Rafał Mironko – WSI Zadanie 4: ID3

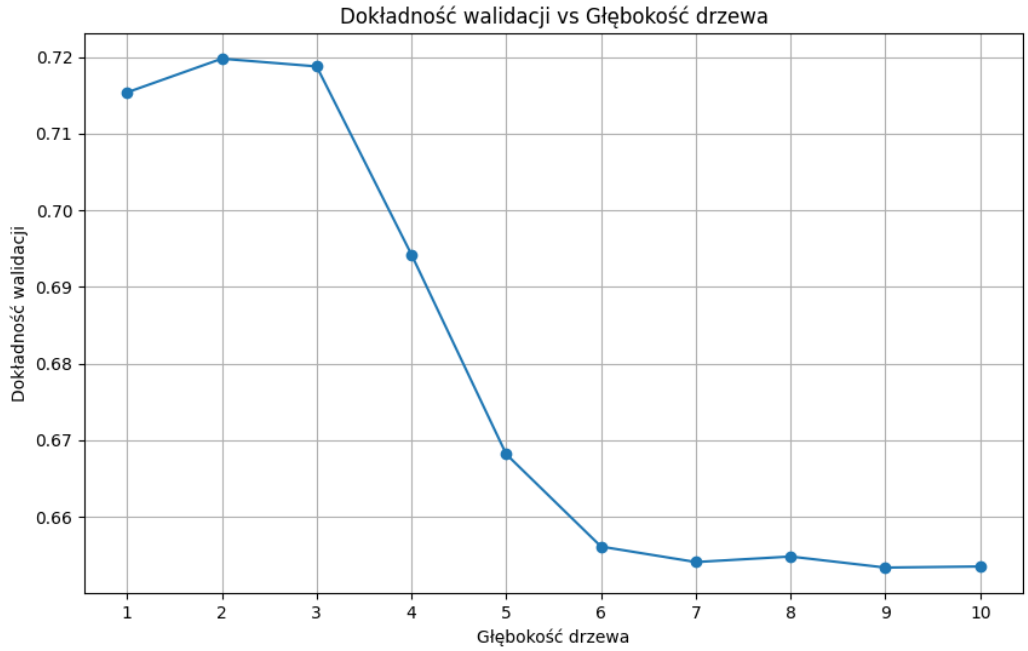
Temat maszynowego uczenia naprawdę mi się spodobał. Praca domowa wymagająca, zabrała sporo czasu, ale wszystko udało się ukończyć.

Podobnie jak poprzednio – nie mam pewności, czy jest sens rozpisywać się na 10 stron raportu. Zamiast tego, pokażę kilka wykresów i powiem, jaki parametr szczególnie zaskoczył mnie podczas testów.

Zgodnie z poleceniem, wykonam testy dla różnych głębokości drzewa [1; 10] – uwzględnię jednak też przygotowanie danych.

Po napisaniu pierwszej wersji kodu, uruchamiam go i czekam na wynik. Mieli się, mieli, czekam 10 minut, przerywam program. Po napisaniu kilku debugujących logów okazało się, że po 10 minutach jestem dopiero na głębokości 7. Dziwne.

Oczywiście, zamiast przyjrzeć się danym bliżej, postanawiam rozwiązać problem siłą. Przerobiłem kod tak, że teraz różne głębokości wykonują się wielowątkowo. Tym razem po trzech minutach doczekałem się wyniku:



Nie ukrywam, nie jestem z niego zadowolony. Czekałem tyle czasu tylko po to, żeby zobaczyć, że głębokości ponad 4 są beznadziejne?

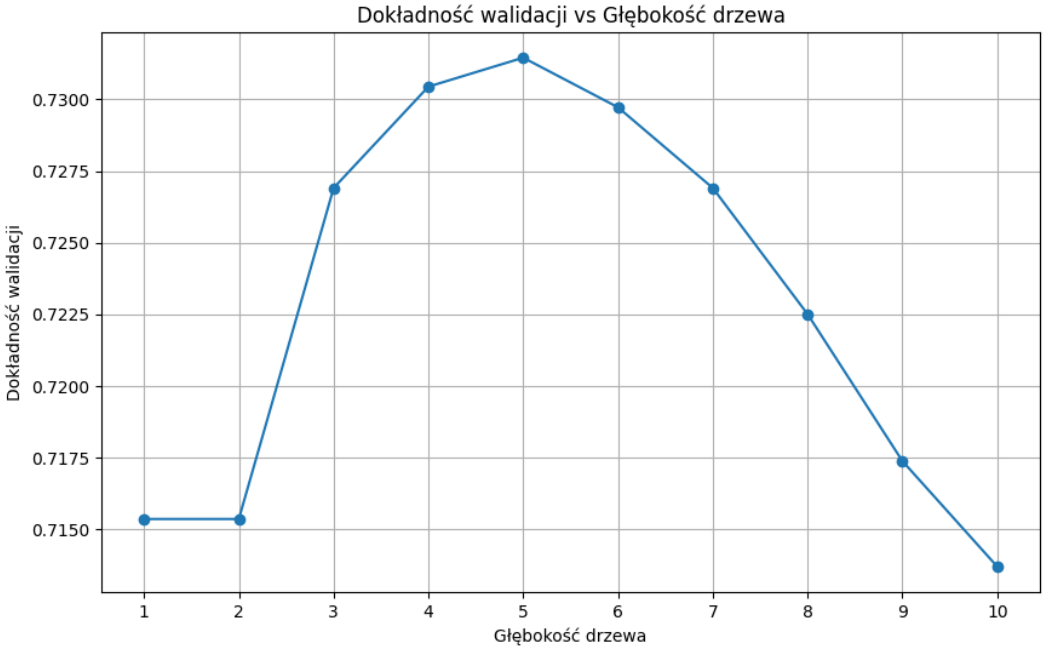
Co ciekawe, nawet chatgpt nie był w stanie mi pomóc. Szukałem rozwiązania, przeglądałem stack overflow, nawet przeczytałem rozdział „Sztucznej Inteligencji dla Inżynierów” o uczeniu maszynowym. W przeciwieństwie do poprzednich rozdziałów jednak, ten nie okazał się zbyt przydatny.

Odpowiedź przyszła z całkiem nieoczekiwanego źródła. Niedawno kupiłem sobie „Uczenie maszynowe z użyciem Scikit-Learn, Kera i TensorFlow”. Już nie chciało mi się szukać rozwiązania

tego problemu, postanowiłem sobie po prostu poczytać ciekawą książkę o uczeniu maszynowym. I...

Był tam rodział omawiający ID3. A w nim informacja o tym, że dla niewielkich rozmiarów danych (poniżej 100 000 rekordów), mających sporo atrybutów, na ogół nie stosuje się dyskretyzacji danych ciągłych na większą ilość przedziałów niż 5. Ja do tej pory miałem 15.

To mnie na tyle zaciekawiło, że od razu zabrałem się za sprawdzenie. Uruchomiłem algorytm na ilości przedziałów dla każdej dyskretyzowanej danej równej 3. Okazało się, że to było właśnie rozwiązanie mojego problemu.



Teraz algorytm zachowuje się tak, jak można się było tego spodziewać. Wychodzi na to, że przy zbyt wysokiej liczbie opcji, algorytm szybko nadmiernie dopasowuje się do danych treningowych i ulega przeuczeniu. Jak widać, przy zmiejszonej liczbie przedziałów, uległo to poprawie.

I jeszcze jedna rzecz. Algorytm teraz wykonuje się kilka sekund. Zastanawiałem się dlaczego, policzyłem to i okazuje się, że ograniczyłem liczbę kombinacji na dyskretyzowanych przedziałach z 15*15*15*15 do 3*3*3*3. Nic dziwnego, że to pomogło.

Podsumowując, ćwiczenie niesamowicie ciekawe. Szczególnie cieszę się, że odkryłem, że zmiana przekazania algorytmowi danych może aż tak mocno zmienić działanie algorytmu.