**Projekt zaliczeniowy „Titanic” przygotowany na przedmiot Podstawy Uczenia Maszynowego (PUM)**

Autor: Dominik Suszek s23396

Spis treści

[Opis problemu 1](#_Toc153971041)

[Dane 2](#_Toc153971042)

[Sposób rozwiązania problemu 2](#_Toc153971043)

[Dyskusja wyników i ewaluacja modelu 2](#_Toc153971044)

[Podsumowanie 2](#_Toc153971045)

[Załączniki 3](#_Toc153971046)

# Opis problemu

Poniższy raport został opracowany na podstawie zbioru danych dotyczących katastrofy brytyjskiego statku pasażerskiego, kursującego na trasie pomiędzy USA i Europą. Ówcześnie był uważany za największy, najbardziej luksusowy i najbezpieczniejszy statek pasażerski na świecie. Niestety, w nocy z 14 na 15 kwietnia 1912 roku, podczas pierwszego rejsu na trasie Southampton – Cherbourg – Queenstown – Nowy Jork, statek zderzył się z górą lodową i zatonął. Ta tragedia zwróciła uwagę na niedostateczny poziom zabezpieczeń i potrzebę odświeżenia zasad bezpieczeństwa morskiego.

# Dane

Zbiór danych dotyczących katastrofy statku Titanic to bardzo dobrze znany zbiór wykorzystywany podczas nauki podstaw uczenia maszynowego. Zgodnie z rekomendacją, dane zostały pobrane ze strony <http://analityk.edu.pl/wpcontent/uploads/2020/02/titanic.csv>

Zbiór zawiera następujące informacje (nie zawsze kompletne) o wszystkich pasażerach Titanica:

* **survival** – czy pasażer przeżył (1 – tak, 0 – nie)
* **pclass** – w której klasie dany pasażer wykupił bilet na podróż (1 – najlepsza, 2, 3)
* **name** – imię i nazwisko pasażera (razem z tytułem, np. Mr.)
* **sex** – płeć
* **age** – wiek
* **sibsp** – małżonek/małżonka i łączna liczba rodzeństwa, które znajdowało się razem z pasażerem/pasażerką na pokładzie statku
* **parch** – łączną liczba dzieci i/lub rodziców, którzy znajdowali się razem z pasażerem/pasażerką na pokładzie statku
* **ticket** – numer identyfikacyjny biletu
* **fare** – cena biletu
* **cabin** – numer kajuty
* **embarked** – port, w którym dana osoba wsiadła na pokład Titanica (C = Cherbourg, Q = Queenstown, S = Southampton)
* **boat** – numer szalupy, na której dana osoba się ewakuowała
* **body** – numer identyfikacyjny znalezionego ciała

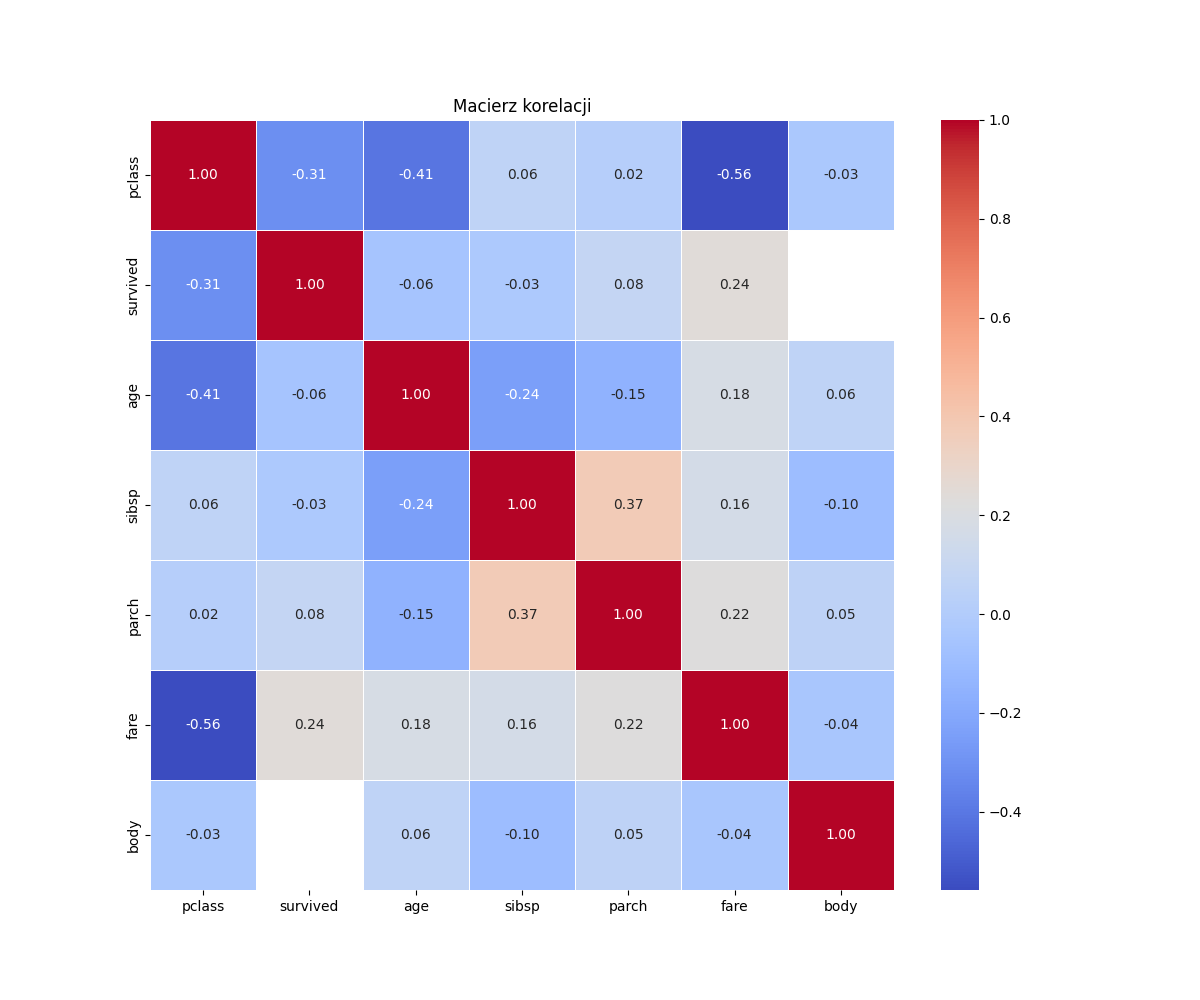
W poniższej tabeli przedstawiono statystyki opisowe zmiennych numerycznych:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **pclass** | **survived** | **age** | **sibsp** | **parch** | **fare** | **body** |
| **count** | 1309 | 1309 | 1046 | 1309 | 1309 | 1308 | 121 |
| **mean** | 2,29 | 0,38 | 29,88 | 0,50 | 0,39 | 33,30 | 160,81 |
| **std** | 0,84 | 0,49 | 14,41 | 1,04 | 0,87 | 51,76 | 97,70 |
| **min** | 1,00 | 0,00 | 0,17 | 0,00 | 0,00 | 0,00 | 1,00 |
| **25%** | 2,00 | 0,00 | 21,00 | 0,00 | 0,00 | 7,90 | 72,00 |
| **50%** | 3,00 | 0,00 | 28,00 | 0,00 | 0,00 | 14,45 | 155,00 |
| **75%** | 3,00 | 1,00 | 39,00 | 1,00 | 0,00 | 31,28 | 256,00 |
| **max** | 3,00 | 1,00 | 80,00 | 8,00 | 9,00 | 512,33 | 328,00 |

W celu zbadania zależności występujących pomiędzy zmiennymi, przygotowano macierz korelacji. Jest to macierz kwadratowa, której elementy są wartościami współczynników korelacji dla wszystkich możliwych par zmiennych losowych.

Zauważono, że pomiędzy zmienną ***survived***, a zmienną ***pclass*** występuje istotna ujemna korelacja. Jest to intuicyjne, ponieważ wraz ze wzrostem wartości zmiennej *pclass* (klasa, w której podróżowała dana osoba) prawdopodobnie malał jej priorytet podczas akcji ewakuacyjnej. W związku z tym, dla większych wartości *pclass* maleje średnia wartość zmiennej *survived* (maleje szansa danej osoby na uratowanie się).

Kolejną istotną korelację zaobserwowano pomiędzy zmiennymi



*Źródła danych, ocena ich wiarygodności*

*Krótka analiza opisowa danych*

*Uzasadnienie: w jaki sposób te dane mogą pomóc rozwiązać problem?*

# Sposób rozwiązania problemu

Pierwszym etapem realizacji projektu było zapoznanie się z danymi i próba zrozumienia zależności występujących pomiędzy zmiennymi. Do tego celu wykorzystano wykres przedstawiający macierz korelacji pomiędzy zmiennymi.

*Krótki opis wybranego modelu wraz z uzasadnieniem*

*Etapy realizacji projektu*

*Miary ewaluacji (oceny jakości) modelu*

# 

# Dyskusja wyników i ewaluacja modelu

*Wyniki modelowania*

# *Ewaluacja modelu*

# Podsumowanie

*Co się udało?*

*Jakie były problemy? Jak je rozwiązaliśmy?*

# *W jaki sposób może być to wykorzystane/rozwinięte w przyszłości?*

# Załączniki

A screenshot of a computer

Description automatically generated

A black and white screen with white text

Description automatically generated