1 Introdução

O avanço dos métodos de machine learning podem ser notados com determinadas características, como por exemplo: acurácia, precisão, sensibilidade e especificidade. Uma característica que vem sendo integrada no grupo de características normalmente utilizado é o entendimento do modelo, ou o quão interpretável ele ou seus resultados são.

Existem duas áreas principais que tem investigado a interpretabilidade dos modelos de machine learning: Explainable Artificial intelligence (XAI) e Interpretable Machine Learning (iML). Essas áreas tem ganho mais destaque à medida que avançamos para a interação mais direta com os seres humanos, como é o caso da GDPR (General Data Protection Regulation), que busca proteger os dados e a privacidade de cidadãos da União Europeia (UE) e do Espaço Econômico Europeu (EEE). O caso da GDPR tem sido particularmente bem discutido conforme referências recentes [1, 2, 3, 4].

Em análise de dados, uma das utilidades da indução de árvores de decisão é prover um modelo de fácil interpretação no problema de domínio. O conceito de árvores de decisão de consenso [5], ainda pouco explorado, mostra um potencial ganho em relação à interpretação de modelos. Neste trabalho são realizados alguns testes com métricas para medição de robustez do conceito de árvores de decisão de consenso, bem como a aplicação de técnicas atuais [6] para interpretabilidade modelo em questão.

2 Metodologia

Foram utilizados *datasets* públicos do UCI Machine Learning Repository [7] e a linguagem de programação Python.

2.1 Análise dos *datasets*

A fim de entender as características dos datasets escolhidos, definimos certos indicadores que podem ajudar tanto na abordagem de desenvolvimento, quanto na interpretação dos resultados obtidos:

1. Quantidade instâncias
2. Quantidade atributos
3. Quantidade atributos numéricos (contínuos)
4. Quantidade atributos categóricos (mesmo aqueles que sejam codificados com números)
5. Quantidade classes (se formos trabalhar com classificação não binária)
6. Porcentagem de exemplos na classe minoritária
7. Porcentagem de exemplos na classe majoritátia

8. Porcentagem de missing values por variável (1o quartil)
9. Porcentagem de missing values por variável (2o quartil)
10. Porcentagem de missing values por variável (3o quartil)

Os indicadores 4 e 5 tem o propósito de evidenciar o desbalanceamento entre as classes, enquanto os indicadores 6, 7 e 8 servem para demonstrar a distribuição de *missing values* em cada *dataset*, esses 3 indicadores são calculados da seguinte maneira:

1. Obter um vetor com a porcentagem de instâncias com *missing values* em cada *dataset*
2. Calcular os quantis 0.25, 0.50 e 0.75 sobre esse vetor (esses seriam os quartis indicados acima)

2.2 Aplicação do modelo de *consensus decision trees*

Realizamos a aplicação do modelo afim de verificar o seu comportamento com os *datasets* escolhidos, que provém de diferentes categorias de fontes, e por sua pouca exploração até o momento.

2.3 Aplicação de técnicas atuais dos campos de XAI e iML

Enfim, realizamos a aplicação de técnicas dos campos de XAI e iML, afim de comparar os resultados obtidos de *consensus decision trees*, tendo como base principalmente os métodos mais atuais e relevantes na área [8, 6].

Referências

[1] L. Edwards and M. Veale, "Enslaving the algorithm: From a "right to an explanation"to a "right to better decisions"?," *SSRN Electronic Journal*, 01 2017.

[2] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI Magazine*, vol. 38, pp. 50–57, 2017.

[3] A. D. Selbst and J. Powles, "Meaningful information and the right to explanation," *International Data Privacy Law*, vol. 7, pp. 233–242, 12 2017.

[4] L. Edwards and M. Veale, "Slave to the algorithm? why a right to explanationn is probably not the remedy you are looking for," *SSRN Electronic Journal*, vol. 16, 12 2017.

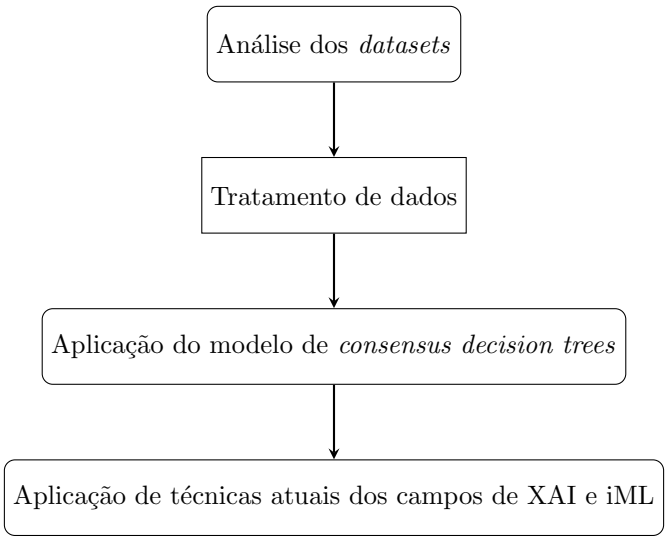

Figura 1: Ilustração do processo

[5] B. Kavšek, N. Lavrač, and A. Ferligoj, "Consensus decision trees: Using consensus hierarchical clustering for data relabelling and reduction," in *Machine Learning: ECML 2001* (L. De Raedt and P. Flach, eds.), (Berlin, Heidelberg), pp. 251–262, Springer Berlin Heidelberg, 2001.

[6] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An approach to evaluating interpretability of machine learning," *CoRR*, vol. abs/1806.00069, 2018.

[7] D. Dua and C. Graff, "UCI machine learning repository," 2017.

[8] M. Robeer, "Contrastive explanation for machine learning," 2018.