Obraz zawierający Czcionka, tekst, logo, symbol

Opis wygenerowany automatycznie

**Wydział Zarządzania**

Projekt zaliczeniowy 2

*Regresja logistyczna i redukcja wielowymiarowości dla danych z przeżywalności katastrofy RMS Titanic.*

Przedmiot: **Statystyczna Analiza Danych**

Autor: Wiktoria Arendarczyk (412133)

Kierunek studiów: Informatyka i Ekonometria

Spis treści

[1. Wstęp 4](#_Toc156555217)

[2. Opis danych 4](#_Toc156555218)

[Źródło i krótkie objaśnienie 4](#_Toc156555219)

[Charakterystyka danych 4](#_Toc156555220)

[3. Metodyka 6](#_Toc156555221)

[ANOVA 6](#_Toc156555222)

[Redukcja wielowymiarowości 6](#_Toc156555223)

[Regresja logistyczna 7](#_Toc156555224)

[Poziom istotności i P-value 7](#_Toc156555225)

[4. ANOVA 8](#_Toc156555226)

[Testowanie założeń 8](#_Toc156555227)

[ANOVA Welcha 8](#_Toc156555228)

[5. Redukcja wielowymiarowości 9](#_Toc156555229)

[Testowanie założeń 9](#_Toc156555230)

[Analiza głównych składowych (PCA – Principal Component Analysis) 10](#_Toc156555231)

[6. Regresja logistyczna 12](#_Toc156555232)

[Postać modelu 12](#_Toc156555233)

[Interpretacja współczynników modelu 13](#_Toc156555234)

[Diagnostyka modelu na danych treningowych 14](#_Toc156555235)

[Diagnostyka modelu na danych testowych 15](#_Toc156555236)

[Odpowiedź na pytanie badawcze 16](#_Toc156555237)

[7. Zakończenie 16](#_Toc156555238)

[Bibliography 16](#_Toc156555239)

Spis wykresów

[Wykres 1 Histogram (wraz z gęstością) wieku 5](file:///C:\Users\Admin\Desktop\Projekt_SAD_2_Arendarczyk\Projekt_SAD_2_Arendarczyk.docx#_Toc156555197)

[Wykres 2 Procent przeżycia katastrofy w zależności od klasy i płci 5](file:///C:\Users\Admin\Desktop\Projekt_SAD_2_Arendarczyk\Projekt_SAD_2_Arendarczyk.docx#_Toc156555198)

[Wykres 3 Wykres pudełkowy wieku w zależności od klasy 6](file:///C:\Users\Admin\Desktop\Projekt_SAD_2_Arendarczyk\Projekt_SAD_2_Arendarczyk.docx#_Toc156555199)

[Wykres 4 Macierz korelacji dla zmiennych (Spearman) 9](file:///C:\Users\Admin\Desktop\Projekt_SAD_2_Arendarczyk\Projekt_SAD_2_Arendarczyk.docx#_Toc156555200)

[Wykres 5 Wykres osypiska 10](file:///C:\Users\Admin\Desktop\Projekt_SAD_2_Arendarczyk\Projekt_SAD_2_Arendarczyk.docx#_Toc156555201)

[Wykres 6 Wykres kolumnowy procentu wyjaśnionej wariancji 10](file:///C:\Users\Admin\Desktop\Projekt_SAD_2_Arendarczyk\Projekt_SAD_2_Arendarczyk.docx#_Toc156555202)

[Wykres 7 Wykres ładunków czynnikowych 11](file:///C:\Users\Admin\Desktop\Projekt_SAD_2_Arendarczyk\Projekt_SAD_2_Arendarczyk.docx#_Toc156555203)

[Wykres 8 Macierz konfuzji 15](file:///C:\Users\Admin\Desktop\Projekt_SAD_2_Arendarczyk\Projekt_SAD_2_Arendarczyk.docx#_Toc156555204)

[Wykres 9 Krzywa ROC 15](file:///C:\Users\Admin\Desktop\Projekt_SAD_2_Arendarczyk\Projekt_SAD_2_Arendarczyk.docx#_Toc156555205)

# Wstęp

Celem projektu jest próba dopasowania modelu regresji logistycznej do badanych danych w celu określenia zależności między zmiennymi. Przeprowadzone zostaną odpowiednie testy statystyczne oraz sprawdzę, czy mój model prognozuje moje przeżycie w katastrofie RMS Titanic.

Dodatkowo przeprowadzona zostanie jednoczynnikowa *ANOVA*, która sprawdzi, czy średni wiek pasażerów różnych klas różni się istotnie. Zastosowana zostanie redukcja wielowymiarowości *(PCA)* w celu spojrzenia na dane z innej perspektywy.

# Opis danych

## Źródło i krótkie objaśnienie

Dane o Titanicu są dostępne w wielu źródłach, a nawet i w pakietach R. Moje dane pochodzą ze strony kaggle. [[1]](#footnote-1) Titanic był brytyjskim transatlantyckim okrętem, który podczas swojego dziewiczego rejsu na trasie Wielka Brytania (Southampton) – Stany Zjednoczone (Nowy Jork) zderzył się z górą lodową w nocy z 14 na 15 kwietnia 1912 roku. Spośród 2208 pasażerów i załogi, w katastrofie zginęło 1496 osób. Tak wysoka liczba ofiar spowodowana była - poza rzecz jasna przedziurawieniem kadłuba i zatonięciem statku - niewystarczającą liczbą szalup oraz ich niewłaściwym załadunkiem i powolnym opuszczaniem. Ewakuacja przebiegała hierarchicznie od najwyższych klas i kobiet oraz dzieci, do najniższych klas i mężczyzn. Niska temperatura wody i powietrza i późne przybycie na miejsce zdarzenia jednostek odpowiadających na sygnał SOS spowodowały, że ktokolwiek kto znalazł się w wodzie miał niskie szanse na przeżycie do przypłynięcia pomocy. Pierwszy statek przybył z pomocą dla podróżnych około dwie godziny po zatonięciu Titanica.

## Charakterystyka danych

Dane przeze mnie badane posiadają poniższe zmienne:

* **Survival** – Informacja binarna na temat tego, czy pasażer przeżył,
* **Pclass** – Klasa biletu (1, 2, 3), gdzie najwyższa klasa to klasa pierwsza,
* **Sex** – Płeć pasażera,
* **Age** – Wiek pasażer w latach,
* **Sibsp** – Liczba rodzeństwa / małżonków podróżującego na pokładzie Titanica,
* **Parch** – Liczba rodziców / dzieci podróżującego na pokładzie Titanica,
* **Ticket** – Numer biletu (numer identyfikacyjny pasażera),
* **Fare** – Cena biletu,
* **Obraz zawierający tekst, zrzut ekranu, diagram, Prostokąt

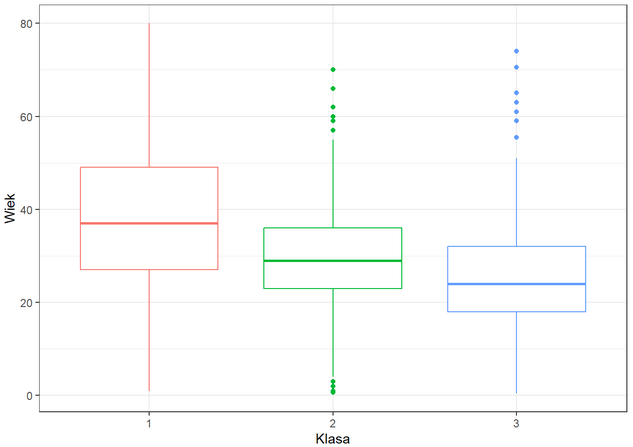
  Opis wygenerowany automatycznie**Obraz zawierający diagram, Wykres, zrzut ekranu, tekst

  Opis wygenerowany automatycznie**Embarked** – Port wypłynięcia (Cherbourg, Queenstown, Southampton).

Wykres 1 Histogram (wraz z gęstością) wieku

Wykres 2 Procent przeżycia katastrofy w zależności od klasy i płci

Liczba pasażerów w zestawie danych to 891. Wszystkie dane były kompletne poza wiekiem, gdzie brakowało około 20% obserwacji. Dane nie były *„Missing Completely at Random”*, gdyż większość braków koncentrowała się w najuboższej grupie. Z tego też powodu postanowiłam usunąć te obserwacje, gdyż nie miałam wystarczających informacji do przeprowadzenia bezpiecznej imputacji. Finalna liczba pasażerów to 712 osób.

Dodatkowo, zmienna *Age* jest ujemnie skorelowana z *Pclass*, to znaczy, że wiek, w jakimś stopniu, zależy od klasy pasażera. Jest to logiczne, gdyż bilet klasy pierwszej kosztował znacznie więcej niż bilet klasy trzeciej, więc można wnioskować, że na drogie bilety mogli pozwolić sobie głównie ludzie zamożni i starsi. Dla klasy pierwszej średnia wynosiła 38 lat, dla klasy drugiej 30 lat, dla klasy trzeciej 25 lat. Dla tego fragmentu danych przeprowadzona zostanie ANOVA, która sprawdzi, czy średnie w grupach różnią się od siebie istotnie.

*Kreska pozioma w środku pudełka* – mediana;

*Pudełko* – 1 i 3 kwartyl (rozstęp ćwiartkowy IQR);

*Wąsy* – wartości do 1.5 IQR;

*Kropki* – wartości odstające.

*Kreska pozioma w środku pudełka* – mediana;

*Pudełko* – 1 i 3 kwartyl (rozstęp ćwiartkowy IQR);

*Wąsy* – wartości do 1.5 IQR;

*Kropki* – wartości odstające.

Wykres 3 Wykres pudełkowy wieku w zależności od klasy

Jeśli chodzi o ogólne statystyki to około 60% pasażerów nie przeżyła podróży, a około 40% pasażerów przeżyła podróż. Największy procent przeżycia miały kobiety z pierwszej klasy – najmniejszy zaś mężczyźni z klasy trzeciej.

# Metodyka

## ANOVA

Analiza wariancji to kolekcja modeli statystycznych, która ogólnie odpowiada na pytanie jak zmienne objaśniające kategoryczne wpływają na średnie zmiennej objaśnianej w danych grupach. W moim przypadku przeprowadzona zostanie jednoczynnikowa ANOVA, czyli analiza przeprowadzona dla jednego niezależnego czynnika grupującego. Model ten bada wpływ jednej zmiennej objaśniającej kategorycznej (tu: klasa biletu) na średnią zmiennej objaśnianej (tu: wiek) w danych grupach. Założenia, które powinny być spełnione dla ANOVY to rozkład normalny oraz równość wariancji w grupach.

## Redukcja wielowymiarowości

Jest to proces zmniejszania liczby badanych zmiennych głównie w celu ich kompresji oraz przyspieszenia obliczeń. Najczęściej stosowaną metodą jest *Principal Component Analysis* (PCA), którą stosować można do zmiennych ilościowych oraz porządkowych możliwych do przybliżenia jako ilościowe (interwałowe). Dla przykładu, dane, które określają trzy zmienne ciągłe można zrzutować do przestrzeni dwuwymiarowej, którą już wizualizuje się prościej.

## Regresja logistyczna

Regresja służy do budowania funkcji, która w sposób matematyczny określa relacje zachodzące między zmienną objaśnianą, a zmiennymi objaśniającymi. Cechą regresji jest możliwość predykcji wartości zmiennej objaśnianej za pomocą znanych wartości zmiennych objaśniających – niemniej jednak nie każdy model daje wyniki zgodne z rzeczywistością.

Regresja logistyczna to jedna z metod regresji używana w statystyce w przypadku, gdy zmienna objaśniana przyjmuje wartości na skali dychotomicznej – czyli tylko dwie wartości, zazwyczaj 0-1. Ogólny wzór na regresję logistyczną można przedstawić następująco:

P(x) – prawdopodobieństwo zdarzenia, βi – wartość parametru dla i-tej zmiennej, Xi – wartość i-tej zmiennej.

Można zauważyć, interpretacja parametrów modelu różnić się będzie od interpretacji parametrów najprostszej regresji liniowej, której równanie występuje w środku funkcji logistycznej. Z interpretacją parametrów modelu wiąże się pojęcie *odds* [pl. szanse] i *odds ratio* [pl. iloraz szans] – jest to stosunek szans wystąpienia danego zdarzenia w jednej grupie do szansy jego wystąpienia w innej grupie. Jeśli chcemy interpretować wartość parametru, to musimy obliczyć , gdzie to współczynnik parametru. Wynik zwróci iloraz szans, który można już zinterpretować zakładając zasadę *ceteris paribus*, która mówi, że pozostałe czynniki pozostają bez zmian.

## Poziom istotności i P-value

W analizach przyjęty zostanie standardowy poziom istotności na poziomie . Oznacza to, że prawdopodobieństwo popełnienia błędu pierwszego rodzaju (uznania prawdziwej hipotezy zerowej za fałszywą) wynosi 5%.

P-value to prawdopodobieństwo uzyskania rezultatów przynajmniej takich jak zaobserwowane zakładając, że hipoteza zerowa jest prawdziwa i tylko przypadek jest tego powodem. Przyjmuje ono wartości od 0 do 1 i jest towarzyszem każdego testu statystycznego. Im mniejsze p-value tym bardziej niestandardowe byłyby dane zakładając, że każde założenie jest prawdziwe – w tym założenie prawdziwości hipotezy zerowej. Jeżeli spełnione są założenia danego testu, to mała wartość p-value (mniejsza od przyjętego poziomu istotności) sugeruje, że istnieją przesłanki do odrzucenia hipotezy zerowej.

# ANOVA

## Testowanie założeń

Obraz zawierający tekst, Czcionka, zrzut ekranu, numer

Opis wygenerowany automatycznieObraz zawierający tekst, paragon, Czcionka, zrzut ekranu

Opis wygenerowany automatycznieW celu przeprowadzenia ANOVY zacząć należy od sprawdzenia założeń. Do sprawdzenia normalności rozkładu użyję *testu Shapiro-Wilka*, którego hipoteza zerowa stanowi, że rozkład jest normalny. Do sprawdzenia równości wariancji w grupach użyję *testu Levene’a*, którego hipoteza zerowa stanowi, że wariancje w grupach są równe.

Tabela 1 Test Levene'a dla ANOVY

Tabela 1 Test Levene'a dla ANOVY

Tabela 2 Test Shapiro-Wilka dla ANOVY

Tabela 2 Test Shapiro-Wilka dla ANOVY

Oba testy sugerują odrzucenie hipotezy zerowej, gdyż p-value jest bliskie zera. Oznacza to, że najprawdopodobniej zmienna *Age* nie ma rozkładu normalnego. Dodatkowo, *test homogeniczności wariancji Levene’a* sugeruje, że wariancje nie są stałe między grupami.

## ANOVA Welcha

Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznieW tym przypadku najlepszym rozwiązaniem będzie użycie *ANOVY Welcha*, która nie wymaga spełnienia założenia równości wariancji między grupami oraz równości wielkości grup. Dodatkowo, z racji, że wielkości grup są wystarczająco duże, naruszenie założenia normalności nie powinno stać na przeszkodzie do przeprowadzenia *ANOVY Welcha*, ale należy mieć to na uwadze.

Tabela 3 Wyniki dla ANOVY Welcha

Tabela 3 Wyniki dla ANOVY Welcha

Tabela 4 Test post-hoc Gamesa-Howella dla ANOVY

Tabela 4 Test post-hoc Gamesa-Howella dla ANOVY

*Obraz zawierający tekst, Czcionka, linia, zrzut ekranu

Opis wygenerowany automatycznieANOVA Welcha* zwróciła p-value na poziomie bliskim zera, co sugeruje, że należy odrzucić hipotezę zerową stanowiącą, że średnie w grupach są równe. Przeprowadzony test *post-hoc Gamesa-Howella* (który nie wymaga homogeniczności wariancji), sugeruje, że wszystkie średnie różnią się od siebie istotnie. **Biorąc pod uwagę wykres pudełkowy, który został wcześniej wyrysowany, wyniki *ANOVY Welcha* oraz testów post-hoc można stwierdzić, że rzeczywiście średnie lat w grupach klasy biletowej różnią się od siebie istotnie.**

# 5. Redukcja wielowymiarowości

## Testowanie założeń

Obraz zawierający tekst, zrzut ekranu, kwadrat, diagram

Opis wygenerowany automatycznieObraz zawierający tekst, Czcionka, linia, numer

Opis wygenerowany automatycznieObraz zawierający tekst, zrzut ekranu, paragon, Czcionka

Opis wygenerowany automatyczniePierwszym założeniem niezbędnym do spełnienia jest to, że dane są skorelowane na tyle, że macierz korelacji różni się istotnie od macierzy jednostkowej. Posłuży do tego *test sferyczności Bartletta*, którego hipoteza zerowa stanowi, że macierz korelacji jest macierzą jednostkową. Odrzucenie hipotezy zerowej sugeruje, że zastosowanie *analizy głównych składowych* (PCA) zwróci sensowne wyniki. Dodatkowo policzone zostaną *współczynniki KMO* (Kaiser-Meyer-Olkin), które przyjmują wartości od 0 do 1 dla każdej zmiennej. Współczynniki te wyrażają siłę korelacji cząstkowej pomiędzy zmiennymi. Generalnie przyjmuje się, że wyniki powyżej 0.5 są akceptowalne. (Vafakhah & Kheirfam, 2015)

Wykres 4 Macierz korelacji dla zmiennych (Spearman)

Tabela 5 Test Bartletta dla PCA

Tabela 5 Test Bartletta dla PCA

Tabela 6 Współczynniki KMO dla PCA

Tabela 6 Współczynniki KMO dla PCA

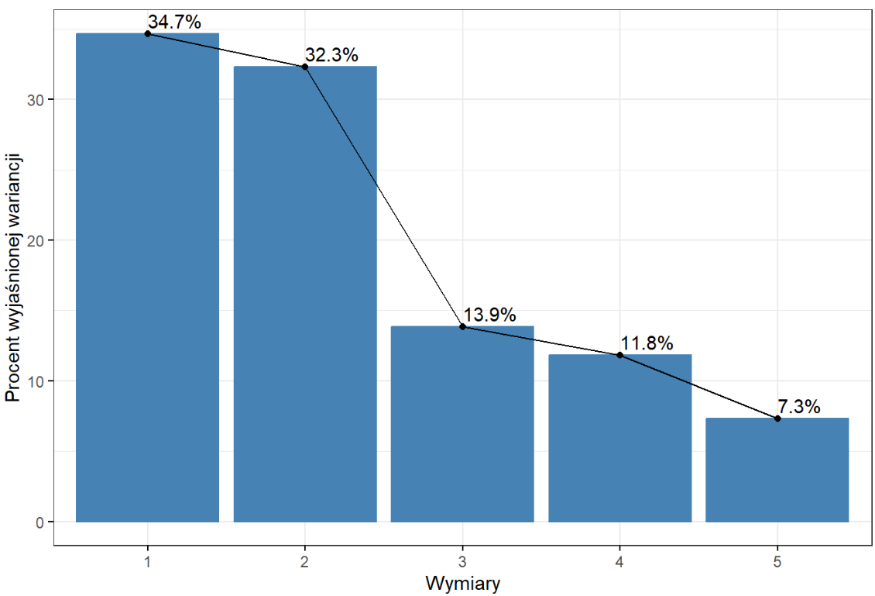
*Test sferyczności* zwrócił p-value bliskie zera, co znaczy, że istnieją przesłanki do odrzucenia hipotezy zerowej. *Współczynniki KMO* są powyżej wartości 0.5, niemniej jednak tylko nieznacznie, co oznacza, że PCA może nie dawać bardzo dobrych wyników, ale można ją przeprowadzić. Macierz korelacji zbudowana została za pomocą metody *Spearmana*, gdyż lepiej radzi sobie ona z obecnością zmiennych ciągłych na skali dyskretnej.

Na podstawie wyników testów oraz samej macierzy korelacji można stwierdzić, że założenia do przeprowadzenia *analizy głównych składowych* są spełnione.

## Analiza głównych składowych (PCA – Principal Component Analysis)

Redukcja przeprowadzona została na pięciu zmiennych: *Age*, *Fare*, *Pclass*, *SibSp* i *Parch*. W PCA poszukujemy jak najmniej komponentów, które wyjaśniają jak najwięcej wariancji danych.

Obraz zawierający diagram, linia, Wykres, stok

Opis wygenerowany automatycznieW przypadku naszych danych już dwa wymiary (składowe) wyjaśniają około 67% wariancji i widać wyraźny skok między nimi, a następną składową (Rys.5). Oznacza to, że pierwsze dwa komponenty mogą reprezentować dane, aczkolwiek ta reprezentacja nie będzie tak dobra, jak gdyby skumulowany procent wariancji wynosił około 90%. Niemniej jednak wykres osypiska (Rys. 6) również sugeruje użycie tylko dwóch składowych, co akcentowane jest charakterystycznym spadkiem, także do dalszych rozważań przyjmę użycie dwóch składowych.

Wykres 5 Wykres osypiska

Wykres 6 Wykres kolumnowy procentu wyjaśnionej wariancji

Przed wstępną interpretacją pragnę zaznaczyć, że redukcja wielowymiarowa powoduje utratę jasności interpretacji nowo powstałych składowych. **Składowe da się opisać w sposób analizy ich wewnętrznej struktury, jednakże nie są one tak naprawdę interpretowalne.**

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie*Tabela 7* prezentuje ładunki składowych dla wariantu dwukomponentowego, co pozwala na względne ich opisanie. Komponent pierwszy ma duże pozytywne wartości dla zmiennej *Age* i *Fare*, i małe negatywne wartości dla pozostałych zmiennych, w tym *Pclass*. Komponent drugi notuje duże pozytywne wartości dla trzech ostatnich zmiennych – *Fare*, *Sibsp* i *Parch*.

Tabela 7 Ładunki składowych (wartości wektora własnego) dla PCA

Tabela 7 Ładunki składowych (wartości wektora własnego) dla PCA

Obraz zawierający tekst, diagram, linia, Wykres

Opis wygenerowany automatycznieIm większe wartości, tym większy udział w wyjaśnianiu danej składowej przez daną zmienną. Różne znaki mówią o różnym kierunku oddziaływania danej zmiennej na składową.

Wykres 7 Wykres ładunków czynnikowych

*Współrzędne końca wektora –* odpowiadające ładunki czynnikowe zmiennych*,*

*Długość wektora –* zasób informacyjny zmiennej pierwotnej jaki niosą składowe główne wyznaczające układ współrzędnych,

*Znak współrzędnych końca wektora* – wskazuje na dodatnią lub ujemną korelację zmiennej pierwotnej i składowych głównych tworzących układ współrzędnych,

*Kąt między wektorami* – skorelowanie zmiennych pierwotnych.

*Współrzędne końca wektora –* odpowiadające ładunki czynnikowe zmiennych*,*

*Długość wektora –* zasób informacyjny zmiennej pierwotnej jaki niosą składowe główne wyznaczające układ współrzędnych,

*Znak współrzędnych końca wektora* – wskazuje na dodatnią lub ujemną korelację zmiennej pierwotnej i składowych głównych tworzących układ współrzędnych,

*Kąt między wektorami* – skorelowanie zmiennych pierwotnych.

Wektory zostały pokolorowane w zależności od stopnia wyjaśnienia zmiennej przez dwie składowe – im większa wartość cos2, tym bliżej wektor jest koła korelacji i tym wyższy stopień wyjaśnienia, gdzie maksymalna wartość wynosi 1.

Można zauważyć, że zmienne *Pclass* oraz *Fare* są wyjaśniane w najwyższym stopniu, a *Age* oraz *Sibsp* w najmniejszym. Oznacza to, że jeśli zależałoby nam najbardziej na utrzymaniu wysokiego poziomu wyjaśnienia zmienności dla *Pclass* i *Fare*, to redukcja do dwóch składowych byłaby dobrym posunięciem. Jeśli natomiast zależałoby nam najbardziej na zmiennej *Age*, to wypadałoby rozważyć redukcję do większej ilości składowych – np. trzech, gdyż sama trzecia składowa wyjaśnia *Age* w około 31%.

# 6. Regresja logistyczna

## Postać modelu

Wybrany przeze mnie model regresji logistycznej przyjmuje *Age*, *Sex*, *Pclass* i *SibsSp* jako zmienne objaśniające zmienną *Survived*. Dla tego modelu będę prowadzić dalsze wnioskowanie (diagnostyka, poprawność predykcji na zbiorze testowym).

Został on przeze mnie wybrany jako najlepszy kandydat po zapoznaniu się z danymi. Brałam pod uwagę korelacje zmiennych między sobą oraz ich relacje z zmienną Survived, którą badamy. Na podstawie wstępnej analizy danych można było zauważyć zależność między przeżyciem pasażera, a jego klasą biletu (*Pclass*) i płcią (*Sex*) [Rys. 2]. Jest to sytuacja logiczna, gdyż na Titanicu najpierw ewakuowano kobiety oraz dzieci z każdej z klas, poczynając od klasy pierwszej. Dodatkowo brać pod uwagę można wiek pasażera, gdyż dzieci i osoby młodsze niezależnie od płci i klasy miały większą szansę na przeżycie od osób starszych.

Zmienna *Fare* ma wysoką korelację z *Pclass* [Rys. 4], więc w celu uniknięcia współliniowości zostanie ona pominięta. Dodatkowo, patrząc na nią z punktu logicznego – im większa cena biletu tym lepsze zakwaterowanie w klasie pierwszej. Nie dawało to jednak gwarancji na przeżycie, gdyż ewakuacja klasy pierwszej była prowadzona niezależnie od ceny biletu – wołani byli wszyscy pasażerowie.

W celu dalszej redukcji mało wnoszących zmiennych przeprowadzona została *Backward Stepwise Regression (BSR)* na zbiorze treningowym (571 obserwacji), która ustala zmienne dające model o najlepszym wskaźniku AIC. Redukcja liczebności zmiennych objaśniających jest potrzebna, gdyż zmniejsza problem współliniowości oraz *overfitting*, czyli przeuczenie modelu na danych treningowych.

Tabela 8 Kroki przeprowadzone przez BSR, gdzie wartości pod zmiennymi to wartości p-value

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Wybór | *Age* | *SibSp* | *Parch* | *Sex\_male* | *Pclass\_21* | *Pclass\_31* | *Embarked\_Q* | *Embarked\_S* |
| BSR | \*\*\* | 0.003 | **X** | \*\*\* | \*\*\* | \*\*\* | 0.06 | 0.02 |
| Mój | \*\*\* | 0.001 | **X** | \*\*\* | \*\*\* | \*\*\* | **X** | **X** |

Postanowiłam wybrać model zredukowany (bez zmiennych Embarked), gdyż ich dodanie lub usunięcie powoduje jedynie nieznaczną zmianę AIC i pseudo R2 oraz modele miały podobne zdolności predykcyjne (macierz konfuzji różniła się tylko 1 obserwacją na korzyść modelu z Embarked). Dodatkowo, port startowy ma mały wkład interpretacyjny i logiczny, gdyż z historycznego punktu widzenia nie było znaczenia z jakiego portu wsiadało się na Titanic.

## Interpretacja współczynników modelu

Pierwotnie, współczynniki regresji logistycznej podane są w *log-odds. Log-odds* to logarytm z *odds ratio*, a pojęcie to zostało już przybliżone w części metodyki. Wartości *odds ratio* są łatwiej interpretowalne, dlatego zostały przeliczone zgodnie ze wzorem podanym w części metodycznej.

Błąd standardowy dla parametrów oraz przedziały ufności mówią jakie wartości mogą przyjmować estymatory, gdyby faktycznie określały całą populację problemu, a nie tylko próbę.

Tabela 9 Współczynniki dla parametrów modelu [odds ratio]

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Zmienna | *(Intercept)* | *Age* | *SibSp* | *Sex\_male* | *Pclass\_21* | *Pclass\_31* |
| Wartość  (95% CI) | 66.68  (24.53 - 181.27) | 0.96  (0.94 – 0.98) | 0.65  (0.5 - 0.85) | 0.09  (0.06 - 0.14) | 0.25  (0.13 - 0.48) | 0.07  (0.04 - 0.14) |
| Błąd standardowy | 33.24 | 0.01 | 0.09 | 0.02 | 0.08 | 0.02 |

Punkt przecięcia (intercept) interpretowalny jest wtedy, gdy wszystkie inne zmienne w modelu przyjmują wartość zerową – można wyliczyć wtedy bezpośrednie prawdopodobieństwo przeżycia, która wynosi około 0.99. Oznacza to, że dla noworodka płci żeńskiej, bez rodzeństwa, klasy pierwszej prawdopodobieństwo przeżycia katastrofy wynosi 99%.

W regresji logistycznej, współczynnik dla danej zmiennej wyraża zmianę w *odds* [pl. szansa] zmiennej objaśnianej w związku z zmianą wartości o jedną jednostkę przez zmienną objaśniającą, zakładając zasadę *ceteris paribus*. Dla przykładu, wzrost o jednostkę w *Age* spowoduje pomnożenie dotychczasowych szans przez 0.96, czyli spadek o 4%. W przypadku *odds* *ratio* w celu określenia wielkości efektu należy zwracać uwagę na odległość wartości od jedynki. Dla przykładu, gdyby pasażer był z klasy trzeciej, to wartość dotychczasowych *odds* zmniejszy się o 93%. Zauważyć można, że istnieje silna ujemna zależność pomiędzy byciem mężczyzną oraz osobą z klasy trzeciej, a przeżywalnością.

Model zbudowany jest w taki sposób, że wszystkie zmiany zmiennych binarnych (z 0 na 1) oraz wzrosty zmiennych ciągłych spowodują spadek bazowych szans.

## Diagnostyka modelu na danych treningowych

W regresji logistycznej sprawdzić należy współliniowość zmiennych, którą można przybliżać pomocą wartości *Variance Inflation Factor* (VIF).

VIF mierzy, jak bardzo wariancja oszacowanego współczynnika regresji jest zwiększona z powodu występowania współliniowości w modelu. Oblicza się go dla każdej zmiennej predykcyjnej, regresując ją na wszystkie inne zmienne predykcyjne. Gdy wartość w danej zmiennej przekracza 5, można mówić o potencjalnym zagrożeniu współliniowości.

Tabela 10 Wartości VIF dla zmiennych

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Zmienna | *Age* | *SibSp* | *Sex\_male* | *Pclass\_21* | *Pclass\_31* |
| Wartość | 1.48 | 1.17 | 1.09 | 1.8 | 2.14 |

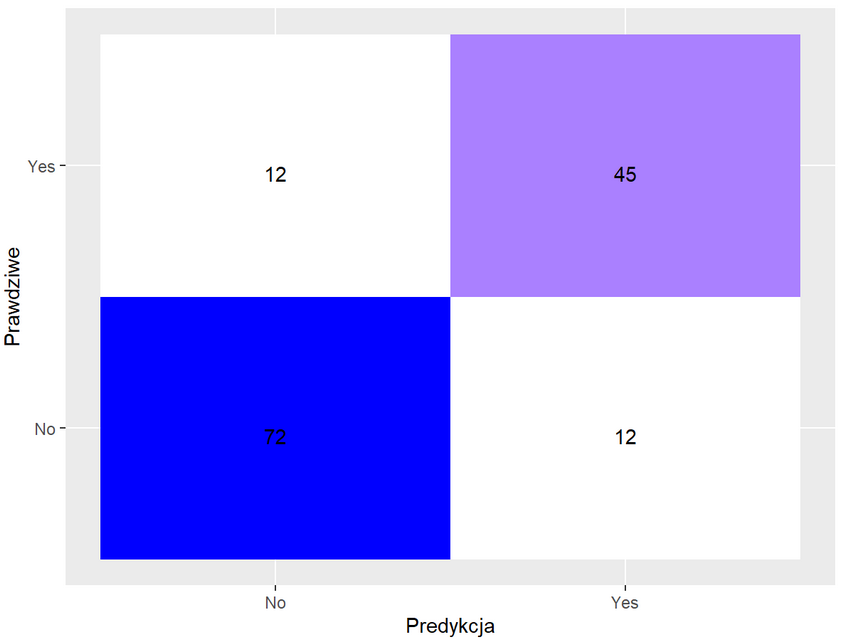
Pierwiastek z danej wartości VIF wskazuje o ile zwiększa się błąd standardowy w porównaniu, gdyby zmienna miała zero korelacji z innymi zmiennymi objaśniającymi. Dla naszych danych można stwierdzić, że współliniowość nie będzie stanowiła dużego problemu.

Policzone zostały dwa wskaźniki Pseudo R2 – McFaddena (0.318), Cox&Snella (0.349) oraz Nagelkerke (0.471).

*Pseudo R2 McFaddena* przyjmuje wartości od wartości nieco mniejszych od 0 do 1, aczkolwiek wartość 0.2 – 0.4 przyjmowana jest już często za dobre dopasowanie modelu (Lee, 2013). *Pseudo R2 McFaddena* wskazuje jaka proporcja zmienności wyjaśniana jest przez model w stosunku do modelu bez zmiennych.

*Pseudo R2 Cox&Snella* przyjmuje wartości podobne do *Pseudo R2 McFaddena*, ale jego wartość maksymalna jest mniejsza od 1 i tak naprawdę zależy od danych. Jest to ciężkie w interpretacji, toteż wprowadzona została korekcja tego podejścia w *Pseudo R2 Nagelkerke*. Oba te podejścia wskazują na różnice między danym modelem, a modelem bez zmiennych – tak jak w *Pseudo R2 McFaddena* – aczkolwiek osiągane jest to za pomocą innych obliczeń.

## Diagnostyka modelu na danych testowych

W celu określenia jakości predykcji modelu należy porównać predykcje do aktualnych wartości danych testowych. Macierz konfuzji ilustruje na ile dokładny jest model oraz jakiego typu błędy popełnia. Klasą pozytywną jest wynik „Yes” dla *Survived*.

Wykres 8 Macierz konfuzji

Obraz zawierający tekst, linia, diagram, Wykres

Opis wygenerowany automatycznieMacierz konfuzji pokazuje, że model popełnia po 12 błędów z klas *false positive* i *true negative*. Na podstawie tych wyników można określić dokładność, czyli stosunek poprawnych predykcji do predykcji ogółem – 0.8298 (95% CI: 0.7574 – 0.8878). W zależności od tego, na czym najbardziej zależy nam w predykcji należy skupić się na dodatkowych wyznacznikach – specyficzność (0.8571) oraz czułości (0.7895). Czułość to stosunek prawdziwych pozytywów do prawdziwych pozytywów i fałszywych negatywów, zaś specyficzność to stosunek prawdziwych negatywów do prawdziwych negatywów i fałszywych pozytywów. W tej sytuacji zależy nam, aby model miał jak najmniej fałszywych pozytywów, natomiast większa tolerancja jest do fałszywych negatywów – czyli ważniejszy jest dla nas wynik specyficzności.

Wykres 9 Krzywa ROC

Interpretacja kompromisów między czułością, a specyficznością obrazowana jest przez krzywą *ROC (Receiver operating characteristics)*. Na podstawie krzywej ROC można wyznaczyć punkt odcięcia prawdopodobieństwa predykcji, który satysfakcjonuje nasze potrzeby. Domyślny punkt odcięcia prawdopodobieństwa to 0.5, czyli predykcje >= 0.5 uznaje się za przeżycie katastrofy.

W przypadku, gdybyśmy chcieli zachować specyficzność na poziomie 0.9167, a czułość na 0.6842, to musielibyśmy zastosować odcięcie na poziomie 0.58.

## Odpowiedź na pytanie badawcze

Na początku zostało postawione pytanie: Czy przeżyłabym katastrofę Titanica? Model regresji logistycznej poproszony został o predykcję na zestawie parametrów: *Age*: 21; *Pclass\_21*: 1; *Pclass\_31*: 0; *Sex\_male*: 0; *SibSp*: 0.

**Model zwrócił predykcję na poziomie 0.868. Oznacza to, że przewiduje, że przeżyłabym katastrofę Titanica zakładając standardowe odcięcie na poziomie 0.5.**

# Zakończenie

Celem projektu było przeprowadzenie ANOVY, PCA oraz regresji logistycznej na danych w celu znalezienia różnorodnych zależności między zmiennymi. Znaleziono zależność między wiekiem, a klasą biletu – im droższy bilet (wyższa klasa) tym wyższa była średnia wieku w grupach. Potwierdzone zostały ogólne przekonania, że największą szansę na przeżycie miały młode kobiety z klasy pierwszej – dowiedziałam się również, że model prognozuje moje przeżycie w katastrofie.

# Bibliography

Lee, D., 2013. A Comparison of Choice-based Landscape Preference Models between British and Korean Visitors to. *Life Science Journal, 10(2).*

Vafakhah, M. i Kheirfam, H., 2015. Assessment of some homogeneous methods for the regional analysis of suspended sediment yield in the south and southeast of the Caspian Sea. *Journal of Earth System Science*, August.

1. *Titanic – Machine Learning from Disaster*, Kaggle, <https://www.kaggle.com/c/titanic/data>, Dostęp: 06.12.2023r. [↑](#footnote-ref-1)