaginięcie w niej. Stąd 1 grupa to przedział 0-10, a kolejne wyznaczane są przez 2 i 3 kwantyl.

Code

Dla tak przygotowanych zmiennych zbudowano model i otrzymano takie oceny jego jakości:

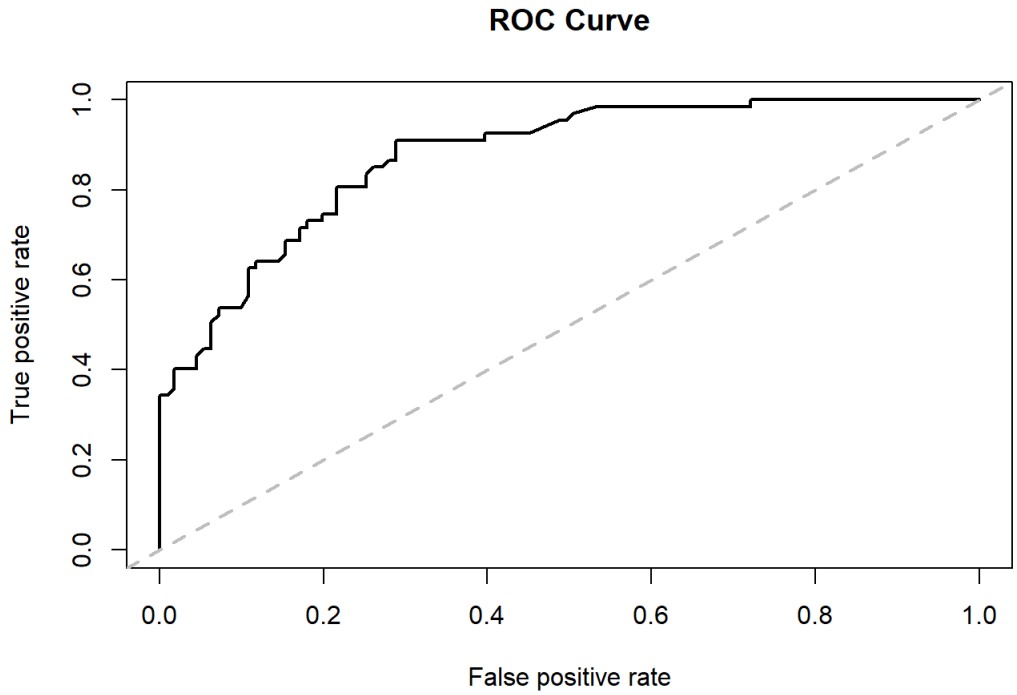
confusionMatrix:

Code

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction dead alive
##      dead    89    17
##      alive   22    50
##
##                Accuracy : 0.7809
##                  95% CI : (0.7129, 0.8393)
##     No Information Rate : 0.6236
##     P-Value [Acc > NIR] : 5.057e-06
##
##                   Kappa : 0.5401
##  Mcnemar's Test P-Value : 0.5218
##
##             Sensitivity : 0.7463
##             Specificity : 0.8018
##          Pos Pred Value : 0.6944
##          Neg Pred Value : 0.8396
##              Prevalence : 0.3764
##          Detection Rate : 0.2809
##    Detection Prevalence : 0.4045
##       Balanced Accuracy : 0.7740
##
##        'Positive' Class : alive
##
```

AUC i krzywa ROC:

Code



Code

```
## [[1]]
## [1] 0.875084
```

Widzimy, że podany model osiągnął skuteczność prognoz na poziomie 78% oraz AUC wynoszące około 0.875. Są to dość dobre wyniki, jednak troszeczkę gorsze od wcześniej użytych metod.

W celu sprawdzenia jakie czynniki wpływały na to jak klasyfikowano obserwacje przedstawione zostaną tablice zawierające

prawdopodobieństwa warunkowe, o których pisaliśmy wcześniej. Na ich podstawie postaramy się sprawdzić, co mogło wpływać na to czy dany pasażer przeżył katastrofę Titanica. Szukamy takich prawdopodobieństw, które będą dla wartości danej zmiennej będą charakteryzować się dużym ilorazem jeśli chodzi o prawdopodobieństwo dla grupy która przeżyła / prawdopodobieństwa dla grupy która nie przeżyła - będzie to oznaczać, że dla danej wartości zmiennej wystąpiła dysproporcja, wskazująca, że cecha pozytywnie wpływała na szanse na przeżycie katastrofy. Analogicznie, jeśli ten iloraz będzie mały to cecha negatywnie wpływała na te szanse. Będziemy szukać ilorazów w okolicach 2 (im większych tym lepiej) lub mniejszych od 1/2 (analogicznie - im mniejszy tym lepiej, bo to oznacza dużą dysproporcje w obrębie danej grupy obserwacji), ale uwzględnić należy również sensowność potencjalnych wniosków oraz to jak liczne były dane grupy - iloraz równy 2 da nam zarówno 0.04/0.02, jak i 0.5/0.25, lecz w pierwszym przypadku będzie on obliczony jedynie na podstawie 6% obserwacji ze zbioru, a w drugiej - 75%. Stąd im większe będą wartości prawdopodobieństw tym bardziej należy zwrócić uwagę na daną cechę, tym mniej przypadku w takich wynikach.

Pclass

	dead	alive
1st	0.15	0.39
2nd	0.17	0.27
3rd	0.67	0.34

Sex

	dead	alive
female	0.15	0.67
male	0.85	0.33

SibSp

	dead	alive
0	0.74	0.64
1	0.17	0.33
2/+	0.09	0.04

Age

	dead	alive
A	0.05	0.09
B	0.45	0.41
C	0.24	0.26
D	0.26	0.23

Parch

	dead	alive
0	0.82	0.69
1	0.09	0.19
2/+	0.09	0.12

Fare

	dead	alive
A	0.32	0.14
B	0.29	0.20
C	0.22	0.30
D	0.14	0.27
E	0.03	0.09

Embarked

	dead	alive
C	0.13	0.24
Q	0.09	0.09
S	0.79	0.67

Widzimy, że jeśli chodzi o zmienną Pclass, podróżujący 1szą klasą mieli zdecydowanie większe szanse na przeżycie, niż podróżujący gorszymi klasami. Jeśli ktoś nie przeżył katastrofy, to aż 67% spośród nich podróżowało 3 klasą.
Podobnie dużą różnicę widać jeśli chodzi o płeć - spośród ludzi, którzy zginęli, aż 85% to mężczyźni, podczas gdy aż 67% osób, które przeżyły to kobiety.
Gdy spojrzymy na zmienną Fare, widzimy, że wraz ze wzrostem opłaty (kolejne litery alfabetu to coraz większe opłaty) iloraz prawdopodobieństw jest coraz bardziej na korzyść kolumny alive - stąd wniosek, że im więcej ktoś zapłacił za rejs tym większe miał szanse na przeżycie katastrofy.
Widzimy również, że w przypadku zmiennej Age, jedynie w najmłodszej grupie wiekowej widać większą różnicę - prawdopodobieństwo, że ktoś należy do tej grupy pod warunkiem, że przeżył jest niemal dwukrotnie większe, niż że do niej należy po warunkiem, że nie przeżył, a pozostałe grupy wydają się nie wskazywać na żadną zależność - stąd w kolejnym modelu można wyciągnąć wniosek, że wiek wpływał na przeżywalność. W tym wypadku model wskazuje, że osoba z najmłodszej grupy miała większe szanse na przeżycie niż z jakiejś grupy starszej, co wydaje się bardzo logiczne - często to właśnie życie najmłodszych ratowane jest w pierwszej kolejności.

Porównanie modeli

Poniżej przedstawiono wartości metryk accuracy (dokładność) oraz AUC jakie osiągnęły poszczególne modele.

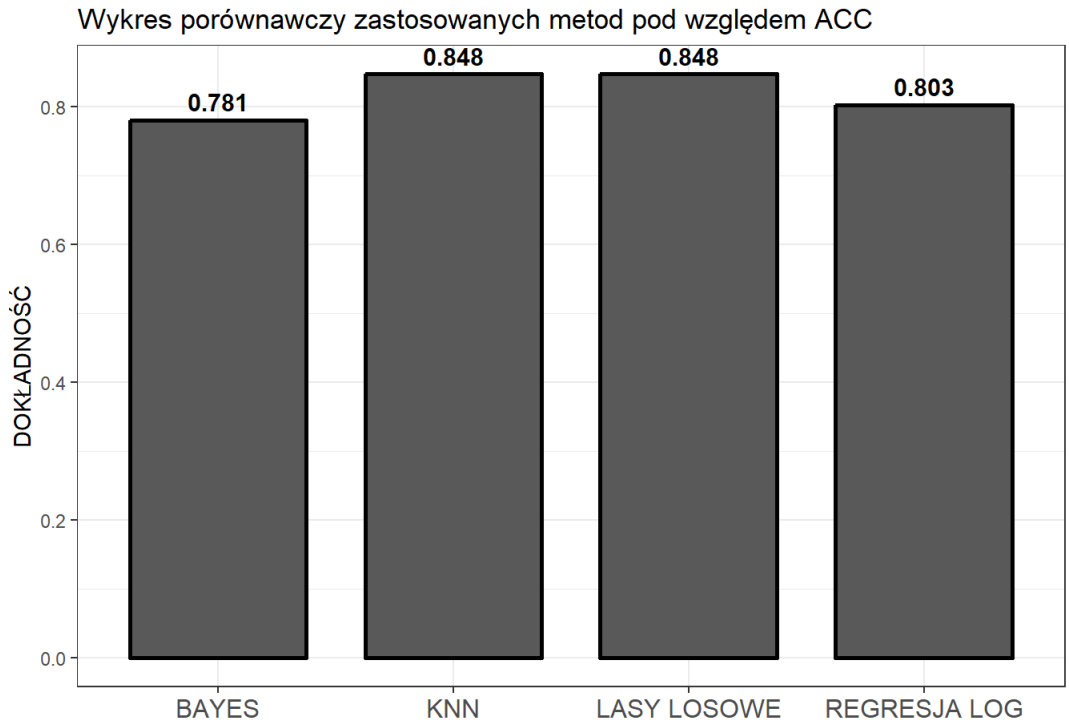
W formie tabelarycznej:

Code

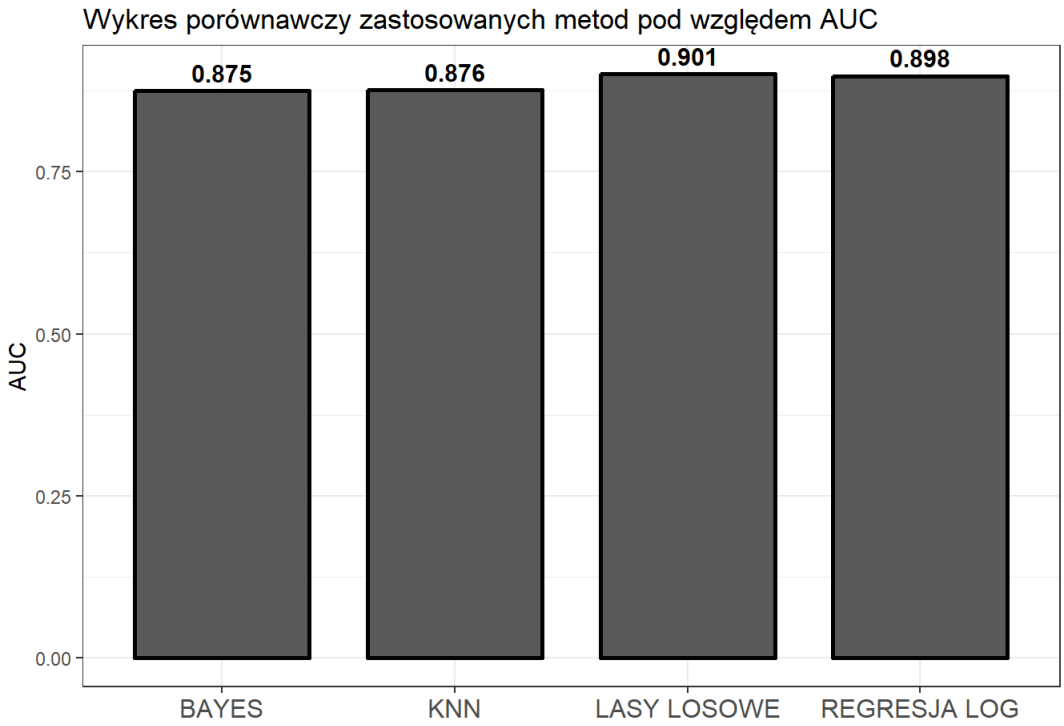
	AUC	ACC
REGRESJA LOG	0.898	0.803
KNN	0.876	0.848
LASY LOSOWE	0.901	0.848
BAYES	0.875	0.781

Oraz w formie graficznej:

Code



Code



Zarówno pod względem AUC jak i skuteczności predykcji najlepszym okazał się model oparty na algorytmie lasów losowych - jako jedyny osiągnął AUC większe od 0.9, a także skuteczność na poziomie 84%. Co zaskakujące równie dobrze spisał się pod tym kątem model KNN, lecz słabiej wypadł pod kątem AUC. Pozostałe modele (regresja logistyczna i naiwny klasyfikator Bayesa) wypadły trochę słabiej, jednak dostarczają one więcej informacji pod kątem analizy czynników wpływających na to czy pasażer przeżył katastrofę Titanica czy też nie. Jednak pod kątem prognozowania, w tym badaniu, jako najlepszy należy wskazać **model oparty na algorytmie lasów losowych** , choć model oparty na KNN okazał się być naprawdę nieznacznie gorszy.

Podsumowanie, wnioski, pomysły

Jednym z celów badania było sprawdzenie jakie czynniki wpływały na przeżywalność katastrofy. Na podstawie badania można wyciągnąć następujące wnioski:

- zdecydowanie większe szanse na przeżycie katastrofy miały kobiety, aniżeli mężczyźni,
- im gorszą klasą podróżował dany pasażer tym malały jego szanse na przeżycie katastrofy,
- w najmłodszej grupie wiekowej (mamy na myśli dzieci i młodzież) szanse na przeżycie były większe niż u pasażerów, którzy byli starsi od nich,
- wg modelu regresji logistycznej negatywnie na szanse przeżycie wpływał również fakt podróżowania razem z rodzeństwem lub małżonkiem / małżonką - możliwe, że osoby te ze względu na panującą na pokładzie panikę (lub jakieś sytuacje losowe) były rozdzielone i próbowały się odszukać, przez co traciły cenny czas i inni docierali do szalup przed nimi,
- wydaje się również, że wraz ze wzrostem wysokości opłat za rejs szanse na przeżycie były coraz większe - co brzmi dość sensownie, bo wydaje się, że Ci którzy zapłacili najwięcej mogli być potraktowani bardziej ‘priorytetowo’,
- zmienne Parch i Embarked wydają się nie mieć istotnego wpływu co jest zgodne z naszą intuicją - niezbyt widzimy powód, żeby port z którego podróż rozpoczął dany pasażer miał znaczenie w kontekście przeżywalności katastrofy. Z kolei jednak zmienna Parch mogłaby mieć analogiczne znaczenie co zmienna SibSp, jednak nie zostało to wykazane w żadnym modelu. Należy jednak pamiętać, że zmienna SibSp okazała się istotna tylko w przypadku modelu regresji logistycznej, stąd, na podstawie badania, jej znaczenie jest mniejsze niż pozostałych zmiennych,
- jak wspomniano we wcześniejszej części, pod kątem rezultatów klasyfikacji, najlepszym modelem okazał się model oparty na algorytmie lasów losowych.

W ramach projektu przygotowaliśmy dane i stworzyliśmy w oparciu o nie 4 modele. Podczas wykonywania tych czynności napotkaliśmy na różne problemy, które opisaliśmy w ramach tego referatu i przedstawiliśmy nasze pomysły na rozwiązanie sytuacji. Jednak nie są to oczywiście jedyne możliwe rozwiązania. Patrząc z perspektywy na przeprowadzone badanie udało nam się wpaść na kilka pomysłów, co można byłoby poprawić, gdzie spróbować czegoś innego, jakie inne rozwiązania mogłyby (lub nie) sprawdzić się lepiej. Oto nasze wnioski i pomysły:

- skoro zaskakująco dobrze zadziałał model KNN, być może jego rozbudowana wersja, czyli ważna metoda KNN dałaby jeszcze lepsze rezultaty,
- analogicznie, w przypadku regresji liniowej zastosować można było inne metody wyboru zmiennych objaśniających do modelu - jak

chociażby metoda Hellwiga czy metoda krokowa wprzód lub inne,

- jeśli chodzi o model lasów losowych, pole do popisu wydaje się jeszcze większe, szczególnie ze względu na możliwość szukania najlepszych parametrów- czy to dot. pojedynczych drzew tworzących model czy to ogólnych parametrów modelu. Wykorzystaliśmy wartości domyślne w większości parametrów, ale próba zbudowania wielu modeli i sprawdzenia jakie parametry będą najlepsze pozwoliłaby prawdopodobnie dodatkowo poprawić skuteczność modelu,
- w przypadku modelu Bayesa, możliwe, że zmienne numeryczne należało pogrupować w inny sposób i to poprawiłoby wyniki - z perspektywy czasu zastanawia czy skoro dla najmłodszej grupy widać było pewną różnicą jeśli chodzi o szanse na przeżycie, to być może dla grupy osób najstarszych również taka różnica występowała i należało wydzielić najstarszą grupę od wyższego wieku niż 36 lat,
- zbiór danych na jakim pracowaliśmy zawierał 891 obserwacji - nie jest to zbiór mały, ale liczba obserwacji nie jest też jakoś bardzo duża. Stąd być może sensowne byłoby skorzystanie z walidacji krzyżowej w celu otrzymania wyników bardziej miarodajnych i wiarygodnych. Zdecydowaliśmy jednak, że zbiór jest na tyle liczny, że 'tradycyjny' podział zbioru (poparty sprawdzeniem czy proporcje w zbiorach są zachowane) będzie wystarczająco dobry,
- w kontekście skuteczności modeli, możliwe, że przy użyciu innego punktu odcięcia niż standardowe 0.5 wyniki byłyby lepsze - w celu jego wyznaczenia można byłoby użyć np. kryterium Youden'a,
- rzecz jasna uwzględnienie w badaniu kolejnych modeli umożliwiłoby jeszcze dokładniejszą analizę i być może pojawiłyby się nowe, ciekawe wnioski dot. czynników wpływających na przeżywalność katastrofy Titanica,
- po stworzeniu klasyfikatora Bayesowskiego zastanawialiśmy się również czy transformacja zmiennych Parch i SibSp tak jak na potrzeby tej metody, nie byłaby korzystna również w kontekście pozostałych metod. Być może również któraś ze zmiennych, jakie usunęliśmy ze zbioru na etapie przygotowania danych miała cenne informacje o których nie pomyśleliśmy i należało je w jakiś sposób wykorzystać - czy to w aktualnej czy w zmienionej formie. Może też uzupełnieni braków w zmiennej Cabin jednak nie byłoby złym pomysłem, mimo aż 80% deficytu.