[PSZT] Drzewo decyzyjne ID-3 “grzyby z ruletką”

1) Autorzy:

a) Rafał Lewanczyk
b) Kacper Biegajski

https://github.com/rafallewanczyk/id3-grzyby

2) Treść zadania:

a) Budowa dwóch drzew decyzyjnych przy pomocy algorytmu ID-3, jedno standardowe drugie posługujące się ruletką w celu wyboru testu
b) Analiza działania oraz dopasowania do danych obu drzew, porównanie ich zachowań

3) Wkład w projekt:

a) Rafał Lewanczyk:
   i) Implementacja standardowego algorytmu ID-3
   ii) Przeprowadzanie testów
b) Kacper Biegajski:
   i) Implementacja ruletki
   ii) Wyświetlanie drzewa
   iii) Obsługa danych wejściowych
   iv) Walidacja krzyżowa

4) Decyzje projektowe:

a) Przy każdym uruchomieniu dane są wczytywane z pliku oraz przed wykonaniem walidacji mieszane
b) Aby program zadziałał prawidłowo plik zawierający dane musi w pierwszym wierszu zawierać listę atrybutów na której klasa jest jako pierwsza oraz w kolejnych wierszach stany oddzielone przecinkiem

5) Wykorzystane biblioteki:

a) python 3.7.3
b) biblioteka networkx - reprezentacja grafu
c) biblioteka pandas - operacje na danych

reszta modułów zawarta w pliku *requirements.txt*

6) Wykonane instrukcje:

a) python main.py -d data/expanded -k 3
b) python main.py -d data/expanded -k 6
c) python main.py -d data/expanded -k 10

7) Cel:

Celem tego eksperymentu jest zbadanie jakości drzewa decyzyjnego zbudowanego przez ruletkowy wybór testów oraz porównanie go do standardowego drzewa decyzyjnego zbudowanego przy pomocy algorytmu ID3.

8) Teza:

Drzewo zbudowane przy pomocy standardowego algorytmu popełnia mniej błędów.

9) Wyniki eksperymentów

W naszych eksperymentach badamy współczynnik błędu *q* jaki popełniają drzewa. Wykorzystujemy do tego celu k-krotną walidację krzyżową

$$q = \frac{ilość\ popełnionych\ błędów}{ilość\ walidowanych\ rekordów} \qquad \overline{q} = \frac{1}{k}\sum_{i=1}^{k} q_i$$

Tabele przedstawiają współczynnik błędu dla k-tej iteracji walidacji krzyżowej dla poszczególnych algorytmów.

Dla k=3:

k	**ID3 standard**	**ID3 ruletka**
1	0.0	0.0
2	0.0	0.0
3	0.0	0.00215
avg	0.0	0.00072

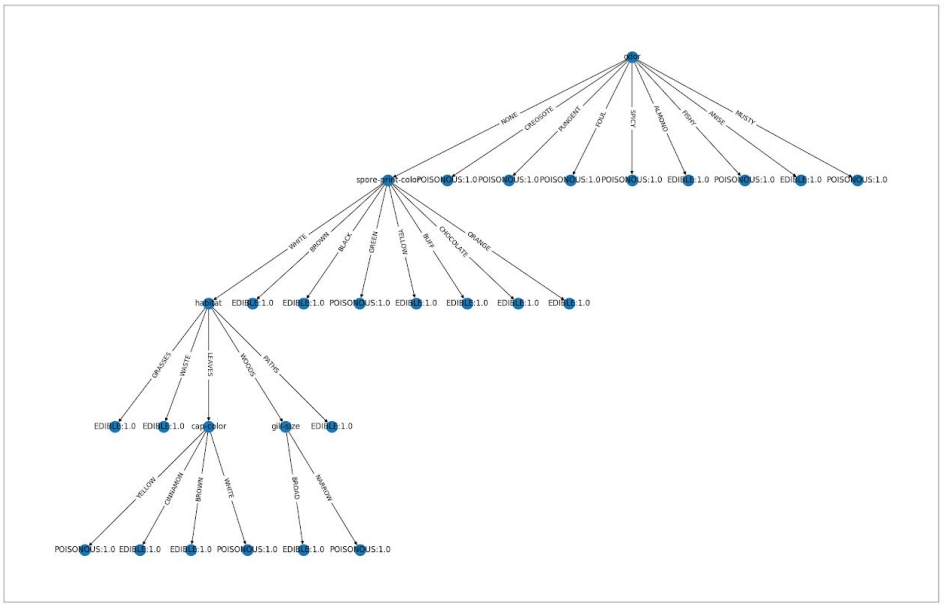
Dla k = 6

k	**ID3 standard**	**ID3 ruletka**
1	0.0	0.0
2	0.0	0.0
3	0.0	0.0
4	0.0	0.0
5	0.0	0.00215
6	0.0	0.0
avg	0.0	0.00035

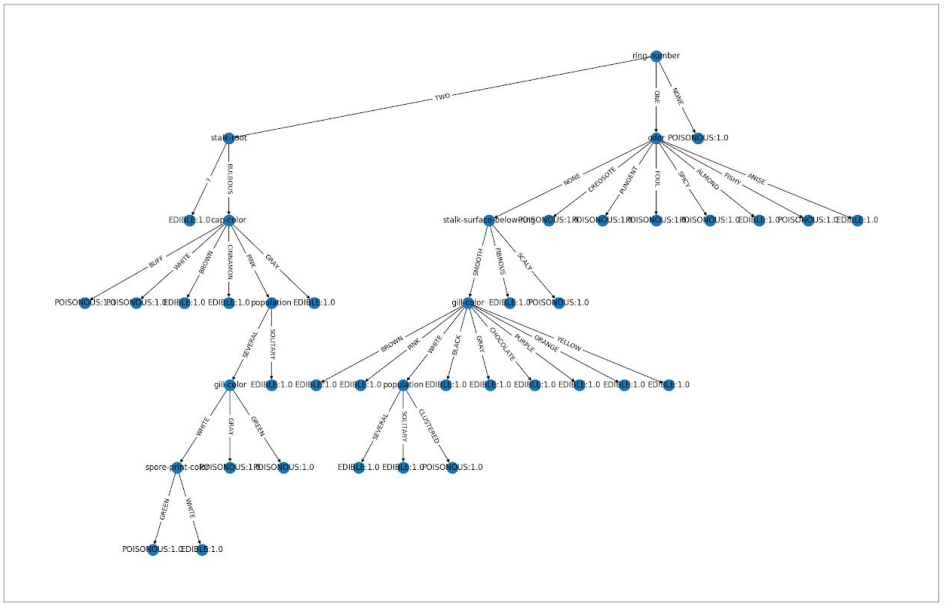
Dla k=10:

k	**ID3 standard**	**ID3 ruletka**
1	0.0	0.0
2	0.0	0.0
3	0.0	0.0
4	0.0	0.00238
5	0.0	0.0
6	0.0	0.0
7	0.0	0.0
8	0.0	0.00238
9	0.0	0.0
10	0.0	0.0
avg	0.0	0.00047

Jedno z drzew wygenerowanych przez standardowy algorytm ID3



Jedno z drzew wygenerowanych przez algorytm ID3 z ruletką



10) Omówienie wyników eksperymentu:

a) Standardowy algorytm nie popełnił żadnych błędów
b) Standardowy algorytm buduje minimalne drzewo natomiast z bardzo dużym prawdopodobieństwem wygeneruje drzewo nieoptymalne z większą ilością wierzchołków (na przykładzie normalny algorytm - 28 wierzchołków, ruletkowy - 42). Pogarsza to czas budowy drzewa ruletkowego oraz wykonywania na nim zapytań.
c) Algorytm posługujący się ruletką popełnia nieznaczne błędy.
d) Widać lekki wzrost błędów algorytmu ruletkowego przy większych danych treningowych (przejście z k=6 na k=10)

11) Wnioski

a) Standardowy algorytm ID3 radzi sobie lepiej od implementacji z ruletką.
b) Można spodziewać się wraz ze wzrostem danych treningowych mniejszej ilości błędów, jednak w algorytmie z ruletką mamy czynnik losowy, przez co nie zawsze będzie to spełnione.
c) Standardowy algorytm ID3 działa szybciej.