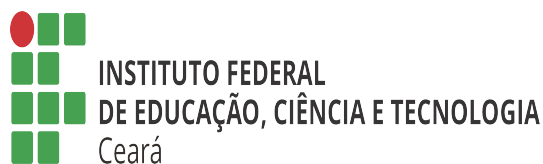


Instituto Federal de Educação, Ciência e Tecnologia do Ceará



Reconhecimento de Padrões

Relatório do Trabalho 1 - KNN e DMC

Aluno	Pedro Wilson Félix Magalhães Neto
Professor	Ajalmar Rego da Rocha Neto

Fortaleza, 13 de Abril de 2024

Conteúdo

1	Objetivo	1
2	Métodos Utilizados	1
2.1	KNN (K-Nearest Neighbors)	1
2.2	DMC (Distância Mínima ao Centro)	1
3	Procedimento Experimental	1
3.1	Resultados do KNN	2
3.2	KNN aplicado na IRIS Flower	2
3.3	KNN aplicado ao Artificial	4
3.4	KNN aplicado ao dataset Coluna	7
3.5	Resultados do DMC	9
3.6	DMC aplicado na IRIS Flower	9
3.7	DMC aplicado na Artificial	11
3.8	DMC aplicado ao dataset Coluna	13
4	Análise Comparativa	15
5	Conclusão	16

1 Objetivo

O objetivo deste trabalho é explorar e aplicar técnicas de reconhecimento de padrões para resolver problemas específicos. Este relatório apresenta uma análise comparativa entre dois algoritmos de classificação. Esses algoritmos foram aplicados aos conjuntos de dados Iris, column3C e artificial, para classificar diferentes espécies de flores, problemas na coluna e por último uma base artificial com dataset de dados aleatórios. Esses padrões podem ser utilizados para classificação, detecção de anomalias, segmentação, entre outros.

2 Métodos Utilizados

Para o experimento em questão foram utilizados dois modelos (KNN e DMC) para cada conjunto de dados.

2.1 KNN (K-Nearest Neighbors)

O algoritmo KNN é um método de classificação baseado em instâncias que classifica novos pontos de dados com base nos k pontos de dados mais próximos no espaço de características. A classe mais comum entre os vizinhos mais próximos é atribuída ao ponto de dados de teste.

2.2 DMC (Distância Mínima ao Centro)

O algoritmo DMC classifica pontos de dados com base na distância ao centro de cada classe no espaço de características. A classe cujo centro é o mais próximo do ponto de dados de teste é atribuída a esse ponto.

3 Procedimento Experimental

Os algoritmos KNN e DMC foram implementados e testados nos conjuntos de dados dividindo-os em conjuntos de treinamento e teste, com uma proporção de 80% para treinamento e 20% para teste, e em seguida feito um holdout com crossvalidation de 20 realizações, utilizando o parâmetro `random_state` (parâmetro que controla a aleatoriedade no processo de divisão dos dados em conjuntos de treinamento e teste durante a divisão) igual a 4, computando a acurácia e o desvio padrão para análise posterior.

3.1 Resultados do KNN

3.2 KNN aplicado na IRIS Flower

Após execução do algoritmo KNN no conjunto de dados da IRIS [1], os seguintes resultados foram obtidos:

Tabela 1: Conjunto de dados de treinamento IRIS

sepal-length	sepal-width	petal-length	petal-width	class
5.4	3.	4.5	1.5	1
4.6	3.2	1.4	0.2	0
6.7	2.5	5.8	1.8	2
4.9	3.	1.4	0.2	0
5.	2.3	3.3	1.	1
6.7	3.3	5.7	2.5	2
7.2	3.2	6.	1.8	2
5.8	2.6	4.	1.2	1
6.7	3.1	4.7	1.5	1
5.1	3.8	1.6	0.2	0
7.7	3.	6.1	2.3	2
5.	3.4	1.5	0.2	0
6.7	3.1	4.4	1.4	1
5.4	3.7	1.5	0.2	0
...
5.6	2.7	4.2	1.3	1
7.2	3.6	6.1	2.5	2
6.3	2.9	5.6	1.8	2
6.6	2.9	4.6	1.3	1
6.4	3.1	5.5	1.8	2
7.	3.2	4.7	1.4	1
6.3	2.3	4.4	1.3	1
6.5	3.	5.8	2.2	2
7.2	3.	5.8	1.6	2
7.7	2.8	6.7	2.	2

Tabela 2: Conjunto de dados de Teste IRIS

sepal-length	sepal-width	petal-length	petal-width	class
6.4	2.8	5.6	2.1	2
5.7	3.8	1.7	0.3	0
7.4	2.8	6.1	1.9	2
7.6	3.	6.6	2.1	2
7.3	2.9	6.3	1.8	2
6.	2.9	4.5	1.5	1
6.	2.7	5.1	1.6	1
5.8	4.	1.2	0.2	0
5.4	3.9	1.7	0.4	0
6.3	2.8	5.1	1.5	2
5.	3.	1.6	0.2	0
4.8	3.4	1.6	0.2	0
4.8	3.	1.4	0.1	0
6.1	2.9	4.7	1.4	1
5.7	2.5	5.	2.	2
4.9	3.1	1.5	0.1	0
5.	2.	3.5	1.	1
4.7	3.2	1.3	0.2	0
4.8	3.4	1.9	0.2	0
6.3	2.7	4.9	1.8	2
5.	3.2	1.2	0.2	0
6.7	3.3	5.7	2.1	2
6.2	2.2	4.5	1.5	1
5.	3.4	1.6	0.4	0
4.7	3.2	1.6	0.2	0
5.1	3.8	1.5	0.3	0
4.5	2.3	1.3	0.3	0
5.4	3.9	1.3	0.4	0
5.4	3.4	1.7	0.2	0
5.8	2.7	5.1	1.9	2

- Melhor K encontrado foi o 10, que proporcionou a maior acurácia média em 20 realizações.
- Acurácia média: 96,83%
- Desvio padrão da acurácia: 0.0324465372232
- Matriz de Confusão:

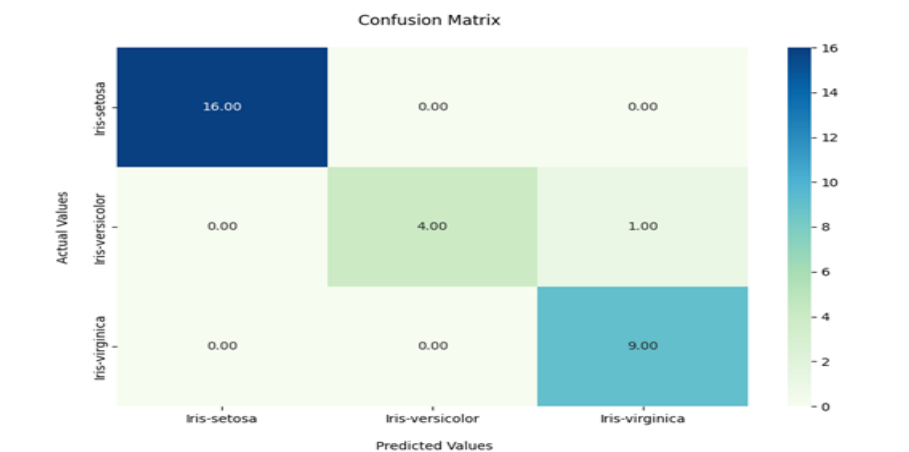


Figura 1: Matriz de Confusão IRIS KNN

- Superfície de decisão:

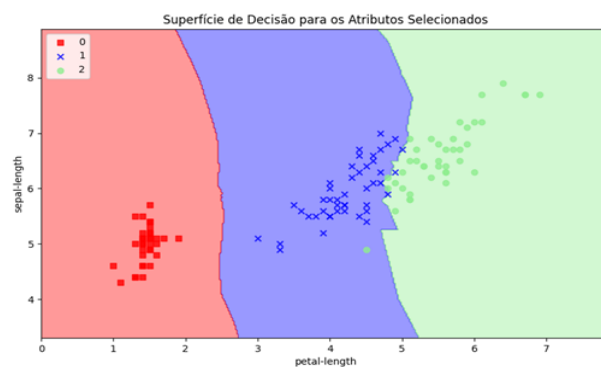


Figura 2: Superfície de decisão KNN IRIS

3.3 KNN aplicado ao Artificial

Após execução do algoritmo KNN no conjunto de dados da Artificial, os seguintes resultados foram obtidos:

Tabela 3: Conjunto de dados de treinamento Artificial KNN

Dados A	Dados B	Classes
-0.004268	-0.334478	0
0.021354	-0.450545	0
-0.420009	0.062253	0
0.977228	0.317345	0
-0.172149	0.035077	0
0.066048	-0.619702	0
1.027527	-0.094554	0
-0.363976	0.118806	0
0.463479	-0.071397	0
-0.46755	0.772365	0
1.256943	0.428881	0
0.102484	0.878226	0
0.260112	-0.386065	0
0.988671	0.494784	0
0.104755	1.308494	0
-0.189939	-0.092242	0
0.925954	0.925959	1
0.108662	0.442478	0
0.326031	1.294497	0
1.076515	0.394968	1
1.114283	0.48637	0
1.114359	-0.204005	0
-0.036571	0.904783	0
0.171564	0.259908	0
1.157075	0.956277	1
1.499391	1.242684	1
0.848472	-0.058711	0
0.712858	0.55339	1
0.679715	1.099374	1
-0.145667	1.334291	0
1.204817	1.481624	1
0.853454	0.852723	1

Tabela 4: Conjunto de dados de teste Artificial KNN

Dados A	Dados B	Classes
-0.265384	0.902218	0
1.029017	-0.628524	0
0.851539	1.171573	1
-0.190276	0.585742	0
-0.214062	1.193429	0
0.233524	1.054191	0
0.650146	-0.378274	0
0.454533	0.822189	1

- Melhor K encontrado foi o 3, que proporcionou a maior acurácia média em 20 realizações.
- Acurácia média: 93,75%
- Desvio padrão da acurácia: 0.073950997
- Matriz de Confusão:

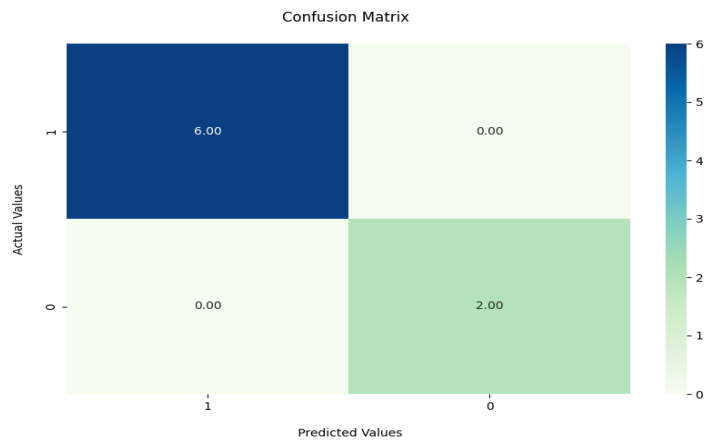


Figura 3: Matriz de Confusão Artificial I

- Superfície de decisão:

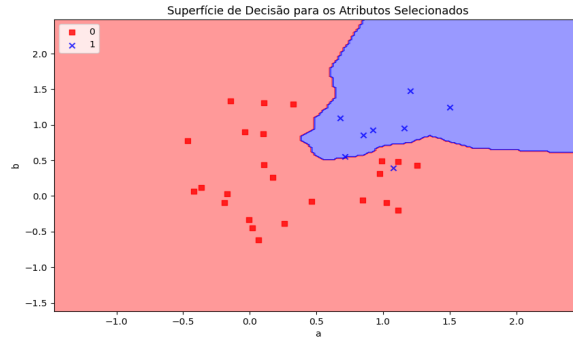


Figura 4: Superfície de decisão Artificial I

3.4 KNN aplicado ao dataset Coluna

Após execução do algoritmo KNN no conjunto de dados da Coluna [2], os seguintes resultados foram obtidos:

Tabela 5: Conjunto de dados de treinamento Coluna

pelvic-incidence	pelvic-tilt	lumbarlor-dosis-angle	sacral-slope	pelvic-radius	degree-spondylolisthesis	class
95.48	46.55	59.	48.93	96.68	77.28	2
48.92	19.96	40.26	28.95	119.32	8.03	0
35.88	1.11	43.46	34.77	126.92	-1.63	1
...
58.83	37.58	125.74	21.25	135.63	117.31	2
61.41	25.38	39.1	36.03	103.4	21.84	2
80.07	48.07	52.4	32.01	110.71	67.73	2

Tabela 6: Conjunto de dados de teste Coluna

pelvic-incidence	pelvic-tilt	lumbar-lordosis-angle	sacral-slope	pelvic-radius	degree-spondylolisthesis	class
5.8100e+01	1.4840e+01	7.9650e+01	4.3260e+01	1.1359e+02	5.0240e+01	2
5.6450e+01	1.9440e+01	4.3580e+01	3.7000e+01	1.3919e+02	-1.8600e+00	1
9.2030e+01	3.5390e+01	7.7420e+01	5.6630e+01	1.1572e+02	5.8060e+01	2
4.4220e+01	1.5100e+00	4.6110e+01	4.2710e+01	1.0863e+02	4.2810e+01	2
...
5.0750e+01	2.0240e+01	3.7000e+01	3.0520e+01	1.2234e+02	2.2900e+00	1
7.0250e+01	1.0340e+01	7.6370e+01	5.9910e+01	1.1924e+02	3.2670e+01	2
8.1100e+01	2.4790e+01	7.7890e+01	5.6310e+01	1.5184e+02	6.5210e+01	2
6.1450e+01	2.2690e+01	4.6170e+01	3.8750e+01	1.2567e+02	-2.7100e+00	1

- Melhor K encontrado foi o 7, que proporcionou a maior acurácia média em 20 realizações.
- Acurácia média: 84,1935%
- Desvio padrão da acurácia: 0.03328
- Matriz de Confusão:

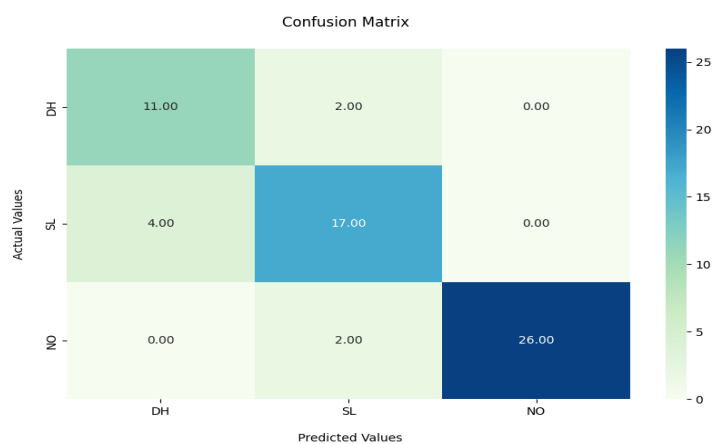


Figura 5: Matriz de Confusão Coluna

- Superfície de decisão:

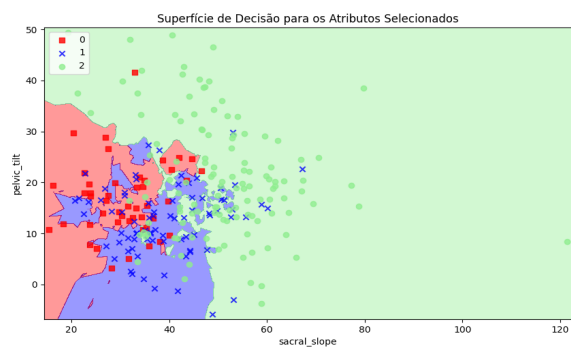


Figura 6: Superfície de decisão Coluna

3.5 Resultados do DMC

3.6 DMC aplicado na IRIS Flower

Após execução do algoritmo DMC no conjunto de dados da IRIS, os resultados foram obtidos através dos dados selecionados para treinamento e teste abaixo:

Tabela 7: Conjunto de dados de treinamento IRIS DMC

sepal-length	sepal-width	petal-length	petal-width	class
4.3	3.	1.1	0.1	0
5.8	2.7	3.9	1.2	1
5.7	3.	4.2	1.2	1
4.8	3.4	1.6	0.2	0
6.7	3.1	5.6	2.4	2
5.5	2.4	3.8	1.1	1
4.9	3.1	1.5	0.1	0
...
6.3	3.4	5.6	2.4	2
6.	2.7	5.1	1.6	1
7.7	3.	6.1	2.3	2
6.1	2.9	4.7	1.4	1
6.2	2.2	4.5	1.5	1
6.4	3.1	5.5	1.8	2
6.8	3.	5.5	2.1	2
5.1	3.7	1.5	0.4	0
4.6	3.4	1.4	0.3	0
5.	3.5	1.3	0.3	0
6.1	2.8	4.7	1.2	1
6.9	3.1	5.4	2.1	2
6.3	2.5	4.9	1.5	1

Tabela 8: Conjunto de dados de Teste IRIS DMC

sepal-length	sepal-width	petal-length	petal-width	class
6.5	3.	5.2	2.	2
6.4	2.7	5.3	1.9	2
6.	2.9	4.5	1.5	1
5.1	3.3	1.7	0.5	0
6.5	2.8	4.6	1.5	1
5.8	2.7	4.1	1.	1
5.	3.2	1.2	0.2	0
4.8	3.1	1.6	0.2	0
6.5	3.	5.8	2.2	2
6.8	3.2	5.9	2.3	2
5.3	3.7	1.5	0.2	0
7.9	3.8	6.4	2.	2
4.6	3.1	1.5	0.2	0
5.4	3.4	1.5	0.4	0
7.7	3.8	6.7	2.2	2
6.1	2.6	5.6	1.4	2
5.2	2.7	3.9	1.4	1
6.4	2.9	4.3	1.3	1
6.5	3.2	5.1	2.	2
5.8	2.8	5.1	2.4	2
5.1	3.8	1.5	0.3	0
6.3	2.5	5.	1.9	2
5.	3.5	1.6	0.6	0
6.4	3.2	5.3	2.3	2
6.	3.4	4.5	1.6	1
5.8	4.	1.2	0.2	0
5.9	3.2	4.8	1.8	1
5.6	3.	4.5	1.5	1
5.8	2.7	5.1	1.9	2
6.7	3.1	4.4	1.4	1

- Acurácia média: 93,66%
- Desvio padrão da acurácia: 0.0433333
- Matriz de Confusão:

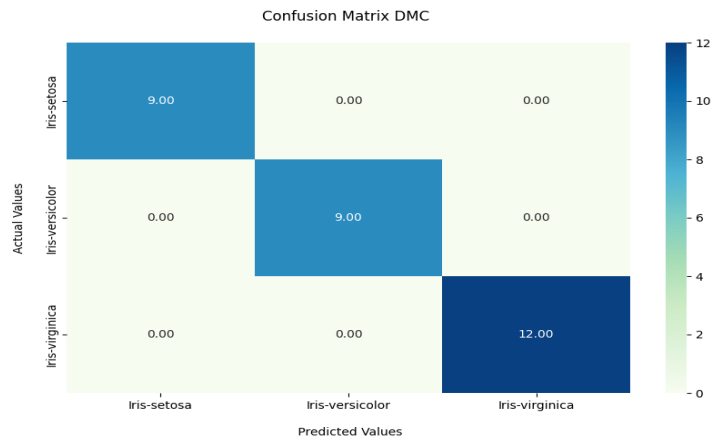


Figura 7: Matriz de Confusão IRIS DMC

- Superfície de decisão:

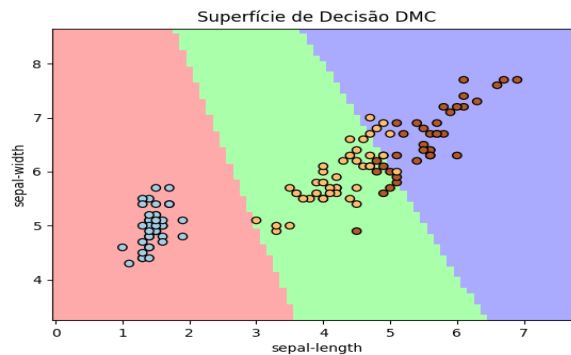


Figura 8: Superfície de decisão IRIS DMC

3.7 DMC aplicado na Artificial

Após execução do algoritmo DMC no conjunto de dados Artificial, os resultados foram obtidos através dos dados selecionados para treinamento e teste abaixo:

Tabela 9: Conjunto de dados de treinamento Artificial DMC

Dados A	Dados B	Classes
1.256943	0.428881	0
0.104755	1.308494	0
-0.363976	0.118806	0
0.463479	-0.071397	0
1.027527	-0.094554	0
-0.214062	1.193429	0
0.260112	-0.386065	0
-0.172149	0.035077	0
...
0.326031	1.294497	0
-0.145667	1.334291	0
0.650146	-0.378274	0
0.853454	0.852723	1
-0.265384	0.902218	0
0.712858	0.55339	1

Tabela 10: Conjunto de dados de teste DMC Artificial

Dados A	Dados B	Classes
0.454533	0.822189	1
1.204817	1.481624	1
-0.420009	0.062253	0
0.108662	0.442478	0
0.925954	0.925959	1
0.679715	1.099374	1
-0.190276	0.585742	0
-0.004268	-0.334478	0

- Acurácia média: 74,37%
- Desvio padrão da acurácia: 0.08361930100162282
- Matriz de Confusão:

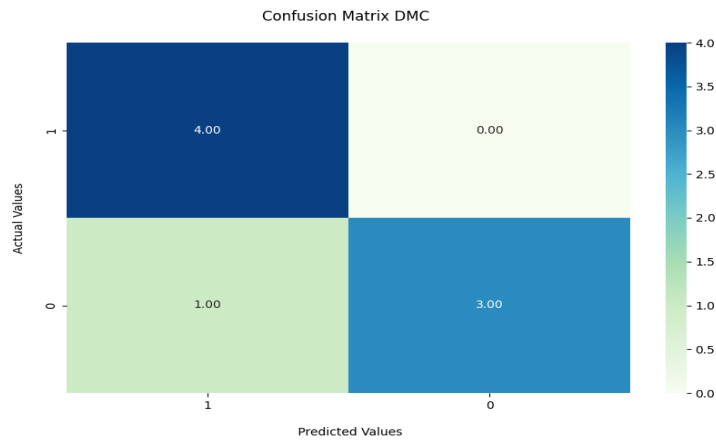


Figura 9: Matriz de Confusão Artificial DMC

- Superfície de decisão:

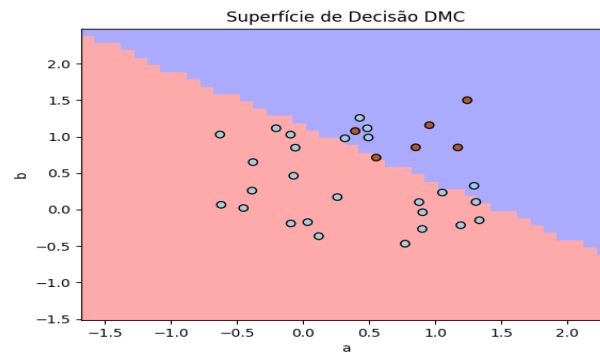


Figura 10: Superfície de decisão Artificial DMC

3.8 DMC aplicado ao dataset Coluna

Após execução do algoritmo DMC no conjunto de dados da Coluna3C, os resultados foram obtidos através dos dados selecionados para treinamento e teste abaixo:

Tabela 11: Conjunto de dados de treinamento Coluna DMC

pelvic-incidence	pelvic-tilt	lumbarlor-dosis-angle	sacral-slope	pelvic-radius	degree-spondylolisthesis	class
69.78	13.78	58.	56.	118.93	17.91	2
36.13	22.76	29.	13.37	115.58	-3.24	0
69.76	19.28	48.5	50.48	96.49	51.17	2
...
89.83	22.64	90.56	67.2	100.5	3.04	1
89.68	32.7	83.13	56.98	129.96	92.03	2
63.17	6.33	63.	56.84	110.64	42.61	2

Tabela 12: Conjunto de dados de treinamento Coluna DMC

pelvic-incidence	pelvic-tilt	lumbarlor-dosis-angle	sacral-slope	pelvic-radius	degree-spondylolisthesis	class
57.04	0.35	49.2	56.69	103.05	52.17	2
48.03	3.97	58.34	44.06	125.35	35.	2
85.58	30.46	78.23	55.12	114.87	68.38	2
40.56	17.98	34.	22.58	121.05	-1.54	0
54.6	21.49	29.36	33.11	118.34	-1.47	1
72.64	18.93	68.	53.71	116.96	25.38	2
51.62	15.97	35.	35.66	129.39	1.01	1
38.66	12.99	40.	25.68	124.91	2.7	0
95.38	24.82	95.16	70.56	89.31	57.66	2
40.25	13.92	25.12	26.33	130.33	2.23	0
63.03	22.55	39.61	40.48	98.67	-0.25	0
35.88	1.11	43.46	34.77	126.92	-1.63	1
47.81	10.69	54.	37.12	125.39	-0.4	1
...
57.15	16.49	42.84	40.66	113.81	5.02	1
55.84	28.85	47.69	27.	123.31	2.81	0
47.32	8.57	35.56	38.75	120.58	1.63	1
39.36	7.01	37.	32.35	117.82	1.9	1
44.25	1.1	38.	43.15	98.27	23.91	2
43.12	13.82	40.35	29.3	128.52	0.97	1

- Acurácia média: 75,72%
- Desvio padrão da acurácia: 0.06620275198
- Matriz de Confusão:

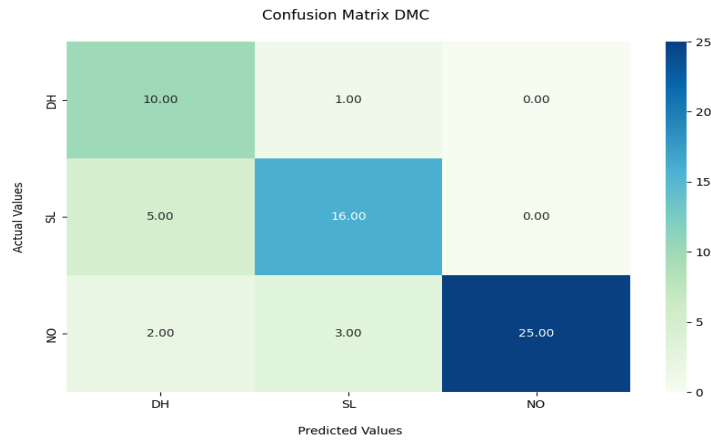


Figura 11: Matriz de Confusão Coluna DMC

- Superfície de decisão:

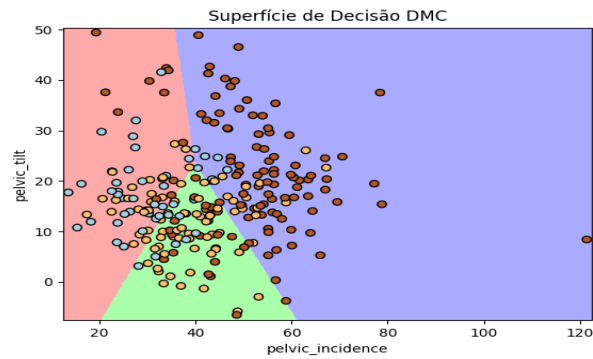


Figura 12: Superfície de decisão Coluna DMC

4 Análise Comparativa

Ambos os algoritmos KNN e DMC obtiveram resultados promissores na classificação dos 3 datasets utilizados. No entanto, o KNN demonstrou uma acurácia ligeiramente superior em comparação com o DMC. Ao examinar

as matrizes de confusão, observamos que ambos os algoritmos classificaram corretamente a maioria das instâncias. No entanto, o DMC teve um número ligeiramente maior de erros de classificação, especialmente entre as espécies Versicolor e Virginica no caso do dataset da IRIS por exemplo.

Tabela 13: Resultados de Acurácia e Desvio Padrão

Algoritmo	Acurácia Média	Desvio Padrão da Acurácia
KNN IRIS	96,83%	0,03244
DMC IRIS	93,66%	0,04333
KNN Artificial I	93,75%	0,07395
DMC Artificial I	74,37%	0,08361
KNN Coluna3C	84,19%	0,03328
DMC Coluna3C	75,72%	0,06620

5 Conclusão

Com base nos resultados obtidos, podemos concluir que o algoritmo KNN tende a ter um desempenho ligeiramente melhor do que o DMC na classificação do conjunto de dados Iris. No entanto, a escolha entre esses algoritmos pode depender das características específicas do problema e das preferências do usuário.

Referências

- [1] *Iris Dataset*. <https://archive.ics.uci.edu/ml/datasets/Iris>. [Accessed: 2024-04-04].
- [2] *Vertebral Column Dataset*. <https://archive.ics.uci.edu/ml/datasets/Vertebral+Column>. [Accessed: 2024-04-04].