**WYŻSZA SZKOŁA INFORMATYKI**

**I ZARZĄDZANIA**

Z SIEDZIBĄ W RZESZOWIE

Sprawozdanie

Sztuczna inteligencja

**KNN**

Prowadzący: dr inż. Mariusz Wrzesień Wykonawca: Oleh Korsunskyi w58957

Rzeszów 2019

# Opis problemu

Celem laboratorium jest sprawdzenie poprawności klasyfikacji nieznanych obiektów, zamieszczonych w zbiorze testowym, przy pomocy algorytmu kNN z różnymi parametrami. Ten algorytm odnosi się do metod opartych no podobieństwie i polega na odnalezieniu podobnych obiektów dla konkretnego przypadku. Do tego celu wykorzystane zostaną trzy metryki, przy pomocy których będą obliczone odległości pomiędzy obiektami i zostaną sprawdzone błędy podczas prognozowania klas dla poszczególnych metryk metodą N-krotnej-skrośnej-walidacji. Wyniki zostaną poddane trzem metodą głosowania: metodzie normalnej, metodzie sumy odwrotności kwadratów odległości oraz sumy odległości. Również wszystkie wyliczenia będą robione dla jednego, trzech oraz pięciu sąsiadów.

# Badane zbiory danych

Dla naszego problemu wykorzystujemy zbiór danych Iris który składa się z 150 obiektów rozdzielonych na 3 klasy: iris-setosa, iris-verginica, iris-versicolor. Przy czym każda klasa jest reprezentowana taką samą liczbą przypadków, czyli 50. Każdy przypadek jest opisany wektorem, który składa się z 4 cech: petal\_width, petal\_length, sepal\_width, sepal\_length.

# Obliczenia

Skoro dla sprawdzania poprawności poszczególnych parametrów kNN wykorzystujemy walidacje krzyżową to zbiór pierwotny najpierw jest potasowany, żeby uniknąć nieprawidłowej oceny, a później zgodnie z algorytmem jest podzielony na N równych części. W naszym przypadku N=10. Potem na i-tym kroku i-ty zbiór będzie wykorzystywany jako zbiór testowy, a suma innych zbiorów jako zbiór uczący. Więc dostaniemy 10 ocen dla każdej wariacji parametrów. Każda ocena to procent błędnie klasyfikowanych obiektów ze zbioru testowego. Na koniec musimy policzyć średnią arytmetyczną tych wszystkich błędów, to i będzie nasza N-krotna-skrośnia-walidacja.

Program, w którym wykonane obliczenia dla poszczególnych parametrów AffinitySEEKER. Program w którym wykonane obliczenia walidacji krzyżowej i dokładności – Excel. Po zaimportowaniu zbioru uczącego oraz testującego przystąpiono do obliczania błędów, za pomocą trzech metryk, dla trzech metod głosowania, dla trzech parametrów k.

Pierwszą z tych metryk jest metryka City Block. Jej wzór przedstawia się następująco:

Następną użytą metryką jest odległość Euklidesa, czyli odległość geometryczna w przestrzeni wielowymiarowej, dzięki której można wyliczyć najkrótszą drogę:

Ostatnią użytą jest metryka Minkowskiego. Jest podobna do metryki Euklidesa tylko zamiast 2 potęgi mamy p potęgę.

Parametry:

n – liczba wektorów(cech) Xi – i-ta cecha obiektu nieznanego Zi – i-ta cecha obiektu znanego ze zbioru p – potęga w metryce Minkowskiego, dzięki niej metryka Minkowskiego jest bardziej ogólna niż metryka Euklidesa

Również stosujemy trzy metody głosowania: zwykłą, sumę odwrotności kwadratów odległości oraz sumę odległości.

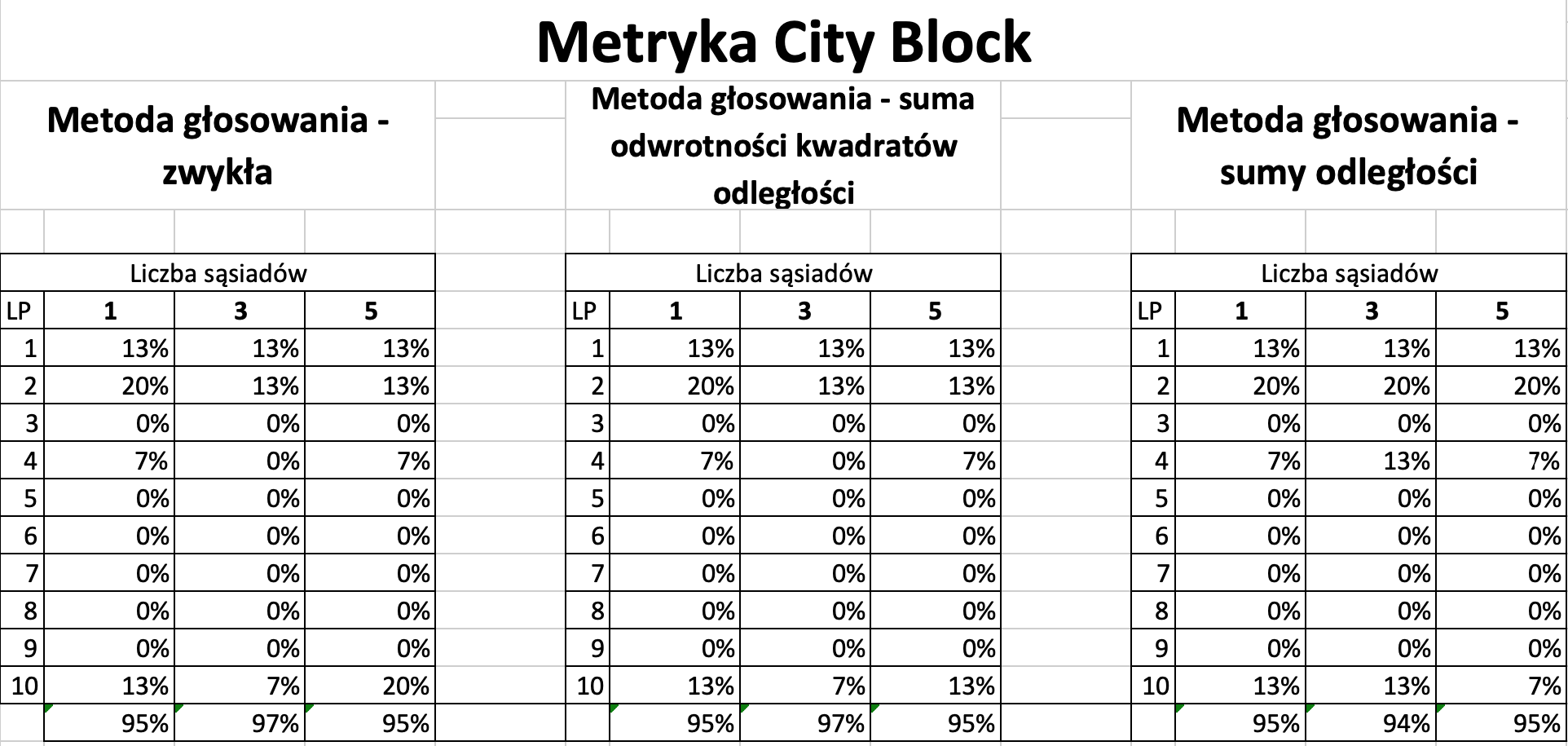
W zwykłym głosowaniu po prostu wybieramy klasę większości. Mamy uporządkowane obiekty od najbliższego do najdalszego do naszego obiektu klasyfikowanego. I mówimy, że nasz obiekt klasyfikowany jest tej klasy, obiektów której jest najwięcej wśród k najbliższych sąsiadów.

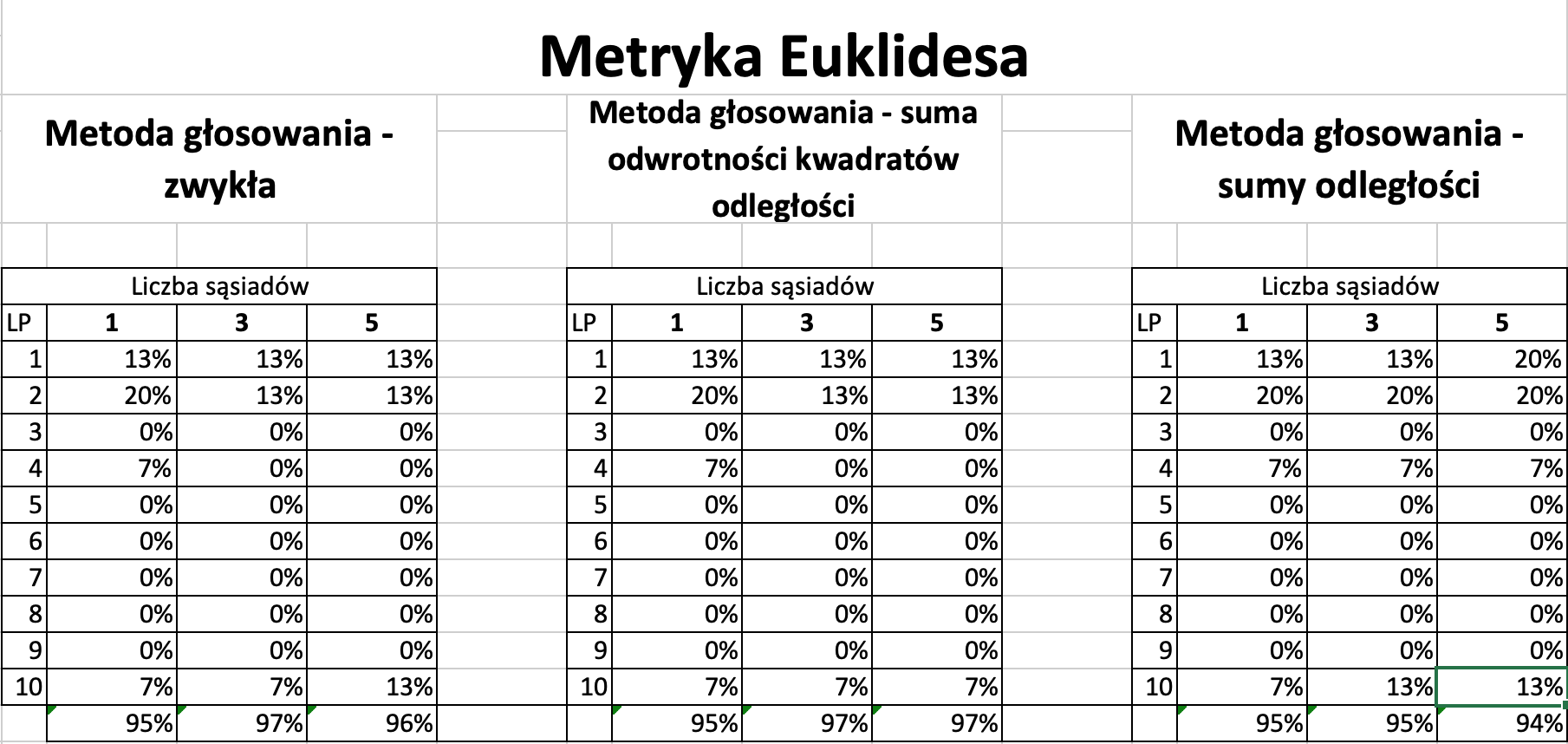
W metodzie odwrotności kwadratów odległości liczymy sumę wyników odwrotności kwadratów odległości wśród k obiektów dla poszczególnych klas. Której klasy wynik jest większy do tej klasy i przydzielamy nasz obiekt. Tę metodę przedstawia poniższy wzór, gdzie d to odległość między obiektem klasyfikowanym a obiektem danej klasy.

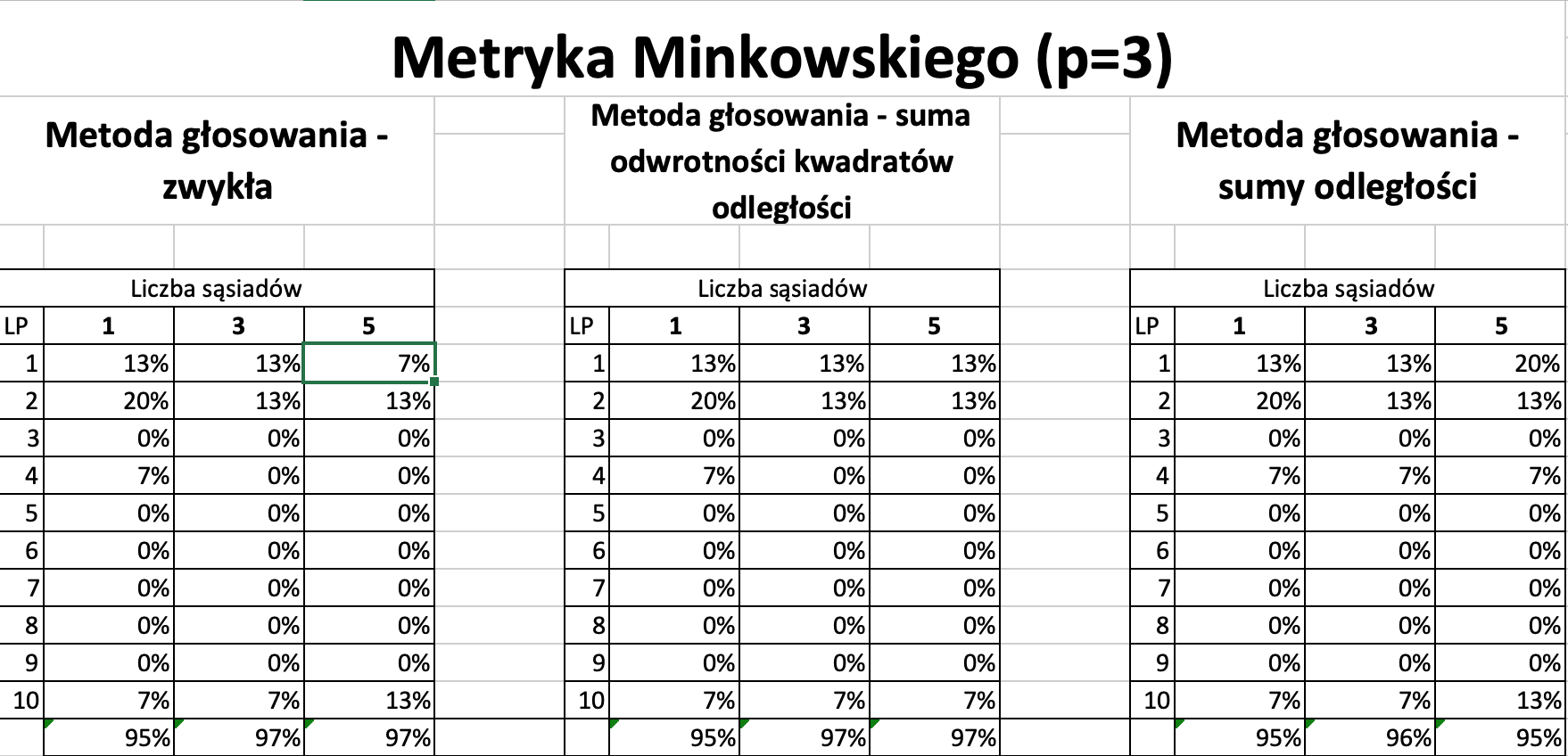
Metoda sumy odległości polega na tym, że liczymy sumę odległości wszystkich obiektów danej klasy wśród obiektów k, a później liczymy średnią arytmetyczną. W tym przypadku wynik której klasy jest mniejszy do tej klasy przydzielamy nasz obiekt.

Wszystkie obliczenia robimy dla 3 wartości parametru k: 1, 3, 5, który reprezentuje liczbę najbliższych sąsiadów

# Wyniki







# Analiza wyników

Za pomocą programu AffinitySEEKER zostało przeprowadzone badanie dokładności klasyfikacji obiektów na podstawie zbioru Iris metodą N-krotnej-skrośnej-walidacji. Do tego celu został wykorzystany algorytm kNN z wykorzystaniem takich metryk: CityBlock, Euklidesa, Minkowsiego(p=3). Również podczas badania były wykorzystywane 3 metody głosowania: zwykła, suma odwrotności kwadratów odległości, suma odległości. Podczas obliczeń była wybierana taka liczba **k** najbliższych sąsiadów: 1, 3, 5. W sumie uzyskałem 27 wyników dokładności. Na podstawie tych wyników wybiorę najlepsze połączenia parametrów algorytmu kNN do klasyfikowania kwiatów irisa, które muszą być stosowane do tego problemu. Również wybiorę najgorsze połączenie parametrów, które lepiej nie stosować do klasyfikowania kwiatów irisa. Najlepsze wyniki dokładności - 97%, uzyskane przy stosowaniu algorytmu kNN z takimi trójkami parametrów:

|  |  |  |
| --- | --- | --- |
| Metryka | Metoda Głosowania | Liczba sąsiadów |
| City Block | Zwykła | 3 |
| City Block | Suma odwrotności kwadratów odległości | 3 |
| Euklidesa | Zwykła | 3 |
| Euklidesa | Suma odwrotności kwadratów odległości | 3 |
| Euklidesa | Suma odwrotności kwadratów odległości | 5 |
| Minkowskiego | Zwykła | 3 |
| Minkowskiego | Zwykła | 5 |
| Minkowskiego | Suma odwrotności kwadratów odległości | 3 |
| Minkowskiego | Suma odwrotności kwadratów odległości | 5 |

Najgorsze wyniki dokładności -94%, uzyskane przy stosowaniu algorytmu kNN z takimi trójkami parametrów:

|  |  |  |
| --- | --- | --- |
| Metryka | Metoda głosowania | Liczba sąsiadów |
| City Block | Sumy odległości | 3 |
| Euklidesa | Sumy odległości | 5 |

# Wnioski badawcze

Na podstawie przeprowadzanego badania można zauważyć, że najwięcej na wynik klasyfikacji zbioru Iris wpływa metoda głosowania i liczba sąsiadów, a metryka w tym przypadku nie jest ważna. Metoda głosowania suma odwrotności kwadratów odległości oraz metoda zwykła razem z liczbą sąsiadów = 3 pokazują najlepsze wyniki niezależnie od metryki, dlatego te dwa połączenia są najlepsze dla klasyfikowania kwiatów irisa.

Natomiast najgorszym wyborem dla zbioru Iris będzie stosowanie metody głosowania suma odległości, w szczególności razem z k = 3 lub 5.

# Wnioski

Algorytm kNN jest jednym z najłatwiejszych algorytmów uczenia maszynowego w pierwotnej postaci. Natomiast ten algorytm bardzo dobrze radzi sobie ze zbiorami które są łatwe do klasyfikowania i które są opisane za pomocą liczbowych atrybutów. Ten algorytm posiada dużą liczbę wariacji z wykorzystaniem różnych metryk, różnych metod głosowania, i oczywiście wyborów liczby sąsiadów. A w przypadku niektórych metryk, są nawet parametry w samej metryce, które też można trenować, przykładem jest potęga w metryce Minkowskiego. Ale i to jeszcze nie koniec. kNN jest podstawą dla innych bardziej zaawansowanych algorytmów uczenia maszynowego, takich jak Parzen–Rosenblatt window, który dodaje wagi na podstawie odległości. Algorytm kNN i jego uogólnienia bez wątpliwości są przydatnym narzędziem w uczeniu maszynowym.