FAZENDO DATA SCIENCE NO SANEAMENTO: UM ESTUDO DE CASO

Gustavo de Souza Groppo

Graduado em Economia pela Universidade Federal de Viçosa (UFV), Especialista em Finanças pela Universidade Federal de Minas Gerais (CEPEAD/UFMG), Mestre em Economia Aplicada pela Universidade de São Paulo (ESALQ/USP) e Doutor em Saneamento pela Universidade Federal de Minas Gerais (DESA/UFMG). Analista de Desenvolvimento Tecnológico | Cientista de Dados - COPASA-MG**.**

Endereço: Rua Maranhão, 987, apto 601 - Funcionários - Belo Horizonte - MG - CEP: 30150-331 - Brasil - Tel: (31) 97195-9955 - e-mail: gustavo.groppo@gmail.com

RESUMO: A ciência de dados ambientais é um campo interdisciplinar emergente, que aborda efetivamente a complexidade dentro dos sistemas ambientais e pode fornecer soluções promissoras para o monitoramento de estações de tratamento de esgoto (ETEs). Inúmeros modelos de *Machine Learning* (ML) têm sido desenvolvidos e aplicados para fazer previsões nos mais diversos fins. Contudo, pouca atenção tem sido dada ao aprimoramento das capacidades preditivas e de explicabilidade destes modelos de ML, representando uma barreira entre pesquisa e prática, visto que os profissionais se abstêm de usar tais modelos devido à falta de explicabilidade. Visando superar essas questões propõe-se o emprego da *Data Science Trajectory* (DST), juntamente com o XGBoost e com o SHAP para avaliar um conjunto de variáveis, suas dependências e suas interações com a DBOe e, assim, compreender os mecanismos de remoção de poluentes, fundamental para o controle da qualidade do efluente. O uso de métodos como o SHAP, pode gerar insights extras tanto na comparação entre modelos, quanto na sua interpretação, podendo ser aplicado para quaisquer fins.

PALAVRAS-CHAVE: *Data Science*, *Soft Computing*, Predição da demanda biológica de oxigênio

1. INTRODUÇÃO:

A ciência de dados ambientais é um campo interdisciplinar emergente, que aborda efetivamente a complexidade dentro dos sistemas ambientais e pode fornecer soluções promissoras para o monitoramento de estações de tratamento de esgoto (ETEs) [Cheng *et al.*, 2019].

A instrumentação, o controle e a automação em constante atualização vem produzindo uma quantidade enorme de dados. Este aumento substancial que vem acontecendo tem mostrado a necessidade de encontrar ferramentas para a extração de informações relevantes, o que ficou conhecido como *Knowledge Discovery in Data-bases* (KDD) [Fayyad *et al.*, 1996]. Em 2000 foi desenvolvida uma metodologia de mineração de dados padrão para a indústria denominada *Cross-Industry Standard Process for Data Mining* (CRISP- DM) [Wirth e Hipp, 2000], amplamente utilizada no desenvolvimento de projetos de descoberta de conhecimento [Martínez-Plumed *et al*., 2021].

Não obstante, segundo Martínez-Plumed *et al*. (2021), a principal diferença entre a mineração de dados pretérita e a ciência de dados hoje é que a primeira é orientada a objetivos e se concentra no processo, enquanto a segunda é exploratória e orientada a dados. Nesse contexto Martínez-Plumed *et al*. (2021) propuseram o novo modelo de trajetórias de ciência de dados denominado DST. Este modelo representa uma revisão importante do CRISP-DM original, visto que a DST permite a flexibilidade que a ciência de dados do século XXI exige, podendo incorporar metodologias atuais no desenvolvimento e implantação de projetos de ciência de dados [Martínez-Plumed *et al*., 2021].

O *"dilema de ser rico em dados e pobre em informações"* é atribuído à falta de metodologia para selecionar o algoritmo certo para um determinado caso, bem como a ausência de procedimentos prototípicos de processamento de dados padrão e de cientistas de dados ambientais [Gibert *et al.*, 2018].

Diante desse cenário, proponho a aplicação da trajetória de ciência de dados (DST), juntamente com o algoritmo *e**X**treme **G**radient **B**oosting* (XGBoost) [Chen e Guestrin, 2016] e o método ***SH**apley **A**ditive ex**P**lanations* (SHAP) [Lundberg e Lee, 2017], objetivando realizar a predição da demanda biológica de oxigênio efluente (DBOe) de uma ETE UASB, com lodo ativado, resultando na melhor compreensão de como cada um dos principais fatores, impactam significativamente na operação.

2. OBJETIVO

Devido à complexidade intrínseca dos processos das ETEs é sempre um desafio responder de forma rápida e adequada, a dinâmica das mudanças que ocorrem no sistema, visando garantir a qualidade do efluente tratado. O método de *machine learning* (ML) foi empregado para modelar tal processo, a fim de evitar algumas deficiências dos modelos mecanicistas convencionais. Entretanto, existe uma lacuna na literatura sobre modelagem de ML, visto que a grande maioria dos estudos vêm focado na previsão ou construção de sensores suaves, sem interpretar os modelos, objetivando entender como os principais parâmetros podem ser influenciados ou controlados [Wang *et al.*, 2021].

O objetivo deste trabalho será avaliar a influência de um conjunto de variáveis monitoradas ao longo do processo de tratamento de esgoto, em sistemas UASB, com lodo ativado, sobre a variável demanda biológica de oxigênio efluente (DBOe), empregando o método SHAP que interpretará como o algoritmo XGBoost extrai os efeitos de cada variável utilizada sobre a DBOe, além de construir algumas análises exploratórias empregando a DST, visando melhorar o controle da qualidade do efluente tratado.

3. ESTUDOS CORRELATOS

Alguns estudos vêm empregando tais métodos para analisar e melhor entender como os modelos de ML, interpretados localmente, são boas alternativas para modelar, interpretar e visualizar fenômenos e processos complexos. Li (2022) conduziu experimentos de simulação que comparam o XGBoost explicado pelo SHAP, ao *Spatial Lag Model* (SLM) e Regressão Geograficamente Ponderada em Multiescala (MGWR), no nível do parâmetro, sugerindo que os modelos de ML interpretados localmente são boas alternativas para os modelos de estatística espacial, visto que apresentam um melhor desempenho quando efeitos espaciais e não espaciais complexos (por exemplo, não linearidades, interações) coexistem e são desconhecidos.

Wang *et al.* (2022) apresentaram um framework atualizado, envolvendo três modelos, baseados em árvore interpretáveis , quais sejam: (Random Forest, XGBoost e LightGBM), o SHAP para avaliar o Total de Sólidos Suspensos no efluente (TSSe) e o Fosfato no efluente (PO4e), objetivando compreender os mecanismos de remoção de poluentes em Estações de Tratamento de Efluentes (ETEs), consequentemente ajudar no controle da qualidade do efluente.

4. METODOLOGIA

4.1. ÁREA DE ESTUDO E DADOS UTILIZADOS

O estudo foi realizado para a estação de tratamento de esgoto por lodos ativados Betim Central na Região Metropolitana de Belo Horizonte (RMBH). Para a análise e predição do DBOe utilizaram um conjunto de 19 (dezenove) variáveis chave que são monitoradas ao longo do processo de tratamento, além de incluir a variável pluviometria, haja vista sua considerável influência no processo, dado que parte da contribuição pluviométrica é lançada nas redes coletoras de esgoto, impactando no processo de tratamento. A Figura 1 a seguir apresenta uma foto aérea da estação com a indicação dos pontos onde foram realizadas as coletas para as análises das variáveis consideradas no estudo, assim como na Tabela 1 são indicadas as variáveis, sua descrição e os respectivos pontos de coleta.



Figura 1 – Foto aérea ETE Betim Central com indicação dos pontos de monitoramento. Fonte: arquivo dos autores

O conjunto de dados é composto por 84 observações (01/2015 a 12/2021), que representam as medias mensais para cada uma das variáveis empregadas no presente estudo. As medidas descritivas deste conjunto de variáveis estão apresentadas na Tabela 2.

Tabela 1 – Indicação dos pontos de monitoramento onde foi coletado o esgoto para a realização das análises das variáveis consideradas no estudo

Variável	Descrição	Ponto de coleta
vaz_a	Vazão Afluente (L/s)	Medidor de vazão final
dbo_a	DBO (mg/L)	Esgoto Bruto
dqo_a	DQO (mg/L)	Esgoto Bruto
sst_a	SST (mg/L)	Esgoto Bruto
alc_a	Alcalinidade (mg/L)	Esgoto Bruto
pH_a	pH	Esgoto Bruto
ssed_a	Sólidos Sedimentáveis (mL/L)	Esgoto Bruto
temp_a	Temperatura (°C)	Esgoto Bruto
ntk_a	Nitrogênio Total (mg/L)	Esgoto Bruto
naa_a	Nitrogênio Total Amoniacal (mg/L)	Esgoto Bruto
pluv	Pluviosidade (mm chuva/mês)	Estação Pluviométrica
rel_dqo_dbo	Adimensional	-
dbo_e	DBO (mg/L)	Esgoto Tratado
dqo_e	DQO (mg/L)	Esgoto Tratado
sst_e	SST (mg/L)	Esgoto Tratado
ntk_e	Nitrogênio Total (mg/L)	Esgoto Tratado
naa_e	Nitrogênio Total Amoniacal (mg/L)	Esgoto Tratado
od_ta	Oxigênio Dissolvido (mg/L)	Tanques de aeração
sstta	Sólidos suspensos totais (mg/L)	Tanques de aeração
ssvta	Sólidos suspensos voláteis (mg/L)	Tanques de aeração

Tabela 2 – Medidas descritivas das séries históricas das variáveis

	Minimo	Média	Mediana	Desvio padrão	Primeiro quartil	Terceiro quartil	Coeficiente variação
vaz_a	302,00	427,42	421,00	54,05	393,00	464,00	12,65%
dbo_a	106,40	290,84	288,30	90,44	238,70	349,32	31,10%
dqo_a	192,90	711,84	687,70	243,82	566,50	831,00	34,25%
sst_a	90,00	382,11	356,00	154,68	275,00	446,00	40,48%
alc_a	66,00	239,12	240,00	45,31	204,50	271,50	18,95%
ph_a	6,40	7,35	7,40	0,40	7,19	7,50	5,51%
ssed_a	1,90	3,21	3,00	0,91	2,72	3,50	28,42%
temp_a	21,70	25,40	25,50	1,80	24,00	26,40	7,08%
ntk_a	15,30	51,51	49,34	30,71	39,95	55,85	59,62%
n_a_a	13,20	30,98	32,40	6,65	26,83	35,80	21,48%
pluv	0,00	117,62	68,90	136,76	10,00	176,50	116,27%
rel_dqo_dbo	1,54	2,46	2,45	0,40	2,25	2,64	16,32%
dbo_e	1,30	15,93	12,20	10,82	8,30	21,12	67,94%
dqo_e	9,56	54,19	49,40	24,06	39,55	63,00	44,41%
sst_e	4,00	17,14	14,40	11,87	8,75	20,97	69,26%
ntk_e	0,40	20,27	16,70	14,50	8,16	29,50	71,52%
n_a_e	0,39	16,60	13,30	12,96	5,35	26,30	78,07%
od_ta	0,53	1,03	1,00	0,25	0,87	1,18	24,30%
sstta	1.083,95	7.315,83	6.074,50	4.009,12	5.010,83	9.154,59	54,80%
ssvta	834,30	4.915,25	4.347,50	2.621,14	3.514,83	5.714,99	53,33%

Os dados passaram por um pré processamento onde foram normalizados usando a técnica de mínimo – máximo $[x_{Scaled} = x - min(x)/max(x) - min(x)]$, onde x_{Scaled} é o valor normalizado, x é o valor atual, $min(x)$ e $max(x)$ compreendem o menor e o maior valor presente em cada uma das colunas, respectivamente. Este conjunto de dados foram divididos em 70% para treinar os modelos e 30% para testá-los.

Empregou-se o *software* R (The R Foundation 2022) para realizer a predição do dbo_e.

4.2. *DATA SCIENCE TRAJECTORY (DST):*

O *Cross-Industry Standard Process for Data Mining* (CRISP- DM) é uma metodologia de mineração de dados padrão da indústria [Wirth e Hipp, 2000] sendo a metodologia analítica mais amplamente utilizada no desenvolvimento de projetos de descoberta de conhecimento [Martínez-Plumed *et al*., 2021]. O modelo CRISP-DM consiste na compreensão de negócios, na compreensão de dados, na preparação de dados, na construção, na avaliação e na implantação do modelo. Essa é uma metodologia capaz de transformar os dados da empresa em conhecimento e informações de gerenciamento. A Figura 2 mostra as seis fases do modelo de processo CRISP-DM e suas interações.

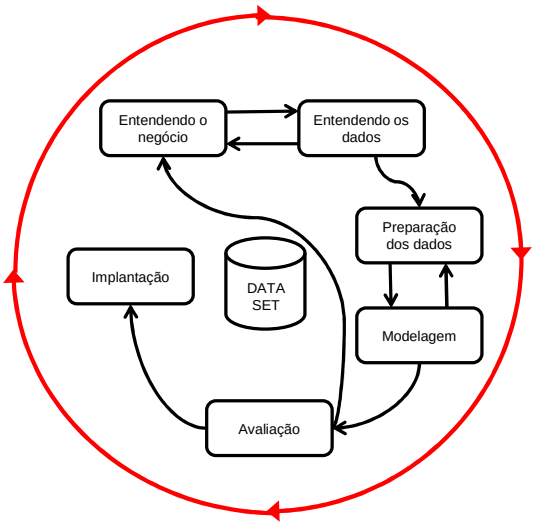


Figura 2 – The CRISP-DM process model of data mining [Adaptado Wirth e Hipp, 2000].

Qualquer projeto de mineração de dados (DM) começa com a definição da meta do projeto (objetivo) que está incluída na compreensão do negócio. Na fase de compreensão dos dados as hipóteses de informações ocultas relacionadas ao objetivo são formadas com base na experiência e em suposições qualificadas. Na fase de preparação os dados relevantes são coletados e pré-processados. Na fase de modelagem, um fluxo de trabalho de DM é construído para encontrar as configurações desejadas de parâmetros para os algoritmos selecionados e para executar a tarefa de mineração de dados nos dados pré-processados. Na fase de avaliação o modelo treinado é testado em relação a um conjunto de dados reais e os resultados da DM são avaliados de acordo com o objetivo de negócio. O conjunto de teste segue as etapas desenvolvidas nas fases de preparação e de modelagem. Após a avaliação do modelo estudado, tem-se a sexta, e última fase, que é pôr em produção (implantação).

A principal diferença entre a mineração de dados há vinte anos e a ciência de dados hoje é que a primeira é orientada a objetivos e se concentra no processo, enquanto a segunda é orientada a dados e é exploratória [Martínez-Plumed *et al.*, 2021]. Nesse contexto Martínez-Plumed *et al.* (2021) propuseram o novo modelo de trajetórias de ciência de dados denominado DST. Este modelo representa uma revisão importante do CRISP-DM original, visto que ele ainda representa uma das trajetórias mais comuns na ciência de dados, que vão dos dados ao conhecimento quando há um objetivo de negócio claro que se traduz em um objetivo de DM. Contudo, a DST permite a flexibilidade que a ciência de dados do século XXI exige, podendo incorporar as metodologias novas no desenvolvimento e implantação de projetos de ciência de dados [Martínez-Plumed *et al.*, 2021].

A ciência de dados é fundamentalmente exploratória e pode incluir algumas das seguintes atividades: exploração de objetivos de negócios que podem ser alcançados de forma orientada por dados; exploração de fonte de dados; exploração de valores que podem ser extraídos dos dados; exploração de resultados da ciência de dados aos objetivos do negócio; exploração narrativa que extraem histórias valiosas dos dados; e exploração de produto onde são encontradas novas maneiras de transformar o valor extraído dos dados em um serviço que forneça algo novo e valioso para usuários e clientes [Martínez-Plumed *et al.*, 2021].

Um projeto de ciência de dados bem-sucedido segue uma trajetória através de um espaço como o retratado em Figura 3. Diferentemente do CRISP-DM, a DST não possui setas dado que as atividades não devem ser realizadas em qualquer ordem pré-determinada [Martínez-Plumed *et al.*, 2021], sendo de responsabilidade do líder (es) de projeto decidir qual passo será dado a seguir. Mesmo que o mapa DST contenha todas as fases do CRISP-DM, estes não são necessariamente executados na ordem padrão. As atividades orientadas para o objetivo são intercaladas com atividades exploratórias.

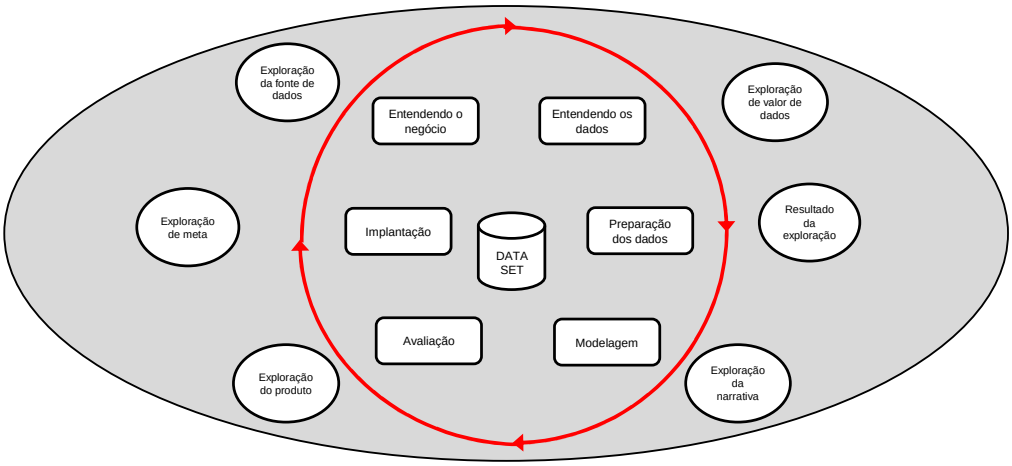


Figura 3 – O mapa de DST, contendo o círculo externo de atividades exploratórias, o círculo interno de atividades CRISP-DM (ou direcionadas a um objetivo) e, no centro, as atividades de gerenciamento de dados [Adaptado de Martínez-Plumed *et al.*, 2021].

A figura 4 representa a trajetória através do mapa DST onde a meta é estabelecida como uma primeira etapa de forma orientada a dados (exploração de meta) e dados relevantes são então explorados para extrair conhecimento valioso (exploração de valor de dados). As atividades clássicas do CRISP-DM são realizadas para limpar e transformar os dados (transformação de dados) que serão usados para treinar um determinado modelo de aprendizado de máquina (modelagem). Finalmente, o produto e / ou apresentação do usuário final mais apropriado é explorado (exploração do produto), a fim de transformar o valor extraído dos dados em um produto valioso para os usuários.

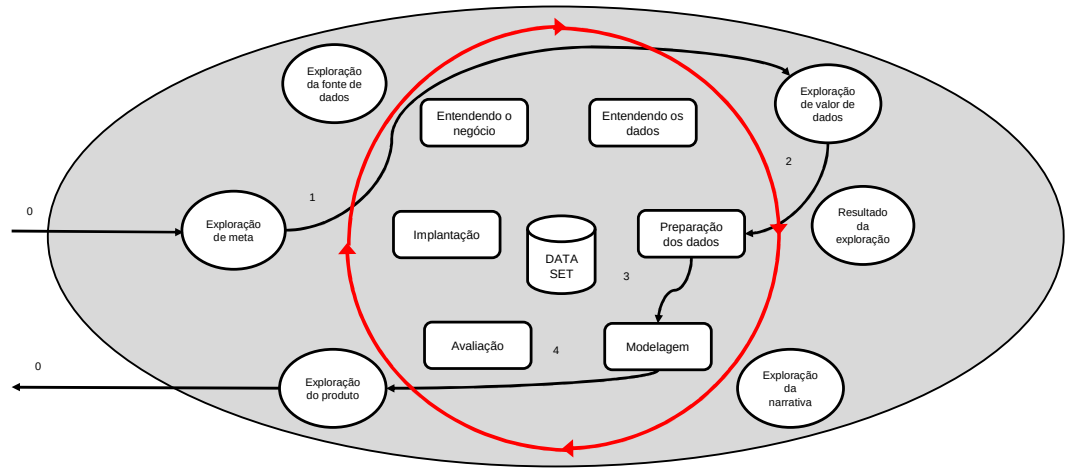


Figura 4 – Exemplo de trajetória através de um projeto de ciência de dados [Adaptado de Martínez-Plumed *et al.*, 2021].

4.3. *EXTREME GRADIENT BOOSTING (XGBOOST):*

O XGBoost é a abreviação de e**X**treme **G**radient **Boost**ing, algoritmo proposto por Chen e Guestrin, (2016). Esse algoritmo é uma implementação eficiente e escalonável da estrutura de aumento de gradiente proposto por Friedman, (2001) e Friedman *et al.*, (2000), e tem se tornando cada vez mais popular, não apenas para classificação, mas também para problemas de regressão devido ao seu alto desempenho [e.g. Chen e Guestrin, 2016; Lei *et al.*, 2019; Li *et al.*, 2020; Madhuri *et al.*, 2021]. A computação paralela, distribuída, *out-of-core* com reconhecimento de cache torna o algoritmo mais de dez vezes mais rápido do que os modelos populares usados no aprendizado de máquina e de aprendizado profunda. Ainda, segundo Chen e Guestrin, (2016), o algoritmo fornece resultados de última geração em muitos problemas de desafios de mineração de dados.

Deixe a saída de uma arvore ser

$$f(x) = w_q(x_i) \tag{1}$$

em que x é o vetor de entrada e w_q é o *score* correspondente da folha q. A saída de um conjunto de árvores k será

$$y_i = \sum_{k=1}^{k} f_k(x_i) \tag{2}$$

O algoritmo XGBoost tenta minimizar a função objetivo J na etapa t

$$J(t) = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{t-1} + f_t(x_i)\right) + \sum_{i=1}^{t} \Omega\left(f_i\right) \tag{3}$$

em que o primeiro termo contém a função de perda do treino L (por exemplo, erro quadrático médio) entre a classe real y e a saída $\hat{y}$ para as n amostras e o segundo termo é o termo de regularização, que controla a complexidade do modelo e ajuda a evitar o *overfitting*.

No XGBoost, a complexidade é definida como:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{4}$$

em que T é o número de folhas, γ é a pseudo-regularização do hiperparâmetro, dependendo de cada conjunto de dados e λ é a norma L_2 para os pesos das folhas.

Usando gradientes para a aproximação de segunda ordem da função de perda e encontrar os pesos ideais de w, o valor ideal da função objetivo é:

$$J(t) = -\frac{1}{2}\sum_{j=1}^{T} \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda} + \gamma T \tag{5}$$

em que $g_i = \partial_{\hat{y}^{t-1}} L(y, \hat{y}^{t-1})$ e $h_i = \partial^2_{\hat{y}^{t-1}} L(y, \hat{y}^{t-1})$ são as estatísticas de gradiente sobre a função de perda e I é o conjunto de folhas.

Os parâmetros de ajuste mais importantes do XGBoost são: (1) nrounds, que é o número máximo de iterações de *boost*. (2) max_depth é a profundidade máxima de uma árvore individual. (3) min_child_weight é a soma mínima do peso da instância necessária em um nó folha. (4) gama é a redução de perda mínima necessária para fazer uma partição adicional em um nó folha da árvore. (5) subamostra é a proporção de subamostra das instâncias ou linhas de treinamento. (6) learning_rate é usado durante a atualização para evitar *overfitting*. O ajuste do XGBoost é complicado pois a alteração de qualquer um dos parâmetros pode afetar os valores ideais dos outros. Isto posto, a maioria dos estudos usa o valor padrão dos parâmetros para modelagem, e poucos estudos descreveram os detalhes do processo de ajuste dos parâmetros do XGBoost [Li *et al.*, 2020].

A abordagem de pesquisa em grade, usada para encontrar a melhor combinação de parâmetros, foi a mesma empregada por Li *et al.* (2020). A faixa de parâmetros para procurar o melhor conjunto de combinação é a seguinte: max_depth é de 2 a 10 com 2 etapas, min_child_weight é de 1 a 5 com 1 etapa, gamma é de 0 a 0,4 com 0,1 etapa, subamostra é de 0,6 a 1 com 0,1 passo e os valores da taxa de aprendizagem são 0,01, 0,05, 0,1, 0,2 e 0,3.

4.4. *SHAPLEY ADDITIVE EXPLANATIONS (SHAP):*

Segundo Li (2022), a inteligência artificial (AI) e o ML, anteriormente considerados, abordagens *black box*, estão se tornando mais interpretáveis, como resultado dos recentes avanços em *eXplainable AI* (XAI) [maiores detalhes ver Gunning e Aha, 2019], em particular, a interpretação local através de métodos como SHAP.

O SHAP é um dos métodos de atribuição de *features* aditivas de classe e é baseado na unificação da teoria dos jogos [Shapley, 1953], que visa distribuir de forma justa as contribuições dos jogadores quando eles alcançam coletivamente um determinado resultado e explicações locais [Lundberg e Lee, 2017]. Os valores de Shapley

podem ser usados em ML para quantificar a contribuição de cada *feature* no modelo que fornece coletivamente a previsão [Strumbelj e Kononenko, 2014]. SHAP cria entradas simplificadas z' mapeando x para z' através de $x = h_x(z')$. Com base em z', o modelo original $f(x)$ pode ser aproximado com uma função linear de variáveis binárias:

$$f(x) = g(z') = \emptyset_0 + \sum_{i=1}^{M} \emptyset_i z_i' \quad (6)$$

em que $z' \in \{0,1\}^M$, M é o número de *features* de entrada, $\emptyset_0 \in \mathbb{R}$, o z_i' representa as *features* observadas ($z_i' = 1$) ou desconhecidas ($z_i' = 0$), e $\emptyset_i$'s são os valores atribuídos as *features*.

De acordo com Lundberg *et al*. (2019), o SHAP é o único método que possui todas as três propriedades desejáveis: precisão local, ausência e consistência e, portanto, é uma forte motivação para usar valores SHAP para atribuição das *features* do conjunto de árvores.

Para calcular valores SHAP:

$$f_x(S) = f\big(h_x(z')\big) = E[f(x)|x_s] \quad (7)$$

em que S é o conjunto de índices diferentes de zero em z' e $E[f(x)|x_s]$ é o valor esperado da função condicionada a um subconjunto S das *features* de entrada.

Os valores SHAP combinam essas expectativas condicionais com os valores Shapley clássicos da teoria dos jogos para atribuir valores ϕ_i a cada *feature*:

$$\emptyset_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(M - |S| - 1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)] \quad (8)$$

em que N é o conjunto de todos os *features* de entrada

O método Tree SHAP, usado neste trabalho, aproveita estruturas de árvore de decisão para calcular o valor $E[f(x)|x_s]$ de forma eficiente. Dado o número de árvores T e o número máximo de folhas em qualquer árvore L, a complexidade original do cálculo de $E[f(x)|x_s]$ é $O(TL2^M)$. Dada a profundidade máxima de qualquer árvore D, a complexidade de calcular $E[f(x)|x_s]$ com Tree SHAP é $O(TLD^2)$, o que diminui a complexidade computacional de um nível exponencial de alta ordem para um nível quadrático [Lundberg e Lee, 2019].

O Tree SHAP é um modelo baseado em árvore treinado com os dados de entrada X usados para treinar o modelo ($N \times M$ matriz de N instâncias e M *features*). Estes geram uma matriz $N \times M$ dos valores SHAP. Cada valor indica o quanto a *feature* contribui para a previsão da instância correspondente.

5. RESULTADOS E DISCUSSÕES:

As correlações estatísticas entre os recursos (x) e o alvo (y) são apresentadas na Figura 5. A DBOe tem uma correlação negativa bem fraca com a temperatura do afluente. Já a pluviometria tem uma correlação positiva moderada com a vazão que tem uma correlação negativa fraca com a DBOa, enquanto que a temperatura tem uma correlação positiva fraca com a pluviometria. Embora as correlações estatísticas sejam benéficas para uma análise preliminar, elas não capturam as dependências não lineares combinatórias entre os multivariados x e y que são extremamente importantes em estudos baseados em ML.

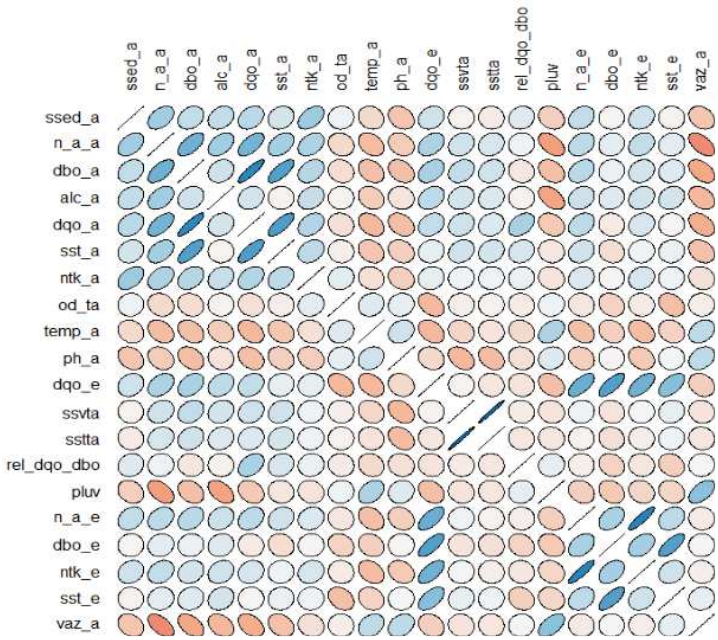


Figura 5 – Matriz de Correlação entre as variáveis empregadas na modelagem

A Figura 6 apresenta a função densidade de probabilidade da vazão expressa em (L/s) por mês, boxplot da DBOa e da DBOe expressa em (mg/L) por mês. Na subfigura mais à esquerda, observamos como a vazão da ETE varia em função das estações do ano, impactando na concentração da DBOa que chega na estação. Em períodos chuvosos observamos uma redução das concentrações da DBOa, indicando também que há um grande número de ligações clandestinas de água pluvial nas redes coletoras de esgoto, que impactam no processo de tratamento.

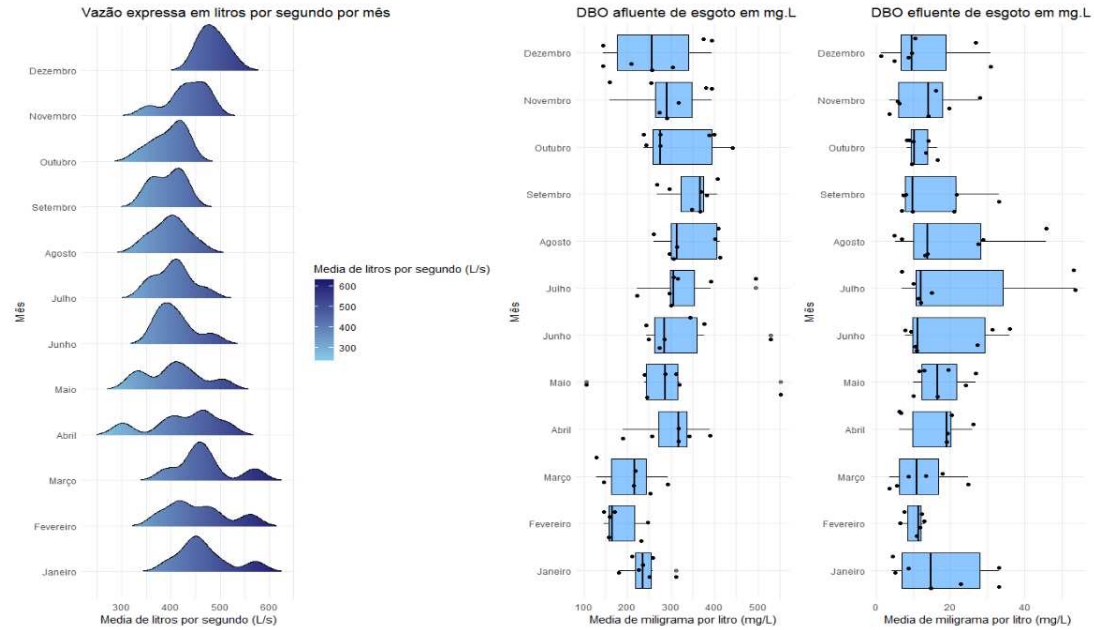


Figura 6 – Função densidade de probabilidade da vazão expressa em (L/s) por mês, boxplot do DBOa e do DBOe expressa em (mg/L) por mês

A precisão do modelo será avaliada empregando as métricas, coeficiente de determinação (R^2), raiz quadrada do erro médio quadrático (RMSE) e erro absoluto médio (MAE). Os resultados obtidos tiveram um resultado

satisfatório. A combinação das covariáveis preditoras explicou 76% da variação do DBOe, com um RMSE de 0,115 e um MAE na ordem de 0,08.

A figura 7 apresenta o modelo estimado empregando o XGBoost versus os dados realizado para a DBOe.

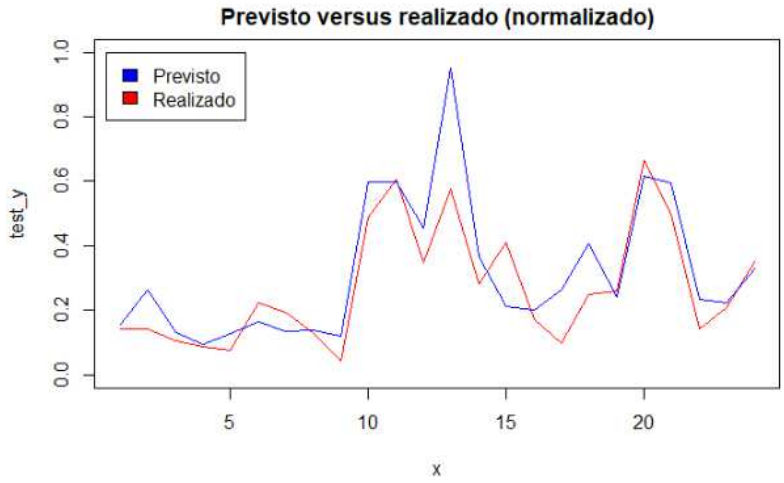


Figura 7 – Dados reais *versus* previsto utilizando o modelo XGBoost.

Com o objetivo de inferir a importância das variáveis que possuem um maior peso na predição da DBOe, empregou-se o método SHAP. Este método atribuirá pesos diferentes às variáveis, conforme sua percepção de padrão de acontecimentos da variável preditiva. Os resultados estão apresentados nas Figuras 8 e 9.

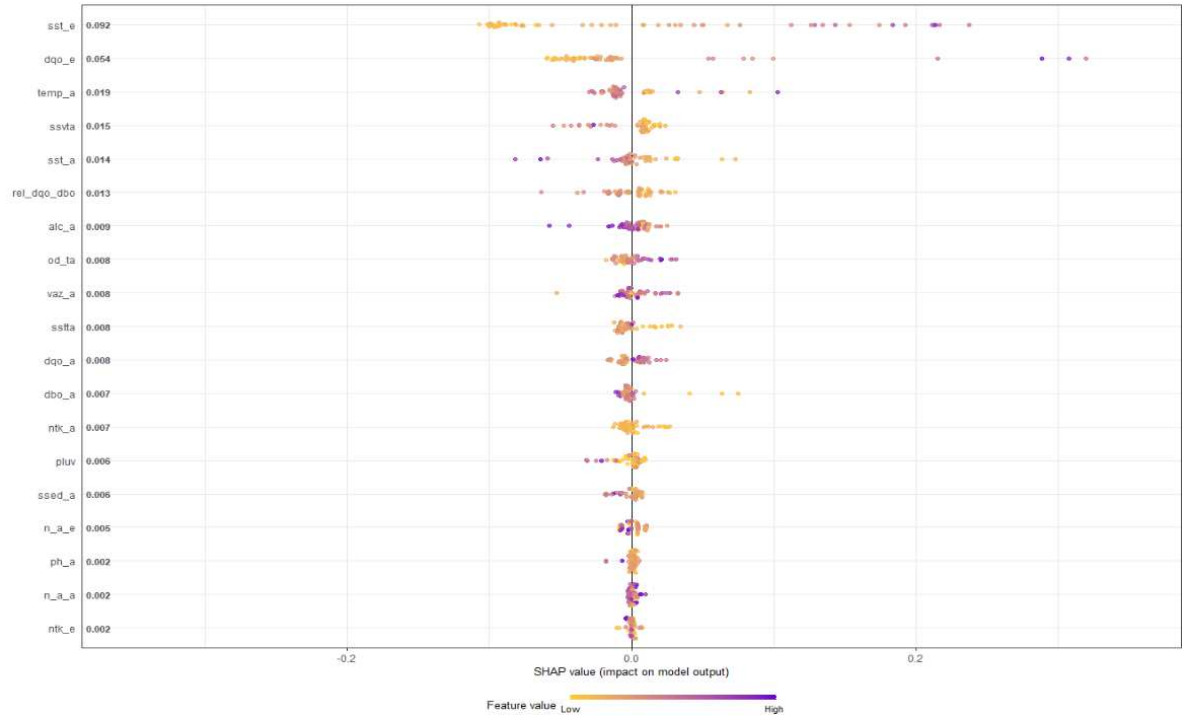


Figura 8 – Importância do SHAP para o modelo XGBoost.

As variáveis que possuem os maiores pesos na predição do valor da DBOe foram o sst_e e a dqo_e. Tal constatação é importante no acompanhamento operacional de uma estação de tratamento pois o tempo de duração para a determinação em laboratório destes parâmetros é bem inferior ao observado para a DBO, o que permite que ações corretivas sejam implementadas de forma mais ágil caso seja identificada a possibilidade de aumento das concentrações de DBO no efluente tratado da planta, contribuindo assim para uma melhor avaliação do desempenho de estações de tratamento de esgoto.

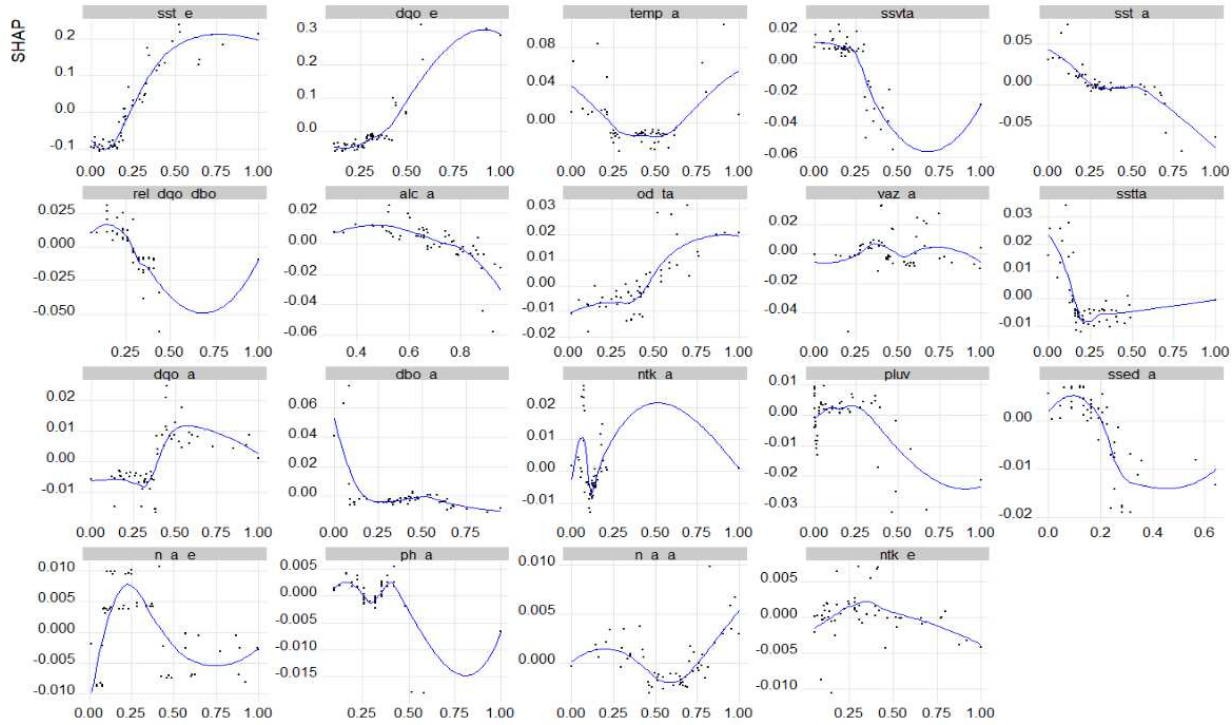


Figura 9 – Gráficos de dependência SHAP das *features*

A Figura 8 mostra a distribuição dos valores SHAP de todas as instâncias para cada variável, mostrando o impacto individual e orientado (positivo ou negativo) destas instâncias na predição. Já a Figura 9 mostra a distribuição dos valores SHAP de todas as instâncias para cada recurso. Exemplificando, para o caso da variável pluviometria, os valores médios mensais variam de 0 a 640 mm/mês. O primeiro quartil foi de 10 mm, a mediana de 68,90 mm e o terceiro quartil de 176,50 mm. Para valores acima do primeiro quartil um aumento das chuvas impactam negativamente no DBOe.

6. CONCLUSÕES:

Os pesquisadores estão avaliando inúmeros modelos de ML para fazer previsões (regressão e classificação) para diversos fins. Contudo, pouca atenção tem sido dada ao aprimoramento das capacidades preditivas e de explicabilidade destes modelos de ML, representando uma barreira entre pesquisa e prática, visto que os profissionais se abstêm de usar tais modelos, devido à falta de explicabilidade. Visando superar essas questões aplico a DST, juntamente com o XGBoost e com o SHAP para avaliar as influências dos recursos, as dependências e as interações com a DBOe.

A ideia subjacente é explicar a previsão de uma instância, calculando a contribuição de um recurso para a previsão da DBOe, identificando dependências e interações entre recursos e a DBOe, garantindo que os detalhes implícitos sejam descobertos e examinados minuciosamente, bem como permitindo selecionar um modelo genuinamente confiável e robusto para interpretação. O uso de um método de explicação preciso, robusto e granular, como o SHAP, pode gerar *insights* extras tanto na comparação entre modelos, quanto na sua interpretação, podendo ser aplicado para quaisquer fins.

7. REFERÊNCIAS:

1. Chen, T., Guestrin, C. *XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining. ACM*, 785-794, 2016.

2. Cheng, T., Dairi, A., Harrou, F., Sun, Y., Leiknes, *T. Monitoring Influent Conditions of Wastewater Treatment Plants by Nonlinear Data-Based Techniques. IEEE Access*, 7, 108827-108837, 2019.

3. Fayyad, U., Shapiro, G.P., Smyth, P. From data mining Knowledge Discovery in Databases. *AI Magazine*, 17, 3, 1996.

4. Friedman, J.H., Hastie, T., Tibshirani, R. *Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics*, 28(2), 337-407, 2000.

5. Friedman, J.H. *Greedy function approximation: a gradient boosting machine. Annals of Statistics*, 1189-1232, 2001.

6. Gibert, K., Horsburgh, J.S., Athanasiadis, I.N., Holmes G. *Environmental data science. Environmental Modelling Software*, 106, 4-12, 2018.

7. Gunning, D., Aha, D. *DARPA's explainable artificial intelligence (XAI) program. AI Magazine*, 38(3), 50-57, 2019.

8. Lei, Y., Jiang, W., Jiang, A., Zhu, Y., Niu, H., Zhang, S. *Fault diagnosis method for hydraulic directional valves integrating PCA and XGBoost. Processes*, 7(589), 1-12, 2019.

9. Li, Y., Li, M., Li, C., Liu, Z. *Forest aboveground biomass estimation using Landsat 8 and Sentinel-1A data with machine learning algorithms. Scientific Reports*, 10, 9952, 2000.

10. Li, Z. *Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. Computers, Environment and Urban Systems*, 96, 101845, 2022.

11. Lundberg, S.M., Lee, S.I. *A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems*, 30, 4768-4777, 2017.

12. Lundberg, S.M., Erion, G.G., Lee, S.I. *Consistent individualized feature attribution for tree ensembles. arXiv*, 1802.03888v3, 2019.

13. Madhuri, R., Sistla, S., Raju, K.S. *Application of machine learning algorithms for flood susceptibility assessment and risk management. Journal of Water and Climate Change*, 12(6), 2608-2623, 2021.

14. Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M.J., Flac, P. *CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Transactions on Knowledge and Data Engineering*, 33(8), 3048-3061, 2021.

15. Shapley, L. *A Value for n-Person Games*. *In*: Kuhn, H.; Tucker, A., Eds., Contributions to the Theory of Games II, Princeton University Press, Princeton, 307-317, 1953.

16. Strumbelj, E., Kononenko, I. *Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems*, 41(3), 647-665, 2014.

17. Wang, D., Thunéll, S., Lindberg, U., Jiang, L., Trygg, J., Tysklind, M. *Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based machine learning methods. Journal of Environmental Management*, 301, 113941, 2022.

18. Wirth, R., Hipp, J. *CRISP-DM: Towards a standard process model for data mining. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Springer-Verlag: London, UK, 2000.