Klasyfikator oparty o drzewo decyzyjne

Łukasz Odwrot 218283

13.03.2018

Spis treści

1 Wstęp **2**

2 Badane zbiory **2**

3 Porównanie metod kroswalidacji **3**

4 Badane parametry **4**

5 Wpływ parametrów na zbiór wine **10**

6 Wpływ parametrów na zbiór glass **15**

7 Wpływ parametrów na zbiór diabetes **20**

8 Optymalne drzewa decyzyjne dla zbiorów danych **25**

9 Porównanie wyników z metodą "Naive Bayes" **27**

10 Wnioski **28**

1 Wstęp

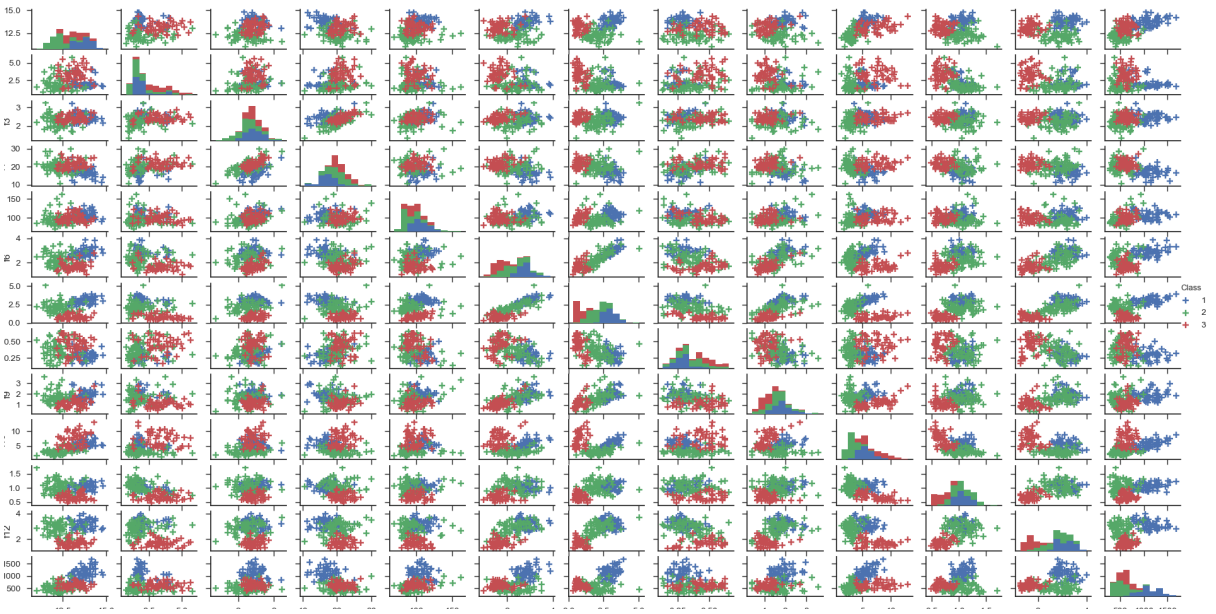
Drzewo decyzyjne, to struktura składająca się z:

- Węzłów zawierających pojedyncze atrybuty
- Liści przechowujących informacje o dopasowanej klasie
- Krawędzi łączących węzły i liście

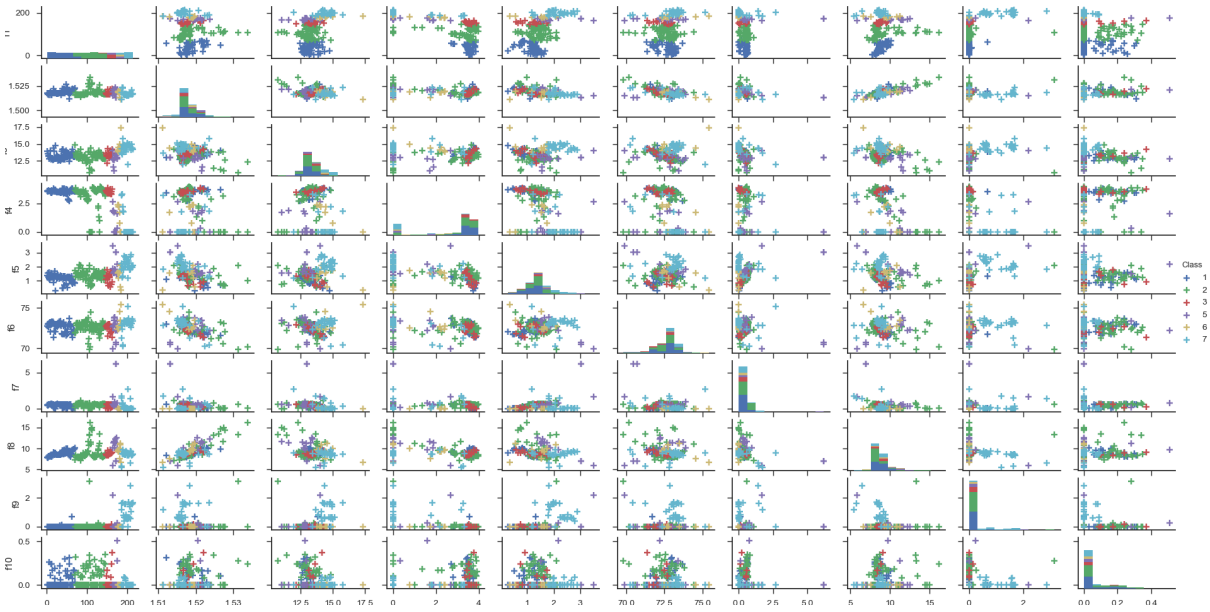
Poczynając od węzła budującego korzeń, bazując na wartościach atrybutów wybieramy kolejne węzły leżące niżej. Dojście do liścia oznacza przyporządkowanie.
Algorytm C4.5 jest rozszerzeniem prostego algorytmu ID3, pozbywając się jednocześnie wielu z jego wad. W czasie budowy drzewa decyzyjnego mogą wystąpić atrybuty o nieznanej wartości. Przyrost informacji jest wtedy obliczany jedynie dla atrybutów ze zdefiniowanymi wartościami. W czasie klasyfikacji również mogą wystąpić atrybuty o nieznanej wartości, a klasyfikacja odbywa się poprzez wyliczenie prawdopodobieństwa. Atrybuty mogą mieć wartości ciągłe. Występuje również mechanizm przycinania drzewa, zapobiegający zjawisku overfitingu.

2 Badane zbiory

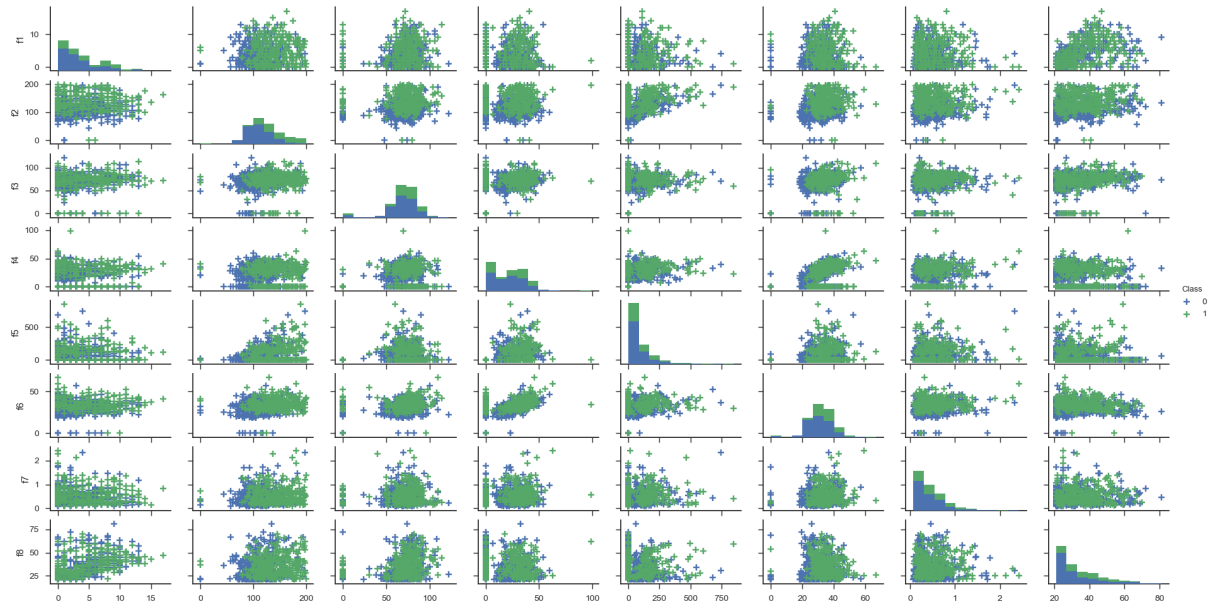
Rozkłady cech dla poszczególnych klas przedstawiono na poniższych rysunkach.



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)

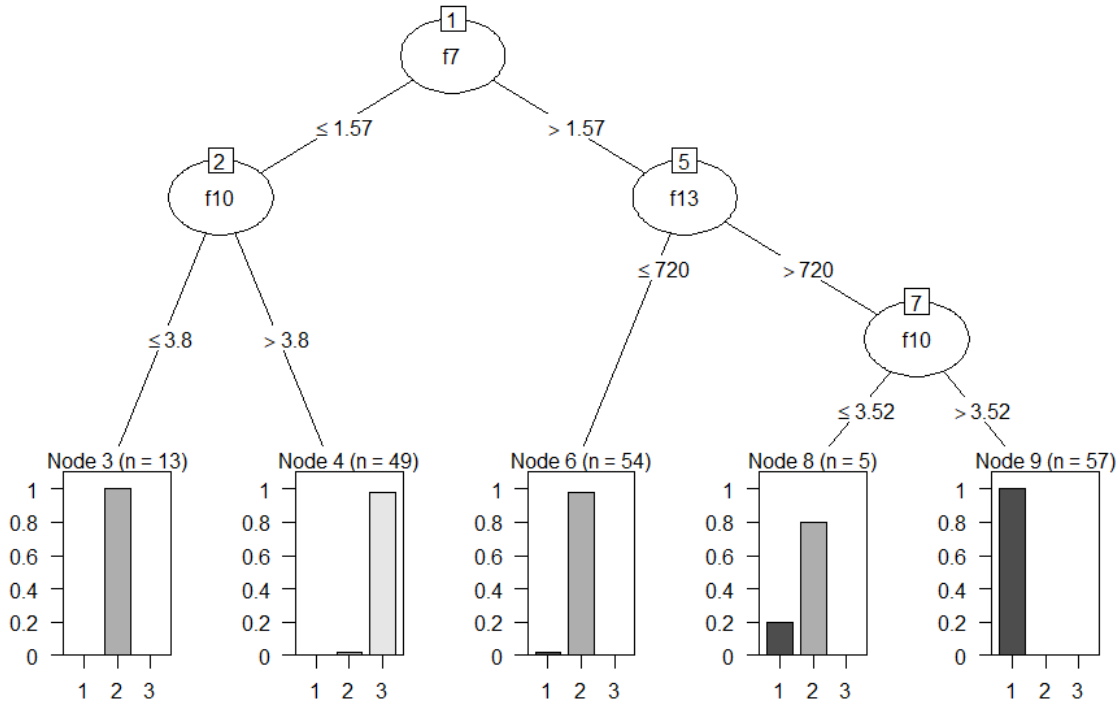
Do porównania wyników użyty zostanie parametr fscore.

Folds	Stratified	Normall no randomize	Normall randomize
WINE			
2	0.8702254	0.2159056	0.8991312
3	0.8813016	0.179399	0.9150061
4	0.9326211	0.6095417	0.9105371
5	0.9272085	0.7563068	0.9087973
6	0.9437285	0.8759431	0.9097962
7	0.9438199	0.8265988	0.9437564
8	0.9380819	0.904744	0.938011
9	0.9380761	0.9104553	0.9323077
10	0.9154446	0.8706026	0.8985881
GLASS			
2	0.6392635	0.1014686	0.7088985
3	0.6457611	0.03486039	0.701681
4	0.675482	0.2136425	0.7084925
5	0.6774104	0.2178138	0.659168
6	0.7275745	0.2381391	0.6300857
7	0.6519988	0.2785823	0.6405154
8	0.6906709	0.3186959	0.6425939
9	0.7124522	0.2739046	0.7037276
10	0.6788223	0.515942	0.655208
DIABETES			
2	0.7654584	0.8065448	0.8108632
3	0.7949952	0.7971602	0.8008256
4	0.8078049	0.8	0.8079602
5	0.7935549	0.7980198	0.781655
6	0.8059406	0.8036999	0.8027478
7	0.8023483	0.8086359	0.7988281
8	0.7967644	0.7976072	0.8104449
9	0.8251208	0.810757	0.8189739
10	0.7935549	0.8083416	0.8050193

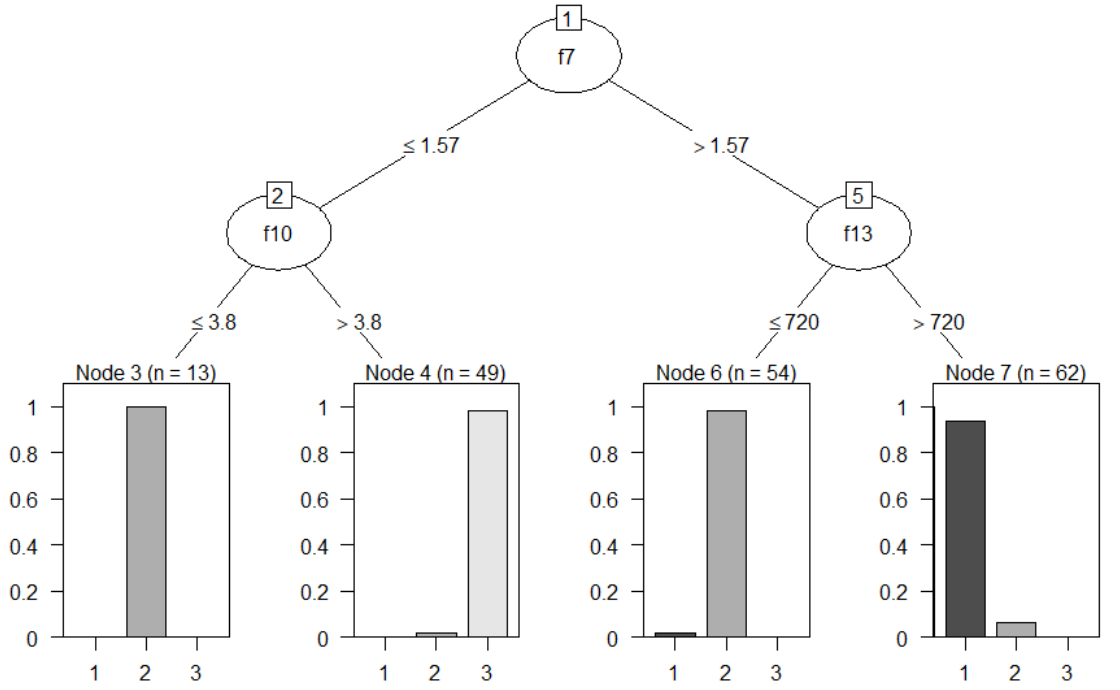
4 Badane parametry

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")

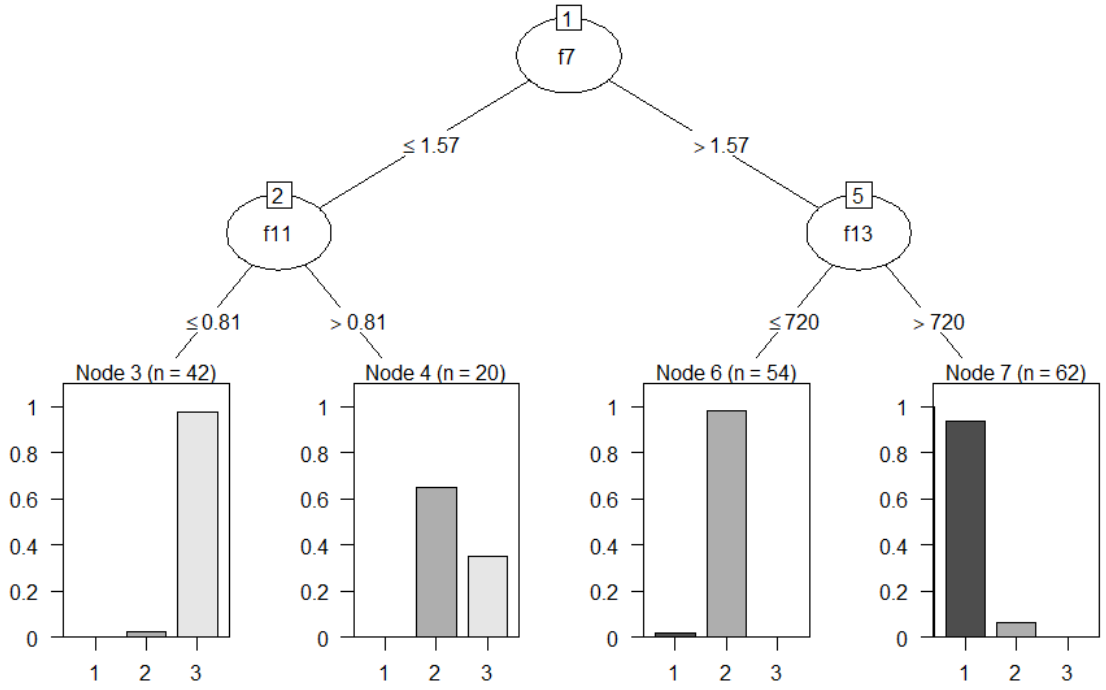
MinCases - określa minimalną ilość próbek w danym liściu.



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

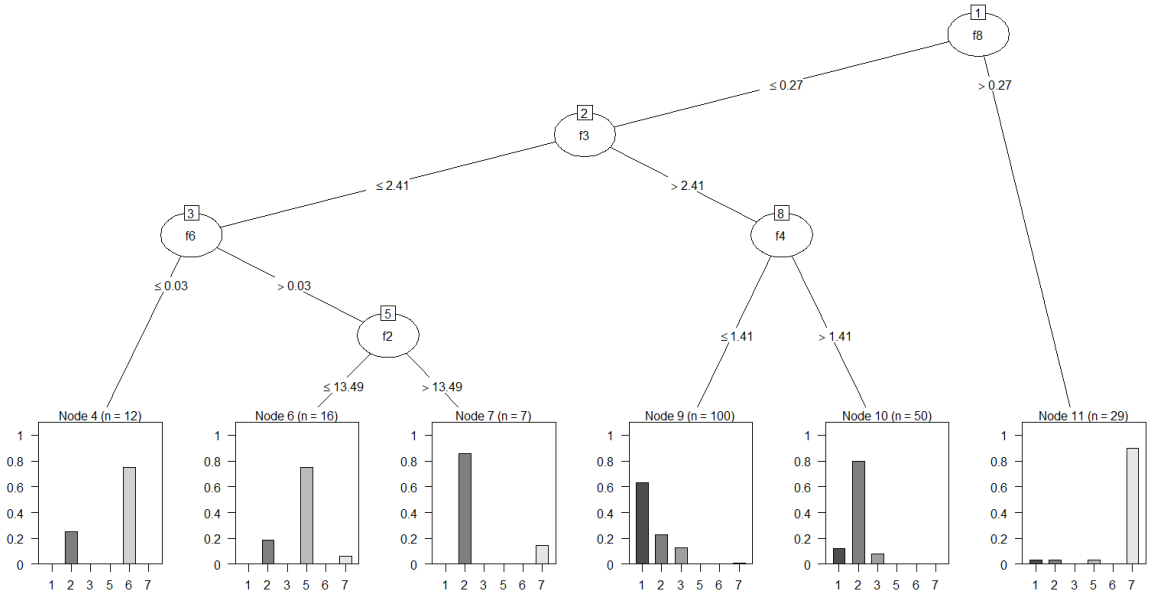


Rysunek 5: Drzewo klasyfikacji instancji wine przy ustawieniu parametru minCases = 10

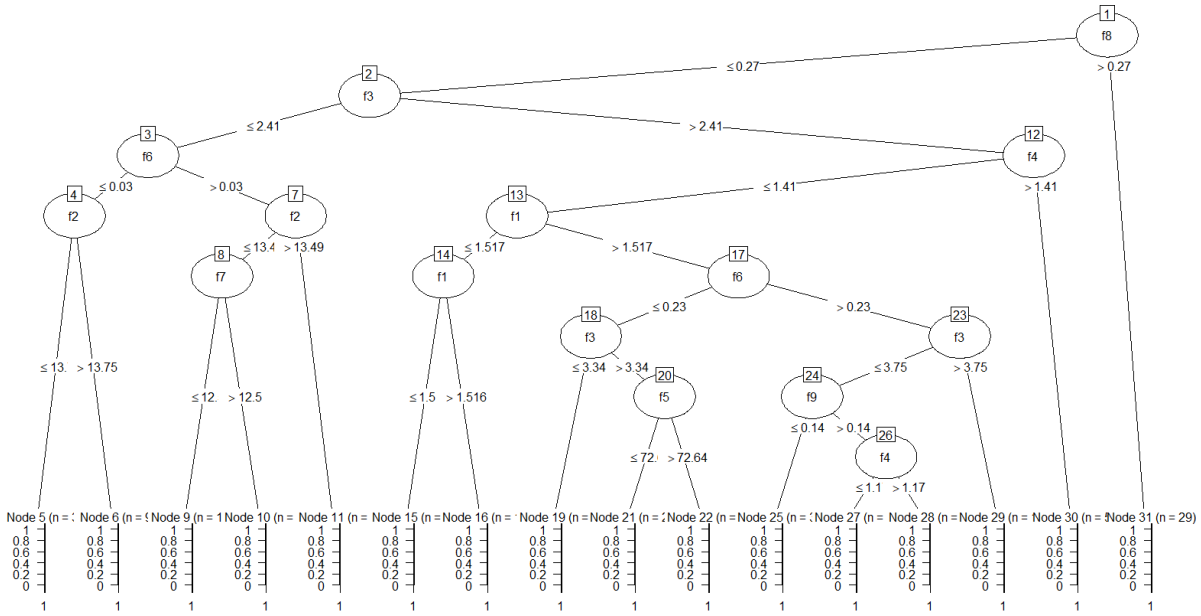


Rysunek 6: Drzewo klasyfikacji instancji wine przy ustawieniu parametru minCases = 20

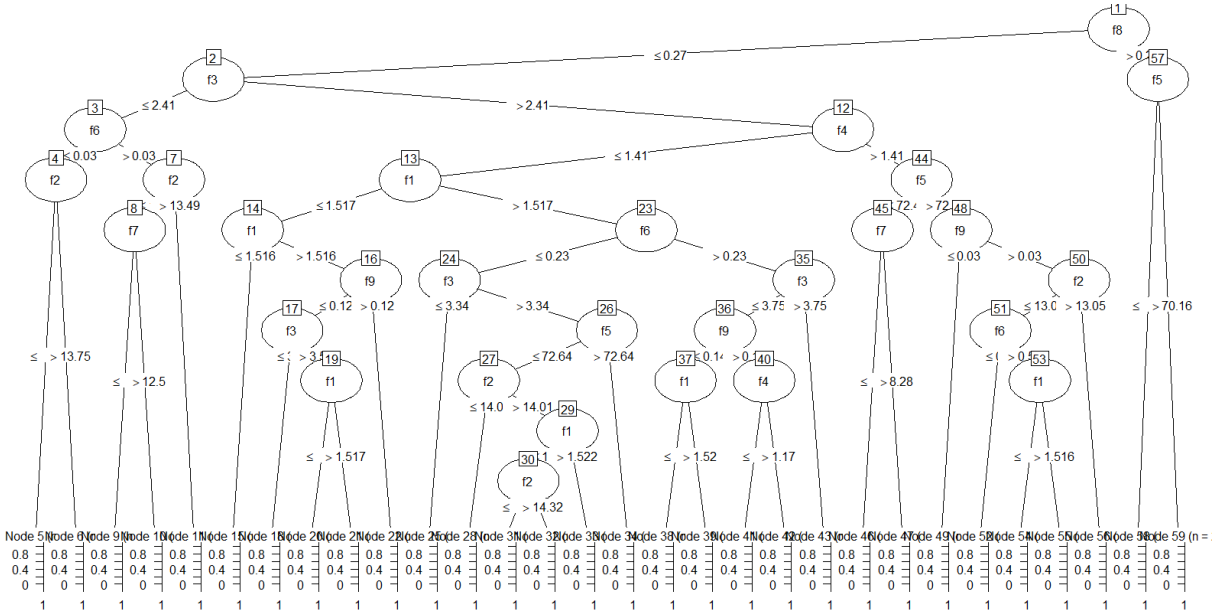
CF - confidence factor - parametr przyjmujący wartość z zakresu od 0 do 1. Wraz ze zmniejszeniem wartości, zwiększy się ilość przycinanych węzłów. Pomaga zapobiegać zbyt dużemu rozrostowi drzewa, oraz zjawisku przeuczenia.



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

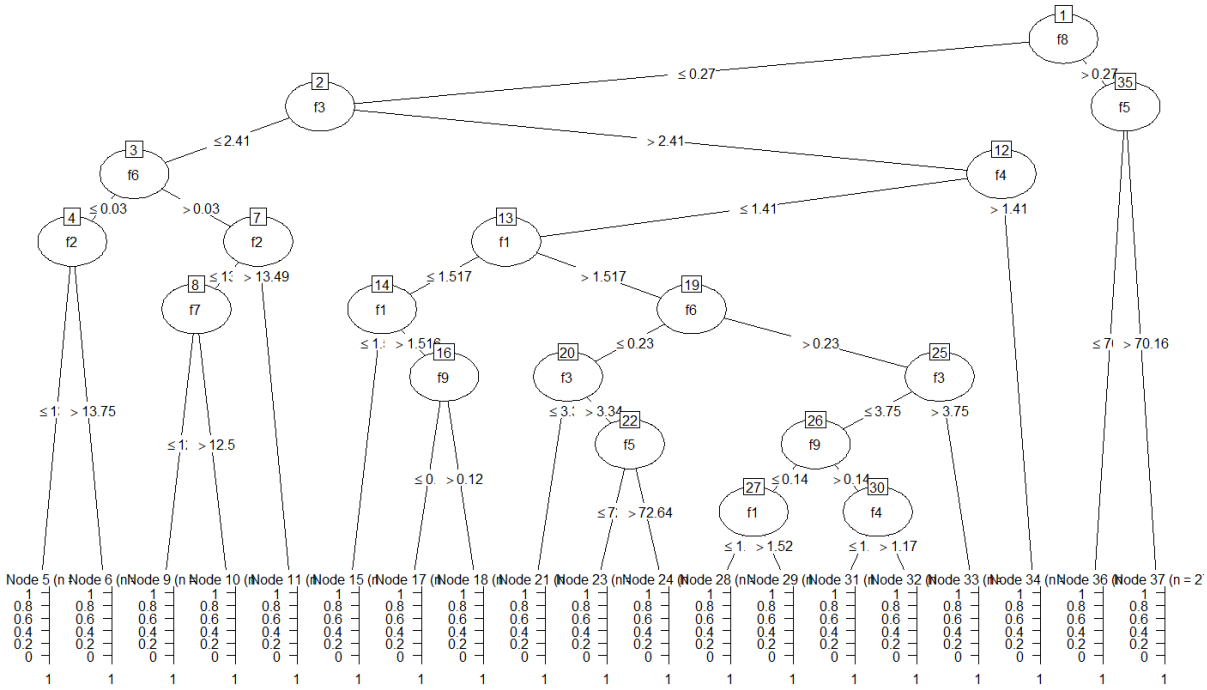


Rysunek 8: Drzewo klasyfikacji instancji glass przy ustawieniu parametru CF = 0.1

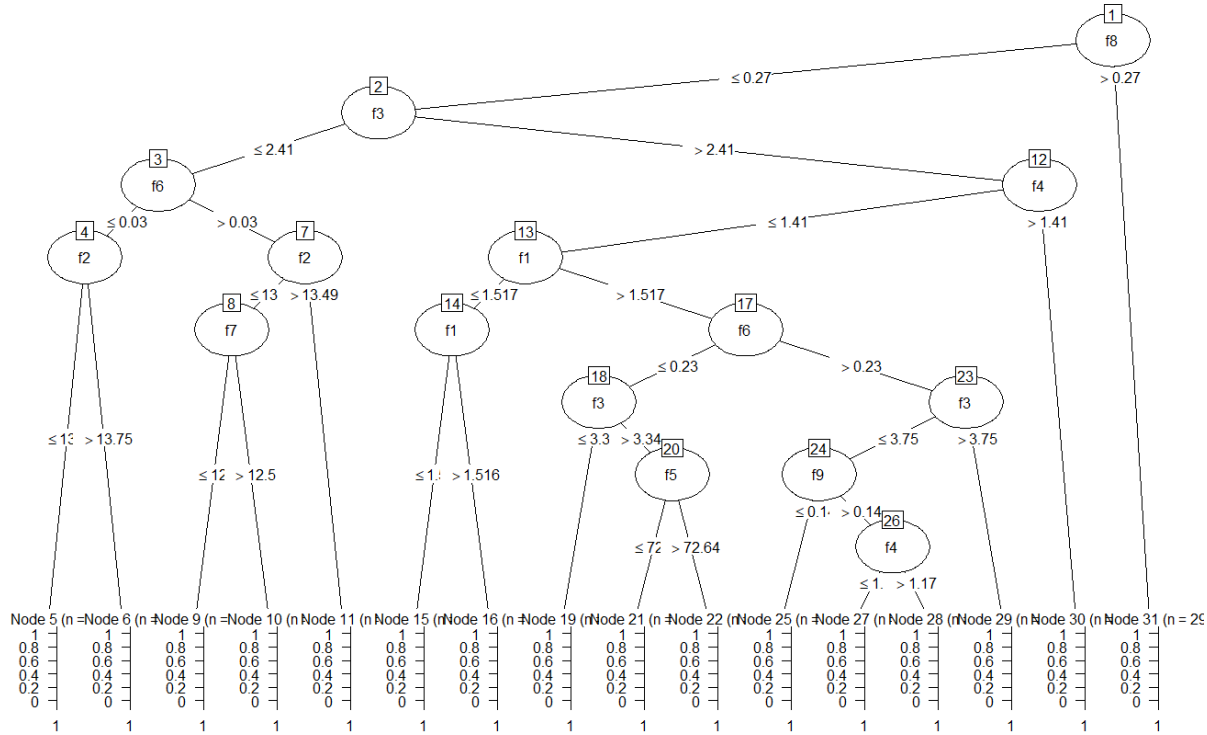


Rysunek 9: Drzewo klasyfikacji instancji glass przy ustawieniu parametru CF = 1

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



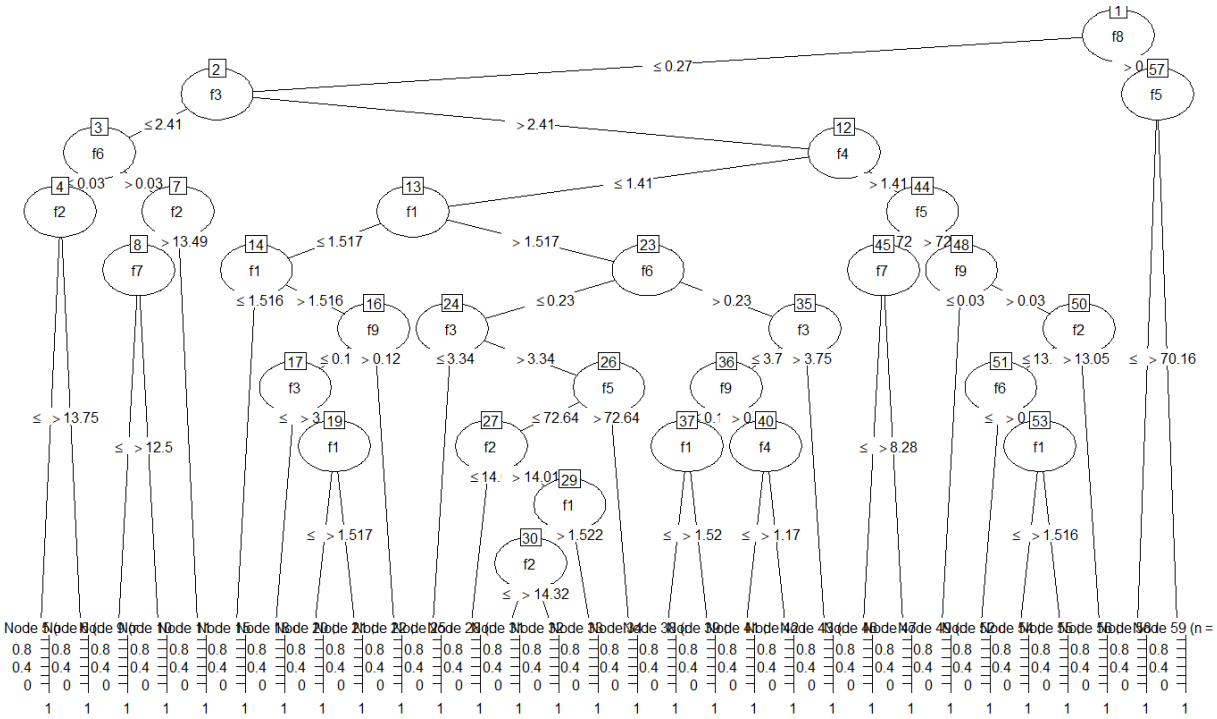
Rysunek 10: Drzewo klasyfikacji instancji glass przy ustawieniu parametru noGlobalPruning = TRUE



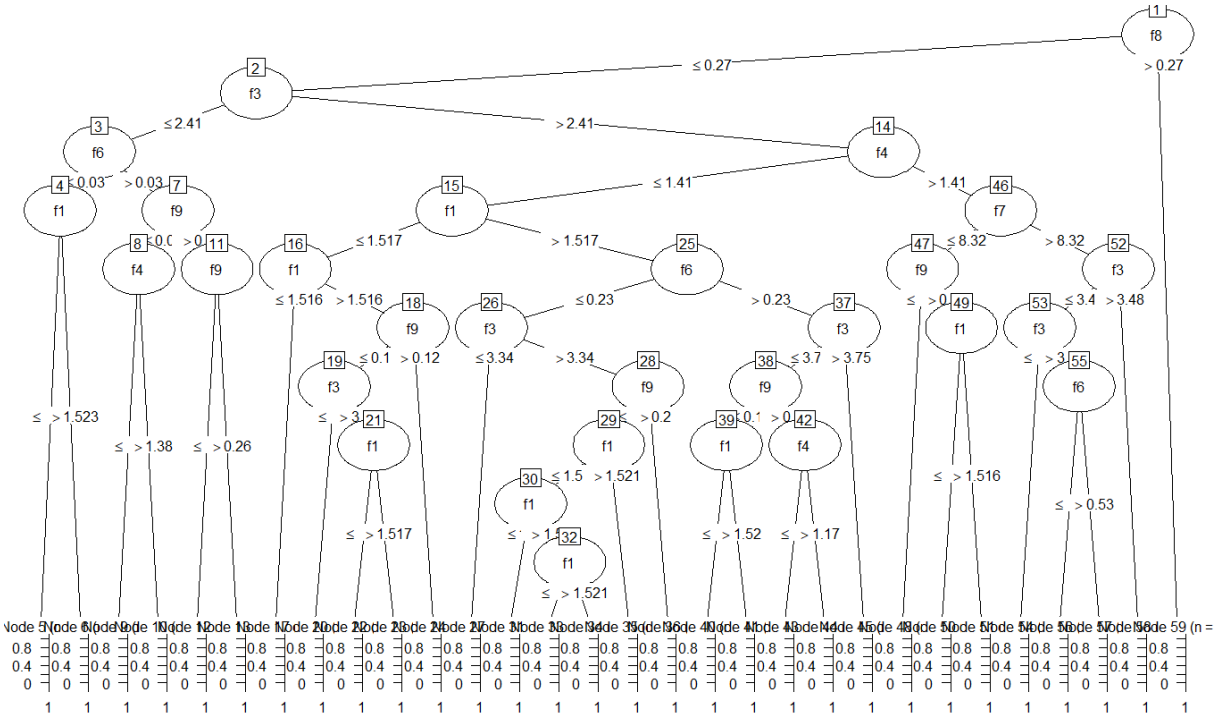
Rysunek 11: Drzewo klasyfikacji instancji glass przy ustawieniu parametru noGlobalPruning = FALSE

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

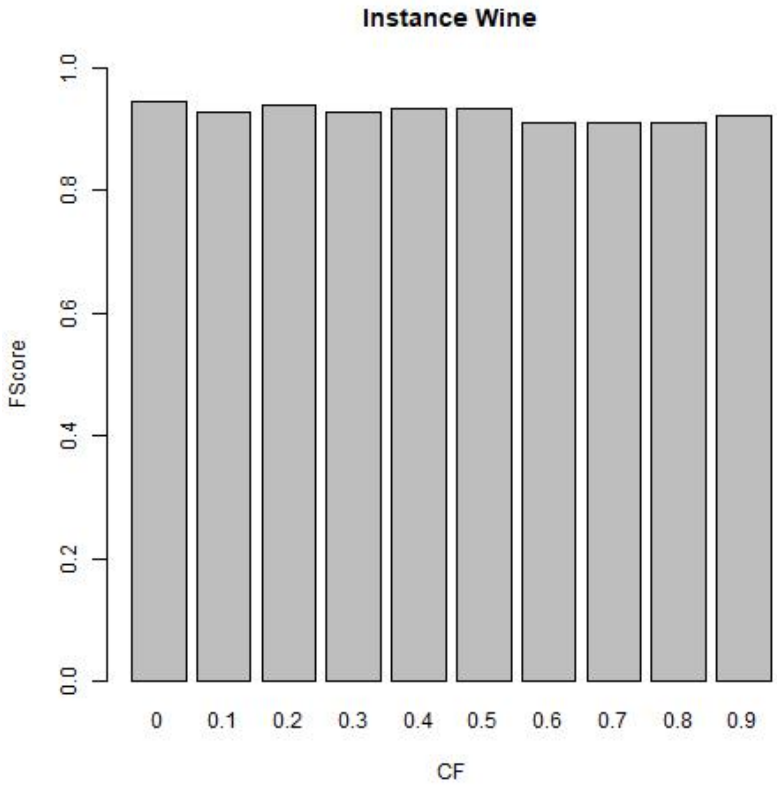
5 Wpływ parametrów na zbiór wine

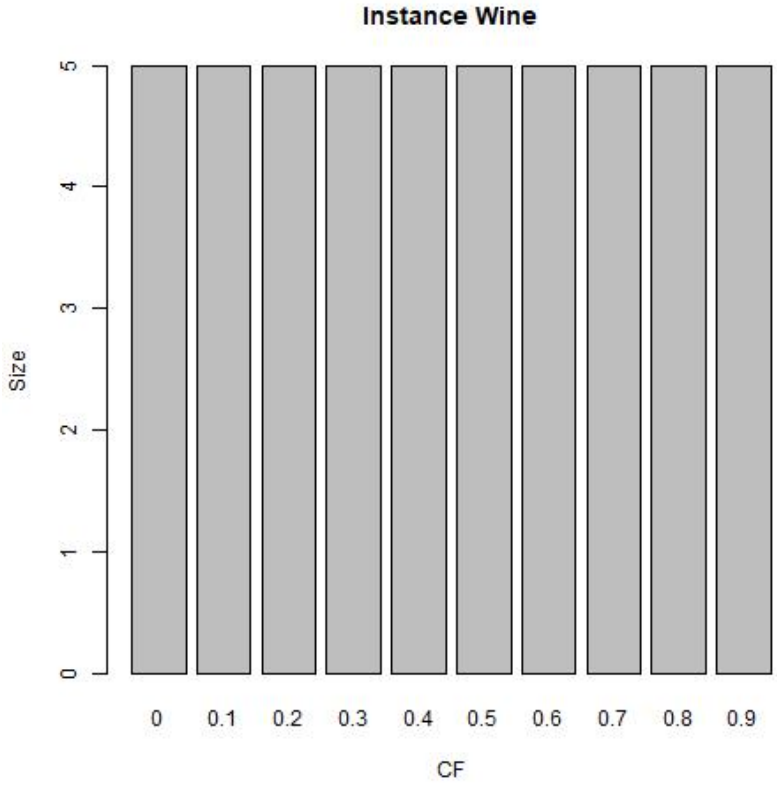
vecCF <dbl>	vecCFFscore <dbl>
0.0	0.9438202
0.1	0.9270759
0.2	0.9380969
0.3	0.9268195
0.4	0.9320197
0.5	0.9324953
0.6	0.9100742
0.7	0.9098478
0.8	0.9096366
0.9	0.9212156

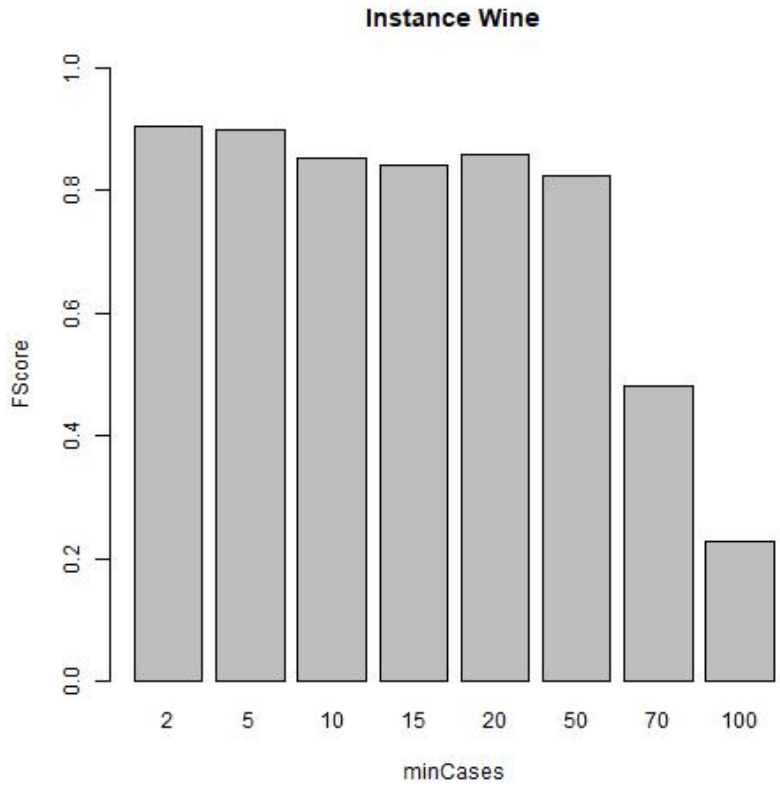
vecMinCases <dbl>	vecMinCasesFscore <dbl>
2	0.9049640
5	0.8989442
10	0.8530562
15	0.8408337
20	0.8575344
50	0.8238553
70	0.4807056
100	0.2274717

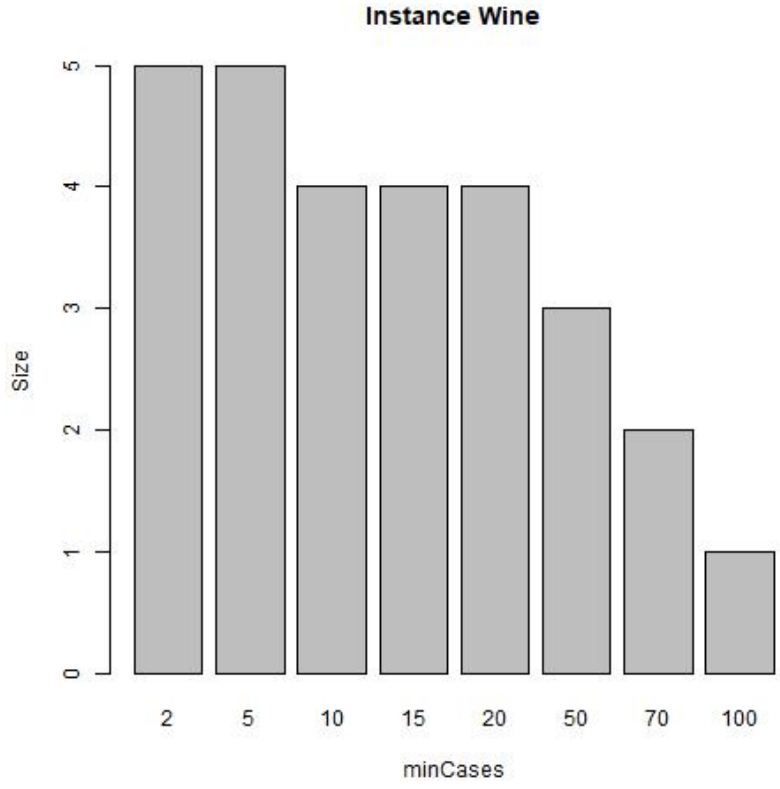
vecNoGlobalPruning <lgl>	vecNoGlobalPruningFscore <dbl>
TRUE	0.9324470
FALSE	0.9324402

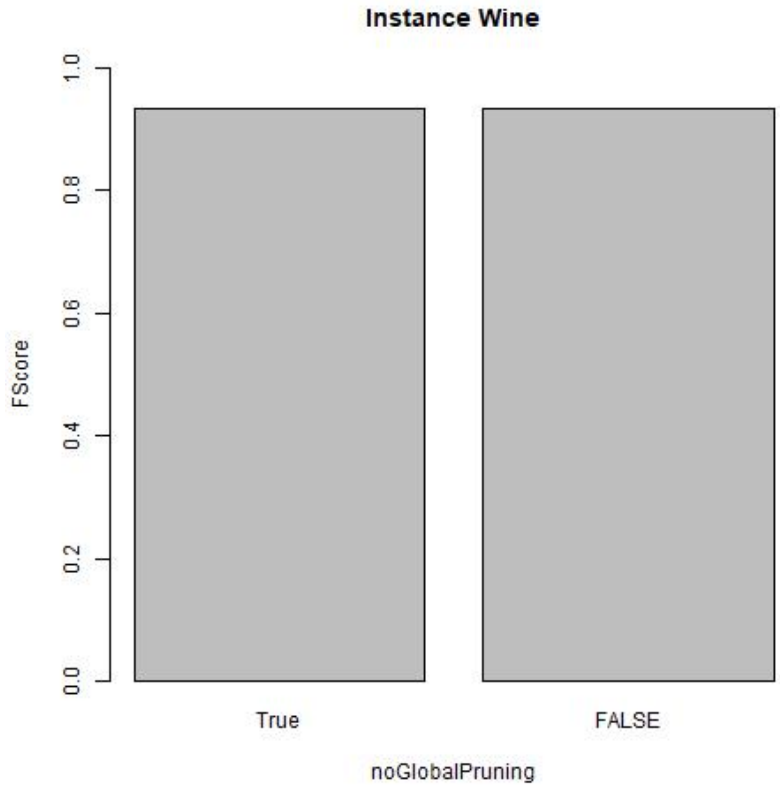
vecWinnow <lgl>	vecWinnowFscore <dbl>
TRUE	0.9098298
FALSE	0.9606070

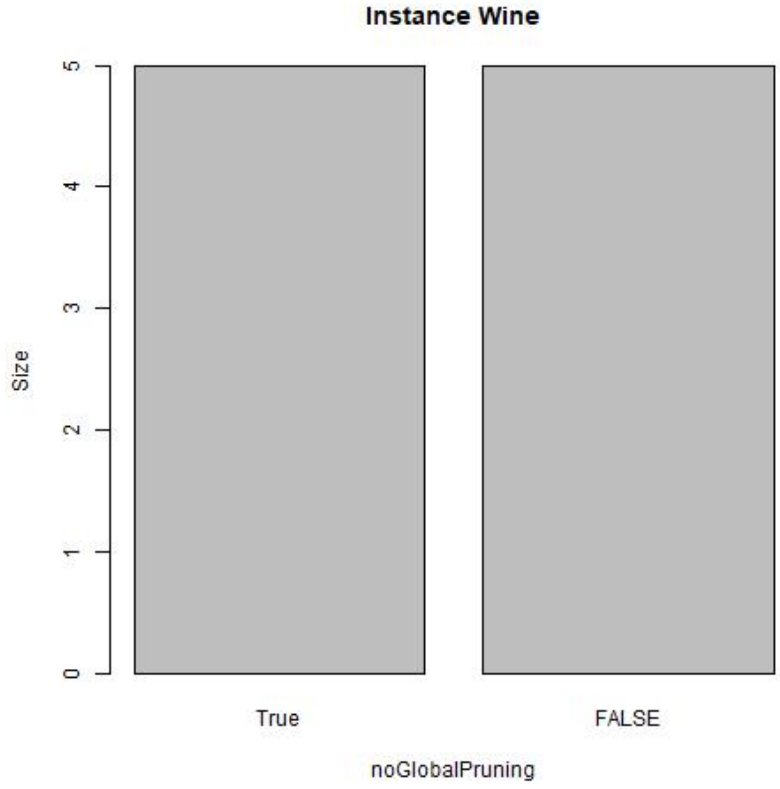

Instance Wine
FScore
0.0
0.2
0.4
0.6
0.8
1.0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
CF

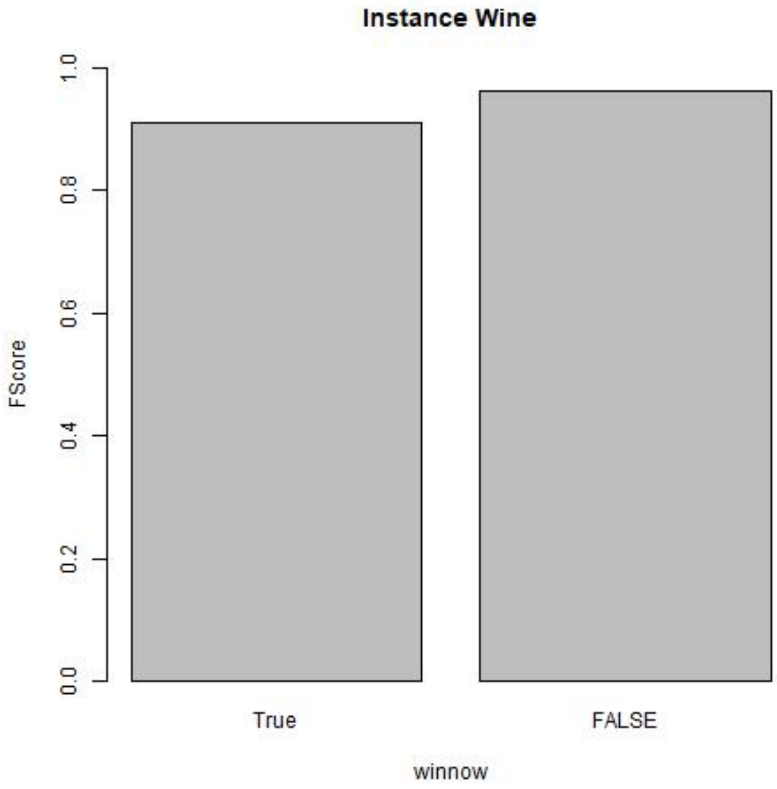

Instance Wine
Size
0
1
2
3
4
5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
CF

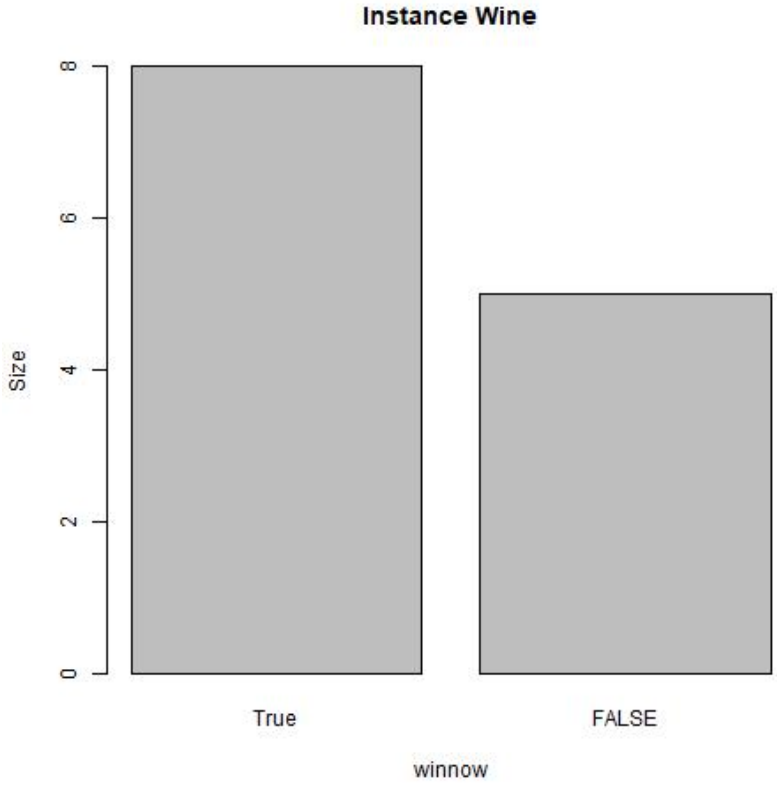

Instance Wine
FScore
0.0
0.2
0.4
0.6
0.8
1.0
2
5
10
15
20
50
70
100
minCases


Instance Wine
Size
0
1
2
3
4
5
2
5
10
15
20
50
70
100
minCases


Instance Wine
FScore
0.0
0.2
0.4
0.6
0.8
1.0
True
FALSE
noGlobalPruning


Instance Wine
Size
0
1
2
3
4
5
True
FALSE
noGlobalPruning


Instance Wine
FScore
0.0
0.2
0.4
0.6
0.8
1.0
True
FALSE
winnow


Instance Wine
Size
0
2
4
6
8
True
FALSE
winnow

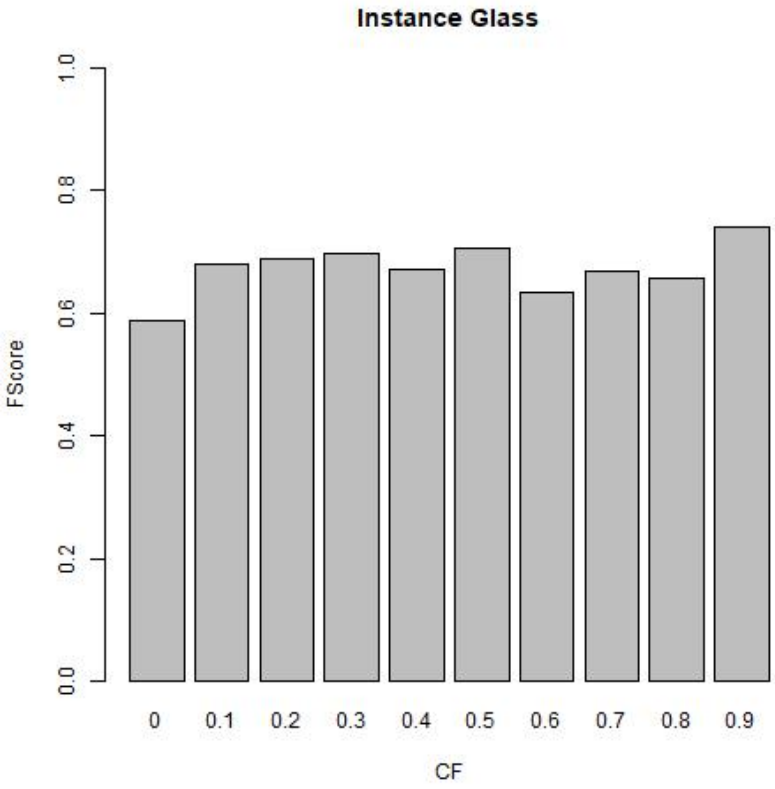
6 Wpływ parametrów na zbiór glass

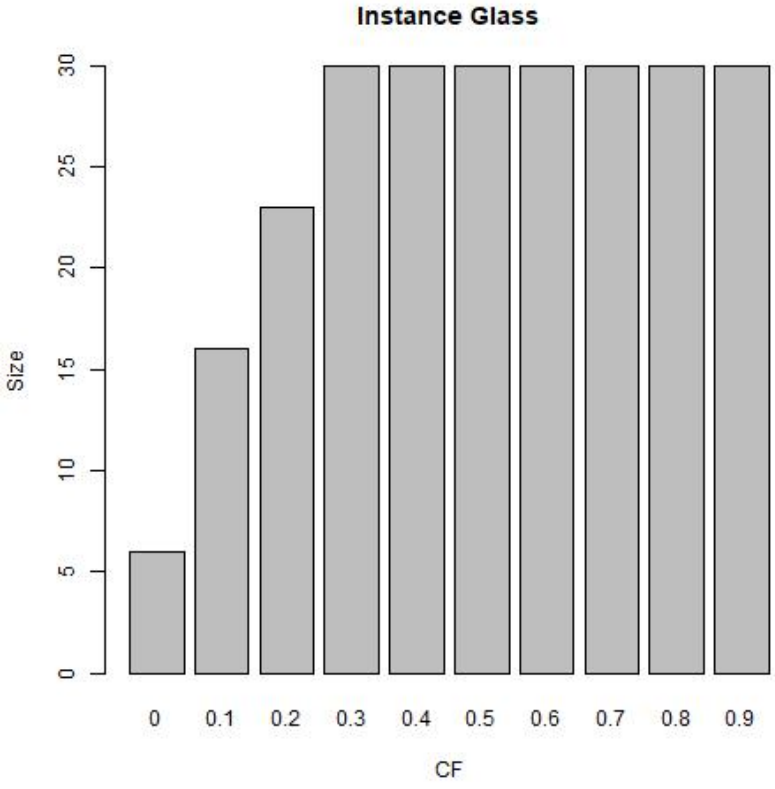
vecCF <dbl>	vecCFFscore <dbl>
0.0	0.5874974
0.1	0.6791418
0.2	0.6876832
0.3	0.6975695
0.4	0.6696331
0.5	0.7050780
0.6	0.6329307
0.7	0.6663450
0.8	0.6561494
0.9	0.7390340

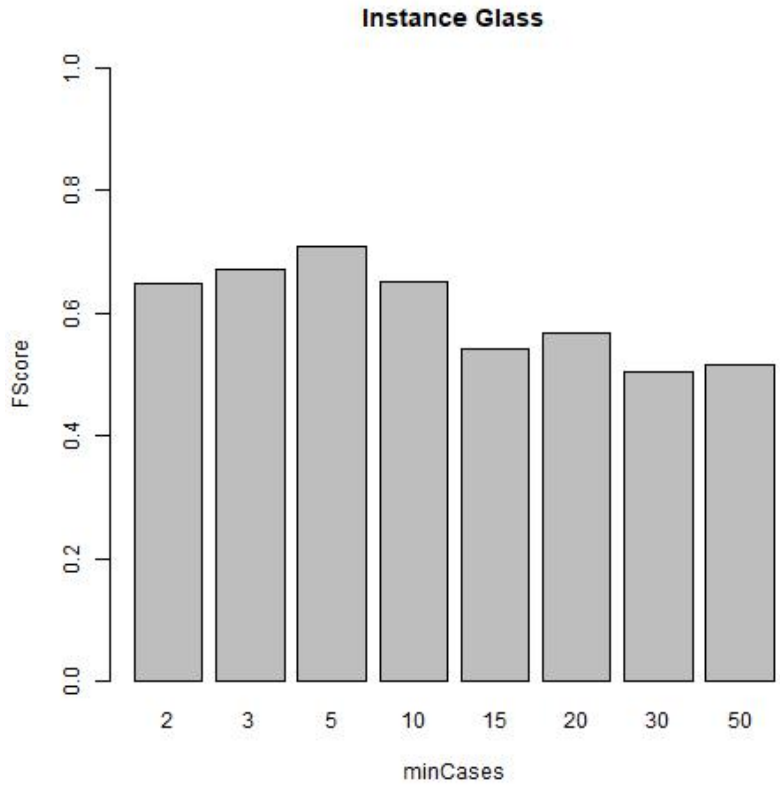
vecMinCases <dbl>	vecMinCasesFscore <dbl>
2	0.6485750
3	0.6716577
5	0.7079362
10	0.6512506
15	0.5423596
20	0.5678074
30	0.5036682
50	0.5165069

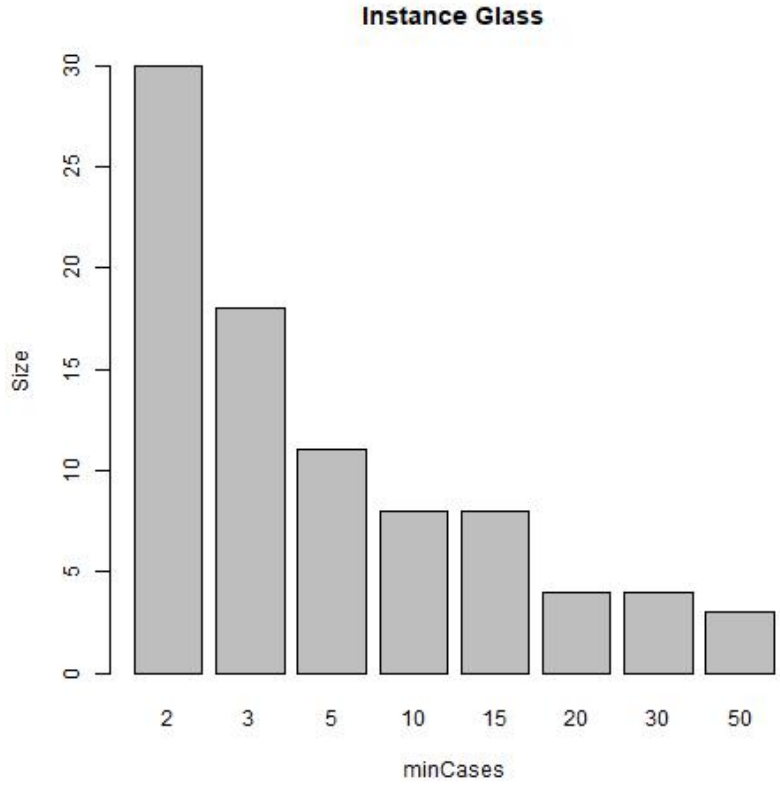
vecNoGlobalPruning <lgl>	vecNoGlobalPruningFscore <dbl>
TRUE	0.6995008
FALSE	0.6920270

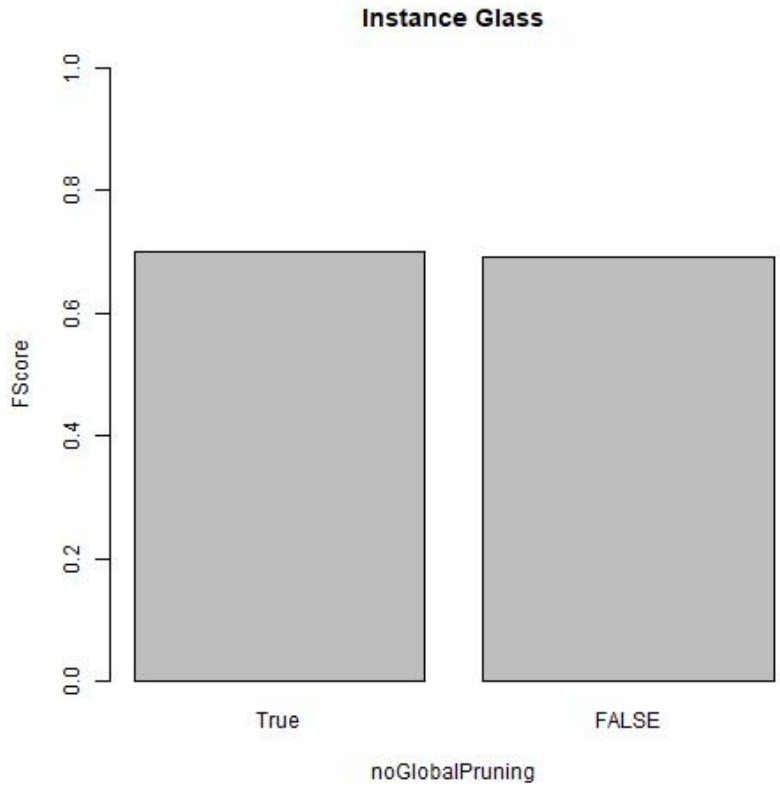
vecWinnow <lgl>	vecWinnowFscore <dbl>
TRUE	0.7047031
FALSE	0.6480832

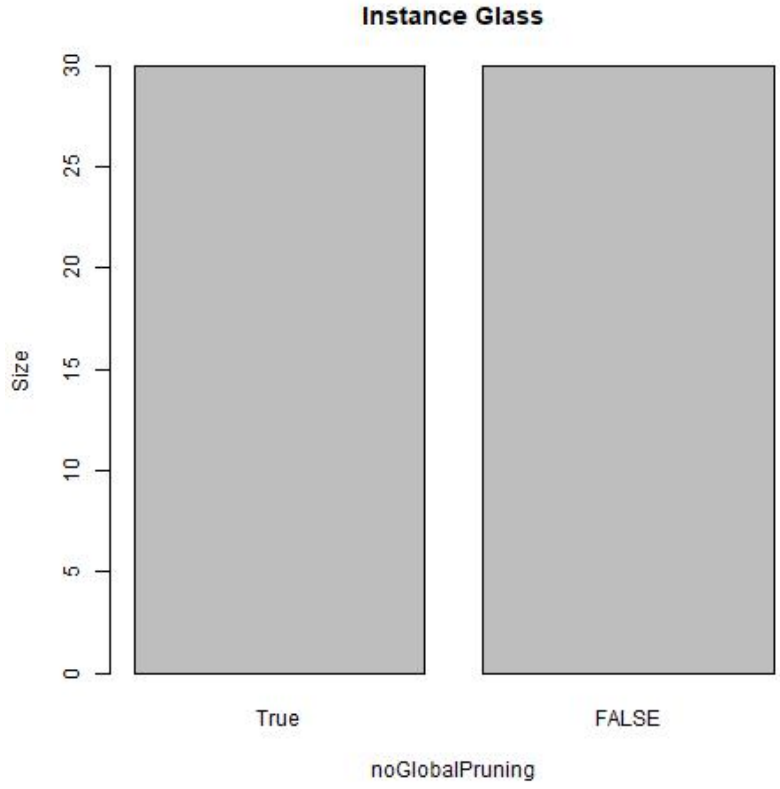

Instance Glass
FScore
0.0
0.2
0.4
0.6
0.8
1.0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
CF

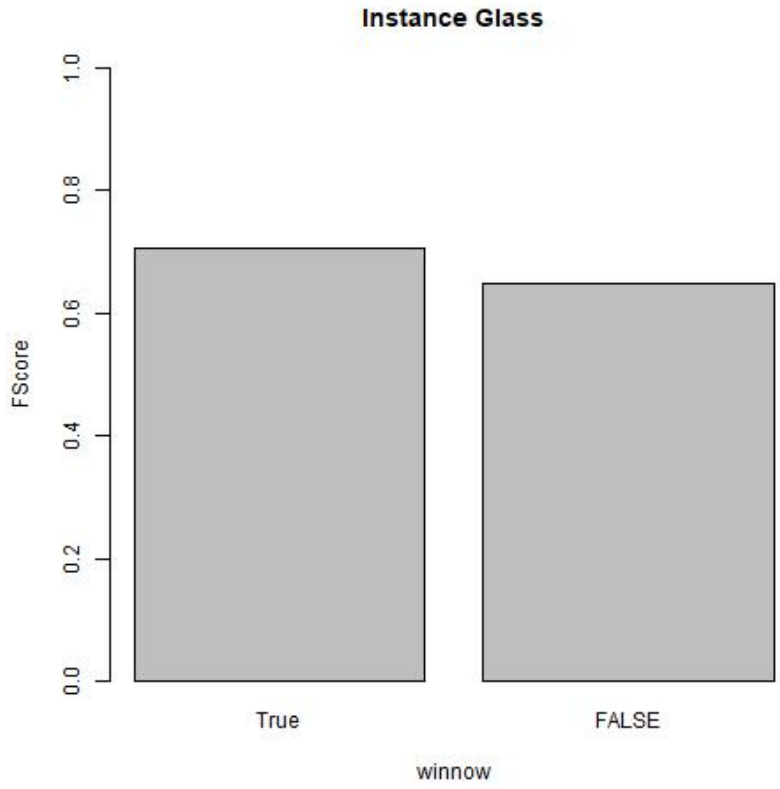

Instance Glass
Size
0
5
10
15
20
25
30
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
CF

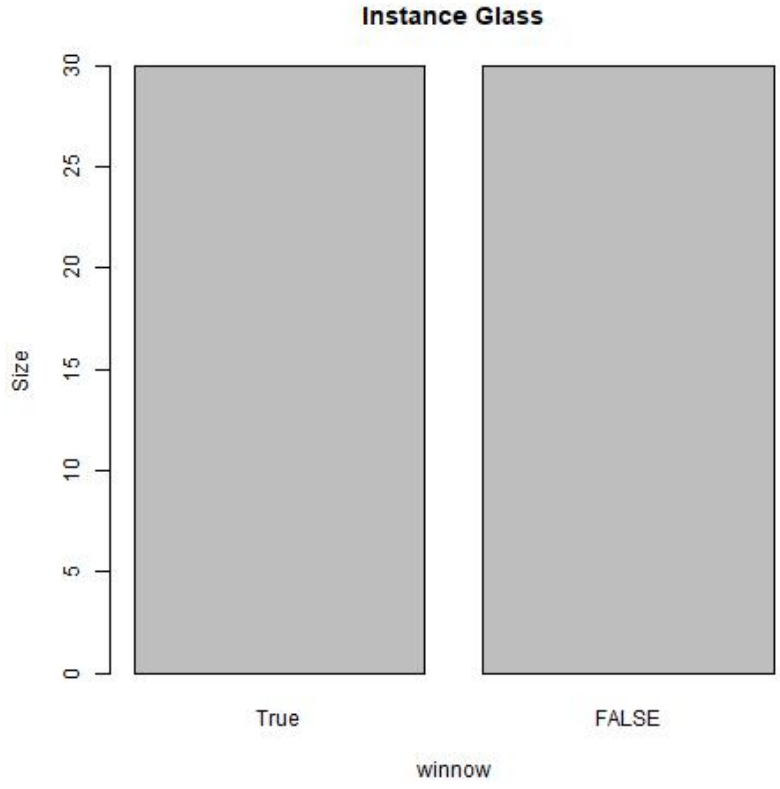

Instance Glass
FScore
0.0
0.2
0.4
0.6
0.8
1.0
2
3
5
10
15
20
30
50
minCases


Instance Glass
Size
0
5
10
15
20
25
30
2
3
5
10
15
20
30
50
minCases


Instance Glass
FScore
0.0
0.2
0.4
0.6
0.8
1.0
True
FALSE
noGlobalPruning


Instance Glass
Size
0
5
10
15
20
25
30
True
FALSE
noGlobalPruning


Instance Glass
FScore
0.0
0.2
0.4
0.6
0.8
1.0
True
FALSE
winnow


Instance Glass
Size
0
5
10
15
20
25
30
True
FALSE
winnow

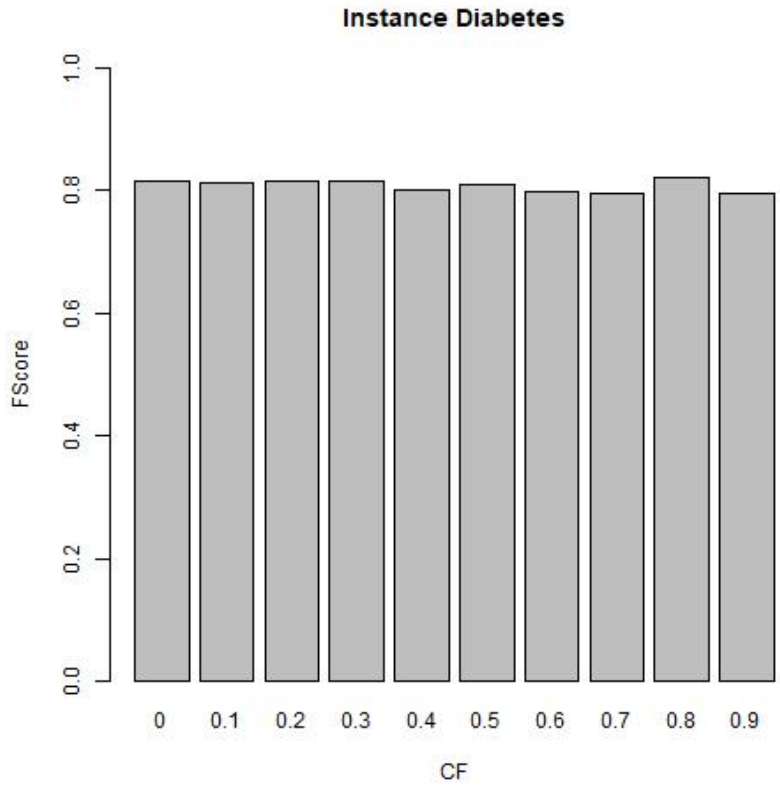
7 Wpływ parametrów na zbiór diabetes

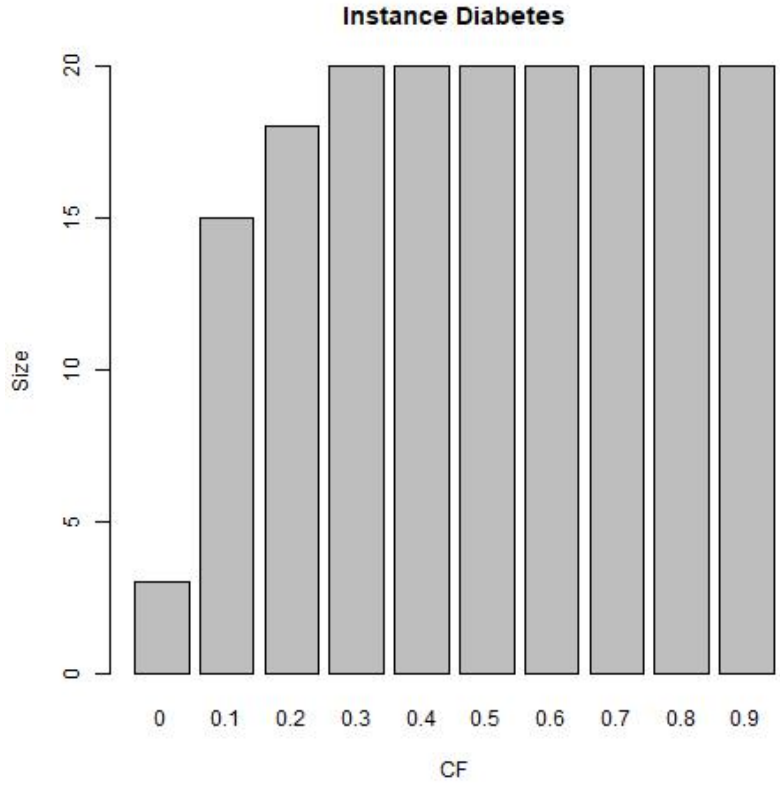
vecCF <dbl>	vecCFFscore <dbl>
0.0	0.8150943
0.1	0.8109696
0.2	0.8148855
0.3	0.8134255
0.4	0.8007737
0.5	0.8096677
0.6	0.7976190
0.7	0.7939698
0.8	0.8192534
0.9	0.7931727

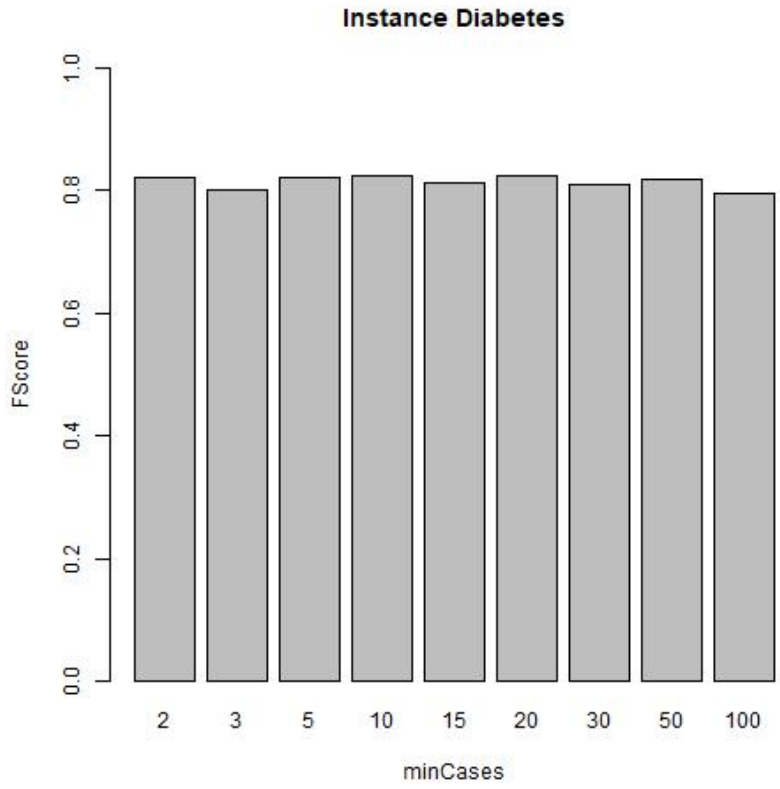
vecMinCases <dbl>	vecMinCasesFscore <dbl>
2	0.8200590
3	0.8000000
5	0.8196078
10	0.8231884
15	0.8123195
20	0.8235294
30	0.8085520
50	0.8174905
100	0.7933723

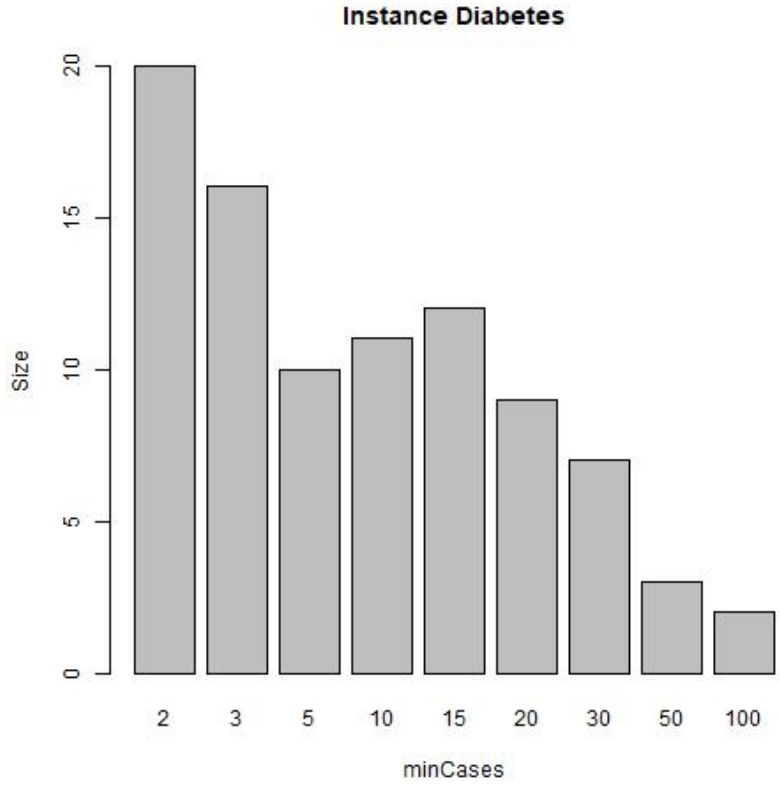
vecNoGlobalPruning <lgl>	vecNoGlobalPruningFscore <dbl>
TRUE	0.804000
FALSE	0.810757

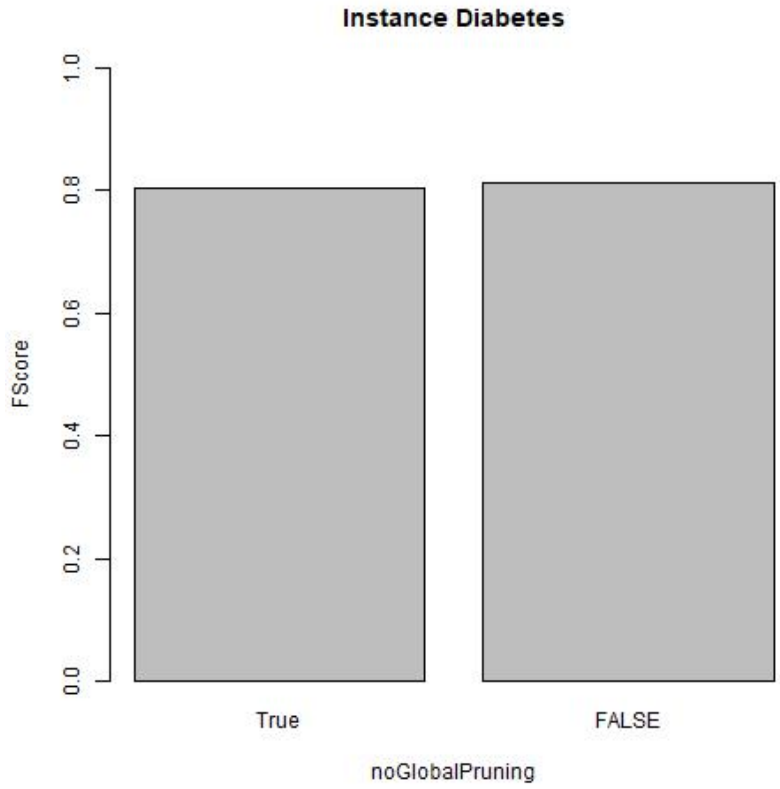
vecWinnow <lgl>	vecWinnowFscore <dbl>
TRUE	0.8038462
FALSE	0.7880000

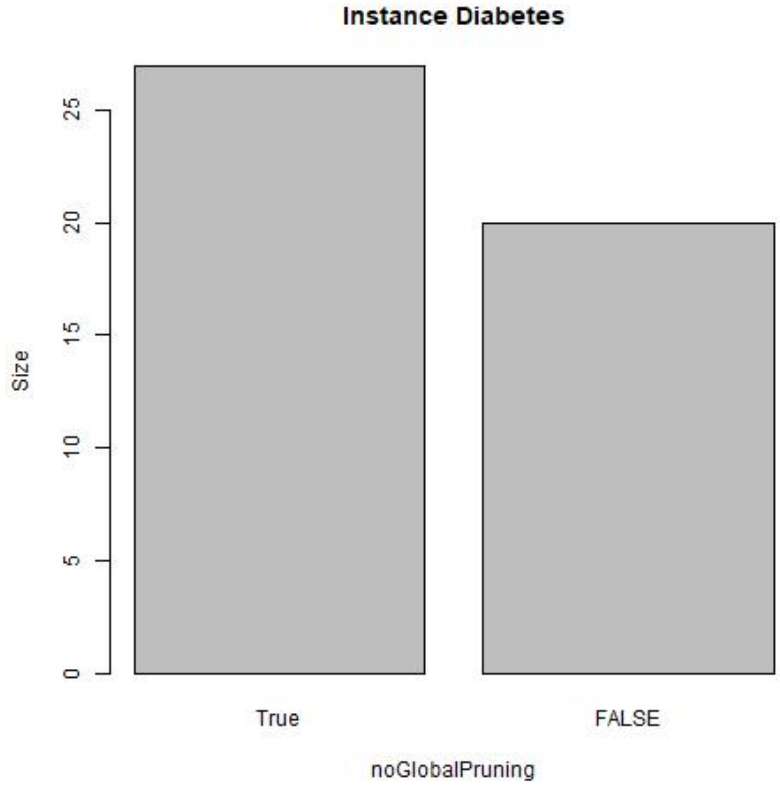

Instance Diabetes
FScore
0.0
0.2
0.4
0.6
0.8
1.0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
CF

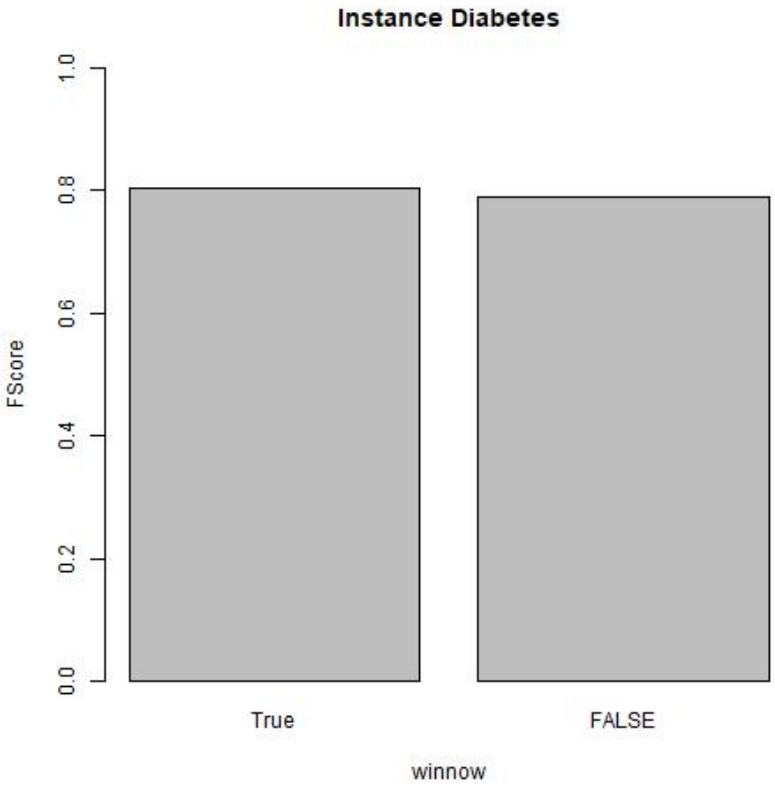

Instance Diabetes
Size
0
5
10
15
20
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
CF

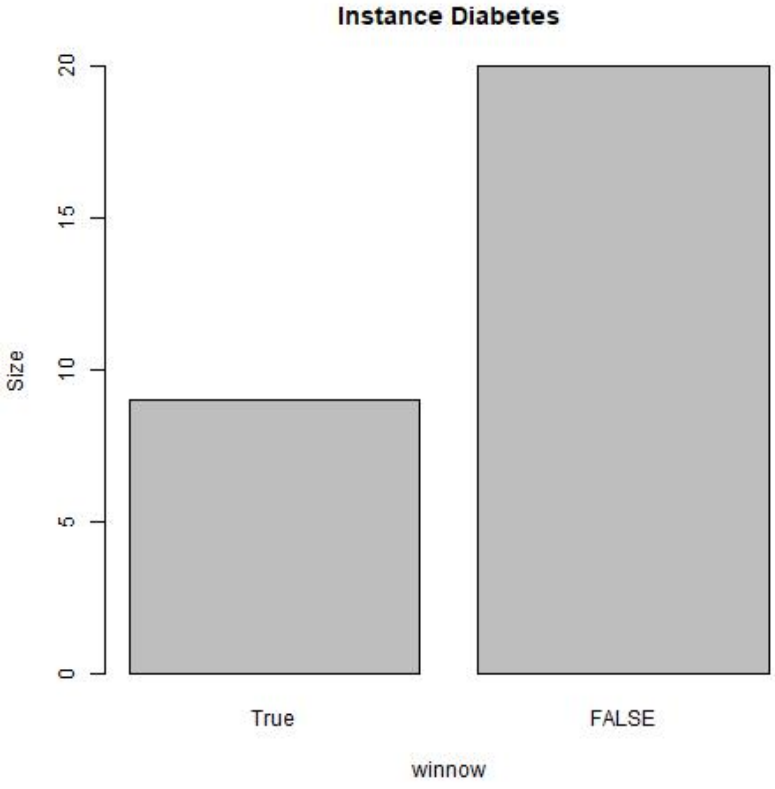

Instance Diabetes
FScore
0.0
0.2
0.4
0.6
0.8
1.0
2
3
5
10
15
20
30
50
100
minCases


Instance Diabetes
Size
0
5
10
15
20
2
3
5
10
15
20
30
50
100
minCases


Instance Diabetes
FScore
0.0
0.2
0.4
0.6
0.8
1.0
True
FALSE
noGlobalPruning


Instance Diabetes
Size
0
5
10
15
20
25
True
FALSE
noGlobalPruning

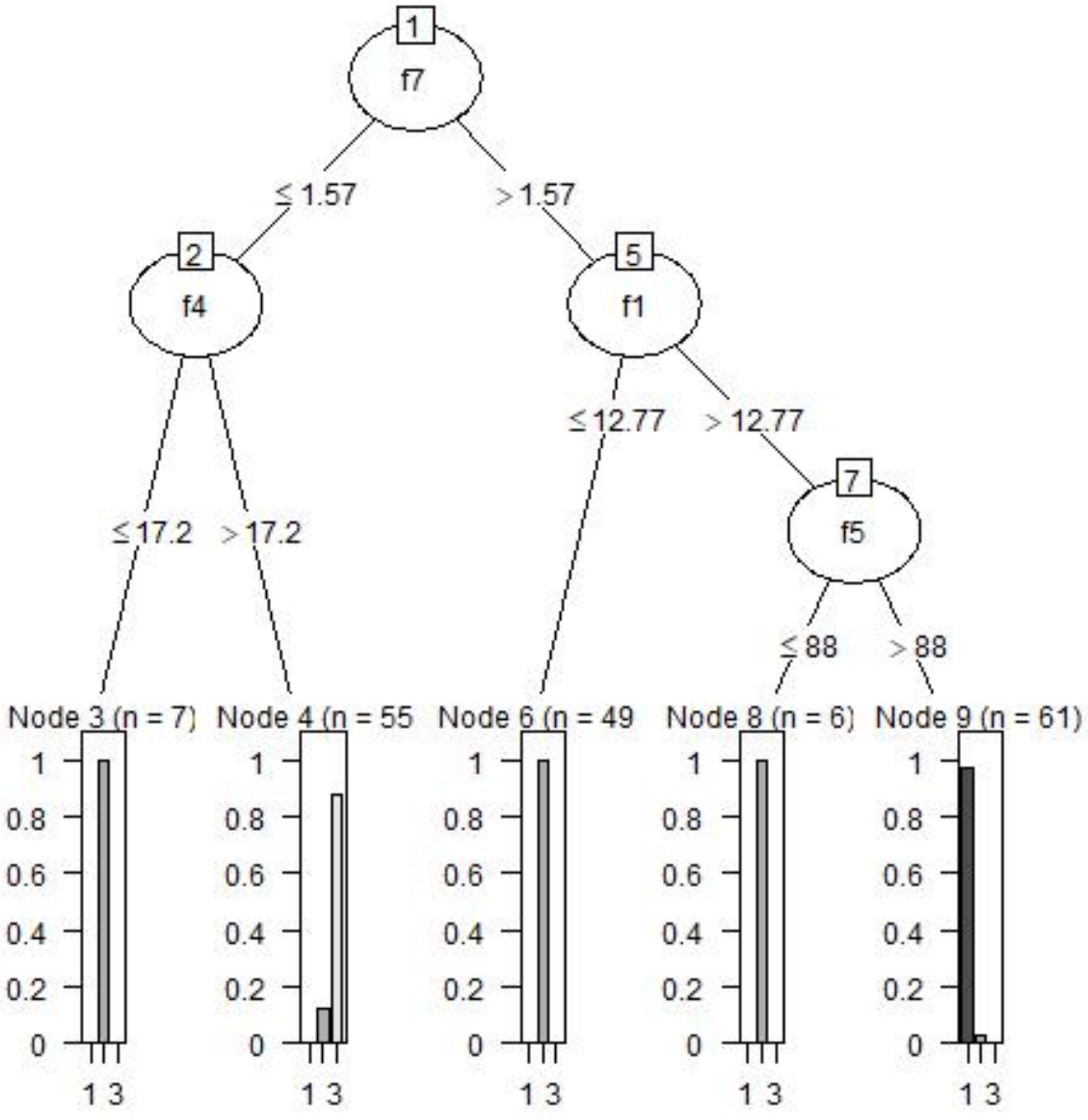

Instance Diabetes
FScore
0.0
0.2
0.4
0.6
0.8
1.0
True
FALSE
winnow


Instance Diabetes
Size
0
5
10
15
20
True
FALSE
winnow

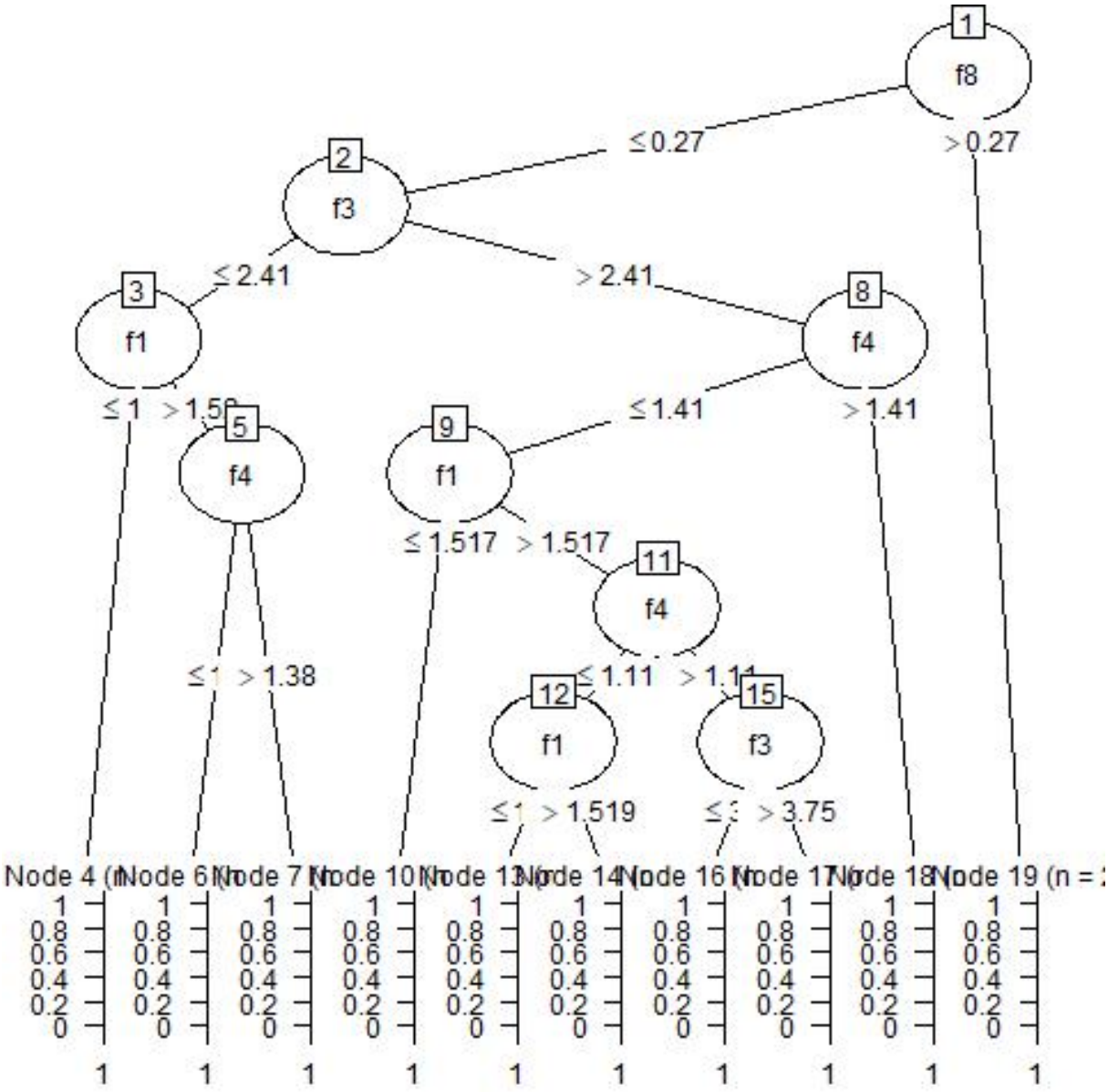
8 Optymalne drzewa decyzyjne dla zbiorów danych

Dla każdej instancji danych wybrany został zbiór parametrów, który powinien dążyć do polepszenia jakości klasyfikacji oraz minimalizacji drzewa, priorytezując jakość. Wyniki zostały przedstawione poniżej.

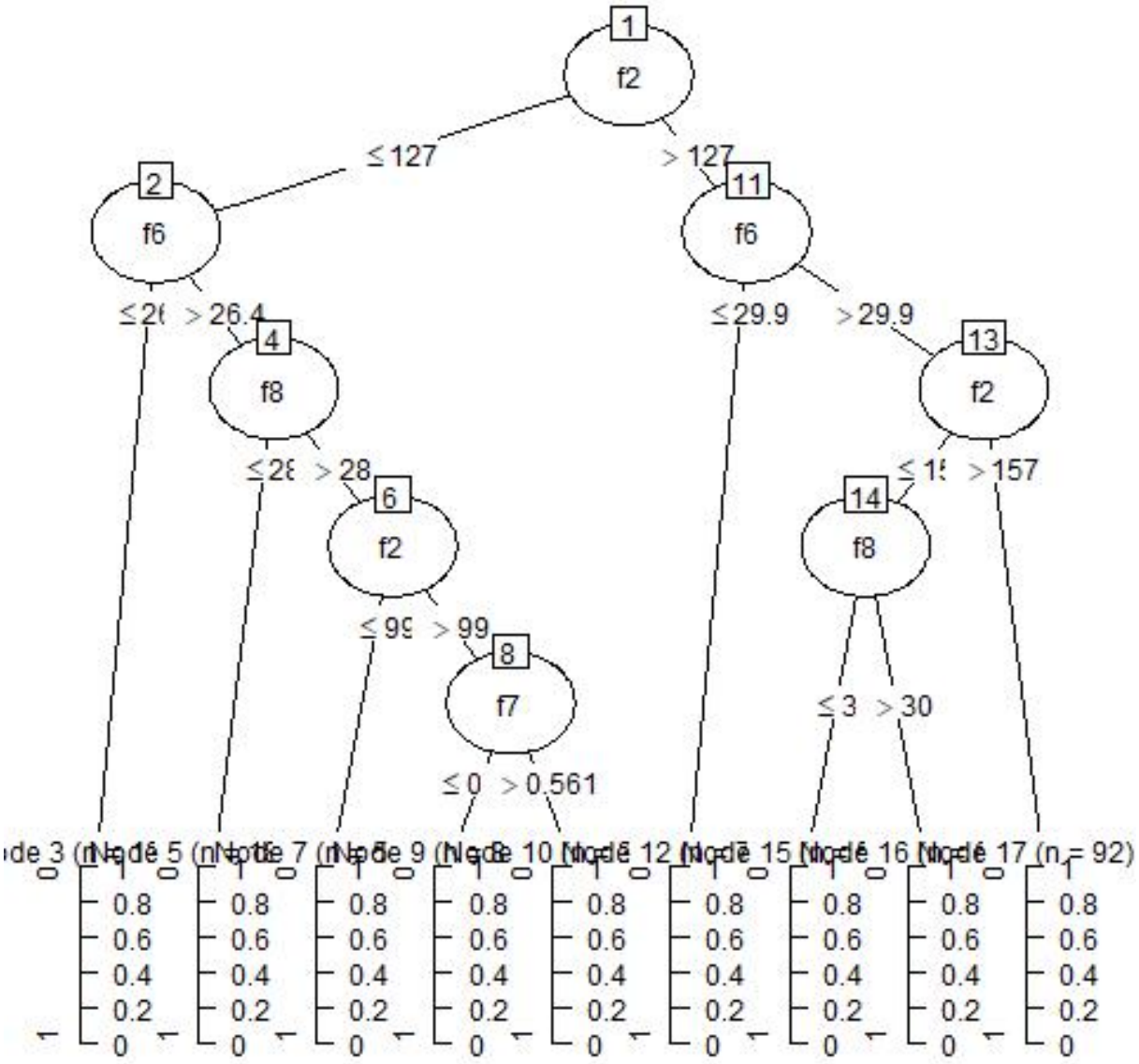
Zbiór	CF	minCases	noGlobalPruning	winnow	F1
Wine	0	5	TRUE	FALSE	0.932
Glass	0.1	5	FALSE	TRUE	0.691
Diabetes	0.1	20	FALSE	FALSE	0.816



Rysunek 14: Optymalne drzewo klasyfikujące dla instancji wine



Rysunek 15: Optymalne drzewo klasyfikujące dla instancji glass

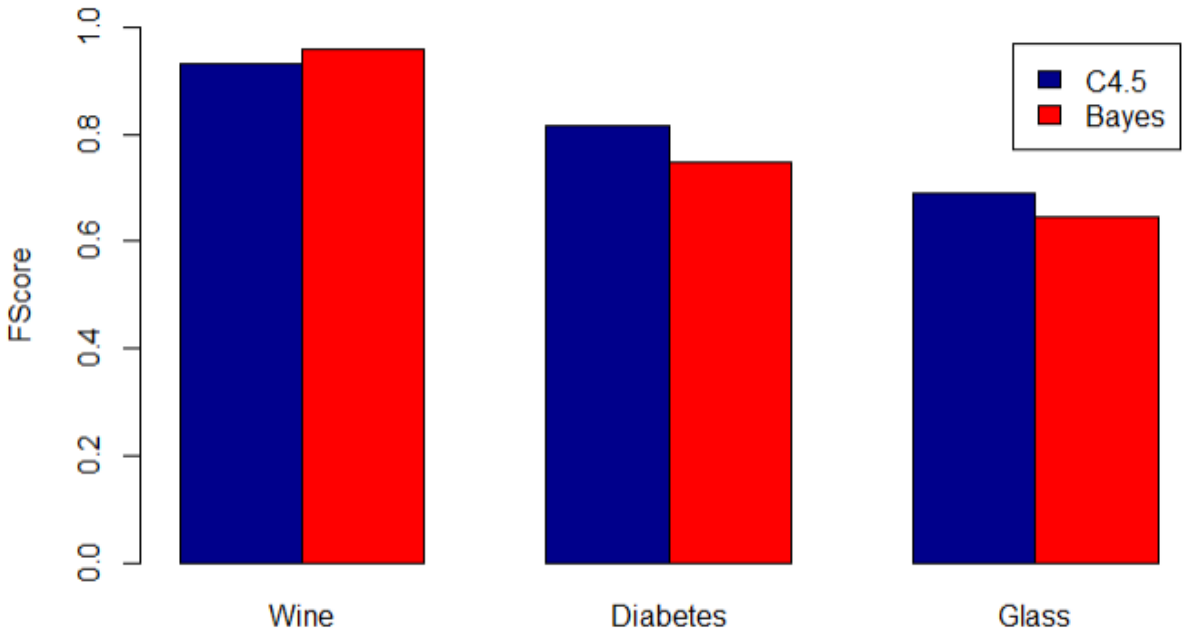


Rysunek 16: Optymalne drzewo klasyfikujące dla instancji diabetes

9 Porównanie wyników z metodą "Naive Bayes"

Do porównania wyników między rezultatami otrzymanymi poprzez użycie drzewa decyzyjnego, a metodą "Naive Bayes" posłuży miara F1. Porównane zostaną najlepsze rezultaty.

Zbiór	C4.5	Naive Bayes
Wine	0.932	0.957
Glass	0.691	0.646
Diabetes	0.816	0.748



Rysunek 17: Zestawienie wyników dla różnych klasyfikatorów

10 Wnioski

Zaimplementowane drzewa decyzyjne w przypadku dwóch zbiorów dały lepsze rezultaty. Dzięki odpowiedniej parametryzacji istnieje możliwość ograniczenia rozmiaru drzewa oraz polepszenia jakości klasyfikacji.// Z powodu jasnych zasad klasyfikacji drzewa decyzyjne są przychylniej akceptowane przez specjalistów w wielu branżach niż deep learning, ponieważ łatwo można prześledzić każdą decyzję podejmowaną przez drzewo.// Testy kroswalidacji po raz kolejny wykazały, że najlepszą metodą jest kroswalidacja stratyfikowana. Przed jej użyciem warto jednak zapoznać się z danymi wejściowymi, ponieważ w przypadku podziału na więcej foldów niż liczności najmniej licznej klasy wyniki mogą ulec pogorszeniu.