# WSI – sprawozdanie

Pryimak Andrii-Stepan 336173

## Implementacja Drzewa ID3

### Wprowadzenie

* : Implementacja klasyfikatora drzewa decyzyjnego ID3 dla atrybutów nominalnych i testów tożsamościowych.
* Zbiory danych: [Breast cancer](http://archive.ics.uci.edu/ml/datasets/Breast+Cancer) oraz [mushroom](https://archive.ics.uci.edu/ml/datasets/mushroom).
* Cel: Ocena dokładności klasyfikatora i macierzy pomyłek na podanych zbiorach danych oraz analiza wyników.

### Zestawy danych

|  |  |  |
| --- | --- | --- |
|  | Breast cancer | Mushroom |
| Rozmiar | 286 | 8124 |
| Liczba atrybutów | 9 | 22 |
| Liczba klas | 2 | 2 |
| Entropy | 0.877 | 0.999 |
| Proporcja klas | 'no-r.-e.' / 'r.-e ' : 201 / 85 | 'e'/ 'p' : 4208 / 3916 |

\*r.-e. = recurrence-events

### Wyniki

Przeprowadzono 10 pomiarów z randomowym podziałem zestawów na testujący oraz uczący. Zbiór uczący do zbioru testującego były podzielone w 60% do 40%

|  |  |  |
| --- | --- | --- |
|  | Breast cancer | Mushroom |
| Avarage Accuracy | 0.568 | 1.0 |
| Min Accuracy | 0.460 | 1.0 |
| Max Accuracy | 0.678 | 1.0 |

Macierz pomyłek:

|  |  |  |  |
| --- | --- | --- | --- |
| Predicted | no-r.-e. | r.-e. | NaN |
| Actual |  |  |  |
| no-r.-e. | 55.0 | 17.2 | 8.3 |
| r.-e. | 19.8 | 10.4 | 4.3 |

|  |  |  |
| --- | --- | --- |
| Predicted | e | p |
| Actual |  |  |
| e | 1688.0 | 0.0 |
| p | 0.0 | 1562.0 |

Przy zmianie rozmiaru danych do 300 program z grzybami myli się więcej ale ma efektywność 97 procent. Z min wynikiem 87%

Przy zmianie liczby kolumn do 9 i liczbie wpisów 8124 program też dobrze sobie radzi

### Podsumowanie

Wyniki klasyfikatora ID3 na dwóch różnych zbiorach danych, dotyczących raka piersi i grzybów, pokazują znaczące różnice w dokładności klasyfikacji.

Grzyby

* Dokładność: 100%
* Macierz pomyłek: Klasyfikator poprawnie sklasyfikował wszystkie próbki, co oznacza, że nie popełnił żadnych błędów.
* Interpretacja: Wysoka dokładność klasyfikatora na zbiorze danych dotyczących grzybów sugeruje, że dane te są łatwiejsze do klasyfikacji. To wynikać z bardziej jednoznacznych i wyraźnych wzorców oraz dużo większej liczby atrybutów oraz przykładowych danych.

Rak Piersi

* Dokładność: 56.8%
* Macierz pomyłek: Klasyfikator miał trudności z poprawnym klasyfikowaniem próbek, co widać po liczbie błędnych klasyfikacji.
* Interpretacja: Niższa dokładność klasyfikatora na zbiorze danych dotyczących raka piersi wynika z kilku czynników:
  + Mniejsza ilość danych: Mniejsza liczba próbek może prowadzić do gorszej wydajności modelu, ponieważ model ma mniej danych do nauki.
  + Wyższa entropia: Dane dotyczące raka piersi mogą być bardziej złożone i zawierać więcej zmienności, co utrudnia modelowi naukę i predykcję.

### Wnioski

* Grzyby: Klasyfikator ID3 działa bardzo dobrze na zbiorze danych dotyczących grzybów, osiągając 100% dokładności.
* Rak Piersi: Klasyfikator ID3 ma trudności z klasyfikacją danych dotyczących raka piersi, co wynikaa z mniejszej ilości danych oraz wyższej entropii w danych.