Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska

Wstęp do sztucznej inteligencji

Raport z laboratorium 4.

Kacper Bugała

Warszawa, 2022Z

1. Opis rozwiązania

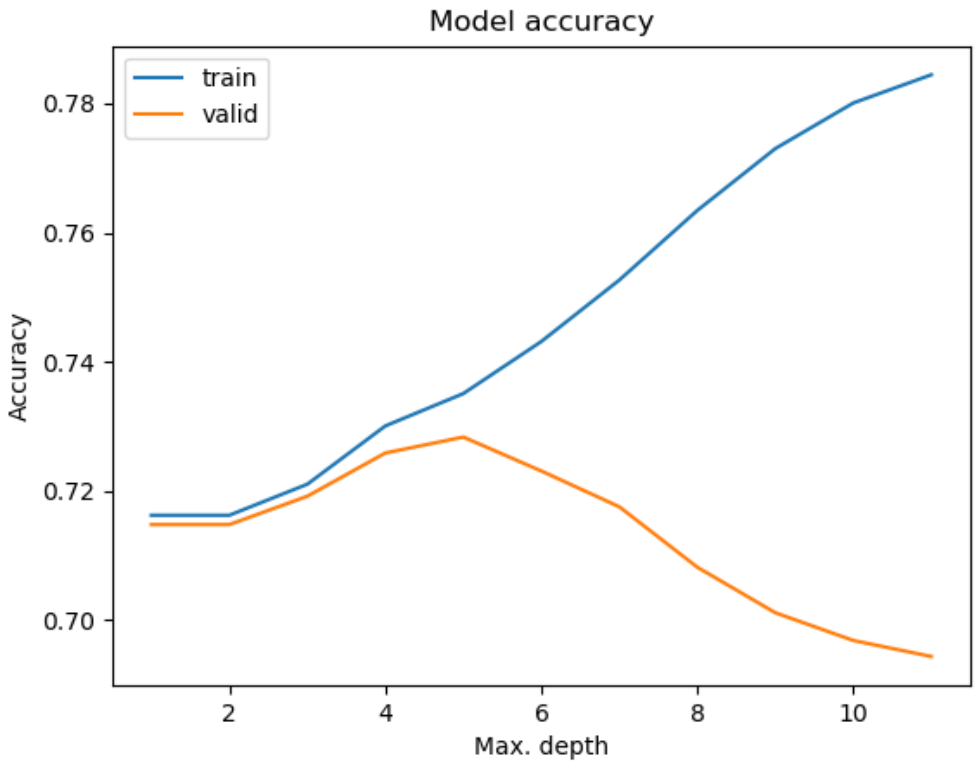
Zaimplementowano algorytm ID3 do generowania drzew decyzyjnych. Implementacja została przetestowana na dostarczonej bazie danych. Baza zawiera informacje medyczne dotyczące pacjentów (11 różnych atrybutów), a także informacje czy cierpi na choroby układu krążenia – **cardio** (określane jako *boolean*). Atrybut **cardio** wybrany został jako wartość klasy rekordu. Zbiór podzielono na trzy podzbiory - trenujący, walidacyjny i testowy. Stworzony model uczy się na zbiorze trenującym, lecz aby nie dopuścić do *overfittingu* modelu do tych danych, jakość klasyfikacji oceniana jest także na zbiorze walidacyjnym. Rozmiary tych podzbiorów ustalono kolejno na 60% - 20% - 20% całej bazy wiedzy. Algorytm ID3 mówi o tym, że w kolejnych głębokościach drzewa, tworzony węzeł dzieli pozostałe dane na podstawie wartości atrybutu o największej wartości ***info gain***. Parametr ten wyliczany jest na podstawie entropii zbioru. Maksymalna głębokość drzewa dla tej bazy danych wynosi 11, czyli dokładnie tyle co liczba atrybutów. Część danych została poddana dyskretyzacji na podstawie danych medycznych, oraz w wyniku eksperymentów.

Oczekiwanym rezultatem jest znalezienie głębokości drzewa, która klasyfikować będzie z największą precyzją. Podczas procesu uczenia, analizowane są zarówno precyzja na zbiorze trenującym jak i na zbiorze walidacyjnym. Wraz z początkowym wzrostem głebokości spodziewana jest tendencja wzrostowa dla obu tych precyzji. Typowe dla drzewa decyzyjnego jest fakt, że od pewnej głębokości drzewa rośnie już jedynie precyzja na zbiorze treningowym, a na walidacyjnym zaczyna spadać. Optymalna głębokość to ta, która ma największą precyzję na zbiorze walidacyjnym. Gdy ta precyzja zaczyna spadać, oznacza to overfitting modelu do danych treningowych.

2. Badanie wpływu zmiany głębokości

Przeprowadzono eksperymenty, w celu zbadania wpływu zmiany głębokości drzewa na jakość modelu. Analizowano **głębokości od 1 do 11**, czyli wszystkie możliwe dla tej bazy danych. dla każdej głębokości symulację powtarzano 20-krotnie, a wyniki uśredniono. Wyniki takiego eksperymentu pozwolą stwierdzić, jaka jest optymalna głębokość drzewa, bez występowania overfittingu.

Wyniki eksperymentu przedstawiono na wykresie 2.1



Rys. 2.1. Wykres precyzji modelu w procesie uczenia

Precyzja na zbiorze walidacyjnym rośnie do maksymalnej głębokości równej 5. Wykres jasno pokazuje, że zwiększanie maksymalnej głębokości powyżej tej wartości sprawia, że model bardzo dobrze klasyfikuje na zbiorze trenującym, lecz dzieje się to w wyniku overfittingu. Dokładne wyniki dla kolejnych głębokości przedstawiono w tabeli 2.2.

	depths										
	1	2	3	4	5	6	7	8	9	10	11
train	0.715	0.715	0.72	0.729	0.734	0.742	0.75	0.761	0.772	0.779	0.783
val	0.717	0.717	0.72	0.729	0.733	0.727	0.721	0.71	0.702	0.699	0.697

Rys. 2.2. Dokładne wyniki eksperymentu dla różnych głębokości

Ustalono maksymalną głębokość drzewa na 5, a następnie sprawdzono jakość modelu na zbiorze testowym (rys. 2.3):

```
Setting 'max_depth' = 5 for this solver
TEST_SET:
Depth = 5
Accuracy: 0.7211774325429272
```

Rys. 2.3. Dokładność na zbiorze testowym

Na zbiorze testowym udało się uzyskać dokładność ponad 72%, co jest bardzo dobrym wynikiem.

3. Wnioski

Zaimplementowany model drzewa decyzyjnego z algorytmem ID3 spełnił swoje zadanie. Analiza procesu doboru maksymalnej głębokości drzewa pokazała problem overfittingu i potwierdziła słuszność stosowania modelu walidacyjnego na etapie uczenia. Uzyskany wynik na zbiorze testowym jest zadowalająco uzykując **72%** dokładności oceny stanu zdrowia pacjenta, na podstawie innych jego danych. Badanie potwierdziło jednak fakt, że nie da się uzyskać 100% skuteczności dla podanego zbioru. Duży wpływ na jakość modelu ma dyskretyzacja danych wejściowych. Eksperymenty przeprowadzono dla kilku różnych zasad podziału, zaprezentowane wyniki są przeprowadzone na najlepszym znalezionym podziale.