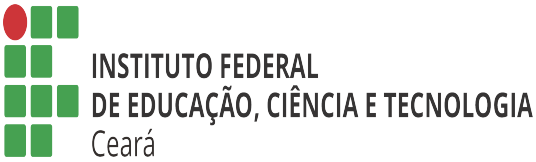
Instituto Federal de Educação, Ciência e Tecnologia do Ceará



Reconhecimento de Padrões

Relatório do Trabalho 3 - Classificador Naive Bayes

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

Fortaleza, 23 de Abril de 2024

Conteúdo

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 os quatro algoritmos de classificação. Esses algoritmos foram aplicados aos conjunto de dados Iris[3], column3C [4], artificial2, Breast Cancer[1]; Dermatology [2] 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

Utilizamos para estudo e implementação, o Classificador Naive Bayes(CNB) para analisar os conjuntos de dados que serão apresentados ao longo desse relatório. Ao final detalharemos a sua respectiva analise comparativa com os modelos do trabalho anterior KNN , DMC , CBGM e a conclusão.

2.1 CNB (Classificador Naive Bayes)

O Classificador Naive Bayes (CNB) é um modelo de aprendizado de máquina supervisionado baseado no teorema de Bayes com a suposição "ingênua"de independência condicional entre os atributos. Essa suposição simplifica o modelo, permitindo que ele seja treinado e usado de forma eficiente, especialmente em conjuntos de dados com muitos atributos.

2.2 Funcionamento

- **Teorema de Bayes**: O CNB é fundamentado no teorema de Bayes, que descreve a probabilidade condicional de um evento ocorrer, dado o conhecimento prévio de outros eventos relacionados. Matematicamente, o teorema de Bayes é expresso como:

$$P(C_k|x_1, x_2, ..., x_n) = \frac{P(x_1, x_2, ..., x_n|C_k) \cdot P(C_k)}{P(x_1, x_2, ..., x_n)}$$

- **Suposição de Independência Condicional**: O "ingênuo"na Classificador Naive Bayes refere-se à suposição de que os atributos são independentes entre si, dado o valor da classe. Isso significa que a presença

ou ausência de um atributo não está relacionada à presença ou ausência de outro atributo, uma vez que a classe é conhecida.

$$P(x_i|C_k, x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_n) = P(x_i|C_k)$$

- **Treinamento**: Durante a fase de treinamento, o CNB estima as probabilidades a priori das classes $P(C_k)$ e as probabilidades condicionais dos atributos para cada classe $P(x_i|C_k)$. Isso é feito contando a frequência de ocorrência de cada classe e de cada valor de atributo para cada classe nos dados de treinamento.

- **Classificação**: Para classificar uma nova instância x, o CNB calcula a probabilidade posterior de cada classe dada a instância usando o teorema de Bayes. Como o denominador da fórmula de Bayes é constante para todas as classes, pode ser ignorado na comparação das probabilidades posteriores. A classe predita é aquela com a maior probabilidade posterior.

$$P(C_k|x_1, x_2, ..., x_n) \propto P(C_k) \cdot \prod_{i=1}^{n} P(x_i|C_k)$$

O Classificador Naive Bayes é um modelo simples e eficiente que funciona bem em muitos problemas de classificação, especialmente quando a suposição de independência condicional entre os atributos é razoável. Ele é amplamente utilizado em aplicações práticas, como classificação de documentos de texto, detecção de spam e diagnóstico médico.

3 Procedimento Experimental

Esse experimento será realizado com o algoritmo estudado CNB, nos conjuntos de dados dividindo-os em treinamento e teste, com uma proporção de 80% para treinamento e 20% para teste, e em seguida feito um holdout com 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), sempre em seu melhor valor , que será analisado e exibido durante a execução, outros aspectos como a acurácia e o desvio padrão serão salvos para análise posterior.

3.1 Classificador Naive Bayes

3.2 CNB aplicado ao dataset IRIS Flower

Após execução do algoritmo CNB no conjunto de dados da IRIS [3], os seguintes resultados foram obtidos:

- Melhor acurácia ao utilizar randon_state 42, acontece na realização 1 em 20 realizações.
- Acurácia média: 1,00%
- Desvio padrão da acurácia: 0.01
- Matriz de Confusão 1 escolhida por exibir a melhor acurácia:

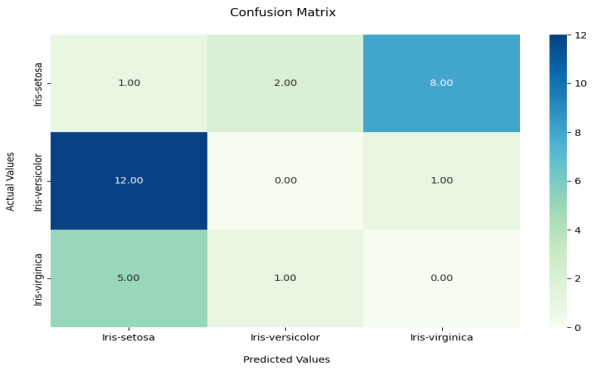


Figura 1: Matriz de Confusão IRIS

- Superfície de decisão:

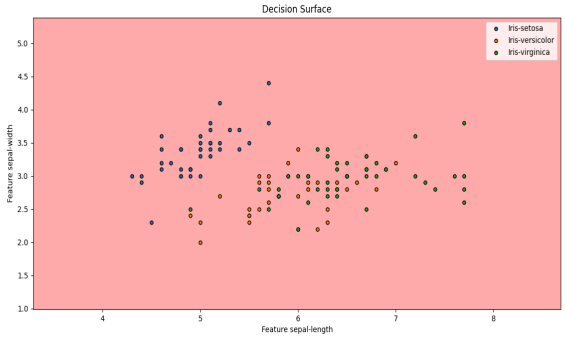


Figura 2: Superfície de decisão IRIS

- Gaussianas:

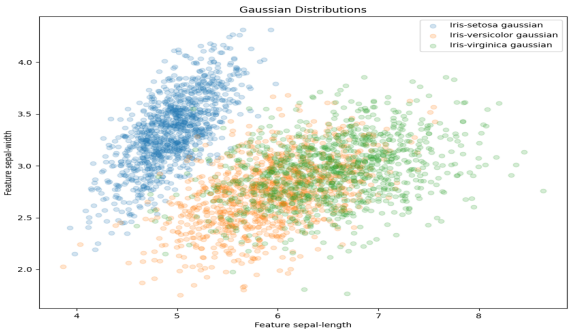


Figura 3: Gaussianas IRIS

- Plot do conjunto de dados de treinamento e teste:

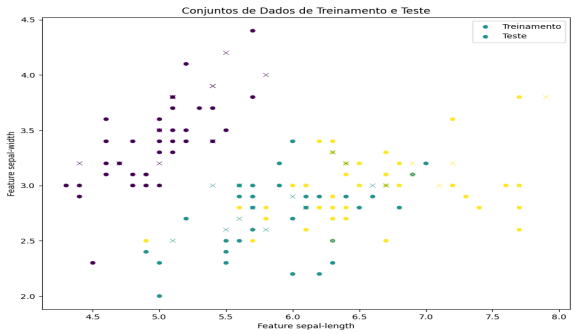


Figura 4: Conjunto de dados IRIS

3.3 CNB aplicado ao Artificial 2

Após execução do algoritmo CNB no conjunto de dados da Artificial2, os seguintes resultados foram obtidos:

- Melhor acurácia ao utilizar randon_state 42, acontece na realização 6 em 20 realizações.
- Acurácia média: 12%
- Desvio padrão da acurácia: 0.12
- Matriz de Confusão 6 escolhida por exibir a melhor acurácia:

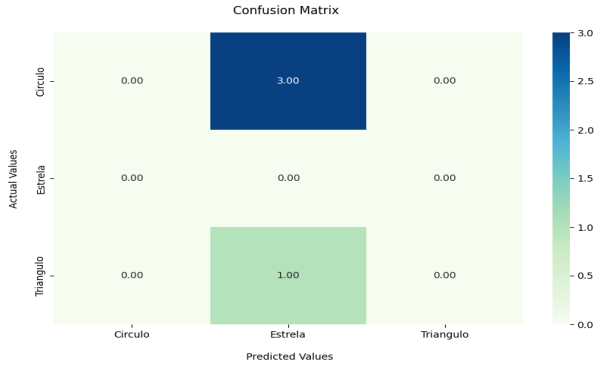


Figura 5: Matriz de Confusão Artificial II

- Superfície de decisão:

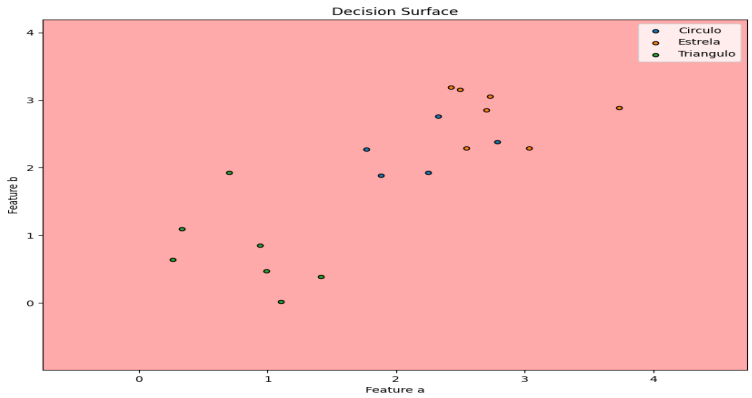


Figura 6: Superfície de decisão Artificial II

- Gaussianas:

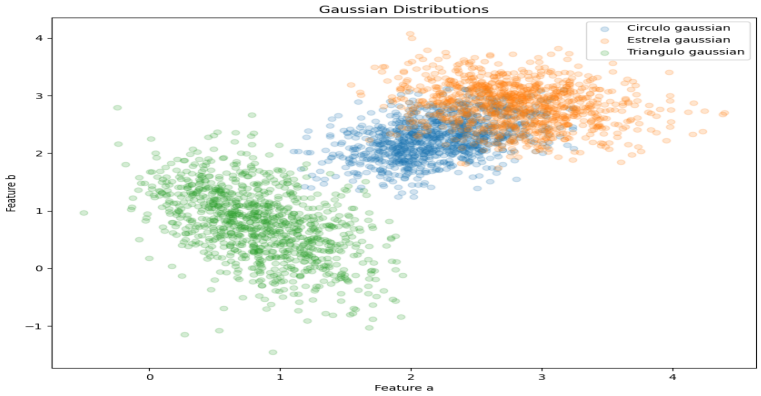


Figura 7: Gaussianas Artificial II

- Conjunto de dados de treinamento e teste:

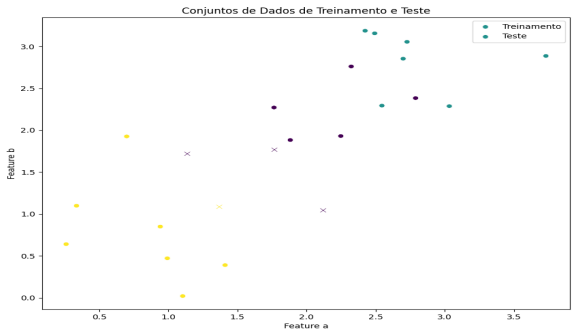


Figura 8: Conjunto de dados Artificial II

3.4 CNB aplicado ao dataset Coluna3C

Após execução do algoritmo CNB no conjunto de dados da Coluna [4], os seguintes resultados foram obtidos:

- Melhor acurácia ao utilizar randon_state 42, acontece na realização 13 de 20 realizações.
- Acurácia média: 42,00%
- Desvio padrão da acurácia: 0.06
- Matriz de Confusão da realizacao 13:

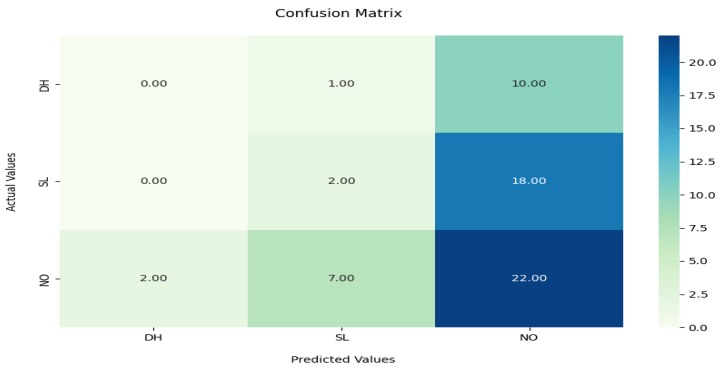


Figura 9: Matriz de Confusão Coluna

- Superfície de decisão:

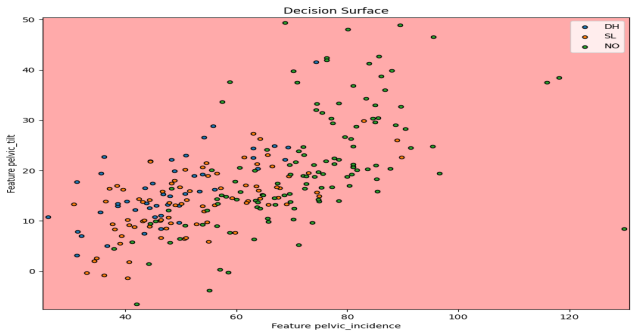


Figura 10: Superfície de decisão Coluna

- Gaussianas:

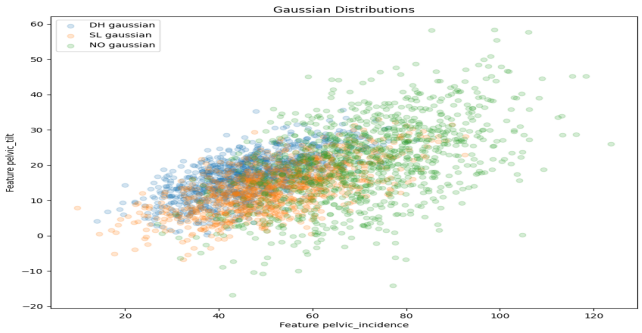


Figura 11: Gaussianas Coluna3C

- Conjunto de dados de treinamento e teste:

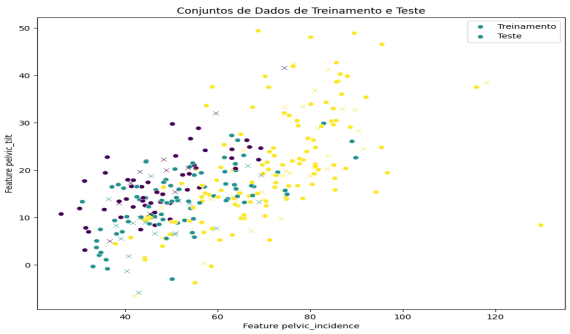


Figura 12: Conjunto de dados Coluna3C

3.5 CNB aplicado ao dataset Breast-Câncer

Após execução do algoritmo CNB no conjunto de dados da Breast-Câncer [1], os seguintes resultados foram obtidos:

- Melhor acurácia ao utilizar randon_state 42, acontece na realização 19
- Acurácia média: 70%
- Desvio padrão da acurácia: 0.05
- Matriz de Confusão da realizacao 19:

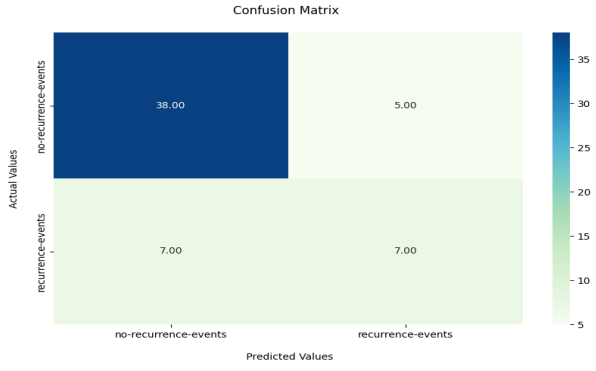


Figura 13: Matriz de Confusão Breast-Cancer

- Superfície de decisão:

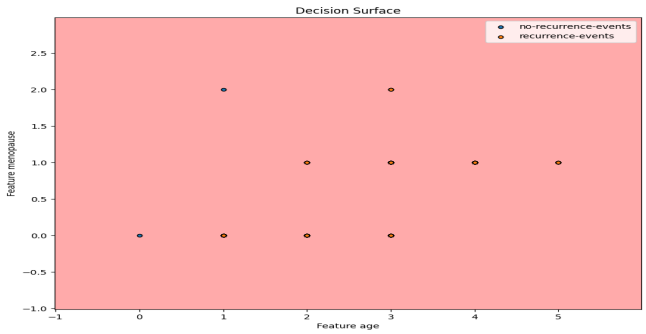


Figura 14: Superfície de decisão Breast-Cancer

- Gaussianas:

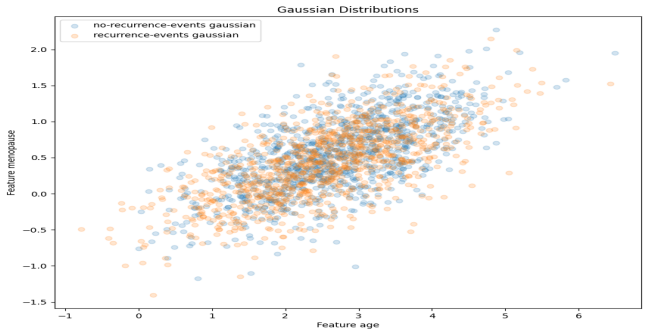


Figura 15: Gaussianas Breast-Cancer

- Conjunto de dados de treinamento e teste:

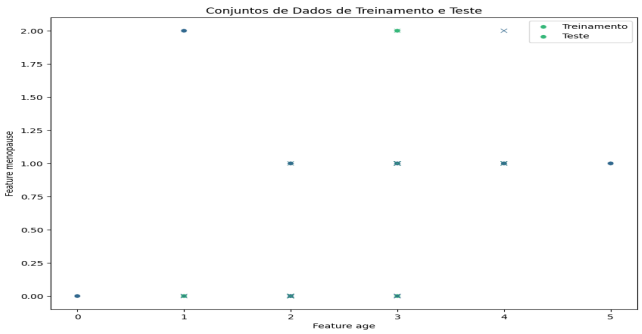


Figura 16: Conjunto de dados Breast-Cancer

3.6 CNB aplicado ao dataset Dermatology

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

- Melhor acurácia ao utilizar randon_state 42, acontece na realização 7 de 20 realizações.
- Acurácia média: 23%
- Desvio padrão da acurácia: 0.06
- Matriz de Confusão da realizacao 7:

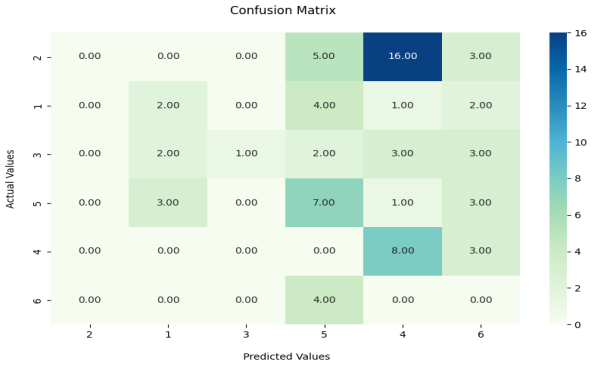


Figura 17: Matriz de Confusão Dermatology

- Superfície de decisão:

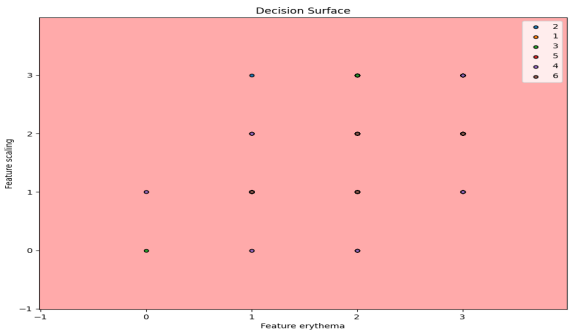


Figura 18: Superfície de decisão Dermatology

- Gaussianas:

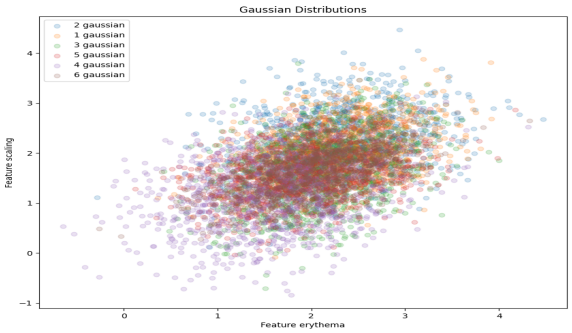


Figura 19: Gaussianas Dermatology

- Conjunto de dados de treinamento e teste:

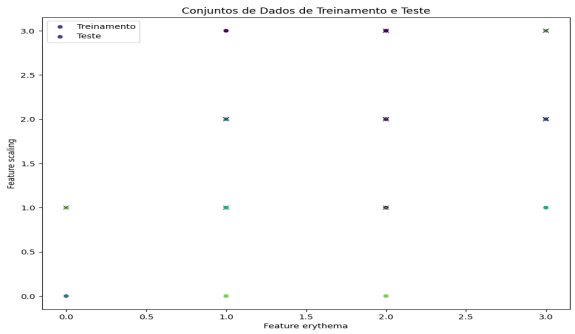


Figura 20: Conjunto de dados Dermatology

4 Análise Comparativa

Para comparar os modelos KNN, DMC, CBGM e CNB em diferentes conjuntos de dados, podemos analisar a acurácia média e o desvio padrão da acurácia para cada modelo em cada conjunto de dados. Também podemos examinar as matrizes de confusão para uma realização específica de cada modelo, que nos dá informações sobre como cada modelo está performando em termos de classificação de diferentes classes.

Aqui está uma tabela resumindo os resultados:

Tabela 1: Comparação de Modelos de Classificação

Conjunto de Dados	Modelo	Acurácia Média	Desvio Padrão da Acurácia
Iris	KNN	96.83%	0.0324
Iris	DMC	93.66%	0.0433
Iris	CBGM	97.33%	0.0133
Iris	CNB	1.00%	0.01
Artificial I	KNN	93.75%	0.0740
Artificial I	DMC	74.37%	0.0836
Artificial II	CBGM	70.00%	0.1871
Artificial II	CNB	12.00%	0.12
Coluna3C	KNN	84.19%	0.0333
Coluna3C	DMC	75.72%	0.0662
Coluna3C	CBGM	83.87%	0.0456
Coluna3C	CNB	42.00%	0.06
Cancer	CBGM	69.47%	0.04520
Câncer	CNB	70.00%	0.05
Dermatológico	CBGM	89.59%	0.0320
Dermatology	CNB	23.00%	0.06

5 Conclusão

Com base nos resultados apresentados na tabela de comparação de modelos de classificação, podemos tirar várias conclusões:

1. **Desempenho Variado dos Modelos**: Os diferentes modelos (KNN, DMC, CBGM e CNB) apresentam desempenhos variados em diferentes conjuntos de dados. Por exemplo, o CBGM alcança altas taxas de acurácia média em conjuntos de dados como Iris e Coluna3C, enquanto o CNB tem um desempenho muito baixo no conjunto de dados IRIS. Os modelos tiveram desempenhos variados nos conjuntos de dados artificiais. Por exemplo, o KNN teve um desempenho muito bom no Artificial I, enquanto o CNB teve um desempenho muito baixo no Artificial II. Tanto o CBGM quanto o CNB apresentaram desempenhos competitivos no conjunto de dados Coluna3C, com acurácias médias de 83.87% e 42.00%, respectivamente. Tanto o CBGM quanto o CNB tiveram desempenhos semelhantes no conjunto de dados Breast-Câncer, com acurácias médias de cerca de 70%. O CNB teve um desempenho muito inferior no conjunto de dados Dermatology, com uma acurácia média de apenas 23%.

2. **Sensibilidade ao Conjunto de Dados**: O desempenho de cada modelo varia de acordo com o conjunto de dados. Por exemplo, o CNB tem um desempenho muito alto no conjunto de dados Iris, mas um desempenho muito baixo no conjunto de dados Artificial II.

3. **Influência do Algoritmo**: O algoritmo de classificação escolhido (KNN, DMC, CBGM ou CNB) tem um impacto significativo no desempenho do modelo em cada conjunto de dados. Alguns algoritmos podem ser mais adequados para conjuntos de dados específicos do que outros.

4. **Importância da Avaliação**: É importante avaliar o desempenho de diferentes modelos em vários conjuntos de dados para determinar qual é o mais adequado para uma tarefa específica de classificação. A acurácia média e o desvio padrão da acurácia são métricas úteis para comparar o desempenho dos modelos.

5. **Necessidade de Ajuste de Parâmetros**: Em alguns casos, pode ser necessário ajustar os parâmetros do modelo para melhorar seu desempenho em um conjunto de dados específico. Isso pode ser especialmente verdadeiro para algoritmos mais sensíveis a parâmetros, como o KNN.

Em resumo, a escolha do modelo de classificação mais adequado depende de vários fatores, incluindo o conjunto de dados, as características dos dados e os objetivos da tarefa de classificação. É importante realizar uma avaliação abrangente dos diferentes modelos para tomar uma decisão informada sobre qual modelo usar em uma aplicação específica.

Referências

[1] *Breast Cancer*. https://archive.ics.uci.edu/dataset/14/breast+cancer. [Accessed: 2024-04-20].

[2] *Dermatology*. https://archive.ics.uci.edu/dataset/33/dermatology. [Accessed: 2024-04-13].

[3] *Iris Dataset*. https://archive.ics.uci.edu/ml/datasets/Iris. [Accessed: 2024-04-04].

[4] *Vertebral Column Dataset*. https://archive.ics.uci.edu/ml/datasets/Vertebral+Column. [Accessed: 2024-04-04].