Preprocessing

Karol Klimek

1. Wstęp

Preprocessing to wstępne przetwarzanie danych polegające na czyszczeniu, przekształcaniu i integrowaniu danych. Ma on na celu przygotowanie danych do analizy poprzez poprawę ich jakości.

1.1. Czyszczenie danych

W zbiorach danych, zwłaszcza dużych, istnieje możliwość występowania błędów, szumów, duplikatów, braków, danych nieistotnych lub ich niewłaściwy format. Czyszczenie danych w preprocessingu polega na usunięciu lub zmodyfikowaniu tych nieprawidłowości, co pozwala na poprawną analizę.

1.2. Transformacja danych

Przed przeprowadzeniem analizy danych należy sprawdzić czy ich format jest poprawny oraz przeglądnąć zbiór, aby upewnić się czy nie występują wartości, które mogłyby zmienić wyniki analizy. Transformacja danych pozwala na upewnienie się, że wyżej wymienione błędy nie wystąpią w zbiorze naszych danych.

1.3. Redukcja danych

Do analizy danych często nie są konieczne wszystkie dostępne w zbiorze obserwacje. Redukcja danych sprawia, że do analizy zostawimy jak największą liczbę danych przy zmniejszeniu ich złożoności.

2. Preprocessing danych na pliku *titanic_new.csv*

	Passenger.Id	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	1	0	3	Braund, Mr. Owen Harris	male	22.00	1	0	A/5 21171	'7.25'		S
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1	0	PC 17599	'71.2833'	C85	C
3	3	1	3	Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 3101282	'7.925'		S
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803	'53.1'	C123	S
5	5	0	3	Allen, Mr. William Henry	male	35.00	0	0	373450	'8.05'		S
6	6	0	3	Moran, Mr. James	male	*NA*	0	0	330877	'8.4583'		Q
7	7	0	1	McCarthy, Mr. Timothy J	male	54.00	0	0	17463	'51.8625'	E46	s
8	8	0	3	Palsson, Master. Gosta Leonard	male	2.00	3	1	349909	'21.075'		s
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00	0	2	347742	'11.1333'		S
10	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.00	1	0	237736	'30.0708'		C
11	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.00	1	1	PP 9549	'16.7'	G6	s
12	12	1	1	Bonnell, Miss. Elizabeth	female	58.00	0	0	113783	'26.55'	C103	S
13	13	0	3	Saundercock, Mr. William Henry	male	20.00	0	0	A/5. 2151	'8.05'		s
14	14	0	3	Andersson, Mr. Anders Johan	male	39.00	1	5	347082	'31.275'		S
15	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.00	0	0	350406	'7.8542'		S
16	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55.00	0	0	248706	'16'		s
17	17	0	3	Rice, Master. Eugene	male	2.00	4	1	382652	'29.125'		Q
18	18	1	2	Williams, Mr. Charles Eugene	male	*NA*	0	0	244373	'13'		S
19	19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31.00	1	0	345763	'18'		S
20	20	1	3	Masselmani, Mrs. Fatima	female	*NA*	0	0	2649	'7.225'		C
21	21	0	2	Fynney, Mr. Joseph J	male	35.00	0	0	239865	'26'		S
22	22	1	2	Beesley, Mr. Lawrence	male	34.00	0	0	248698	'13'	D56	s
23	23	1	3	McGowan, Miss. Anna "Annie"	female	15.00	0	0	330923	'8.0292'		Q
25	25	0	3	Palsson, Miss. Torborg Danira	female	8.00	3	1	349909	'21.075'		S

Pierwsze 25 obserwacji z titanic_new.csv

Objaśnienie znaczenia kolumn:

- Passenger.id : numer pasażera
- Survived : 0 –> pasażer nie przeżył; 1-> pasażer przeżył
- Pclass : numer klasy (1-najwyższa, 2 lub 3)
- Name : Imię i nazwisko
- Sex : płeć
- Age : wiek w latach
- SibSp : liczba rodzeństwa lub współmałżonków pasażera na pokładzie
- Parch : liczba rodziców lub dzieci pasażera na pokładzie
- Ticket : numer biletu
- Fare : opłata za bilet
- Cabin : zajmowana kabina
- Embarked : port wsiadania (C = Cherbourg, Q = Queenstown, S = Southampton)

Zbiór danych *titanic_new.csv* zawiera 12 zmiennych oraz 891 obserwacji. Przed wykonaniem preprocessingu należy oglądnąć dane, zadać pytanie czy wszystkie wartości są konieczne do wykonania analizy oraz znaleźć potencjalne błędy w zbiorze.

Na pierwszy rzut oka można zauważyć występowanie braków danych (kolumny Age, Cabin) oraz kolumn danych zbędnych do przeprowadzenia analizy (np. Name, Ticket).

Czyszczenie danych rozpoczynamy od usunięcia danych nieistotnych. Zbędne do analizy będą kolumny:

- Passenger.id
- Name
- Ticket
- Fare
- Cabin
- Embarked

Usunięcia danych dokonamy za pomocą funkcji select() z pakietu dplyr. Sprawi to utworzenie nowego zbioru danych lepiej przygotowanego do analizy.

```
4  data=select(data, -Passenger.Id)
5  data=select(data, -Name)
6  data=select(data, -Ticket)
7  data=select(data, -Fare)
8  data=select(data, -Cabin)
9  data=select(data, -Embarked)
```

Po usunięciu zbędnych kolumn nasz zbiór danych wygląda następująco:

	Survived	Pclass	Sex	Age	SibSp	Parch
1	0	3	male	22.00	1	0
2	1	1	female	38.00	1	0
3	1	3	female	26.00	0	0
4	1	1	female	35.00	1	0
5	0	3	male	35.00	0	0
6	0	3	male	*NA*	0	0
7	0	1	male	54.00	0	0
8	0	3	male	2.00	3	1
9	1	3	female	27.00	0	2
10	1	2	female	14.00	1	0

Następnie należy usunąć szumy. Można zauważyć ich występowanie w kolumnie Survived:

24	111	1	male	28.00	0	0
45	-1	3	female	19.00	0	0

Kolumna Survived powinna przyjmować wartości 0 lub 1, zatem wartości między innymi 111 oraz -1 są niepoprawne i należy je usunąć. Do tego użyjemy funkcji filter().

```
11  data=filter(data, Survived %in% c(0,1))
```

Zbiór danych zmniejszył się z 891 obserwacji do 887.

Można również zauważyć, że niektóre obserwacje mają niepoprawnie wpisaną płeć (różną od ‘male’ lub ‘female’). Usuniemy je analogicznie jak wcześniej.

```
12  data=filter(data, Sex=='female' | Sex=='male')
```

Zbiór danych zmniejszył się z 887 obserwacji do 879.

W zbiorze nie występują inne szumy, więc przechodzimy do sprawdzenia poprawności typu danych.

```
> typeof(data$Survived)
[1] "integer"
> typeof(data$Pclass)
[1] "integer"
> typeof(data$Age)
[1] "double"
> typeof(data$Sex)
[1] "character"
> typeof(data$SibSp)
[1] "integer"
> typeof(data$Parch)
[1] "integer"
```

Widać, że typy zmiennych są poprawne, więc należy się przyjrzeć czy w zbiorze występują wartości brakujące.

Brakujące wartości występują w kolumnie Age, zatem zastąpimy je wartością średnią wieku pasażerów, których wiek jest podany.

```
> ave_age = mean(data$Age, na.rm = TRUE)
> data$Age[is.na(data$Age)] = ave_age
> ave_age
[1] 29.73959
```

W ten sposób nie usuwamy dużej liczby obserwacji, ponieważ wielu z nich brakowało wieku.

Tak wygląda obecny zbiór danych:

	Survived	Pclass	Sex	Age	SibSp	Parch
1	0	3	male	22.00000	1	0
2	1	1	female	38.00000	1	0
3	1	3	female	26.00000	0	0
4	1	1	female	35.00000	1	0
5	0	3	male	35.00000	0	0
6	0	3	male	29.72559	0	0
7	0	1	male	54.00000	0	0
8	0	3	male	2.00000	3	1
9	1	3	female	27.00000	0	2
10	1	2	female	14.00000	1	0
11	1	3	female	4.00000	1	1
12	1	1	female	58.00000	0	0
13	0	3	male	20.00000	0	0
14	0	3	male	39.00000	1	5
15	0	3	female	14.00000	0	0

Na tym zakończyliśmy preprocessing danych. Należy mieć na uwadze, że mógłby on zostać wykonany w inny sposób, natomiast uważam że wyżej przeze mnie przedstawiony jest najbardziej optymalny.

3. Analiza danych na podstawie *titanic_new.csv*

3.1. Statystyki opisowe

3.1.1. Miary występowania

table(data$Survived)

	Var1	Freq
1	0	549
2	1	338

table(data$Pclass)

	Var1	Freq
1	1	214
2	2	184
3	3	489

table(data$Sex)

	Var1	Freq
1	female	308
2	male	571

table(data$SibSp)

	Var1	Freq
1	0	598
2	1	209
3	2	27
4	3	15
5	4	18
6	5	5
7	8	7

table(data$Parch)

	Var1	Freq
1	0	669
2	1	115
3	2	80
4	3	5
5	4	4
6	5	5
7	6	1

3.1.2 Miary położenia

```
> mean(data$Age)
[1] 29.7368
> mean(data$SibSp)
[1] 0.5244596
> mean(data$Parch)
[1] 0.3833902
```

Średnia wieku pasażerów wynosi około 30 lat. Średnia wartość liczby rodzeństwa/współmałżonka wynosi ok. 0.5 a liczby dzieci/rodziców 0.4, co oznacza, że niewiele osób podróżowało z najbliższymi.

```
> kurtosis(data$Age)
[1] 0.9599045
```

Niewielka wartość kurtozy dla wieku świadczy o skoncentrowaniu wyników wokół średniej.

3.1.3. Miary zmienności

```
> sd(data$Age)
[1] 12.99274
```

Wartość odchylenia standardowego pokazuje, że wartości są rozproszone wokół średniej.

```
> min(data$Age)
[1] 0.42
> max(data$Age)
[1] 80
```

Najmłodsza osoba na pokładzie statku miała nieco ponad 5 miesięcy, natomiast najstarsza 80 lat.

3.1.4. Miary asymetrii

```
> skewness(data$Age)
[1] 0.4452924
> skewness(data$SibSp)
[1] 3.696436
> skewness(data$Parch)
[1] 2.731659
```

Skośność dla wieku jest niewielka, co pokazuje że centrum rozkładu znajduje się blisko średniej. Natomiast wartości skośności dla liczby osób z rodziny wskazują na przesunięcie centrum w prawą stronę względem średniej.

3.2. Relacje między zmiennymi

W pozostawionych danych istnieje moim zdaniem szansa na korelację. W celu sprawdzenia moich przypuszczeń należy policzyć współczynnik korelacji między poszczególnymi zmiennymi.

```
> cor(data$Survived, data$Age)
[1] -0.0656355
> cor(data$Survived, data$Pclass)
[1] -0.3409262
> cor(data$Survived, data$SibSp)
[1] -0.0322073
> cor(data$Survived, data$Parch)
[1] 0.08383042
```

Po przeprowadzeniu testów Pearsona na poszczególnych parach zmiennych można wywnioskować, że jedynie korelacja między Survived i Pclass oraz Survived i Parch jest istotna statystycznie.

```
> cor.test(data$Survived, data$Pclass)

        Pearson's product-moment correlation

data:  data$Survived and data$Pclass
t = -10.74, df = 877, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3980766 -0.2811397
sample estimates:
       cor
-0.3409262
```

```
> cor.test(data$Survived, data$Parch)

        Pearson's product-moment correlation

data:  data$Survived and data$Parch
t = 2.4913, df = 877, p-value = 0.01291
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.0178047 0.1491282
sample estimates:
       cor
0.08383042
```

Obie wartości współczynnika korelacji są zbliżone do zera, natomiast można na ich podstawie wywnioskować, że szanse przetrwania rosły wraz ze wzrostem liczby rodziców/dzieci oraz rosły wraz ze spadkiem numeru klasy pasażera (czyli wzrostem jego klasy, ponieważ klasa 1 jest najwyższa, klasa 3 najniższa), co można rozumieć tak, że pasażerowie wyższych klas mieli większe szanse przeżycia np. dzięki temu, że byli bardziej priorytetowi w procesie ewakuacji.

Dokonam jeszcze analizy współczynnika korelacji między wiekiem a liczbą osób z rodziny.

```
> cor.test(data$Age,data$SibSp)

        Pearson's product-moment correlation

data:  data$Age and data$SibSp
t = -7.09, df = 877, p-value = 2.76e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2944226 -0.1693134
sample estimates:
       cor
-0.2328311

> cor.test(data$Age, data$Parch)

        Pearson's product-moment correlation

data:  data$Age and data$Parch
t = -5.3132, df = 877, p-value = 1.367e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2399169 -0.1117748
sample estimates:
      cor
-0.176594
```

Obie pary zmiennych są istotne statystycznie, zatem można wysnuć wniosek że wraz ze wzrostem wieku malała liczba rodzeństwa/współmałżonków i dzieci/rodziców.

4. Podsumowanie

Przeprowadzony preprocessing przygotował dane i umożliwił ich poprawną analizę. Jest to bardzo istotna czynność, którą należy wykonać aby analiza danych była możliwa i aby jej wyniki odzwierciedlały rzeczywistość.