Aluno: João Vitor de Camargo
Matrícula: 00274722

Relatório de Atividade Prática: Algoritmo KNN (K-Nearest Neighbors)

Análise dos dados normalizados

Utilizando valores de K = [1, 3, 5, 7, 9, 11, 13, 15], os seguintes resultados foram obtidos:

K	**Acurácia (em %)**
1	92,10526316
3	95,61403509
5	94,73684211
7	94,73684211
9	93,85964912
11	93,85964912
13	94,73684211
15	93,85964912

Com os resultados, é possível observar que a menor acurácia foi obtida quando utilizando apenas um vizinho. Em cenários em que apenas um vizinho é utilizado, *outliers* podem acabar influenciando a decisão incorretamente, como demonstrado em aula.

Por outro lado, a maior acurácia foi obtida com 3 vizinhos e a partir disso, o resultado parece estagnar entre 93% e 95%. Isso leva a acreditar que para o conjunto de dados em questão, aumentar o valor de K não trará resultados mais precisos. Na verdade, testes adicionais nos levam a acreditar no contrário. Com K = 105, a acurácia ainda se mantém perto de 94%. A partir de 155, ocorreu uma leve queda para 92%, menor valor até então. Todavia, à medida que K aumenta, a acurácia diminui. Com 205, 305 e 405, os valores obtidos são 89%, 75% e 63%. Isso mostra que um K grande mais (que representa uma parcela muito grande do conjunto) acaba diminuindo a relevância da proximidade. Considerando que o conjunto tem cerca de 450 entradas, um K igual a 405 acaba fazendo com que a maioria do conjunto geral vença e que, portanto, a proximidade acabe sendo praticamente irrelevante.

Análise dos dados não normalizados

Quando analisando a grandeza dos valores, é possível notar que a falta da normalização influenciará (provavelmente de forma negativa) nos resultados. Exemplo:

21.8	101.2	718.9	0.09	0.2	0.14	0.06
27.27	105.9	733.5	0.1	0.32	0.37	0.11
32.19	86.12	487.7	0.18	0.33	0.14	0.13
21.08	92.8	599.5	0.15	0.22	0.18	0.12
19.35	80.78	433.1	0.13	0.39	0.34	0.08
26.37	161.2	1780	0.13	0.24	0.27	0.18

Na amostra acima, é possível ver que, enquanto alguns atributos variam entre 0 e 0.5 (4 últimas colunas), outros podem chegar a casa de milhares (terceira coluna). Isso faz com que a relevância de alguns atributos seja quase nula, visto que outros irão influenciar milhares de vezes mais no cálculo da distância.

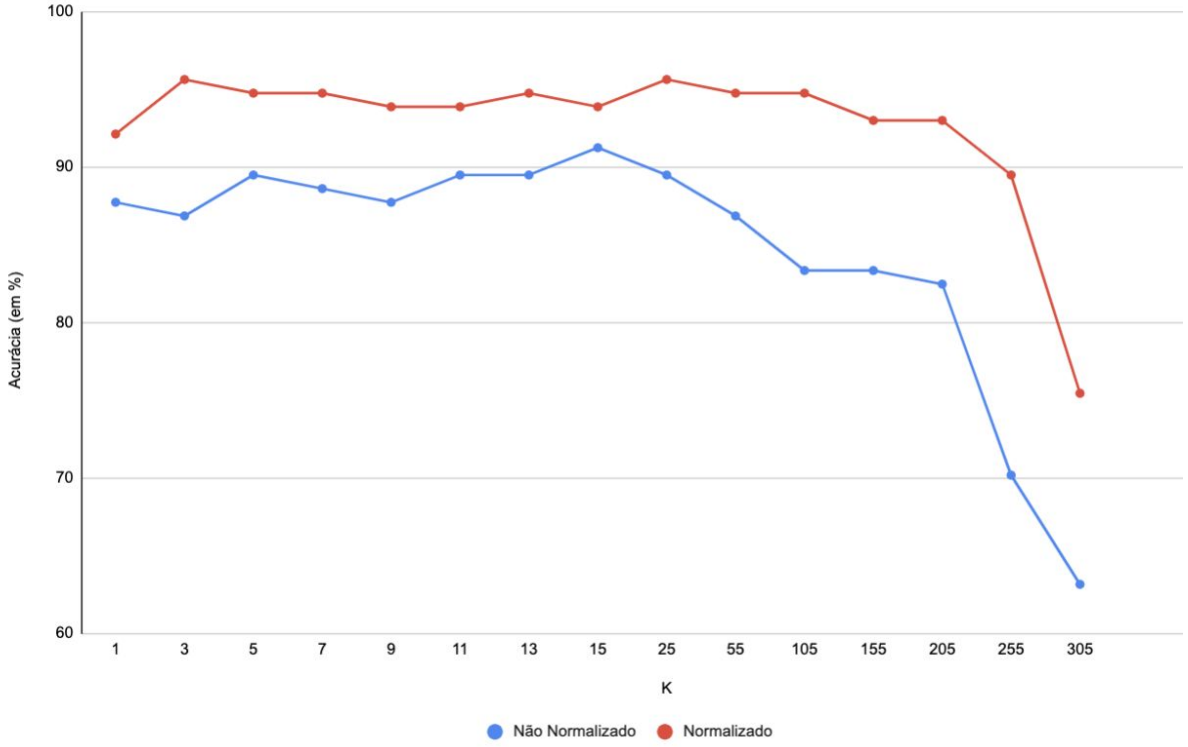
Utilizando valores de K = [1, 3, 5, 7, 9, 11, 13, 15], os seguintes resultados foram obtidos:

K	**Acurácia (em %)**
1	87,71929825
3	86,84210526
5	89,47368421
7	88,59649123
9	87,71929825
11	89,47368421
13	89,47368421
15	91,22807018

Observando apenas esses resultados, a conclusão inicial é de que, para o conjunto não normalizado, o aumento do valor de K é benéfico e aumenta a acurácia. Todavia, embora a maior acurácia tenha de fato sido obtida com K = 15 (a maior na análise inicial), o aumento não se mantém. Com K = 25, a acurácia é 89%, enquanto para os normalizados, foi de 95%. O valor também começa a cair mais rápido sem a normalização: a acurácia para K = 205 e 255 para o conjunto não normalizado foi de 82% e 70%, a do conjunto normalizado foi 92% e 89%, respectivamente.

Comparação: dados normalizados e não normalizados

O gráfico abaixo faz uma comparação da acurácia obtida (em %) variando o valor de K para os conjuntos de dados normalizados e não normalizados:



Primeiramente é possível ver que o desempenho do conjunto não normalizado se manteve sempre inferior ao normalizado, mostrando assim a importância da normalização. Além disso, é também importante notar que a medida que K aumentou, a acurácia do conjunto não normalizado diminui mais rapidamente, como é possível ver com K = 55, por exemplo.