Relatório de Análise e Classificação do dataset Fetal Health

Resumo do Dataset:

O treinamento de classificação tem como objetivo analisar dados de cardiotocografia fetal (CTG) e construir modelos de aprendizado de máquina capazes de **classificar a saúde fetal** em três categorias:

- **Normal (1)**
- **Suspeito (2)**
- **Patológico (3)**

O dataset foi importado do arquivo CSV contendo 2126 registros e 22 variáveis.

link: https://www.kaggle.com/datasets/andrewmvd/fetal-health-classification

Foram realizadas visualizações iniciais:

- **Distribuição das classes** usando countplot.
- **Boxplots** das variáveis numéricas para verificar outliers.
- **Pairplot** e **scatter matrix** para explorar correlações entre features.

Pré Processamento

Tratamento de valores ausentes

- Função trata_faltantes aplicada:
- **Colunas numéricas**: preenchidas com a média da coluna.
- **Colunas categóricas**: preenchidas com a moda.

Limpeza de valores inválidos

- Função limpar_valores_invalidos_fetal aplicada para remover registros com valores fora do intervalo esperado:

Remoção de duplicados

Todas as linhas duplicadas exatas foram removidas para evitar redundância de informação.

Normalização

Todas as colunas numéricas foram **normalizadas usando Z-score**, preservando a escala relativa entre as variáveis e garantindo melhor desempenho dos modelos baseados em distância e gradiente.

Visualização de Dados

- **Boxplots gerais** das features para identificar distribuições e outliers.
- **Pairplots** por categoria de saúde fetal para entender relações entre variáveis.
- **Scatter matrix** das principais variáveis, confirmando correlações entre medidas de CTG, como acelerações, decelerations e variabilidade.

Treinamento de Modelos de Classificação

Algoritmos Testados:

Foram testados três algoritmos clássicos de classificação:

- **Random Forest**: ensemble de árvores de decisão, robusto a outliers e capaz de estimar importância de features.

  Taxa de acerto: 95,2%

- **Logistic Regression**: modelo linear para classificação multinomial.

  Taxa de acerto: 88,8%

- **K-Nearest Neighbors (KNN)**: baseado em similaridade de observações.

Taxa de acerto: 88,8%

Pipeline

Cada modelo foi treinado em um pipeline que incluiu:

- **StandardScaler**: normalização das features.
- **Classificador**: algoritmo correspondente.

Avaliação foi feita usando:

- **Acurácia**
- **Relatório de classificação**
- **Matriz de confusão**

Resultados iniciais

- Todas as métricas foram calculadas no conjunto de teste (20% dos dados).
- Random Forest apresentou a melhor performance geral, com alta acurácia e bom equilíbrio entre classes.

Seleção de Features

A importância das features foi calculada usando o **atributo feature_importances_ do Random Forest**.

As features menos importantes puderam ser descartadas para simplificar o modelo sem perda significativa de performance.

Uma tabela visual foi gerada mostrando:

- **Nome da feature**
- **Valor de importância**

Esse procedimento permitiu identificar quais variáveis mais contribuem para a classificação da saúde fetal.

Retreinamento com Features Selecionadas

O modelo Random Forest foi re-treinado usando apenas as **features mais importantes**.

- Novas métricas de avaliação foram geradas:
    - **Relatório de classificação**
    - **Acurácia**
    - **Matriz de confusão**