WSI - ćwiczenie 6.

Regresja i klasyfikacja

grupa 101

Wykonał: Adrian Pruszyński

1. Treść Zadania

W ramach szóstego ćwiczenia należy zaimplementować drzewo decyzyjne indukowane algorytmem ID3.

Należy umożwliwić ustawienie maksymalnej głębokości drzewa podczas jego tworzenia. Jeżeli węzeł na tej głębokości nie pozwala na jednoznaczną klasyfikację, powinien stać się liściem zawierającym najczęstszą klasę.

Następnie należy przetestować zaimplementowany klasyfikator z użyciem zbioru danych titanic https://web.stanford.edu/class/archive/cs/cs109/ cs109.1166/stuff/titanic.csv.

Należy uwzględnić podział zbioru na treningowy, walidacyjny i testowy. Uwaga! Niektóre atrybuty mogą się nie nadawać do zastosowania (na przykład imię i nazwisko), a niektore atrybuty mają wartości ciągłe (na przykład wiek), które należy podzielić na zakresy i potraktować te zakresy jako atrybuty dyskretne.

2. Założenia

Dane dzielone są na 3 podzbiory trenujący, walidacyjny, testowy wielkości odpowiednio 60%, 20%, 20% całego zbioru.

Do eksperymentu dołączyłem własną propozycję podziału na zakresy zarówno dla atrybutu Age jak i Fare, zakresy przedstawione są poniżej.

Age: [0, 6), [6, 12), [12, 16), [16, 25), [25, 35), [35, 45), [45, 60), [60, 400)

Fare: [0, 15), [15, 25), [25, 60), [60, 100), [100, 150), [150, 300), [300, 1000)

Zakładam, że poziom korzenia to poziom 1

3. Prezentacja wyników działania algorytmu

Tabela 1 Przedstawia dokładność klasyfikacji dla różnych parametrów podziału i głębokości na zbiorze walidacyjnym.

Depth / Age classes – Fare classes	Custom	8 - 8	8 - 4	4 - 8	4 - 4
2	0,785	0,785	0,785	0,785	0,785
3	0,785	0,785	0,791	0,785	0,791
5	0,785	0,791	0,791	0,802	0,797
7	0,723	0,746	0,740	0,746	0,734

Wyniki na zbiorze testowym dla 3 najlepszych wyników z tabeli 1

Tabela 2 Dokładność dla wybranych najlepszych wartości z tabeli 1

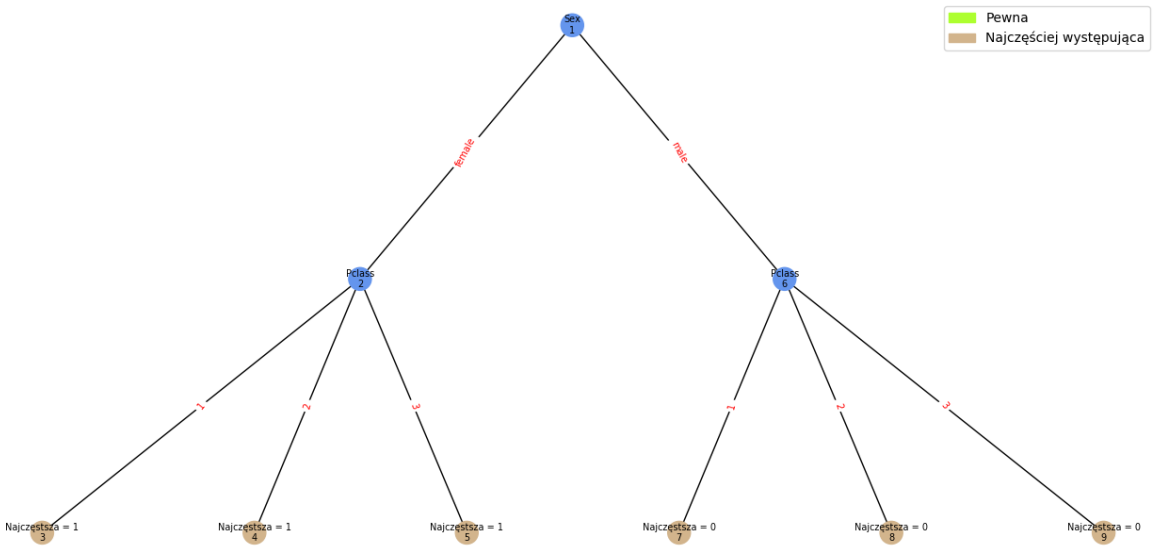
Depth / Age classes – Fare classes	4 - 8	4 - 4
3	x	0.787
5	0.725	0.730

Jak widać w tabeli 3 wynik dla najlepszego modelu wybranego dla zbioru walidacyjnego nie pokrywają się najlepszym wynikiem dla zbioru testowego w tabeli 2.

4. Analiza działania algorytmu

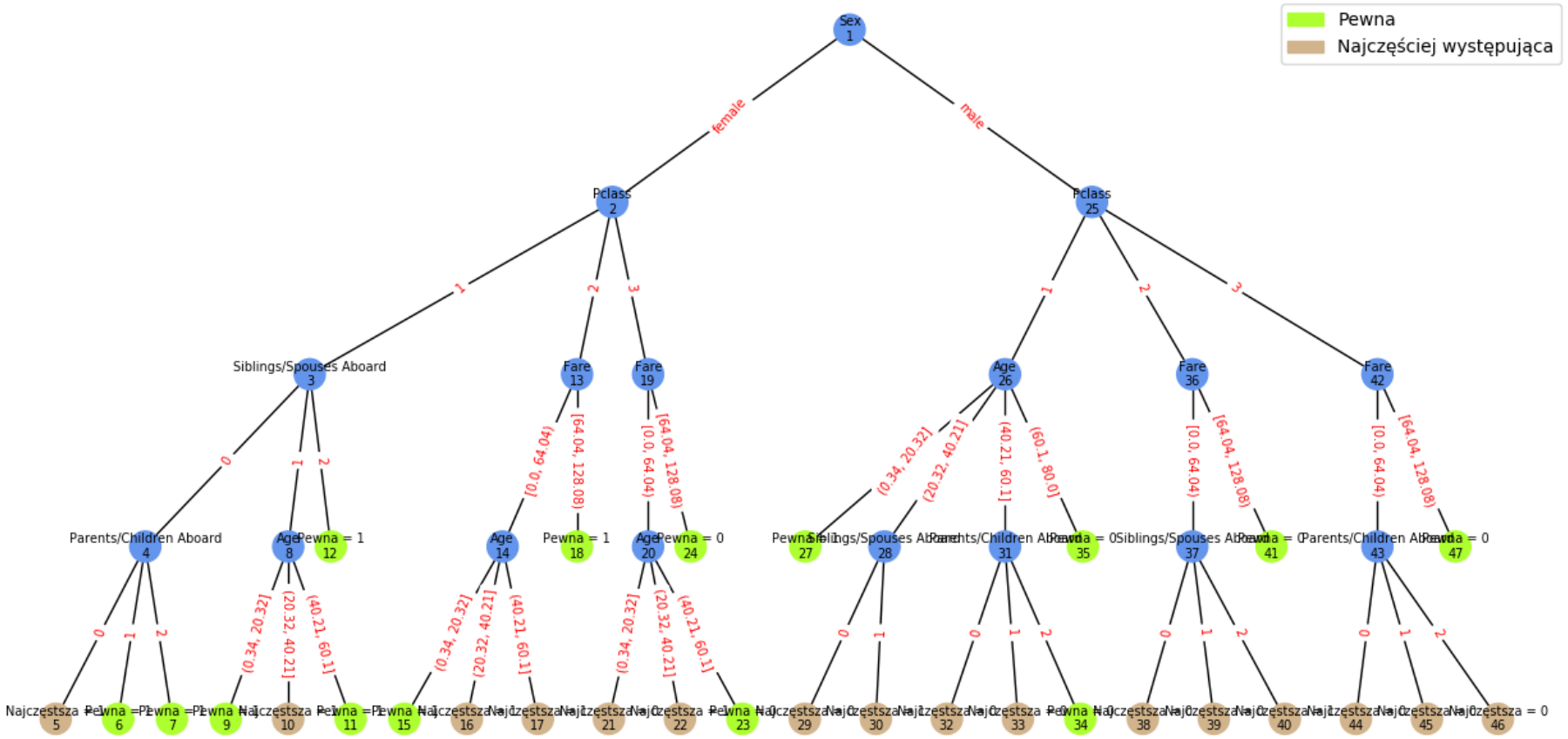
Jak łatwo możemy zauważyć w tabeli 1 dla niektórych różnych głębokości oraz takiego samego podziału na klasy możemy zaobserwować takie same wartości dokładności, może być to spowodowane sytuacją, gdy wszystkie liście mające tego samego rodzica mają taką samą wartość. Taką sytuację widać to na rysunku 1

Rysunek 1



Takiej sytuacji jak na rysunku 1 zapobiega przycinanie drzew.

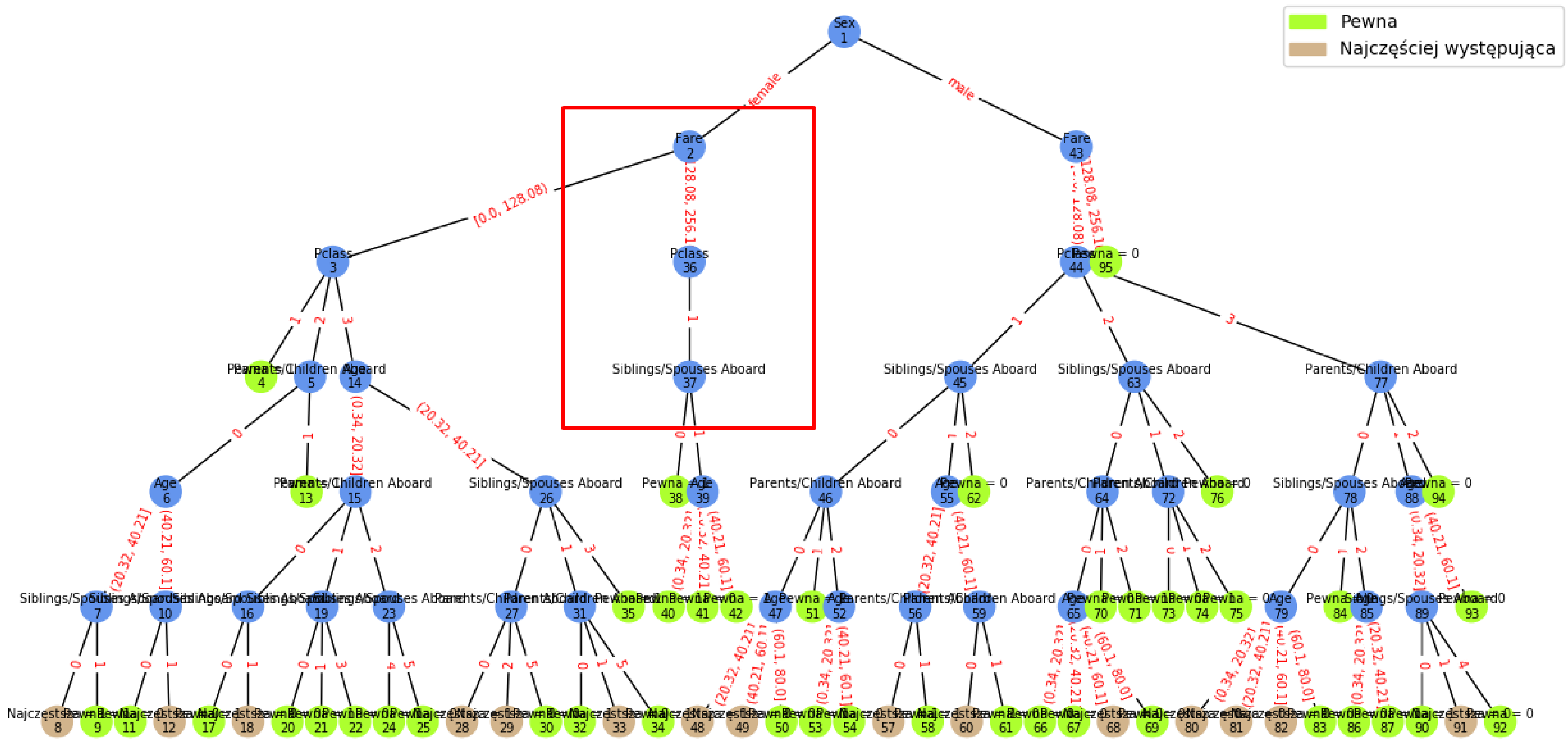
Rysunek 2



Na rysunku 2 widzimy drzewo o głębokości 5 dla parametrów podziału Age 4 Fare 8 dla którego osiągnięto najwyższą dokładność klasyfikacji równą 0,802.

Dla złego podziału na zbiór uczący, walidacyjny i testowy np.: 20%, 40%, 40% możemy zaobserwować sytuację, gdzie wiele z podzbiorów powstałych z podziału w zbiorze uczącym nie będzie miało wszystkich możliwych wartości klas. Jest to spowodowane zbyt małym zbiorem uczącym dokładność klasyfikacji dla takiego drzewa = 0.696

Rysunek 3



5. Wnioski i podsumowanie

Drzewo zbudowane przy pomocy algorytmu ID3 jest podatne na powstanie nadmiarowych liści (rysunek 1), które można usunąć przy pomocy przycinania np. algorytm C4.5.

Zbyt mały zbór danych uczących może prowadzić do budowy modelu niższej jakości jak na rysunku 3.

Drzewo pozwala na klasyfikację na poziomie dokładności około 78% zależnie od dobranych parametrów może zostać ona zwiększona.

Odpowiedni dobór podziałów na klasy dla danych (przygotowanie danych) ma duże znaczenie dla jakości zbudowanego modelu.

Dobór odpowiedniej głębokości drzewa poza samą szybkością ma również wpływ na dokładność klasyfikacji.

Zbyt mała głębokość prowadzi do utraty części informacji natomiast zbyt duża znacznie zwiększa uszczegółowienie modelu. W konsekwencji model może stać się idealnie dopasowany do zbioru uczącego, na którym był tworzony, a co za tym idzie nie jest zdolny do uogólnienia i wykorzystania go na niezależnych zbiorach testowych. (przeuczenie modelu)