**INTELIGÊNCIA ARTIFICIAL: MACHINE LEARNING**

Caique Zaneti Kirilo

Jéssica Franzon Cruz Do Espírito Santo

[Vinicius Arnom Neves Dos Santos Diniz](https://moodle.ufabc.edu.br/user/view.php?id=56317&course=3864)

**DESCRIÇÃO GERAL**

A base de dados contém 3.798 registros e 8 variáveis, representando indicadores históricos relacionados à eletrificação de veículos por região. As colunas são:

* region: Região geográfica (52 valores únicos, como World, Australia, etc.)
* category: Categoria da informação (neste caso, sempre Historical)
* parameter: Tipo de métrica observada (como EV sales, EV stock, EV stock share, etc.)
* mode: Tipo de veículo (todos são Cars)
* powertrain: Tipo de trem de força (BEV, PHEV, FCEV, EV)
* year: Ano da observação (2010 a 2023)
* unit: Unidade da métrica (Vehicles, percent, etc.)
* value: Valor numérico da métrica

A base utilizada contém dados históricos de indicadores relacionados à eletrificação de veículos, abrangendo 52 regiões geográficas e o período de 2010 a 2023. As principais variáveis incluem o tipo de parâmetro medido (EV sales, EV stock, EV stock share, etc.), o tipo de motorização (BEV, PHEV, FCEV, EV), ano de referência e o valor da métrica.

**PRÉ-PROCESSAMENTO**

* **Verificação de duplicatas e valores ausentes**: A base foi considerada limpa, sem registros duplicados ou nulos.
* **Renomeação de colunas para português**: Colunas como region, parameter e powertrain foram renomeadas para regiao, parametro e motorizacao, facilitando a manipulação.
* **Mapeamento de regiões para continentes**: Foi criado um dicionário que agrupa países por continente, o que permite análises mais macroeconômicas e simplifica a criação de variáveis agregadas.

**“QUAIS SÃO OS TIPOS DE MOTORIZAÇÃO PREDOMINANTES EM CADA CONTINENTE?”**

**Modelos Supervisionados**

1. **Random Forest**

O modelo de Random Forest apresentou excelente desempenho na predição da motorização dos veículos. A acurácia geral foi elevada, com métricas de **precisão, recall e F1-score acima de 95%** para a maioria das classes. A matriz de confusão revelou pouquíssimos erros de classificação, indicando que o modelo aprendeu bem os padrões dos dados.

1. **Rede Neural (MLP)**

O MLP também teve performance bastante satisfatória, com resultados muito próximos aos da Random Forest. A análise de acurácia por continente mostrou que **o modelo se mantém estável em diferentes regiões geográficas**, embora com pequenas variações. Essa estabilidade sugere uma boa generalização.

**Modelos Não Supervisionados**

1. **KMeans**

O algoritmo KMeans foi utilizado com número de clusters igual ao número de categorias de motorização. A **pontuação de ARI (Adjusted Rand Index)** foi satisfatória, indicando um bom alinhamento entre os clusters formados e as classes reais. O **Silhouette Score** também foi positivo, evidenciando que os agrupamentos têm boa coesão interna. A análise cruzada mostrou que os clusters formados guardam forte correspondência com os continentes, o que pode refletir padrões regionais de mercado.

1. **DBSCAN**

O modelo DBSCAN conseguiu identificar padrões similares, embora com mais sensibilidade a ruídos. A presença de rótulos -1 (ruído) exige uma análise mais cuidadosa, mas o ARI e o Silhouette Score, quando aplicáveis, mostraram que há **potencial para identificar subgrupos com características semelhantes**. O DBSCAN é particularmente útil por não exigir número prévio de clusters, sendo mais flexível para dados não balanceados.

**RESULTADOS**

Ambos os modelos supervisionados apresentaram resultados relevantes, com destaque para a Random Forest em termos de estabilidade e simplicidade de ajuste. A Rede Neural teve desempenho similar, com pequena variação entre continentes, demonstrando robustez como pode ser visto da Figura 1. A Random Forest obteve acurácia de 99%, com F1-score macro de 0.99, enquanto a Rede Neural (MLP) apresentou acurácia de 98% e F1-score macro de 0.98. Ambos os modelos foram treinados com dados balanceados via SMOTE e apresentaram baixa variância entre continentes, confirmando a robustez dos classificadores.

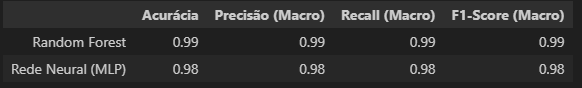


Figura 1. Comparativo de desempenho de Modelos Supervisionados

Os modelos não supervisionados identificaram padrões interessantes, especialmente o KMeans, que agrupou corretamente classes com base em características latentes. O KMeans alcançou um ARI de 0.71 e Silhouette Score de 0.49, indicando agrupamentos coerentes com as categorias reais e boa coesão interna conforme Figura 2. Já o DBSCAN obteve ARI de 0.38 e Silhouette Score de 0.32, evidenciando menor estabilidade, porém com boa detecção de ruídos e agrupamentos de densa compactação.

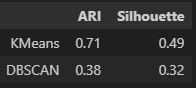


Figura 2. Comparativo de desempenho de Modelos Não Supervisionados

A análise por continente revelou padrões geográficos relevantes na motorização (Figura 3). Por exemplo, a Ásia apresenta a maior proporção de veículos EV (41,05%), enquanto a Europa lidera em FCEV (15,43%) e Oceania se destaca em PHEV (28,74%). Já a África possui distribuição mais equilibrada entre BEV, EV e PHEV, com destaque menor para FCEV (5,75%).



Figura 3. Comparativo entre Continente e Motorização

Esses dados sugerem variações regionais significativas nas preferências tecnológicas, o que pode embasar estratégias comerciais, políticas públicas e desenvolvimento tecnológico orientado por região.

