POLITECHNIKA WROCŁAWSKA
WYDZIAŁ INFORMATYKI I ZARZĄDZANIA

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

Indukcja drzew decyzyjnych C5.0 w R

Dokumentacja ćwiczenia nr 2

AUTOR:
Adam Dłubak

PROWADZĄCY:
Dr inż. Paweł Myszkowski

WROCŁAW, 2018

Spis treści

Rozdział 1

Zbiory Badawcze

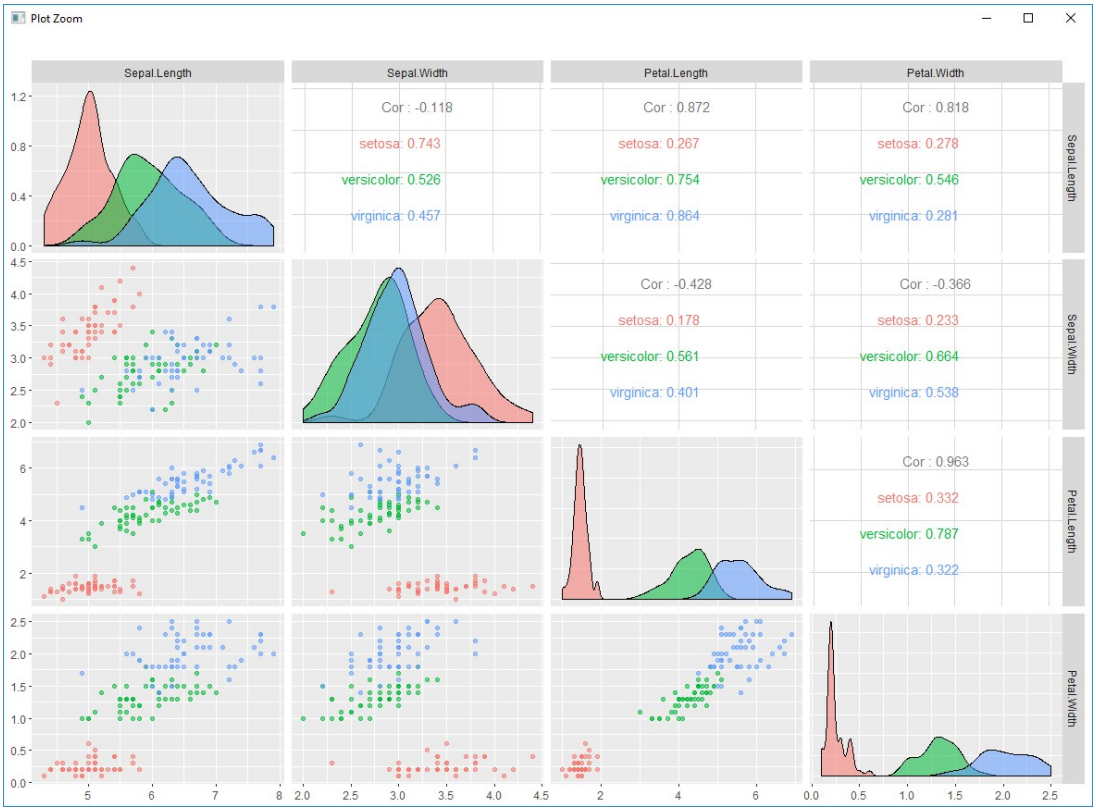
W ramach realizacji ćwiczenia, badania zostały oparte o 1 zbiór wykorzystany do wstępnej weryfikacji oraz 3 zbiory danych testowych, które uwzględniały również dane z wartościami ciągłymi. Wykorzystane zbiory danych to:

1.1. Iris Dataset

Zestaw pomiarów kwiatów irysa, udostępniony po raz pierwszy przez Ronalda Fishera w roku 1936. Jeden z najbardziej znanych zbiorów, a zarazem bardzo prosty i użyteczny. Zbiór irysów składa się z 4 wartości pomiarów jego płatków (szerokości i długość) oraz klasy do jakiej należy. W tym ćwiczeniu został on wykorzystany jako zbiór do wstępnej weryfikacji poprawności wykonywanych operacji na algorytmie C5.0.

- Liczba atrybutów: 4
- Rodzaj atrybutów: wartości typu Float
- Liczba instancji: 150
- Liczba klas: 3

Rys. 1.1: Rozłożenie ilościowe poszczególnych klas zbioru Iris

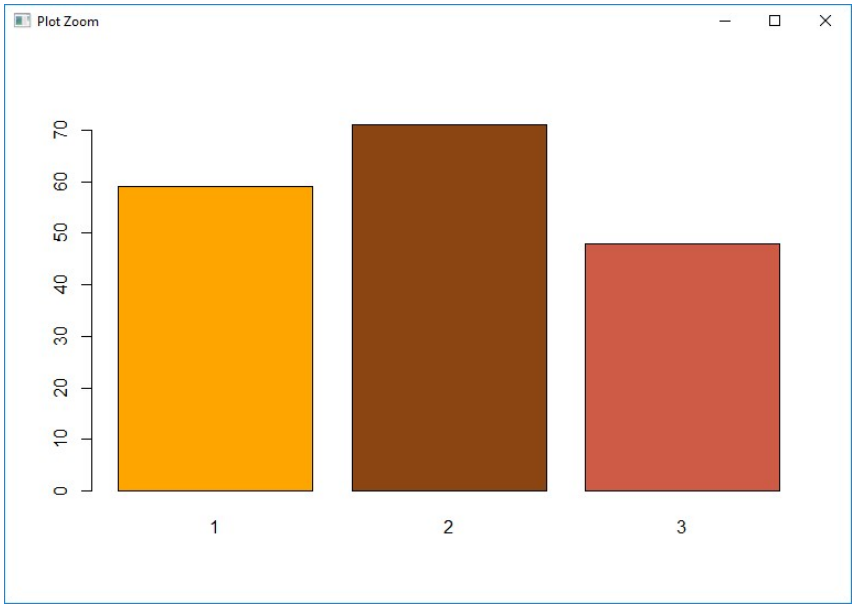


Rys. 1.2: Zestawienie zależności atrybutów i klas zbioru Iris

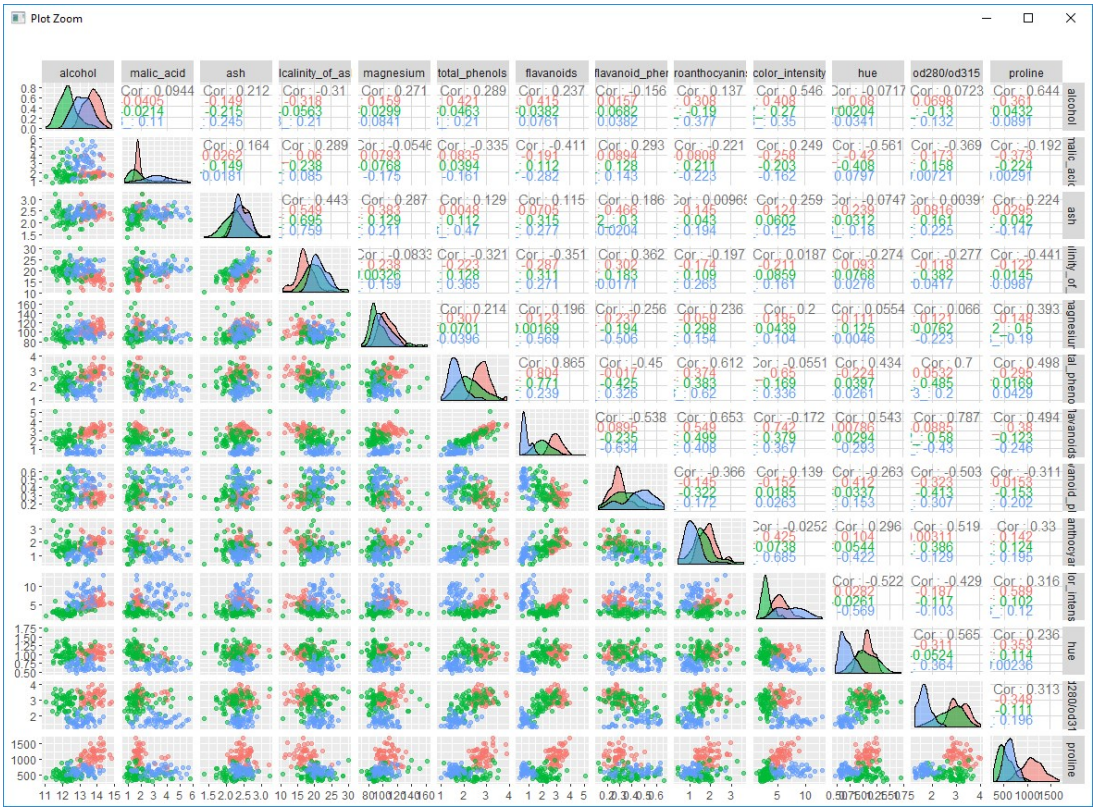
1.2. Wine Dataset

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 Float i Intiger.
- Liczba instancji: 178.
- Liczba klas: 3



Rys. 1.3: Rozłożenie ilościowe poszczególnych klas zbioru Wine

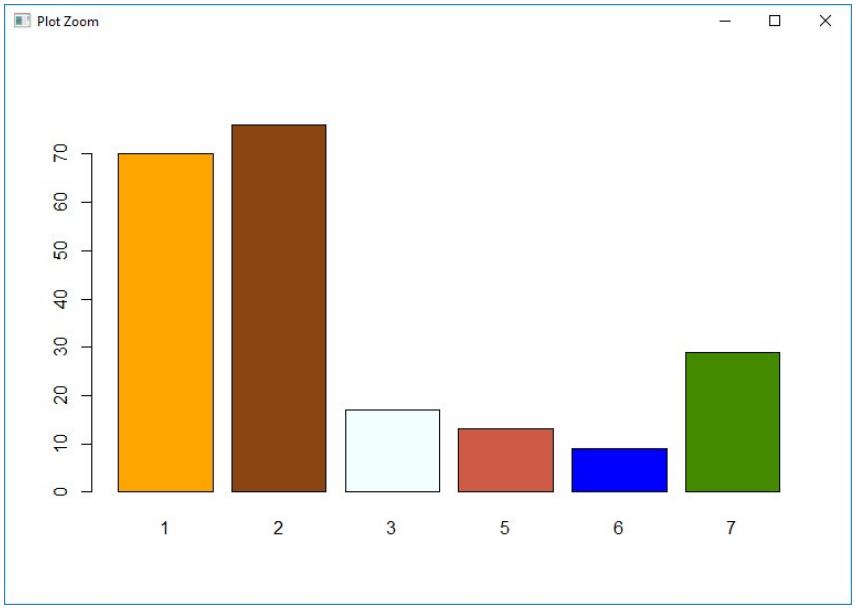


Rys. 1.4: Zestawienie zależności atrybutów i klas zbioru Wine

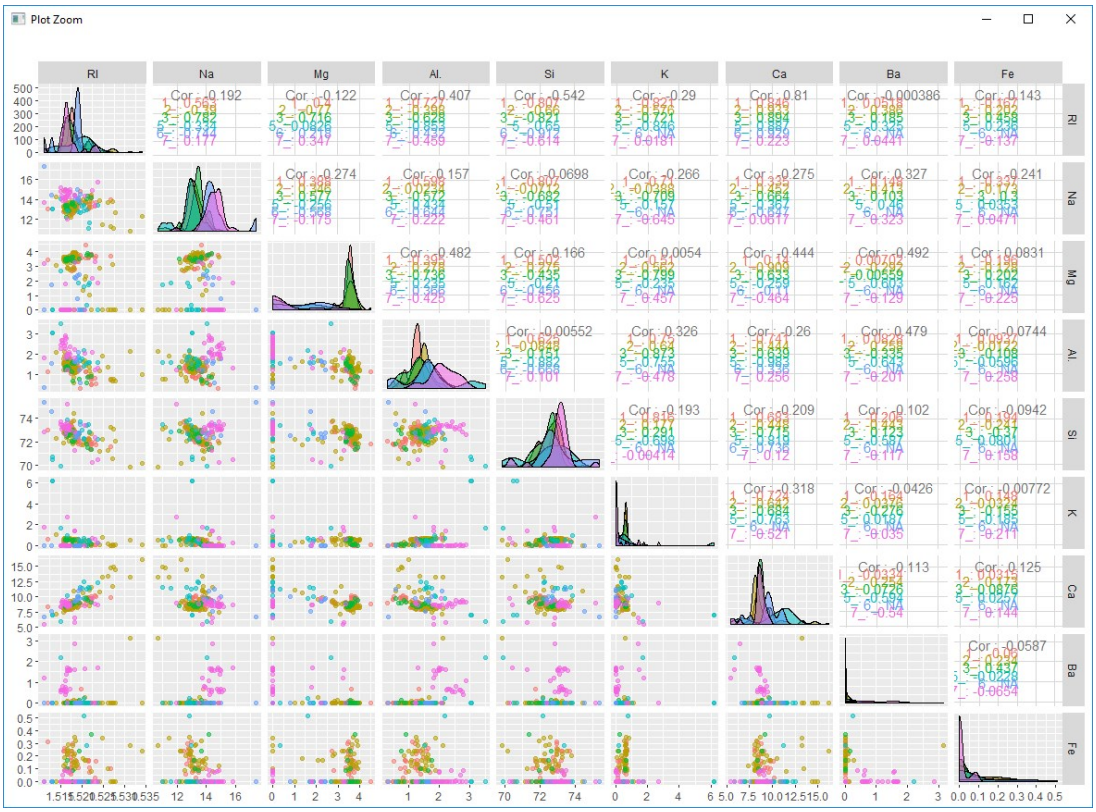
1.3. Glass Dataset

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: 9.
- Rodzaj atrybutów: realistyczne, ciągłe.
- Liczba instancji: 214.
- Liczba klas: 7, występuje zróżnicowanie w ilości instancji w klasie, ponadto w zbiorze nie występuje krotka o klasie nr 4.



Rys. 1.5: Rozłożenie ilościowe poszczególnych klas zbioru Glass

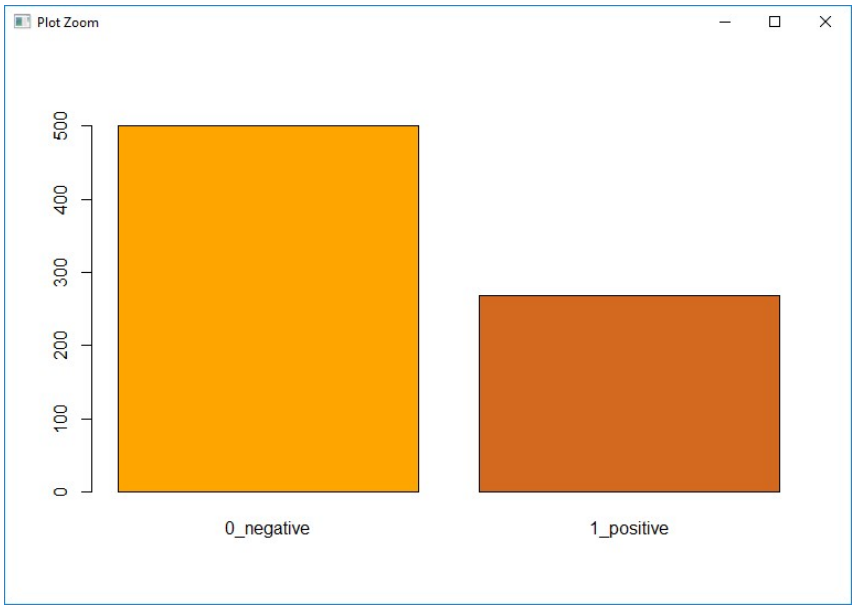


Rys. 1.6: Zestawienie zależności atrybutów i klas zbioru Glass

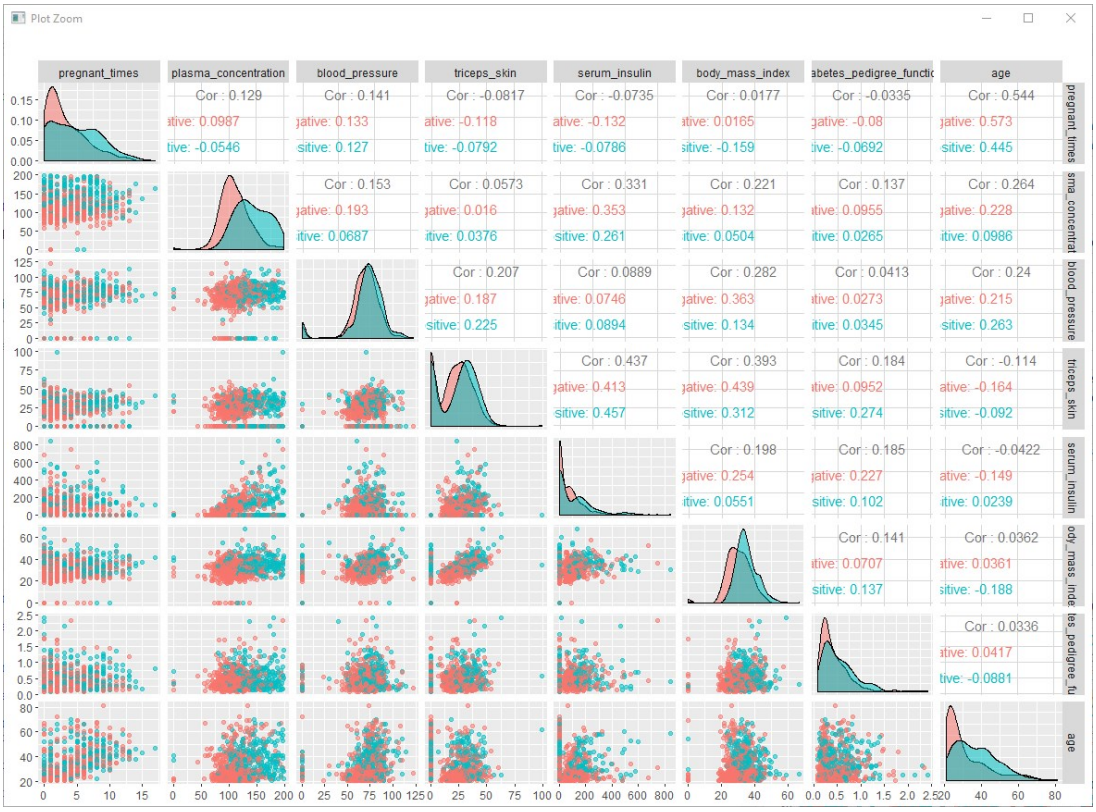
1.4. Pima Indians Diabetes Dataset

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*. Zawiera on dane dotyczące zachorowań na cukrzycę wśród kobiet z indiańskiego plemienia Pima. Każdy z 768 obiektów zbioru opisany jest przy pomocy 8 cech zawierających następujące informacje: ile razy pacjentka była w ciąży, test tolerancji glukozy, ciśnienie rozkurczowe, grubość zagięcia skóry, poziom insuliny, masę ciała, czy ktoś w rodzinie był chory na cukrzycę oraz wiek pacjentki. Każdy z obiektów przynależy do jednej z dwóch klas. Pierwsza klasa oznacza, że pacjentka nie choruje na cukrzycę, a druga klasa oznacza, że dana kobieta jest diabetykiem.

- Liczba atrybutów: 8.
- Rodzaj atrybutów: realistyczne, ciągłe i typu Intiger (wiek i ilość dotychczasowych ciąż).
- Liczba instancji: 768.
- Liczba klas: 2 - wartość 1 (pozytywna) lub 0 (negatywna).



Rys. 1.7: Rozłożenie ilościowe poszczególnych klas zbioru Diabetes



Rys. 1.8: Zestawienie zależności atrybutów i klas zbioru Diabetes

Rozdział 2

Wstęp teoretyczny

2.1. Algorytm C4.5

Algorytm C4.5 zaproponowany przez Quinlana (Quinlan, 1996) jest udoskonaloną wersją wcześniejszego algorytmu ID3 (Quinlan, 1986). W porównaniu do algorytmu ID3 poprawione zostało m.in. kryterium podziału, tak aby uzyskiwane podziały dla większych zbiorów danych generowały mniejszy błąd klasyfikacji i możliwa była klasyfikacja obiektów z brakującymi wartościami atrybutów. W algorytmie ID3 jako kryterium podziału stosowana była reguła zysku informacji, natomiast w C4.5 reguła względnego zysku.

Ponadto w algorytmie C4.5 wprowadzono przycinanie. Początkowo była to podstawowa metoda przycinania pesymistycznego (*ang. pessimistic pruning*), która następnie podlegała stopniowym udoskonaleniom (*ang. error-based pruning*). Podczas procesu uczenia się oraz klasyfikacji istnieje możliwość pracy z obiektami nieposiadającymi wartości wszystkich atrybutów (dane z brakującymi wartościami atrybutów), dodatkowo algorytm C4.5 dostosowany jest do pracy z ciągłymi wartościami atrybutów.

C4.5 w przypadku niepełnego zestawu danych, na podstawie znanych wartości wyznacza najbardziej prawdopodobną wartość brakującej cechy i na tej podstawie dokonuje klasyfikacji.
W przypadku analizowania ciągłego rozkładu wartości dokonuje on dyskretyzacji danych. Realizowane jest to w następujący sposób :

- Zbiór wartości z zestawu treningowego dzieli się na podzbiory zawierające wartości < Ai oraz > Ai , gdzie oczywiście Ai jest jedna z wartości analizowanej cechy. Dla każdego z takich podziałów wyznaczany jest zysk informacyjny. Ostatecznie wybierany jest taki podział który generuje największy zysk informacji.

Dla przeciwdziałania występującej w ID 3 możliwości przerostu drzewa , C4.5 zawiera funkcję przycinania drzewa która ma następujące cechy:

- przycinanie jest realizowane wstecznie (zaczyna się od liści)
- mając dany węzeł nie będący liściem i jego poddrzewo obliczana jest w heurystyczny sposób wartość przewidywanego błędu dla aktualnego poddrzewa.

- obliczana jest wartość przewidywanego błędu dla sytuacji, gdyby rozpatrywane poddrzewo zastąpić pojedynczym liściem z kategorią najpopularniejszą wśród liści.
- porównane zostają te dwie wartości i ewentualnie dokonuje się zamiany poddrzewa na pojedynczy liść.

2.2. Algorytm C5.0

Zarówno C4.5, jak i C5.0 mogą tworzyć klasyfikatory wyrażone jako drzewa decyzyjne lub zestawy reguł. W wielu aplikacjach preferowane są zestawy reguł, ponieważ są prostsze i łatwiejsze do zrozumienia niż drzewa decyzyjne, ale metody zestawu reguł C4.5 są powolne i wymagają pamięci. C5.0 zawiera nowe algorytmy generowania zestawów reguł, a według literatury poprawa jest znaczna. Poniżej przedstawione zostaną przykładowe dane zaczerpnięte z artykułu "*Is See5/C5.0 Better Than C4.5?*" (*link: http://rulequest.com/see5-comparison.html*)

- **Dokładność** - Zestawy reguł C5.0 mają zauważalnie niższe wskaźniki błędów w niewidocznych przypadkach dla różnych zestawów danych.
- **Prędkość** - C5.0 jest znacznie szybszy; wykorzystuje różne algorytmy i jest wysoce zoptymalizowany. Na przykład C4.5 potrzebował więcej niż osiem godzin, aby znaleźć zestaw reguł dla dla zbioru Forest, ale C5.0 wykonał zadanie w czasie poniżej trzech minut.
- **Pamięć** - C5.0 zwykle używa o rząd wielkości mniej pamięci niż C4.5 podczas konstruowania zestawu reguł. W przypadku zbioru danych Forest, C4.5 potrzebuje więcej niż 3 GB, natomiast C5.0 wymaga mniej niż 200 MB.

C5.0 przynosi również wiele ulepszeń i nowości w porównaniu do algorytmu C4.5. Kilka z nich zostanie wymienionych poniżej: **Boosting** - opierając się na badaniach Freunda i Schapire'a, jest to ekscytująca nowość, która nie ma odpowiednika w C4.5. Boosting to technika generowania i łączenia wielu klasyfikatorów w celu poprawy dokładności predykcyjnej.

C5.0 zawiera kilka nowych udogodnień, takich jak **zmienne koszty błędnej klasyfikacji**. W C4.5 wszystkie błędy są traktowane jako równe, ale w zastosowaniach praktycznych niektóre błędy klasyfikacji są poważniejsze niż inne. C5.0 pozwala na zdefiniowanie oddzielnego kosztu dla każdej przewidywanej / rzeczywistej pary klas; jeśli ta opcja jest używana, C5.0 konstruuje klasyfikatory, aby zminimalizować oczekiwane koszty błędnej klasyfikacji zamiast współczynników błędów.

C5.0 ma kilka **nowych typów danych** oprócz tych dostępnych w C4.5, w tym daty, czasy, znaczniki czasu, uporządkowane atrybuty dyskretne i etykiety na okładki. Oprócz brakujących wartości, C5.0 umożliwia zanotowanie wartości jako nie dotyczy. Ponadto C5.0 zapewnia możliwości definiowania nowych atrybutów jako funkcji innych atrybutów.

C5.0 jest również łatwiejszy w użyciu. Opcje zostały uproszczone i rozszerzone - na przykład do **próbkowania** i **walidacji krzyżowej** - a programy C4.5 do generowania drzew decyzyjnych i zestawów reguł zostały połączone w jeden program.

2.2.1. Parametry algorytmu

Biblioteka C5.0 udostępnia metodę *C5.0Control()*, która umożliwia dostosować różne parametry kontrolujące dopasowanie algorytmu C5.0.

- **subset** - wartość logiczna; czy model powinien oceniać grupy dyskretnych predyktorów dla podziałów? C5.0 domyślnie ustawia ten parametr na FALSE, co oznacza, że żadne próby grupowania nie zostaną ocenione na etapie wzrostu drzewa.

- **bands** - liczba całkowita z zakresu od 2 do 1000. Jeśli parametr rbm ustawiony jest na TRUE, model porządkuje reguły według ich wpływu na współczynnik błędu i grupuje reguły na określoną liczbę zakresów. Modyfikuje to wynik tak, aby wpływ na poziom błędu był widoczny dla grup reguł w zakresie. Jeśli ta opcja zostanie wybrana, a reguły = FALSE, zostanie wyświetlone ostrzeżenie i reguły zostaną zmienione na TRUE.

- **winnow** - wartość logiczna: czy używać funkcji predyktorowania (np. wybór funkcji)?

- **noGlobalPruning** - wartość logiczna; odpowiada za przełączanie czy ma ostatecznie zostać wykonane globalne przycięcie w celu uproszczenia struktury drzewa.

- **CF** - liczba z zakresu [0, 1] dla współczynnika ufności.

- **minCases** - jest liczbą całkowitą odpowiadającą najmniejszej liczbie próbek, które należy umieścić w co najmniej dwóch podziałach.

- **fuzzyThreshold** - wartość logiczna służąca do oceny możliwych zaawansowanych podziałów danych.

- **sample** - wartość z zakresu [0, 999], która określa losową proporcję danych używaną do szkolenia modelu. Domyślnie wszystkie próbki są używane do trenowania modelu. Próbki niewykorzystywane do treningu są używane do oceny dokładności modelu na wyjściu.

- **seed** - liczba całkowita dla liczb losowych ziaren kodu w C.

- **earlyStopping** - wartość logiczna określająca, czy powinna zostać wykorzystana wewnętrzna metoda zatrzymania boostingu (wzmocnienia).

- **label** - Etykieta dla wyjścia danych ustawionych parametrów.

2.3. Miary jakości klasyfikatora

- **Trafność** (*ang. Accuracy*) - określa, jaka część prognozowanych etykiet jest zgodna z rzeczywistymi wynikami. Oznacza to procent poprawnie sklasyfikowanych etykiet.

- **Precyzja** (*ang. Precision*) - określa liczbę adekwatnych elementów w zbiorze wyników. W kontekście klasyfikacji jest to liczba poprawnych etykiet z wszystkich zbiorów klasyfikowanych etykiet. Wyniki są uśredniane dla wszystkich etykiet.

- **Czułość** (*ang. Recall*)- określa liczbę poprawnych wyników względem liczby wszystkich poprawnych etykiet. W kontekście klasyfikacji jest to liczba poprawnie sklasyfikowanych etykiet w zbiorze podzielona przez łączną liczbę etykiet ze zbioru. Wyniki są uśredniane.

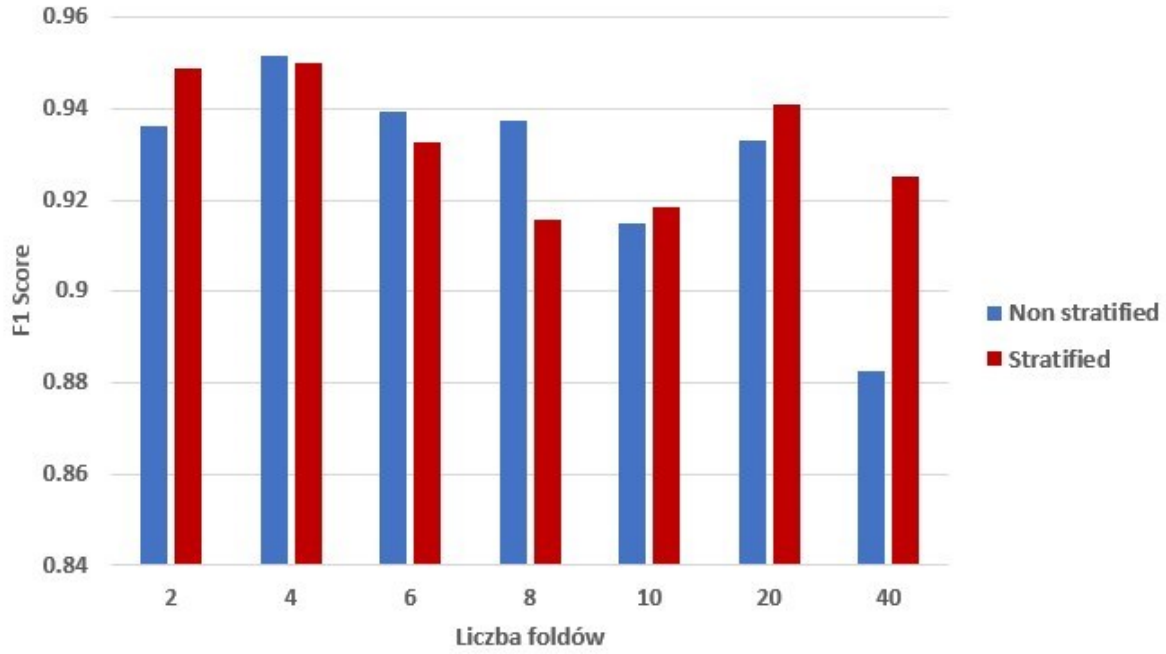
- **Wskaźnik F1** (*ang. F1 Score*)- jest to średnia harmoniczna precyzji i czułości. Najczęściej stosowana jest dla niezrównoważonych zbiorów danych w celu ustalenia, czy klasyfikator działa dobrze dla wszystkich klas.

Rozdział 3

Badania i analiza wyników

3.1. Kroswalidacja - ”zwykła” i stratyfikowana

3.1.1. Wine Dataset

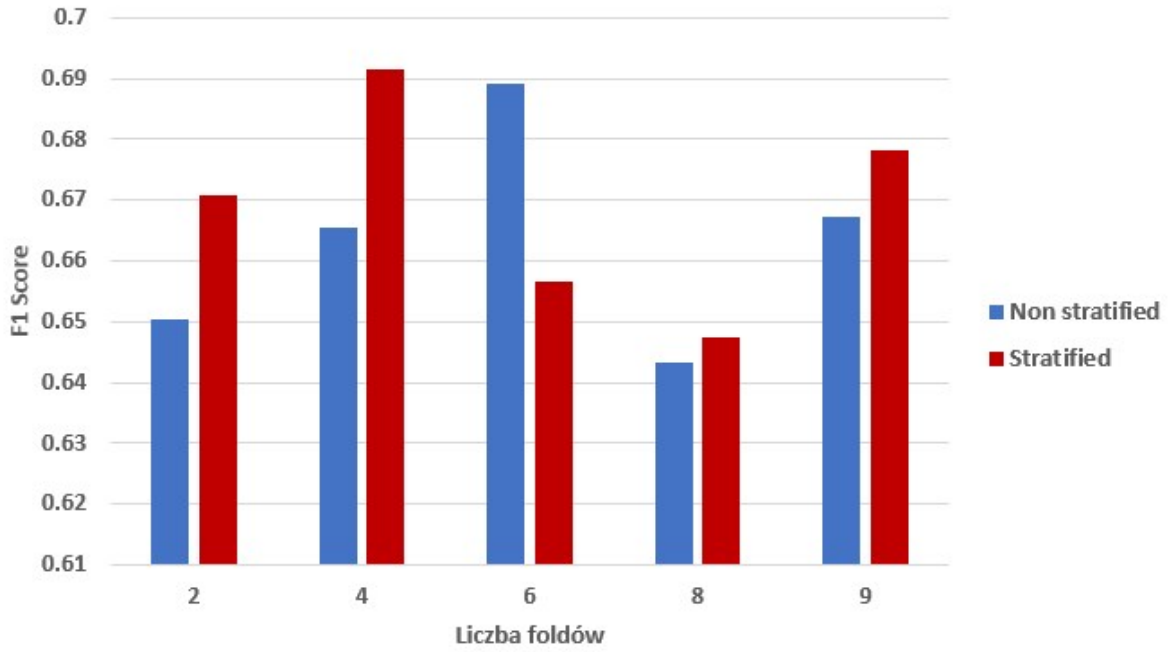


Rys. 3.1: Porównanie kroswalidacji ”zwykłej” i stratyfikowanej dla zbioru Wine

Tab. 3.1: Porównanie kroswalidacji ”zwykłej” i stratyfikowanej (F1 Score) dla zbioru Wine

Liczba Foldów	**2**	**4**	**6**	**8**	**10**	**20**	**40**
K-Fold	0.936	0.951	0.939	0.937	0.915	0.933	0.883
Stratified K-Fold	0.949	0.950	0.933	0.916	0.918	0.941	0.925

3.1.2. Glass Dataset

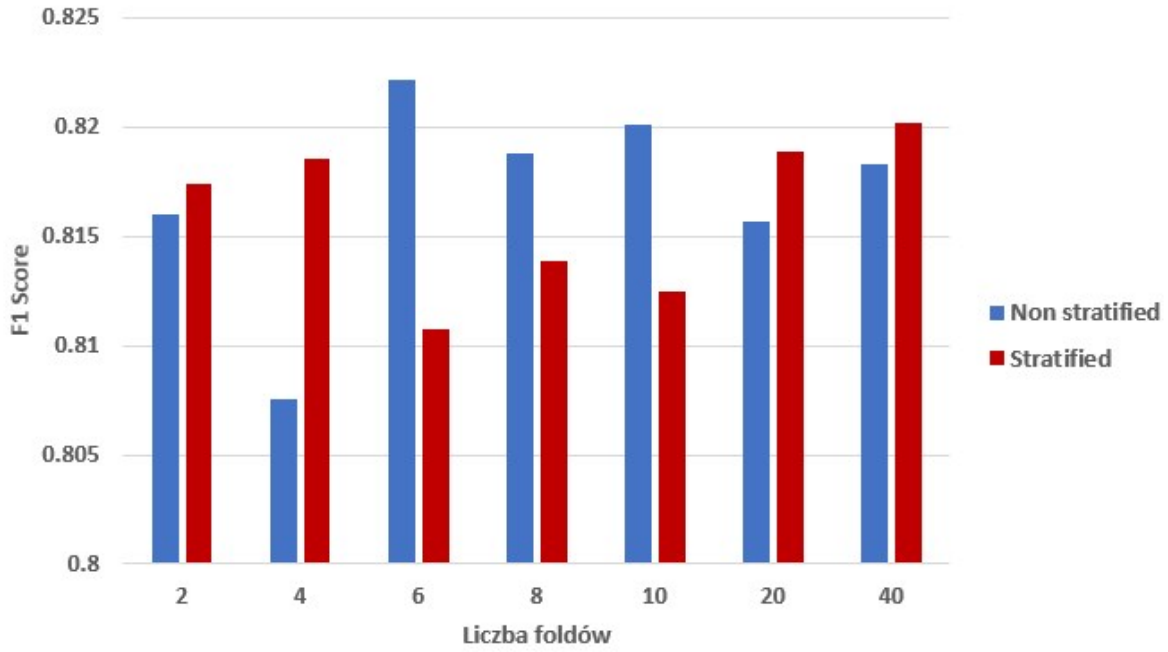


Rys. 3.2: Porównanie kroswalidacji ”zwykłej” i stratyfikowanej dla zbioru Glass

Tab. 3.2: Porównanie kroswalidacji ”zwykłej” i stratyfikowanej (F1 Score) dla zbioru Glass

Liczba Foldów	**2**	**4**	**6**	**8**	**9**
K-Fold	0.650	0.666	0.689	0.643	0.667
Stratified K-Fold	0.671	0.691	0.657	0.647	0.678

3.1.3. Pima Indians Diabetes Dataset



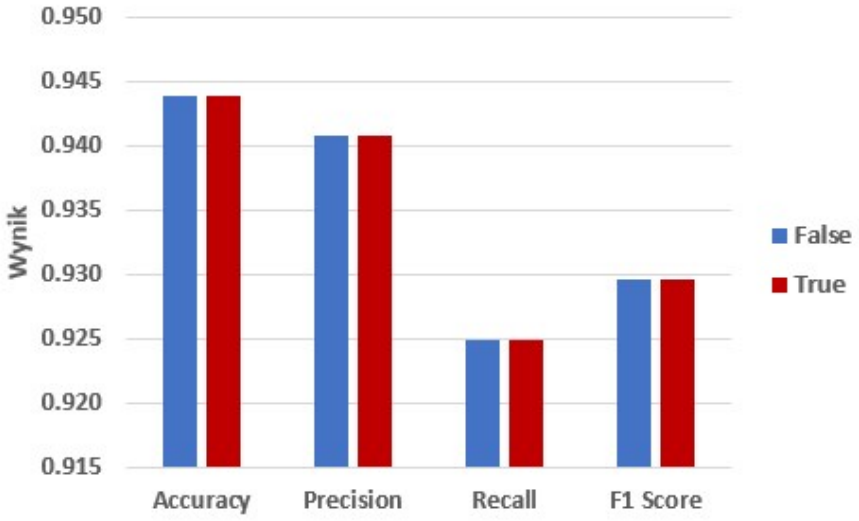
Rys. 3.3: Porównanie kroswalidacji ”zwykłej” i stratyfikowanej dla zbioru P.I.D.

Tab. 3.3: Porównanie kroswalidacji ”zwykłej” i stratyfikowanej (F1 Score) dla zbioru P.I.D.

Liczba Foldów	**2**	**4**	**6**	**8**	**10**	**20**	**40**
K-Fold	0.816	0.808	0.822	0.819	0.820	0.816	0.818
Stratified K-Fold	0.817	0.819	0.811	0.814	0.812	0.819	0.820

3.2. Analiza parametrów algorytmu dla różnych zbiorów danych

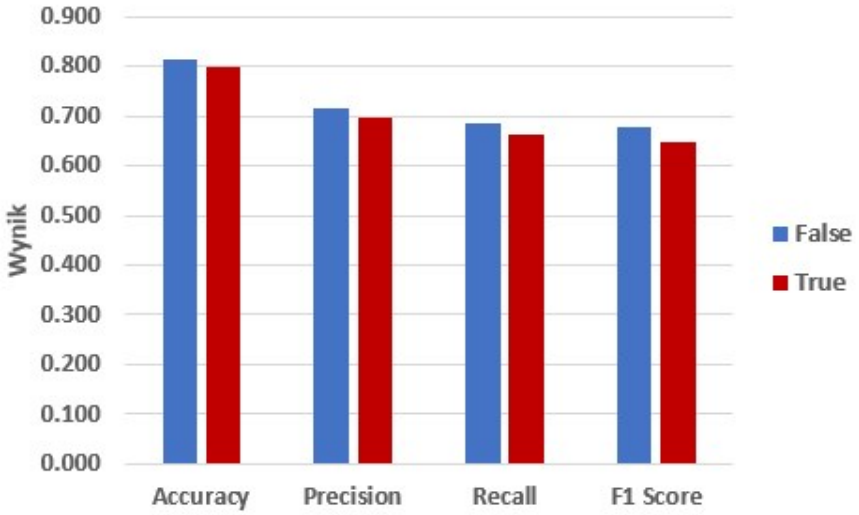
3.2.1. Parametr *noGlobalPruning*



Rys. 3.4: Porównanie parametru *noGlobalPruning* dla zbioru Wine

Tab. 3.4: Porównanie parametru *noGlobalPruning* dla zbioru Wine

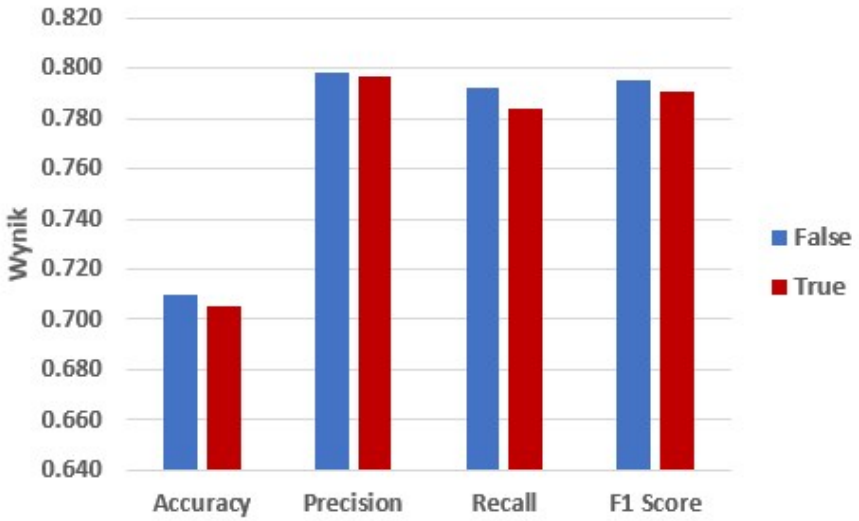
	False	**True**
Accuracy	0.944	0.944
Precision	0.941	0.941
Recall	0.925	0.925
F1 Score	0.930	0.930
Tree Size	6	6



Rys. 3.5: Porównanie parametru *noGlobalPruning* dla zbioru Glass

Tab. 3.5: Porównanie parametru *noGlobalPruning* dla zbioru Glass

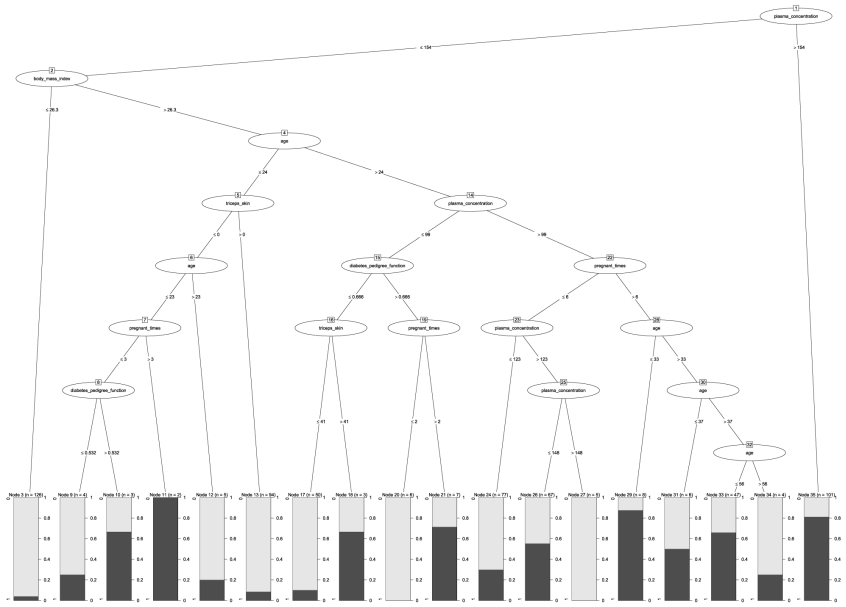
	False	**True**
Accuracy	0.813	0.799
Precision	0.717	0.697
Recall	0.687	0.663
F1 Score	0.679	0.647
Tree Size	22	19



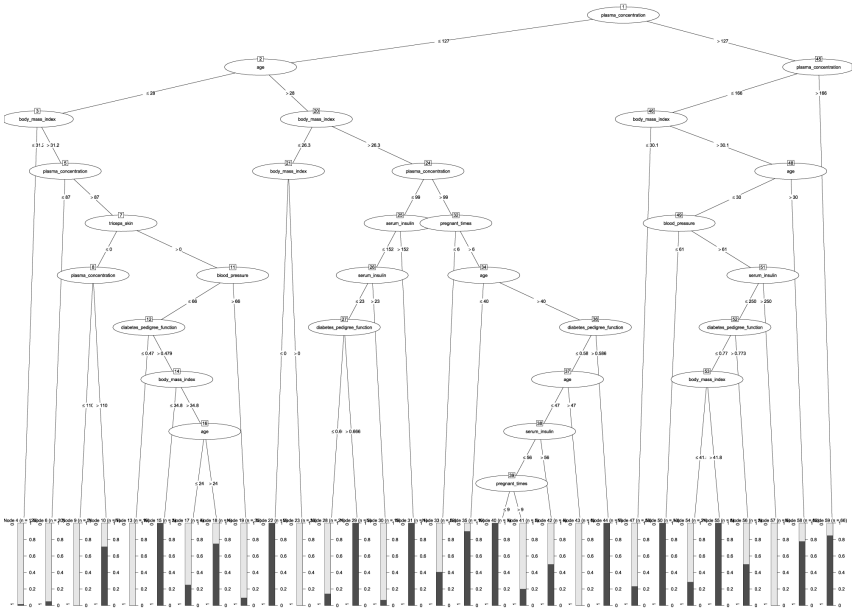
Rys. 3.6: Porównanie parametru *noGlobalPruning* dla zbioru P.I.D.

Tab. 3.6: Porównanie parametru *noGlobalPruning* dla zbioru P.I.D.

	False	**True**
Accuracy	0.709	0.705
Precision	0.798	0.797
Recall	0.792	0.784
F1 Score	0.795	0.790
Tree Size	30	18

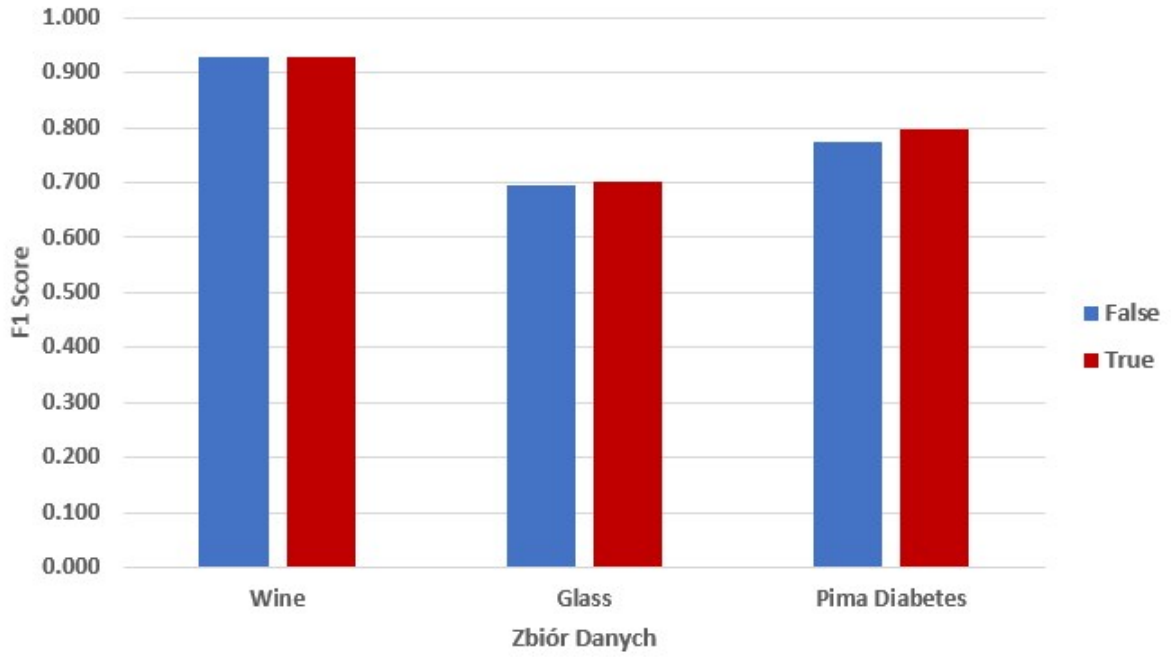


Rys. 3.7: Wizualizacja drzewa z parametrem *noGlobalPruning* = True dla zbioru P.I.D.



Rys. 3.8: Wizualizacja drzewa z parametrem *noGlobalPruning* = False dla zbioru P.I.D.

3.2.2. Parametr *FuzzyThreshold*

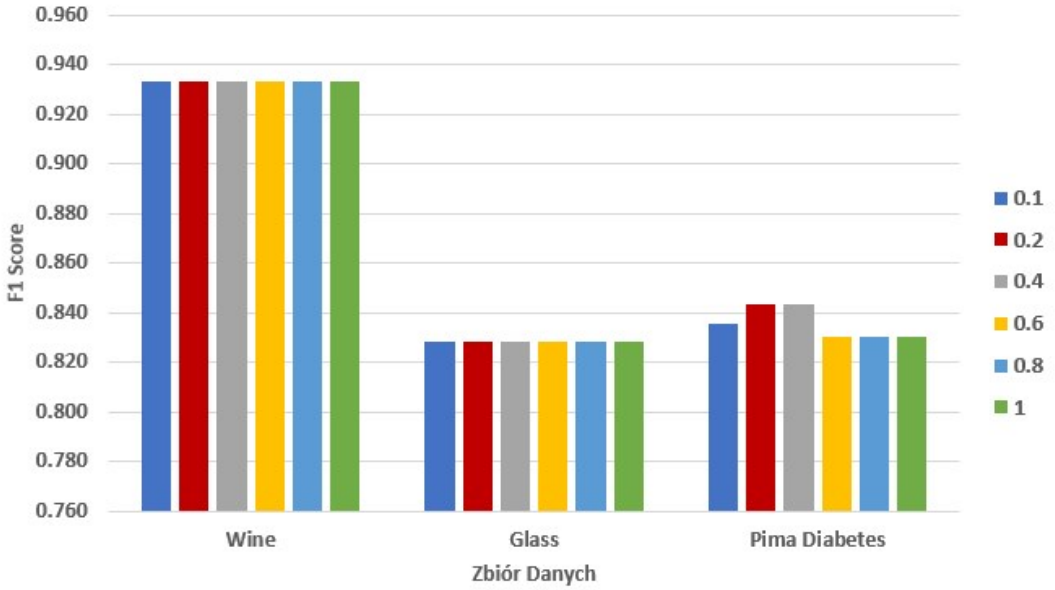


Rys. 3.9: Porównanie parametru *fuzzyThreshold* (F1 Score) dla wszystkich zbiorów

Tab. 3.7: Porównanie parametru *fuzzyThreshold* dla wszystkich badanych zbiorów

	Wine		**Glass**		**Pima Diabetes**	
	False	**True**	**False**	**True**	**False**	**True**
Accuracy	0.945	0.945	0.834	0.841	0.722	0.735
Precision	0.939	0.939	0.672	0.683	0.827	0.828
Recall	0.925	0.925	0.736	0.744	0.728	0.768
F1 Score	0.928	0.928	0.694	0.703	0.774	0.797
Tree Size	6	6	18	18	22	22

3.2.3. Parametr *CF*

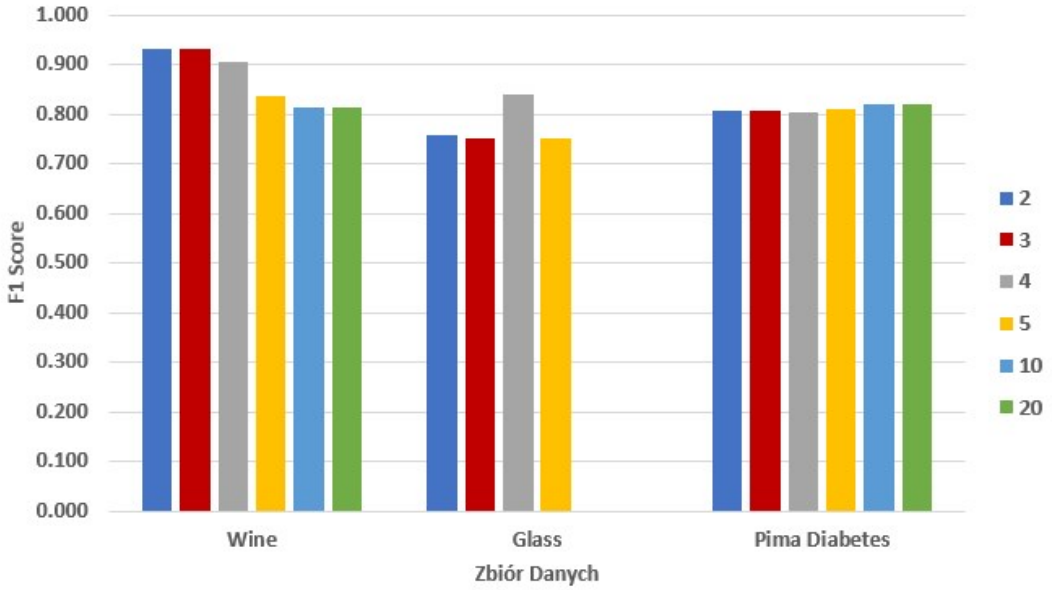


Rys. 3.10: Porównanie parametru *CF* (F1 Score) dla wszystkich zbiorów

Tab. 3.8: Porównanie parametru *CF* dla wszystkich badanych zbiorów

Value	**0.1**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
	Wine					
Accuracy	0.923	0.923	0.923	0.923	0.923	0.923
Precision	0.923	0.923	0.923	0.923	0.923	0.923
Recall	0.968	0.968	0.968	0.968	0.968	0.968
F1 Score	0.933	0.933	0.933	0.933	0.933	0.933
Tree Size	5	5	5	5	5	5
	Glass					
Accuracy	0.980	0.980	0.980	0.980	0.980	0.980
Precision	0.500	0.500	0.500	0.500	0.500	0.500
Recall	0.250	0.250	0.250	0.250	0.250	0.250
F1 Score	0.829	0.829	0.829	0.829	0.829	0.829
Tree Size	12	22	22	22	22	22
	Pima Indians Diabetes					
Accuracy	0.767	0.760	0.760	0.748	0.748	0.748
Precision	0.839	0.824	0.824	0.820	0.820	0.820
Recall	0.832	0.864	0.864	0.840	0.840	0.840
F1 Score	0.835	0.844	0.844	0.830	0.830	0.830
Tree Size	13	22	23	31	31	31

3.2.4. Parametr *minCases*



Rys. 3.11: Porównanie parametru *minCases* (F1 Score) dla wszystkich zbiorów

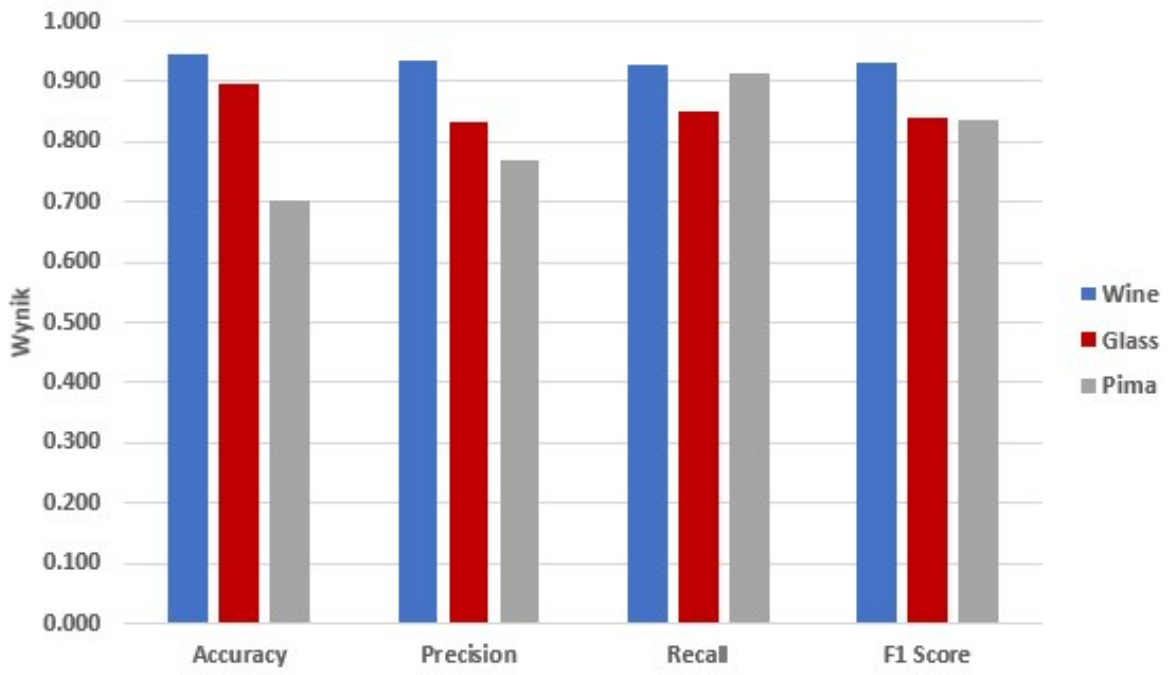
Tab. 3.9: Porównanie parametru *minCases* dla wszystkich badanych zbiorów

Value	**2**	**3**	**4**	**5**	**10**	**20**
	Wine					
Accuracy	0.946	0.946	0.921	0.875	0.860	0.860
Precision	0.935	0.935	0.937	0.843	0.814	0.814
Recall	0.929	0.929	0.893	0.834	0.814	0.814
F1 Score	0.931	0.931	0.905	0.838	0.814	0.814
Tree Size	6	6	5	4	3	3
	Glass					
Accuracy	0.858	0.840	0.900	0.840	0.818	0.638
Precision	0.819	0.820	0.833	0.820	N/A	N/A
Recall	0.773	0.739	0.850	0.739	0.711	0.378
F1 Score	0.758	0.750	0.840	0.750	N/A	N/A
Tree Size	19	15	13	12	6	4
	Pima Indians Diabetes					
Accuracy	0.743	0.743	0.721	0.732	0.726	0.726
Precision	0.831	0.831	0.806	0.815	0.802	0.802
Recall	0.784	0.784	0.800	0.808	0.840	0.840
F1 Score	0.807	0.807	0.803	0.811	0.820	0.820
Tree Size	11	10	8	9	6	6

3.3. Najlepsze uzyskane wyniki

Tab. 3.10: Wykorzystane parametry do uzyskania najlepszych rezultatów

	Wine	**Glass**	**Pima Diabetes**
noGlobalPruning	True	False	False
fuzzyThreshold	True	False	True
CF	0.25	0.25	0.30
minCases	3	4	10

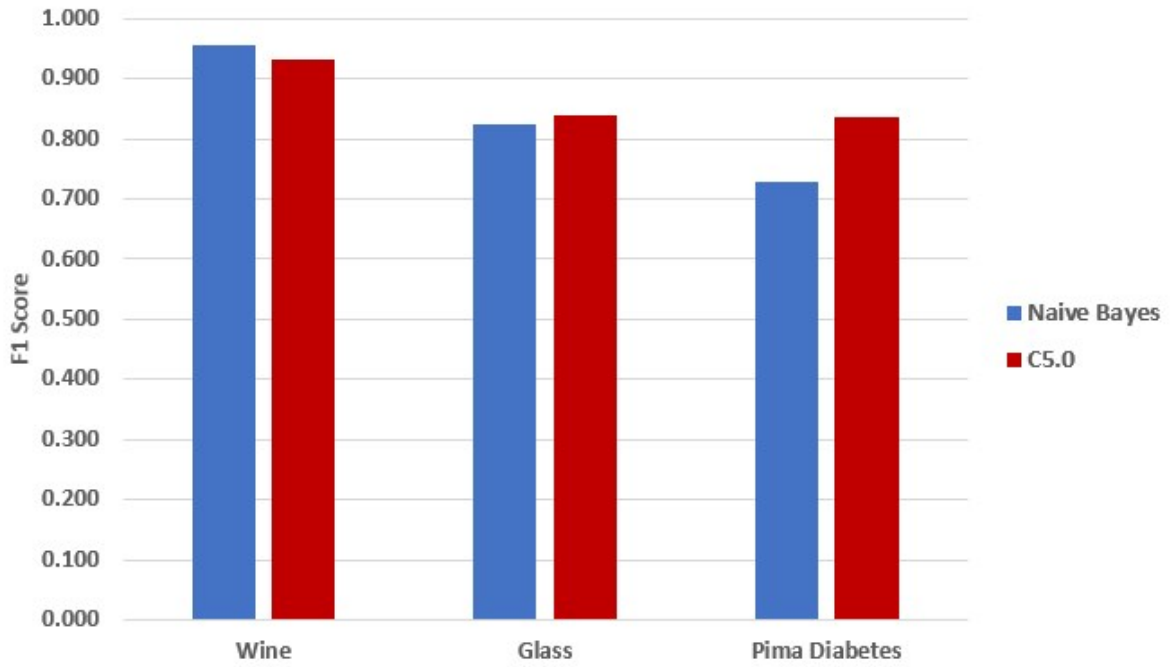


Rys. 3.12: Przedstawienie najlepszych uzyskanych wyników dla badanych zbiorów

Tab. 3.11: Przedstawienie najlepszych uzyskanych wyników dla badanych zbiorów

	Wine	**Glass**	**Pima Diabetes**
Accuracy	0.946	0.897	0.702
Precision	0.935	0.834	0.770
Recall	0.929	0.852	0.912
F1 Score	0.931	0.840	0.835
Tree Size	6	13	7

3.4. Porównanie z Naiwnym Bayesem



Rys. 3.13: Porównanie najlepszych wyników C5.0 z najlepszymi uzyskanymi wynikami algorytmu Naiwnego Bayesa

Tab. 3.12: Porównanie najlepszych wyników C5.0 z najlepszymi uzyskanymi wynikami algorytmu Naiwnego Bayesa

	Wine	**Glass**	**Pima Diabetes**
Naive Bayes	0.957	0.825	0.730
C5.0	0.931	0.840	0.835

Spis rysunków

Spis tabel