UNIversidade federal do Paraná

MATEUS FELIPE DE CÁSSIO FERREIRA

RELATÓRIO: IMPACTOS DA REPRESENTAÇÃO



curitiba

2021mateus felipe de cássio ferreira

relatório: impactos da representação

Relatório apresentado como requisito parcial à conclusão da disciplina CI171 – Aprendizagem de Máquina, no Curso de Bacharelado em Informática Biomédica, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Orientador: Prof. Dr. Luiz Eduardo Soares de Oliveira

curitiba

2021

SUMÁRIO

[1 INTRODUÇÃO 3](#_Toc73302207)

[2 METODOLOGIA 4](#_Toc73302208)

[3 RESULTADOS 6](#_Toc73302209)

[4 CONCLUSÃO 11](#_Toc73302210)

# INTRODUÇÃO

Este relatório busca apresentar os resultados obtidos referentes ao primeiro Laboratório da disciplina de CI171 – Aprendizagem de Máquina, que consiste em identificar os impactos de uma representação de um conjunto de dados, em um vetor de características, na acurácia final de um classificador. Assim, este relatório tomou como base o arquivo *digits.py*, que extrai a representação de uma base de dados contendo imagens de dígitos inteiros manuscritos, e o arquivo *knn.py,* que implementa o algoritmo do classificador kNN (*k Nearest Neighbors*), que é um classificador baseado em instâncias e distâncias.

O objetivo deste relatório é o de propor o tamanho da imagem, com seu respectivo vetor de características, que maximiza a acurácia desse classificador.

# METODOLOGIA

Assim, como mencionado anteriormente, este relatório utilizou arquivos *digits.py* e *knn.py* para identificar os impactos que a alteração da representação do conjunto de dados gera sobre o resultado da acurácia do classificador.

Assim, utilizou-se uma base de dados, já rotulada, de 2000 imagens de dígitos inteiros manuscritos. Essa base de dado foi separada entre treinamento e validação, ambas contendo 50% do total de imagens no banco de dados.

No arquivo *digits.py* foi determinado um conjunto de 10 pares de valores [X,Y], sem um critério previamente definido, que irão compor os novos tamanhos no redimensionamento das imagens do banco de dados, sendo:

* features\_1.txt = [20,10];
* features\_2.txt = [10,20];
* features\_3.txt = [20,20];
* features\_4.txt = [10,10];
* features\_5.txt = [100,100];
* features\_6.txt = [90,50];
* features\_7.txt = [80,30];
* features\_9.txt = [120,70];
* features\_8.txt = [140,70];
* features\_10.txt = [110,60].

Esse redimensionamento dos dados é importante visto que o vetor de características que será extraído dessas figuras precisa ter o mesmo tamanho para que a classificação ocorra da maneira adequada.

Com relação ao número de vizinhos mais próximos, parâmetro fundamental na construção de um classificador kNN, o valor de **K** variou de 1 à 10. Esse valor é um dos parâmetros da função *KNeighborsClassifier*() que constrói um kNN.

Por fim, foram utilizadas quatro diferentes métricas para o cálculo de distância para a classificação, que foram:

* *euclidean;*
* *manhattan;*
* *chebyshev;*
* *minkowski.*

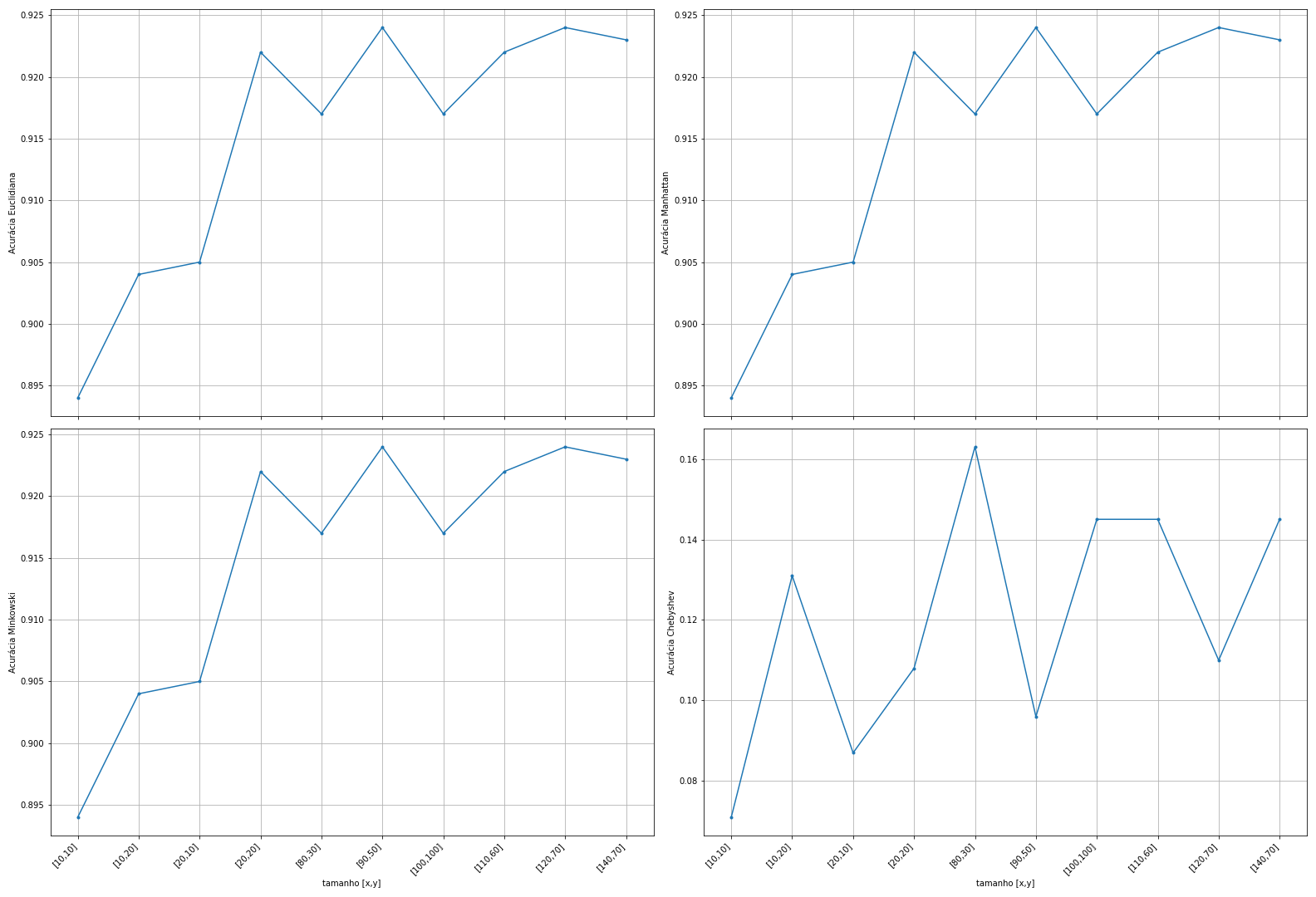
Uma vez que se trata de um problema balanceado, em que o número de representações de todas as classes é bastante semelhante, a métrica utilizada para avaliar o melhor conjunto de parâmetros será a medida de acurácia.

# RESULTADOS

O primeiro passo foi determinar, no conjunto de vetores [X,Y] do redimensionamento das imagens, qual seria o conjunto que maximizaria a *performance* do classificador, tomando um valor de **K** de 3 e variando a métrica de distância. A Figura 1 apresenta o resultado, para cada par de valor [X,Y] que obtivemos um melhor resultado de acurácia.

Curiosamente, o resultado entre das métricas “*euclidean”, “manhattan” e “minkowski”* obtiveram resultados idênticos de acurácia. Nesses gráficos da Figura 1 nota-se que para o valor de [X,Y] = [90,50] e [X,Y] = [120,70] a acurácia é maximizada, com um valor de 0.924.

Por outro lado, a métrica “*chebyshev”* apresentou um nível de acurácia baixíssimo para todos os valores de [X,Y].

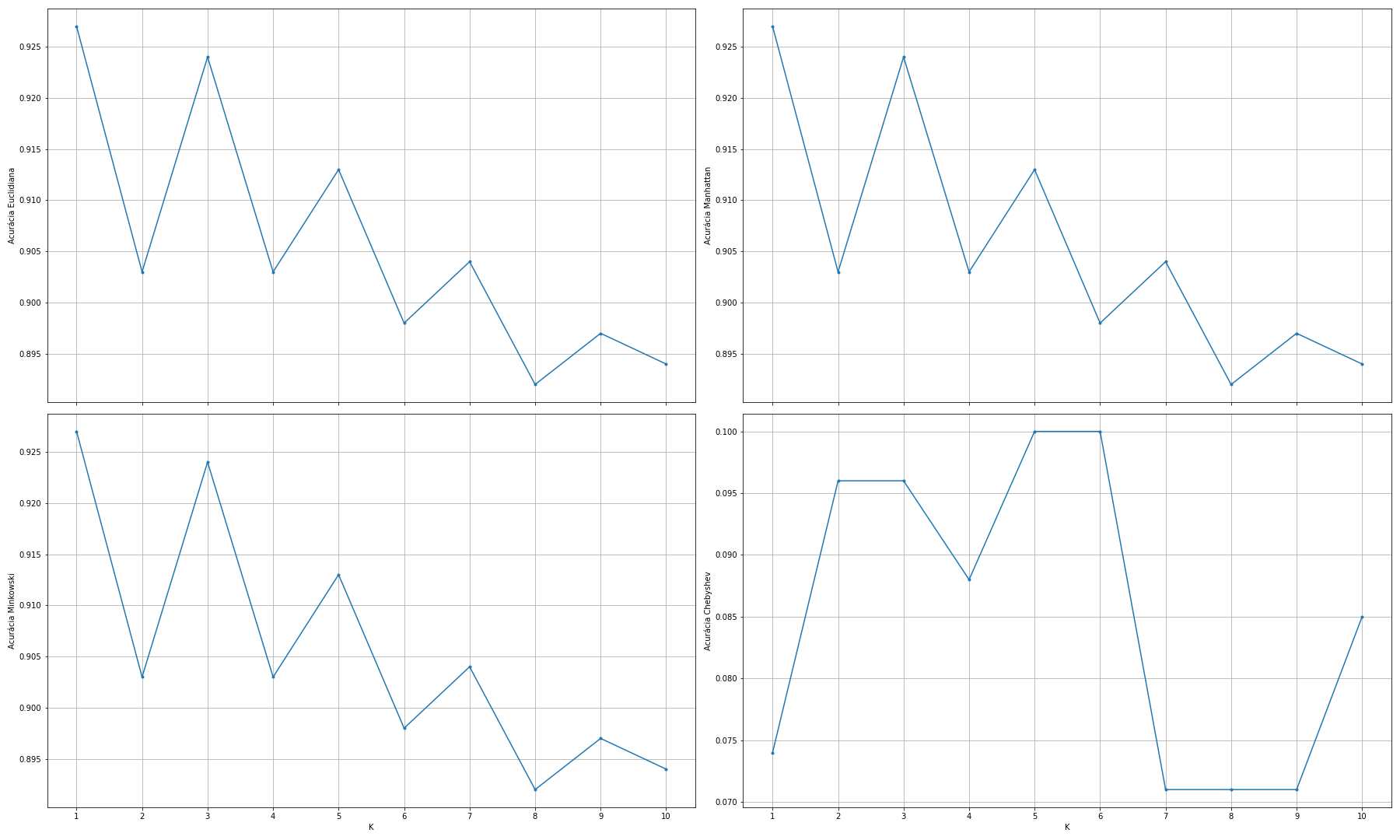
FIGURA 1 – COMPARATIVO DA ACURÁCIA ENTRE OS DIFERENTES PARES DE VALORES [X,Y] E DIFERENTES MÉTRICAS DE DISTÂNCIA ABORDADOS

FONTE: O autor (2021).

Assim, tomando o resultado anterior como uma base, a próxima etapa consistiu em buscar os parâmetros de **K** e a métrica de distância que conseguisse maximizar ainda mais a acurácia do classificador.

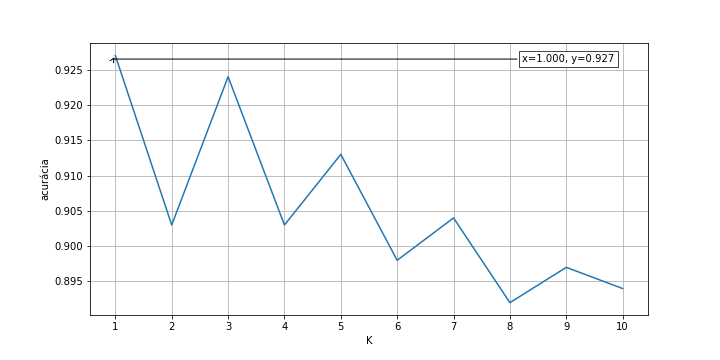
A Figura 2 apresenta um comparativo das acurácias obtidas, para cada valor de **K** (que varia de 1 à 10), levando em consideração a métrica de distância escolhida.

Assim como na figura anterior, a Figura 2 apresentou os mesmos resultados de acurácia para cada uma das três métricas (“*euclidean”, “manhattan” e “minkowski”)*. Congruente com o resultado anterior, a métrica “*chebyshev”* apresentou um nível de acurácia muito baixo, mesmo variando o número de vizinhos buscados (**K**).

FIGURA 2 – COMPARATIVO DA ACURÁCIA ENTRE OS DIFERENTES VALORES DE K E DIFERENTES MÉTRICAS DE DISTÂNCIA ABORDADOS

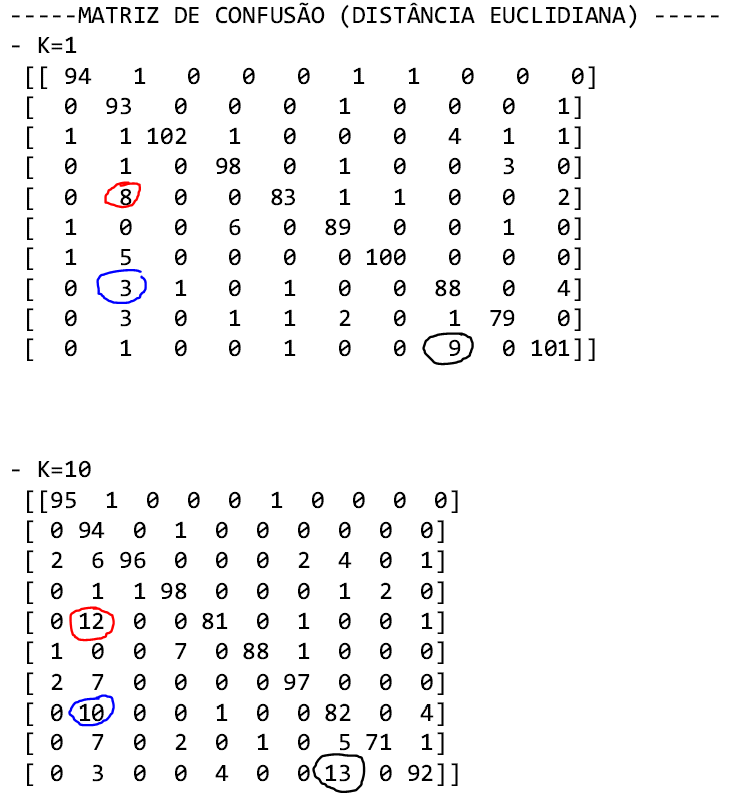
FONTE: O autor (2021).

A Figura 3 mostra, em detalhes, um gráfico da métrica de *“euclidean”,* com diferentes valores de **K**. Uma vez que essa figura é semelhante às métricas *“manhattan” e “minkowski”* percebe-se que, para o valor de **K = 1,** a acurácia atingiu o valor de 0.927, que é o valor máximo obtido tomando um tamanho de dimensionamento das imagens [X,Y] = [90,50]. Assim como esperado, a acurácia é reduzida quando utilizamos valores pares para **K**.

FIGURA 3 – ACURÁCIA PARA DIFERENTES VALORES DE K

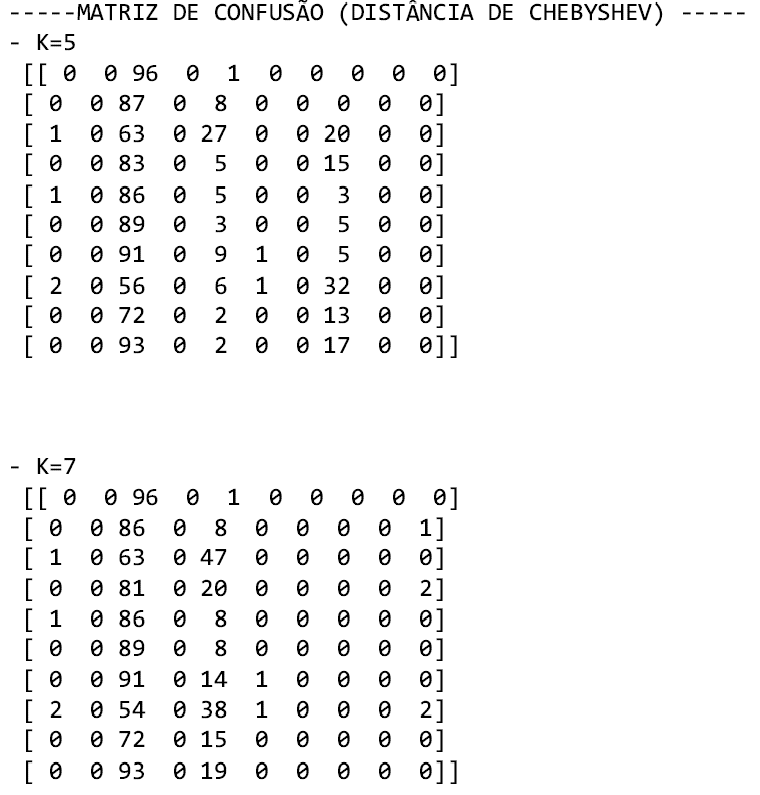
FONTE: O autor (2021).

Com relação às matrizes de confusão, percebe-se que, tomando o valor de **K = 1** quando comparado ao valor de **K = 10,** o classificador cometeu maiores erros ao classificar alguns números específicos. A Figura 4 apresenta as matrizes de confusão para o valor máximo de acurácia (**K = 1**) e o valor mínimo (**K = 10**). Percebe-se que o classificador começou a cometer maiores erros e confundir a classificação entre os números (1,4), em vermelho; (1,7), em azul e (7,9) em preto.

FIGURA 4 – MATRIZES DE CONFUSÃO PARA DIFERENTES VALORES DE K UTILIZANDO A MÉTRICA *euclidean*

FONTE: O autor (2021).

Por outro lado, a Figura 5 apresenta a matriz de confusão para o ponto de maior e menos acurácia do modelo que utiliza a métrica “*chebyshev”.* Como esperado, a matriz de confusão não apresenta, em sua diagonal principal, os maiores valores que representam os acertos do classificador, o que tornou a sua acurácia baixa quando comparado com as outras métricas descritas acima.

FIGURA 5 – MATRIZES DE CONFUSÃO PARA DIFERENTES VALORES DE K UTILIZANDO A MÉTRICA *chebyshev*

FONTE: O autor (2021).

# CONCLUSÃO

A partir dos resultados do experimento nota-se que existe um significativo impacto na escolha do tamanho da imagem e, consequentemente, no tamanho do vetor de características que são extraídas daquela imagem e do banco de dados como um todo. Este relatório descreveu que para um valor de redimensionamento [X,Y] = [90,50] e para o valor de **K = 1**, a acurácia do classificador kNN é maximizada, com um valor de 0.927.