Sprawozdanie

Indukcyjne Metody Analizy Danych

Ćwiczenie 4. Algorytm klasyfikacji k-najbliższych sąsiadów

Autor: Paweł Mielniczuk

Spis treści:

1. Opis działania algorytmu KNN

Knn – Knn jest modelem o działaniu instancyjnym/leniwym. Oznacza to, że wszystkie obliczenia zostają dokonane dopiero podczas klasyfikacji. KNN, czyli k nearest neighbors, może zostać użyty zarówno do regresji jak i klasyfikacji. Pseudo algorytm może być opisany jakos:

Kroki wstępne:

1. Załadowanie zbioru danych
2. Przypisanie wartości k

Kroki predykcji

Dla każdej instancji ze zbioru treningowego:

1. Oblicz dystans (np. Euklidesowy) pomiędzy nową instancją, a wszystkimi ze zbioru
2. Posortuj dystanse od najmniejszej do największej
3. Wybierz k pierwszych wyników
4. Wybierz najczęściej występującą klasę z danych odległości (głosowanie)
5. Zwróć wynik głosowania

Jednym z problemów algorytmu KNN jest odpowiednie dobranie parametru k. Jeżeli dany parametr będzie zbyt niski pojawi się overfitting. Natomiast w przypadku zbyt dużego k granice klas stają się bardziej gładkie a błąd predykcji, ponieważ nie mamy zbyt dokładnych granic.

2. Wprowadzenie do zbiorów danych

Podczas analizy i implementacji użyte zostały cztery zbiory danych. Zbiory podzielone są na dwie części. Pierwszą z nich są cechy, dokładnie wektor, cech oraz etykiety mówiące o przynależności wektora cech do konkretnej klasy.

Wszystkie zbiory dostępne są do pobrania ze strony
https://archive.ics.uci.edu/ml/datasets.html

Zbiory danych zostały ściągnięte i załadowane przy użyciu biblioteki *pandas* lub bezpośrednio załadowane za pomocą biblioteki *scikit-learn*.

Zbiory danych:

- Iris data set
- Wine data set
- Glass identification data set
- Pima diabetes data set
- Seeds dataset

Ciekawostką jest, że w trakcie badania klasyfikatora i tworzenia sprawozdania ostatni ze zbiorów *Pima diabetes* został usunięty ze strony UCI ze przez ograniczenie uprawnień do udostępniania danego zbioru.

I'm sorry, the dataset "pima indians diabetes" does not appear to exist.

A note from the donor regarding Pima Indians Diabetes data:

"Thank you for your interest in the Pima Indians Diabetes dataset. The dataset is no longer available due to permission restrictions."

Rysunek 1 Wiadomość ze strony https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes mówiąca o braku dalszego dostępu do danego zbioru.

Poniżej zaprezentowano opis zbiorów. Opis ten pomoże w zrozumieniu danych, które będą analizowane. Dobre zrozumienie danych z którymi się pracuje jest niezbędną częścią do poprawnego przeprowadzenia badań.

Zbiór Iris

Jest to prawdopodobnie jeden z najbardziej znanych i podstawowych zbiorów danych przy problemach klasyfikacji i rozpoznawania wzorców.

Zbiór składa się ze 150 instancji, podzielonych na 3 równe zbiory po 50 klas każda.

Definicje atrybutów:

- Sepal – zielony płatek u dołu kielicha służący do ochrony kwiatu w trakcie kwitnięcia,
- Petal – płatek kwiatu, służący do przyciągania uwagi ptaków i insektów

Cechy zbioru zawierają cztery informacje:

1. Sepal length in cm
2. Sepal width in cm
3. Petal length in cm
4. Petal width in cm

Ostatnią, piątą kolumną jest klasa mówiąca o typie irysa. Możliwe są trzy klasy:

1. Iris Setosa
2. Iris Versicolour
3. Iris Virginica

Zbiór Wine

Zbiór ten został skonstruowany w wyniku analizy składu chemicznego win stworzonych w tym samym rejonie Włoch lecz przy użyciu trzech różnych odmian uprawnych.

Zbiór składa się ze 178 instancji.

Definicje atrybutów oraz cechy zbioru:

1. Alcohol – alkohol
2. Malic acid – kwas jabłkowy
3. Ash – popiół
4. Alkalinity of ash – alkaliczność popiołu
5. Magnesium – magnez
6. Total phenols – całkowita zawartość fenoli
7. Flavonoids – flawonoidy
8. Nonflavanoid phenols – fenole nieflawonowe
9. Proanthocyanidins – proantocyjanidyny
10. Color intensity, intensywność koloru
11. Hue – odcień
12. OD280/OD315 of diluted wines - OD280 / OD315 rozcieńczonych win
13. Proline – Proline

Pierwszy atrybut w pliku zawierającym dane jest identyfikatorem klasy od 1 do 3.

Rozłożenie instancji klas jest następujące:

- Klasa 1 – 59 instancji,
- Klasa 2 – 71 instancji,
- Klasa 3 – 48 instancji.

Zbiór Glass identification

Zbiór powstał poprzez analizę składu chemicznego badanego szkła aby określić typ powstałego szkła oraz jego przeznaczenie.

Zbiór składa się z 214 instancji podzielonych na 6 klas.

Rozłożenie instancji klas jest następujące:

- Klasa 1 – 70 instancji,
- Klasa 2 – 76 instancji,
- Klasa 3 – 17 instancji,
- Klasa 4 - 13,
- Klasa 5 - 9,
- Klasa 6 - 29.

Definicje atrybutów oraz cechy zbioru:

1. Id – numer porządkowy
2. Refractive index – współczynnik załamania światła
3. Sodium – sód
4. Magnesium – magnez
5. Aluminium – glin
6. Silicon – krzem
7. Potassium – potas
8. Calcium – wapń
9. Barium – bar
10. Iron – żelazo

Zbiór Seeds

Zbiór reprezentuje atrybuty 3 różnych typów zbóż.

Zbiór składa się z 210 instancji podzielonych na 3 klasy.

Rozłożenie instancji klas jest następujące:

- Klasa 1 (Kama) – 70 instancji,
- Klasa 2 (Rosa) – 70 instancji,
- Klasa 3 (Canadian) – 70 instancji,

Definicje atrybutów oraz cechy zbioru. Wszystkie atrybuty są miarami nasion zboża:

1. Area – pole
2. Perimeter – obwód
3. Compactness – ścisłość
4. Length of kernel – długość nasiona
5. Width of kernel – szerokość nasiona
6. Asymmetry coefficient – współczynnik asymetrii
7. Length of kernel groove - długość rowka nasiona

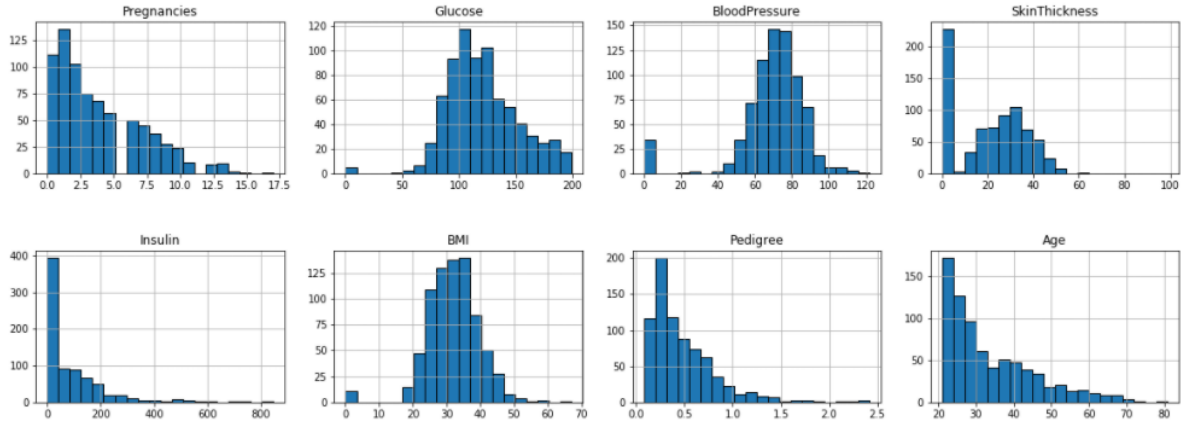
Zbiór Pima diabetes

Celem zbioru jest umożliwienie zdiagnozowania czy dany pacjent ma cukrzycę, bazując na diagnostykach zamieszczonych w cechach zbioru. Wszyscy pacjenci przebadani byli kobietami mającymi przynajmniej 21 lat oraz byli pochodzenia indiańskiego plemienia Pima.

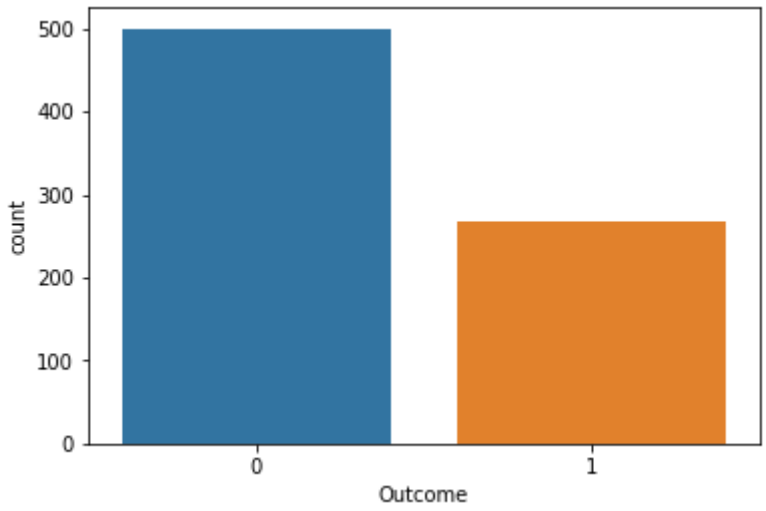
Zbiór składa się z 768 instancji posiadających dwie możliwe klasy 1 – oznaczające że zbadana osoba jest chora na cukrzycę, 0 – oznaczające że dana osoba nie jest chora na cukrzycę.

Definicje atrybutów oraz cechy zbioru:

1. Pregnancies – liczba ciąży
2. Glucose – poziom glukozy
3. Blood ressure – ciśnienie krwi
4. Skin thickness – grubość skóry
5. Insulin – poziom insuliny
6. BMI – body mass index
7. Diabetes pedigree function – funkcja pedigree
8. Age – wiek



Rysunek 2 Histogramy atrybutów danych zbioru Pima diabetes



Rysunek 3 Rozkład klas

3. Implementacja

Do implementacji użyty został język programowania Python wraz z bibliotekami takimi jak: pandas, sklearn, numpy, matplotlib

Parametry badane w knn

N (number of votes) – Podstawowy parametr, który odpowiada za liczbę sąsiednich punktów, które biorą udział w głosowaniu.

Weights – Parametr używany podczas głosowania, przyjmuje 2 wartości: ‘uniform’ i ‘distance’. W pierwszej wersji podczas głosowania wszystkie głosy są tak samo ważne i na żadne z nich nie są nakładane wagi. W przypadku ‘distance’ głosy są ważone z wagą równą odwrotności ich odległości do badanego punktu. Oznacza to, że głosy instancji, które są bliżej nowego punktu są ważniejsze podczas klasyfikacji.

Metric – Ten parametr mówi o tym w jaki sposób obliczane są odległości do badanego punktu. Zbadane zostaną 3 miary mierzenia odległości:

1. **Euclidean** – pierwiastek z sumy kwadratów różnic między współrzędnymi, linia prosta pomiędzy dwoma punktami,
2. **Manhattan** – suma bezwzględnych różnic współrzędnych punktów,
3. **Chebyshev** – maksimum bezwzględnych różnic między współrzędnymi punktów.

4. Analiza wyników zbiorów danych

a. Analiza wyników zbioru Iris dataset

Stratified Crossvalidation

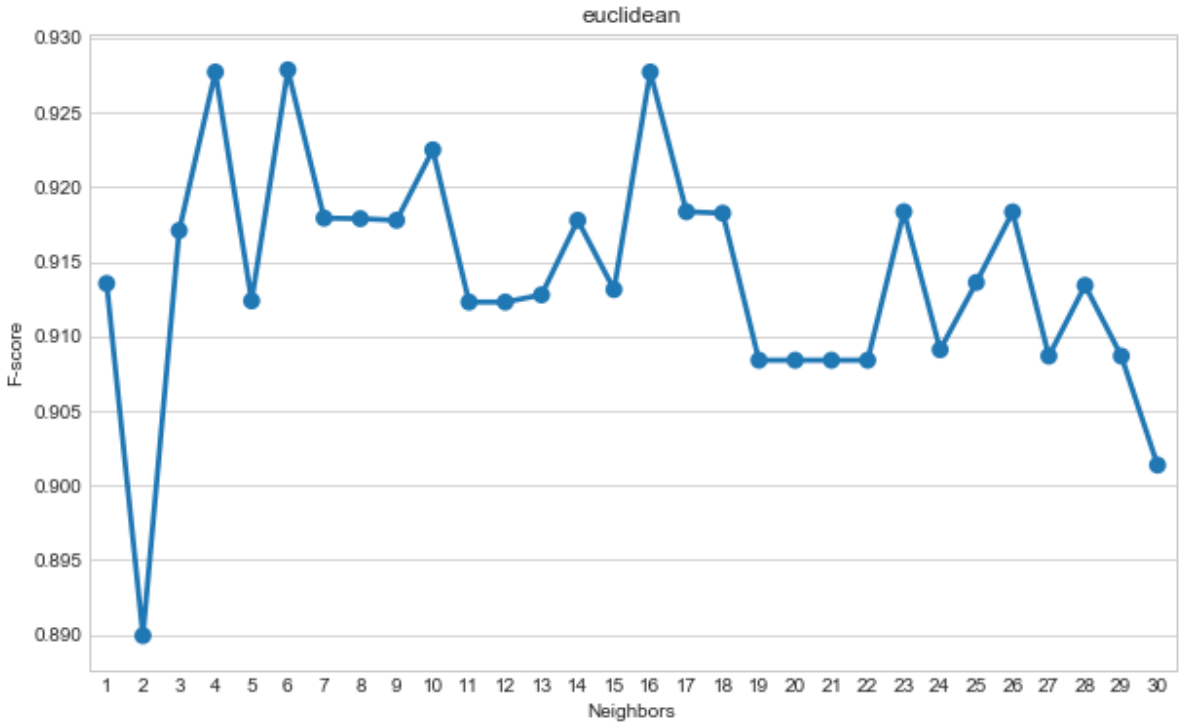
Metryka odl.	Waga	Średni F-score	Średni Accuracy	Najlepszy F-score	Optymalne K
Euklidean	Uniform	0.901	0.905	0.928	6
Euklidean	Distance	0.918	0.919	0.933	8
Manhattan	Uniform	0.907	0.910	0.928	12
Manhattan	Distance	0.912	0.914	0.928	12
Chebyshev	Uniform	0.927	0.929	0.933	11
Chebyshev	Distance	0.927	0.929	0.933	11

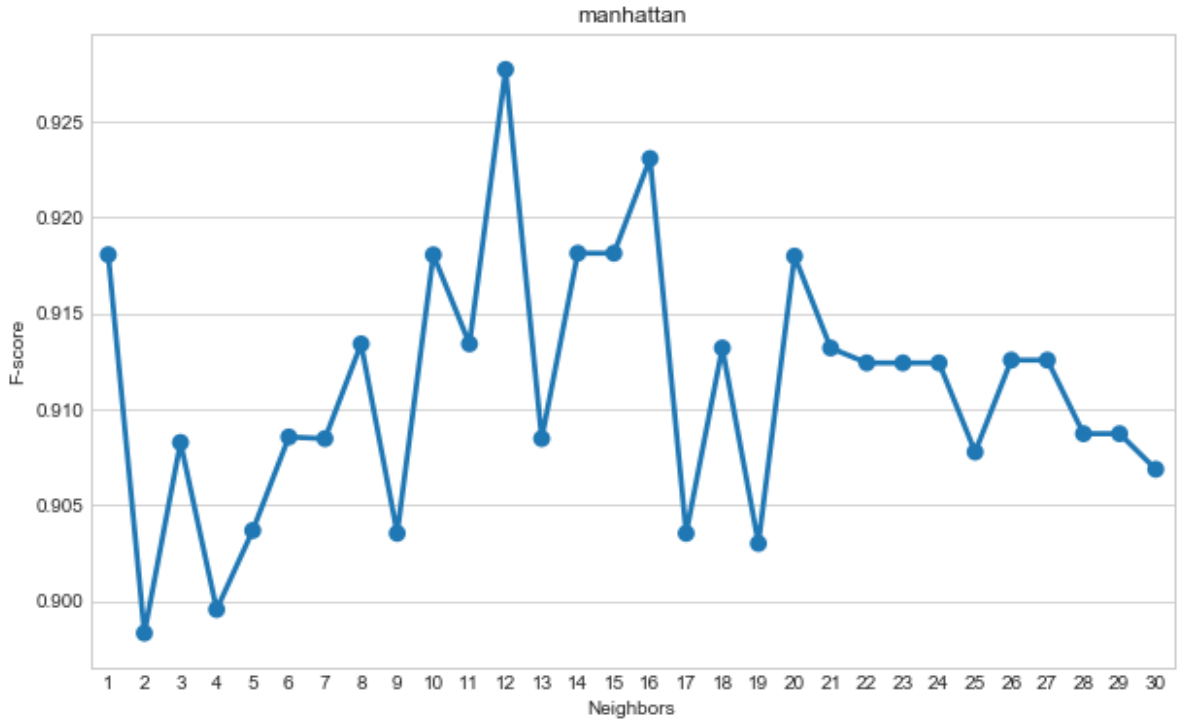
Crossvalidation

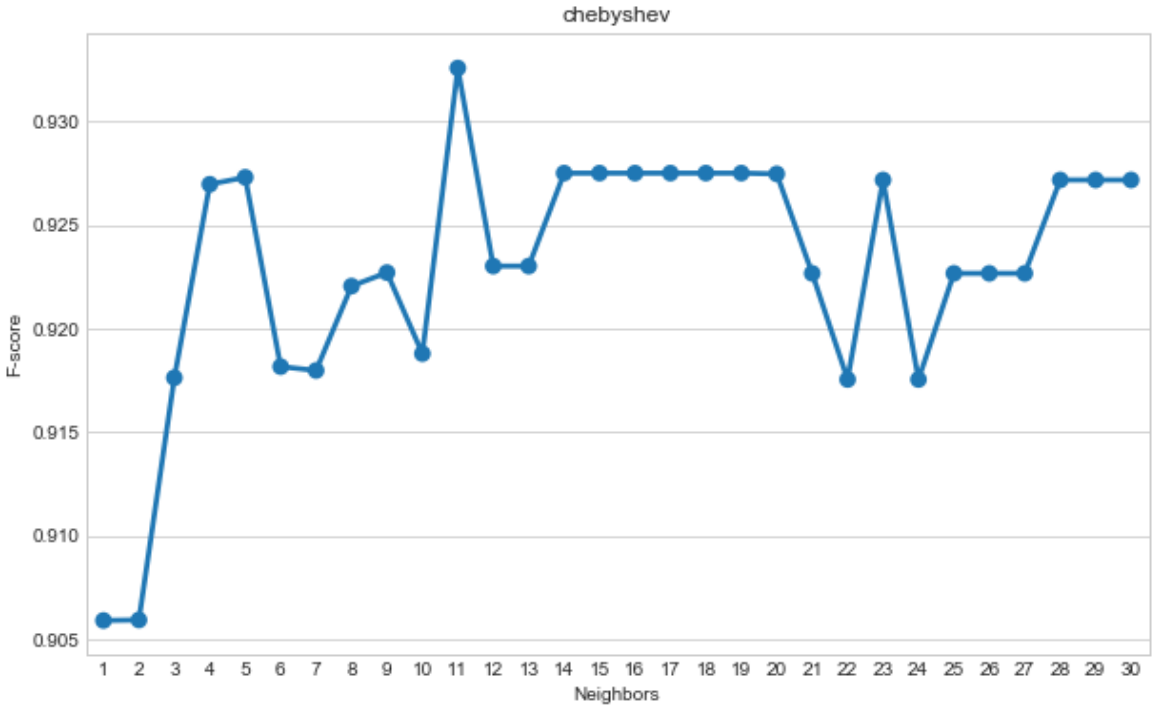
Metryka odl.	Waga	Średni F-score	Średni Accuracy	Najlepszy F-score	Optymalne K
Euklidean	Uniform	0.694	0.920	0.837	6
Euklidean	Distance	0.732	0.933	0.788	5
Manhattan	Uniform	0.688	0.900	0.786	3
Manhattan	Distance	0.732	0.933	0.788	25
Chebyshev	Uniform	0.689	0.900	0.839	1
Chebyshev	Distance	0.786	0.947	0.839	1

W przypadku analizy zbioru Iris widać zdecydowaną przewagę stratyfikowanej kroswalidacji w porównaniu ze zwykłą. W przypadku SKF różnica pomiędzy f-score a accuracy nie była duża. Najlepsze wyniki dla SKF uzyskano przy użyciu odległości euklidesowej i bez wag nałożonych na głosowanie. Najgorszą miarą okazało się być Manhattan wraz z brakiem wag, jednakże o niewiele. Dla odległości euklidesowych w SKF optymalne K były bardzo wysokie, a w przypadku zwykłej KF niskie.

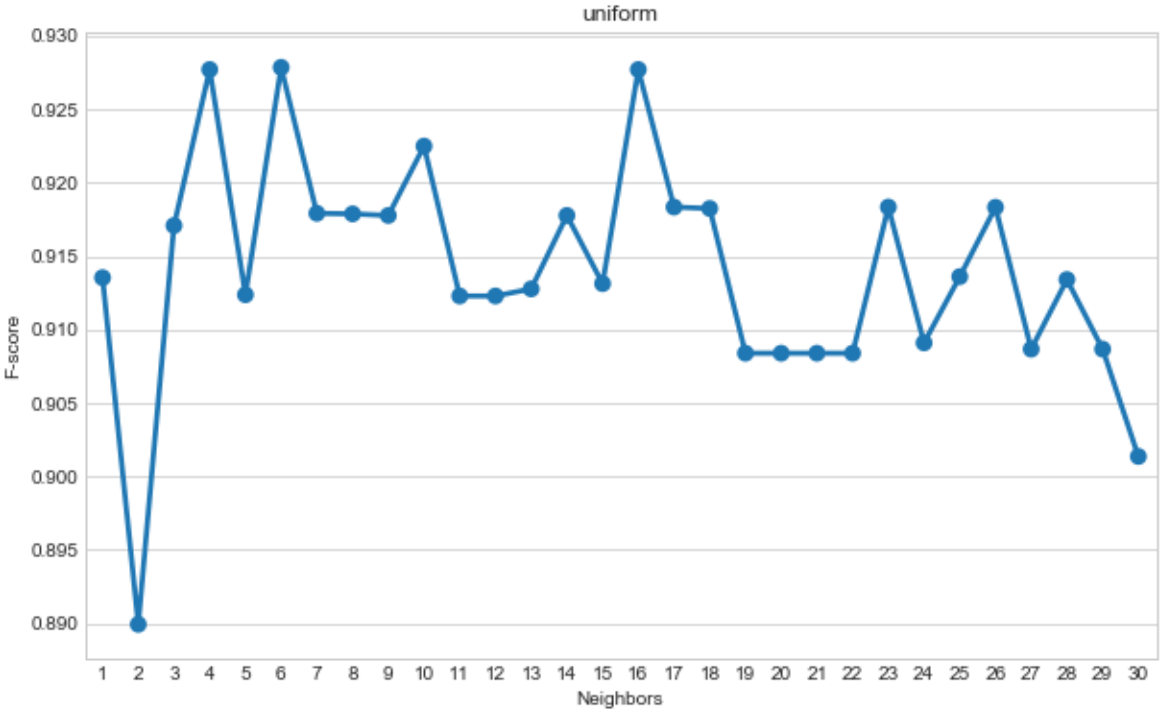
Analiza algorytmów miar odległości, przy pominięciu wag, kroswalidacji stratyfikowanej z folds=10.

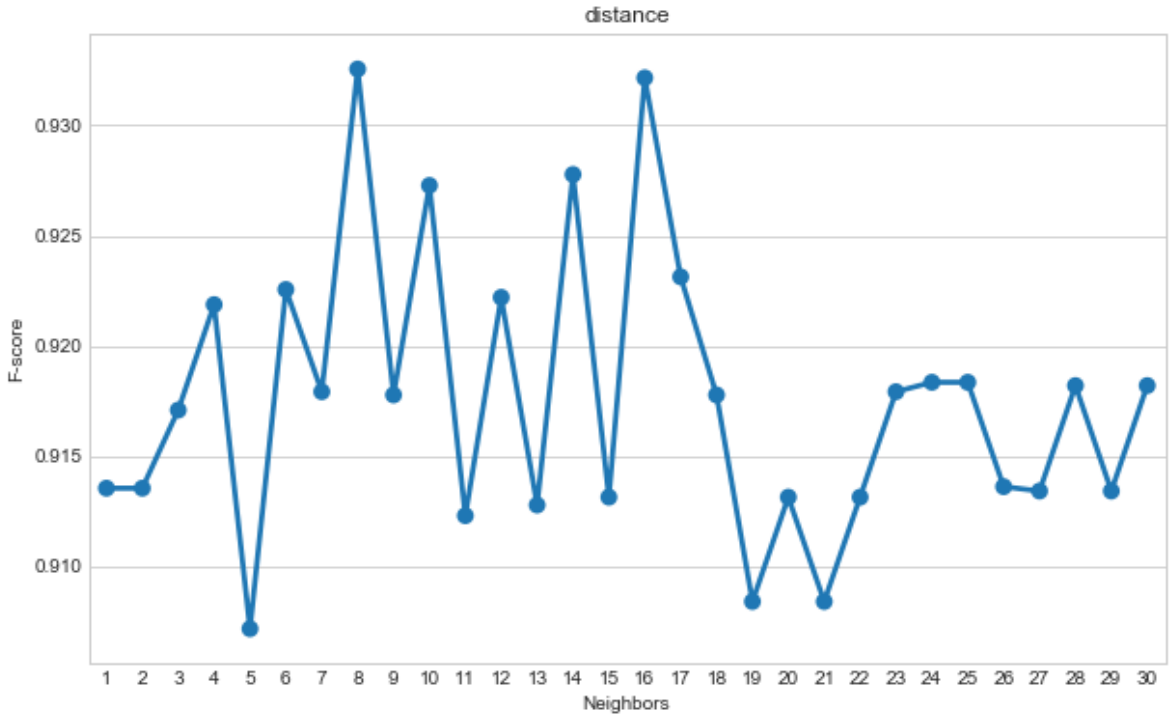






Analiza wyników klasyfikacji z nałożeniem lub brakiem wag na głosy, przy pomocy kroswalidacji stratyfikowanej z fold=10 i euklidesową miarą odległości.





b. Analiza wyników zbioru Pima Diabetes dataset

Stratified Crossvalidation

Metryka odl.	Waga	Średni F-score	Średni Accuracy	Najlepszy F-score	Optymalne K
Euklidean	Uniform	0.692	0.751	0.719	5
Euklidean	Distance	0.708	0.758	0.719	15
Manhattan	Uniform	0.684	0.749	0.712	15
Manhattan	Distance	0.694	0.749	0.714	21
Chebyshev	Uniform	0.669	0.737	0.704	11
Chebyshev	Distance	0.691	0.749	0.713	13

Crossvalidation

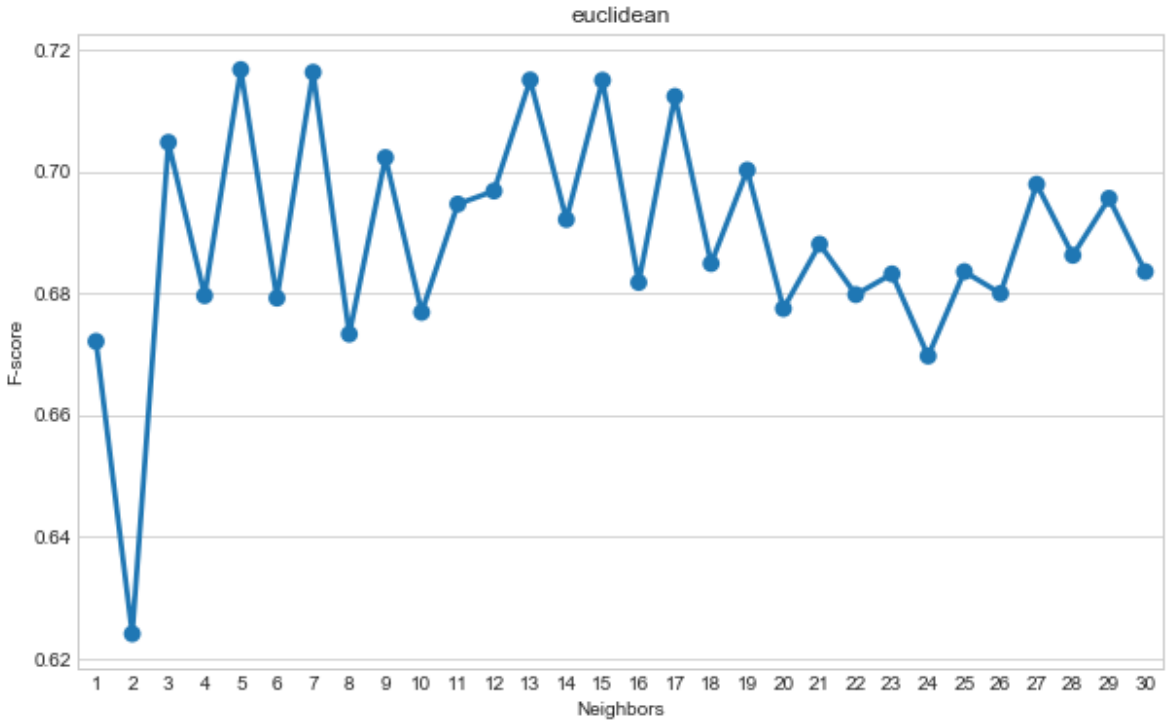
Metryka odl.	Waga	Średni F-score	Średni Accuracy	Najlepszy F-score	Optymalne K
Euklidean	Uniform	0.682	0.749	0.714	7
Euklidean	Distance	0.705	0.758	0.723	15
Manhattan	Uniform	0.689	0.754	0.709	23
Manhattan	Distance	0.693	0.753	0.717	15
Chebyshev	Uniform	0.660	0.734	0.692	11
Chebyshev	Distance	0.688	0.747	0.708	11

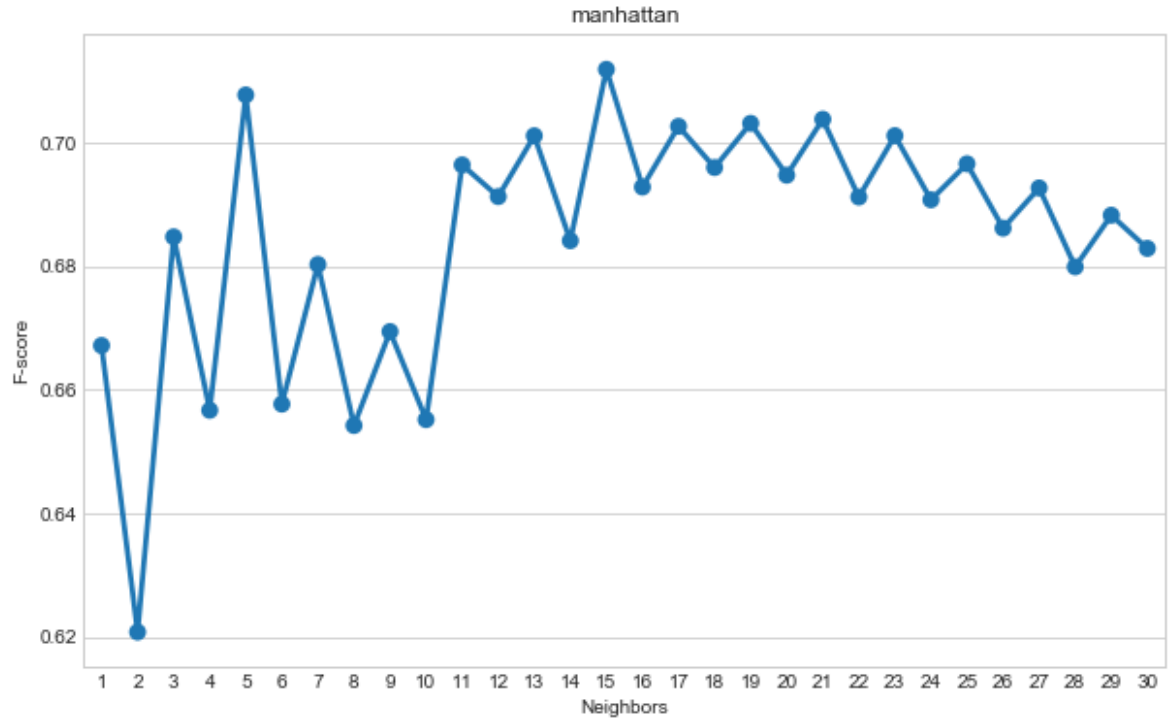
Dla zbioru Pima diabetes wyniki kroswalidacji zwykłej i stratyfikowanej były bardzo zbliżone do siebie. Najlepsze wyniki uzyskano tak jak w przypadku Iris dataset dla

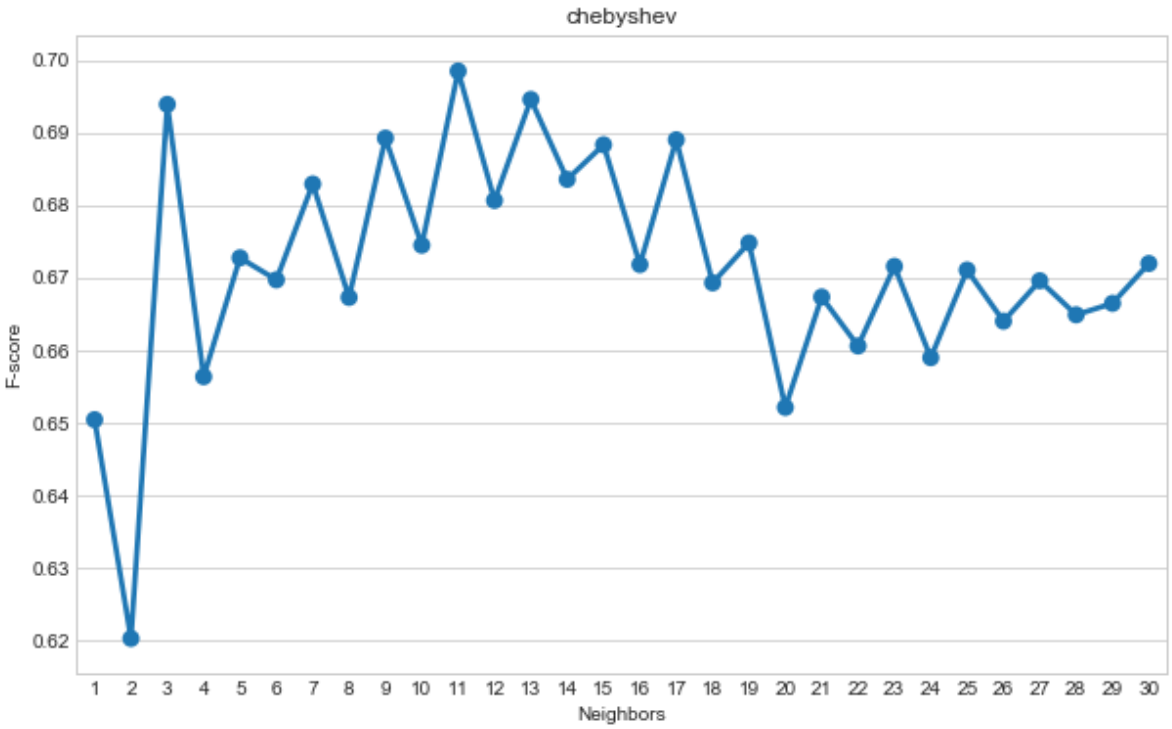
odległości euklidesowej jednakże tym razem z uwzględnieniem wag. Optymalne K wyniosło 15. Najgorsze okazało się połączenie odległości Chebysheva wraz z brakiem wag, które przyniosło spadek o około 5%.

Klasyfikator	**F-score**
Naive Bayes	0.75
C45	0.83
KNN	0.71

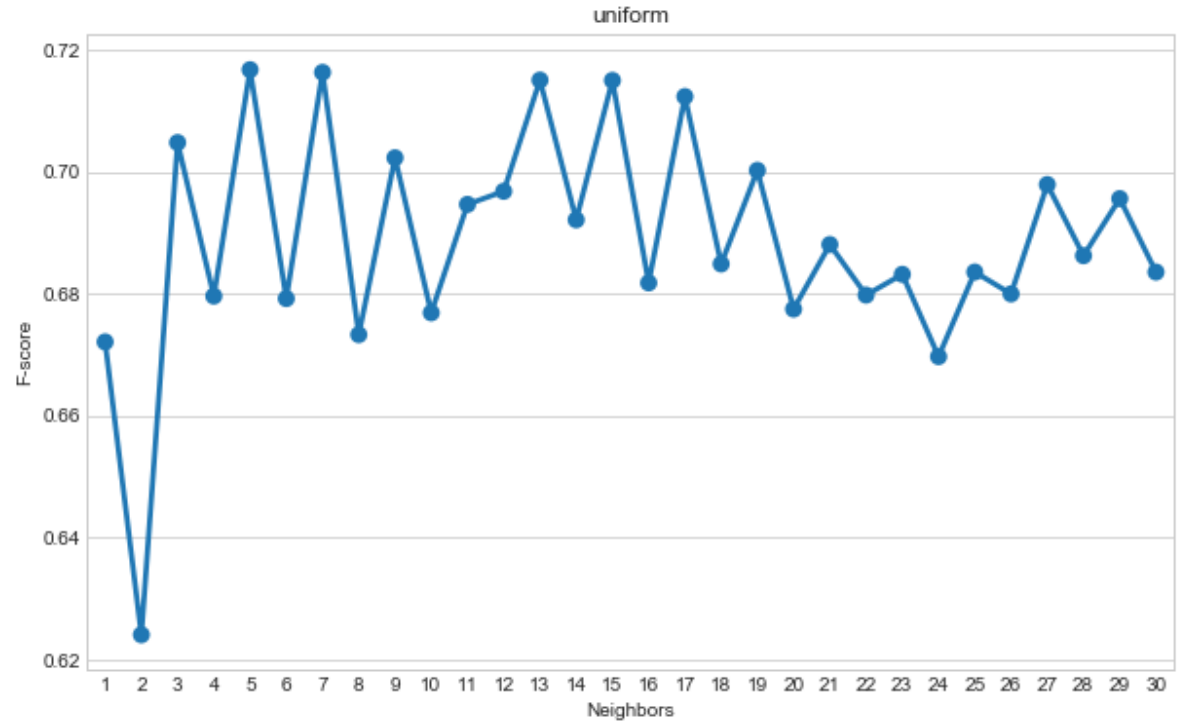
Analiza algorytmów miar odległości, przy pominięciu wag, kroswalidacji stratyfikowanej z folds=10.

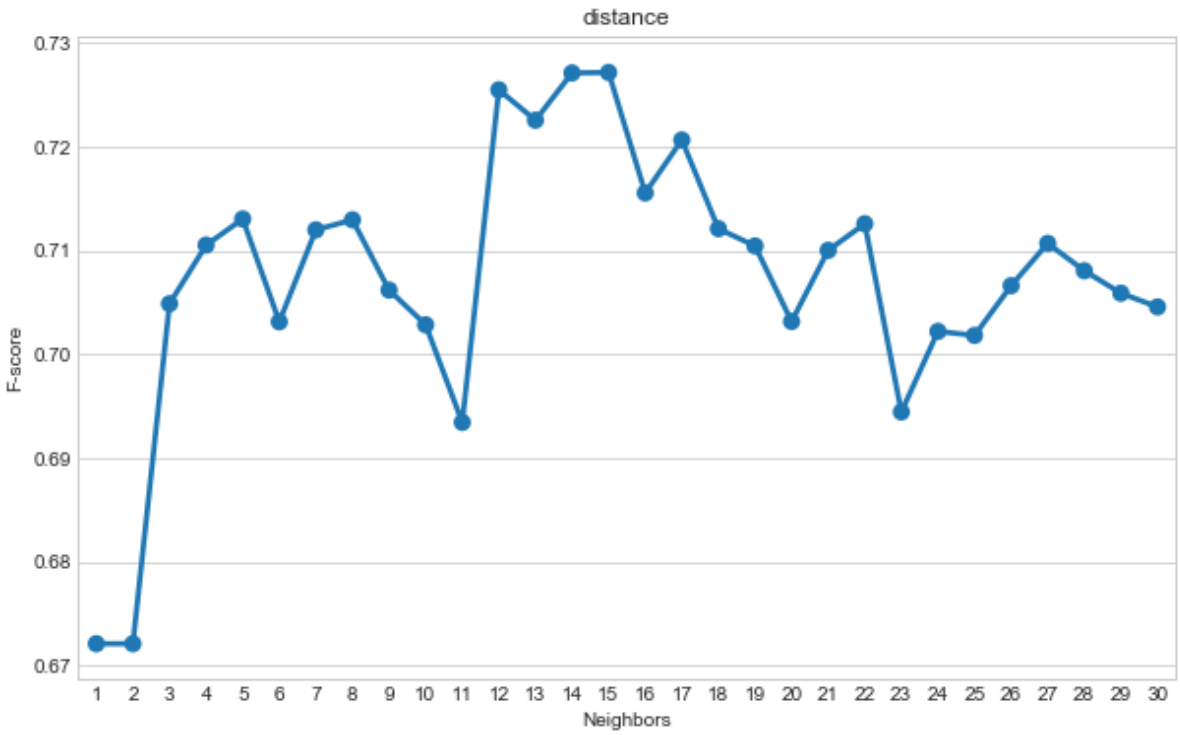






Analiza wyników klasyfikacji z nałożeniem lub brakiem wag na głosy, przy pomocy kroswalidacji stratyfikowanej z fold=10 i euklidesową miarą odległości.





c. Analiza wyników zbioru Glass dataset

Stratified Crossvalidation

Metryka odl.	Waga	Średni F-score	Średni Accuracy	Najlepszy F-score	Optymalne K
Euklidean	Uniform	0.348	0.623	0.600	1
Euklidean	Distance	0.433	0.635	0.600	1
Manhattan	Uniform	0.349	0.622	0.591	1

Manhattan	Distance	0.462	0.650	0.621	3
Chebyshev	Uniform	0.323	0.578	0.567	1
Chebyshev	Distance	0.351	0.588	0.584	3

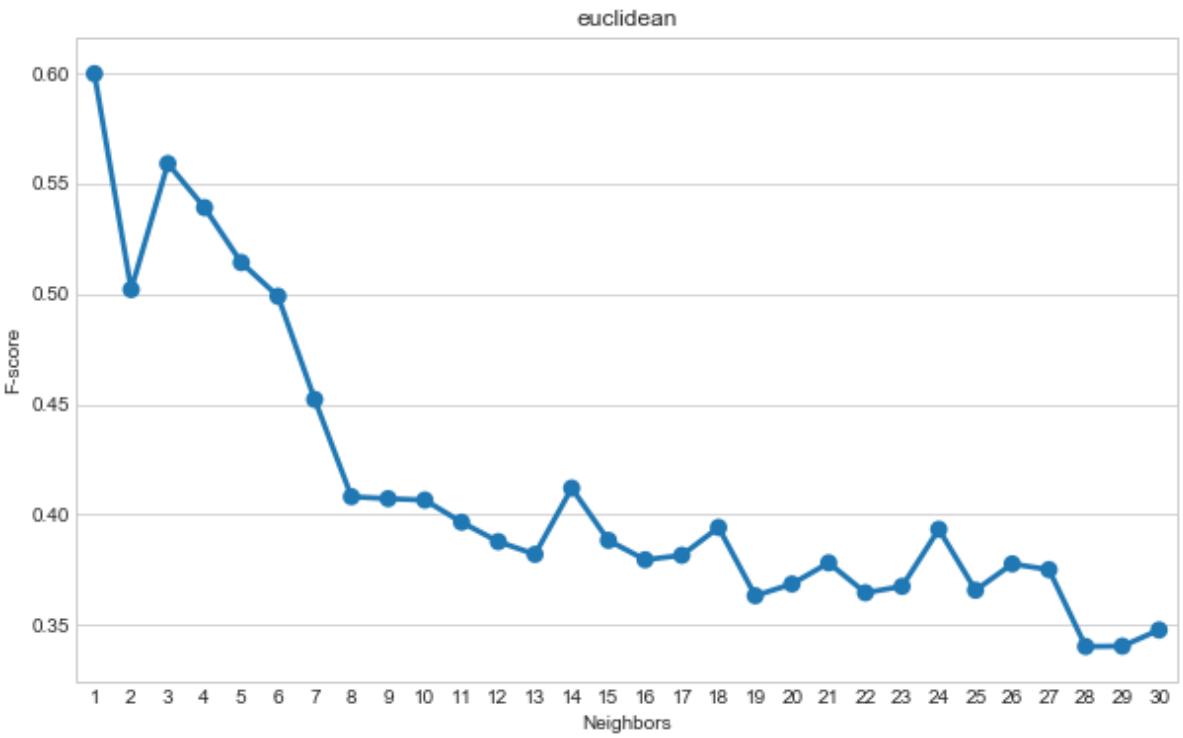
Crossvalidation

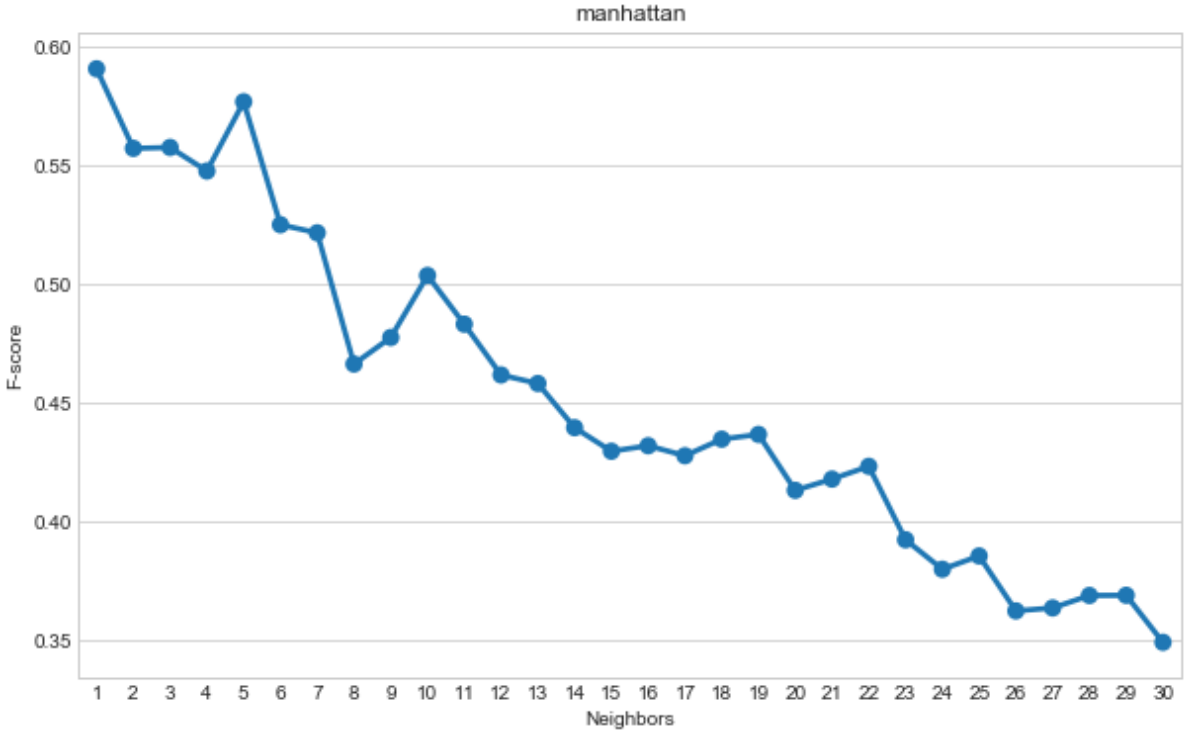
Metryka odl.	Waga	Średni F-score	Średni Accuracy	Najlepszy F-score	Optymalne K
Euklidean	Uniform	0.195	0.301	0.291	7
Euklidean	Distance	0.238	0.355	0.303	10
Manhattan	Uniform	0.221	0.346	0.327	5
Manhattan	Distance	0.268	0.429	0.297	11
Chebyshev	Uniform	0.172	0.249	0.292	2
Chebyshev	Distance	0.224	0.327	0.286	6

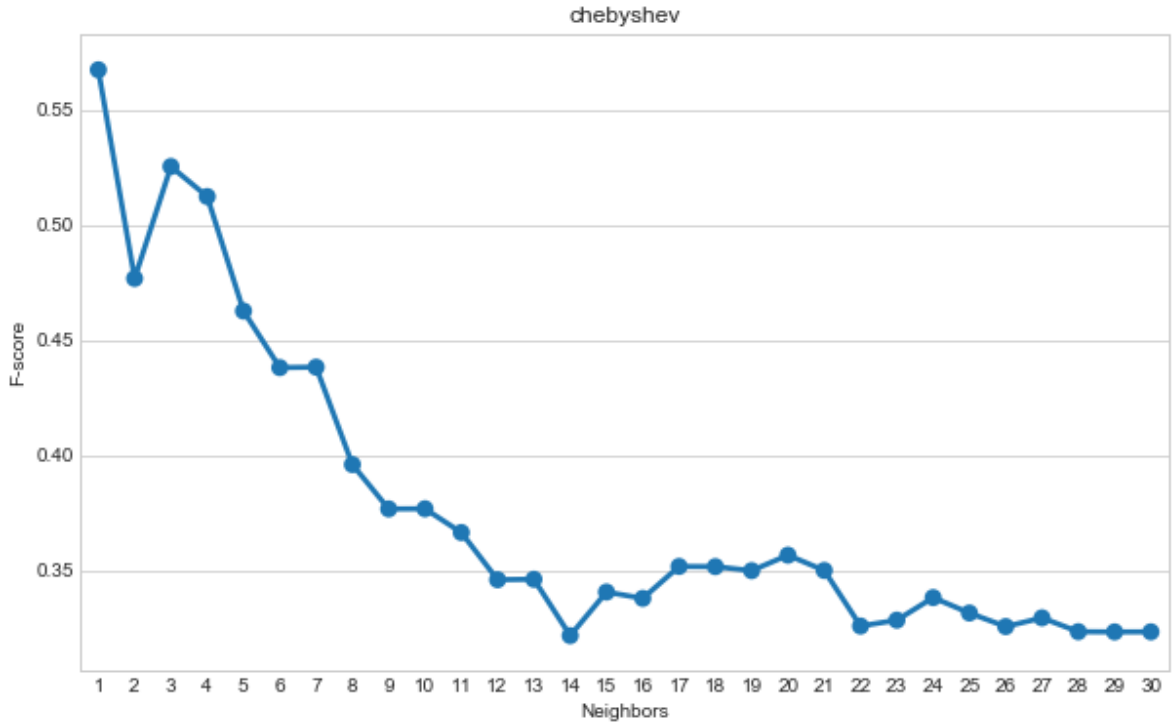
Dla zbioru Glass dataset uzyskano bardzo słabe wyniki w przypadku klasyfikacji za pomocą KNN. Wyniki kroswalidacji stratyfikowanej utrzymywały się w okolicach 30-40%, a bez stratyfikacji około 17-27%. Stratyfikacja przyniosła zysk około 15%. Najlepsza okazała się po raz pierwszy miara odległości Manhattan połączona z brakiem wag na głosy. Najgorsza okazała się miara odległości Chebyshev'a, która zarówno dla stratyfikacji i jej braku przynosiła bardzo słabe wyniki.

Klasyfikator	F-score
Naive Bayes	0.57
C45	0.78
KNN	0.46

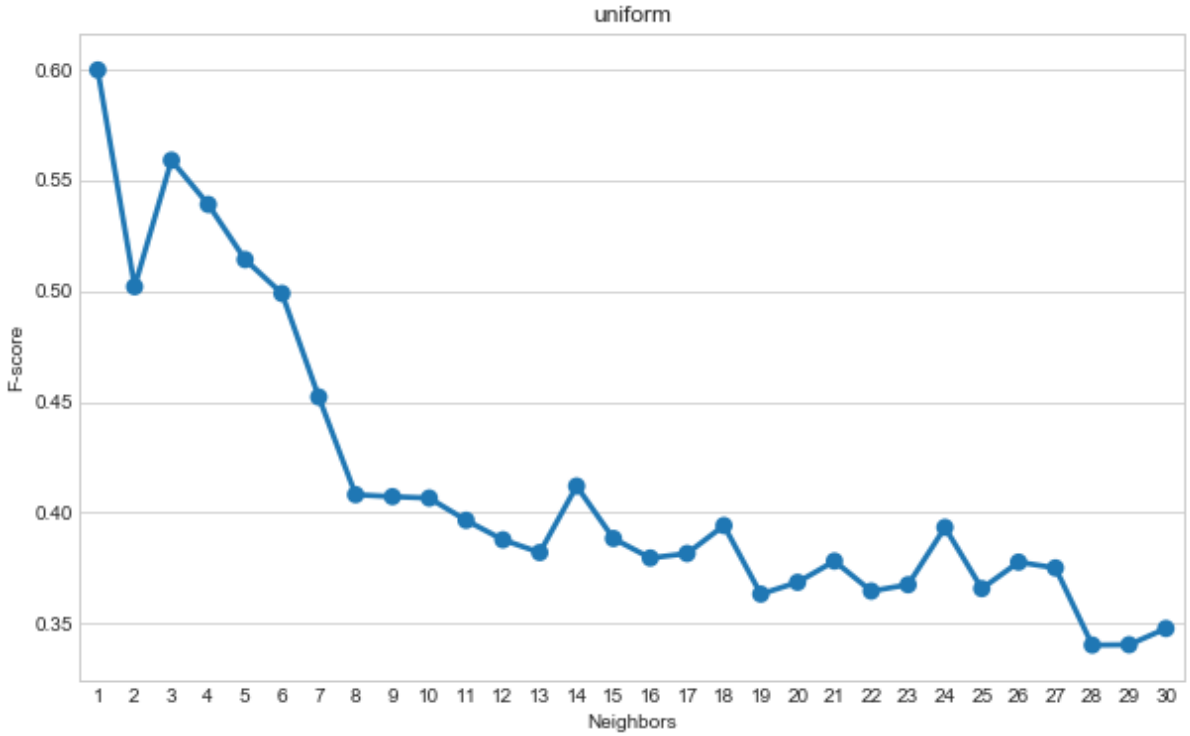
Analiza algorytmów miar odległości, przy pominięciu wag, kroswalidacji stratyfikowanej z folds=10.

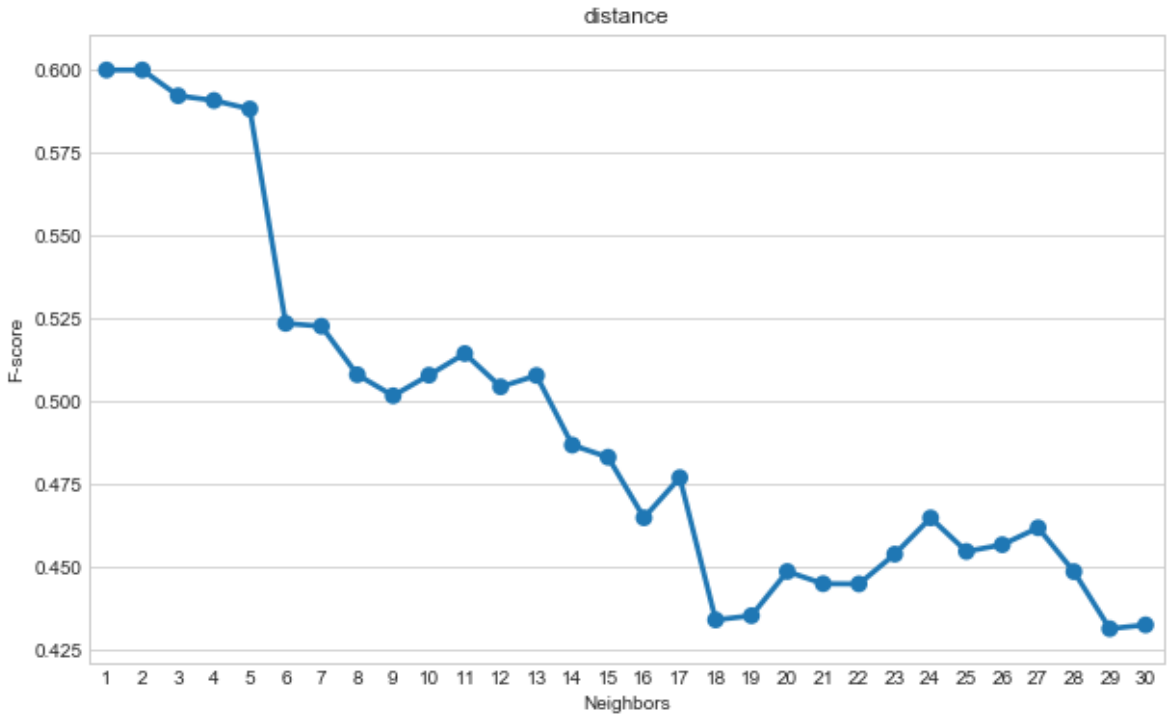

euclidean
F-score
0.60
0.55
0.50
0.45
0.40
0.35
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Neighbors


manhattan
F-score
0.60
0.55
0.50
0.45
0.40
0.35
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Neighbors



Analiza wyników klasyfikacji z nałożeniem lub brakiem wag na głosy, przy pomocy kroswalidacji stratyfikowanej z fold=10 i euklidesową miarą odległości.





d. Analiza wyników zbioru Wine dataset

Stratified Crossvalidation

Metryka odl.	Waga	Średni F-score	Średni Accuracy	Najlepszy F-score	Optymalne K
Euklidean	Uniform	0.963	0.962	0.978	20
Euklidean	Distance	0.968	0.967	0.978	21
Manhattan	Uniform	0.956	0.956	0.980	11
Manhattan	**Distance**	**0.978**	**0.978**	**0.980**	**11**
Chebyshev	Uniform	0.933	0.932	0.950	23
Chebyshev	Distance	0.940	0.938	0.952	20

Crossvalidation

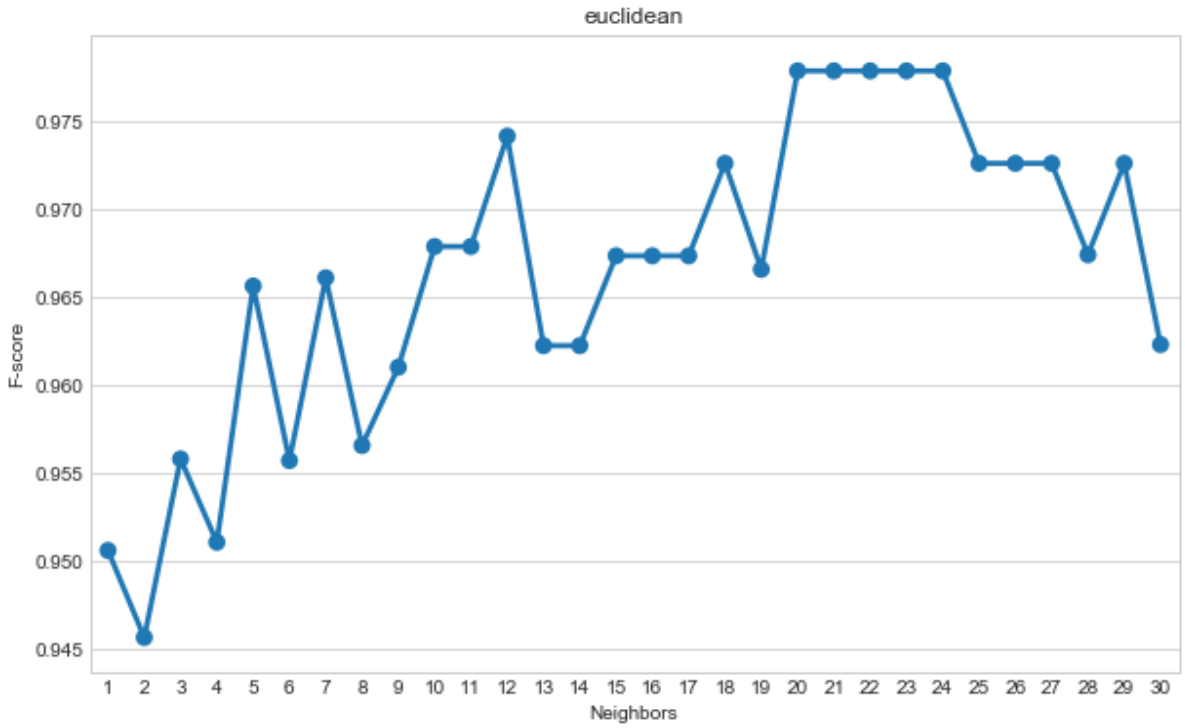
Metryka odl.	Waga	Średni F-score	Średni Accuracy	Najlepszy F-score	Optymalne K
Euklidean	Uniform	0.718	0.944	0.792	2
Euklidean	Distance	0.725	0.961	0.779	7
Manhattan	Uniform	0.764	0.939	0.839	11
Manhattan	Distance	0.765	0.944	0.839	11
Chebyshev	Uniform	0.581	0.883	0.753	3
Chebyshev	Distance	0.643	0.905	0.753	3

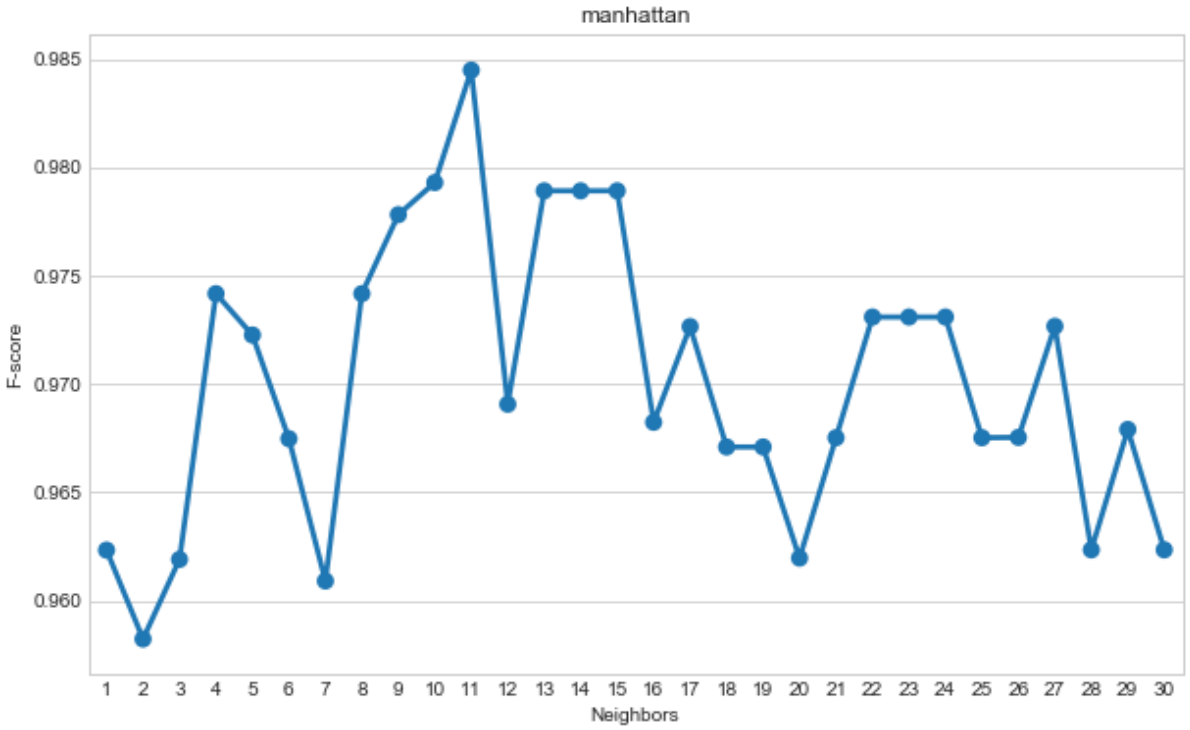
W przypadku Wine dataset uzyskano bardzo dobre wyniki dla SKF, były one na poziomie 93-98%. Zwykła kroswalidacja przyniosła wyniki na poziomie 58-77%, co jest spadkiem średnio o ponad 20%. Najlepsza miara odległości była to odległość Manhattan wraz z ustawieniem wag na głosy. Dodatkowo dla stratyfikowanej

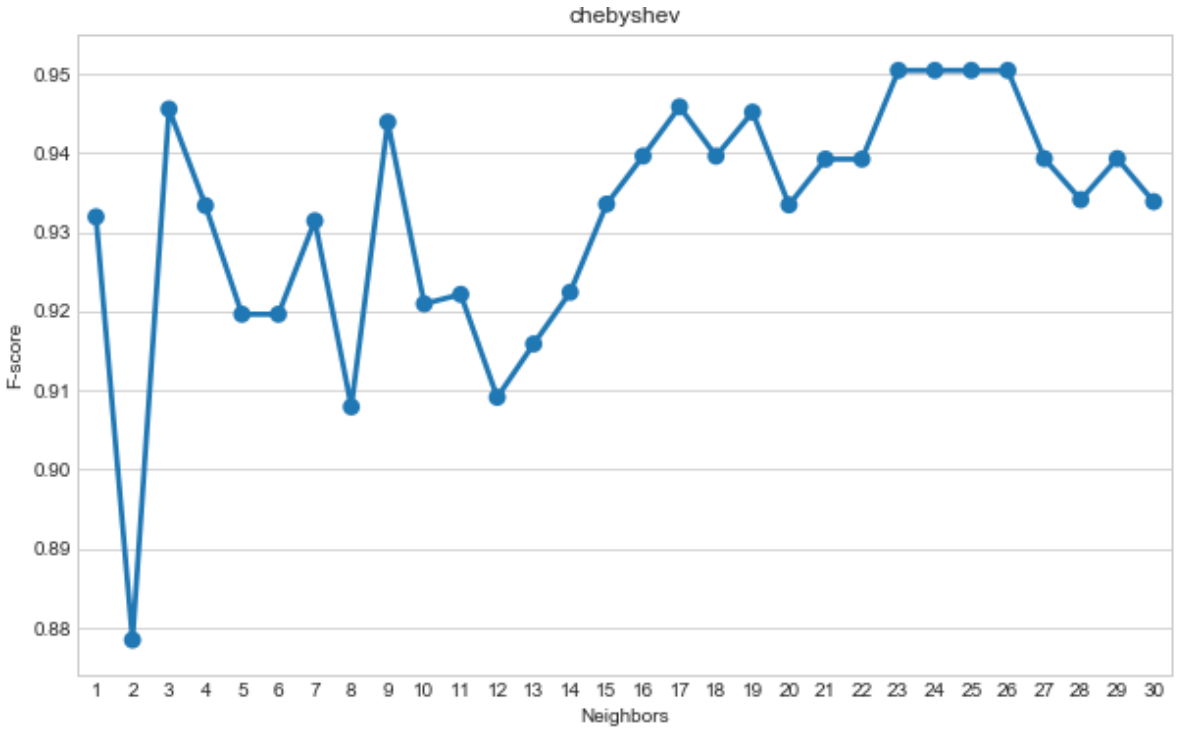
kroswalidacji optymalne parametry głosujących sąsiadów były o wiele wyższe niż w przypadku jej braku.

Klasyfikator	**F-score**
Naive Bayes	0.95
C45	0.97
KNN	0.98

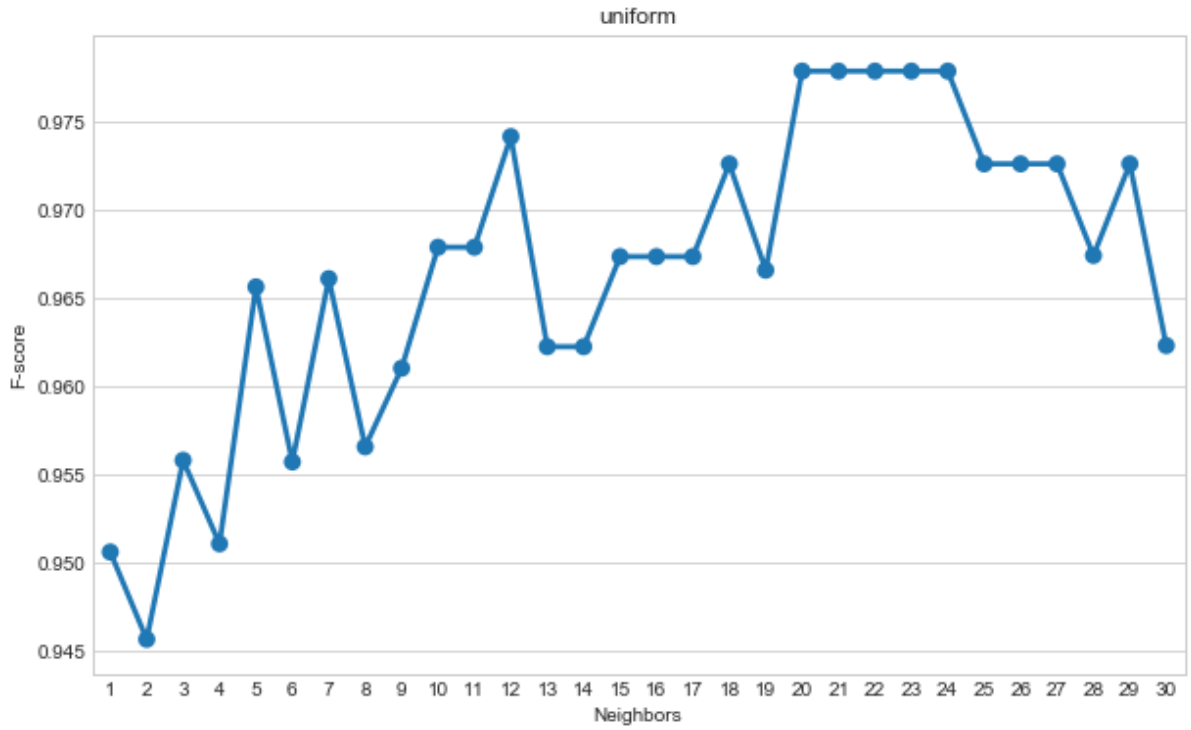
Analiza algorytmów miar odległości, przy pominięciu wag, kroswalidacji stratyfikowanej z folds=10.

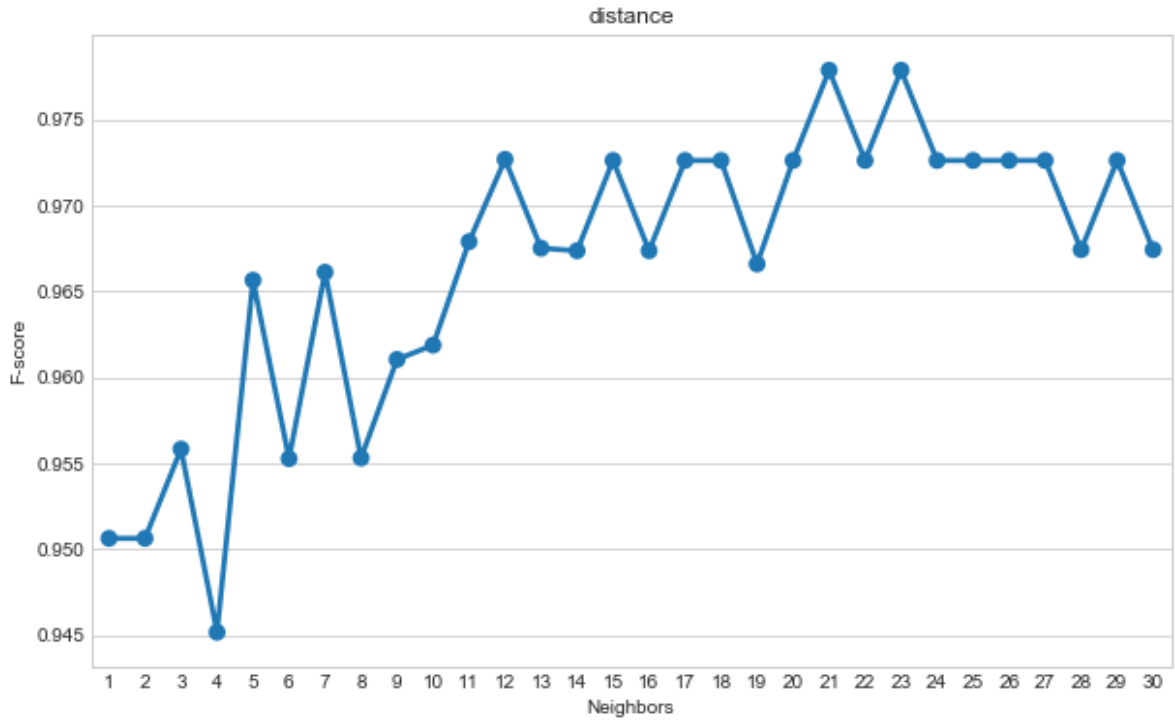






Analiza wyników klasyfikacji z nałożeniem lub brakiem wag na głosy, przy pomocy kroswalidacji stratyfikowanej z fold=10 i euklidesową miarą odległości.





e. Analiza wyników zbioru Seeds dataset

Stratified Crossvalidation

Metryka odl.	Waga	Średni F-score	Średni Accuracy	Najlepszy F-score	Optymalne K

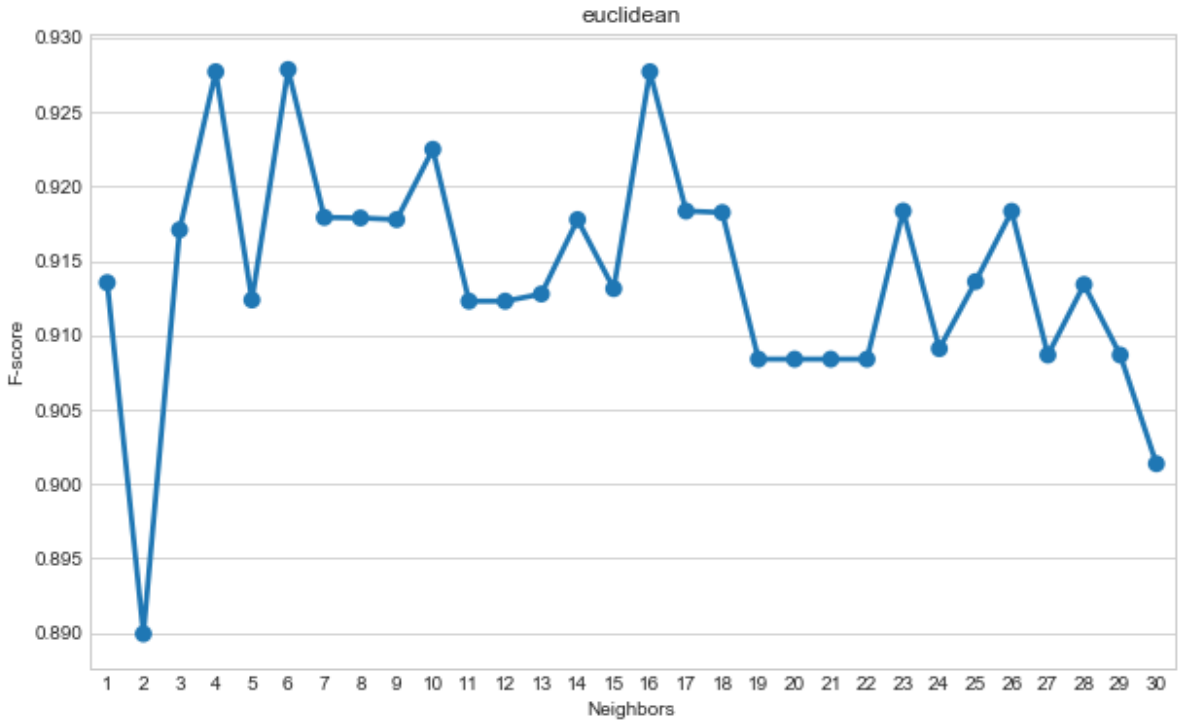
Euklidean	Uniform	0.901	0.905	0.928	6
Euklidean	Distance	0.918	0.919	0.933	8
Manhattan	Uniform	0.907	0.910	0.928	12
Manhattan	Distance	0.912	0.914	0.928	12
Chebyshev	Uniform	0.927	0.929	0.933	11
Chebyshev	Distance	0.927	0.928	0.933	14

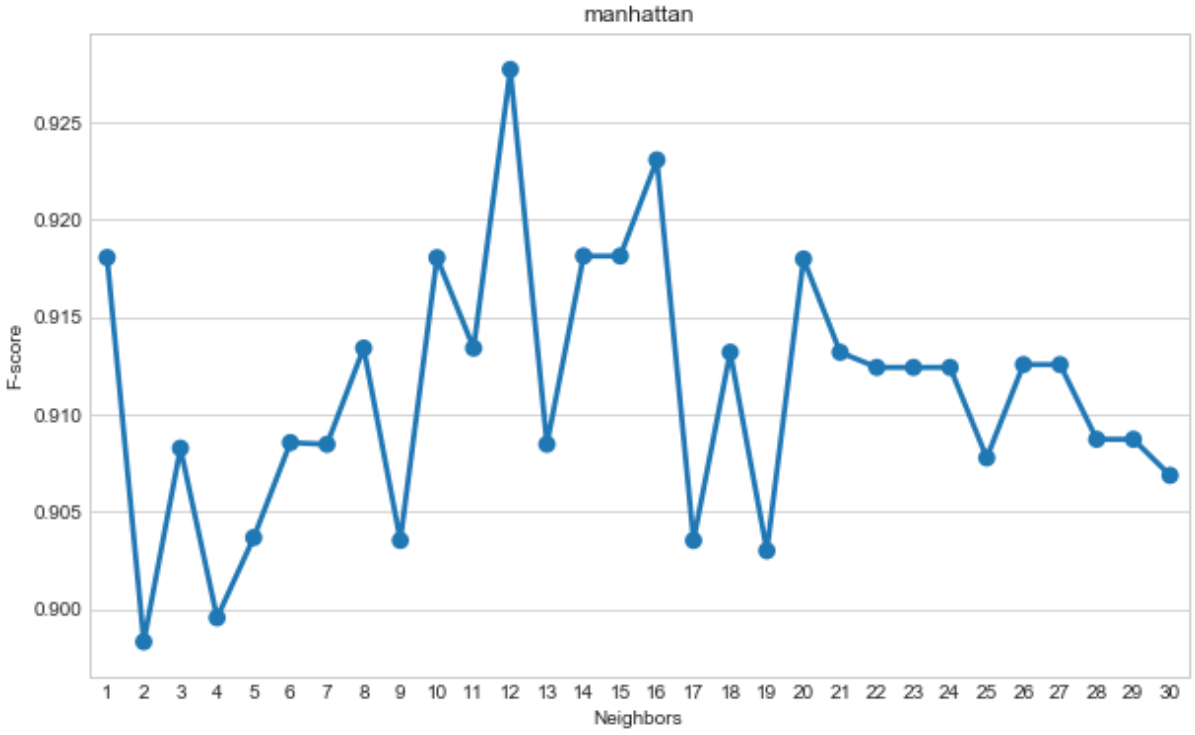
Crossvalidation

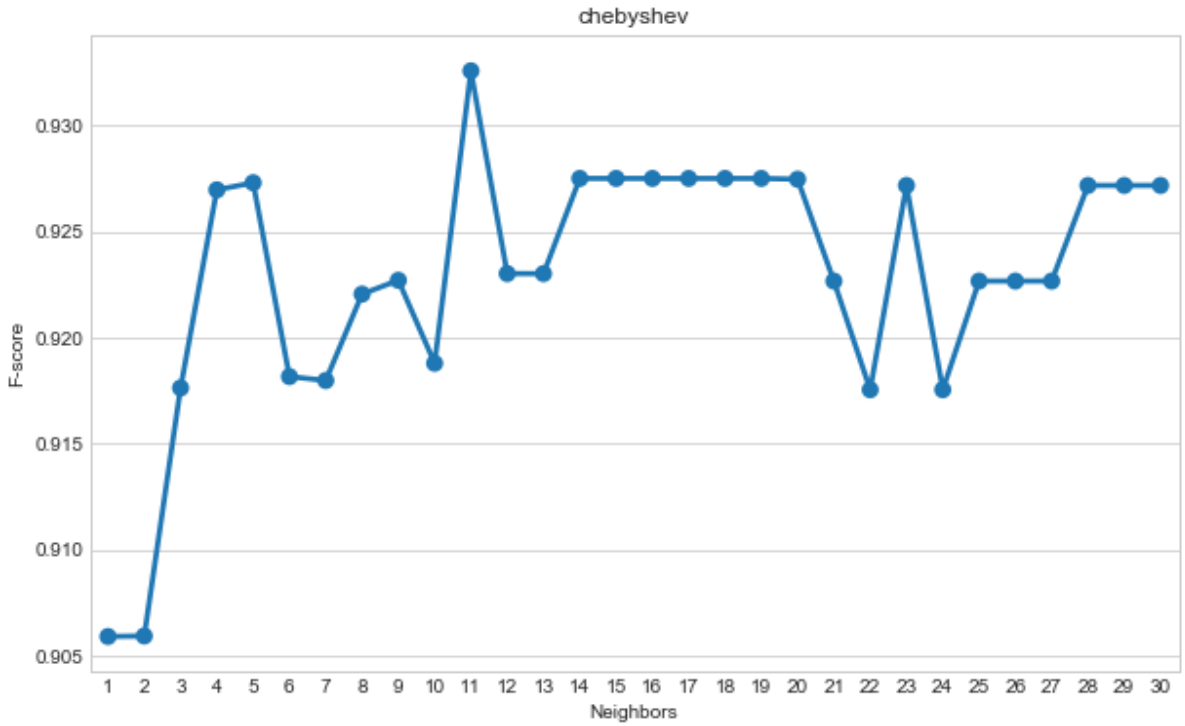
Metryka odl.	Waga	Średni F-score	Średni Accuracy	Najlepszy F-score	Optymalne K
Euklidean	Uniform	0.550	0.876	0.557	12
Euklidean	Distance	0.551	0.876	0.555	12
Manhattan	Uniform	0.596	0.862	0.607	11
Manhattan	Distance	0.599	0.871	0.607	11
Chebyshev	Uniform	0.506	0.900	0.558	11
Chebyshev	Distance	0.506	0.900	0.558	11

Dla zbioru Seeds uzyskano o wiele wyższe wyniki dla kroswalidacji stratyfikowanej. Dla SKF były one w granicach 91%, a w przypadku zwykłej KF około 50-59%. Jest to zysk stratyfikacji na poziomie 40%. Dodatkowo w przypadku stratyfikowanej kroswalidacji wszystkie wyniki bez względu na ustawienie wag lub ich braku i miarę odległości były bardzo bliskie sobie. Z wyjątkiem odległości euklidesowej SKF i KF dawały podobne optymalne K. Dla zwykłej kroswalidacji f-score bardzo odbiegał od miary accuracy. Najlepsze wyniki dla SKF uzyskano dla odległości Chebyshev'a. Nałożenie wag na głosy nie miało wpływu na wyniki. Jest to jedyny zbiór dla, którego ta odległość dawała najlepsze wyniki, a najgorsze uzyskano dla odległości euklidesowej.

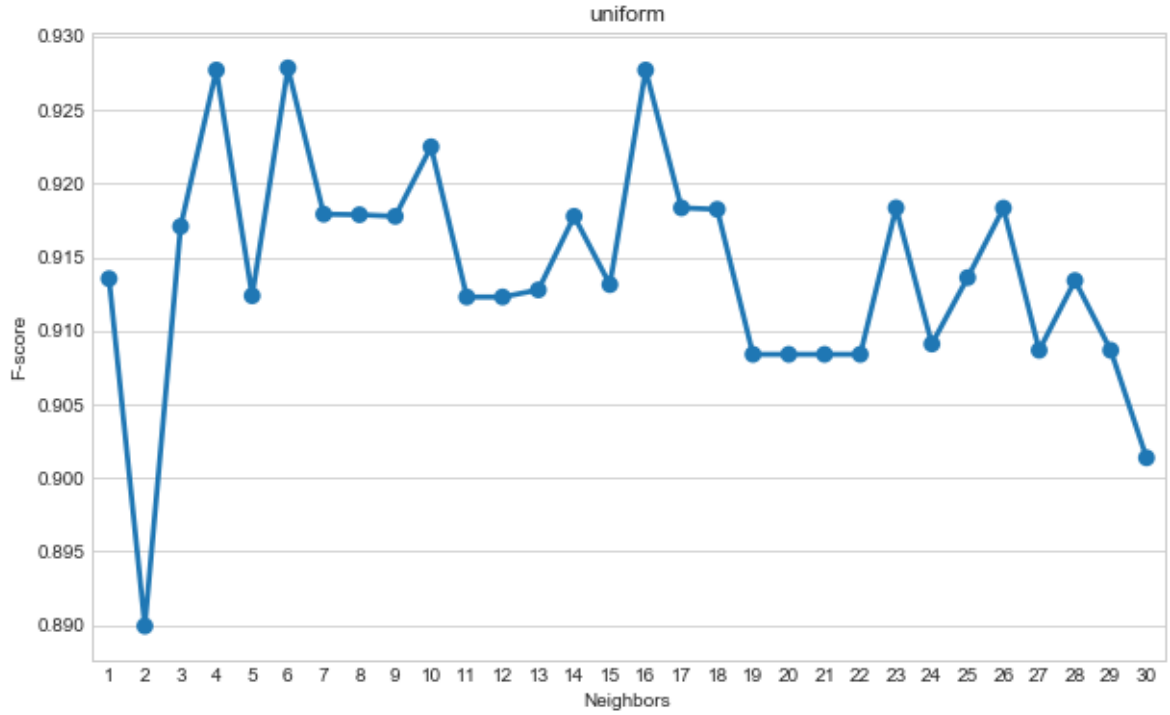
Analiza algorytmów miar odległości, przy pominięciu wag, kroswalidacji stratyfikowanej z folds=10.

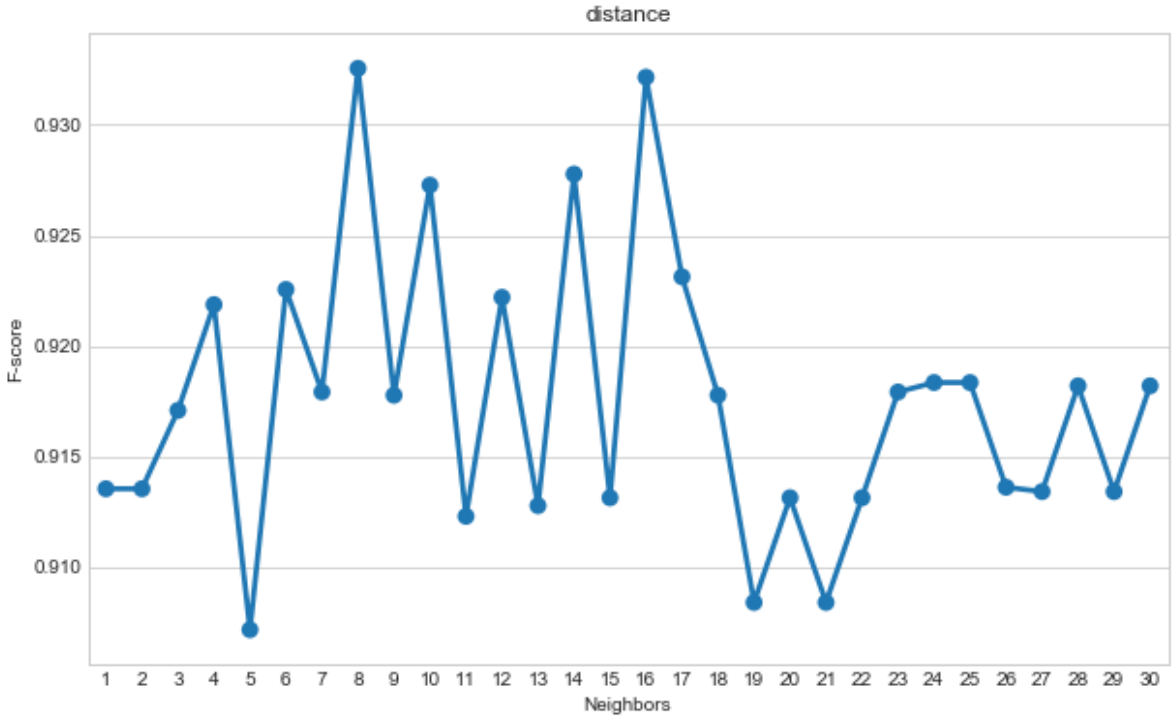

euclidean
F-score
Neighbors
0.930
0.925
0.920
0.915
0.910
0.905
0.900
0.895
0.890
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30


manhattan
F-score
Neighbors
0.925
0.920
0.915
0.910
0.905
0.900
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30



Analiza wyników klasyfikacji z nałożeniem lub brakiem wag na głosy, przy pomocy kroswalidacji stratyfikowanej z fold=10 i euklidesową miarą odległości.





5. Podsumowanie

W każdym zbiorze stratyfikacja kroswalidacji dawała przewagę ponad zwykłą kroswalidacją. W niektórych przypadkach była ona aż na poziomie 40%. Najczęsciej najlepsze wyniki uzyskiwano dla odległości euklidesowych lub odległości Manhattan. **Zdecydowanie ważniejsze okazało się, którą miarę odległości wybierzemy do obliczenia głosów, niż to czy nałożymy na te głosy wagi.** Tylko w przypadku zbioru Seeds najlepsze wyniki uzyskano dla odległości Chebyshev'a, w innych zbiorach dawała ona najczęściej najgorsze wyniki.

Nie było możliwym stwierdzenie jednej optymalnej miary odległości dla zbiorów, jako że każdy ze zbiorów dawał bardzo różne wyniki dla różnych miar. Dobrą miarą wyjściową była jednak odległość euklidesowa. Wyniki f-score i accuracy w przypadku kroswalidacji stratyfikowanej były do siebie najczęściej zbliżone, a w przypadku zwykłej kroswalidacji były najczęściej odległe.

KNN dawał podobne wyniki jak w przypadku poprzednio badanych