

Relatório de Classificação Supervisionada: Tentativa de Engenharia Reversa de Marcas

Disciplina: Inteligência Artificial (FGA0221)

Professor: Fabiano Araujo Soares, Dr

Aluno: Marcos Antonio Teles de Castilhos

1. Definição do Problema

O objetivo deste experimento foi investigar se marcas de robótica possuem "assinaturas de engenharia" distintas. A hipótese inicial era de que fabricantes diferentes priorizam trade-offs diferentes (ex: a *iRobot* poderia focar em sensores, enquanto a *Xiaomi* foca em bateria/custo). Utilizou-se um modelo de classificação supervisionada para tentar prever a **Marca** do robô baseando-se exclusivamente em suas **Especificações Técnicas**.

2. Metodologia

- **Algoritmo:** *Random Forest Classifier* (n=100), escolhido por sua robustez em lidar com dados não-lineares e capacidade de medir a importância das features.
- **Variáveis Preditoras (X):** Todas as especificações técnicas (Sução, Bateria, Capacidade, Ruído, Sensores, etc.).
- **Variável Alvo (y):** Marca (Brand).
- **Protocolo:** Divisão 80/20 (Treino/Teste) com semente aleatória fixa para reprodutibilidade.

3. Resultados Experimentais

O modelo apresentou desempenho **nulo**, indistinguível de uma escolha aleatória.

- **Acurácia Global: 0.09 (9%).**

- Considerando que existem 10 classes balanceadas, a probabilidade de acerto ao acaso é de 10%. O modelo não conseguiu aprender nenhum padrão acima do acaso.
- **Precisão e Recall:** Próximos de 0.00 para a maioria das classes. O modelo falhou em identificar consistentemente qualquer uma das marcas (Ecovacs e Eufy tiveram 0% de acerto).

3.1. Análise de Importância das Features (Feature Importance)

A análise da contribuição de cada variável para a decisão da árvore confirmou a ausência de sinal nos dados.

| Feature | Importância (%) |
|------------------|-----------------|
| Suction Power | 9.3% |
| Dustbin Capacity | 9.2% |
| Battery Life | 8.9% |
| Charging Time | 8.6% |
| Weight | 8.4% |

Interpretação: Em um cenário real, esperaríamos que 1 ou 2 features dominassem (ex: "Sistema de Navegação" seria decisivo para distinguir marcas high-end). No entanto, a distribuição de importância é **plana/uniforme**. Todas as variáveis contribuem com cerca de 8-9%, o que indica que o modelo estava apenas dividindo ruído aleatório.

4. Diagnóstico Crítico dos Dados

Este experimento serviu como a validação final da qualidade do dataset sintético utilizado. Os resultados comprovam que:

1. **Ausência de Identidade de Marca:** O algoritmo gerador dos dados não implementou perfis distintos para cada marca. Uma *iRobot* no dataset tem estatisticamente as mesmas especificações de uma *Proscenic*.
2. **Independência Estatística Total:** A variável **Brand** é estatisticamente independente das variáveis de **Specs**. Não há correlação, linear ou não-linear, que permita a classificação.

5. Conclusão

O experimento de classificação falhou em prever as marcas, mas teve sucesso em **auditar a integridade dos dados**. Conclui-se que o dataset é composto por ruído branco estruturado, onde os rótulos (Labels) não possuem relação causal ou correlacional com os atributos (Features).