

# [PSZT-U] Magic mushrooms

Piotr Frątczak  
(300207)

Bartosz Świtalski  
(300279)

8 stycznia 2021

## 1 Opis problemu

Implementacja przewidywania jadalności grzyba<sup>1</sup> za pomocą algorytmu ID3.

## 2 Decyzje projektowe

- dane trenujące pochodzą ze [strongy](#)

## 3 Cele eksperymentu

Implementacja algorytmu konstruowania drzewa decyzyjnego ID3 z testami binarnymi.

## 4 Użycie

```
/magic-mushrooms$
```

```
cd magic-mushrooms/  
python3 main.py <plik> <indeks> <separator>  
<proporcja>
```

### Oznaczenia argumentów

**plik** - nazwa pliku umieszczonego w katalogu `data/`, np. *agaricus-lepiota.data*

**indeks** - indeks pozycji klasyfikatora w liście atrybutów, np. 0

**separator** - separator wartości atrybutów w pliku, np. ,

**proporcja** - proporcja danych z pliku użytych jako zbiór trenujący, np. 0.8

---

<sup>1</sup>rząd pieczarkowców ([agaricales](#))

## Komentarz do użycia

Po uruchomieniu skryptu na wyjściu pojawi się informacja o proporcjach podziału zbioru uczącego oraz o skuteczności zdolności algorytmu do przewidywania jadalności grzyba (w zakresie  $[0; 1]$ ).

## 5 Testowanie

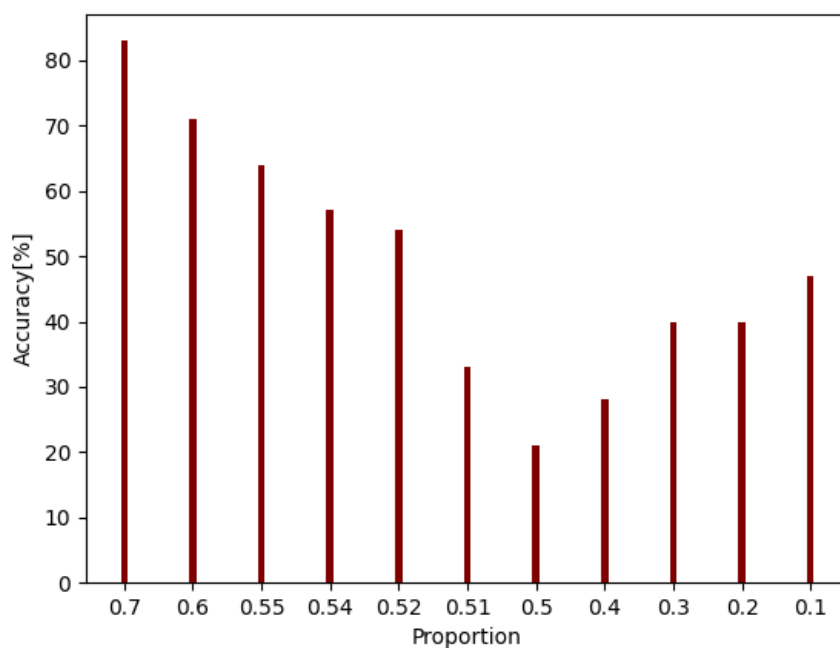
- Ustalona proporcja zbioru trenującego do zbioru testującego to 80% : 20%
- Po zbadaniu dokładności dla ustalonej proporcji testowaliśmy wpływ wielkości zbioru trenującego na dokładność klasyfikacji.
- Testowaliśmy również predykcję bez uwzględnienia najbardziej wartościowych atrybutów (tych, dla których przyrost informacji był największy) na dokładność klasyfikacji.

## 6 Wyniki

Dla standardowych proporcji dokładność klasyfikacji wyniosła  $\approx 97\%$ .

### Zmniejszanie rozmiaru zbioru trenującego

proporcja zbiór trenujący:zbiór testujący	dokładność klasyfikacji
70% : 30%	$\approx 83\%$
60% : 40%	$\approx 71\%$
55% : 45%	$\approx 64\%$
54% : 46%	$\approx 57\%$
52% : 48%	$\approx 54\%$
51% : 49%	$\approx 33\%$
50% : 50%	$\approx 21\%$
40% : 60%	$\approx 28\%$
30% : 70%	$\approx 40\%$
20% : 80%	$\approx 40\%$
10% : 90%	$\approx 47\%$



## Analiza dokładności bez wybranych atrybutów

Dla standardowych proporcji (80 : 20) badaliśmy wpływ nieuwzględnienia kolejno 5 *najlepszych* atrybutów (pod względem przyrostu informacji).

5 najlepszych atrybutów to: *gill-attachment* (indeks 6), *gill-spacing* (7), *gill-size*(8), *gill-color* (9) oraz *cap-color* (3).

ignorowany atrybut	ranga atrybutu	dokładność klasyfikacji z pominięciem atrybutów
<i>gill-attachment</i>	1	$\approx 97\%$
<i>gill-spacing</i>	2	$\approx 98\%$
<i>gill-size</i>	3	$\approx 97\%$
<i>gill-color</i>	4	$\approx 97\%$
<i>cap-color</i>	5	$\approx 89\%$

Tablica 1: Porównanie dokładności klasyfikacji wg ignorowania kolejnych *najlepszych* atrybutów.

Przebadaliśmy także przypadki usunięcia kolejno  $n$  *najlepszych* atrybutów. Po zignorowaniu pojedynczego atrybutu sprawdzaliśmy, jaki atrybut został wytypowany na kolejny *najlepszy* i w kolejnym teście dodawaliśmy go do ignorowanych itd, sprawdzając tym samym kolejny *najlepszy* atrybut itd.

liczba ignorowanych <i>najlepszych</i> atrybutów	dokładność klasyfikacji z pominięciem atrybutu
2	$\approx 98\%$
3	$\approx 53\%$
4	$\approx 20\%$
5	$\approx 56\%$

Tablica 2: Porównanie dokładności klasyfikacji wg ignorowania  $n$  kolejnych *najlepszych* atrybutów.

## 7 Wnioski

Dla typowych proporcji 80 : 20 wynik dokładności klasyfikacji jest zgodny z oczekiwanym - wynosi blisko 100%. Również zgodnie z oczekiwaniami, w miarę zmniejszania liczby elementów zbioru trenującego dokładność maleje i przy proporcjach 50 : 50 można zauważyć, że występują duże wahania losowe, co oznacza, że algorytm nie jest już dokładny.


Analiza dokładności klasyfikacji przy ignorowaniu kolejnych *najlepszych* atrybutów wykazała, że brak pojedynczego atrybutu praktycznie nie wiąże się z jakąkolwiek utratą dokładności. Cztery najbardziej znaczące (*najlepsze*) atrybuty po zignorowaniu nie spowodowały żadnej utraty dokładności.

Z kolei zbadanie wpływu ignorowania *n* *najlepszych* atrybutów wykazało, że dla danego zbioru testowego (*agaricus-lepiota.data*) zignorowanie więcej niż dwóch *najlepszych* atrybutów na raz prowadzi do utraty dokładności.

## 8 Podsumowanie

Projekt wprowadzający w tematykę uczenia maszynowego oraz drzew decyzyjnych. Dzięki własnej implementacji algorytmu ID3 poznano istotę konstrukcji drzew decyzyjnych. Zastosowanie testów binarnych pozwoliło na zapoznanie się z przykładową reprezentacją modelu drzewa decyzyjnego.

## 9 Powiązane linki

 [Repozytorium projektowe](#)