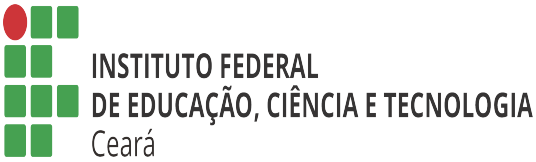
Instituto Federal de Educação, Ciência e Tecnologia do Ceará



Reconhecimento de Padrões

Relatório do Trabalho 2 - Classificador Bayesiano Gaussiano Multivariado

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

Fortaleza, 16 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 três 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 Bayesiano Gaussiano Multivariado(CBGM) 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 e a conclusão.

2.1 CBGM (Classificador Bayesiano Gaussiano Multivariado)

O Classificador Bayesiano Gaussiano Multivariado (CBGM) é um modelo de classificação probabilístico baseado no teorema de Bayes. Ele assume que os atributos do conjunto de dados de entrada são distribuídos normalmente (ou gaussianamente) e estima as probabilidades condicionais de cada classe dada uma observação utilizando distribuições normais multivariadas.

2.2 Funcionamento

- **Teorema de Bayes**: O classificador é construído com base no teorema de Bayes, que descreve a relação entre a probabilidade condicional de uma classe dada uma observação e a distribuição das características dessa observação.

- **Distribuições Normais Multivariadas**: O CBGM assume que as características do conjunto de dados de entrada são distribuídas normalmente. Isso significa que as observações de cada classe são modeladas como amostras de uma distribuição normal multivariada, onde cada dimensão representa um atributo.

- **Estimação de Parâmetros**: Para cada classe, o CBGM estima os parâmetros de sua distribuição normal multivariada, incluindo a média e a matriz de covariância dos atributos. Isso é feito a partir dos dados de treinamento.

- **Classificação**: Dada uma nova observação, o classificador calcula a probabilidade de pertencer a cada classe com base nas distribuições normais multivariadas estimadas e no teorema de Bayes. A classe mais provável é então atribuída à observação.

O Classificador Bayesiano Gaussiano Multivariado CBGM é uma extensão do Classificador Bayesiano Gaussiano (CBG), que assume que as características são independentes entre si. Ao considerar as correlações entre os atributos, o CBGM pode capturar informações mais complexas sobre os dados e, muitas vezes, apresenta melhor desempenho em conjuntos de dados onde as características estão correlacionadas.

3 Procedimento Experimental

Esse experimento será realizado com o algoritmo estudado CGBM, 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 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), 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 Bayesiano Gaussiano Multivariado

3.2 CBGM aplicado na IRIS Flower

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

- Melhor acurácia ao utilizar randon_state 42, acontece na realização 5 em 20 realizações.

- Acurácia média: 97,33%

- Desvio padrão da acurácia: 0.0133333333333

- Matriz de Confusão 5 escolhida por exibir a melhor acurácia:

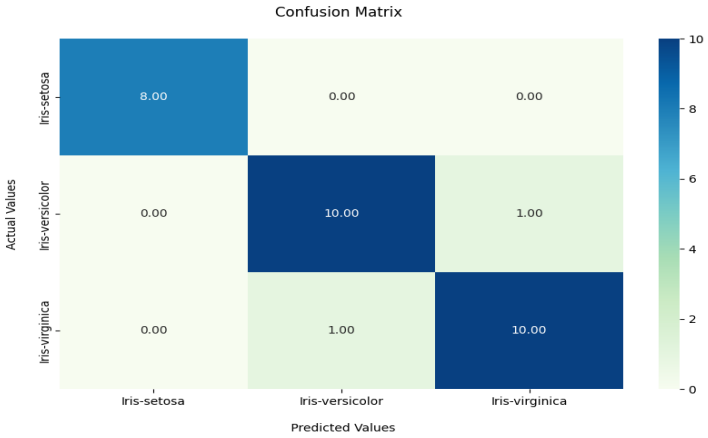


Figura 1: Matriz de Confusão IRIS

- Superfície de decisão:

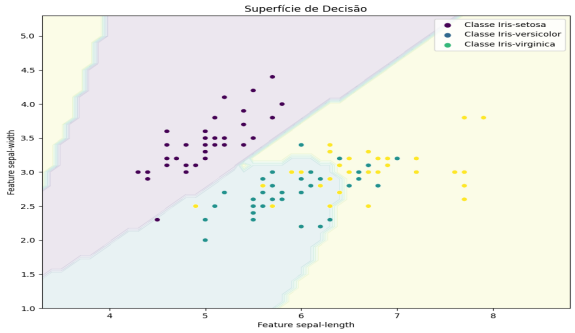


Figura 2: Superfície de decisão IRIS

- Gaussianas:

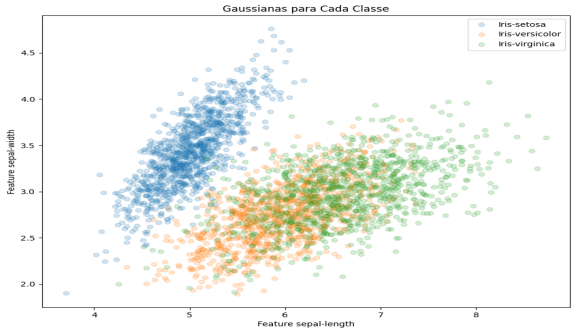


Figura 3: Gaussianas IRIS

- Conjunto de dados de treinamento e teste:

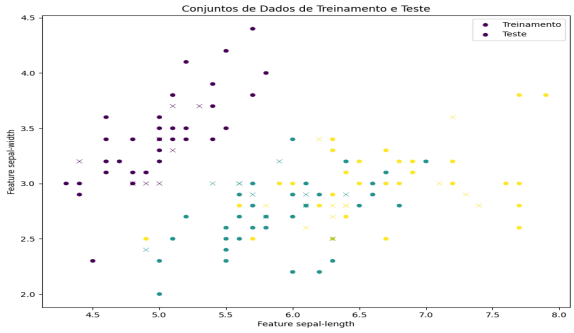


Figura 4: Conjunto de dados IRIS

3.3 CBGM aplicado ao Artificial 2

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

- Melhor acurácia ao utilizar randon_state 42, acontece na realização 3 em 20 realizações.
- Acurácia média: 70%
- Desvio padrão da acurácia: 0.1870828693
- Matriz de Confusão 3 escolhida por exibir a melhor acurácia:

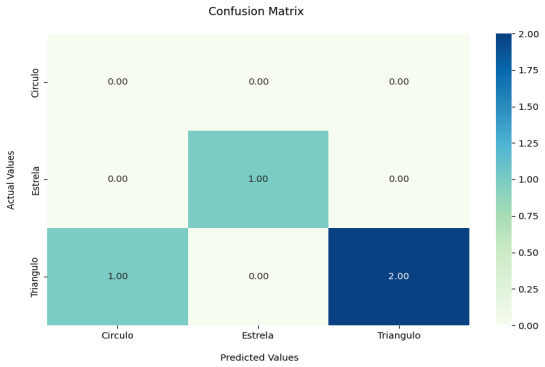


Figura 5: Matriz de Confusão Artificial II

- Superfície de decisão:

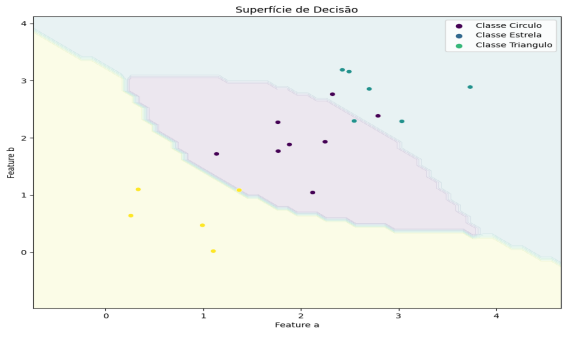


Figura 6: Superfície de decisão Artificial II

- Gaussianas:

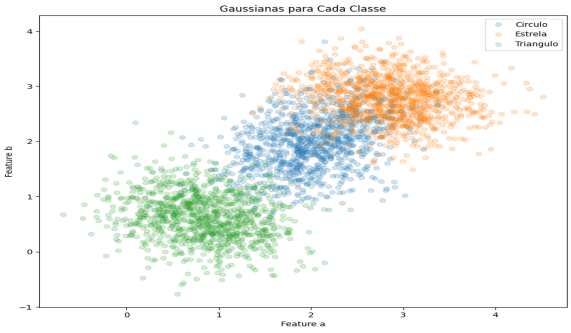


Figura 7: Gaussianas Artificial II

- Conjunto de dados de treinamento e teste:

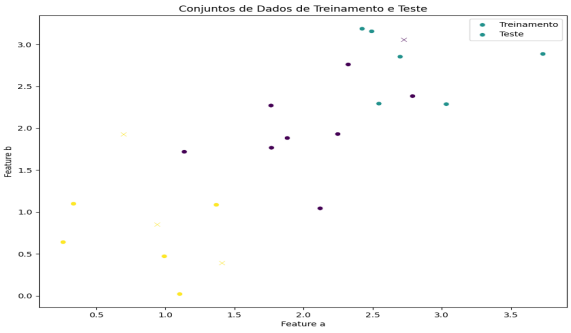


Figura 8: Conjunto de dados Artificial II

3.4 CBGM aplicado ao dataset Coluna3C

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

- Melhor acurácia ao utilizar randon_state 42, acontece na realização 5 de 20 realizações.
- Acurácia média: 83,87%
- Desvio padrão da acurácia: 0.045619792
- Matriz de Confusão da realizacao 5:

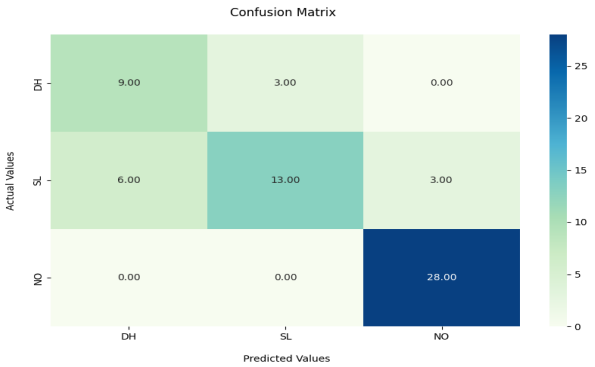


Figura 9: Matriz de Confusão Coluna

- Superfície de decisão:

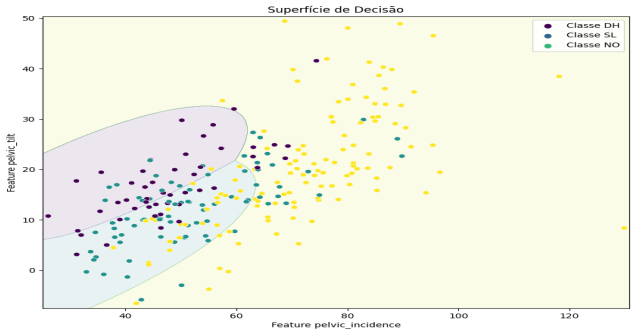


Figura 10: Superfície de decisão Coluna

- Gaussianas:

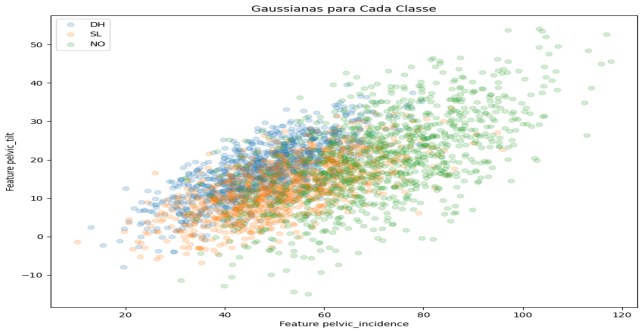


Figura 11: Gaussianas Coluna3C

- Conjunto de dados de treinamento e teste:

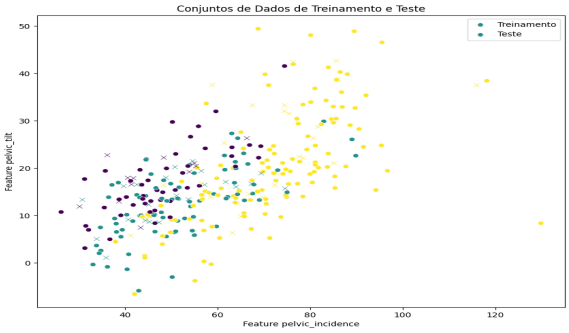


Figura 12: Conjunto de dados Coluna3C

3.5 CBGM aplicado ao dataset Breast-Câncer

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

- Melhor acurácia ao utilizar randon_state 42, acontece na realização 4
- Acurácia média: 95,75%
- Desvio padrão da acurácia: 0.0117402647
- Matriz de Confusão da realizacao 4:

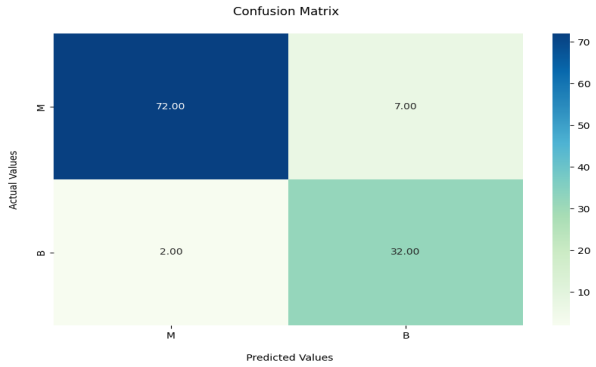


Figura 13: Matriz de Confusão Breast-Cancer

- Superfície de decisão:

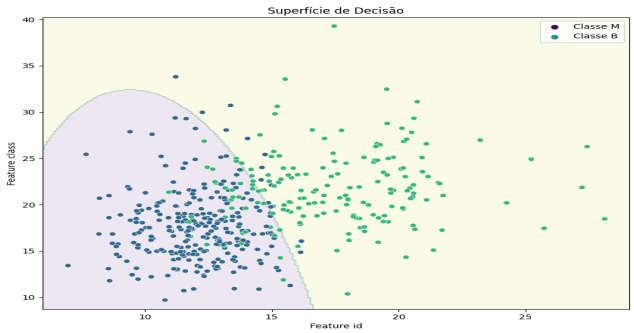


Figura 14: Superfície de decisão Breast-Cancer

- Gaussianas:

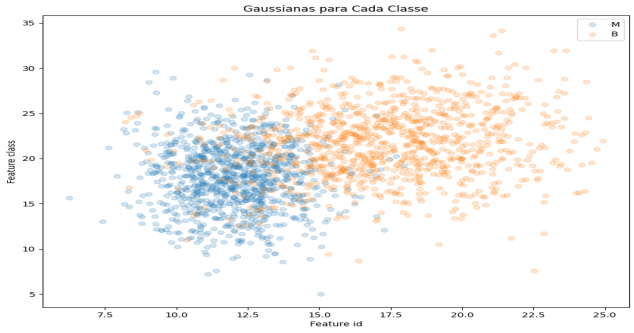


Figura 15: Gaussianas Breast-Cancer

- Conjunto de dados de treinamento e teste:

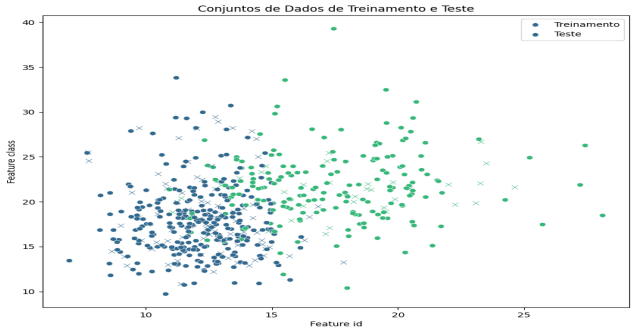


Figura 16: Conjunto de dados Breast-Cancer

3.6 CBGM aplicado ao dataset Dermatology

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

- Melhor acurácia ao utilizar randon_state 42, acontece na realização 1 de 20 realizações.
- Acurácia média: 89,58%
- Desvio padrão da acurácia: 0.031950421
- Matriz de Confusão da realizacao 1:

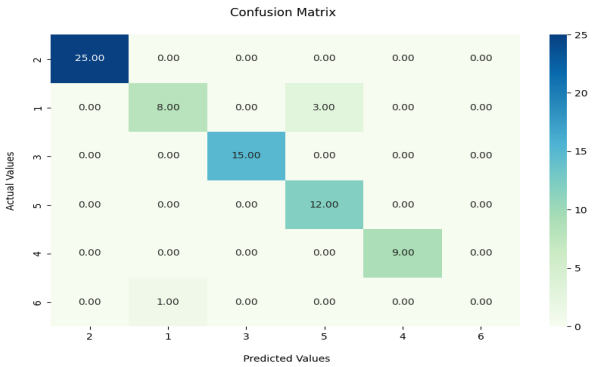


Figura 17: Matriz de Confusão Dermatology

- Superfície de decisão:

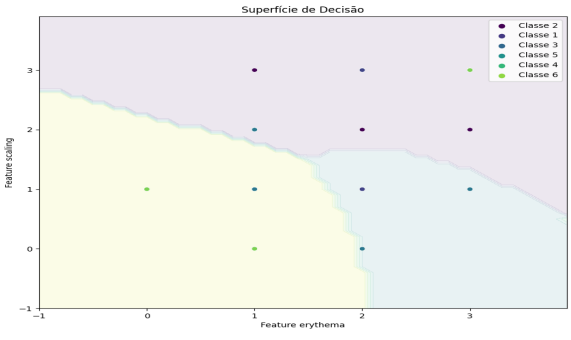


Figura 18: Superfície de decisão Dermatology

- Gaussianas:

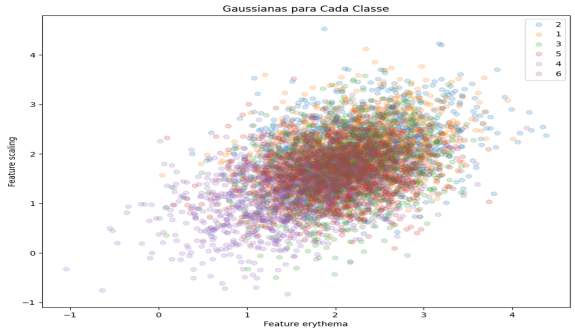


Figura 19: Gaussianas Dermatology

- Conjunto de dados de treinamento e teste:

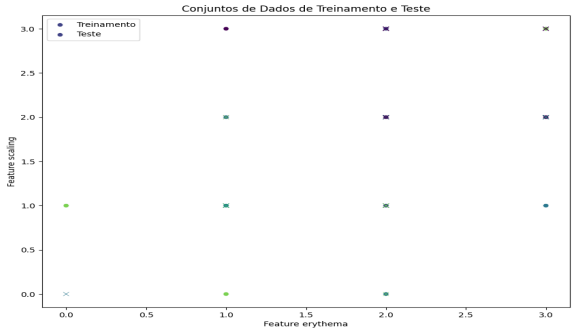


Figura 20: Conjunto de dados Dermatology

4 Análise Comparativa

Para comparar os modelos KNN, DMC e CBGM 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
Artificial I	KNN	93.75%	0.0740
Artificial I	DMC	74.37%	0.0836
Artificial II	CBGM	70.00%	0.1871
Coluna3C	KNN	84.19%	0.0333
Coluna3C	DMC	75.72%	0.0662
Coluna3C	CBGM	83.87%	0.0456
Cancer	CBGM	95.75%	0.0117
Dermatológico	CBGM	89.59%	0.0320

5 Conclusão

Com base nesses resultados, podemos observar que, em geral, o modelo CBGM apresentou uma performance melhor ou comparável em relação aos modelos KNN e DMC na maioria dos conjuntos de dados. No entanto, é importante considerar outros aspectos, como tempo de treinamento e interpretabilidade do modelo, ao fazer uma escolha entre os diferentes algoritmos. Além disso, a escolha do modelo ideal pode variar dependendo das características específicas do conjunto de dados e dos objetivos do problema.

Referências

[1] *Breast Cancer Wisconsin (Diagnostic)*. https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic. [Accessed: 2024-04-13].

[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].