Indukcyjne metody analizy danych
Ćwiczenie 1

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

Prowadzący: dr inż. Paweł Myszkowski
Student: Piotr Bielak, 218137
WT 17:05

Wrocław, 13 marca 2018r.

Spis treści

1 Wprowadzenie

1.1 Cel ćwiczenia

Celem ćwiczenia było poznanie tzw. naiwnego klasyfikatora Bayesa oraz zbadanie i ocena jego działania na 3 określonych zbiorach danych. W trakcie badań należało uwzględnić różne metody dyskretyzacji danych i kroswalidacji oraz zaobserwować wpływ tych parametrów na wartości zadanych metryk.

1.2 Klasyfikator Bayesowski

Typowym zagadnieniem w uczeniu maszynowym jest zadanie klasyfikacji. Należy ono do grupy tzw. **zadań uczenia nadzorowanego**, czyli zakłada istnienie zbioru danych, w którym każda instancja (wektor cech) jest oznaczona odpowiednią etykietą (*klasa*). Narzędzie, które jest uczone na takim zbiorze, a następnie używane do przyporządkowywania etykiet do nowych instancji, nazywa się **klasyfikatorem**. W tym ćwiczeniu użyty zostanie **naiwny klasyfikator Bayesowski** (ang. *Naive Bayes Classifier*). Jest on oparty o twierdzenie Bayesa:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

gdzie:
X - wektor cech danej instancji,
Y - klasa.

Powyższy zapis odczytujemy jako prawdopodobieństwo przynależności instancji X do klasy Y. Ważne jest, że ten klasyfikator zakłada niezależność wszystkich atrybutów (cech), co w większości przypadków się nie sprawdza (stąd nazwa *naiwny*). Stąd w powyższym wzorze, człon $P(X|Y)$ można zastąpić iloczynem prawdopodobieństw:

$$P(X|Y) = \prod_{i=1}^{n} P(X_i|Y)$$

Problem jaki się tutaj pojawia, to sytuacja w której jedno z prawdopodobieństw $P(X_i|Y) = 0$, wtedy cały iloczyn się również wyzeruje. W celu przecidziałania temu, stosuje się tzw. wygładzanie danych – dla metody Laplace'a zwiększa się częstość występowania danego atrybutu. Klasa przypisywana przez klasyfikator dla danej instancji jest dobierana w taki sposób, aby prawdopodobieństwo $P(Y|X)$ przyjęło największą sposród możliwych wartości.

Można wyróżnić 2 główne typy klasyfikatorów Bayesowskich:

- **Gaussowski naiwny Bayes** – atrybuty przyjmują wartości ciągłe oraz zakłada się, że każdy atrybut posiada rozkład normalny;
- **wielomianowy naiwny Bayes** – atrybuty przyjmują wartości dyskretne; parametrami przyjętego tutaj rozkładu wielomianowego (prawdopodobieństwami) są wektory postaci $(P(X_1|Y_i), P(X_2|Y_i), ..., P(X_n|Y_i))$ dla każdej klasy Y_i.

1.3 Dyskretyzacja

Często w różnych zbiorach danych atrybuty są zdefiniowane jako wartości ciągłe, co utrudnia pracę z nimi. Stosuje się zbieg dyskretyzacji, który jest określona jako funkcja: $f : R \rightarrow N$ (ew. w dziedzinę liczb całkowitych), która dla poszczególnych wartości atrybutów przypisuje im odpowiednie, dyskretne wartości.

Algortym ten działa najczęściej w oparciu o tzw. kubełkowanie. Tworzona jest odpowiednia liczba kubełków (przedziałów wartości $[x_i, x_{i+1}), [x_{i+1}, x_{i+2}), ...$). Następnie każda wartość danego atrybutu jest przypisywana do odpowiedniego przedziału. Po zakończeniu tej procedury, zamiast posługiwać się konkretną wartością atrybuty, zostają one zastąpione za pomocą np. numerów/etykiet kubełków.

W ćwiczeniu zostały wykorzystane następujące metody dyskretyzacji:

- **Equal-width binning** – zakłada się tutaj, że szerokość kubełka/przedziału jest stała, a parametrem który się ustawia jest liczba tych kubełków;

$$binwidth = \frac{x_{max} - x_{min}}{\#bins}$$

- **Equal-frequency binning** – szerokości kubełków mogą być różne, ale powinny być tak dobrane, aby w każdym z nich mieściło się mniej więcej po równo wartości atrybutów; parametrem tutaj również jest liczba kubełków;

- **CAIM (Class-Attribute Interdependence Maximization)** – w przeciwieństwie do poprzednich metod dyskretyzacji, które należą do grupy metod bez nadzoru, ta metoda pochodzi z grupy metod nazdorowanych (z naczycielem); korzysta ona z całego zbioru danych (atrybuty wraz z klasami) i próbuje zmaksymalizować zależność między atrybutami danej klasy, przy jednoczesnej minimalizacji liczby etykiet ("kubełków"); dokładny opis działania tej metody został podany w [1].

1.4 Kroswalidacja

W celu lepszej oceny jakości działania (uczenia) klasyfikatora stostuje się kroswalidację (sprawdzian krzyżowy). Zakłada ona, że zbiór danych dzielimy na podzbiory: zbiór danych uczących i zbiór danych testowych/walidacyjnych. W ćwiczeniu zostały wykorzystane dwie metody:

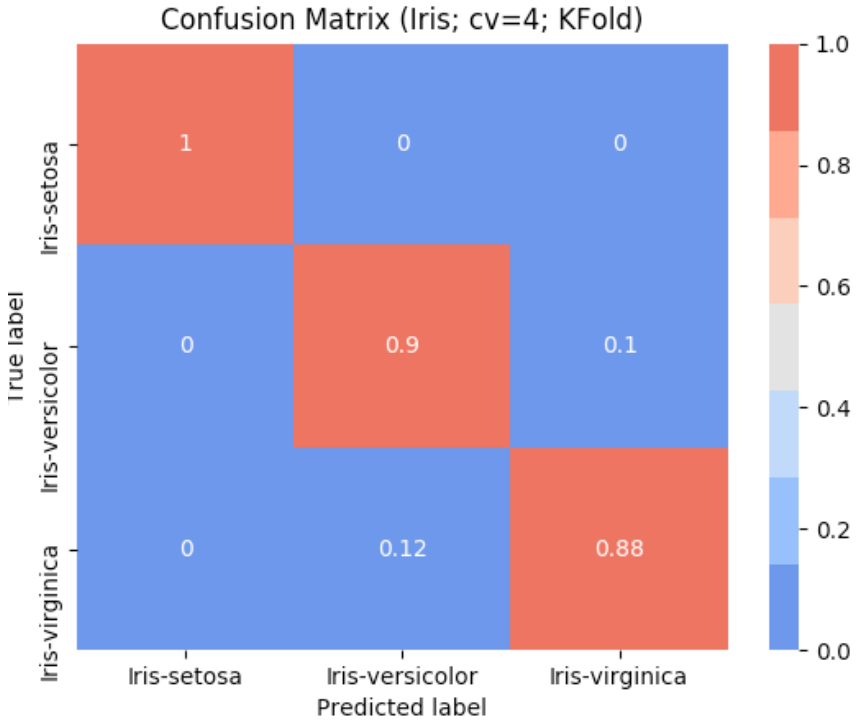
- **K-Fold** – zbiór danych jest dzielony na K pozbiorów, z których każdy kolejno jest przyjmowany jako zbiór testowy, natomiast pozostałe służą do nauki modelu; metoda może być dość czasochłonna przy dużej wartości parametru K, jako że przeprowadzanych jest kolejno K przebiegów metody;

- **Stratified K-Fold** – metoda ta jest bardzo podobna do poprzedniej, jednak w przeciwieństwie do niej gwarantuje, że podczas podziałów podzbiorów, w każdym z nich zostanie zachowana proporcja instancji różnych klas, zgodnie z proporcją istniejącą w całkowitym zbiorze danych.

1.5 Metryki

Jako miary (metryki) oceny jakości klasyfikatora zostały zastosowane następujące miary:

- **Accuracy** (dokładność) – stosunek liczby prawidłowo zaklasyfikowanych instancji do liczby wszystkich instancji,
$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$
- **Precision** (precyzja) – stosunek liczby prawidłowo zaklasyfikowanych pozytywnych instancji do liczby wszystkich instancji zaklasyfikowanych jako pozytywne,
$$Precision = \frac{TP}{TP + FP}$$
- **Recall** – stosunek liczby prawidłowo zaklasyfikowanych pozytywnych instancji do liczby wszystkich poprawnie zaklasyfikowanych,
$$Recall = \frac{TP}{TP + FN}$$
- **F1-Score** – ważona średnia wartości Precision oraz Recall; uwzględnia zarówno błędne pozytywy jak i błędne negatywy, dzięki czemu wnosi więcej informacji niż Accuracy,
$$F1 = \frac{Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$
- **Confusion Matrix** (macierz konfuzji) – jest to macierz, która prezentuje dla wszystkich dostępnych klas w danych zbiorze danych, jak często instancje zostały klasyfikowane jako poczególne klasy; komórka i, j oznacza zatem, że instancja z klasy i została zaklasyfikowana jako j, stąd idealna macierz konfuzji powinna zawierać niezerowe wartości tylko na przekątnej (prawidłowa klasyfikacja); w kolejnym rozdziale, zaprezentowane macierze, zostały znormalizowane.



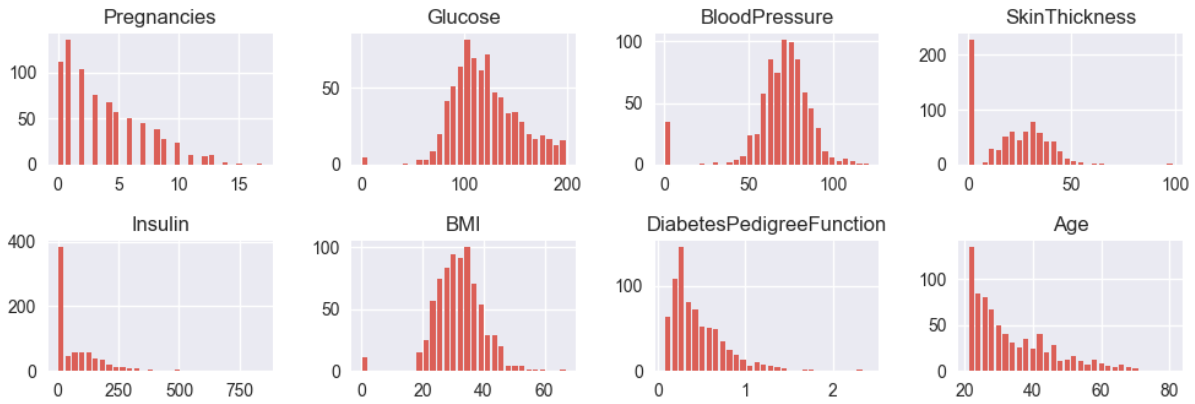
Rysunek 1: Przykładowa (znormalizowana) macierz konfuzji.

2 Eksperyment

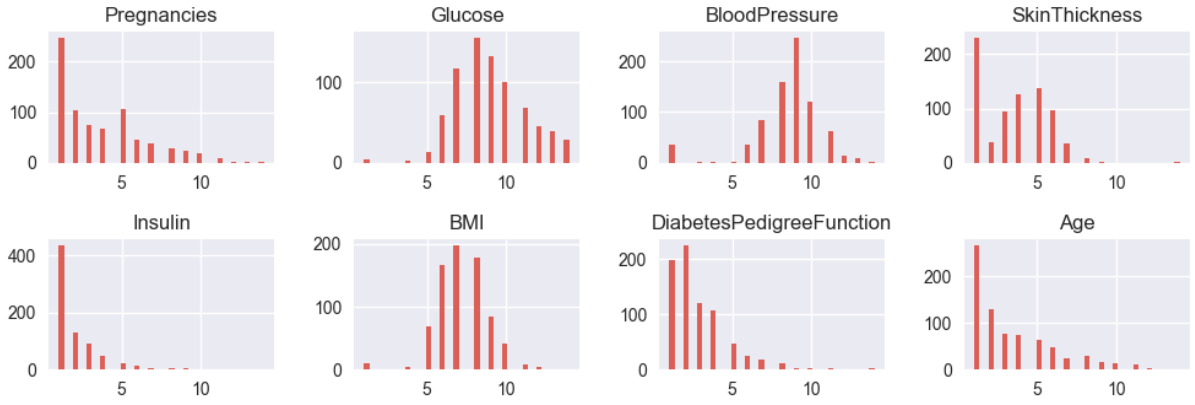
2.1 Założenia

2.2 Wyniki dyskretyzacji

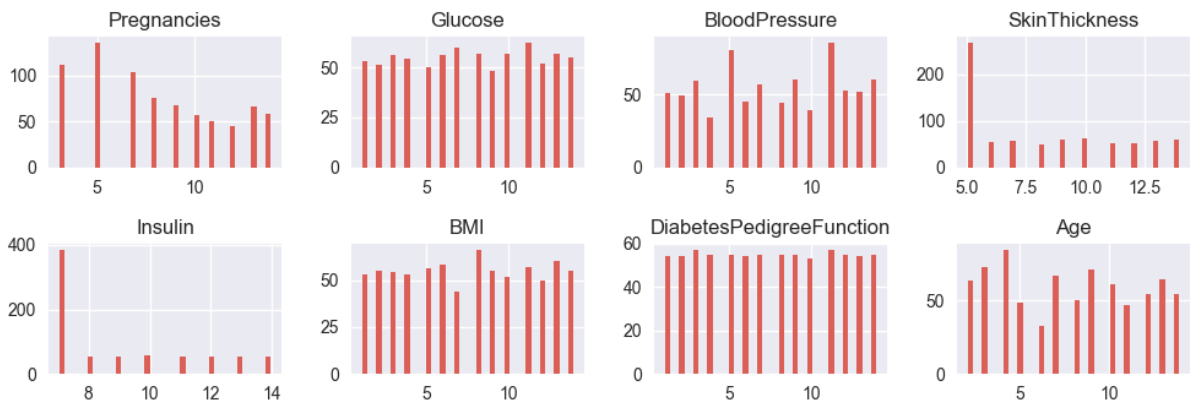
2.2.1 Zbiór danych - "Diabetes"



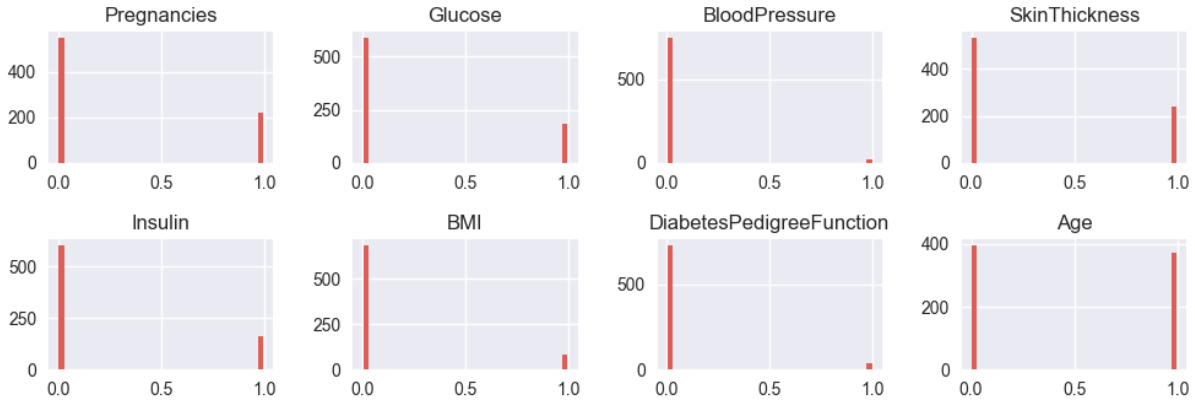
Rysunek 2: Rozkłady atrybutów zbioru "Diabetes" – brak dyskretyzacji.



Rysunek 3: Rozkłady atrybutów zbioru "Diabetes" – dyskretyzacja "equal-width".

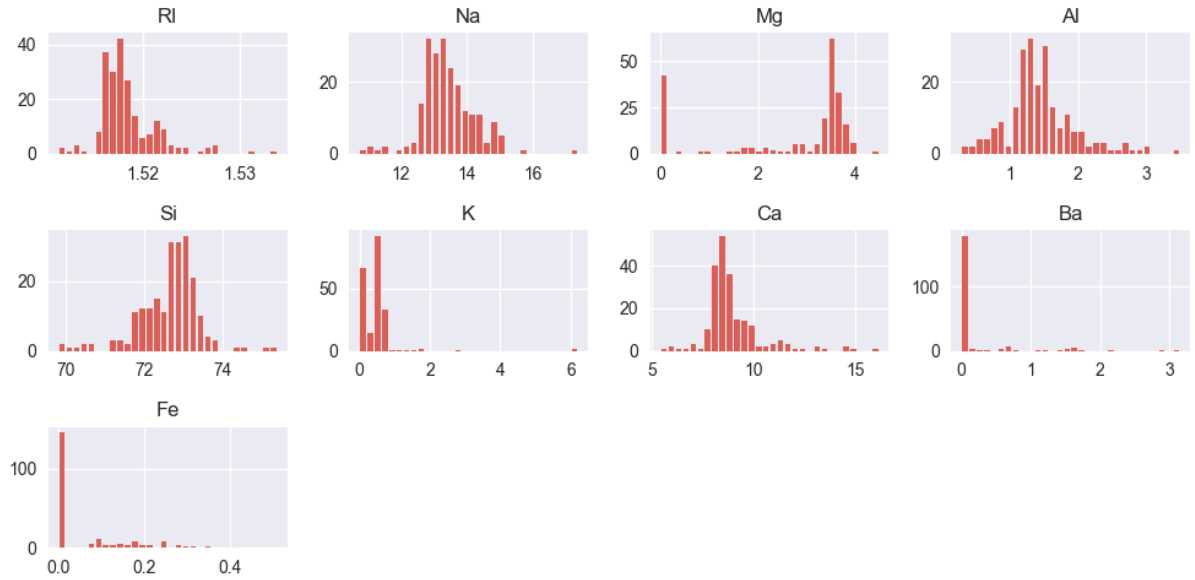


Rysunek 4: Rozkłady atrybutów zbioru "Diabetes" – dyskretyzacja "equal-frequency".

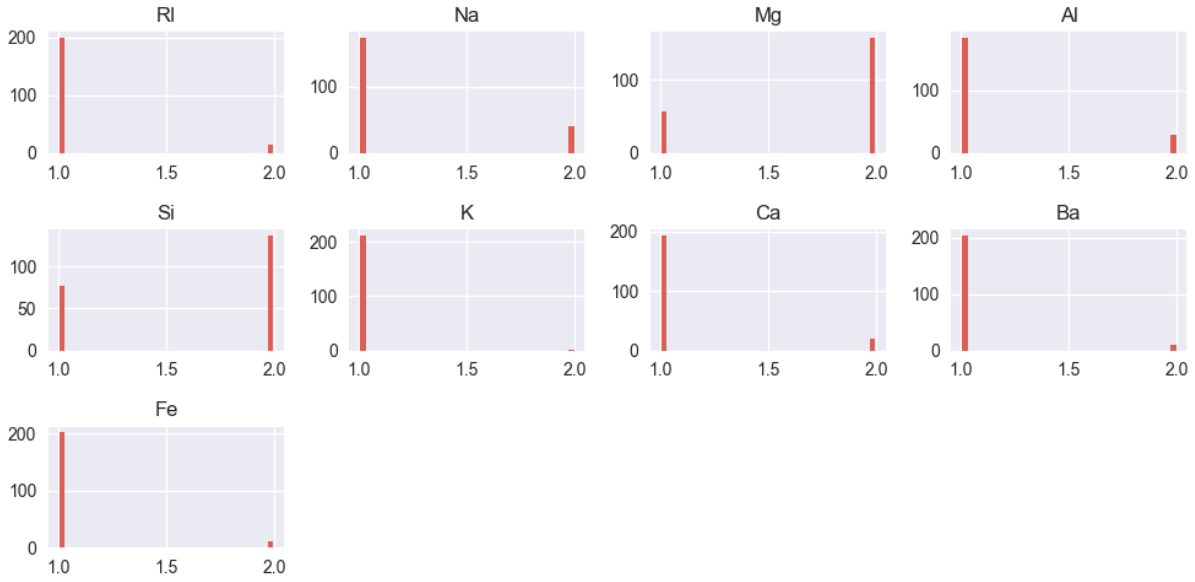


Rysunek 5: Rozkłady atrybutów zbioru "Diabetes" – dyskretyzacja "CAIM".

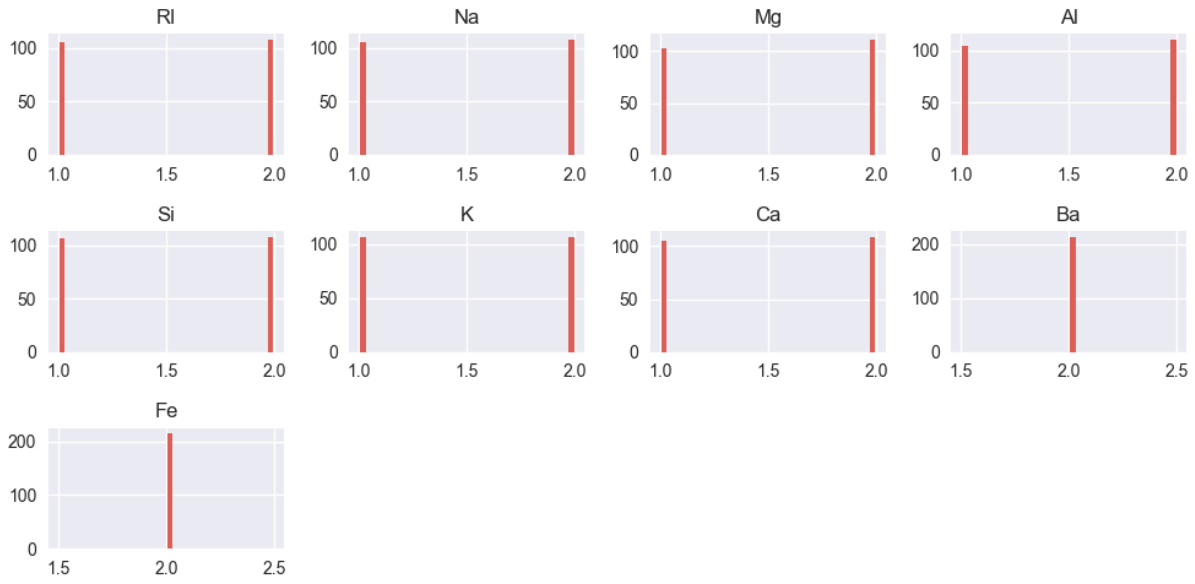
2.2.2 Zbiór danych - "Glass"



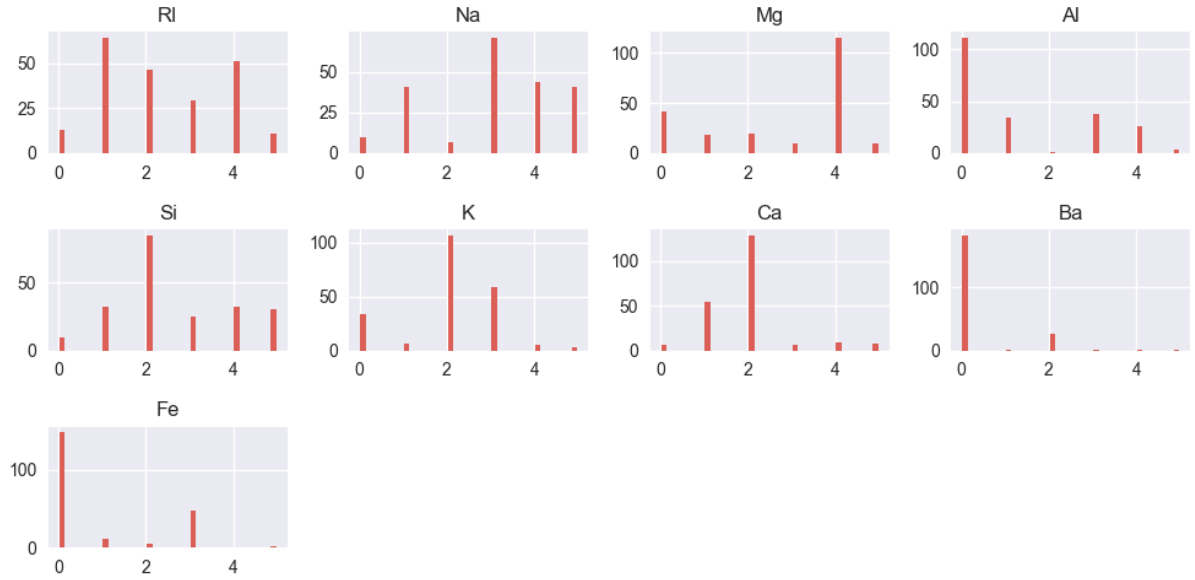
Rysunek 6: Rozkłady atrybutów zbioru "Glass" – brak dyskretyzacji.



Rysunek 7: Rozkłady atrybutów zbioru "Glass" – dyskretyzacja "equal-width".

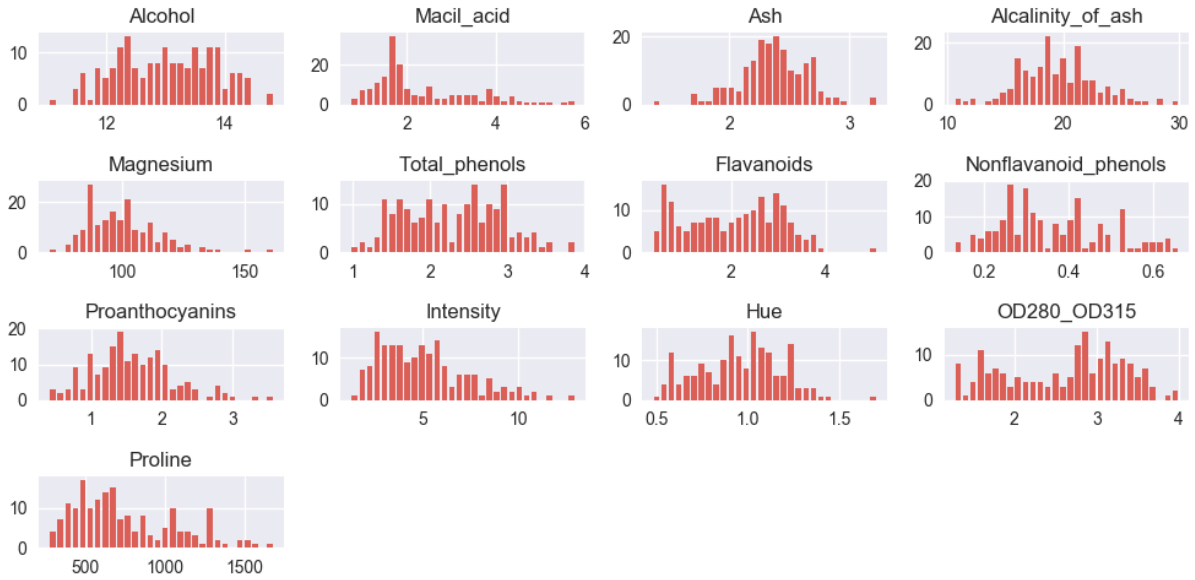


Rysunek 8: Rozkłady atrybutów zbioru "Glass" – dyskretyzacja "equal-frequency".

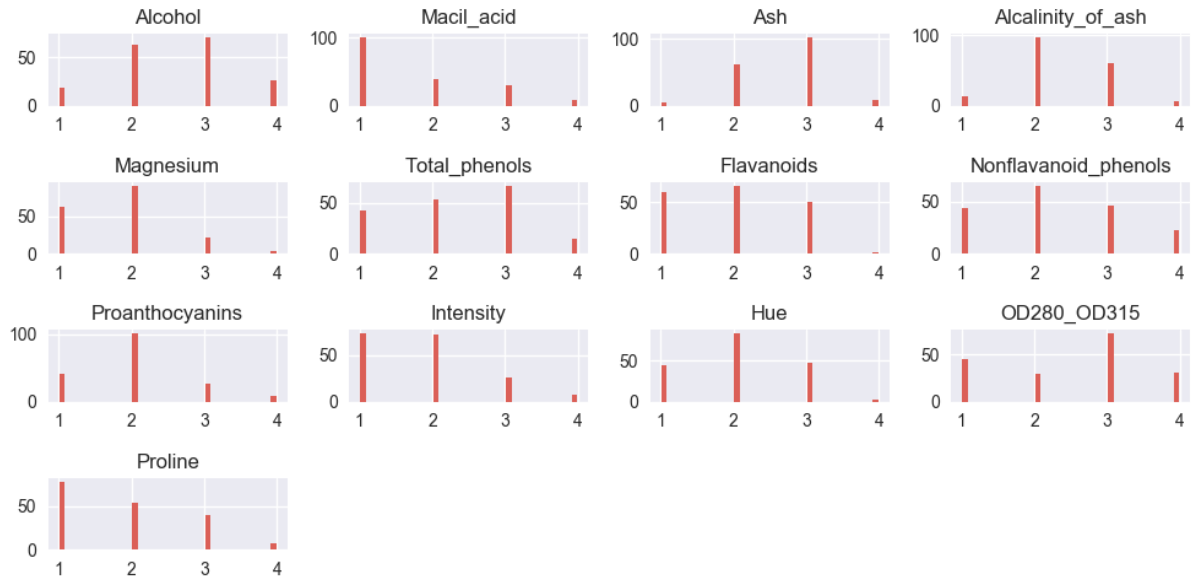


Rysunek 9: Rozkłady atrybutów zbioru "Glass" – dyskretyzacja "CAIM".

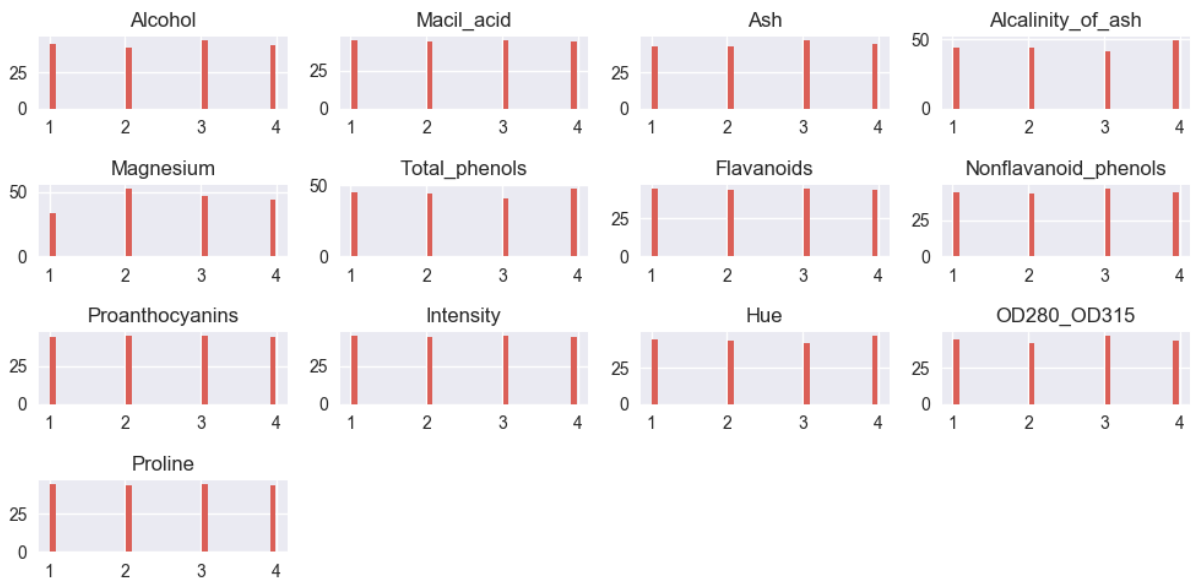
2.2.3 Zbiór danych - "Wine"



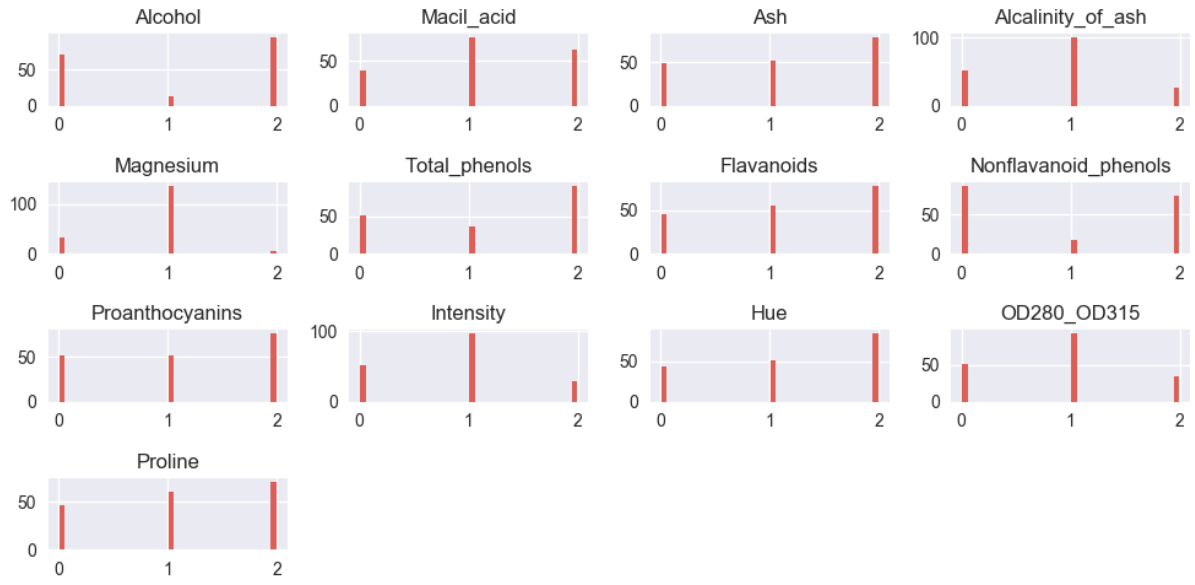
Rysunek 10: Rozkłady atrybutów zbioru "Wine" – brak dyskretyzacji.



Rysunek 11: Rozkłady atrybutów zbioru "Wine" – dyskretyzacja "equal-width".



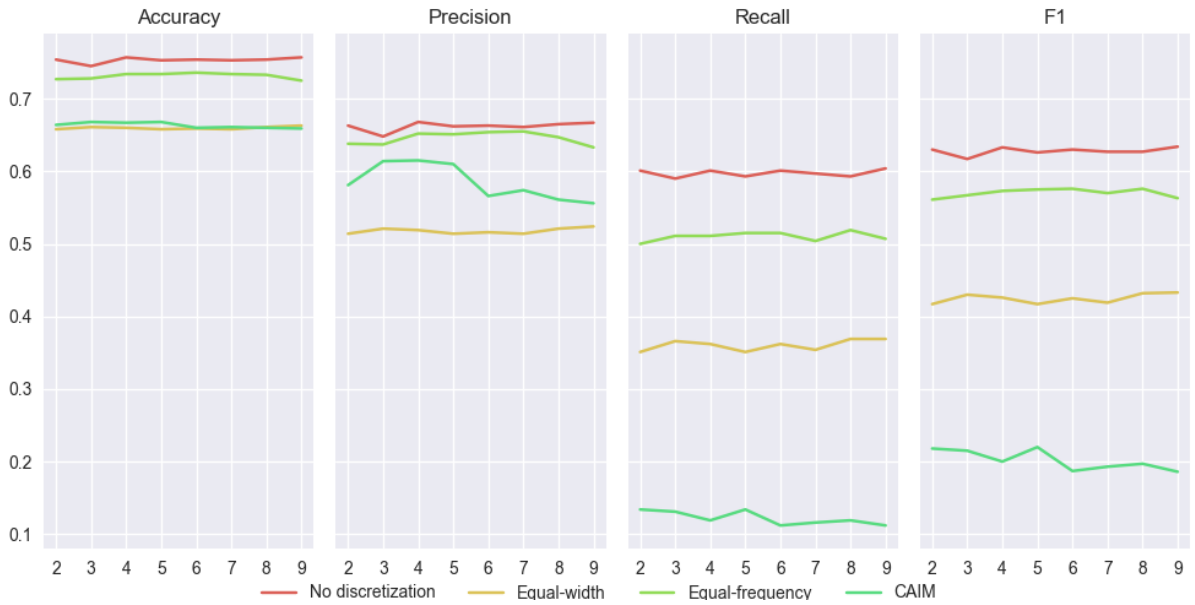
Rysunek 12: Rozkłady atrybutów zbioru "Wine" – dyskretyzacja "equal-frequency".



Rysunek 13: Rozkłady atrybutów zbioru "Wine" – dyskretyzacja "CAIM".

2.3 Wyniki kroswalidacji

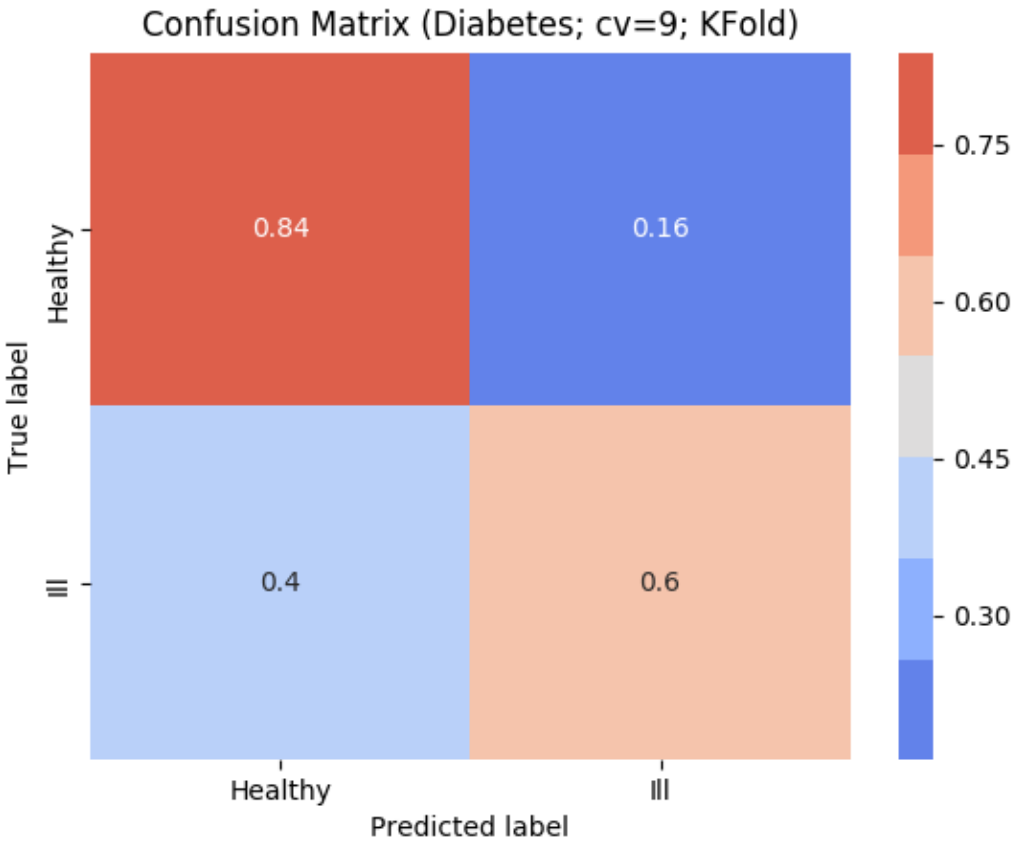
2.3.1 Zbiór danych – "Diabetes"



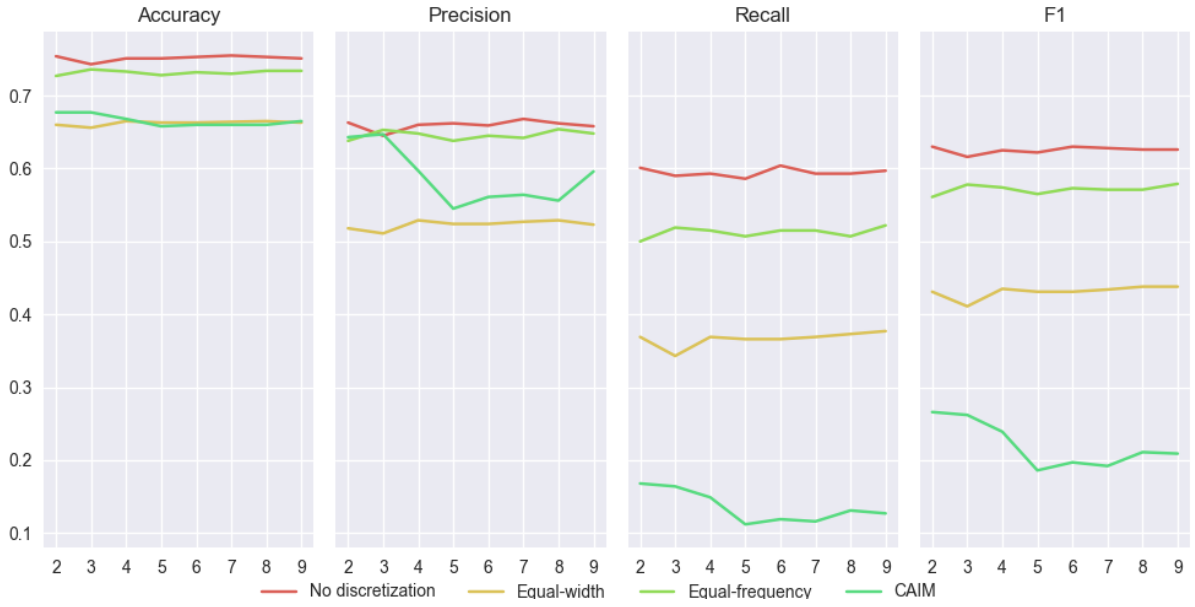
Rysunek 14: Wykresy wartości metryk dla zbioru "Diabetes" – krowalidacja zwykła.

Tabela 1: Wartości metryk dla zbioru "Diabetes" – krowalidacja zwykła.

Metoda dyskr.	**Metryka**	**CV**							
		2	3	4	5	6	7	8	9
Brak	Accuracy	0.754	0.745	0.757	0.753	0.754	0.753	0.754	0.757
	Precision	0.663	0.648	0.668	0.662	0.663	0.661	0.665	0.667
	Recall	0.601	0.59	0.601	0.593	0.601	0.597	0.593	0.604
	F1	0.63	0.617	0.633	0.626	0.63	0.627	0.627	0.634
Equal-width	Accuracy	0.658	0.661	0.66	0.658	0.659	0.658	0.661	0.663
	Precision	0.514	0.521	0.519	0.514	0.516	0.514	0.521	0.524
	Recall	0.351	0.366	0.362	0.351	0.362	0.354	0.369	0.369
	F1	0.417	0.43	0.426	0.417	0.425	0.419	0.432	0.433
Equal-freq	Accuracy	0.727	0.728	0.734	0.734	0.736	0.734	0.733	0.725
	Precision	0.638	0.637	0.652	0.651	0.654	0.655	0.647	0.633
	Recall	0.5	0.511	0.511	0.515	0.515	0.504	0.519	0.507
	F1	0.561	0.567	0.573	0.575	0.576	0.57	0.576	0.563
CAIM	Accuracy	0.664	0.668	0.667	0.668	0.66	0.661	0.66	0.659
	Precision	0.581	0.614	0.615	0.61	0.566	0.574	0.561	0.556
	Recall	0.134	0.131	0.119	0.134	0.112	0.116	0.119	0.112
	F1	0.218	0.215	0.2	0.22	0.187	0.193	0.197	0.186



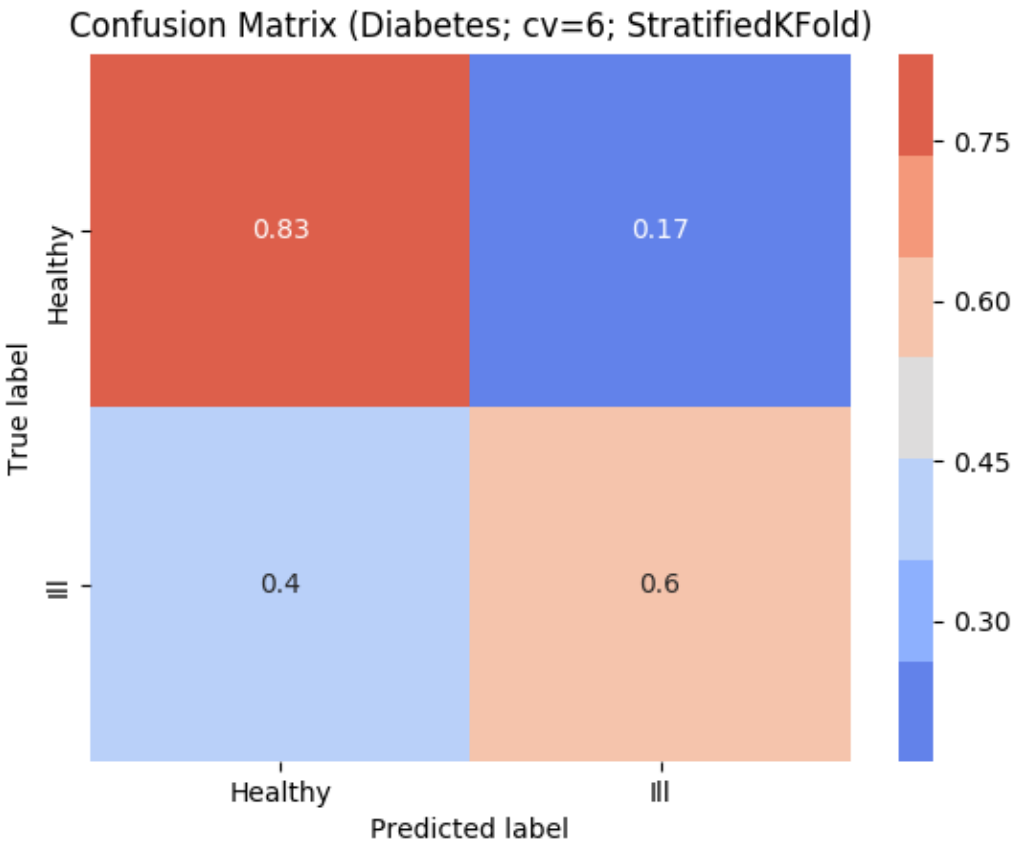
Rysunek 15: Macierz konfuzji dla najlepszej wartości F1 – kroswalidacja zwykła.



Rysunek 16: Wykresy wartości metryk dla zbioru "Diabetes" – krowalidacja stratyfikowana.

Tabela 2: Wartości metryk dla zbioru "Diabetes" – krowalidacja stratyfikowana.

Metoda dyskr.	**Metryka**	**CV**							
		2	3	4	5	6	7	8	9
Brak	Accuracy	0.754	0.743	0.751	0.751	0.753	0.755	0.753	0.751
	Precision	0.663	0.645	0.66	0.662	0.659	0.668	0.662	0.658
	Recall	0.601	0.59	0.593	0.586	0.604	0.593	0.593	0.597
	F1	0.63	0.616	0.625	0.622	0.63	0.628	0.626	0.626
Equal-width	Accuracy	0.66	0.656	0.665	0.663	0.663	0.664	0.665	0.663
	Precision	0.518	0.511	0.529	0.524	0.524	0.527	0.529	0.523
	Recall	0.369	0.343	0.369	0.366	0.366	0.369	0.373	0.377
	F1	0.431	0.411	0.435	0.431	0.431	0.434	0.438	0.438
Equal-freq	Accuracy	0.727	0.736	0.733	0.728	0.732	0.73	0.734	0.734
	Precision	0.638	0.653	0.648	0.638	0.645	0.642	0.654	0.648
	Recall	0.5	0.519	0.515	0.507	0.515	0.515	0.507	0.522
	F1	0.561	0.578	0.574	0.565	0.573	0.571	0.571	0.579
CAIM	Accuracy	0.677	0.677	0.668	0.658	0.66	0.66	0.66	0.665
	Precision	0.643	0.647	0.597	0.545	0.561	0.564	0.556	0.596
	Recall	0.168	0.164	0.149	0.112	0.119	0.116	0.131	0.127
	F1	0.266	0.262	0.239	0.186	0.197	0.192	0.211	0.209



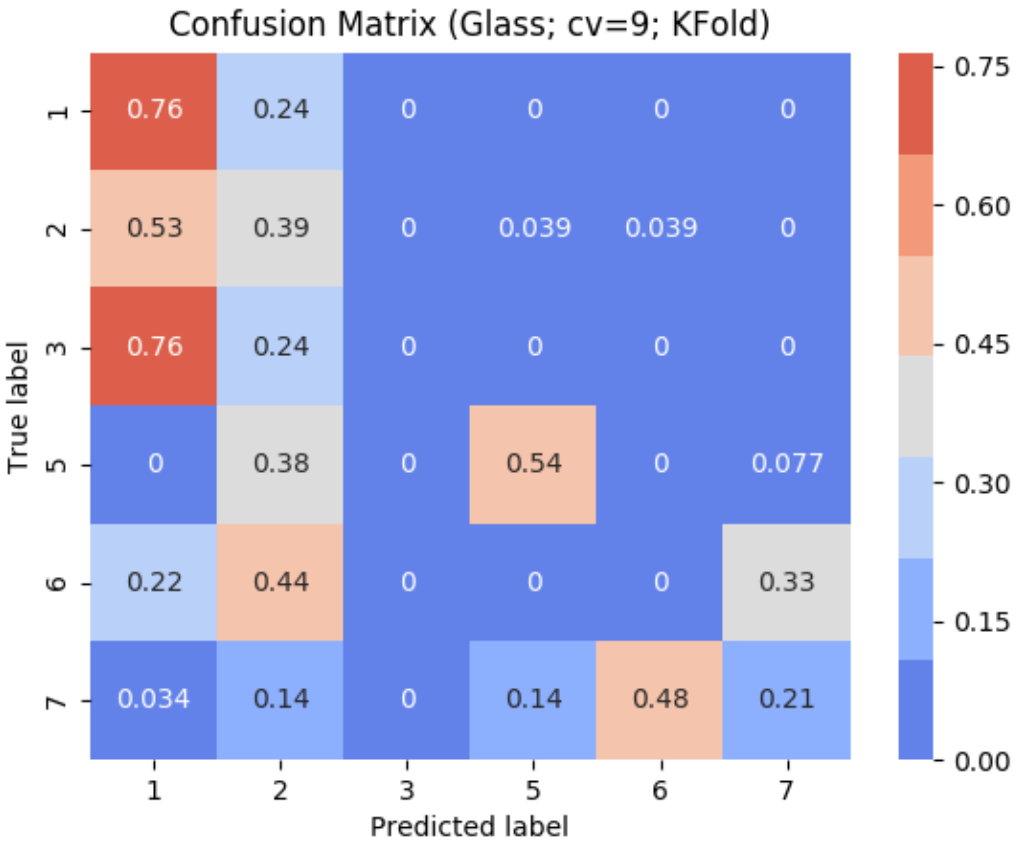
Rysunek 17: Macierz konfuzji dla najlepszej wartości F1 – kroswalidacja stratyfikowana.

2.3.2 Zbiór danych – "Glass"

Rysunek 18: Wykresy wartości metryk dla zbioru "Glass" – krowalidacja zwykła.

Tabela 3: Wartości metryk dla zbioru "Glass" – krowalidacja zwykła.

Metoda dyskr.	**Metryka**	**CV**							
		2	3	4	5	6	7	8	9
Brak	Accuracy	0.093	0.234	0.136	0.201	0.121	0.28	0.173	0.257
	Precision	0.039	0.076	0.062	0.137	0.153	0.208	0.098	0.263
	Recall	0.044	0.11	0.067	0.139	0.124	0.208	0.086	0.207
	F1	0.041	0.09	0.064	0.125	0.133	0.188	0.086	0.214
Equal-width	Accuracy	0.173	0.014	0.089	0.084	0.019	0.037	0.056	0.023
	Precision	0.058	0.004	0.03	0.022	0.005	0.014	0.015	0.006
	Recall	0.081	0.007	0.045	0.039	0.009	0.019	0.026	0.011
	F1	0.067	0.005	0.036	0.028	0.006	0.016	0.019	0.008
Equal-freq	Accuracy	0.159	0.014	0.075	0.103	0.009	0.079	0.103	0.042
	Precision	0.054	0.004	0.023	0.026	0.003	0.027	0.027	0.011
	Recall	0.075	0.007	0.038	0.048	0.004	0.039	0.048	0.02
	F1	0.063	0.005	0.029	0.034	0.003	0.032	0.034	0.014
CAIM	Accuracy	0.215	0.047	0.257	0.285	0.248	0.383	0.439	0.449
	Precision	0.054	0.014	0.091	0.125	0.137	0.166	0.208	0.342
	Recall	0.101	0.022	0.126	0.17	0.218	0.206	0.223	0.316
	F1	0.071	0.017	0.106	0.142	0.162	0.179	0.207	0.308

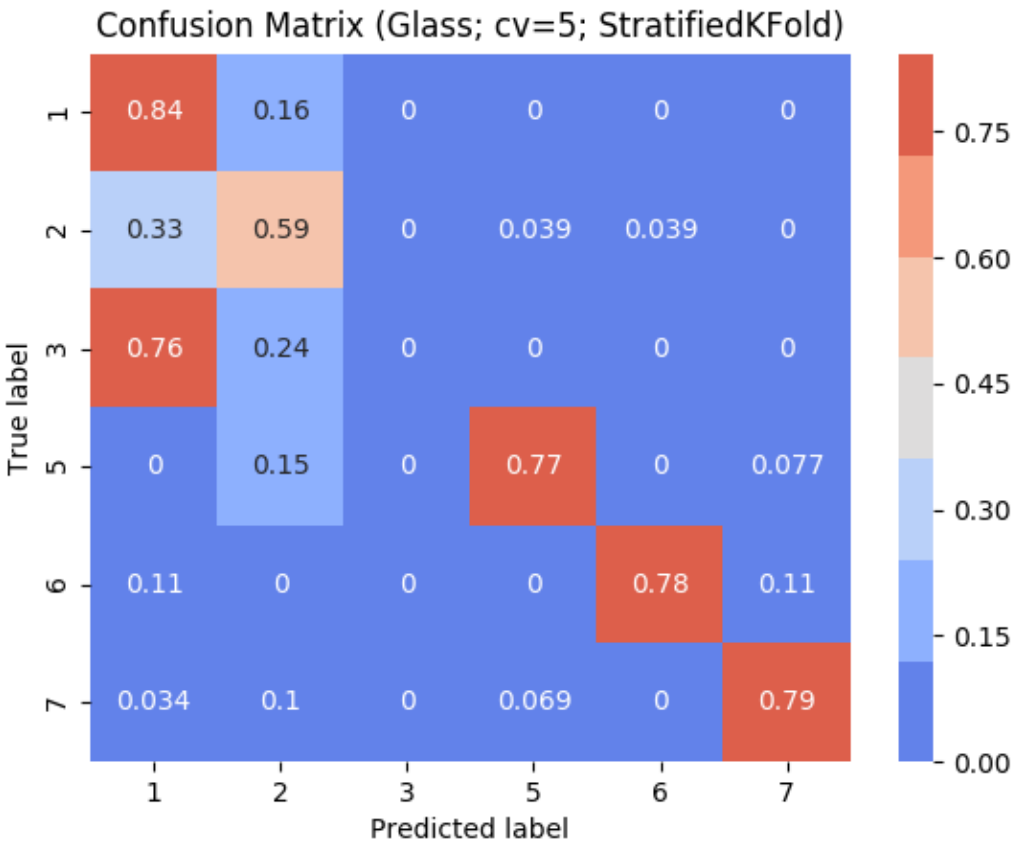


Rysunek 19: Macierz konfuzji dla najlepszej wartości F1 – kroswalidacja zwykła.

Rysunek 20: Wykresy wartości metryk dla zbioru "Glass" – krowalidacja stratyfikowana.

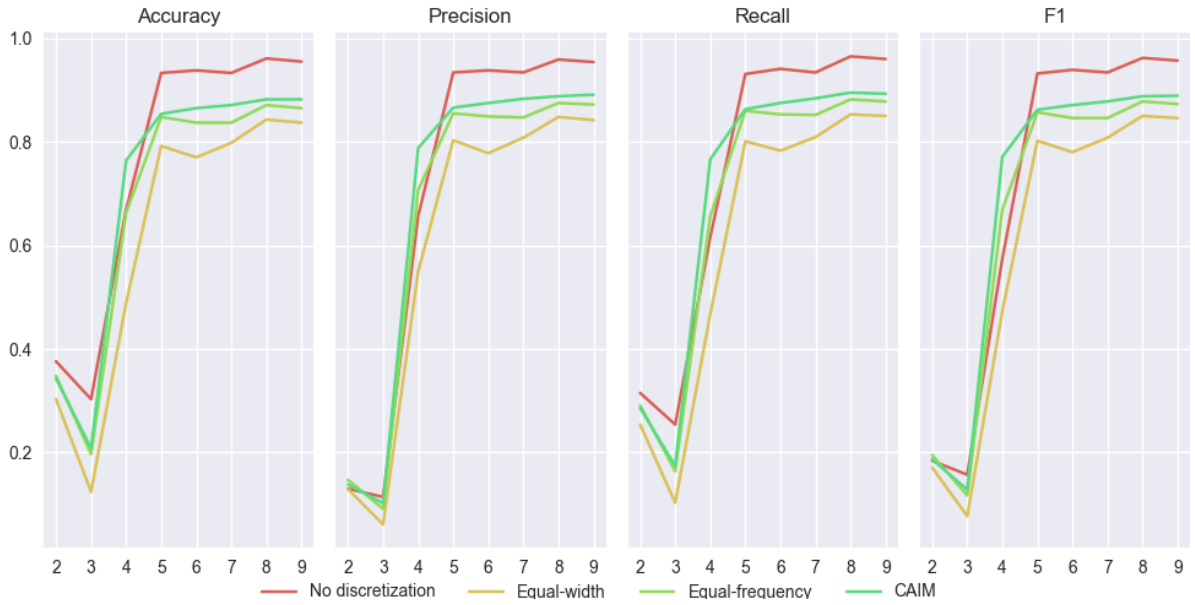
Tabela 4: Wartości metryk dla zbioru "Glass" – krowalidacja stratyfikowana.

Metoda dyskr.	**Metryka**	**CV**							
		2	3	4	5	6	7	8	9
Brak	Accuracy	0.364	0.374	0.397	0.332	0.416	0.393	0.421	0.444
	Precision	0.461	0.464	0.443	0.4	0.427	0.407	0.384	0.45
	Recall	0.379	0.474	0.465	0.418	0.438	0.452	0.437	0.499
	F1	0.386	0.439	0.426	0.388	0.415	0.416	0.392	0.454
Equal-width	Accuracy	0.36	0.36	0.364	0.35	0.364	0.364	0.374	0.369
	Precision	0.226	0.295	0.226	0.059	0.226	0.226	0.227	0.227
	Recall	0.172	0.176	0.178	0.164	0.178	0.178	0.19	0.184
	F1	0.099	0.143	0.109	0.087	0.109	0.109	0.129	0.12
Equal-freq	Accuracy	0.341	0.421	0.411	0.397	0.416	0.374	0.397	0.407
	Precision	0.12	0.146	0.145	0.139	0.146	0.131	0.139	0.142
	Recall	0.163	0.205	0.2	0.193	0.203	0.181	0.193	0.198
	F1	0.124	0.168	0.165	0.159	0.167	0.148	0.159	0.163
CAIM	Accuracy	0.664	0.673	0.673	0.673	0.664	0.654	0.664	0.668
	Precision	0.58	0.598	0.602	0.596	0.591	0.579	0.596	0.59
	Recall	0.593	0.619	0.608	0.629	0.625	0.599	0.607	0.614
	F1	0.581	0.604	0.599	0.607	0.602	0.584	0.597	0.598



Rysunek 21: Macierz konfuzji dla najlepszej wartości F1 – kroswalidacja stratyfikowana.

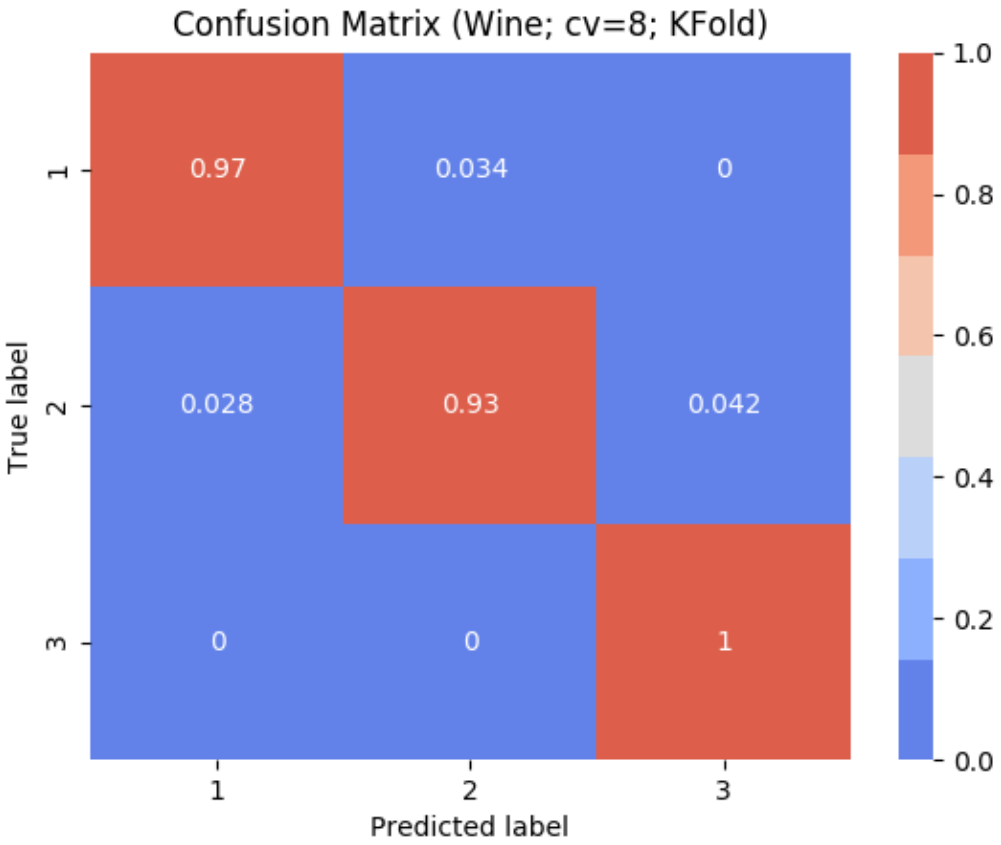
2.3.3 Zbiór danych – "Wine"



Rysunek 22: Wykresy wartości metryk dla zbioru "Wine" – krowalidacja zwykła.

Tabela 5: Wartości metryk dla zbioru "Wine" – krowalidacja zwykła.

Metoda dyskr.	**Metryka**	**CV**							
		2	3	4	5	6	7	8	9
Brak	Accuracy	0.376	0.303	0.669	0.933	0.938	0.933	0.961	0.955
	Precision	0.13	0.114	0.657	0.934	0.938	0.934	0.959	0.954
	Recall	0.315	0.254	0.616	0.931	0.941	0.934	0.965	0.96
	F1	0.184	0.157	0.574	0.932	0.939	0.934	0.962	0.957
Equal-width	Accuracy	0.303	0.124	0.489	0.792	0.77	0.798	0.843	0.837
	Precision	0.129	0.061	0.55	0.803	0.778	0.808	0.848	0.842
	Recall	0.254	0.103	0.467	0.801	0.783	0.809	0.853	0.85
	F1	0.171	0.077	0.473	0.802	0.78	0.808	0.85	0.846
Equal-freq	Accuracy	0.348	0.197	0.663	0.848	0.837	0.837	0.871	0.865
	Precision	0.147	0.091	0.707	0.855	0.849	0.847	0.875	0.872
	Recall	0.291	0.164	0.656	0.86	0.853	0.852	0.882	0.878
	F1	0.195	0.117	0.669	0.857	0.846	0.846	0.878	0.873
CAIM	Accuracy	0.343	0.208	0.764	0.854	0.865	0.871	0.882	0.882
	Precision	0.138	0.102	0.788	0.866	0.875	0.883	0.888	0.891
	Recall	0.286	0.174	0.766	0.863	0.875	0.884	0.895	0.893
	F1	0.187	0.128	0.771	0.862	0.871	0.878	0.888	0.889

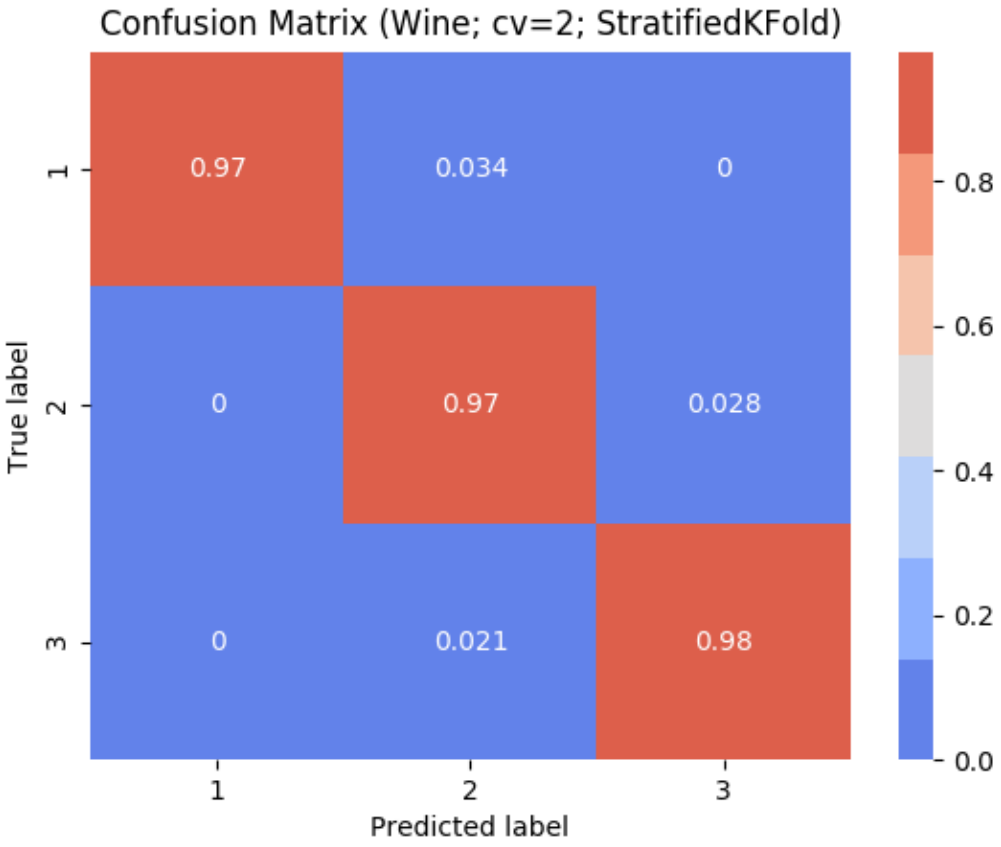


Rysunek 23: Macierz konfuzji dla najlepszej wartości F1 – kroswalidacja zwykła.

Rysunek 24: Wykresy wartości metryk dla zbioru "Wine" – krowalidacja stratyfikowana.

Tabela 6:

Metoda dyskr.	**Metryka**	**CV**							
		2	3	4	5	6	7	8	9
Brak	Accuracy	0.972	0.961	0.961	0.955	0.961	0.961	0.961	0.955
	Precision	0.973	0.961	0.959	0.954	0.96	0.959	0.959	0.954
	Recall	0.972	0.964	0.965	0.96	0.964	0.965	0.965	0.96
	F1	0.972	0.962	0.962	0.957	0.962	0.962	0.962	0.957
Equal-width	Accuracy	0.899	0.899	0.904	0.904	0.893	0.899	0.904	0.899
	Precision	0.902	0.903	0.908	0.908	0.897	0.902	0.907	0.902
	Recall	0.91	0.909	0.914	0.914	0.904	0.908	0.913	0.908
	F1	0.903	0.904	0.909	0.909	0.899	0.904	0.909	0.904
Equal-freq	Accuracy	0.893	0.888	0.91	0.899	0.904	0.91	0.916	0.916
	Precision	0.901	0.896	0.917	0.906	0.911	0.916	0.92	0.92
	Recall	0.905	0.898	0.921	0.908	0.915	0.92	0.924	0.924
	F1	0.9	0.894	0.915	0.904	0.91	0.915	0.92	0.92
CAIM	Accuracy	0.916	0.899	0.904	0.916	0.899	0.91	0.904	0.904
	Precision	0.924	0.908	0.911	0.922	0.906	0.915	0.91	0.91
	Recall	0.926	0.911	0.917	0.928	0.91	0.92	0.915	0.915
	F1	0.92	0.905	0.911	0.921	0.905	0.915	0.91	0.91



Rysunek 25: Macierz konfuzji dla najlepszej wartości F1 – kroswalidacja stratyfikowana.

3 Wnioski

-

4 Bibliografia