

# Algoritmos de agrupamento: K-Means, Fuzzy C-Means e Gaussian Mixture Model (GMM)

ATIVIDADE 3 – AGRUPAMENTOS

DISCIPLINA: INTELIGÊNCIA COMPUTACIONAL

PROFESSOR ALISSON MARQUES DA SILVA

MATHEUS COELHO E RODRIGO SILVA

# Introdução

- ▶ Este trabalho investiga a aplicação de três algoritmos de agrupamento amplamente utilizados - K-Means, Fuzzy C- Means (FCM) e Gaussian Mixture Models (GMM) - em uma abordagem de classificação supervisionada. O objetivo principal é avaliar a eficácia desta metodologia em diferentes tipos de conjuntos de dados, identificando cenários onde a técnica se mostra mais adequada.

# Metodologia

- ▶ A metodologia consiste em utilizar algoritmos de agrupamento para fins de classificação supervisionada. Para cada classe presente nos conjuntos de dados, foi treinado um modelo de agrupamento, com o número de clusters ( $k$ ) otimizado pelo método do Cotovelo. O processo de classificação ocorre associando cada amostra de teste ao cluster mais próximo de cada modelo de classe, atribuindo a classe do modelo cujo cluster apresentar maior similaridade. Antes do treinamento, todos os atributos foram normalizados utilizando o método StandardScaler, e os rótulos das classes foram codificados numericamente via LabelEncoder, garantindo compatibilidade com os algoritmos de agrupamento.

# Por que utilizar o método do cotovelo?

- ▶ No nosso trabalho, a abordagem é de agrupamento supervisionado. Nós já sabemos quais são as classes (0 e 1 para o Adult; 0 a 6 para o Dry Bean). O que não sabemos é a complexidade interna de cada uma dessas classes.
- ▶ Pense assim: A classe "Renda > 50K" é um grupo homogêneo e compacto, que pode ser bem representado por um único centro (cluster)? Ou será que essa classe é mais complexa, formada por subgrupos distintos (ex: um subgrupo de jovens com alta escolaridade e outro de pessoas mais velhas com muito tempo de carreira)?
- ▶ É para responder a essa pergunta que usamos o método do cotovelo.


# Algoritmos de classificação utilizados

- ▶ K-Means: O algoritmo K-Means particiona os dados em  $k$  clusters, minimizando a soma das distâncias quadráticas entre os pontos e os centróides dos clusters. O algoritmo é iterativo e converge para um mínimo local da função objetivo.
- ▶ Fuzzy C-Means (FCM): O FCM é uma extensão fuzzy do K-Means, onde cada ponto pode pertencer a múltiplos clusters com diferentes graus de pertinência. Isso permite uma modelagem mais flexível de dados com sobreposição entre classes.
- ▶ Gaussian Mixture Model (GMM): O GMM modela os dados como uma mistura de distribuições gaussianas, sendo capaz de capturar formas mais complexas de clusters. Utiliza o algoritmo Expectation-Maximization (EM) para estimar os parâmetros.

# Resultados e discussão

- ▶ A. Resultados no Dataset Adult: No conjunto “Adult”, todos os algoritmos apresentaram desempenho limitado, com acurácia média próxima de 50%. Isso indica que as classes presentes nesse dataset possuem alta sobreposição, dificultando a separação por métodos baseados em agrupamento.
- ▶ Em contraste, no conjunto “Dry Bean”, a abordagem de agrupamento supervisionado mostrou-se altamente eficaz. O algoritmo GMM obteve a melhor acurácia média (91,32%), superando K-Means (89,75%) e FCM (89,81%).



- 
- ▶ Número de Clusters Usado por Classe (na última repetição - Adult):
    - ▶ - Classe '0.0': 6 clusters
    - ▶ - Classe '1.0': 4 clusters
  - ▶ Os dados dentro de cada classe são muito heterogêneos e complexos.

- ▶ Número de Clusters Usado por Classe (na última repetição - DryBean):
  - ▶ - Classe '0.0': 3 clusters
  - ▶ - Classe '1.0': 3 clusters
  - ▶ - Classe '2.0': 3 clusters
  - ▶ - Classe '3.0': 4 clusters
  - ▶ - Classe '4.0': 3 clusters
  - ▶ - Classe '5.0': 4 clusters
  - ▶ - Classe '6.0': 4 clusters
- ▶ Dados muito mais homogêneos e bem definidos

Matriz de Confusão - GMM (Dataset Adult - Acurácia 52,13%)

Classe Verdadeira

Renda <= 50K

Renda > 50K

Renda <= 50K

Renda > 50K

Classe Prevista

2949

3929

897

1274

3500

3000

2500

2000

1500

1000

DP da acurácia =  
0.1209



Matriz de Confusão - FuzzyCMeans (Dataset Adult - Acurácia: 49,42%)

Classe Verdadeira



Renda <= 50K

Renda > 50K

Classe Prevista

DP da acurácia =  
0.0221

Matriz de Confusão - KMeans (Dataset Adult - Acurácia: 52,55%)

Classe Verdadeira

Renda <= 50K

Renda > 50K

Renda <= 50K

Renda > 50K

Classe Prevista

5440

1438

1709

462

5000

4000

3000

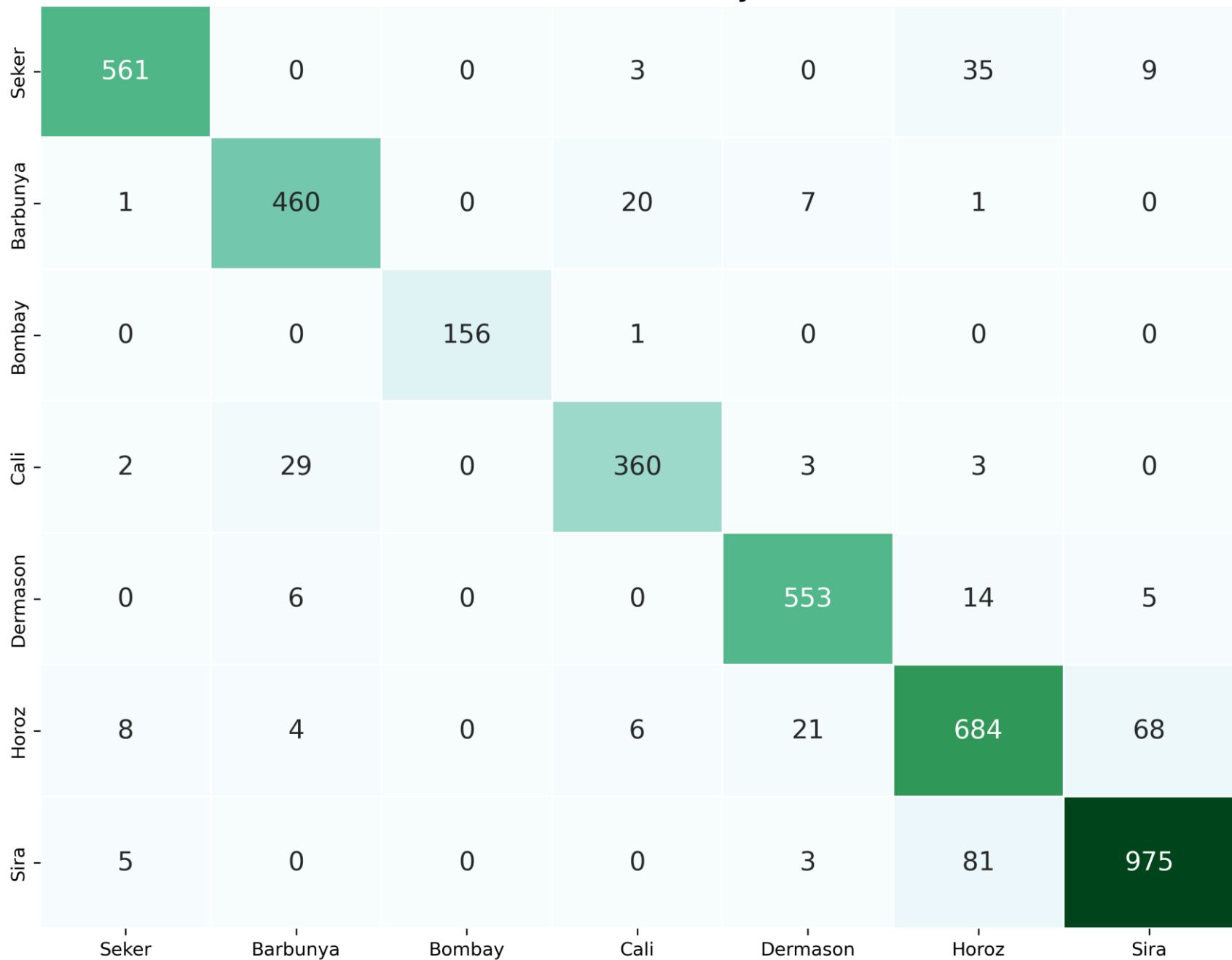
2000

1000

DP da acurácia =  
0.0996

Matriz de Confusão - GMM (Dataset Dry Bean - Acurácia 91,32%)

Classe Verdadeira



800

600

400

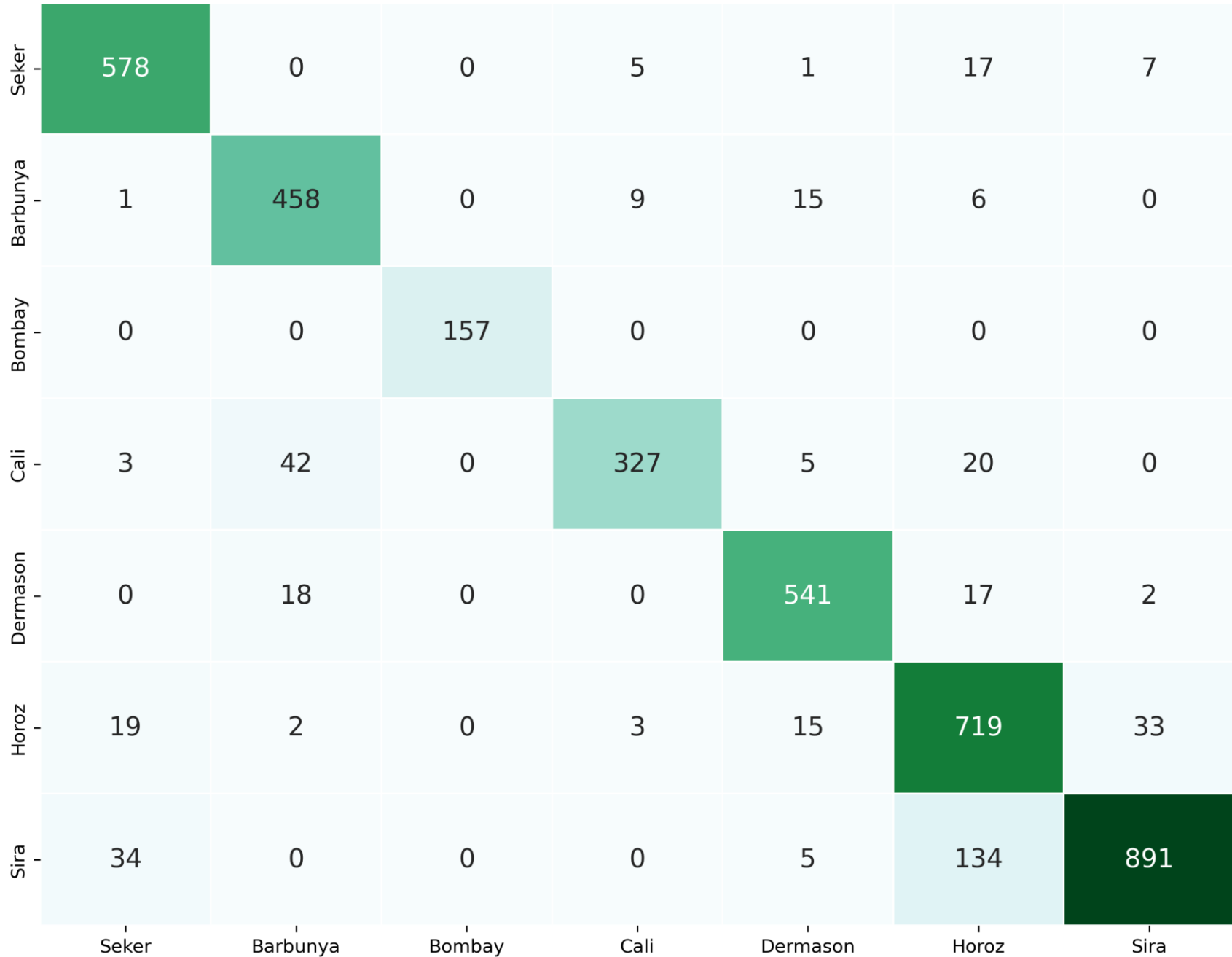
200

0

DP da acurácia  
= 0.0040

Matriz de Confusão - FuzzyCMeans (Dataset Dry Bean - Acurácia 89,81%)

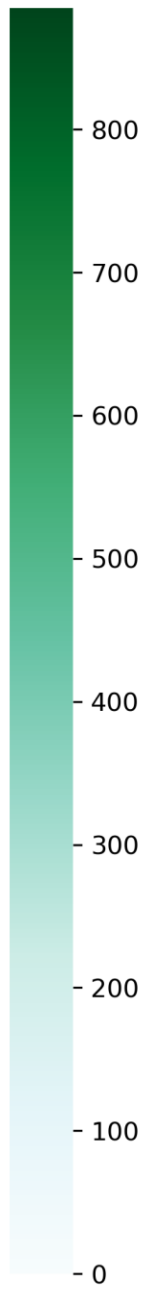
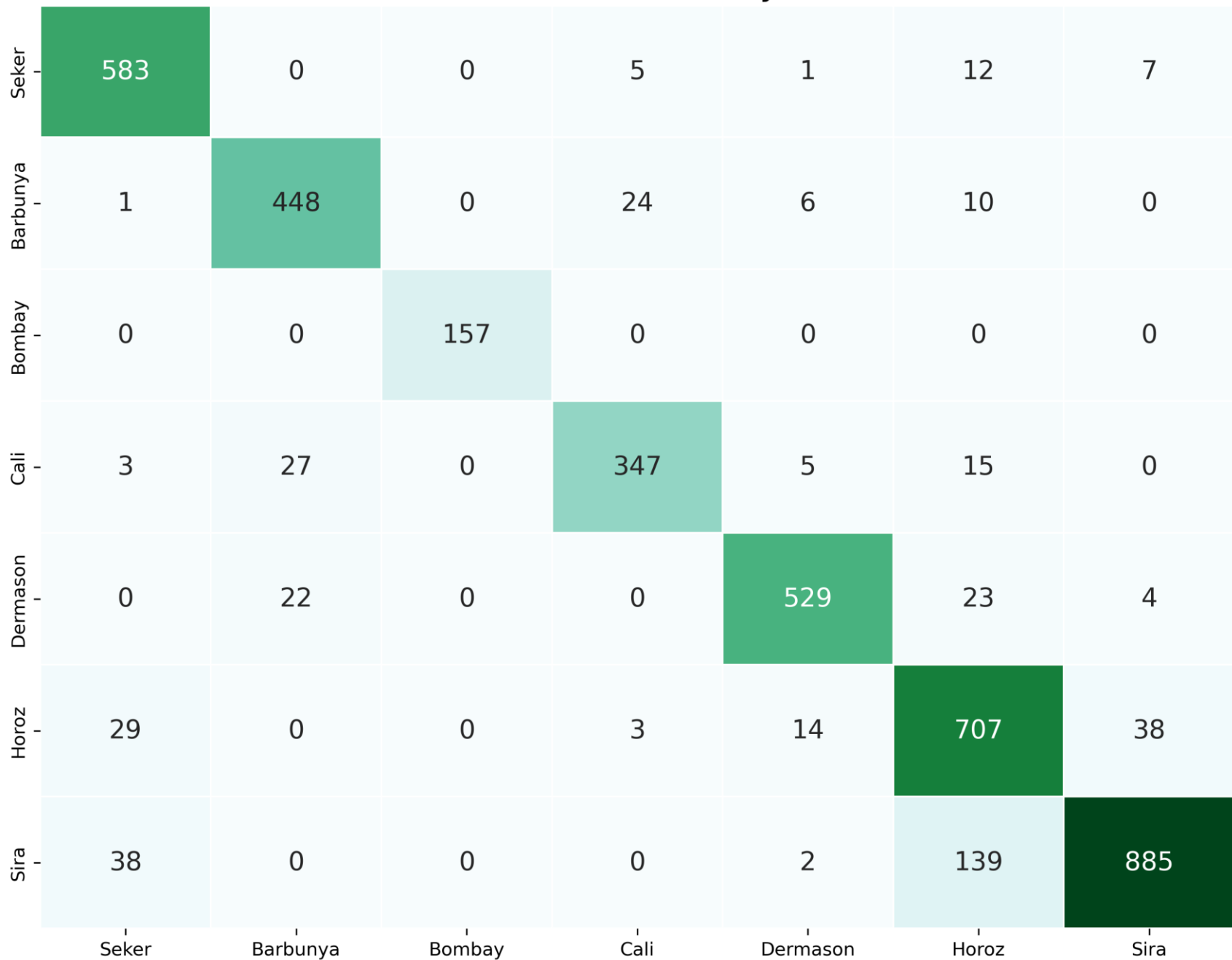
Classe Verdadeira



DP da acurácia =  
0.0046

Matriz de Confusão - KMeans (Dataset Dry Bean - Acurácia 89,75%)

Classe Verdadeira



DP da acurácia  
= 0.0042

# Referências

- ▶ [1] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297, 1967.
- ▶ [2] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press, 1981.
- ▶ [3] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” Journal of the Royal Statistical Society, vol. 39, no. 1, pp. 1–38, 1977.
- ▶ [4] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” Journal of Computational and Applied Mathematics, vol. 20, pp. 53–65, 1987.