Marzena Nowakowska, Wydział Zarządzania i Modelowania Komputerowego, PŚk

Temat 8. Wybrane aspekty eksploracji danych
Biblioteka *pandas* oraz biblioteki wspomagające

Przykładowa eksploracja danych i modelowanie za pomocą klasyfikacyjnych drzew decyzyjnych są realizowane na zbiorze *titanic.csv* zawierającym dane jednostkowe o pasażerach tragicznego rejsu statku "Titanic". Zbiór jest opisany przez następujące cechy (zmienne):

- **survival** – czy pasażer przeżył (1-tak, 0-nie)
- **pclass** – clasa, którą podróżował (1-najlepsza)
- **name** – imię, nazwisko
- **sex** – płeć
- **age** – wiek
- **sibsp** – liczba rodzeństwa /małżonków na pokładzie (*Number of Siblings/Spouses Aboard*)
- **parch** – liczba dzieci lub rodziców
- **ticket** – numer biletu
- **fare** – cena biletu
- **cabin** – numer kajuty
- **embarked** – port, w którym osoba wsiadła na pokład (C = Cherbourg; Q = Queenstown; S = Southampton)
- **boat** – numer łódki, którą osoba była ewakuowana
- **body** – numer identyfikacyjny znalezionego ciała

Skopiować plik *titanic.csv* do swojego folderu z danymi.

Zad. 1.
Pobrać do pliku *Zad08_01* treść programu *Zad08-TitanicExploration.py*. Wykonać analizę programu. Odpowiedzieć na pytania zawarte w komentarzach wyróżnione wyrazami pisanymi wielkimi literami.

Zad. 2.
Pobrać do pliku *Zad08_02* treść programu *Zad08-TitanicTree.py*. Wykonać analizę programu.
Przygotować odpowiedzi na pytania:

- jak działa metoda *replace* (linia 34)
- co oznacza i jak działa wyrażenie *df.isnull().sum()* (linia 51)
- objaśnić parametry *how* i *inplace* metody *dropna* (linia 54)
- co jest wynikiem wywołania: *train_test_split(df, test_size=0.2)* (linia 57)
- sprawdzić działanie różnych wersji wykresu drzewa decyzyjnego; zdefiniować wywołanie wg uznania (poszukać wskazówek i Internecie)

Zad. 3.
Utworzyć w pliku *Zad08_03* kopię programu *Zad08_02*. Wprowadzić we właściwym miejscu podany niżej fragment programu kategoryzujący zmienne ilościowe (numeryczne). Zbudować drzewo decyzyjne, w którym oryginalne wartości są zastąpione ich kodowanymi odpowiednikami (dlaczego nie potrzeba modyfikować instrukcji przypisania do X?). Utworzyć wykres uzyskanego drzewa definiując swoje parametry wykresu. Sprawdzić, na ile drzewo wynikowe różni się od tego z zad. 2.

```
# Kategoryzacja cech ilościowych; wykorzystanie właściwości 'loc' obiektu DataFrame
df.loc[ df['fare'] <= 7.91, 'fare'] = 0
df.loc[(df['fare'] > 7.91) & (df['fare'] <= 14.454), 'fare'] = 1
df.loc[(df['fare'] > 14.454) & (df['fare'] <= 31), 'fare'] = 2
df.loc[ df['fare'] > 31, 'fare'] = 3

df.loc[ df['age'] <= 16, 'age'] = 0
df.loc[(df['age'] > 16) & (df['age'] <= 32), 'age'] = 1
df.loc[(df['age'] > 32) & (df['age'] <= 48), 'age'] = 2
df.loc[(df['age'] > 48) & (df['age'] <= 64), 'age'] = 3
df.loc[ df['age'] > 64, 'age'] =4
```

Zad. 4.
Doczytać w Internecie, jak można sprawdzić jakość drzewa klasyfikującego wykorzystując zbiór testowy. Uzupełnić program *Zad08_03* o ten element rozwiązania.