KNN

Maciej Urbaniak 200842

May 15, 2018

1 Wstęp

Nowy obiekt w algorytmie KNN oblicza odległości od innych grup za pomocą wybranej metryki (euklidesa,manhattan itd.). Następnie poprzez wybrany sposób głosowania zostaje zaklasyfikowany do konkretnej grupy.

Odległość euklidesowa w n-wymiarowej przestrzeni:

$$\sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

Odległość manhattan w n-wymiarowej przestrzeni:

$$\sum_{i=1}^{n} |p_i - q_i|$$

Distance Weighted Nearest Neighbors "wartość" głosu sąsiada zależy od odległości nowego obiektu im dalej według wybranej metryki tym jego głos ma mniejsze znaczenie.

Większościowe opcja która ma więcej głosów zwycięża.

Równoprawne każdy sąsiad ma taką samą wagę.

2 Parametry

Aby polepszyć działanie kNN stosuje się standaryzację lub normalizację danych. Zastosowanie ich powoduje że wszystkie wymiary dla których liczona jest odległość posiadają jednakową istotność. W przeciwnym wypadku mogło by dojść do sytuacji w której pojedynczy wymiar zdominował by inne wymiary.

Standaryzacja to doprowadzenie danych w których wartość średnia poszczególnej cechy ma wartość 0 a odchylenie standardowe = 1

Wartość distance oznacza jaka odległość została wybrana: 1 dla Manhattan oraz 2 dla Euklidesa.

2.1 Wine

Kroswalidacja k=2 skalowana najlepsze wyniki:
kmax = 13
distance = 1
Accuracy = 0.9550505
Mean F1 = 0.9559354

Kroswalidacja k=3 skalowana najlepsze wyniki:
kmax = 13
distance = 1
Accuracy = 0.9497175
Mean F1 = 0.9515044

Kroswalidacja k=5 skalowana najlepsze wyniki:
kmax = 13
distance = 1
Accuracy = 0.9720635
Mean F1 = 0.9736866

Kroswalidacja k=10 skalowana najlepsze wyniki:
kmax = 8
distance = 1
Accuracy = 0.9718954
Mean F1 = 0.9728361

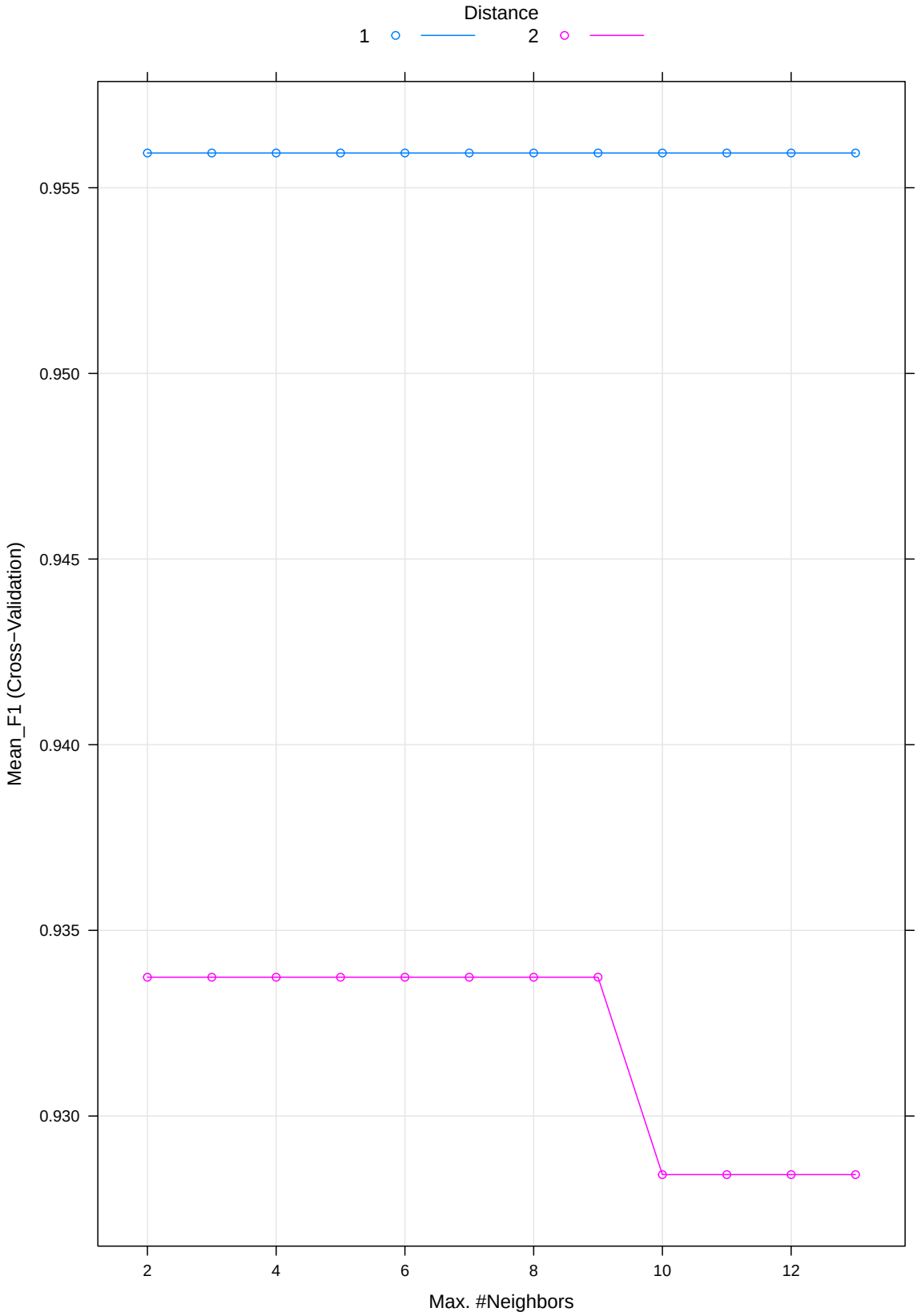

Figure 1: K=2 Wine skalowana

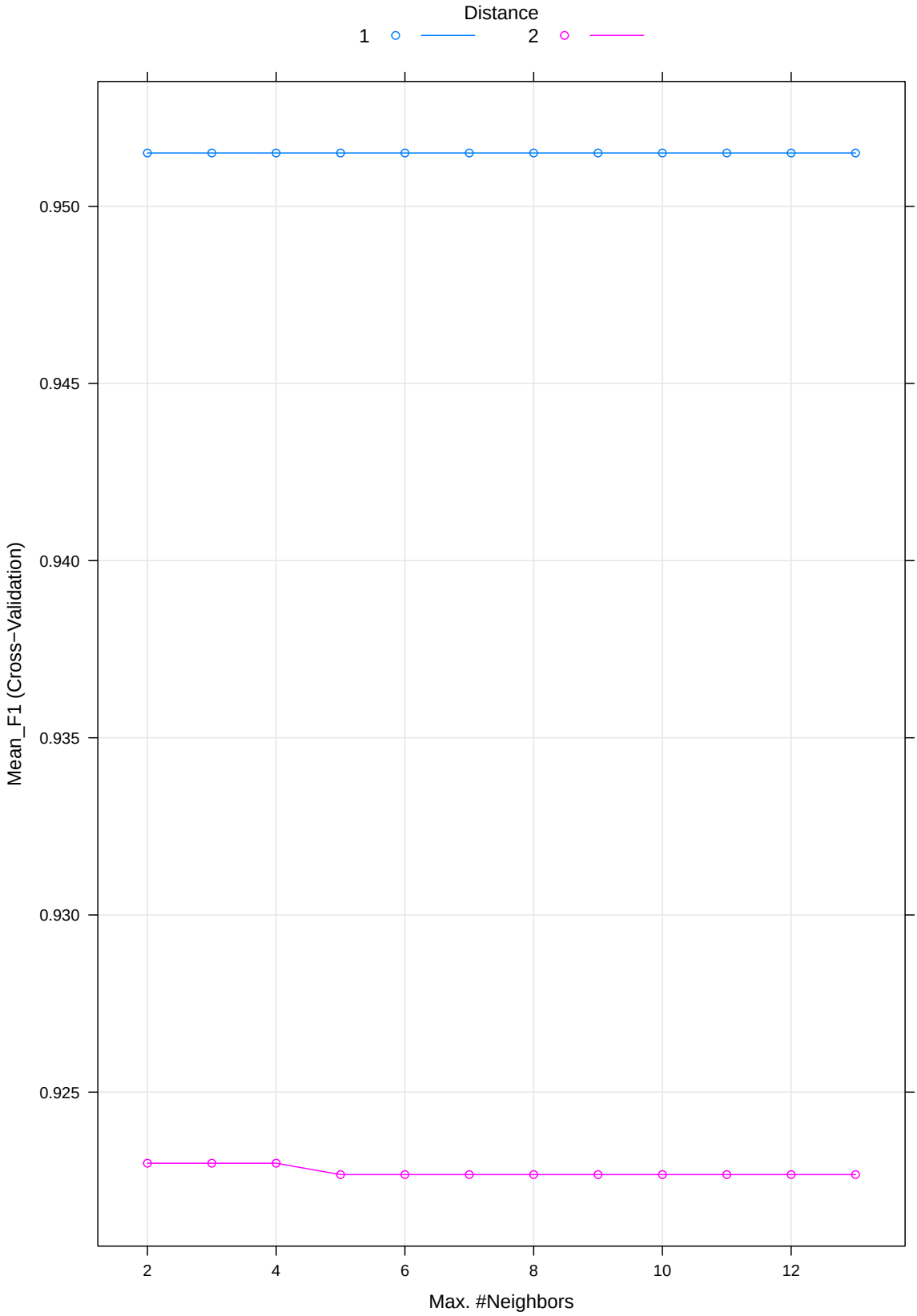

Figure 2: K=3 Wine skalowana

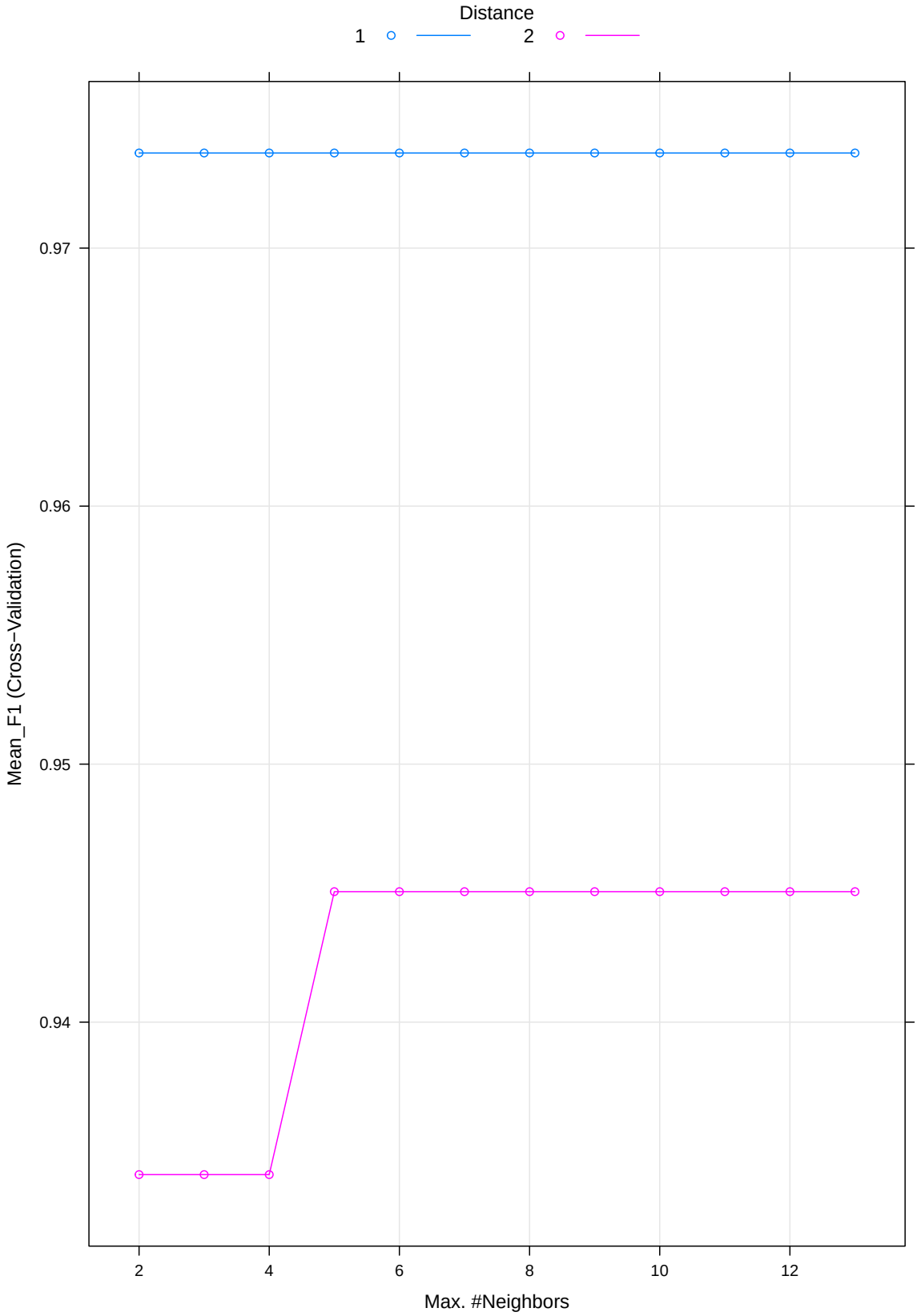

Figure 3: K=5 Wine skalowana

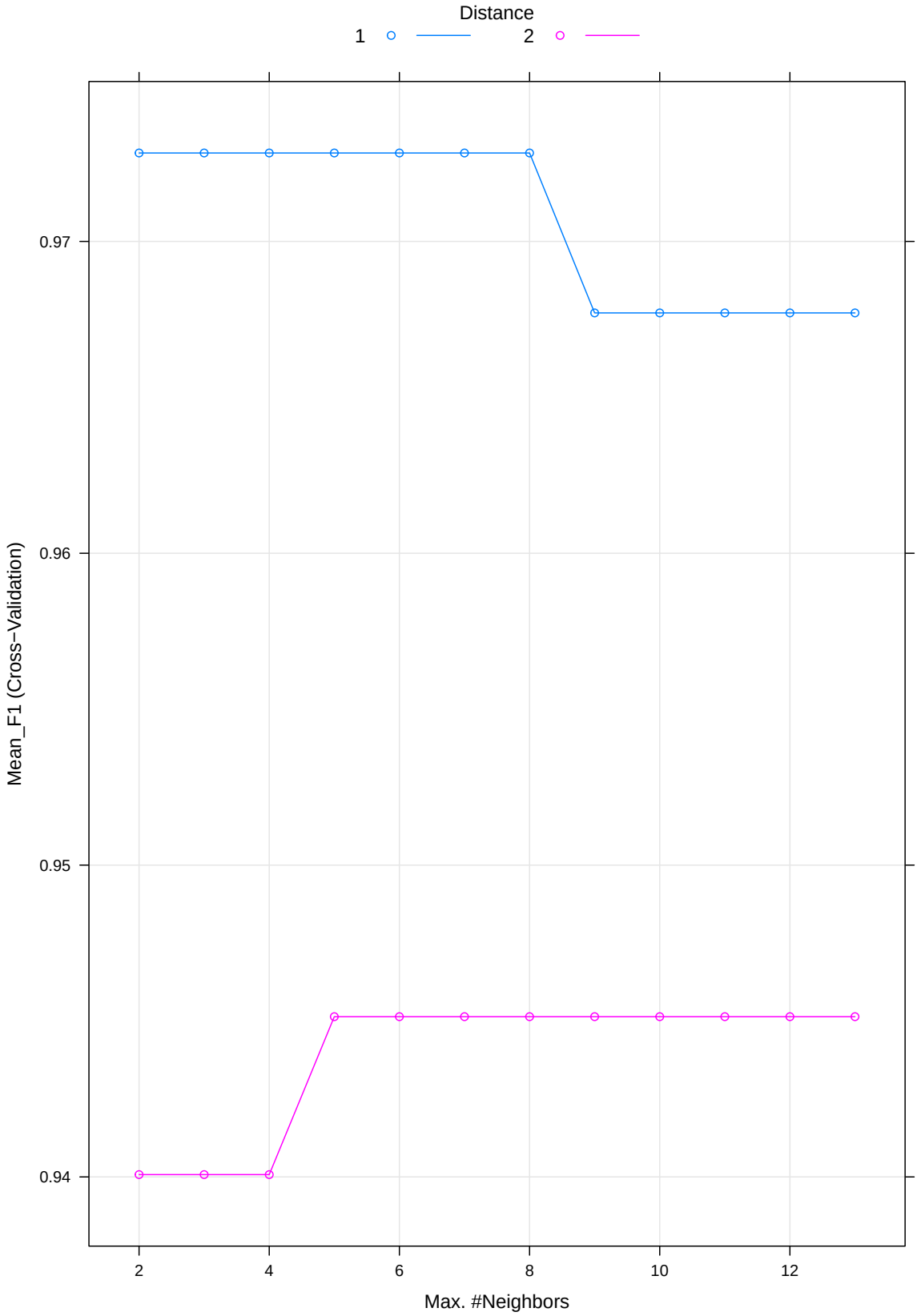

Figure 4: K=10 Wine skalowana

2.2 Glass

Kroswalidacja k=2 skalowana najlepsze wyniki:
kmax = 4
distance = 1
Accuracy = 0.7289720
Mean F1 = 0.6862793

Kroswalidacja k=3 skalowana najlepsze wyniki:
kmax = 4
distance = 1
Accuracy = 0.7756260
Mean F1 = 0.7479251

Kroswalidacja k=5 skalowana najlepsze wyniki:
kmax = 13
distance = 1
Accuracy = 0.7523810
Mean F1 = 0.7419061

Kroswalidacja k=10 skalowana najlepsze wyniki:
kmax = 13
distance = 1
Accuracy = 0.7480519
Mean F1 = 0.8120151

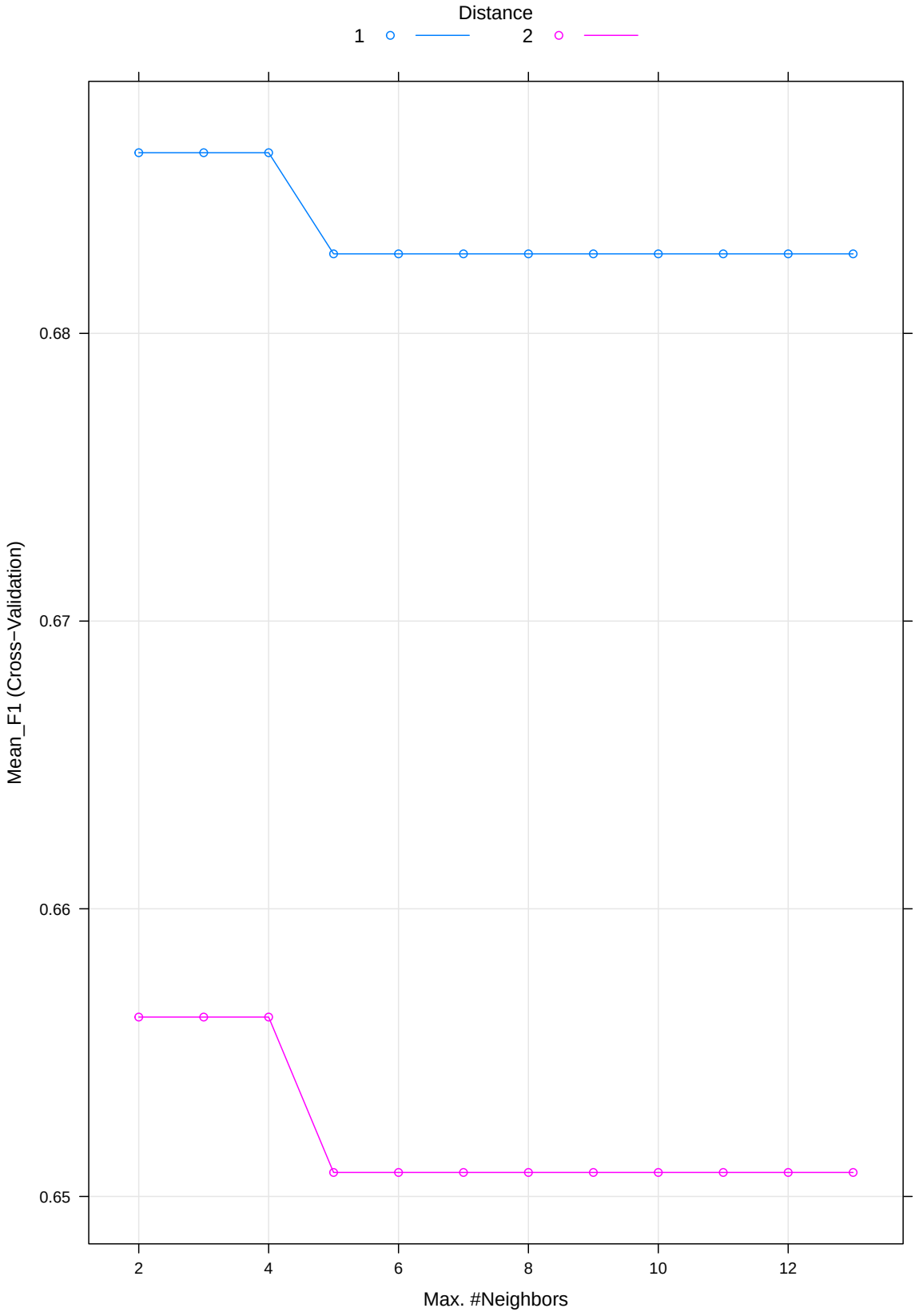

Figure 5: K=2 Glass skalowana

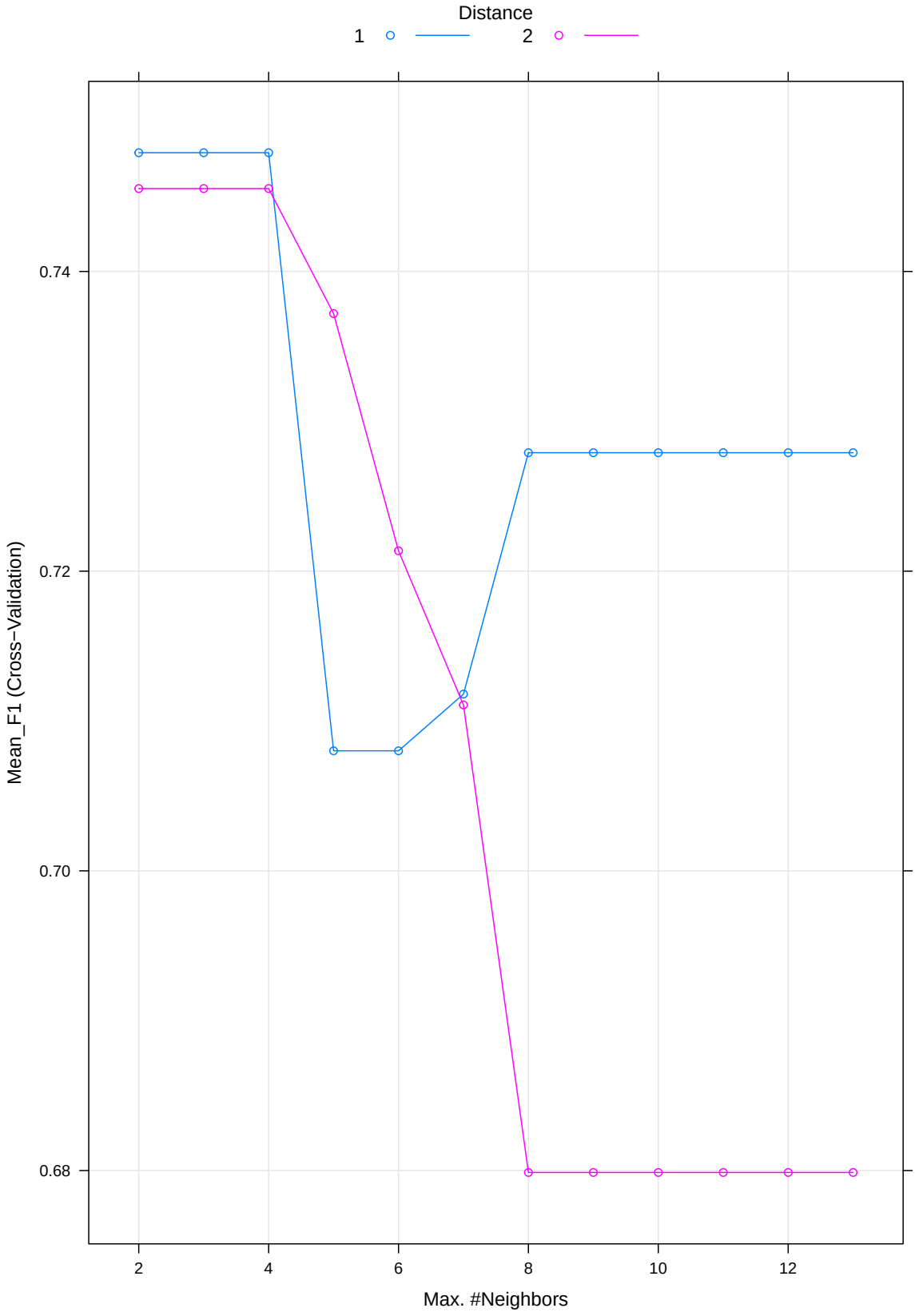

Figure 6: K=3 Glass skalowana

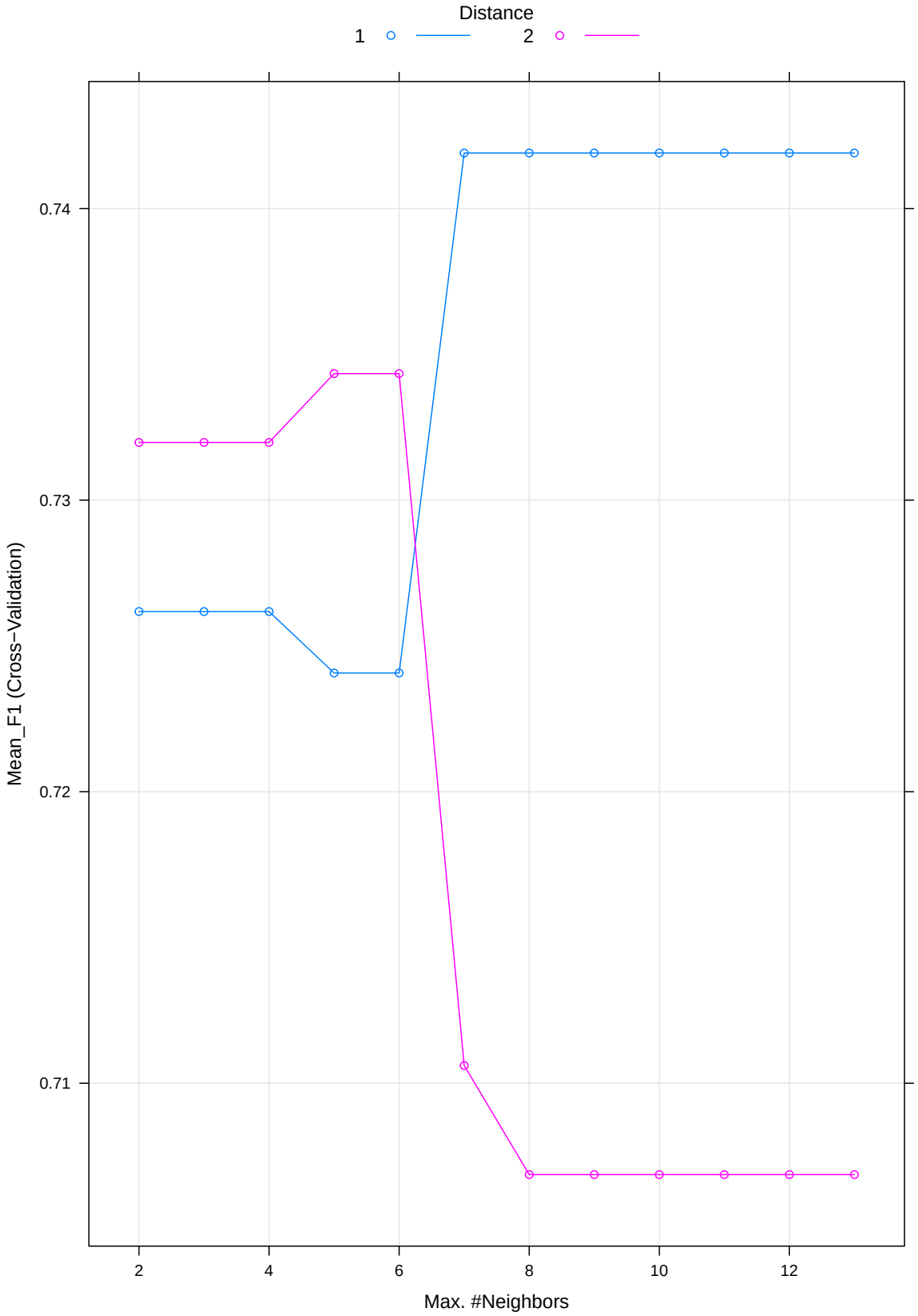

Figure 7: K=5 Glass skalowana

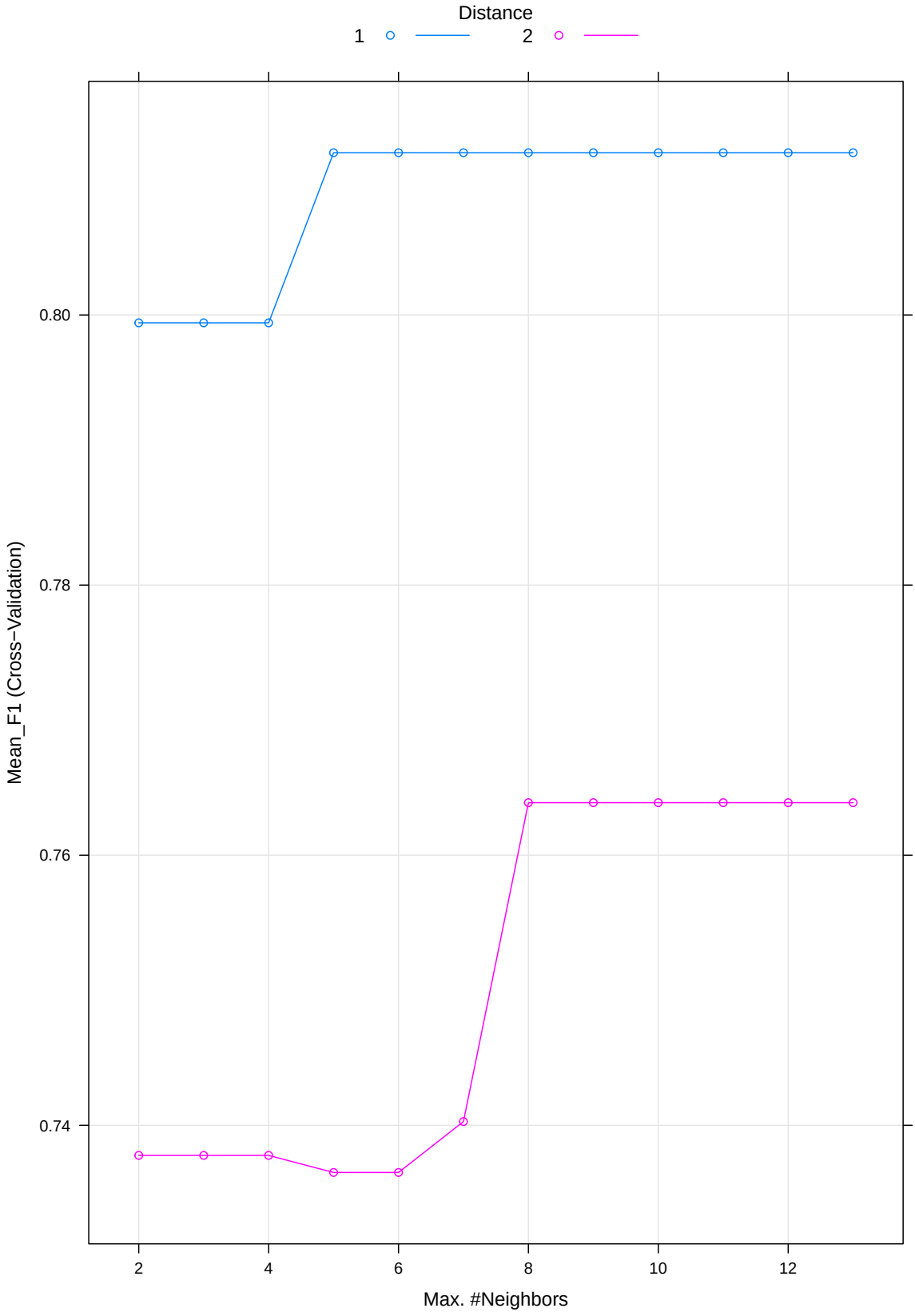

Figure 8: K=10 Glass skalowana

2.3 Diabetes

Kroswalidacja k=2 skalowana najlepsze wyniki:
kmax = 13
distance = 2
Accuracy = 0.7447917
Mean F1 = 0.8123013

Kroswalidacja k=3 skalowana najlepsze wyniki:
kmax = 13
distance = 2
Accuracy = 0.7304688
Mean F1 = 0.8022949

Kroswalidacja k=5 skalowana najlepsze wyniki:
kmax = 12
distance = 2
Accuracy = 0.7304049
Mean F1 = 0.7985658

Kroswalidacja k=10 skalowana najlepsze wyniki:
kmax = 9
distance = 2
Accuracy = 0.7487526
Mean F1 = 0.8139676

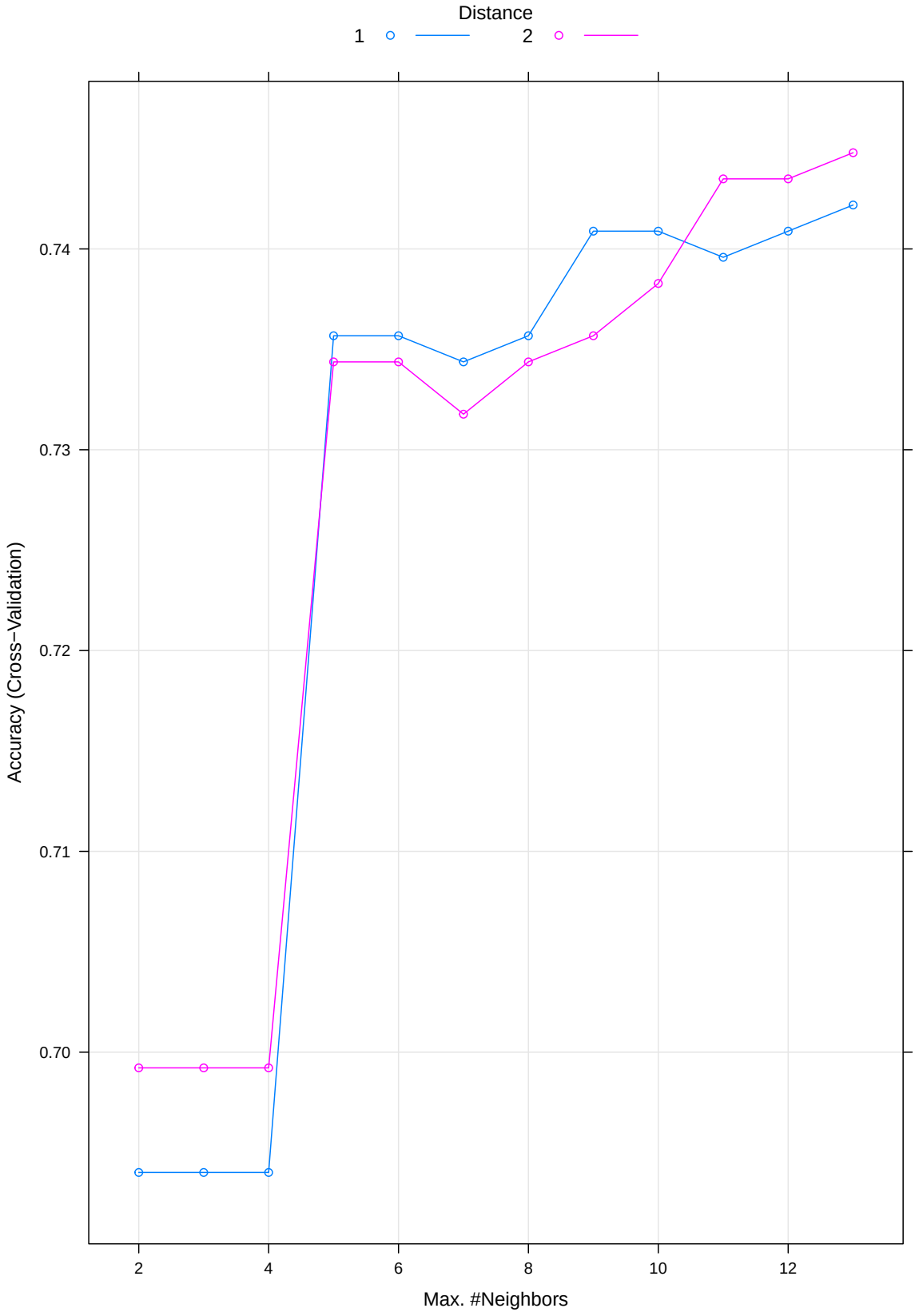

Figure 9: K=2 Diabetes skalowana

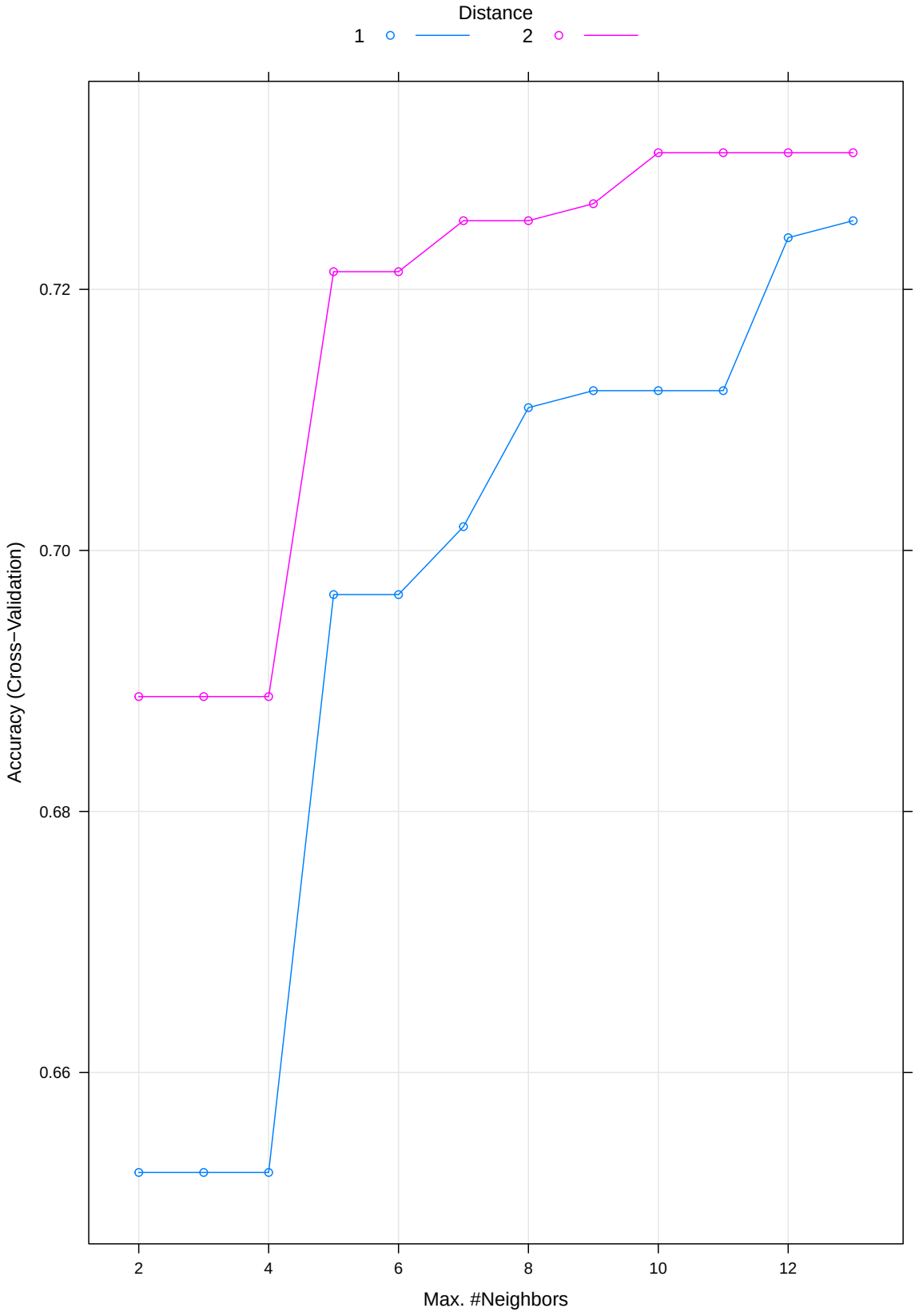

Figure 10: K=3 Diabetes skalowana

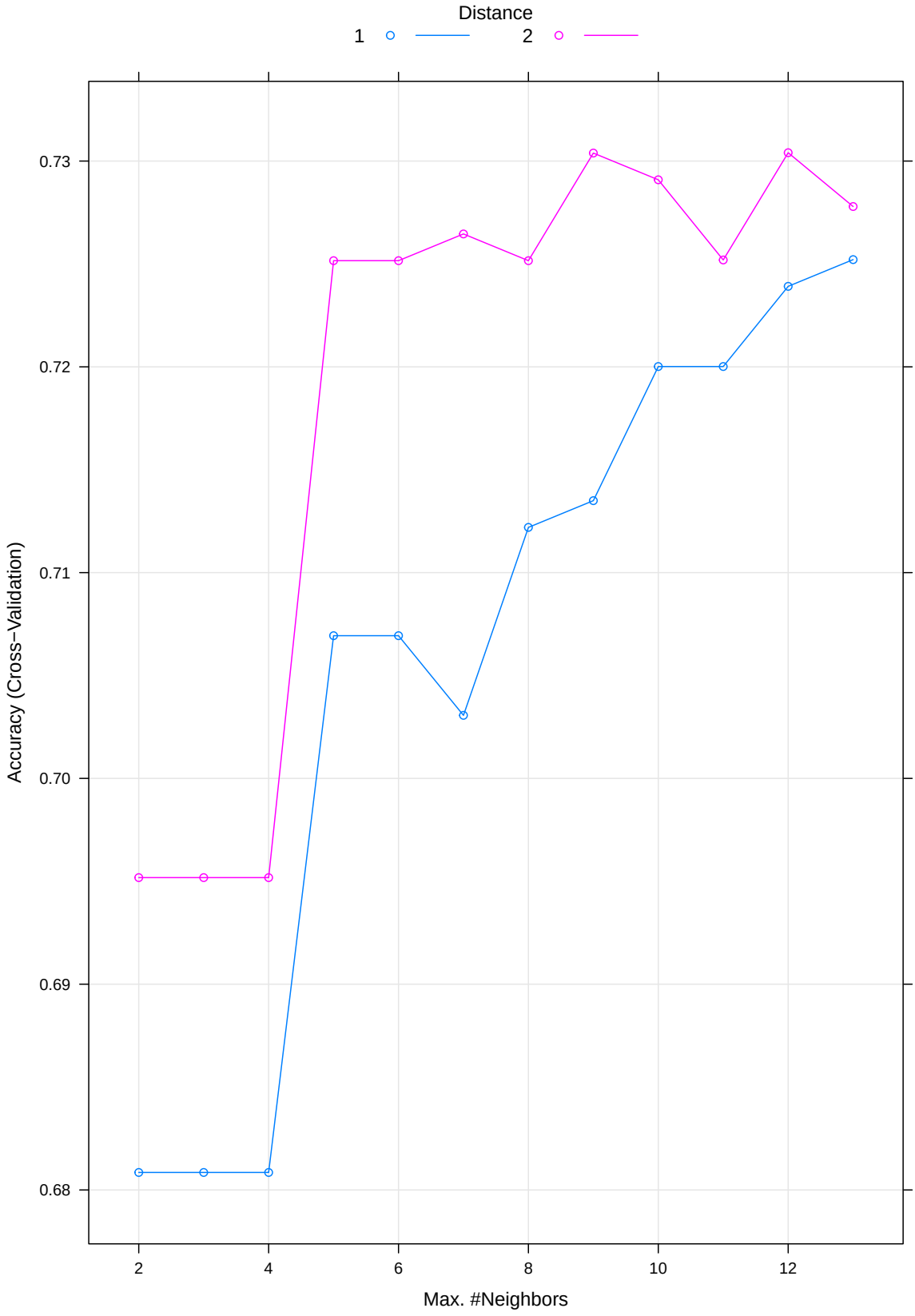

Figure 11: K=5 Diabetes skalowana

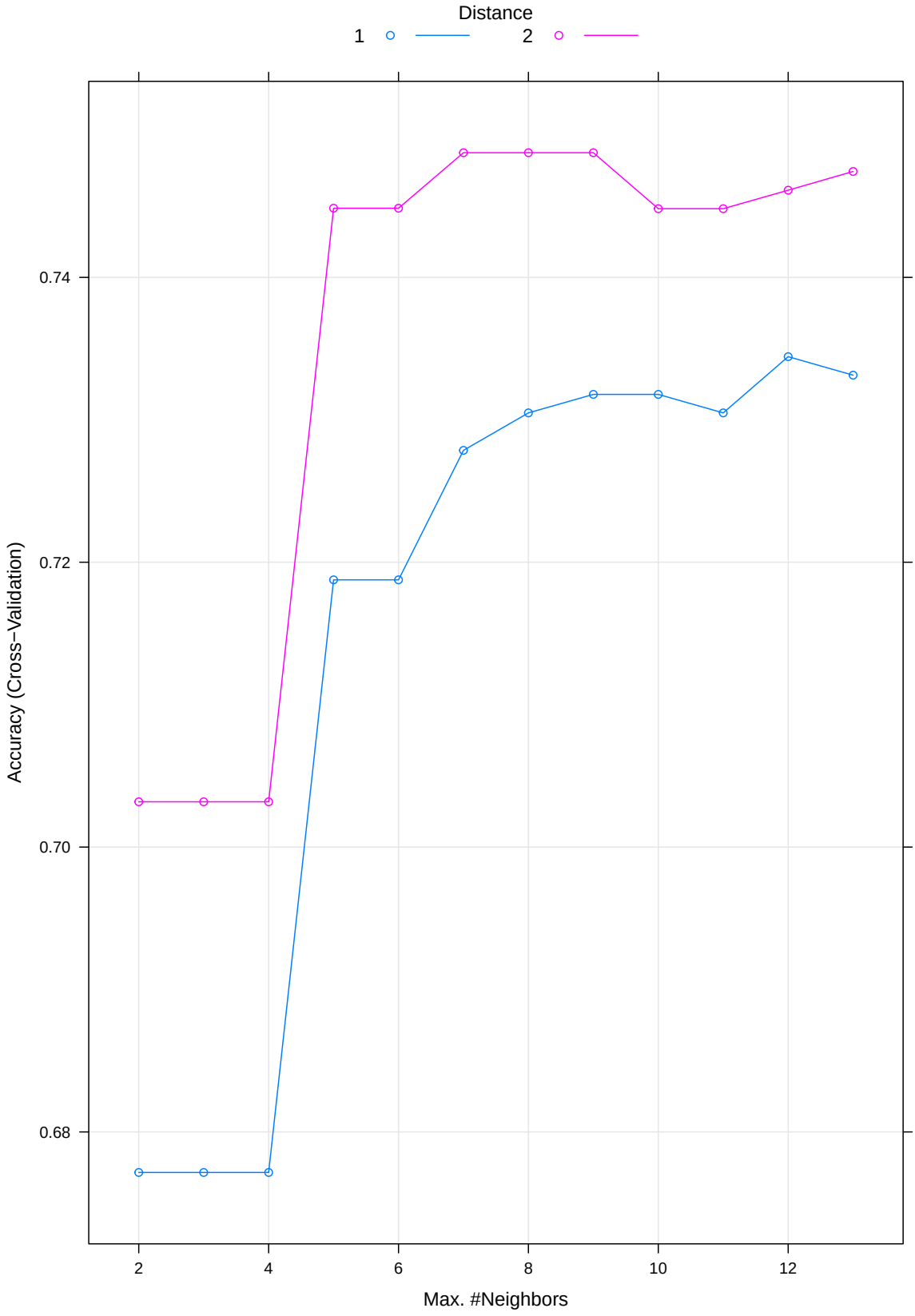

Figure 12: K=10 Diabetes skalowana

2.4 Seeds

Table 1: Seeds Kroswalidacja stratyfikowana dla k=2

kmax	distance	Accuracy	Mean F1
2	1	0.9380952	0.9375234
2	1	0.9380952	0.9375234
2	2	0.9333333	0.9327281
3	1	0.9380952	0.9375234
3	2	0.9333333	0.9327281
4	1	0.9380952	0.9375234
4	2	0.9333333	0.9327281
5	1	0.9333333	0.9327859
5	2	0.9333333	0.9327281
6	1	0.9333333	0.9327859
6	2	0.9333333	0.9327281
7	1	0.9333333	0.9327859
7	2	0.9333333	0.9327281
8	1	0.9333333	0.9330992
8	2	0.9333333	0.9327281
9	1	0.9333333	0.9330992
9	2	0.9333333	0.9327281
10	1	0.9333333	0.9330992
10	2	0.9285714	0.9280020
11	1	0.9333333	0.9330992
11	2	0.9285714	0.9280020
12	1	0.9333333	0.9330992
12	2	0.9285714	0.9280020
13	1	0.9333333	0.9330992
13	2	0.9285714	0.9280020

Kroswalidacja k=2 skalowana najlepsze wyniki:
kmax = 6
distance = 1
Accuracy = 0.9523810
Mean F1 = 0.9516511

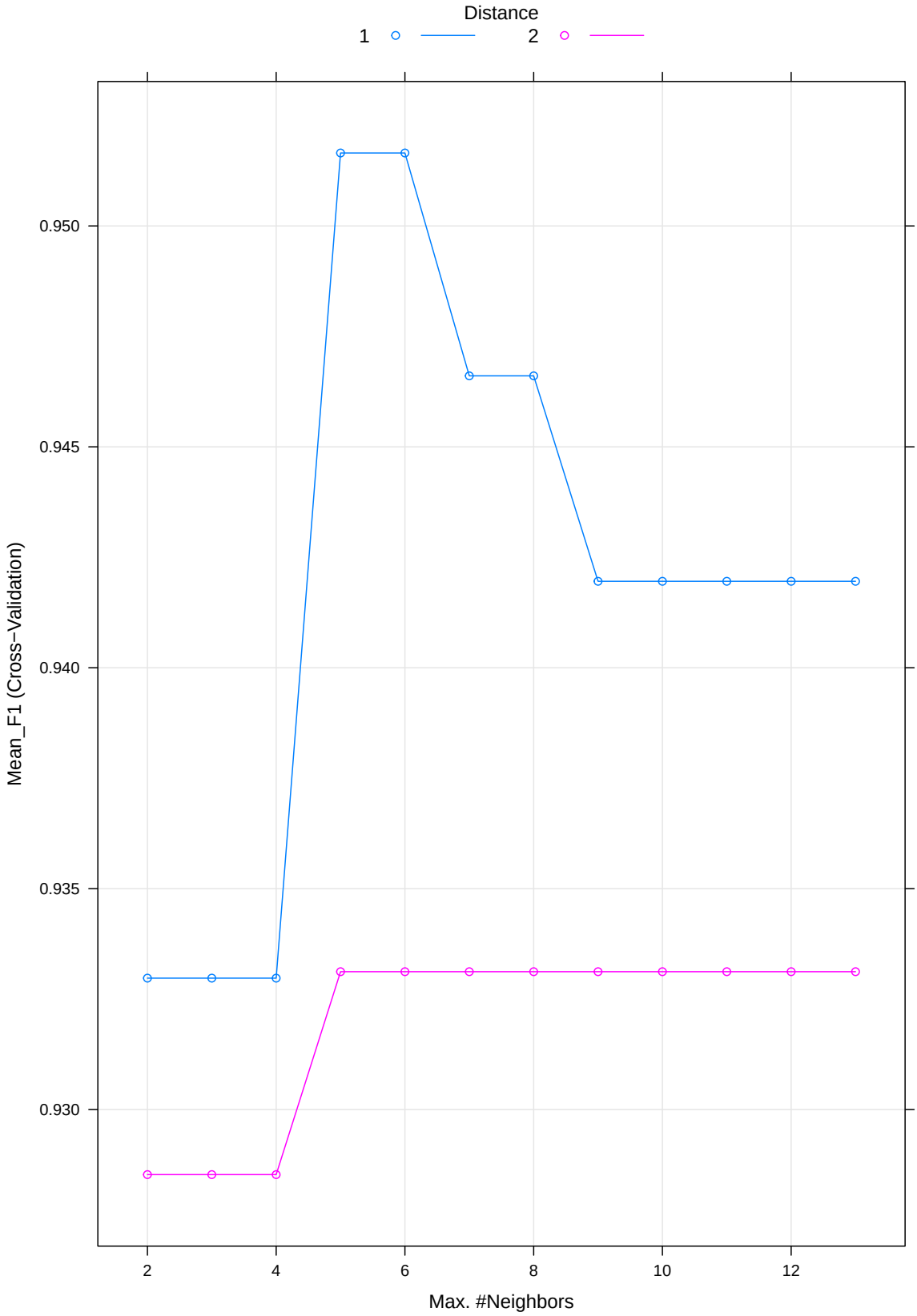

Figure 13: K=2 Seeds skalowana

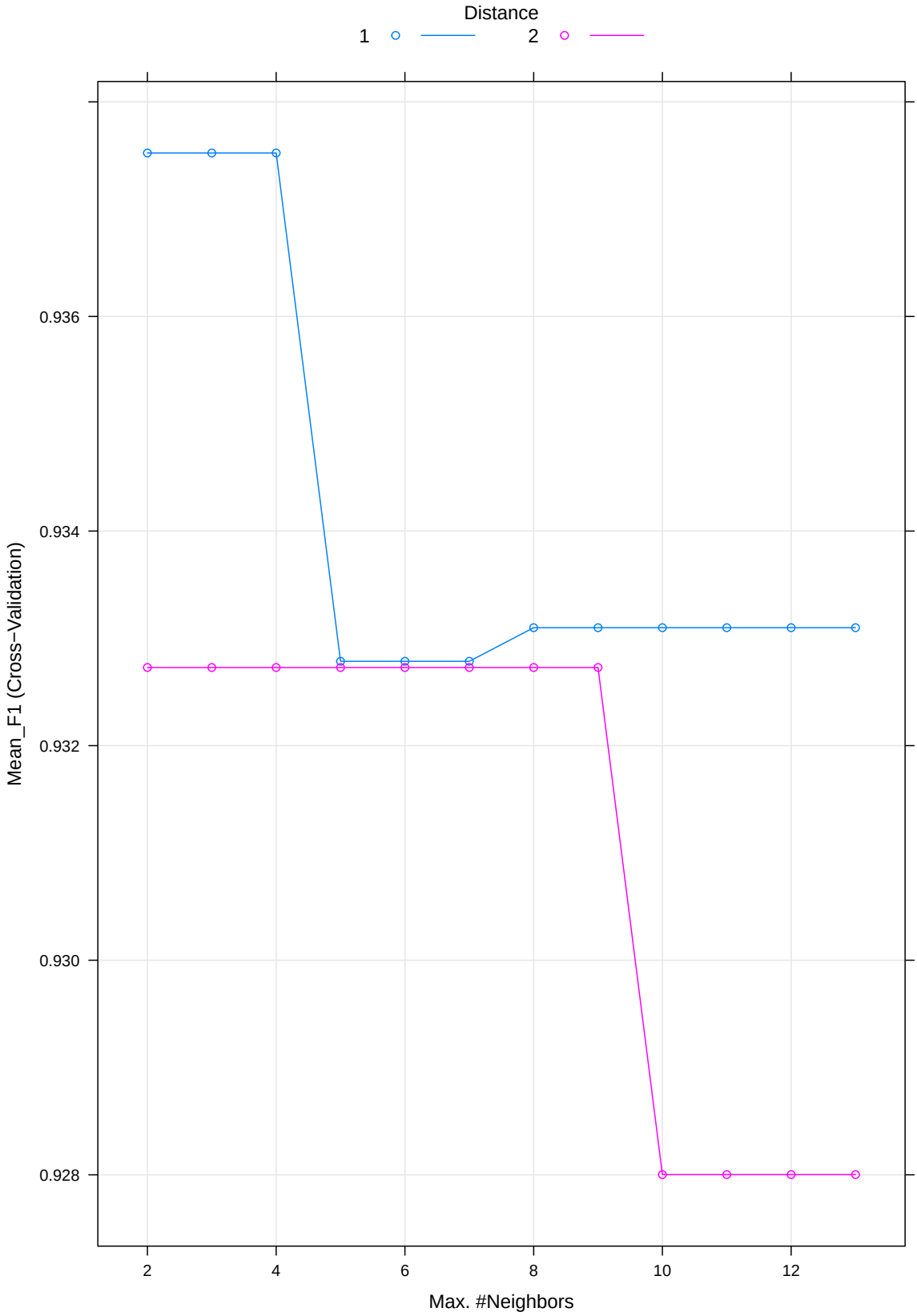

Figure 14: K=2 Seeds

3 Wnioski

Porównanie najlepszych:
Wine
Kroswalidacja k=5 skalowana najlepsze wyniki:
kmax = 13
distance = 1
Accuracy = 0.9720635
Mean F1 = 0.9736866
Najlepszy Bayes: 0.9616
Najlepszy C4.5: 0.955
Glass
Kroswalidacja k=10 skalowana najlepsze wyniki:
kmax = 13
distance = 1
Accuracy = 0.7480519
Mean F1 = 0.8120151
Najlepszy Bayes: 0.4417
Najlepszy C4.5: 0.650
Diabetes
Kroswalidacja k=10 skalowana najlepsze wyniki:
kmax = 9
distance = 2
Accuracy = 0.7487526
Mean F1 = 0.8139676
Najlepszy Bayes: 0.7551
Najlepszy C4.5: 0.711

Algorytm prosty w implementacji i szybko się uczy ale czas predykcji jest istotnie dłuższy wraz ze wzrostem danych oraz ich wymiarów. Liczba K w KNN wpływa na wrażliwość na wartości odstające im większa tym bardziej odporna gdyż obiekt potrzebuje wymaga większej ilości głosów sąsiadów aby dołączyć do wybranej grupy.