Universidade Federal de Juiz de Fora
Faculdade de Engenharia & Instituto de Ciências Exatas
Bacharelado em Engenharia Computacional

Revisão sistemática de métodos computacionais baseados na biologia e métodos regressivos para a previsão de carbono total

Matheus dos Reis Casarim

JUIZ DE FORA
AGOSTO, 2025

Revisão sistemática de métodos computacionais baseados na biologia e métodos regressivos para a previsão de carbono total

Matheus dos Reis Casarim

Universidade Federal de Juiz de Fora

Faculdade de Engenharia & Instituto de Ciências Exatas

MAC

Bacharelado em Engenharia Computacional

Orientador: Leonardo Goliatt da Fonseca

JUIZ DE FORA

AGOSTO, 2025

Revisão sistemática de métodos computacionais baseados na biologia e métodos regressivos para a previsão de carbono total

Matheus dos Reis Casarim

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO FACULDADE DE ENGENHARIA & INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM ENGENHARIA COMPUTACIONAL.

Aprovada por:

Leonardo Goliatt da Fonseca
Doutorado em Modelagem Computacional

Samuel da Costa Alves Basilio
Mestrado em Ciência da Computação

Lahis Souza de Assis
Doutorado em Modelagem Computacional

JUIZ DE FORA

22 DE AGOSTO, 2025

Resumo

Com o passar dos anos, a criação de modelos computacionais vem se intensificando. Com isso, os modelos não heurísticos e heurísticos surgem ou se modificam. Muitos estudos acadêmicos vêm sendo feitos e publicados sobre diversos temas utilizando modelos computacionais como base comparativa, porém os estudos acabam escolhendo somente alguns modelos, devido à grande quantidade existente, não fazendo uma comparação de todos os modelos entre si, mas sim um agrupamento escolhido por cada autor. Assim na tentativa de ampliar as comparações feitas, esse estudo tem por objetivo buscar em 3 bases de dados consolidadas (*Scopus*, *WebofScience* e *ScienceDirect*) as referências, na tentativa de comparar os modelos usados para resolver o problema de previsão do *Total Organic Carbon* (TOC). Para tal comparação, buscou-se analisar uma métrica de qualidade, R^2, e três métricas que medem erros e desvios do resultado ideal ou esperado, MSE, MAE e $RMSE$. Com essas métricas, foram comparados diversos modelos utilizados nos estudos de previsão do TOC, na tentativa de responder à pergunta: "A previsão do total de carbono orgânico pode ser obtida de forma indireta, por meio de dados de logs de poços, utilizando os métodos baseados em eventos naturais ou biológicos, e obter uma precisão maior com um menor tempo de execução em comparação aos métodos de aprendizado de máquina baseados em modelos regressivos?".

Palavras-chave: Revisão sistemática, total de carbono orgânico, heurístico, não heurístico, comparativo, metodologias preditivas.

Abstract

Over the years, the development of computational models has intensified. As a result, non-heuristic and heuristic models have emerged or changed. Many academic studies have been conducted and published on various topics using computational models as a basis for comparison. However, these studies tend to select only a few models based on the sheer number of models available. Instead of comparing all models, they instead compare a grouping chosen by each author. Therefore, in an attempt to broaden the comparisons made, this study aims to search for references in three consolidated databases (*Scopus*, *Web of Science*, and *ScienceDirect*) to compare the models used to solve the problem of predicting *Total Organic Carbon* (TOC). For this comparison, we analyzed a quality metric, R^2, and three metrics that measure errors and deviations from the ideal or expected result: MSE, MAE, and $RMSE$. Using these metrics, several models used in TOC prediction studies were compared in an attempt to answer the question: "Can the prediction of total organic carbon be obtained indirectly, through well log data, using methods based on natural or biological events, and achieve greater accuracy with a shorter execution time compared to machine learning methods based on regressive models?"

Keywords: Systematic review, total organic carbon, heuristic, non-heuristic, comparative, predictive methodologies.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

ANFIS	Adaptive Neuro-Fuzzy Inference System
ACO	Ant Colony Optmization
ABC	Artificial Bee Colony
ANN	Artificial Neural Networks
BPANN	Back Propagation Artificial Neural Network
BRR	Bayesian Regularized Regression
CMR	Clustering Methods Regression
CNN	Convolutional Neural Network
DEDNN	Deep Encoder-Decoder Neural Network
DE	Differential Evolution
DDLOGR	Dual-Difference ΔlogR method
DCMF	Dynamic Committee Machine with Fuzzy-c-Means Clustering
ENLM	Elastic Net Linear Model
ELNN	Elman Neural Network
xNES	Exponential Natural Evolution Strategies
XGB	Extreme Gradient Boosting
ELM	Extreme Learning Machine
FFNN	Feedforward Neural Network
FCDN	Fully Connected Deep Network
FL	Fuzzy Logic
GPR	Gaussian Process Regression
g-GMDH	Generalized Group Method of Data Handling Neural Network
GR	Generalized Regression
GA	Genetic Algorithm
GB	Gradient Boosting
GWO	Grey Wolf Optimization
GMDH	Group Method of Data Handling

KNN	k-Nearest Neighbor
LR	Linear Regression
LOGR	ΔlogR method
MLP	Multi-layer Perceptron Neural Network
MRA	Multiple Regression Analysis
MARS	Multivariate Adaptive Regression Splines
NN	Neural Network
NGB	Neutral Gradient Boosting
PSO	Particle Swarm optimization
RBF	Radial Basis Function
RF	Random Forests
RR	Ridge Regression
SAA	Simulated Annealing Algorithm
SVR	Support Vector Regression
TOC	Total Organic Carbon

1 Introdução

Com o passar dos anos vem se notando a necessidade da criação e da manutenção de modelos de previsões. Tais modelos computacionais têm uma grande variedade atuando desde previsões climáticas a biológicas, como efeitos de remédios e doenças, são muito utilizados para auxílio na tomada de decisões, por terem um custo direto baixo e uma precisão consideravelmente elevada. Estes modelos de previsões no Brasil ganharam mais holofotes após as enchentes na região sul do país, pois já havia modelos mostrando essa possibilidade muitos anos antes dos últimos desastres que ocorreram no início do ano de 2024. O projeto governamental Brasil2040[1], montado em 2015, tinha o objetivo de "Estimar como as mudanças climáticas afetariam os setores econômicos em diferentes horizontes e sugerir estratégias de prevenção e adaptação dos diferentes sistemas que poderiam ser afetados.". Utilizando-se de modelos climáticos globais foi possível notar que esta região enfrentaria anomalias em precipitações. Estes tipos de modelos tentam usar cenários já existentes para prever futuros cenários hipotéticos e seus desdobramentos, para auxiliar na tomada de decisões e na tentativa de amenizar ou evitar os piores cenários apresentados pelos modelos montados. Esse caso mostra a importância do estudo de modelos computacionais e de estudos para sua manutenção.

No modelo utilizado para a predição de Carbono Orgânico Total ou, em inglês, *Total Organic Carbon* (TOC), temos a utilização de modelos regressivos com os dados coletados por meio de amostras e outros meios, na tentativa de estimar a viabilidade de uma região para a extração do petróleo. A vantagem da utilização desses modelos computacionais em comparação aos experimentais laboratoriais seria o menor custo e tempo (SAPORETTI et al., 2022; REIS et al., 2023; WOOD, 2020; WANG et al., 2014). Porém essa vantagem carrega uma desvantagem que envolve a maior incerteza na resposta final do que a incerteza inerente aos experimentos laboratoriais.

Assim, vários modelos são utilizados para tentar atingir uma resposta igualmente

[1] <https://www.agroicone.com.br/$res/arquivos/pdf/160727143013_BRASIL-2040-Resumo-Executivo.pdf> |Paginas 10-15

precisa comparada com os modelos laboratoriais. Com isso os modelos computacionais acabam sendo divididos em duas categorias sendo elas heurísticas e não heurísticas, essas categorias têm vários tipos de modelos como Extreme gradient boosting estudados nos artigos (WOOD, 2023; LIU; TIAN; CHEN, 2021), e o K-Nighbors Nearest estudados nos artigos (GOLIATT; SAPORETTI; PEREIRA, 2023; HANDHAL et al., 2020; REN et al., 2023).

Vendo a grande diversidade de modelos e métodos usados para tentar prever o TOC este artigo propõe-se a realizar um estudo sobre os artigos já produzidos sobre esse tema, a fim de fazer comparações dos resultados obtidos dos artigos coletados juntamente com os metadados dos mesmos para determinar qual dos modelos apresenta resultados mais consistentes, quais foram os mais citados e utilizados. Assim, pretende-se responder à seguinte questão: dentre os modelos computacionais coletados quais forneceriam os melhores resultados.

2 Metodologia

Para a obtenção dos artigos estudados foi adotado o método PICOt, que é utilizado na área médica para estudos de caso sugerindo possíveis intervenções para os mesmos Santos, Pimenta e Nobre (2007). Apesar de esse método ser amplamente utilizado na área da medicina ele pode ser uma ferramenta útil em estudos revisionais de diversas áreas, incluindo análise de problemas computacionais, como a previsão de TOC já que o estudo pode ser colocado dentro dos parâmetros da metodologia PICOt. Os parâmetros podem ser visualizados na Tabela 2.1 e sua ideia é tentar abordar uma pergunta como "A previsão do total de carbono orgânico pode ser obtida de forma indireta, por meio de dados de logs de poços, utilizando os métodos baseados em eventos naturais ou biológicos, e obter uma precisão maior com um menor tempo de execução em comparação aos métodos de aprendizado de máquina baseados em modelos regressivos?", e com a pergunta montada verificar dentro da bibliografia disponível as intervenções propostas e compará-las de forma mais ampla.

Tabela 2.1: Visualização do modelo PICOt utilizado no estudo

Letra	**Significado**	**Referencial na pergunta**
P	Grupo estudado (Population)	"A previsão do total de carbono orgânico"
I	Intervenção (Intervention)	"métodos baseados em eventos naturais ou biológicos"
C	Comparação (Comparison)	"métodos de aprendizado de maquina baseados em modelos regressivos"
O	Resultados (Outcome)	"obter uma precisão maior e em menor tempo de execução"
t	Tipo de estudo realizado (Type of Study)	Revisão sistemática

Agora, com a pergunta montada, é necessário pegar um conjunto de artigos com potencial para responder esse questionamento e, para isso, foram consultadas 3 grandes base de dados para pesquisa de artigos científicos sendo essas *Scopus*[2], *WebofScience*[3]

[2]<https://www.scopus.com>
[3]<https://www.webofscience.com/wos/>

e *ScienceDirect*[4] resultando em 818 artigos totais retornados. Esse conjunto de artigos foi encontrado utilizando um mesmo conjunto de palavras-chave nas três bases de dados contendo a seguinte regra:

("toc" **OR** "total organic carbon")
AND
("machine learning" **OR** "deep learning" **OR** "optimization")
AND
("petroleum" **OR** "oil" **OR** "shale")

O período de publicação foi definido para artigos produzidos entre 2014 - 2024. A divisão dos artigos por base de dados se deu da seguinte maneira: *Scopus*(184), *Science Direct*(489), *WebofScience*(145) e 3 artigos usados como base, esses 3 artigos pegos como base foram escolhidos previamente antes de serem feitas a busca na base de dados, sendo usados como guia dentro do tema.

Os 3 artigos escolhidos previamente são: (SAPORETTI et al., 2022; GOLIATT et al., 2024; GOLIATT; SAPORETTI; PEREIRA, 2023). Um dos motivos para a escolha desses artigos como base além dos três tratarem do tema de predição de TOC, terem uma boa quantidade de metodologias computacionais abordadas e por ter um acesso direto aos metadados desses artigos. As metodologias computacionais que tiveram dados retirados desses três artigos foram: **3 conjuntos de dados** referentes a *Linear Regression* (LR), **2 conjuntos de dados** referente a metodologia *Support Vector Regression* (SVR) e **1 conjunto de dado** referente as seguintes metodologias *Artificial Bee Colony* (ABC), *Differential Evolution* (DE), *Extreme Learning Machine* (ELM), *Elastic Net Linear Model* (ENLM), *Exponential Natural Evolution Strategies* (xNES), *Genetic Algorithm* (GA), *Grey Wolf Optimization* (GWO), *k-Nearest Neighbor* (KNN), *Multi-layer Perceptron Neural Network* (MLP), *Multivariate Adaptive Regression Splines* (MARS), *Particle Swarm optimization* (PSO) e *Random Forests* (RF).

Apesar dessa primeira busca ter retornado um número grande de artigos foi preciso utilizar alguns passos para refinar melhor o grupo de artigos encontrados, pois a relação das palavras-chave encontradas no total dos 821 artigos retornados representado

[4] <https://www.sciencedirect.com>

pela Figura 2.1, acaba sendo bastante ampla contendo vários artigos com foco em outras áreas que não respondem de maneira direta a pergunta feita inicialmente, como, ter foco no gás de xisto ou tentar prever a porosidade do solo. Com isso foram adotados os passos descritos pela Tabela 2.2. Esses passos reduziram drasticamente o número total de artigos e, com isso, a relação das palavras-chave, representadas pela Figura 2.2. Com esse número reduzido de artigos obtemos um conjunto com maior potencial para responder à pergunta inicial, tendo em vista que os artigos que não focavam diretamente no TOC foram retirados.

Tabela 2.2: Passos seguidos e total de artigos restantes

Passo	Total de artigos	Ação tomada
0	3	Definição da base de artigos usados para tratar do tema
1	821	Busca nas bases de dados adicionando aos 3 artigos bases
2	614	Retirada dos artigos que são duplicados já que foi utilizada mais de uma base de dados
3	614	Classificação inicial dos artigos não duplicados
4	57	Retirada de artigos que buscam a porosidade do solo, foco maior em gás ou não foco na pergunta inicial
5	47	Exclusão dos artigos não rastreados ou escritos com idioma chinês
6	38	Exclusão dos artigos com erros, retratação ou utilizando metodologias de qualificação pouco usual
7	38	Classificação do conteúdo e agrupamento dos artigos restantes

A grande diminuição do número de artigos do passo 3 para o 4 se deu pela retirada dos artigos que não tiveram o foco principal na predição do TOC de maneira mais direta.

Com o número reduzido de 38 artigos únicos, cujos dados foram agrupados por metodologias computacionais utilizadas ou citadas na predição do TOC buscou-se avaliar se os passos seguidos tiveram sucesso no refinamento dos possíveis dados produzidos pelos artigos. Essas metodologias referenciadas pelos artigos estão disponíveis na Tabela 2.3, a qual representa a visualização dos artigos referentes ao último passo da Tabela 2.2. Nota-se uma grande variedade de metodologias computacionais citadas pelos autores e com o potencial de verificar se as metodologias inspiradas em eventos biológicos\naturais poderiam ser superiores às regressivas.

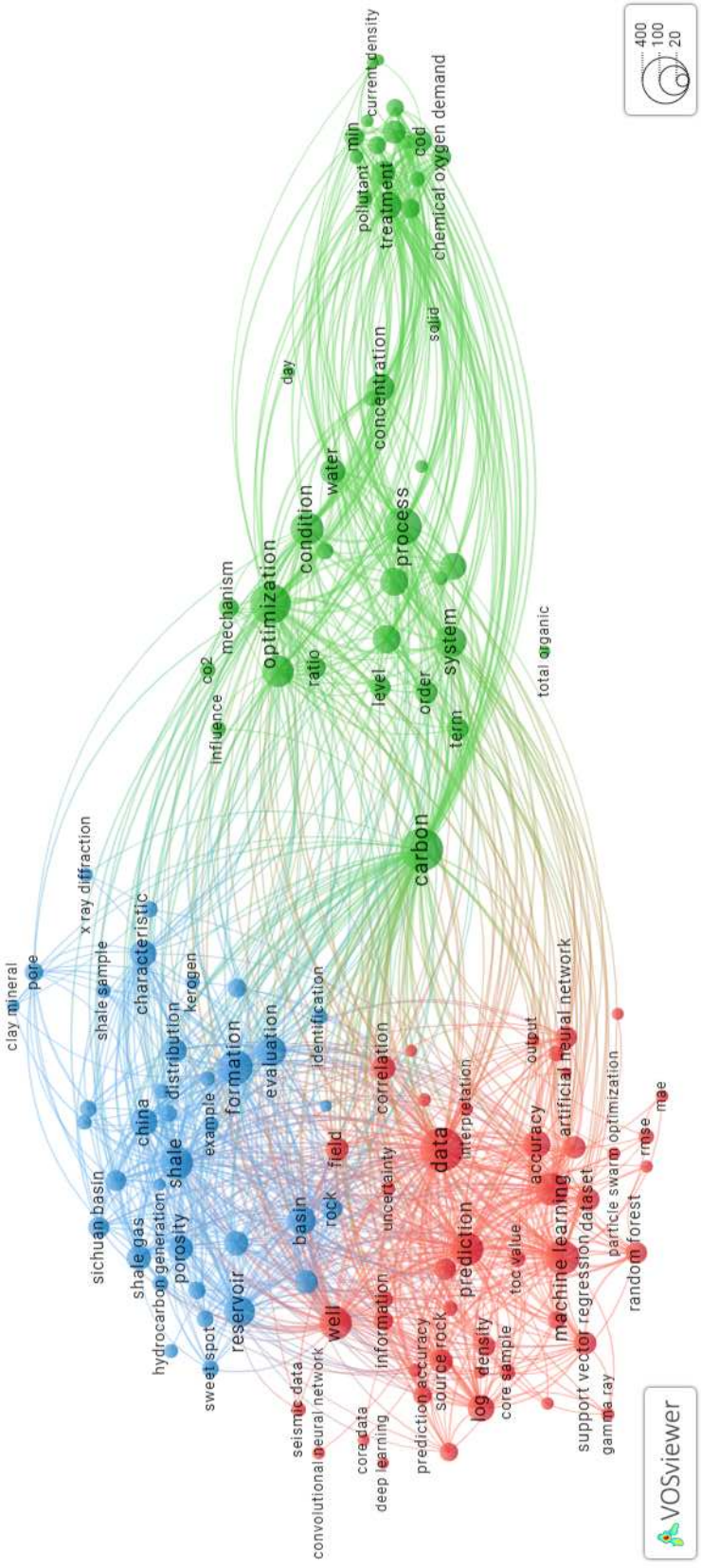

Figura 2.1: Mapa de aparições e relação de palavras-chave dos 821 artigos

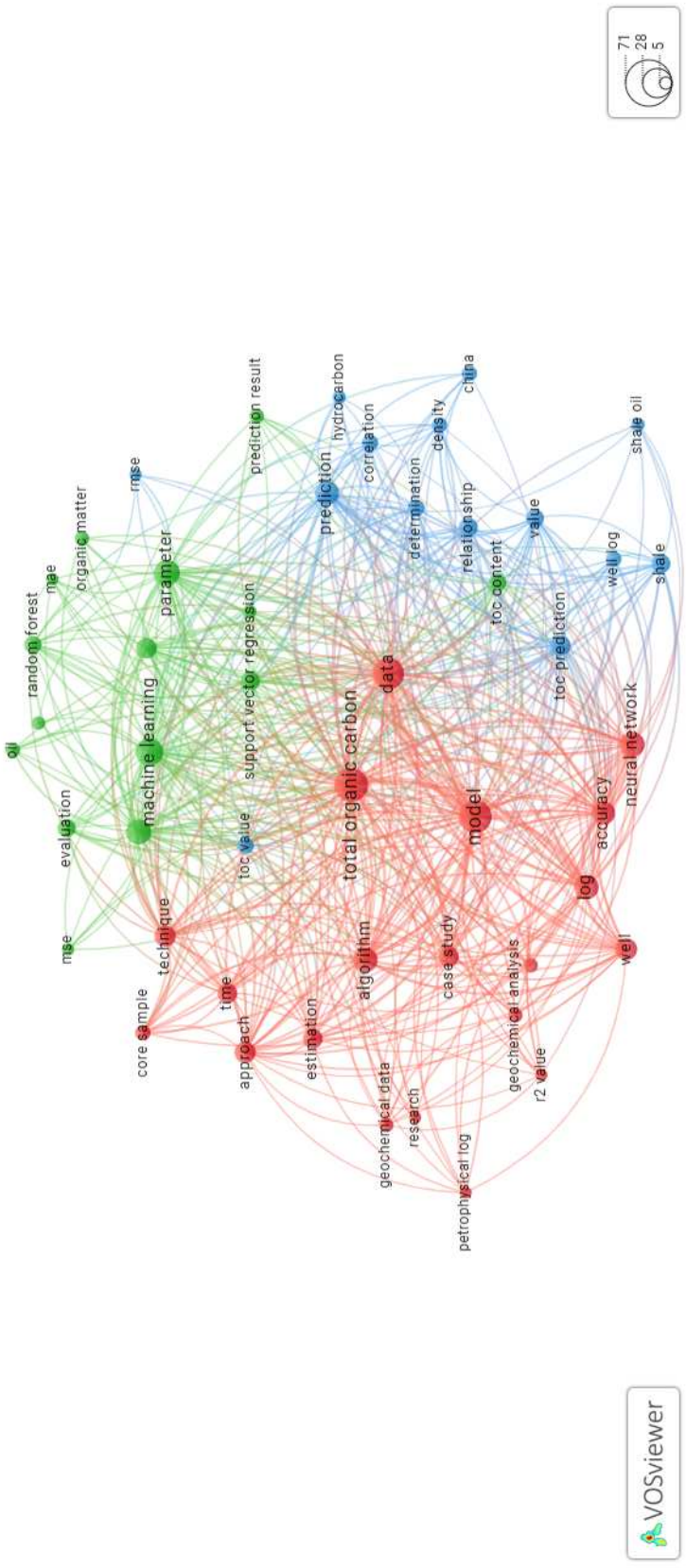

Figura 2.2: Mapa de aparições e relação de palavras-chave dos 38 artigos

Tabela 2.3: Descrição dos artigos

Artigo	Metodologia citada	Tipo da metodologia
(SAPORETTI et al., 2022) *	ANFIS, ANN, BRR, ELM, GPR, kNN, RF, GB, SVR, ENLM	Não Heurístico (ANFIS, ANN, ELM, kNN, RF) Heurístico (BRR, GPR, GB, SVR, ENLM)
(GOLIATT et al., 2024) *	ELM, SVR, MARS, GA, PSO, DE, ABC, GWO, xNES, ENLM	Não Heurístico (ELM, GA, PSO, ABC, GWO) Heurístico (SVR, MARS, DE, xNES, ENLM)
(GOLIATT; SAPORETTI; PEREIRA, 2023) *	KNN, LR, MLP, RF, RR, SVR	Não Heurístico (KNN, MLP, RF) Heurístico(LR, RR, SVR)
(REIS et al., 2023)	RF, MLP, SVR, XGB, NGB	Não Heurístico (RF, MLP) Heurístico (SVR, XGB, NGB)
(AMOSU; IMSALEM; SUN, 2021)	SVR, ELM	Não Heurístico (ELM) Heurístico (SVR)
(WOOD, 2020)	SVR	Heurístico(SVR)
(ASANTE-OKYERE; MARFO; ZIGGAH, 2023)	MARS, RF, GB	Não Heurístico (RF) Heurístico (MARS, GB)
(SIDDIG; IBRAHIM; ELKATATNY, 2021)	SVR, RF, LOGR	Não Heurístico (RF) Heurístico (SVR, LOGR)
Nyakilla et al. (2022)	SVR, GPR	Heurístico (SVR, GPR)
(MKONO et al., 2023)	SVR, GR, g-GMDH	Não Heurístico (g-GMDH) Heurístico (SVR, GR)
(HASSAN et al., 2023)	ANN	Não Heurístico (ANN)
(ZHENG; WU; HOU, 2021)	LOGR, SVR, ANN, FCDN	Não Heurístico (ANN, FCDN) Heurístico (LOGR, SVR)
(LIU; TIAN; CHEN, 2021)	XGB, LOGR, RF, SVR, KNN, LR	Não Heurístico (RF ,KNN) Heurístico (XGB, LOGR, SVR)

Artigo	Metodologia citada	Tipo da metodologia
(SUN et al., 2023)	SVR, XGB, RF, LOGR	Não Heurístico (RF) Heurístico (SVR, XGB,LOGR)
(WOOD, 2023)	SVR, XGB, RF, KNN, ENLM	Não Heurístico (RF, KNN) Heurístico (SVR, XGB, ENLM)
(ZHANG; WU; WU, 2023)	MRA, LR, BPANN, GB, LOGR	Não Heurístico (BPANN) Heurístico (MRA, LR, GB, LOGR)
(Asgari Nezhad; MORADZADEH; KAMALI, 2018)	GPR, ANN, LOGR	Não Heurístico (ANN) Heurístico (GPR, LOGR)
(ABDIZADEH et al., 2017)	ACO, NN, LOGR	Não Heurístico (ACO, NN) Heurístico (LOGR)
(HANDHAL et al., 2020)	RF, KNN, BPANN, SVR	Não Heurístico (RF, KNN, BPANN) Heurístico (SVR)
(RUI et al., 2019)	PSO, SVR, LOGR	Não Heurístico (PSO) Heurístico (SVR, LOGR)
(WANG et al., 2019)	CNN, BPANN, LOGR	Não Heurístico (CNN, BPANN) Heurístico (LOGR)
(TABATABAEI et al., 2015)	ACO, GA, BPANN, ANN	Não Heurístico (ACO, GA, BPANN, ANN)
(WANG; PENG, 2018)	SAA, GA, BPANN, SVR, PSO	Não Heurístico (GA, BPANN, PSO) Heurístico (SAA)
(RUI et al., 2020)	LOGR, GPR	Heurístico (LOGR, GPR)
(SHALABY et al., 2019)**	ANN, LOGR	Não Heurístico (ANN) Heurístico (LOGR)
(ZHU et al., 2019)	LOGR, DDLOGR, CMR, FL	Não Heurístico (CMR, FL) Heurístico (LOGR, DDLOGR)
(JOHNSON et al., 2018)	ANN	Não Heurístico (ANN)
(GHARAVI et al., 2022)	RF, KNN, LR, GB	Não Heurístico (RF, KNN) Heurístico (LR, GB)
(ASANTE-OKYERE; ZIGGAH; MARFO, 2021)	CNN, LOGR, DDLOGR	Não Heurístico (CNN) Heurístico (LOGR, DDLOGR)

Artigo	Metodologia citada	Tipo da metodologia
Safaei-Farouji e Kadkhodaie (2022)	MLP, RF, SVR, GWO, GA, PSO, RBF	Não Heurístico (MLP, RF, GWO, GA, PSO) Heurístico (SVR, RBF)
(ZHANG et al., 2022)	LOGR, CNN, DEDNN	Não Heurístico (CNN, DEDNN) Heurístico (LOGR)
(BOLANDI; KADKHODAIE; FARZI, 2017)	KNN, ANN, SVR, LOGR, FL	Não Heurístico (KNN, ANN, FL) Heurístico (SVR, LOGR)
(BARHAM et al., 2021)	ANN, FFNN, BPANN	Não Heurístico (ANN, FFNN, BPANN)
(BAI; TAN, 2021)	ELM, RBF, BPANN, DCMF, GR, ELNN	Não Heurístico (ELM, BPANN, DCMF) Heurístico (RBF, GR, ELNN)
(REN et al., 2023)	KNN	Não Heurístico (KNN)
(SHALABY et al., 2020)	RF, SVR, LR, GPR, BRR	Não Heurístico (RF) Heurístico (SVR, LR, GPR, BRR)
(HU et al., 2021)	LOGR, ANN, BPANN	Não Heurístico (ANN, BPANN) Heurístico (LOGR)
(MULASHANI et al., 2021)	ANN, LOGR, BRR, SVR, DDLOGR, GMDH	Não Heurístico (ANN) Heurístico (LOGR, BRR, SVR, DDLOGR, GMDH)

As metodologias estão como abreviações para facilitar a visualização

Artigos com * foram pré adicionados no estudo

Artigos com ** com overtraining relatado pelo autor

2.1 Comparação dos dados

Esses 38 artigos restantes mostrados na Tabela 2.2 tiveram seus dados analisados e os resultados fornecidos das experiências feitas pelos autores utilizando a fase de treinamento da metodologia computacional abordada, somente para pegar o número de iterações usadas para o treinamento da mesma, já as métricas qualitativas, como a precisão ou o erro produzido foram coletados na última fase, sendo chamada pelos autores de verificação ou comprovação, a fim de ranquear essas metodologias e ver qual delas apresentaria o resultado mais consistente e em menor tempo de execução.

Para ficar mais claro o dado esperado dos artigos seguiriam com o formato: $Dado = (\mathcal{M}, \mathcal{Q})$ onde o $\mathcal{M}$ representa o nome de uma metodologia preditiva para o problema de predição do TOC e $\mathcal{Q}$ seria o conjunto de dados que seria possível extrair junto do $\mathcal{M}$. Esse conjunto representado pelo $\mathcal{Q}$ é representado por uma Tupla, em linguagem computacional, que seria um conjunto de dois dados fixos que

podem ser representados pela seguinte formato: $\mathcal{Q} = (Name, Value)$, onde $Name$ faz referencia a uma métrica qualitativa adotada pelo autor e $Value$ seria o valor correspondente dessa métrica qualitativa. Sendo assim cada artigo produz um grande conjuntos de dados representados pelo *Dado*, pois os mesmos muitas vezes acabam utilizando vários modelos computacionais para fazer a comparações entre as mesmas.

Sendo assim o conjunto $\mathcal{M}$ esta contido nas metodologias referenciadas pela Tabela 2.2, já o conjunto $\mathcal{Q}$ representa todas as possíveis métricas qualitativas adotadas pelos autores na produção dos seus artigos onde os dados produzidos na forma de matriz de confusão não foi considerados onde esse tipo de dado se apresentou em somente um artigo (AMOSU; IMSALEM; SUN, 2021).

2.1.1 Métricas adotadas para a comparações

Como observado na Tabela 2.3 há uma grande variedade de metodologias computacionais adotadas pelos autores, o mesmo pode ser observado nas métricas qualitativas, que podem ser observadas na Tabela 2.4. Com essa grande variedade torna-se necessário fazer escolhas de quais métricas o estudo vai considerar para as comparações entre as metodologias computacionais, sendo assim, foram escolhidas as 4 métricas que foram mais utilizadas pelos autores, e com isso teríamos 1 métrica para a aferição da precisão e 3 para aferição do erro produzido pela metodologia.

Tabela 2.4: Número de dados das metodologias de qualificação

Metodologia	Quantidade de dados
* R^2	115
$MAPE$	35
$MARE$	10
WI	10
* MAE	88
* $RMSE$	64
$NRMSE$	8
Nº Runs	32
$PEARSON$	5
$SPEARMAN$	5
* MSE	67
APD	3
$AAPD$	3
SD	3
$AAPE$	3
MRE	24

Métrica marcada com * foi adotada para o estudo comparativo

Essas métricas adotadas foram usadas no estudo por apresentar um número maior de aparições dentro dos artigos usados, representando o conjunto $\mathcal{Q}$, sendo métricas estatísticas usadas para aferir a qualidade de um grupo de dados em comparação a uma resposta desejada. A descrição do método matemático dessas metodologias qualificativas utilizadas nas comparações entre as metodologias preditivas estam descritas abaixo.

R-quadrado ou *R-squared* R^2 que tem por objetivo aferir a precisão obtida pelas metodologias computacionais cujo seu valor pode variar entre 0 e 1. Quanto mais próximo de 1 melhor é a predição feita.

$R^2 = \sum_{i=1}^{N}(y_i - \hat{y}_i) / \sum_{i=1}^{N}(y_i - \bar{y})$

Erro médio absoluto ou *Mean Absolute Error* (MAE) tem por objetivo aferir a média absoluta dos erros onde quanto mais próxima de 0 melhor a qualidade da predição.

$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$

Erro quadrático médio ou *Mean Squared Error* (MSE), que tem por objetivo aferir o erro médio ao quadrado em comparação aos valores reais, onde quanto mais próxima de 0 melhor a qualidade da predição.

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

Raiz quadrada do erro médio absoluto ou *Root Mean Squared Error* ($RMSE$) este sendo a raiz quadrada da métrica anterior, MSE.

$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$

Onde temos o N que representa o número total de amostras dentro do conjunto de dados, o y_i representa o valor real do dado da i-ésima amostra, o $\hat{y}_i$ representa o valor que o modelo propõe para a mesma i-ésima amostra e o $\bar{y}$ representa a média dos dados observados em y_i ou seja ele pode ser descrito também como:

$\bar{y} = \frac{1}{N}\sum_{i=1}^{N}(y_i)$

Com as metodologias qualitativas escolhidas torna-se necessário a divisão entre grupos para o estudo comparativo. Essa divisão sera feita sobre as metodologias computacionais que representam o grupo $\mathcal{M}$ e a quantidade de dados obtida pode ser verificado na Tabela 2.5, onde mostra a quantidade de conjuntos $\mathcal{Q}$ dentro do conjunto das metodologias obtidas para o estudo, ou seja, representa somente a quantidade de aparições das métricas qualitativas marcadas na Tabela 2.4.

O primeiro grupo é formado pelas metodologias preditivas baseadas em eventos biológicos/naturais. Alguns deles que tiveram dados usados nas comparação são *Genetic Algorithm* (GA), *Grey Wolf Optimization* (GWO), *Artificial Neural Networks* (ANN), entre outros. Esses métodos foram separados nesse grupo pois sua inspiração se dá em eventos que ocorrem ordinariamente na natureza ou representam comportamentos que ocorrem naturalmente no ambiente também recebem a terminologia de alguns autores de algoritmos bioinspirados.

O segundo grupo é formado por metodologias preditivas que utilizam métodos regressivos, tais como *Clustering Methods Regression* (CMR), *Gaussian Process Regression* (GPR), *Linear Regression* (LR), entre outros. Esses métodos foram separados nesse grupo pois utilizam modelos puramente matemáticos para regredir os parâmetros da função de predição de TOC.

Esse 3 grupo criado não era esperado no início do estudo mas como houve um grande número de dados que não se enquadraram no primeiro grupo (Métodos baseados em eventos biológicos/naturais)

Tabela 2.5: Número de dados usados das metodologias encontradas

Metodologia	Quantidade do conjunto de dados Ω
ABC	1
ACO	5
ANN	17
BPANN	12
BRR	1
CMR	4
CNN	8
DDLOGR	4
DE	1
ELM	1
ENLM	1
FCDN	1
FL	6
GA	5
GB	2
GBM	1
GPR	8
GWO	1
g-GMDH	2
KNN	7
LOGR	17
LR	6
MARS	3
MLP	2
MRA	3
NGB	1
PSO	3
RBF	1
RF	9
SVR	22
XGB	4
xNES	1
	Quantidade do conjunto de dados Ω
Total	160

nem no segundo grupo (Métodos regressivos) sendo considerados métodos híbridos os mesmos foram agrupados e serão analisados como um grupo a parte.

2.2 Metodologias computacionais utilizadas pelos autores e dados usados para o treinamento

Essa seção será utilizada para explicar sobre as principais metodologias computacionais observadas na Tabela 2.5 e comentar brevemente sobre os dados usados por elas, as metodologias que são variações serão agrupadas em uma metodologia base e será descrito suas peculiaridades. As metodologias serão descritas caso tenha uma maior relevância para o estudo proposto, ou seja, tendo um número razoável de dados coletados e sendo usada por mais de um artigo.

2.2.1 Dados usados pelas metodologias computacionais

Os métodos mais citados para a coleta de dados para prever o TOC sendo estas *Pyrolysis*, muito utilizada em laboratórios para aferição do total de carbono e hidrocarbonetos presentes em uma amostra. Os métodos a seguir foram bastante utilizados nos modelos computacionais *Spectral Gamma Ray log*, *Resistivity log*, *Interval Transit Time* e *Neutron Porosity*. Onde a *Pyrolysis* acaba sendo o mais custoso para ser produzido e os demais acabam sendo mais baratos e medindo de maneira indireta o TOC.

2.2.2 Métodos específicos

Os métodos específicos usados pelos autores para resolver o problema de TOC são: $\Delta logR$ *method* LOGR *Dual-Difference* $\Delta logR$ *method* DDLOGR. Onde o LOGR foi usado por 10 artigos, e o DDLOGR sendo uma variação do LOGR acabou sendo usados por apenas 2 artigos. Sendo métodos mais consolidados nesta área, sendo muito usado como comparativo entre outros modelos computacionais.

2.2.3 Artificial Neural Networks (ANN) e suas variações

Essa metodologia foi abordada por 9 artigos tendo uma variação abordada por 6 artigos sendo esta *Back Propagation Artificial Neural Network* BPANN e varias outras que foram abordadas por vários artigos. Sua principal característica é simular o processo neural humano para resolver problemas não lineares Johnson et al. (2018), Barham et al. (2021). Uma descrição visual pode ser vista na Imagem 2.3 onde estaria representado uma ANN totalmente conectada. As entradas dos dados são representadas pelos nós $N_1, N_2, N_3...N_x$, os pesos da função representados pelos $F_1, F_2, F_3...F_y$ e as saídas seriam representadas pelos nós $O_1, O_2, O_3...O_z$.

Suas variações são modificações da estrutura do ANN base alternando a quantidade de ligações dos nós, a quantidade de representações dos nós dos dados iniciais, funções intermediarias ou até mesmo a quantidade de saídas. As variações que tiveram maior relevância tanto no número de aparições quanto na quantidade de artigos diferentes que usaram a mesma são: ANN, BPANN, *Convolutional Neural Network* CNN. Outras como *Generalized Group Method of Data Handling Neural Network* g-GMDH,

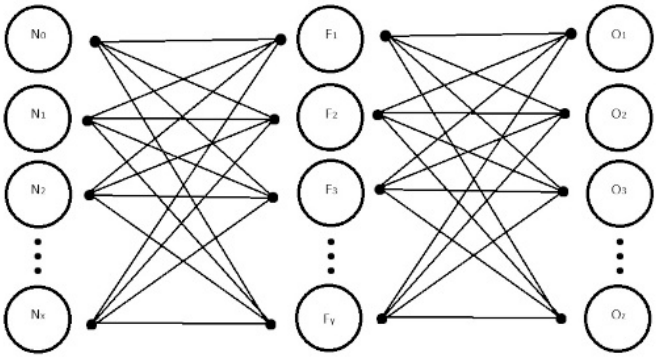

Figura 2.3: Rede Neural totalmente conectada

Fully Connected Deep Network FCDN, *Elman Neural Network* ELN, entre outras tiveram poucos dados ou não apresentaram dados validos para o estudo.

2.2.4 Modelos regressivos e suas variações

As metodologias regressivas foram abordadas por vários artigos e as que mais foram usadas foram *Support Vector Regression* SVR e *Clustering Methods Regression* CMR, vale notar alguns autores utilizam a terminologia *Support Vector Machine*. O funcionamento da metodologia SVR é utilizar um conjunto de dados e transformar os mesmos em uma função mais linear possível, na tentativa de linearizar todo o grupo de dados a partir de uma amostra menor.

A metodologia CMR seria a junção de vários algoritmos individuais para fazer a previsão final do TOC. Para esse estudo os métodos individuais não foram contabilizados na Tabela 2.5, sendo agrupados somente como CMR. Outras métodos que foram amplamente utilizados pelos autores foram: Bayesian Regularized Regression BRR, Gaussian Process Regression GPR e Linear Regression LR. Onde esses foram muito usados pelo conjunto de metodologias CMR, mas também foram usadas de forma única.

2.2.5 Métodos de otimização

As metodologias de otimização que mais se destacou nos estudos abordados foi *Genetic Algorithm* GA, valendo mencionar outras metodologia que apesar de não ter tido muitos dados coletados destaca-se pela consistência dos dados sendo esta *Ant Colony Optmization* ACO.

As duas metodologias citadas a cima são métodos bio-inspirados. Onde o funcionamento da metodologia GA é inspirada no funcionamento da seleção natural, em que o resultado é buscado de forma onde é repassado as informações entre as iterações alterando os coeficientes da equação de acordo com os processos de transmissão de D.N.A (ácido desoxirribonucleico) usando de mutação, nos coeficientes da equação, cruzamento e seleção no conjunto de resposta dado, afim de melhorar a predição da equação gerada na predição de TOC Goliatt et al. (2024), Safaei-Farouji e Kadkhodaie (2022).

A metodologia ACO seria inspirada no comportamento das formigas para encontrar uma rota

de um ponto ao outro. Onde encontram o caminho mais curto por uma trilha de feromônios que reforçam as trilhas mais percorridas, que acaba sendo as rotas mais eficientes Abdizadeh et al. (2017).

2.2.6 Boosting e árvore

Essas metodologias não foram agrupadas nos outros grupos mas destacaram-se entre os artigos estudados sendo elas: *Extreme Gradient Boosting* e *Gradient Boosting* representando os métodos *Boosting* e *Random Forests* RF representando os métodos de árvore. As três metodologias citadas são métodos de *ensemble*, que significa que eles combinam múltiplas árvores de decisão, de classificação e de regressão, para melhorar a precisão da função de predição de TOC Gharavi et al. (2022), Sun et al. (2023).

3 Resultados e discussão

Após os passos seguidos da Tabela 2.2 chegamos ao total de 38 artigos que apresentaram um total de 202 conjuntos de dados retirados dos quais 1 é relacionado a matriz de confusão, obtido no artigo (AMOSU; IMSALEM; SUN, 2021), 35 estão relacionados a medidas para determinação do TOC de maneira indireta, 6 dados não foram especificados sendo utilizados termos amplos para defini-los não sendo possível colocar-los como R ou R^2 sendo esses artigos (Asgari Nezhad; MORADZADEH; KAMALI, 2018; GHARAVI et al., 2022; REN et al., 2023), resultando assim um total de 160 conjuntos de dados distribuídos em 32 metodologias de predição Tabela 2.5 e 16 métodos qualitativos Tabela 2.4.

O número de aparições das metodologias qualitativas acabou por não ter uma boa distribuição de aparições, sendo a metodologia qualitativa com maior uso pelos autores R^2 com **115 dados**, e a menor foi APD com **3 dados** obtidos onde mesmo a metodologia mais utilizada pelos autores acaba não tendo uma aparição maior que 75% do conjunto total dos dados coletados podendo ser observados na Tabela 2.4.

Os dados das metodologias preditivas acabaram seguindo o mesmo padrão de dispersão das metodologias qualitativas tendo sua divisão exposta na Tabela 2.5. Com isso é notório a preferência por algumas metodologias preditivas sendo elas SVR, com **21 conjuntos de dados** retirados, ANN e LOGR, com **17 conjuntos de dados** retirados.

O mesmo não se aplica à localização dos dados utilizados nos estudos onde dos 38 artigos, observados na Tabela 2.3, 18 artigos utilizaram informações de poços coletados da China seguido por 4 do Canada. Alguns desses artigos adotaram uma validação cruzada utilizando-se de mais de um poço, sendo assim a Figura 3.1 apresenta a soma maior que o numero de artigos. Vale ressaltar que alguns estudos acabam por utilizar a mesma área geográfica nos seus estudos, utilizando-se de validação cruzada de um mesmo poço variando somente a quantidade de dados.

As tabelas que discriminam a quantidade de dados utilizada no estudo são Tabela 3.1,Tabela 3.2 e Tabela 3.3. Com as quantidades de dados expostos fica evidente que a comparação em relação ao tempo não poderá ser respondida da melhor maneira possível, pois o número de iterações para o treinamento disponível no conjunto de dados coletados é baixa, sendo **32 conjunto de dados** que apresentam essa informação. Para sanar esse problema será utilizada as informações de iterações de artigos que fizeram o estudo para encontrar o melhor numero de iteração para o problema abordado, com isso nem todas as metodologias computacionais pode ser aferido o numero de iterações. A impossibilidade de expandir essa métrica de um estudo de uma metodologia computacional específica para outras não estudadas é devido aos metadados diferentes inerentes de cada metodologia.

Figura 3.1: Localidade dos poços usados pelos artigos estudados

Tabela 3.1: Grupo 1 (Métodos baseados em eventos biológicos\naturais)

Metodologia	**Métricas qualitativas**			
	R^2	MAE	$RMSE$	MSE
ABC	1	1	1	0
ACO	5	0	0	5
ANN	15	1	4	3
BPANN	9	4	5	5
CNN	8	2	0	2
DE	1	1	1	0
FCDN	1	0	0	0
FL	1	4	0	5
GA	5	1	1	4
GWO	1	1	1	0
g-GMDH	1	2	0	2
MLP	1	1	1	0
PSO	3	1	3	0
xNES	1	1	1	0
	R^2	MAE	$RMSE$	MSE
Total	53	20	18	26

3.1 Análise dos resultados das metodologias

Com os dados agrupados os mesmos foram classificados entre os grupos e depois comparados os melhores de cada grupo, utilizando as métricas citadas na Tabela 2.4 e essa comparação foi feita por média simples e a maior variação da média encontrada de cada métrica qualitativa, representada pelo conjunto $\mathcal{Q}$, de cada metodologia computacional, representada pelo conjunto $\mathcal{M}$, com o objetivo de encontrar o método

Tabela 3.2: Grupo 2 (Métodos regressivos)

Metodologia	**Métricas qualitativas (Quantidade)**			
	R^2	MAE	$RMSE$	MSE
CMR	0	4	0	4
ENLM	1	1	1	0
GPR	4	7	1	8
LR	5	5	4	1
MARS	3	1	3	2
MRA	1	2	3	0
SVR	21	11	14	7
	R^2	MAE	$RMSE$	MSE
Total	35	31	26	22

Tabela 3.3: Grupo 3 (Métodos não agrupados no Grupo 1 nem no Grupo 2

Metodologia	**Métricas qualitativas (Quantidade)**			
	R^2	MAE	$RMSE$	MSE
BRR	1	1	0	0
DDLOGR	0	4	0	4
ELM	1	1	1	0
GB	0	2	2	0
GBM	1	0	1	1
KNN	7	7	7	0
LOGR	6	14	4	10
NGB	0	0	0	0
RBF	0	1	0	1
RF	8	5	4	2
XGB	3	2	1	1
	R^2	MAE	$RMSE$	MSE
Total	27	37	20	19

que sofreu menos variação pelos autores e poder dizer qual deles apresenta melhores resultados para o problema proposto de predição de TOC. Sendo assim a Tabela 3.4 foi montada para expor os resultados obtidos para as comparações feitas, onde as expressões usadas estão descritas abaixo onde N representa o total de dados coletados e x_i o dado coletado.

- Média = $\frac{1}{N}\sum_{i=1}^{N} x_i$
- Variação = $MAX(\left|x_i - \frac{\sum_{i=1}^{N} x_i}{N}\right|)$

Com os dados expostos na Tabela 3.4 é possível fazer as comparações mais aprofundadas. Apesar de ”Variação”estar em módulo o sinal dela foi colocado para representar o valor que mais se afastou da média e os valores em negrito seriam os melhores valores de acordo com cada métrica qualitativa. Entretanto o valor da variação que esta em negrito refere-se a menor variação da média e não diretamente a melhor média ou ao valor bruto que ela representa. Esse valor bruto faz referencia ao dado retirado dos artigos onde na tabela temos representado do seguinte modo $Dado_B = x(-)$ onde o x seria um próprio

Tabela 3.4: Dados usados na comparação

Metodologia	**Média(Variação)**			
	R^2	MAE	$RMSE$	MSE
ABC	0.7840(-)	0.3430(-)	0.2210(-)	-
ACO	0.8836(-0.2357)	-	-	0.0101(0.0099)
ANN	0.7089(-0.5519)	0.0322(-)	0.2007(1.2845)	0.2324(1.0058)
BPANN	0.8147(-0.2996)	0.4459(0.3737)	0.7748(1.2845)	0.2631(1.0058)
BRR	**0.9630**(-)	3.2500(-)	-	0.0101(-)
CMR	-	0.9435(0.3355)	-	2.0813(1.4168)
CNN	0.7474(-0.2374)	0.4250(-0.0850)	-	0.4250(-0.0850)
DDLOGR	-	0.6310(0.1660)	-	0.5970(-0.2620)
DE	0.7910(-)	0.3320(-)	0.2130(-)	-
ELM	0.8140(-)	0.3110(-)	0.4710(-)	-
ENLM	0.8140(-)	0.3150(-)	0.4700(-)	-
FCDN	0.8900(-)	-	-	-
FL	0.9425(-)	0.7580(-0.1710)	-	0.9622(-0.8403)
GA	0.9006(-0.1166)	0.3380(-)	0.2160(-)	**0.0094**(**-0.0017**)
GB	-	**0.0314**(**0.0001**)	**0.0392**(**-0.0087**)	-
GBM	0.6200(-)	-	0.6600(-)	0.4300(-)
GPR	0.8089(0.1431)	0.9548(0.8672)	0.5629(-)	2.6234(4.6376)
GWO	0.7860(-)	0.3230(-)	0.2140(-)	-
g-GMDH	0.9336(-)	0.9151(-0.4601)	-	1.3797(-0.9755)
KNN	0.8896(-0.1494)	0.3165(0.7065)	0.5141(0.7539)	-
LOGR	0.5219(-0.437)	0.9568(1.2032)	0.3837(0.6863)	4.2378(12.9892)
LR	0.8174(0.1326)	1.4782(1.8318)	1.2670(0.2530)	0.2423(-)
MARS	0.6967(-0.1367)	0.3430(-)	0.5800(0.1200)	0.3900(-0.1000)
MLP	0.7957(-)	0.9350(-)	1.1890(-)	-
MRA	0.9346(-)	0.3407(-0.0349)	0.4139(0.2033)	-
NGB	-	-	-	-
PSO	0.8835(-0.0995)	0.3400(-)	0.5973(0.6082)	-
RBF	-	0.7040(-)	-	0.6813(-)
RF	0.8529(-0.3029)	1.1101(2.2199)	0.8010(0.3890)	0.5455(-0.1355)
SVR	0.7435(-0.6135)	0.8016(1.7591)	0.8383(0.8560)	1.5242(7.6161)
XGB	0.8568(**0.0567**)	0.4284(-0.2017)	0.7700(-)	0.1071(-)
xNES	0.7720(-)	0.3340(-)	0.2210(-)	-

Variação representada com -"significa que essa métrica tem somente um dado.

dado retirado dos artigos e outro modo seria $Dado_B = x(y)$ onde o x representa a média desses valores, o y representa a maior variação obtida dos valores em relação à média dos valores e $x + y$ representa um dos valores obtidos dos artigos e este seria o dado que mais se desviou da média.

Apesar disso não podemos tirar as conclusões diretamente dos valores em negrito da Tabela 3.4, pois os mesmos acabam não levando em conta a quantidade de aparições no estudo onde temos a método BRR como melhor R^2, no valor de 0.9630, entretanto esse dado acaba não tendo comparativos entre ele próprio, ou seja, não há mais representantes desse método e consideraremos como um caso particular de sucesso. O mesmo é valido para o MLP ao qual apresenta o segundo pior resultado para $RMSE$, com

o valor de 1.1890, somente atrás do LR, com valor 1.2670, onde seria considerado um caso particular de fracasso. Nesse mesma explicação o LR apresenta um resultado muito elevado na métrica $RMSE$ mas o mesmo apresenta 4 dados retirados dos artigos onde assim temos uma grande média. Contudo essa média elevada se dá devido a um artigo que apresentou 3 dados dos 4 coletados e esse artigo em especifico seria (GOLIATT; SAPORETTI; PEREIRA, 2023), onde teria vários métodos estacados para obter o resultado e alguns do seus modelos foram considerados como LR.

Aprofundando mais nas comparações feitas esse caso apresentado anteriormente ocorre em alguns dos dados apresentados, onde um artigo aborda sozinho uma metodologia computacional variando seus metadados, isso acaba por inflar o numero de aparições de algumas metodologias computacionais. Sendo assim fazer algumas comparações isoladas utilizando somente os dados apresentados na Tabela 3.4 não representam os resultados em sua totalidade onde também deve-se avaliar a distribuição das metodologias por artigos, metodologias que apresentam o valor de iteração no período de treinamento, entre outros fatores foram considerados para fazer o ranqueamento das metodologias computacionais.

Com essas considerações feitas podemos analisar os dados dos três grupos formados para o estudo, assim temos no Grupo 1, representados na Tabela 3.1, 2 metodologias se destacaram pela quantidade de aparições sendo elas ANN e BPANN com 17 e 12 conjunto de dados respectivamente. O BPANN acaba tendo um conjunto de dados mais consistentemente bons entre suas aparições sendo o melhor dentro do Grupo 1, onde essa metodologia seria uma modificação na própria ANN que por sua vez apresentando uma maior quantidade de dados podemos inferir que sua média do R^2 tenderia a um bom resultado por volta de $R^2 \approx 0.7$ já o BPANN teria por volta de $R^2 \approx 0.8$.

Analisando o Grupo 2, representados pela Tabela 3.2, tivemos um método que se destacou tanto por sua quantidade de aparições quanto por sua consistência nas métricas qualitativas de erro sendo ele o SVR com um total de 22 conjunto de dados, tendo a maior quantidade de dados com R^2 com a media tendo próximo da ANN, porem suas métricas de erro acabaram destacando-se por serem bem menores que outras metodologias de seu grupo com uma boa quantidade de dados.

Analisando o Grupo 3, com suas metodologias representadas pela tabela 3.3, tivemos a maior aparição dos métodos Logaritmos sendo amplamente usado por diversos autores como comparativo de outras metodologias não tendo o foco direto no estudo, isso pode ter influenciado nas médias de seus dados acabando ficando pior que a segunda metodologia que mais apresentou conjunto de dados RF, com 16 dados para LOGR e 9 para RF, onde a média de suas métricas de R^2 tiveram ≈ 0.25 de vantagem para RF obtendo uma boa media de R^2 dentro do Grupo 3.

3.2 Conclusão

Após essas análises temos que, em maioria, as melhores metodologias foram as que mais foram estudadas pelos autores dentro desse tema em específico, sendo elas BPANN, SVR e RF. O BPANN se destacou

devido a uma boa média de R^2 e as métricas de erro apesar de elevadas não tiveram grande variação como o ANN, já o SVR se destacou devido à quantidade de aparições e não por não ter grandes variações nos estudos com foco nele e o RF por uma boa consistência de dados, porém o LOGR apresenta melhores médias nas métricas de erros.

Para tentar responder o questionamento do tempo os artigos estudados não apresentam essa métrica diretamente, entretanto podemos adotar a medida de complexidade temporal para tentar responder essa pergunta. Adotando essa métrica temos que utilizar algumas inferências sobre os hiperparâmetros que serão adotados, sendo que serão adotados os dados dos artigos que fizeram os estudos sobre esses aspectos . Com isso proposto podemos montar um cenário onde essas variáveis foram adotadas seguindo alguns artigos (ZHANG; WU; WU, 2023; HANDHAL et al., 2020) como exemplo, que fizeram testes sobre os hiperparâmetros para adotar-los dentro da conta e usá-los como se todos os artigos tivessem usado essas configurações nos seus estudos, fazendo assim uma homogeneização dos hiperparâmetros. Substituindo os valores mencionados pelos artigos e usando como número fixo o $N = 100$ e $E = 10^3$. As equações de complexidade temporal das 3 metodologias consideradas como as melhores dentro dos seus respectivos grupos estão descritas abaixo, onde E representa o número de épocas e W o número de conexões entre os nós da metodologia BPANN, o T seria a representação do número de arvores adotadas na metodologia RF e N representa o número de amostras estudadas por todas as metodologias.

- BPANN $O(E * N * W) \Rightarrow O(10^3 * 100 * 40) \Rightarrow O(4000000)$
- SVR $O(N^3) \Rightarrow O(100^3) \Rightarrow O(1000000)$
- RF $O(T * N * logN) \Rightarrow O(37 * 100 * log100) \Rightarrow O(7400)$

Neste cenário teríamos RF, SVR, BPANN do menor tempo de execução para o maior respectivamente. Entretanto devido a que esses parâmetros usados podem não representar um cenário ideal pois esses dados usados não foram uma média dos artigos tendo em vista que dentro dos dados coletados houve uma grande variação nas iterações, no período do treinamento das metodologias, quanto no equipamento usado para executar esses modelos e esses metadados coletados não obtiveram um valor aceitável para ser levado em consideração para essa análise.

É notória a falta de informações complementares para a realização mais apurada das comparações propostas sendo talvez algo relacionado somente a essa área específica sendo preciso outras análises em mais áreas de conhecimento que usam metodologias preditivas para saber se essa falta de informações seria somente nesta ou ocorre em mais áreas de conhecimento.

O link de acesso para a ferramenta usada na revisão dos artigos ainda não está disponível devido à criação do projeto como privado no *Rayyan*[5] onde o pedido de mudança para um projeto público está em análise do suporte deles. Caso queira ver o projeto feito na ferramenta entre em contato pelo e-mail matheuscasarim@hotmail.com juntamente com o seu e-mail para poder ser convidado para o projeto.

[5]<https://www.rayyan.ai/>

Bibliografia

ABDIZADEH, H. et al. Estimation of total organic carbon from well logs and seismic sections via neural network and ant colony optimization approach: a case study from the mansuri oil field, sw iran. *Geopersia*, Univrsity Of Tehran Press, v. 7, n. 2, p. 255–266, 2017. ISSN 2228-7817. Disponível em: <https://geopersia.ut.ac.ir/article_61899.html>.

AMOSU, A.; IMSALEM, M.; SUN, Y. Effective machine learning identification of toc-rich zones in the eagle ford shale. *Journal of Applied Geophysics*, v. 188, p. 104311, 2021. ISSN 0926-9851. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0926985121000586>.

ASANTE-OKYERE, S.; MARFO, S. A.; ZIGGAH, Y. Y. Estimating total organic carbon (toc) of shale rocks from their mineral composition using stacking generalization approach of machine learning. *Upstream Oil and Gas Technology*, v. 11, p. 100089, 2023. ISSN 2666-2604. Disponível em: <https://www.sciencedirect.com/science/article/pii/S266626042300004X>.

ASANTE-OKYERE, S.; ZIGGAH, Y. Y.; MARFO, S. A. Improved total organic carbon convolutional neural network model based on mineralogy and geophysical well log data. *Unconventional Resources*, v. 1, p. 1–8, 2021. ISSN 2666-5190. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666519021000017>.

Asgari Nezhad, Y.; MORADZADEH, A.; KAMALI, M. R. A new approach to evaluate organic geochemistry parameters by geostatistical methods: A case study from western australia. *Journal of Petroleum Science and Engineering*, v. 169, p. 813–824, 2018. ISSN 0920-4105. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0920410516308439>.

BAI, Y.; TAN, M. Dynamic committee machine with fuzzy-c-means clustering for total organic carbon content prediction from wireline logs. *Computers & Geosciences*, v. 146, p. 104626, 2021. ISSN 0098-3004. Disponível em: <https://www.sciencedirect.com/science/article/pii/S009830042030604X>.

BARHAM, A. et al. Predicting the maturity and organic richness using artificial neural networks (anns): A case study of montney formation, ne british columbia, canada. *Alexandria Engineering Journal*, v. 60, n. 3, p. 3253–3264, 2021. ISSN 1110-0168. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1110016821000405>.

BOLANDI, V.; KADKHODAIE, A.; FARZI, R. Analyzing organic richness of source rocks from well log data by using svm and ann classifiers: A case study from the kazhdumi formation, the persian gulf basin, offshore iran. *Journal of Petroleum Science and Engineering*, v. 151, p. 224–234, 2017. ISSN 0920-4105. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0920410517300189>.

GHARAVI, A. et al. Application of machine learning techniques for identifying productive zones in unconventional reservoir. *International Journal of Intelligent Networks*, v. 3, p. 87–101, 2022. ISSN 2666-6030. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666603022000094>.

GOLIATT, L. et al. Performance of evolutionary optimized machine learning for modeling total organic carbon in core samples of shale gas fields. *Petroleum*, v. 10, n. 1, p. 150–164, 2024. ISSN 2405-6561. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2405656123000354>.

GOLIATT, L.; SAPORETTI, C.; PEREIRA, E. Super learner approach to predict total organic carbon using stacking machine learning models based on well logs. *Fuel*, v. 353, p. 128682, 2023. ISSN 0016-2361. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0016236123012954>.

HANDHAL, A. M. et al. Prediction of total organic carbon at rumaila oil field, southern iraq using conventional well logs and machine learning algorithms. *Marine and Petroleum Geology*, v. 116, p. 104347, 2020. ISSN 0264-8172. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0264817220301306>.

HASSAN, A. et al. Prediction of total organic carbon in organic-rich shale rocks using thermal neutron parameters. *ACS Omega*, v. 8, n. 5, p. 4790–4801, 2023. Disponível em: <https://doi.org/10.1021/acsomega.2c06918>.

HU, S. et al. Quantitative interpretation of toc in complicated lithology based on well log data: A case of majiagou formation in the eastern ordos basin, china. *Applied Sciences*, v. 11, n. 18, 2021. ISSN 2076-3417. Disponível em: <https://www.mdpi.com/2076-3417/11/18/8724>.

JOHNSON, L. M. et al. Geochemical property modelling of a potential shale reservoir in the canning basin (western australia), using artificial neural networks and geostatistical tools. *Computers & Geosciences*, v. 120, p. 73–81, 2018. ISSN 0098-3004. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0098300417307914>.

LIU, X.; TIAN, Z.; CHEN, C. Total organic carbon content prediction in lacustrine shale using extreme gradient boosting machine learning based on bayesian optimization. *Geofluids*, Wiley Online Library, v. 2021, n. 1, p. 6155663, 2021.

MKONO, C. N. et al. Deep learning integrated approach for hydrocarbon source rock evaluation and geochemical indicators prediction in the jurassic - paleogene of the mandawa basin, se tanzania. *Energy*, v. 284, p. 129232, 2023. ISSN 0360-5442. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0360544223026269>.

MULASHANI, A. K. et al. Group method of data handling (gmdh) neural network for estimating total organic carbon (toc) and hydrocarbon potential distribution (s1, s2) using well logs. *Natural Resources Research*, Springer, v. 30, n. 5, p. 3605–3622, 2021.

NYAKILLA, E. E. et al. Evaluation of source rock potentiality and prediction of total organic carbon using well log data and integrated methods of multivariate analysis, machine learning, and geochemical analysis. *Natural Resources Research*, Springer, v. 31, n. 1, p. 619–641, 2022.

REIS, M. A. A. d. A. d. et al. Source rock evaluation from rock to seismic data: An integrated machine-learning-based work flow and application in the brazilian presalt (santos basin). *Minerals*, v. 13, n. 9, 2023. ISSN 2075-163X. Disponível em: <https://www.mdpi.com/2075-163X/13/9/1179>.

REN, Y. et al. Characteristics, classification and knn-based evaluation of paleokarst carbonate reservoirs: A case study of feixianguan formation in northeastern sichuan basin, china. *Energy Geoscience*, v. 4, n. 3, p. 100156, 2023. ISSN 2666-7592. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666759223000021>.

RUI, J. et al. Toc content prediction based on a combined gaussian process regression model. *Marine and Petroleum Geology*, v. 118, p. 104429, 2020. ISSN 0264-8172. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0264817220302129>.

RUI, J. et al. Total organic carbon content prediction based on support-vector-regression machine with particle swarm optimization. *Journal of Petroleum Science and Engineering*, v. 180, p. 699–706, 2019. ISSN 0920-4105. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0920410519305625>.

SAFAEI-FAROUJI, M.; KADKHODAIE, A. Application of ensemble machine learning methods for kerogen type estimation from petrophysical well logs. *Journal of Petroleum Science and Engineering*, v. 208, p. 109455, 2022. ISSN 0920-4105. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0920410521010986>.

SANTOS, C. M. d. C.; PIMENTA, C. A. d. M.; NOBRE, M. R. C. The pico strategy for the research question construction and evidence search. *Revista Latino-Americana de Enfermagem*, Escola de Enfermagem de Ribeirão Preto / Universidade de São Paulo, v. 15, n. 3, p. 508–511, Jun 2007. ISSN 0104-1169. Disponível em: <https://doi.org/10.1590/S0104-11692007000300023>.

SAPORETTI, C. et al. Hybrid machine learning models for estimating total organic carbon from mineral constituents in core samples of shale gas fields. *Marine and Petroleum Geology*, v. 143, p. 105783, 2022. ISSN 0264-8172. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0264817222002616>.

SHALABY, M. R. et al. Integrated toc prediction and source rock characterization using machine learning, well logs and geochemical analysis: Case study from the jurassic source rocks in shams field, nw desert, egypt. *Journal of Petroleum Science and Engineering*, v. 176, p. 369–380, 2019. ISSN 0920-4105. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0920410519300622>.

SHALABY, M. R. et al. Thermal maturity and toc prediction using machine learning techniques: case study from the cretaceous–paleocene source rock, taranaki basin, new zealand. *Journal of Petroleum Exploration and Production Technology*, Springer, v. 10, p. 2175–2193, 2020.

SIDDIG, O.; IBRAHIM, A. F.; ELKATATNY, S. Application of various machine learning techniques in predicting total organic carbon from well logs. *Computational Intelligence and Neuroscience*, Wiley Online Library, v. 2021, n. 1, p. 7390055, 2021.

SUN, J. et al. Prediction of toc content in organic-rich shale using machine learning algorithms: Comparative study of random forest, support vector machine, and xgboost. *Energies*, v. 16, n. 10, 2023. ISSN 1996-1073. Disponível em: <https://www.mdpi.com/1996-1073/16/10/4159>.

TABATABAEI, S. M. E. et al. A hybrid stochastic-gradient optimization to estimating total organic carbon from petrophysical data: A case study from the ahwaz oilfield, sw iran. *Journal of Petroleum Science and Engineering*, v. 127, p. 35–43, 2015. ISSN 0920-4105. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0920410515000297>.

WANG, G. et al. Identifying organic-rich marcellus shale lithofacies by support vector machine classifier in the appalachian basin. *Computers & Geosciences*, v. 64, p. 52–60, 2014. ISSN 0098-3004. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0098300413003002>.

WANG, H. et al. An improved neural network for toc, s1 and s2 estimation based on conventional well logs. *Journal of Petroleum Science and Engineering*, v. 176, p. 664–678, 2019. ISSN 0920-4105. Disponível em: <https://www.sciencedirect.com/science/article/pii/S092041051930110X>.

WANG, P.; PENG, S. A new scheme to improve the performance of artificial intelligence techniques for estimating total organic carbon from well logs. *Energies*, v. 11, n. 4, 2018. ISSN 1996-1073. Disponível em: <https://www.mdpi.com/1996-1073/11/4/747>.

WOOD, D. A. Total organic carbon predictions from lower barnett shale well-log data applying an optimized data matching algorithm at various sampling densities. *Pure and Applied Geophysics*, Springer, v. 177, n. 11, p. 5451–5468, 2020.

WOOD, D. A. Predicting total organic carbon from few well logs aided by well-log attributes. *Petroleum*, v. 9, n. 2, p. 166–182, 2023. ISSN 2405-6561. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2405656122000657>.

ZHANG, H.; WU, W.; WU, H. Toc prediction using a gradient boosting decision tree method: A case study of shale reservoirs in qinshui basin. *Geoenergy Science and Engineering*, v. 221, p. 111271, 2023. ISSN 2949-8910. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0920410522011238>.

ZHANG, W. et al. A deep encoder-decoder neural network model for total organic carbon content prediction from well logs. *Journal of Asian Earth Sciences*, v. 240, p. 105437, 2022. ISSN 1367-9120. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1367912022003686>.

ZHENG, D.; WU, S.; HOU, M. Fully connected deep network: An improved method to predict toc of shale reservoirs from well logs. *Marine and Petroleum Geology*, v. 132, p. 105205, 2021. ISSN 0264-8172. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0264817221003081>.

ZHU, L. et al. An improved method for evaluating the toc content of a shale formation using the dual-difference δlogr method. *Marine and Petroleum Geology*, v. 102, p. 800–816, 2019. ISSN 0264-8172. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0264817219300315>.