Uma Revisão sobre o uso de Frameworks de Interpretabilidade em Aprendizado de Máquina

Ivo de Abreu Araújo[1]

[1]Programa de Pós-graduação em Ciência da Computação (PPGCC) - Instituto de Ciências Exatas e Naturais (ICEN)
Universidade Federal do Pará (UFPA) – Belém, PA – Brasil

Abstract. *Machine learning models have enabled intelligent solutions in various sectors and applications of society due to their robust prediction capabilities coming from their learning processes. Thus, understanding complex model decisions become essential for confidence in the results. Thus this paper highlights a review with the objective of analyzing the use of interpretability frameworks in black-box models. The results obtained after the analysis of 143 studies confirm that interpretability in models has been consolidating through frameworks such as LIME and SHAP that are able to map possible factors that implicate in the predictive results.*

Resumo. *Os modelos de aprendizado de máquina têm possibilitado soluções inteligentes em vários setores e aplicações da sociedade devido suas capacidades de predições robustas provenientes de seus processos de aprendizagem. Dessa maneira, entender decisões de modelos complexos torna-se essencial para a confiança nos resultados. Assim este artigo destaca uma revisão com o objetivo de analisar o uso de frameworks de interpretabilidade em modelos de caixas pretas. Os resultados obtidos de após a análise de 143 estudos confirmam que a interpretabilidade em modelos vem consolidando-se por meio de frameworks como o LIME e SHAP que conseguem mapear possíveis fatores que implicam nos resultados preditivos.*

1. Introdução

Nos últimos anos a ascensão do uso de modelos de aprendizado de máquina tem provido avanços em diversas áreas em que as tomadas de decisões são críticas, tais como na medicina, no monitoramento de trânsito, sistemas judiciais, nas áreas de educação e finanças. Tendo em vista as problemáticas definidas os modelos são capazes de provêr soluções robustas com alto poder preditivo [Jordan and Mitchell 2015], [Yang et al. 2018].

No entanto, a necessidade de entender as razões que levaram as decisões de algoritmos é cada vez mais importante, especialmente quando são considerados caixas pretas como as redes neurais e *Random Forest*, os quais podem ter *performances* elevadas nos resultados, porém podem conduzir a conclusões equivocadas devida a falta de transparência que é inerente à esses modelos [Gilpin et al. 2018].

A interpretabilidade tem sido um fator relevante no contexto de modelos para revelar as compreensões que levaram as predições, e desta forma geram um grau maior de confiabilidade e usabilidade nos algoritmos que estão sendo utilizados. Por exemplo, sistemas médicos que não entregam interpretabilidade sobre os modelos de caixa preta

aplicados tendem a não gerar uma confiança elevada para o usuário, mesmo tendo altas taxas de acurácia, mas em contrapartida baixa interpretabilidade quando comparados com modelos lineares [Goodman and Flaxman 2017].

Considerando a importância de se entender as decisões dos modelos de aprendizado de máquina, este trabalho tem como propósito apresentar os resultados de uma revisão sobre *frameworks* de interpretabilidade de modelos conhecidos como caixas pretas. O foco principal concentrou-se em analisar e entender os principais *frameworks* que estão sendo utilizados para interpretar e abrir as caixas pretas que produzem as altas predições, podemos citar que o LIME (*Local Interpreatable Model-Agnostic Explanations*) e SHAP (*Shapley Additive Explanation*) foram as abordagens mais presentes na revisão.

O presente artigo está estruturado com as seguintes seções. A Seção 2 apresenta uma fundamentação teórica. Na Seção 3 temos os procedimentos de construção do protocolo da revisão sistemática. A Seção 4 apresenta a execução do protocolo da revisão sistemática. Na seção 5 é abordado sobre as respostas das questões da pesquisa. Na Seção 6 temos os resultados obtidos e as discussões. Por fim, na Seção 7 temos as Conclusões.

2. Fundamentação Teórica

2.1. Modelos de Caixas Pretas e Caixas Brancas

A capacidade de compreender modelos de aprendizado de máquina está relacionada como os resultados são passados e compreendidos pelos humanos. Deste modo, no panorama da interpretabilidade os modelos são categorizados em dois grupos.

O primeiro grupo é formado por modelos que geram soluções de fácil entendimento como as árvores de decisão, regras e regressão linear os quais são modelos de caixa branca, pois permitem conhecer indícios que contribuíram para o processo decisório das predições sendo assim, auto-interpretáveis, no entanto, são considerados não tão robustos em termos de acurácia de acordo com a complexidade do problema [Ponce and de Lourdes Martinez-Villaseñor 2017], [Luštrek et al. 2016].

O segundo grupo compreende as redes neurais, SVM (*Support Vector Machine)*, modelos *ensemble* tais como do tipo *Random Forest*, os quais trabalham como caixas pretas provisionando previsões mais precisas, porém sacrificando a interpretabilidade devido as decisões implícitas que são produzidas internamente. Dessa forma, é entregue o resutaldo, mas não as razões que levaram aos mesmos [Zhang et al. 2019].

2.2. Frameworks de Interpretabilidade

Os *frameworks* de interpretabilidade são soluções que visam entregar interpretabilidade para modelos de aprendizado de máquina, especialmente os considerados caixas pretas. Assim temos nesta seção uma breve definição do LIME e SHAP. Os *frameworks* de interpretabilidade atuam de duas maneiras, localmente e globalmente. Segundo [Molnar 2019] a interpretabilidade global está associada com um processo de compreender todo o modelo. Já a interpretabilidade local visa explicar uma parte do modelo que corresponde ao comportamento com uma predição.

O LIME é um *framework* de interpretabilidade de modelos de aprendizado de máquina que desenvolve interpretações por meio da perturbação sobre os dados de entrada utilizando um modelo linear que possibilita mapear e entender os dados que influenciam os resultados das previsões ou até mesmo ruídos [Ribeiro et al. 2016]. Desta

maneira, modelos de caixa preta podem ser desmistificados e compreendidos através de um distúrbio de dados que conduz a um modelo local interpretável.

O SHAP (*Shapley Additive Explanation*) é outra abordagem que atua no âmbito da interpretação local e global de previsões de modelos, sendo utilizado especialmente em modelos de caixas pretas e funciona de forma semelhante à teoria do jogo, no qual as *features* são consideradas jogadores que recebem um atributo chamado valor SHAPY, o qual contribui para que interpretações de modelos sejam efetuadas e explicadas de forma gráfica à usuários, desta maneira, as caixas pretas são abertas e exploradas[Lundberg and Lee 2017].

3. Estruturação do Protocolo

Esta seção compreende as etapas e recursos utilizados para construir o protocolo da revisão. As seguintes questões foram definidas para a pesquisa. Q1. Quais são os *frameworks* existentes na literatura que contribuem para o processo de interpretação de algoritmos de aprendizado de máquina? Q2. Quais as pesquisas mais relevantes sobre o desempenho de *frameworks* de interpretabilidade de modelos de aprendizado de máquina? Q3. No que diz respeito à interpretabilidade, quais os algoritmos de aprendizado de máquina estão sendo investigados? Q4. Quais são as áreas investigadas quando se avalia a interpretabilidade de algoritmos de aprendizado de máquina?

3.1. Formulação da String de Busca

Nesta etapa foram definidas 15 *strings* de busca para serem utilizadas nas bases de dados IEEE, ACM Digital e ScienceDirect. Termos como *"machine learning interpretability"* foram especificados juntamente com sinônimos com o objetivo de capturar estudos relevantes para a pesquisa. Na Tabela 1 temos as *strings* de busca que foram preparadas.

Tabela 1. Strings de Busca Escolhidas. Elaborada pelo autor.

("framework of interpretability of machine learning") OR ("tool of interpretability in machine learning") OR ("tools of explainability of machine learning models") OR ("framework of explainability of machine learning") OR ("interpretable machine learning tool") OR ("interpretable machine learning") OR ("explanable machine learning tools") OR ("explainable frameworks of machine learning") OR ("explainability of machine learning models") OR ("performance of framework of interpretability machine learning") OR ("machine learning interpretability techniques") OR ("decipherable machine learning") OR ("understandable machine learning") OR ("explicable machine learning")

4. Execução do Protocolo

4.1. Seleção dos Estudos

Para a busca dos estudos, foram realizadas pesquisas nos três repositórios escolhidos por meio das *strings* prédefinidas que possibilitaram filtrar os trabalhos candidatos pelos seguintes critérios de inclusão: presença de alguma parte da *sring* no título ou *abstract*; estudos publicados de 2015 a 2020; leitura de *abstract* para analisar se o artigo usa ou propõe um *framwework* de interpretabilidade. Após a aplicação dos filtros nos estudos o número reduziu de 143 para um total de 26.

Após o primeiro filtro resultaram 131 artigos, sendo 82 do IEEE filtrados e 49 artigos do ACM Digital. Todos os artigos selecionados nesta etapa foram estruturados

numa base de dados do *software* JabRef, o qual disponibiliza informações importantes sobre os artigos que podem ser acessadas posteriormente. Dos 26 estudos do filtro final 18 são originários da plataforma IEEE e 8 artigos provenientes do ACM Digital.

5. Respostas às Questões da Pesquisa

5.1. Q1. Quais são os frameworks existentes na literatura que contribuem para o processo de interpretação de algoritmos de aprendizado de máquina?

Concernente a primeira questão, foram encontrados os *frameworks* mais utilizados para a interpretação de modelos de aprendizado de máquina. Na Figura 1 temos um gráfico que representa os resultados da análise, no qual o LIME foi o *framework* mais utilizado tendo aparecido em 19 estudos. O SHAP foi o segundo *framework* mais frequente e apareceu em 3 estudos. Ambos os *frameworks* são considerados agnósticos, pois permitem interpretar qualquer modelo. Os demais *frameworks* como Georgias, *Anchors* estiveram presentes em apenas um estudo. Apenas dois trabalhos exploraram modelos tanto num contexto global como local [Barredo-Arrieta et al. 2019], [Messalas et al. 2019]. Assim, nota-se que houve uma prevalência da interpretabildiade local nos estudos.

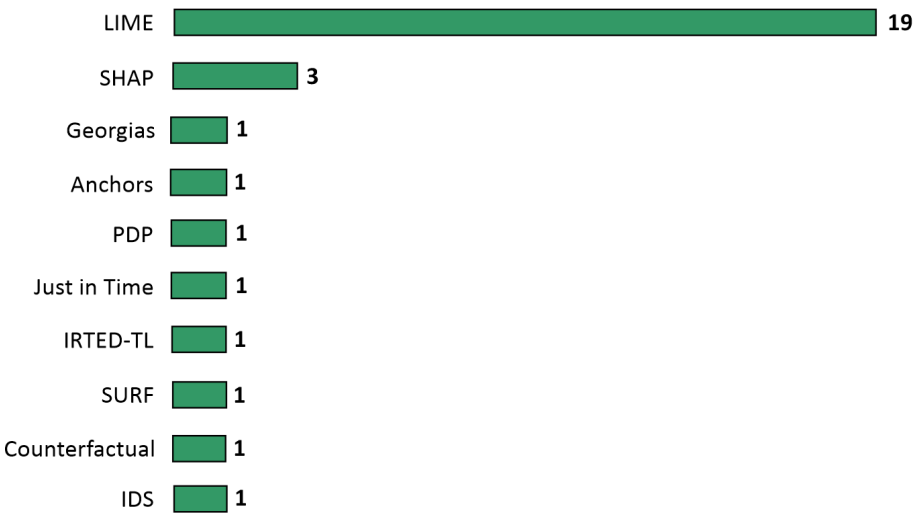

Figura 1. *Frameworks* de Interpretabilidade filtrados. Legendas: *Partial Dependence Plots* (PDP); *Inter-Region Tax Evasion Detection method based on Transfer Learning* (IRTED-TL); *Speeded-Up Robust Features* (SURF); *Interpretable Decision Sets* (IDS). Elaborada pelo autor.

5.2. Q2. Quais as pesquisas mais relevantes sobre o desempenho de frameworks de interpretabilidade de modelos de aprendizado de máquina?

Para a segunda questão alguns artigos relevantes como o estudo de [Barredo-Arrieta et al. 2019] que descreve a *performance* do SHAP *framework* aplicado nos modelos de *Random Forest* e Redes Neurais, os quais foram utilizados para a predição de tráfego em tempo real. Os resultados obtidos com o SHAP foram demonstrados por meio de uma escala gráfica de distribuição das variáveis consideradas importantes para as predições. Três *frameworks* de intepretabilidade (LIME, SHAP, *Anchors*) foram utilizados numa *Random Forest* e fizeram interpretações num *dataset* de mortalidade e diabetes e foram avaliados por meio de métricas de interpretabilidade

como identidade, estabilidade, separabilidade, similaridade, e tempo de execução. No *dataset* de interpretabilidade ambos LIME, Anchors e SHAP tiveram um desempenho satisfatório na métrica de separabilidade [El Shawi et al. 2019].

Em [Monteiro de Aquino and Cozman 2019] é desenvolvido um estudo com a proposta de comparar o desempenho de dois *frameworks* de interpretabilidade, o LIME e PDP. As interpretações coletadas por ambos *frameworks* foram inseridas num questionário e aplicadas a um grupo de usuários que atribuíram respostas relacionadas às interpretações produzidas. No geral, o LIME recebeu mais *feedbacks* positivos dos usuários devido a forma de exposição das interpretações serem mais amigáveis.

5.3. Q3. No que diz respeito à interpretabilidade, quais os algoritmos de aprendizado de máquina estão sendo investigados?

Com base nos estudos filtrados, obtivemos as respostas para a terceira questão apresentadas na Figura 2 com um gráfico que demonstra a porcentagem de algoritmos que foram utilizados a partir dos 26 estudos selecionados.

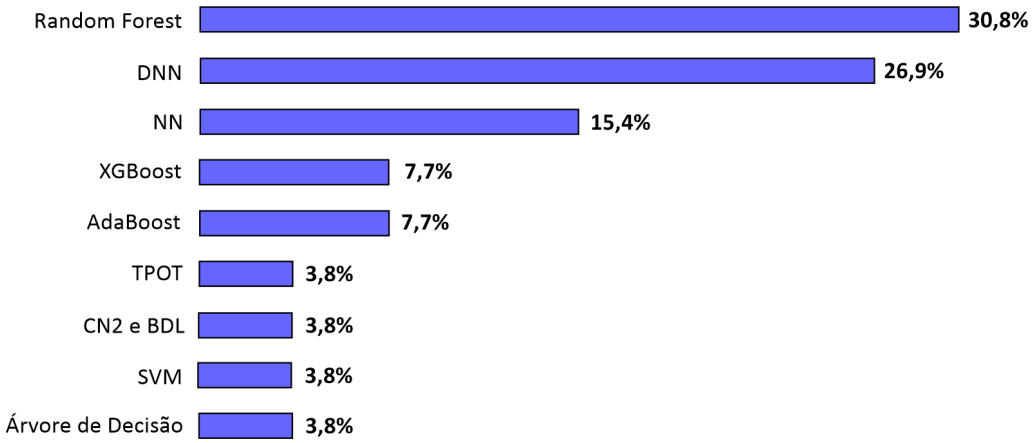

Figura 2. Algoritmos de aprendizado de máquina utilizados nos artigos selecionados. Fontes: *Deep Neural Network* (DNN); CN2 *Induction Algorithm* (CN2); *Bayesian Decision Lists* (BDL); *Neural Network* (NN); *Tree-based Pipeline Optimization Tool* (TPOT). Elaborada pelo autor.

O tipo de algoritmo mais utilizado de acordo com os estudos obtidos foi do tipo *Random Forest* com um percentual de 30,7%, seguido por Redes Neurais (15,3%) e as Redes Neurais Convolucionais com 11,5%. Podemos notar que 92% dos algoritmos selecionados são do tipo caixa preta. Este resultado demonstra que uma solução mais robusta como uma caixa preta aliada com a interpretabilidade têm sido buscada por meio de *frameworks* como o LIME.

5.4. Q4. Quais são as áreas investigadas quando se avalia a interpretabilidade de algoritmos de aprendizado de máquina?

A respeito dos contextos em que estão sendo utilizados os *frameworks* de interpretabilidade. Há várias áreas em que as soluções de aprendizado de máquina estão sendo importantes de serem robustas e interpretáveis para que a confiança nas soluções seja maior, como por exemplo na área médica e judicial. Na Figura 3 temos as áreas em que estão sendo aplicados os *frameworks* de interpretabilidade e percebe-se que não apenas na área

da saúde é importante entender o modelo utilizado, mas outras áreas que envolvem o fator humano é cada vez mais importante apoiar decisões de forma mais transparente.

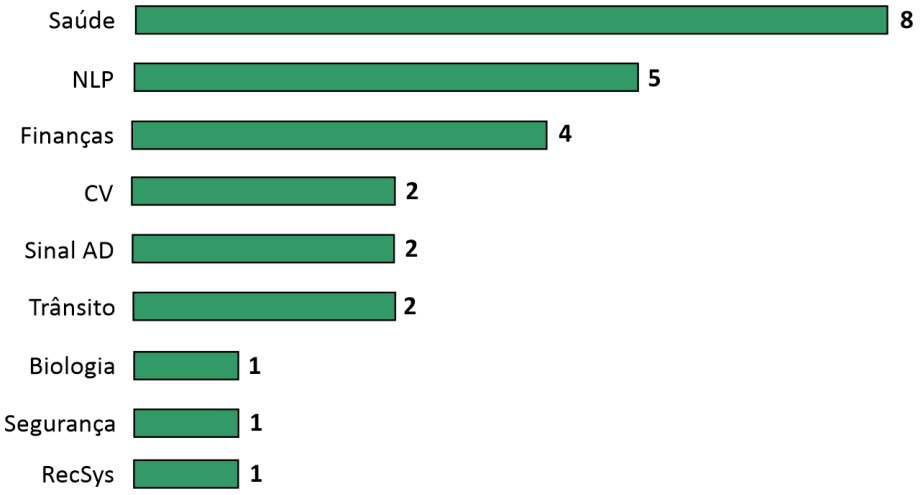

Figura 3. Áreas exploradas pelos frameworks de interpretabilidade. Legendas: *Natural Language Processing* (NLP); *Computer Vision* (CP); Sinal Analógico/Digital (AD); *Recommender Systems* (RecSys). Elaborada pelo autor.

Ainda na Figura 3 temos um total de 9 áreas que foram beneficiadas pelos *frameworks* de interpretabilidade. No geral, o campo da saúde foi o que mais ficou em evidência, sendo parte de 8 estudos (30%), notamos que o campo de finanças também vem sendo fundamental o entendimento das razões das predições. É notável também que diversos outros segmentos têm buscado elucidar importantes decisões a partir de modelos de caixas pretas interpretados. Estudos conduzidos por [El Shawi et al. 2019], [Lakkaraju et al. 2016] [Mothilal et al. 2020] que exploraram a interpretabilidade em aprendizado de máquina em mais de uma área de atuação.

6. Resultados e Discussões

Nesta seção temos sumarizado os resultados gerais e discussões obtidas com a revisão, os quais complementam para uma visão mais holística das informações contempladas. Na Figura 4 podemos notar a relação da quantidade de artigos distribuídos por ano. É possível perceber que no ano de 2019 houve um destaque no número de 16 publicações, o qual corresponde a 61% dos artigos selecionados. Assim, podemos observar que há um crescimento na busca por interpretabilidade em aprendizado de máquina devido a importância de entender modelos. É relevante mencionar que não houve interseção de estudos entre as bases selecionadas e que não houve nenhum trabalho com caráter investigativo como desta revisão.

O presente trabalho permite constatar que conhecer soluções de interpretabilidade como *frameworks* citados na seção 5.1 são uma alternativa interessante quando decisões de algoritmos de caixa preta impactam na vida humana ou até mesmo precisam ser melhor compreendidas por cientistas de dados e especialistas que visam ter um domínio mais seguro das ações de um modelo [Malhi et al. 2019]. Assim, temos consequentemente uma possibilidade maior de confiança em resultados de aprendizado de máquina que entregam interpretabilidade aliada a um nível de desempenho considerável de uma caixa preta. Outro ponto a se destacar é o foco no uso de *frameworks* para interpretações locais. Logo é

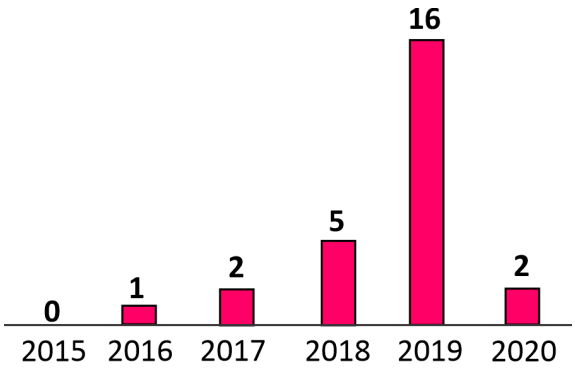

Figura 4. Etapa de seleção dos estudos. Elaborada pelo autor.

possível perceber que tem havido uma demanda maior para interpretações em predições individuais, pois entender uma predição específica é menos complexo do que entender o comportamento global de um modelo.

7. Conclusões

Neste artigo foi conduzida uma revisão da literatura que destacou o uso de *frameworks* de interpretabilidade para interpretar decisões de modelos de aprendizado de máquina. O uso de *frameworks* como soluções de interpretabilidade de modelos de aprendizado de máquina são interessantes em domínios de decisões críticas, pois o entendimento de predições de caixas pretas promove ações mais assertivas, como, por exemplo, identificar anomalias no modelo que afetam a predição para corrigi-lás de modo a mitigá-las e melhorar o desempenho do modelo. Uma vez que a acurácia pode não ser suficiente para compreender uma predição, pois quando um cientista de dados cria um solução de aprendizado de máquina, ele está confiando mesmo que implicitamente no modelo que pode ser tornar mais transparente com o uso de *frameworks* de interpretabilidade.

Por fim, por meio deste estudo foi cumprido os propósitos estabelecidos inicialmente com a apresentação dos *frameworks* de interpretabilidade envolvidos em aplicações de aprendizado de máquina que visam tornar as suas soluções mais significativas e transparentes para as decisões futuras. Uma limitação encontrada no trabalho é o número de apenas 10 *frameworks* de interpretabilidade terem sido filtrados. Como trabalhos futuros deseja-se uma escolha maior de bases e *strings* de busca a fim de encontrar novas propostas de interpretabilidade de modelos. Pretende-se também fazer uma avaliação da interpretabilidade de *frameworks* em modelos de caixa preta com o objetivo de investigar se há alguma semelhança em ambos nas interpretações.

Referências

Barredo-Arrieta, A., Laña, I., and Del Ser, J. (2019). What lies beneath: A note on the explainability of black-box machine learning models for road traffic forecasting. In *Proc. IEEE Intelligent Transportation Systems Conf. (ITSC)*, pages 2232–2237.

El Shawi, R., Sherif, Y., Al-Mallah, M., and Sakr, S. (2019). Interpretability in healthcare a comparative study of local machine learning interpretability techniques. In *Proc. IEEE 32nd Int. Symp. Computer-Based Medical Systems (CBMS)*, pages 275–280.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *Proc. IEEE 5th Int. Conf. Data Science and Advanced Analytics (DSAA)*, pages 80–89.

Goodman, B. and Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57.

Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260.

Lakkaraju, H., Bach, S. H., and Leskovec, J. (2016). Interpretable Decision Sets: A Joint Framework for Description and Prediction. KDD '16, pages 1675–1684, San Francisco, California, USA. Association for Computing Machinery.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774.

Luštrek, M., Gams, M., Martinčić-Ipšić, S., et al. (2016). What makes classification trees comprehensible? *Expert Systems with Applications*, 62:333–346.

Malhi, A., Kampik, T., Pannu, H., Madhikermi, M., and Främling, K. (2019). Explaining machine learning-based classifications of in-vivo gastral images. In *2019 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7. IEEE.

Messalas, A., Kanellopoulos, Y., and Makris, C. (2019). Model-agnostic interpretability with shapley values. In *Proc. Systems and Applications (IISA) 2019 10th Int. Conf. Information, Intelligence*, pages 1–7.

Molnar, C. (2019). *Interpretable machine learning: A Guide for Making Black Box Models Explainable*. Leanpub.

Monteiro de Aquino, R. and Cozman, F. (2019). Natural language explanations of classifier behavior. In *Proc. IEEE Second Int. Conf. Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 239–242.

Mothilal, R. K., Sharma, A., and Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. FAT* '20, pages 607–617, Barcelona, Spain. Association for Computing Machinery.

Ponce, H. and de Lourdes Martinez-Villaseñor, M. (2017). Interpretability of artificial hydrocarbon networks for breast cancer classification. In *Proc. Int. Joint Conf. Neural Networks (IJCNN)*, pages 3535–3542.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?"explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Yang, C., Rangarajan, A., and Ranka, S. (2018). Global model interpretation via recursive partitioning. In *Proc. IEEE 20th Int. Conf. High Performance Computing and Communications; IEEE 16th Int. Conf. Smart City; IEEE 4th Int. Conf. Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1563–1570.

Zhang, W., Zhou, Y., and Yi, B. (2019). An interpretable online learner's performance prediction model based on learning analytics. In *Proceedings of the 2019 11th International Conference on Education Technology and Computers*, pages 148–154.