Comparação de Métodos de Explicação para Predições Individuais de Random Forests

Objetivo

O objetivo deste projeto é comparar dois métodos de explicação para predições individuais de modelos Random Forests, com base na métrica de **fidelity**. Os métodos a serem comparados são:

- **Explicações abdutivas** geradas por meio do **solver Z3**;
- **Explicações LIME** (Local Interpretable Model-agnostic Explanations).

Descrição da Tarefa

1. Implementar um método para obter explicações abdutivas para predições de instâncias pelo modelo random forest. Você deve implementar o método usando o solver Z3.
2. Escolher pelo menos **5 conjuntos de dados** com **no mínimo 8 atributos** cada e com variação na quantidade de atributos.
3. Treinar um **modelo Random Forest** de classificação para cada conjunto de dados.
4. Selecionar um subconjunto de **instâncias de teste** (mínimo recomendado: 30 ou máximo do conjunto de teste).
5. Para cada instância de teste:
   - Gerar uma **explicação abdutiva** usando o solver **Z3**.
   - Gerar uma **explicação com o LIME** contendo um número de atributos igual ao tamanho da explicação abdutiva **mais 1**.
6. Avaliar a **fidelity** de cada explicação com base no comportamento do modelo Random Forest.
7. Para cada dataset, reportar:
   - **Fidelity média e desvio padrão** de cada método.
   - **Tamanho médio e desvio padrão** das explicações abdutivas.
   - **Tempo médio e desvio padrão** para obter as explicações.

Fidelity

A métrica de **fidelity** deve ser definida da seguinte forma:

- Gerar uma **vizinhança local** da instância explicada (ex.: pequenas perturbações nas features).
- Calcular a proporção de vezes em que a explicação reproduz corretamente a predição do modelo para essa vizinhança.
- Fidelity = $\frac{\text{número de acertos da explicação}}{\text{quantidade de vizinhos}}$

A mesma vizinhança deve ser usada para avaliar os dois métodos.

Conjuntos de Dados Sugeridos

Os alunos podem utilizar datasets de domínio público, como:

- UCI Adult Income (14 atributos)
- Breast Cancer Wisconsin (30 atributos)
- Heart Disease (13 atributos)

O aluno deve dar preferência a conjuntos de dados com temática semelhante ao do seu projeto de pesquisa do mestrado. O mais recomendado seriam conjuntos de dados que serão de fato utilizados na pesquisa.

Entrega

- Link do Google Colab, contendo:
  - Código-fonte completo;
  - Resultados experimentais em tabelas;
  - Exemplos de explicação fornecidos por ambos os métodos em instâncias variadas de datasets variados;
  - Breve discussão dos resultados;