Application of Machine Learning Ensembles to Oil Production Prediction

Letícia M. B. de Lucena[1], **Isaac Lima Oliveira Filho**[2], **Amanda Gondim de Oliveira**[1]

[1]Universidade Federal Rural do Semi-Árido (UFERSA)
Mossoró – RN – Brasil

[2]Universidade do Estado do Rio Grande do Norte (UERN)
Mossoró – RN – Brasil

leticia.lucena@alunos.ufersa.edu.br

isaacoliveira@uern.br, amandagondim@ufersa.edu.br

Abstract. *This work investigates the application of machine learning models, both individual and ensemble, for predicting daily oil production. The problem was formulated as supervised regression, with the daily oil volume as the target. A preprocessing step was carried out with a focus on avoiding data leakage, including the removal of highly correlated variables and the use of pipelines. Five individual models and different ensemble approaches, such as Bagging, Voting, and Stacking, were tested. The best performance was achieved by a heterogeneous committee with k-NN and Decision Tree in Bagging, reaching* R^2 *of* 0.9897 *and MAE of* 52.9 *Sm³, demonstrating the performance gain from combining diverse models.*

Resumo. *Este trabalho investiga a aplicação de modelos de aprendizado de máquina, tanto individuais quanto em ensemble, para prever a produção diária de petróleo. O problema foi formulado como regressão supervisionada, tendo como alvo o volume diário de petróleo. Foi realizada uma etapa de pré-processamento com foco em evitar vazamento de dados, incluindo a remoção de variáveis altamente correlacionadas e o uso de pipelines. Foram testados cinco modelos individuais e diferentes abordagens de ensemble, como Bagging, Voting e Stacking. O melhor desempenho foi obtido por um comitê heterogêneo com k-NN e Árvore de Decisão em Bagging, alcançando* R^2 *de* $0,9897$ *e MAE de* $52,9$ *Sm³, evidenciando o ganho de desempenho com a combinação de modelos diversos.*

1. Introdução

A indústria do petróleo e gás é um dos pilares da economia global, atendendo a mais de 50% das necessidades globais de energia, com o petróleo sendo um recurso fundamental para diversos setores, como os de energia, transporte e produção industrial [Salimovna 2024]. Além disso, o lucro produzido pela exportação de petróleo tem sido usado de forma inteligente pelos países para o investimento em diversos setores, como saúde e infraestrutura [Pop 2018].

A previsão precisa da produção de petróleo é essencial para o gerenciamento de reservatórios, planejamento do desenvolvimento de campos e avaliação econômica. Um

dos métodos utilizados é a Análise de Curvas de Declínio (ACD), amplamente utilizada para estimar a produção de óleo, gás e água ao longo do tempo. Contudo, ela apresenta limitações importantes ao lidar com comportamentos complexos e não lineares dos reservatórios reais, o que pode resultar em previsões menos precisas[Dike et al. 2025].

Nesse contexto, métodos baseados em aprendizado de máquina surgem como alternativas promissoras, capazes de capturar relações mais complexas entre variáveis operacionais e a produção, aumentando a precisão das estimativas [Daramola et al. 2024]. Por isso, este trabalho propõe a aplicação e avaliação de modelos individuais e métodos *ensemble* de *Machine Learning* (ML) para a predição da produção diária de petróleo, utilizando a base pública *Volve Production Data*.

Este artigo está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados; em seguida, a Seção 3 aborda a metodologia utilizada; a Seção 4 expõe os resultados e discussões; por fim, a Seção 5 apresenta as considerações finais.

2. Trabalhos relacionados

[Calvette 2019] investiga a aplicação de técnicas de ML e abordagens *data-driven* na previsão da produção de óleo, gás e água em poços inteligentes, utilizando uma base real e outra sintética. Os resultados indicam que os modelos de *deep learning* com redes neurais recorrentes superam significativamente as abordagens tradicionais, mostrando reduções do erro RMSE e MAE de 66% e 79%, respectivamente. A autora concluiu que técnicas avançadas de *deep learning* aumentam a precisão preditiva e a eficiência na modelagem da produção em reservatórios inteligentes.

[Samad et al. 2025] apresenta uma abordagem baseada em *machine learning* para prever a produção futura de um campo de petróleo, usando dados reais do campo Volve, na Noruega. Foram testados três modelos que alcançaram R^2 de mais de 0,99 nos testes. O estudo destaca as vantagens do *machine learning* sobre métodos tradicionais, como análise de declínio e simulações de reservatório, especialmente diante da complexidade e variabilidade dos parâmetros envolvidos na produção.

[Ghodsi et al. 2025] propõe a aplicação de *Stacking Ensemble Learning* para a previsão da taxa de produção de petróleo, utilizando um conjunto de modelos de *machine learning* otimizados por técnicas de ajuste de hiperparâmetros. Os resultados indicam que o modelo de *ensemble* supera tanto modelos individuais quanto correlações empíricas tradicionais, apresentando melhor desempenho preditivo e maior capacidade de generalização. O estudo evidencia, assim, a eficácia de estratégias de combinação de modelos na melhoria da precisão em problemas de previsão de produção em reservatórios.

3. Metodologia

Essa seção aborda os procedimentos metodológicos adotados neste trabalho, desde a fonte dos dados usados até o treinamento dos modelos de ML.

3.1. Base de dados

A base de dados utilizada neste trabalho foi a Volve Data Production, disponibilizada de forma pública pela empresa Equinor[1]. Essa base contém dados sobre a produção de petróleo no campo Volve, que operou entre 2008 e 2016.

[1]Disponível em: Volve Data Village

A base *Volve Production Data* contém 15.634 instâncias e 24 atributos, distribuídos entre poços produtores e poços injetores de água. Os dados vão de setembro de 2007 a dezembro de 2016, ano em que as atividades foram encerradas. Cada instância corresponde a um registro de um único poço em determinado dia. A variável-alvo é `BORE_OIL_VOL`, que representa o volume diário de óleo produzido, expresso em metros cúbicos padrão (Sm^3).

3.2. Pré-processamento

A primeira etapa do pré-processamento foi remover todas as colunas que não possuíam valores preditivos: 'DATEPRD', 'WELL_BORE_CODE', 'NPD_WELL_BORE_CODE', 'NPD_WELL_BORE_NAME', 'NPD_FIELD_CODE', 'NPD_FIELD_NAME', 'NPD_FACILITY_CODE', 'NPD_FACILITY_NAME', 'FLOW_KIND', 'AVG_CHOKE_UOM', 'BORE_WI_VOL'. Essas colunas se tratavam apenas de códigos e rótulos de identificação.

Em seguida, como pode ser visto na Figura 1 foi calculada a correlação de Pearson entre os atributos restantes para identificar variáveis que possibilitassem o vazamento de dados durante a predição. Com isso, foi possível observar que a variável `BORE_GAS_VOL` era 99% relacionada com nosso atributo alvo, sendo assim, ela foi removida.

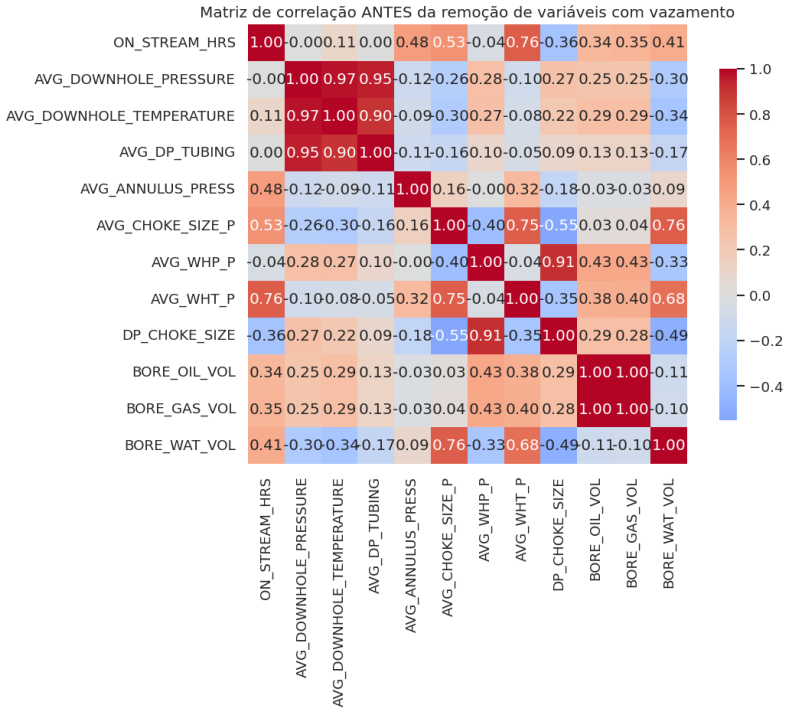


Figura 1. Matriz de correlação entre os atributos

As demais correlações com `BORE_OIL_VOL`, após a exclusão de `BORE_GAS_VOL`, são apresentadas na Tabela 1. As variáveis de maior correlação linear com a produção de óleo foram a pressão na cabeça do poço (`AVG_WHP_P`, $r = 0{,}431$) e a temperatura na cabeça do poço (`AVG_WHT_P`, $r = 0{,}379$), seguidas pelas horas de operação (`ON_STREAM_HRS`, $r = 0{,}339$). A variável `BORE_WAT_VOL` apresentou correlação fraca e negativa ($r = -0{,}116$).

Tabela 1. Correlação entre os atributos preditores finais e a variável-alvo BORE_OIL_VOL

	Correlação
AVG_WHP_P	0.4306
AVG_WHT_P	0.3795
ON_STREAM_HRS	0.3385
DP_CHOKE_SIZE	0.2891
AVG_DOWNHOLE_TEMPERATURE	0.2889
AVG_DOWNHOLE_PRESSURE	0.2479
AVG_DP_TUBING	0.1263
AVG_CHOKE_SIZE_P	0.0251
BORE_WAT_VOL	-0.1163
AVG_ANNULUS_PRESS	-0.0302
WELL_TYPE_WI	-0.0309

O tratamento de valores ausentes foi realizado em dois níveis: linhas com cinco ou mais sensores ausentes (de um total de oito) foram eliminadas por carregarem informação insuficiente para a modelagem; já os demais valores ausentes foram preservados para imputação posterior. A imputação por mediana foi realizada exclusivamente dentro do *Pipeline* scikit-learn, ajustada apenas com os dados de treinamento, garantindo que não houvesse vazamento de informações para o conjunto de teste. Foram removidas ainda 33 instâncias duplicadas detectadas na base.

Não foram eliminados *outliers* de variáveis operacionais ou de produção, pois picos e quedas abruptas eram esperados no contexto de produção e carregam informação real do reservatório.

O único atributo categórico remanescente, `WELL_TYPE` (com categorias *OP* e *WI*), foi convertido por codificação *one-hot*, gerando a variável binária `WELL_TYPE_WI`. Tal atributo carrega a informação operacional relacionada ao tipo de poço: OP, sendo os poços produtores; WI, sendo os poços injetores. A base final pré-processada contém 9.122 instâncias e 12 atributos, incluindo a variável-alvo.

A normalização das variáveis numéricas, necessária para algoritmos sensíveis à escala (k-NN, SVR, MLP), é realizada por `StandardScaler` dentro do *Pipeline*, calculada exclusivamente nos dados de treinamento de cada partição, nunca sobre a base completa.

Além da pré-processada, foram geradas quatro bases adicionais para análise comparativa:

- **Base 1 — PCA-95**: componentes principais que retêm 95% da variância explicada (6 componentes);
- **Base 2 — PCA-90**: componentes principais que retêm 90% da variância (5 componentes);
- **Base 3 — PCA-2D**: apenas as duas primeiras componentes principais (fins de visualização);
- **Base 4 — Reduzida por Importância**: três atributos selecionados com base na importância extraída da Árvore de Decisão (`AVG_WHT_P`, `AVG_WHP_P` e

BORE_WAT_VOL), responsáveis por mais de 84% da importância acumulada.

3.3. Modelos de Machine Learning

O problema foi abordado como regressão supervisionada. Como base, foram escolhidos cinco modelos individuais avaliados: K-Nearest Neighbors Regressor (k-NN), Decision Tree Regressor, Linear Regression, MLP Regressor e Support Vector Regressor (SVR). A Figura 2 exemplifica visualmente o processo de treinamento dos modelos individuais até os comitês e as métricas de avaliação.

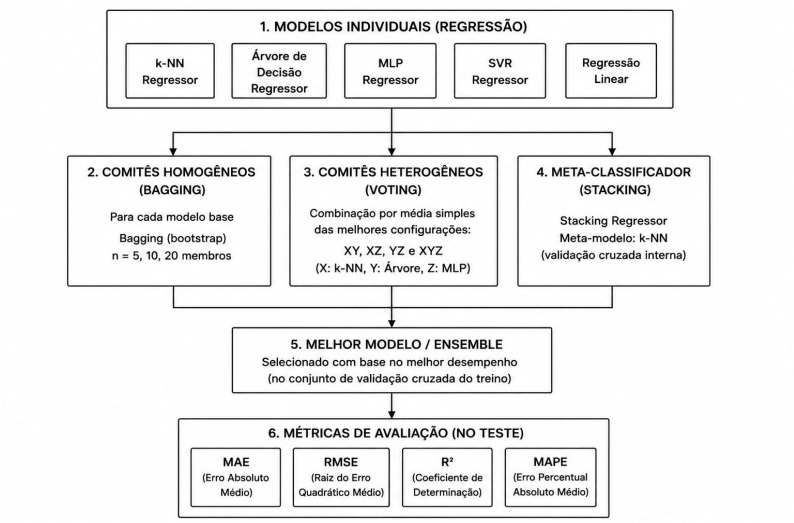


Figura 2. Fluxograma sobre os modelos utilizados

Para todos os modelos, a otimização de hiperparâmetros foi realizada via GridSearchCV com validação cruzada K-Fold ($k = 5$) aplicada ao conjunto de treinamento, com o *Pipeline* completo encapsulado dentro da busca, garantindo que o pré-processamento seja recalculado em cada *fold* sem contaminação do conjunto de validação interna. Os hiperparâmetros utilizados podem ser observados na Tabela 2.

Tabela 2. Hiperparâmetros avaliados e melhores valores encontrados.

Modelo	Hiperparâmetro	Espaço de busca	Melhor valor
k-NN	n_neighbors	{3, 5, 10, 15}	3
k-NN	weights	{uniform, distance}	distance
Árvore de Decisão	max_depth	{4, 8, 12, None}	None
Árvore de Decisão	min_samples_leaf	{1, 5, 10}	5
MLP	hidden_layer_sizes	{(50,), (30, 15)}	(50,)
MLP	alpha	$\{10^{-3}, 10^{-4}\}$	10^{-4}
MLP	max_iter / early_stopping	{150 / True}	–
SVR	C	{10, 100}	100
SVR	epsilon	{0.1, 1.0}	1.0
SVR	kernel	{rbf}	rbf
Regressão Linear	–	sem hiperparâmetros	–

3.4. Métodos Ensemble

Foram criados comitês homogêneos para cada um dos cinco algoritmos individuais, foi construído um `BaggingRegressor` com amostragem com reposição (*bootstrap*), testando três tamanhos de comitê: 5, 10 e 20 estimadores base.

Também foram implementados comitês heterogêneos com os três algoritmos com melhor desempenho individual: k-NN (X), Árvore de Decisão (Y) e MLP (Z). Eles foram combinados em quatro estruturas de comitê heterogêneo via `VotingRegressor`: XY, XZ, YZ e XYZ. A combinação é feita por média das predições de algoritmos estruturalmente distintos, buscando redução de erros correlacionados. Adicionalmente, foi avaliado o efeito de aplicar Bagging ao membro Árvore de Decisão dentro do melhor comitê heterogêneo, com 5, 10 e 20 estimadores, para mensurar o benefício da diversidade estrutural combinada à diversidade amostral.

Um meta-modelo foi treinado sobre as predições dos dois melhores estimadores base, usando `StackingRegressor`. O *stacking* utiliza validação cruzada interna para gerar as predições dos estimadores base no treinamento, evitando o sobreajuste do meta-modelo.

Em todas as configurações *ensemble*, o pré-processamento foi encapsulado no mesmo *Pipeline*, garantindo que imputação, padronização e transformações sejam recalculadas a cada *fold* da validação cruzada.

3.5. Pipeline

A avaliação experimental seguiu um protocolo rigoroso para garantir comparabilidade e ausência de vazamento de dados entre as etapas.

Divisão dos dados: a base foi dividida em conjuntos de treinamento (80%, 7.297 instâncias) e teste (20%, 1.825 instâncias), com `random_state = 42` fixo para reprodutibilidade.

Avaliação hold-out: cada modelo treinado no conjunto completo de treinamento foi avaliado no conjunto de teste, com o *Pipeline* ajustado exclusivamente nos dados de treinamento.

Validação cruzada: além do hold-out, todos os modelos foram avaliados por validação cruzada K-Fold sobre o conjunto de treinamento, com o *Pipeline* completo encapsulado e recalculado em cada *fold*, com o objetivo de verificar a capacidade preditiva das variáveis operacionais.

Métricas de desempenho: foram adotadas cinco métricas complementares para caracterizar diferentes aspectos do erro de predição: Erro Absoluto Médio (MAE), Raiz do Erro Quadrático Médio (RMSE), Coeficiente de Determinação (R^2), Erro Percentual Absoluto Médio (MAPE) e Desvio Padrão do erro de predição.

4. Resultados e Discussões

Nesta seção são apresentados e discutidos os resultados dos modelos, tanto de forma individual quanto em relação aos comitês.

4.1. Desempenho dos algoritmos individuais

Os cinco algoritmos individuais foram avaliados tanto em hold-out quanto em validação cruzada ($k = 5$) sobre o conjunto de treinamento. A Tabela 3 resume os resultados no conjunto de teste.

Tabela 3. Desempenho dos algoritmos individuais no conjunto de teste (*hold-out*).

Modelo	MAE	RMSE	R^2	MAPE	DP erro
k-NN	56,49	148,10	0,9869	14,7%	136,90
Árvore de Decisão	89,75	233,87	0,9674	30,8%	215,97
MLP	223,94	329,41	0,9353	137,5%	241,58
SVR	199,23	390,38	0,9092	35,8%	335,72
Regressão Linear	655,27	868,13	0,5508	319,4%	569,45

O k-NN obteve o melhor desempenho em todas as métricas principais, com $R^2 \approx 0{,}987$ e MAE $\approx 56{,}5$ Sm³. A Árvore de Decisão ficou em segundo lugar, com $R^2 \approx 0{,}967$, desempenho notável considerando sua natureza interpretável e sua capacidade de operar sem padronização. MLP e SVR apresentaram resultados intermediários, com R^2 entre 0,90 e 0,94, enquanto a Regressão Linear obteve $R^2 \approx 0{,}55$, evidenciando que as relações entre as variáveis operacionais e a produção de óleo são predominantemente não lineares.

Os resultados de validação cruzada ($k = 5$) foram consistentes com o hold-out, com desvios padrão baixos entre os *folds*, indicando estabilidade dos modelos. O k-NN obteve CV R^2 médio de $0{,}9855 \pm 0{,}0042$ e a Árvore de Decisão, $0{,}9713 \pm 0{,}0052$.

4.1.1. Importância dos atributos e base reduzida

A análise de *feature importance* da Árvore de Decisão revelou que apenas três variáveis concentram mais de 84% da importância total: temperatura na cabeça do poço (`AVG_WHT_P`, importância $\approx 0{,}370$), pressão na cabeça do poço (`AVG_WHP_P`, $\approx 0{,}362$) e volume de água produzida (`BORE_WAT_VOL`, $\approx 0{,}115$). Esse resultado é consistente com a matriz de correlação sem vazamento, que apontava `AVG_WHP_P` e `AVG_WHT_P` como as variáveis mais correlacionadas com o alvo entre as genuinamente disponíveis.

A avaliação dos modelos sobre a Base 4 (reduzida, apenas os três atributos mais importantes) mostrou uma queda de desempenho de aproximadamente 7 a 9 pontos percentuais de R^2 em relação à base completa (e.g., k-NN: $R^2 = 0{,}9162$ vs. $0{,}9869$; Árvore de Decisão: $0{,}8933$ vs. $0{,}9674$). Isso indica que, embora as três variáveis selecionadas sejam as mais informativas individualmente, os demais atributos agregam poder preditivo relevante e complementar, não devendo ser descartados em aplicações onde o desempenho é prioritário.

4.2. Comitês homogêneos

Os comitês homogêneos via `BaggingRegressor` produziram ganhos expressivos em relação aos algoritmos individuais, especialmente para a Árvore de Decisão, que é um modelo de alta variância. A Tabela 4 apresenta os principais resultados.

Tabela 4. Desempenho dos comitês homogêneos com melhor configuração (hold-out).

Modelo (comitê)	MAE	RMSE	R^2	MAPE	DP erro
k-NN (Bagging, $n = 20$)	57,99	148,57	0,9868	14,6%	136,78
k-NN (Bagging, $n = 10$)	58,45	148,88	0,9868	14,5%	136,93
k-NN (Bagging, $n = 5$)	58,96	153,67	0,9859	14,6%	141,91
Árvore (Bagging, $n = 10$)	66,14	165,67	0,9836	26,5%	151,89
Árvore (Bagging, $n = 20$)	65,68	167,50	0,9833	28,6%	154,08
MLP (Bagging, $n = 5$)	164,93	269,20	0,9568	42,1%	212,75
SVR (Bagging, $n = 5$)	199,91	390,77	0,9090	37,5%	335,76
Regressão Linear (Bagging, $n = 20$)	655,01	868,39	0,5505	318,2%	570,14

A Árvore de Decisão se beneficiou significativamente do Bagging: com $n = 10$ membros, o R^2 passou de $0{,}9674$ (individual) para $0{,}9836$, aproximando-se do k-NN. O k-NN, por sua vez, mostrou ganhos menores com o Bagging, pois já apresentava baixa variância em sua configuração individual. O ganho do Bagging aumenta com o tamanho do comitê, mas tende a estabilizar entre $n = 10$ e $n = 20$. A Regressão Linear não se beneficiou do Bagging, confirmando que seu desempenho inferior se deve ao viés do modelo e não à variância.

4.3. Comitês heterogêneos e stacking

Os comitês heterogêneos via `VotingRegressor` foram construídos com as três melhores configurações individuais: k-NN (X), Árvore de Decisão (Y) e MLP (Z). A Tabela 5 apresenta os resultados.

Tabela 5. Desempenho dos comitês heterogêneos e stacking (hold-out).

Modelo	MAE	RMSE	R^2	MAPE	DP erro
Heterogêneo XY (Voting)	60,23	153,30	0,9860	21,0%	140,97
Heterogêneo XYZ (Voting)	80,61	154,45	0,9858	20,0%	131,75
Heterogêneo XZ (Voting)	95,62	163,55	0,9841	25,4%	132,68
Heterogêneo YZ (Voting)	107,30	197,63	0,9767	24,8%	165,96
Heterogêneo XY + Bagging Árvore ($n = 20$)	52,90	131,49	0,9897	20,5%	120,38
Stacking (meta-modelo: k-NN)	61,71	147,48	0,9870	27,1%	133,95

O comitê heterogêneo XY (k-NN + Árvore de Decisão) obteve o melhor desempenho entre as combinações por votação, com $R^2 \approx 0{,}986$. A adição do MLP ao comitê (XYZ) não proporcionou ganho expressivo em R^2, mas reduziu o desvio padrão do erro de predição, indicando maior estabilidade. A combinação mais eficaz foi o comitê heterogêneo XY com a Árvore de Decisão expandida via Bagging ($n = 20$), que atingiu $R^2 = 0{,}9897$ e MAE $= 52{,}90$ Sm^3, sendo o melhor resultado dentre todas as abordagen e demonstrando que a diversidade estrutural entre algoritmos e a diversidade amostral do Bagging atuam de forma complementar.

O Stacking com k-NN como meta-aprendiz obteve desempenho comparável ao comitê XY por votação ($R^2 \approx 0{,}987$), sem superá-lo de forma expressiva nesta base.

Isso sugere que, quando os estimadores base têm qualidade muito distinta entre si (k-NN consistentemente superior à Árvore e MLP), a média simples por votação já pondera adequadamente as previsões.

A validação cruzada ($k = 5$) dos comitês heterogêneos mostrou desvios padrão entre *folds* menores do que os modelos individuais correspondentes, confirmando a maior estabilidade das predições dos *ensembles*.

4.4. Análise comparativa geral

A Tabela 6 consolida os dez melhores modelos por R² em hold-out, considerando todas as categorias avaliadas.

Tabela 6. Top 10 modelos por R² (hold-out) — consolidado.

Modelo	MAE	RMSE	R²	Categoria
Het. XY + Bagging Árvore ($n = 20$)	52,90	131,49	0,9897	Ensemble heterogêneo
Stacking (meta-modelo: k-NN)	61,71	147,48	0,9870	Stacking
k-NN (individual)	56,49	148,10	0,9869	Individual
k-NN (Bagging, $n = 20$)	57,99	148,57	0,9868	Ensemble homogêneo
k-NN (Bagging, $n = 10$)	58,45	148,88	0,9868	Ensemble homogêneo
Heterogêneo XY (Voting)	60,23	153,30	0,9860	Ensemble heterogêneo
k-NN (Bagging, $n = 5$)	58,96	153,67	0,9859	Ensemble homogêneo
Heterogêneo XYZ (Voting)	80,61	154,45	0,9858	Ensemble heterogêneo
Heterogêneo XZ (Voting)	95,62	163,55	0,9841	Ensemble heterogêneo
Árvore (Bagging, $n = 10$)	66,14	165,67	0,9836	Ensemble homogêneo

4.5. Teste estatístico

Para verificar as diferenças de desemprenho entre os modelos, foram realizados testes estatísticos sobre os resultados dos modelos. A Tabela 7 mostra os resultados do teste de Friedman aplicado para comparar os modelos. A hipótese nula (H_0) é que todos os modelos possuem resultados equivalentes.

Tabela 7. Resultados do teste de Friedman.

Comparação	χ^2	p-valor	Conclusão
5 modelos individuais	39,28	$6{,}10 \times 10^{-8}$	Rejeita H_0 ($p < 0{,}01$)
5 melhores comitês	35,84	$3{,}12 \times 10^{-7}$	Rejeita H_0 ($p < 0{,}01$)
Todos (10 modelos)	87,16	$6{,}03 \times 10^{-15}$	Rejeita H_0 ($p < 0{,}01$)

O teste de Friedman, aplicado aos resultados da validação cruzada K-Fold ($k = 10$), indicou diferenças significativas entre os modelos individuais ($\chi^2 = 39{,}28$; $p = 6{,}10 \times 10^{-8}$), os cinco melhores comitês ($\chi^2 = 35{,}84$; $p = 3{,}12 \times 10^{-7}$) e as dez configurações ($\chi^2 = 87{,}16$; $p = 6{,}03 \times 10^{-15}$). Como pode ser visto na Tabela 8, o teste de Wilcoxon confirmou a superioridade do k-NN sobre a Árvore de Decisão ($W = 0{,}0$; $p = 0{,}0020$) e do comitê Het. XY + Bagging AD ($n = 20$) sobre o k-NN Bagging ($W = 3{,}0$; $p = 0{,}0098$). Entretanto, não houve diferença significativa entre o melhor modelo individual, k-NN (MAE = 52,59 Sm³), e o melhor comitê, Het. XY + Bagging AD (MAE = 50,69 Sm³), ($W = 14{,}0$; $p = 0{,}1934$).

Tabela 8. Resultados do teste de Wilcoxon pareado.

Comparação	W	p-valor	Conclusão
2 melhores individuais	0,0	0,0020	k-NN superior ($p < 0{,}01$)
2 melhores comitês	3,0	0,0098	Het. XY+Bag superior ($p < 0{,}01$)
Melhor indiv. vs. melhor comitê	14,0	0,1934	Sem diferença significativa

5. Conclusão

Este trabalho apresentou um estudo sistemático de predição da produção diária de óleo do campo de Volve (Mar do Norte) a partir de variáveis operacionais e de reservatório, empregando técnicas de ML para regressão supervisionada. O trabalho foi conduzido visando à identificação objetiva de vazamento de dados por correlação, uso de *Pipelines* anti-vazamento de pré-processamento e validação estatística dos resultados.

A remoção do atributo `BORE_GAS_VOL`, identificado como variável de vazamento por correlação $|r| \approx 0{,}999$ com o alvo, é importante, pois, sem ela, os modelos aprenderiam uma relação quase tautológica, inflando artificialmente o desempenho e tornando as predições inúteis em cenários de uso real. Com a remoção, os resultados refletem a capacidade de cada algoritmo em inferir a produção de óleo a partir de condições operacionais disponíveis.

Dentre os algoritmos individuais, o k-NN apresentou o melhor desempenho, com $R^2 \approx 0{,}987$ em hold-out e $R^2 \approx 0{,}986$ em validação cruzada ($k = 5$), seguido pela Árvore de Decisão ($R^2 \approx 0{,}967$). A Regressão Linear, utilizada como modelo paramétrico de referência, obteve $R^2 \approx 0{,}55$, confirmando a natureza predominantemente não linear do problema.

Os métodos *ensemble* demonstraram valor incremental: o Bagging reduziu a variância da Árvore de Decisão, elevando seu R^2 de 0,967 para 0,984; o comitê heterogêneo combinando k-NN e Árvore de Decisão com Bagging ($n = 20$) atingiu o melhor resultado geral ($R^2 = 0{,}9897$, MAE $= 52{,}9$ Sm³), evidenciando que diversidade estrutural e diversidade amostral atuam de forma complementar.

A análise de *feature importance* e a matriz de correlação sem vazamento convergem para a mesma conclusão: pressão e temperatura na cabeça do poço (`AVG_WHP_P` e `AVG_WHT_P`) são as variáveis mais informativas para a predição de produção, seguidas pelo volume de água produzida (`BORE_WAT_VOL`). Esse resultado é fisicamente interpretável: as condições de pressão e temperatura na superfície refletem diretamente a energia do fluido ao longo da coluna de produção, e o avanço da frente de água sinaliza o declínio natural do campo.

É importante ressaltar que a produção diária de petróleo possui a característica temporal e, ao utilizar o K-Fold, essa propriedade não foi explorada. Por isso, para trabalhos futuros, propõe-se a utilização de modelos de séries temporais que aproveitem explicitamente a estrutura temporal e de painel dos dados. A incorporação de variáveis de decline natural do reservatório e de indicadores de manutenção pode também ampliar a capacidade preditiva do modelo em horizontes mais longos.

6. Sobre o uso de ferramentas de IA

Em conformidade com as diretrizes de integridade acadêmica e transparência científica, declara-se que as ferramentas de inteligência artificial ChatGPT (OpenAI) e Claude AI (Anthropic) foram utilizadas no desenvolvimento deste trabalho. O uso de tais ferramentas se deu como suporte na revisão gramatical do texto, no refinamento estético e na otimização do *layout* LaTeX e fluxogramas apresentados. Ressalta-se que a concepção metodológica, a parametrização dos cenários experimentais, a execução dos treinamentos e a análise crítica dos resultados foram realizadas integralmente pelos autores, mantendo a originalidade e a responsabilidade intelectual do conteúdo científico.

Agradecimentos
Este trabalho foi realizado com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), processo nº 132492/2026-1.

Referências

Calvette, T. (2019). *Aplicação de técnicas de machine learning e data driven em poços inteligentes de petróleo*. PhD thesis, PUC-Rio.

Daramola, G. O., Jacks, B. S., Ajala, O. A., and Akinoso, A. (2024). Ai applications in reservoir management: Optimizing production and recovery in oil and gas fields. *Computer Science IT Research Journal*.

Dike, H., Adams, D. E., Mosobalaje, O., Oluwasanmi, O., Olaniyan, D., and Bolu, I. S. (2025). Oil-well production forecasting: A machine learning approach. In *SPE Nigeria Annual International Conference and Exhibition*.

Ghodsi, M., Vaziri, P., Kanaani, M., and Sedaee, B. (2025). Optimizing oil production forecasts in iranian oil fields: a comprehensive analysis using ensemble learning techniques. *Journal of Petroleum Exploration and Production Technology*, 15(4):63.

Pop, V. (2018). Petroleum – strategic resource for world economy. *Studia Universitatis „Vasile Goldis" Arad – Economics Series*, 28:70 – 85.

Salimovna, S. (2024). The importance of the world oil and gas industry in the economy of countries. *American Journal of Economics and Business Management*.

Samad, A., Khan, I. M., Rahaman, M. S., Sakib, A., and Islam, M. A. (2025). Data-driven approach to predict future oil production of an oil field using machine learning techniques. *SCS Eng*, 3:68–73.