UNIVERSIDADE FEDERAL DO PARANÁ

SETOR DE TECNOLOGIA

CURSO DE INFORMÁTICA BIOMÉDICA

ANNA CAROLINE BOZZI

IMPACTOS DA REPRESENTAÇÃO

CURITIBA

2020

Introdução

Dado os arquivos digits.py, knn.py, juntamente com os dados digits, que compostos trata-se de um sistema de classificação de dígitos, são 2000 imagens de números inteiros, das quais é utilizado 50% para teste e 50% validação pelo método KNeighborsClassifier. Foi realizada uma análise dado diferentes tamanhos de normalização para as imagens, presente na função rawpixel do arquivo digits.py, e também para os diferentes tamanhos houve variação de K, números de vizinhos mais próximos, para cada comparação, presente na função main do arquivo knn.py.

Métodos

Foram variados os taanhos de normalização em:

- 20 x 10
- 40 x 10
- 80 x 40
- 100 x 50
- 140 x 70
- 200 x 100

Ou seja, foram realizados testes para vetores de 200 características até 20.000.

Para cada variação de normalização houve variação no número de vizinhos mais próximos para comparação, k, de 3 à 30.

As métricas de distâncias utilizadas para a classificação foram:

- euclidean
- manhattan
- minkowski

Nesses casos a medida de comparação utilizada será a acurácia, já que trata-se de um caso problema balanceado.

Resultados e Análise

1. Métrica Euclidiana

Dentre os valores de normalização para as imagens, a acurácia máxima observada para essa métrica foi com normalização de 140x70 e K = 3, conforme observa-se no gráfico da Fig.1. Ao manter a normalização e variar K é possível observar que há um decréscimo na acurácia até K = 30 onde há a acurácia mínima observada. A matriz de Confusão na Fig.2 mostra as frequências de classificação de cada classe desse modelo para a máxima acurácia observada. E na Fig.3 a matriz de confusão para a mínima acurácia observada.

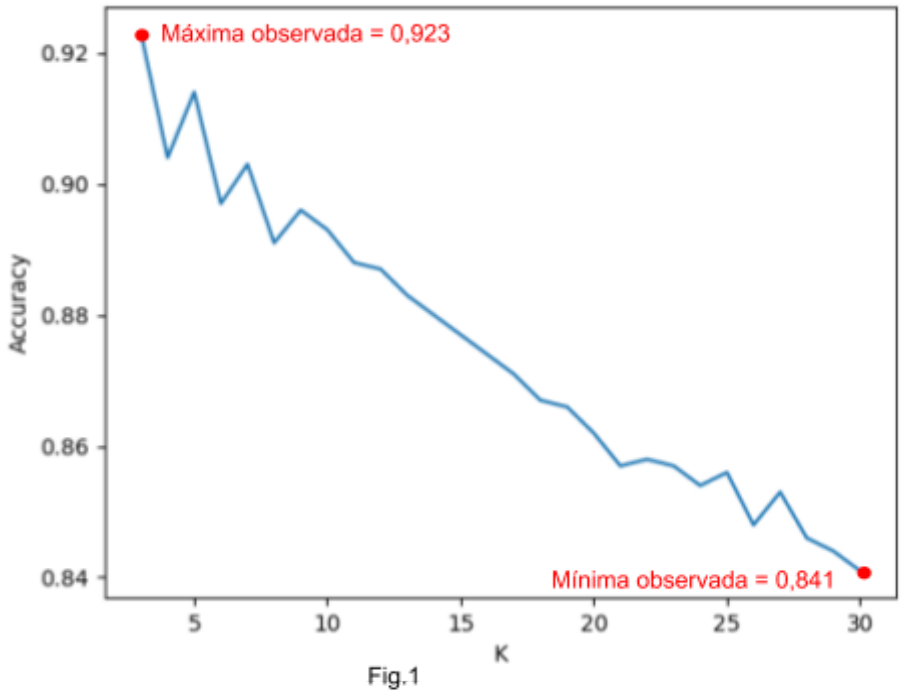

Fig.1
Relação entre K e acurácia para a métrica Euclidiana com normalização de 140x70.

```
    0   1   2   3   4   5   6   7   8   9
0 [ 96   0   0   0   0   1   0   0   0   0]
1 [  0  95   0   0   0   0   0   0   0   0]
2 [  0   4 102   0   1   0   1   2   1   0]
3 [  0   1   0  98   0   1   0   1   2   0]
4 [  0  10   2   0  82   0   0   0   0   1]
5 [  1   0   0   4   0  91   1   0   0   0]
6 [  2   5   0   0   0   0  99   0   0   0]
7 [  1   8   0   0   0   0   0  85   0   3]
8 [  0   3   0   3   0   1   0   1  78   1]
9 [  0   1   0   0   5   0   0   8   1  97]
```

Fig.2
Matriz de confusão da métrica Euclidiana com k=3 e normalização 140x70.

```
    0   1   2   3   4   5   6   7   8   9
0 [94   2   0   0   0   1   0   0   0   0]
1 [ 0  93   0   1   0   0   1   0   0   0]
2 [ 2  11  85   2   1   1   3   5   1   0]
3 [ 0   1   1  94   0   2   0   3   2   0]
4 [ 0  14   1   0  74   0   1   1   0   4]
5 [ 1   3   0   8   0  84   1   0   0   0]
6 [ 0  10   0   0   2   0  94   0   0   0]
7 [ 0  14   0   0   1   0   0  79   0   3]
8 [ 1   9   0   4   0   4   0   5  63   1]
9 [ 0   7   0   0   5   0   0  19   0  81]
```

Fig.3
Matriz de confusão para a métrica Euclidiana com k=30 e normalização 140x70.

Observando as matrizes e o gráfico é possível verificar que mesmo no pior caso temos uma acurácia de 84%. Verificando assim que o aumento a variação de K apresentou um padrão para todas as variações de Normalização, como pode-se observar seguir nas Fig.4, Fig.5, Fig.6 ,Fig.7 ,Fig.8 os gráficos correspondentes aos outros valores de normalização para a métrica Euclidiana.

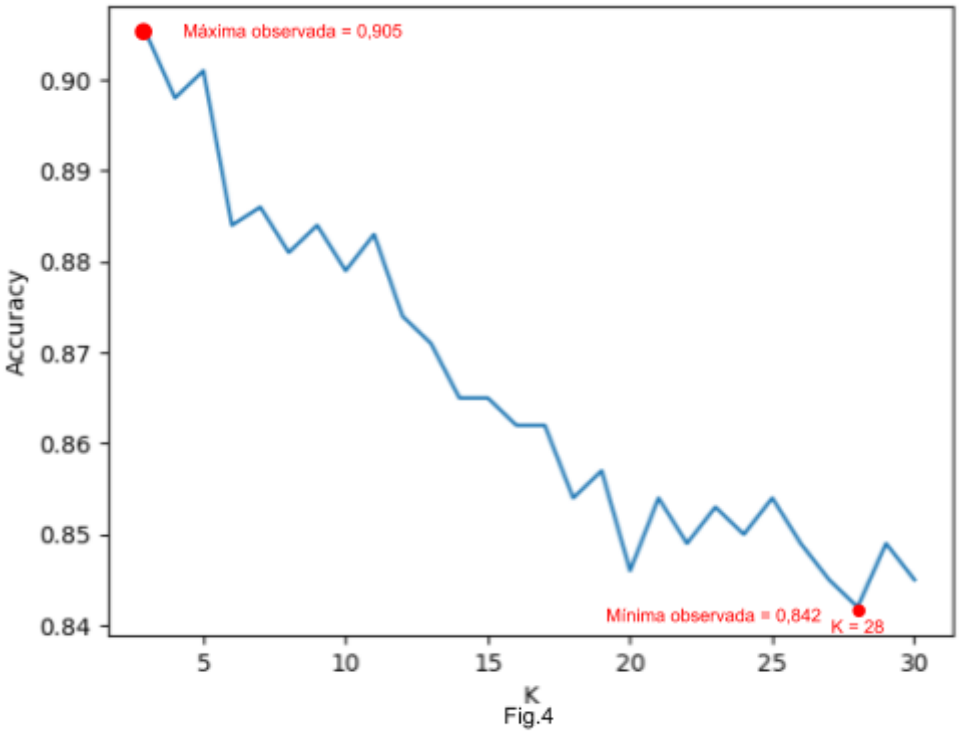

Fig.4
Relação entre K e acurácia para a métrica Euclidiana com normalização de 20x10.

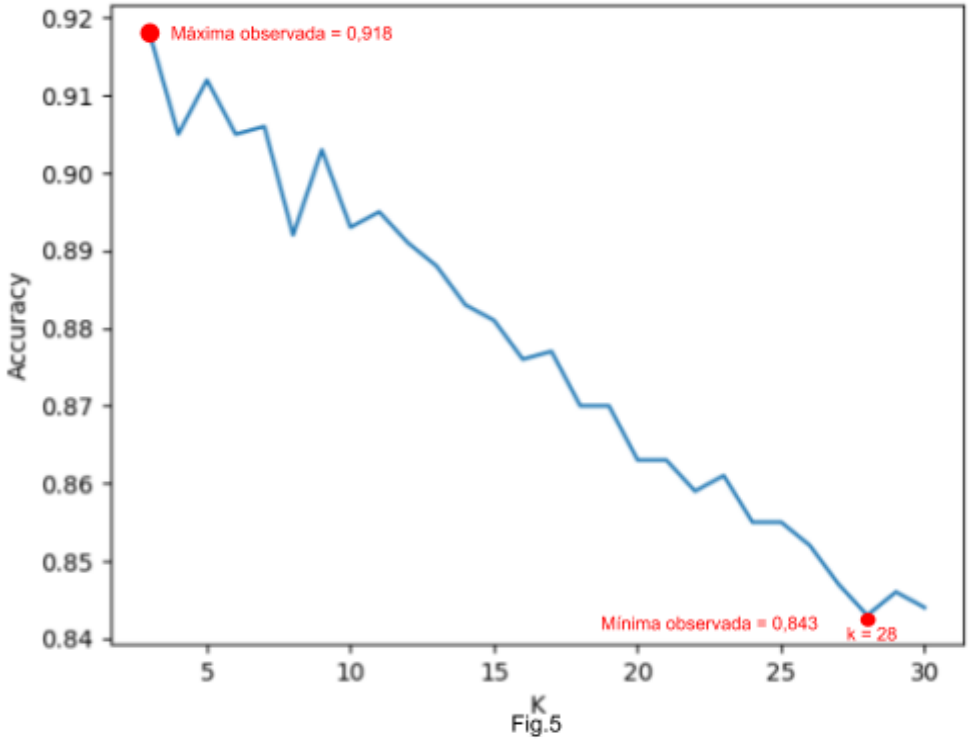

Fig.5
Relação entre K e acurácia para a métrica Euclidiana com normalização de 40x20.

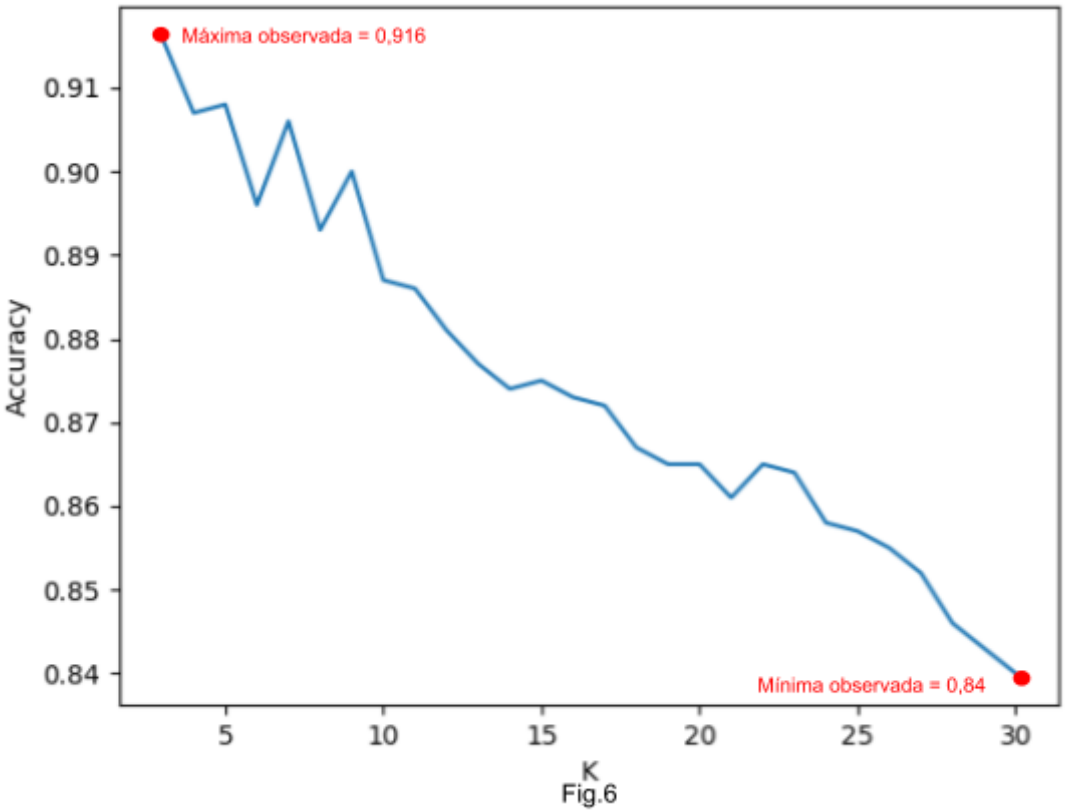

Fig.6
Relação entre K e acurácia para a métrica Euclidiana com normalização de 80x40.

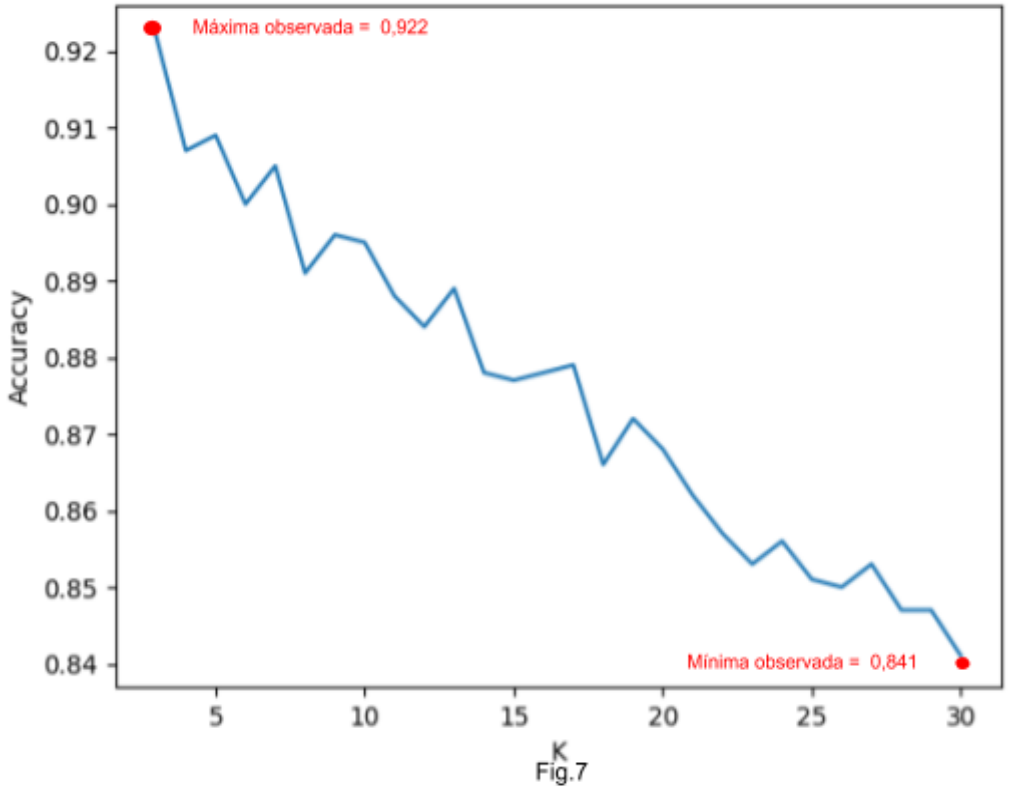

Fig.7
Relação entre K e acurácia para a métrica Euclidiana com normalização de 100x50.

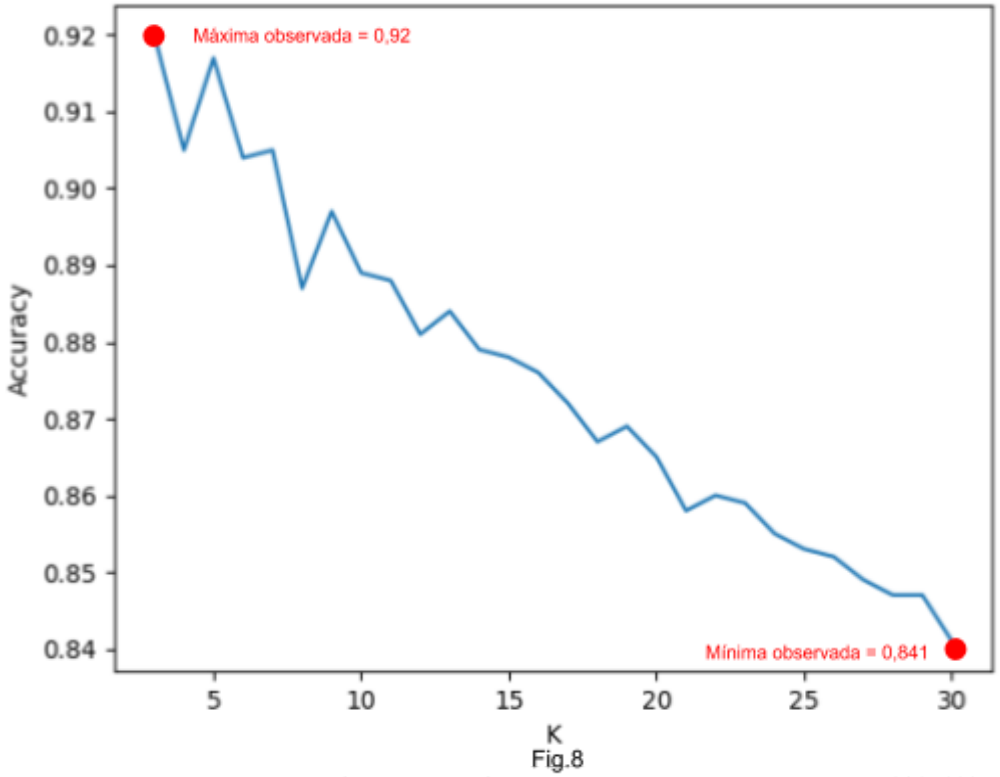

Fig.8

Relação entre K e acurácia para a métrica Euclidiana com normalização de 200x100.

2. Métrica Manhattan

Para a métrica Manhattan a normalização que maximizou a acurácia foi 140x70 para K = 3, assim como aconteceu na Euclidiana, conforme observa-se no gráfico da Fig.9. Ao manter a normalização e variar K é possível observar que há um decréscimo na acurácia até K = 30 onde há a acurácia mínima observada. A matriz de Confusão na Fig.10 mostra as frequências de classificação de cada classe desse modelo para a máxima acurácia observada. E na Fig.11 a matriz de confusão para a mínima acurácia observada.

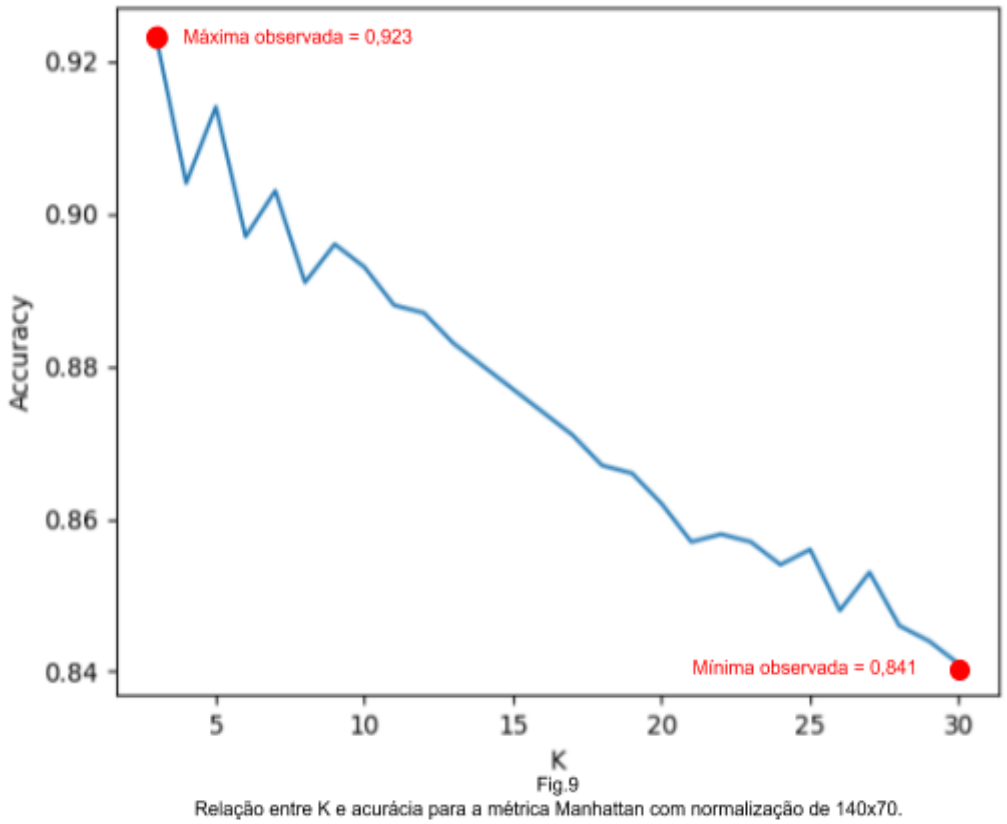

Fig.9
Relação entre K e acurácia para a métrica Manhattan com normalização de 140x70.

	0	1	2	3	4	5	6	7	8	9
0	96	0	0	0	0	1	0	0	0	0
1	0	95	0	0	0	0	0	0	0	0
2	0	4	102	0	1	0	1	2	1	0
3	0	1	0	98	0	1	0	1	2	0
4	0	10	2	0	82	0	0	0	0	1
5	1	0	0	4	0	91	1	0	0	0
6	2	5	0	0	0	0	99	0	0	0
7	1	8	0	0	0	0	0	85	0	3
8	0	3	0	3	0	1	0	1	78	1
9	0	1	0	0	5	0	0	8	1	97

Fig.10
Matriz de confusão da métrica Euclidiana com k=3 e normalização 140x70.

	0	1	2	3	4	5	6	7	8	9
0	94	2	0	0	0	1	0	0	0	0
1	0	93	0	1	0	0	1	0	0	0
2	2	11	85	2	1	1	3	5	1	0
3	0	1	1	94	0	2	0	3	2	0
4	0	14	1	0	74	0	1	1	0	4
5	1	3	0	8	0	84	1	0	0	0
6	0	10	0	0	2	0	94	0	0	0
7	0	14	0	0	1	0	0	79	0	3
8	1	9	0	4	0	4	0	5	63	1
9	0	7	0	0	5	0	0	19	0	81

Fig.10
Matriz de confusão da métrica Euclidiana com k=30 e normalização 140x70.

Mesmo no pior caso, a Matriz da direita, temos um valor de acurácia de 84.1%, demonstrando que mesmo com alta variação de normalização e K, a classificador mantém uma constância. A seguir nas Fig.11, Fig.12, Fig.13 ,Fig.14 ,Fig.15 os gráficos correspondentes aos outros valores de normalização para a métrica Manhattan.

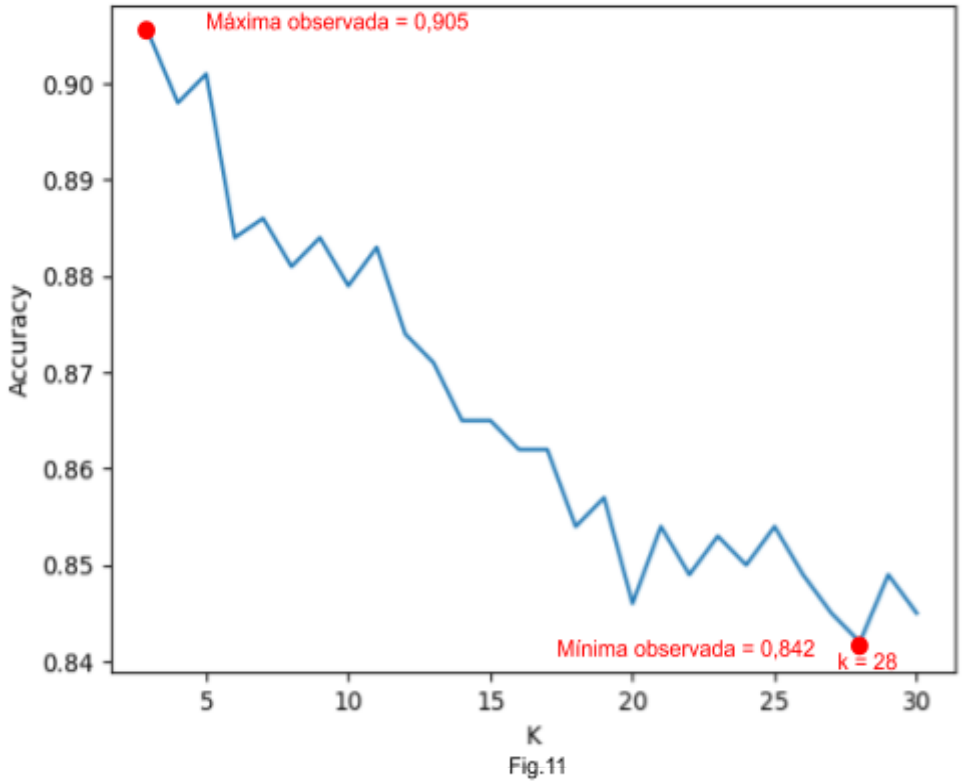

Fig.11
Relação entre K e acurácia para a métrica Manhattan com normalização de 20x10.

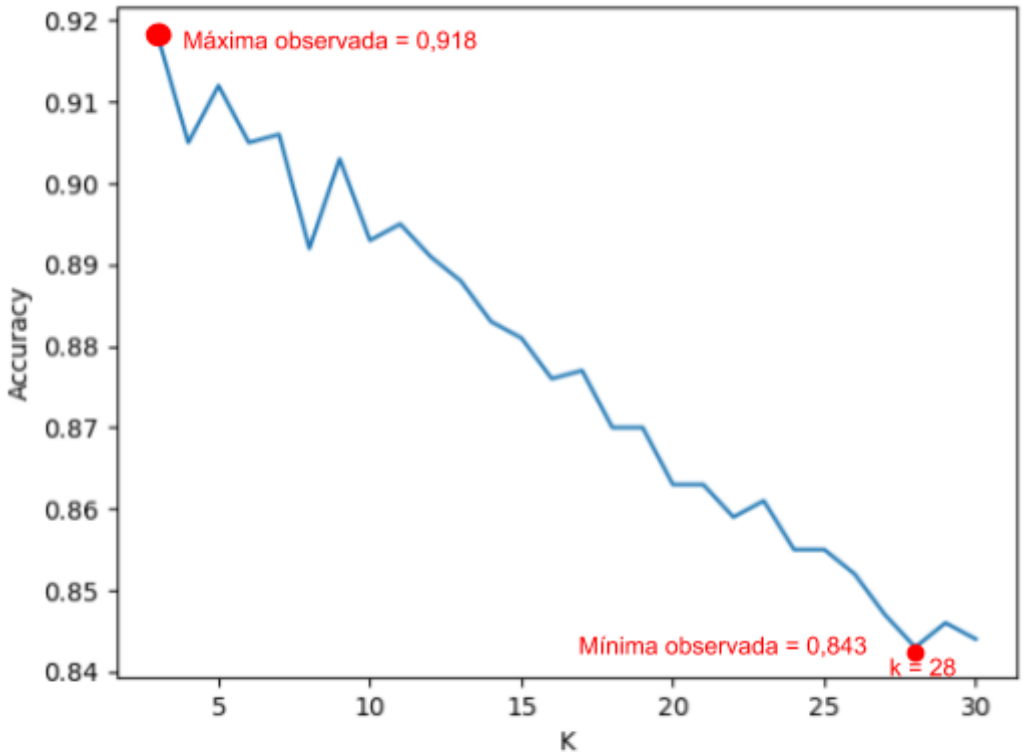

Fig.12
Relação entre K e acurácia para a métrica Manhattan com normalização de 40x20.

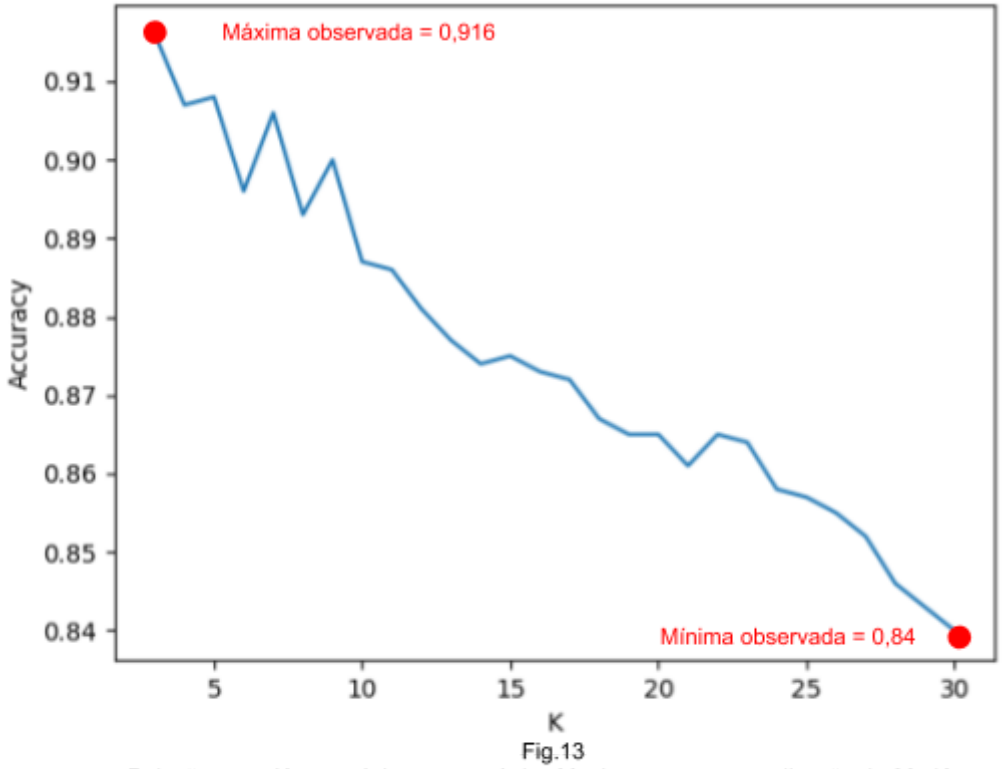

Fig.13
Relação entre K e acurácia para a métrica Manhattan com normalização de 80x40.

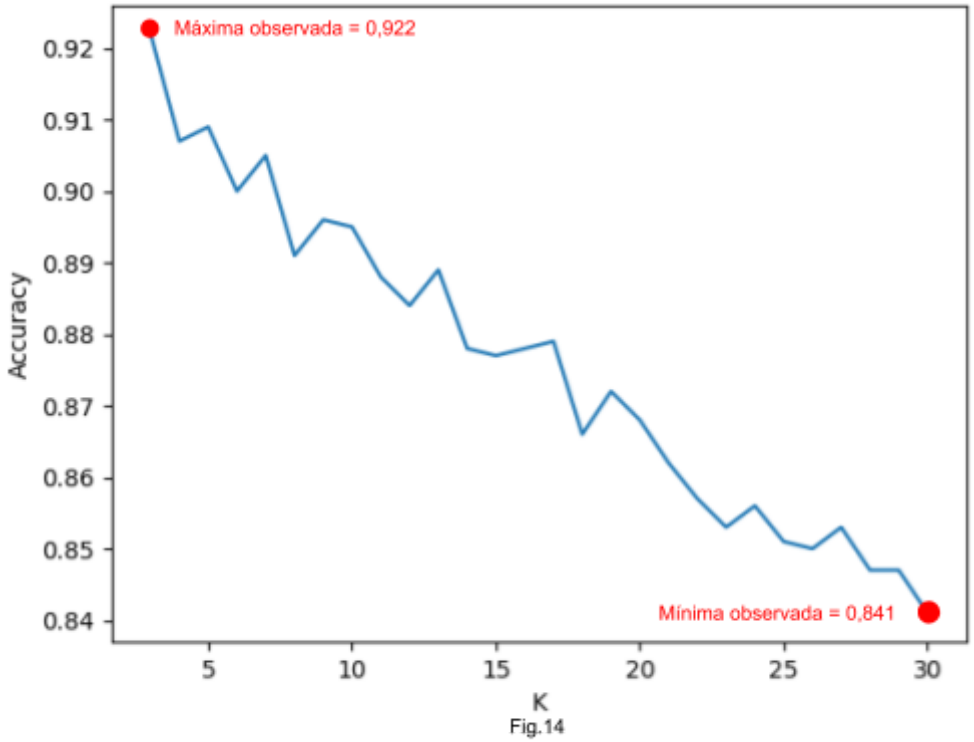

Fig.14
Relação entre K e acurácia para a métrica Manhattan com normalização de 100x50.

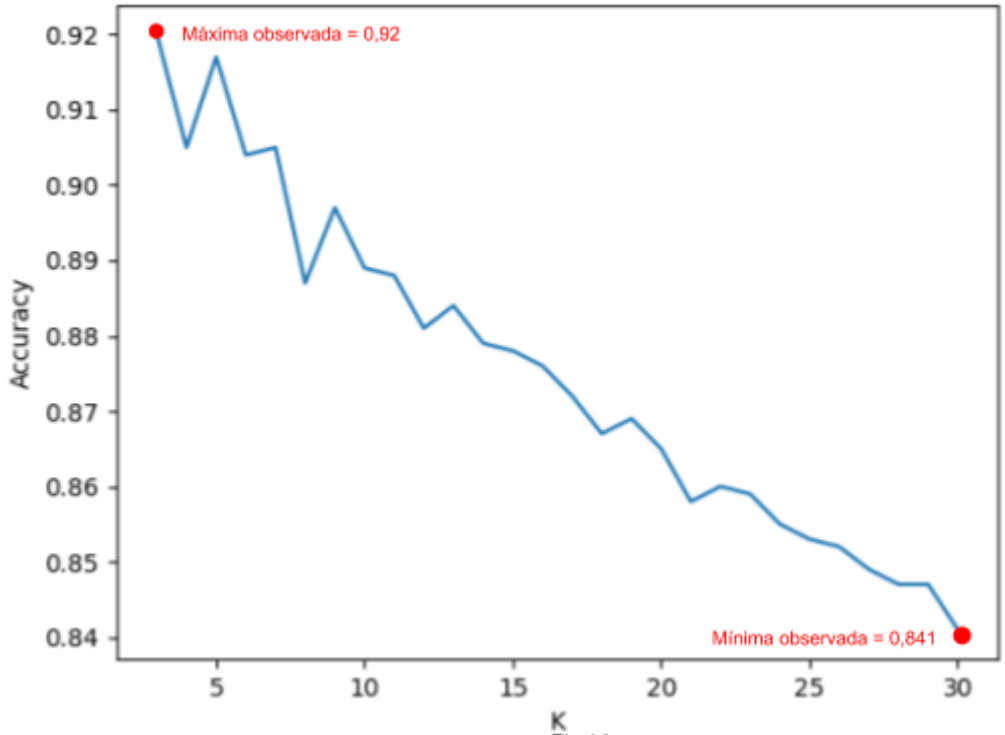

Fig.14
Relação entre K e acurácia para a métrica Manhattan com normalização de 200x100.

3. Métrica Minkowski

Segue a apresentação da última métrica utilizada, que apresentou todos os mesmos resultados já obtidos anteriormente pelas outras duas métricas. A normalização que apresentou maximização da acurácia foi a de 140X70.

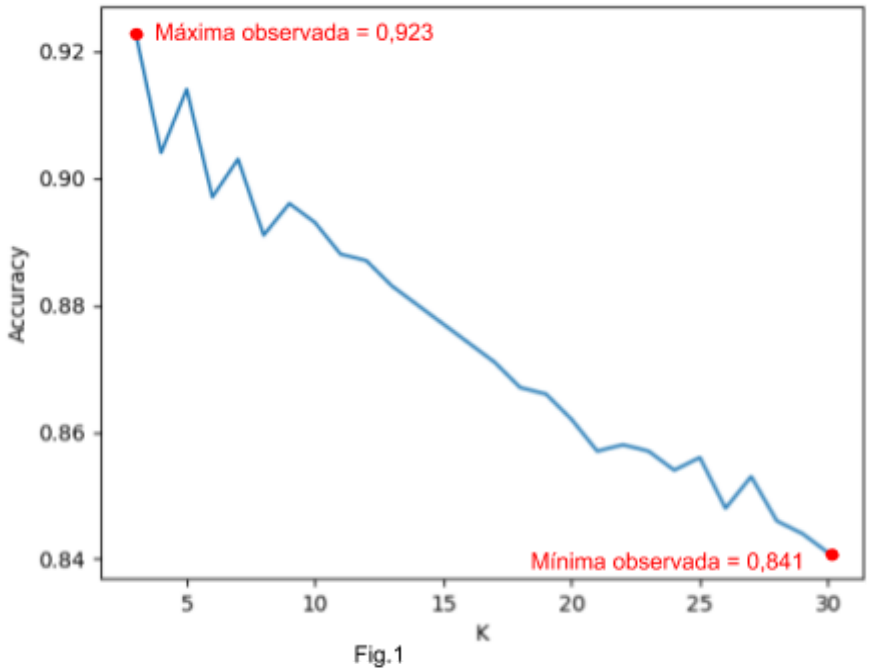

Fig.1
Relação entre K e acurácia para a métrica Euclidiana com normalização de 140x70.

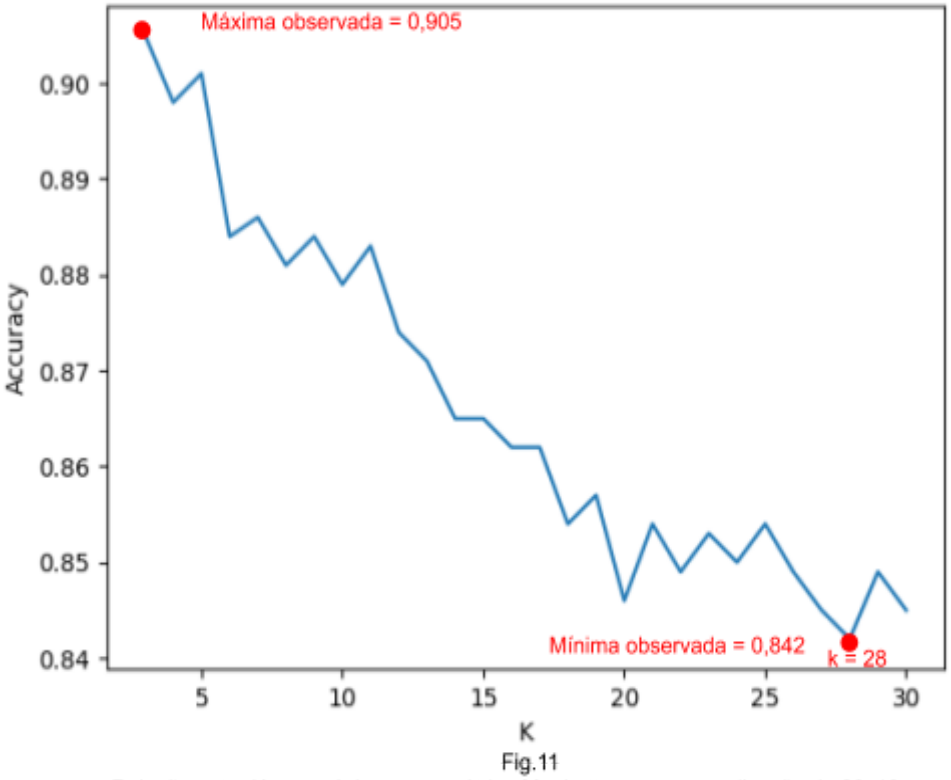

Fig.11
Relação entre K e acurácia para a métrica Manhattan com normalização de 20x10.

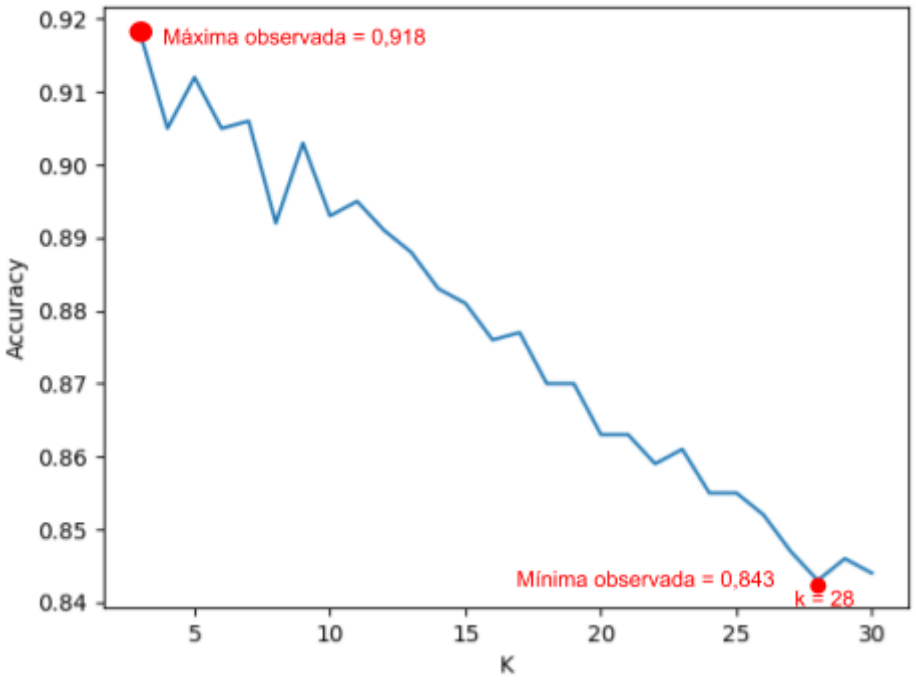

Fig.12
Relação entre K e acurácia para a métrica Manhattan com normalização de 40x20.

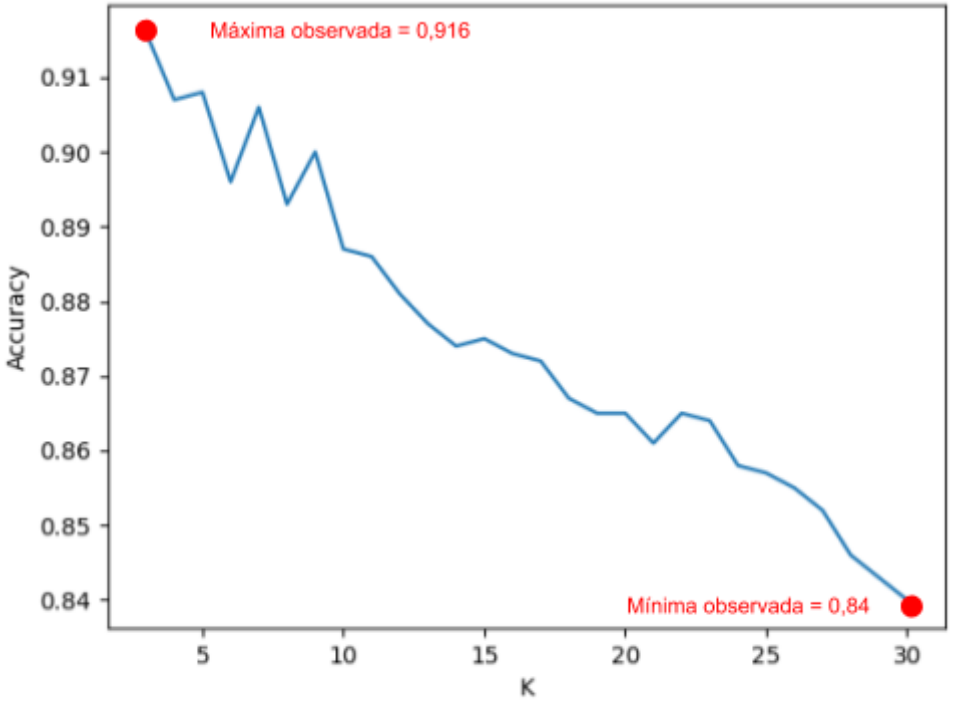

Fig.13
Relação entre K e acurácia para a métrica Manhattan com normalização de 80x40.

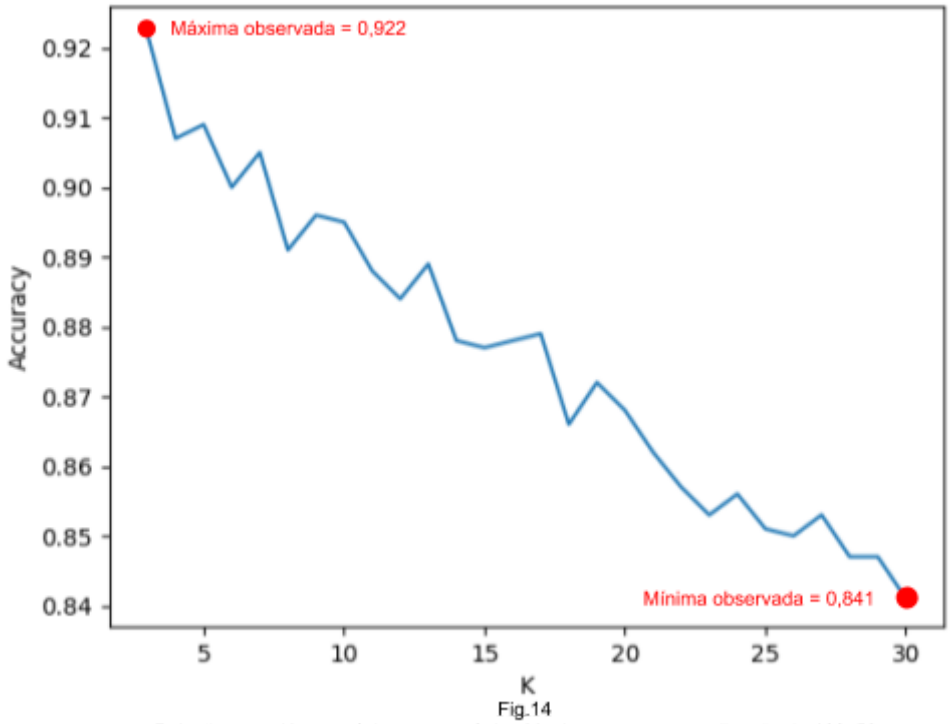

Fig.14
Relação entre K e acurácia para a métrica Manhattan com normalização de 100x50.

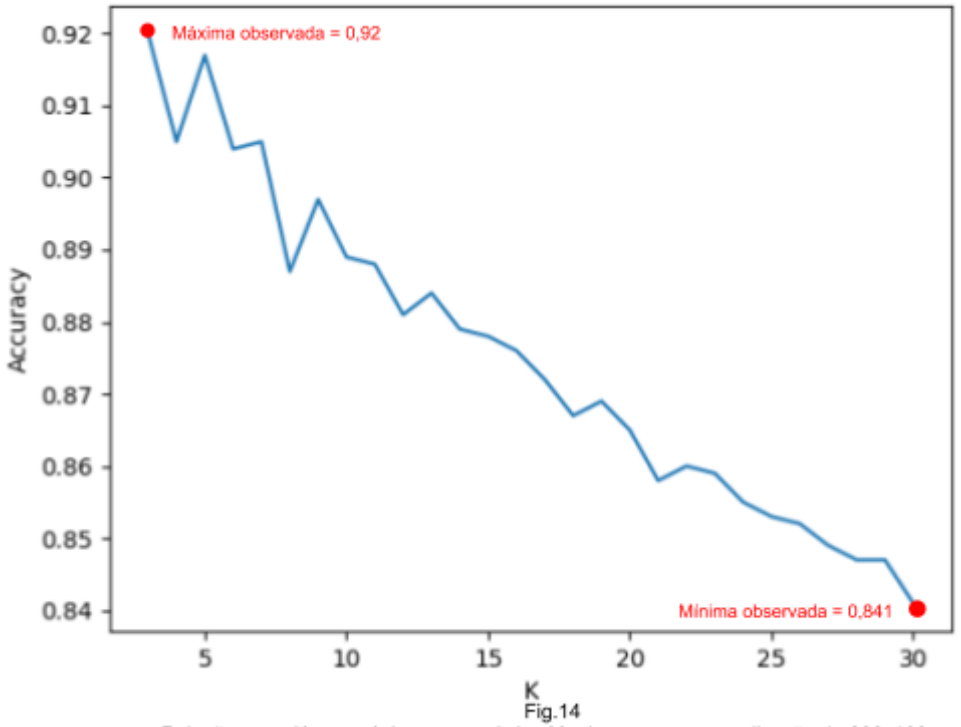

Fig.14
Relação entre K e acurácia para a métrica Manhattan com normalização de 200x100.

Conclusão

Apresentado todos os testes e resultado é possível verificar que para as 3 métricas do cálculo das distâncias, todos os resultados observados em relação a acurácia, foram iguais, assim como as matrizes de Confusão para todos os casos de variação tanto de K quanto normalização, dado o problema de dados balanceados. Para esse caso problema foi notável que pequenos valores de normalização a acurácia não era considerável. A grande variação detectada foi em relação ao tempo de execução para as diferentes variações de normalização. Foi utilizado o comando *time* na execução do programa. A mais rápida foi de 941 segundos, correspondente à normalização de 20x10. A mais lenta foi 37 minutos e 52 segundos, correspondente a normalização de 200x100, a que foi observada com maximização da acurácia, 140x70, foi de 16 minutos e 18 segundos.