

Indukcyjne metody analizy danych

Ćwiczenie 2

Indukcja drzew decyzyjnych C4.5 w R

Prowadzący: dr inż. Paweł Myszkowski

Student: Piotr Bielak, 218137

WT 17:05

Wrocław, 10 kwietnia 2018r.

Spis treści

| | | |
|----------|---|-----------|
| 1 | Wprowadzenie | 3 |
| 1.1 | Cel ćwiczenia | 3 |
| 1.2 | Algorytm C4.5 | 3 |
| 2 | Analiza zbiorów danych | 5 |
| 2.1 | Zbiór danych – "Diabetes" | 5 |
| 2.2 | Zbiór danych – "Glass" | 6 |
| 2.3 | Zbiór danych – "Wine" | 8 |
| 3 | Eksperyment | 10 |
| 3.1 | Założenia | 10 |
| 3.2 | Badanie parametrów algorytmu C4.5 | 10 |
| 3.3 | Wyniki krosvalidacji | 14 |
| 3.3.1 | Zbiór danych – "Diabetes" | 14 |
| 3.3.2 | Zbiór danych – "Glass" | 18 |
| 3.3.3 | Zbiór danych – "Wine" | 22 |
| 4 | Wnioski | 26 |

1 Wprowadzenie

1.1 Cel ćwiczenia

Celem ćwiczenia było zapoznanie się z algorytmem C4.5, służącym do budowy drzew decyzyjnych. Należało również zbadać i ocenić jego działanie na 3 określonych zbiorach danych. W trakcie badań należało uwzględnić różne parametry samego algorytmu oraz metody krosvalidacji, a następnie zaobserwować wpływ tych parametrów na wartości zadanych metryk. Ostatnim krokiem było porównanie działania algorytmu z klasyfikatorem naiwnego Bayesa.

1.2 Algorytm C4.5

Drzewo decyzyjne to klasyfikator, który dzieli dane rekurencyjnie na podzbiory za pomocą określonych reguł (węzłów decyzyjnych). Należy ono do grupy algorytmów nadzorowanego uczenia maszynowego (*supervised learning*) i może być używane zarówno dla danych dyskretnych, jak i ciągłych w celach klasyfikacyjnych i regresyjnych. Najpopularniejsze algorytmy budowy drzew decyzyjnych to: ID3, C4.5, C5.0, CART i wiele innych. W ramach tego ćwiczenia omówiony i zbadany zostanie algorytm C4.5, który stanowi rozszerzenie podstawowego algorytmu ID3. Podstawowe różnice między tymi algorytmami to:

- ID3 radzi sobie tylko z danymi kategorycznymi, natomiast C4.5 obsługuje również dane ciągłe,
- C4.5 radzi sobie z brakującymi danymi,
- C4.5 używa algorytmów przycinania drzewa (*error based pruning*).

Wspólnymi cechami obu algorytmów są: podatność na wartości odstające (*outliers*) oraz kryterium używane podczas podziału zbioru danych w węzłach drzewa (zysk informacyjny).

Zysk informacyjny i entropia

Entropia jest miarą określającą nieuporządkowanie danych i dla zmiennej losowej X o wartościach x_1, x_2, \dots, x_n określona jest wzorem:

$$E(X) = - \sum_{i=1}^n p(x_i) \cdot \log_2 p(x_i),$$

korzystając z tej definicji można określić tzw. **zysk informacyjny**:

$$IG(X, A) = E(X) - \sum_{i \in \text{values}(A)} \frac{|\{x \in X \mid \text{value}(x, A) = i\}|}{|X|} E(\{x \in X \mid \text{value}(x, A) = i\})$$

Lista kroków algorytmu C4.5

Dla określonego zbioru danych D , algorytm C4.5 jest zdefiniowany w następujący sposób:

- K.1. $Tree = \{\}$
- K.2. Jeśli osiągnięto warunek końcowy, to zakończ algorytm.
- K.3. Dla każdego atrybutu w zbiorze danych oblicz zysk informacyjny.
- K.4. a_{best} = wybierz najlepszy atrybut względem obliczonych w K.3. wartości.
- K.5. Dołącz do drzewa $Tree$ węzeł decyzyjny dla atrybutu a_{best} .
- K.6. D_v = podzbiory wynikające z podziału zbioru D za pomocą wartości atrybutu a_{best} .
- K.7. Dla każdego $d \in D_v$ wykonuj:
 - K.7.1. $Tree_v = C4.5(d)$
 - K.7.2. Dołącz $Tree_v$ do odpowiedniej gałęzi w węźle decyzyjnym.
- K.8. Zwróć $Tree$.

Metody przycinania drzewa (prunning)

W celu uniknięcia przeuczenia (*overfitting*) oraz poprawienia jakości generalizacji drzewa, stosuje się metody tzw. przycinania (*prunning*). Można je podzielić na dwie grupy:

- na etapie budowy drzewa (*pre-prunning*)
- po zakończeniu procesu budowy drzewa (*post-prunning*)

W przypadku metod *pre-prunning* istnieje możliwość przedwczesnego zatrzymania algorytmu budowy drzewa i znacznego pogorszenia wydajności drzewa (w skrajnych przypadkach korzeń może zostać w ogóle nie rozwinięty). Stąd też preferowane są metody *post-prunning*. Najpopularniejszymi algorytmami tutaj są:

- zastępowanie poddrzew (*subtree replacement*) – wybrane poddrzewo jest zastępowane pewną wartością znajdującą się w nim; ważne jest jednak, że rozważane są wszystkie poddrzewa w ramach danego poddrzewa i dopiero wtedy podejmowana jest decyzja,
- wznoszenie poddrzew (*subtree raising*) – określony węzeł jest usuwany i jeden z jego potomków (również węzeł decyzyjny) jest umieszczany na jego miejscu; Wszystkie instancje umieszczone w ramach poddrzewa wyznaczonego przez usuwany węzeł, są ponownie rozmieszczane w wynikowym poddrzewie,
- *reduced error prunning* – rozpoczynając od liści, każdy węzeł jest zastępowany najczęściej występującą klasą w ramach tego węzła; jeśli nie pogorszy to błędu klasyfikacji (accuracy), to zmiana jest zachowywana.

2 Analiza zbiorów danych

2.1 Zbiór danych – "Diabetes"

| Nazwa klasy | Liczba instancji | % instancji |
|-------------|------------------|-------------|
| 1 (chory) | 500 | 65 (%) |
| 0 (zdrowy) | 268 | 35 (%) |

Tabela 1: Udział procentowy klas w zbiorze "Diabetes".

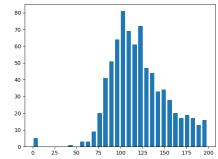
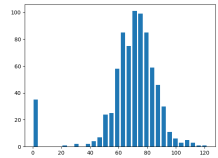
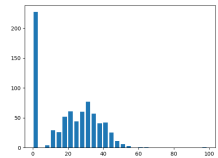
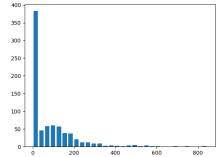
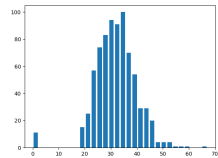
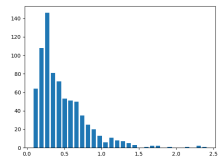
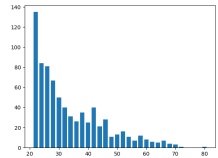
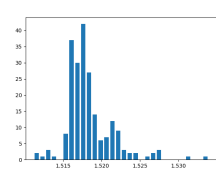
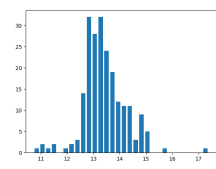
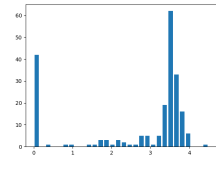
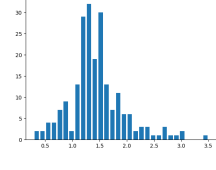
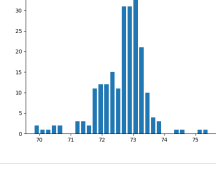
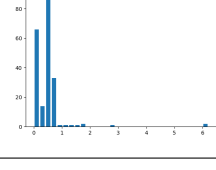
| Nazwa atrybutu | Min | Max | Średnia | Ochyl. stand. | Rozkład |
|--------------------------|------|------|---------|---------------|---|
| Glucose | 0 | 199 | 120.89 | 31.95 |  |
| BloodPressure | 0 | 122 | 69.11 | 19.34 |  |
| SkinThickness | 0 | 99 | 20.54 | 15.94 |  |
| Insulin | 0 | 846 | 79.80 | 115.17 |  |
| BMI | 0 | 67.1 | 31.99 | 7.88 |  |
| DiabetesPedigreeFunction | 0.08 | 2.42 | 0.47 | 0.33 |  |
| Age | 21 | 81 | 33.24 | 11.75 |  |

Tabela 2: Atrybuty zbioru danych "Diabetes".

2.2 Zbiór danych – "Glass"

| Nazwa klasy | Liczba instancji | % instancji |
|--|------------------|-------------|
| 1 (building_windows_float_processed) | 70 | 33 (%) |
| 2 (building_windows_non_float_processed) | 76 | 36 (%) |
| 3 (vehicle_windows_float_processed) | 17 | 8 (%) |
| 4 (vehicle_windows_non_float_processed) | 0 | 0 (%) |
| 5 (containers) | 13 | 6 (%) |
| 6 (tableware) | 9 | 4 (%) |
| 7 (headlamps) | 29 | 13 (%) |

Tabela 3: Udział procentowy klas w zbiorze "Glass".

| Name | Min | Max | Mean | Std | Distribution |
|------|-------|-------|-------|------|--|
| RI | 1.51 | 1.53 | 1.52 | 0.00 |  |
| Na | 10.73 | 17.38 | 13.41 | 0.81 |  |
| Mg | 0.00 | 4.49 | 2.68 | 1.44 |  |
| Al | 0.29 | 3.50 | 1.44 | 0.50 |  |
| Si | 69.81 | 75.41 | 72.65 | 0.77 |  |
| K | 0.00 | 6.21 | 0.50 | 0.65 |  |

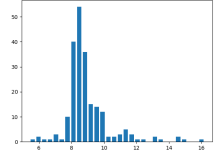
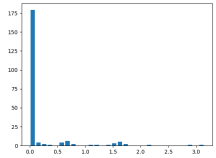
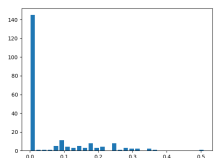
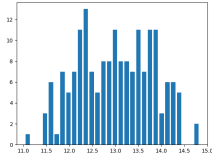
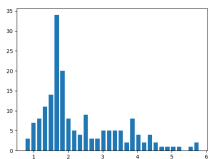
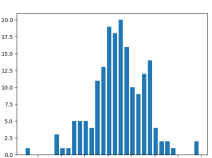
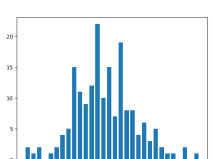
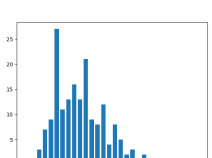
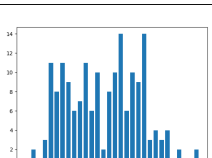
| | | | | | |
|----|------|-------|------|------|--|
| Ca | 5.43 | 16.19 | 8.96 | 1.42 |  |
| Ba | 0.00 | 3.15 | 0.18 | 0.50 |  |
| Fe | 0.00 | 0.51 | 0.06 | 0.10 |  |

Tabela 4: Atrybuty zbioru danych "Glass".

2.3 Zbiór danych – "Wine"

| Nazwa klasy | Liczba instancji | % instancji |
|-------------|------------------|-------------|
| 1 (Class 1) | 59 | 33 (%) |
| 2 (Class 2) | 71 | 40 (%) |
| 3 (Class 3) | 48 | 27 (%) |

Tabela 5: Udział procentowy klas w zbiorze "Wine".

| Name | Min | Max | Mean | Std | Distribution |
|-------------------|-------|--------|-------|-------|---|
| Alcohol | 11.03 | 14.83 | 13.00 | 0.81 |  |
| Macil_acid | 0.74 | 5.80 | 2.34 | 1.11 |  |
| Ash | 1.36 | 3.23 | 2.37 | 0.27 |  |
| Alcalinity_of_ash | 10.60 | 30.00 | 19.49 | 3.33 |  |
| Magnesium | 70.00 | 162.00 | 99.74 | 14.24 |  |
| Total_phenols | 0.98 | 3.88 | 2.30 | 0.62 |  |

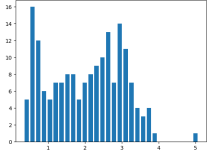
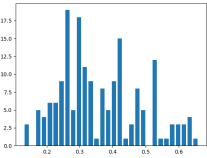
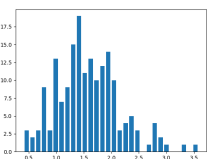
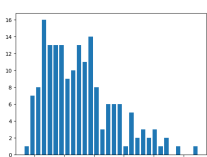
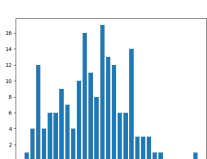
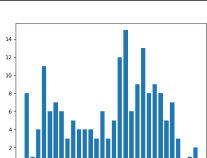
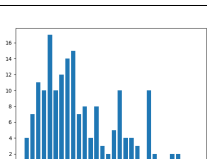
| | | | | | |
|----------------------|--------|---------|--------|--------|---|
| Flavanoids | 0.34 | 5.08 | 2.03 | 1.00 |  |
| Nonflavanoid_phenols | 0.13 | 0.66 | 0.36 | 0.12 |  |
| Proanthocyanins | 0.41 | 3.58 | 1.59 | 0.57 |  |
| Intensity | 1.28 | 13.00 | 5.06 | 2.31 |  |
| Hue | 0.48 | 1.71 | 0.96 | 0.23 |  |
| OD280_OD315 | 1.27 | 4.00 | 2.61 | 0.71 |  |
| Proline | 278.00 | 1680.00 | 746.89 | 314.02 |  |

Tabela 6: Atrybuty zbioru danych "Wine".

3 Eksperyment

3.1 Założenia

Eksperyment został podzielony na dwie fazy. Pierwsza służyła do zbadania parametrów algorytmu C4.5, natomiast druga miała na celu ocenę jakości działania drzewa decyzyjnego dla wybranych zbiorów danych (**Diabetes**, **Glass** oraz **Wine**). Podobnie jak w przypadku algorytmu klasyfikatora Bayesa została tutaj również zastosowana krowalidacja zwykła oraz stratyfikowana i zostały obliczone miary *accuracy*, *precision*, *recall* oraz *F1*.

Szczegółowe wyniki (wykresy, tabelki, wizualizacje drzew) tego eksperymentu są przedstawione w kolejnych podrozdziałach.

3.2 Badanie parametrów algorytmu C4.5

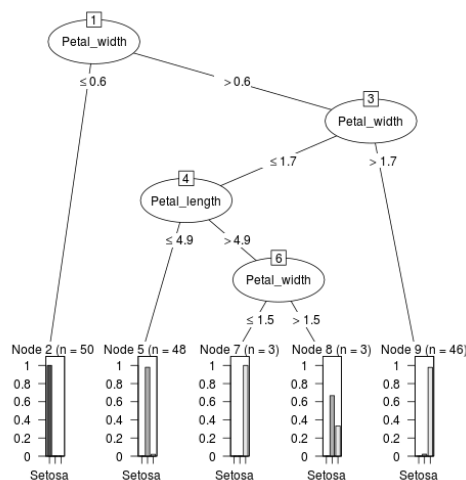
W implementacji ćwiczenia wykorzystano bibliotekę *rWeka*. Spośród dostępnych tutaj parametrów algorytmu C4.5 zostały wybrane i zbadane następujące:

| Nazwa parametru | Wybrane wartości | Opis |
|-----------------------------------|----------------------|--|
| Reduced error pruning | RE = { TRUE, FALSE } | czy przeprowadzać przycinanie drzewa metodą "reduced error" |
| Number of folds for RE pruning | NBF = { 2, 10 } | liczba podziałów danych (podzbiorów) używanych podczas przycinania "reduced error" |
| Min. number of instances per leaf | NBINST = { 1, 10 } | min. liczba instancji w liściu |
| Confidence threshold for pruning | CONF = { 0.01, 0.4 } | próg ufności dla przycinania drzewa |

Tabela 7: Zbadane parametry algorytmu C4.5.

Dodatkowo zostało zbadane zachowanie drzewa dla domyślnych wartości parametrów (ustalonych przez autorów biblioteki *rWeka*). Poniższe rysunki obrazują drzewa decyzyjne dla zbioru "Iris" przy zastosowaniu powyższych opcji konfiguracyjnych.

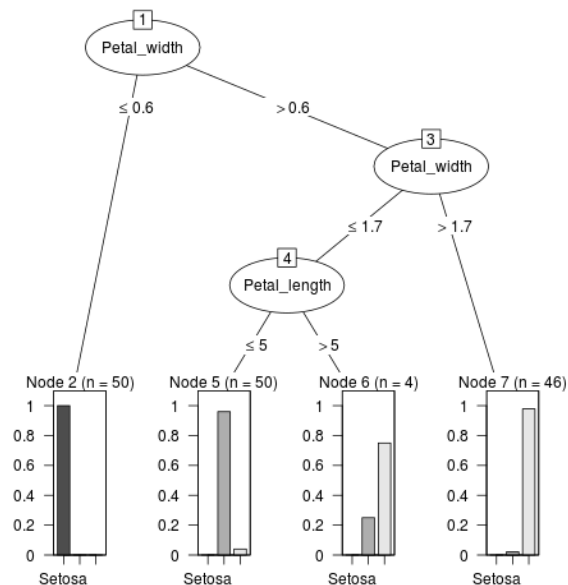
Domyślne opcje konfiguracyjne



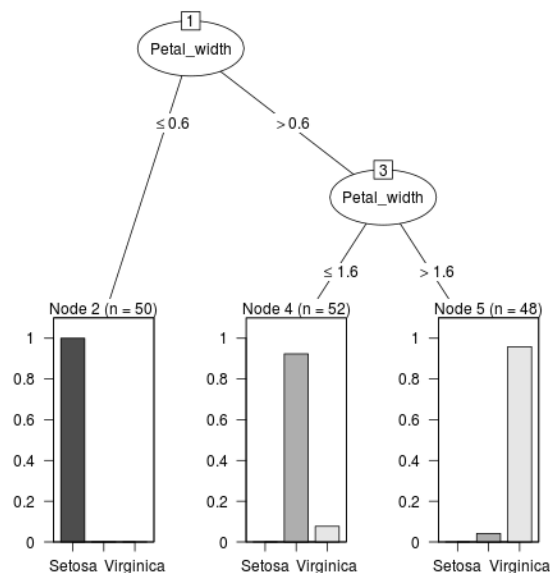
Rysunek 1: Drzewo dla domyślnej konfiguracji.

Przycinanie "Reduced error"

Zastosowanie przycinania drzewa metodą *reduced error* pozwoliło zmniejszyć głębokość otrzymanego drzewa. Dodatkowo można zauważyć, że zastosowanie większej liczby podziałów zbioru danych (foldy) pozwalało zgeneralizować drzewo do 2 reguł.



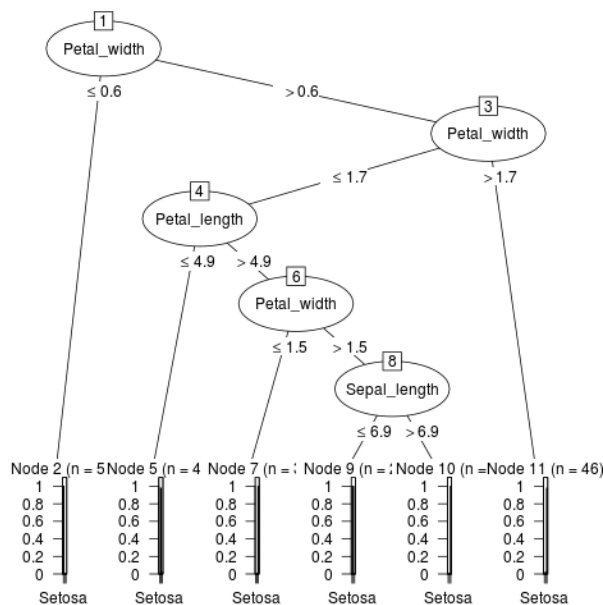
Rysunek 2: Drzewo dla RE = TRUE oraz NBF = 2.



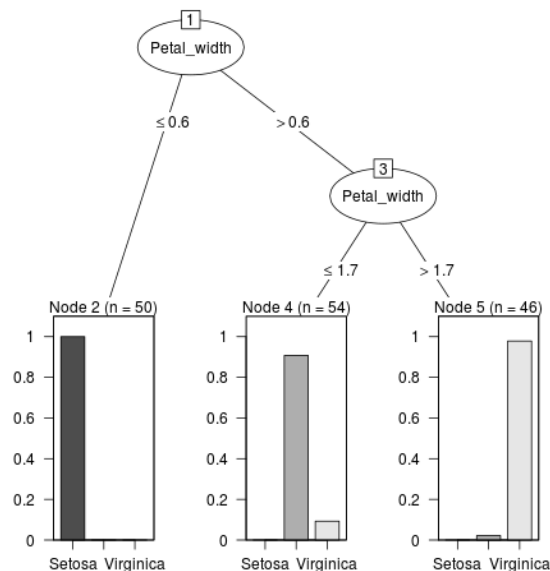
Rysunek 3: Drzewo dla RE = TRUE oraz NBF = 10.

Min. liczba instancji w liściu

Parametr określający minimalną liczbę instancji w liściu drzewa decyzyjnego znacząco wpływa na odporność drzewa na przeuczenie (*overfitting*). W przypadku ustalenia tego parametru na wartość równą jeden, ryzyko przeuczenia jest bardzo wysokie, dodatkowo można zauważyć, że głębokość drzewa wzrosła (głębokość równa 5) i jest znacznie większa niż w przypadku ustalenia parametru na wartość 10 (głębokość 2).



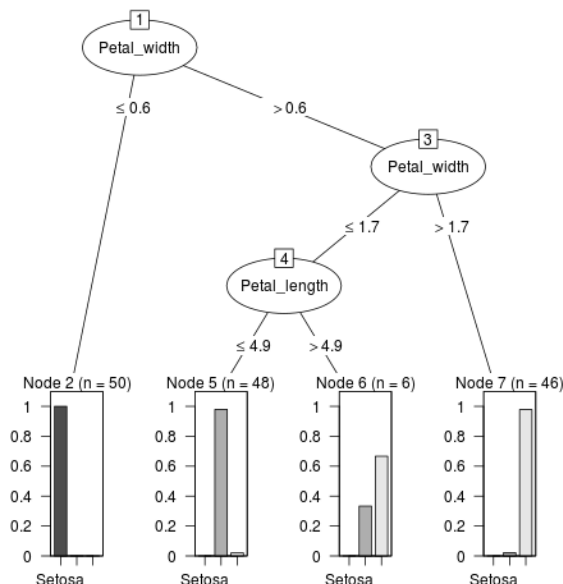
Rysunek 4: Drzewo dla NBINST = 1.



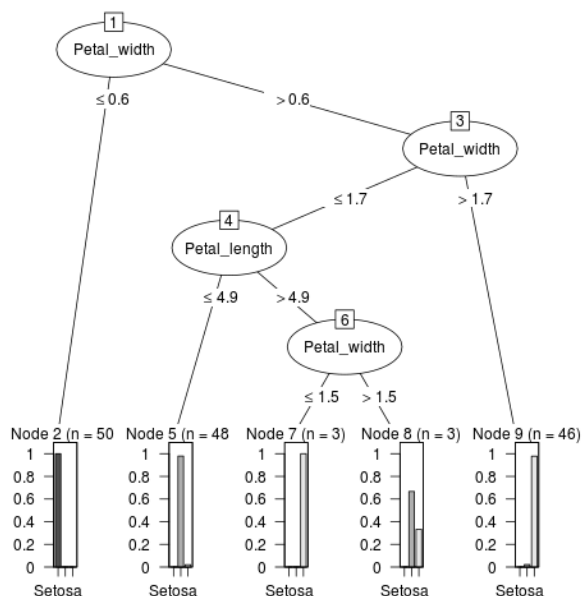
Rysunek 5: Drzewo dla NBINST = 10.

Próg ufności

Zastosowanie wbudowanej metody przycinania (zamiast *reduced error*), również pozwoliło ograniczyć głębokość drzewa, jednak efekty nie są tak dobre jak w przypadku tamtej metody. Dla parametru progu ufności równego 0.4 otrzymano drzewo identyczne jak w przypadku zastosowania domyślnych parametrów algorytmu. Natomiast zastosowanie bardzo niskiego progu ufności (0.01) pozwalało ograniczyć głębokość drzewa o jeden (prawie identyczny z drzewem otrzymanym dla metody *reduced error* z 2 podziałami).



Rysunek 6: Drzewo dla $\text{CONF} = 0.01$.



Rysunek 7: Drzewo dla $\text{CONF} = 0.4$.

3.3 Wyniki krosvalidacji

Poniżej zostały przedstawione wyniki zastosowania krosvalidacji dla wybranych zbiorów danych. W ramach danego procesu krosvalidacji, wyznaczono wartości miar oceny jakości klasyfikatora. Dodatkowo zostały zamieszczone tabelki z dokładnymi wartościami tych miar. Parametrami każdego procesu krosvalidacji są:

- liczba podzbiorów (foldów), zmieniająca się od 2 do 9 ze skokiem 1,
- zestaw opcji konfiguracyjnych:
 - (C1) domyślna konfiguracja,
 - (C2) RE = TRUE, NBF = 10, NBINST = 10,
 - (C3) CONF = 0.01, NBINST = 10.

3.3.1 Zbiór danych – "Diabetes"

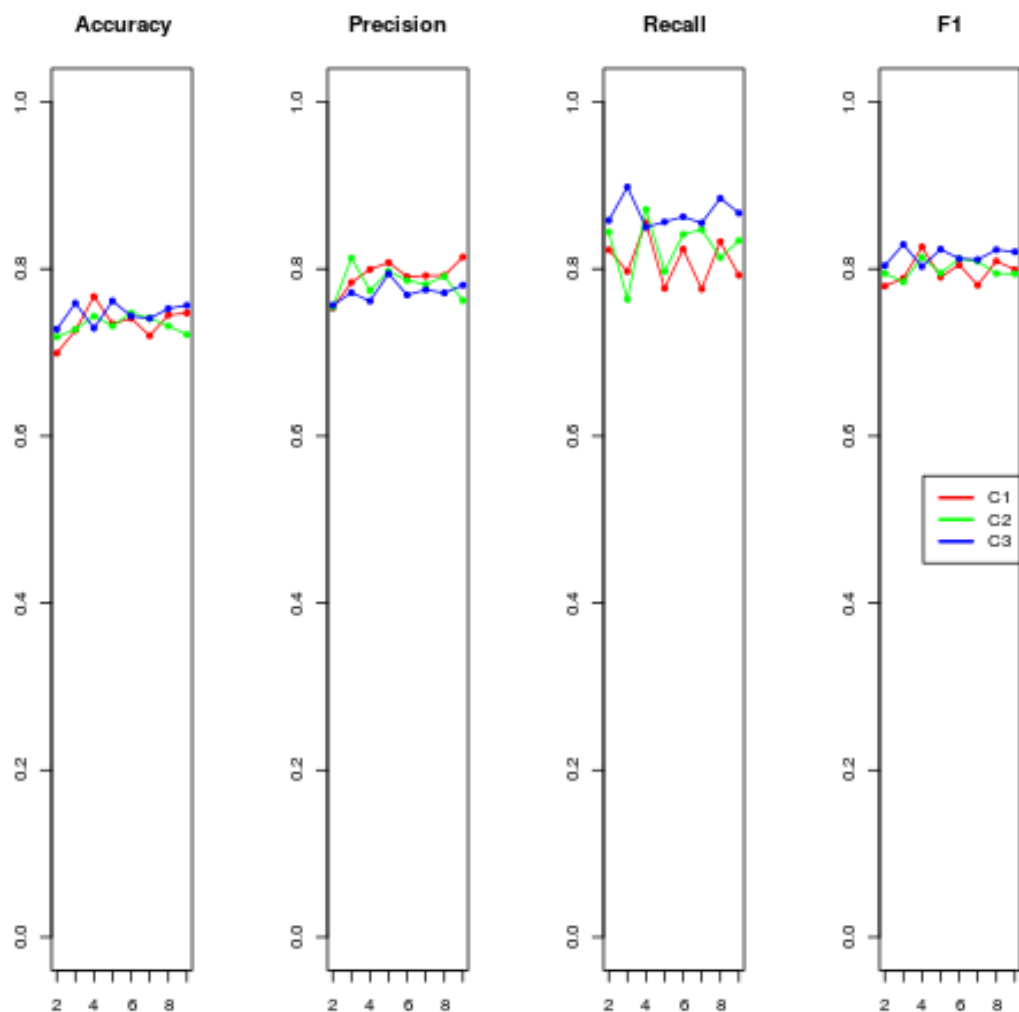
Krosvalidacja zwykła

| Klasyfikator | Accuracy | Precision | Recall | F1 | Komentarz |
|--------------|----------|-----------|--------|------|---------------------------|
| Bayes | 0.76 | 0.67 | 0.60 | 0.63 | Brak dyskretyzacji; K = 9 |
| C4.5 | 0.76 | 0.77 | 0.90 | 0.83 | C3; K = 3 |

Tabela 8: Najlepsze wyniki dla klas. Bayesowskiego i drzewa C4.5 (względem F1).



Rysunek 8: Wartości metryk dla klasyfikatora Bayesowskiego.

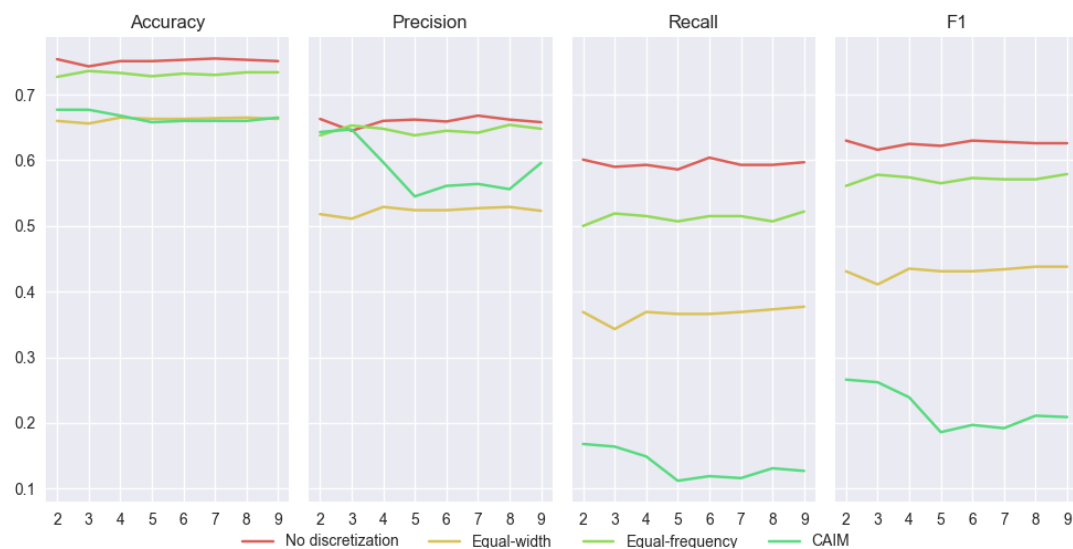


Rysunek 9: Wartości metryk dla drzewa decyzyjnego C4.5.

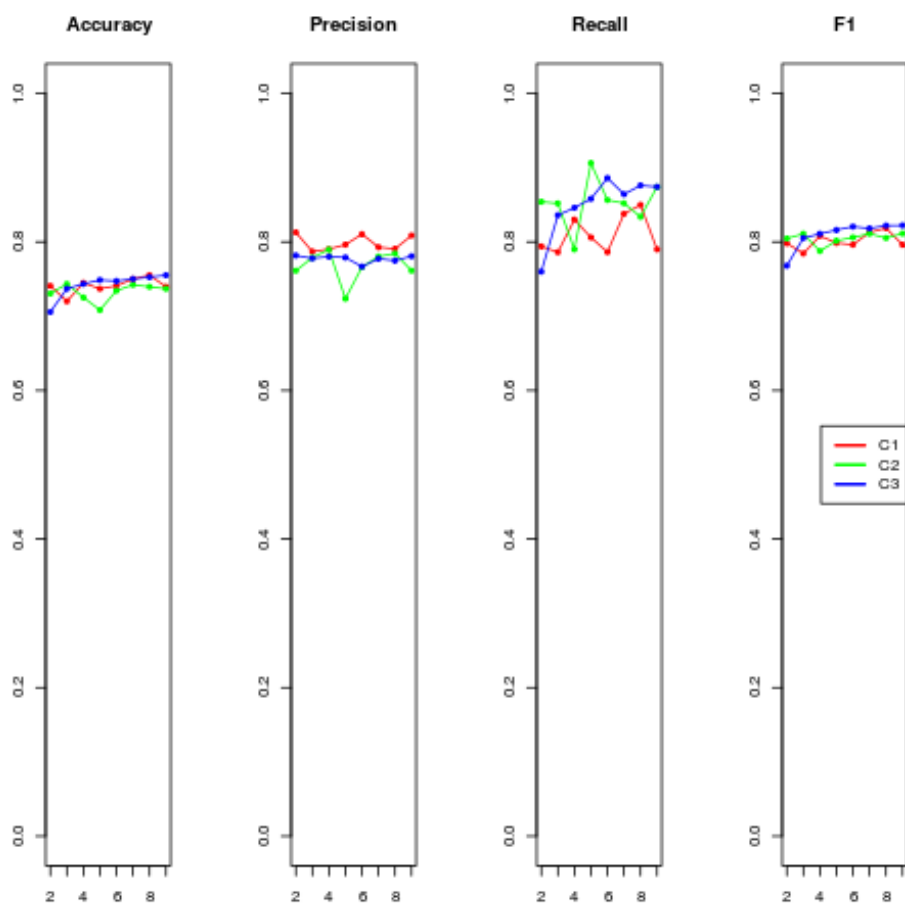
| Konfiguracja | Metryka | Liczba foldów | | | | | | | |
|--------------|-----------|---------------|------|------|------|------|------|------|------|
| | | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| C1 | Accuracy | 0.70 | 0.73 | 0.77 | 0.73 | 0.74 | 0.72 | 0.74 | 0.75 |
| C1 | Precision | 0.75 | 0.78 | 0.80 | 0.81 | 0.79 | 0.79 | 0.79 | 0.81 |
| C1 | Recall | 0.82 | 0.80 | 0.86 | 0.78 | 0.82 | 0.78 | 0.83 | 0.79 |
| C1 | F1 | 0.78 | 0.79 | 0.83 | 0.79 | 0.80 | 0.78 | 0.81 | 0.80 |
| C2 | Accuracy | 0.72 | 0.73 | 0.74 | 0.73 | 0.75 | 0.74 | 0.73 | 0.72 |
| C2 | Precision | 0.75 | 0.81 | 0.77 | 0.80 | 0.79 | 0.78 | 0.79 | 0.76 |
| C2 | Recall | 0.84 | 0.76 | 0.87 | 0.80 | 0.84 | 0.85 | 0.81 | 0.83 |
| C2 | F1 | 0.79 | 0.78 | 0.81 | 0.80 | 0.81 | 0.81 | 0.79 | 0.79 |
| C3 | Accuracy | 0.73 | 0.76 | 0.73 | 0.76 | 0.74 | 0.74 | 0.75 | 0.76 |
| C3 | Precision | 0.76 | 0.77 | 0.76 | 0.79 | 0.77 | 0.78 | 0.77 | 0.78 |
| C3 | Recall | 0.86 | 0.90 | 0.85 | 0.86 | 0.86 | 0.86 | 0.88 | 0.87 |
| C3 | F1 | 0.80 | 0.83 | 0.80 | 0.82 | 0.81 | 0.81 | 0.82 | 0.82 |

Tabela 9: Dokładne wartości metryk dla drzewa decyzyjnego C4.5.

Krowalidacja stratyfikowana



Rysunek 10: Wartości metryk dla klasyfikatora Bayesowskiego.



Rysunek 11: Wartości metryk dla drzewa decyzyjnego C4.5.

| Klasyfikator | Accuracy | Precision | Recall | F1 | Komentarz |
|--------------|----------|-----------|--------|------|---------------------------|
| Bayes | 0.75 | 0.66 | 0.60 | 0.63 | Brak dyskretyzacji; K = 6 |
| C4.5 | 0.75 | 0.77 | 0.89 | 0.82 | C3; K = 6 |

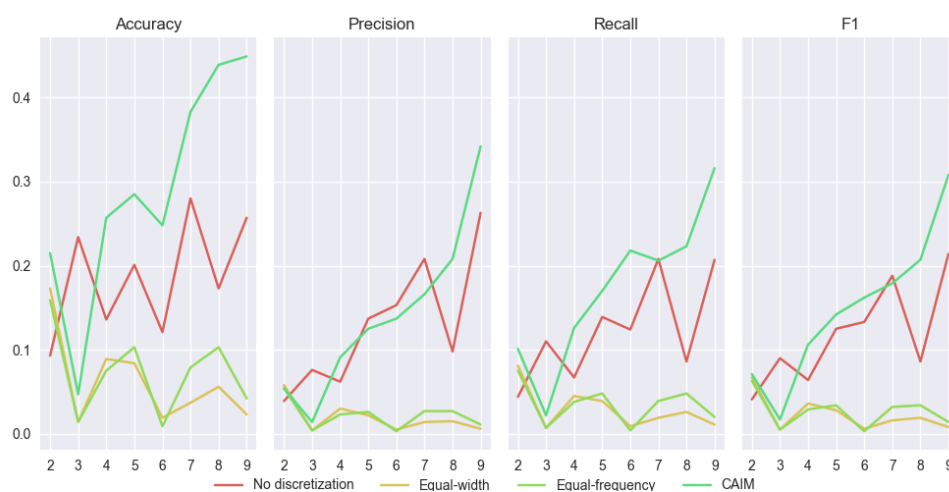
Tabela 10: Najlepsze wyniki dla klas. Bayesowskiego i drzewa C4.5 (względem F1).

| | | Liczba foldów | | | | | | | |
|--------------|-----------|---------------|------|------|------|------|------|------|------|
| Konfiguracja | Metryka | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| C1 | Accuracy | 0.74 | 0.72 | 0.74 | 0.74 | 0.74 | 0.75 | 0.76 | 0.74 |
| C1 | Precision | 0.81 | 0.79 | 0.79 | 0.80 | 0.81 | 0.79 | 0.79 | 0.81 |
| C1 | Recall | 0.79 | 0.79 | 0.83 | 0.81 | 0.79 | 0.84 | 0.85 | 0.79 |
| C1 | F1 | 0.80 | 0.78 | 0.81 | 0.80 | 0.80 | 0.81 | 0.82 | 0.80 |
| C2 | Accuracy | 0.73 | 0.74 | 0.73 | 0.71 | 0.73 | 0.74 | 0.74 | 0.74 |
| C2 | Precision | 0.76 | 0.78 | 0.79 | 0.72 | 0.77 | 0.78 | 0.78 | 0.76 |
| C2 | Recall | 0.85 | 0.85 | 0.79 | 0.91 | 0.86 | 0.85 | 0.83 | 0.87 |
| C2 | F1 | 0.80 | 0.81 | 0.79 | 0.80 | 0.81 | 0.81 | 0.81 | 0.81 |
| C3 | Accuracy | 0.71 | 0.74 | 0.74 | 0.75 | 0.75 | 0.75 | 0.75 | 0.76 |
| C3 | Precision | 0.78 | 0.78 | 0.78 | 0.78 | 0.77 | 0.78 | 0.77 | 0.78 |
| C3 | Recall | 0.76 | 0.84 | 0.85 | 0.86 | 0.89 | 0.86 | 0.88 | 0.87 |
| C3 | F1 | 0.77 | 0.81 | 0.81 | 0.82 | 0.82 | 0.82 | 0.82 | 0.82 |

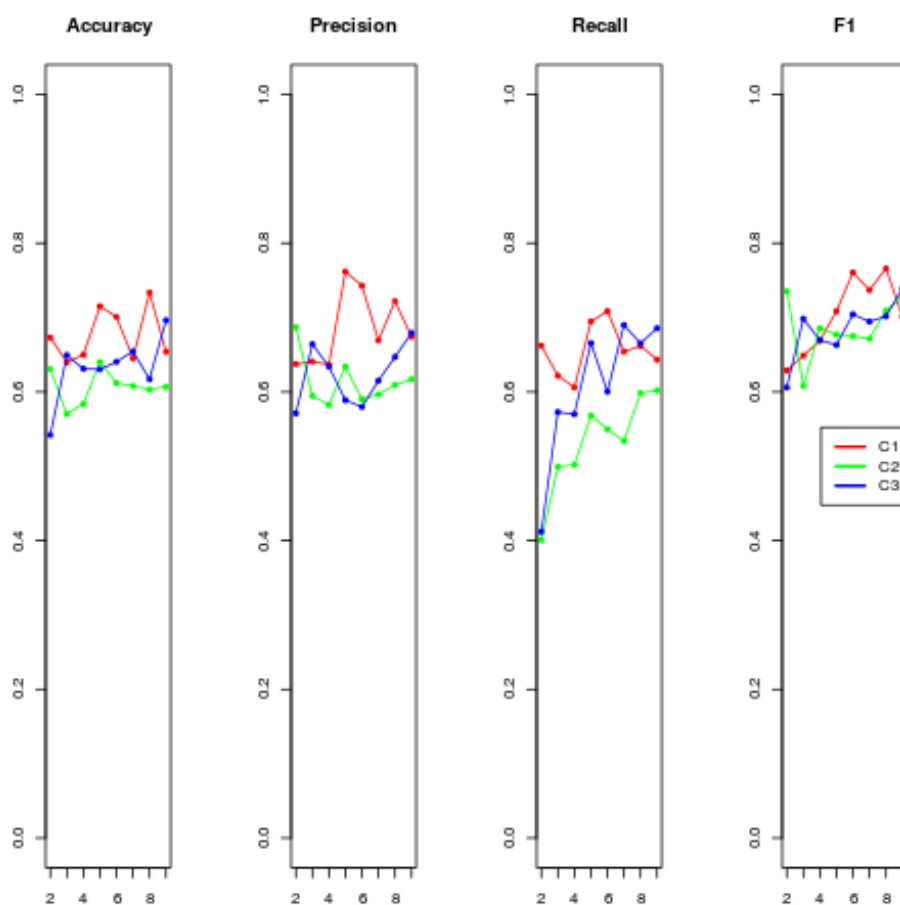
Tabela 11: Dokładne wartości metryk dla drzewa decyzyjnego C4.5.

3.3.2 Zbiór danych – "Glass"

Kroswalidacja zwykła



Rysunek 12: Wartości metryk dla klasyfikatora Bayesowskiego.



Rysunek 13: Wartości metryk dla drzewa decyzyjnego C4.5.

| Klasyfikator | Accuracy | Precision | Recall | F1 | Komentarz |
|--------------|----------|-----------|--------|------|-------------|
| Bayes | 0.44 | 0.34 | 0.32 | 0.31 | CAIM; K = 9 |
| C4.5 | 0.73 | 0.72 | 0.66 | 0.77 | C1; K = 8 |

Tabela 12: Najlepsze wyniki dla klas. Bayesowskiego i drzewa C4.5 (względem F1).

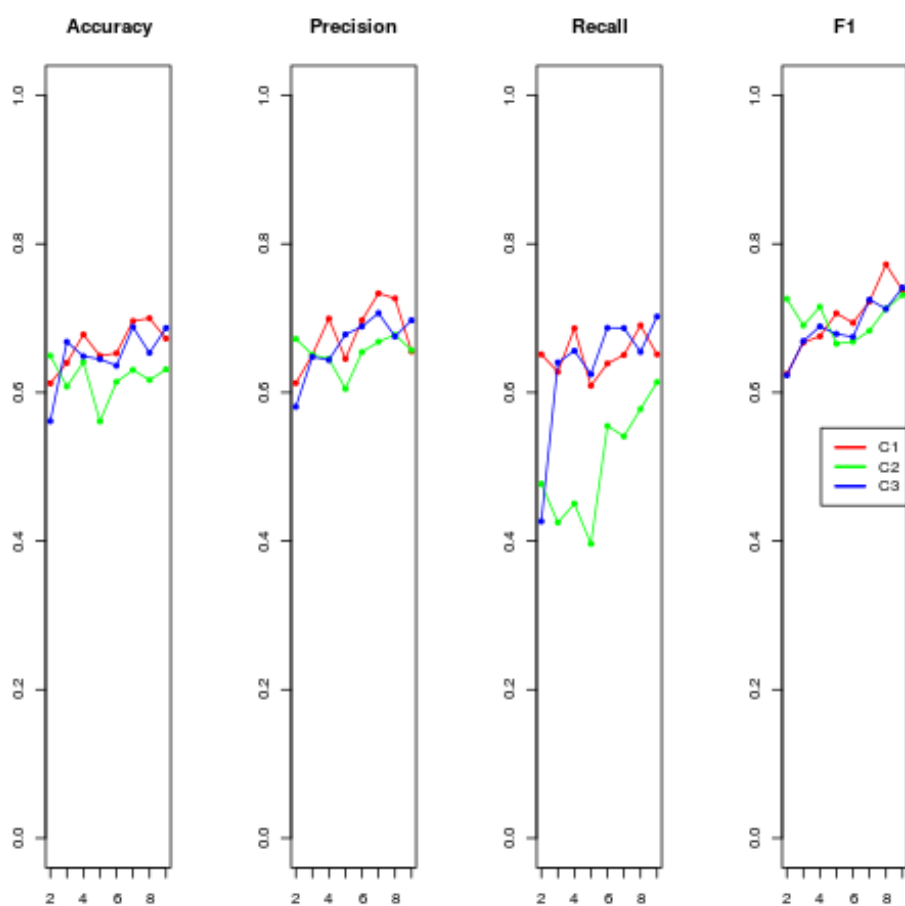
| | | Liczba foldów | | | | | | | |
|--------------|-----------|---------------|------|------|------|------|------|------|------|
| Konfiguracja | Metryka | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| C1 | Accuracy | 0.67 | 0.64 | 0.65 | 0.72 | 0.70 | 0.65 | 0.73 | 0.65 |
| C1 | Precision | 0.64 | 0.64 | 0.64 | 0.76 | 0.74 | 0.67 | 0.72 | 0.67 |
| C1 | Recall | 0.66 | 0.62 | 0.61 | 0.69 | 0.71 | 0.65 | 0.66 | 0.64 |
| C1 | F1 | 0.63 | 0.65 | 0.67 | 0.71 | 0.76 | 0.74 | 0.77 | 0.70 |
| C2 | Accuracy | 0.63 | 0.57 | 0.58 | 0.64 | 0.61 | 0.61 | 0.60 | 0.61 |
| C2 | Precision | 0.69 | 0.59 | 0.58 | 0.63 | 0.59 | 0.60 | 0.61 | 0.62 |
| C2 | Recall | 0.40 | 0.50 | 0.50 | 0.57 | 0.55 | 0.53 | 0.60 | 0.60 |
| C2 | F1 | 0.74 | 0.61 | 0.69 | 0.68 | 0.68 | 0.67 | 0.71 | 0.73 |
| C3 | Accuracy | 0.54 | 0.65 | 0.63 | 0.63 | 0.64 | 0.65 | 0.62 | 0.70 |
| C3 | Precision | 0.57 | 0.66 | 0.63 | 0.59 | 0.58 | 0.62 | 0.65 | 0.68 |
| C3 | Recall | 0.41 | 0.57 | 0.57 | 0.67 | 0.60 | 0.69 | 0.67 | 0.69 |
| C3 | F1 | 0.61 | 0.70 | 0.67 | 0.66 | 0.70 | 0.69 | 0.70 | 0.74 |

Tabela 13: Dokładne wartości metryk dla drzewa decyzyjnego C4.5.

Krowalidacja stratyfikowana



Rysunek 14: Wartości metryk dla klasyfikatora Bayesowskiego.



Rysunek 15: Wartości metryk dla drzewa decyzyjnego C4.5.

| Klasyfikator | Accuracy | Precision | Recall | F1 | Komentarz |
|--------------|----------|-----------|--------|------|-------------|
| Bayes | 0.67 | 0.60 | 0.63 | 0.61 | CAIM; K = 5 |
| C4.5 | 0.70 | 0.73 | 0.69 | 0.77 | C1; K = 8 |

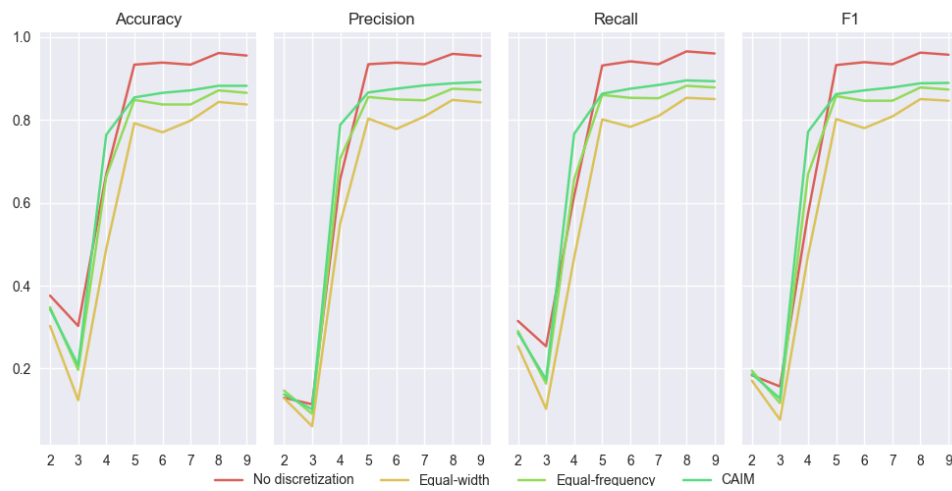
Tabela 14: Najlepsze wyniki dla klas. Bayesowskiego i drzewa C4.5 (względem F1).

| | | Liczba foldów | | | | | | | |
|--------------|-----------|---------------|------|------|------|------|------|------|------|
| Konfiguracja | Metryka | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| C1 | Accuracy | 0.61 | 0.64 | 0.68 | 0.65 | 0.65 | 0.70 | 0.70 | 0.67 |
| C1 | Precision | 0.61 | 0.65 | 0.70 | 0.65 | 0.70 | 0.73 | 0.73 | 0.66 |
| C1 | Recall | 0.65 | 0.63 | 0.69 | 0.61 | 0.64 | 0.65 | 0.69 | 0.65 |
| C1 | F1 | 0.62 | 0.67 | 0.68 | 0.71 | 0.69 | 0.72 | 0.77 | 0.74 |
| C2 | Accuracy | 0.65 | 0.61 | 0.64 | 0.56 | 0.61 | 0.63 | 0.62 | 0.63 |
| C2 | Precision | 0.67 | 0.65 | 0.65 | 0.60 | 0.65 | 0.67 | 0.68 | 0.66 |
| C2 | Recall | 0.48 | 0.42 | 0.45 | 0.40 | 0.55 | 0.54 | 0.58 | 0.61 |
| C2 | F1 | 0.73 | 0.69 | 0.72 | 0.67 | 0.67 | 0.68 | 0.71 | 0.73 |
| C3 | Accuracy | 0.56 | 0.67 | 0.65 | 0.64 | 0.64 | 0.69 | 0.65 | 0.69 |
| C3 | Precision | 0.58 | 0.65 | 0.64 | 0.68 | 0.69 | 0.71 | 0.68 | 0.70 |
| C3 | Recall | 0.43 | 0.64 | 0.66 | 0.62 | 0.69 | 0.69 | 0.65 | 0.70 |
| C3 | F1 | 0.62 | 0.67 | 0.69 | 0.68 | 0.67 | 0.72 | 0.71 | 0.74 |

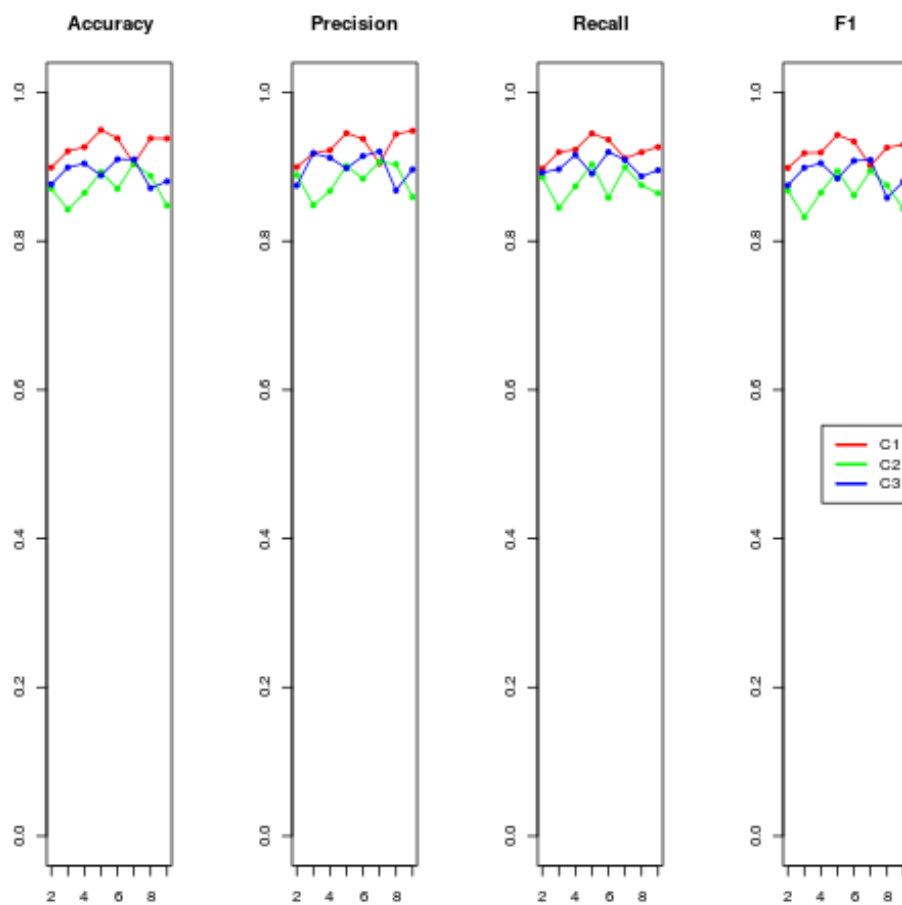
Tabela 15: Dokładne wartości metryk dla drzewa decyzyjnego C4.5.

3.3.3 Zbiór danych – "Wine"

Kroswalidacja zwykła



Rysunek 16: Wartości metryk dla klasyfikatora Bayesowskiego.



Rysunek 17: Wartości metryk dla drzewa decyzyjnego C4.5.

| Klasyfikator | Accuracy | Precision | Recall | F1 | Komentarz |
|--------------|----------|-----------|--------|------|-----------------------------|
| Bayes | 0.96 | 0.96 | 0.97 | 0.96 | Brak dyskretyzacji; $K = 8$ |
| C4.5 | 0.95 | 0.95 | 0.94 | 0.94 | C1; $K = 5$ |

Tabela 16: Najlepsze wyniki dla klas. Bayesowskiego i drzewa C4.5 (względem F1).

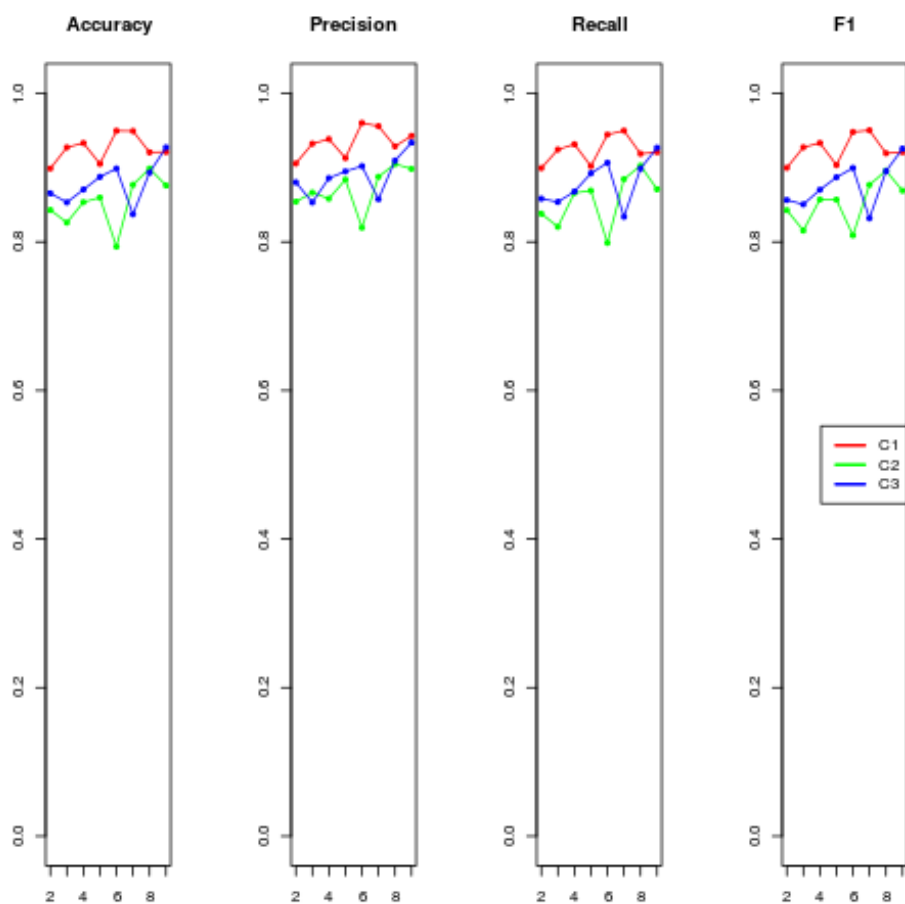
| | | Liczba foldów | | | | | | | |
|--------------|-----------|---------------|------|------|------|------|------|------|------|
| Konfiguracja | Metryka | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| C1 | Accuracy | 0.90 | 0.92 | 0.93 | 0.95 | 0.94 | 0.90 | 0.94 | 0.94 |
| C1 | Precision | 0.90 | 0.92 | 0.92 | 0.95 | 0.94 | 0.90 | 0.94 | 0.95 |
| C1 | Recall | 0.90 | 0.92 | 0.92 | 0.94 | 0.94 | 0.91 | 0.92 | 0.93 |
| C1 | F1 | 0.90 | 0.92 | 0.92 | 0.94 | 0.93 | 0.90 | 0.93 | 0.93 |
| C2 | Accuracy | 0.87 | 0.84 | 0.86 | 0.89 | 0.87 | 0.90 | 0.89 | 0.85 |
| C2 | Precision | 0.89 | 0.85 | 0.87 | 0.90 | 0.88 | 0.91 | 0.90 | 0.86 |
| C2 | Recall | 0.89 | 0.84 | 0.87 | 0.90 | 0.86 | 0.90 | 0.88 | 0.86 |
| C2 | F1 | 0.87 | 0.83 | 0.87 | 0.89 | 0.86 | 0.89 | 0.88 | 0.84 |
| C3 | Accuracy | 0.88 | 0.90 | 0.90 | 0.89 | 0.91 | 0.91 | 0.87 | 0.88 |
| C3 | Precision | 0.87 | 0.92 | 0.91 | 0.90 | 0.91 | 0.92 | 0.87 | 0.90 |
| C3 | Recall | 0.89 | 0.90 | 0.92 | 0.89 | 0.92 | 0.91 | 0.89 | 0.90 |
| C3 | F1 | 0.87 | 0.90 | 0.90 | 0.88 | 0.91 | 0.91 | 0.86 | 0.88 |

Tabela 17: Dokładne wartości metryk dla drzewa decyzyjnego C4.5.

Kroswalidacja stratyfikowana



Rysunek 18: Wartości metryk dla klasyfikatora Bayesowskiego.



Rysunek 19: Wartości metryk dla drzewa decyzyjnego C4.5.

| Klasyfikator | Accuracy | Precision | Recall | F1 | Komentarz |
|--------------|----------|-----------|--------|------|-----------------------------|
| Bayes | 0.97 | 0.97 | 0.97 | 0.97 | Brak dyskretyzacji; $K = 2$ |
| C4.5 | 0.95 | 0.96 | 0.95 | 0.95 | C1; $K = 7$ |

Tabela 18: Najlepsze wyniki dla klas. Bayesowskiego i drzewa C4.5 (względem F1).

| | | Liczba foldów | | | | | | | |
|--------------|-----------|---------------|------|------|------|------|------|------|------|
| Konfiguracja | Metryka | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| C1 | Accuracy | 0.90 | 0.93 | 0.93 | 0.91 | 0.95 | 0.95 | 0.92 | 0.92 |
| C1 | Precision | 0.91 | 0.93 | 0.94 | 0.91 | 0.96 | 0.96 | 0.93 | 0.94 |
| C1 | Recall | 0.90 | 0.92 | 0.93 | 0.90 | 0.94 | 0.95 | 0.92 | 0.92 |
| C1 | F1 | 0.90 | 0.93 | 0.93 | 0.90 | 0.95 | 0.95 | 0.92 | 0.92 |
| C2 | Accuracy | 0.84 | 0.83 | 0.85 | 0.86 | 0.79 | 0.88 | 0.90 | 0.88 |
| C2 | Precision | 0.85 | 0.87 | 0.86 | 0.88 | 0.82 | 0.89 | 0.90 | 0.90 |
| C2 | Recall | 0.84 | 0.82 | 0.87 | 0.87 | 0.80 | 0.88 | 0.90 | 0.87 |
| C2 | F1 | 0.84 | 0.82 | 0.86 | 0.86 | 0.81 | 0.88 | 0.90 | 0.87 |
| C3 | Accuracy | 0.87 | 0.85 | 0.87 | 0.89 | 0.90 | 0.84 | 0.89 | 0.93 |
| C3 | Precision | 0.88 | 0.85 | 0.89 | 0.89 | 0.90 | 0.86 | 0.91 | 0.93 |
| C3 | Recall | 0.86 | 0.85 | 0.87 | 0.89 | 0.91 | 0.83 | 0.90 | 0.93 |
| C3 | F1 | 0.86 | 0.85 | 0.87 | 0.89 | 0.90 | 0.83 | 0.90 | 0.93 |

Tabela 19: Dokładne wartości metryk dla drzewa decyzyjnego C4.5.

4 Wnioski

- Drzewa decyzyjne prezentują wyniki uczenia w sposób bardziej przystępny dla człowieka. Można bardzo łatwo prześledzić proces podejmowania decyzji w wygenerowanym drzewie – dla danej instancji należy wybierać odpowiednie gałęzie w zależności od wartości atrybutów instancji oraz warunków określonych w węzłach drzewa.
- Dzięki zastosowaniu różnych metod przycinania drzewa można ograniczyć głębokość drzewa i pozwolić na lepszą generalizację zbioru danych.
- Dla zbioru danych "Diabetes" pozwalało otrzymać stosunkowo lepsze wyniki. Dla krosvalidacji zwykłej: Precision równe 0.76 zamiast 0.66, Recall: 0.90 zamiast 0.6 oraz F1: 0.83 zamiast 0.63, natomiast Accuracy było na takim samym poziomie (dane dla najlepszych wyników obu algorytmów względem miary F1). Zastosowanie krosvalidacji stratyfikowanej na tym zbiorze danych nie spowodowało dużych zmian i wyniki są porównywalne (prawie identyczne jak w przypadku krosvalidacji zwykłej).
- W przypadku zbioru danych "Glass" przy zastosowaniu krosvalidacji zwykłej, zastosowanie algorytmu drzewa C4.5 pozwalało uzyskać znacząco lepsze wyniki – prawie 2 razy lepsze (patrz: Tabela 12). Natomiast przy zastosowaniu krosvalidacji stratyfikowanej różnice nie były tak duże; należy również pamiętać, że w tej sytuacji klasyfikator Bayesowski (z metodą dyskretyzacji CAIM) uzyskał również lepsze wyniki niż dla krosvalidacji zwykłej.
- Dla zbioru danych "Wine" algorytm C4.5 uzyskał nieznacznie gorsze wyniki (patrz: Tabela 16). Należy jednak zauważyć, że oba algorytmy uzyskiwały, dla tego zbioru danych, bardzo dobre osiągi (rzędu 90%).
- Wpływ opcji konfiguracyjnych (parametrów) algorytmu C4.5 jest uzależniony od zbioru danych. Dla zbioru "Diabetes" zróżnicowanie wyników osiągniętych przez algorytm jest dość małe w zależności od paramterów. Największe zróżnicowanie pojawia się w przypadku zbioru "Glass".