

UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS/FACULDADE DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM
COMPUTACIONAL

Jonathan do Amaral Braz

Estimativa do índice de compressão de solos argilosos moles: Uma abordagem de
aprendizado de máquina automatizado

Juiz de Fora

2025

Jonathan do Amaral Braz

**Estimativa do índice de compressão de solos argilosos moles: Uma abordagem de
aprendizado de máquina automatizado**

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Modelagem Computacional.

Orientador: Prof. Dr. Leonardo Goliatt da Fonseca

Juiz de Fora

2025

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Braz, Jonathan do Amaral.

Estimativa do índice de compressão de solos argilosos moles : Uma abordagem de aprendizado de máquina automatizado / Jonathan do Amaral Braz. – 2025.

67 f. : il.

Orientador: Leonardo Goliatt da Fonseca

Dissertação (Mestrado) – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas/Faculdade de Engenharia. Programa de Pós-Graduação em Modelagem Computacional, 2025.

1. Aprendizado de máquina automatizado. 2. Regressão. 3. Índice de compressão. I. Fonseca, Leonardo Goliatt da, orient. II. Título.

Jonathan do Amaral Braz

Estimativa do índice de compressão de solos argilosos moles: Uma abordagem de aprendizado de máquina automatizado

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Modelagem Computacional. Área de concentração: Modelagem Computacional.

Aprovada em 25 de abril de 2025.

BANCA EXAMINADORA

Prof. Dr. Leonardo Goliatt da Fonseca - Orientador

Universidade Federal de Juiz de Fora

Prof. Dr. Heder Soares Bernardino

Universidade Federal de Juiz de Fora

Prof.^a Dr.^a Katia Vanessa Bicalho

Universidade Federal do Espírito Santo

Juiz de Fora, 06/06/2025.



Documento assinado eletronicamente por **KATIA VANESSA BICALHO, Usuário Externo**, em 06/06/2025, às 22:40, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Heder Soares Bernardino, Professor(a)**, em 07/06/2025, às 00:44, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Leonardo Goliatt da Fonseca, Professor(a)**, em 07/06/2025, às 08:05, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no Portal do SEI-Uffj (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **2442935** e o código CRC **65BCC7F0**.

À memória de Minerva.

AGRADECIMENTOS

Agradeço à minha mãe, Presciliana, pelo papel essencial na minha trajetória pessoal e profissional até aqui.

Ao meu namorado, Gabriel, pela parceria e pelo companheirismo que eu desejei e de que eu precisei neste processo. Para além deste trabalho, agradeço pela oportunidade de experienciar formas profundas de amor e pego emprestadas as palavras da poeta Matilde Campilho para dizer que “foi, e ainda é, uma aventura tremenda”.

Ao restante da minha família, pelo amor, apoio e compreensão de ausências: meu pai, Celso; meus irmãos, Celso Júnior e Priscila; meus sobrinhos, Isadora e Álvaro; meu padrasto, João; meus cunhados, Sandra e Felipe; minha ex-babá e segunda mãe, Aparecida; e minha madrinha, Ilza.

Aos meus amigos, também pelo amor e especialmente pelos momentos de alívio cômico e dramático. Em ordem alfabética: Camila Benevenuto, Isadora Tavares, João Chiavegatto, Júlia Gaione, Laís Cerqueira, Laise Lutz, Lívia Mazzeu, Maria Cláudia Brasil, Mariana Dias, Matheus Hotz, Rafael Paschoalino, Raquel Costa, Sarah Kiffer, Vinícius Gomes e Wylker Simões.

À minha terapeuta, Hortênsia Vieira. Pelas trocas nas experiências simultâneas de pós-graduação e por ter sido uma bússola, entre outras questões da alma, no processo de me tornar mestre.

À Professora Tatiana Rodriguez. Querida amiga, orientadora da graduação e principal mentora acadêmica. Por quase uma década de parceria, que é parte essencial da minha construção como pesquisador e futuro professor.

Ao meu orientador, Professor Leonardo Goliatt, pelo suporte, paciência e constante incentivo. Aos membros da banca, Professora Katia Bicalho e Professor Heder Bernardino, pela disposição em contribuir com este trabalho.

Ao Programa de Pós-Graduação em Modelagem Computacional (PPGMC), por ampliar meu panorama de ferramentas de pesquisa e pelo espaço de trocas com alunos e professores de áreas diferentes. Agradeço também aos amigos feitos durante o mestrado, em ordem alfabética: Alexsandro Pattiele, Ana Clara Verly, Eduardo Santos, Lucas Teixeira e Thiago Esterci.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo financiamento crucial ao desenvolvimento deste trabalho e das conseqüentes publicações. Agradeço também à Universidade Federal de Juiz de Fora (UFJF), pela contribuição na minha formação profissional e pessoal.

Por fim, agradeço à minha gata de estimação, Minerva, pela companhia fiel de uma

década que, apesar de ter se encerrado na primeira semana de mestrado, continuou viva na memória pelos dois anos que se seguiram até aqui.

“Ouça, Virgínia, é preciso amar o inútil. Criar pombos sem pensar em comê-los, plantar roseiras sem pensar em colher as rosas, escrever sem pensar em publicar, fazer coisas assim, sem esperar nada em troca. A distância mais curta entre dois pontos pode ser a linha reta, mas é nos caminhos curvos que se encontram as melhores coisas.” (Lygia Fagundes Telles, *Ciranda de Pedra*, p. 135)

RESUMO

O recalque é um fenômeno de grande interesse na engenharia geotécnica, relacionado à deformação vertical de solos e importante na prevenção de problemas estruturais em construções. Para a sua quantificação por meio de cálculo, é necessária uma propriedade chamada índice de compressão, que pode ser obtida por ensaios de laboratório. Entretanto, esses procedimentos podem ser custosos em termos de tempo, equipamento e mão de obra especializada. Por outro lado, a literatura tem apresentado, nas últimas décadas, equações empíricas e modelos de aprendizado de máquina para a estimativa de propriedades de solos e resultados de ensaios geotécnicos. Além disso, nos últimos anos, o aprendizado de máquina automatizado (AutoML) tem surgido como uma alternativa para treinar e combinar diversos modelos com procedimentos mais simples. Dessa forma, o objetivo deste trabalho é desenvolver modelos de AutoML para estimar índices de compressão de solos e comparar suas performances. Para isso, foram selecionados quatro bancos de dados com amostras de solos de lugares diferentes. Esses conjuntos passaram por pré-processamento de dados e, então, foram aplicados a três modelos: AutoGluon, FLAML e H2O. As métricas de avaliação usadas foram R^2 , $RMSE$ e o tempo de execução. Ao final, constatou-se que os três modelos apresentaram valores próximos de R^2 e $RMSE$, tendo o FLAML apresentado os valores mais satisfatórios na maioria dos casos. O H2O apresentou resultados próximos, mas com tempo de execução consideravelmente menor. Concluiu-se que, apesar de terem estratégias diferentes, os três modelos apresentam desempenho semelhante diante das mesmas condições.

Palavras-chave: Aprendizado de máquina automatizado; Regressão; Índice de compressão.

ABSTRACT

Settlement is a phenomenon of great interest in geotechnical engineering, related to the vertical deformation of soils and important in preventing structural issues in constructions. To perform the calculation, a property called compression index is required, which can be obtained through laboratory tests. However, these procedures can be costly in terms of time, equipment and specialized labor. On the other hand, in recent decades, literature has presented empirical equations and machine learning models to estimate soil properties and results from geotechnical tests. Furthermore, in recent years, automated machine learning (AutoML) has emerged as an alternative to train and stack several models with simpler procedures. Therefore, this study aims to develop AutoML models for the prediction of soil compression indexes and to compare their performances. In order to do so, four databases with soil samples from different places were selected. These sets underwent data preprocessing and were then applied to three models: AutoGluon, FLAML and H2O. The evaluation metrics used were R^2 , $RMSE$ and execution time. In the end, it was found that the three models presented similar R^2 and $RMSE$ values, with FLAML showing the most satisfactory values in most cases. H2O produced similar results but with a considerably shorter execution time. It was concluded that, despite having different strategies, the three models perform similarly under the same conditions.

Keywords: Automated machine learning; Regression; Compression index.

LISTA DE ILUSTRAÇÕES

Figura 1 – Esquema de componentes de uma célula edométrica	19
Figura 2 – Prensa de adensamento	20
Figura 3 – Gráfico de deformação por tempo em um ensaio de adensamento	21
Figura 4 – Gráfico de índice de vazios em relação à tensão axial de ensaios de adensamento	22
Figura 5 – Efeito de descarregamento seguido de recarregamento de uma amostra de solo	23
Figura 6 – Esquema da estratégia de <i>stacking</i> multicamada do AutoGluon	29
Figura 7 – Esquema da estratégia de busca do FLAML	32
Figura 8 – Matriz de correlação entre as variáveis de Benbouras et al. (2019)	44
Figura 9 – Matriz de correlação entre as variáveis de Kalantary e Kordnaeij (2012)	45
Figura 10 – Matriz de correlação entre as variáveis de Pham et al. (2024)	45
Figura 11 – Matriz de correlação entre as variáveis de Uzer (2024)	46
Figura 12 – Boxplots para os resultados de R^2 de Benbouras et al. (2019)	47
Figura 13 – Boxplots para os resultados de $RMSE$ de Benbouras et al. (2019)	48
Figura 14 – Boxplots para os resultados de tempo de Benbouras et al. (2019)	49
Figura 15 – Boxplots para os resultados de R^2 de Kalantary e Kordnaeij (2012)	51
Figura 16 – Boxplots para os resultados de $RMSE$ de Kalantary e Kordnaeij (2012)	52
Figura 17 – Boxplots para os resultados de tempo de Kalantary e Kordnaeij (2012)	53
Figura 18 – Boxplots para os resultados de R^2 de Pham et al. (2024)	54
Figura 19 – Boxplots para os resultados de $RMSE$ de Pham et al. (2024)	55
Figura 20 – Boxplots para os resultados de tempo de Pham et al. (2024)	56
Figura 21 – Boxplots para os resultados de R^2 de Uzer (2024)	58
Figura 22 – Boxplots para os resultados de $RMSE$ de Uzer (2024)	59
Figura 23 – Boxplots para os resultados de tempo de Uzer (2024)	60

LISTA DE TABELAS

Tabela 1 – Resultado da revisão de literatura	15
Tabela 2 – Equações empíricas para estimativa do índice de compressão	25
Tabela 3 – Algoritmos de aprendizado de máquina usados pelo AutoGluon	28
Tabela 4 – Algoritmos de aprendizado de máquina usados pelo FLAML	31
Tabela 5 – Algoritmos de aprendizado de máquina usados pelo H2O	34
Tabela 6 – Algoritmos passados como hiperparâmetros nos modelos de AutoML	39
Tabela 7 – Estatísticas dos dados de Benbouras et al. (2019)	42
Tabela 8 – Estatísticas dos dados de Kalantary e Kordnaeij (2012)	42
Tabela 9 – Estatísticas dos dados de Pham et al. (2024)	43
Tabela 10 – Estatísticas dos dados de Uzer (2024)	43
Tabela 11 – Amostras retiradas de Pham et al. (2024)	43
Tabela 12 – Amostras retiradas de Uzer (2024)	43
Tabela 13 – Métricas de desempenho para o banco de dados de Benbouras et al. (2019)	47
Tabela 14 – Estatísticas de R^2 do banco de dados de Benbouras et al. (2019)	48
Tabela 15 – Estatísticas de $RMSE$ do banco de dados de Benbouras et al. (2019)	49
Tabela 16 – Estatísticas de tempo do banco de dados de Benbouras et al. (2019)	50
Tabela 17 – Métricas de desempenho para o banco de dados de Kalantary e Kordnaeij (2012)	50
Tabela 18 – Estatísticas de R^2 do banco de dados de Kalantary e Kordnaeij (2012)	51
Tabela 19 – Estatísticas de $RMSE$ do banco de dados de Kalantary e Kordnaeij (2012)	52
Tabela 20 – Estatísticas de tempo do banco de dados de Kalantary e Kordnaeij (2012)	53
Tabela 21 – Métricas de desempenho para o banco de dados de Pham et al. (2024)	54
Tabela 22 – Estatísticas de R^2 do banco de dados de Pham et al. (2024)	55
Tabela 23 – Estatísticas de $RMSE$ do banco de dados de Pham et al. (2024)	56
Tabela 24 – Estatísticas de tempo do banco de dados de Pham et al. (2024)	57
Tabela 25 – Métricas de desempenho para o banco de dados de Uzer (2024)	57
Tabela 26 – Estatísticas de R^2 do banco de dados de Uzer (2024)	58
Tabela 27 – Estatísticas de $RMSE$ do banco de dados de Uzer (2024)	59
Tabela 28 – Estatísticas de tempo do banco de dados de Uzer (2024)	60
Tabela 29 – Tabela resumo dos resultados	61

SUMÁRIO

1	Introdução	13
1.1	Contextualização	13
1.2	Objetivo	14
1.3	Contribuição	14
1.4	Estrutura do trabalho	15
2	Compressibilidade dos solos	17
2.1	Recalques por carregamento dos solos	17
2.2	Ensaio de compressão edométrica e cálculo de recalque	18
2.3	Estimativa do índice de compressão por modelos matemáticos	24
3	Aprendizado de máquina automatizado	27
3.1	AutoGluon	27
3.2	FLAML	30
3.3	H2O	33
4	Materiais e métodos	36
4.1	Coleta de dados	36
4.2	Pré-processamento e análise dos dados	37
4.3	Especificações de implementação	38
4.4	Métricas de desempenho	40
5	Resultados	42
5.1	Pré-processamento e análise dos dados	42
5.2	Métricas de desempenho	46
6	Conclusões e trabalhos futuros	62
6.1	Conclusões	62
6.2	Trabalhos futuros	63
	REFERÊNCIAS	64

1 Introdução

1.1 Contextualização

O recalque de uma camada de solo é um fenômeno de grande interesse para a engenharia geotécnica, frequentemente relacionado à manutenção da estabilidade de estruturas e à prevenção de patologias da construção (Caputo, 2006). O domínio desse fenômeno por meio de cálculos exige o entendimento da relação entre a carga recebida pelo solo e a deformação vertical que ele sofre. Para quantificar essa relação, há um indicador chamado índice de compressão, que está presente nas equações propostas para estimativa de recalque e, portanto, é considerado uma informação essencial nesse tipo de investigação (Das, 2011).

O índice de compressão é calculado com base em resultados de procedimentos de laboratório, como o ensaio de compressão triaxial e o ensaio de compressão edométrica. Entretanto, esses procedimentos são custosos, tanto em termos de tempo, quanto em termos de equipamento e mão de obra especializada – o ensaio de compressão edométrica, por exemplo, tem duração de 24 horas para cada um dos carregamentos que devem ser feitos e requer o uso de uma prensa de adensamento específica (Das, 2011). Dadas essas questões, é notável na literatura a produção de equações empíricas que se propõem a estimar o índice de compressão para grupos de solos específicos (Kalantary e Kordnaeij, 2012).

Além disso, na última década, também tem sido notável a produção acadêmica que faz uso de aprendizado de máquina para estimar propriedades físicas e hidráulicas dos solos. Destacam-se aqui trabalhos relacionados à estimativa de índices de consistência com base em outros índices de consistência e trabalhos relacionados à estimativa da permeabilidade de solos com base na granulometria do material. Além disso, a literatura também apresenta uma série de modelos que se propõem a estimar o índice de compressão de solos com base em suas propriedades físicas. Tal informação, somada à produção de equações empíricas, permite inferir a relevância deste objeto de pesquisa (Bicalho et al., 2022).

Apesar da popularidade recente do aprendizado de máquina, o desenvolvimento de trabalhos relacionados a esse tema apresenta alguns desafios, como a implementação sucessiva de diversos modelos e a busca pela melhor configuração de hiperparâmetros para cada um deles. Nesse sentido, o aprendizado de máquina automatizado, ou AutoML, surge como uma alternativa, tanto para os usuários iniciantes, quanto para os mais experientes, no sentido de oferecer implementações simples para processos complexos que ocorrem de forma automatizada (Erickson et al., 2020).

1.2 Objetivo

O objetivo deste trabalho é estimar o índice de compressão de solos argilosos moles por meio de algoritmos de aprendizado de máquina automatizado. Para isso, tem-se como objetivos secundários: a seleção de bancos de dados, a partir de trabalhos da literatura, contendo características dos solos e seus respectivos índices de compressão; a definição das variáveis de entrada para o desenvolvimento dos modelos; o treinamento e a avaliação de algoritmos de aprendizado de máquina automatizado, seguido da comparação entre seus desempenhos.

1.3 Contribuição

Para o desenvolvimento deste trabalho, foi feita uma revisão da literatura acerca da aplicação de aprendizado de máquina na estimativa de índice de compressão de solos, a fim de se identificar contribuições que pudessem ser adicionadas a este campo de pesquisa. Executou-se uma busca nas bases de dados Scopus e Web of Science com a seguinte configuração de palavras-chave: (“Machine learning” OR “Automated machine learning” OR “AutoML”) AND (“Compression index”) AND (“Soil”). Os resultados encontrados nas bases Scopus e Web of Science foram submetidos a análise de co-ocorrência de palavras-chave no *software* VOSViewer. A Tabela 1 apresenta a frequência de ocorrência de palavras-chave relacionadas a algoritmos de inteligência computacional apresentadas pelo VOSViewer. Percebe-se, pelos resultados apresentados, que nas bases de dados consultadas não há aplicações de aprendizado de máquina automatizado para esse tipo de problema.

Tabela 1 – Resultado da revisão de literatura

Algoritmo	Frequência
Neural Networks	9
Random Forest	6
Adaptive Boosting	3
Support Vector Machine	3
Decision Tree	2
Gradient Boosting	2
Extreme Gradient Boosting	2
Particle Swarm Optimization	2
Adaptive Neuro-Fuzzy Inference	1
Ant Colony Optimization	1
Bayesian Optimization	1
Extreme Learning Machine	1
Fuzzy Inference	1
Fuzzy Neural Networks	1
Fuzzy Systems	1
Gene Expression Programming	1
Genetic Programming	1
Light Gradient Boosting Machine	1
Manta Ray Foraging Optimization	1

Fonte: Autor (2025)

Foi executada também uma busca na base de dados ScienceDirect, cuja forma de exportação dos dados não possui compatibilidade com o VOSViewer. A configuração usada na busca foi: (“Automated machine learning”OR “AutoML”) AND (“Compression index”) AND (“Soil”). A pesquisa não apresentou resultados. É possível afirmar, então, que este trabalho preenche uma lacuna, tanto para os estudos envolvendo AutoML, quanto para os estudos de estimativa do índice de compressão de solos.

Além disso, há contribuição para as Metas Nacionais dos Objetivos de Desenvolvimento Sustentável (ODS), que são adequações consolidadas pelo Instituto de Pesquisa Econômica Aplicada às propostas da Assembleia Geral das Nações Unidas (IPEA, 2018). A categoria correspondente à contribuição deste trabalho é a ODS 9, “Construir infraestruturas resilientes, promover a industrialização inclusiva e sustentável e fomentar a inovação”. Tal classificação é feita considerando que este é um trabalho científico em que há o desenvolvimento de tecnologias relacionadas à construção civil por meio de pesquisa científica.

1.4 Estrutura do trabalho

Este trabalho é dividido em seis capítulos, sendo o Capítulo 1 esta introdução.

No Capítulo 2 são introduzidos conceitos básicos da mecânica dos solos relacionados

à compressibilidade. Discute-se as características básicas do recalque e são apresentados os ensaios geotécnicos e as equações necessárias para a quantificação desse fenômeno. É apresentado também o índice de compressão, elemento central neste trabalho, e equações empíricas para estimativa desse parâmetro.

No Capítulo 3 são apresentados conceitos gerais sobre aprendizado de máquina automatizado. Em seguida, são discutidos em mais detalhes os três modelos usados neste trabalho, destacando suas principais características de tratamento de dados, treinamentos de modelos e estratégias de busca para obter os melhores resultados.

O Capítulo 4 contém a metodologia para a aplicação dos modelos de aprendizado de máquina automatizado, incluindo a seleção de diversos bancos de dados e os devidos tratamentos. Especifica-se também as decisões tomadas para a implementação dos modelos, a fim de garantir a reprodutibilidade do trabalho e a padronização desses modelos, para comparação e também para maior eficiência. Ao final do capítulo, são apresentadas as métricas de desempenho usadas para avaliar os desempenhos.

No Capítulo 5 são apresentados os resultados. A partir da análise de correlação entre as variáveis de cada banco de dados, são selecionadas as variáveis de entrada de forma padronizada para os modelos. Em seguida, são apresentadas as métricas de desempenho obtidas de cada modelo e os resultados são comparados.

No Capítulo 6 são feitas conclusões sobre o trabalho desenvolvido, bem como uma análise de possibilidades de continuação da pesquisa, na forma de sugestão para trabalhos futuros.

2 Compressibilidade dos solos

Neste capítulo são apresentados conceitos básicos de mecânica dos solos fundamentais para a compreensão do objeto de pesquisa. No Item 2.1 é explicado o conceito de recalque e as diferentes formas de ocorrência deste fenômeno para tipos diferentes de solos. No Item 2.2 são apresentados procedimentos de laboratório para a obtenção do índice de compressão e métodos de cálculo para a quantificação de recalques. Por fim, no Item 2.3 são apresentadas formas de estimativa do índice de compressão por meio de modelos matemáticos encontrados na literatura.

2.1 Recalques por carregamento dos solos

Um fenômeno de interesse da engenharia geotécnica é a deformação vertical dos solos, que acontece quando esses materiais estão sujeitos a tensões causadas pela construção de fundações rasas ou pela execução de aterros. Essas deformações podem ocorrer devido à deformação das partículas de solo, ao deslocamento dessas partículas ou à expulsão de água dos espaços vazios que existem na estrutura do material (Das, 2011). Tal comportamento é chamado de recalque e sua importância está relacionada à estabilidade de edificações: conhecido o comportamento de um solo sob determinadas cargas, é possível evitar que haja o deslocamento vertical de uma construção de forma desigual, evitando assim o surgimento de patologias e problemas estruturais na edificação (Caputo, 2006).

As deformações podem ser organizadas em três tipos. O primeiro, denominado recalque elástico ou imediato, ocorre diretamente após a aplicação de cargas e é causado pela deformação elástica dos grãos, sem alterar a quantidade de água na estrutura do material. Por outro lado, o segundo e o terceiro podem se desenvolver em intervalos maiores de tempo após o carregamento. O segundo é chamado recalque por adensamento primário e é provocado pela expulsão de água dos poros da estrutura. O terceiro, denominado recalque por compressão secundária, é causado pelo rearranjo das partículas de solo (Das, 2011).

O comportamento do material nessas situações depende da sua constituição e da presença de água em sua estrutura. Solos com predominância de areia, que são grãos com diâmetro maior, tendem a se deformar mais rápido, tanto na condição não saturada, quanto na condição saturada – nesta, por apresentarem maior facilidade no escoamento de água em seus espaços vazios. Pode-se dizer, então, que o recalque elástico e o recalque por adensamento ocorrem simultaneamente nesse tipo de material. Solos com predominância de argila, que são grãos mais finos, se deformam mais rápido quando não há grandes quantidades de água. Por outro lado, estes tomam um tempo maior para completar esse processo quando estão saturados, por terem capacidade de drenagem menor. Conseqüentemente, em muitos casos, o recalque causado por adensamento é maior que o

recalque imediato (Pinto, 2006).

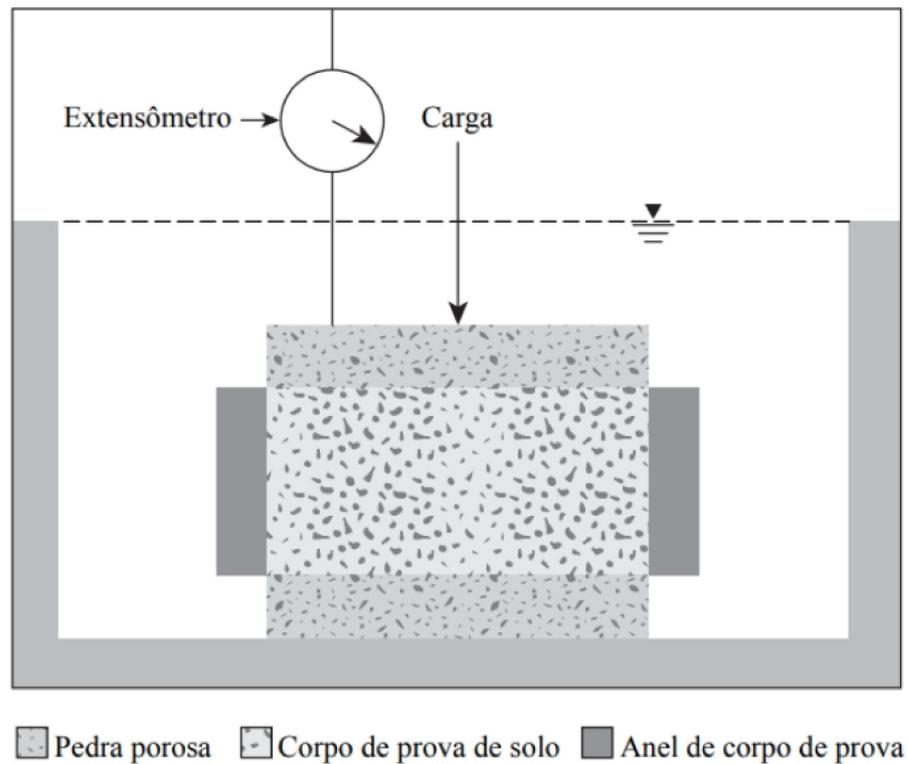
Esse comportamento pode ser expresso por parâmetros obtidos através de procedimentos de laboratório, como o ensaio de compressão axial e o ensaio de compressão edométrica. A partir desses parâmetros, é possível determinar o recalque de um solo, sem usar diretamente os resultados do ensaio (Pinto, 2006). Na Seção 2.2 é elaborado em mais detalhes o ensaio de compressão edométrica, por ser o procedimento que gerou os dados usados neste trabalho. Serão apresentados também os tipos de resultados que são gerados pelo ensaio, assim como os parâmetros que podem ser obtidos a partir de tais resultados e a metodologia para o cálculo de recalque com base nesses parâmetros.

2.2 Ensaio de compressão edométrica e cálculo de recalque

O ensaio de compressão edométrica, também conhecido como ensaio de adensamento, consiste na compressão vertical de uma amostra de solo confinada em uma célula, de forma a impedir qualquer deformação horizontal. Esse ensaio tem o objetivo de simular o comportamento do material em situações de campo, quando uma camada de solo sofre compressão por outras camadas aterradas acima. O ensaio também é considerado representativo para situações em que há carregamento sobre áreas menores, como a construção de fundações rasas (Pinto, 2006).

Uma das principais diretrizes para a execução de ensaios de compressão edométrica é a norma D2435 da ASTM (2020). Na etapa de preparação desse procedimento, uma amostra indeformada de solo é posicionada em um anel rígido, ajustado a uma célula de compressão edométrica. Acima e abaixo dessa amostra, são posicionadas pedras porosas que permitem a saída de água do solo. O sistema descrito está apresentado na Figura 1.

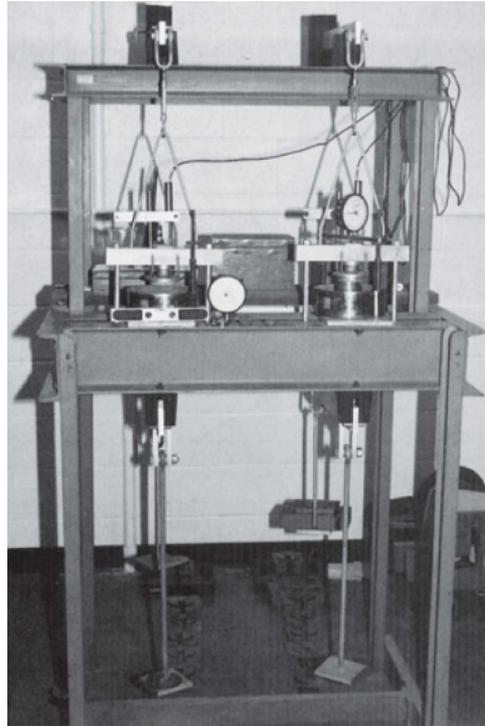
Figura 1 – Esquema de componentes de uma célula edométrica



Fonte: Das (2011)

A célula edométrica é colocada em uma prensa, como a que está ilustrada na Figura 2, para que seja aplicada uma carga axial sobre a amostra por um braço de alavanca. A deformação do solo é medida por um extensômetro em determinados intervalos de tempo, até o fim do ensaio para o carregamento específico. O ensaio pode ser encerrado quando não houver mais variações de deformação ou quando houver passado um período de 24 horas. Em seguida, dobra-se o valor da carga e o procedimento de ensaio é repetido (Das, 2011; Pinto, 2006).

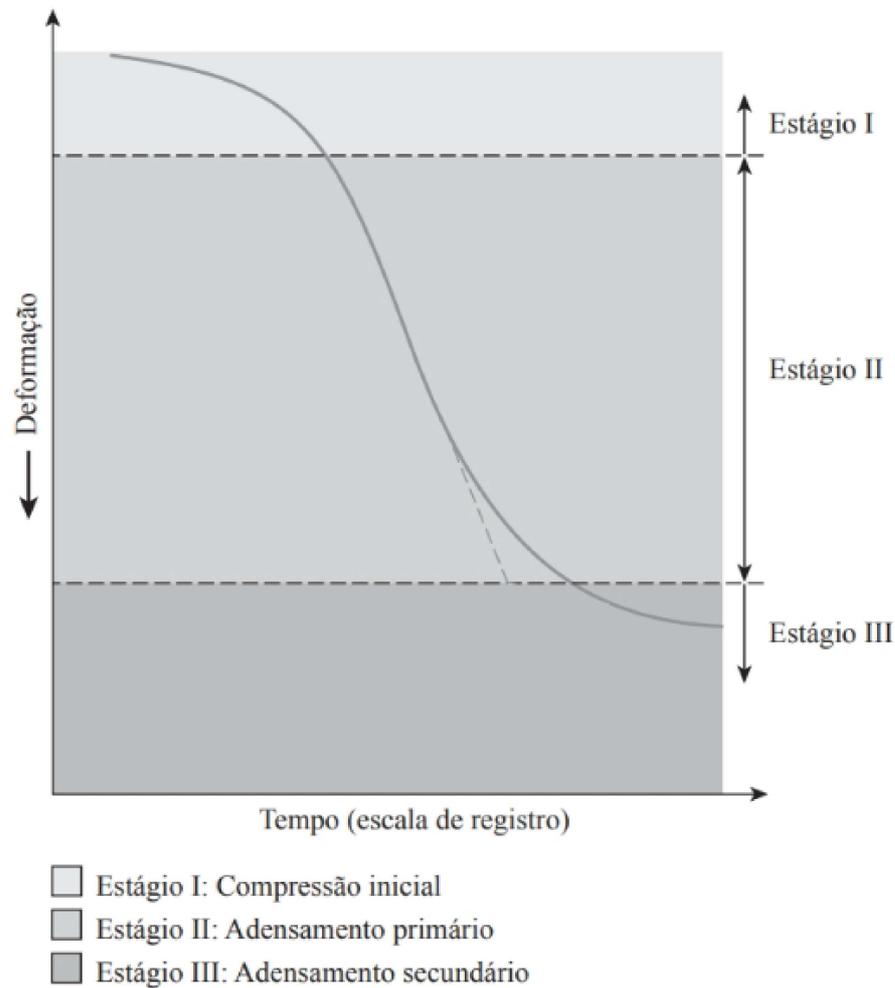
Figura 2 – Prensa de adensamento



Das (2011)

Os resultados iniciais do ensaio se apresentam no formato de variação da deformação ao longo de um intervalo de tempo para uma determinada carga, como está apresentado na Figura 3. Neste exemplo, são destacadas as três etapas citadas na Seção 2.1: o Estágio I de compressão inicial, causado imediatamente pelo carregamento; o Estágio II de adensamento primário, em que há a expulsão de água dos poros da amostra; e o Estágio III de compressão secundária, em que há um reajuste na estrutura do solo (Das, 2011).

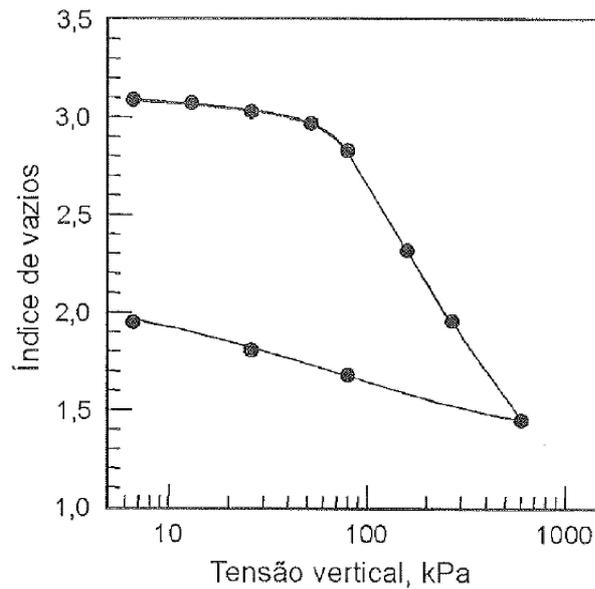
Figura 3 – Gráfico de deformação por tempo em um ensaio de adensamento



Fonte: Das (2011)

Feita a aplicação de diversas cargas, o resultado final relaciona a variação dessas cargas com a variação da altura da amostra, como ilustrado na Figura 4. Nesse caso, representa-se a variação de altura em função da variação do índice de vazios – matematicamente definido como a razão entre o volume de vazios e o volume de partículas de solo de uma amostra. Quando o eixo das abscissas é apresentado em escala logarítmica, é possível perceber que, a partir de um valor de pressão aplicada, o índice de vazios varia de forma linear às tensões. Esse trecho da curva é denominado reta virgem (Pinto, 2006).

Figura 4 – Gráfico de índice de vazios em relação à tensão axial de ensaios de adensamento



Fonte: Pinto (2006)

A partir dos resultados apresentados, é possível extrair um parâmetro chamado índice de compressão, descrito pela Equação 2.1, que consiste na inclinação da reta virgem. Fisicamente, é possível afirmar que resultados maiores de índice de compressão indicam solos mais compressíveis, enquanto valores menores apontam para solos menos compressíveis (Pinto, 2006).

$$C_c = \frac{e_1 - e_2}{(\log \bar{\sigma}_2 - \log \bar{\sigma}_1)} \quad (2.1)$$

em que:

C_c : índice de compressão do solo;

e_1 : índice de vazios inicial da amostra;

e_2 : índice de vazios final da amostra;

$\bar{\sigma}_2$: tensão final do trecho;

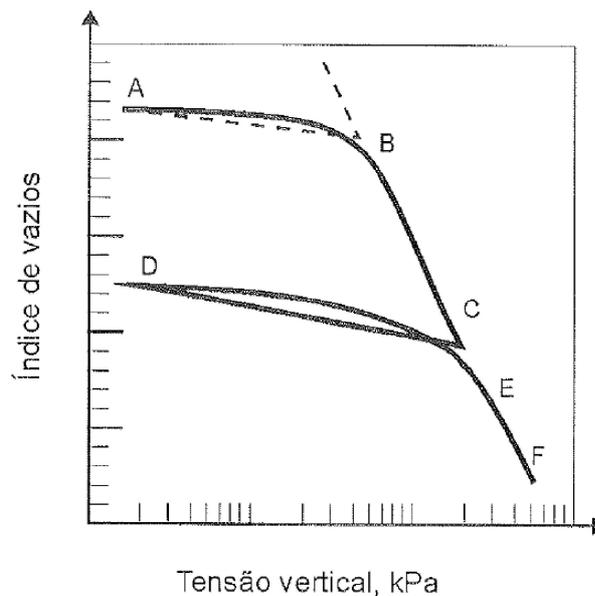
$\bar{\sigma}_1$: tensão inicial do trecho.

A tensão a partir da qual se inicia a reta virgem está relacionada à história geológica dos solos. Qualquer camada de solo pode ter sido submetida à mesma tensão do momento em que é coletada uma amostra para ensaio de compressão edométrica. Porém, é possível que a mesma camada possa ter sido submetida a tensões maiores, que podem ter sido removidas por processos geológicos naturais ou por ação antrópica. Quando uma amostra de solo fino é coletada, ela passa por um processo de descompressão, ou seja, há um

aumento do seu índice de vazios. Durante a progressão de cargas no ensaio, a compressão do solo acontece com pouca variação até que seja atingida a tensão máxima a que esse solo estava submetido no momento da coleta. A partir desse valor de tensão, a relação entre o índice de vazios e a tensão aplicada é linear (Das, 2011).

É possível verificar em laboratório se a tensão máxima sofrida por aquele solo foi igual ou maior à tensão atuante no momento da amostragem. Quando a diminuição da tensão resulta em um processo de aumento do seu índice de vazios, como ilustrado na Figura 5, é possível concluir que a sobrecarga máxima aplicada no ensaio não é maior que a tensão máxima a que o solo foi submetido no passado – esses são denominados argilas normalmente adensadas. Por outro lado, quando a sobrecarga máxima aplicada no ensaio é maior que a tensão máxima submetida ao solo no passado, dá-se ao material o nome de argilas sobreadensadas. A tensão máxima a que o solo já foi submetido em sua história geológica é chamada de tensão de pré-adensamento (Das, 2011).

Figura 5 – Efeito de descarregamento seguido de recarregamento de uma amostra de solo



Fonte: Pinto (2006)

O conhecimento de argilas normalmente adensadas ou sobreadensadas é importante na etapa do cálculo de recalque, uma vez que cada uma delas requer uma equação diferente. Para argilas normalmente adensadas, o recalque pode ser calculado a partir da Equação 2.2 (Pinto, 2006).

$$\rho = \frac{C_c \cdot H_1}{1 + e_1} \cdot \log \frac{\bar{\sigma}_2}{\bar{\sigma}_1} \quad (2.2)$$

em que:

ρ : recalque da camada de solo;

C_c : índice de compressão;

H_1 : altura da camada de solo;

e_1 : índice de vazios inicial;

$\bar{\sigma}_2$: tensão final da camada de solo após o carregamento;

$\bar{\sigma}_1$: tensão inicial da camada de solo.

No caso do recalque em argilas sobreadensadas, é importante considerar o trecho anterior à tensão de pré-adensamento, em que a inclinação da curva é representada pelo índice de recompressão. O valor desse índice também pode ser obtido por resultados de ensaios e varia entre 10 e 20% do valor do índice de compressão. O cálculo do recalque é, então, descrito pela Equação 2.3 (Pinto, 2006).

$$\rho = \frac{H}{1 + e_1} \cdot \left(C_r \cdot \log \frac{\bar{\sigma}_a}{\bar{\sigma}_i} + C_c \cdot \log \frac{\bar{\sigma}_f}{\bar{\sigma}_a} \right) \quad (2.3)$$

em que:

ρ : recalque da camada de solo;

H_1 : altura da camada de solo;

e_1 : índice de vazios inicial;

C_c : índice de compressão;

C_r : índice de recompressão;

$\bar{\sigma}_a$: tensão de pré-adensamento do solo;

$\bar{\sigma}_i$: tensão inicial;

$\bar{\sigma}_f$: tensão final.

2.3 Estimativa do índice de compressão por modelos matemáticos

Com base nas informações apresentadas na Seção 2.2, pode-se dizer que o ensaio de compressão edométrica é custoso, tanto em termos de tempo, quanto em termos de equipamento. A literatura apresenta há décadas equações empíricas para estimar o índice de compressão de solos, facilitando assim a estimativa de recalques. A Tabela 2 apresenta alguns modelos. É importante destacar que, de maneira geral, as equações são desenvolvidas para grupos específicos de solos. Como exemplo, cita-se o modelo desenvolvido por Hough (1957) para solos coesivos inorgânicos, como siltes, argilas siltosas

e argilas. Além disso, há a possibilidade de modelos serem desenvolvidos com base no material encontrado em regiões específicas, como o modelo de Koppula (1981) formulado para solos de Chicago.

Nota-se que alguns usam apenas uma variável de entrada, enquanto outros apresentam múltiplas variáveis. Três dessas variáveis são propriedades chamadas índices físicos e correlacionam pesos e volumes das três fases de uma amostra de solo – as partículas sólidas e os vazios preenchidos por água e ar. O primeiro apresentado, ω_n , é a relação entre o peso de água e o peso de partículas sólidas em uma amostra; o índice de vazios, e_0 , é a relação entre o volume de vazios e o volume de sólidos; e a massa específica dos sólidos, G_s , que é a relação entre a massa e o volume das partículas de um solo (Pinto, 2006). Por fim, a quarta e a quinta variável estão relacionadas ao índice de consistência que um solo apresenta em função da quantidade de água presente nele. Essas variáveis são: o limite de plasticidade, LL , que indica o teor de umidade em que o solo passa do estado líquido para o estado plástico; e o índice de plasticidade, IP , que indica a faixa de umidade em que um solo se encontra no estado plástico (Das, 2011).

Tabela 2 – Equações empíricas para estimativa do índice de compressão

Equação	Referência
$C_c = 0.01 \cdot \omega_n - 0.05$	Azzouz et al. (1976)
$C_c = 0.01 \cdot \omega_n$	Koppula (1981)
$C_c = 0.013 \cdot \omega_n - 0.115$	Park e Lee (2011)
$C_c = 0.54 \cdot e_0 - 0.19$	Nishida (1956)
$C_c = 0.43 \cdot e_0 - 0.11$	Cozzolino (1961)
$C_c = 0.75 \cdot e_0 - 0.38$	Sower (1970)
$C_c = 0.49 \cdot e_0 - 0.11$	Park e Lee (2011)
$C_c = 0.4 \cdot (e_0 - 0.25)$	Azzouz et al. (1976)
$C_c = 0.15 \cdot e_0 + 0.01077$	Bowles (1989)
$C_c = 0.287 \cdot e_0 - 0.015$	Ahadiyan et al. (2008)
$C_c = 0.3 \cdot (e_0 - 0.27)$	Hough (1957)
$C_c = 1.02 - 0.95 \cdot e_0$	Gunduz (2007)
$C_c = 0.014 \cdot (IP + 3.6)$	Sridharan e Nagaraj (2000)
$C_c = 0.014 \cdot IP$	Tiwari e Ajmera (2012)
$C_c = 0.0028 \cdot IP - 0.0052$	Bello et al. (2019)
$C_c = \frac{IP}{74}$	Wroth e Wood (1978)
$C_c = 1.325 \cdot IP$	Wroth e Wood (1978)
$C_c = 0.02 + 0.014 \cdot IP$	Nacci et al. (1975)
$C_c = 0.0082 \cdot IP + 0.0915$	Solanki (2012)

Tabela 2 (Continuação)

Equação	Referência
$C_c = 0.0086 \cdot (IP + 24.2674)$	Vinod (2010)
$C_c = 0.007 \cdot IP + 0.01$	Akayuli e Ofosu (2013)
$C_c = 0.014 \cdot IP + 0.165$	Yoon (2004)
$C_c = -0.0049 \cdot IP + 0.2882$	Salih (2020)
$C_c = 0.0038 \cdot IP + 0.22$	Dway e Thant (2014)
$C_c = 0.0058 \cdot (IP + 13.776)$	Laskar e Pal (2012)
$C_c = 0.003 \cdot IP - 0.081$	Nesamatha e Arumairaj (2015)
$C_c = 0.007 \cdot IP + 0.04$	Rashed et al. (2017)
$C_c = 0.017 \cdot IP + 0.18$	Shakih et al. (2014)
$C_c = 0.0082 \cdot IP + 0.0475$	Jain et al. (2015)
$C_c = 0.013 + 0.020 \cdot IP$	Kootahi e Moradi (2017)
$C_c = 0.006 \cdot (LL - 9)$	Azzouz et al. (1976)
$C_c = \frac{LL-13}{109}$	Mayne (1980)
$C_c = 0.009 \cdot (LL - 10)$	Terzaghi e Peck (1967)
$C_c = 0.014 \cdot LL - 0.168$	Park e Lee (2011)
$C_c = 0.0046 \cdot (LL - 9)$	Bowles (1989)
$C_c = 0.2343 \cdot \left(\frac{LL}{100}\right) \cdot G_s$	Nagaraj e Murthy (1985)
$C_c = 0.2926 \cdot \left(\frac{LL}{100}\right) \cdot G_s$	Park e Lee (2011)
$C_c = 0.009 \cdot \omega_n + 0.005 \cdot LL$	Koppula (1981)
$C_c = 0.009 \cdot \omega_n + 0.002 \cdot LL - 0.1$	Azzouz et al. (1976)
$C_c = 0.4 \cdot (e_0 + 0.001 \cdot \omega_n - 0.25)$	Azzouz et al. (1976)
$C_c = -0.156 + 0.411 \cdot e_0 + 0.00058 \cdot LL$	Al-Khafaji e Andersland (1992)
$C_c = -0.023 + 0.271 \cdot e_0 + 0.001 \cdot LL$	Ahadiyan, Ebne e Bajestan (2008)
$C_c = 0.37 \cdot (e_0 + 0.003 \cdot LL + 0.0004 \cdot \omega_n - 0.34)$	Azzouz et al. (1976)
$C_c = -0.404 + 0.341 \cdot e_0 + 0.006 \cdot \omega_n + 0.004 \cdot LL$	Yoon (2006)
$C_c = 0.141 \cdot G_s^{1.2} \cdot \left(\frac{1+e_0}{G_s}\right)^{2.38}$	Herrero (1983)

Fonte: Kalantary e Kordnaeij (2012); Uzer (2024)

3 Aprendizado de máquina automatizado

As aplicações e pesquisas relacionadas a aprendizado de máquina apresentaram um crescimento significativo na última década – como apresentado brevemente no Capítulo 1. Entretanto, o desempenho dos modelos depende de decisões de projeto. Isso pode apresentar certo empecilho para usuários menos experientes, por exigir conhecimento relacionado à escolha dos algoritmos certos para cada tipo de problema e aos hiperparâmetros que configuram cada algoritmo, devendo esse processo ser repetido a cada aplicação de um modelo diferente (Hutter et al., 2019). Além disso, com o desenvolvimento constante de novas técnicas, a incorporação de novas práticas e ferramentas pode ser um desafio para profissionais da área (Erickson et al., 2020).

O aprendizado de máquina automatizado oferece uma solução para essas questões: os usuários podem fornecer os dados e indicar a variável alvo; o modelo, por sua vez, é encarregado de executar técnicas de pré-processamento, assim como de treinar e combinar diversos algoritmos para encontrar os modelos com melhores desempenhos (Hutter et al., 2019). Essa técnica apresenta tanto a facilidade para usuários menos experientes abordarem diversos modelos, quanto a oportunidade de reduzir processos repetitivos para usuários também mais experientes (Erickson et al., 2020).

A seguir, são apresentados três modelos publicados na literatura recente: o AutoGluon, o FLAML e o H2O. Um breve panorama de cada trabalho será apresentado e, em seguida, as características básicas de cada modelo serão discutidas, com foco nas especificações que cada modelo possui para pré-processamento de dados, treinamento de algoritmos de aprendizado de máquina e estratégias de busca para encontrar os modelos mais satisfatórios.

3.1 AutoGluon

Erickson et al. (2020) propuseram um framework de código aberto em Python. O objetivo era que este modelo de AutoML, o AutoGluon, fosse capaz de trabalhar com dados brutos diretamente e que usasse a combinação de diversos algoritmos de aprendizado de máquina para obter métricas de desempenho mais satisfatórias. Além disso, o trabalho destaca como diferencial desse modelo a capacidade de interrupção e retomada de seus processos. Sua eficiência foi validada por meio de um benchmark, em que foram aplicados 50 testes de classificação e regressão ao AutoGluon e a outros modelos públicos e comerciais: TPOT, H2O, AutoWEKA, auto-sklearn e GoogleML Tables. Os resultados apresentados indicaram que o AutoGluon superou os demais modelos na maior parte dos casos.

A estrutura básica de implementação é apresentada no código a seguir. A função `fit()` apresenta diversos parâmetros que especificam o processo de treinamento dos

modelos: a otimização dos hiperparâmetros dos algoritmos, definida pelo parâmetro `hyperparameter_tune`; a estratégia de *stacking*, pelo parâmetro `auto_stack`; o limite de tempo, por `time_limits`; e a métrica de avaliação, por `eval_metric`.

```

1 from autogluon import TabularPrediction as task
2 predictor = task.fit("train.csv", label="class")
3 predictions = predictor.predict("test.csv")

```

Como citado anteriormente, uma das características principais do AutoGluon é a sua capacidade de trabalhar com dados brutos. Quando não há especificações por parte do usuário, o modelo recebe um conjunto de dados e reconhece o tipo de problema a ser solucionado, de acordo com o tipo de dados que a coluna da variável alvo contém. Valores de string indicariam um problema de classificação, enquanto valores numéricos indicariam um problema de regressão. Além disso, o modelo também é capaz de fazer o processamento dos dados, organizado em duas etapas. Na primeira, é feito um pré-processamento chamado pelos autores de agnóstico, em que há a classificação dos dados como numéricos, categóricos, textuais ou data/hora. Quando o modelo não consegue reconhecer o tipo de dado de uma coluna, ela é descartada, sendo considerada de pouco valor preditivo. Na segunda etapa, o conjunto é tratado em relação aos dados faltantes. Em vez de remover as respectivas amostras ou preencher as lacunas com a média ou a mediana, é criada a categoria "Desconhecido", permitindo que o modelo lide com a falta de informação. Outra função executada automaticamente é a normalização dos dados, que é explicada em mais detalhes no Capítulo 4.

Os dados são treinados por um conjunto de algoritmos de aprendizado de máquina, passados como hiperparâmetros na função `fit()`, que estão apresentados na Tabela 3. Os autores ressaltam que o AutoGluon treina os modelos de forma sequencial, em contraponto com a maioria dos modelos de AutoML disponíveis no momento da publicação do trabalho, que treinam seus modelos em paralelo. A ordem dos algoritmos pode ser alterada, assim como podem ser escolhidos quais algoritmos serão usados no treinamento.

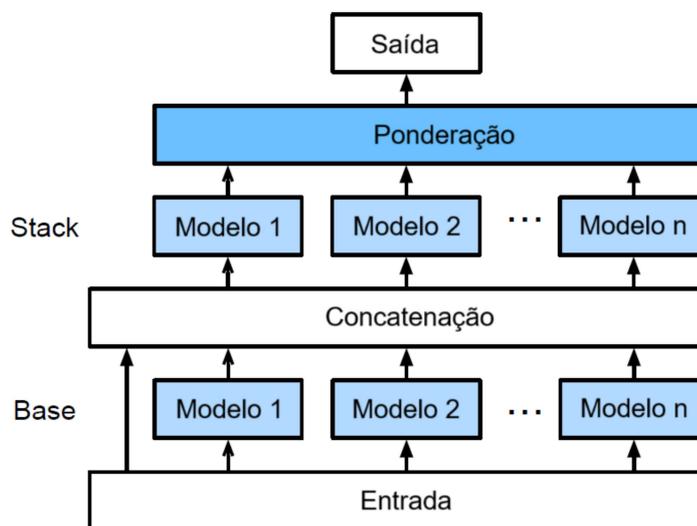
Tabela 3 – Algoritmos de aprendizado de máquina usados pelo AutoGluon

Parâmetro	Algoritmo
NN	Neural Networks
GBM	LightGBM
CAT	CatBoost
XGB	XGBoost
FASTAI	FastAI
XT	Extra Trees
KNN	K-Neighbors
RF	Random Forest

Fonte: Erickson et al. (2020)

Cada modelo de AutoML apresenta sua própria combinação de estratégias de busca de modelos com as métricas mais satisfatórias. No caso do AutoGluon, os autores destacam duas características: a validação cruzada e o *stacking*. A validação cruzada, que também será explicada em mais detalhes no Capítulo 4, é passada como hiperparâmetro e deve ser ativada com especificação da quantidade de dobras. O *stacking* é a prática de combinar as previsões de algoritmos de aprendizado de máquina para obter modelos melhores. O processo de *stacking* do AutoGluon é descrito como multicamadas e suas etapas são ilustradas na Figura 6. Primeiramente, os conjuntos de dados são passados como parâmetros de entrada e os algoritmos de aprendizado de máquina são treinados individualmente. Em seguida, as previsões feitas pelos algoritmos são concatenadas aos dados e os conjuntos são passados para a próxima camada. Tal camada é constituída de múltiplos modelos treinados para combinar as previsões dos modelos da camada anterior, chamados *stackers*. Esses, por sua vez, são passados para a próxima camada e o processo é repetido múltiplas vezes. Por fim, a última camada agrega as previsões dos *stackers* de forma ponderada, o que melhora a precisão do modelo e também evita o sobreajuste de dados.

Figura 6 – Esquema da estratégia de *stacking* multicamada do AutoGluon



Fonte: Adaptado de Erickson et al. (2020)

O AutoGluon usa os mesmos tipos de modelos e hiperparâmetros, tanto na camada base, quanto nas camadas de *stacking*. Essa característica contrasta com as formas de *stacking* tradicionais, que geralmente usam modelos mais simples nas camadas superiores. Os autores ressaltam que essa técnica pode ser vista de forma análoga ao aprendizado profundo, em que cada neurônio é, na verdade, um modelo de aprendizado de máquina completo. Além disso, os *stackers* recebem, não só as previsões da camada anterior, como também os dados originais. Ao incluir os dados originais em cada camada, o método cria

algo semelhante às conexões de salto usadas em redes neurais. Isso permite que os modelos de camadas superiores tenham acesso aos dados originais durante todo o treinamento, garantindo a robustez do modelo.

O tempo total de execução, que é pré-estabelecido pelo usuário, é dividido igualmente entre as camadas de *stacking*. Antes de treinar cada modelo, o AutoGluon estima o tempo necessário e, caso não haja tempo suficiente, ele pula para a próxima camada. Além disso, cada modelo é salvo logo após o treinamento, garantindo uma tolerância a falhas de hardware, interrupção da rede ou tempo limite excedido. Dessa forma, o modelo pode ser retomado caso o processo seja interrompido.

3.2 FLAML

Wang et al. (2021) desenvolveram um framework de código aberto em Python, sob a justificativa de ser uma linguagem popular na área de ciência de dados. Tal modelo, chamado Fast and Lightweight AutoML (FLAML), se propõe a preencher uma lacuna relacionada à capacidade das bibliotecas tradicionais de apresentarem boas métricas de desempenho diante de orçamentos computacionais menores. Os recursos usados para alcançar esse objetivo são estratégias de busca baseadas em estimativas de custo para melhoria do modelo. Para validar o desempenho do FLAML, o trabalho apresenta um benchmark com outros modelos de código aberto: auto-sklearn, H2O e TPOT. Foram aplicados 53 conjuntos de dados, sendo 39 para problemas de classificação e 14 para problemas de regressão. Os resultados apresentados indicaram melhor desempenho do FLAML em 60 a 85% das tarefas, quando fornecidos orçamentos iguais para todos os modelos.

A estrutura básica de implementação é apresentada no código a seguir. A classe principal, `AutoML()`, recebe os parâmetros `time_budget` para definir o orçamento de tempo, `metric` para escolher métricas de desempenho, `task` para especificar o tipo de tarefa, `n_jobs` para ativar paralelização e `estimator_list` para determinar quais algoritmos de aprendizado de máquina serão usados no processo.

```

1 from flaml import AutoML
2 automl = AutoML()
3 automl.fit(X_train, y_train, task='classification')
4 prediction = automl.predict(X_test)

```

Em relação ao pré-processamento dos dados, o trabalho ressalta que o modelo realiza uma amostragem progressiva para conjuntos de dados com menos de 10.000 amostras. Isso quer dizer que o modelo começa seu processo de amostragem com grupos menores e gradativamente aumenta o tamanho desses grupos, de acordo com a necessidade. Além disso, para problemas de classificação, o FLAML faz uso de amostragem estratificada,

garantindo que a distribuição seja representativa do conjunto completo.

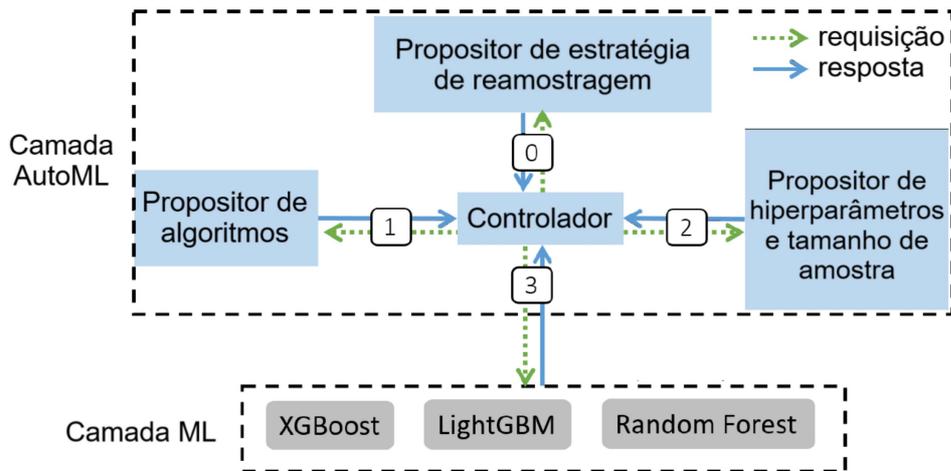
O processo de amostragem do modelo aponta para o processo iterativo da etapa de treinamento dos dados na busca pelos modelos mais adequados. Para essa etapa, o modelo conta com uma arquitetura dividida em duas camadas: a camada ML, que contém os algoritmos de aprendizado de máquina a serem usados, e a camada AutoML, que contém quatro elementos essenciais à análise. Esses elementos são: o controlador; o propositor de algoritmos; o propositor de hiperparâmetros e tamanho de amostras; e o propositor de estratégia de reamostragem. A cada iteração, o propositor de estratégia de reamostragem escolhe o método de validação, podendo ser validação cruzada ou *holdout*. Em seguida, o propositor de algoritmo seleciona um dos algoritmos na camada ML e o propositor de hiperparâmetros e tamanho de amostras define os respectivos parâmetros do algoritmo, por meio de busca aleatória, e o tamanho da amostra a ser usada. Por fim, o controlador executa e avalia o treinamento, observando o desempenho e o orçamento computacional. O processo é repetido até o esgotamento do tempo disponível. A Figura 7 ilustra o esquema descrito e a Tabela 4 apresenta os algoritmos usados pelo modelo. O trabalho também destaca que não há dependência de meta-aprendizado no processo.

Tabela 4 – Algoritmos de aprendizado de máquina usados pelo FLAML

Parâmetro	Algoritmo
<code>lgb</code>	LightGBM
<code>xgboost</code>	XGBoost
<code>xgb_limitdepth</code>	XGBoost Limit Depth
<code>rf</code>	Random Forest
<code>extra_tree</code>	Extra Trees
<code>histgb</code>	Hist Gradient Boosting
<code>castboost</code>	CatBoost
<code>kneighbor</code>	K-Neighbors

Fonte: Wang et al. (2021)

Figura 7 – Esquema da estratégia de busca do FLAML



Fonte: Adaptado de Wang et al. (2021)

Segundo os autores, a estratégia de busca pelos melhores modelos do FLAML apresenta uma inovação: o conceito de Custo Estimado para Melhoria (ECI). Esse é um indicador dinâmico que é calculado a cada iteração de duas formas diferentes, dependendo do caso em que um algoritmo de aprendizado de máquina se encontra, quando comparado com os demais.

No primeiro caso, quando um algoritmo tem a melhor métrica de desempenho até então, o ECI é obtido pelo mínimo entre dois valores: o ECI_1 e o ECI_2 . O ECI_1 , apresentado na Equação 3.1, estima o custo para que o algoritmo em questão encontre um conjunto de hiperparâmetros que apresente métricas melhores do que as atuais. O ECI_2 , apresentado na Equação 3.2, quantifica o custo para que esse algoritmo – com o conjunto atual de hiperparâmetros e custo computacional já consumido – treine um grupo amostral maior. No segundo caso, quando o algoritmo não tem a melhor métrica de desempenho até então, o ECI é obtido pela Equação 3.3. A partir desses indicadores, o FLAML escolhe qual algoritmo será explorado na próxima iteração. A probabilidade de um algoritmo ser escolhido é inversamente proporcional ao seu ECI .

$$ECI_1 = \max(K_0 - K_1, K_1 - K_2) \quad (3.1)$$

$$ECI_2 = c \cdot \kappa_l \quad (3.2)$$

$$ECI = \max\left(\frac{(\tilde{\epsilon}_l - \tilde{\epsilon}^*)(K_0 - K_2)}{\delta}, \min(ECI_1, ECI_2)\right) \quad (3.3)$$

em que:

- K_0 : custo total gasto no algoritmo de aprendizado de máquina até o momento;
- K_1 : custo total gasto no algoritmo até o momento, em que a melhor configuração foi encontrada;
- K_2 : custo total gasto no algoritmo até o momento, em que a segunda melhor configuração foi encontrada;
- c : fator de multiplicação;
- κ_l : custo da última tentativa para o algoritmo;
- $\tilde{\epsilon}_l$: melhor erro atual do algoritmo;
- $\tilde{\epsilon}^*$: melhor erro atual entre todos os algoritmos;
- δ : redução de erro entre as duas melhores configurações do algoritmo.

Assim como outros modelos de aprendizado de máquina automatizado, o FLAML também possui a função de *stacking*, para combinar algoritmos e obter métricas mais satisfatórias. É um recurso desativado por padrão para manter a sobrecarga baixa, mas há a opção de habilitá-lo quando o armazenamento e o orçamento computacional não são uma preocupação.

3.3 H2O

LeDell e Poirier (2020) propuseram um framework de código aberto com API em R, Python, Java e Scala, sob a justificativa de que possa ser usado por equipes diversas da área de ciência de dados. O objetivo era que o modelo, chamado H2O, fosse capaz de gerar resultados precisos e competitivos, por meio da paralelização do treinamento e do *stacking* de algoritmos. Como resultado, é apresentada uma lista de classificação dos melhores modelos. Para verificar a eficiência do modelo, foi executado um benchmark com outros modelos de AutoML: Auto-WEKA, auto-sklearn, TPOT e AutoGluon-Tabular. Foram usados 44 conjuntos de dados de classificação binária e multiclasse. Embora nenhum modelo superasse os demais de forma consistente, os resultados indicaram que o H2O se destacou em diversas tarefas e apresentou um desempenho competitivo, especialmente em termos de escalabilidade e de velocidade.

A seguir, é apresentada a estrutura básica de implementação do modelo em R e em Python. Os autores ressaltam que a interface foi projetada para ter o mínimo possível de parâmetros, de forma que o usuário precisa indicar o conjunto de dados

pelo parâmetro `training_frame`, a variável alvo por `y`, o orçamento computacional por `max_runtime_secs` e o número total de modelos treinados por `max_models`.

```

1 aml = H2OAutoML(max_runtime_secs=3600)
2 aml.train(y="response_colname", training_frame=train)
3 aml <- h2o.automl(y="response_colname", training_frame=train,
  max_runtime_secs=3600)

```

Em termos de pré-processamento de dados, o H2O apresenta como recursos: imputação automática, preenchendo valores ausentes (de forma não especificada no trabalho); normalização, quando necessário; e codificação one-hot para problemas de classificação. Os modelos baseados em árvores suportam variáveis categóricas, permitindo que os dados sejam tratados em sua forma original.

O modelo realiza treinamento e validação cruzada, diante da especificação de valores maiores que 1 para o parâmetro `nfolds`. Os algoritmos usados pelo H2O estão apresentados na Tabela 5. De forma padrão, são aplicadas: três configurações pré-determinadas e uma grade aleatória de XGBoost; cinco configurações de GBM; uma grade fixa de GLM; uma configuração de DRF; e uma configuração de Fully Connected Deep Neural Network. Os modelos previamente especificados têm o objetivo de fornecer padrões eficientes para cada algoritmo e, a partir dos intervalos definidos para os hiperparâmetros de alguns modelos, usa-se a busca aleatória. A ordem de treinamento dos algoritmos, que pode ser alterada pelo usuário, é definida para começar com modelos que, em geral, fornecem bons resultados, como XGBoost e GLM. Gradativamente, então, são inseridos outros modelos, para aumentar a diversidade na etapa de *stacking*.

Tabela 5 – Algoritmos de aprendizado de máquina usados pelo H2O

Parâmetro	Algoritmo
XGBoost	XGBoost
GBM	Gradient Boosting Machine
GLM	Generalized Linear Model
DRF	Distributed Random Forest
XRT	Extremely Randomized Trees
Deep Learning	Fully Connected Deep Neural Network

Fonte: LeDell e Poirier (2020)

Após o treinamento dos modelos individualmente, são construídos dois modelos de *stacked ensemble*, uma classe de algoritmos composta por algoritmos individuais previamente treinados. Como dito anteriormente, a busca aleatória para alguns algoritmos produz um conjunto diverso de modelos a serem usados de base. O H2O produz, então, dois ensembles: o *Todos os Modelos* e o *Melhor da Família*. O *Melhor da Família* é otimizado para casos de uso de produção – que necessitam de menor custo – e contém

seis ou menos modelos de base. Conseqüentemente, ele geralmente apresenta desempenho ligeiramente inferior ao *Todos os Modelos*. Em ambos os casos, o treinamento é feito usando validação cruzada, conceito que será melhor explicado no Capítulo 4.

Por fim, como resultado da execução do H2O, é apresentado um *leaderboard*: uma lista com classificação dos modelos de acordo com suas métricas de desempenho ou outros atributos, como tempo de execução.

4 Materiais e métodos

Neste capítulo é descrita a metodologia adotada para o desenvolvimento da pesquisa, bem como os materiais usados. No Item 4.1 são apresentados os critérios usados para a seleção dos bancos de dados e estes são brevemente descritos. No Item 4.2 são apresentadas as ferramentas de pré-processamento usadas em tais conjuntos. No Item 4.3 são descritas as especificações para a implementação dos algoritmos de aprendizado de máquina automatizado. Por fim, no Item 4.4 são apresentadas as métricas de desempenho usadas para comparação de performances.

4.1 Coleta de dados

A seguir, são apresentadas informações referentes aos dados usados neste trabalho e aos seus trabalhos de origem. Estabeleceu-se como critério de escolha que as publicações apresentassem acesso aberto aos dados e que tivessem um banco com pelo menos 300 amostras. Este número é justificado por tentativas anteriores de treinar o H2O com bancos de dados e por obter um aviso do modelo de que era necessário um mínimo de 200 amostras na etapa de treino para que pudesse apresentar métricas satisfatórias.

Benbouras et al. (2019) propuseram modelos que pudessem estimar o índice de compressão de solos usando redes neurais artificiais, programação genética e análise de regressão múltipla. Para isso, foi desenvolvido um banco de dados com 373 amostras referentes a 68 furos de sondagem feitos em projetos geotécnicos na cidade de Argel, na Argélia. As amostras do trabalho têm como variável alvo o índice de compressão, C_c (adimensional), e como variáveis independentes: densidade única, ρ_h (g/cm³); teor de umidade, W (%); índice de vazios inicial, e_0 (adimensional); teor de finos, FC (%); limite de liquidez, WL (%); índice de plasticidade, PI (%); e o tipo de solo (argila, margam, silte, areia e pedregulho).

Kalantary e Kordnaeij (2012) desenvolveram modelos de estimativa do índice de compressão usando redes neurais artificiais e análise de regressão para gerar equações empíricas, comparando os resultados obtidos com os resultados obtidos por equações empíricas selecionadas da literatura. Para isso, foi construído um banco de dados com 400 amostras de ensaios de adensamento realizados em 125 locais de construção na província de Mazandaran, no Irã. Assim como no trabalho anterior, as amostras têm como variável alvo o índice de compressão, C_c (adimensional), e as variáveis independentes são: teor de umidade natural, ω_n (%); limite de liquidez, LL (%); índice de plasticidade, PI (%); índice de vazios inicial, e_0 (adimensional); e peso específico dos sólidos, G_s (g/cm³).

Pham et al. (2024) tiveram como objetivo desenvolver uma equação que pudesse estimar o valor do índice de compressão de um solo utilizando regressão simbólica. O

trabalho contou com o desenvolvimento de um banco de dados contendo 1877 amostras, provenientes de 113 registros de perfurações feitas na cidade de Ho Chin Minh, na China. A variável alvo de cada amostra é o índice de compressão, C_c (adimensional), como nos demais trabalhos, e as variáveis independentes deste banco de dados são: frações de argila, silte, areia e pedregulho (%); densidade úmida, ρ (g/cm³); densidade seca, ρ_d (g/cm³); massa específica, G_s (g/cm³); teor de umidade, w (%); porosidade, n (%); índice de vazios inicial, e_0 (adimensional); limite de liquidez, LL (%); limite de plasticidade, PL (%); e índice de plasticidade, PI (%).

Uzer (2024) propôs um modelo para estimar o índice de compressão de solos finos, comparando a implementação de redes neurais artificiais com modelos de regressão tradicionais. Para isso, foi construído um banco de dados com base em resultados de ensaios desenvolvidos na Universidade Técnica de Istambul, na Turquia, nos períodos de 1959 a 1982 e de 1993 a 2010. Os dados inicialmente coletados passaram por um processo de tratamento em que *outliers* foram desconsiderados, resultando em um conjunto de 458 amostras. A variável alvo das amostras é o índice de compressão, C_c (adimensional), e as variáveis independentes são: teor de umidade natural, w_n (%); limite de liquidez, LL (%); limite de plasticidade, LP (%); e índice de vazios inicial, e_0 (adimensional).

4.2 Pré-processamento e análise dos dados

O processamento de dados feitos neste trabalho incluiu quatro etapas, que serão explicadas a seguir: análise de correlação, seleção de variáveis de entrada e divisão dos dados para as etapas de treinamento e teste.

Para a primeira etapa, foi preciso entender a influência que as variáveis de entrada exercem sobre a variável independente. Para isso, foi usada a matriz de correlação pela função `corr()`, da biblioteca Pandas. A matriz de correlação apresenta como parâmetro padrão, para cada par de variáveis, o coeficiente de correlação de Pearson, apresentado na Equação 4.1 (Bruce e Bruce, 2019).

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)s_x s_y} \quad (4.1)$$

em que:

x_i : variável a ser correlacionada com a variável y ;

\bar{x} : média de valores das amostras de x ;

y_i : variável a ser correlacionada com a variável x ;

\bar{y} : média de valores das amostras de y ;

n : número de pares de observações;

s_x : desvio padrão da média de x ;

s_y : desvio padrão da média de y .

Este coeficiente quantifica a relação que o aumento ou a diminuição de uma variável de entrada tem no aumento ou na diminuição da variável de saída. Além disso, também quantifica a relação que o aumento da entrada tem com a diminuição da saída (ou vice-versa). Quando ambas as variáveis crescem de forma diretamente proporcional, elas são consideradas positivamente correlacionadas, e seu coeficiente de correlação varia entre 0 e 1. Por outro lado, quando elas crescem de forma inversamente proporcional, elas são consideradas negativamente correlacionadas, e seu coeficiente varia entre -1 e 0 (Bruce e Bruce, 2019).

A etapa seguinte está relacionada à seleção das variáveis de entrada. A partir da matriz de correlação, pode-se identificar quais atributos possuem maior e menor influência em um banco de dados específico. Tal ferramenta pode orientar a seleção de um conjunto específico de variáveis para serem usadas no modelo. Para o caso de bancos de dados com uma grande quantidade de amostras e de atributos, essa técnica pode otimizar o tempo de processamento dos modelos, assim como pode melhorar seus desempenhos (Gerón, 2021).

Por fim, a última etapa está relacionada à divisão dos dados para as etapas de treinamento e teste. Neste trabalho, adotou-se a indicação de Gerón (2021) na divisão de 80% do banco de dados para a etapa de treinamento – e de validação, quando há – e 20% para a etapa de teste. A divisão foi feita com a função `train_test_split`, da biblioteca `scikit-learn`, adotando a proporção observada. Para garantir a reprodutibilidade dessa etapa, foram usadas 30 sementes aleatórias – variando de 0 a 29, de acordo com o índice da respectiva execução.

4.3 Especificações de implementação

A implementação dos modelos de AutoML foi feita considerando algumas especificações, tanto para obter resultados menos enviesados, quanto para fins de comparação dos modelos entre si. Tais especificações serão discutidas a seguir: normalização dos dados, padronização dos algoritmos, *stacking*, validação cruzada e execuções independentes.

De maneira geral, algoritmos de aprendizado de máquina podem não performar muito bem quando suas variáveis de entrada possuem atributos de escalas muito diferentes entre si (Gerón, 2021) – no caso dos dados usados neste trabalho, um exemplo é o limite de liquidez de Kalantary e Kordnaeij (2012) com variação de 39.82% a 81.00%, enquanto o índice de vazios inicial varia de 0.21 a 0.63. Isso acontece porque os atributos com maiores escalas têm uma tendência a exercer maior influência no resultado final, por conta de sua alta variação em torno da média. Para evitar este problema, foi usada a técnica da

padronização ou normalização dos dados (Bruce e Bruce, 2019). Os modelos de AutoML usados oferecem recursos de normalização dos dados de entrada como hiperparâmetros em suas funções de treino. Dessa forma, foi necessário apenas verificar se as funções de treinamento tinham como padrão a função de normalização ativada e, caso não tivessem, o recurso seria ativado manualmente.

Como apresentado no Capítulo 3, cada um dos modelos faz uso de diversos algoritmos de regressão passados como hiperparâmetros para as estimativas. Ao longo do estudo da documentação oficial do H2O, constatou-se que, para garantir a reprodutibilidade do modelo, era necessário excluir o Deep Learning da lista de hiperparâmetros. Optou-se, então, pela padronização dos modelos de AutoML em relação aos algoritmos de regressão, sendo usados apenas aqueles que são comuns aos três modelos. Justifica-se essa escolha como uma possibilidade de investigar qual a eficiência dos três modelos diante de condições semelhantes. A Tabela 6 apresenta as especificações de hiperparâmetros passadas em cada modelo. De maneira geral, são variações dos algoritmos Gradient Boosting Machine, XGBoost, Random Forest e Extremely Randomized Trees.

Tabela 6 – Algoritmos passados como hiperparâmetros nos modelos de AutoML

Modelo	AutoGluon	FLAML	H2O
Algoritmos	GBM	lgbm	GBM
	XGB	xgboost	XGB
	RF	rf	RF
	XT	extra_tree	XRT

Fonte: Autor (2025)

Ainda em relação aos algoritmos, como também explicado no Capítulo 3, os três modelos usam *stacking*, ou seja, combinam múltiplos algoritmos para gerar resultados mais acurados. Optou-se por usar essa ferramenta, verificando se os modelos tinham como padrão a função ativada e, caso não tivessem, o recurso seria ativado manualmente.

Outras técnicas importantes usadas neste trabalho são a validação cruzada e as execuções independentes, como ferramenta para evitar o sobreajuste dos modelos aos dados disponíveis. A validação cruzada é uma técnica que consiste em dividir o conjunto de dados em partes iguais e reservar uma para testar a capacidade preditiva do modelo, enquanto as demais partes são usadas para treinamento. O processo é repetido até que cada uma das partes tenha sido usada como conjunto de teste. Dessa forma, é possível otimizar a extração de informação dos bancos de dados, gerando um modelo mais robusto (Gerón, 2021). Verificou-se que os modelos de AutoML usados possuem a validação cruzada como opção e, para este trabalho, foi adotada a divisão dos conjuntos de dados em 5 partes iguais. Além disso, foi implementado manualmente um loop de 30 execuções independentes, em que os conjuntos de dados de treino e teste foram selecionados aleatoriamente por meio de

uma semente aleatória – que, para este trabalho, teve como valores arbitrados os números das execuções, para garantir reprodutibilidade.

No processo de estudo das documentações oficiais, descobriu-se também que eram precisos outros artifícios para garantir a reprodutibilidade das implementações. Além da exclusão do algoritmo Deep Learning, como descrito anteriormente, o H2O requer a determinação de um valor para o parâmetro `max models`, referente ao número máximo de modelos individuais treinados, para o qual foi arbitrado um valor máximo de 20. Verificou-se também que o H2O e o FLAML requerem uma semente aleatória e foi arbitrado o valor de 42. Por fim, foi necessário também especificar um tempo máximo de execução para o H2O e, portanto, foi arbitrado o orçamento computacional de 120 segundos para todos os modelos, a fim de se manter a padronização.

4.4 Métricas de desempenho

As etapas de teste têm seu desempenho avaliado quantitativamente pelas chamadas métricas de desempenho. Para este trabalho, foram usados o coeficiente de determinação, a raiz quadrada do erro médio e o tempo de execução de cada um dos modelos. As métricas obtidas foram apresentadas e comparadas no Capítulo 5 e esses resultados foram discutidos.

O coeficiente de determinação, ou R^2 , é uma métrica que está apresentada na Equação 4.2. Seus resultados variam de 0 a 1 e medem a proporção de variação dos dados. Quanto mais próximo de 1 é o coeficiente, mais bem ajustado o modelo é em relação aos dados usados, assim como quanto mais próximo de 0 é o coeficiente, menos ajustado é o modelo Bruce e Bruce (2019).

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4.2)$$

em que:

y_i : valor real de y ;

\hat{y}_i : valor previsto de y ;

n : número de amostras.

A raiz quadrada do erro quadrático médio, ou *RMSE*, é uma métrica representada pela Equação 4.3 e indica um valor médio de erro apresentado nas estimativas de um modelo. Em contraponto com o erro quadrático médio, o MSE, tem a vantagem de ser mais sensível a possíveis outliers e tem resultados mais fáceis de serem interpretados, por estarem na mesma dimensão que os valores estimados (Bruce e Bruce, 2019).

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (4.3)$$

em que:

y_i : valor real de y ;

\hat{y}_i : valor previsto de y ;

n : número total de amostras.

5 Resultados

Neste capítulo são apresentados os resultados obtidos. O Item 5.1 apresenta estatísticas dos bancos de dados, bem como a matriz de correlação entre as variáveis que os compõem. O Item 5.2, por sua vez, apresenta as estatísticas das métricas de desempenho usadas para comparação de performance dos modelos.

5.1 Pré-processamento e análise dos dados

Embora os bancos de dados apresentem grandes diferenças em quantidades e tipos de variáveis independentes, optou-se como primeiro tratamento a padronização em relação a esses atributos. Essa padronização foi feita considerando apenas as variáveis que os quatro bancos têm em comum e descartando as demais. A decisão é justificada pela intenção de facilitar a comparação de desempenho de cada modelo entre os bancos de dados. Dessa forma, as variáveis que foram mantidas são: teor de umidade natural, W (%); limite de liquidez, LL (%); índice de plasticidade, IP (%); e índice de vazios inicial, e_0 (adimensional). As Tabelas 7, 8, 9 e 10 apresentam informações estatísticas sobre os atributos e a variável alvo, respectivamente, para Benbouras et al. (2019), Kalantary e Kordnaeij (2012), Pham et al. (2024) e Uzer (2024).

Tabela 7 – Estatísticas dos dados de Benbouras et al. (2019)

Estatística	W (%)	LL (%)	IP (%)	e_0	C_c
Média	21.97	50.30	26.19	0.63	0.16
Desvio padrão	5.52	9.18	7.47	0.15	0.05
Mínimo	8.00	23.00	4.00	0.28	0.01
25%	18.00	43.00	22.00	0.53	0.13
50%	21.44	51.00	26.00	0.61	0.16
75%	25.07	57.00	29.00	0.72	0.19
Máximo	42.11	75.00	69.00	1.55	0.46

Fonte: Autor (2025)

Tabela 8 – Estatísticas dos dados de Kalantary e Kordnaeij (2012)

Estatística	W (%)	LL (%)	IP (%)	e_0	C_c
Média	28.60	39.82	18.59	0.77	0.21
Desvio padrão	7.80	9.90	8.58	0.18	0.08
Mínimo	10.20	24.00	3.00	0.36	0.05
25%	23.72	33.00	12.00	0.66	0.16
50%	27.55	38.00	17.00	0.74	0.19
75%	31.40	45.00	24.00	0.83	0.25
Máximo	70.00	81.00	50.00	1.88	0.63

Fonte: Autor (2025)

Tabela 9 – Estatísticas dos dados de Pham et al. (2024)

Estatística	W (%)	LL (%)	IP (%)	e_0	C_c
Média	31.11	38.29	15.92	0.91	0.22
Desvio padrão	22.16	17.07	10.55	0.58	0.29
Mínimo	11.90	16.80	2.99	0.40	0.02
25%	18.30	24.40	5.90	0.57	0.06
50%	21.22	32.20	14.70	0.66	0.10
75%	29.88	50.47	24.10	0.88	0.19
Máximo	98.92	95.36	51.71	2.82	1.91

Fonte: Autor (2025)

Tabela 10 – Estatísticas dos dados de Uzer (2024)

Estatística	W (%)	LL (%)	IP (%)	e_0	C_c
Média	35.16	53.31	26.31	1.02	0.31
Desvio padrão	11.99	15.84	11.46	0.33	0.15
Mínimo	13.00	23.00	5.00	0.45	0.06
25%	26.00	41.00	17.00	0.78	0.21
50%	32.00	50.00	24.00	0.96	0.27
75%	41.00	64.00	35.00	1.21	0.40
Máximo	70.00	99.00	62.00	2.02	0.75

Fonte: Autor (2025)

Optou-se por excluir algumas amostras dos bancos de dados de Pham et al. (2024) e de Uzer (2024), que estão apresentadas, respectivamente, nas Tabelas 11 e 12. A escolha de exclusão dos dados seguiu o critério estabelecido de manter amostras que apresentassem apenas valores percentuais entre 0 e 100.

Tabela 11 – Amostras retiradas de Pham et al. (2024)

W (%)	LL (%)	IP (%)	e_0	C_c
102.1	103.4	37.7	2.78	0.69
101.2	100.2	35.7	2.69	0.97

Fonte: Autor (2025)

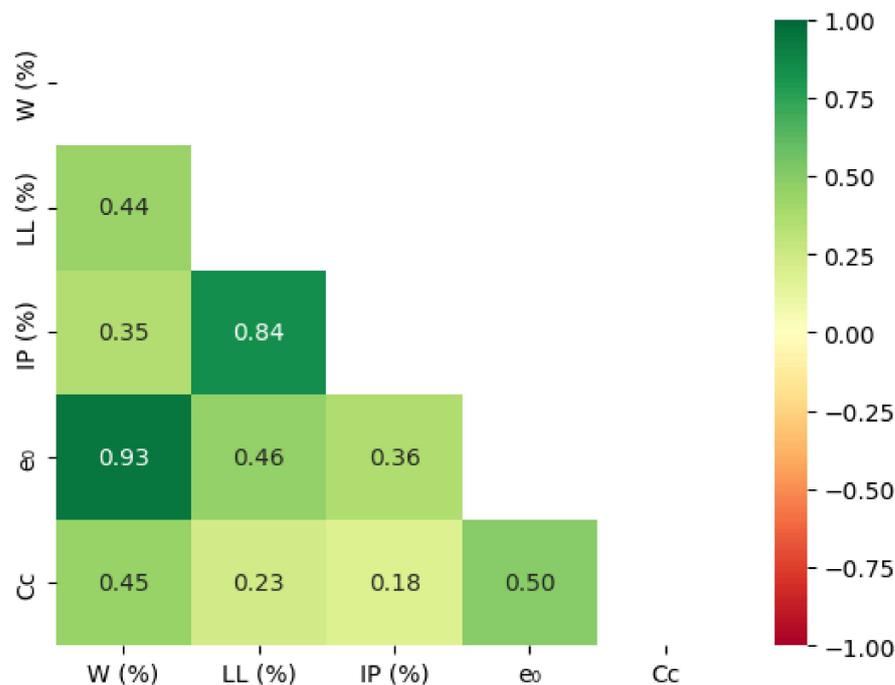
Tabela 12 – Amostras retiradas de Uzer (2024)

W (%)	LL (%)	IP (%)	e_0	C_c
14.00	19.00	-6.00	0.46	0.12
52.00	103.00	64.00	1.55	0.45
51.00	102.00	73.00	1.55	0.50

Fonte: Autor (2025)

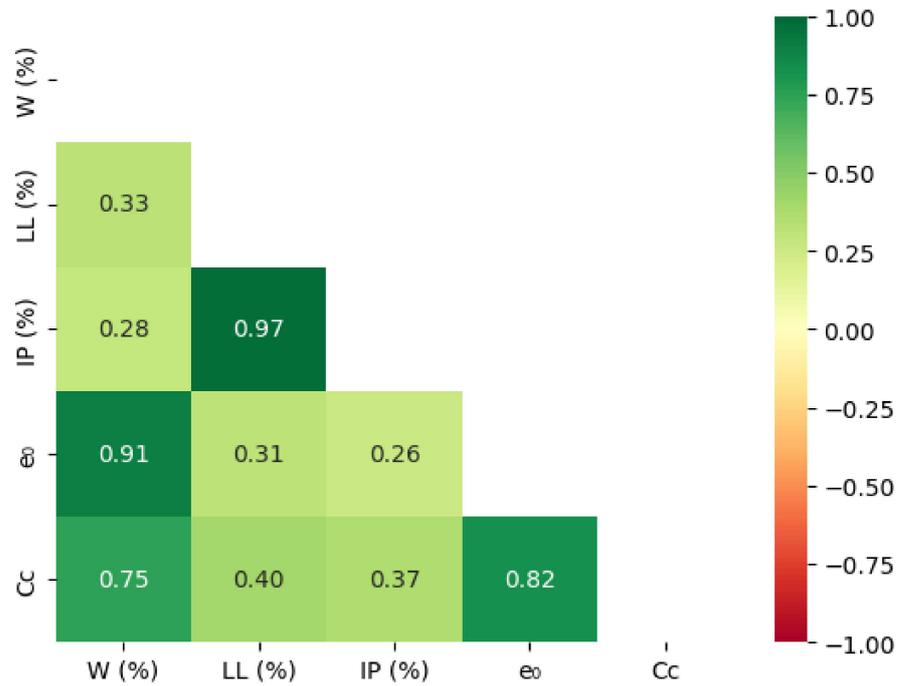
A seguir, são apresentadas as matrizes de correlação entre as variáveis dos bancos de dados. As Figuras 8, 9, 10 e 11 correspondem, respectivamente, aos resultados de Benbouras et al. (2019), Kalantary e Kordnaeij (2012), Pham et al. (2024) e Uzer (2024). Pode-se perceber que, em todos os bancos de dados, os atributos são positivamente correlacionados com a variável alvo, o índice de compressão, C_c . Além disso, a sequência dos atributos com maiores coeficientes aos atributos com menores coeficientes é a mesma em todas as matrizes: a propriedade mais correlacionada é o índice de vazios inicial, e_0 , seguido do teor de umidade, W , do limite de liquidez, LL , e do índice de plasticidade, IP . O índice de vazios, enquanto característica que mais influencia no índice de compressão dos solos, é coerente com a teoria apresentada no Capítulo 2, pois o índice de compressão está relacionado a um fenômeno em que há a diminuição de espaços vazios e o rearranjo de partículas sólidas em amostras.

Figura 8 – Matriz de correlação entre as variáveis de Benbouras et al. (2019)



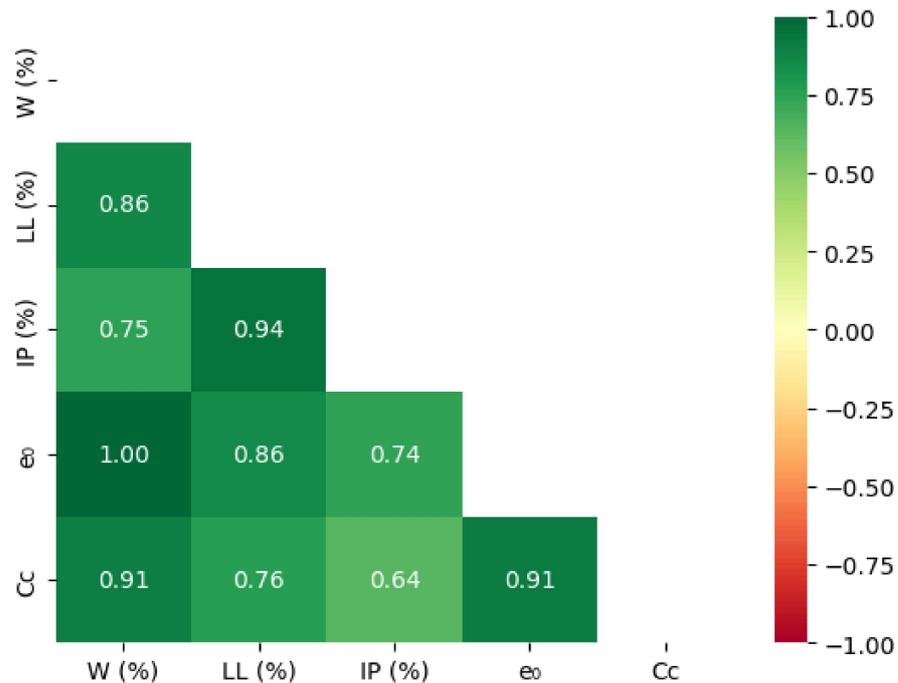
Fonte: Autor (2025)

Figura 9 – Matriz de correlação entre as variáveis de Kalantary e Kordnaeij (2012)



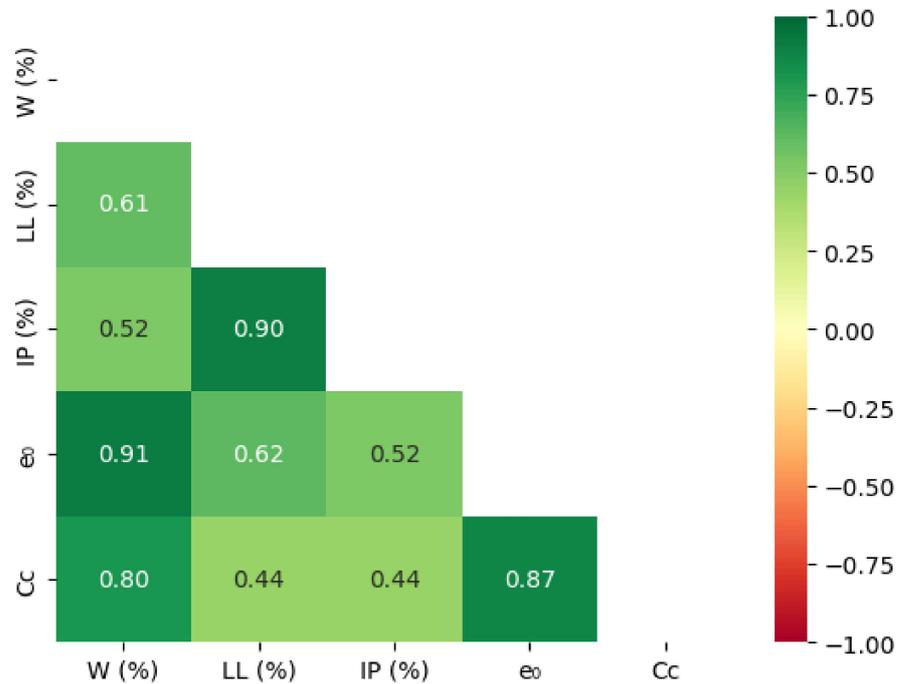
Fonte: Autor (2025)

Figura 10 – Matriz de correlação entre as variáveis de Pham et al. (2024)



Fonte: Autor (2025)

Figura 11 – Matriz de correlação entre as variáveis de Uzer (2024)



Fonte: Autor (2025)

Em relação à seleção de variáveis, observou-se que a padronização dos bancos de dados manteve dois atributos associados à consistência dos solos, o limite de liquidez e o índice de plasticidade. O limite de liquidez apresentou em três das matrizes maior coeficiente de correlação. Dessa forma, sob a justificativa de otimização de variáveis, optou-se por escolher como entrada para os modelos o teor de umidade inicial, W (%), o limite de liquidez, LL (%) e o índice de vazios inicial, e_0 (%).

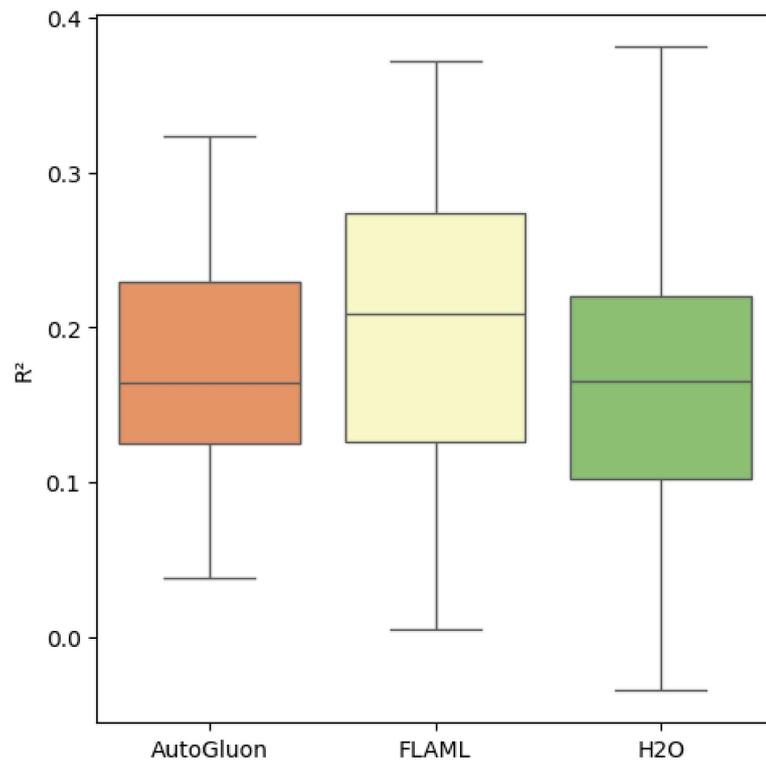
5.2 Métricas de desempenho

A Tabela 13 apresenta as métricas de desempenho para o banco de dados de Benbouras et al. (2019). Tais resultados são apresentados na forma de média e desvio padrão das 30 execuções independentes. Para ilustrar a comparação entre os modelos, as Figuras 12, 13 e 14 apresentam boxplots com dados estatísticos, respectivamente, para R^2 , $RMSE$ e tempo. Analogamente, as Tabelas 14, 15 e 16 descrevem numericamente as estatísticas apresentadas nos boxplots.

Tabela 13 – Métricas de desempenho para o banco de dados de Benbouras et al. (2019)

Modelo	Métrica	Média	Desvio padrão
AutoGluon	R^2	0.1793	0.0736
	RMSE	0.0488	0.0048
	Tempo	51.2560	7.1070
FLAML	R^2	0.2004	0.0931
	RMSE	0.0482	0.0056
	Tempo	121.8494	0.8670
H2O	R^2	0.1699	0.0974
	RMSE	0.0491	0.0049
	Tempo	30.1223	6.4278

Fonte: Autor (2025)

Figura 12 – Boxplots para os resultados de R^2 de Benbouras et al. (2019)

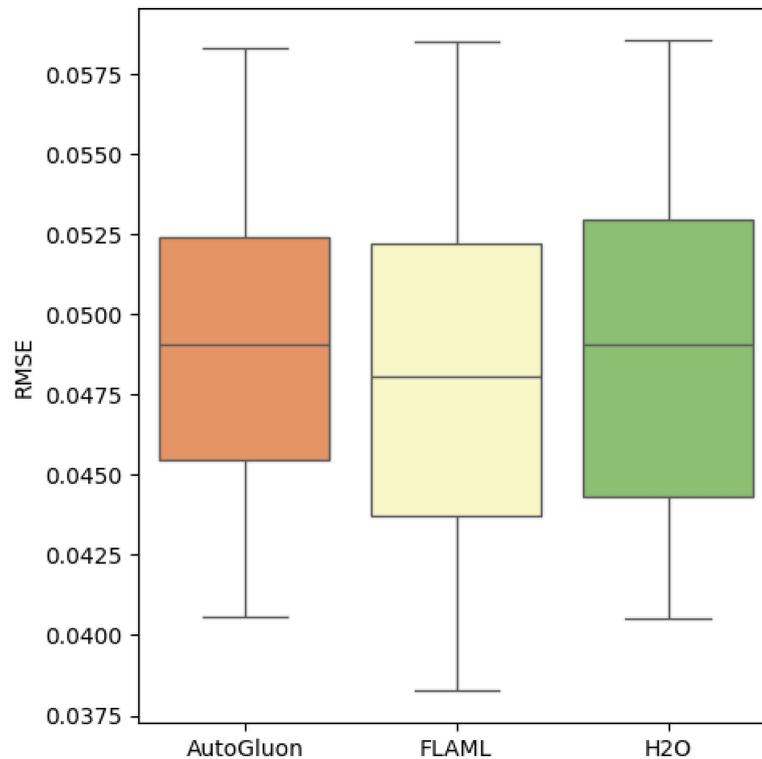
Fonte: Autor (2025)

Tabela 14 – Estatísticas de R^2 do banco de dados de Benbouras et al. (2019)

Estatística	AutoGluon	FLAML	H2O
Média	0.1793	0.2004	0.1699
Desvio padrão	0.0736	0.0931	0.0974
Mínimo	0.0383	0.0052	-0.0343
25%	0.1249	0.1260	0.1027
50%	0.1640	0.2089	0.1653
75%	0.2295	0.2737	0.2203
Máximo	0.3232	0.3718	0.3814

Fonte: Autor (2025)

De maneira geral, os valores de média e desvio padrão obtidos para R^2 foram relativamente próximos entre os modelos, sendo o FLAML o modelo com maior média, seguido do AutoGluon e do H2O. No boxplot e nas estatísticas, nota-se que o AutoGluon possui menor variabilidade, de maneira geral. O FLAML, apesar de ter menor amplitude que o H2O entre os valores mínimo e máximo, tem maior amplitude entre os quartis. Não foram observados *outliers*.

Figura 13 – Boxplots para os resultados de $RMSE$ de Benbouras et al. (2019)

Fonte: Autor (2025)

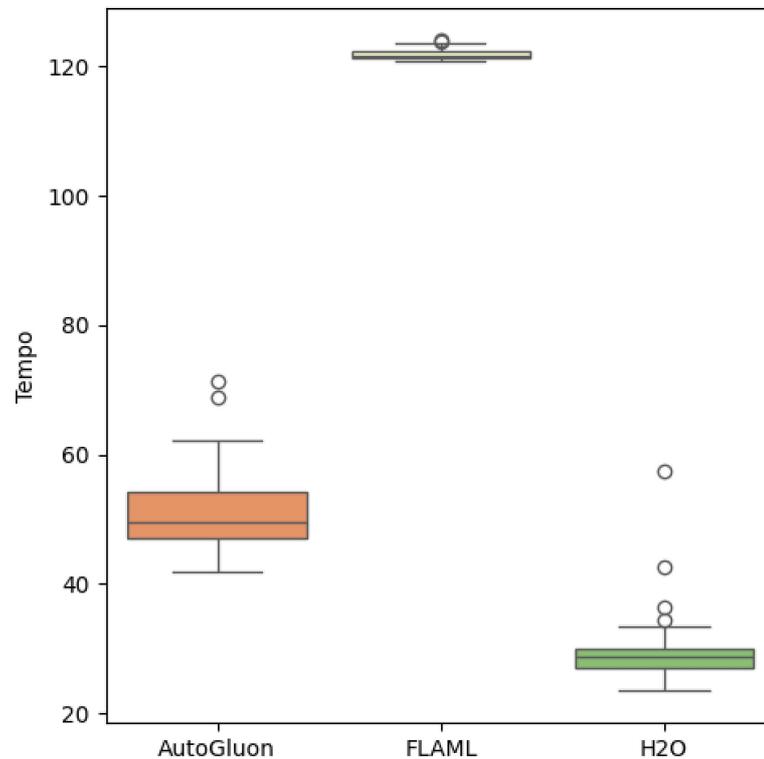
Tabela 15 – Estatísticas de *RMSE* do banco de dados de Benbouras et al. (2019)

Estatística	AutoGluon	FLAML	H2O
Média	0.0488	0.0482	0.0491
Desvio padrão	0.0048	0.0056	0.0049
Mínimo	0.0406	0.0383	0.0405
25%	0.0455	0.0437	0.0443
50%	0.0491	0.0481	0.0490
75%	0.0524	0.0522	0.0530
Máximo	0.0583	0.0585	0.0585

Fonte: Autor (2025)

A mesma proximidade de média e desvio padrão pode ser percebida para os valores de *RMSE*, sendo o FLAML o modelo com menor média, seguido do AutoGluon e do H2O. Nos boxplots e nas estatísticas, nota-se que o FLAML possui maior variação, tanto entre os quartis, quanto entre o mínimo e o máximo. O AutoGluon, apesar de ter variação entre mínimo e máximo próxima à do H2O, possui menor variação entre os quartis. Não foram observados *outliers*.

Figura 14 – Boxplots para os resultados de tempo de Benbouras et al. (2019)



Fonte: Autor (2025)

Tabela 16 – Estatísticas de tempo do banco de dados de Benbouras et al. (2019)

Estatística	AutoGluon	FLAML	H2O
Média	51.2560	121.8494	30.1223
Desvio padrão	7.1070	0.8670	6.4278
Mínimo	41.7832	120.8830	23.5334
25%	46.9554	121.2526	26.9164
50%	49.5162	121.6039	28.8363
75%	54.1555	122.2352	29.8495
Máximo	71.3553	123.9707	57.4281

Fonte: Autor (2025)

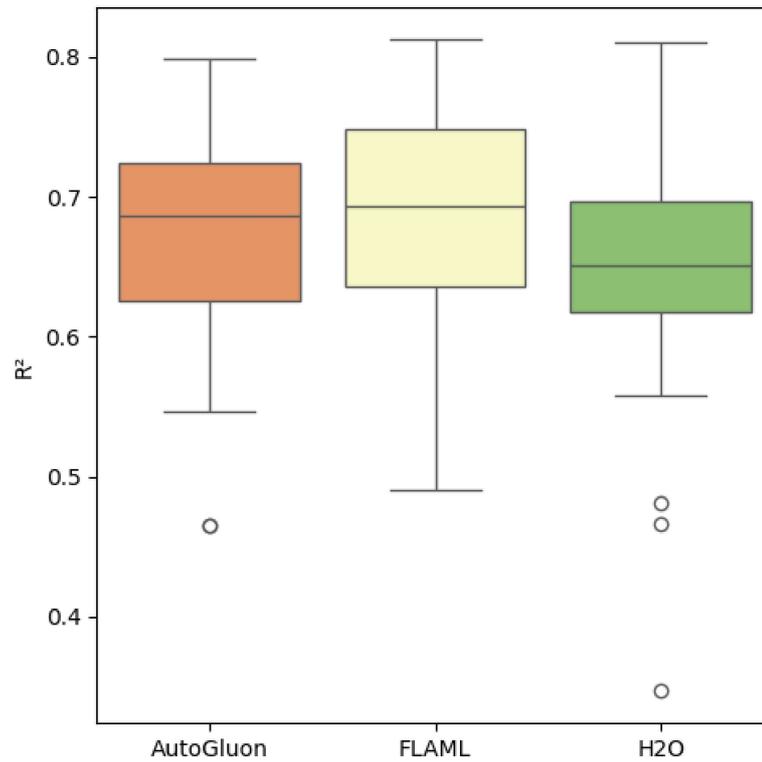
Por outro lado, observa-se uma grande diferença nos resultados do tempo de processamento, sendo o H2O o modelo com menor tempo gasto, seguido do AutoGluon e do FLAML, que apresentou valores próximos ao orçamento computacional fornecido. O FLAML apresentou menor amplitude de maneira geral, seguido do H2O e do AutoGluon. Foram observados *outliers* para os três modelos, com o H2O apresentando valores mais distantes da distribuição.

A Tabela 17 apresenta as mesmas métricas de desempenho, agora para o banco de dados de Kalantary e Kordnaei (2012), também na forma de média e desvio padrão das 30 execuções independentes. De forma análoga ao caso anterior, as Figuras 15, 16 e 17 ilustram boxplots com as estatísticas, respectivamente, para R^2 , $RMSE$ e tempo. As Tabelas 18, 19 e 20 apresentam as estatísticas representadas nos boxplots.

Tabela 17 – Métricas de desempenho para o banco de dados de Kalantary e Kordnaei (2012)

Modelo	Métrica	Média	Desvio padrão
AutoGluon	R^2	0.6688	0.0812
	RMSE	0.0434	0.0045
	Tempo	48.8981	5.0793
FLAML	R^2	0.6770	0.0869
	RMSE	0.0426	0.0032
	Tempo	122.0627	1.1404
H2O	R^2	0.6403	0.0909
	RMSE	0.0452	0.0045
	Tempo	30.5905	5.8975

Fonte: Autor (2025)

Figura 15 – Boxplots para os resultados de R^2 de Kalantary e Kordnaeij (2012)

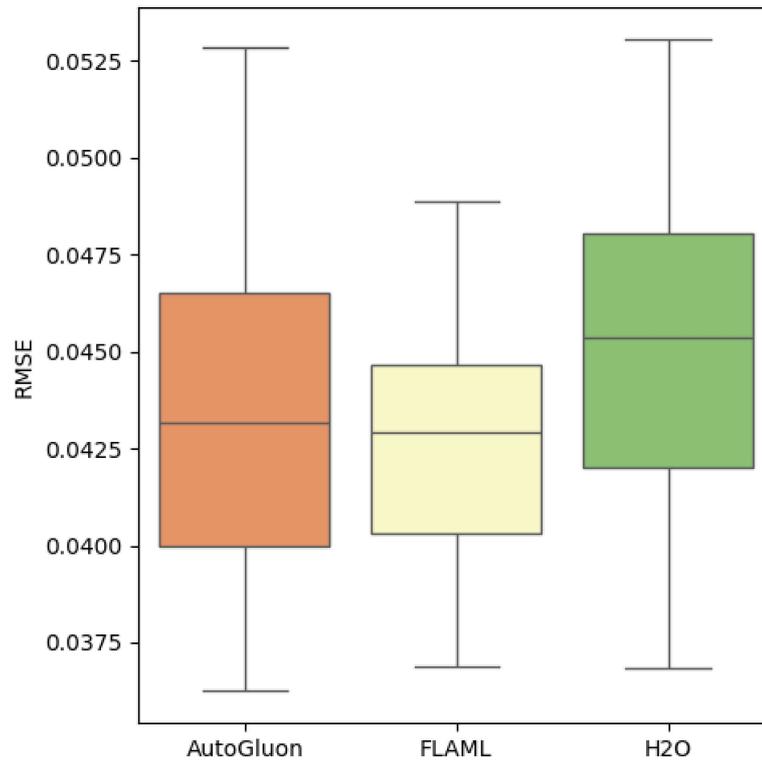
Fonte: Autor (2025)

Tabela 18 – Estatísticas de R^2 do banco de dados de Kalantary e Kordnaeij (2012)

Estatística	AutoGluon	FLAML	H2O
Média	0.6688	0.6770	0.6403
Desvio padrão	0.0812	0.0869	0.0909
Mínimo	0.4654	0.4900	0.3473
25%	0.6258	0.6359	0.6180
50%	0.6867	0.6928	0.6510
75%	0.7239	0.7484	0.6962
Máximo	0.7988	0.8121	0.8102

Fonte: Autor (2025)

Neste caso, os valores observados para R^2 também são próximos, sendo novamente a maior média apresentada pelo FLAML, seguido do AutoGluon e do H2O. A variação entre os quartis foi relativamente próxima para os três modelos. Destaca-se a presença de *outliers* para o AutoGluon e para o H2O, tendo o H2O valores consideravelmente abaixo nos valores médios.

Figura 16 – Boxplots para os resultados de $RMSE$ de Kalantary e Kordnaeij (2012)

Fonte: Autor (2025)

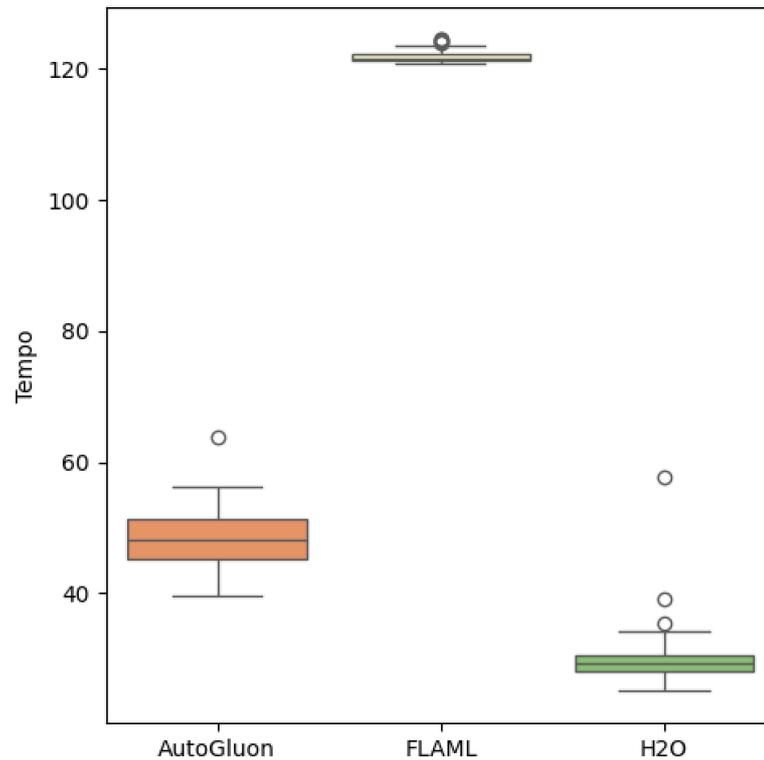
Tabela 19 – Estatísticas de $RMSE$ do banco de dados de Kalantary e Kordnaeij (2012)

Estatística	AutoGluon	FLAML	H2O
Média	0.0434	0.0426	0.0452
Desvio padrão	0.0045	0.0032	0.0045
Mínimo	0.0363	0.0369	0.0368
25%	0.0400	0.0403	0.0420
50%	0.0432	0.0429	0.0454
75%	0.0465	0.0447	0.0480
Máximo	0.0528	0.0489	0.0530

Fonte: Autor (2025)

Para o $RMSE$, o FLAML apresentou a menor média de valores, seguido do H2O e do AutoGluon. Nos boxplots e nas estatísticas, nota-se menor variabilidade de maneira geral para o FLAML, enquanto o AutoGluon e o H2O apresentam proximidade, tanto em relação aos valores mínimos e máximos, quanto em relação aos quartis. Não foram observados *outliers*.

Figura 17 – Boxplots para os resultados de tempo de Kalantary e Kordnaeij (2012)



Fonte: Autor (2025)

Tabela 20 – Estatísticas de tempo do banco de dados de Kalantary e Kordnaeij (2012)

Estatística	AutoGluon	FLAML	H2O
Média	48.8981	122.0627	30.5905
Desvio padrão	5.0793	1.1404	5.8975
Mínimo	39.6926	120.8887	25.1434
25%	45.3193	121.2803	28.0068
50%	48.1239	121.6317	29.3743
75%	51.4515	122.3483	30.6401
Máximo	63.8658	124.4557	57.7292

Fonte: Autor (2025)

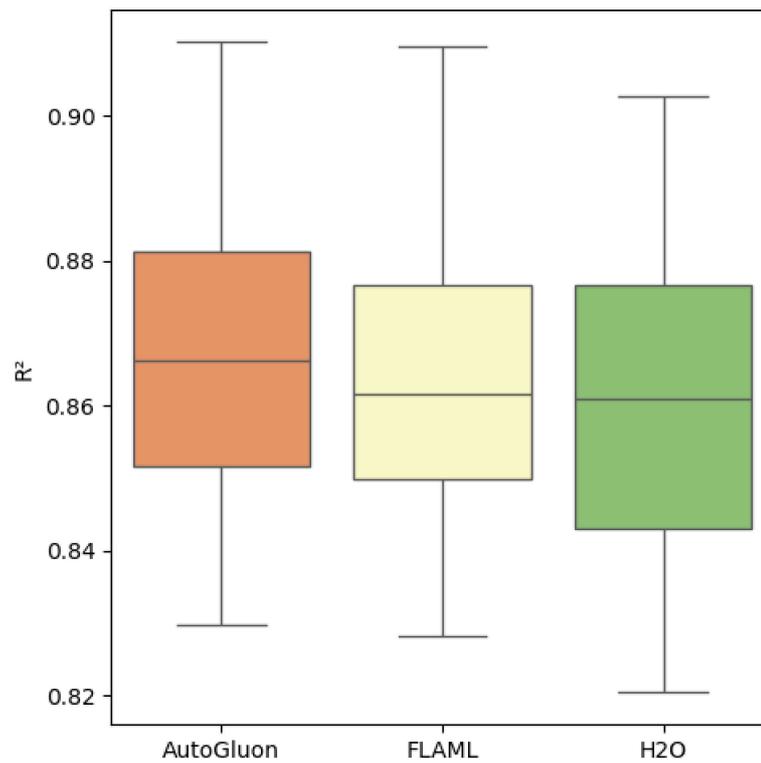
Em termos de tempo de execução, o H2O novamente apresentou o menor gasto de tempo, seguido do AutoGluon e do FLAML – que, assim como no caso anterior, se aproximou do tempo total disponível. Foram observados *outliers* nos três casos, novamente com o H2O apresentando valores que se distanciam mais da distribuição apresentada.

A Tabela 21 apresenta as métricas de desempenho para o banco de dados de Pham et al. (2024) na forma de média e desvio padrão das 30 execuções independentes. As Figuras 18, 19 e 20 ilustram boxplots com as estatísticas, respectivamente, para R^2 , $RMSE$ e tempo e as Tabelas 22, 23 e 24 descrevem numericamente as estatísticas.

Tabela 21 – Métricas de desempenho para o banco de dados de Pham et al. (2024)

Modelo	Métrica	Média	Desvio padrão
AutoGluon	R ²	0.8676	0.0206
	RMSE	0.1036	0.0089
	Tempo	54.3106	3.4887
FLAML	R ²	0.8630	0.0186
	RMSE	0.1055	0.0085
	Tempo	125.4575	5.3038
H2O	R ²	0.8603	0.0210
	RMSE	0.1065	0.0090
	Tempo	53.4534	3.5962

Fonte: Autor (2025)

Figura 18 – Boxplots para os resultados de R^2 de Pham et al. (2024)

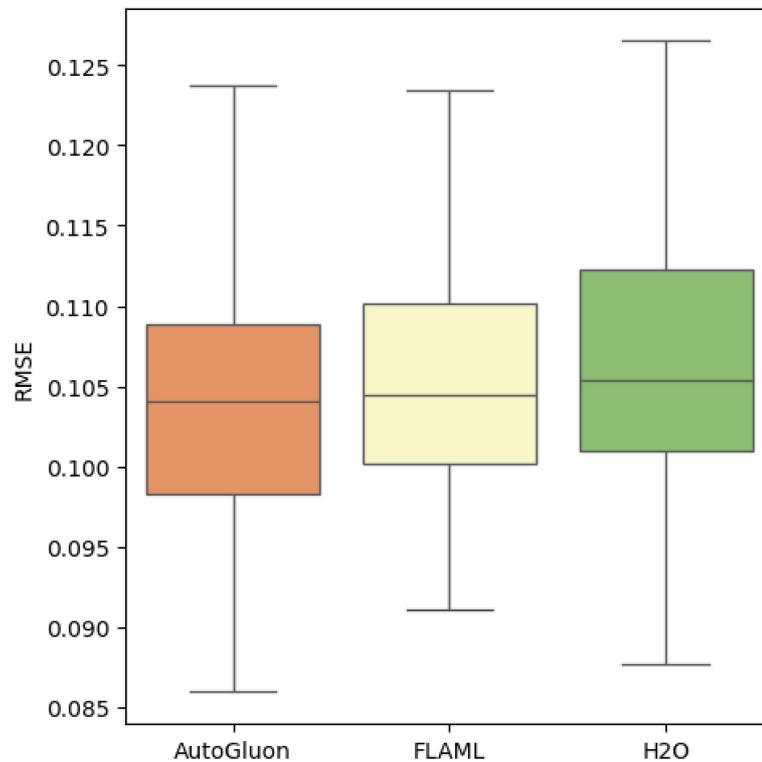
Fonte: Autor (2025)

Tabela 22 – Estatísticas de R^2 do banco de dados de Pham et al. (2024)

Estatística	AutoGluon	FLAML	H2O
Média	0.8676	0.8630	0.8603
Desvio padrão	0.0206	0.0186	0.0210
Mínimo	0.8297	0.8283	0.8205
25%	0.8516	0.8499	0.8431
50%	0.8663	0.8616	0.8611
75%	0.8813	0.8766	0.8767
Máximo	0.9102	0.9097	0.9027

Fonte: Autor (2025)

Para este banco de dados, os modelos também apresentaram médias de R^2 próximas entre si, variando a partir da terceira casa decimal. O modelo com maiores valores é o AutoGluon, seguido do FLAML e do H2O. Nos boxplots e nas estatísticas, nota-se que há menor variabilidade entre valores mínimo e máximo para o FLAML, enquanto o AutoGluon e o H2O apresentam essa amplitude de forma semelhante. Os três modelos também apresentam variações próximas entre quartis. Não foram observados *outliers*.

Figura 19 – Boxplots para os resultados de $RMSE$ de Pham et al. (2024)

Fonte: Autor (2025)

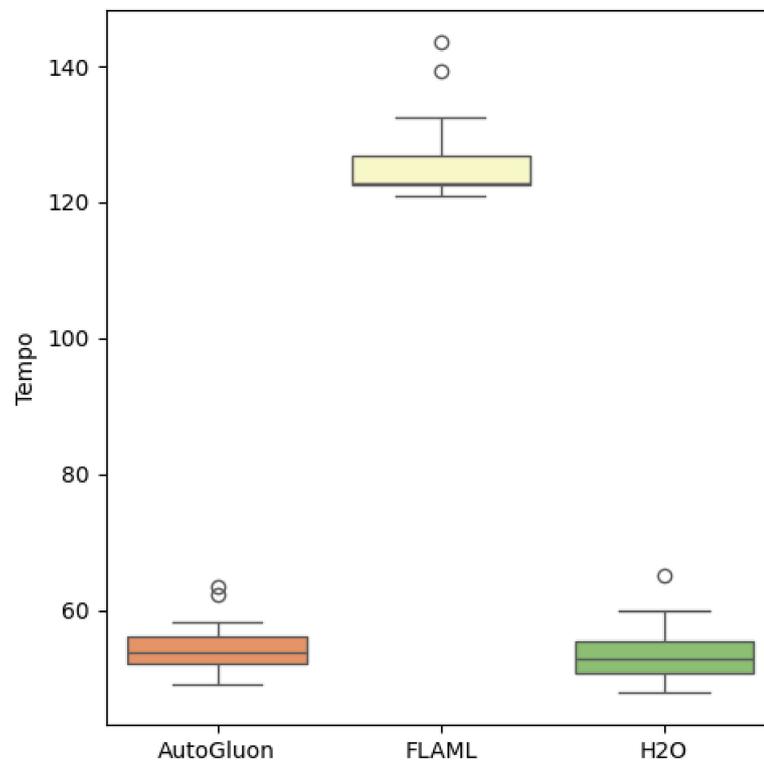
Tabela 23 – Estatísticas de *RMSE* do banco de dados de Pham et al. (2024)

Estatística	AutoGluon	FLAML	H2O
Média	0.1036	0.1055	0.1065
Desvio padrão	0.0089	0.0085	0.0090
Mínimo	0.0860	0.0911	0.0877
25%	0.0982	0.1002	0.1010
50%	0.1041	0.1044	0.1053
75%	0.1089	0.1102	0.1122
Máximo	0.1237	0.1234	0.1265

Fonte: Autor (2025)

Para o *RMSE*, os valores encontrados também foram próximos entre si, também variando a partir da terceira casa decimal. O menor valor foi apresentado pelo AutoGluon, seguido do FLAML e do H2O. A distribuição entre o primeiro e o terceiro quartil nos boxplots é relativamente semelhante entre os três modelos. A variação entre mínimo e máximo do FLAML, assim como nos casos anteriores, é menor que a dos demais, que são relativamente próximas. Não foram observados *outliers*.

Figura 20 – Boxplots para os resultados de tempo de Pham et al. (2024)



Fonte: Autor (2025)

Tabela 24 – Estatísticas de tempo do banco de dados de Pham et al. (2024)

Estatística	AutoGluon	FLAML	H2O
Média	54.3106	125.4575	53.4534
Desvio padrão	3.4887	5.3038	3.5962
Mínimo	48.9976	121.0133	47.9250
25%	52.1295	122.5110	50.8337
50%	53.8842	122.7970	52.8049
75%	56.0631	126.8746	55.4175
Máximo	63.4668	143.4789	65.0458

Fonte: Autor (2025)

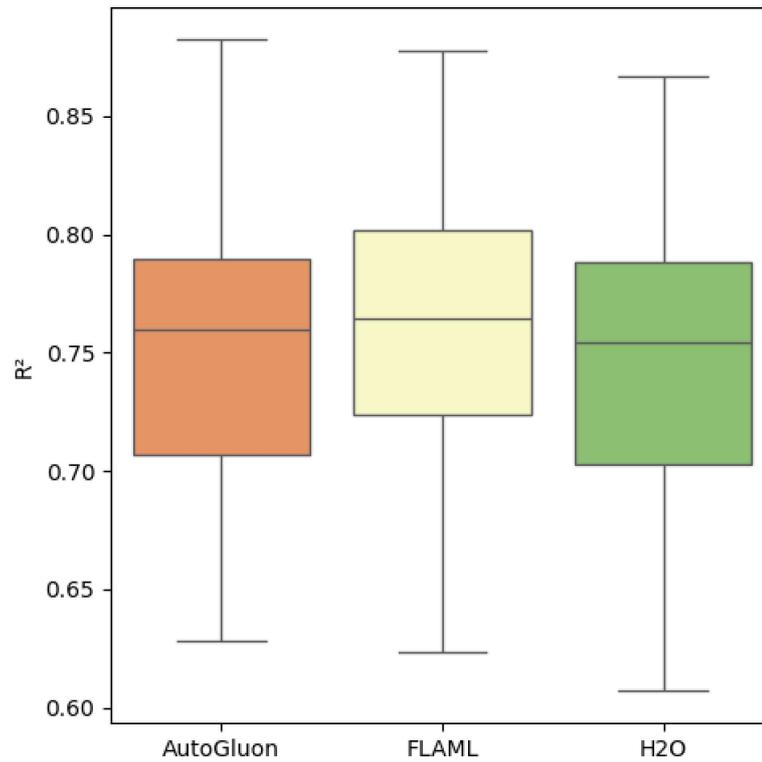
Em relação ao tempo de execução, o H2O apresentou a menor média de tempo gasto, seguido por uma diferença relativamente pequena do AutoGluon e por uma diferença relativamente maior do FLAML – que novamente consumiu valores próximos ao orçamento computacional em suas execuções. Foram observados *outliers* para os três casos, com o FLAML dessa vez apresentando valores mais distantes de sua distribuição.

A Tabela 25 apresenta as métricas para o banco de dados de Uzer (2024), também na forma de média e desvio padrão das 30 execuções independentes. As Figuras 21, 22 e 23 apresentam boxplots com dados estatísticos, respectivamente, para R^2 , $RMSE$ e tempo. As Tabelas 26, 27 e 28, respectivamente, descrevem as estatísticas ilustradas.

Tabela 25 – Métricas de desempenho para o banco de dados de Uzer (2024)

Modelo	Métrica	Média	Desvio padrão
AutoGluon	R^2	0.7493	0.0621
	RMSE	0.0735	0.0094
	Tempo	49.9298	4.5789
FLAML	R^2	0.7623	0.0616
	RMSE	0.0716	0.0096
	Tempo	122.0263	1.0720
H2O	R^2	0.7460	0.0649
	RMSE	0.0740	0.0096
	Tempo	37.0519	4.4360

Fonte: Autor (2025)

Figura 21 – Boxplots para os resultados de R^2 de Uzer (2024)

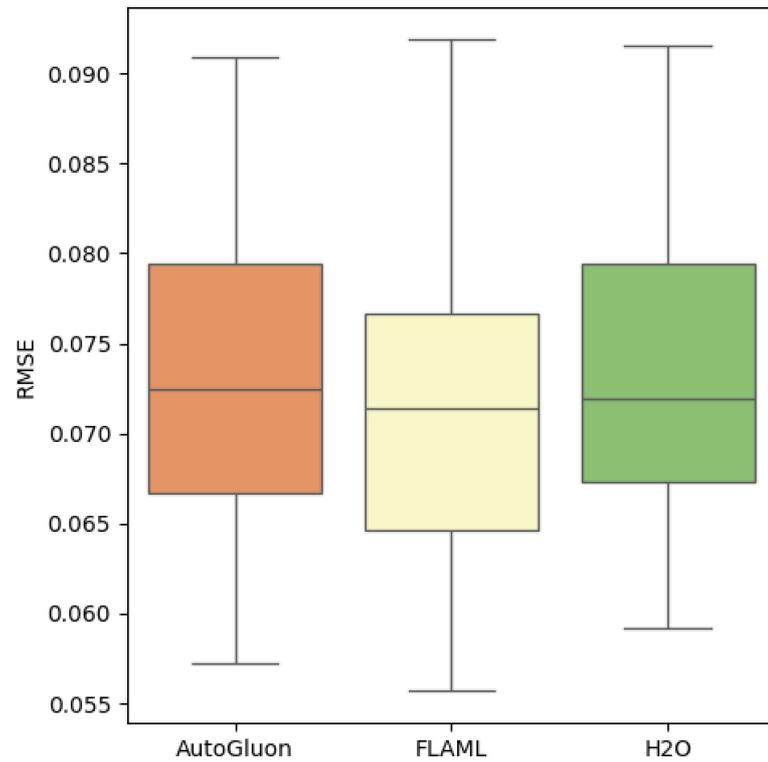
Fonte: Autor (2025)

Tabela 26 – Estatísticas de R^2 do banco de dados de Uzer (2024)

Estatística	AutoGluon	FLAML	H2O
Média	0.7493	0.7623	0.7460
Desvio padrão	0.0621	0.0616	0.0649
Mínimo	0.6279	0.6237	0.6072
25%	0.7068	0.7237	0.7028
50%	0.7595	0.7648	0.7546
75%	0.7898	0.8019	0.7884
Máximo	0.8823	0.8780	0.8668

Fonte: Autor (2025)

Para os valores de R^2 , assim como nos casos anteriores, também há uma proximidade entre os valores médios, sendo o FLAML o modelo com maior média, seguido do AutoGluon e do H2O. A amplitude de variação é relativamente próxima para os três modelos, tanto em relação aos valores mínimos e máximos, quanto em relação aos quartis. Não foram observados *outliers*.

Figura 22 – Boxplots para os resultados de $RMSE$ de Uzer (2024)

Fonte: Autor (2025)

Tabela 27 – Estatísticas de $RMSE$ do banco de dados de Uzer (2024)

Estatística	AutoGluon	FLAML	H2O
Média	0.0735	0.0716	0.0740
Desvio padrão	0.0094	0.0096	0.0096
Mínimo	0.0572	0.0557	0.0592
25%	0.0667	0.0646	0.0673
50%	0.0725	0.0714	0.0719
75%	0.0794	0.0766	0.0794
Máximo	0.0909	0.0919	0.0916

Fonte: Autor (2025)

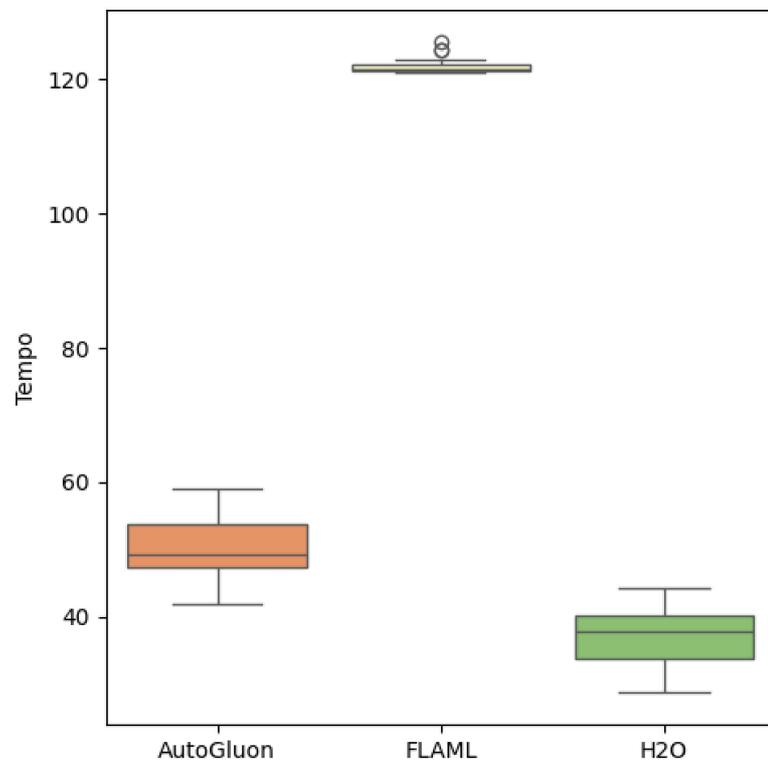
Para o $RMSE$, o FLAML também apresentou menores valores de média, seguido do AutoGluon e do H2O. As variações entre mínimo e máximo, assim como entre primeiro e terceiro quartis, são relativamente próximas para os três modelos e não foram observados *outliers*.

Tabela 28 – Estatísticas de tempo do banco de dados de Uzer (2024)

Estatística	AutoGluon	FLAML	H2O
Média	49.9298	122.0263	37.0519
Desvio padrão	4.5789	1.0720	4.4360
Mínimo	41.9108	120.9860	28.7307
25%	47.3233	121.3888	33.7994
50%	49.2126	121.6460	37.7792
75%	53.7124	122.1726	40.2688
Máximo	58.9462	125.5173	44.2709

Fonte: Autor (2025)

Figura 23 – Boxplots para os resultados de tempo de Uzer (2024)



Fonte: Autor (2025)

Quanto ao tempo de processamento, o comportamento dos modelos é semelhante aos casos anteriores, tendo o H2O o menor custo computacional, seguido do AutoGluon e do FLAML, que novamente teve valores próximos do orçamento computacional disponível. Foram observados *outliers* apenas para o H2O e os valores são próximos da distribuição.

A Tabela 29 organiza um resumo dos resultados encontrados, dividindo por banco de dados e classificando o melhor modelo de acordo com o maior valor médio de R^2 , o menor valor médio de $RMSE$ e o menor valor médio de tempo de execução. Com base nos

resultados apresentados anteriormente e nas respectivas discussões, é possível perceber que os modelos apresentaram valores médios de R^2 e $RMSE$ próximos entre si. Observa-se também que o FLAML foi considerado o melhor modelo de forma consistente para essas métricas, com exceção de R^2 e de $RMSE$ para os dados de Pham et al. (2024). Por outro lado, o tempo de execução apresenta diferenças bem delimitadas. O H2O, também de forma consistente, apresentou o menor tempo de execução para todos os modelos, sendo coerente com a teoria apresentada no Item 3.3 deste trabalho, a respeito da preocupação ao desenvolver um modelo para uso de produção com menor custo computacional. O FLAML também se mostrou coerente com a teoria apresentada no Item 3.2 deste trabalho, uma vez que o modelo usou todo o orçamento disponível a fim de encontrar os melhores ajustes – tendo sido bem-sucedido nos casos apresentados.

Tabela 29 – Tabela resumo dos resultados

Banco de dados	Métrica	Melhor modelo
Benbouras et al. (2019)	R^2	FLAML
	RMSE	FLAML
	Tempo	H2O
Kalantary e Kordnaeij (2012)	R^2	FLAML
	RMSE	FLAML
	Tempo	H2O
Pham et al. (2024)	R^2	AutoGluon
	RMSE	AutoGluon
	Tempo	H2O
Uzer (2024)	R^2	FLAML
	RMSE	FLAML
	Tempo	H2O

Fonte: Autor (2025)

Em relação à capacidade de ajuste dos modelos a cada banco de dados, percebe-se que os modelos apresentaram, de maneira geral, os maiores valores de R^2 para o de Pham et al. (2024), seguido de Uzer (2024), Kalantary e Kordnaeij (2012) e Benbouras et al. (2019). Por outro lado, os modelos apresentaram, de maneira geral, menores valores de $RMSE$ para o de Kalantary e Kordnaeij (2012), seguido de Benbouras et al. (2019), Uzer (2024) e Pham et al. (2024). Pode-se dizer, então, que não há uma convergência em relação a um banco de dados apresentar melhores ajustes que outros.

É importante ressaltar também que os modelos apresentaram *outliers* apenas nas estatísticas de tempo de execução, com exceção dos resultados de R^2 para o banco de Kalantary e Kordnaeij (2012). Tais informações podem indicar uma confiabilidade dos modelos em termos de acurácia nas estimativas. Por outro lado, tais resultados também indicam menor consistência em relação ao consumo de orçamento computacional.

6 Conclusões e trabalhos futuros

Neste capítulo são apresentadas conclusões acerca da pesquisa desenvolvida, no Item 5.1, e são sugeridos trabalhos futuros para sua continuação, no Item 5.2.

6.1 Conclusões

Foram selecionados quatro bancos de dados com mínimo de 300 amostras. O critério de decisão para este tamanho foi definido após tentativas anteriores de treinar os modelos com bancos de dados menores e o recebimento de avisos do H2O de que era necessário um mínimo de 200 amostras na etapa de teste para que ele pudesse apresentar métricas de desempenho satisfatórias. Os conjuntos selecionados contêm amostras de solos de diferentes lugares do mundo – Argélia, China, Irã e Turquia – e apresentaram uma grande diferença em número e em tipos de variáveis. Para fins de comparação, foi feita então uma padronização desses dados e foram escolhidos para pré-processamento e análise de correlação o teor de umidade natural, o limite de liquidez, o índice de plasticidade e o índice de vazios inicial. Com base no coeficiente de correlação de Pearson, foram selecionadas as variáveis de entrada. O índice de plasticidade foi excluído, por ser a variável menos correlacionada com o índice de compressão e por haver outra variável também relacionada à consistência dos solos com maior coeficiente de correlação.

Os algoritmos usados para este trabalho foram o AutoGluon, o FLAML e o H2O, que apresentam diferentes estratégias de busca para encontrar os melhores modelos. A fim de comparar os desempenhos, suas características comuns foram padronizadas: algoritmos usados, orçamento computacional, validação cruzada e uso de *stacking*. Para melhor aproveitamento das amostras disponíveis, foram feitas 30 execuções independentes, variando a seleção de amostras por meio de controle de semente aleatória. As métricas de desempenho escolhidas para análise foram R^2 , $RMSE$ e tempo de execução.

De maneira geral, percebeu-se que o FLAML apresentou os melhores valores de R^2 e de $RMSE$ para todos os bancos de dados, com exceção de um deles, que teve o AutoGluon como melhor modelo para essas métricas. Em relação ao tempo de execução, o H2O apresentou os menores valores em todas as análises, enquanto o FLAML apresentou o maior tempo, próximo ao valor total disponível. Os resultados são coerentes com as características apresentadas sobre os modelos no Capítulo 3, em que se destaca que o FLAML faz uso de todo o tempo fornecido para encontrar um modelo com maior capacidade de ajuste, enquanto o H2O tem como um de seus objetivos principais encontrar um modelo que tenha o menor custo computacional possível.

Conclui-se que, apesar das diferentes estratégias de busca dos modelos, os três modelos apresentam métricas de desempenho próximas sob condições semelhantes. En-

tretanto, é importante considerar as prioridades do usuário. Com base nos resultados obtidos, pode-se dizer que o FLAML tem maior probabilidade de apresentar métricas mais satisfatórias, porém consumindo todo o orçamento computacional disponível. Por outro lado, o H2O tem maior probabilidade de apresentar métricas competitivas em um tempo de execução menor que o máximo fornecido. Por fim, pode-se dizer também que o AutoGluon é capaz de oferecer resultados com métricas de desempenho e tempo de execução competitivos, sem necessariamente se destacar positiva ou negativamente, quando comparado com os demais.

6.2 Trabalhos futuros

Ao longo do desenvolvimento deste trabalho, contemplou-se a possibilidade de expansão das análises, além de outras questões a serem investigadas acerca deste mesmo objeto de pesquisa. Portanto, sugere-se para trabalhos futuros:

- Analisar os mesmos bancos de dados usando outros modelos de aprendizado de máquina automatizado e comparar com os resultados obtidos neste trabalho, a fim de reforçar os resultados obtidos ou apresentar modelos com desempenhos melhores;
- Analisar outros bancos de dados usando estes mesmos modelos de aprendizado de máquina automatizado e endossar a análise realizada, a fim de se executar um benchmark mais complexo;
- Comparar os resultados obtidos pelos modelos de aprendizado de máquina automatizado com resultados obtidos pelas equações empíricas apresentadas, a fim de comparar a eficiência de ambas as metodologias;
- Por fim, como constatado que os modelos são desenvolvidos com base em grupos específicos de solos, sugere-se que seja aplicada esta mesma metodologia para bancos de dados compostos por solos brasileiros.

REFERÊNCIAS

- AHADIYAN, J.; EBNE, J. R.; BAJESTAN, M. S. Prediction determination of soil compression index, C_c , in Ahwaz region. **Journal of Geotechnical and Geoenvironmental Engineering**, v. 134, n. 10, p. 1488–1497, 2008.
- AKAYULI, C. A.; OFOSU, B. Empirical Model for Estimating Compression Index from Physical Properties of Weathered Birimian Phyllites. **Electronic Journal of Geotechnical Engineering**, v. 18, p. 6135–6144, 2013.
- AL-KHAFAJI, A. W.; ANDERSLAND, O. B. Equations for Compression Index Approximation. **Journal of Geotechnical Engineering**, v. 118, n. 1, 1992.
- ASTM INTERNATIONAL. **D2435**: Standard Test Methods for One-Dimensional Consolidation Properties of Soils Using Incremental Loading. West Conshohocken, Estados Unidos, 2020.
- AZZOUZ, A. S.; KRIZEK, R. J.; COROTIS, R. B. Regression Analysis of Soil Compressibility. **Soils and Foundations**, v. 16, n. 2, p. 19–29, 1976.
- BELLO, A. A.; OWOSENI, J. O.; FATOYINBO, I. O. Evaluation of plasticity and consolidation characteristics of migmatite–gneiss-derived laterite soils. **SN Applied Sciences**, v. 1, p. 1-11, 2019.
- BENBOURAS, M. A.; MITICHE, R. K.; ZEDIRA, H.; PETRISOR, A. I.; MEZOUAR, N.; DEIBICHE, F. A new approach to predict the compression index using artificial intelligence methods. **Marine Georesources & Geotechnology**, v. 37, n. 6, p. 285–300, 2019.
- BICALHO, K. V.; MORESCHI, M.; KUSTER, J. S. H.; OLIVEIRA FILHO, A. G.; HISATUGU, W. Estudo do uso da Técnica de Redes Neurais Artificiais em Engenharia Geotécnica. In: XX Congresso Brasileiro de Mecânica dos Solos e Engenharia Geotécnica, Campinas, Brasil, 2022. **Anais...** Campinas: Galoá, 2022, p. 434–447.
- BOWLES, J. E. **Physical and Geotechnical Properties of Soils**. Nova York, Estados Unidos: McGraw-Hill Book Company, 1989.
- BRUCE, P.; BRUCE, A. **Estatística Prática para Cientistas de Dados - 50 Conceitos Essenciais**. Rio de Janeiro, Brasil: Alta Books, 2019.
- CAPUTO, H. P. **Mecânica dos Solos e Suas Aplicações**. 6. Ed. Rio de Janeiro, Brasil: LTC, 2006.
- COZZOLINO, V. M. Statistical forecasting of compression index. In: 5th International Conference on Soil Mechanics, 1961. **Anais...** p. 51–53.
- DAS, B. M. **Fundamentos de Engenharia Geotécnica**. 7. Ed. São Paulo, Brasil: Cengage Learning, 2011.
- DWAY, S. M. M.; THANT, D. A. A. Soil Compression Index Prediction Model for Clayey Soils. **International Journal of Science, Engineering and Technology Research**, v. 3, n. 11, p. 2458–2462, 2014.

- ERICKSON, N.; MUELLER, J.; SHIRKOV, A.; ZHANG, H.; LARROY, P.; LI, M.; SMOLA, A. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. **arXiv preprint arXiv: 2003.05605**, 2020.
- GERÓN, A. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn, Keras e TensorFlow**. 2. Ed. Rio de Janeiro, Brasil: Alta Books, 2021.
- GUNDUZ, Z.; ARMAN, H. Possible relationships between compression and recompression indices of a low-plasticity clayey soil. **Arabian Journal for Science and Engineering**, v. 32, n. 2, p. 1709–190, 2007.
- HERRERO, O. R. Universal compression Index Equation. **Journal of Geotechnical Engineering Division of the American Society of Civil Engineers**, v. 109, n. 10, p. 1179–1200, 1983.
- HOUGH, B. K. **Basic Soils Engineering**. Nova York, Estados Unidos: The Ronald Press Company, 1957.
- HUTTER, F.; KOTTHOFF, L.; VANSCHOREN, J. Automated Machine Learning: Methods, Systems, Challenges. Cham, Suíça: Springer Nature, 2019.
- INSTITUTO DE PESQUISA ECONÔMICA APLICADA. Agenda 2030 - ODS: Metas Nacionais dos Objetivos de Desenvolvimento Sustentável. Disponível em: https://portalantigo.ipea.gov.br/agencia/index.php?option=com_content&view=article&id=33895&Itemid=433. Acesso em: 27 fev. 2025.
- JAIN, V. K.; DIXIT, M.; CHITRA, R. Correlation of Plasticity Index and Compression Index of Soil. **International Journal of Innovative Engineering and Technology**, v. 5, p. 263–270, 2015.
- KALANTARY, F.; KORDNAEIJ, A. Prediction of compression index using artificial neural network. **Scientific Research and Essays**, v. 7, n. 31, p. 2835–2848, 2012.
- KOOTAHI, K.; MORADI, G. Evaluation of Compression Index of Marine Fine-Grained Soils by the Use of Index Tests. **Marine Georesources and Geotechnology**, v. 35, p. 548–570, 2017.
- KOPPULA, S. D. Statistical Estimation of Compression Index. **Geotechnical Testing Journal**, v. 4, n. 2, p. 68–73, 1981.
- LASKAR, A.; PAL, S. K. Geotechnical Characteristics of Two Different Soils and Their Mixture and Relationships between Parameters. **Electronic Journal of Geotechnical Engineering**, v. 17, p. 2821–2832, 2012.
- LEDELL, E.; POIRIER, S. H2O AutoML: Scalable Automatic Machine Learning. In: 7th ICML Workshop on Automated Machine Learning, 2020. **Anais...**
- MAYNE, P. W. Cam-Clay Predictions of Undrained Strength. **Journal of Geotechnical Engineering Division**, v. 106, n. 6, p. 1219–1242, 1980.
- NACCI, V. A.; WANG, M.; DEMARS, K. Engineering Behavior of Calcareous Soils. In: Civil Engineering in the Oceans III, ASCE Specialty Conference, 1975. **Anais...** p. 9–12.

NAGARAJ, T. S.; MURTHY, B. R. S. Prediction of the Preconsolidation Pressure and Compression Index of Soils. **Geotechnical Testing Journal**, v. 8, n. 4, p. 199–202, 1985.

NESAMATHA, R.; ARUMAIRAJ, P. Numerical Modeling for Prediction of Compression Index from Soil Index Properties. **Electronic Journal of Geotechnical Engineering**, v. 20, p. 4369–4378, 2015.

NISHIDA, Y. A Brief Note on Compression Index of Soils. **Journal of the Soil Mechanics and Foundations Division**, v. 82, n. 3, p. 1–14, 1956.

PARK, H. I.; LEE, S. R. Evaluation of the compression index of soils using an artificial neural network. **Computers and Geotechnics**, v. 38, n. 4, p. 472–481, 2011.

PHAM, K.; NGUYEN, K.; LIM, K.; KIM, Y.; CHOI, H. A generalized formula for predicting soil compression index using multi-evolutionary algorithm. **Engineering Geology**, v. 343, 2024.

PINTO, C. S. **Curso Básico de Mecânica dos Solos em 16 Aulas**. 3. Ed. São Paulo, Brasil: Oficina de Textos, 2006.

RASHED, K.; SALIH, N.; ABDALLA, T. Correlation of Consistency and Compressibility Properties of Soils in Sulaimani City. **Sulaimani Journal of Engineering Sciences**, v. 4, p. 146–155, 2017.

SALIH, N. B. Geotechnical Characteristics Correlations for Fine-Grained Soils. **IOP Conference Series: Materials Science and Engineering**, v. 737, n. 1, 2020.

SHAIKH, M.; AHSAN, K.; ALI MOLLA, K. Development of Strength and Compressibility Correlations of Cohesive Soils of Some Regions in Khulna City. **International Journal of Advanced Structural and Geotechnical Engineering**, v. 3, p. 242–245, 2014.

SOLANKI, C. Quick Computation of Settlement for Shallow Foundations of Alluvial Deposits. In: do International Conference on Chemical, Civil and Environment Engineering, Dubai, Emirados Árabes Unidos, 2012. **Anais...** p. 24–25.

SOWER, G. B. **Introductory Soil Mechanics and Foundation**. 3. Ed. Londres, Reino Unido: The Macmillan Company of Collier-Macmillan Ltd., 1970.

SRIDHARAN, A.; NAGARAJ, H. B. Compressibility behaviour of remoulded, fine-grained soils and correlation with index properties. **Canadian Geotechnical Journal**, v. 37, p. 712–722, 2000.

TERZAGHI, K.; PECK, R. B. **Soil Mechanics in Engineering Practice**. 2. Ed. Nova York, Estados Unidos: Wiley, 1967.

TIWARI, B.; AJMERA, B. New Correlation Equations for Compression Index of Remolded Clays. **Journal of Geotechnical and Geoenvironmental Engineering**, v. 138, n. 6, p. 757–762, 2012.

UZER, A. U. Accurate Prediction of Compression Index of Normally Consolidated Soils Using Artificial Neural Networks. **Buildings**, v. 14, n. 9, 2024.

VINOD, P.; BINDU, J. Compression Index of Highly Plastic Clays — an Empirical Correlation. **Indian Geotechnical Journal**, v. 40, n. 3, p. 174–180, 2010.

WANG, C.; WU, Q.; WEIMER, M.; ZHU, E. FLAML: A Fast and Lightweight AutoML Library. In: 4th Conference on Machine Learning and Systems (MLSys), 2021. **Anais...** p. 434–447.

WROTH, C. P.; WOOD, D. M. The Correlation of Index Properties with Some Basic Engineering Properties of Soils. **Canadian Geotechnical Journal**, v. 15, p. 137–145, 1978.

YOON, G. L. et al. Empirical Correlations of Compression Index for Marine Clay from Regression Analysis. **Canadian Geotechnical Journal**, v. 41, n. 6, p. 1213–1221, 2004.

YOON, G. L.; KIM, B. T. Regression analysis of compression index for Kwangyang marine clay. **Journal of Civil Engineering**, v. 10, n. 6, p. 415–418, 2006.