UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I - CAMPINA GRANDE
CENTRO CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE GRADUAÇÃO EM BACHARELADO EM ESTATÍSTICA

JOÃO VITOR ANDRADE ALVES DE SOUZA

USO DO MODELO DE MACHINE LEARNING PARA PREDIZER PROPRIEDADES REOLÓGICAS DE FLUIDOS ARGILOSOS

CAMPINA GRANDE - PB
2023

JOÃO VITOR ANDRADE ALVES DE SOUZA

USO DO MODELO DE MACHINE LEARNING PARA PREDIZER PROPRIEDADES REOLÓGICAS DE FLUIDOS ARGILOSOS

Trabalho de Conclusão de Curso Bacharelado em Estatística apresentado ao Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba como requisito parcial à obtenção do título de Bacharelado em Estatística.

Área de concentração: Estatística

Orientador: Prof. Dr. Tiago Almeida de Oliveira

CAMPINA GRANDE - PB
2023

S729u Souza, Joao Vitor Andrade Alves de.
Uso do modelo de machine learning para predizer propriedades reógicas de fluidos argilosos [manuscrito] / Joao Vitor Andrade Alves de Souza. - 2023.
32 p. : il. colorido.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2023.
"Orientação : Prof. Dr. Tiago Almeida de Oliveira, Coordenação do Curso de Estatística - CCT. "

1. Aprendizado de máquina. 2. Superfície resposta. 3. Coeficiente de determinação. I. Título

21. ed. CDD 519.5

Elaborada por Geovani S. de Oliveira - CRB - 15/1009

Biblioteca Central
BC/UEPB

JOÃO VITOR ANDRADE ALVES DE SOUZA

USO DO MODELO DE MACHINE LEARNING PARA PREDIZER PROPRIEDADES REOLÓGICAS DE FLUIDOS ARGILOSOS

Trabalho de Conclusão de Curso Bacharelado em Estatística apresentado ao Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba como requisito parcial à obtenção do título de Bacharelado em Estatística.

Área de concentração: Estatística

Aprovado em: 05/12/2023

BANCA EXAMINADORA

Documento assinado digitalmente
gov.br
TIAGO ALMEIDA DE OLIVEIRA
Data: 07/12/2023 18:21:36-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Tiago Almeida de Oliveira (Orientador)
Universidade Estadual da Paraíba (UEPB)

Documento assinado digitalmente
gov.br
RENALLE CRISTINA ALVES DE MEDEIROS NASCII
Data: 07/12/2023 18:18:05-0300
Verifique em https://validar.iti.gov.br

Profa. Dra. Renalle Cristina Alves de Medeiros Nascimento
Universidade Federal Rural de Pernambuco (UFRPE)

Documento assinado digitalmente
gov.br
JANAINA APARECIDA CEZARIO
Data: 07/12/2023 16:17:09-0300
Verifique em https://validar.iti.gov.br

Profa. Ms. Janaína Aparecida Cezário
Universidade Estadual da Paraíba (UEPB)

Dedico a Deus, aos meus familiares, aos meus amigos e aos meus colegas de sala.

AGRADECIMENTOS

Primeiramente, agradeço a Deus, por toda sabedoria e inspiração, por me guiar e fortalecer ao longo deste ciclo acadêmico.

Quero agradecer a José Jackson, pois foi o que acompanhou esta jornada mais de perto, nos momentos mais conturbados e felizes, me incentivando e dando coragem.

Quero expressar minha profunda gratidão à meu orientador, Tiago Almeida de Oliveira e a professora Ana Patricia Bastos Peixoto de Oliveira, cuja orientação sábia foi fundamental para o desenvolvimento deste trabalho.

À meus pais Sandra e Vitornilson, meus irmãos: Giovana, José Gabriel e Maria Júlia e meus avôs: João e Rita, por seu constante apoio, amor e compreensão. Seu encorajamento foi o alicerce que me sustentou nos momentos desafiadores.

Aos amigos, colegas, familiares e todos os professores do departamento agradeço por compartilharem comigo suas experiências e por serem uma fonte valiosa de apoio emocional durante todo o processo.

Agradeço a banca examinadora professora Renalle e Janaína , pela atenção.

Agradeço profundamente a todas as pessoas que participaram como voluntárias neste ciclo.

“Não é o mais forte que sobrevive, nem o mais inteligente, mas o que melhor se adapta às mudanças.” - Charles Darwin

RESUMO

A argila é um material composto em sua base por argilominerais é versátil e usada em cosméticos, na contrução civil, em casas sustentáveis, artesanatos. Essencial na engenharia de petróleo, atua como base para fluidos duráveis tem origem através de erosões natural de rochas, sendo usada de diversas formas, assim tendo uma mistura flexível de compostos como barro, areia e lodo. Sua maleabilidade permite melhorias, incluindo o uso de materiais como composições de fluidos a base de argila, variando a concentração de Goma Xantana, Carboximeltilcelulose e lubrificante, submetidos a diferentes temperaturas de envelhecimento. Assim aplicando-se algoritmo de *Machine Learning* para predizer como as variáveis irão se comportar no ajuste do modelo. A implementação de técnicas de aprendizagem de máquinas mostra uma vasta gama de avanços significativos para compreensão dos estudo. Diante disso, foram comparados dois modelos estatístico o primeiro de Análise de Superfície de Resposta e o segundo de *Machine Learning*, sendo utilizados como comparativo a métrica de coeficiente de determinação (R^2), os resultados das variáveis respostas tanto para VA(cP) Viscosidade Aparente em ML $(0,99999998)$ e SR $(88,86)$, como para VP(cP) Viscosidade Plástica em ML $(0,99999985)$ e SR$(86,22)$, o resultado apresentados em *Machine Learning* mostraram ter sido melhores, pois apresentaram valores mais próximo de 1, indicando um melhor ajuste do modelo.

Palavras-chave: aprendizado de máquina. superfíce resposta. coeficiente de determinação.

ABSTRACT

Clay is a material composed primarily of clay minerals and is versatile, used in cosmetics, civil construction, sustainable housing, and crafts. Essential in petroleum engineering, it serves as a base for durable fluids, originating from natural rock erosions and used in various forms, creating a flexible mixture of compounds such as clay, sand, and mud. Its malleability allows for improvements, including the use of materials like clay-based fluid compositions, varying the concentration of Xanthan Gum, Carboxymethylcellulose, and lubricant, subjected to different aging temperatures. Applying a Machine Learning algorithm to predict how variables will behave in model adjustment, the implementation of machine learning techniques shows a wide range of significant advances for understanding the study. In this context, two statistical models were compared: the first being Response Surface Analysis and the second being Machine Learning, using the coefficient of determination (R^2) as a metric for comparison. The results for the response variables, both for Apparent Viscosity in cP $(0,99999998)(88,86)$ and SR $(88,86)$, as well as Plastic Viscosity in cP $(0,99999985)$ and SR $(86,22)$, showed that the results presented in Machine Learning were better. This is because they were closer to 1, indicating a better fit of the model.

Keywords: machine learning. response surface. coefficient of determination.

SUMÁRIO

Página

1 INTRODUÇÃO

A argila é um material com vasta gama de utilização, desde cosméticos, artesanatos e em larga escala pelas indústrias, na fabricação de utensílios de uso doméstico como louças e cerâmicas. Tem na composição os argilominerais como elemento principal, podendo conter quartzo, silicatos, óxidos, dentre outros componentes que alteram as suas propriedades.. É muito utilizada na Engenharia, notadamente na engenharia de petróleo como base para fluídos e componentes de alta durabilidade.

A análise da argila e de sua superfície é interessante para ressaltar as características e suas propriedades para que se tenha uma base de conhecimento para tomada de decisões levando aos fins desejados. A argila é produto da erosão da rocha, que ocorreu devido a acontecimentos naturais, e é um material com diversas formas de utilização, que podem diferir de local para local (Labarta, 2015). E por ser um material flexível, quando agregado algum material, pode auxiliar a modificar e até melhorar suas propriedades, um destes é a goma xantana.

De acordo com Andrade, Chaves e Incer (2008), a goma xantana é um biopolímero o produzido pela bactéria Xanthomonas campestris, e que é conhecido como espessante, estabilizante e emulsificante, muito utilizado em massa na indústria alimentícia. A goma também se destaca por melhorar a qualidade reológica dos fluidos argilosos, quando adicionado à mistura.

O Machine Learning é um subcampo da Inteligência Artificial tendo diversas ramificações pela Ciência da Computação e a Ciência de dados. Inicialmente empregado para o reconhecimento de padrões e em aprimoramento contínuo, hoje, ele se tornou a base das aplicações em inteligência artificial. Uma técnica proeminente nesse domínio é o Gradient Boosting, que gradualmente aprimora modelos de previsões fracos para minimizar erros anteriores, resultando em um gradiente de erros progressivamente reduzido. Essa abordagem é eficaz na resolução de problemas de regressão e classificação, culminando na construção de árvores de decisões robustas. Por meio de análises aprofundadas, é possível tomar decisões mais assertivas com base nesses modelos aprimorados.

Em contrapartida, a metodologia de superfície de resposta (MSR) é uma coleção de técnicas matemáticas e estatísticas que são úteis para modelagem e análise nas aplicações em que a resposta de interesse seja influenciada por várias variáveis e o objetivo seja otimizar essa resposta, utilizando para isso inferência clássica (Cecon e Silva, 2011).

Quando se analisa comparativamente os modelos XGBoost para duas variáveis respostas simultâneas em um mesmo modelo, podemos observar o desempenho ante a análise em superfície resposta, e conseguir assim diferenciar os resultados de cada modelo detalhadamente. Este tipo de abordagem tem o objetivo de avaliar o impacto de cada ajuste de acordo com desempenho do Erro Quadrático Médio.

Ao introduzir modelos estatísticos e análises regressivas aplicadas a conjuntos de dados específicos, a escolha de um modelo de ajuste varia conforme os tipos de dados e as respostas desejadas. Contudo, a eficácia de cada modelo só se concretiza ao optar pelo que possui a mais alta taxa de acurácia, garantindo assim resultados satisfatórios.

Assim, este estudo tem como propósito realizar uma análise comparativa do modelo XGBoost para duas variáveis resposta simultâneas, e também explorar a análise em superfície resposta univariada. O objetivo é identificar o modelo mais eficaz para a análise de regressão na seleção, fazendo uma singela contribuição a literatura nacional que até então não havia cruzado informações entre esses dois modelos estatísticos. Essa abordagem se revela pertinente para orientar futuras pesquisas e estudos.

2 REFERENCIAL TEÓRICO

Neste capítulo, será abordado a base teórica dos algoritmos de aprendizagem de máquina, serão discutidas as estruturas básicas dos principais algoritmos, bem como a validação do modelo e as principais técnicas utilizadas no aprendizado supervisionado para a subárea de regressão.

2.1 Machine Learning

Entende-se que, o aprendizado de máquina, traduzido do inglês que é *Machine Learning*, é um braço da inteligência artificial que emprega diversas técnicas da probabilidade, da estatística e de otimização que possibilitam o aprendizado de computadores para realizar detecção de padrões difíceis de discernir os parâmetros analisados de dados passados (Mathues e Mendonça, 2020).

De acordo com Frajacomo (2020), os algoritmos de *Machine Learning* supervisionados retiram seu aprendizado a partir de conjuntos de dados com amostras já rotuladas. Posteriormente, a análise dos dados pelo algoritmo faz com que ele evolua, e se torne cada vez mais eficiente em resolver a tarefa de classificação ou regressão modelada. Logo, este processo de melhoria do algoritmo, é conhecido como 'treino de um algoritmo'. Podemos avaliar um algoritmo de Machine Learning em um processo chamado de 'teste de um algoritmo', e neste processo, o algoritmo irá predizer o rótulo de amostras desconhecidas por ele. As suas predições serão comparadas com os verdadeiros rótulos dessas amostras, e que chegando no final deste processo, o número de classificações verdadeiramente positivas, negativas, falsamente positivas e falsamente negativas permitirá o algoritmo a ser avaliado utilizando diferentes métricas ou métricas para o caso de estudos de predição com variáveis contínuas.

Na prática de uso de Machine Learning é necessário que diferentes algoritmos sejam utilizado na busca por solucionar problemas relacionados a dados. Os profissionais da área de dados deixam explícito que não existe um único algoritmo que resolva todos os problemas, e que para cada situação vai utilizar mais adequado a ela, a depender do tipo de modelo, quais variáveis serão utilizadas, dentre outros, e obtendo assim o melhor para cada circunstância (Mahesh, 2018).

Segundo Chollet (2021), na criação de um sistema de aprendizado de máquina é feito o treinamento, em vez de ser feita a programação explicitamente. Sendo alimentado com um número de exemplos com de tamanhos relevantes para uma dada tarefa ou problema a ser resolvido, logo o sistema recupera padrões estatísticos naqueles exemplos, permitindo assim que estabeleça regras por si mesma, que automatizam as tarefas e tragam resultado satisfatório.

2.2 Algoritmos

Esta sessão nos trás os algortmos que foram utilizados no modelo de *Machine Learning*, dando destaque no *XGboost*. Os algoritmos são essenciais para poder realizar as análises preditivas, tendo o Xgboost como a solução de quando se trabalha com regressão e classificação, por fim esta sessão será visto algoritmos da classe de arvores de decisão em que se irá aumentando a complexidade até se apresentar o XGboost.

2.2.1 Árvore de decisão

A árvore de decisão é um modelo voltado para classificação ou regressão, baseado na execução de sucessivas partições binárias de uma amostra, buscando a construção de subamostras internamente mais homogêneas. Cada subamostra particionada recebe o nome de nó e cada resultado final identificado recebe o nome de folha. Para iniciar uma árvore, começa-se do primeiro nó, e verificando se a condição nele imposta é ou não atendida. Sendo caso afirmativo, prossegue-se à esquerda e em caso negativo, à direita. (Lorençatto, 2022). Na Figura 1 é ilustrada uma árvore decisão.

Figura 1 – Ilustração da Árvore de Decisão.

Fonte: Sigmoidal, 2019

2.2.2 *Gradient Boosting*

Sabe-se que o *Gradient Boosting* é uma técnica de aprendizado de máquina que através da combinação de vários modelos de aprendizado fracos, consegue formar um modelo mais forte e preciso. O objetivo do *Gradient Boosting* é melhorar a precisão do modelo, reduzindo o erro de viés e variância. Os algoritmos de Boosting visam melhorar o poder de previsão utilizando uma sequência de modelos fracos, cada um compensando os pontos fracos de seus antecessores.

O método de Boosting consiste em, iterativamente, utilizar estimadores fracos para realizar estimações e em seguida adicionar esta estimativa de forma ponderada a uma estimação gerada por um estimador forte. Sendo assim, é realizado um gradiente de modelos, reduzindo os mais fraco e permitindo que os fortes permaneçam. Neste método as árvores que são criadas, tem entre 8 e 32 folhas, dependendo do tamanho da base de dados. O Gradient Boosting também necessita de uma função de perda diferenciável para poder avaliar o quão precisa e utilizável é a estimativa (Spolador, 2021).

2.2.3 XGBOOST

O *XGBoost*, nada mais é do que a redução do termo *eXtreme Gradient Boosting*, cujo é uma biblioteca que fornece algoritmos do aprendizado de máquina, usando as estruturas de *gradient boosting*, como árvore de decisão, random forest e análise de regressão. Pode-se entender o XGBoost como um algoritmo de aprendizado de máquina, tal qual que, usa árvores de decisão para fazer previsões. Amplamente conhecido por sua eficácia e escalabilidade, e é usado em muitas aplicações de aprendizado de máquina, incluindo regressão, classificação e ranking, e é um sistema de boosting de árvore escalável chamado XGBoost, que é amplamente usado por cientistas de dados para conseguir resultados de qualidade em muitos trabalho em aprendizado de máquina(Chen e Guestrin, 2016).

O XGBoost implementa em sua biblioteca técnicas dos processos de *Bagging* e de *Boosting*. O *Bagging*, de acordo com o (IBM, 2021) é uma técnica de ensacamento, também é conhecida como agregação de *bootstrap*, é o método de aprendizado de conjunto comumente usado para diminuindo a variância dentro de um banco de dados desordenados. No processo, uma amostra aleatória de dados é selecionada em um conjunto de treinamento com substituição, o que significa que pontos de dados individuais podem ser escolhidos mais de uma vez. Logo após gerar várias amostras de dados, esses modelos fracos são treinados independentemente e, dependendo do tipo de tarefa, ou seja, regressão ou classificação, a média ou a maioria dessas previsões produzem uma estimativa mais precisa.

A previsão para uma instância i em um modelo XGBoost com K árvores é dada por:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i),$$

em que, K é o número total de árvores no modelo, f_k respresenta $k - esima$ árvore, já x_i são os atributos da $i - esima$ instancia.

Seja,

$$f_k(x) = w_{q(x)},$$

na qual, $q(x)$ é uma função que mapeia uma instância x para um dos J nós da árvore e w é o vetor de pesos atribuído a cada nó da árvore.

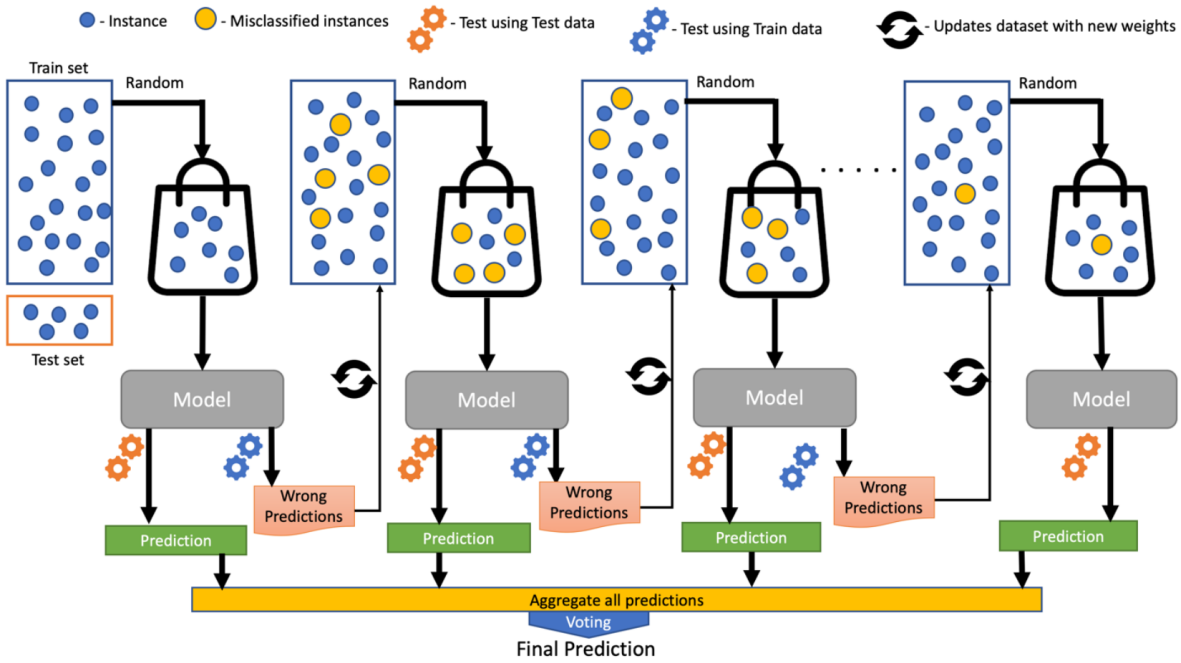
A função de perda $\mathcal{L}(\phi)$ é dada por:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

em que, l é a função de perda que mede a discrepância entre as previsões $\hat{y}_i$ e os rótulos reais y_i e para $\Omega(f_k)$ é a função de regularização que penaliza complexidade, geralmente sendo a soma dos quadrados dos pesos nos nós da árvore.

Na Figura 2, observa-se uma representaçao do *XGboost*, em que recebe algoritmos fracos e transforma em modelo fortes.

Figura 2 – Ilustração de XGboost.



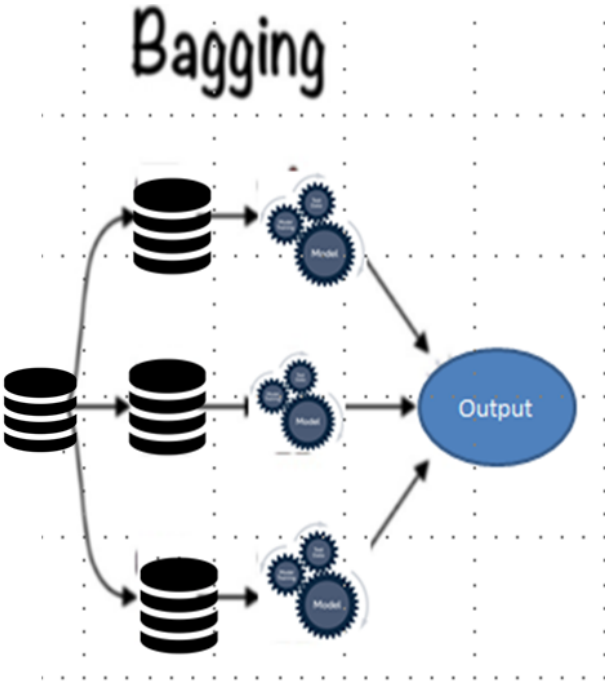
Fonte: dzone, 2020

2.2.4 *Boosting*

O *Boosting* que é uma técnica de impulsionamento de aprendizado de máquina, que combina vários modelos de aprendizado fracos para formar modelos mais fortes e preciosos. Ele é um algoritmo que impulsiona a árvore de decisão, em que impulsionar faz referência à técnica de aprendizado de conjunto muitos modelos de forma sequencial. A cada novo modelo que este algoritmo gera, tenta corrigir as deficiências do modelo anterior (Silveira, 2021).

A representação do funcionamento do *Bagging* é representada na Figura 2, nos quais são treinando vários modelos independentes em subconjuntos de forma aleatória, gerando combinações de previsões por média.

Figura 3 – Ilustração de Bagsing.



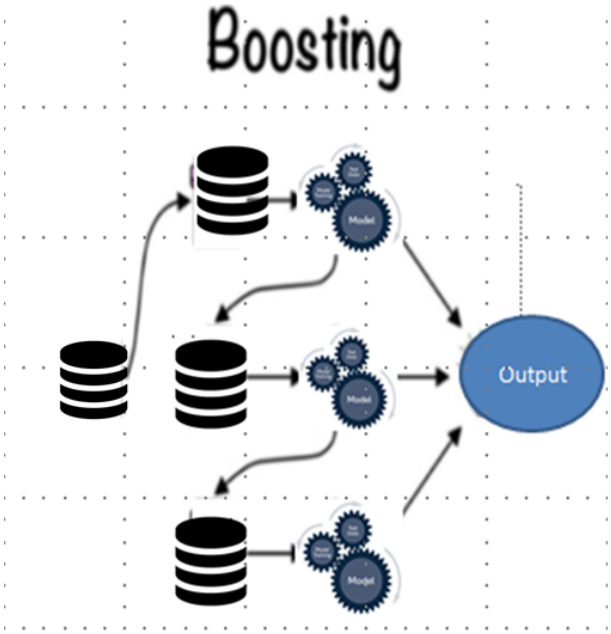
Fonte: shiksha, 2023

2.2.5 *Bagging*

O *Bagging* é uma técnica de ensacamento, também é conhecida como agregação de bootstrap, é o método de aprendizado de conjunto comumente usado para reduzir a variância dentro de um conjunto de dados barulhento. No processo de Bagging, uma amostra aleatória de dados é selecionada em um conjunto de treinamento com substituição, o que significa que pontos de dados individuais podem ser escolhidos mais de uma vez. Logo após gerar várias amostras de dados, esses modelos fracos são treinados independentemente e, dependendo do tipo de tarefa, ou seja, regressão ou classificação, a média ou a maioria dessas previsões produzem uma estimativa mais precisa (IBM, 2021).

Na figura 4, observa-se uma performance como Boosting trabalha, tendo técnicas em que os modelos fracos serão treinados, com uma melhora nos modelos classificados anteriormente, sempre visando a correção e os erros.

Figura 4 – Ilustração de Boosting.



Fonte: shiksha, 2023

2.2.6 Multioutputregressor

O multiouputregressor é uma ferramenta que é utilizada quando tem duas ou mais variáveis alvos no modelo, são utilizadas conjuntamente. Existem muitos algoritmos utilizados em aprendizagem de máquinas em que só aceitam uma variável, assim prevendo apenas um valor. Outros algoritmos irão suportar mais de uma saída, trazendo a regressão de multisaídas, tendo regressão linear e árvores de decisões. Outrossim existirá modelos que irão solucionar, servindo como uma alternativa para utilizar alguns algoritmos que antes não conseguiria fazer previsões com mais de uma saída (Brownlee, 2020).

A regressão é uma solução de problema de modelagem preditiva que busca a previsão de um valor numérico. Seja por exemplo, prever número de vendas, um peso, tamanho, uma quantidade de cliques em site, são problemas de regressão. Geralmente, apenas um valor numérico é encontrado, dadas as variáveis de entrada. Alguns problemas de regressão, necessitam da previsão de dois ou mais valores, como por exemplo, prevendo coordenada x e y. Esses problemas são nomeados de regressão de saída múltipla e regressão multi-saída ((Brownlee, 2020).

2.2.7 Superfície Resposta

A superfície resposta (MSR), é um conjunto de técnicas avançadas trabalhas em análise de regressões, também tem um papel importante quando te trata de análise de planejamento experimental, ajudando na compreensão e na influência de alguns fatores na variável alvo, ou seja variável Resposta (Y) (Soares, 2016).

A forma geral de um modelo de superfície de resposta de segunda ordem para duas variáveis de entrada $(x1, x2)$ pode ser expressa como:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$$

em que, são representados Y como variável resposta, β_0 coeficiente linear, β_1, β_2 são os coeficientes das variáveis de entrada $(x1, x2)$, β_{11}, β_{22} são os coeficientes das componentes quadráticas, β_{12} o coeficiente da interação entre e ϵ o erro aleatório.

2.3 Seleção e validação do modelos

No uso de modelos estatísticos e de machine learning uma importante etapa é o uso de métricas de avaliação. Neste sentido, a validação cruzada para permitir a reprodutibilidade dos modelos de Machine Learning e as métrica de qualidade de ajuste para possíveis comparações com as técnicas clássicas de estatística, são importantes.

2.3.1 Validação Cruzada

A técnica de validação cruzada ou *cross-validation* é empregada no campo de *machine learning,* desempenhando um papel fundamental quando se compara e seleciona modelos em cenários de modelagem preditiva. Antes de criar um modelo de aprendizagem de máquinas, é essencial realizar uma etapa de alocação, separando uma porção dos dados para testes e reservando a outra parte para validação. Essa prática é feita para evitar o overfitting, uma situação em que o modelo se ajusta excessivamente aos dados de treinamento, prejudicando sua capacidade de generalização para dados desconhecidos (Santos, 2022).

2.3.2 Erro Quadrático Médio (MSE)

O Erro Quadrático Médio conhecido também como (MSE), pode ser descrito, como a média dos erros quadráticos associados a cada valor previsto, ou seja é representado pela média dos quadrados de todas as discrepâncias entre os valores previstos e os valores reais (Liberal, 2018).

$$EQM = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

em que, n são os números de observações, y_i é o valor observado e $\hat{y}_i$ o valor previsto pelo modelo

2.3.3 Coeficiente de Determinação (R^2)

O R^2 ou coeficiente de determinação se é dado em porcentagem e tem a função de medir a proporção da variação da variável resposta (y), quando se dá por (x) variável

explicativa. R^2 tem um suporte de (0 a 1) e se explica que quanto mais próximo de 1 melhor será o desempenho de uma variável com a outra, pois as nuvens de pontos que serão apresentadas no diagrama estará bem próximo a reta de regressão. É uma medida que é positiva pois se trata da correlação de Pearson ao quadro (Martins, 2018).

$$R^2 = 1 - \frac{\text{SQR}}{\text{SQT}}$$

em que, SQR é soma dos quadrados dos resíduos, e SQT por soma total dos quadrados.

3 MATERIAIS E MÉTODOS

Os dados utilizados neste estudo foram obtidos da tese Nascimento (2013), que envolveu a análise de 19 experimentos de composições de fluidos a base de argila, variando a concentração de Goma Xantana, Carboximeltilcelulose e lubrificante, submetidos a diferentes temperaturas de envelhecimento. As variáveis alvo incluem a Viscosidade Aparente (VA) em centipoises (cP) e a Viscosidade Plástica (VP) em centipoises (cP). As variáveis explicativas são o Limite de Escoamento (LE) em Newton por metro quadrado (N/m^2), a Força Gel (FG) em Newton por metro quadrado (N/m^2) e Leitura do Viscosímetro a 3rpm $L3$.

Tabela 1 – Dados referentes aos tipos de fluídos em função da Viscosidade aparente (VA), Viscosidade plástica (VP), Limite de Escoamento (LE), Força Gel (FG) e Leitura do Viscosímetro (L3)

Fluidos	VA	VP	LE	FG	L3
FB1	42,5	16,0	53,0	10,0	39,0
FB2	34,0	15,0	38,0	51,0	26,0
FB3	46,5	15,0	63,0	8,0	38,0
FB4	46,5	17,0	59,0	52,0	38,0
FB5	53,5	15,0	77,0	11,0	46,0
FB6	48,0	20,0	56,0	57,0	36,0
FB7	60,0	17,0	86,0	10,0	53,0
FB8	50,5	22,0	57,0	53,0	34,0
FB9	41,0	15,0	52,0	13,0	38,0
FB10	42,5	15,0	55,0	58,0	35,0
FB11	50,5	17,0	67,0	3,0	44,0
FB12	56,0	18,0	76,0	61,0	45,0
FB13	55,0	19,0	72,0	18,0	43,0
FB14	51,5	19,0	65,0	47,0	40,0
FB15	60,0	23,0	74,0	27,0	46,0
FB16	59,5	23,0	73,0	65,0	46,0
FB17	55,0	19,0	72,0	21,0	44,0
FB18	55,5	21,0	69,0	22,0	40,0
FB19	56,0	19,0	74,0	22,0	47,0

Fonte: Nascimento, 2013.

Com base nesses dados, foram construídos dois modelos: um modelo de Análise de Superfície de Resposta e um modelo XGBoost, usando o MultiOutputRegressor. Neste sentido, a metodologia, empregada para as análise seguiu os seguintes passos:

i) Separação de Variáveis: Realizou-se a distinção entre variáveis Explicativas (X) e a variável de resposta (y).

ii) Ajuste do Modelo de Superfície de Resposta Original: O modelo original foi ajustado utilizando todas as amostras disponíveis.

iii) Treinamento do Xgboost: O modelo foi ajustado utilizando *crosvalidation.*

iv) Utilização do MultiOutputRegressor: Para lidar com saídas multivariadas, empregou-se a técnica MultiOutputRegressor.

v) Estimação com o Modelo Original de Superficie resposta: Foram efetuadas previsões utilizando o modelo original sobre o conjunto de dados completos.

vi) Uso da Validação Cruzada: O processo de validação cruzada foi aplicado para avaliar a robustez e generalização do modelo no caso do XGboost.

vii) Cálculo dos resíduos do modelo original: Os resíduos, representativos dos erros do modelo original, foram calculados para uma análise mais aprofundada.

viii) Avaliação do Desempenho do Modelo Original tanto de Superficie Resposta como XGboost: O desempenho do modelo original foi avaliado com base em métricas relevantes, como R^2 (coeficiente de determinação) e erro médio quadrático.

ix) Superfície de Resposta (Original): Foi gerada a superfície de resposta para o modelo XGboost, proporcionando uma comparação visual e a visualização gráfica da superfície para melhor compreensão e interpretação dos resultados.

A realização desses passos permite entender como o modelo XGboost se comporta. Foram encontradas áreas onde ele pode ser aprimorado, e avaliou-se de forma crítica os resultados obtidos. Essa abordagem detalhada é crucial para ter certeza de que os resultados que estamos apresentando em nossa pesquisa são válidos e confiáveis.

Foram empregados métodos da estatística descritiva, incluindo o cálculo de médias, medianas, valores máximos e mínimos, bem como a criação de gráficos de correlação e dispersão, para explorar os dados coletados. Além disso, foi aplicado testes estatísticos apropriados para avaliar a normalidade das variáveis que seriam objeto de estudo em nossa análise. Para o desempenho e a confiabilidade da modelagem, foi utilizado o erro quadrático médio (MSE) e o coeficiente de determinação (R^2).

A normalidade das variáveis foi avaliada por meio de testes estatísticos (shapiro-wilks) com o objetivo de garantir a robustez das análises subsequentes. Esta etapa foi fundamental para a aplicação do modelo de regressão. Para o modelo XGBOOST foram utilizadas diversas bibliotecas, para que fossem realizadas as análises propostas. As bibliotecas incluem: *pandas, numpy, matplotlib.pyplot, mpl-toolkits.mplot3d, sklearn.multioutput, sklearn.metrics* (importando as funções R^2 score e $mean_squared_error$), *xgboost e sklearn.linear-model.*

4 RESULTADOS E DISCUSSÃO

Na Tabela 2, mostraremos as medidas de posição que foram métricas fundamentais para obtermos uma análise preliminar dos dados, oferecendo uma compreensão abrangente da centralidade.

Tabela 2 – Medidas de posições e dispersão para as variáveis Viscosidade aparente (VA), Viscosidade plástica (VP), Limite de Escoamento (LE), Força Gel (FG) e Limite de Escoamento (L3)

-	VA(cP)	VP(cP)	LE(N/m)	FG(N/m)	L3 (3rpm)
Média	51,52	18,15	65,15	32,05	40,94
1º Q	47,25	15,50	56,50	12,00	38,00
2ºQ - Mediana	53,50	18,00	67,00	22,00	40,00
3º Q	56,00	19,50	73,50	52,50	45,50
Desvio Padrão	6,85	2,73	11,49	21,58	6,05

Fonte: Elaborado pelo autor, 2023.

Ao analisar a Tabela 2, observa-se medidas de posições, notoriamente tem-se a média, mediana e os quartis, a fim de compreender a variável VA. para a média tem-se aproximadamente $51,52$ cP. A mediana, situada em $53,50$ cP . Os quartis tem-se o primeiro quartil, com $47,25$ cP. Enquanto isso, o terceiro quartil de $56,00$ cP abrangendo 75% das observações abaixo.

Com média de aproximadamente $18,15$ cP, servindo como um ponto central representativo das observações para a variável VP. Já a mediana, tem-se 18 cP. Ao explorarmos os quartis, observamos que o primeiro quartil, com $15,5$ cP, o terceiro quartil, com 19 cP.

A média da variável LE esta situada em aproximadamente $65,15$ N/m. A mediana, ou segundo quartil, fixada em $67,00$ N/m. Ao observar os quartis, obtém-se que o primeiro quartil, com $56,50$ N/m. Já o terceiro quartil, representando 75% da distribuição e com $73,50$ N/m.

A média da variável FG está situada em aproximadamente $32,05$ N/m. A mediana, ou segundo quartil, fixada em $22,00$ N/m. Ao observar os quartis, obtém-se que o primeiro quartil, com $12,00$ N/m, o terceiro quartil, representando 75% da distribuição e com $52,50$ N/m.

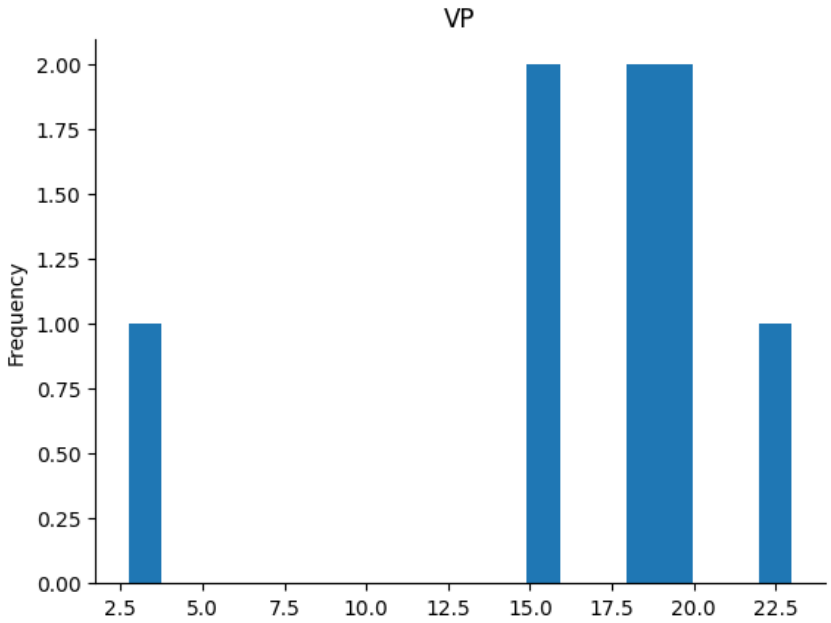
Já a média da variável L3, observada em aproximadamente $40,94$ 3rpm. A mediana, fixada em $40,00$ 3rpm,a mediana assume um papel essencial como divisor central do conjunto de dados. Ao explorar os quartis, percebe-se que o primeiro quartil, com $38,00$ 3rpm, com o terceiro quartil, representando 75% da distribuição e com $45,50$ 3rpm .

Como medida de medida de dispersão, considerando o desvio padrão como uma métrica de avaliação quanto a variabilidade dos dados. Para a variável VA, observamos um desvio padrão de $6,856$ cP, enquanto para VP, esse valor é de $2,733$ cP. As variáveis a seguir

com os respectivos valores de desvio padrão sendo elas LE com $11,49$ N/m, FG com $21,58$ N/m e L3 com $6,05$ 3rpm.

Os gráficos do comportamento das variáveis VA e VP estão representados na figura 5. Para a variável VA, a frequência tende a ser 1, por este motivo o gráfico não encontra-se no texto. Já para a variável VP, as frequências tem uma oscilação maior entre as classes, como demonstra o gráfico de barras abaixo, no intervalo que se compreende de $2,5$ à 5, obten-se a frequência de 1, entre 15 à $17,5$ uma frequência de 2, entre $17,5$ à 20 também de 2 e $22,5$ em diante frequência de 2.

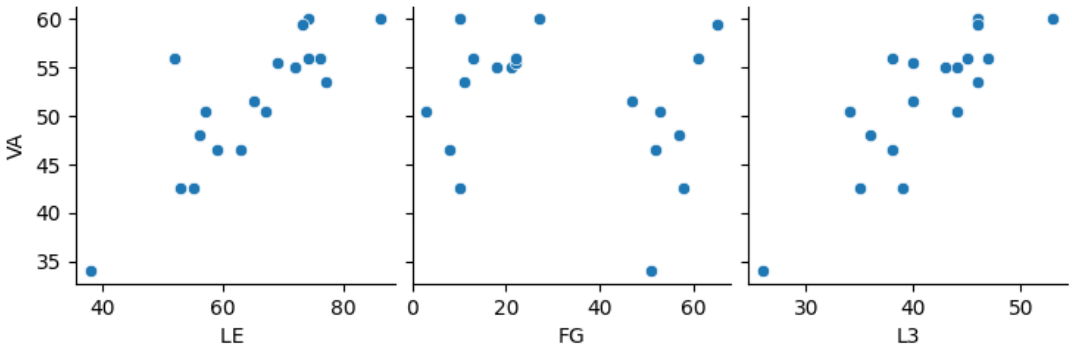
Figura 5 – Gráfico do comportamento das 19 observações em classe em relações a Viscosidade Plástica (VP).



Fonte: Elaborado pelo autor, 2023.

Na Figura 6 observa-se três gráficos estão representando as dispersões das variáveis, para $VA \times LE$, temos uma associação linear positiva, para $VA \times FG$ não temos associação linear pois os pontos estão todos espalhados, já para $VA \times L3$ também temos uma associação linear positiva.

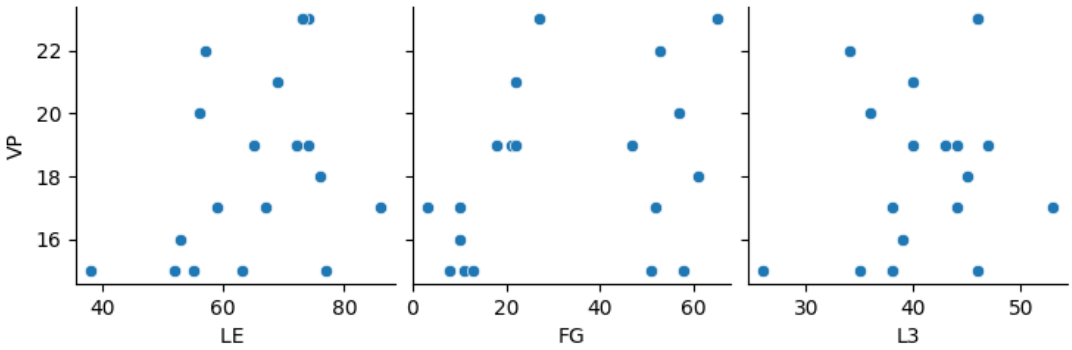
Figura 6 – Gráficos de dispersões das variáveis Limite de Escoamento (LE), Força Gel (FG) e Limite de Escoamento (L3)em relação a Viscosidade aparente (VA)



Fonte: Elaborado pelo autor, 2023.

Os gráficos apresentados na figura 7, estão representando as dispersões de variável dependente em relação a variável dependente, para $VP \times LE$, temos uma associação linear levemente positiva, para $VP \times FG$ não temos associação linear pois os pontos estão todos espalhados, já para $VA \times L3$ também não temos associação linear.

Figura 7 – Gráficos de dispersões das variáveis Limite de Escoamento (LE), Força Gel (FG) e Limite de Escoamento (L3)em relação a Viscosidade Plástica (VP).



Fonte: Elaborado pelo autor, 2023.

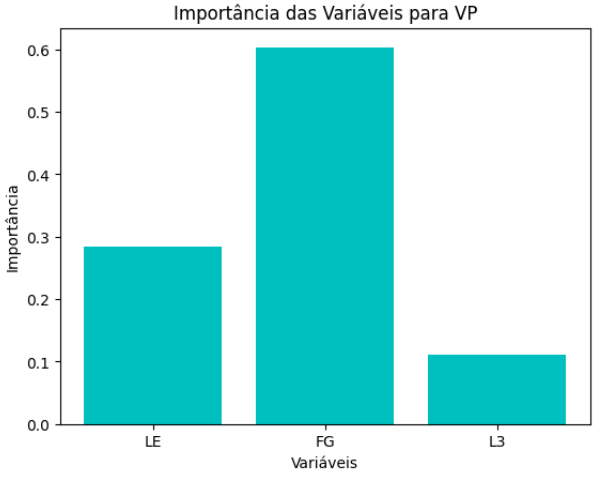
Na figura 8, tem-se o quanto de importância cada variável explicativa (LE, FG E L3), têm relevância na variável alvo (VA). Neste caso,pode se observar que a variável mais importaten na explicação é a variável LE, representando mais de 80%, bem abaixo estão as demais variáveis explicativas, L3 explica aproximadamente 15% e a FG é a que menos explica com aproximadamente 5%.

Figura 8 – Gráfico de representação das variáveis explicativas Limite de Escoamento (LE), Força Gel (FG) e Limite de Escoamento (L3) em relação a Viscosidade aparente (VA).

Fonte: Elaborado pelo autor, 2023.

A figura 9, tem-se a importância cada variável explicativa (LE, FG E L3), na relevância da variável alvo (VP), neste caso dá para observar-se que a mais explicativa é a LE, representando aproximadamente 60%, seguida da FG explicando aproximadamente 30% e a L3 sendo a que menos explica com aproximadamente 20%.

Figura 9 – Gráfico de representação das variáveis explicativas Limite de Escoamento (LE), Força Gel (FG) e Limite de Escoamento (L3) em relação a Viscosidade Plática (VP).



Fonte: Elaborado pelo autor, 2023.

A análise para a normalidade foram realizadas através do teste de Shapiro-Wilk, com um nível de significância de 5%, é possível afirmar estatisticamente, que todas as variáveis trás um padrão de uma distribuição normal, pois os valores de p são superiores a $0,05$. na tabela 3 mostraremos os P valores de cada variável respectiva.

Tabela 3 – Resultados do P-valor no teste de normalidade, para as variáveis Viscosidade aparente (VA), Viscosidade plástica (VP), Limite de Escoamento (LE), Força Gel (FG) e Limite de Escoamento (L3).

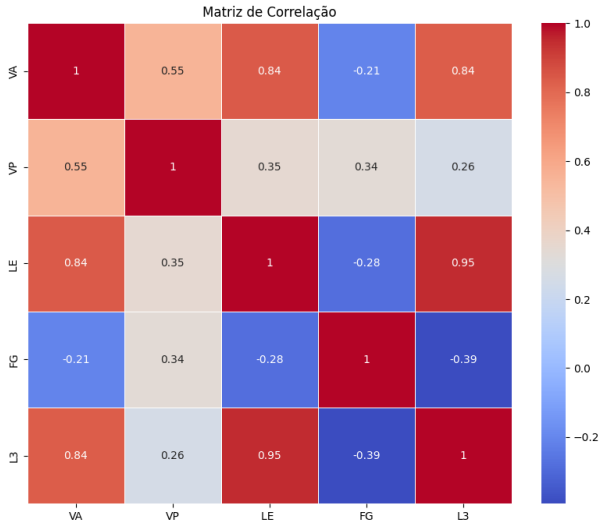
Variável	P valor
VA	0,9190
VP	0,9044
LE	0,9626
FG	0,8732

Fonte: Elaborado pelo autor, 2023.

Em uma análise sobre o gráfico de correlação observa-se uma matriz de ordem 4 por 4. deve-se destacar a forte correlação positiva de aproximadamente 95% entre as variáveis L3 e LE. Além disso, observamos correlações alta entre VA e LE, bem como entre L3 e VA, registrando cerca de 84% em ambas as associações. Porém, identificamos uma correlação de baixa e pouco representativa entre as variáveis VA e FG, apresentando um resultado de $-0,21$. Esta correlação negativa sugere uma relação fraca entre as duas variáveis, indicando que variações em uma variável não estão fortemente associadas às variações na outra.

Na Figura 10, tem-se o Gráfico de correlação para a variável resposta VA.

Figura 10 – Gráfico de Correlação das variáveis explicativas Limite de Escoamento (LE), Força Gel (FG) e Limite de Escoamento (L3) em relação a Viscosidade Aparente (VA).



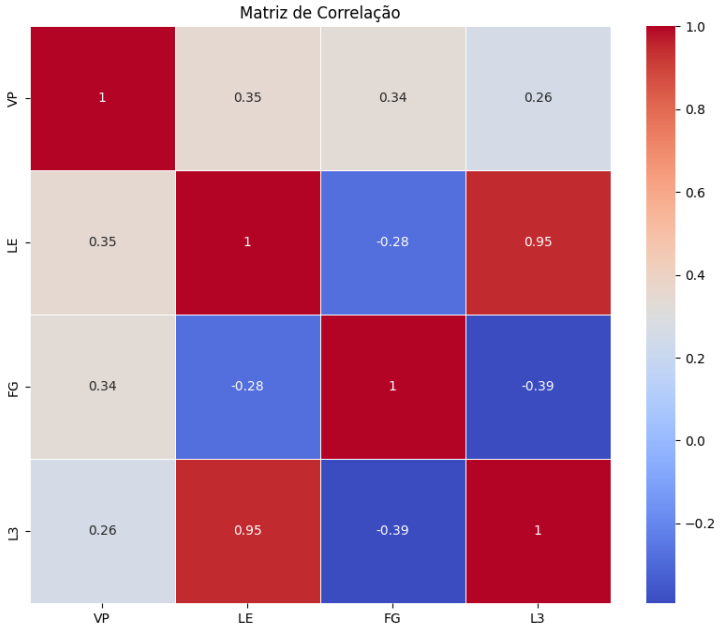
Fonte: Elaborado pelo autor, 2023.

Esta análise mostra uma matriz quadrada de ordem 4 por 4, onde a diagonal principal é composta por valores 1, pois cada variável está correlacionada consigo mesma. Já para as relações destacadas, merece um olhar diferenciado para a correlação positiva de cerca de 95% entre L3 e LE. No entanto, observa-se que algumas correlações não apresentam

uma representatividade tão expressiva. Como exemplo, a relação entre VP e L3 mostra uma correlação de $0,26$, indicando uma associação mais baixa entre essas variáveis.

O Gráfico de correlação para a variável resposta (Y) VP, está apresentado na Figura 11. Já o gráfico para correlação entre as variáveis LE, FG, L3 e VP, mostra uma correlação forte entre LE e L3, e uma correlação pouco representativa entre as demais.

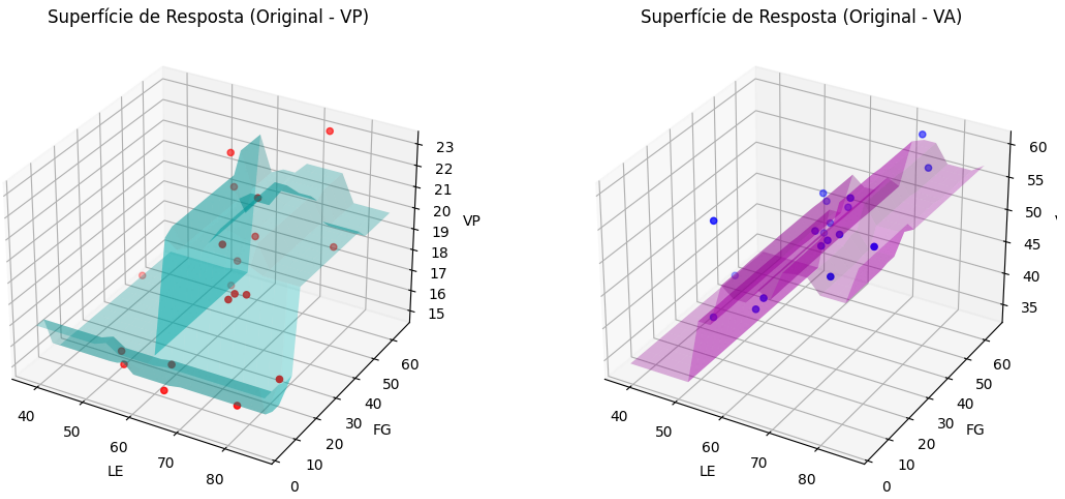
Figura 11 – Gráfico de Correlação das variáveis explicativas Limite de Escoamento (LE), Força Gel (FG) e Limite de Escoamento (L3) em relação a Viscosidade Plástica (VP).



Fonte: Elaborado pelo autor, 2023.

Por fim na Figura 12, mostra-se dois gráficos de análise em superfície resposta, em que tem-se, todos os dados em forma de pontos, ele é de forma tridimensional, tem-se os seus pontos de máximo.

Figura 12 – Gráficos de Superfície Resposta obtida pelo XGboost referente as variáveis Viscosidade Aparente (VA) e Viscosidade Plástica (VP).



Fonte: Elaborado pelo autor, 2023.

Na tabela 4 estão apresentados os resultados para o Erro Quadrático Médio (MSE) e os coeficientes de determinação para o modelo Machine Learning e Superfície de resposta para as variáveis VA eVP. Neste sentido, os valores foram considerados baixos, $VA(7,5921870)$ e $VP(1,0591529)$, indicando um bom poder de previsão. Para o Coeficiente de Determinação R^2 do modelo de Machine Learning apresentou os resultados de $0,99999998$ para (VA) E $0,99999985$ para (VP). Já para a superfície reposta, que servira de comparação com o modelo XGboost os resultados foram os seguintes (VA) $0,8886$ e (VP) $0,8622$. Assim pode-se afirmar que o modelo XGboost apresenta uma métrica de Coeficiente de Determinação mais alta, bem próxima de 1, portanto para este conjunto de dados em especifico, este modelo tem melhor desempenho.

Tabela 4 – Métricas: Erro quadrático médio(MSE) e coeficiente de determinação R^2 apresentadas nos modelos de (ML) *Machine Learning* e Superfície Resposta, para as variáveis Viscosidade aparente (VA), Viscosidade plástica (VP).

Variável	(MSE) - ML	R^2 - ML	R^2 - S. Resposta
VA	7,5921870	0,99999998	0,8886
VP	1,0591529	0,99999985	0,8622

Fonte: Elaborado pelo autor, 2023.

Segundo (Ghattas e Manzon, 2023) , os resultados apontam que todos os modelos ML teve um desempenho melhor em relação aos modelos polinomiais da abordagem RSM. Os autores observaram que para tamanhos maiores de conjuntos de treinamento, os modelos de ML não lineares foram mais precisos. Isto também foi observado neste trabalho, em que o modelo de *Machine Learning* foi superior para as duas variáveis, tanto para viscosidade

aparente, quanto para a viscosidade plástica, tendo como fonte de comparação a métrica de coeficiente de determinação.

De acordo com (Zhang e Wu, 2020) explica que para o polinomial de 2ª ordem, ambos modelos lineares de aprendizagem de ML, LASSO e GLM, é superior o método RSM, visto que os dois modelos permitem a seleção de influenciadores na formulação do modelo e, eliminando os fatores de influência que interferem na estimativa de coeficientes precisos. Visto que mais uma vez pode-se ter embasamento que muitos dos modelos de *Machine Learning* se tornam superiores, em relação à análise de superfície resposta.

Conforme Kumari, et al.(2023), o resultado experimental do desempenho de biossorção do MB foi testado e considerado superior ao uso da técnica de otimização RSM (Método de Superfície Resposta) baseada em CCD (Desenho Composto Central) e RNA (Redes Neurais Artificiais), que produziu um valor R^2 de $0,9945$, resultado que corrobora o instrumento do meu estudo/presente estudo, que traz um coeficiente de R^2 superior, cujo conseguimos um resultado de R^2 de $0,99999998$ para a variável Viscosidade Aparente e de $0,99999985$ para a variável Viscosidade Plástica, ressaltando o uso de XGBoost em relação ao Método de Superfície Resposta.

5 CONCLUSÃO

Dado o exposto, esclarece-se que, no início do processo de análise, não foi realizada nenhuma filtragem ou limpeza dos dados, uma vez que o banco de dados não apresentava valores discrepantes nem faltantes. Em outras palavras, os dados utilizados foram os mesmos encontrados na tese de Nascimento (2013).

Ao considerar o pressuposto de normalidade, todas as variáveis, de acordo com o teste de Shapiro-Wilk, apresentaram p-valores superiores ao nível de significância estabelecido em $0,05$. Isso significa que todas elas seguiram uma distribuição normal.

No que diz respeito às correlações, observa-se que, nos dados referentes a (VA e LE) e (VA e L3), há uma correlação de $0,84$. Para VP e suas variáveis, não foi encontrada nenhuma correlação consideravelmente forte.

Diante da apresentação dos modelos de Análise de Superfície Resposta versus Machine Learning (XGboost) o Coeficiente de Determinação R^2 foi a métrica de comparação entre ambos, e apoiadas pelo desenvolvimento da análise dos modelos, foram possíveis observar que o modelo de XGboost tem um comportamento que melhor ajusta os dados em que foram testados, no qual conseguimos um resultado de $0,99999998$ para a variável Viscosidade Aparente e de $0,99999985$ para a variável Viscosidade Plástica, enquanto na Análise de Superfície Resposta para (VA) foi de $88,86$ e para o atributo (VP) foi de $86,22$.

Para futuras pesquisas pode-se ser realizado o estudo comparativo com maior quantidade de dados para a confirmação do modelo para grandes amostras, e pode-se trazer para o ajuste do modelo o Random Forest, que só é possível para o modelo XGbosst com quantidade de dados maior, pode-se utilizar desse método para teste e treino, refinando o modelo.

REFERÊNCIAS

ANDRADE, C. L. S. d.; CHAVES, F. H. L.; INCER, M. A. E. Trabalho de Conclusão de Curso, *Um estudo sobre a goma xantana: análise das aplicações e do mercado*. Rio de Janeiro: [s.n.], 2008. 69 f. Disponível em: ⟨https://pantheon.ufrj.br/handle/11422/17568⟩.

BROWNLEE, J. *How to Develop Multi-Output Regression Models with Python*. 2021. Disponível em: ⟨https://machinelearningmastery.com/multi-output-regression-models-with-python/⟩. Acessado em: 15 nov. 2023.

CECON, P. R.; SILVA, A. R. d. *Introdução à Metodologia de Superfícies de Resposta*. Viçosa, 2011. Disponível em: ⟨https://arsilva.weebly.com/uploads/2/1/0/0/21008856/apostila_-_superfcie_de_resposta.pdf⟩.

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*. San Francisco, California, USA: [s.n.], 2016. p. 785–794. Disponível em: ⟨https://dl.acm.org/doi/abs/10.1145/2939672.2939785⟩.

CHOLLET, F. *Deep Learning with Python, Second Edition*. Simon and Schuster, 2021. Disponível em: ⟨https://books.google.com.br/books?hl=pt-BR&lr=&id=mjVKEAAAQBAJ&oi=fnd&pg=PR9&dq=Chollet+(2021)&ots=Ag7UxH-HWj&sig=YIJj_rHyU9Y0TRn8dMJVbXzmm7M#v=onepage&q=Chollet%20(2021)&f=false⟩.

DZONE. *XGBoost: A Deep Dive Into Boosting*. 2020. Disponível em: ⟨https://dzone.com/articles/xgboost-a-deep-dive-into-boosting⟩. Acessado em: 22 nov. 2023.

FRAJACOMO, H. C. Trabalho de Conclusão de Curso, *Seleção de SNPs utilizando Random Forests*. São Carlos: [s.n.], 2020. Universidade Federal de São Carlos, Graduação em Ciência da Computação. Disponível em: ⟨https://repositorio.ufscar.br/bitstream/handle/ufscar/15891/TCC_Henrique_Final.pdf?sequence=1&isAllowed=y⟩.

GHATTAS, B.; MANZON, D. Machine learning alternatives to response surface models. *Mathematics*, v. 11, n. 15, p. 3406, 2023. Disponível em: ⟨https://www.mdpi.com/2227-7390/11/15/3406⟩.

IBM. *¿Qué es bagging? — IBM*. 2021. Disponível em: ⟨https://www.ibm.com/mx-es/topics/bagging⟩. Acessado em: 21 nov. 2023.

LABARTA, A. *Análisis del Impacto de Metodologías Activas en la Educación Superior*. Dissertação (Trabalho de Conclusão de Curso) — Universitat Politècnica de Catalunya, 2015. 1 arquivo PDF (102 p.). Disponível em: ⟨https://upcommons.upc.edu/bitstream/handle/2117/79605/TFG%20alberto%20gaudo.pdf⟩.

LIBERAL, P. P. Trabalho de Conclusão de Curso, *Técnicas de Previsão na Produção Agrícola: Uso de Modelos de Séries Temporais para Projetar a Produção de uma Fazenda Cocoicultora*. João Pessoa: [s.n.], 2018. 50 p. Curso de Tecnologia em Análise e Desenvolvimento de Sistemas, Centro de Ciências Exatas e Tecnológicas, Universidade Federal da Paraíba. Disponível em: ⟨https://repositorio.ufpb.br/jspui/bitstream/123456789/13337/1/PPL09112018.pdf⟩.

LORENCATTO, A. A. Trabalho de Conclusão de Curso, *Análise do efeito das condições climáticas na previsão de curto prazo da demanda energética utilizando o método XGBOOST*. 2022. 45 f. Disponível em: ⟨https://repositorio.unesp.br/server/api/core/bitstreams/84f706fe-9f30-4213-9aeb-fe4b912d1ce2/content⟩.

MAHESH, B. Machine learning algorithms - a review. *International Journal of Science and Research (IJSR)*, v. 9, n. 1, p. 1–4, jan 2020. Disponível em: ⟨https://www.ijsr.net/archive/v9i1/ART20203995.pdf⟩.

MARTINS, E. Coeficiente de determinação. *Revista de Ciência Elementar*, v. 6, n. 1, p. 024, 2018. Disponível em: ⟨https://casadasciencias.org/ojs/index.php/rce/article/view/439⟩.

MATEUS, F. M. Q.; MENDONçA, M. d. C. *Machine Learning na Melhoria de Processos Internos: Estudos de Caso na Indústria de Varejo Brasileira*. Rio de Janeiro: UFRJ/Escola Politécnica, 2020. 98 p. Disponível em: ⟨http://www.repositorio.poli.ufrj.br/monografias/monopoli10031889.pdf⟩.

NASCIMENTO, R. C. A. d. M. *Estudo do fenômeno de prisão diferencial e da estabilidade térmica de fluidos argilosos*. Tese (Tese de Doutorado) — Universidade Federal de Campina Grande, Campina Grande, 2013. 158 f. Disponível em: ⟨http://dspace.sti.ufcg.edu.br:8080/xmlui/bitstream/handle/riufcg/2625/RENALLE%20CRISTINA%20ALVES%20DE%20MEDEIROS%20NASCIMENTO%20-%20TESE%20PPG-CEMat%202013.pdf?sequence=3&isAllowed=y⟩.

ONLINE, S. *Random Forest Algorithm: Python Code*. 2023. Disponível em: ⟨:https://www.shiksha.com/online-courses/articles/random-forest-algorithm-python-code/.⟩. Acessado em: 31 out. 2023.

SANTOS, V. L. d. Trabalho de Conclusão de Curso, *Uso de Machine Learning para identificação de solicitação de teste de confirmação em projeto de teste de software*. Recife: [s.n.], 2022. 33 f. Disponível em: ⟨https://repository.ufrpe.br/bitstream/123456789/4211/1/tcc_victorleuthierdossantos.pdf⟩.

SIGMOIDAL. *XGBoost: aprenda algoritmo de machine learning em Python*. 2019. Disponível em: ⟨https://sigmoidal.ai/xgboost-aprenda-algoritmo-de-machine-learning-em-python/⟩. Acessado em: 2 nov. 2023.

SILVEIRA, M. R. *Detecção de domínios maliciosos por meio de DNS passivo utilizando XGBoost*. Dissertação (Dissertação de Mestrado) — Universidade Estadual Paulista, Instituto de Biociências, Letras e Ciências Exatas, São José do Rio Preto, 2021. 67 f. Disponível em: ⟨https://repositorio.unesp.br/handle/11449/202882⟩.

SOARES, S. S. *Aplicação da metodologia de superfície de resposta na avaliação de fatores correlacionados a periodontite em Índios Kiriri do nordeste do Brasil*. Dissertação (Dissertação de Mestrado) — Universidade Federal da Bahia, Faculdade de Odontologia, Salvador, 2016. 70 f. Disponível em: ⟨https://repositorio.ufba.br/bitstream/ri/21737/1/Disserta%c3%a7%c3%a3o_ODONTO_Susana%20Silva%20Soares.pdf⟩.

SOUSA, L. R. R. d. Trabalho de Conclusão de Curso, *Utilização de Aprendizagem de Máquina para o Desenvolvimento de um Modelo Computacional para Previsão de Risco de Dengue em Palmas - TO*. Palmas/TO: [s.n.], 2018. 49 p. Curso de Sistemas de Informação, Centro Universitário Luterano de Palmas.

SPOLADOR, R. H. Trabalho de Conclusão de Curso, *Aplicação do Método de Gradient Boosting*. Niterói: [s.n.], 2021. 63 p. Universidade Federal Fluminense, Graduação em Estatística, Instituto de Matemática e Estatística. Disponível em: ⟨https://estatistica.uff.br/wp-content/uploads/sites/33/2021/05/tcc_20202_RodolofoHauret_117054008.pdf⟩.

ZHANG, Y.; WU, Y. Introducing machine learning models to response surface methodologies. In: KAYAROGANAM, P. (Ed.). *Response Surface Methodology in Engineering Science*. IntechOpen, 2021. Disponível em: ⟨https://www.intechopen.com/chapters/76805#tab2⟩.