WSI Laboratorium 4 Raport

Algorytm ID3

Miłosz Andryszczuk 331355

1. Wykorzystane biblioteki

Do przeprowadzenia eksperymentu zostały wykorzystane następujące biblioteki:

- matplotlib
  do rysowania wykresów
  https://matplotlib.org/stable/index.html
- numpy
  do obsługi operacji matematycznych
  https://numpy.org/doc/
- pandas
  do przetwarzania danych w formacie tabelarycznym
  https://pandas.pydata.org/docs/
- scikit-learn
  do podziału danych na zbiory oraz oceny dokładności modelu
  https://scikit-learn.org/stable/

2. Opis eksperymentu

W ramach tego eksperymentu zbadano działanie algorytmu ID3, który pozwala na budowanie drzew decyzyjnych wykorzystywanych do klasyfikacji danych.

Algorytm analizuje dane wejściowe i tworzy drzewo decyzyjne, wybierając atrybuty, które maksymalizują przyrost informacji na każdym poziomie drzewa. Proces jest powtarzany rekurencyjnie, aż do osiągnięcia określonej głębokości drzewa lub spełnienia warunku stopu.

Celem eksperymentu było zbadanie jakości klasyfikacji wykonywanej przez algorytm ID3 dla różnych wartości głębokości drzewa, liczby przedziałów w dyskretyzacji cech oraz rozmiarów zbioru treningowego. Porównano również różnicę pomiędzy dokładnością na zbiorach treningowym i walidacyjnym, aby ocenić wpływ przeuczenia.

3. Przebieg Eksperymentu

3.1. Wpływ głębokości drzewa decyzyjnego na dokładność

W pierwszej części eksperymentu przeanalizowano wpływ maksymalnej głębokości drzewa na dokładność klasyfikacji. Głębokość drzewa była testowana w przedziale [1;15]. Dla każdej głębokości algorytm ID3 trenowano na zbiorze treningowym i oceniano na zbiorze walidacyjnym. Na tej podstawie wyznaczono najlepszą wartość głębokości, która unika przeuczenia (over-fitting). Dla tej głębokości algorytm uruchomiono ponownie na zbiorze testowym, aby wyliczyć ostateczną dokładność.

Porównano dokładność algorytmu na zbiorze treningowym i walidacyjnym. Wyniki przedstawiono na wykresach, z których jeden ilustruje dokładności na obu zbiorach, a drugi różnicę między nimi, aby uwidocznić zjawisko przeuczenia.

3.2. Wpływ liczby przedziałów dyskretyzacji cech na dokładność

W kolejnej części zbadano wpływ liczby przedziałów (wartość q) używanych do dyskretyzacji cech ciągłych na jakość klasyfikacji. Liczba przedziałów była testowana w zakresie od 2 do 20.

Wartość q oznacza, na ile grup zostaje podzielony zakres danej cechy – na przykład przy q=2 każda cecha ciągła może przyjąć tylko dwie wartości dyskretne, co odpowiada podziałowi na dwie równe grupy danych.

Dla każdej wartości q trenowano algorytm ID3 z optymalną głębokością drzewa i oceniano dokładność na zbiorach treningowym i walidacyjnym. Wyniki przedstawiono na wykresie obrazującym zależność między liczbą przedziałów a dokładnością.

3.3. Wpływ rozmiaru zbioru treningowego na jakość klasyfikacji

W ostatniej części eksperymentu przeanalizowano wpływ rozmiaru zbioru treningowego na jakość klasyfikacji. Proporcja danych treningowych była zmieniana od 10% do 90% całego zbioru, przy zachowaniu stałej głębokości drzewa.

Dokładności na zbiorach treningowym i walidacyjnym dla różnych rozmiarów zbioru treningowego przedstawiono na wykresie. Wyniki umożliwiły ocenę wpływu ilości danych na zdolność algorytmu do generalizacji.

4. Wyniki

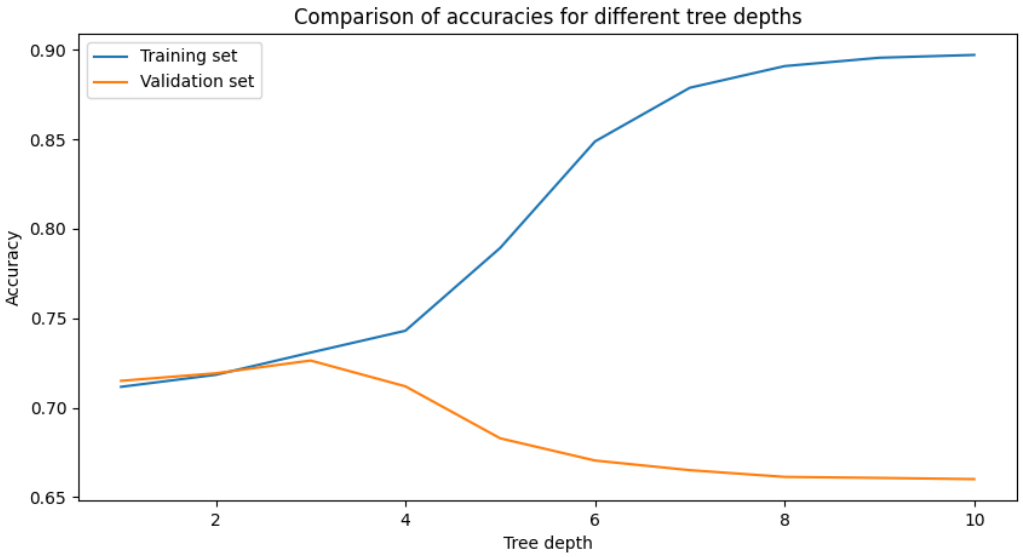
4.1. Wpływ głębokości drzewa decyzyjnego na dokładność

Na osi poziomej pierwszego wykresu znajduje się maksymalna głębokość drzewa, a na osi pionowej – dokładność klasyfikacji na zbiorach treningowym i walidacyjnym. Drugi wykres ilustruje różnicę między dokładnością na zbiorze treningowym a walidacyjnym, co pozwala na ocenę przeuczenia (over-fitting) w zależności od głębokości drzewa.

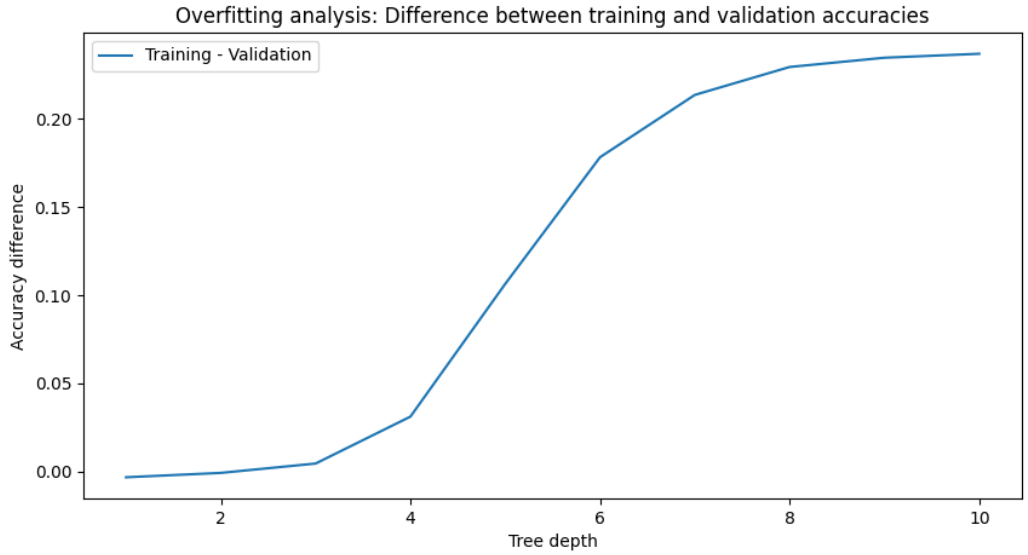
Pierwszy wykres wykazuje, że większa głębokość drzewa prowadzi do wyraźnego wzrostu dokładności na zbiorze treningowym, podczas gdy dokładność na zbiorze walidacyjnym stabilizuje się i zaczyna spadać przy większych głębokościach. Dla głębokości od 1 do około 3 dokładność na obu zbiorach rośnie, co wskazuje na poprawę modelu. Jednak dla głębokości większej niż 3 widoczny jest efekt przeuczenia – model dopasowuje się do danych treningowych – zapamiętuje je, zamiast uczyć się reguł, które w nich występują.

Najlepszą dokładność dla zbioru walidacyjnego zapewnia głębokość równa 3.

Algorytm z tą głębokością dla nieznanego zbioru testowego wykazywał dokładność równą około 72,67%.



Drugi wykres uwidacznia tę zależność – różnica między dokładnością na zbiorach treningowym i walidacyjnym staje się coraz większa dla rosnącej głębokości drzewa. Szczególnie dla głębokości większej niż 3 różnica ta znacząco wzrasta. Mimo to w pewnym momencie zaczyna się stabilizować, co oznacza, że nawet kiedy występuje zjawisko przeuczenia to algorytm wciąż nie działa źle, ale jednak znacznie gorzej niż dla niższych wartości głębokości.

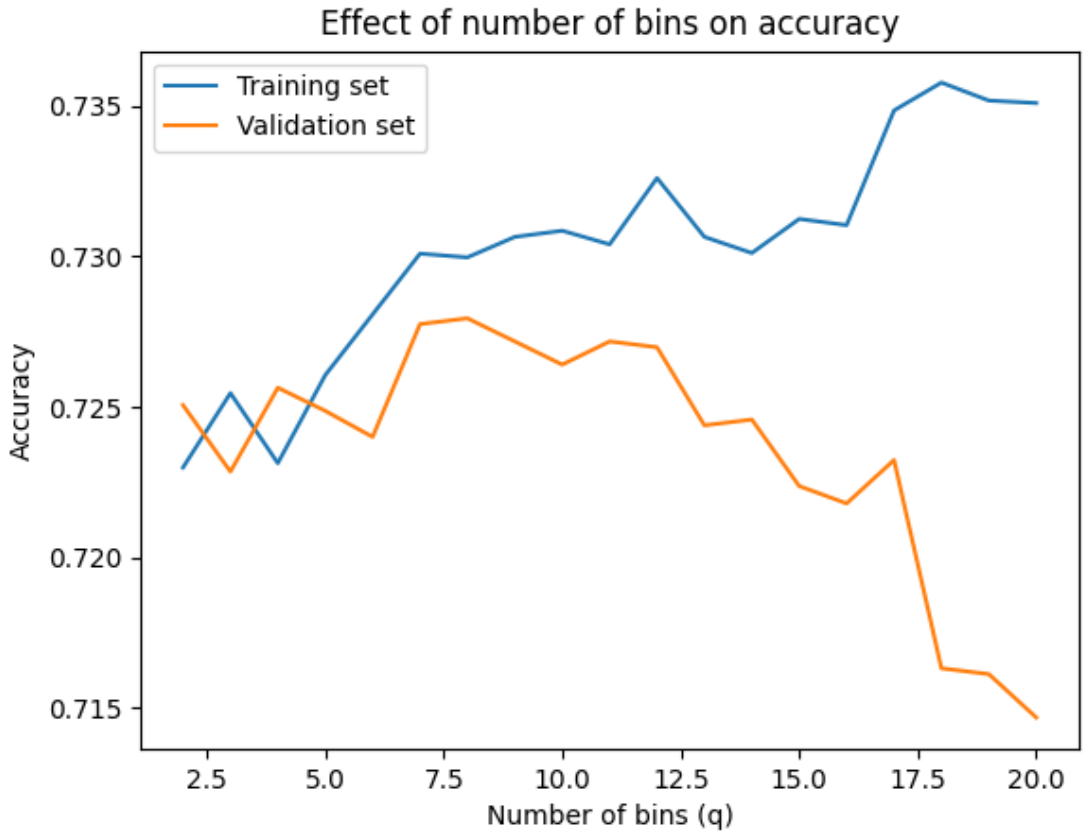


4.2. Wpływ liczby przedziałów dyskretyzacji cech na dokładność

Na osi poziomej wykresu przedstawiono liczbę przedziałów (q), na które podzielono każdą cechę ciągłą w procesie dyskretyzacji, natomiast na osi pionowej znajduje się dokładność klasyfikacji na zbiorach treningowym i walidacyjnym.

Wykres pokazuje, że zmiana liczby przedziałów wpływa na dokładność modelu. Dokładność na zbiorze treningowym wzrasta wraz ze zwiększaniem liczby przedziałów. Przy większej liczbie przedziałów (q) algorytm ID3 ma więcej możliwości podziału danych na podstawie cech, co pozwala na lepsze rozróżnienie przykładów w zbiorze treningowym. Model staje się bardziej złożony, dzięki czemu “zapamiętanie” danych treningowych jest łatwiejsze.

W przypadku zbioru walidacyjnego maksymalna dokładność jest osiągana dla około q=10, a następnie dokładność zaczyna spadać. Spadek dokładności na zbiorze walidacyjnym przy większej liczbie przedziałów może być wynikiem przeuczenia modelu, ponieważ przy dużych wartościach q cechy stają się zbyt szczegółowe, co prowadzi do gorszej generalizacji. Model dopasowuje się zbyt ściśle do danych treningowych, uwzględniając nawet drobne różnice czy przypadkowe wzorce, które nie występują w rzeczywistych danych testowych czy walidacyjnych.



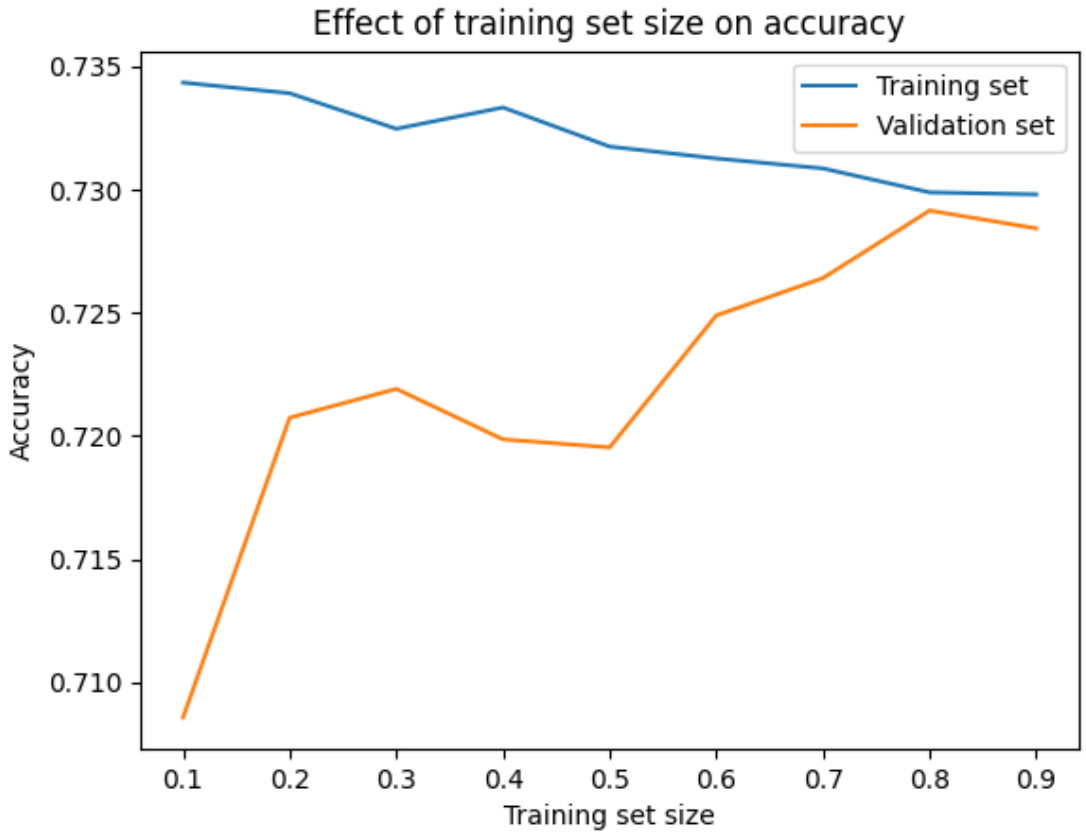
4.3. Wpływ rozmiaru zbioru treningowego na jakość klasyfikacji

Na osi poziomej wykresu przedstawiono rozmiar zbioru treningowego jako ułamek całego dostępnego zbioru danych, natomiast na osi pionowej znajduje się dokładność klasyfikacji na zbiorach treningowym i walidacyjnym.

Wykres pokazuje, że wraz ze wzrostem rozmiaru zbioru treningowego dokładność na zbiorze walidacyjnym początkowo wyraźnie wzrasta, osiągając maksimum w okolicach proporcji 70%-80%. Natomiast dokładność na zbiorze treningowym pozostaje na względnie stałym poziomie lub nawet nieznacznie spada.

Spadek dokładności na zbiorze treningowym może wynikać z tego, że większy zbiór treningowy oznacza bardziej zróżnicowane dane, co może utrudniać modelowi perfekcyjne dopasowanie do wszystkich przykładów. Jednocześnie, większy zbiór treningowy prowadzi do lepszego uogólnienia modelu, co skutkuje wzrostem dokładności na zbiorze walidacyjnym.

Przy bardzo małych rozmiarach zbioru treningowego (np. 10%-20%) model nie jest w stanie odpowiednio nauczyć się wzorców w danych, co prowadzi do niskiej dokładności na zbiorze walidacyjnym.



5. Wnioski

Wyniki eksperymentu wykazały, że maksymalna głębokość drzewa decyzyjnego w algorytmie ID3 ma kluczowe znaczenie dla jakości klasyfikacji. Zwiększanie głębokości drzewa poprawiało dokładność na zbiorze treningowym, ale prowadziło również do przeuczenia, co objawiało się spadkiem dokładności na zbiorze walidacyjnym przy większych głębokościach.

Eksperyment z liczbą przedziałów w dyskretyzacji cech wykazał, że nadmierna liczba przedziałów może prowadzić do przeuczenia, podczas gdy zbyt mała liczba skutkuje niedostatecznym uchwyceniem szczegółów w danych. Wyniki wskazują, że poza głębokością również optymalny wybór liczby przedziałów jest kluczowy dla uzyskania wysokiej jakości klasyfikacji.

Badanie wpływu rozmiaru zbioru treningowego pokazało, że większe zbiory treningowe poprawiają dokładność na zbiorze walidacyjnym, szczególnie przy początkowym wzroście rozmiaru. Jednakże, po przekroczeniu pewnego progu (około 70%-80% dostępnych danych), korzyści z dalszego zwiększania rozmiaru stają się marginalne, a dokładność stabilizuje się.

Optymalna głębokość drzewa pozwalała na zrównoważenie precyzji modelu i jego zdolności do generalizacji, a tym samym uniknięcie zjawiska przeuczenie (over-fitting).