POLITECHNIKA WROCŁAWSKA
WYDZIAŁ INFORMATYKI I ZARZĄDZANIA

KIERUNEK: INFORMATYKA
SPECJALNOŚĆ: DANOLOGIA
KURS: INDUKCYJNE METODY ANALIZY DANYCH

Klasyfikator oparty na twierdzeniu Bayesa przy naiwnym założeniu o wzajemnej niezależności atrybutów

Dokumentacja ćwiczenia nr 1

AUTOR:
Adam Dłubak

PROWADZĄCY:
Dr inż. Paweł Myszkowski

WROCŁAW, 2018

Rozdział 1

Opis algorytmu

Naiwny klasyfikator Bayesa jest to technika klasyfikacji oparta na teorii Bayesa z oszacowaniem niezależności pomiędzy wskaźnikami. Oznacza to, iż naiwny klasyfikator Bayesa zakłada, że obecność każdej poszczególnej cechy w kategorii jest niezwiązana z obecnością żadnej innej cechy. Przykładowo pojazd może być uznany za samochód jeśli ma cztery koła, silnik i kierownicę. Nawet jeśli te cechy zależą od siebie nawzajem albo od istnienia innych cech, wszystkie one niezależnie zwiększają prawdopodobieństwo, że ten pojazd to samochód i to właśnie dlatego w nazwie występuje słowo „naiwny".

Naiwny model Bayesa jest łatwy do zbudowania i szczególnie przydatny dla bardzo dużych zestawów danych. Naiwny Bayes, pomimo swojej prostoty, jest znany z przewyższania nawet bardzo wyrafinowanych metod klasyfikacji.

W języku programowania *Python* w bibliotece *Scikit learn* są trzy typy modeli Naiwnego Bayesa:

- **Gaussa** - jest wykorzystywany w klasyfikacji i zakłada, że cechy wynikają z rozkładu normalnego.
- **Wielomianowy** - używa się go do odrębnych wyliczeń. Przykładowo, jeżeli występuje problem z klasyfikacją tekstu. Tutaj można rozważyć próby Bernoullego, które są jeden krok dalej i zamiast „słowo występujące w tekście", należy „policz jak często słowo występuje w tekście", można pomyśleć o tym jak o „ilość razy gdy liczba wyników x(i) jest zaobserwowana przez n prób".
- **Bernoullego** - dwumianowy model przydaje się jeśli cechy wektorów są binarne (np. zera i jedynki). Jedna aplikacja będzie klasyfikacją tekstu z modelem „torbą słów", gdzie jedynki i zera są „słowami występującymi w tekście" i odpowiednio „słowami niewystępującymi w tekście".

Wzór, z którego korzysta nawiny Bayes jest uproszczony, gdyż pomija prawdopodobieństwo wystąpienia obserwacji (mianownik z twierdzenia Bayesa). Wynika to z tego, że przyjmuje się to samo prawdopodobieństwo wystąpienia zdarzenia w klasach. Dzięki temu zabiegowi etap uczenia jest bardzo szybki - proces wymaga tylko zliczania wystąpień.

Rozdział 2

Wykorzystane dane

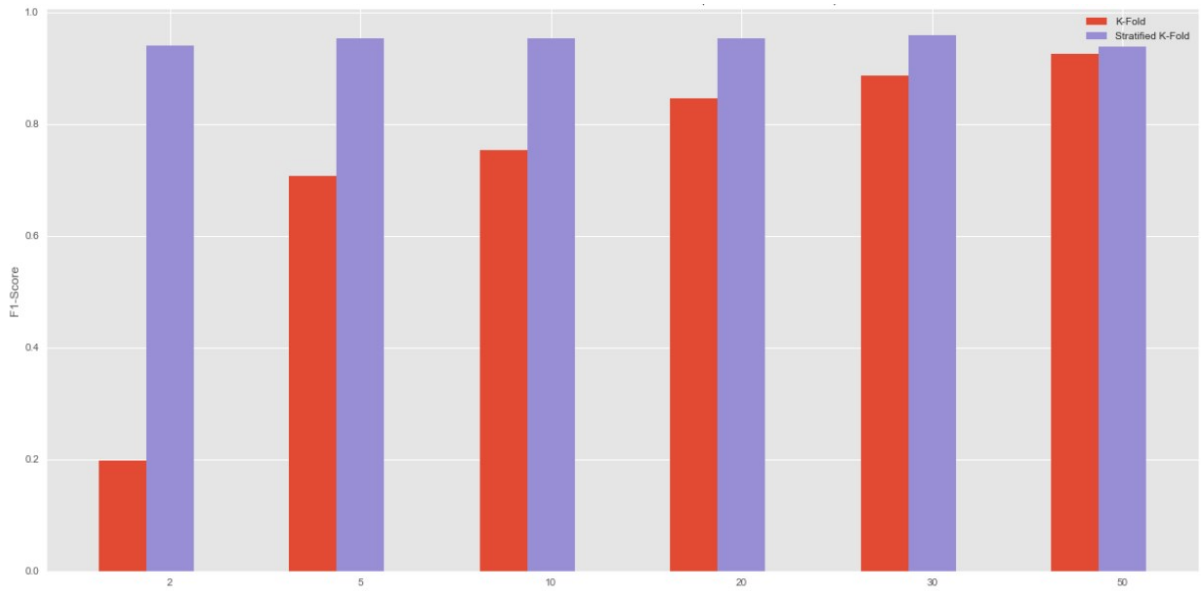
W ramach realizacji ćwiczenia, badania zostały oparte o 3 zbiory danych testowych, które uwzględniały również dane z wartościami ciągłymi. Wykorzystane zbiory danych to:

- **Wine dataset**
  Zbiór danych jest wynikiem analizy chemicznej win uprawianych w tym samym regionie we Włoszech, ale uzyskanych z trzech różnych odmian. W analizie określono 13 składników znalezionych w każdym z trzech rodzajów win.
  - Liczba atrybutów: 13.
  - Rodzaj atrybutów: wartości typu Integer.
  - Liczba instancji: 178.
  - Liczba klas: 3, występuje zróżnicowanie w ilości instancji w klasie.
- **Glass dataset**
  Zbiór danych powstał w wyniku motywacji badania dochodzeń kryminologicznych. Poprawne zidentyfikowanie rodzaju szkła znalezionego na miejscu przestępstwa, na postawie jego składu pozwala na użycie go jako dowodu w sprawie.
  - Liczba atrybutów: 10.
  - Rodzaj atrybutów: realistyczne, ciągłe.
  - Liczba instancji: 214.
  - Liczba klas: 7, występuje zróżnicowanie w ilości instancji w klasie.
- **Diabetes (Pima Indians Diabetes) dataset**
  Zbiór danych Pima Indian Diabetes przewidywanie wystąpienie cukrzycy w oparciu o badania diagnostyczne. Pochodzi on z *National Institute of Diabetes and Digestive and Kidney Diseases*. Celem jego stworzenia jest przewidywanie, czy pacjent cierpi na cukrzycę w oparciu o badania diagnostyczne.
  - Liczba atrybutów: 9.
  - Rodzaj atrybutów: realistyczne, ciągłe.
  - Liczba instancji: 768.
  - Liczba klas: 2.

Rozdział 3

Badania i Analiza Danych

3.1. Kroswalidacja i stratyfikowana kroswalidacja



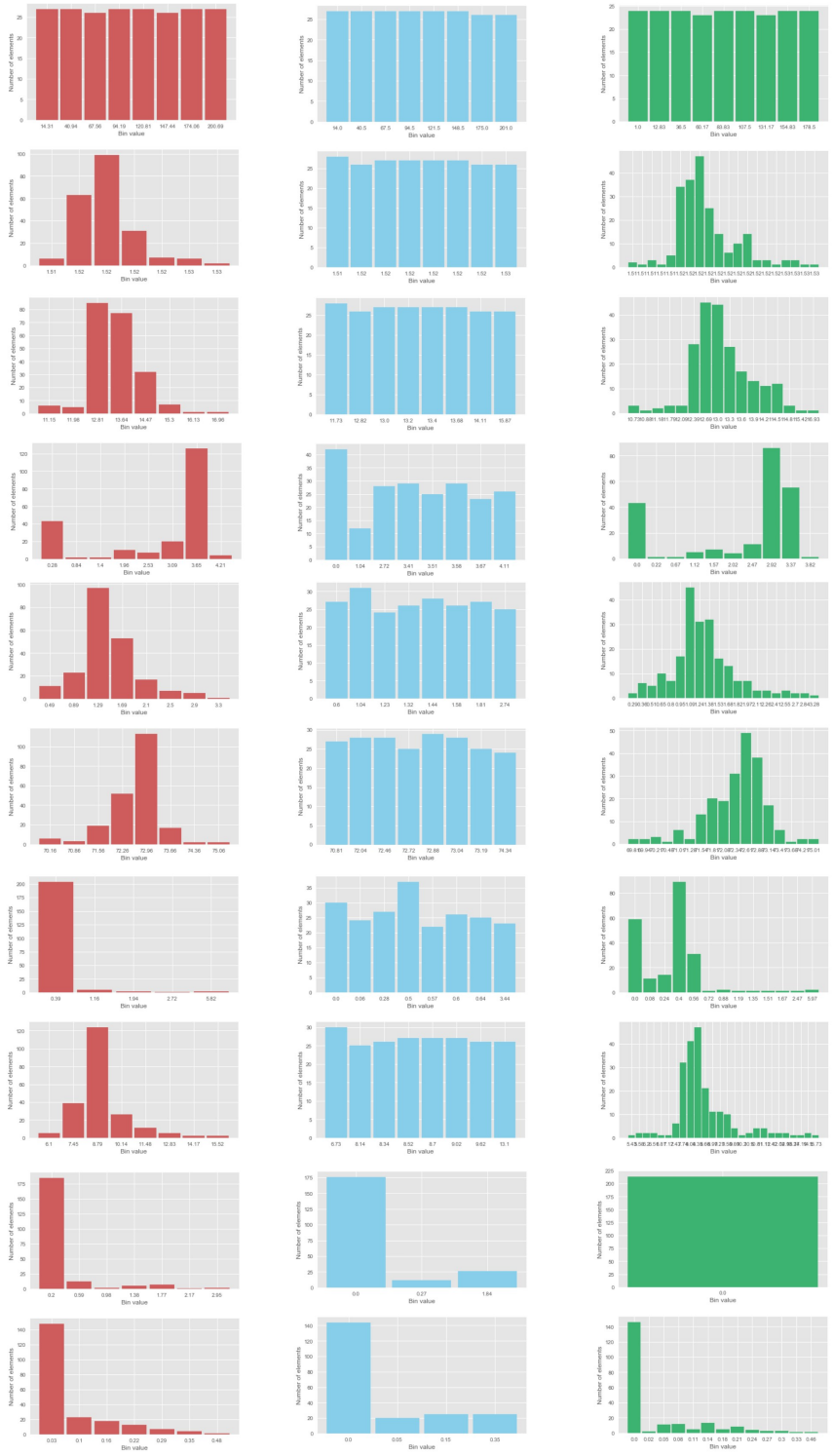
Rys. 3.1: K-Fold vs Stratified K-Fold for Wine Dataset (Gaussian Classifier)

Tab. 3.1: K-Fold vs Stratified K-Fold for Wine Dataset (Gaussian Classifier

Liczba Foldów	**2**	**5**	**10**	**20**	**30**	**50**
K-Fold	0.350	0.614	0.694	0.747	0.776	0.789
Stratified K-Fold	0.545	0.773	0.813	0.815	0.819	0.790

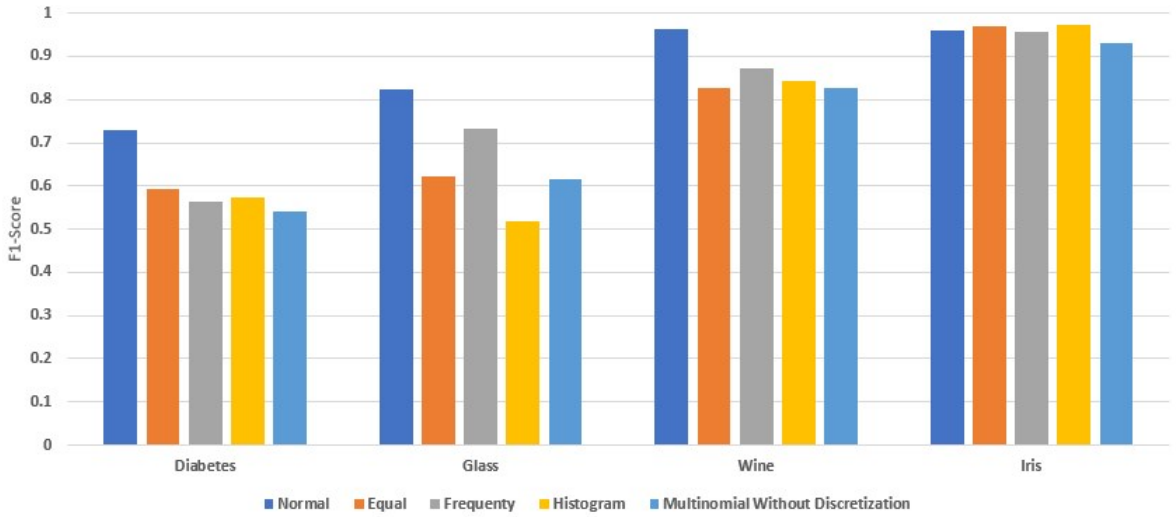
Przedstawione powyżej dane dokładnie wskazują na przewagę sposobu stratyfikowania danych w foldach. Dzięki temu uzyskiwane wyniki są stanowczo lepsze (szczególnie dla małej liczby foldów), niż w przypadku, gdy dane w nich zostaną rozmieszone losowo. Dopiero w bardzo dużej liczbie foldów (30 i więcej), ta znaczna różnica rozpoczyna się zacierać jednak nadal jest widoczna.

3.2. Dyskretyzacja



Rys. 3.2: Dyskretyzacja zbioru Glass

3.3. Analiza sposobów dyskretyzacji



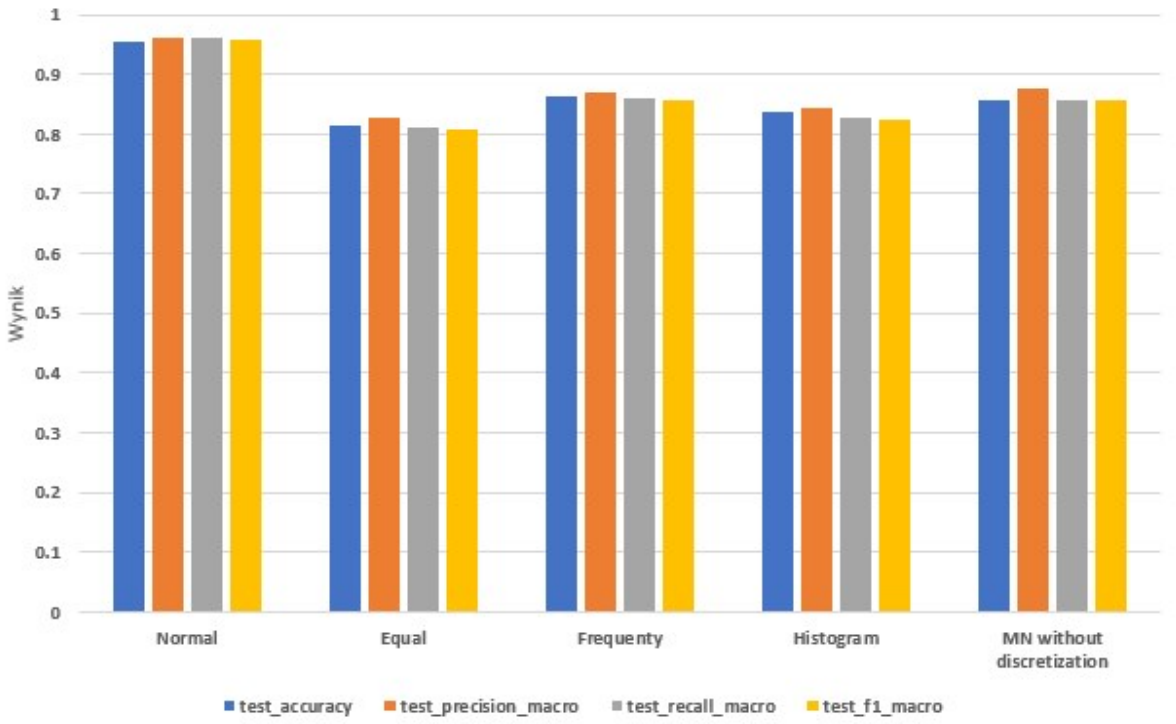
Rys. 3.3: Porównanie sposobów dyskretyzacji względem różnych zbiorów danych

Tab. 3.2: Porównanie sposobów dyskretyzacji względem różnych zbiorów danych

	Diabetes	**Glass**	**Wine**	**Iris**
Normal	0.730	0.825	0.962	0.961
Equal	0.593	0.622	0.827	0.970
Frequenty	0.563	0.732	0.871	0.957
Histogram	0.572	0.517	0.843	0.972
Multinomial Without Discretization	0.541	0.616	0.826	0.931

W zestawieniu tym porównane zostały wszystkie analizowane 4 zbiory danych oraz 3 sposoby dyskretyzacji. Ponadto został przeprowadzony test wykorzystujący klasyfikator Gaussowski oraz klasyfikator Multinomial na danych, które nie zostały zdyskretyzowane.

Powołując się na treści literatury, z przebadanych danych powinno wynikać, iż klasyfikator oparty o założenie rozkładu normalnego danych powinien radzić sobie z ich kwalifikacją najgorzej, jednakże według przeprowadzonych danych jest wręcz przeciwnie. Zachowuje się on nad wyraz dobrze. Wynikać to może z nie do końca poprawnej implementacji dyskretyzacji danych. Biorąc jednak pod uwagę jedynie sposoby dyskretyzacji danych, widać iż nie ma jednego, najskuteczniejszego sposobu dyskretyzacji danych. W zależności od zbioru na którym prowadzone są badania i rozkładzie jego danych, różnie radzą sobie sposoby dyskretyzacji. W przypadku zbioru Glass zdecydowanie najlepszy okazał się sposób dyskretyzacji z uwzględnieniem równości występować atrybutów w każdym kubełku. Może to wynikać z bliskości rozkładu atrybutów tego zbioru do rozkładu normalnego. W przypadku zbioru Diabetes przykładowo, najlepszy okazał się sposób podziału na X równych kubełków.



Rys. 3.4: Zestawienie uzyskiwanych wyników na zbiorze Wine

Tab. 3.3: Zestawienie uzyskiwanych wyników na zbiorze Wine

	Normal	**Equal**	**Freqeunty**	**Histogram**	**MN without discr.**
Accuracy	0.956	0.815	0.863	0.837	0.858
Precision Macro	0.962	0.827	0.871	0.843	0.876
Recall Macro	0.960	0.810	0.860	0.827	0.856
F1 Macro	0.957	0.809	0.858	0.824	0.855