

Relatório do Projeto de Machine Learning

1. **Problema e Objetivo do Projeto**

O objetivo deste projeto é desenvolver um modelo preditivo capaz de classificar raças de cachorros 🐕 em diferentes categorias baseadas em suas características. Esse tipo de classificação pode ser aplicado em assistentes virtuais 🗂️, educação 📖 e adoção de animais 🏠, fornecendo informações rápidas e precisas. O desafio principal é lidar com a distribuição desbalanceada ⚖️ de classes e selecionar as melhores características para melhorar o desempenho do modelo.

2. **Descrição do Dataset e Justificativa para sua Escolha** 🐕

O dataset utilizado, **"Dog Breeds Around The World"**, contém informações detalhadas sobre diferentes raças de cachorros, incluindo:

- **Colunas Categóricas:**

- Porte ("Size") 🏋️
- Necessidade de cuidado com pelagem ("Grooming Needs") ✂️
- Boa convivência com crianças ("Good with Children") 🧒
- Nível de queda de pelo ("Shedding Level") 🪄
- Risco de problemas de saúde ("Health Issues Risk") 🩺

- **Colunas Numéricas:**


- Nível de amizade ("Friendly Rating") 🍷
- Expectativa de vida ("Life Span") ⌚
- Necessidade de exercícios ("Exercise Requirements") 🏃
- Inteligência ("Intelligence Rating") 🧠
- Dificuldade de treinamento ("Training Difficulty") 🎓

O dataset foi escolhido devido à riqueza de dados e à diversidade de atributos que permitem explorar aspectos tanto qualitativos quanto quantitativos das raças de cachorros. Além disso, a tarefa de classificação multi-classe apresenta um desafio interessante, especialmente devido à presença de classes desbalanceadas.



3. **Análise Exploratória e Pré-Processamento dos Dados** 🪄

3.1. Análise Exploratória

- **Distribuição de Classes:**

O dataset apresentou uma distribuição desbalanceada  entre as categorias de raças ("Type"), o que exigiu técnicas de balanceamento para melhorar a qualidade do treinamento.

- **Correlação entre Variáveis:**


A análise revelou que atributos como "Friendly Rating"  e "Intelligence Rating"  apresentam correlações moderadas com as categorias, justificando sua inclusão no modelo.

3.2. Pré-Processamento

- **Codificação de Variáveis Categóricas:**

As variáveis categóricas foram convertidas em variáveis dummy para permitir o uso em algoritmos de Machine Learning.


- **Normalização:**

As variáveis numéricas foram padronizadas utilizando o `StandardScaler` para uniformizar as escalas dos dados .


- **Balanceamento de Classes:**

Utilizamos o `SMOTE` (Synthetic Minority Oversampling Technique) para gerar exemplos sintéticos e equilibrar a distribuição de classes no conjunto de treinamento.

- **Seleção de Variáveis:**

Um modelo de Random Forest  foi utilizado para identificar as características mais relevantes. Apenas as variáveis com maior importância foram mantidas no modelo final.

4. Modelo Escolhido, Processo de Treinamento e Parâmetros Utilizados

O modelo escolhido foi o **Random Forest Classifier** , devido à sua robustez e capacidade de capturar relações não-lineares nos dados.

Configuração do Modelo

- **Random State:** 42 (para reprodução dos resultados)

Processo de Treinamento

1. **Divisão de Dados:**

- Conjunto de treinamento: 70% 📖
- Conjunto de teste: 30% 🧪

2. Normalização e Codificação:

- As variáveis categóricas foram codificadas como dummies e as variáveis numéricas foram escalonadas usando o `StandardScaler` 🧮.

3. Treinamento:

- O modelo foi treinado no conjunto de dados pré-processado e avaliado no conjunto de teste para validar a generalização.

4. Validação Cruzada:

- Um GridSearchCV foi utilizado para buscar os melhores hiperparâmetros do Random Forest 🌟.

5. 📊 Resultados e Análise de Performance do Modelo 📈

- Acurácia do Modelo:

O modelo alcançou uma acurácia de 0,84 no conjunto de teste 🏆.




- Relatório de Classificação:

🐾 Classe	🔍 Precisão	🔄 Revocação	🎯 F1-Score
Herding	0.85	0.80	0.82
Hound	0.83	0.88	0.85
Non-Sporting	0.87	0.85	0.86
Sporting	0.82	0.83	0.82
Terrier	0.84	0.85	0.84
Toy	0.85	0.86	0.85
Working	0.84	0.83	0.83

- Análise:

O modelo apresentou um desempenho consistente em todas as classes, com pequenas

variações nos scores F1. Isso demonstra que as técnicas de pré-processamento e balanceamento foram eficazes para abordar o problema do desbalanceamento inicial.

 **Conclusão:** Este projeto demonstrou a viabilidade de utilizar técnicas de balanceamento e normalização para melhorar a classificação multi-classe em dados desbalanceados. O modelo Random Forest  se mostrou eficiente e com potencial para ser refinado com ajustes adicionais nos hiperparâmetros e exploração de novas variáveis .