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

Łukasz Odwrot 218283

13.03.2018

Spis treści

1 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:

$$P(y \mid x_1, ..., x_n) = \frac{P(y)P(x_1, ...x_n \mid y)}{P(x_1, ..., x_n)}$$

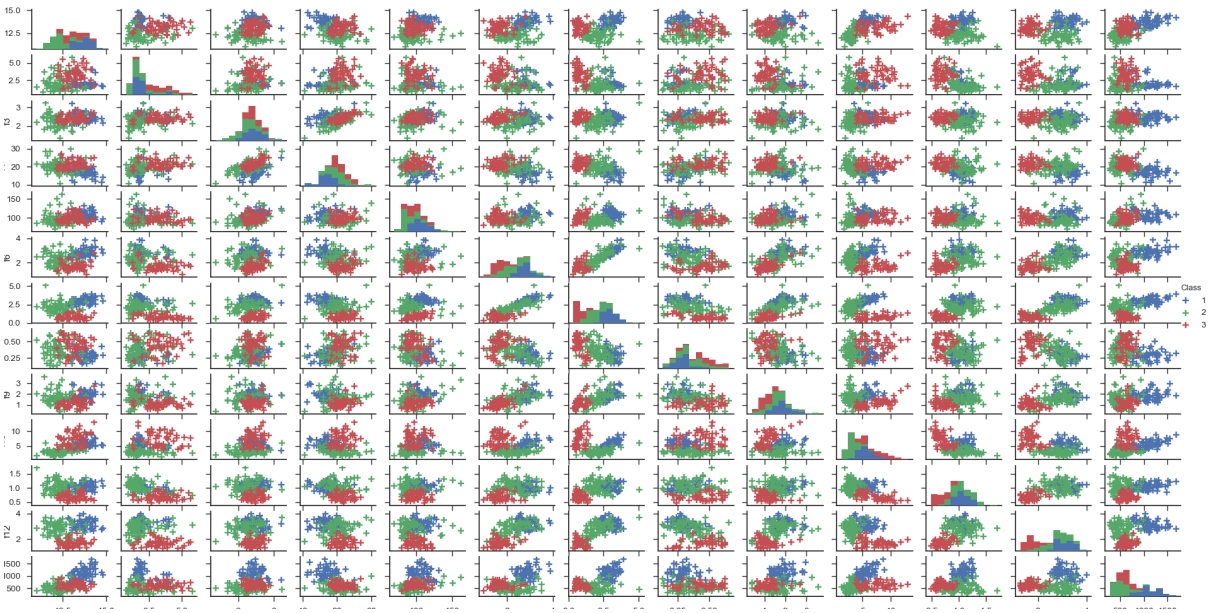
Korzystając z założenia o niezależności zdarzeń i przekształceń można dojść do wzoru:

$$\hat{y} = arg \max_y P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

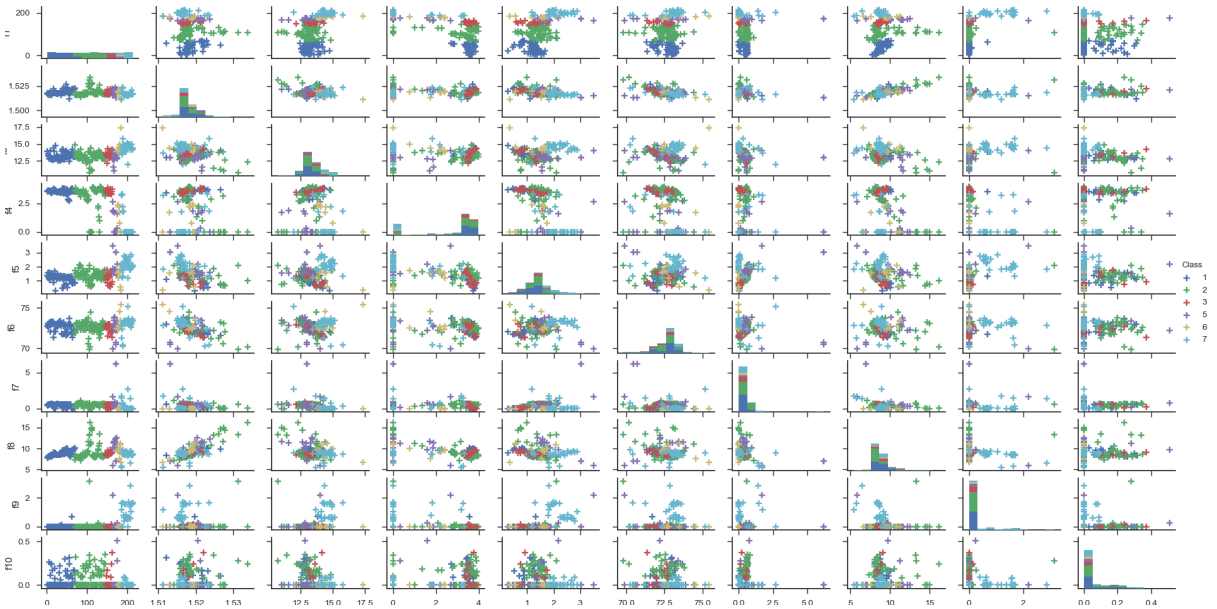
Dzięki takiemu mechanizmowi na podstawie ciągu uczącego można wytrenować klasyfikator, a następnie wykorzystać go do klasyfikacji nowych obiektów. Do badania jakości uzyskanych klasyfikatorów użyte zostaną następujące mechanizmy: Confusion Matrix, ccuracy, Precision, Recall, Fscore. Badania zostaną przeprowadzone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

2 Badane zbiory

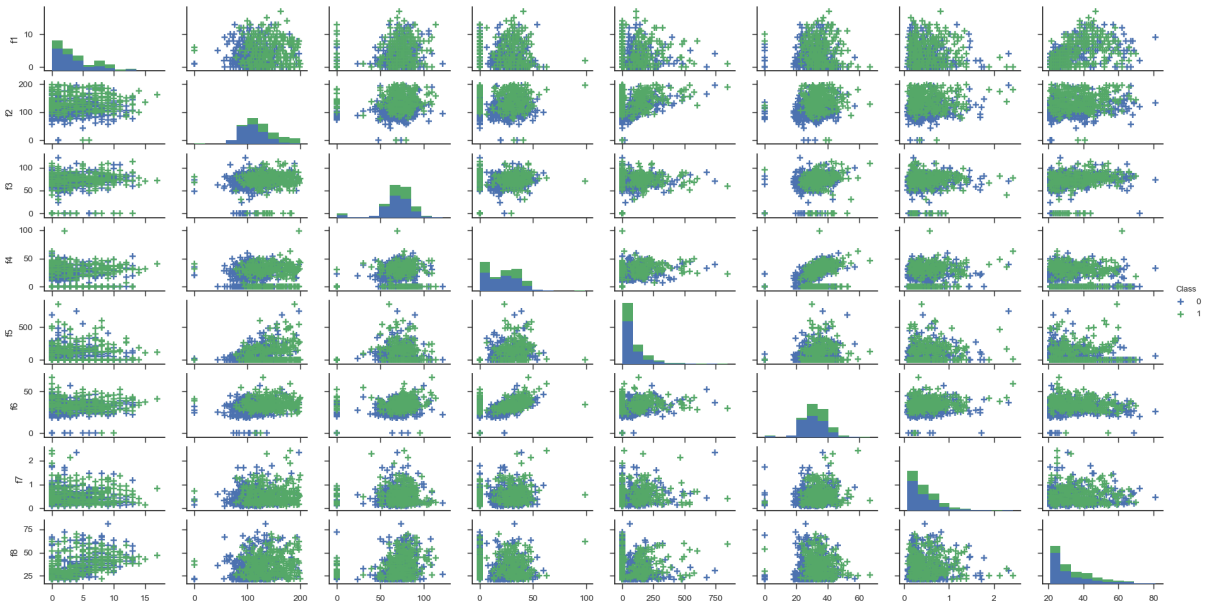
Rozkłady cech dla poszczególnych klas przedstawiono na poniższych rysunkach.



Rysunek 1: Rozkład cech dla zbioru Wine



Rysunek 2: Rozkład cech dla zbioru Glass



Rysunek 3: Rozkład cech dla zbioru Diabetes

3 Porównanie metod kroswalidacji

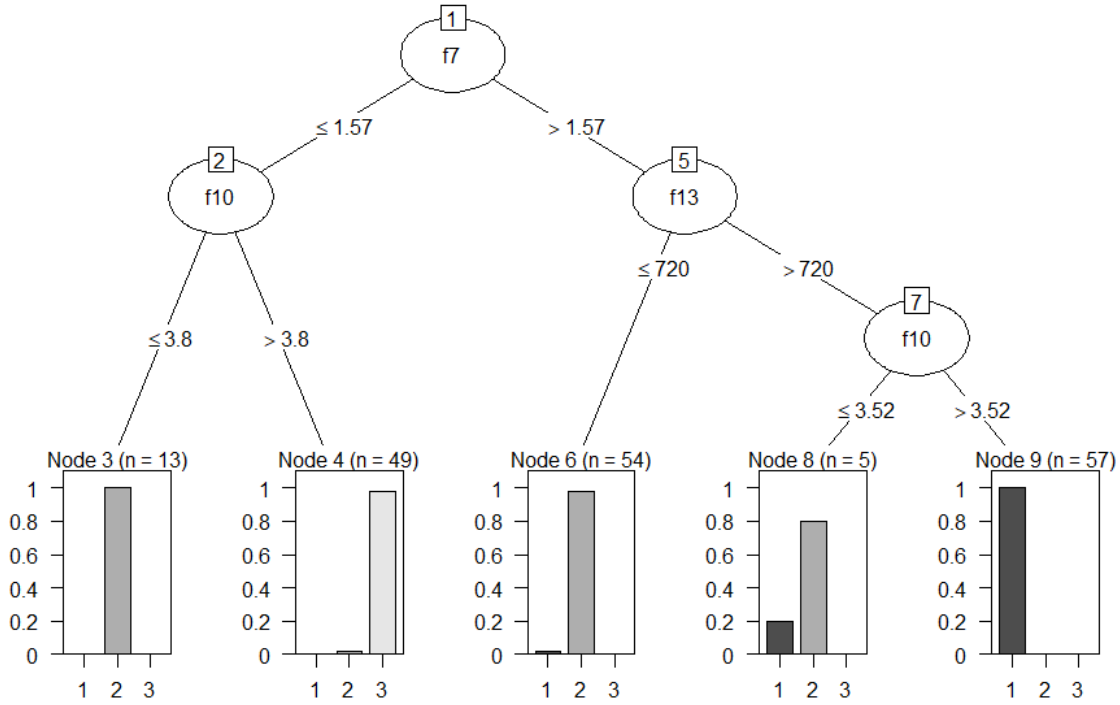
Zaimplementowane zostały dwie metody kroswalidacji:
- Niestratyfikowana (ręcznie)
- Stratyfikowana (z pomocą biblioteki caret)

4 Badane parametry

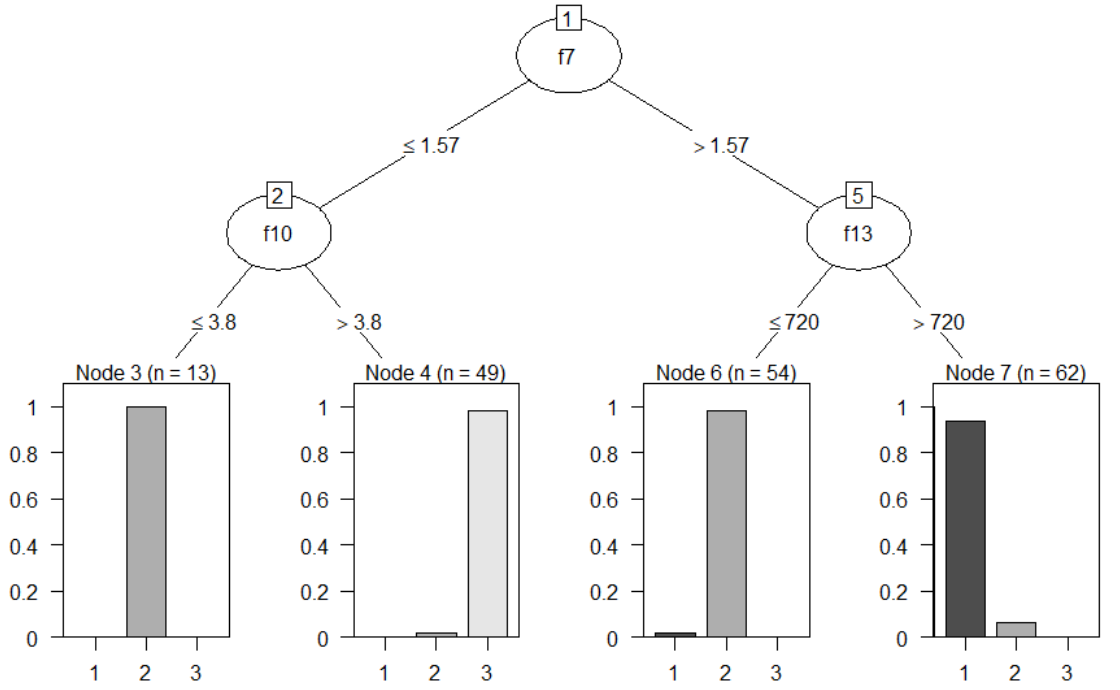
Do zbadania wybrano wpływ modyfikacji 4 różnych parametrów na drzewo. Domyślna konfiguracja parametrów drzewa wygląda następująco: (subset = TRUE, bands = 0, winnow = FALSE, noGlobalPruning = FALSE, CF = 0.25, minCases = 2, fuzzyThreshold

= FALSE, sample = 0, seed = sample.int(4096, size = 1) -1L, earlyStopping = TRUE, label = "outcome")

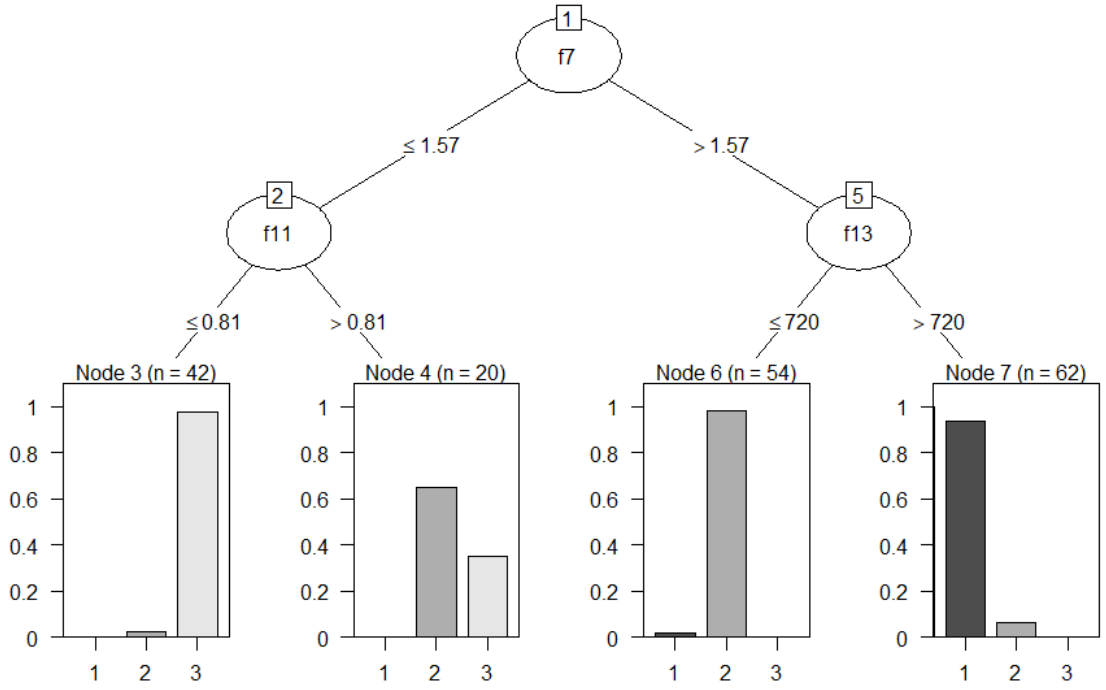
MinCases - określa minimalną ilość próbek w danym liściu.



Rysunek 4: Drzewo klasyfikacji instancji wine przy ustawieniu parametru minCases = 5

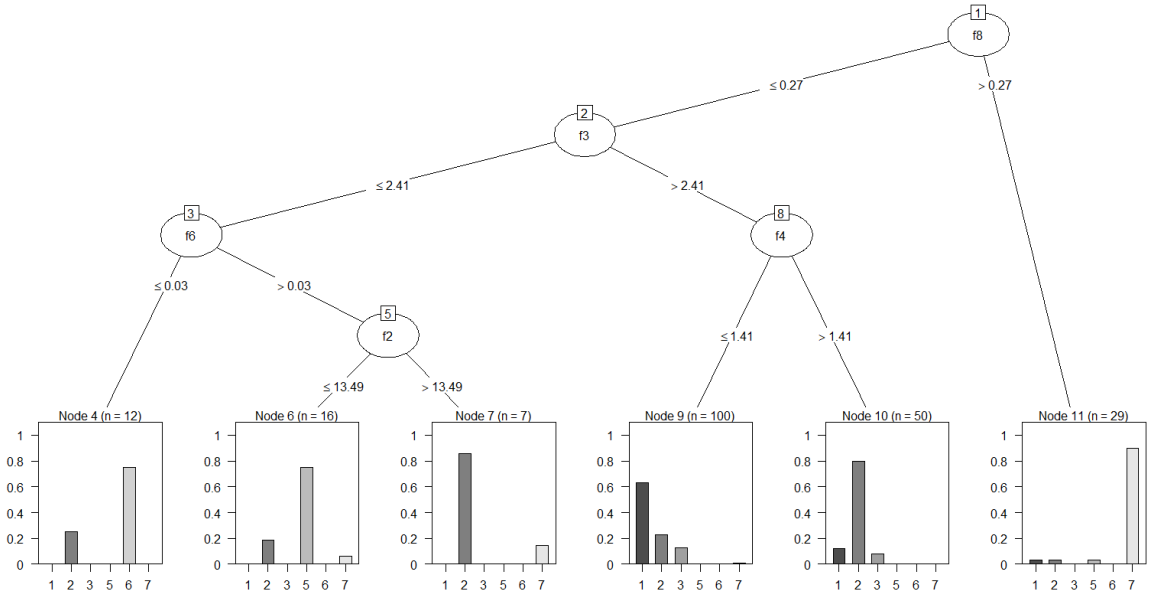


Rysunek 5: Drzewo klasyfikacji instancji wine przy ustawieniu parametru minCases = 10

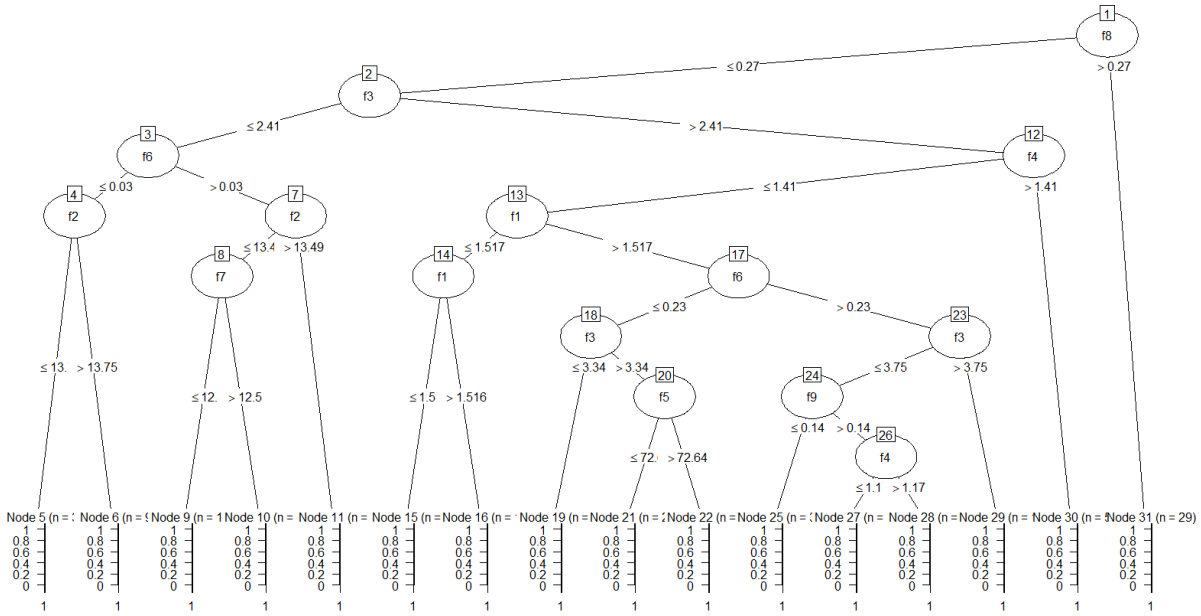


Rysunek 6: Drzewo klasyfikacji instancji wine przy ustawieniu parametru minCases = 20

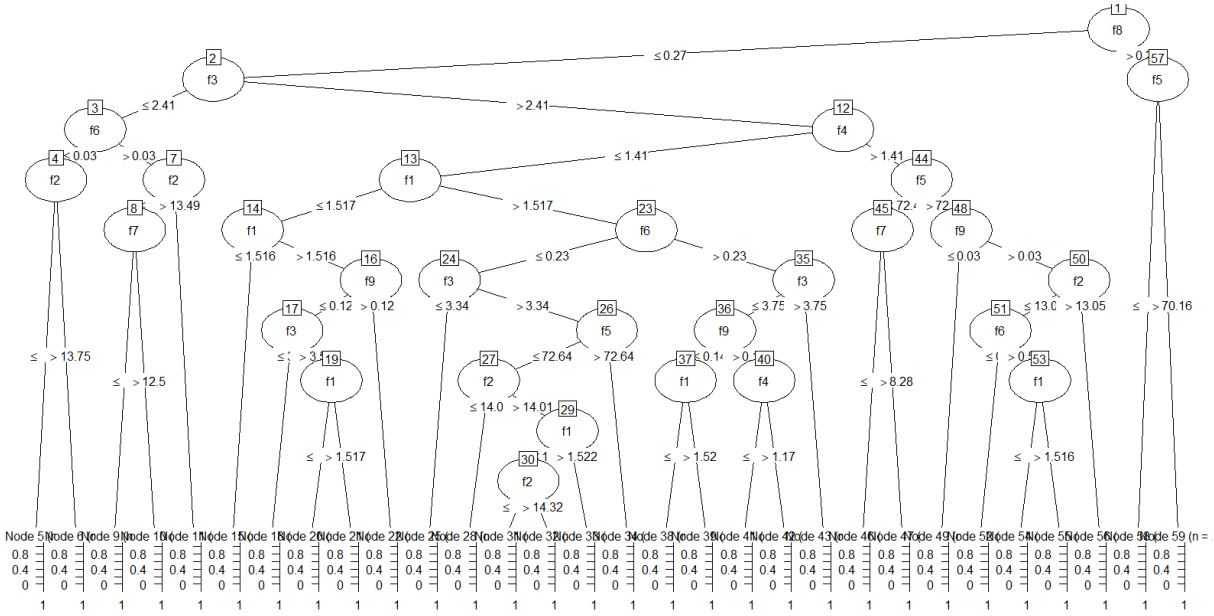
CF - confidence factor - parametr przyjmujący wartość z zakresu ¡0, 1¿. Wraz ze zmniejszeniem wartości, zwiększy się ilość przycinanych węzłów. Pomaga zapobiegać zbyt dużemu rozrostowi drzewa, oraz zjawisku przeuczenia.



Rysunek 7: Drzewo klasyfikacji instancji glass przy ustawieniu parametru CF = 0

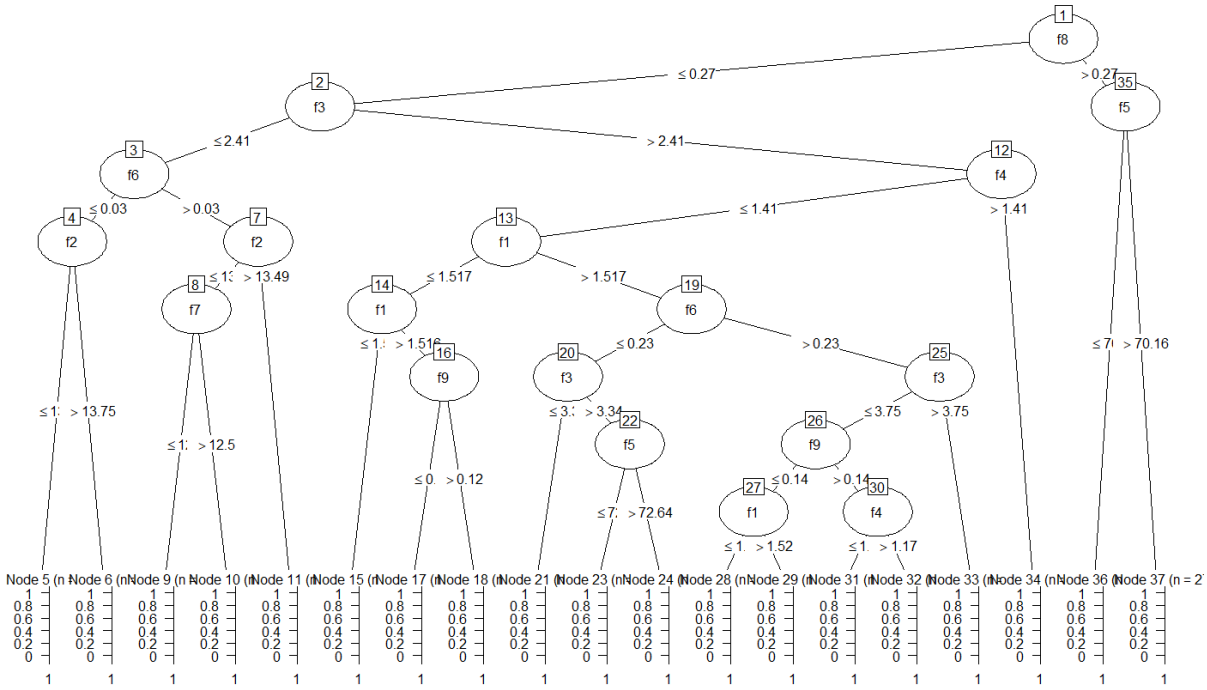


Rysunek 8: Drzewo klasyfikacji instancji glass przy ustawieniu parametru CF = 0.1

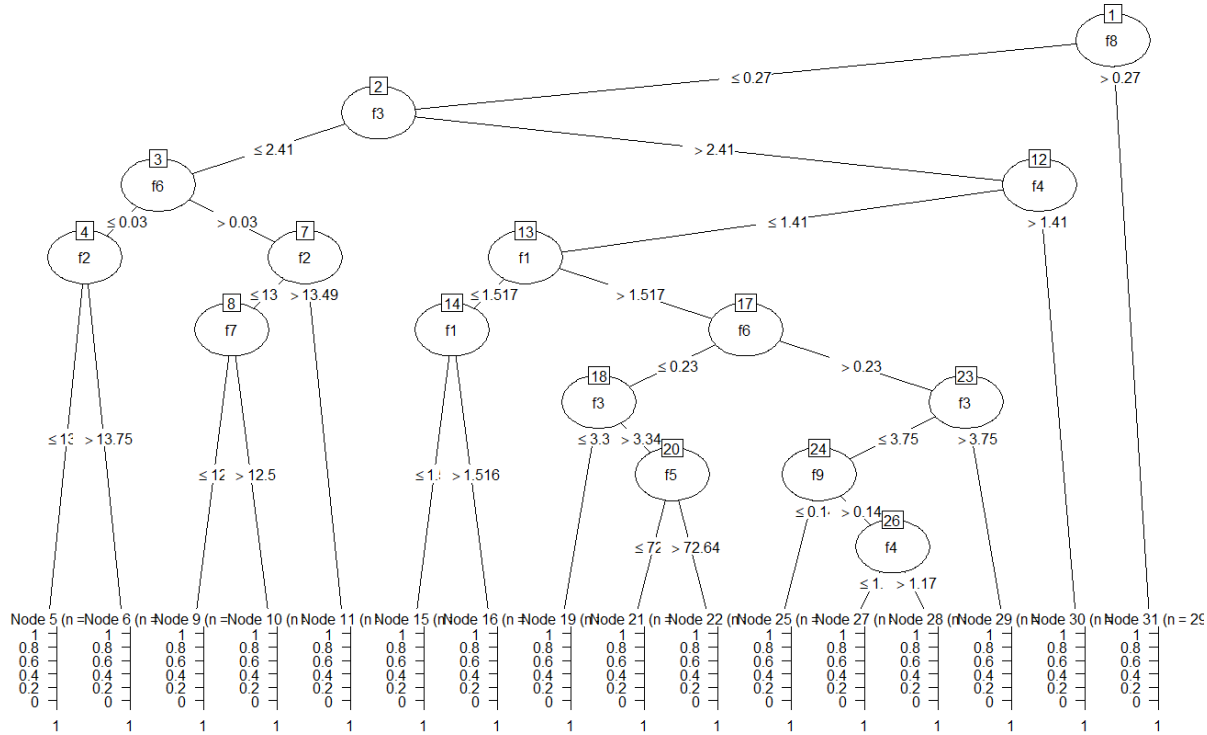


Rysunek 9: Drzewo klasyfikacji instancji glass przy ustawieniu parametru CF = 1

noGlobalPruning - Przełącznik informujący czy powinien zostać wykonany końcowy krok odpowiedzialny za dodatkowe przycinanie drzewa w celu zmniejszenia jego rozmiaru.



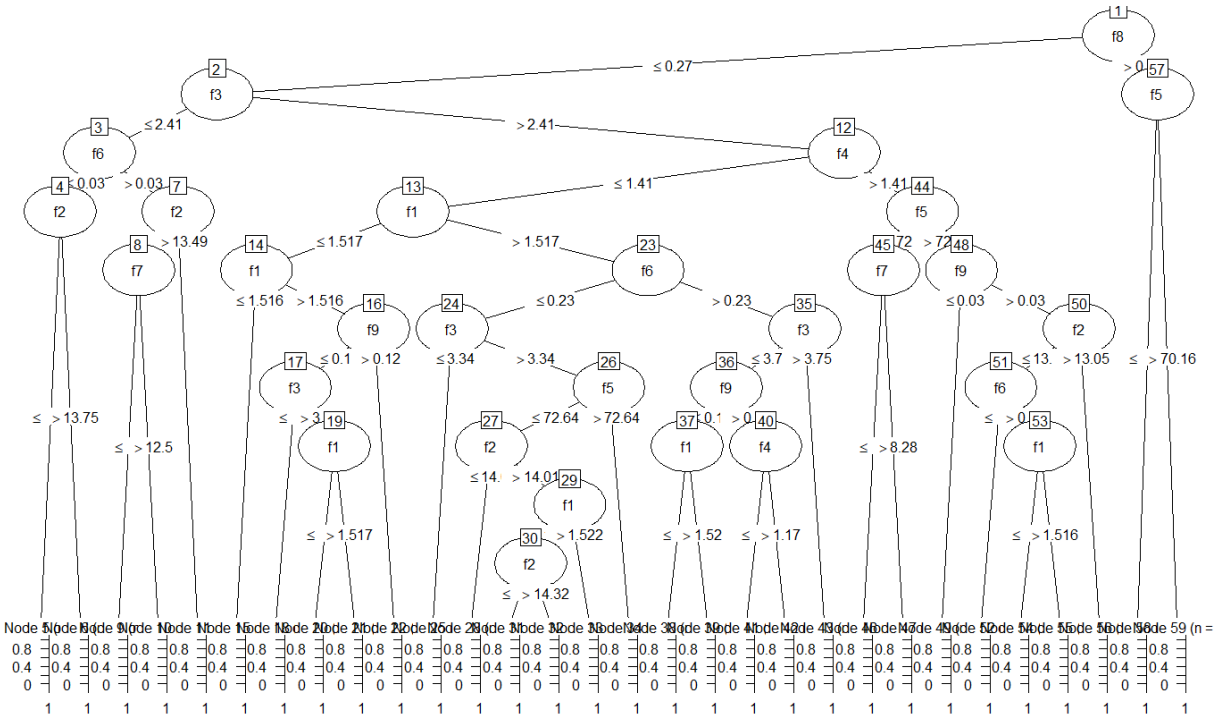
Rysunek 10: Drzewo klasyfikacji instancji glass przy ustawieniu parametru noGlobalPruning = TRUE



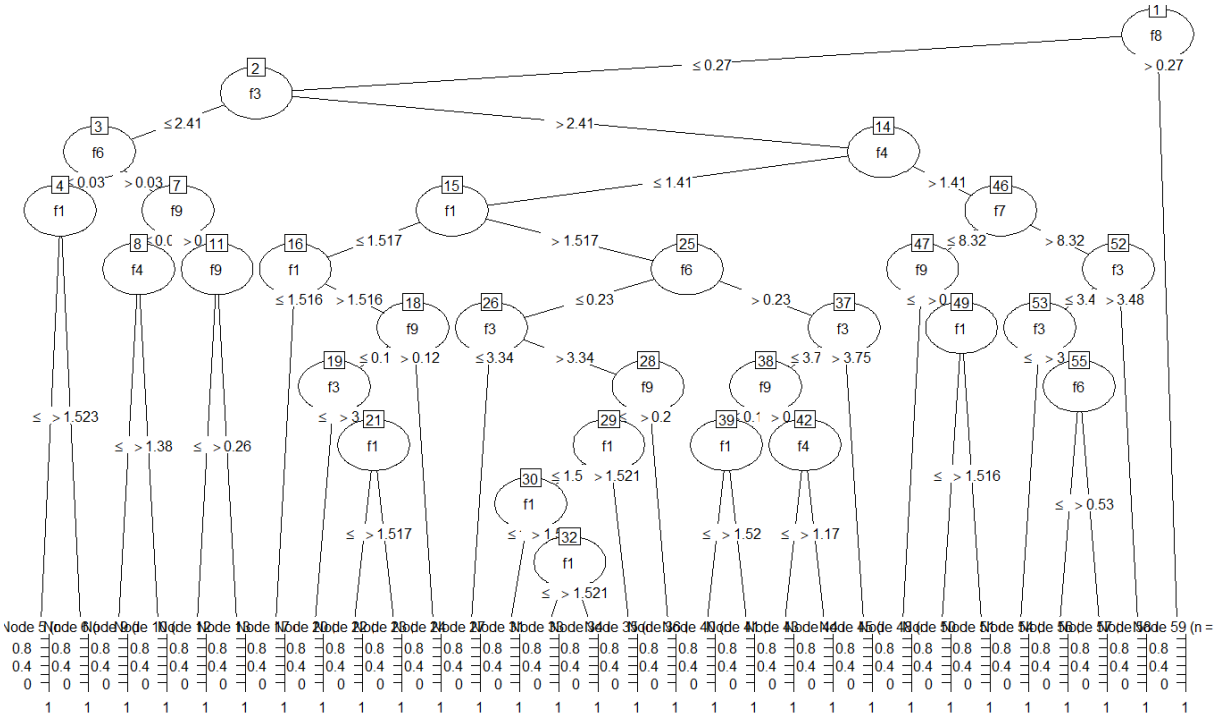
Rysunek 11: Drzewo klasyfikacji instancji glass przy ustawieniu parametru noGlobalPruning = FALSE

winnowing - Czy powinien zostać zaaplikowany krok przesiewania. Polega on na wstępnej selekcji cech, które mają być wykorzystane do późniejszego modelowania drzewa. Dane zostają rozdzielone na dwie części i dopasowany zostaje inicjacyjny model. Każdy pre-

dyktor (przesłanka) jest kolejno eliminowana i sprawdzany jest wpływ takiej operacji na drzewo. Predyktory są oznaczane w zależności od tego czy wpływają one na zwiększenie ilości generowanych błędów.



Rysunek 12: Drzewo klasyfikacji instancji glass przy ustawieniu parametru winnow = FALSE



Rysunek 13: Drzewo klasyfikacji instancji glass przy ustawieniu parametru winnow = FALSE