MushroomDT

June 10, 2018

Zaadowanie bibliotek

```
In [1]: import pandas as pd
        import numpy as np
        import seaborn as sns
        %matplotlib inline
        from matplotlib import pyplot as plt

        #Ustalenie stylu wykresów jako ggplot
        # plt.style.use('ggplot')

        from sklearn import tree
```

Wczytanie danych

```
In [2]: # Ustalenie cieki do datasetu
        filename_mushrooms = './agaricus-lepiota.csv'

        # Wczytanie datasetu jako dataframe
        mushrooms_dataframe = pd.read_csv(filename_mushrooms, sep=";")

        # Wywietlenie dataframe
        display(mushrooms_dataframe)
```

```
    classes cap-shape cap-surface cap-color bruises odor gill-attachment  \
0         p         x           s         n       t    p               f
1         e         x           s         y       t    a               f
2         e         b           s         w       t    l               f
3         p         x           y         w       t    p               f
4         e         x           s         g       f    n               f
5         e         x           y         y       t    a               f
6         e         b           s         w       t    a               f
7         e         b           y         w       t    l               f
8         p         x           y         w       t    p               f
9         e         b           s         y       t    a               f
10        e         x           y         y       t    l               f
11        e         x           y         y       t    a               f
12        e         b           s         y       t    a               f
```

13	p	x	y	w	t	p	f
14	e	x	f	n	f	n	f
15	e	s	f	g	f	n	f
16	e	f	f	w	f	n	f
17	p	x	s	n	t	p	f
18	p	x	y	w	t	p	f
19	p	x	s	n	t	p	f
20	e	b	s	y	t	a	f
21	p	x	y	n	t	p	f
22	e	b	y	y	t	l	f
23	e	b	y	w	t	a	f
24	e	b	s	w	t	l	f
25	p	f	s	w	t	p	f
26	e	x	y	y	t	a	f
27	e	x	y	w	t	l	f
28	e	f	f	n	f	n	f
29	e	x	s	y	t	a	f
...	...	...	...	...	...	...	...
8094	e	b	s	g	f	n	f
8095	p	x	y	c	f	m	f
8096	e	k	f	w	f	n	f
8097	p	k	y	n	f	s	f
8098	p	k	s	e	f	y	f
8099	e	k	f	w	f	n	f
8100	e	f	s	n	f	n	a
8101	p	k	s	e	f	s	f
8102	e	x	s	n	f	n	a
8103	e	k	s	n	f	n	a
8104	e	k	s	n	f	n	a
8105	e	k	s	n	f	n	a
8106	e	k	s	n	f	n	a
8107	e	x	s	n	f	n	a
8108	p	k	y	e	f	y	f
8109	e	b	s	w	f	n	f
8110	e	x	s	n	f	n	a
8111	e	k	s	w	f	n	f
8112	e	k	s	n	f	n	a
8113	p	k	y	e	f	y	f
8114	p	f	y	c	f	m	a
8115	e	x	s	n	f	n	a
8116	p	k	y	n	f	s	f
8117	p	k	s	e	f	y	f
8118	p	k	y	n	f	f	f
8119	e	k	s	n	f	n	a
8120	e	x	s	n	f	n	a
8121	e	f	s	n	f	n	a
8122	p	k	y	n	f	y	f
8123	e	x	s	n	f	n	a

	gill-spacing	gill-size	gill-color	...	stalk-surface-below-ring	\
0	c	n	k	...	s	
1	c	b	k	...	s	
2	c	b	n	...	s	
3	c	n	n	...	s	
4	w	b	k	...	s	
5	c	b	n	...	s	
6	c	b	g	...	s	
7	c	b	n	...	s	
8	c	n	p	...	s	
9	c	b	g	...	s	
10	c	b	g	...	s	
11	c	b	n	...	s	
12	c	b	w	...	s	
13	c	n	k	...	s	
14	w	b	n	...	f	
15	c	n	k	...	s	
16	w	b	k	...	s	
17	c	n	n	...	s	
18	c	n	n	...	s	
19	c	n	k	...	s	
20	c	b	k	...	s	
21	c	n	n	...	s	
22	c	b	k	...	s	
23	c	b	w	...	s	
24	c	b	g	...	s	
25	c	n	n	...	s	
26	c	b	n	...	s	
27	c	b	w	...	s	
28	c	n	k	...	s	
29	w	n	n	...	s	
...	...	...	...	...	...	
8094	w	b	g	...	s	
8095	c	b	y	...	y	
8096	w	b	w	...	s	
8097	c	n	b	...	k	
8098	c	n	b	...	k	
8099	w	b	w	...	k	
8100	c	b	o	...	s	
8101	c	n	b	...	s	
8102	c	b	y	...	s	
8103	c	b	y	...	s	
8104	c	b	y	...	s	
8105	c	b	y	...	s	
8106	c	b	o	...	s	
8107	c	b	y	...	s	
8108	c	n	b	...	s	

8109	w	b	w	...	s
8110	c	b	o	...	s
8111	w	b	p	...	s
8112	c	b	o	...	s
8113	c	n	b	...	k
8114	c	b	y	...	y
8115	c	b	y	...	s
8116	c	n	b	...	k
8117	c	n	b	...	s
8118	c	n	b	...	s
8119	c	b	y	...	s
8120	c	b	y	...	s
8121	c	b	n	...	s
8122	c	n	b	...	k
8123	c	b	y	...	s

	stalk-color-above-ring	stalk-color-below-ring	veil-type	veil-color	\
0	w	w	p	w	
1	w	w	p	w	
2	w	w	p	w	
3	w	w	p	w	
4	w	w	p	w	
5	w	w	p	w	
6	w	w	p	w	
7	w	w	p	w	
8	w	w	p	w	
9	w	w	p	w	
10	w	w	p	w	
11	w	w	p	w	
12	w	w	p	w	
13	w	w	p	w	
14	w	w	p	w	
15	w	w	p	w	
16	w	w	p	w	
17	w	w	p	w	
18	w	w	p	w	
19	w	w	p	w	
20	w	w	p	w	
21	w	w	p	w	
22	w	w	p	w	
23	w	w	p	w	
24	w	w	p	w	
25	w	w	p	w	
26	w	w	p	w	
27	w	w	p	w	
28	w	w	p	w	
29	w	w	p	w	
...	...	...	...	...	

8094	w	w	p	w
8095	c	c	p	w
8096	w	w	p	w
8097	p	p	p	w
8098	w	p	p	w
8099	w	w	p	w
8100	o	o	p	n
8101	p	w	p	w
8102	o	o	p	n
8103	o	o	p	n
8104	o	o	p	o
8105	o	o	p	n
8106	o	o	p	o
8107	o	o	p	o
8108	p	w	p	w
8109	w	w	p	w
8110	o	o	p	o
8111	w	w	p	w
8112	o	o	p	n
8113	p	p	p	w
8114	c	c	p	w
8115	o	o	p	o
8116	p	w	p	w
8117	p	w	p	w
8118	p	w	p	w
8119	o	o	p	o
8120	o	o	p	n
8121	o	o	p	o
8122	w	w	p	w
8123	o	o	p	o

	ring-number	ring-type	spore-print-color	population	habitat
0	o	p	k	s	u
1	o	p	n	n	g
2	o	p	n	n	m
3	o	p	k	s	u
4	o	e	n	a	g
5	o	p	k	n	g
6	o	p	k	n	m
7	o	p	n	s	m
8	o	p	k	v	g
9	o	p	k	s	m
10	o	p	n	n	g
11	o	p	k	s	m
12	o	p	n	s	g
13	o	p	n	v	u
14	o	e	k	a	g
15	o	p	n	y	u

16	o	e	n	a	g
17	o	p	k	s	g
18	o	p	n	s	u
19	o	p	n	s	u
20	o	p	n	s	m
21	o	p	n	v	g
22	o	p	n	s	m
23	o	p	n	n	m
24	o	p	k	s	m
25	o	p	n	v	g
26	o	p	n	n	m
27	o	p	n	n	m
28	o	p	k	y	u
29	o	p	n	v	d
...	...	...	...	...	...
8094	t	p	w	n	g
8095	n	n	w	c	d
8096	t	p	w	n	g
8097	o	e	w	v	l
8098	o	e	w	v	d
8099	t	p	w	s	g
8100	o	p	b	v	l
8101	o	e	w	v	p
8102	o	p	n	c	l
8103	o	p	o	c	l
8104	o	p	n	v	l
8105	o	p	y	v	l
8106	o	p	n	v	l
8107	o	p	n	c	l
8108	o	e	w	v	l
8109	t	p	w	n	g
8110	o	p	n	v	l
8111	t	p	w	n	g
8112	o	p	b	v	l
8113	o	e	w	v	d
8114	n	n	w	c	d
8115	o	p	o	v	l
8116	o	e	w	v	l
8117	o	e	w	v	d
8118	o	e	w	v	d
8119	o	p	b	c	l
8120	o	p	b	v	l
8121	o	p	b	c	l
8122	o	e	w	v	l
8123	o	p	o	c	l

[8124 rows x 23 columns]

Zbiór danych ma 8124 wiersze i 23 kolumny (pierwsza kolumna to atrybut decyzyjny, a pozostae 22 kolumny to atrybuty warunkowe). W celu dalszego zbadania datasetu i weryfikacji typów danych kategorycznych w kadej kolumnie, wypisano unikalne wartoci kadej kolumny. Sprawdzono równie, czy zbiór danych zawiera brakujce wartoci lub niepotrzebne kolumny.

```
In [3]: print("Liczba rónych wartoci atrybutów dla kadej kolumny:")
        for x in mushrooms_dataframe.columns:
            x_unique = mushrooms_dataframe[x].unique()
            print("{:>25}: {:>2} {}".format(x, x_unique.shape[0], x_unique))
```

```
Liczba rónych wartoci atrybutów dla kadej kolumny:
                  classes:  2 ['p' 'e']
                cap-shape:  6 ['x' 'b' 's' 'f' 'k' 'c']
              cap-surface:  4 ['s' 'y' 'f' 'g']
                cap-color: 10 ['n' 'y' 'w' 'g' 'e' 'p' 'b' 'u' 'c' 'r']
                  bruises:  2 ['t' 'f']
                     odor:  9 ['p' 'a' 'l' 'n' 'f' 'c' 'y' 's' 'm']
          gill-attachment:  2 ['f' 'a']
             gill-spacing:  2 ['c' 'w']
                gill-size:  2 ['n' 'b']
               gill-color: 12 ['k' 'n' 'g' 'p' 'w' 'h' 'u' 'e' 'b' 'r' 'y' 'o']
              stalk-shape:  2 ['e' 't']
               stalk-root:  5 ['e' 'c' 'b' 'r' '?']
 stalk-surface-above-ring:  4 ['s' 'f' 'k' 'y']
 stalk-surface-below-ring:  4 ['s' 'f' 'y' 'k']
   stalk-color-above-ring:  9 ['w' 'g' 'p' 'n' 'b' 'e' 'o' 'c' 'y']
   stalk-color-below-ring:  9 ['w' 'p' 'g' 'b' 'n' 'e' 'y' 'o' 'c']
               veil-type:  1 ['p']
              veil-color:  4 ['w' 'n' 'o' 'y']
             ring-number:  3 ['o' 't' 'n']
               ring-type:  5 ['p' 'e' 'l' 'f' 'n']
        spore-print-color:  9 ['k' 'n' 'u' 'h' 'w' 'r' 'o' 'y' 'b']
              population:  6 ['s' 'n' 'a' 'v' 'y' 'c']
                 habitat:  7 ['u' 'g' 'm' 'd' 'p' 'w' 'l']
```

Zauwaono, e sporód 22 atrybutów warunkowych, jedynie ‘veil-type’ zawiera tylko jedn warto “p”. Zatem atrybut ten nie zapewnia adnej wartoci dodanej do klasyfikatora. Podjto decyzj o usuniciu tej kolumny - utworzono generyczny kod usuwajcy wszystkie kolumny zawierajce jedn warto.

```
In [4]: print("Rozmiar mushrooms_dataframe przed usuniciem: ",mushrooms_dataframe.shape)

        # Usunicie kolumn zawierajcych jedn wartoc
        for col in mushrooms_dataframe.columns.values:
            col_unique = mushrooms_dataframe[col].unique()
            if len(col_unique) == 1:
                print("Usunito kolumn '{}',która zawiera tylko jedn warto: {}".format(col, col
                mushrooms_dataframe = mushrooms_dataframe.drop(col, 1)
```

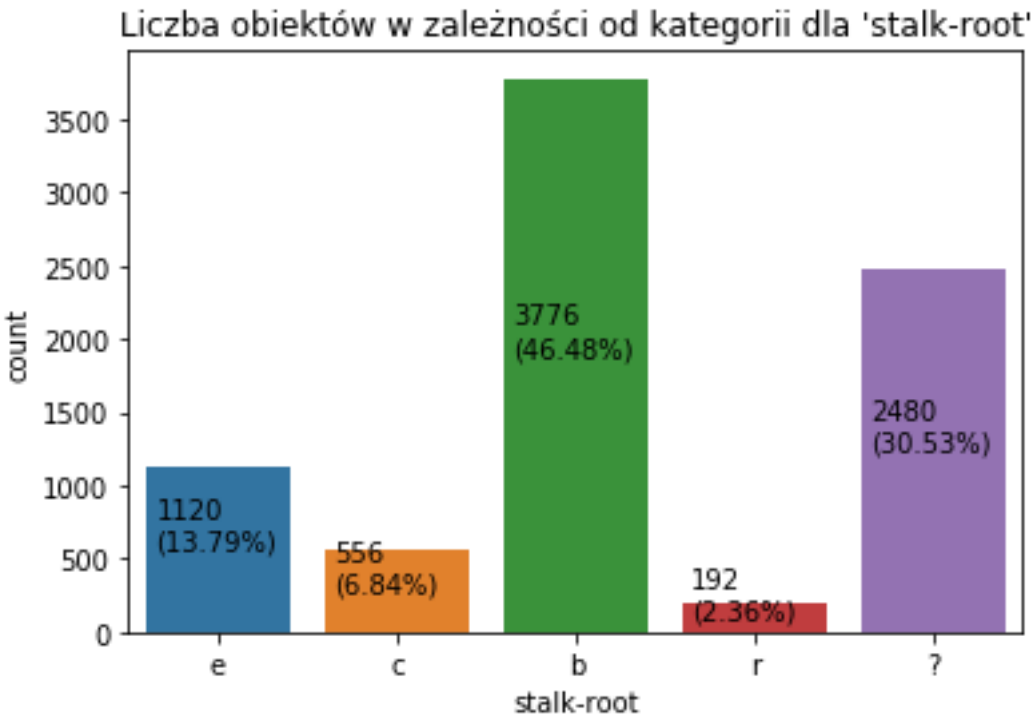
```
        print("Rozmiar mushrooms_dataframe po usuniciu: ",mushrooms_dataframe.shape)
```

```
Rozmiar mushrooms_dataframe przed usuniciem:  (8124, 23)
Usunito kolumn 'veil-type',która zawiera tylko jedn warto: p
Rozmiar mushrooms_dataframe po usuniciu:  (8124, 22)
```

Stwierdzono równie, e kolumna ‘stalk-root’ zawiera brakujce wartoci. Zbadano udzia brakujcych wartoci w zbiorze - utworzono generyczny kod badajcy udziay brakujych wartoci.

```
In [44]: for x in mushrooms_dataframe.columns:
             x_unique = mushrooms_dataframe[x].unique()
             if '?' in x_unique:
                 column = mushrooms_dataframe[x]
                 column_count = column.count()
                 column_value_count = column.value_counts()

                 print("Liczba obiektów w zalenoci od kategorii i ich udzia procentowy dla klas
                 stat = column_value_count.to_frame()
                 stat['percent'] = 100. * column_value_count / column_count
                 print(stat)

                 fig = plt.figure()
                 fig.patch.set_facecolor('xkcd:white')
                 ax = sns.countplot(x=x, data=mushrooms_dataframe)
                 ax.set_title("Liczba obiektów w zalenoci od kategorii dla '{}'".format(x))

         #          for p in ax.patches:
         #              height = p.get_height()
         #              ax.text(p.get_x()+0.25, height+ 3, 'n=%.0f'%(height))

                 for p in ax.patches:
                     ax.annotate('{:.0f}\n({:.2f}%)'.format(p.get_height(), 100. * p.get_height
                 plt.show()
```

```
Liczba obiektów w zalenoci od kategorii i ich udzia procentowy dla klasy 'stalk-root':

   stalk-root    percent
b        3776  46.479567
?        2480  30.526834
e        1120  13.786312
c         556   6.843919
r         192   2.363368
```



Moliwe dziaania do podjcia w pzypadku wystpowania brakujcych danych to m.in. usunicie kolumn lub wierszy zawierajcych brakujce dane, wypeninie brakujcych wartoci inn wartocia np. z poprzedniej lub nastpnej komórki. Stwierdzono, e udzial procentowy brakujcych wartoci ('?') dla atrybutu 'stalk-root' wynosi ponad 30,5%. Podjeto decyzj o usuniciu tej kolumny.

Utworzono genryczny kod oczyszczajcy zbiór danych z kolumn zawierajcych brakujce wartoci powyej zadanego progu procentowego (25%).

```
In [6]: for x in mushrooms_dataframe.columns:
            x_unique = mushrooms_dataframe[x].unique()
            if '?' in x_unique:
                print("Number of rows with missing values in column '{}': {}".format(x, mushroo

        mushrooms_dataframe_dropped_rows = mushrooms_dataframe.copy(deep=True)
        for x in mushrooms_dataframe.columns:
            mushrooms_dataframe_dropped_rows = mushrooms_dataframe_dropped_rows[mushrooms_data
        print("mushrooms_dataframe_dropped_rows: ",mushrooms_dataframe_dropped_rows.shape)

        # Próg procentowy usuwania kolumn z brakujcymi wartociami
        drop_percentage = 0.25

        mushrooms_dataframe_dropped_cols = mushrooms_dataframe.copy(deep=True)
        for col in mushrooms_dataframe_dropped_cols:
            mushrooms_dataframe_dropped_cols.loc[mushrooms_dataframe_dropped_cols[col] == '?',

        for col in mushrooms_dataframe_dropped_cols.columns.values:
```

```
            no_rows = mushrooms_dataframe_dropped_cols[col].isnull().sum()
            percentage = no_rows / mushrooms_dataframe_dropped_cols.shape[0]
            if percentage >= drop_percentage:
                del mushrooms_dataframe_dropped_cols[col]
                print("Column {} contains {} missing values. This is {} percent. Dropping this
```

```
Number of rows with missing values in column 'stalk-root': 2480
mushrooms_dataframe_dropped_rows:  (5644, 22)
Column stalk-root contains 2480 missing values. This is 0.3052683407188577 percent. Dropping tl
```