Titanic_cardinal

April 10, 2025

1 Titanic - cardinality

Kardynalność cech, zwana także licznością cech, odnosi się do liczby różnych wartości, jakie może przyjąć dana cecha w zbiorze danych. Innymi słowy, jest to liczba unikalnych kategorii lub wartości, które może przyjąć dana zmienna.

Na przykład, jeśli mamy cechę "kolor oczu" i w naszym zbiorze danych występują tylko trzy różne wartości: "niebieski", "zielony" i "brązowy", to kardynalność tej cechy wynosi 3.

Kardynalność cech jest ważnym czynnikiem przy analizie danych, ponieważ może wpłynąć na wybór odpowiednich technik przetwarzania danych i modelowania. Na przykład, cechy o dużej kardynalności mogą wymagać specjalnego podejścia, takiego jak kodowanie one-hot, aby być odpowiednio przetworzone przez modele uczenia maszynowego. Natomiast cechy o niskiej kardynalności mogą być łatwiejsze do przetwarzania i interpretacji.

[52]:
```
import pandas as pd
import numpy as np
import arff
import matplotlib.pyplot as plt
```

[53]:
```
titanic = arff.load(open('Titanic.arff', 'r'))
print(titanic.keys())
attributes = titanic["attributes"]
data = titanic["data"]
df = pd.DataFrame(data = data, columns = [x[0] for x in attributes])
df.replace("?", np.nan, inplace=True)
display(df.head(10))
```

```
dict_keys(['description', 'relation', 'attributes', 'data'])
```

	pclass	survived	name	sex
0	1.0	1	Allen, Miss. Elisabeth Walton	female
1	1.0	1	Allison, Master. Hudson Trevor	male
2	1.0	0	Allison, Miss. Helen Loraine	female
3	1.0	0	Allison, Mr. Hudson Joshua Creighton	male
4	1.0	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female
5	1.0	1	Anderson, Mr. Harry	male
6	1.0	1	Andrews, Miss. Kornelia Theodosia	female
7	1.0	0	Andrews, Mr. Thomas Jr	male
8	1.0	1	Appleton, Mrs. Edward Dale (Charlotte Lamson)	female

```
9     1.0        0                          Artagaveytia, Mr. Ramon    male

       age  sibsp  parch    ticket      fare    cabin embarked  boat   body  \
0  29.0000    0.0    0.0     24160  211.3375       B5        S     2    NaN
1   0.9167    1.0    2.0    113781  151.5500  C22 C26        S    11    NaN
2   2.0000    1.0    2.0    113781  151.5500  C22 C26        S  None    NaN
3  30.0000    1.0    2.0    113781  151.5500  C22 C26        S  None  135.0
4  25.0000    1.0    2.0    113781  151.5500  C22 C26        S  None    NaN
5  48.0000    0.0    0.0     19952   26.5500      E12        S     3    NaN
6  63.0000    1.0    0.0     13502   77.9583       D7        S    10    NaN
7  39.0000    0.0    0.0    112050    0.0000      A36        S  None    NaN
8  53.0000    2.0    0.0     11769   51.4792     C101        S     D    NaN
9  71.0000    0.0    0.0  PC 17609   49.5042     None        C  None   22.0

                         home.dest
0                     St Louis, MO
1  Montreal, PQ / Chesterville, ON
2  Montreal, PQ / Chesterville, ON
3  Montreal, PQ / Chesterville, ON
4  Montreal, PQ / Chesterville, ON
5                    New York, NY
6                      Hudson, NY
7                     Belfast, NI
8              Bayside, Queens, NY
9              Montevideo, Uruguay
```

1.1 Zad 1

Sprawdź liczebność poszczególnych etykiet dla danych zmiennych jakościowych

[54]:
```
display(df.columns)
cardinal_col = ["pclass", "survived", "sex", "age", "sibsp", "parch", "cabin",␣
 ↪"embarked", "boat", "home.dest", "fare" ]
for i in cardinal_col:
    unique_vals = df[i].nunique(dropna=True)
    print(f"Liczba etykiet zmiennej {i}: {unique_vals}")
```

```
Index(['pclass', 'survived', 'name', 'sex', 'age', 'sibsp', 'parch', 'ticket',
       'fare', 'cabin', 'embarked', 'boat', 'body', 'home.dest'],
      dtype='object')

Liczba etykiet zmiennej pclass: 3
Liczba etykiet zmiennej survived: 2
Liczba etykiet zmiennej sex: 2
Liczba etykiet zmiennej age: 98
Liczba etykiet zmiennej sibsp: 7
Liczba etykiet zmiennej parch: 8
Liczba etykiet zmiennej cabin: 186
Liczba etykiet zmiennej embarked: 3
```

```
Liczba etykiet zmiennej boat: 27
Liczba etykiet zmiennej home.dest: 369
Liczba etykiet zmiennej fare: 281
```

1.2 Zad 2

Wyświetl z użyciem funkcji print liczbę wszystkich pasażerów.

[55]:
```
print("Liczba wszystkich pasażerów: {}".format(df.shape[0]))
```

```
Liczba wszystkich pasażerów: 1309
```

1.3 Zad 3

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

Łączna liczba pasażerów wynosi 1309. Jesteśmy w stanie podzielić zmienne na te o małej kardynalności oraz te o dużej kardynalności. Zmienne o małej kardynalności: - pclass - survived - sex - sibsp - parch - embarked

Zmienne o dużej kardynalności: - age - cabin - boat - home.dest - fare

1.4 Zad 4

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

[56]:
```
cabin_cardinal = df["cabin"].unique()
print(f"Unikalne etykiety kolumny kabiny to: {cabin_cardinal}")
number_of_cabin_cardinal = len(cabin_cardinal)
print(f"Liczba unikalnych etykiet kolumny kabiny wynosi:␣
 ↪{number_of_cabin_cardinal}")
```

```
Unikalne etykiety kolumny kabiny to: ['B5' 'C22 C26' 'E12' 'D7' 'A36' 'C101'
None '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 kolumny kabiny wynosi: 187
```

1.5 Zad 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

[57]:
```
df["CabinReduced"] = df["cabin"].astype(str).str[0]

df.loc[df["CabinReduced"] == "<", "CabinReduced"] = pd.NA # w przypadku␣
 ↪pojawienia sie tego znaku zamieniamy na nan

CabinReduced = df["CabinReduced"].unique()
print("Unikalne etykiety kabin:", CabinReduced)
display(df[["cabin", "CabinReduced"]].head(20))
```

```
Unikalne etykiety kabin: ['B' 'C' 'E' 'D' 'A' 'N' 'T' 'F' 'G']
```

	cabin	CabinReduced
0	B5	B
1	C22 C26	C
2	C22 C26	C
3	C22 C26	C
4	C22 C26	C
5	E12	E
6	D7	D
7	A36	A
8	C101	C
9	None	N
10	C62 C64	C
11	C62 C64	C
12	B35	B
13	None	N
14	A23	A
15	None	N
16	B58 B60	B
17	B58 B60	B
18	D15	D
19	C6	C

1.6 Zad 6

Wyświetl liczbę etykiet dla zmiennych z ptk 5. O ile procent zredukowano kardynalność zbioru zmiennej opisującej kabiny?

[58]:
```
cabin = df["cabin"].nunique()
cabin_after_reduce = df["CabinReduced"].nunique()
print(f"Liczba etykiet dla zmiennej cabin: {cabin}")
print(f"Liczba etykiet dla zmiennej CabinReduced: {cabin_after_reduce}")
percent_diff = (1 - cabin_after_reduce / cabin) * 100
print(f"Redukcja kardynalności zbioru zmiennej opisującej kabiny nastąpiła o␣
  ↪{percent_diff:.2f} %")
```

```
Liczba etykiet dla zmiennej cabin: 186
Liczba etykiet dla zmiennej CabinReduced: 9
Redukcja kardynalności zbioru zmiennej opisującej kabiny nastąpiła o 95.16 %
```

1.7 Zad 7

Uzasadnij dlaczego dokonujesz redukcji akurat tej zmiennej. Jak to wpływa na przyszłe analizy. Czy powoduje jakieś negatywne skutki?

Dokonujemy redukcji zmiennej cabin ponieważ ma ona bardzo wysoką kardynalność. Może to utrudniać dalszą analizę, w tym m.in np tworzenie wykresów. Każda kabina zawiera informacje np "E12" itp co można zredukować do oznaczenia poprzez sam pierwszy znak - pierwszą literę.

Zredukowana zmienna cabin sprawia, że praca nad analizą tych danych będzie bardziej przejrzysta. Nie zmienia to faktu, że mając mniej typów tej zmiennej jesteśmy w stanie określić np, czy typ kabiny miał wpływ na liczbę osób, któe zginęły w katastrofie.

Natomiast, jeśli chcielibyśmy przeprowadzić bardzo szczegółową analizę w oparciu na typ kabiny, to lepiej mieć więcej typów, wtedy jesteśmy w stanie dostrzec więcej zależności i detali, któe mogłyby sięukryć, używając zredukowanej zmiennej. Nie mamy wtedy dokładnej informacji o lokalizacji danego pasażera.