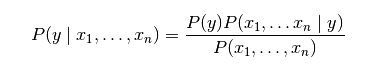
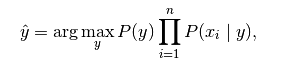
Wstęp

Naiwny klasyfikator bayesowski to prosty klasyfikator probabilistyczny oparty o twierdzenie Bayesa i założeniu o niezależności zmiennych losowych. Dla danej klasy obiektu y i wektora cech X na podstawie twierdzenia Bayesa prawdziwy jest wzór:



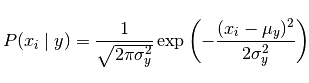
Korzystając z założenia o niezależności zdarzeń i przekształceń możemy dojść do wzoru:



Dzięki takiemu mechanizmowi na podstawie ciągu uczącego można wytrenować klasyfikator, a następnie wykorzystać go do klasyfikacji nowych obiektów. Do badania jakości uzyskanych klasyfikatorów użyte zostaną następujące mechanizmy: *Confusion Matrix*, ccuracy, Precision, Recall, Fscore. Badania zostaną przeprowadzone na trzech zbiorach danych: Iris, Wine oraz Diabetes.

Implementacja klasyfikatora i problem wygładzania

Na podstawie zaprezentowanych wcześniej wzorów można stwierdzić, że w przypadku, gdy dana kombinacja cechy i wartości nie wystąpi w zbiorze uczącym, wyzeruje ona prawdopodobieństwo klasyfikacji do danej klasy przy wystąpieniu cechy w czasie klasyfikacji właściwej. Z tym zjawiskiem można poradzić sobie poprzez wygładzenie danych, czyli eliminację zerowych prawdopodobieństw lub założenie, że dane mają rozkład normalny. W tym wypadku można skorzystać ze wzoru, który likwiduje zerowe prawdopodobieństwa.



Używanym klasyfikatorem będzie naive\_bayes z biblioteki sklearn. Do klasyfikatora zadajemy wektory z cechami obiektów ciągu uczącego oraz klasy, z których pochodzą. Następnie po zadaniu nowego wektora cech, otrzymamy wektor przyporządkowani do konkretnej klasy.

Metody dyskretyzacji

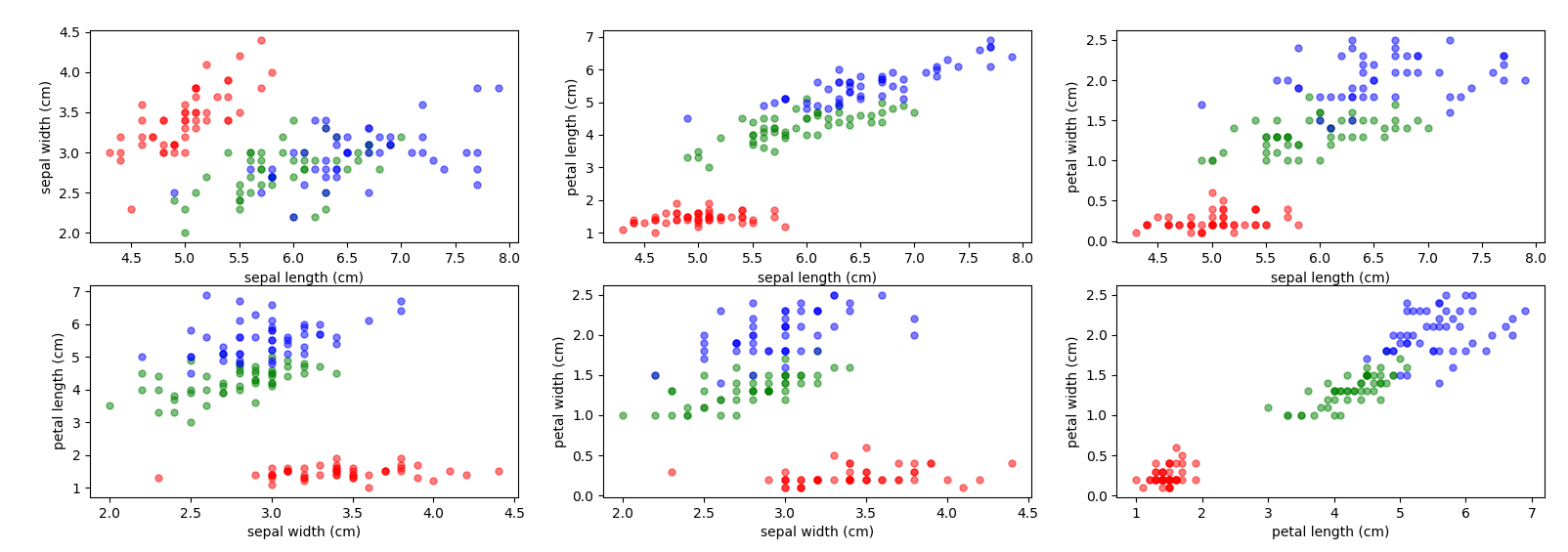
Jednym z celów zadania jest zbadanie wpływu dyskretyzacji danych na jakość klasyfikatora. W programie zaimplementowany zostały trzy rodzaje dyskretyzacji.

*equal width intervals -* podział zakresu wartości atrybutu ciągłego na $ k$ przedziałów o jednakowej szerokości

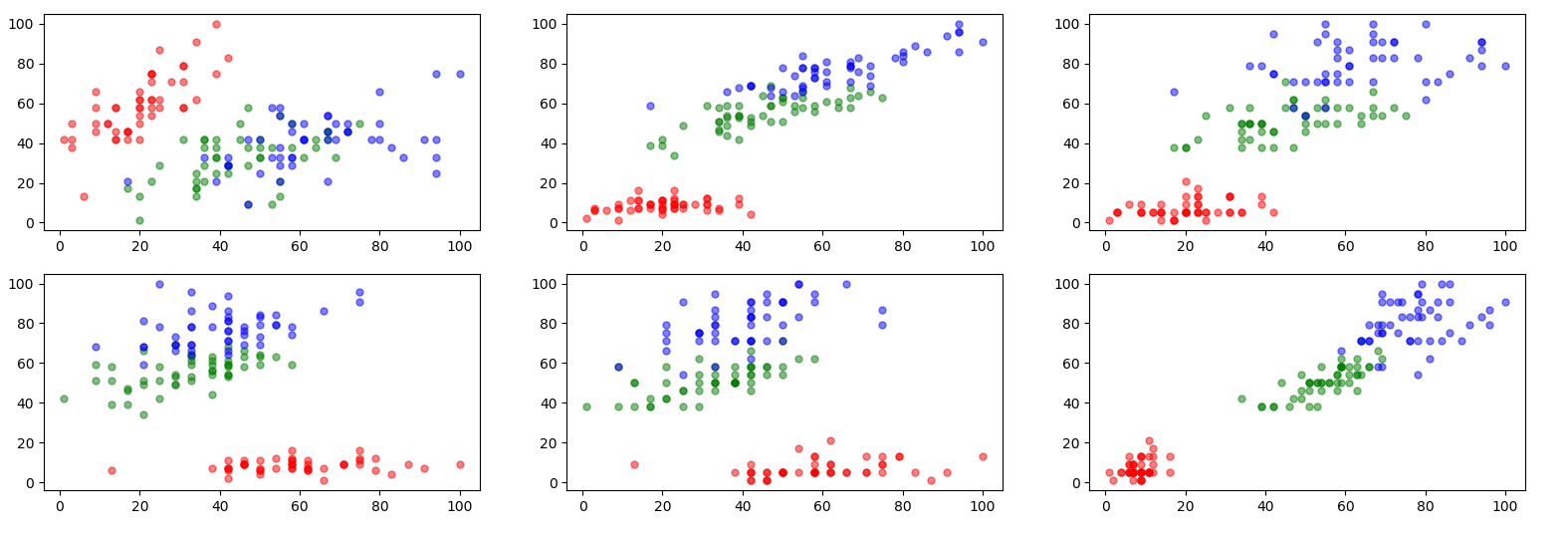
*equal frequency intervals -* podział zakresu wartości atrybutu ciągłego na $ k$ przedziałów, z których każdemu odpowiada możliwie tyle samo przykładów ze zbioru trenującego

quantile transform – konwersja danych do postaci rozkładu normalnego i przypisanie, użyta zostanie funkcja dostępna w pakiecie sklearn

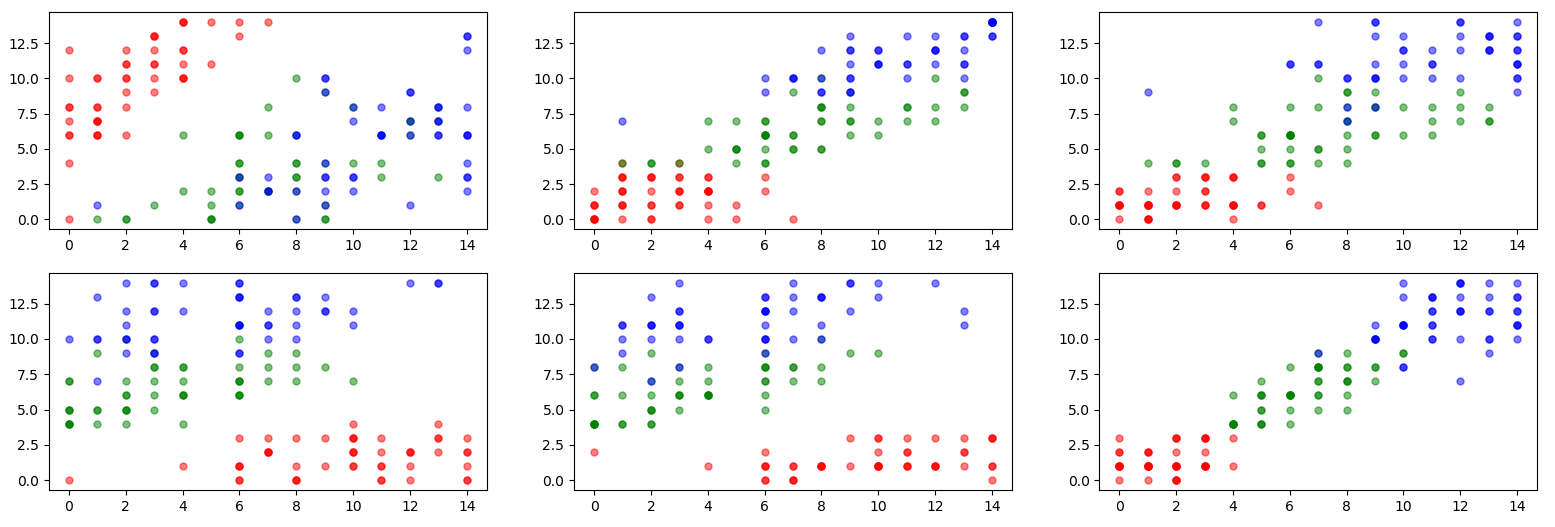
Działanie poszczególnych metod dyskretyzacji przedstawiono poniżej



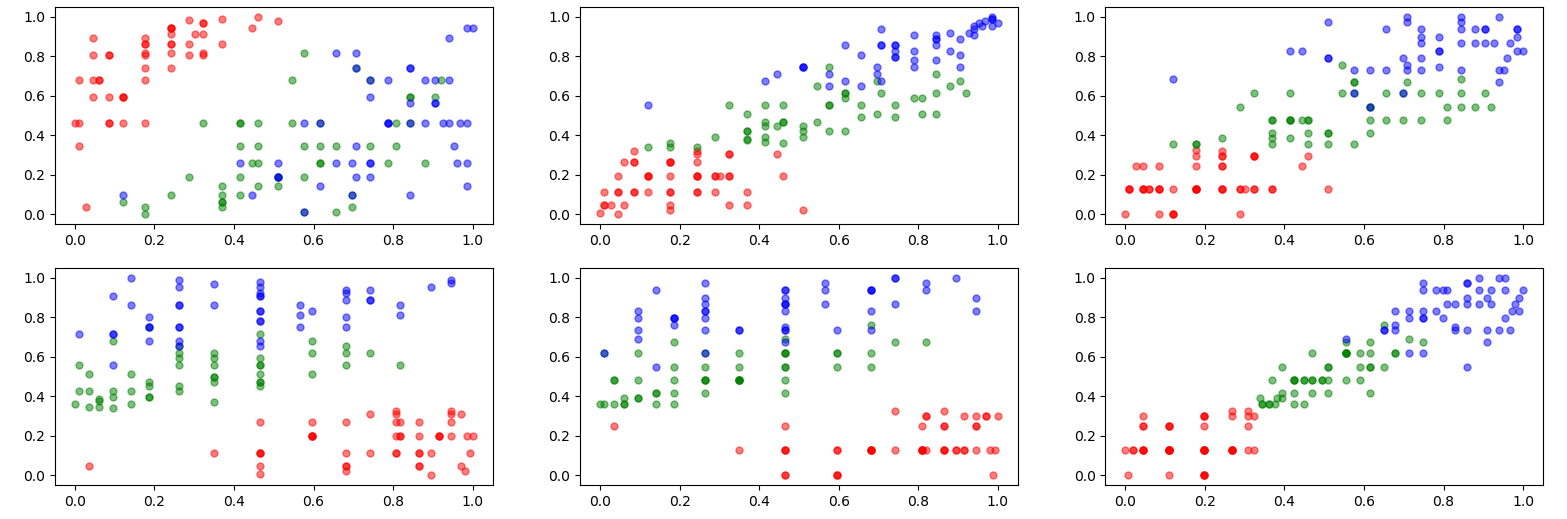
Brak dyskretyzacji



*equal width intervals (100 przedziałów)*



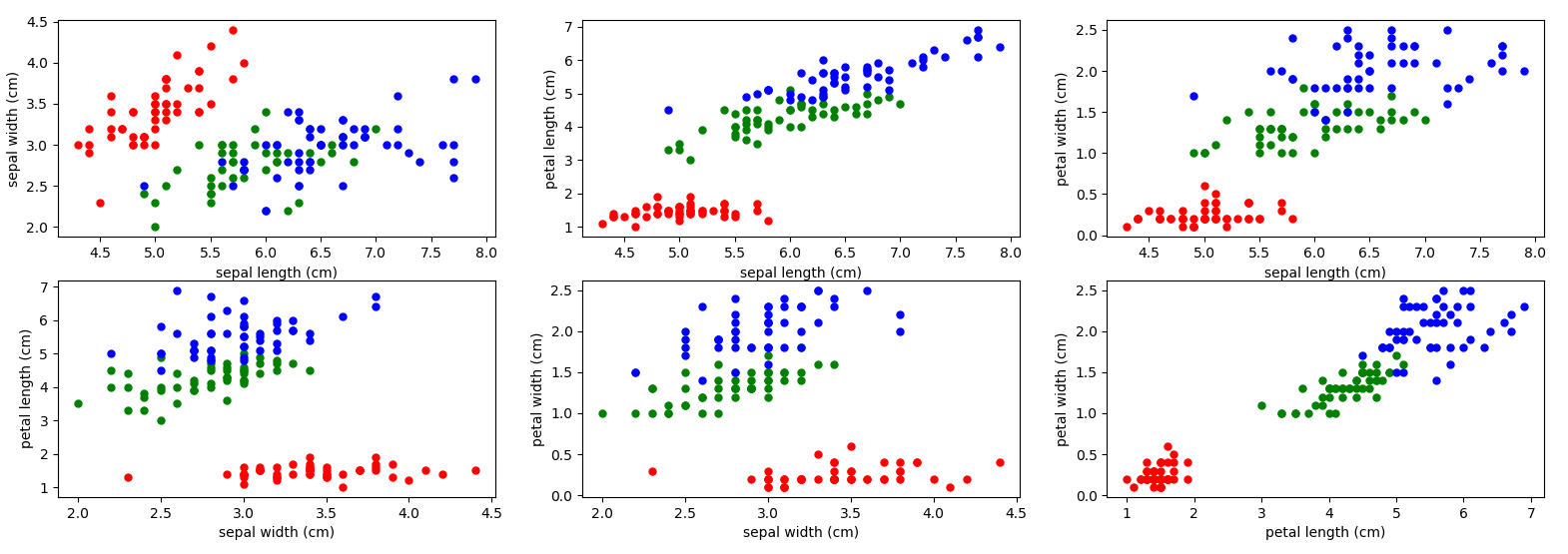
*equal frequency intervals (15 grup)*



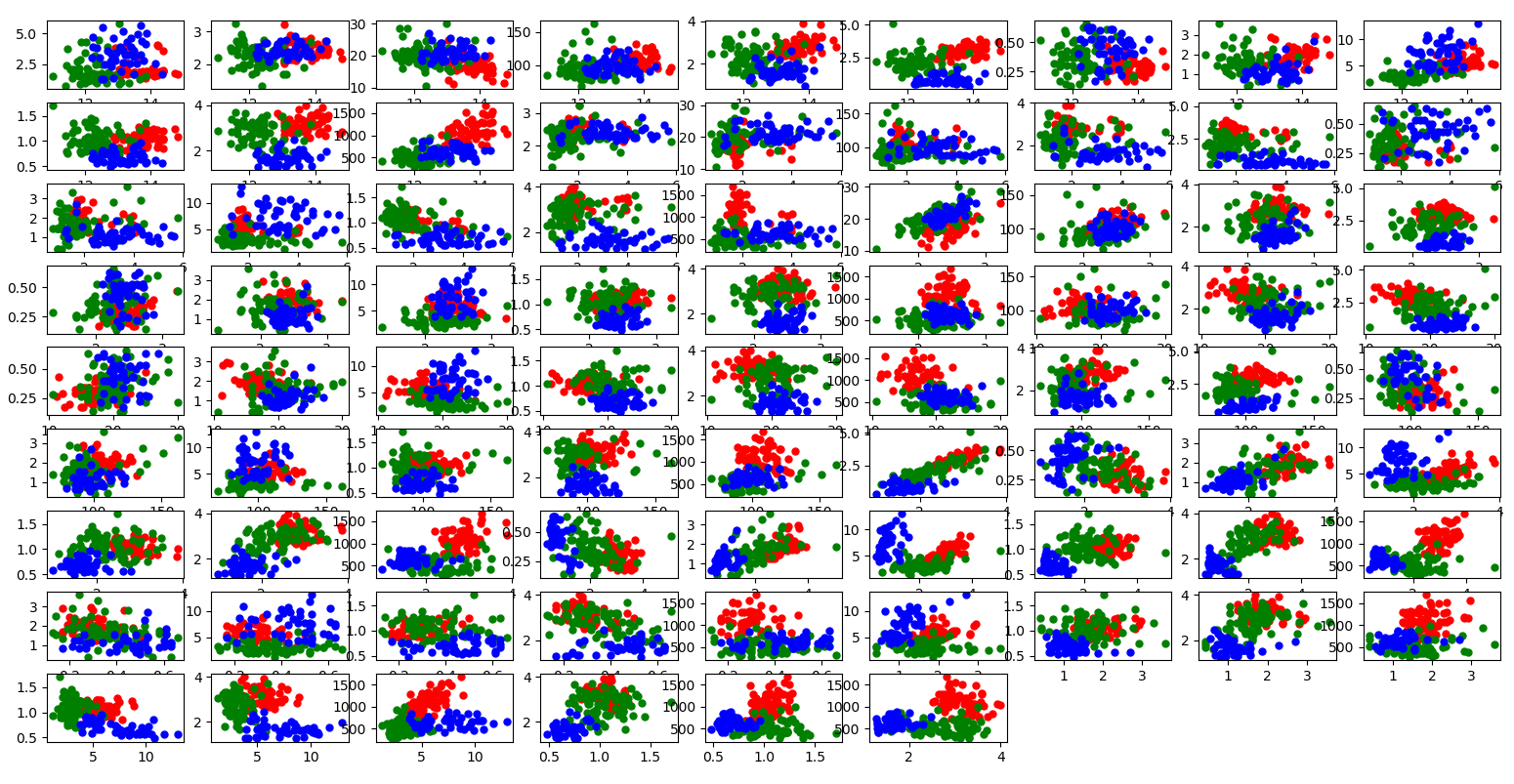
quantile transform

Przedstawienie omawianych zbiorów

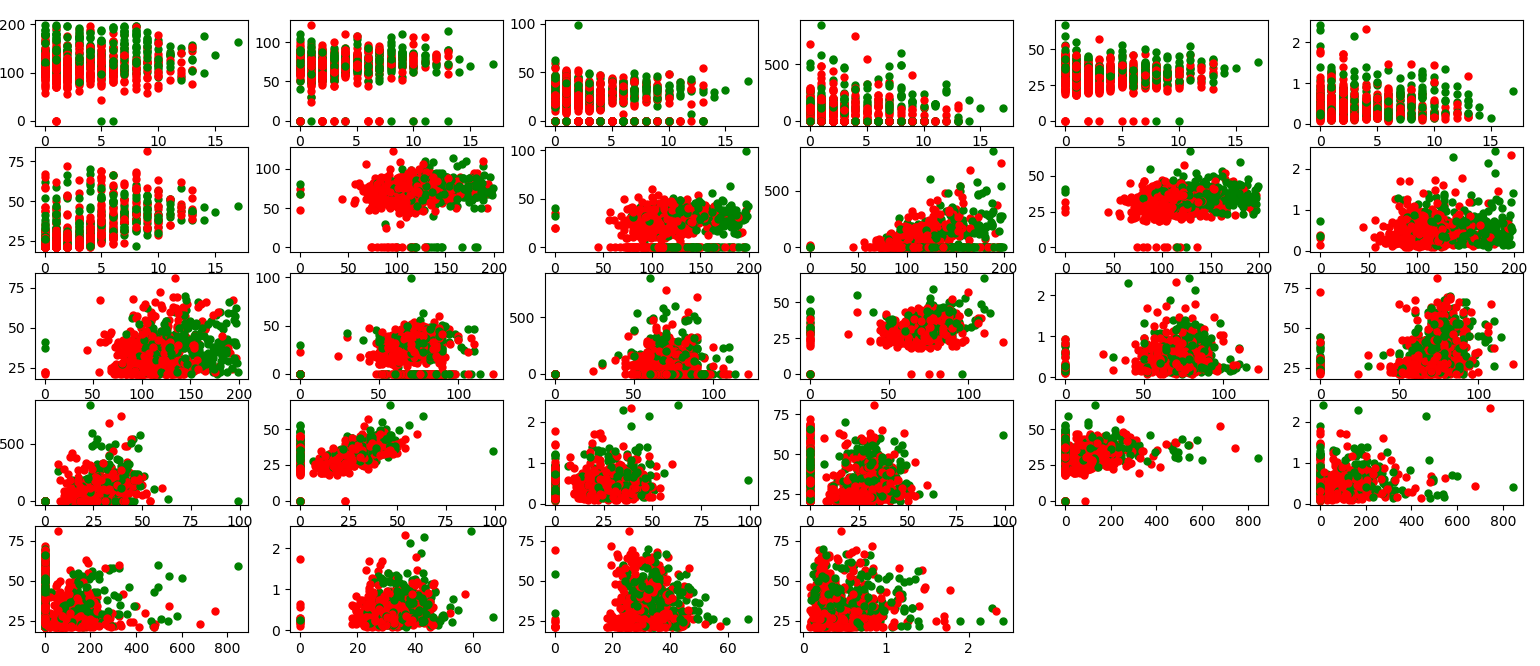
Iris



Wine



Diabetes

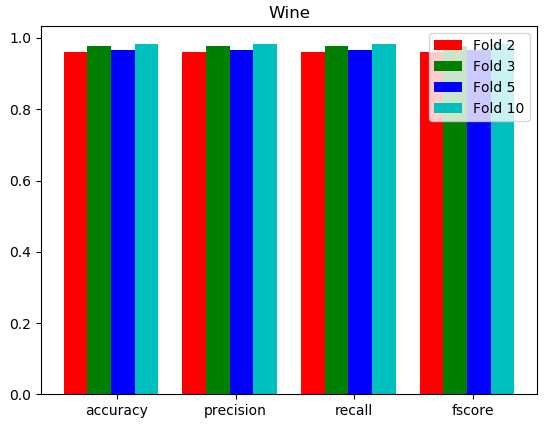


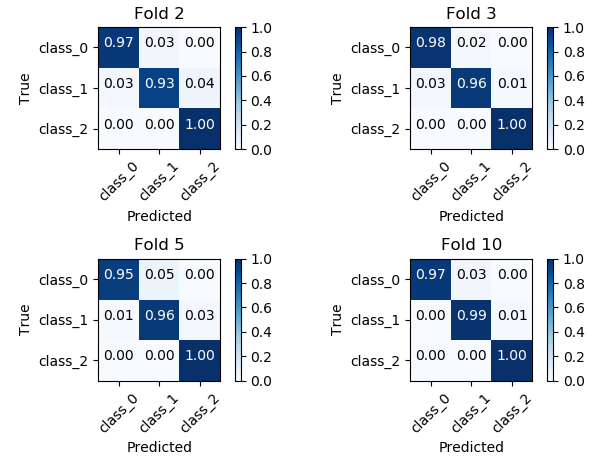
**Porównanie działania algorytmu przy różnych podziałach danych i sposobie liczenia prawdopodobieństwa**

Do oceny klasyfikatora zostanie użyta metoda X-fold. Polega ona na tym, że zbiór dzielimy na X w miarę możliwości równych części. Każda część zawiera możliwie tyle samo danych z każdej klasy. Jedna część zostanie użyta do oceny klasyfikatora, a pozostałe wejdą w skład zbioru trenującego. Następnie X razy zmianie ulegnie część do klasyfikacji, a cały proces zostanie powtórzony.

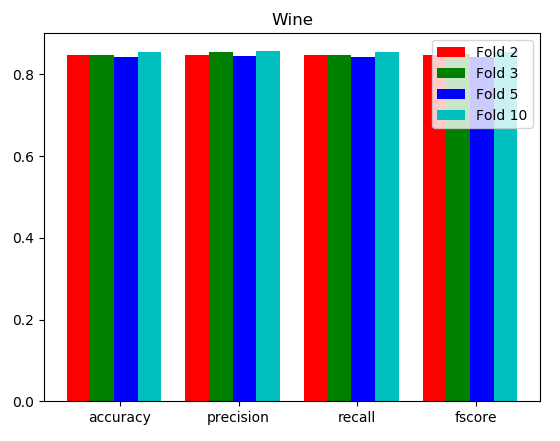
Przetestowana zostanie metoda uczenia poprzez podział zbioru na 3, 5 i 10 części. Ponadto przetestowane zostaną dwie metody liczenia prawdopodobieństwa – Gausian oraz Multinomial z wygładzaniem laplaca.

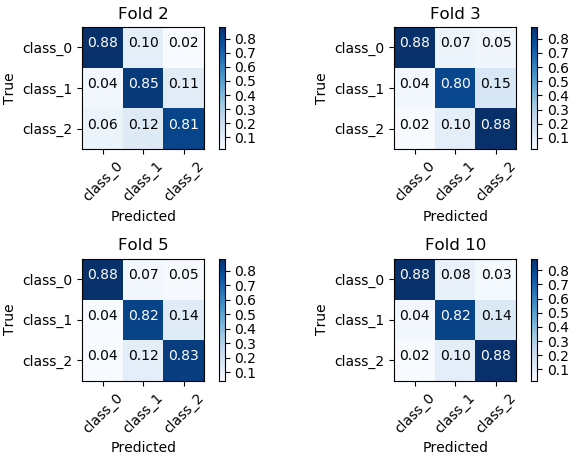
Gausian:





Multinomial:



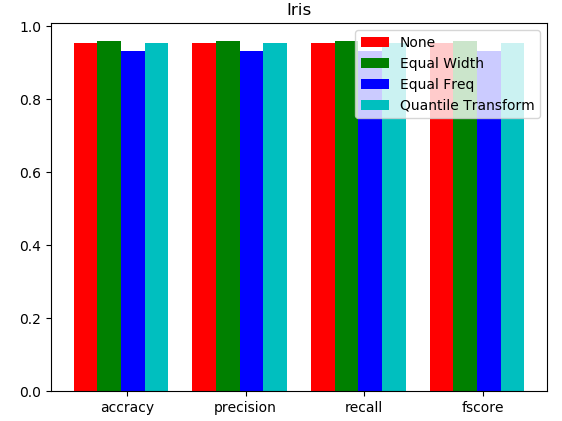
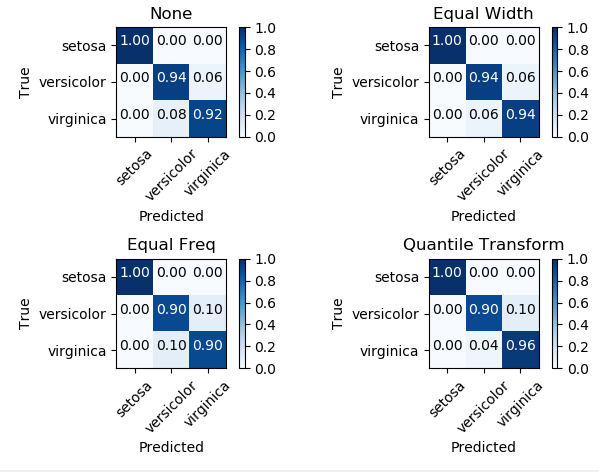


Można zauważyć, że zmiana ilości podziałów nie ma wpływu na wynik klasyfikacji, co jest spowodowane rozkładem badanych danych. Natomiast zmiana metody liczenia prawdopodobieństwa na sprawiła, że wyniki klasyfikatora uległy znaczącemu pogorszeniu.

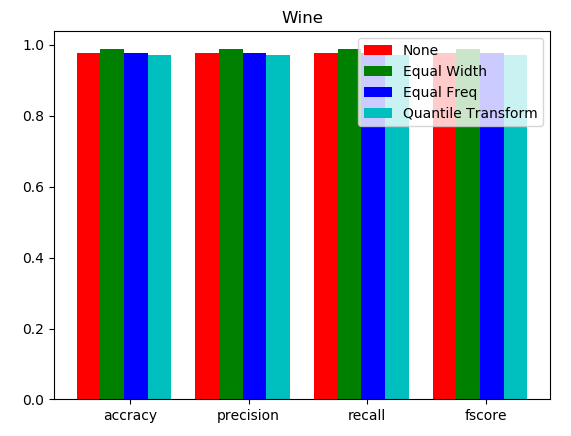
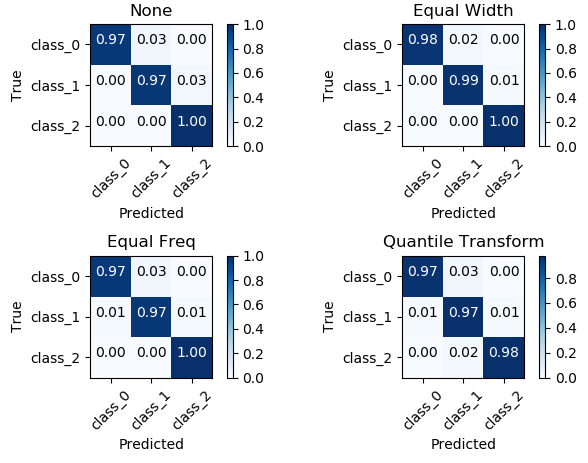
W późniejszych eksperymentach użyta zostanie metoda 10-fold do walidacji oraz Gausian do liczenia prawdopodobieństwa.

**Porównanie działania metod dyskretyzacji dla badanych zbiorów**

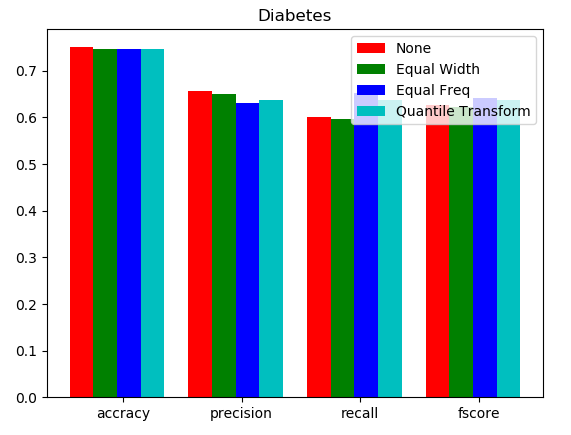
Zbiór Iris

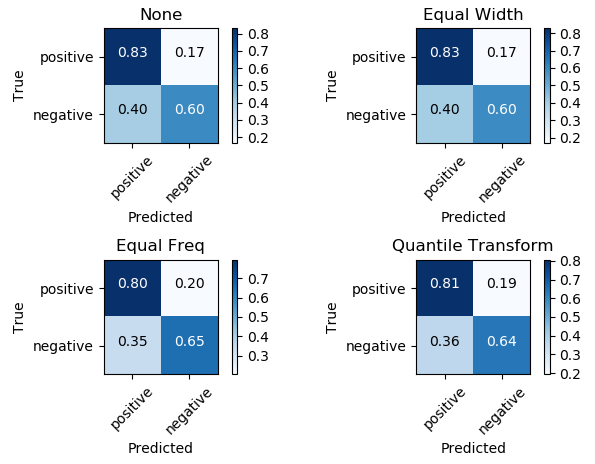
 

wine

Diabetes





Na podstawie przeprowadzonych obserwacji możemy wyciągnąć wniosek, że zaimplementowane metody dyskretyzacji nie wpływają znacząco na zmianę działania klasyfikatora dla żadnego ze zbiorów. Spowodowane jest to rozkładem badanych zbiorów danych.