Aprendizado Supervisionado e Não-Supervisionado Aplicado em Dados de Aerogeradores, Sinais EMG e Reconhecimento Facial

Pedro Lucas Farias de Melo

I. INTRODUÇÃO

O aprendizado de máquina é uma área essencial na inteligência artificial moderna, permitindo a construção de modelos capazes de aprender padrões a partir de dados. Entre suas categorias principais estão o aprendizado supervisionado e o não supervisionado. No aprendizado supervisionado, modelos são treinados com base em dados rotulados, permitindo realizar tarefas como regressão e classificação. Já no aprendizado não supervisionado, não há rótulos explícitos, e o objetivo é descobrir estruturas e padrões ocultos nos dados, como agrupamentos ou representações reduzidas. Este trabalho busca aplicar e avaliar técnicas representativas de cada paradigma, abordando três problemas distintos:

1) Regressão linear aplicada ao dataset de aerogeradores.
2) Classificação de sinais EMG faciais utilizando KNN.
3) Redução de dimensionalidade e clusterização aplicada a imagens faciais.

II. METODOLOGIA

A. Aprendizado Supervisionado Usando Regressão

O primeiro problema se utiliza um dataset **"aerogerador.dat"**, composto por medidas de velocidade do vento (variável independente) e potência gerada (variável dependente).

As etapas consistem em:

1) Visualização inicial dos dados através de gráfico de dispersão.
2) Organização dos dados em matrizes:
   - $X \in R^{N \times p}$ contendo a velocidade do vento.
   - $y \in R^{N \times 1}$ contendo a potência gerada.
3) Implementação de um modelo de Regressão Linear pelo método dos mínimos quadrados ordinários (biblioteca scikit-learn).
4) Validação com 500 rodadas de hold-out (80% treino, 20% teste).
5) Métricas de desempenho: MSE (Mean Squared Error) e MAE (Mean Absolute Error).
6) Cálculo estatístico final: média, desvio padrão, valor máximo e mínimo das métricas.

B. Aprendizado Supervisionado Usando Classificação

O segundo problema utiliza o dataset **"EMG-Dataset.csv"**, contendo $N = 50000$ amostras de sinais coletados por sensores faciais.

Características dos dados utilizados:

- Duas variáveis independentes:
  - Sensor 1 (Corrugador do Supercílio).
  - Sensor 2 (Zigomático Maior).
- Variável dependente: classe de expressão facial, com 5 categorias (Neutro, Sorriso, Sobrancelhas levantadas, Surpreso, Rabugento).

Etapas que foram utilizadas:

1) Organização dos dados:
   - para MQO: $X \in R^{N \times p}, y \in R^{N \times C}$;
   - para Bayesiano: $X \in R^{p \times N}, y \in R^{C \times N}$
2) Visualização inicial dos dados com gráfico de dispersão colorido pelas classes.
3) Treinamento do modelo K-Nearest Neighbors (KNN).
4) Seleção do hiperparâmetro K utilizando k-fold cross-validation com diferentes valores de K.
5) Validação com 500 rodadas aleatórias (80% treino, 20% teste).

6) Métricas de desempenho: acurácia e matriz de confusão.
7) Extração das métricas estatísticas (média, desvio padrão, máximo e mínimo da acurácia).

C. Paradigma Não-Supervisionado - Redução de Dimensionalidade

O terceiro problema proposto envolve a análise de imagens faciais, com 640 amostras de dimensões 128 x 120, totalizando 15.360 variáveis independentes por imagem. Para lidar com a alta dimensionalidade, foram aplicados três métodos. O primeiro deles foi o uso do PCA (principal Component Analysis) para lidar com projeções considerando variâncias de 90%, 80% e 75%. O segundo método foi se utilizando do algoritmo t-SNE (t-Distributed Stochastic Neighbor Embedding) para reduzir a dimensionalidade não linear se utilizando principalmente da visualização de dados de alta dimensão em um espaço de baixa dimensão que geralmente é 2D ou 3D. E o último método é se aproveitando do algoritmo de redução de dimensionalidade baseado nas técnicas de aprendizado de variedades e ideias de análise topológicas de dados, chamado UMAP onde se projetarar diferentes dimensões (3, 15, 55 e 101).

D. Paradigma Não-Supervisionado - Clusterização

Após a redução de dimensionalidade aplicando os métodos citados anteriormente, se deve aplicar os algoritmos de clusterização. Os utilizados algoritmos de clusterização foram K-means e K-medoids. Em seguida foi realizado testes com diferentes valores de K. Sua avaliação foi usando Índice de Dunn para escolher o número ideal de clusters mostrando isso em um gráfico 2D ou 3D de acordo com o número de dimensões da projeção.

III. Resultados

A. Aprendizado Supervisionado

1) Regressão Linear – Aerogerador:

- **Visualização inicial:** Gráfico de dispersão da velocidade do vento vs. potência gerada.

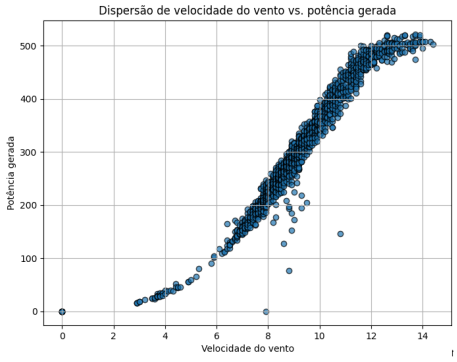

Fig. 1. Gráfico de dispersão da potência gerada pelo aerogerador em função da velocidade do vento. Cada ponto representa uma amostra do dataset "aerogerador.dat".

- **Desempenho do modelo:** Tabela de estatísticas de MSE e MAE após 500 rodadas de amostragem aleatória.

TABLE I
ESTATÍSTICAS DE DESEMPENHO DO MODELO DE REGRESSÃO LINEAR

Métrica	Média	Desvio-Padrão	Maior Valor	Menor Valor
MSE	798.302141	165.195013	1371.823766	456.260910
MAE	18.402390	0.755098	20.420843	16.266246

O modelo captura o padrão da potência em função da velocidade do vento, embora possíveis outliers ou grande variação nos erros possam afetar o desempenho.

2) Classificação – Sinais EMG Faciais (K-NN):

- **Visualização inicial:** Scatter plot 2D dos sinais EMG por classes

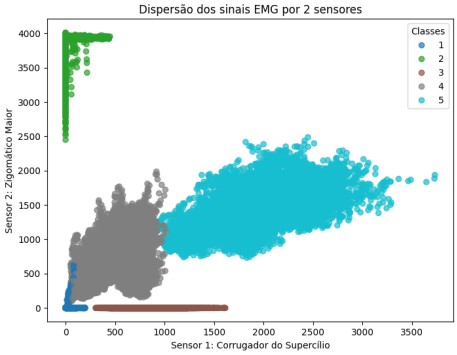

Fig. 2. Dispersão das amostras de sinais EMG para os dois sensores faciais, coloridas de acordo com as cinco classes de expressão facial.

- **Escolha do hiperparâmetro K:** Gráfico ou

tabela com acurácia média para cada valor de K.

TABLE II
ACURÁCIA MÉDIA PARA DIFERENTES VALORES DE K

K	Acurácia Média
1	0.99920
7	0.99896
11	0.99886
17	0.99866
23	0.99840
39	0.99824
101	0.99816
501	0.99782
1001	0.99622

O melhor valor de K encontrado foi 1 de acordo com a tabela acima.

- **Validação aleatória K-NN (500 rodadas) :**

TABLE III
ESTATÍSTICAS VALIDAÇÃO ALEATÓRIA K-NN DE 500 RODADAS

Métrica	Estatísticas
Média	0.999174
Desvio-padrão	0.000269
Maior valor	0.999800
Menor valor	0.998200

- Melhor caso:

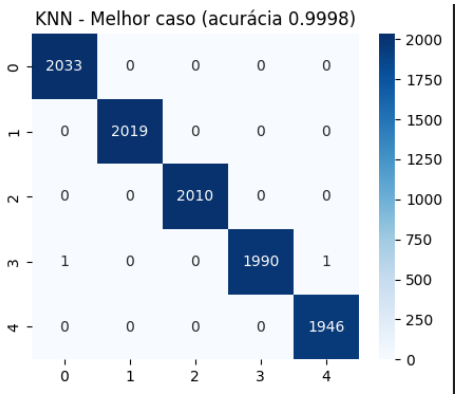

Fig. 3. Matriz de confusão correspondente ao melhor caso de validação aleatória do modelo K-NN com K=1, mostrando acurácia máxima obtida em 500 rodadas.

- Pior caso:

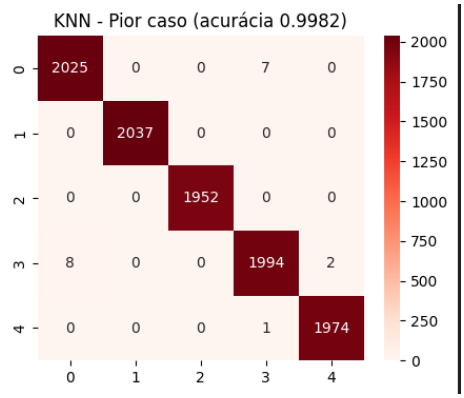

Fig. 4. Matriz de confusão correspondente ao pior caso de validação aleatória do modelo K-NN com K=1, mostrando acurácia mínima obtida em 500 rodadas.

3) Classificação Bayesiana:

- **Validação aleatória do modelo Bayesiano (500 rodadas):**

TABLE IV
ESTATÍSTICAS VALIDAÇÃO ALEATÓRIA DO MODELO BAYESIANO DE 500 RODADAS

Métrica	Estatísticas
Média	0.989005
Desvio-padrão	0.001125
Maior valor	0.992100
Menor valor	0.985700

- Melhor caso:

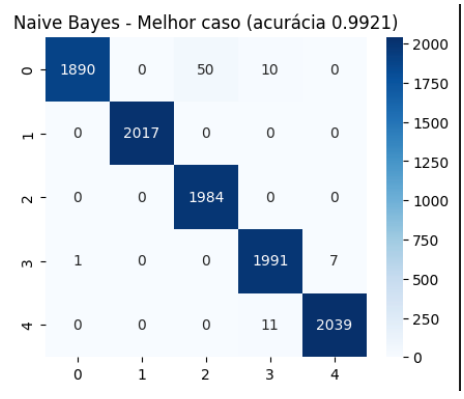

Fig. 5. Matriz de confusão do melhor caso do modelo Gaussian Naive Bayes após 500 rodadas de validação aleatória, evidenciando a acurácia máxima.

- Pior caso:

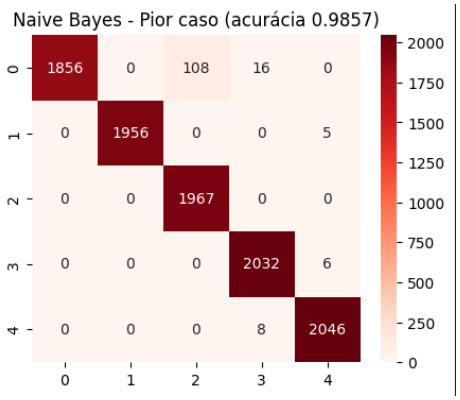

Fig. 6. Matriz de confusão do pior caso do modelo Gaussian Naive Bayes após 500 rodadas de validação aleatória, evidenciando a acurácia mínima.

B. Aprendizado Não-Supervisionado – Imagens de Rostos

1) Redução de Dimensionalidade:

- **t-SNE:** Scatter plot 2D

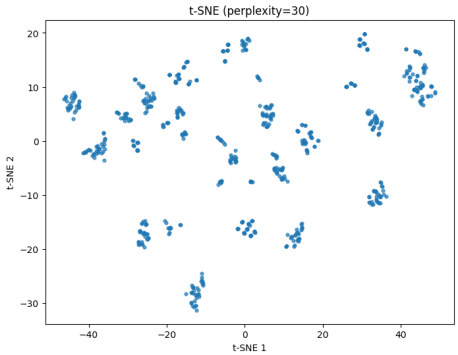

Fig. 7. Projeção 2D do conjunto de imagens faciais utilizando t-SNE, evidenciando a separação dos padrões de rostos em espaço bidimensional.

- **PCA:** Scatter plots 2D para 90%, 80% e 75% de variância
  - **90% de variância**

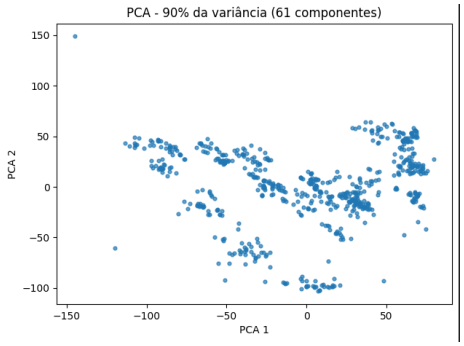

Fig. 8. Projeção das imagens faciais em 2D utilizando PCA com 90% da variância preservada.

  - **80% de variância**

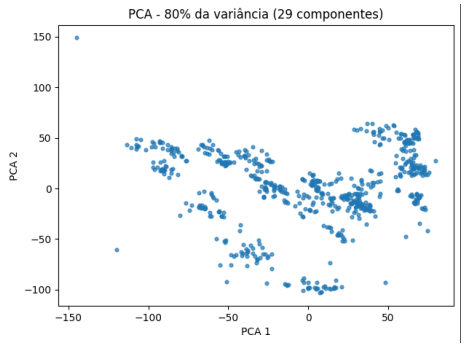

Fig. 9. Projeção das imagens faciais em 2D utilizando PCA com 80% da variância preservada.

  - **75% de variância**

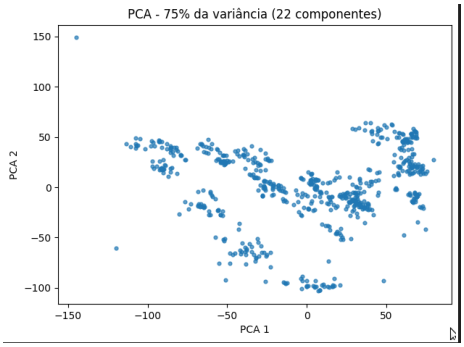

Fig. 10. Projeção das imagens faciais em 2D utilizando PCA com 75% da variância preservada.

- **UMAP:** Projeção 2D e 3D para dimensões 3, 15, 55 e 101
  - **3 dimensões:**

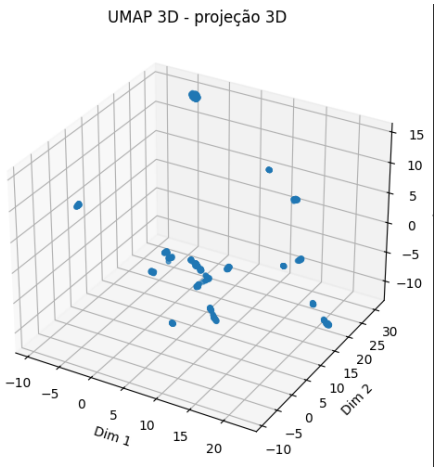

Fig. 11. Projeção das imagens faciais em 3 dimensões utilizando UMAP, mostrando agrupamentos iniciais.

– **15 dimensões:**

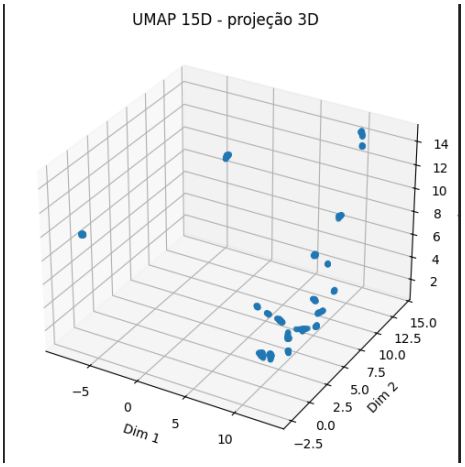

Fig. 12. Projeção das imagens faciais em 15 dimensões utilizando UMAP, reduzida para visualização em 2D.

– **55 dimensões:**

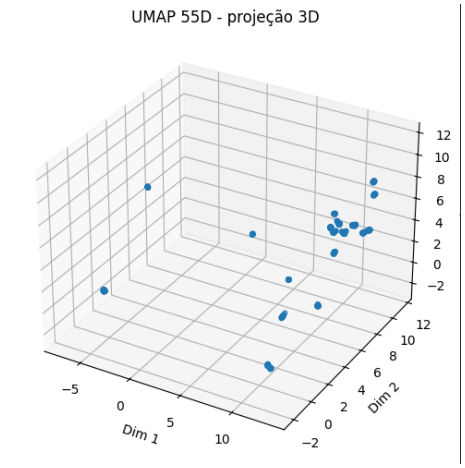

Fig. 13. Projeção das imagens faciais em 55 dimensões utilizando UMAP, reduzida para visualização em 2D.

– **101 dimensões:**

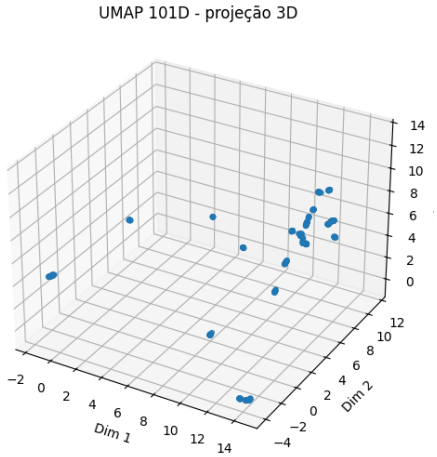

Fig. 14. Projeção das imagens faciais em 101 dimensões utilizando UMAP, reduzida para visualização em 2D.

2) Clusterização – K-means e K-medoids:

– **Índice de Dunn:**

∗ Resultados usando K-means:

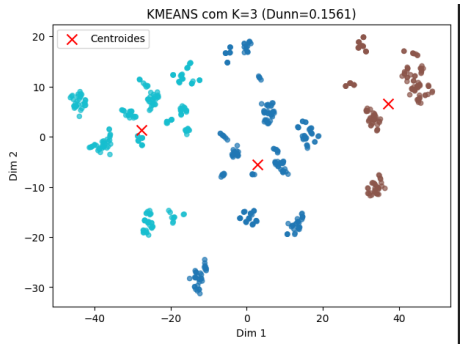

Fig. 15. Resultado da clusterização com K-means para K=3 após redução de dimensionalidade. Cada ponto representa uma imagem projetada e colorida pelo cluster atribuído.

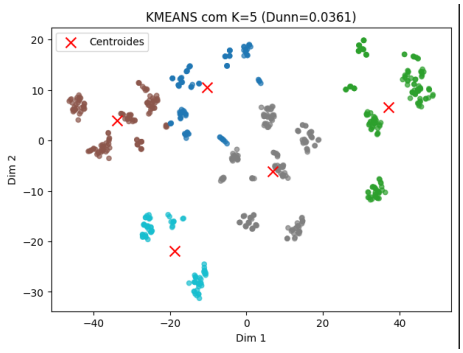

Fig. 16. Resultado da clusterização com K-means para K=5 após redução de dimensionalidade.

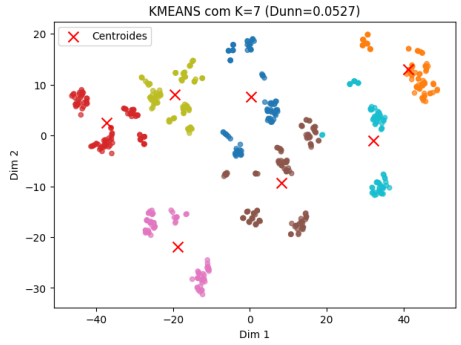

Fig. 17. Resultado da clusterização com K-means para K=7 após redução de dimensionalidade.

∗ Resultados usando K-medoids:

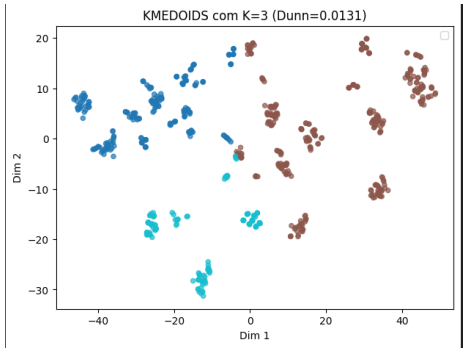

Fig. 18. Resultado da clusterização com K-medoids para K=3 após redução de dimensionalidade. Cada ponto representa uma imagem projetada e colorida pelo cluster atribuído.

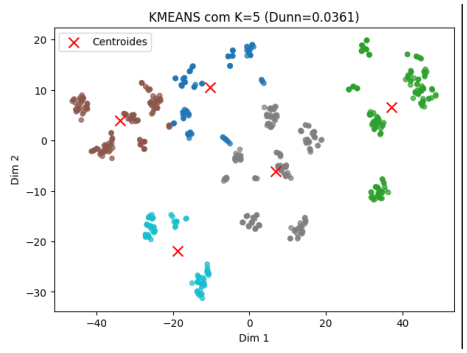

Fig. 19. Resultado da clusterização com K-medoids para K=5 após redução de dimensionalidade.

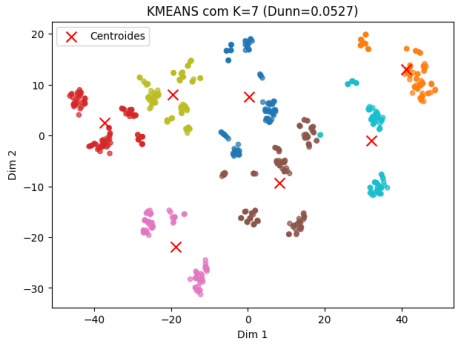

Fig. 20. Resultado da clusterização com K-medoids para K=7 após redução de dimensionalidade.

- Comparação dos índice de DUNN entre K-MEANS e K-MEDOIDS:

TABLE V
VALORES DO ÍNDICE DE DUNN PARA K-MEANS E K-MEDOIDS

K	K-means Dunn	K-medoids Dunn
3	0.15608044993236098	0.013112439925528156
5	0.03606069105917359	0.002963912988607878
7	0.05272609406172854	0.0032991700898502407

IV. CONCLUSÃO

Neste trabalho, foram exploradas diferentes abordagens de aprendizado de máquina aplicadas a sinais de EMG faciais e à regressão de potência de aerogeradores. Para a tarefa de regressão, o modelo de regressão linear demonstrou ser capaz de estimar com precisão a relação entre a velocidade do vento e a potência gerada, conforme evidenciado pelas métricas MSE e MAE obtidas após 500 rodadas de validação aleatória. A análise estatística mostrou consistência nos resultados, com médias baixas de erro e desvios-padrão reduzidos.

Na tarefa de classificação de sinais EMG, foram aplicados dois modelos distintos: K-NN e Gaussian Naive Bayes. Ambos os modelos apresentaram bom desempenho na identificação das expressões faciais, embora o K-NN tenha apresentado ligeira vantagem em acurácia média, enquanto o Bayesiano apresentou maior consistência em algumas classes específicas. A utilização de validação cruzada e amostragem aleatória permitiu avaliar a robustez dos modelos, e as matrizes de confusão associadas aos melhores e piores casos

forneceram insights adicionais sobre possíveis áreas de confusão entre classes.

Adicionalmente, a aplicação de métodos de redução de dimensionalidade (PCA, t-SNE e UMAP) facilitou a visualização e compreensão dos dados de alta dimensionalidade, além de auxiliar na escolha do número ideal de clusters para algoritmos não supervisionados, como K-means e K-medoids.

REFERENCES

[1] Abu-Mostafa, Yaser S.; Magdon-Ismail, Malik; Lin, Hsuan-Tien. *Learning from data*. New York: AML-Book, 2012.

[2] Bishop, Christopher M.; Nasrabadi, Nasser M. *Pattern recognition and machine learning*. New York: Springer, 2006.

[3] Faceli, Katti; Lorena, Ana Carolina; Garcia, Rodrigo Fernandes de Mello; Gama, João. *Inteligência artificial: uma abordagem de aprendizado de máquina*. 2. ed. Rio de Janeiro: LTC, 2023. 400 p.

[4] Haykin, Simon. *Neural networks and learning machines*. 3. ed. Índia: Pearson Education, 2009.

[5] Marsland, Stephen. *Machine learning: an algorithmic perspective*. Boca Raton: Chapman and Hall/CRC, 2011.

[6] Netto, Amilcar; Maciel, Francisco. *Python para data science: machine learning descomplicado*. Rio de Janeiro: Alta Books, 2021.

[7] Rezende, Solange Oliveira (org.). *Sistemas inteligentes: fundamentos e aplicações*. 1. reimp. São Paulo: Manole, 2005.

[8] Dunn, Joseph C. *Well-separated clusters and optimal fuzzy partitions*. Journal of Cybernetics, v. 4, n. 1, p. 95-104, 1974.

[9] MacQueen, James. *Some methods for classification and analysis of multivariate observations*. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. University of California Press, 1967. p. 281-298.

[10] Misuraca, Michelangelo; Spano, Maria; Balbi, Simona. *BMS: An improved Dunn index for Document Clustering validation*. Communications in Statistics - Theory and Methods, v. 48, n. 20, p. 5036-5049, 2019.

[11] Shirkhorshidi, Ali Seyed; Aghabozorgi, Saeed; Wah, Teh Ying. *A comparison study on similarity and dissimilarity measures in clustering continuous data*. PLoS One, v. 10, n. 12, p. e0144059, 2015.

[12] Webb, Andrew R. *Statistical pattern recognition*. John Wiley & Sons, 2003.