Análise de Explicabilidade de um Modelo de Aprendizado de Máquina para Aplicações Industriais

Ramon Gomes Durães * **Turíbio Tanus Salis** *
Frederico Gualberto Ferreira Coelho * **Antônio de Pádua Braga** *

* *Escola de Engenharia, Universidade Federal de Minas Gerais, MG, Brasil (e-mails: ramonduraes, turibiotanussalis, fredgfc, apbraga@ufmg.br).*

Abstract: Machine learning models are used in numerous applications and, for some of them, it is desirable to be able to explain the output of the models. The search for model explainability lead to the development of *Tree SHAP*: a framework for tree-based models that calculates the contribution of each input feature to the predictions. In this paper, we train a regression model that predicts the yield value of seamless steel tubes at the end of the heat treatment process. We then perform a model explainability analysis to highlight the gains of such analysis in industrial applications.

Resumo: Modelos de aprendizado de máquina estão sendo utilizados em cada vez mais aplicações e, para algumas delas, é desejável que as predições feitas pelo modelo sejam explicáveis. A busca pela explicabilidade de modelos levou ao desenvolvimento do método *Tree SHAP*, que calcula a contribuição individual de cada característica de entrada de modelos baseados em árvores para o resultado obtido. Neste trabalho, é treinado um modelo de regressão para prever o limite de escoamento de tubos de aço sem costura ao final do processo de tratamento térmico. Em seguida, é feita uma análise de explicabilidade do modelo para ressaltar o valor que este tipo de análise pode agregar em aplicações industriais.

Keywords: Machine Learning, Explainability, Shapley Values, SHAP, Tree SHAP, Industrial Applications.

Palavras-chaves: Aprendizado de Máquina, Explicabilidade, Shapley Values, SHAP, Tree SHAP, Aplicações Industriais.

1. INTRODUÇÃO

O Aprendizado de Máquina (AM) é uma área da inteligência artificial que utiliza algoritmos computacionais para transformar dados empíricos em modelos preditivos (Edgar and Manz, 2017). Alguns tipos de modelos, chamados de "caixa preta", são complexos o suficiente para que suas predições não sejam interpretáveis por humanos (Petch et al., 2022). Entretanto, a crescente utilização de modelos de AM na resolução de problemas evidenciou aplicações para as quais não se deve confiar cegamente nas predições geradas, a exemplo de aplicações médicas (Kundu, 2021).

O estudo de explicabilidade de modelos culminou em métodos que auxiliam na extração de conhecimento dos modelos de diferentes formas, cada uma com seus pontos fortes e fracos. Em Belle and Papantonis (2021), os autores citam alguns desses métodos e os categorizam em relação a transparência, critério de avaliação e tipo de explicações.

Atualmente, um dos mais relevantes métodos para explicar modelos de aprendizado de máquina é o *SHapley Additive exPlanations (SHAP)* (Lundberg and Lee, 2017). Ele é inspirado no conceito de *Shapley Values* advindo da Teoria de Jogos, desenvolvido por Shapley (1951), que busca quantificar a contribuição de cada jogador para o resultado final de um jogo. No contexto de AM, essa busca se traduz em calcular a contribuição de cada variável de entrada do modelo para a saída gerada. Dois pontos positivos deste método relevantes para este trabalho são sua facilidade de utilização e sua capacidade de gerar gráficos que viabilizam a extração de conhecimento do modelo e suas predições de forma visual e intuitiva.

Apesar de sua popularidade no meio acadêmico, a adoção de métodos como o *SHAP* ainda não é comum nas indústrias de manufatura. Em realidade, pesquisas de mercado que analisam a adoção de inteligência artificial em empresas, como a realizada por Loukides (2022), da O'Reilly Media Inc., mostram que o setor de manufatura é um dos que menos investe em IA. Glover (2021), da Tech Monitor, argumenta que a falta de expertise dos profissionais e utilização de tecnologias antigas representam barreiras para a adoção de IA. A Boston Consulting Group (BCG) analisou especificamente a indústria siderúrgica, adicionando a cultura das empresas e a maturidade em dados como desafios a serem superados (Rodriguez et al., 2021).

Neste trabalho, é apresentado um estudo de caso da indústria siderúrgica que utiliza dados reais do processo de tratamento térmico de tubos de aço sem costura. É treinado um modelo de regressão para prever o limite

de escoamento, uma propriedade mecânica dos tubos de aço, ao final do tratamento térmico. Em seguida, utiliza-se uma variação do método *SHAP* específica para modelos baseados em árvore, o *Tree SHAP* (Lundberg et al., 2020), para extrair conhecimento da aplicação por meio da explicação do modelo treinados. Destrinchamos a análise dos gráficos gerados pelo método a fim de ressaltar o valor que este tipo de análise pode agregar em aplicações industriais.

O restante deste trabalho está organizado como segue: na Seção 2 é feita uma revisão bibliográfica a respeito do método *SHAP*, suas origens e características. A Seção 3 descreve o processo estudado neste trabalho. A Seção 4 apresenta e discute os resultados do estudo de casos. A Seção 5 encerra o trabalho fazendo as considerações finais.

2. REVISÃO BIBLIOGRÁFICA

2.1 Shapley Values

Na teoria de jogos, define-se como um *jogo cooperativo* um conjunto de circunstâncias nas quais dois ou mais jogadores contribuem para alcançar um resultado final. Neste contexto, os *Shapley Values* foram propostos por Shapley (1951) como uma forma justa de quantificar a contribuição marginal de cada membro do time para o resultado final alcançado. Em outras palavras, os *Shapley Values* são o "pagamento" para cada jogador de acordo com sua contribuição para o resultado do time.

Matematicamente, considere um time T com p jogadores que obteve um valor final $v = v(T)$ em um jogo. O *Shapley Value* $\varphi_m(v)$ atribuído a cada jogador m é definido como:

$$\varphi_m(v) = \frac{1}{p} \sum_S \frac{[v(S) \cup \{m\} - v(S)]}{\frac{p-1}{k(S))}} \tag{1}$$

em que S representa todos os subconjuntos do time $T = 1, 2, 3...p$ que podem ser construídos excluindo o jogador de interesse m; $k(S)$ é o tamanho de S, $v(S)$ é o valor atingido pelo sub-time S e $v(S \cup m)$ é o valor atingido pelo sub-time S após a adição do jogador m. De maneira intuitiva, o *Shapley Value* de cada jogador é medido adicionando-o e removendo-o de todos os possíveis subconjuntos dos demais jogadores e calculando a soma ponderada de sua contribuição nestes subconjuntos.

Os *Shapley Values* possuem a importante propriedade de serem a única forma comprovada matematicamente de calcular a contribuição individual dos jogadores garantindo alguns axiomas importantes Shapley (1951). Entre eles, cita-se:

- *Simetria*: se dois jogadores i e j contribuírem o mesmo tanto para o resultado final, eles devem receber o mesmo valor;
- *Jogadores dummy*: os jogadores que não contribuírem para o resultado final devem receber um valor igual a zero;
- *Monotonicidade*: se um jogador i contribui consistentemente mais para o resultado final que um jogador j, o valor atribuído ao jogador i deve refletir isso e ser maior que o valor atribuído ao jogador j;
- *Linearidade*: Se for possível separar um jogo em duas partes, a soma dos valores alocados a cada jogador em cada etapa do jogo deve ser igual ao valor alocado ao jogador considerando o jogo inteiro;

Em aprendizado de máquina, Rozemberczki et al. (2022) mostram que os *Shapley Values* foram utilizados em diversos contextos. Para este trabalho, destaca-se sua utilização para explicabilidade de modelos, na qual os jogadores equivalem às características de entrada de um modelo de AM e o valor atingido pelo time equivale à saída (ou predição) do modelo. Dessa forma, os *Shapley Values* atribuem a cada característica de entrada sua contribuição individual para a saída gerada. Nota-se que para isso a saída do modelo deve ser um valor escalar, então utiliza-se a probabilidade da classe nos problemas de classificação.

2.2 SHapley Additive ExPlanations (SHAP)

Apesar das suas propriedades e garantias teóricas, na prática o cálculo exato dos *Shapley Values* é custoso: a complexidade de (1) é exponencial com o número de entradas dos modelos de AM (Lundberg et al., 2019). Por isso alguns algoritmos inspirados nessa técnica implementam simplificações para calcular valores aproximados.

É o caso do método *SHapley Additive exPlanations (SHAP)* proposto por Lundberg and Lee (2017). Nele, define-se *SHAP values* como os *Shapley values* da função de esperança condicional do modelo original, o que mantém as propriedades dos axiomas mostrados na Seção 2.1, mantendo também sua complexidade computacional. Entretanto, são apresentadas duas formas de calcular valores aproximados dos *SHAP values* que independem do modelo de AM utilizado. Isso é feito assumindo independência entre as características de entrada e linearidade dos modelos. Dessa forma, estima-se o impacto de variações locais em cada entrada do modelo separadamente, seja por amostragem ("*Shapley sampling values*") ou por linearização utilizando *kernels* ("*Kernel SHAP*").

Apesar da possibilidade de se calcular aproximações dos *SHAP values* de forma agnóstica, o conhecimento da estrutura do modelo de AM utilizado possibilita melhorias no algoritmo. Como exemplo, cita-se o "*Deep SHAP*": uma formulação específica para redes neurais profundas apresentada no artigo original, para a qual não é necessário assumir independência das características nem linearidade do modelo.

2.3 TreeExplainer

Modelos baseados em árvores como *Random Forests* e *Gradient Boosted Trees* têm um longo histórico de utilização em aprendizado de máquina. Eles performam particularmente bem quando aplicados a dados em formato tabular, nos quais as características de entrada são individualmente importantes, sem fortes estruturas temporais ou espaciais. Nestes casos, esses modelos são considerados o estado-da-arte para muitas aplicações, se mostrando consistentemente melhores que regressões lineares ou redes neurais profundas Chen and Guestrin (2016).

Apesar de modelos de árvores simples serem altamente interpretáveis, a combinação de um comitê dessas árvores,

como nas *Random Forests*, melhora a acurácia das predições em detrimento de sua explicabilidade. O método *TreeExplainer* (Lundberg et al., 2020) é proposto para lidar com este problema. Ele é capaz de calcular os *SHAP values* de forma exata, provendo explicações locais ótimas que mantêm as propriedades dos axiomas dos *Shapley Values* mostradas na Seção 2.1. O algoritmo tem complexidade polinomial de baixa ordem, o que representa um grande avanço em relação a outros métodos exatos de cálculo de *Shapley Values* cuja complexidade é exponencial. Isso possibilita análises não só de uma observação mas do conjunto de dados como um todo.

A definição exata do algoritmo utilizado, denominado *Tree SHAP*, foge ao escopo deste trabalho e pode ser encontrada no artigo que propõe o método (Lundberg et al., 2020). Entretanto, vale entender seu funcionamento de forma intuitiva.

Em sua forma mais simples, ainda com complexidade exponencial, o *Tree SHAP* estima os *SHAP values* percorrendo recursivamente o caminho da observação de interesse na estrutura da árvore para cada subconjunto S de características de entrada. Caso o nó atual seja uma folha, é retornado o valor da folha; caso contrário, o algoritmo desce um nível na estrutura da árvore escolhendo o nó subsequente (da esquerda ou da direita) de acordo com o *threshold* do nó. Caso a característica avaliada não esteja presente no subconjunto analisado, utiliza-se uma média ponderada dos valores dos nós subsequentes. A complexidade deste algoritmo é $O(TLM2^M)$, em que T é o número de árvores no comitê, L é o número de folhas e M representa o número de características do modelo.

Um refinamento deste algoritmo permite a obtenção dos mesmos resultados mas com complexidade polinomial de baixa ordem. Para isso, em cada ramificação da árvore no algoritmo descrito acima, é acumulada na memória a proporção de possíveis subconjuntos de características que são levadas para as ramificações subsequentes. Os autores mostram que isso é similar a executar o algoritmo simplificado para todos os 2^M subconjuntos de características simultaneamente. Isso resulta numa complexidade de tempo da ordem $O(TLD^2)$, em que D é a profundidade das árvores, ao troco de um aumento na complexidade de memória para $O(D^2 + M)$.

3. ESTUDO DE CASO

O caso em estudo se refere a uma planta de tratamento térmico para a produção de tubos de aço sem costura. Ela é composta principalmente de um tanque de resfriamento e de um forno de temperamento. Ao longo do processo, os tubos são aquecidos e logo em seguida resfriados abruptamente para conferir a eles determinadas propriedades mecânicas. Devido à formação de austenita, inevitável durante este tratamento (Stein et al., 2005), este processo é também chamado de austenitização. Já na operação de temperamento, o tubo é reaquecido por um intervalo de tempo para ajustar as tensões internas decorrentes do processo anterior. Após passarem por estes processos, os tubos de aço são então submetidos a uma inspeção de qualidade e testes laboratoriais.

Este tratamento térmico altera a microestrutura dos metais, que por sua vez altera suas propriedades mecânicas finais. No caso dos tubos de aço, essas propriedades dependem principalmente da geometria do tubo, da composição química da liga de aço e das condições de tratamento térmico às quais os tubos foram submetidos durante seu processo de fabricação (Gomes et al., 2020). Assim, é possível produzir tubos com diferentes características de acordo com a necessidade de sua aplicação. Nesta planta, são produzidos 3 tipos diferentes de tubos, cujas propriedades atendem às especificações das indústrias de óleo e gás.

Para este estudo, a propriedade mecânica de interesse é o limite de escoamento, também chamado de tensão de cedência, tensão de limite elástico (em Portugal), tensão de escoamento (no Brasil) ou ainda limite de elasticidade aparente. Ele é definido como a tensão máxima (em MPa) que o material suporta, ainda no regime elástico de deformação. A partir deste limite, se houver algum acréscimo de tensão o material passa a sofrer deformação plástica (definitiva) (Carneiro et al., 2021).

Com o objetivo de prever o limite de escoamento de cada tubo produzido, foi desenvolvido um modelo de aprendizado de máquina utilizando dados coletados ao longo de aproximadamente 2 anos de execução do processo descrito. A Tabela 1 contém o nome, descrição e unidade de medida de cada variável do conjunto de dados. São utilizadas informações relativas às propriedades químicas da liga metálica e às características físicas de cada tubo produzido (e.g.: seu diâmetro). Além destas, são utilizadas também variáveis que caracterizam os processos de austenitização e temperamento. Entre elas, cita-se o parâmetro de *Tsuchiyama*: uma medida proposta por Tsuchiyama (2002) para melhor caracterizar os ciclos térmicos dos processos, dado que a temperatura alvo não é atingida instantaneamente ao longo de cada ciclo. De acordo com Gomes et al. (2010), este parâmetro tem boas correlações com as propriedades mecânicas de materiais tratados termicamente.

4. RESULTADOS E DISCUSSÃO

4.1 Materiais

Para a realização do estudo a seguir, foi utilizada a versão 3.9 da linguagem Python. O modelo de aprendizado de máquina é proveniente da biblioteca *scikit-learn* (Pedregosa et al., 2011). Para as análises de explicabilidade, é utilizado o pacote *shap*, disponibilizado pelos autores do método. Os métodos específicos para a implementação do *TreeExplainer* (Lundberg and Lee, 2017; Lundberg et al., 2020) também foram incorporados como módulos do pacote. Ambos os pacotes estão disponíveis através do gerenciador de pacotes *PIP* (*Python Package Index*).

Como explicado na Seção 3, os dados utilizados são provenientes de um processo real de temperamento de tubos de aço sem costura. As variáveis de entrada e saída do modelo estão descritas na Tabela 1. A base de dados contém 834 observações, sendo que cada uma representa um tubo produzido.

Tabela 1. Descrição do conjunto de dados utilizado

Variável de Entrada	**Descrição**	**Unidade**
DIAMETER	Diâmetro do tubo de aço	mm
WALL_THICKNESS	Espessura da parede do tubo de aço	mm
CEQ	Conteúdo de carbono equivalente da liga metálica	%
JEC_AUST_SOAKING_TIME	Tempo de imersão do aço no processo de austenização	s
JEC_AUST_TSUCHIYAMA	Parâmetro de Tsuchiyama no processo de austenização	-
JEC_AUST_TUBE_TEMP_MEAN	Temperatura de saída do tubo no processo de austenização	°C
JEC_REV_SOAKING_TIME	Tempo de imersão do aço no processo de temperamento	s
JEC_REV_TSUCHIYAMA	Parâmetro de Tsuchiyama no processo de temperamento	-
JEC_REV_TUBE_TEMP_MEAN	Temperatura de saída do tubo no processo de temperamento	°C
JEC_TANK_IMERSION_TIME	Tempo de imersão do tubo no tanque de resfriamento	s
JEC_TANK_INT_FLW	Taxa de fluxo de água no tanque de resfriamento	l/s
PCM	Percentual de carbono equivalente, calculado como proposto por Ito and Bessyo (1968)	%
PER_{ELEMENTO}	Composição química da liga metálica (total de 14 elementos). Nomenclatura: PER_{ELEMENTO}, em que {ELEMENTO} é o símbolo do elemento químico (e.g.: "PER_C" para percentual de carbono)	%
Variável de Saída	**Descrição**	**Unidade**
FE_LIMIT	Limite de escoamento dos tubos de aço	MPa

4.2 Treinamento do Modelo

Foi utilizado o modelo *CatBoostRegressor*, da biblioteca *scikit-learn*, para a predição do limite de escoamento dos tubos de aço. O conjunto de dados foi dividido em 70% das observações para treinamento e 30% para validação. A taxa de aprendizado utilizada é de 0.05 e a profundidade máxima de cada árvore treinada é 5. A função de perda utilizada é a raiz do erro quadrático médio (*RMSE*). Esta mesma função é utilizada como métrica de avaliação para a seleção do melhor modelo. O treinamento ocorre ao longo de 2000 iterações e o melhor resultado para o conjunto de validação é atingido na iteração 500, para a qual o *RMSE* é de 12.0945191. Esta performance foi considerada satisfatória para a aplicação.

Nota-se que o foco deste trabalho é a análise de explicabilidade do modelo, não seu processo de treinamento em si. Entretanto, é importante ressaltar que essa análise é altamente dependente de um modelo ajustado adequadamente para o problema em estudo.

4.3 Análise de Explicabilidade

A análise de explicabilidade do modelo treinado é realizada por meio dos gráficos gerados pelo pacote *SHAP*, que possibilitam uma análise visual e intuitiva dos resultados. Eles podem ser obtidos para problemas de regressão e de classificação, mas alguns deles são específicos de para modelos baseados em árvores.

A Figura 1 mostra o gráfico de forças para uma observação do conjunto de teste. O gráfico de forças, introduzido por Lundberg et al. (2018), explica a saída do modelo para uma observação de interesse em relação a um valor base. Este valor base é simplesmente a saída média do modelo para o conjunto de treinamento. Ele exibe a influência do valor de cada característica de entrada (i.e. cada segmento da barra) como se fosse uma força que contribui positiva (em vermelho) ou negativamente (em azul) no deslocamento da saída em relação à média. Desta forma, o gráfico evidencia tanto a direção da contribuição de cada valor de entrada quanto a magnitude desta contribuição, através do tamanho de cada segmento de barra.

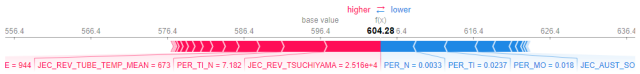


Figura 1. Gráfico de forças para uma das observações do conjunto de teste.

Para o exemplo mostrado na Figura 1, o limite de escoamento predito foi de 604.28. Nota-se que a variável de entrada que mais contribui para o aumento da predição em relação à média é o parâmetro de Tsuchiyama do processo de temperamento (i.e.: JEC_REV_TSUCHIYAMA). A variável que mais contribui para abaixar o limite de escoamento é o percentual de nitrogênio da liga metálica (i.e.: PER_N).

Para melhorar a inspeção visual para uma observação, o gráfico de forças pode ser separado verticalmente para as variáveis de entrada mais relevantes, gerando os gráficos *Waterfall*. A Figura 2 mostra dois exemplos destes gráficos. Neles, as 9 variáveis mais relevantes são mostradas separadamente, enquanto o efeito combinado das demais é mostrado na última barra, na parte inferior. Nota-se que para a observação mostrada à direita, não são mostrados valores que aumentem média da saída e a vairável mais relevante é o do percentual de carbono da liga de aço (i.e.: PER_C).

Percebe-se que a variável mais relevante para o modelo varia de acordo com a observação em análise. Logo não pode-se interpretar a maior barra de cada gráfico mostrado acima como a variável mais relevante para o modelo. Para visualizar as variáveis mais importantes para o modelo, levando em consideração todo o conjunto de dados, utiliza-se o gráfico *beeswarm*. A Figura 3 mostra este gráfico, calculado para todo o conjunto de treinamento do problema em estudo, em que cada ponto é uma observação. No eixo vertical, as características são exibidas separadamente, começando pela mais relevante para o modelo. No eixo horizontal, a posição dos pontos indica a contribuição do valor da entrada para a saída do modelo (i.e. seu *SHAP value*). Os valores das características em cada observação são indicados pela cor de cada ponto: azul para valores

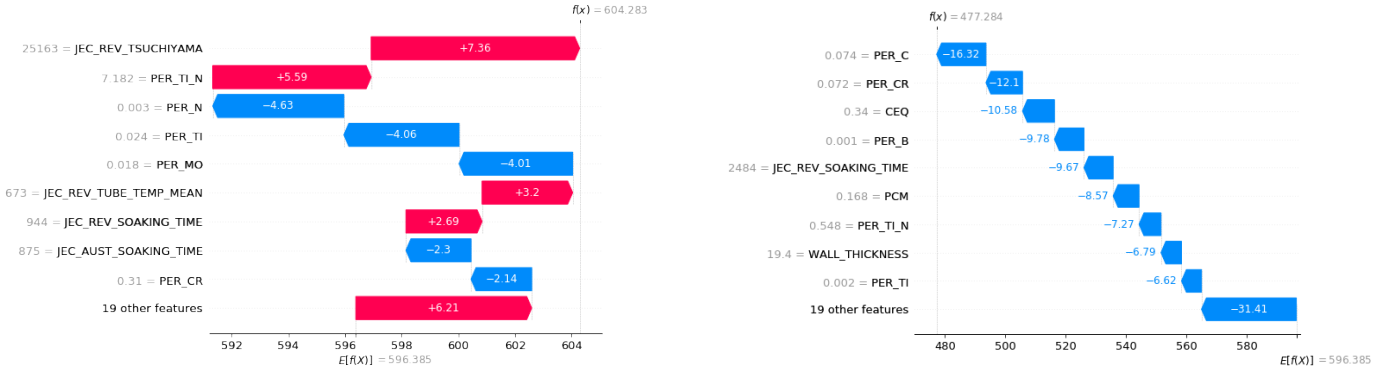


Figura 2. Gráfico Waterfall para duas observações do conjunto de dados.

baixos e vermelho para valores altos em relação à média dos dados.

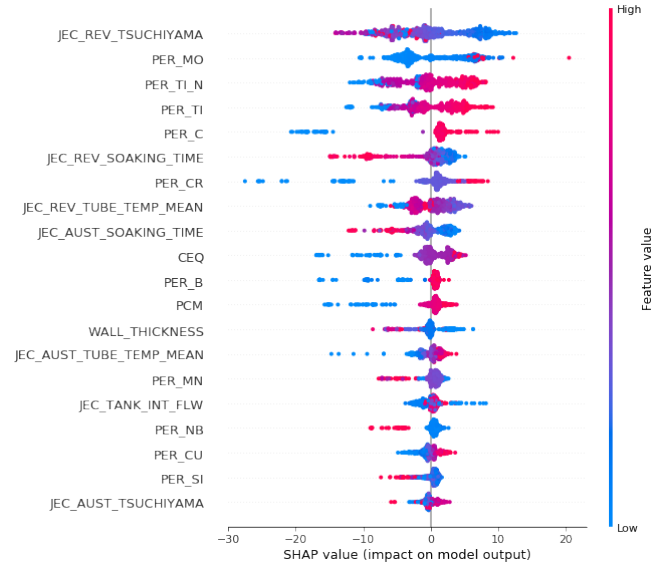


Figura 3. Gráfico Beeswarm para o conjunto de dados de validação.

Para o processo de tratamento térmico em estudo, a Figura 3 destaca o parâmetro de Tsuchiyama do processo de temperamento (i.e.: JEC_REV_TSUCHIYAMA) como a entrada mais relevante para o limite de escoamento dos tubos produzidos. A figura parece mostrar também que, para esta variável, quanto mais baixo o valor do parâmetro de Tsuchiyama, mais ele contribui para o aumento do limite de escoamento dos tubos. Entretanto, não há uma separação clara de cores como no caso do percentual de carbono (i.e.: PER_C), logo não se pode uma relação monotônica entre as variáveis. Esta observação é confirmada pela análise da Figura 4b a seguir.

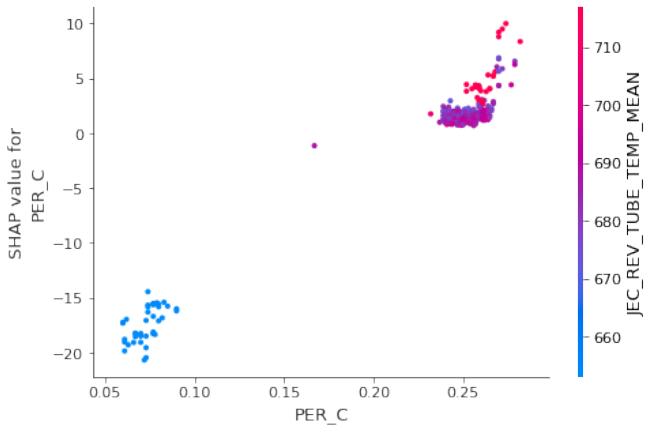
Uma outra forma de analisar todo o conjunto de dados é utilizando o gráfico de dependência. Ele permite a inspeção da influência de uma característica de entrada específica para a saída. No eixo horizontal, são mostrados os valores da variável e no eixo vertical são mostrados os *SHAP values* representando o quanto o conhecimento do valor da variável altera a predição para cada observação. Também é possível exibir efeitos de interação entre duas variáveis quaisquer através das cores dos pontos. A Figura 4 mostra dois gráficos de dependência para o problema em estudo.

Analisando a Figura 4a, percebe-se indícios de uma relação linear entre o percentual de carbono na liga metálica e a saída do modelo: quanto maior o percentual de carbono na liga, maior o limite de escoamento do tubo. Analisando as cores dos pontos, percebe-se também que o percentual de carbono da liga também é correlacionado com a temperatura do tubo no processo de temperamento: quanto maior o valor da variável PER_C, mais os pontos se tornam vermelhos (indicando aumento na temperatura de temperamento).

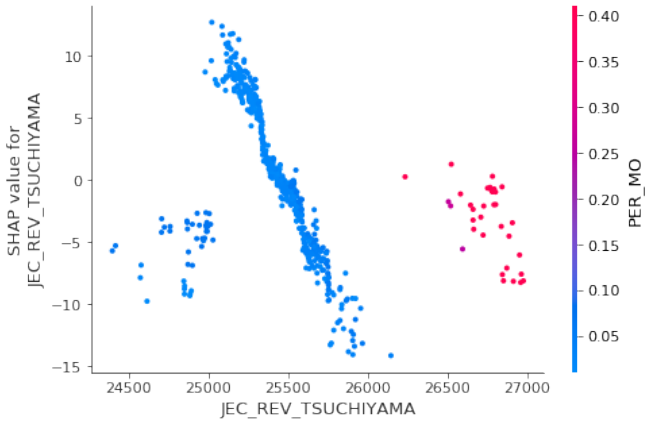
A Figura 4b analisa a variável de entrada destacada como a mais importante para o modelo pela Figura 3: o parâmetro de Tsuchiyama do processo de temperamento (i.e.: JEC_REV_TSUCHIYAMA). Pelas cores dos pontos, percebe-se que há uma interação significativa entre esta variável e o percentual de molibdênio (i.e.: variável PER_MO) da liga metálica: os pontos com percentual de molibdênio acima da média aparecem apenas para os valores mais altos do parâmetro de Tsuchiyama. É interessante notar ainda que esta figura revela que não há uma relação linear entre o parâmetro de Tsuchiyama e a saída do modelo. A contribuição desta variável para a saída do modelo é positiva para uma faixa de valores entre aproximadamente 25000 e 25500. Para valores acima e abaixo desta faixa, a contribuição é negativa.

Por fim, utiliza-se também Gráficos de Decisão para extrair conhecimento do modelo. Eles podem ser gerados para uma ou mais observações do conjunto de dados, sendo que cada observação será representada por uma linha vertical. Cada linha tem uma cor fixa que indica o valor da saída do modelo para aquela observação. Além das linhas coloridas, há sempre uma reta preta no gráfico que indica o valor base do modelo, o valor médio da saída no conjunto de treinamento, que serve como referência para a análise. As linhas então partem do valor de repouso e oscilam, de baixo para cima, acumulando os efeitos de cada variável para a saída do modelo. Exemplos deste tipo de gráfico para o caso em estudo são mostrados na Figura 5.

Na Figura 5a, percebe-se que a saída do modelo para observações distintas percorre caminhos diferentes da média até o valor predito, promovidos por valores distintos das variáveis de entrada. Pode-se verificar também como cada característica influencia a saída do modelo e em qual inten-

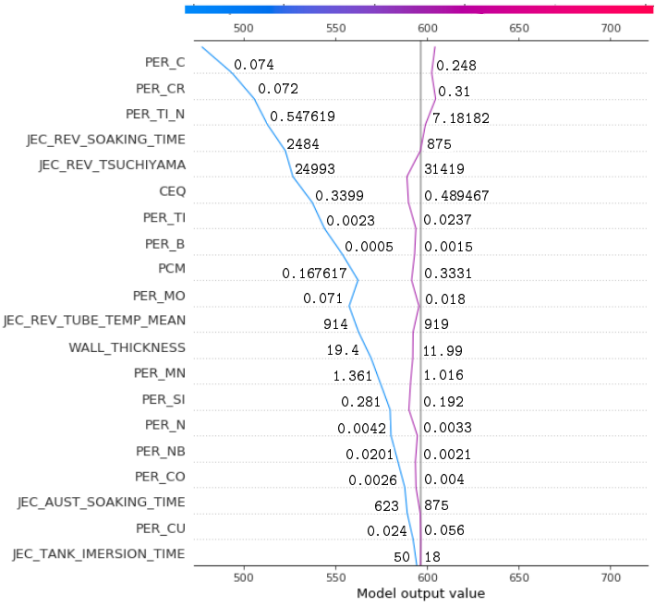


(a) Análise da variável PER_C mostrando interações com a variável JEC_REV_TUBE_TEMP_MEAN.

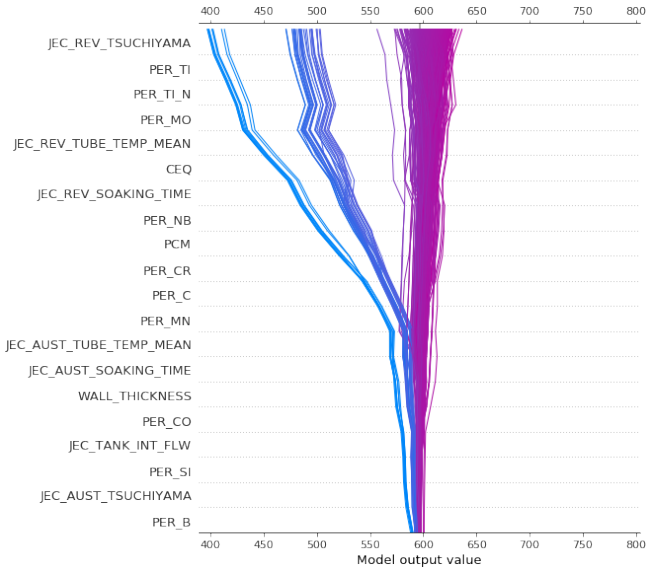


(b) Análise da variável JEC_REV_TSUCHIYAMA mostrando interações com a variável PER_MO.

Figura 4. Gráficos de dependência.



(a) Comparação entre duas observações.



(b) Visualização de todo o conjunto de dados.

Figura 5. Gráficos de decisão.

sidade. Este tipo de análise valida o funcionamento correto do modelo e pode identificar vieses causados por conjuntos de dados desbalanceados, ou não representativos, adotados no treinamento do modelo.

Também é possível gerar o gráfico de decisão para um conjunto de dados, como mostrado na Figura 5b. Percebe-se que há 3 agrupamentos distintos das linhas verticais, indicando que também devem haver 3 configurações similares que produzem tubos de aço sem costura similares. De fato, como discutido na Seção 3, na planta em estudo são produzidos 3 tipos diferentes de tubos de aço, mas esta informação não estava disponível para o modelo.

5. CONCLUSÃO

O estudo da explicabilidade de modelos de aprendizado de máquina está avançando muito rapidamente: os métodos utilizados neste artigo foram todos desenvolvidos entre 5 e 3 anos atrás. Apesar de já difundidos no meio acadêmico, a adoção dessas tecnologias nas indústrias do setor de manufatura é notoriamente lenta, como discutido na Seção 1.

Neste trabalho, treinou-se um modelo de aprendizado de máquina para a predição do limite de escoamento de tubos de aço ao fim do processo de tratamento térmico. Em seguida, foram feitas análises detalhadas de gráficos de explicabilidade geradas pelo método *Tree SHAP* para o modelo treinado. A análise, além de validar o modelo, possibilitou a extração de conhecimentos não triviais a respeito do processo e do modelo em si. Espera-se, com isso, fortalecer a ligação entre a teoria e a prática, incentivando a adoção deste tipo de tecnologia na indústria.

REFERÊNCIAS

Belle, V. and Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in Big Data*, 4. doi:10.3389/fdata.2021.688969. URL https://www.frontiersin.org/article/10.3389/fdata.2021.688969.

Carneiro, M.V., Salis, T.T., Almeida, G.M., and Braga, A.P. (2021). Prediction of Mechanical Properties of Steel Tubes Using a Machine Learning Approach. *Journal of Materials Engineering and Performance*, 30(1), 434–443. doi:10.1007/s11665-020-05345-0.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754. URL http://arxiv.org/abs/1603.02754.

Edgar, T.W. and Manz, D.O. (2017). Chapter 6 - machine learning. In T.W. Edgar and D.O. Manz (eds.), *Research Methods for Cyber Security*, 153–173. Syngress. doi:https://doi.org/10.1016/B978-0-12-805349-2.00006-6. URL https://www.sciencedirect.com/science/article/pii/B9780128053492000066.

Glover, C. (2021). Technical and human factors are restricting AI adoption in manufacturing. URL https://techmonitor.ai/technology/ai-and-automation/ai-adoption-in-manufacturing-google-siemens. Acessado em 09/05/2022.

Gomes, A., Ravetti, M., and Carrano, E. (2020). Multi-objective matheuristic for minimization of total tardiness and energy costs in a steel industry heat treatment line. *Computers & Industrial Engineering*, 151, 106929. doi:10.1016/j.cie.2020.106929.

Gomes, C., Kaiser, A.L., Bas, J.P., Aissaoui, A., and Piette, M. (2010). Predicting the mechanical properties of a quenched and tempered steel thanks to a "tempering parameter". *Revue de Métallurgie*, 107(7-8), 293–302. doi:10.1051/metal/2010061.

Ito, Y. and Bessyo, K. (1968). *Weldability Formula of High Strength Steels: Related to Heat-affected Zone Cracking*. Document // International Institute of Welding. IIW. URL https://books.google.com.br/books?id=vy_xtgAACAAJ.

Kundu, S. (2021). Ai in medicine must be explainable. *Nature Medicine*, 27, 1–1. doi:10.1038/s41591-021-01461-z.

Loukides, M. (2022). AI adoption in the enterprise 2022. URL https://www.oreilly.com/radar/ai-adoption-in-the-enterprise-2022/. Acessado em 09/05/2022.

Lundberg, Scott M.and Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67. doi:10.1038/s42256-019-0138-9. URL https://doi.org/10.1038/s42256-019-0138-9.

Lundberg, S. and Lee, S.I. (2017). A unified approach to interpreting model predictions. doi:10.48550/ARXIV.1705.07874. URL https://arxiv.org/abs/1705.07874.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.I. (2019). Explainable ai for trees: From local explanations to global understanding. doi:10.48550/ARXIV.1905.04610. URL https://arxiv.org/abs/1905.04610.

Lundberg, S.M., Nair, B., Vavilala, M.S., Horibe, M., Eisses, M.J., Adams, T., Liston, D.E., Low, D.K.W., Newman, S.F., Kim, J., et al. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10), 749.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Petch, J., Di, S., and Nelson, W. (2022). Opening the black box: The promise and limitations of explainable machine learning in cardiology. *Canadian Journal of Cardiology*, 38(2), 204–213. doi:https://doi.org/10.1016/j.cjca.2021.09.004. URL https://www.sciencedirect.com/science/article/pii/S0828282X21007030.

Rodriguez, J., Kothiyal, M., Kalvenes, J., Wolfgang, M., Lukic, V., and Nath, G. (2021). Strengthening the steel industry with ai. URL https://www.bcg.com/en-br/publications/2021/value-of-ai-in-steel-industry. Acessado em 09/05/2022.

Rozemberczki, B., Watson, L., Bayer, P., Yang, H.T., Kiss, O., Nilsson, S., and Sarkar, R. (2022). The shapley value in machine learning. doi:10.48550/ARXIV.2202.05594. URL https://arxiv.org/abs/2202.05594.

Shapley, L.S. (1951). *Notes on the N-Person Game - II: The Value of an N-Person Game*. RAND Corporation, Santa Monica, CA. doi:10.7249/RM0670.

Stein, C.R., Hermenegildo, T.F., Araújo, F.G.d.S., and Cota, A.B. (2005). Efeito da rápida austenitização sobre as propriedades mecânicas de um aço SAE1045. *Rem: Revista Escola de Minas*, 58, 51 – 56. doi:10.1590/S0370-44672005000100009. URL http://old.scielo.br/scielo.php?script=sci_arttext&pid=S0370-44672005000100009&nrm=iso.

Tsuchiyama, T. (2002). Netsushori. *Journal of the Japan Society for Heat Treatment*, 42, 163–168.