WSI - ćwiczenie 5.

Modele bayesowskie

grupa 101

Wykonał: Adrian Pruszyński

1. Treść Zadania

W ramach piątego ćwiczenia należy zaimplementować naiwny klasyfikator Bayesa. Korzystając z tego klasyfikatora należy zbadać, który atrybut ze zbioru danych wine (https://archive.ics.uci.edu/ml/datasets/wine) pozwala osiągnąć najlepszą dokładność klasyfikacji. W celu oceny zbioru należy wykorzystać algorytm n-krotnej walidacji krzyżowej.

Następnie należy zbadać, czy dodanie atrybutu o numerze N + 1 pozwala poprawić jakość klasyfikacji, gdzie N oznacza ostatnią cyfrę numeru indeksu (czyli student, którego numer indeksu kończy się na 0 powinien dodać atrybut o numerze 1). Jeżeli atrybut N + 1 jest atrybutem znalezionym w poprzednim etapie, dodany powinien zostać atrybut o numerze N + 2. Numeracja atrybutów jest zgodna z informacjami pod powyższym linkiem, tj.

1. Alcohol
2. Malic acid
3. Ash
4. Alcalinity of ash
5. Magnesium
6. Total phenols
7. Flavanoids
8. Nonflavanoid phenols
9. Proanthocyanins
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline

2. Założenia

Algorytm walidacji krzyżowej został zaimplementowany w najprostszej formie tj. nie zapewnia wyboru podzbiorów tak aby dla każdej klasy wybierać kolejne podzbiory do walidacji krzyżowej w sposób reprezentatywny dla całego zbioru danych uczących. Czyli ilość reprezentantów danej klasy w podzbiorze jest wynikiem losowania reprezentantów z całego zbioru.

3. Prezentacja wyników działania algorytmu

Tabela 1 Przedstawia wyniki poszukiwana najlepszego atrybutu ze względu na dokładność klasyfikacji

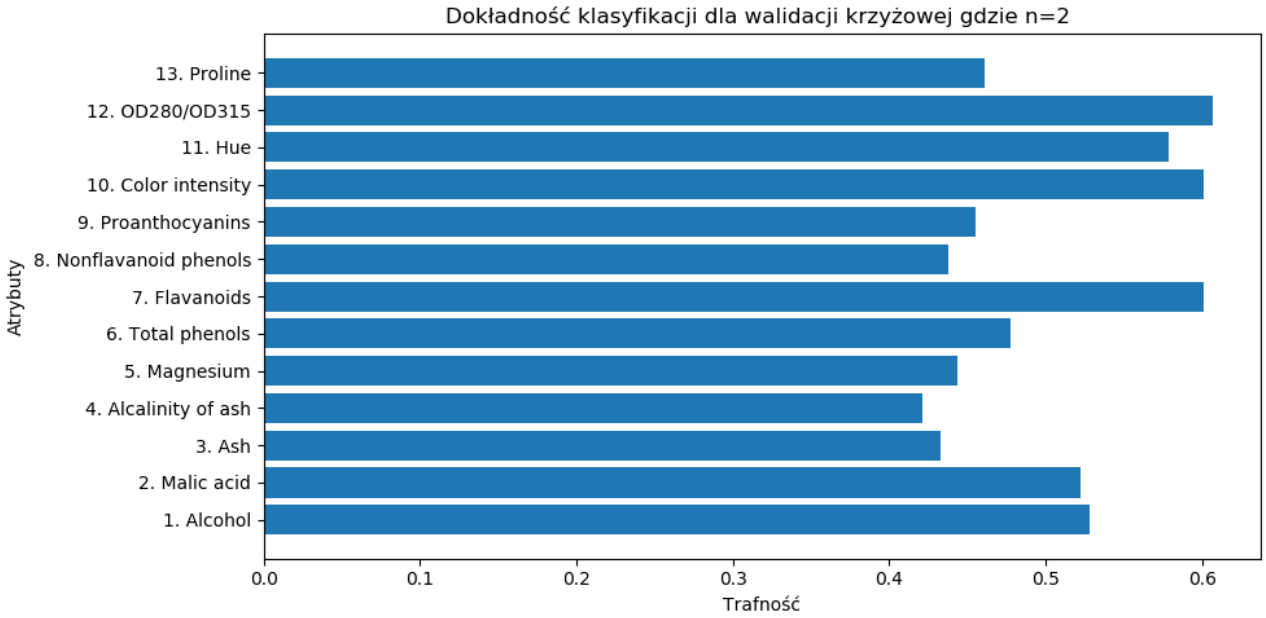
folds	attribute	avg_acc
2	12. OD280/OD315	0,61
5	7. Flavanoids	0,77
10	7. Flavanoids	0,81
15	7. Flavanoids	0,82
20	7. Flavanoids	0,81
30	7. Flavanoids	0,82

Tabela 2 Przedstawia wyniki dołączenia dodatkowego atrybutu do atrybutu zapewniającego największą dokładność klasyfikacji

folds	best_attribute	added_attribute	avg_acc_best	avg_acc_best_with_added
2	12. OD280/OD315	5. Magnesium	0,61	0,60
5	7. Flavanoids	5. Magnesium	0,76	0,84
10	7. Flavanoids	5. Magnesium	0,81	0,87
15	7. Flavanoids	5. Magnesium	0,82	0,84
20	7. Flavanoids	5. Magnesium	0,82	0,85
30	7. Flavanoids	5. Magnesium	0,83	0,85

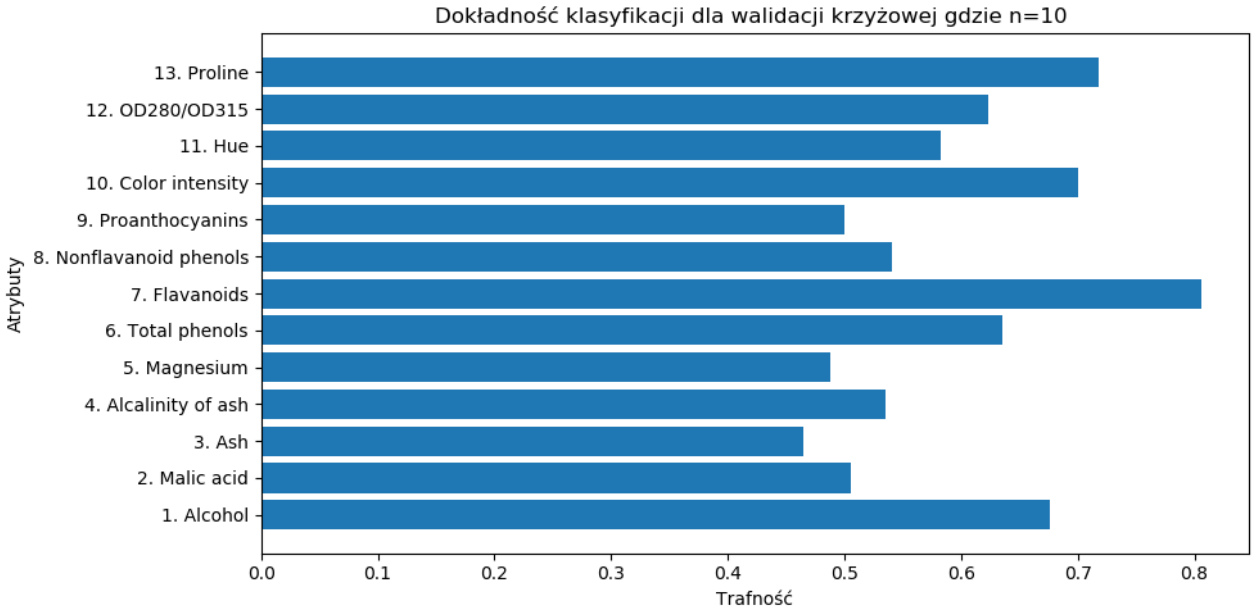
4. Analiza działania algorytmu

Rysunek 1



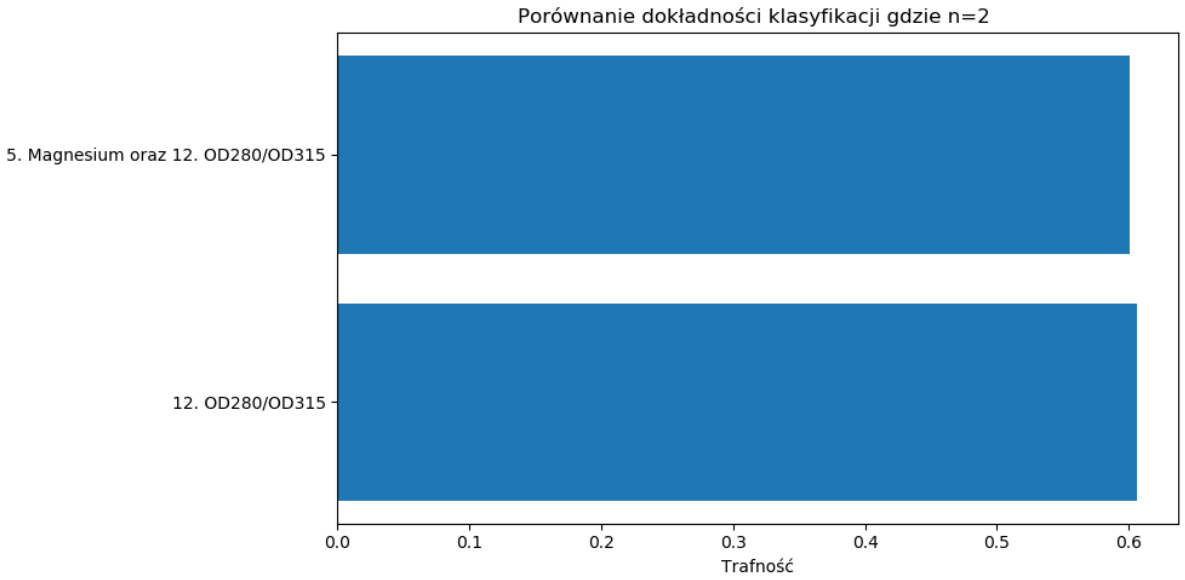
Jak widać na rysunku 1 zbyt niska wartość parametru n dla małego zbioru danych powoduje uszczuplenie zbioru uczącego a w rezultacie budowę słabej jakości modelu. Widzimy to po zestawieniu innych wyników z tabel 1 oraz 2. W efekcie jako atrybut zapewniający najlepszą dokładność klasyfikacji został wybrany inny atrybut niż w pozostałych przypadkach.

Rysunek 2



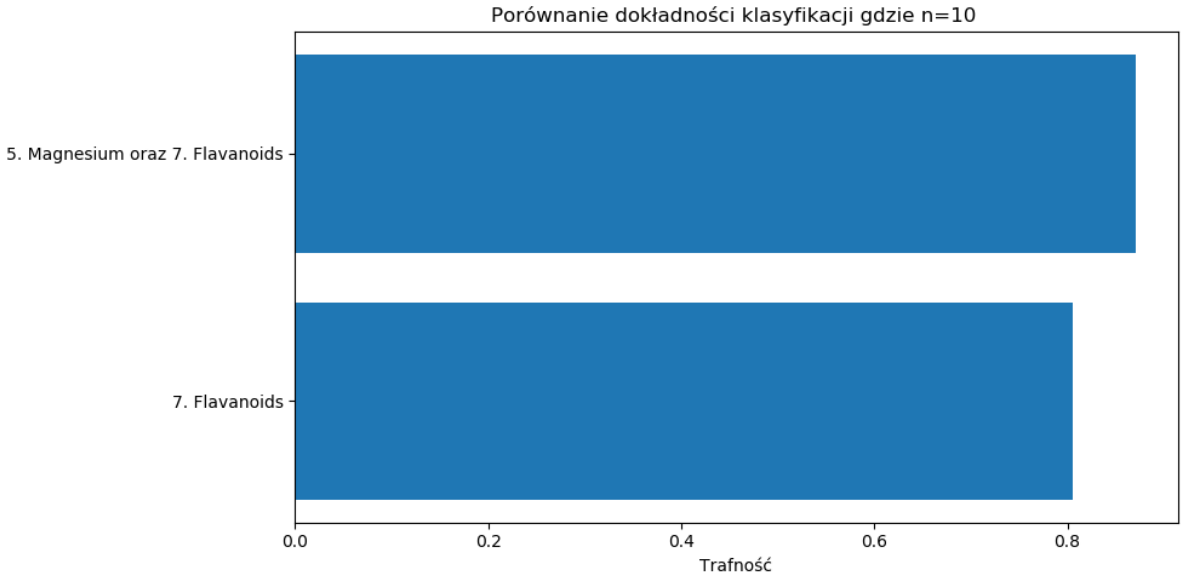
Na rysunku 2 widzimy wyniki dla zalecanej wartości n dla małych zbiorów n=10. Przeglądając wyniki w tabelach 1 oraz 2 widzimy, że otrzymane w ten sposób rezultaty są przystające do reszty wyników.

Rysunek 3



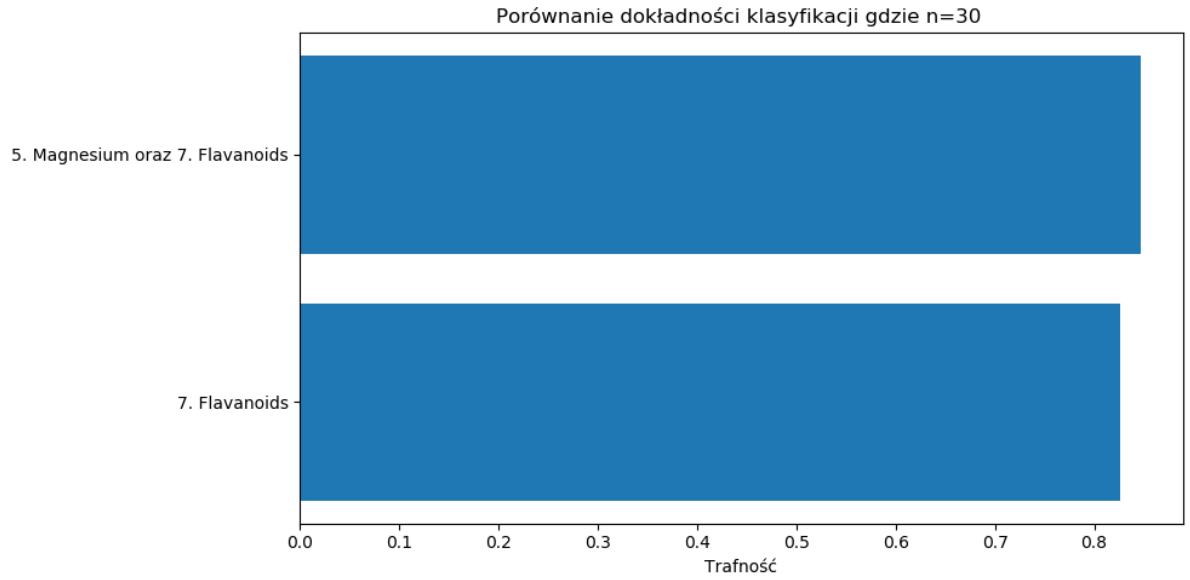
Na rysunku 3 oraz w tabeli 2 zauważyć możemy sytuacje , gdy dodanie kolejnego atrybutu spowodowało spadek trafności klasyfikacji. Warto zauważyć, że jako atrybut zapewniający najlepszą dokładność klasyfikacji został wybrany inny atrybut niż w pozostałych przypadkach.

Rysunek 4



Rysunek 4 wraz z tabelą 2 pokazuje jak wzrasta wartość dokładności klasyfikacji po dodaniu atrybutu dla n = 10.

Rysunek 5



Na rysunku 5 oraz w tabeli 2 dla n = 30 obserwujemy mniejszy wzrost dokładności klasyfikacji po dodaniu kolejnego atrybutu. Może być to spowodowane zbyt małym zbiorem testowym w którym brakuje reprezentantów poszczególnych z klas.

5. Wnioski i podsumowanie

W przypadku n-krotnej walidacji krzyżowej bardzo ważny jest dobór odpowiedniej wartości n do wielkości zbioru danych.

Dla małych zbiorów n powinno być dostatecznie duże by nie uszczuplać w znaczący sposób zbioru uczącego, co mogłoby doprowadzić do zbudowania słabej jakości modelu.

Natomiast zbyt duża wartość n może spowodować niereprezentatywny zbiór testowy w którym może brakować reprezentantów z poszczególnych klas. Efekt ten można minimalizować zapewniając stosunkowy udział reprezentantów klas w zbiorze testowym podobny do tego w zbiorze uczącym.

Dodawanie kolejnych atrybutów może poprawić jakość predykcji, jednak nie jest to pewne, niekiedy efekt może być niezauważalny.