Titanic_cardinality

April 10, 2025

Titanic - cardinality

Dominik Sakłaski, 415120

Załadowanie bibliotek

[1]:
```
import pandas as pd
import numpy as np
```

Wczytanie danych i zmiana wartości “?” na NaN

[2]:
```
path = 'data_titanic.txt'
columns = [
    "pclass",
    "survived",
    "name",
    "sex",
    "age",
    "sibsp",
    "parch",
    "ticket",
    "fare",
    "cabin",
    "embarked",
    "boat",
    "body",
    "home.dest"
]

data = pd.read_csv(path, header=None, names=columns, skiprows=17)

data.index = range(1, len(data) + 1)

data.replace('?', np.nan, inplace=True)
print("20 początkowych wierszy po zamianie '?' na NaN: \n")
print(data.head(20))
```

```
20 początkowych wierszy po zamianie '?' na NaN:

    pclass  survived                                             name  \
1        1         1                    Allen, Miss. Elisabeth Walton
```

2	1	1	Allison, Master. Hudson Trevor
3	1	0	Allison, Miss. Helen Loraine
4	1	0	Allison, Mr. Hudson Joshua Creighton
5	1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)
6	1	1	Anderson, Mr. Harry
7	1	1	Andrews, Miss. Kornelia Theodosia
8	1	0	Andrews, Mr. Thomas Jr
9	1	1	Appleton, Mrs. Edward Dale (Charlotte Lamson)
10	1	0	Artagaveytia, Mr. Ramon
11	1	0	Astor, Col. John Jacob
12	1	1	Astor, Mrs. John Jacob (Madeleine Talmadge Force)
13	1	1	Aubart, Mme. Leontine Pauline
14	1	1	Barber, Miss. Ellen 'Nellie'
15	1	1	Barkworth, Mr. Algernon Henry Wilson
16	1	0	Baumann, Mr. John D
17	1	0	Baxter, Mr. Quigg Edmond
18	1	1	Baxter, Mrs. James (Helene DeLaudeniere Chaput)
19	1	1	Bazzani, Miss. Albina
20	1	0	Beattie, Mr. Thomson

	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	\
1	female	29	0	0	24160	211.3375	B5	S	2	
2	male	0.9167	1	2	113781	151.55	C22 C26	S	11	
3	female	2	1	2	113781	151.55	C22 C26	S	NaN	
4	male	30	1	2	113781	151.55	C22 C26	S	NaN	
5	female	25	1	2	113781	151.55	C22 C26	S	NaN	
6	male	48	0	0	19952	26.55	E12	S	3	
7	female	63	1	0	13502	77.9583	D7	S	10	
8	male	39	0	0	112050	0	A36	S	NaN	
9	female	53	2	0	11769	51.4792	C101	S	D	
10	male	71	0	0	PC 17609	49.5042	NaN	C	NaN	
11	male	47	1	0	PC 17757	227.525	C62 C64	C	NaN	
12	female	18	1	0	PC 17757	227.525	C62 C64	C	4	
13	female	24	0	0	PC 17477	69.3	B35	C	9	
14	female	26	0	0	19877	78.85	NaN	S	6	
15	male	80	0	0	27042	30	A23	S	B	
16	male	NaN	0	0	PC 17318	25.925	NaN	S	NaN	
17	male	24	0	1	PC 17558	247.5208	B58 B60	C	NaN	
18	female	50	0	1	PC 17558	247.5208	B58 B60	C	6	
19	female	32	0	0	11813	76.2917	D15	C	8	
20	male	36	0	0	13050	75.2417	C6	C	A	

	body	home.dest
1	NaN	St Louis, MO
2	NaN	Montreal, PQ / Chesterville, ON
3	NaN	Montreal, PQ / Chesterville, ON
4	135	Montreal, PQ / Chesterville, ON
5	NaN	Montreal, PQ / Chesterville, ON

```
6    NaN                    New York, NY
7    NaN                      Hudson, NY
8    NaN                     Belfast, NI
9    NaN            Bayside, Queens, NY
10    22            Montevideo, Uruguay
11   124                    New York, NY
12   NaN                    New York, NY
13   NaN                   Paris, France
14   NaN                             NaN
15   NaN                   Hessle, Yorks
16   NaN                    New York, NY
17   NaN                    Montreal, PQ
18   NaN                    Montreal, PQ
19   NaN                             NaN
20   NaN                    Winnipeg, MN
```

0. DODATKOWO: Sprawdź liczebność poszczególnych etykiet dla wszystkich zmiennych print('Liczba etykiet zmiennej zmiennaA:{}'.format(len(data. zmiennaA.unique())))

[3]:
```
for column in columns:
    print('Liczba etykiet zmiennej {}:{}'.format(
    column, len(data[column].unique())))
```

```
Liczba etykiet zmiennej pclass:3
Liczba etykiet zmiennej survived:2
Liczba etykiet zmiennej name:1307
Liczba etykiet zmiennej sex:2
Liczba etykiet zmiennej age:99
Liczba etykiet zmiennej sibsp:7
Liczba etykiet zmiennej parch:8
Liczba etykiet zmiennej ticket:929
Liczba etykiet zmiennej fare:282
Liczba etykiet zmiennej cabin:187
Liczba etykiet zmiennej embarked:4
Liczba etykiet zmiennej boat:28
Liczba etykiet zmiennej body:122
Liczba etykiet zmiennej home.dest:370
```

1. Sprawdź liczebność poszczególnych etykiet dla danych zmiennych jakościowych print('Liczba etykiet zmiennej zmiennaA:{}'.format(len(data. zmiennaA.unique())))

[4]:
```
qualitative_variables = [
    "pclass",
    "survived",
    "sex",
    "cabin",
    "embarked",
    "boat",
```

```
    "home.dest"
]

for column in qualitative_variables:
    print('Liczba etykiet zmiennej jakościowej {}:{}'.format(
    column, len(data[column].unique())))
```

```
Liczba etykiet zmiennej jakościowej pclass:3
Liczba etykiet zmiennej jakościowej survived:2
Liczba etykiet zmiennej jakościowej sex:2
Liczba etykiet zmiennej jakościowej cabin:187
Liczba etykiet zmiennej jakościowej embarked:4
Liczba etykiet zmiennej jakościowej boat:28
Liczba etykiet zmiennej jakościowej home.dest:370
```

2. Wyświetl z użyciem funkcji print liczbę wszystkich pasażerów. Wykorzystaj podobny sposób jak w pkt 1.

[5]:
```
print('Liczba wszystkich pasażerów: {}'.format(len(data)))
```

```
Liczba wszystkich pasażerów: 1309
```

3. Skomentuj wyniki otrzymane w punkcie 1 i 2. Podziel zmienne ze względu na dużą i małą moc zbioru (kardynalność).

Ogólna liczba pasażerów: Z analizy danych wynika, że na pokładzie Titanica znajdowało się 1309 pasażerów. Ta liczba stanowi podstawę do głębszego zrozumienia rozmiaru katastrofy oraz umożliwia szczegółowe badania.

Liczba etykiet zmiennych:

- **pclass**: jest to `zmienna jakościowa`, ponieważ klasyfikuje pasażerów do określonych kategorii (klas) opartych na poziomie usług na statku; zmienna posiada 3 unikalne etykiety opisujące klasy biletów na statku.
- **survived**: `zmienna jakościowa`; zawiera 2 unikalne etykiety, które niosą ze sobą informacje czy dany pasażer przeżył katastrofe czy też nie.
- **name**: `nie jest zmienną jakościową`, ponieważ każda nazwa jest unikalna i nie grupuje się w większe kategorie, które miałyby zastosowanie analityczne; zawiera aż 1307 unikalnych etykiet, co jest prawie równe ogólne liczbie pasażerów - jest to typowe dla danych osobowych, gdzie każda nazwa identyfikuje osobę i jest unikalna.
- **sex**: `zmienna jakościowa`, bo klasyfikuje osoby do jednej z dwóch kategorii; zawiera 2 unikalne etykiety (male i female); ta zmienna jest kluczowa w analizach różnicujących wyniki w zależności od płci.
- **age**: `nie jest to zmienna jakościowa`, ponieważ reprezentuje wartości liczbowe (wiek), które można analizować kwantytatywnie (średnia, mediana); posiada 99 unikalnych etykiet;

po wykonaniu grupowania na tej zmiennej mogłaby się ona stać zmienną jakościową; natomiast można przekształcić tą zmienną ilościową na jakościową po wykonaniu np. grupowania na jakieś grupy wiekowe

- **sibsp**: `nie jest to zmienna jakościowa`, tylko ilościowa, bo reprezentuje liczbe rodzeństwa/małżonków podróżujących z pasażerem; zawiera 7 unikalnych etykiet.
- **parch**: `nie jest to zmienna jakościowa`, tylko ilościowa, bo reprezentuje liczbe rodziców/dzieci podróżujących z pasażerem; zawiera 8 unikalnych etykiet.
- **ticket**: : `nie jest zmienną jakościową`; zawieraa 929 unikalnych etykiet; reprezentuje identyfikatory biletów, które nie grupują się w naturalnie rozróżnialne kategorie mające znaczenie analityczne. Obecność dodatkowych liter w niektórych oznaczeniach biletów może wskazywać na specjalne kategorie biletów, co sugeruje, że zmienna ta zawiera ukryte informacje o charakterze biletu. Z tego względu, warto rozważyć redukcję tej zmiennej poprzez stworzenie nowej zmiennej, która będzie skupiała się na ekstrakcji tych zaszytych, istotnych informacji z oryginalnego oznaczenia biletu.
- **fare**: `nie jest zmienną jakościową`, ponieważ reprezentuje cenę biletu w formie liczbowej; zawiera 282 unikalne etykiety.
- **cabin**: `jest zmienną jakościową`, ponieważ lokalizacja kabiny jest kategoryzowana na podstawie oznaczeń, które mogą wskazywać na lokalizację na statku (ciągi znaków, składające się z kombinacji liter i cyfr); zawiera 187 unikalnych etykiet; reprezentuje ona numer / oznaczenie kabiny, w której przebywał pasażer na statku.
- **embarked**: `jest to zmienna jakościowa`, ponieważ miejsce wejścia na statek jest reprezentowane przez kilka kategorii oznaczających porty; zawiera 4 unikalne etykiety (trzy porty oraz jako czwarta wartość to NaN).
- **boat**: `jest to zmienna jakościowa`, ponieważ numery łodzi ratunkowych są kategoryzowane i mogą być analizowane jakościowo w kontekście ewakuacji.; zawiera 28 unikalnych etykiet; oznaczone cyframi lub liczbami co reprezentuje numer łodzi ratunkowej.
- **body**: `nie jest to zmienna jakościowa`, bo każdy numer jest unikalny i służy do celów identyfikacyjnych ofiar katastrofy; zawiera 122 unikalne etykiety.
- **home.dest**: `jest zmienną jakościową`, bo reprezentuje ona cel podróży pasażera i może być analizowane w kontekście geograficznym; zawiera 370 unikalnych etykiet.

UWAGA: jeśli w danej kolumnie znajdują się warości NaN to te wartości są traktowane jak dodatkowa unikalna etykieta, dzięki funkcji .unique().

Wybór zmiennych jakościowych po analizie:

- po przeprowadzeniu analizy zmiennymi jakościowymi są: `pclass, survived, sex, cabin, embarked, boat, home.dest`

Podział zmiennych ze względu na kardynalność:

- `Niska kardynalność`: pclass, survived, sex, embarked
- `Średnia kardynalność`: boat
- `Wysoka kardynalność`: cabin, home.dest

4. Sprawdź, ile unikalnych etykiet ma zmienna mówiąca o kabinie danego pasażera. Użyj takiej funkcji, która zwraca wynik w postaci NumPy array.

[6]:
```
unique_cabins = data['cabin'].unique()

unique_cabins_array = np.array(unique_cabins)

print("Unikalne etykiety kabiny:", unique_cabins_array)
print("Liczba unikalnych etykiet kabiny:", len(unique_cabins_array))
```

```
Unikalne etykiety kabiny: ['B5' 'C22 C26' 'E12' 'D7' 'A36' 'C101' nan 'C62 C64'
'B35' 'A23'
 'B58 B60' 'D15' 'C6' 'D35' 'C148' 'C97' 'B49' 'C99' 'C52' 'T' 'A31' 'C7'
 'C103' 'D22' 'E33' 'A21' 'B10' 'B4' 'E40' 'B38' 'E24' 'B51 B53 B55'
 'B96 B98' 'C46' 'E31' 'E8' 'B61' 'B77' 'A9' 'C89' 'A14' 'E58' 'E49' 'E52'
 'E45' 'B22' 'B26' 'C85' 'E17' 'B71' 'B20' 'A34' 'C86' 'A16' 'A20' 'A18'
 'C54' 'C45' 'D20' 'A29' 'C95' 'E25' 'C111' 'C23 C25 C27' 'E36' 'D34'
 'D40' 'B39' 'B41' 'B102' 'C123' 'E63' 'C130' 'B86' 'C92' 'A5' 'C51' 'B42'
 'C91' 'C125' 'D10 D12' 'B82 B84' 'E50' 'D33' 'C83' 'B94' 'D49' 'D45'
 'B69' 'B11' 'E46' 'C39' 'B18' 'D11' 'C93' 'B28' 'C49' 'B52 B54 B56' 'E60'
 'C132' 'B37' 'D21' 'D19' 'C124' 'D17' 'B101' 'D28' 'D6' 'D9' 'B80' 'C106'
 'B79' 'C47' 'D30' 'C90' 'E38' 'C78' 'C30' 'C118' 'D36' 'D48' 'D47' 'C105'
 'B36' 'B30' 'D43' 'B24' 'C2' 'C65' 'B73' 'C104' 'C110' 'C50' 'B3' 'A24'
 'A32' 'A11' 'A10' 'B57 B59 B63 B66' 'C28' 'E44' 'A26' 'A6' 'A7' 'C31'
 'A19' 'B45' 'E34' 'B78' 'B50' 'C87' 'C116' 'C55 C57' 'D50' 'E68' 'E67'
 'C126' 'C68' 'C70' 'C53' 'B19' 'D46' 'D37' 'D26' 'C32' 'C80' 'C82' 'C128'
 'E39 E41' 'D' 'F4' 'D56' 'F33' 'E101' 'E77' 'F2' 'D38' 'F' 'F G63'
 'F E57' 'F E46' 'F G73' 'E121' 'F E69' 'E10' 'G6' 'F38']
Liczba unikalnych etykiet kabiny: 187
```

5. Zredukuj liczbę cech dla zmiennej opisującej kabiny poprzez zastąpienie obecnych etykiet w formacie LL11 do etykiet zawierających tylko pierwszą literę. Użyj astype(str).str[pozycja]. Nową zmienną nazwij CabinReduced. Wyświetl pierwsze 20 wierszy zbioru danych dla kolumn Cabin i CabinReduced

[7]:
```
data['CabinReduced'] = data['cabin'].astype(str).str[0]
print(data[['cabin', 'CabinReduced']].head(20))
```

```
      cabin CabinReduced
1        B5            B
2   C22 C26            C
3   C22 C26            C
4   C22 C26            C
5   C22 C26            C
6       E12            E
7        D7            D
8       A36            A
9      C101            C
10      NaN            n
11  C62 C64            C
```

```
12  C62 C64           C
13      B35           B
14      NaN           n
15      A23           A
16      NaN           n
17  B58 B60           B
18  B58 B60           B
19      D15           D
20       C6           C
```

6. Wyświetl (jak w pkt 1) liczbę etykiet dla zmiennych z pkt 5. O ile procent zredukowano kardynalność zbioru zmiennej opisującej kabiny?

[8]:
```
columns_reduce = [
    "cabin",
    "CabinReduced"
]

for column in columns_reduce:
    print('Liczba etykiet zmiennej jakościowej {}:{}'.format(
        column, len(data[column].unique())))
```

```
Liczba etykiet zmiennej jakościowej cabin:187
Liczba etykiet zmiennej jakościowej CabinReduced:9
```

Obliczenie procentowej redukcji kardynalnoći zmiannej cabin po jej przekształceniu polega na zliczeniu liczby unikalnych wartości na orginalnej kloumnie cabin oraz na kolumnie przekształconej CabinReduced za pomocą funkcji .nunique(). Następnie obliczamy redukcje kardynalnoći jako różnice pomiędzy liczbą unikalnych wartości kolumny cabin oraz CabinReduced i tą wartość dzielimy liczbę unikalnych etykiet w orginalnej kolumnie cabin. Wszystko mnożymy przez 100, żeby otrzymać wynik procentowy.

[9]:
```
unique_cabin_count = data['cabin'].nunique()
unique_cabin_reduced_count = data['CabinReduced'].nunique()

reduction_percentage = ((unique_cabin_count -
                         unique_cabin_reduced_count) / unique_cabin_count) * 100
print('Zredukowano kardynalność zmiennej cabin o: {:.2f}%'
      .format(reduction_percentage))
```

```
Zredukowano kardynalność zmiennej cabin o: 95.16%
```

7. Uzasadnij dlaczego dokonujesz redukcji akurat tej zmiennej. Jak to wpływa na przyszłe analizy. Czy powoduje jakieś negatywne skutki? Dokonuję redukcji zmiennej opisującej kabiny, ponieważ zawiera ona dużą liczbę unikalnych etykiet, co komplikuje analizy, zaciemnia obraz danych i zwiększa ryzyko przeuczenia modelu uczenia maszynowego. Po wykonaniu redukcji liczba unikalnych wartości spadła ze 187 do 9 kategorii, co stanowi redukcję o około 95%. Taka zmiana znacząco upraszcza analizy i ich interpretację.

Redukcja kardynalności zmiennej opisującej kabiny ma bezpośredni wpływ na przyszłe anal-

izy, przede wszystkim poprzez uproszczenie modelowania danych i zmniejszenie złożoności obliczeniowej. Dzięki redukcji liczby kategorii, modele uczenia maszynowego mogą działać efektywniej, ponieważ mają mniej kategorii do przetwarzania, co zwykle przekłada się na szybsze i bardziej stabilne wyniki. To także ułatwia wizualizacje i interpretację danych, umożliwiając lepsze zrozumienie wzorców i zależności między zmiennymi.

Jednakże, redukcja ta może również prowadzić do utraty istotnych informacji, które mogłyby być wykorzystane w bardziej szczegółowych analizach, takich jak dokładne umiejscowienie kabiny w strukturze statku, co mogło mieć znaczenie podczas ewakuacji. Ponadto, przez agregację danych, różne kabiny o potencjalnie różnych cechach (na przykład różne poziomy luksusu wewnątrz tej samej litery kategorii) są traktowane jako identyczne, co może wprowadzać błędy w analizie.