Documentação do Projeto de Aprendizado de Máquina com Super-Heróis

**Introdução**

Este relatório descreve o desenvolvimento de um projeto de classificação de super-heróis utilizando técnicas de aprendizado de máquina. O objetivo foi classificar personagens como Herói, Anti-Herói ou Vilão com base em atributos como força, inteligência, velocidade, entre outros.

1. Dataset

O dataset contém informações sobre super-heróis, incluindo habilidades, universo (Marvel/DC), armamento, etc. O objetivo é classificar o alinhamento (Alignment) como Good, Bad ou Neutral.

1. Pré-processamento e Análise Exploratória
   * Leitura dos dados: realizada com pandas.read\_csv().
   * Remoção de valores nulos com df.dropna() para evitar problemas na modelagem.
   * Transformação de variáveis categóricas com pd.get\_dummies() (one-hot encoding).
   * Transformação da variável de saída com LabelEncoder().
   * Análise exploratória com describe(), histogramas e gráficos de barras usando matplotlib e seaborn.

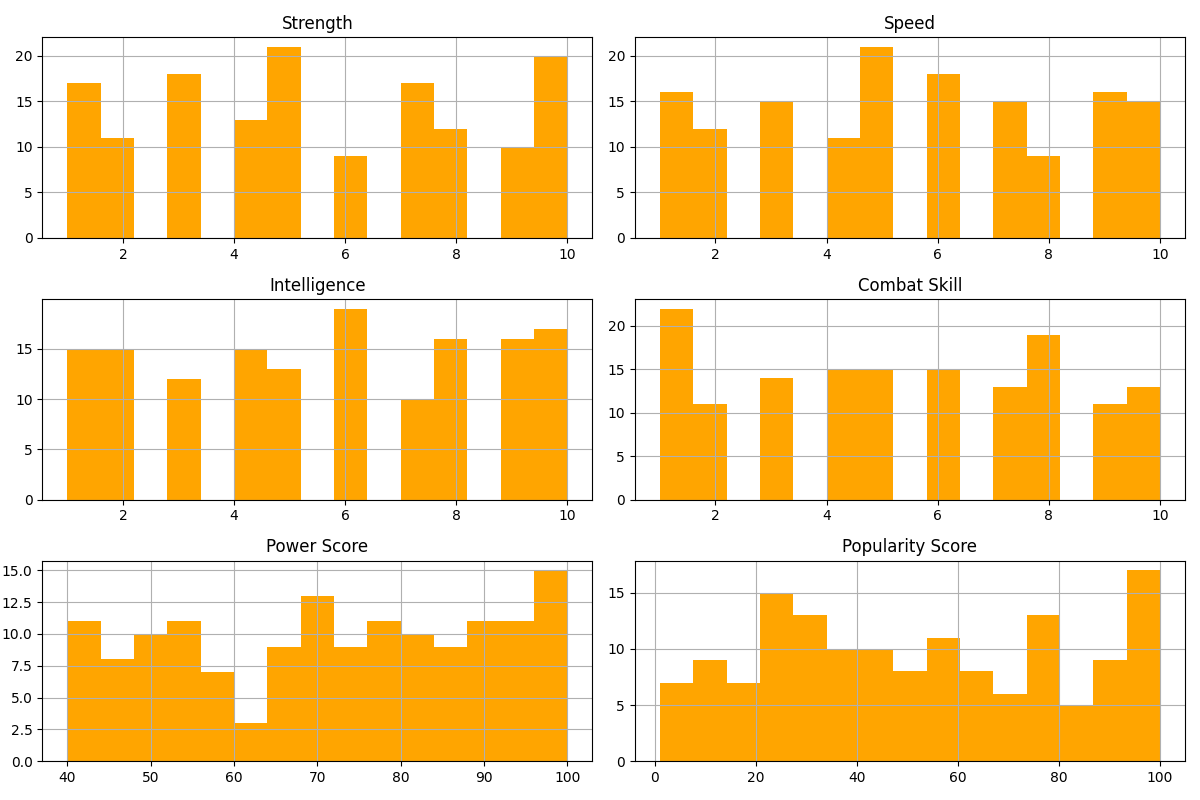
Funções importantes:

* + pd.get\_dummies(): converte colunas categóricas em variáveis binárias (ex: 'Universe' -> 'Universe\_Marvel').
  + LabelEncoder(): converte valores de texto (Good, Bad, Neutral) para inteiros (0, 1, 2).
  + StandardScaler(): normaliza os dados para média 0 e desvio padrão 1.

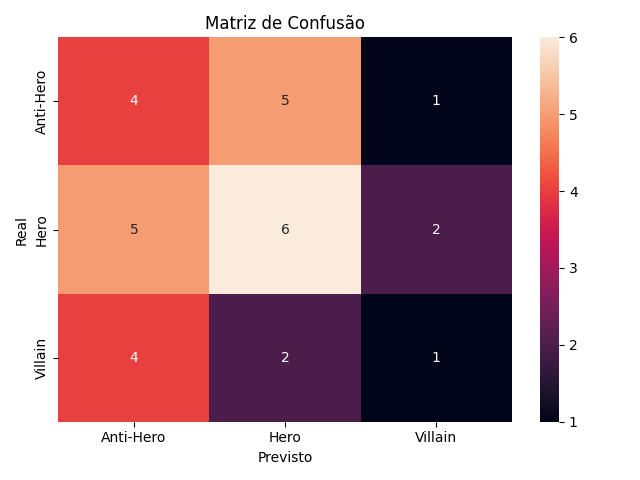
1. Aprendizado de Máquina (ML)
   * Separação de treino e teste com train\_test\_split().
   * Normalização com StandardScaler().
   * Modelo RandomForestClassifier() do scikit-learn para classificação.
   * Avaliação com classification\_report() e confusion\_matrix().
2. Explicabilidade
   * SHAP: usado para visualizar a contribuição de cada variável para as previsões do modelo.
   * LIME: utilizado para interpretar localmente o motivo de uma predição específica.
3. Resultados Gerados
   * Histogramas e gráficos salvos na pasta 'resultados/exploratoria'.
   * Matriz de confusão salva como imagem.
   * Gráfico SHAP e explicação LIME gerados para análise explicativa.

**Explicação dos Gráficos**

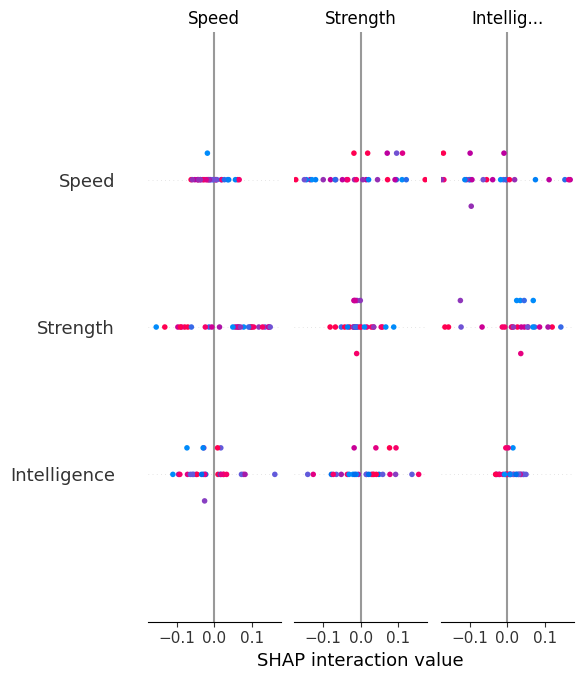
**1.** **Gráficos de Distribuição** (barras): mostram a frequência de valores para atributos como Força, Velocidade e Inteligência. Projeto de Aprendizado de Máquina - Classificação de Super-Heróis Esses gráficos ajudam a entender como os dados estão distribuídos e se há viés em alguma classe



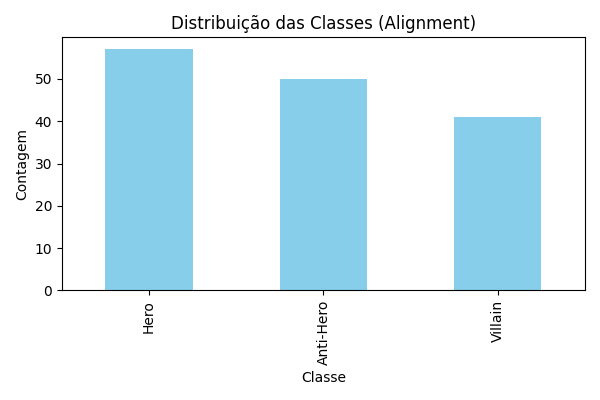
**2. Matriz de Confusão**: compara os rótulos reais com os previstos pelo modelo, evidenciando os acertos e erros por classe. Ajuda a entender onde o modelo está confundindo os personagens, por exemplo, prevendo um herói como anti-herói.



**3. Gráfico SHAP (Interações):** mostra como os atributos interagem entre si e sua contribuição na decisão do modelo. Permite entender, por exemplo, como a combinação entre velocidade e força impacta na classificação.



**4. Distribuição de Classes:** O gráfico apresentado exibe a distribuição das classes de alinhamento dos personagens no conjunto de dados. As três categorias representadas são: Herói, Anti-Herói e Vilão.



**Possíveis Perguntas e Respostas**

Q: O que faz o get\_dummies?

A: Transforma variáveis categóricas em variáveis binárias (0 ou 1).

Q: Para que serve StandardScaler?

A: Para padronizar os dados (média 0, desvio padrão 1).

Q: Por que SHAP e LIME?

A: SHAP oferece explicações globais e locais baseadas em teoria de jogos. LIME fornece explicações locais intuitivas.

Q: Como lidou com dados faltantes?

A: Removendo linhas com NaN nas colunas principais com df.dropna().

Q: Por que usar Random Forest?

A: É robusto, lida com diferentes tipos de dados e funciona bem com ferramentas de explicabilidade.