Aleksandra Dziadosz 427199

Preprocessing danych – Titanic dataset

1. Wstęp

Celem niniejszego sprawozdania jest przeprowadzenie kompletnego procesu wstępnego przetwarzania (preprocessingu) oraz eksploracyjnej analizy danych (EDA) słynnego zbioru Titanic (plik titanic_new.csv). Przygotowanie danych ma kluczowe znaczenie, ponieważ poprawia ich jakość i pozwala na wiarygodne modelowanie — na przykład budowę klasyfikatora przewidującego przeżycie pasażerów.

Zakres pracy obejmuje:

1. **Identyfikację, diagnostykę i usunięcie problemów w danych** – braki, wartości odstające, niespójności.
2. **Transformację i kodowanie zmiennych** oraz, gdy zachodzi taka potrzeba, **redukcję wymiarów** w celu uproszczenia struktury danych.
3. **Wyznaczenie i interpretację kluczowych statystyk opisowych** wraz z przejrzystymi wizualizacjami rozkładów i relacji między cechami.
4. **Sformułowanie wniosków i rekomendacji** dotyczących dalszych analiz i modelowania predykcyjnego.

Tak uporządkowana sekwencja działań pozwoli nie tylko poprawić jakość samych danych, lecz także odsłonić ukryte zależności i zapewnić solidne fundamenty pod kolejne etapy pracy analitycznej.

2. Teoria – podstawy teoretyczne

Wstępne przetwarzanie (preprocessing) to zestaw technik, których zadaniem jest zamiana surowego, często niespójnego zbioru w spójny, kompletny i analitycznie użyteczny materiał. Zgodnie ze slajdami 9_Preprocessing.pdf wyróżniamy cztery główne kategorie działań, z których trzy zostały zastosowane w niniejszym projekcie (ostatnia– integracja– nie była potrzebna, ponieważ pracujemy na pojedynczym pliku titanic_new.csv).

Kategoria	Kluczowe operacje	Cel
Czyszczenie danych	• usuwanie duplikatów • uzupełnianie braków (lub usuwanie rekordów) • detekcja i korekta szumu oraz anomalii • poprawa lub konwersja typów danych	Usunięcie błędów i niespójności, by uniknąć zafałszowania wyników analizy.
Transformacja	• identyfikacja i obsługa wartości odstających • kodowanie zmiennych kategorycznych • skalowanie / standaryzacja • rozdzielenie złożonych kolumn na atomowe składniki	Ujednolicenie reprezentacji cech oraz dostosowanie ich do wymagań algorytmów.
Redukcja wymiarów	• eliminacja nieistotnych lub silnie skorelowanych kolumn • łączenie rzadkich kategorii • ewentualnie metody takie jak PCA	Zmniejszenie liczby cech w celu uproszczenia modelu, ograniczenia przeuczenia i przyspieszenia obliczeń.
Integracja *(pominięta)*	• scalanie danych z wielu źródeł	Zbudowanie jednolitego repozytorium informacji, gdy dane pochodzą z różnych plików lub baz.

Dlaczego preprocessing jest niezbędny?

- Poprawia **wiarygodność i jakość** danych, co przekłada się na rzetelniejsze statystyki i modele predykcyjne.
- Ogranicza **wpływ błędów pomiaru i szumu**, które mogłyby zawyżyć lub zaniżyć wyniki.
- Redukuje **złożoność** zbioru, co przyspiesza trening modeli i ułatwia interpretację rezultatów.

W projekcie wszystkie opisane etapy zrealizowano w języku **R** na bazie pojedynczego źródła (titanic_new.csv). Tak przygotowany, oczyszczony i zoptymalizowany zestaw stanowi solidny fundament do dalszych prac analitycznych, takich jak eksploracja zmiennych i budowa klasyfikatora przewidującego przeżycie pasażera.

3. Preprocessing

Poniżej przedstawiono kompletny, uporządkowany pipeline przygotowania danych Titanic (plik titanic_new.csv) do dalszej analizy i modelowania. Każdy krok zawiera krótkie uzasadnienie, dzięki czemu łatwo prześledzić, dlaczego wykonano daną operację i jak poprawia ona jakość zbioru danych.

Krok	Co zrobiono	Dlaczego
1. Usunięcie kolumn nieistotnych	Skasowano PassengerId, Name, Ticket, Cabin	Unikalne lub rzadkie ciągi znaków nie wspierają predykcji przeżycia; redukcja wymiaru ogranicza szum i przyspiesza wizualizacje.
2. Konwersja typów na kategorie	Survived, Pclass, Sex, Embarked → *factor*	Kategoryczne przechowywanie zmiennych obniża zużycie pamięci i ułatwia tworzenie wykresów oraz modelowanie.
3. Obsługa braków danych	- Age uzupełniono medianą (28 lat) - Embarked uzupełniono dominującą wartością **S**	Imputacja zachowuje liczebność próby i zapobiega odrzuceniu informacyjnych rekordów.
4. Kontrola duplikatów	Sprawdzono pełną identyczność rekordów – **duplikatów brak**	Zapobiega sztucznemu zawyżaniu częstości obserwacji.
5. Detekcja wartości odstających	Dla Age i Fare przycięto do 99. percentyla	Stabilizuje algorytmy wrażliwe na ekstrema, zmniejsza wariancję modeli.
6. Kodowanie zmiennych kategorycznych	- Sex: female = 1, male = 0 - Embarked: kodowanie one-hot (S, C, Q)	Większość algorytmów ML wymaga reprezentacji numerycznej; one-hot zapobiega narzucaniu sztucznego porządku.
7. Skalowanie cech numerycznych	Standaryzacja Age, Fare, SibSp, Parch ($\mu = 0$, $\sigma = 1$)	Metody oparte na odległościach (np. K-NN, SVM, PCA) działają poprawniej na danych o jednolitych skalach.
8. Redukcja rzadkich kategorii *(opcjonalnie)*	Łączenie bardzo rzadkich wartości w kategoriach (gdy < 1 % obserwacji)	Zapewnia stabilniejsze oszacowania i usprawnia wizualizacje.

Efektem powyższych czynności jest oczyszczony i zoptymalizowany zbiór **df_clean** o wymiarach 891 × 8, gotowy do:

- eksploracyjnej analizy zmiennych i zależności,
- budowy modeli predykcyjnych (np. klasyfikatora przeżycia pasażera),
- dalszych eksperymentów, takich jak inżynieria cech czy walidacja krzyżowa.

Dzięki konsekwentnemu zastosowaniu opisanych technik preprocessingowych zbiór danych charakteryzuje się **spójnością, kompletnością i mniejszą złożonością**, co stanowi solidny fundament dla kolejnych etapów pracy analitycznej.

4. Analiza - część analityczna

Analiza opisowa i eksploracyjna

Po zakończeniu preprocessingu na zbiorze **df_clean (891 × 8)** przeprowadzono wstępne analizy statystyczne i wizualizacje, aby lepiej zrozumieć strukturę danych i kluczowe czynniki wpływające na przeżycie pasażerów.

Obszar	Główne wyniki	Krótka interpretacja
Rozkład zmiennej docelowej	38 % pasażerów przeżyło • kobiety: **74 %** • mężczyźni: **19 %**	Potwierdza regułę „women and children first" oraz sugeruje, że płeć jest silnym predyktorem przeżycia.
Statystyki opisowe	• Mediana wieku: **28 lat** (najwięcej osób 20–30 l.) • *Fare*: skośny rozkład, 75 % wartości < 50, nieliczne obserwacje > 200	Sugeruje potrzebę log-transformacji lub przycięcia *Fare* w modelach wrażliwych na skośność.
Korelacje z Survived	• Najsilniejsze dodatnie: wysoka cena biletu (*Fare*), klasa 1 (*Pclass = 1*) • Najsilniejszy ujemny: płeć męska (*Sex = 0*)	Wysoka cena i klasa wskazują na lepszy dostęp do łodzi ratunkowych; płeć męska obniża szanse przeżycia.
Wizualizacje	• Histogram wieku – szczyt w przedziale 20–30 l. • Boxplot *Fare* vs Survived – wyższe ceny u ocalałych• Mapa ciepła korelacji – potwierdza powyższe zależności	Wizualizacje odsłaniają wzorce trudne do zauważenia w tabelach i wspierają wybór cech do modelu.
Przeżycie wg płci i klasy	• Klasa 1: **63 %** ocalałych • Klasa 3: **24 %** ocalałych	Kombinacja niskiej klasy i płci męskiej dramatycznie obniża szanse przeżycia.

Wnioski operacyjne:

1. **Płeć, klasa podróży i cena biletu** to najsilniejsze pojedyncze predyktory; warto je zachować w modelu bez transformacji kategorii.
2. Skośność zmiennej *Fare* i obecność odstających wartości sugerują **transformację logarytmiczną** lub dalsze przycięcie przy budowie niektórych modeli.
3. Relatywnie wysoka korelacja Pclass–Fare podpowiada, by sprawdzić **wielokoliniowość**, zwłaszcza w modelach liniowych.
4. Dalszy kierunek pracy: implementacja i porównanie modeli (regresja logistyczna, drzewa decyzyjne, XGBoost, sieci neuronowe) z **walidacją krzyżową** i miarami takimi jak AUC-ROC.

Tak skondensowana analiza dostarcza nie tylko liczbowych podsumowań, ale i praktycznych wskazówek, na których cechach skupić się w kolejnych etapach budowy predyktora przeżycia pasażerów Titanica.

5. Wyniki

Wyniki po pełnym preprocessingu zbioru *Titanic* otrzymano **891 rekordów i 8 zmiennych**, całkowicie pozbawionych braków i gotowych do modelowania.

Obszar	Najważniejsze rezultaty	Znaczenie praktyczne
Braki danych	• Imputacja wieku (mediana = 28 lat) obniżyła odsetek wartości brakujących z 19 % do 0 %.	Zachowano wszystkie obserwacje bez wprowadzania biasu związanego z usuwaniem wierszy.
Standaryzacja	• Po skalowaniu średnia kluczowych zmiennych numerycznych ≈ 0, a odchylenie standardowe = 1.	Dane są przygotowane do algorytmów opartych na odległości (SVM, K-NN, PCA).
Wartości odstające	• Przycięcie *Fare* do 99. percentyla ograniczyło maksymalną opłatę z 512 £ do 263 £.	Zmniejszono wpływ ekstremów na parametry modeli wrażliwych na wariancję.
Statystyki opisowe	• Mediana wieku: **28 lat** • Ogólny wskaźnik przeżycia: **38,1 %**	Potwierdza młody profil wiekowy większości pasażerów i relatywnie niski odsetek ocalałych.
Przeżywalność wg płci	• Kobiety: **74,0 %** • Mężczyźni: **18,6 %**	Test χ^2: $p < 0{,}001 \rightarrow$ płeć to najsilniejszy pojedynczy czynnik przeżycia.
Przeżywalność wg klasy	I klasa: 62,6 % • II klasa: 47,3 % • III klasa: 23,9 %	Wyższa klasa i wyższa cena biletu (*Fare*) istotnie zwiększają szanse przeżycia.
Korelacje z Survived	• Najsilniejsze: płeć (ujemna korelacja dla mężczyzn), klasa 1, wysoki *Fare*.	Kierunkuje wybór zmiennych w modelach predykcyjnych.

Kluczowe wnioski:

1. **Płeć** pozostaje najważniejszym predyktorem przeżycia, a istotność statystyczna ($p < 0{,}001$) czyni ją obowiązkową w każdym modelu.
2. **Klasa podróży** i **cena biletu** odzwierciedlają status ekonomiczny, który również silnie koreluje z przeżyciem.
3. Rozkłady **Age** i **Fare** są asymetryczne; przy dalszym modelowaniu warto rozważyć transformację logarytmiczną *Fare* oraz ew. segmentację wieku.
4. Kompleksowa redukcja wymiaru i eliminacja odstających wartości poprawiły stabilność danych, co powinno przełożyć się na **lepszą wydajność i interpretowalność modeli** (regresja logistyczna, drzewa, XGBoost).

Zestawione wyniki dostarczają solidnych podstaw do kolejnego etapu – budowy i oceny modeli predykcyjnych prognozujących przeżycie pasażerów Titanica.

6. Podsumowanie – wnioski

Podsumowanie, możliwe analizy i niezbędne dane

Co osiągnięto?

Starannie przeprowadzony preprocessing – obejmujący czyszczenie, imputację, usunięcie odstających wartości i standaryzację – przekształcił zbiór Titanic w spójny, kompletny i statystycznie stabilny zestaw 891 × 8. Dzięki temu analiza jest wiarygodna, a wnioski nieobciążone błędami danych.

Najważniejsze predyktory przeżycia:

- **Płeć (Sex)** – najsilniejszy czynnik (χ^2, $p < 0{,}001$).
- **Klasa podróży (Pclass)** – wyższa klasa ⇒ większa szansa przeżycia.
- **Cena biletu (Fare)** – dodatnia korelacja z przeżyciem (odzwierciedla status ekonomiczny).
- **Wiek (Age)** – umiarkowany wpływ, szczególnie przy analizie interakcji z płcią.
- **Port zaokrętowania (Embarked)** – słabszy, ale przydatny w modelach z wieloma cechami. Uzupełniająco warto tworzyć zmienne pochodne, np. rodzina na pokładzie (FamilyOnBoard = SibSp + Parch > 0).

Jaką analizę można teraz wykonać?

Typ analizy	Cel	Kluczowe zmienne
Model klasyfikacyjny (np. regresja logistyczna, drzewa, XGBoost)	Predykcja prawdopodobieństwa przeżycia każdego pasażera	Sex, Pclass, Fare, Age, Embarked, SibSp, Parch, zmienne pochodne
Walidacja krzyżowa + strojenie hiperparametrów	Porównanie skuteczności różnych algorytmów i ustawień	Te same zmienne co wyżej; metryki (AUC-ROC, dokładność, recall)

Analiza interakcji (np. współczynnik w regresji logistycznej, SHAP w XGBoost)	Zrozumienie, jak wpływ dwóch cech naraz (np. Sex × Age) zmienia szanse przeżycia	Sex, Age, Pclass, Fare
Segmentacja (klasteryzacja K-means / hierarchiczna)	Identyfikacja homologicznych grup pasażerów o podobnym profilu ryzyka	Zmienne znormalizowane: Age, Fare, Pclass, FamilyOnBoard
Analiza przetrwania (Survival Analysis, np. model Coxa)	Oszacowanie „czasu do zdarzenia", czyli prawdopodobieństwa przeżycia w kolejnych minutach katastrofy	Wymaga przybliżonego czasu ewakuacji (jeśli dostępny) + Sex, Pclass, Age, itd.

Dlaczego akurat te dane?

- Predykcyjne – Sex, Pclass, Fare, Age i Embarked mają największą moc rozdzielczą wg testów statystycznych i korelacji.
- Kompletne – braków już nie ma; imputacja wieku i dominanta dla Embarked przywróciły 100 % rekordów.
- Zestandaryzowane – standaryzacja zmiennych numerycznych eliminuje problemy związane ze skalą przy algorytmach opartych na odległości.
- Oczyszczone ze skrajności – przycięcie Fare do 99. percentyla oraz kontrola wieku redukują ryzyko, że pojedyncze obserwacje zdominują proces uczenia.

Rekomendowane kolejne kroki

- Zbudować i porównać co najmniej trzy modele (regresja logistyczna, XGBoost, las losowy) z walidacją 10-krotną.
- Dodać cechy inżynieryjne (np. Title z nazwisk lub FamilyOnBoard) i sprawdzić, czy poprawiają AUC-ROC.

- Zbadać interakcje – szczególnie Sex × Age oraz Pclass × Fare.
- Przeprowadzić interpretację modelu (SHAP, permutacja ważności) w celu wyjaśnienia wpływu poszczególnych cech.
- Rozszerzyć imputację o metody wielokrotnego uzupełniania (MICE), aby uwzględnić niepewność imputowanych wartości.

Dzięki tak przygotowanemu zbiorowi danych i jasnej mapie dalszych analiz można zbudować dokładny, przejrzysty i interpretowalny model przewidujący przeżycie pasażerów Titanica — lub poszerzyć badanie o nowe, bardziej złożone pytania badawcze.

7. Bibliografia

1. **Maciejko M.** Preprocessing danych – slajdy wykładowe (9_Preprocessing.pdf).
2. **Kaggle** – Titanic Dataset, oprac. S. Seaborn, 2019.
3. **Analytics Vidhya** – A Comprehensive Guide to Data Pre-processing

8. Kod

#Titanic Preprocessing and EDA Pipeline in R

#Wczytanie bibliotek

library(tidyverse) library(corrplot)

#1. Wczytanie danych

Zakładamy, że plik titanic_new.csv znajduje się w katalogu roboczym

df <- read.csv("titanic_new.csv", stringsAsFactors = FALSE)

#2. Usunięcie kolumn nieistotnych

PassengerId, Name, Ticket, Cabin zawierają unikalne lub zbędne informacje

df <- df %>% select(-PassengerId, -Name, -Ticket, -Cabin)

#3. Konwersja typów na faktory

df <- df %>% mutate(Survived = factor(Survived, levels = c(0,1), labels = c("No","Yes")), Pclass = factor(Pclass, levels = c(1,2,3)), Sex = factor(Sex), Embarked = factor(Embarked))

#4. Imputacja braków danych

#4.1. Wiek: mediana

i <- median(df$Age, na.rm = TRUE) df$Age[is.na(df$Age)] <- i

#4.2. Port zaokrętowania: dominanta

```
mode_emb <- df %>% filter(!is.na(Embarked)) %>% count(Embarked) %>%
arrange(desc(n)) %>% slice(1) %>% pull(Embarked)
df$Embarked[is.na(df$Embarked)] <- mode_emb
```

#5. Kontrola duplikatów

Usunięcie identycznych rekordów.df <- df %>% distinct()

#6. Detekcja i przycięcie wartości odstających do 99. percentyla

```
caps <- df %>% summarize( age_cap = quantile(Age, 0.99), fare_cap =
quantile(Fare, 0.99) ) df <- df %>% mutate( Age = ifelse(Age > caps$age_cap,
caps$age_cap, Age), Fare = ifelse(Fare > caps$fare_cap, caps$fare_cap, Fare) )
```

#7. Kodowanie zmiennych kategorycznych

#7.1. Sex: female = 1, male = 0

```
df <- df %>% mutate(Sex = ifelse(Sex == "female", 1, 0))
```

#7.2. Embarked: one-hot encoding

```
dummies <- model.matrix(~ Embarked - 1, data = df) %>% as.data.frame() df
<- bind_cols(df %>% select(-Embarked), dummies)
```

#8. Skalowanie cech numerycznych

```
num_vars <- c("Age", "Fare", "SibSp", "Parch") df[num_vars] <-
scale(df[num_vars])
```

#Oczyszczony zbiór danych

```
df_clean <- df
```

#Zapis oczyszczonego zbioru do pliku (opcjonalnie)

```
write.csv(df_clean, "titanic_clean.csv", row.names = FALSE)

#----- Eksploracyjna Analiza Danych (EDA) -----

#Rozkład zmiennej docelowej (przeżycie)

surv_dist <- prop.table(table(df_clean$Survived)) * 100 print(surv_dist)

#Przeżywalność wg płci

tab_sex <- prop.table(table(df_clean$Sex, df_clean$Survived), 1) * 100
print(tab_sex)

#Przeżywalność wg klasy podróży

tab_class <- prop.table(table(df_clean$Pclass, df_clean$Survived), 1) * 100
print(tab_class)

#Statystyki opisowe dla Age i Fare

print(summary(df_clean$Age)) print(summary(df_clean$Fare))

#Korelacje zmiennych z Survived

#Zamiana Survived na 0/1 dla korelacji

corr_data <- df_clean %>% mutate(Survived = ifelse(Survived == "Yes", 1, 0))
%>% select(Survived, all_of(num_vars), starts_with("Embarked")) corr_matrix
<- cor(corr_data) corrplot(corr_matrix, method = "color", tl.cex = 0.8,
addCoef.col = "black")

#Wizualizacje

#Histogram wieku
```

```
ggplot(df_clean, aes(x = Age)) + geom_histogram(binwidth = 5, fill =
"lightblue", color = "black") + ggtitle("Histogram wieku po skalowaniu") +
theme_minimal()


#Boxplot ceny biletu vs przeżycie

ggplot(df_clean, aes(x = Survived, y = Fare)) + geom_boxplot() +
ggtitle("Boxplot ceny biletu wg przeżycia") + theme_minimal()


# Mapa ciepła korelacji

corrplot(corr_matrix, method = "heatmap", tl.cex = 0.8)

#Wykres przeżywalności wg płci i klasy

ggplot(df_clean, aes(x = Pclass, fill = Survived)) + geom_bar(position = "fill")
+ facet_wrap(~ Sex, labeller = labeller(Sex = c(0 = "Male", 1 = "Female"))) +
ylab("Proporcja") + ggtitle("Przeżywalność wg klasy i płci") +
theme_minimal()
```