Politechnika Wrocławska

Indukcyjne Metody Analizy Danych

Ćwiczenie 4 - Algorytm klasyfikacji k-najbliższych sąsiadów

Marcel Cielinski
Index: 236747

prowadzący
dr inż. Paweł Myszkowski

14 maja 2020

Wprowadzenie

1.1 Problem

Celem ćwiczenia było zapoznanie się z metodą klasyfikacji k-najbliższych sąsiadów (ang. k-nearest neighbours, k-nn). Do składowych zadania zalicza się implementację wspomnianego algorytmu w jednym z dwóch wybranych języków: *R* lub *Python*. Następnie należało przeprowadzić badania na czterech, ustalonych na potrzeby ćwiczenia, zbiorach danych. Przedmiotem testów miło być zbadanie wpływu parametrów takich jak: wielkość wartości k (mówiącej o ilości rozpatrywanych sąsiadów), sposób głosowania oraz definicja miary odległości. Zadbać należało także o zastosowanie operacji na danych, które je ujednolicą (standaryzacja lub normalizacja) oraz zbadanie dwóch rodzajów walidacji krzyżowej do oceny skuteczności algorytmu poprzez obserwację metryk mówiących o jakości klasyfikatora. Ostatnim poleceniem jest porównanie otrzymanych rezultatów z wynikami klasyfikatorów z poprzednich ćwiczeń.

1.2 Algorytm *k*-najbliższych sąsiadów

Algorytm k-najbliższych sąsiadów (ang. k-nearest neighbours, k-nn) jest metodą stosowaną zarówno do rozwiązywania zadań regresji, jak i klasyfikacji. Należy ona do grupy algorytmów uczenia „leniwego", tzn. funkcja jest aproksymowana tylko lokalnie oraz proces uczenia jest wykonywany dopiero w momencie, gdy (dla zadania klasyfikacji) chcemy nowemu obiektowi nadać etykietę. Sama klasyfikacja zaś odbywa się na zasadzie głosowania wybranej liczby, najbliższych obiektowi, sąsiadów. O ilości sąsiadów, oddających głos, mówi parametr zamieszczony w nazwie algorytmu - k - jest to równocześnie wskazanie, na bazie jakiego lokalnego zbioru punktów, podejmowana będzie decyzja o etykiecie. Pozostałe dwa parametry, których odpowiedni dobór może być kluczowy dla uzyskania dobrych rezultatów to:

- `sposób głosowania` - mówi o tym jaka waga zostania nadana konkretnej instancji sąsiada - waga posłuży jako stopień w udziale podjęcia decyzji o klasyfikacji rozpatrywanego, nowego punktu. W ćwiczeniu zbadano następujące trzy podejścia ważenia głosów:
  - *uniform* - równouprawnione - wszyscy sąsiedzi oddający głos mają równy wpływ na decyzję
  - *distance* - ważone odległością - wagi odwrotnie proporcjonalne do ich odległości (bliżsi sąsiedzi modelowanego punktu mają większy wpływ na decyzję, niż ustanowieni w większej odległości)
  - *random* - losowe - wagi są przyporządkowywane sąsiadom w sposób losowy
- `definicja miary odległości` - mówi o tym w jaki sposób liczona będzie odległość rozpatrywanych sąsiadów do badanego punktu. W ćwiczeniu zbadano następujące trzy metody pomiaru odległości:
  - *manhattan* - odległość Manhattan - suma bezwzględnych różnic współ-

rzędnych; Wzór 1

$$dist_{manh}(p,q) = \sum_{i=1}^{d} |p_i - q_i| \tag{1}$$

– *euclidean* - odległość euklidesowa - „zwykła" liniowa odległość pomiędzy dwoma punktami w przestrzeni euklidesowej; Wzór 2

$$dist_{eucl}(p,q) = \sqrt{\sum_{i=1}^{d} (p_i - q_i)^2} \tag{2}$$

– *chebyshev* - odległość Czebyszewa - największa różnica współrzędnych; Wzór 3

$$dist_{cheb}(p,q) = \max_{i=1..d} |p_i - q_i| \tag{3}$$

Implementacja

Wybranym przeze mnie środowiskiem jest język *Python*. Do implementacji algorytmu k-najbliższych sąsiadów, który był przedmiotem ćwiczenia, użyto biblioteki uczenia maszynowego *sklearn*. Konkretniej, wykorzystano klasyfikator `KNeighborsClassifier` z pakietu *sklearn.neighbors*. Pozwala on na ustawienie wszystkich interesujących nas parametrów (wielkość wartości k - *n_neighbors*, sposób głosowania - *weights* oraz definicja miary odległości - *metric*) W kodzie wykorzystano także takie biblioteki jak: *pandas*, *matplotlib*, *numpy*, *scipy* oraz *seaborn*.

2.1 Ocena jakości

Do oceny jakości klasyfikatora stosuje się metryki związane, bądź wynikające z macierzy błędów (ang. Confusion matrix). Po wykonaniu zadania klasyfikacji, możliwe jest zbadanie prawidłowości dopasowań, analizując rzeczywiste etykiety. Właśnie w tym celu posługujemy się tablicą pomyłek, która zestawia liczebność instancji prawidłowo i nieprawidłowo oznaczonych przez klasyfikator. Podstawowe takie metryki to:

- *Accuracy* - dokładność - jak dobre są ogólne predykcje klasyfikatora. Jest to stosunek dobrze oznaczonych klas do wszystkich oznaczeń.
- *Precission* - precyzja - mówi o precyzyjności klasyfikatora. Jest to zdolność klasyfikatora do nie oznaczania negatywnych próbek jako pozytywne.
- *Recall* - czułość - jak dobrze klasyfikator radzi sobie ze znajdywaniem pozytywnych próbek. Jest to stosunek próbek dobrze zaklasyfikowanych jako pozytywne przez wszystkie rzeczywiście pozytywne.
- *F1-score* - oznaczane jako *FSC* - jest to w istocie średnia harmoniczna czułości oraz precyzji.

Do badania większości testów posłuży głownie miara *FSC*. Jest to często wykorzystywana miara, która pozwoli na wymiernie dobre porównanie różnych podejść. W implementacji wykorzystany został pakiet *sklearn.metrics*, który udostępnia wszystkie powyżej wymienione metryki.

2.2 Ujednolicenie danych

Ujednolicenie danych polega na zastosowaniu operacji takich jak standaryzacja czy normalizacja. Pierwsza z metod opiera się na przemapowaniu wartości atrybutów w taki sposób, by ich średnia skupiała się na wartości 0, a odchylenie standardowe wynosiło 1. Z kolei normalizacja zamienia wielkości atrybutów w taki sposób, aby instancje zachowały wzajemne proporcje między wartościami atrybutów, jednak znalazły się w zakresie $[0, 1]$. Ustalono, że na potrzeby wszystkich testów, atrybuty zbirów danych będą znormalizowane (za pomocą funkcji dostępnej w pakiecie *sklearn.preprocessing*).

2.3 Kroswalidacja

Kroswalidacja (sprawdzian krzyżowy) jest stosowana w celu lepszej oceny działania modeli, takich jak klasyfikator k-najbliższych sąsiadów. Procedura determinuje podział danych na treningowe i testowe. A jej wykorzystanie tłumaczy się unikaniem overfittingu.

W realizacja zadania wykorzystano dwie metody, dostępne w pakiecie *sklearn.model_selection*. Są nimi:

- `KFold` - kroswalidacja - gdzie całkowity zbiór danych jest dzielony na K równolicznych podzbiorów (parametr *n_splits*), z których kolejno każdy jest traktowany jako zbiór testowy, kiedy model trenowany jest na połączonych (*K-1*) pozostałych podzbiorach. Otrzymuje się wówczas K wyników, z których następnie liczona jest średnia.
- `StratifiedKFold` - kroswalidacja stratyfikowana - jej działanie jest analogiczne do powyżej opisanego, wariantu zwykłego. Różnica polega na tym, że całkowity zbiór jest dzielony (na K części, także parametrem *n_splits*) w miarę możliwości na podzbiory o proporcjonalnym rozkładzie klas - zgodnie z proporcją istniejącą w całościowym zbiorze.

W badaniach ustawiano parametr *shuffle = False* w celu wymiernego zbadania różnicy między kroswalidacją podstawową, a stratyfikowaną. Testowany był także wpływ ilości foldów (K) na otrzymywane wyniki, dla obu metod.

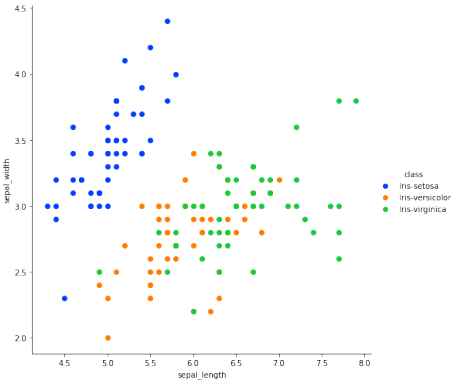
2.4 Zbiory danych

Do walidacji zaimplementowanych funkcjonalności wykorzystano łącznie cztery zbiory danych. Te zbiory, to:

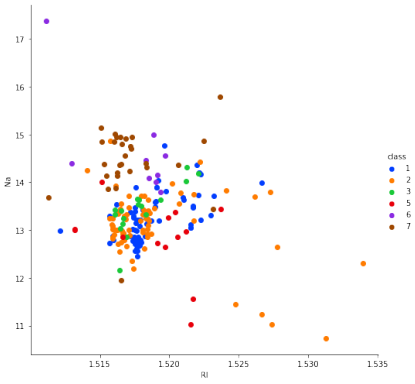
- *iris* - gdzie liczba atrybutów to 4, a klas 3. Zbiór przedstawia odmiany kwiatów Iris, a atrybuty odpowiadają ich cechom.
- *glass* - gdzie liczba atrybutów to 9, a klas 7. Zbiór przedstawia typy szkła, gdzie atrybuty opisują skład chemiczny.

- *wine* - gdzie liczba atrybutów to 13, a klas 3. Zbiór ten zawiera dane o analizie chemicznej win z tego samego regionu, jednak z 3 różnych upraw.
- *seeds* - gdzie liczba atrybutów wynosi 7, a klas 3. Zbiór ten zawiera miary geometrycznych własności 3 różnych odmian ziaren pszenicy.

Na Rysunkach 1, 2, 3, 4 przedstawiono przykładowe rozkłady klas względem wybranych atrybutów. Z kolei w Tabelach 1, 2, 3, 4 zestawiono rozkład klas dla wszystkich badanych zbiorów.



Rysunek 1: Iris



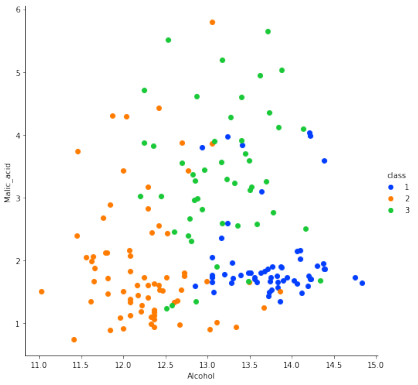
Rysunek 2: Glass

Klasa	Instancje	% rekordów
Setosa	50	33 (%)
Versicolor	50	33 (%)
Virginica	50	33 (%)

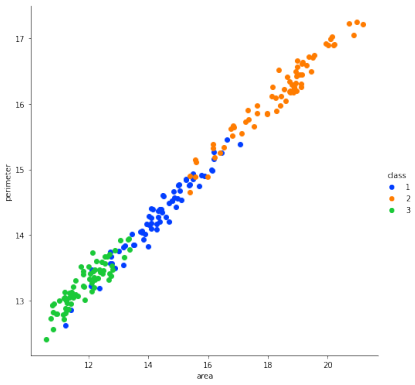
Tablica 1: Klasy zbioru *Iris*

Klasa	Instancje	% rekordów
1	70	33 (%)
2	76	36 (%)
3	17	8 (%)
4	0	0 (%)
5	13	6 (%)
6	9	4 (%)
7	29	13 (%)

Tablica 2: Klasy zbioru *Glass*



Rysunek 3: Wine



Rysunek 4: Seeds

Klasa	Instancje	% rekordów
1	59	33 (%)
2	71	40 (%)
3	48	27 (%)

Tablica 3: Klasy zbioru *Wine*

Klasa	Instancje	% rekordów
1	70	33 (%)
2	70	33 (%)
3	70	33 (%)

Tablica 4: Klasy zbioru *Seeds*

Wyniki eksperymentów

3.1 Zbiór Iris

Poniższe tabelki (Tabela 5, 6, 7, 8, 9, 10) przedstawiają wszystkie przeprowadzone testy na zbiorze *Iris*, z podziałem na badane wartości parametru k, zastosowane schematy głosowania oraz różne definicje miary odległości. Wszystkie testy zrealizowane zostały przy pomocy sprawdzianów krzyżowych (kroswalidacja „zwykła" oraz stratyfikowana). W kolejnych kolumnach, każdej z wyżej wymienionych Tabel, umieszczone zostały odczyty metryk (o których wspomniano w poprzedniej sekcji).

Dodatkowo na rysunkach: 5 oraz 6 przedstawione są wykresy prezentujące graficznie wpływ parametru k, algorytmu k-nn, na osiągane wartości metryki *FSC*.

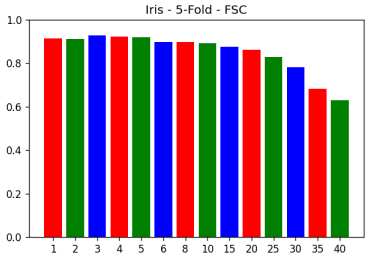
Dla zbioru *Iris*, przy wykorzystaniu algorytmu k-najbliższych sąsiadów, większość z testów osiąga dobre wyniki. Można zaobserwować, że stosowanie kroswalidacji stratyfikowanej przekłada się na wyższe wskazania metryk. Poza tym, zwykle obliczenie euklidesowej miary odległości jest lepszym rozwiązaniem.

k	*ACC*	*PREC*	*REC*	*FSC*
2-Fold				
1	0.32	0.108	0.32	0.162
2	0.327	0.11	0.327	0.164
3	0.313	0.107	0.313	0.159
4	0.327	0.11	0.327	0.164
5	0.32	0.108	0.32	0.162
6	0.32	0.108	0.32	0.162
8	0.32	0.108	0.32	0.162
10	0.313	0.107	0.313	0.159
15	0.307	0.105	0.307	0.156
20	0.293	0.102	0.293	0.151
25	0.3	0.103	0.3	0.154
30	0.293	0.102	0.293	0.151
35	0.28	0.099	0.28	0.146
40	0.28	0.099	0.28	0.146
5-Fold				
1	0.913	0.913	0.913	0.913
2	0.907	0.91	0.907	0.906
3	0.927	0.927	0.927	0.927
4	0.92	0.922	0.92	0.92
5	0.92	0.92	0.92	0.92
6	0.893	0.897	0.893	0.893
8	0.893	0.897	0.893	0.893
10	0.887	0.891	0.887	0.886
15	0.873	0.875	0.873	0.873
20	0.86	0.862	0.86	0.86
25	0.827	0.828	0.827	0.826
30	0.773	0.78	0.773	0.77
35	0.68	0.681	0.68	0.677
40	0.627	0.629	0.627	0.622
10-Fold				
1	0.947	0.947	0.947	0.947
2	0.94	0.943	0.94	0.94
3	0.933	0.933	0.933	0.933
4	0.94	0.941	0.94	0.94
5	0.953	0.953	0.953	0.953
6	0.953	0.954	0.953	0.953
8	0.94	0.94	0.94	0.94
10	0.94	0.94	0.94	0.94
15	0.947	0.947	0.947	0.947
20	0.92	0.92	0.92	0.92
25	0.92	0.922	0.92	0.92
30	0.92	0.922	0.92	0.92
35	0.893	0.897	0.893	0.893
40	0.893	0.899	0.893	0.893

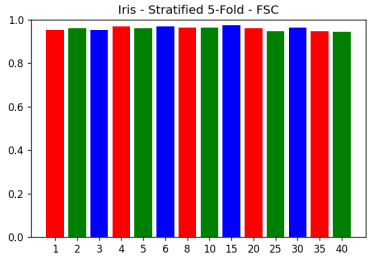
Tablica 5: Wpływ k - zbiór *Iris* - kroswalidacja „zwykła”

k	*ACC*	*PREC*	*REC*	*FSC*
Stratified 2-Fold				
1	0.927	0.928	0.927	0.927
2	0.94	0.945	0.94	0.94
3	0.953	0.953	0.953	0.953
4	0.967	0.968	0.967	0.967
5	0.953	0.953	0.953	0.953
6	0.953	0.953	0.953	0.953
8	0.96	0.962	0.96	0.96
10	0.96	0.962	0.96	0.96
15	0.953	0.953	0.953	0.953
20	0.947	0.948	0.947	0.947
25	0.933	0.937	0.933	0.933
30	0.887	0.898	0.887	0.885
35	0.88	0.886	0.88	0.879
40	0.873	0.88	0.873	0.872
Stratified 5-Fold				
1	0.953	0.953	0.953	0.953
2	0.96	0.96	0.96	0.96
3	0.953	0.953	0.953	0.953
4	0.967	0.968	0.967	0.967
5	0.96	0.96	0.96	0.96
6	0.967	0.968	0.967	0.967
8	0.96	0.96	0.96	0.96
10	0.96	0.96	0.96	0.96
15	0.973	0.974	0.973	0.973
20	0.96	0.96	0.96	0.96
25	0.947	0.947	0.947	0.947
30	0.96	0.962	0.96	0.96
35	0.947	0.947	0.947	0.947
40	0.94	0.943	0.94	0.94
Stratified 10-Fold				
1	0.953	0.953	0.953	0.953
2	0.96	0.96	0.96	0.96
3	0.953	0.953	0.953	0.953
4	0.96	0.96	0.96	0.96
5	0.953	0.953	0.953	0.953
6	0.967	0.968	0.967	0.967
8	0.96	0.96	0.96	0.96
10	0.96	0.96	0.96	0.96
15	0.967	0.967	0.967	0.967
20	0.953	0.953	0.953	0.953
25	0.947	0.947	0.947	0.947
30	0.973	0.975	0.973	0.973
35	0.96	0.962	0.96	0.96
40	0.947	0.948	0.947	0.947

Tablica 6: Wpływ k - zbiór *Iris* - kroswalidacja stratyfikowana



Rysunek 5: Wpływ k - zbiór *Iris* - kroswalidacja „zwykła”



Rysunek 6: Wpływ k - zbiór *Iris* - kroswalidacja stratyfikowana

vs	*ACC*	*PREC*	*REC*	*FSC*
2-Fold				
dist	0.32	0.108	0.32	0.162
uni	0.32	0.108	0.32	0.162
rand	0.22	0.103	0.22	0.14
5-Fold				
dist	0.92	0.92	0.92	0.92
uni	0.92	0.92	0.92	0.92
rand	0.62	0.69	0.62	0.599
10-Fold				
dist	0.953	0.953	0.953	0.953
uni	0.953	0.953	0.953	0.953
rand	0.647	0.734	0.647	0.629

Tablica 7: Wpływ sposobu głosowania - zbiór *Iris* - kroswalidacja „zwykła”

vs	*ACC*	*PREC*	*REC*	*FSC*
Stratified 2-Fold				
dist	0.953	0.953	0.953	0.953
uni	0.953	0.953	0.953	0.953
rand	0.66	0.739	0.66	0.648
Stratified 5-Fold				
dist	0.953	0.953	0.953	0.953
uni	0.96	0.96	0.96	0.96
rand	0.607	0.685	0.607	0.583
Stratified 10-Fold				
dist	0.953	0.953	0.953	0.953
uni	0.953	0.953	0.953	0.953
rand	0.64	0.77	0.64	0.63

Tablica 8: Wpływ sposobu głosowania - zbiór *Iris* - kroswalidacja stratyfikowana

mtr	*ACC*	*PREC*	*REC*	*FSC*
2-Fold				
man	0.313	0.107	0.313	0.159
eucl	0.32	0.108	0.32	0.162
cheb	0.313	0.107	0.313	0.159
5-Fold				
man	0.927	0.927	0.927	0.927
eucl	0.92	0.92	0.92	0.92
cheb	0.913	0.914	0.913	0.913
10-Fold				
man	0.947	0.947	0.947	0.947
eucl	0.953	0.953	0.953	0.953
cheb	0.94	0.94	0.94	0.94

Tablica 9: Wpływ miary odl. - zbiór *Iris* - kroswalidacja „zwykła”

mtr	*ACC*	*PREC*	*REC*	*FSC*
Stratified 2-Fold				
man	0.947	0.947	0.947	0.947
eucl	0.953	0.953	0.953	0.953
cheb	0.947	0.947	0.947	0.947
Stratified 5-Fold				
man	0.947	0.947	0.947	0.947
eucl	0.96	0.96	0.96	0.96
cheb	0.96	0.962	0.96	0.96
Stratified 10-Fold				
man	0.947	0.947	0.947	0.947
eucl	0.953	0.953	0.953	0.953
cheb	0.96	0.96	0.96	0.96

Tablica 10: Wpływ miary odl. - zbiór *Iris* - kroswalidacja stratyfikowana

3.2 Zbiór Glass

Poniższe tabelki (Tabela 11, 12, 13, 14, 15, 16) przedstawiają wszystkie przeprowadzone testy na zbiorze *Glass*, z podziałem na badane wartości parametru k, zastosowane schematy głosowania oraz różne definicje miary odległości. Wszystkie testy zrealizowane zostały przy pomocy sprawdzianów krzyżowych (kroswalidacja „zwykła" oraz stratyfikowana). W kolejnych kolumnach, każdej z wyżej wymienionych Tabel, umieszczone zostały odczyty metryk (o których wspomniano w poprzedniej sekcji).

Dodatkowo na rysunkach: 7 oraz 8 przedstawione są wykresy prezentujące graficznie wpływ parametru k, algorytmu k-nn, na osiągane wartości metryki *FSC*.

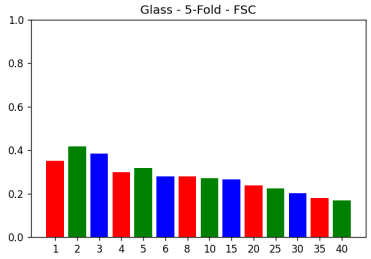
Dla zbioru *Glass*, przy wykorzystaniu algorytmu k-najbliższych sąsiadów, zauważamy duży przyrost jakości klasyfikacji przy zastosowaniu kroswalidacji stratyfikowanej, w zestawieniu z jej podstawową odmianą. Raczej niższe ustawienia parametru k przekładają się na wyższe wskazania metryk. Poza tym, zwykle obliczenie miary odległości *Manhattan* oraz system głosowania *distance* jest lepszym rozwiązaniem.

k	ACC	$PREC$	REC	FSC
2-Fold				
1	0.201	0.063	0.094	0.075
2	0.224	0.066	0.105	0.081
3	0.21	0.066	0.099	0.079
4	0.215	0.063	0.101	0.078
5	0.21	0.06	0.099	0.075
6	0.21	0.065	0.099	0.079
8	0.206	0.058	0.096	0.073
10	0.206	0.056	0.096	0.071
15	0.206	0.054	0.096	0.069
20	0.196	0.053	0.092	0.067
25	0.201	0.053	0.094	0.068
30	0.182	0.048	0.086	0.061
35	0.178	0.046	0.083	0.059
40	0.168	0.045	0.079	0.057
5-Fold				
1	0.35	0.178	0.223	0.19
2	0.416	0.188	0.257	0.211
3	0.383	0.172	0.219	0.189
4	0.299	0.132	0.156	0.142
5	0.318	0.143	0.176	0.155
6	0.28	0.122	0.147	0.133
8	0.28	0.12	0.147	0.131
10	0.271	0.101	0.132	0.114
15	0.266	0.097	0.129	0.11
20	0.238	0.085	0.116	0.097
25	0.224	0.08	0.109	0.092
30	0.201	0.072	0.098	0.083
35	0.164	0.057	0.081	0.067
40	0.164	0.058	0.081	0.067
10-Fold				
1	0.519	0.415	0.405	0.408
2	0.556	0.407	0.403	0.397
3	0.551	0.393	0.386	0.382
4	0.514	0.322	0.319	0.313
5	0.523	0.318	0.324	0.315
6	0.491	0.289	0.304	0.291
8	0.425	0.244	0.231	0.225
10	0.43	0.231	0.229	0.22
15	0.397	0.203	0.209	0.2
20	0.369	0.185	0.191	0.183
25	0.327	0.165	0.171	0.163
30	0.304	0.13	0.156	0.139
35	0.28	0.118	0.145	0.127
40	0.257	0.109	0.134	0.117

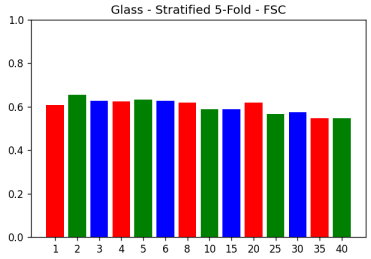
Tablica 11: Wpływ k - zbiór *Glass* - kroswalidacja „zwykła”

k	ACC	$PREC$	REC	FSC
Stratified 2-Fold				
1	0.5	0.487	0.449	0.458
2	0.589	0.583	0.479	0.495
3	0.57	0.491	0.454	0.455
4	0.57	0.509	0.436	0.438
5	0.579	0.427	0.433	0.421
6	0.584	0.447	0.441	0.436
8	0.556	0.403	0.407	0.397
10	0.584	0.409	0.415	0.402
15	0.565	0.326	0.371	0.342
20	0.533	0.271	0.343	0.302
25	0.514	0.26	0.333	0.291
30	0.486	0.248	0.321	0.278
35	0.472	0.25	0.293	0.269
40	0.435	0.239	0.254	0.241
Stratified 5-Fold				
1	0.607	0.59	0.6	0.594
2	0.626	0.653	0.573	0.577
3	0.626	0.587	0.536	0.529
4	0.617	0.625	0.475	0.486
5	0.631	0.484	0.48	0.477
6	0.626	0.471	0.478	0.47
8	0.617	0.455	0.468	0.457
10	0.589	0.378	0.411	0.39
15	0.589	0.446	0.417	0.406
20	0.617	0.539	0.427	0.42
25	0.565	0.328	0.371	0.34
30	0.575	0.292	0.362	0.321
35	0.547	0.281	0.348	0.308
40	0.547	0.278	0.348	0.308
Stratified 10-Fold				
1	0.65	0.613	0.638	0.622
2	0.645	0.632	0.577	0.585
3	0.659	0.647	0.62	0.615
4	0.668	0.64	0.587	0.579
5	0.659	0.529	0.553	0.539
6	0.664	0.577	0.536	0.536
8	0.626	0.471	0.479	0.469
10	0.621	0.45	0.454	0.445
15	0.579	0.373	0.396	0.378
20	0.631	0.401	0.42	0.402
25	0.603	0.476	0.4	0.382
30	0.617	0.477	0.396	0.365
35	0.584	0.295	0.366	0.324
40	0.579	0.291	0.364	0.321

Tablica 12: Wpływ k - zbiór *Glass* - kroswalidacja stratyfikowana



Rysunek 7: Wpływ k - zbiór *Glass* - kroswalidacja „zwykła”



Rysunek 8: Wpływ k - zbiór *Glass* - kroswalidacja stratyfikowana

vs	*ACC*	*PREC*	*REC*	*FSC*
2-Fold				
dist	0.206	0.063	0.096	0.076
uni	0.21	0.06	0.099	0.075
rand	0.164	0.057	0.077	0.065
5-Fold				
dist	0.322	0.156	0.199	0.169
uni	0.318	0.143	0.176	0.155
rand	0.294	0.151	0.198	0.164
10-Fold				
dist	0.551	0.394	0.39	0.384
uni	0.523	0.318	0.324	0.315
rand	0.397	0.329	0.267	0.275

Tablica 13: Wpływ sposobu głosowania - zbiór *Glass* - kroswalidacja „zwykła”

vs	*ACC*	*PREC*	*REC*	*FSC*
Stratified 2-Fold				
dist	0.561	0.465	0.431	0.428
uni	0.579	0.427	0.433	0.421
rand	0.402	0.406	0.306	0.329
Stratified 5-Fold				
dist	0.631	0.539	0.535	0.528
uni	0.631	0.484	0.48	0.477
rand	0.397	0.337	0.262	0.277
Stratified 10-Fold				
dist	0.696	0.641	0.652	0.632
uni	0.659	0.529	0.553	0.539
rand	0.5	0.503	0.418	0.439

Tablica 14: Wpływ sposobu głosowania - zbiór *Glass* - kroswalidacja stratyfikowana

mtr	*ACC*	*PREC*	*REC*	*FSC*
2-Fold				
man	0.215	0.066	0.101	0.079
eucl	0.21	0.06	0.099	0.075
cheb	0.206	0.069	0.096	0.081
5-Fold				
man	0.332	0.152	0.193	0.167
eucl	0.318	0.143	0.176	0.155
cheb	0.262	0.11	0.137	0.121
10-Fold				
man	0.579	0.353	0.379	0.36
eucl	0.523	0.318	0.324	0.315
cheb	0.481	0.294	0.292	0.286

Tablica 15: Wpływ miary odl. - zbiór *Glass* - kroswalidacja „zwykła”

mtr	*ACC*	*PREC*	*REC*	*FSC*
Stratified 2-Fold				
man	0.612	0.463	0.474	0.462
eucl	0.579	0.427	0.433	0.421
cheb	0.519	0.327	0.366	0.343
Stratified 5-Fold				
man	0.664	0.541	0.564	0.547
eucl	0.631	0.484	0.48	0.477
cheb	0.589	0.494	0.449	0.455
Stratified 10-Fold				
man	0.678	0.574	0.613	0.589
eucl	0.659	0.529	0.553	0.539
cheb	0.607	0.557	0.488	0.501

Tablica 16: Wpływ miary odl. - zbiór *Glass* - kroswalidacja stratyfikowana

3.3 Zbiór Wine

Poniższe tabelki (Tabela 17, 18, 19, 20, 21, 22) przedstawiają wszystkie przeprowadzone testy na zbiorze *Wine*, z podziałem na badane wartości parametru k, zastosowane schematy głosowania oraz różne definicje miary odległości. Wszystkie testy zrealizowane zostały przy pomocy sprawdzianów krzyżowych (kroswalidacja „zwykła" oraz stratyfikowana). W kolejnych kolumnach, każdej z wyżej wymienionych Tabel, umieszczone zostały odczyty metryk (o których wspomniano w poprzedniej sekcji).

Dodatkowo na rysunkach: 9 oraz 10 przedstawione są wykresy prezentujące graficznie wpływ parametru k, algorytmu k-nn, na osiągane wartości metryki *FSC*.

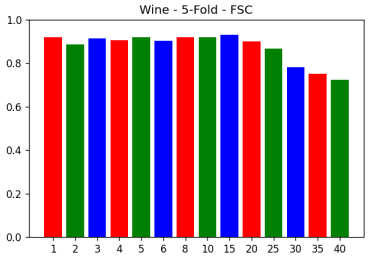
Dla zbioru *Wine*, przy wykorzystaniu algorytmu k-najbliższych sąsiadów, zauważamy spory przyrost jakości klasyfikacji przy zastosowaniu kroswalidacji stratyfikowanej, w zestawieniu z jej podstawową odmianą przy niskich ilościach foldów. Raczej wyższe ustawienia parametru k przekładają się na niewiele lepsze wskazania metryk. Poza tym, zwykle obliczenie miary odległości *Manhattan* daje bardziej obiecujące rezultaty.

k	*ACC*	*PREC*	*REC*	*FSC*
2-Fold				
1	0.354	0.125	0.296	0.176
2	0.354	0.126	0.296	0.176
3	0.354	0.125	0.296	0.176
4	0.348	0.126	0.291	0.176
5	0.348	0.124	0.291	0.174
6	0.337	0.123	0.282	0.171
8	0.337	0.122	0.282	0.17
10	0.343	0.124	0.286	0.173
15	0.331	0.121	0.277	0.168
20	0.326	0.124	0.272	0.17
25	0.331	0.125	0.277	0.173
30	0.303	0.122	0.254	0.164
35	0.32	0.131	0.268	0.176
40	0.315	0.14	0.263	0.183
5-Fold				
1	0.91	0.908	0.918	0.911
2	0.882	0.886	0.886	0.882
3	0.904	0.904	0.912	0.905
4	0.899	0.903	0.906	0.901
5	0.91	0.909	0.919	0.911
6	0.893	0.895	0.903	0.895
8	0.91	0.91	0.919	0.911
10	0.91	0.912	0.918	0.913
15	0.921	0.922	0.93	0.924
20	0.893	0.894	0.899	0.894
25	0.854	0.866	0.843	0.848
30	0.747	0.781	0.714	0.702
35	0.725	0.75	0.689	0.674
40	0.691	0.724	0.658	0.646
10-Fold				
1	0.921	0.92	0.93	0.923
2	0.921	0.922	0.928	0.922
3	0.944	0.941	0.949	0.944
4	0.921	0.923	0.93	0.923
5	0.938	0.936	0.946	0.938
6	0.944	0.943	0.951	0.945
8	0.938	0.937	0.946	0.939
10	0.938	0.937	0.945	0.94
15	0.949	0.948	0.957	0.951
20	0.955	0.953	0.962	0.956
25	0.961	0.958	0.967	0.961
30	0.944	0.944	0.952	0.946
35	0.949	0.949	0.957	0.951
40	0.944	0.943	0.95	0.945

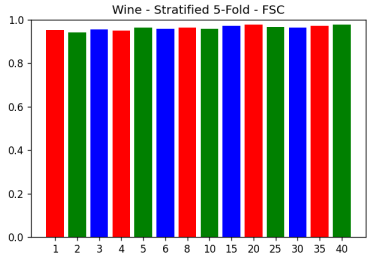
Tablica 17: Wpływ k - zbiór *Wine* - kroswalidacja „zwykła”

k	*ACC*	*PREC*	*REC*	*FSC*
Stratified 2-Fold				
1	0.944	0.943	0.953	0.945
2	0.933	0.935	0.941	0.935
3	0.938	0.937	0.948	0.939
4	0.938	0.939	0.948	0.94
5	0.933	0.933	0.944	0.934
6	0.944	0.945	0.953	0.946
8	0.933	0.935	0.944	0.935
10	0.944	0.945	0.953	0.946
15	0.966	0.964	0.972	0.967
20	0.961	0.958	0.967	0.961
25	0.961	0.959	0.967	0.962
30	0.955	0.954	0.962	0.956
35	0.955	0.954	0.962	0.956
40	0.961	0.962	0.967	0.962
Stratified 5-Fold				
1	0.944	0.943	0.953	0.945
2	0.933	0.933	0.937	0.934
3	0.949	0.947	0.955	0.95
4	0.938	0.939	0.946	0.94
5	0.955	0.952	0.962	0.955
6	0.949	0.95	0.958	0.951
8	0.955	0.954	0.962	0.956
10	0.949	0.949	0.957	0.951
15	0.961	0.96	0.966	0.962
20	0.972	0.971	0.977	0.973
25	0.961	0.959	0.967	0.962
30	0.944	0.945	0.953	0.946
35	0.966	0.964	0.972	0.967
40	0.972	0.971	0.977	0.973
Stratified 10-Fold				
1	0.949	0.949	0.955	0.951
2	0.944	0.945	0.949	0.945
3	0.949	0.947	0.955	0.95
4	0.949	0.95	0.955	0.951
5	0.961	0.958	0.967	0.961
6	0.955	0.954	0.962	0.956
8	0.961	0.959	0.967	0.962
10	0.966	0.966	0.972	0.968
15	0.966	0.964	0.971	0.967
20	0.978	0.976	0.981	0.978
25	0.972	0.969	0.977	0.972
30	0.966	0.964	0.972	0.967
35	0.966	0.964	0.972	0.967
40	0.966	0.964	0.972	0.967

Tablica 18: Wpływ k - zbiór *Wine* - kroswalidacja stratyfikowana



Rysunek 9: Wpływ k - zbiór *Wine* - kroswalidacja „zwykła”



Rysunek 10: Wpływ k - zbiór *Wine* - kroswalidacja stratyfikowana

vs	*ACC*	*PREC*	*REC*	*FSC*
2-Fold				
dist	0.348	0.124	0.291	0.174
uni	0.348	0.124	0.291	0.174
rand	0.27	0.126	0.225	0.162
5-Fold				
dist	0.91	0.909	0.919	0.911
uni	0.91	0.909	0.919	0.911
rand	0.596	0.673	0.598	0.581
10-Fold				
dist	0.938	0.936	0.946	0.938
uni	0.938	0.936	0.946	0.938
rand	0.629	0.748	0.618	0.608

Tablica 19: Wpływ sposobu głosowania - zbiór *Wine* - kroswalidacja „zwykła”

vs	*ACC*	*PREC*	*REC*	*FSC*
Stratified 2-Fold				
dist	0.944	0.943	0.953	0.945
uni	0.933	0.933	0.944	0.934
rand	0.612	0.755	0.629	0.602
Stratified 5-Fold				
dist	0.955	0.952	0.962	0.955
uni	0.955	0.952	0.962	0.955
rand	0.596	0.771	0.605	0.59
Stratified 10-Fold				
dist	0.961	0.958	0.967	0.961
uni	0.961	0.958	0.967	0.961
rand	0.573	0.731	0.573	0.548

Tablica 20: Wpływ sposobu głosowania - zbiór *Wine* - kroswalidacja stratyfikowana

mtr	*ACC*	*PREC*	*REC*	*FSC*
2-Fold				
man	0.354	0.125	0.296	0.176
eucl	0.348	0.124	0.291	0.174
cheb	0.331	0.149	0.277	0.194
5-Fold				
man	0.921	0.923	0.931	0.924
eucl	0.91	0.909	0.919	0.911
cheb	0.837	0.84	0.844	0.839
10-Fold				
man	0.949	0.947	0.955	0.95
eucl	0.938	0.936	0.946	0.938
cheb	0.893	0.894	0.903	0.895

Tablica 21: Wpływ miary odl. - zbiór *Wine* - kroswalidacja „zwykła”

mtr	*ACC*	*PREC*	*REC*	*FSC*
Stratified 2-Fold				
man	0.944	0.944	0.953	0.945
eucl	0.933	0.933	0.944	0.934
cheb	0.91	0.911	0.924	0.912
Stratified 5-Fold				
man	0.966	0.966	0.972	0.968
eucl	0.955	0.952	0.962	0.955
cheb	0.921	0.923	0.93	0.923
Stratified 10-Fold				
man	0.972	0.971	0.977	0.973
eucl	0.961	0.958	0.967	0.961
cheb	0.916	0.915	0.923	0.917

Tablica 22: Wpływ miary odl. - zbiór *Wine* - kroswalidacja stratyfikowana

3.4 Zbiór Seeds

Poniższe tabelki (Tabela 23, 24, 25, 26, 27, 28) przedstawiają wszystkie przeprowadzone testy na zbiorze *Seeds*, z podziałem na badane wartości parametru k, zastosowane schematy głosowania oraz różne definicje miary odległości. Wszystkie testy zrealizowane zostały przy pomocy sprawdzianów krzyżowych (kroswalidacja „zwykła" oraz stratyfikowana). W kolejnych kolumnach, każdej z wyżej wymienionych Tabel, umieszczone zostały odczyty metryk (o których wspomniano w poprzedniej sekcji).

Dodatkowo na rysunkach: 11 oraz 12 przedstawione są wykresy prezentujące graficznie wpływ parametru k, algorytmu k-nn, na osiągane wartości metryki *FSC*.

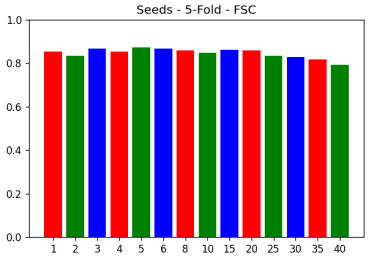
Dla zbioru *Seeds*, przy wykorzystaniu algorytmu k-najbliższych sąsiadów, zauważamy spory przyrost jakości klasyfikacji przy zastosowaniu kroswalidacji stratyfikowanej, w zestawieniu z jej podstawową odmianą przy niskich ilościach foldów. Ustawienia parametru k w okolicach 4-6 przekładają się na najlepsze wskazania metryk. Poza tym, zwykle obliczenie miary odległości euklidesowej daje bardziej obiecujące rezultaty.

k	*ACC*	*PREC*	*REC*	*FSC*
2-Fold				
1	0.31	0.223	0.31	0.259
2	0.305	0.209	0.305	0.248
3	0.31	0.226	0.31	0.261
4	0.31	0.217	0.31	0.255
5	0.314	0.237	0.314	0.27
6	0.314	0.234	0.314	0.268
8	0.314	0.237	0.314	0.27
10	0.314	0.237	0.314	0.27
15	0.314	0.244	0.314	0.275
20	0.295	0.258	0.295	0.276
25	0.295	0.265	0.295	0.279
30	0.29	0.261	0.29	0.275
35	0.29	0.271	0.29	0.28
40	0.29	0.279	0.29	0.284
5-Fold				
1	0.852	0.851	0.852	0.852
2	0.824	0.833	0.824	0.826
3	0.867	0.866	0.867	0.865
4	0.852	0.854	0.852	0.853
5	0.871	0.871	0.871	0.87
6	0.867	0.866	0.867	0.866
8	0.857	0.856	0.857	0.857
10	0.848	0.847	0.848	0.847
15	0.862	0.861	0.862	0.861
20	0.857	0.857	0.857	0.857
25	0.833	0.833	0.833	0.833
30	0.824	0.828	0.824	0.825
35	0.814	0.816	0.814	0.815
40	0.79	0.793	0.79	0.791
10-Fold				
1	0.876	0.877	0.876	0.876
2	0.843	0.855	0.843	0.845
3	0.881	0.881	0.881	0.88
4	0.895	0.896	0.895	0.896
5	0.89	0.89	0.89	0.89
6	0.905	0.906	0.905	0.905
8	0.876	0.876	0.876	0.876
10	0.886	0.885	0.886	0.885
15	0.881	0.881	0.881	0.88
20	0.876	0.876	0.876	0.876
25	0.876	0.876	0.876	0.876
30	0.876	0.877	0.876	0.876
35	0.871	0.872	0.871	0.871
40	0.867	0.867	0.867	0.867

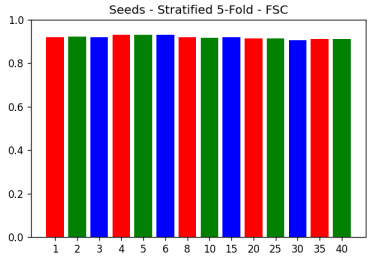
Tablica 23: Wpływ k - zbiór *Seeds* - kroswalidacja „zwykła”

k	*ACC*	*PREC*	*REC*	*FSC*
Stratified 2-Fold				
1	0.919	0.919	0.919	0.919
2	0.91	0.92	0.91	0.911
3	0.919	0.919	0.919	0.919
4	0.929	0.931	0.929	0.929
5	0.929	0.928	0.929	0.928
6	0.919	0.919	0.919	0.919
8	0.919	0.919	0.919	0.919
10	0.914	0.915	0.914	0.915
15	0.914	0.914	0.914	0.914
20	0.91	0.913	0.91	0.91
25	0.905	0.906	0.905	0.905
30	0.9	0.902	0.9	0.901
35	0.9	0.9	0.9	0.9
40	0.9	0.902	0.9	0.901
Stratified 5-Fold				
1	0.919	0.919	0.919	0.919
2	0.914	0.921	0.914	0.916
3	0.914	0.915	0.914	0.913
4	0.919	0.92	0.919	0.919
5	0.924	0.924	0.924	0.923
6	0.929	0.929	0.929	0.929
8	0.919	0.919	0.919	0.919
10	0.91	0.909	0.91	0.909
15	0.919	0.919	0.919	0.919
20	0.91	0.91	0.91	0.909
25	0.914	0.914	0.914	0.914
30	0.905	0.906	0.905	0.905
35	0.91	0.91	0.91	0.91
40	0.91	0.911	0.91	0.91
Stratified 10-Fold				
1	0.914	0.914	0.914	0.914
2	0.89	0.893	0.89	0.891
3	0.919	0.92	0.919	0.918
4	0.929	0.929	0.929	0.929
5	0.914	0.914	0.914	0.914
6	0.929	0.929	0.929	0.928
8	0.919	0.919	0.919	0.919
10	0.924	0.924	0.924	0.924
15	0.914	0.914	0.914	0.914
20	0.91	0.91	0.91	0.909
25	0.914	0.914	0.914	0.914
30	0.905	0.905	0.905	0.905
35	0.91	0.91	0.91	0.909
40	0.91	0.913	0.91	0.91

Tablica 24: Wpływ k - zbiór *Seeds* - kroswalidacja stratyfikowana



Rysunek 11: Wpływ k - zbiór *Seeds* - kroswalidacja „zwykła”



Rysunek 12: Wpływ k - zbiór *Seeds* - kroswalidacja stratyfikowana

vs	*ACC*	*PREC*	*REC*	*FSC*
2-Fold				
dist	0.314	0.237	0.314	0.27
uni	0.314	0.237	0.314	0.27
rand	0.262	0.18	0.262	0.213
5-Fold				
dist	0.867	0.866	0.867	0.865
uni	0.871	0.871	0.871	0.87
rand	0.581	0.694	0.581	0.586
10-Fold				
dist	0.886	0.886	0.886	0.885
uni	0.89	0.89	0.89	0.89
rand	0.586	0.737	0.586	0.584

Tablica 25: Wpływ sposobu głosowania - zbiór *Seeds* - kroswalidacja „zwykła”

vs	*ACC*	*PREC*	*REC*	*FSC*
Stratified 2-Fold				
dist	0.929	0.928	0.929	0.928
uni	0.929	0.928	0.929	0.928
rand	0.59	0.742	0.59	0.591
Stratified 5-Fold				
dist	0.919	0.92	0.919	0.918
uni	0.924	0.924	0.924	0.923
rand	0.643	0.775	0.643	0.646
Stratified 10-Fold				
dist	0.91	0.909	0.91	0.909
uni	0.914	0.914	0.914	0.914
rand	0.605	0.744	0.605	0.606

Tablica 26: Wpływ sposobu głosowania - zbiór *Seeds* - kroswalidacja stratyfikowana

mtr	*ACC*	*PREC*	*REC*	*FSC*
2-Fold				
man	0.31	0.233	0.31	0.266
eucl	0.314	0.237	0.314	0.27
cheb	0.314	0.268	0.314	0.289
5-Fold				
man	0.838	0.838	0.838	0.838
eucl	0.871	0.871	0.871	0.87
cheb	0.862	0.861	0.862	0.861
10-Fold				
man	0.876	0.876	0.876	0.876
eucl	0.89	0.89	0.89	0.89
cheb	0.886	0.887	0.886	0.886

Tablica 27: Wpływ miary odl. - zbiór *Seeds* - kroswalidacja „zwykła”

mtr	*ACC*	*PREC*	*REC*	*FSC*
Stratified 2-Fold				
man	0.914	0.914	0.914	0.914
eucl	0.929	0.928	0.929	0.928
cheb	0.919	0.919	0.919	0.919
Stratified 5-Fold				
man	0.9	0.9	0.9	0.9
eucl	0.924	0.924	0.924	0.923
cheb	0.914	0.914	0.914	0.914
Stratified 10-Fold				
man	0.905	0.905	0.905	0.904
eucl	0.914	0.914	0.914	0.914
cheb	0.929	0.929	0.929	0.929

Tablica 28: Wpływ miary odl. - zbiór *Seeds* - kroswalidacja stratyfikowana

Podsumowanie

Zadanie pozwoliło na zaznajomienie się z algorytmem k-najbliższych sąsiadów oraz również z kolejnymi narzędziami służącymi pracy na danych. Pokazane zostało jak wprowadzać dodatkowe funkcjonalności oraz rozszerzenia,które przekładają się na polepszenie oceny modelu. Wreszcie uświadomiło, że nie istnieje jeden sprawdzony przepis na wszystkie problemy, a tworzenie najlepszych rozwiązań jest związane z odszukaniem właściwego podejścia.

Na koniec, w Tabeli 29 zestawiono wyniki klasyfikatora opartego o algorytm k-nn z innymi, zaimplementowanymi w poprzednich zadaniach metodami, rozwiązującymi zagadnienie przyporządkowywania etykiet.

Classifier (*params*)	*ACC*	*PREC*	*REC*	*FSC*
Zbiór Glass				
MultinominalNB (smothing)	0.52	0.68	0.60	0.50
C4.5 ($C = 0.25$, $M = 2$)	0.68	0.68	0.80	0.74
***k*-nn** ($k = 5$, *distance*, *minkowski*)	0.70	0.64	0.65	0.63
Zbiór Wine				
GaussianNB (smothing)	0.94	0.94	0.95	0.95
C4.5 ($C = 0.25$, $M = 2$)	0.94	0.97	0.97	0.97
***k*-nn** ($k = 20$, *uniform*, *minkowski*)	0.98	0.98	0.98	0.98
Zbiór Seeds				
GaussianNB (smothing)	0.89	0.90	0.90	0.90
C4.5 ($C = 0.10$, $M = 2$)	0.95	0.92	0.93	0.92
***k*-nn** ($k = 6$, *uniform*, *minkowski*)	0.93	0.93	0.93	0.93

Tablica 29: Porównanie klasyfikatorów dla zbiorów *Glass*, *Wine*, *Seeds*