Indukcyjne metody analizy danych
Ćwiczenie 4

Algorytm klasyfikacji k-najbliższych sąsiadów

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

Wrocław, 15 maja 2018r.

Spis treści

1 Wprowadzenie

1.1 Cel ćwiczenia

Celem ćwiczenia było poznanie algorytmu k-najbliższych sąsiadów (k-nn) oraz zbadanie i ocena jego działania na 4 określonych zbiorach danych. W trakcie badań należało uwzględnić różne metody głosowania, metryki odległości oraz liczbę sąsiadów. Należało również zaobserwować wpływ tych parametrów na wartości zadanych miar (Accuracy, Precision, Recall, F1-Score).

1.2 Algorytm KNN

Algorytm ten należy do grupy algorytmów uczenia *leniwego*, tzn. proces uczenia / generalizacji jest wykonywany dopiero w momencie, gdy nowy obiekt ma zostać zaklasyfikowany. Nazwa **knn** wskazuje na najważniejszy parametr tego algorytmu – k, czyli liczbę sąsiadów, którzy są uwzględniani w procesie klasyfikacji nowej instancji. Instancja jest traktowana jako punkt w przestrzeni d-wymiarowej, następnie wyznaczane jest k najbliższych sąsiadów (punktów) w tej przestrzeni (zbiór punktów treningowych), zgodnie z zadaną metryką odległości oraz sposobem głosowania. Zbadane parametry algorytmu zostały opisane poniżej:

- liczba sąsiadów – $n \in \{1..5\}, n \in \mathbb{N}$,
- metryka odległości – Euklidesowa, Manhattan, Czybyszewa,

$$d_{euclidean}(x, y) = \sqrt{(\sum_{i=1}^{d}(x_i - y_i)^2)}$$

$$d_{manhattan}(x, y) = \sum_{i=1}^{d}|x_i - y_i|$$

$$d_{chebyshev}(x, y) = \max_{i=1..d}|x_i - y_i|$$

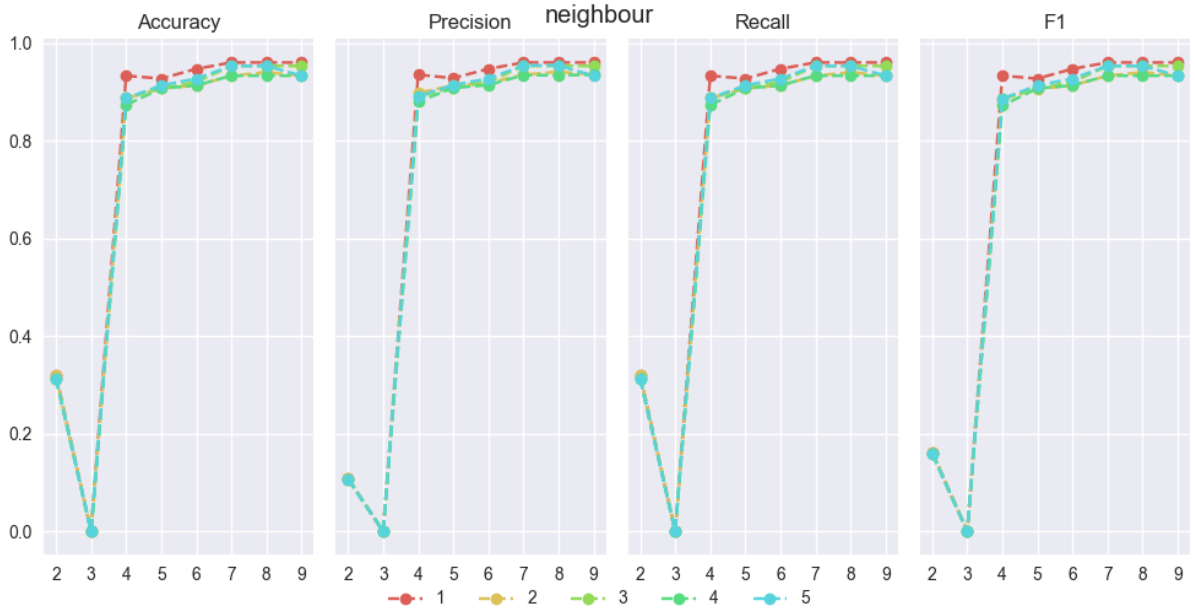
- sposób głosowania – równouprawnione (równe wagi dla odległości), ważone odległością, własne (losowe wagi odległości)

1.3 Eksperytment

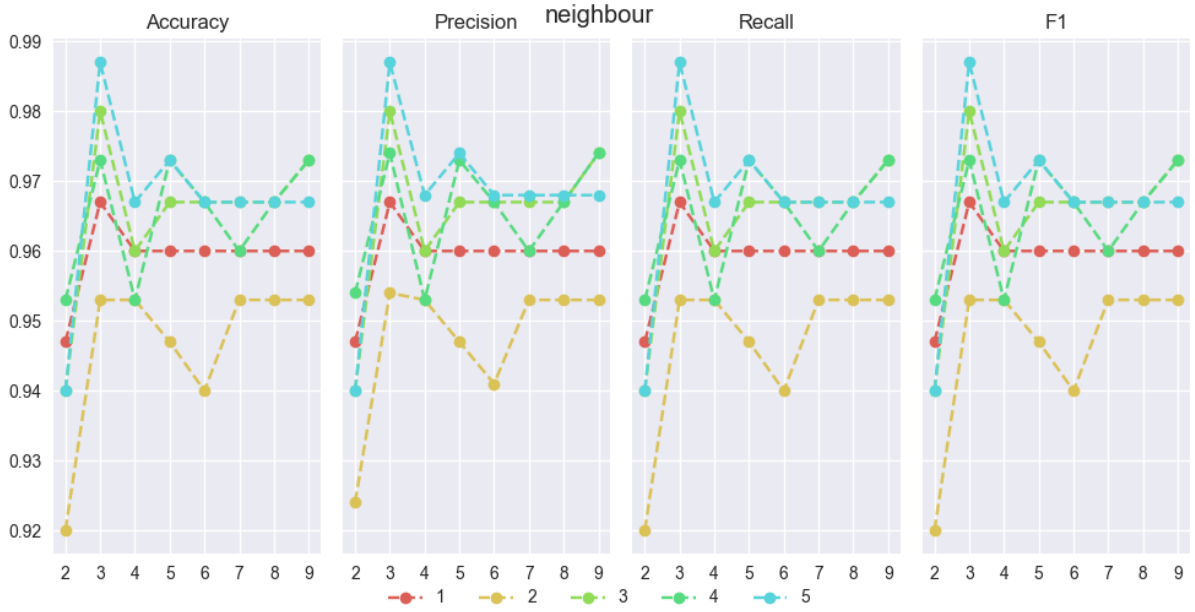
Na początku został zbadany zbiór "Iris". Dla każdego parametru i dla każdej jego wartości został stworzony klasyfikator KNN, który następnie został poddany kroswalidacji (zwykłej i stratyfikowanej) dla liczby foldów od 2 do 9 włącznie (wykresy wpływu wartości parametrów w zależności od liczby foldów). Na tej podstawie została wybrana liczba foldów do testowania kolejnych zbiorów danych. Została ona ustalona na wartość równą 5. Następnie w podobny sposób przebandane zbiory "Diabetes", "Glass", "Seeds" oraz "Wine", tyle że dla ustalonej liczby foldów. Zostały stworzone wykresy wartości miar jakości w zależności od wartości parametrów. Ostatecznie najlepsze wyniki otrzymane tutaj zostały porównane z najlepszymi wynikami dla klasyfikatorów naiwnego Bayesa oraz drzewa C4.5.

2 Wyniki

2.1 Zbiór "Iris"



Rysunek 1: Wykres wartości miar dla zbioru "Iris" dla różnej liczby sąsiadów (kroswalidacja zwykła).



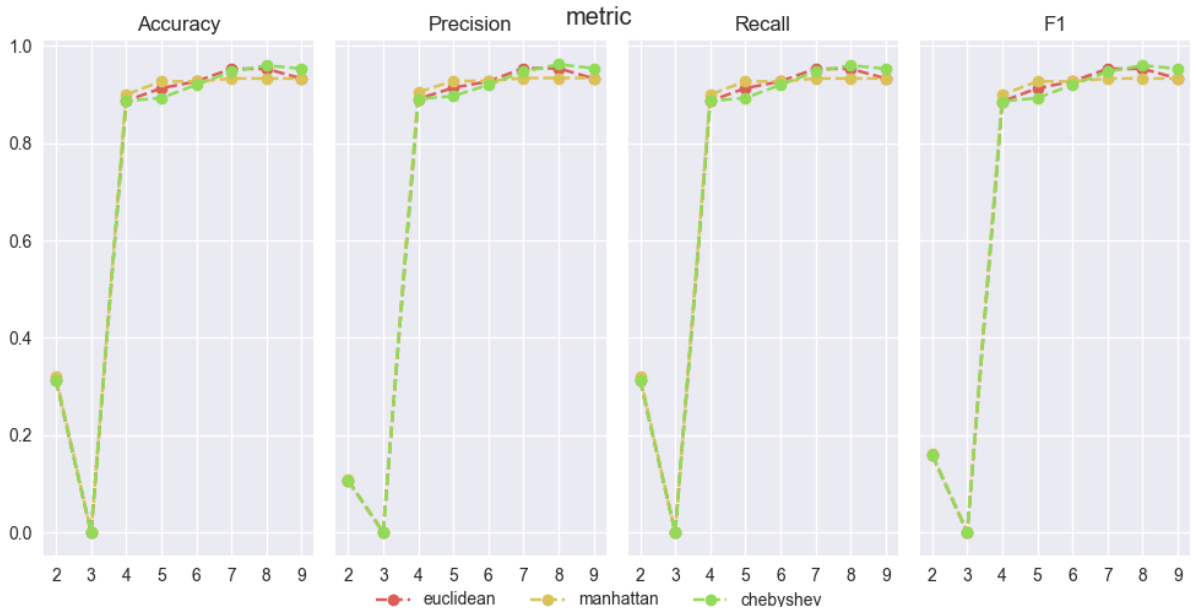
Rysunek 2: Wykres wartości miar dla zbioru "Iris" dla różnej liczby sąsiadów (kroswalidacja stratyfikowana).

Wartość parametru	Metryka	Kroswalidacja							
		K = 2	K = 3	K = 4	K = 5	K = 6	K = 7	K = 8	K = 9
1	Accuracy	0.32	0.0	0.933	0.927	0.947	0.96	0.96	0.96
1	Precision	0.108	0.0	0.935	0.928	0.947	0.96	0.96	0.96
1	Recall	0.32	0.0	0.933	0.927	0.947	0.96	0.96	0.96
1	F1	0.162	0.0	0.933	0.927	0.947	0.96	0.96	0.96
2	Accuracy	0.32	0.0	0.887	0.907	0.913	0.933	0.94	0.933
2	Precision	0.108	0.0	0.898	0.913	0.918	0.935	0.941	0.935
2	Recall	0.32	0.0	0.887	0.907	0.913	0.933	0.94	0.933
2	F1	0.162	0.0	0.885	0.906	0.913	0.933	0.94	0.933
3	Accuracy	0.313	0.0	0.887	0.907	0.92	0.953	0.953	0.953
3	Precision	0.107	0.0	0.891	0.908	0.92	0.953	0.953	0.953
3	Recall	0.313	0.0	0.887	0.907	0.92	0.953	0.953	0.953
3	F1	0.159	0.0	0.886	0.907	0.92	0.953	0.953	0.953
4	Accuracy	0.313	0.0	0.873	0.907	0.913	0.933	0.933	0.933
4	Precision	0.107	0.0	0.88	0.908	0.914	0.934	0.934	0.935
4	Recall	0.313	0.0	0.873	0.907	0.913	0.933	0.933	0.933
4	F1	0.159	0.0	0.872	0.907	0.913	0.933	0.933	0.933
5	Accuracy	0.313	0.0	0.887	0.913	0.927	0.953	0.953	0.933
5	Precision	0.107	0.0	0.891	0.914	0.927	0.954	0.954	0.933
5	Recall	0.313	0.0	0.887	0.913	0.927	0.953	0.953	0.933
5	F1	0.159	0.0	0.886	0.913	0.927	0.953	0.953	0.933

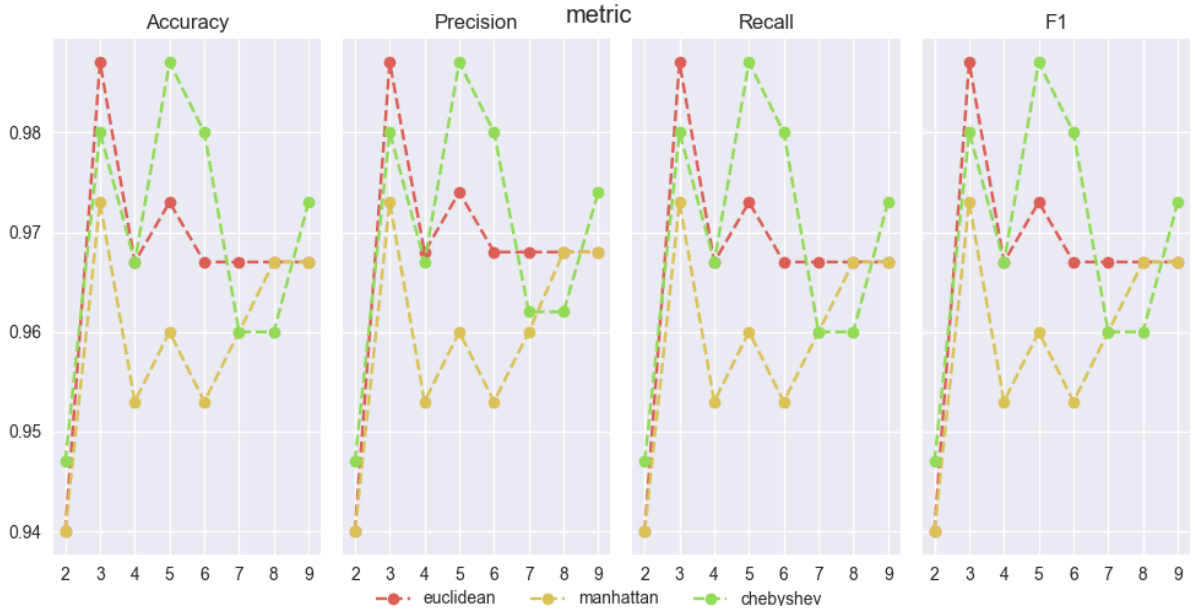
Tabela 1: Wartości miar dla zbioru "Iris" dla różnej liczby sąsiadów (kroswalidacja zwykła).

Wartość parametru	Metryka	Kroswalidacja							
		K = 2	K = 3	K = 4	K = 5	K = 6	K = 7	K = 8	K = 9
1	Accuracy	0.947	0.967	0.96	0.96	0.96	0.96	0.96	0.96
1	Precision	0.947	0.967	0.96	0.96	0.96	0.96	0.96	0.96
1	Recall	0.947	0.967	0.96	0.96	0.96	0.96	0.96	0.96
1	F1	0.947	0.967	0.96	0.96	0.96	0.96	0.96	0.96
2	Accuracy	0.92	0.953	0.953	0.947	0.94	0.953	0.953	0.953
2	Precision	0.924	0.954	0.953	0.947	0.941	0.953	0.953	0.953
2	Recall	0.92	0.953	0.953	0.947	0.94	0.953	0.953	0.953
2	F1	0.92	0.953	0.953	0.947	0.94	0.953	0.953	0.953
3	Accuracy	0.94	0.98	0.96	0.967	0.967	0.967	0.967	0.973
3	Precision	0.94	0.98	0.96	0.967	0.967	0.967	0.967	0.974
3	Recall	0.94	0.98	0.96	0.967	0.967	0.967	0.967	0.973
3	F1	0.94	0.98	0.96	0.967	0.967	0.967	0.967	0.973
4	Accuracy	0.953	0.973	0.953	0.973	0.967	0.96	0.967	0.973
4	Precision	0.954	0.974	0.953	0.973	0.967	0.96	0.967	0.974
4	Recall	0.953	0.973	0.953	0.973	0.967	0.96	0.967	0.973
4	F1	0.953	0.973	0.953	0.973	0.967	0.96	0.967	0.973
5	Accuracy	0.94	0.987	0.967	0.973	0.967	0.967	0.967	0.967
5	Precision	0.94	0.987	0.968	0.974	0.968	0.968	0.968	0.968
5	Recall	0.94	0.987	0.967	0.973	0.967	0.967	0.967	0.967
5	F1	0.94	0.987	0.967	0.973	0.967	0.967	0.967	0.967

Tabela 2: Wartości miar dla zbioru "Iris" dla różnej liczby sąsiadów (kroswalidacja stratyfikowana).



Rysunek 3: Wykres wartości miar dla zbioru "Iris" dla różnych metryk odległości (kroswalidacja zwykła).



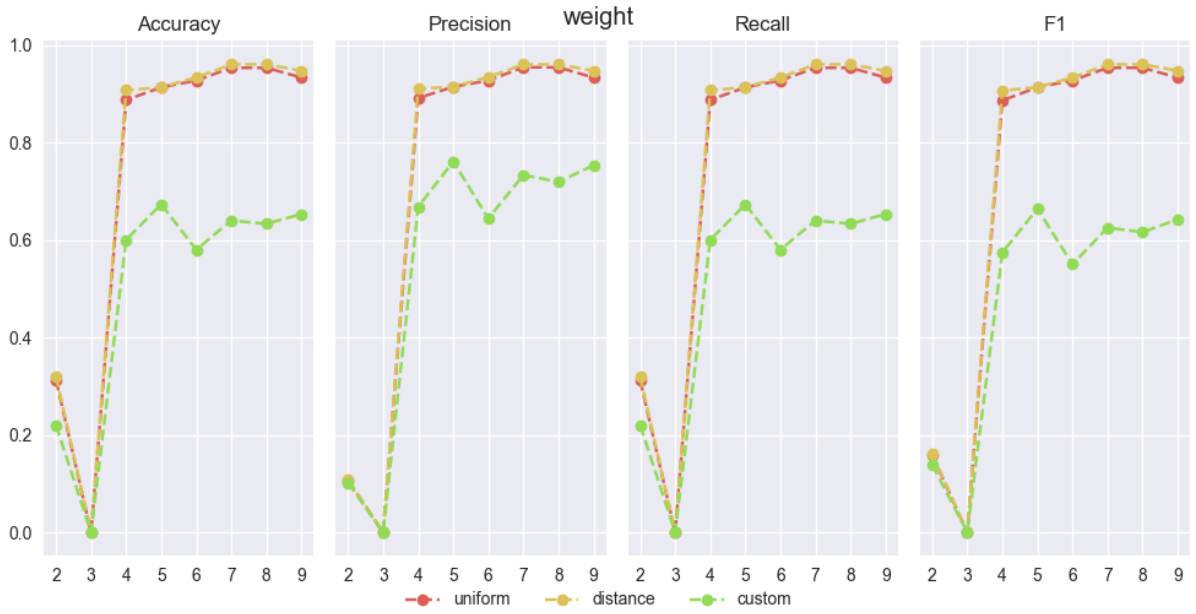
Rysunek 4: Wykres wartości miar dla zbioru "Iris" dla różnych metryk odległości (kroswalidacja stratyfikowana).

Wartość parametru	Metryka	Kroswalidacja							
		K = 2	K = 3	K = 4	K = 5	K = 6	K = 7	K = 8	K = 9
euclidean	Accuracy	0.313	0.0	0.887	0.913	0.927	0.953	0.953	0.933
euclidean	Precision	0.107	0.0	0.891	0.914	0.927	0.954	0.954	0.933
euclidean	Recall	0.313	0.0	0.887	0.913	0.927	0.953	0.953	0.933
euclidean	F1	0.159	0.0	0.886	0.913	0.927	0.953	0.953	0.933
manhattan	Accuracy	0.32	0.0	0.9	0.927	0.927	0.933	0.933	0.933
manhattan	Precision	0.108	0.0	0.905	0.928	0.928	0.934	0.934	0.934
manhattan	Recall	0.32	0.0	0.9	0.927	0.927	0.933	0.933	0.933
manhattan	F1	0.162	0.0	0.9	0.927	0.927	0.933	0.933	0.933
chebyshev	Accuracy	0.313	0.0	0.887	0.893	0.92	0.947	0.96	0.953
chebyshev	Precision	0.107	0.0	0.891	0.897	0.92	0.947	0.962	0.954
chebyshev	Recall	0.313	0.0	0.887	0.893	0.92	0.947	0.96	0.953
chebyshev	F1	0.159	0.0	0.886	0.893	0.92	0.947	0.96	0.953

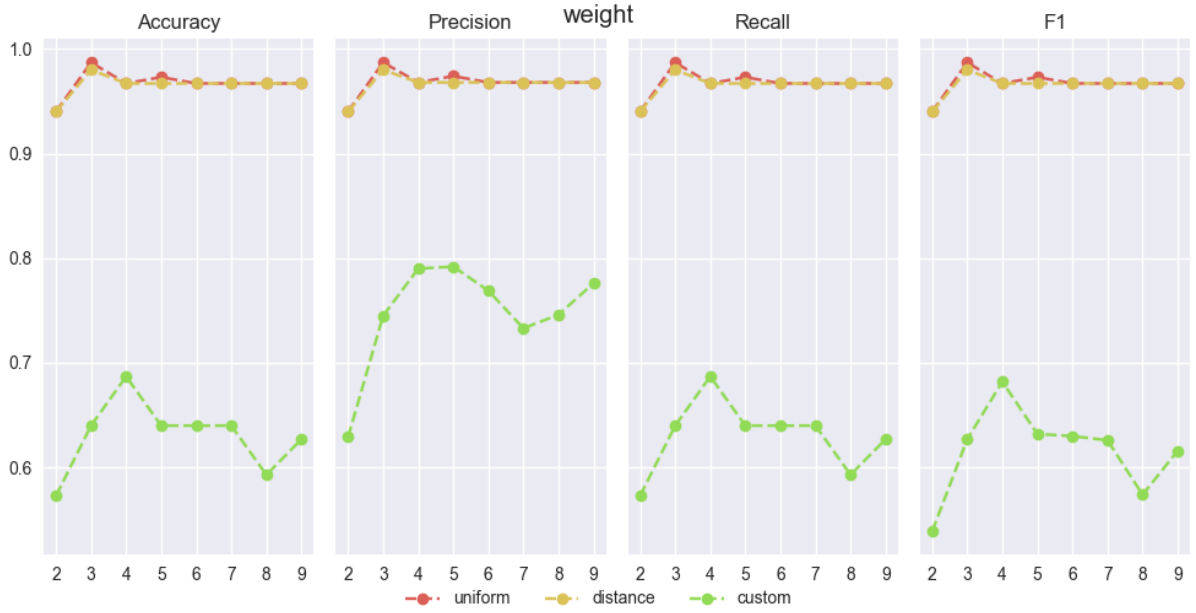
Tabela 3: Wartości miar dla zbioru "Iris" dla różnych metryk odległości (kroswalidacja zwykła).

Wartość parametru	Metryka	Kroswalidacja							
		K = 2	K = 3	K = 4	K = 5	K = 6	K = 7	K = 8	K = 9
euclidean	Accuracy	0.94	0.987	0.967	0.973	0.967	0.967	0.967	0.967
euclidean	Precision	0.94	0.987	0.968	0.974	0.968	0.968	0.968	0.968
euclidean	Recall	0.94	0.987	0.967	0.973	0.967	0.967	0.967	0.967
euclidean	F1	0.94	0.987	0.967	0.973	0.967	0.967	0.967	0.967
manhattan	Accuracy	0.94	0.973	0.953	0.96	0.953	0.96	0.967	0.967
manhattan	Precision	0.94	0.973	0.953	0.96	0.953	0.96	0.968	0.968
manhattan	Recall	0.94	0.973	0.953	0.96	0.953	0.96	0.967	0.967
manhattan	F1	0.94	0.973	0.953	0.96	0.953	0.96	0.967	0.967
chebyshev	Accuracy	0.947	0.98	0.967	0.987	0.98	0.96	0.96	0.973
chebyshev	Precision	0.947	0.98	0.967	0.987	0.98	0.962	0.962	0.974
chebyshev	Recall	0.947	0.98	0.967	0.987	0.98	0.96	0.96	0.973
chebyshev	F1	0.947	0.98	0.967	0.987	0.98	0.96	0.96	0.973

Tabela 4: Wartości miar dla zbioru "Iris" dla różnych metryk odległości (kroswalidacja stratyfikowana).



Rysunek 5: Wykres wartości miar dla zbioru "Iris" dla różnych sposobów głosowania (kroswalidacja zwykła).



Rysunek 6: Wykres wartości miar dla zbioru "Iris" dla różnych sposobów głosowania (kroswalidacja stratyfikowana).

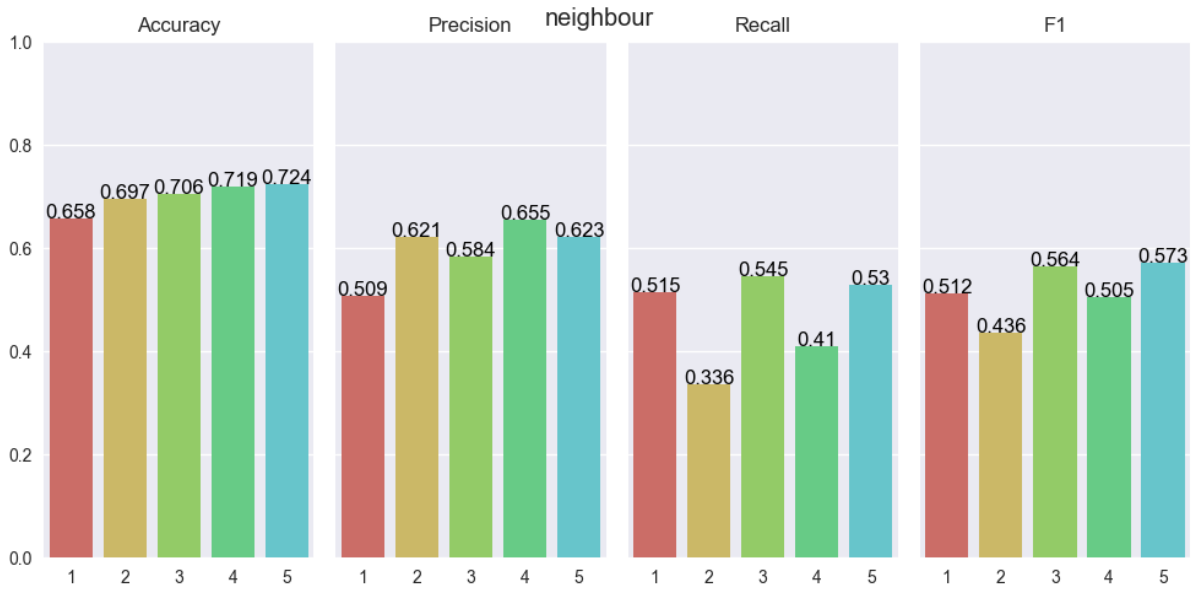
Wartość parametru	Metryka	Kroswalidacja							
		K = 2	K = 3	K = 4	K = 5	K = 6	K = 7	K = 8	K = 9
uniform	Accuracy	0.313	0.0	0.887	0.913	0.927	0.953	0.953	0.933
uniform	Precision	0.107	0.0	0.891	0.914	0.927	0.954	0.954	0.933
uniform	Recall	0.313	0.0	0.887	0.913	0.927	0.953	0.953	0.933
uniform	F1	0.159	0.0	0.886	0.913	0.927	0.953	0.953	0.933
distance	Accuracy	0.32	0.0	0.907	0.913	0.933	0.96	0.96	0.947
distance	Precision	0.108	0.0	0.91	0.914	0.934	0.96	0.96	0.947
distance	Recall	0.32	0.0	0.907	0.913	0.933	0.96	0.96	0.947
distance	F1	0.162	0.0	0.906	0.913	0.933	0.96	0.96	0.947
custom	Accuracy	0.233	0.0	0.607	0.567	0.613	0.593	0.673	0.627
custom	Precision	0.112	0.0	0.681	0.691	0.699	0.729	0.803	0.726
custom	Recall	0.233	0.0	0.607	0.567	0.613	0.593	0.673	0.627
custom	F1	0.152	0.0	0.583	0.535	0.591	0.573	0.67	0.61

Tabela 5: Wartości miar dla zbioru "Iris" dla różnych sposobów głosowania (kroswalidacja zwykła).

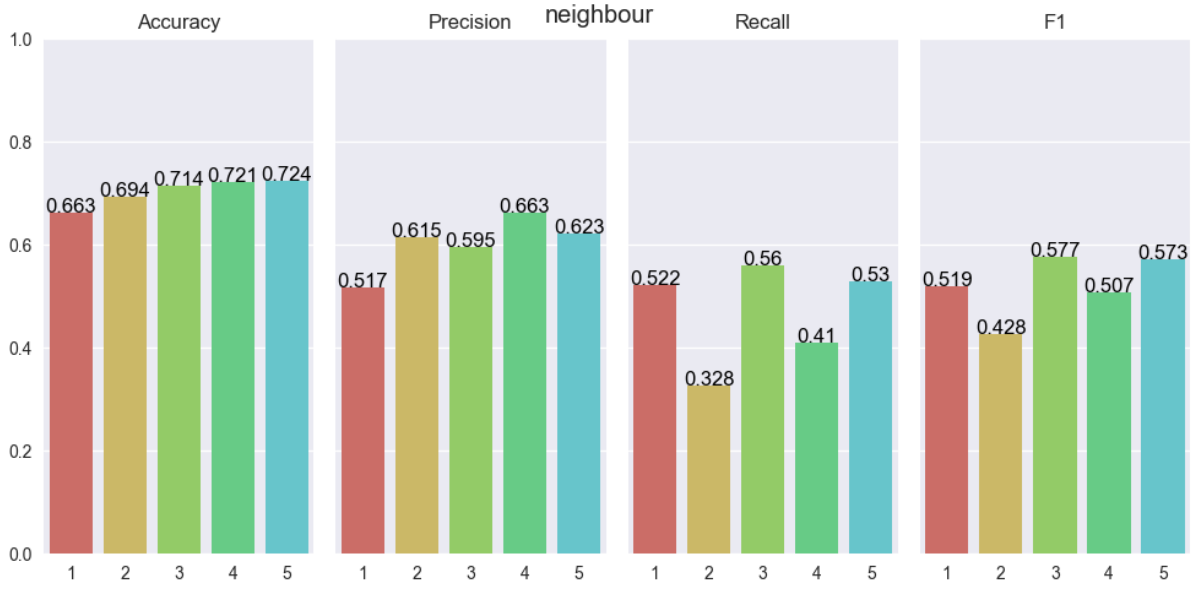
Wartość parametru	Metryka	Kroswalidacja							
		K = 2	K = 3	K = 4	K = 5	K = 6	K = 7	K = 8	K = 9
uniform	Accuracy	0.94	0.987	0.967	0.973	0.967	0.967	0.967	0.967
uniform	Precision	0.94	0.987	0.968	0.974	0.968	0.968	0.968	0.968
uniform	Recall	0.94	0.987	0.967	0.973	0.967	0.967	0.967	0.967
uniform	F1	0.94	0.987	0.967	0.973	0.967	0.967	0.967	0.967
distance	Accuracy	0.94	0.98	0.967	0.967	0.967	0.967	0.967	0.967
distance	Precision	0.94	0.98	0.968	0.968	0.968	0.968	0.968	0.968
distance	Recall	0.94	0.98	0.967	0.967	0.967	0.967	0.967	0.967
distance	F1	0.94	0.98	0.967	0.967	0.967	0.967	0.967	0.967
custom	Accuracy	0.653	0.573	0.687	0.647	0.62	0.68	0.7	0.673
custom	Precision	0.769	0.665	0.778	0.752	0.75	0.764	0.795	0.786
custom	Recall	0.653	0.573	0.687	0.647	0.62	0.68	0.7	0.673
custom	F1	0.644	0.544	0.675	0.634	0.605	0.67	0.692	0.668

Tabela 6: Wartości miar dla zbioru "Iris" dla różnych sposobów głosowania (kroswalidacja stratyfikowana).

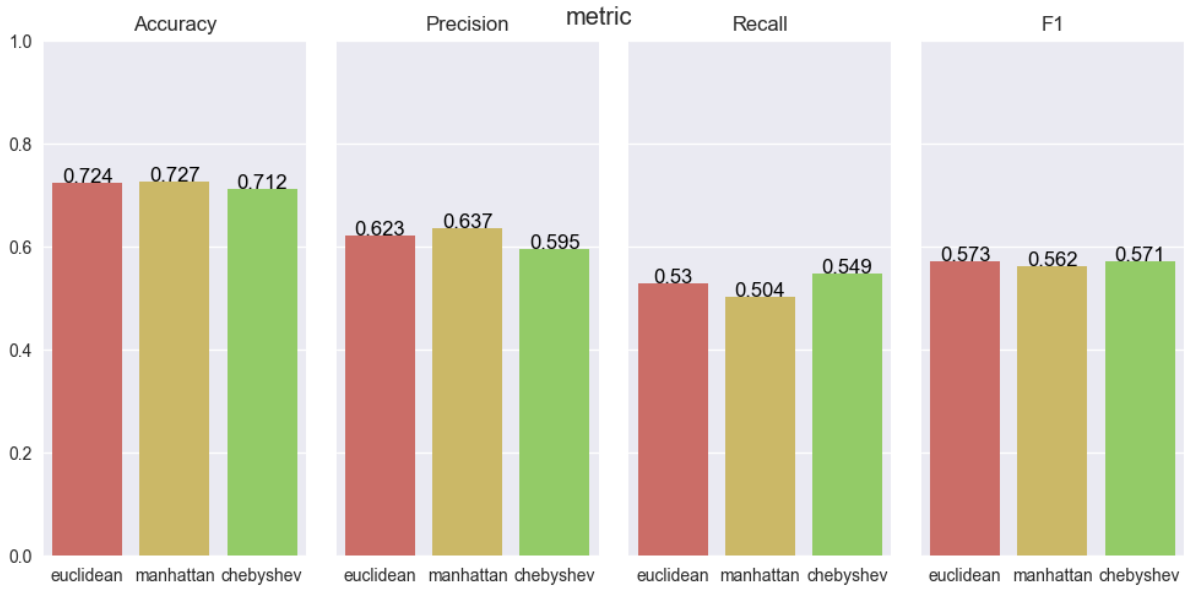
2.2 Zbiór "Diabetes"



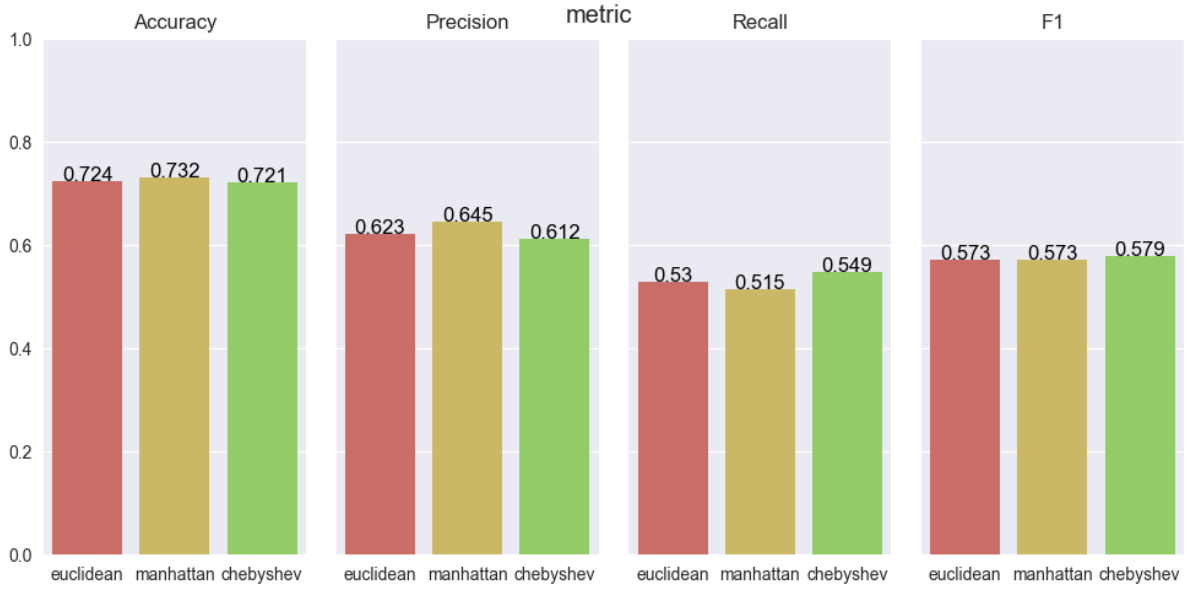
Rysunek 7: Wykres wartości miar dla zbioru "Diabetes" dla różnej liczby sąsiadów (kroswalidacja zwykła).



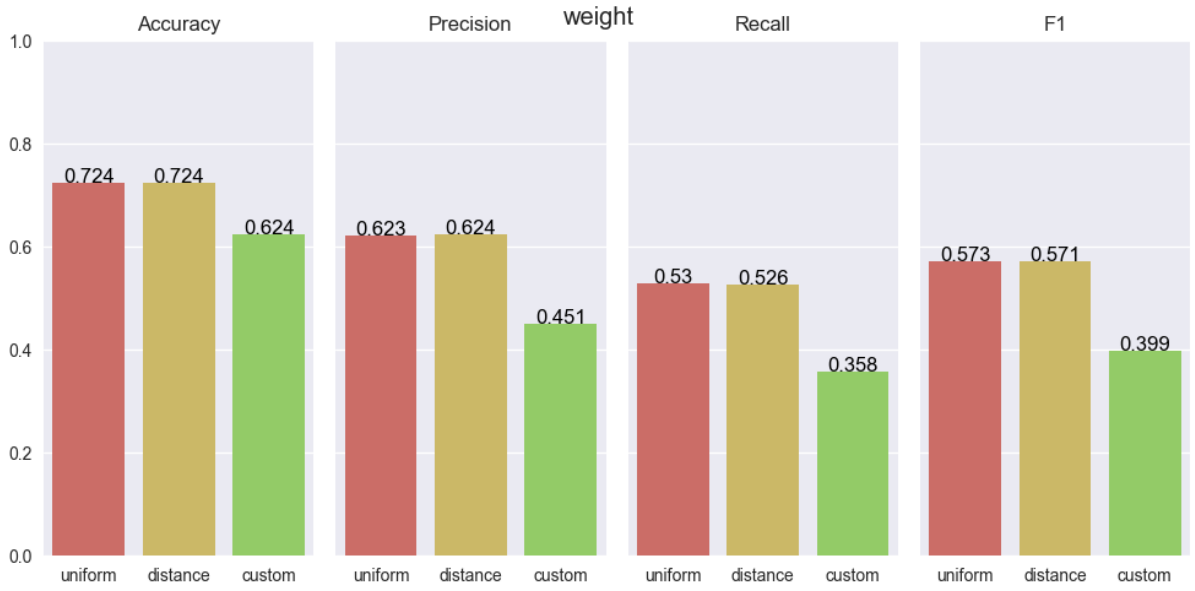
Rysunek 8: Wykres wartości miar dla zbioru "Diabetes" dla różnej liczby sąsiadów (kroswalidacja stratyfikowana).



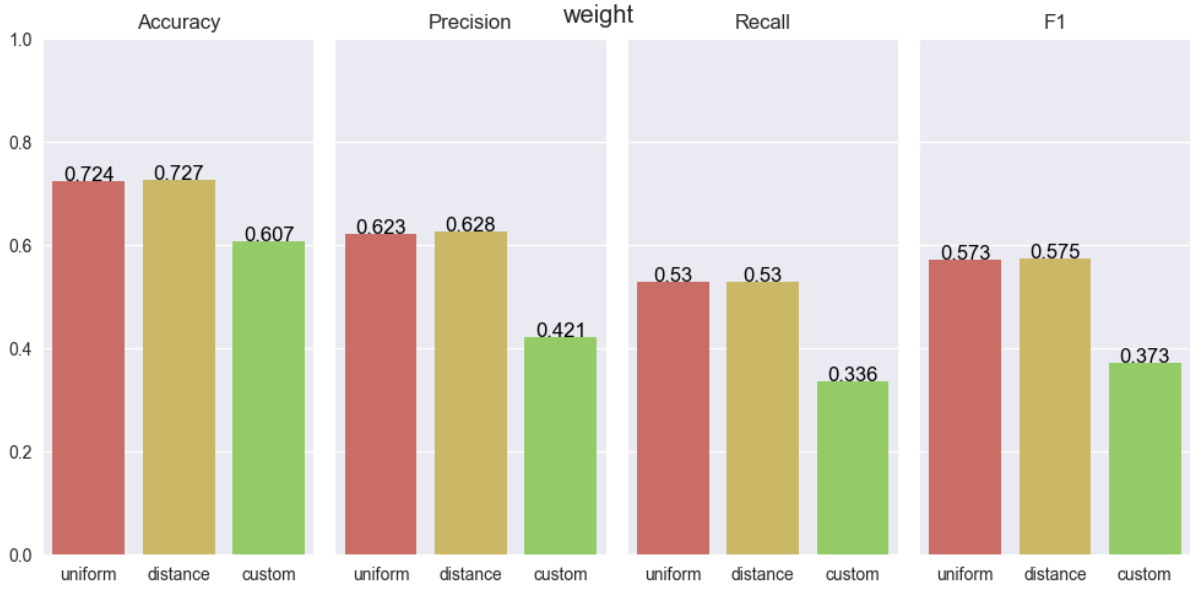
Rysunek 9: Wykres wartości miar dla zbioru "Diabetes" dla różnych metryk odległości (kroswalidacja zwykła).



Rysunek 10: Wykres wartości miar dla zbioru "Diabetes" dla różnych metryk odległości (kroswalidacja stratyfikowana).

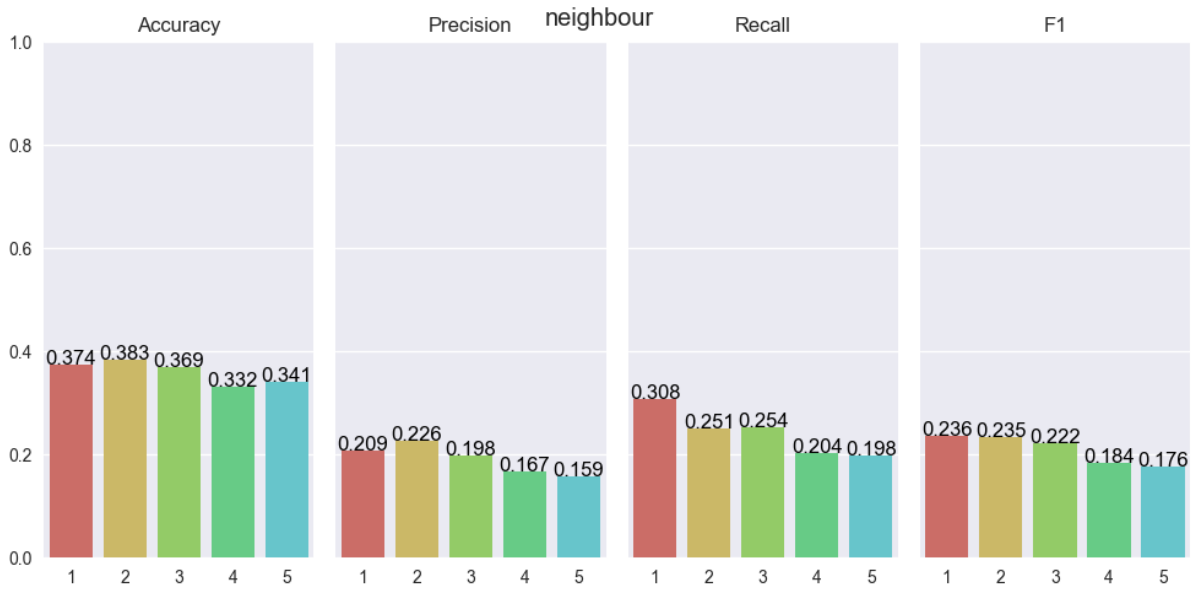


Rysunek 11: Wykres wartości miar dla zbioru "Diabetes" dla różnych sposobów głosowania (kroswalidacja zwykła).

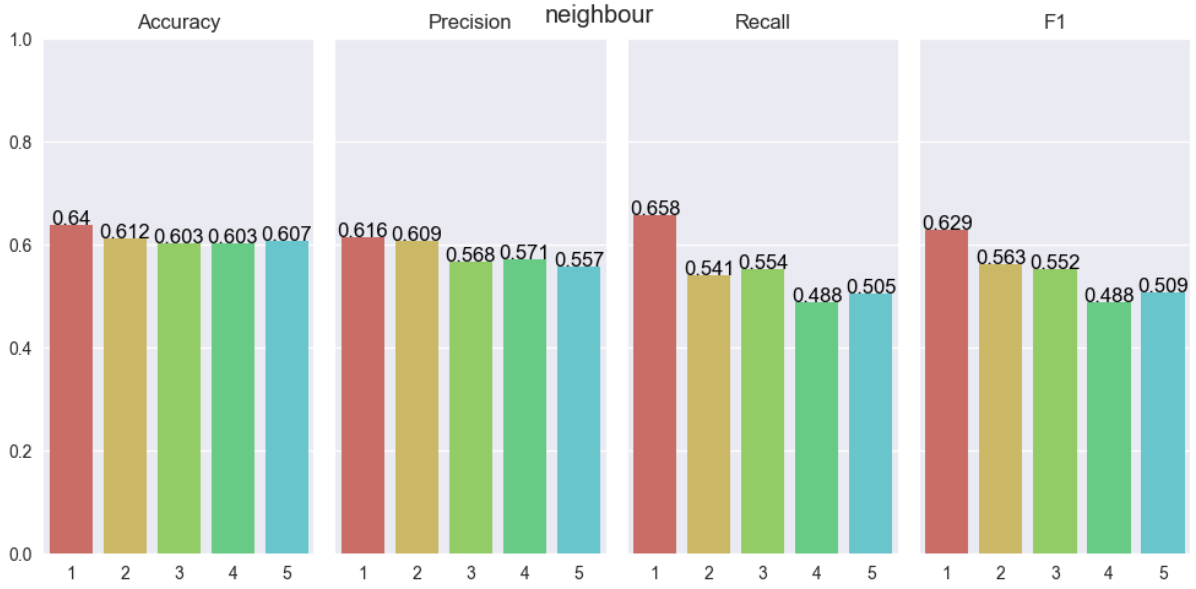


Rysunek 12: Wykres wartości miar dla zbioru "Diabetes" dla różnych sposobów głosowania (kroswalidacja stratyfikowana).

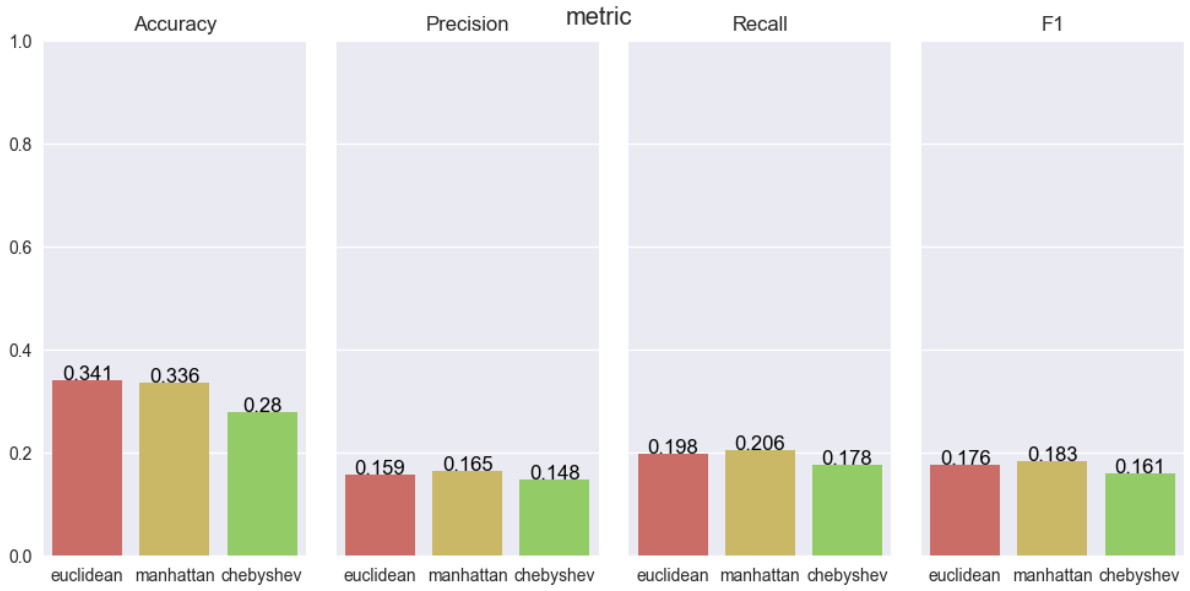
2.3 Zbiór "Glass"



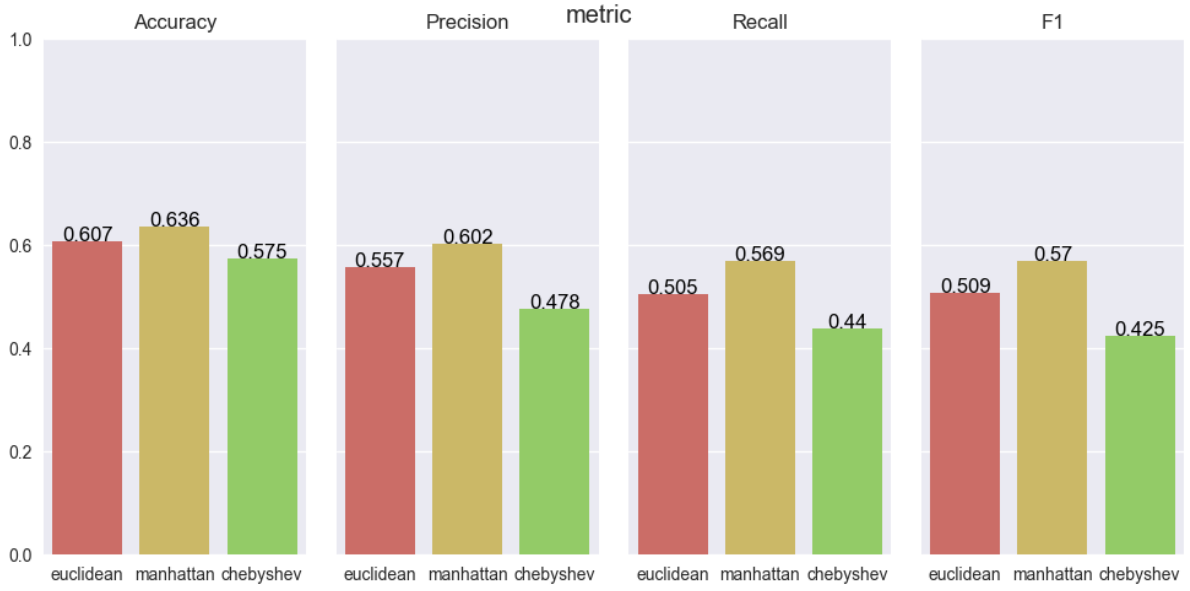
Rysunek 13: Wykres wartości miar dla zbioru "Glass" dla różnej liczby sąsiadów (kroswalidacja zwykła).



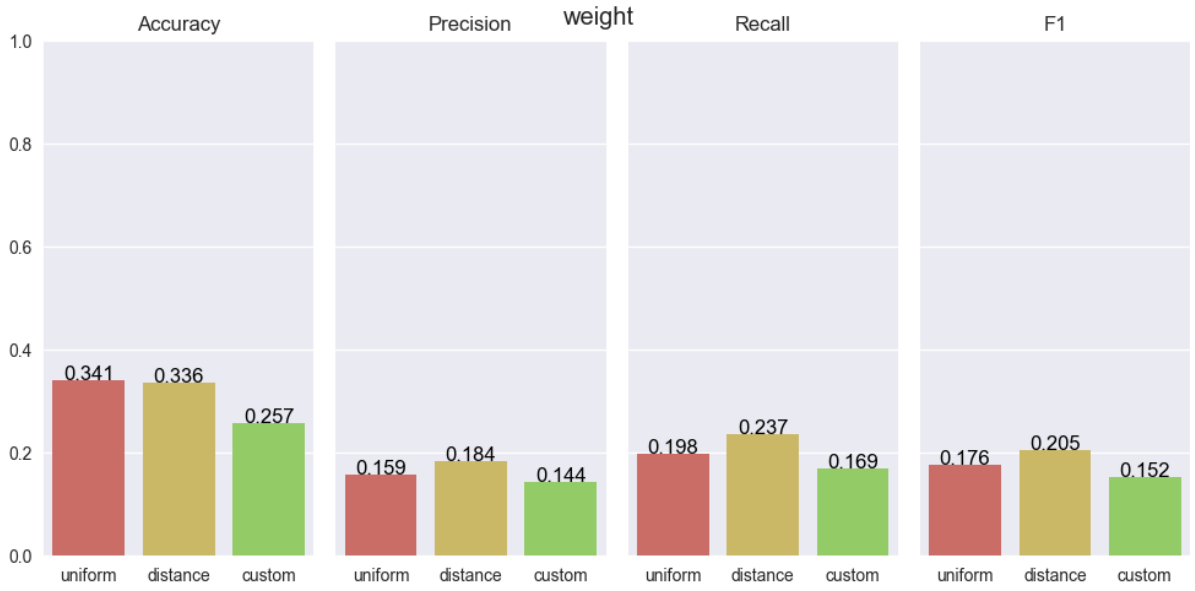
Rysunek 14: Wykres wartości miar dla zbioru "Glass" dla różnej liczby sąsiadów (kroswalidacja stratyfikowana).



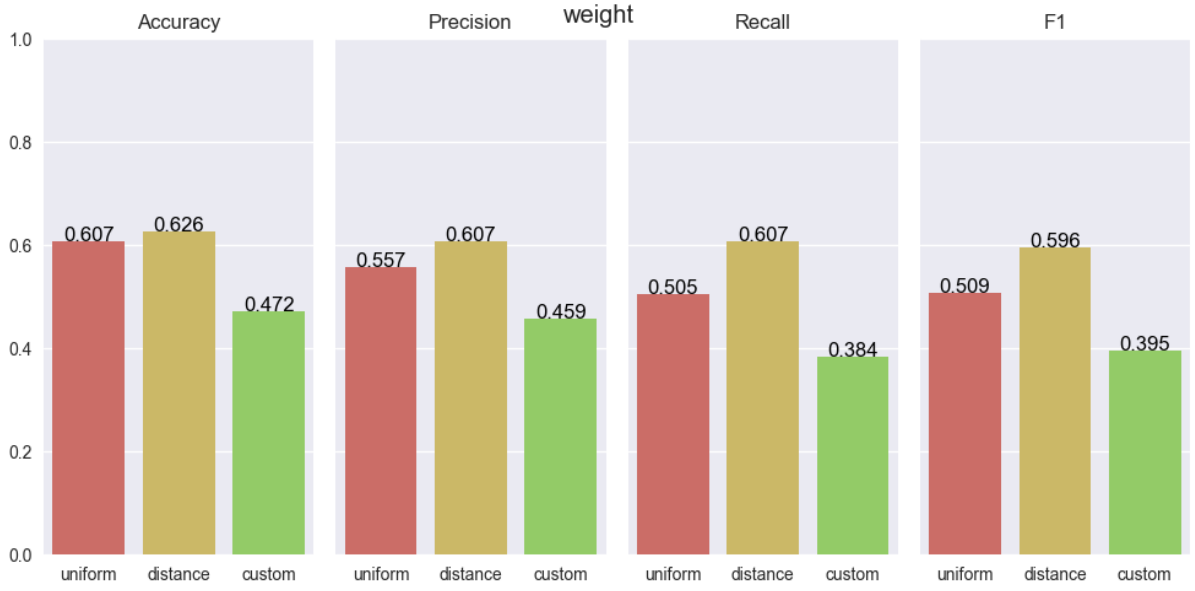
Rysunek 15: Wykres wartości miar dla zbioru "Glass" dla różnych metryk odległości (kroswalidacja zwykła).



Rysunek 16: Wykres wartości miar dla zbioru "Glass" dla różnych metryk odległości (kroswalidacja stratyfikowana).

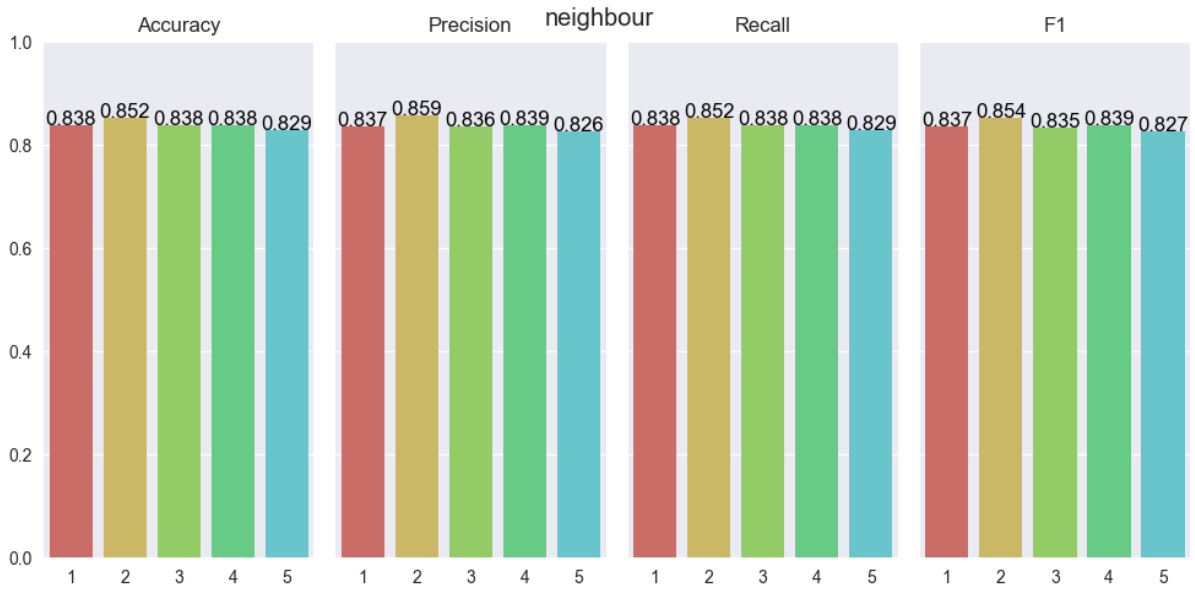


Rysunek 17: Wykres wartości miar dla zbioru "Glass" dla różnych sposobów głosowania (kroswalidacja zwykła).

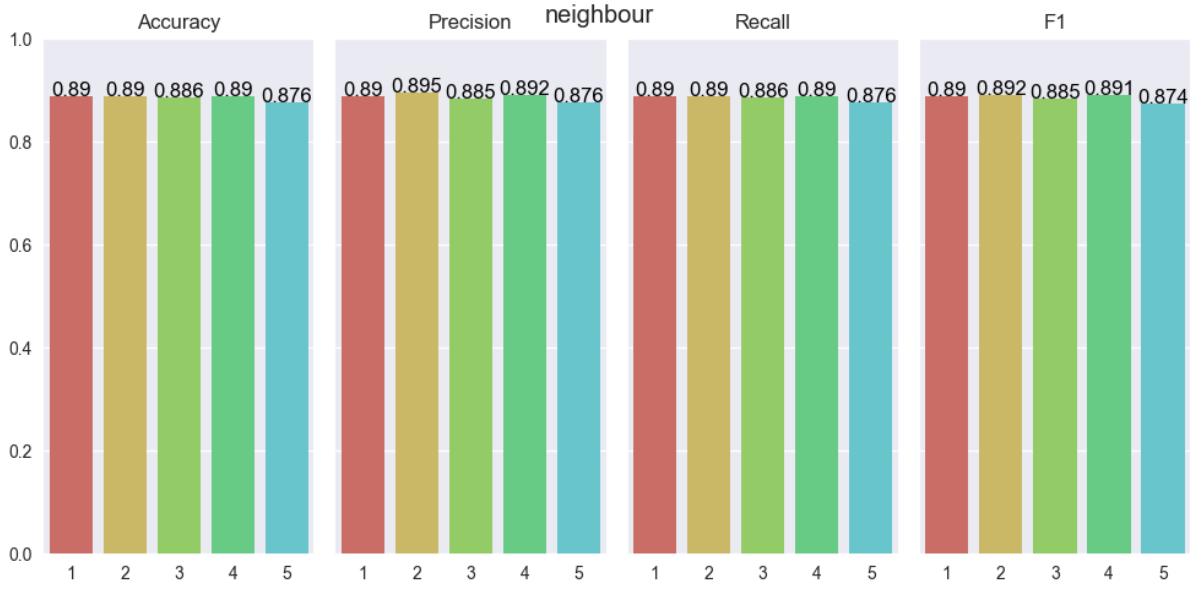


Rysunek 18: Wykres wartości miar dla zbioru "Glass" dla różnych sposobów głosowania (kroswalidacja stratyfikowana).

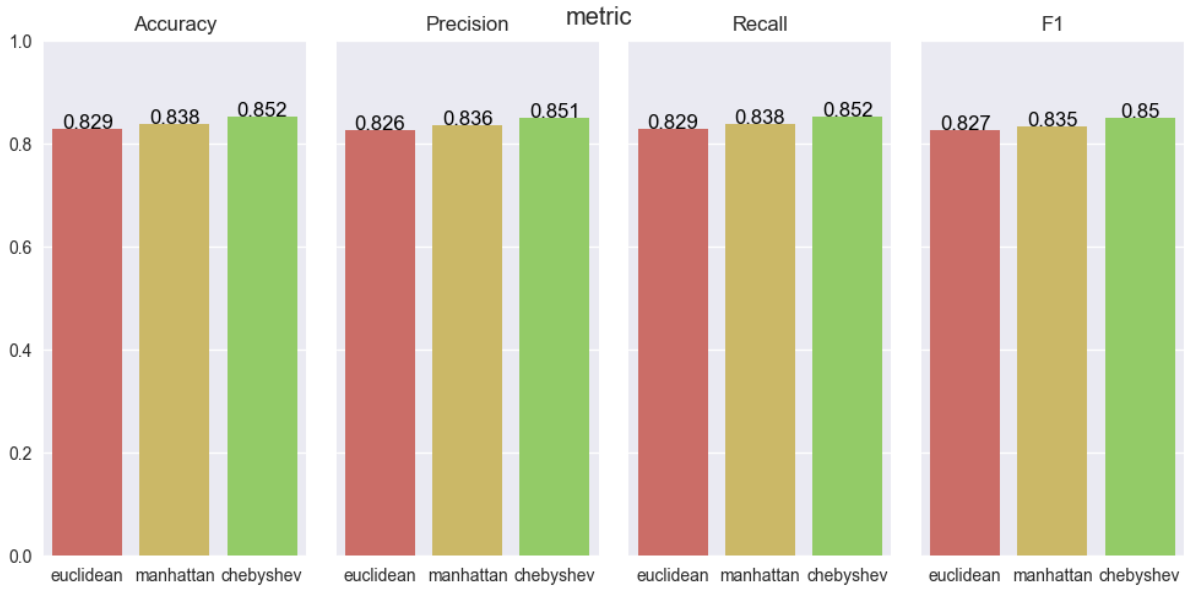
2.4 Zbiór "Seeds"



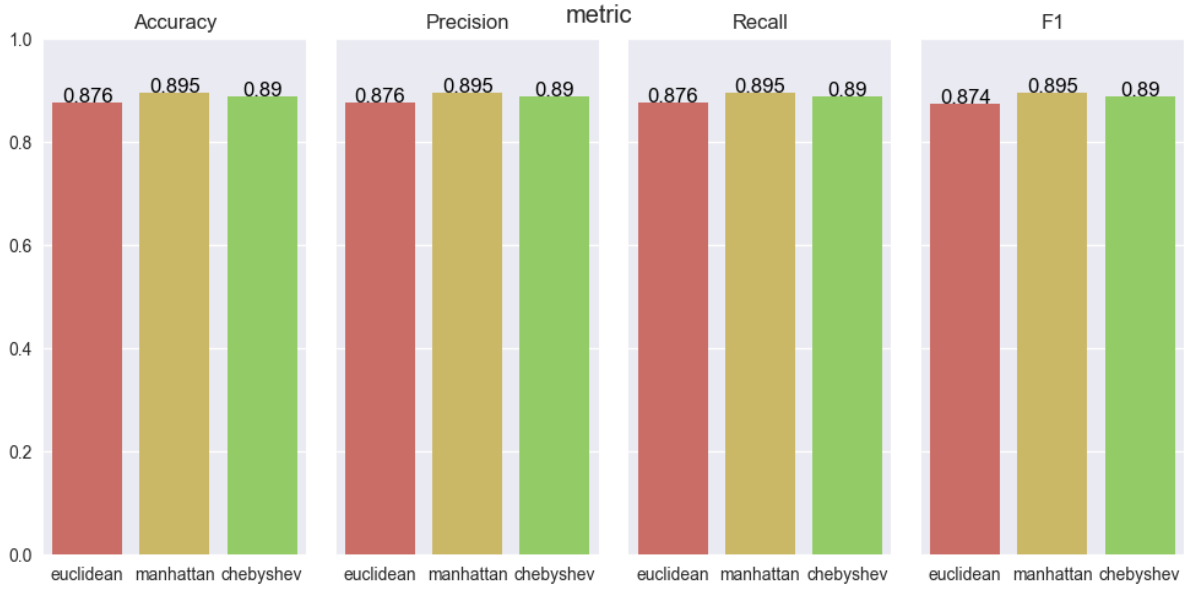
Rysunek 19: Wykres wartości miar dla zbioru "Seeds" dla różnej liczby sąsiadów (kroswalidacja zwykła).



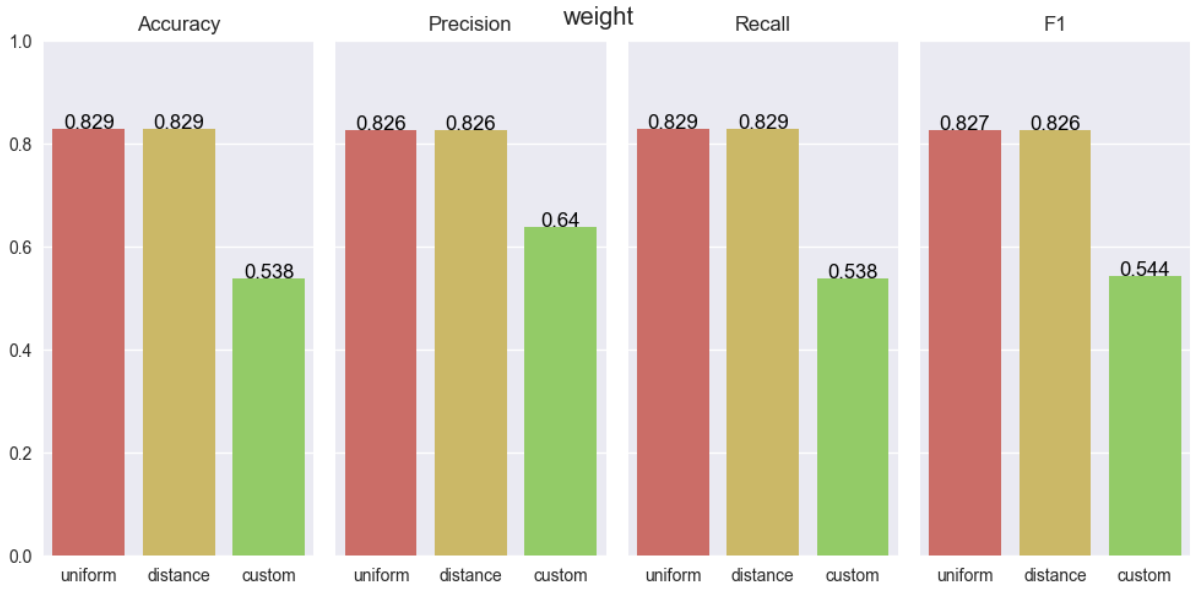
Rysunek 20: Wykres wartości miar dla zbioru "Seeds" dla różnej liczby sąsiadów (kroswalidacja stratyfikowana).



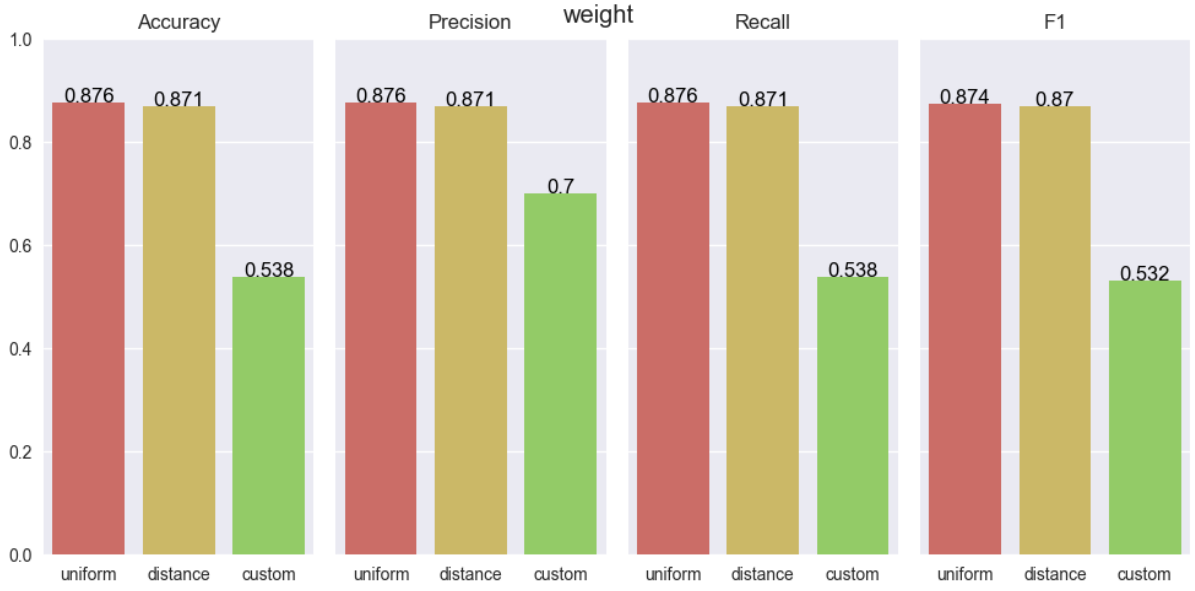
Rysunek 21: Wykres wartości miar dla zbioru "Seeds" dla różnych metryk odległości (kroswalidacja zwykła).



Rysunek 22: Wykres wartości miar dla zbioru "Seeds" dla różnych metryk odległości (kroswalidacja stratyfikowana).

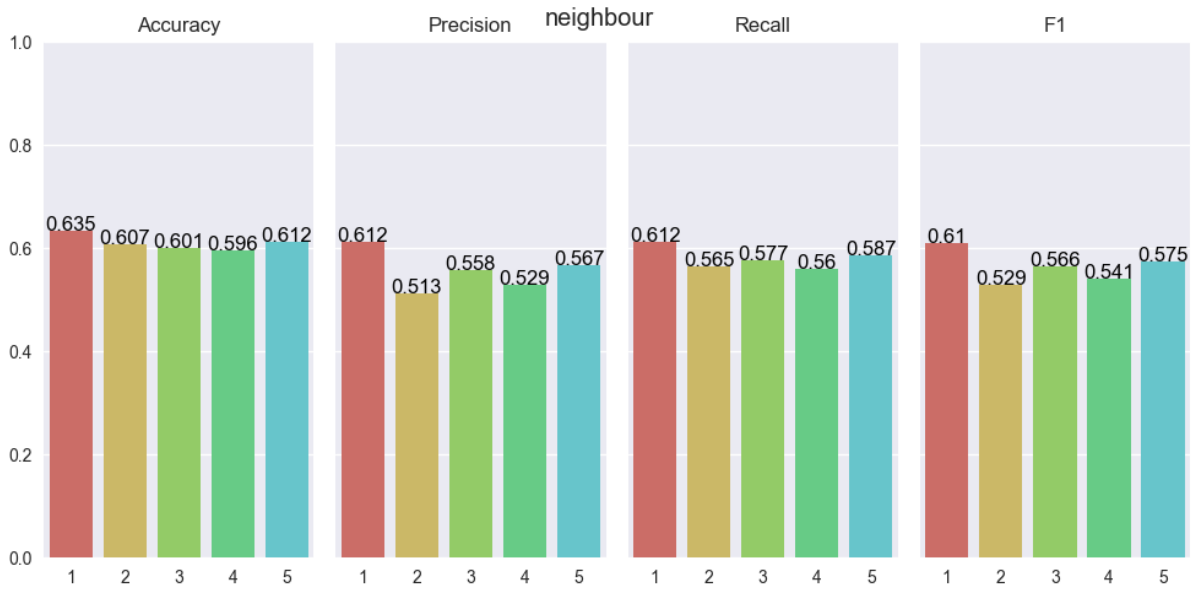


Rysunek 23: Wykres wartości miar dla zbioru "Seeds" dla różnych sposobów głosowania (kroswalidacja zwykła).

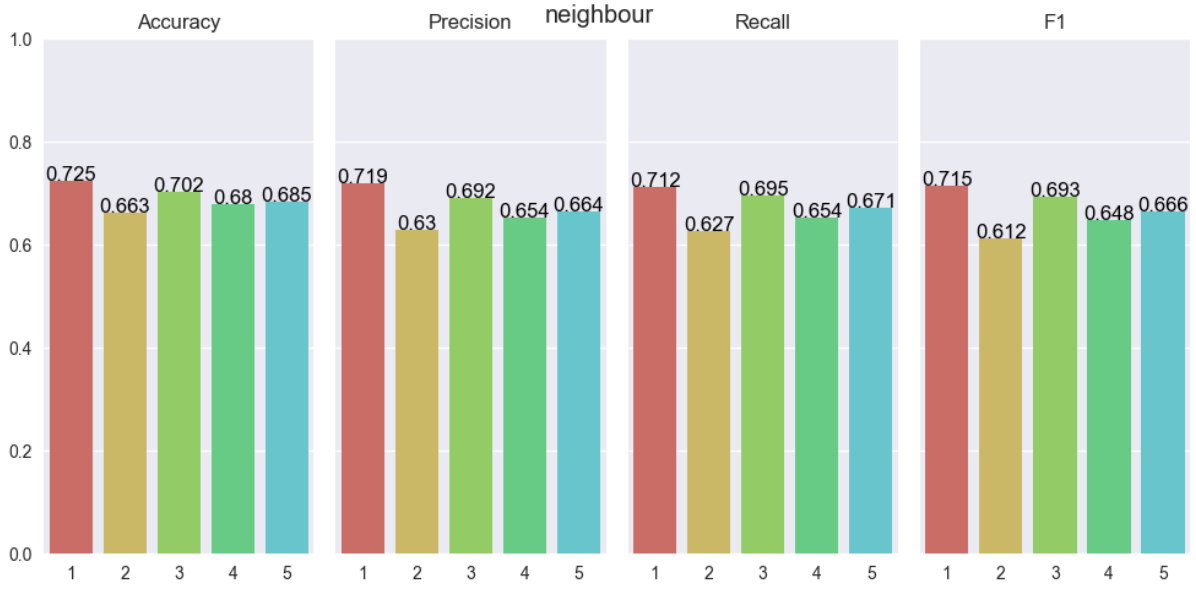


Rysunek 24: Wykres wartości miar dla zbioru "Seeds" dla różnych sposobów głosowania (kroswalidacja stratyfikowana).

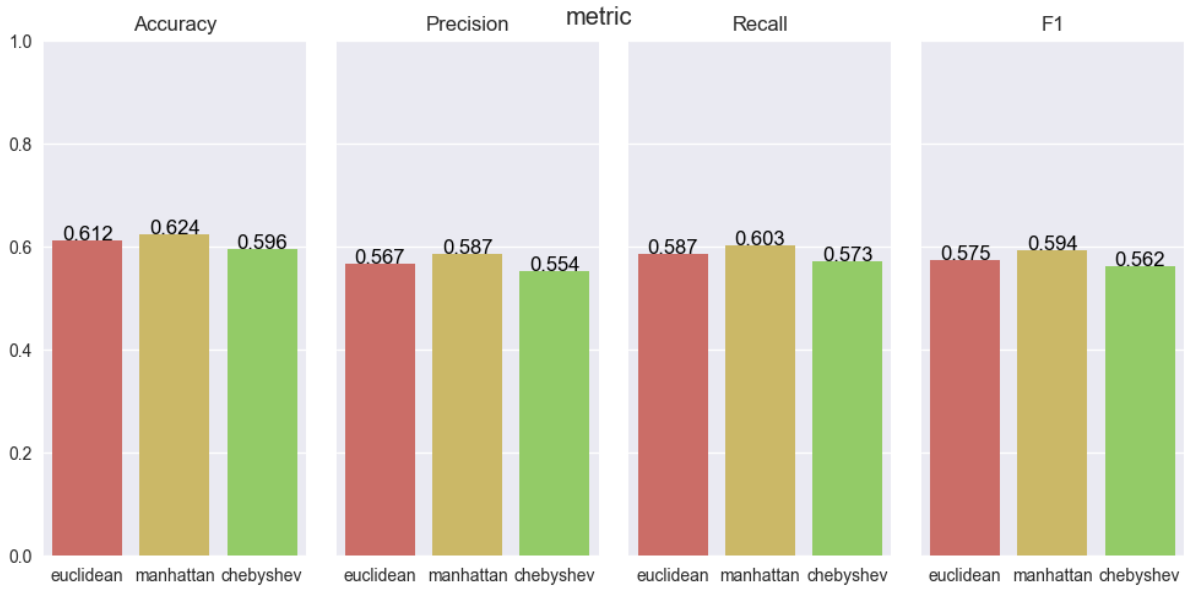
2.5 Zbiór "Wine"



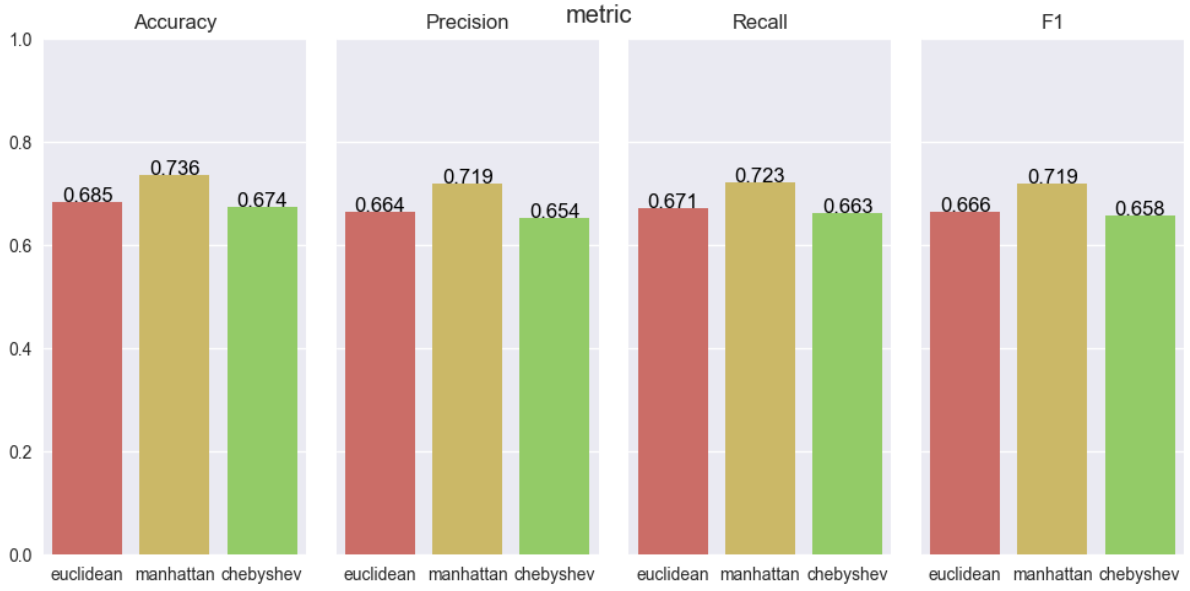
Rysunek 25: Wykres wartości miar dla zbioru "Wine" dla różnej liczby sąsiadów (kroswalidacja zwykła).



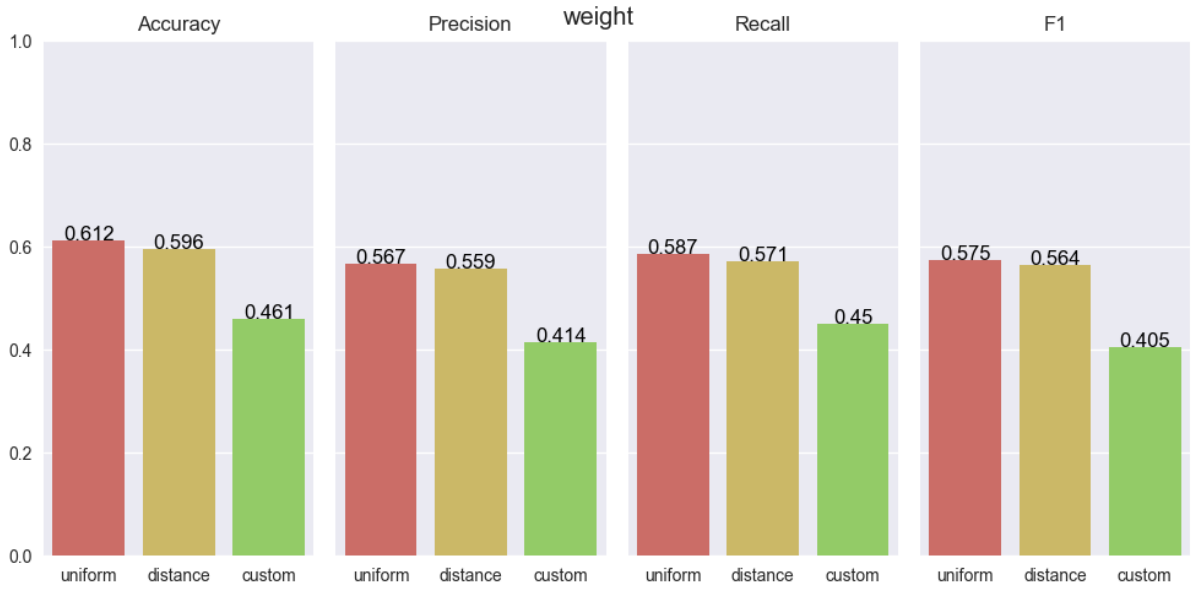
Rysunek 26: Wykres wartości miar dla zbioru "Wine" dla różnej liczby sąsiadów (kroswalidacja stratyfikowana).



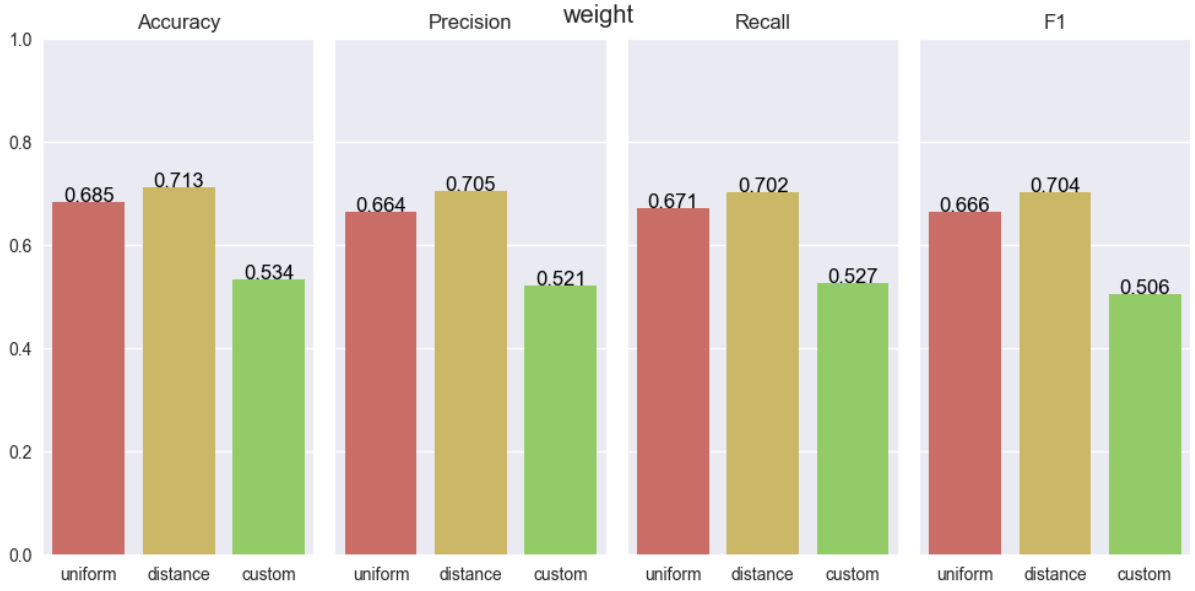
Rysunek 27: Wykres wartości miar dla zbioru "Wine" dla różnych metryk odległości (kroswalidacja zwykła).



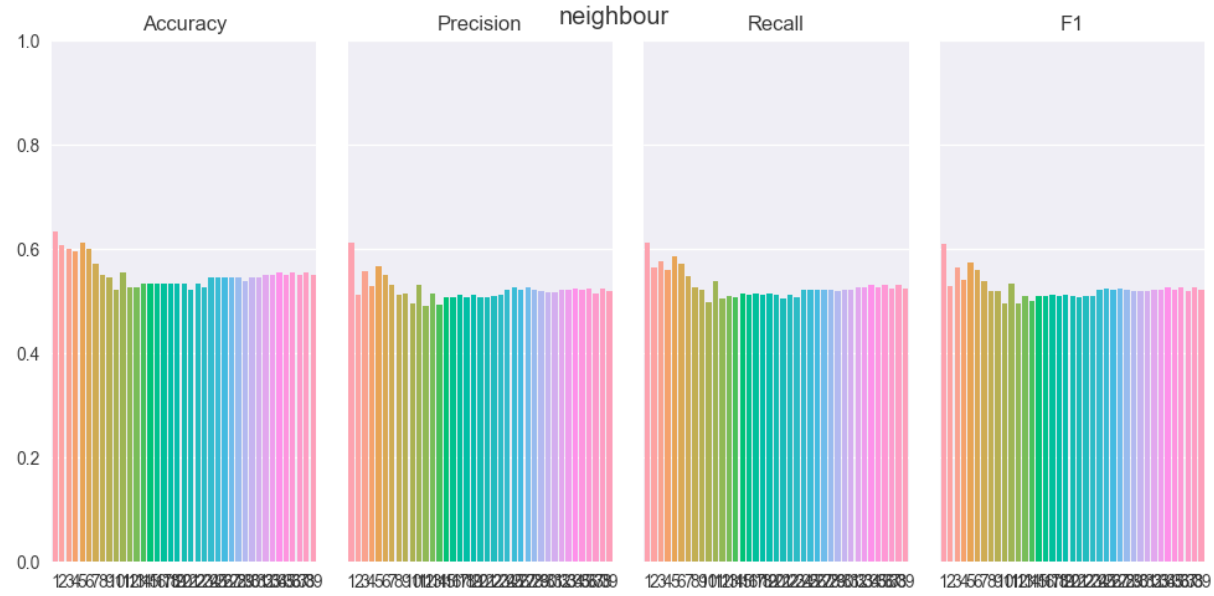
Rysunek 28: Wykres wartości miar dla zbioru "Wine" dla różnych metryk odległości (kroswalidacja stratyfikowana).



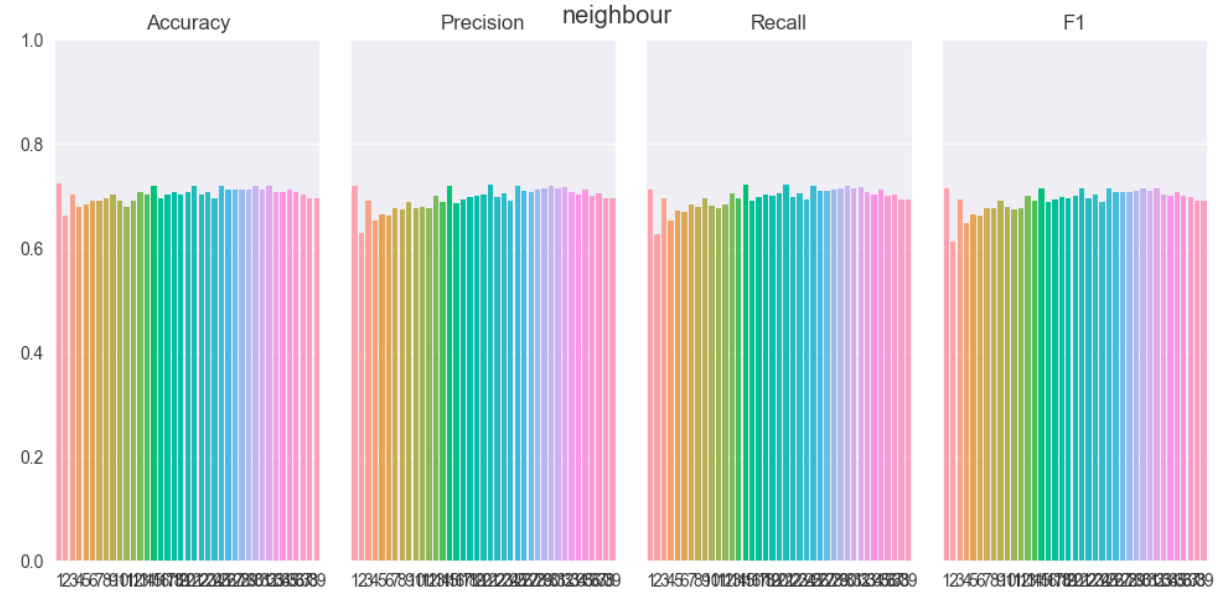
Rysunek 29: Wykres wartości miar dla zbioru "Wine" dla różnych sposobów głosowania (kroswalidacja zwykła).



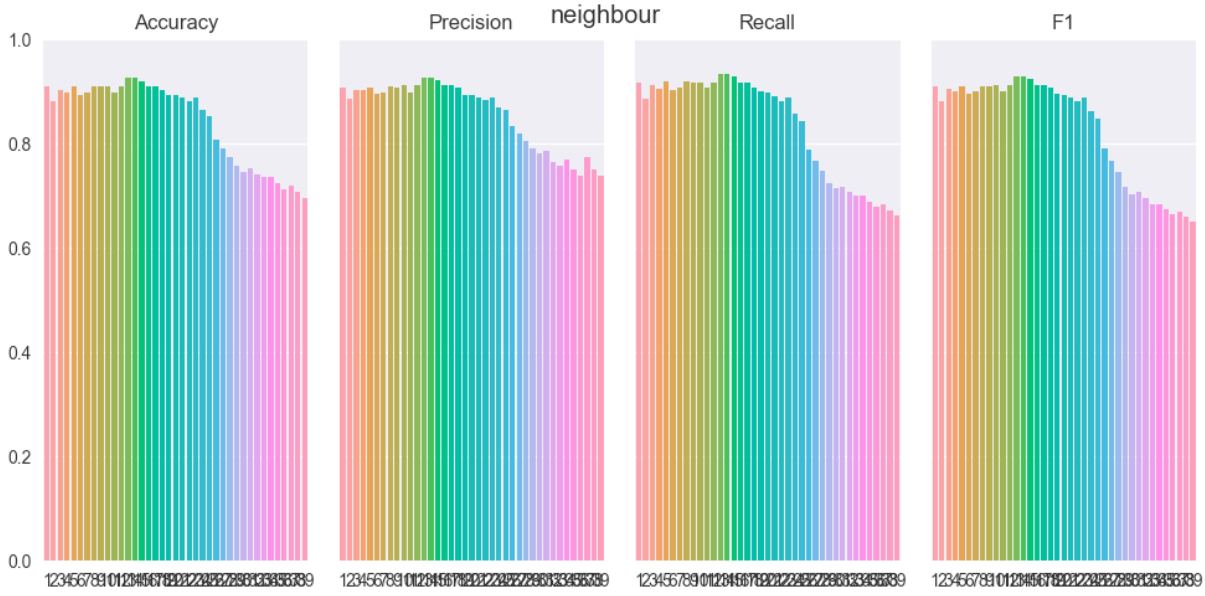
Rysunek 30: Wykres wartości miar dla zbioru "Wine" dla różnych sposobów głosowania (kroswalidacja stratyfikowana).



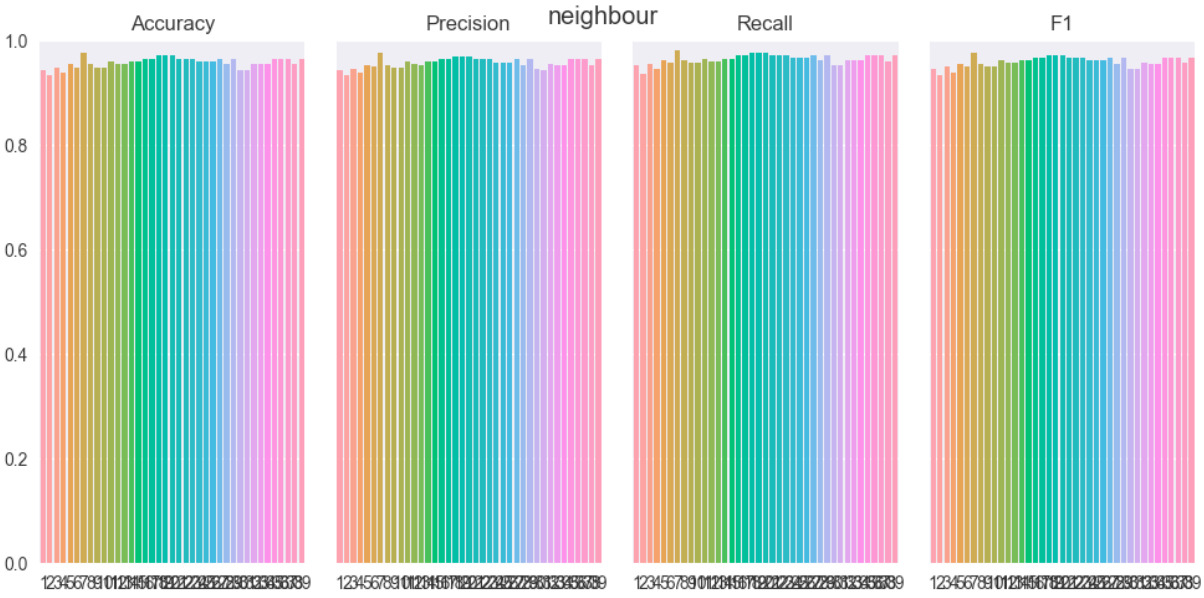
Rysunek 31: Wykres wartości miar dla zbioru "Wine" dla większej liczby sąsiadów (kroswalidacja zwykła, dane nieznormalizowane).



Rysunek 32: Wykres wartości miar dla zbioru "Wine" dla większej liczby sąsiadów (kroswalidacja zwykła, dane znormalizowane).



Rysunek 33: Wykres wartości miar dla zbioru "Wine" dla większej liczby sąsiadów (kroswalidacja stratyfikowana, dane nieznormalizowane).



Rysunek 34: Wykres wartości miar dla zbioru "Wine" dla większej liczby sąsiadów (kroswalidacja stratyfikowana, dane znormalizowane).

3 Porównanie klasyfikatorów

Klasyfikator	Accuracy	Precision	Recall	F1	Komentarz
C4.5	0.75	0.77	0.89	0.82	CV = 6, C3
Naiwny Bayes	0.75	0.66	0.60	0.63	CV = 6, brak dyskr.
KNN	0.71	0.60	0.56	0.58	CV = 5 (strat.), k = 3, euklides, głos. równ.

Tabela 7: Najlepsze wyniki klasyfikatorów dla zbioru "Diabetes".

Klasyfikator	Accuracy	Precision	Recall	F1	Komentarz
C4.5	0.70	0.73	0.69	0.77	CV = 8, C1
KNN	0.64	0.62	0.66	0.63	CV = 5 (strat.), k = 1, euklides, głos. równ.
Naiwny Bayes	0.67	0.60	0.63	0.61	CV = 5, CAIM

Tabela 8: Najlepsze wyniki klasyfikatorów dla zbioru "Glass".

Klasyfikator	Accuracy	Precision	Recall	F1	Komentarz
C4.5	0.94	0.95	0.94	0.94	CV = 8, C1
KNN	0.90	0.90	0.90	0.90	CV = 5 (strat.), k = 5, manhatt., głos. równ.
Naiwny Bayes	0.89	0.89	0.89	0.89	CV = 2, brak dyskr.

Tabela 9: Najlepsze wyniki klasyfikatorów dla zbioru "Seeds".

Klasyfikator	Accuracy	Precision	Recall	F1	Komentarz
Naiwny Bayes	0.97	0.97	0.97	0.97	CV = 2, brak dyskr.
C4.5	0.95	0.96	0.95	0.95	CV = 7, C1
KNN	0.74	0.72	0.72	0.72	CV = 5 (strat.), k = 5, manhatt., głos. równ.

Tabela 10: Najlepsze wyniki klasyfikatorów dla zbioru "Wine".