



T1: Interpretabilidade de Modelos de Aprendizado de Máquina

Prof^a. Dr^a. Daniela Oliveira Ferreira do Amaral

Objetivo:

Dentre as várias propriedades desejáveis em um modelo de aprendizado de máquina, a interpretabilidade é uma das mais importantes. Modelos interpretáveis são mais fáceis de entender, depurar e confiar, o que é essencial em muitas aplicações críticas como nas áreas da saúde e direito.

O objetivo deste trabalho é explorar e compreender o que são modelos interpretáveis e como extrair informações relevantes do processo e aprendizado. Você deverá treinar e interpretar os modelos (**KNN, Naïve Bayes, Árvore de Decisão**) relatando e analisando os resultados encontrados.

GRUPOS: mínimo 2 alunos e máximo de 5.

Atenção: Trabalhos individuais NÃO SERÃO CONSIDERADOS!

Etapas:

1. Escolha do Dataset:

- Os alunos devem escolher um dataset público (por exemplo, do UCI Machine Learning Repository, Kaggle, ou outro) que seja adequado para classificação. O dataset deve:
 - ter um **número razoável de features (pelo menos 5)** para permitir a análise de importância de features;
 - resolver um **problema de classificação**, ou seja, a variável a ser predita deve ser categórica.
- O conjunto de dados deve incluir tanto variáveis numéricas quanto categóricas, e deve ser suficientemente grande para permitir uma análise significativa (pelo menos 100 instâncias).
- Não serão considerados datasets comumente usados em exemplos de aula, como: Iris, Titanic, Adult, Breast Cancer, Wine Quality, etc.
- Exemplos de sites que podem ser utilizados para procurar dados:
 - [UCI Machine Learning Repository](https://archive.ics.uci.edu/).

- [Kaggle Datasets](#)
- [Google Dataset Search](#)

2. Treinamento dos Modelos:

- Você deverá montar o fluxo de treinamento de um modelo de aprendizado de máquina, incluindo:
 - Pré-processamento dos dados (tratamento de valores ausentes, normalização, codificação de variáveis categóricas, etc.).
 - Divisão do dataset em conjuntos de treinamento e teste (80/20 ou 70/30).
 - Treinamento dos modelos KNN, Naïve Bayes e Árvore de Decisão.
 - Avaliação do desempenho dos modelos utilizando métricas apropriadas (acurácia, precisão, recall, F1-score, etc.).
 - Justifique as escolhas feitas durante o pré-processamento e treinamento dos modelos.
- Garanta que os modelos possuem uma performance suficiente para que a análise de interpretabilidade seja significativa.

3. Interpretabilidade dos Modelos:

- Você deverá explicar as decisões de cada um dos modelos, para isso, utilize ferramentas variadas de interpretabilidade para cada um dos modelos. Algumas sugestões são:
 - **Árvore de Decisão:** Analisar a árvore gerada e identificar as features mais importantes.
 - **Naïve Bayes:** Analisar as probabilidades condicionais e discutir como elas influenciam as previsões.
 - **KNN:** Discutir a dificuldade de interpretação do KNN e explorar técnicas como SHAP ou LIME para interpretar as previsões.
- Tenha em mente que ferramentas como Análise de Permutação, SHAP e LIME podem ser usadas independentemente do modelo.

4. Comparação e Análise:

- Comparar a interpretabilidade dos três modelos. Tente responder algumas perguntas:
 - Os resultados fizeram sentido?
 - Os modelos concordaram em quais as variáveis mais relevantes?
- Explicar as ferramentas de interpretabilidade utilizadas nos modelos.
- Discutir as limitações de cada modelo em termos de interpretabilidade.

5. Apresentação:

- Os alunos devem gravar e postar um vídeo de uma breve apresentação (10-15 minutos) para explicar o desenvolvimento do trabalho. A apresentação deve incluir:
 - Descrição do dataset e do problema.
 - Metodologia de treinamento e avaliação dos modelos.

- ♦ Análise de interpretabilidade para cada modelo.
- ♦ Discussão sobre a comparação dos modelos e a importância da interpretabilidade.
- ♦ Conclusões e reflexões finais.

Critérios de Avaliação:

Aspecto	Escolha dos Dados	Pré-Processamento e Treinamento	Interpretabilidade	Análise	Apresentação
Pontos	1	1	2	3	3

Entregáveis:

- ♦ O **código desenvolvido** deve ser enviado via GitHub ou uma pasta compactada no Moodle.
- ♦ O **vídeo da apresentação** deve ser enviado via YouTube (não listado) ou outra plataforma de compartilhamento de vídeo. Utilize um arquivo txt para indicar o link.

Apontamentos Gerais:

- ♦ Justificativas ou discussões, como necessário ao analisar a interpretabilidade, **devem estar presentes no código em formato de README, comentários ou Markdown.**
- ♦ Principais bibliotecas de aprendizado de máquina ou utilitárias, são:
 - **Scikit-learn:** <https://scikit-learn.org/stable/>
 - **Pandas:** <https://pandas.pydata.org/>
 - **NumPy:** <https://numpy.org/>
 - **Matplotlib:** <https://matplotlib.org/>
 - **Seaborn:** <https://seaborn.pydata.org/>
- ♦ Dentre as bibliotecas de interpretabilidade, alguns exemplos, além do próprio Sklearn são:
 - **SHAP (SHapley Additive exPlanations):** <https://github.com/slundberg/shap>
 - **LIME (Local Interpretable Model-agnostic Explanations):** <https://github.com/marcotcr/lime>
 - **ELI5 (Explain Like I'm 5):** <https://github.com/eli5-org/eli5>

Prazo:

- ♦ O trabalho deve ser entregue, impreterivelmente, até **30/09/2025**.