UNIVERSIDADE FEDERAL DA PARAÍBA
Programa de Pós-Graduação em Informática

Aprendizagem de Máquina para Detecção de Áreas de Aluvião: Um Estudo no Semiárido Brasileiro

Daniel Baptista Vio

João Pessoa - PB

2025

UNIVERSIDADE FEDERAL DA PARAÍBA
Programa de Pós-Graduação em Informática

Daniel Baptista Vio

Aprendizagem de Máquina para Detecção de Áreas de Aluvião: Um Estudo no Semiárido Brasileiro

Dissertação apresentada à Universidade Federal da Paraíba como parte dos requisitos para obtenção do título de Mestre em Informática

Orientador: Prof. Dr. Gustavo Henrique Matos Bezerra Motta

Segundo Orientador: Prof. Dr. Jonas Otaviano Praça de Souza

Coorientador: Prof. Dr. Leandro Carlos de Souza

João Pessoa - PB

2025

Catalogação na publicação
Seção de Catalogação e Classificação

V795a Vio, Daniel Baptista.
Aprendizagem de máquina para detecção de áreas de aluvião : um estudo no semiárido brasileiro / Daniel Baptista Vio. - João Pessoa, 2025.
52 f. : il.

Orientação: Gustavo Henrique Matos Bezerra Motta, Jonas Otaviano Praça de Souza.
Coorientação: Leandro Carlos de Souza.
Dissertação (Mestrado) - UFPB/CI.

1. Inteligência artificial - Aprendizado de máquina. 2. Floresta aleatória. 3. Depósitos aluvionares. 4. Regiões semiáridas. 5. Redução de dados. I. Motta, Gustavo Henrique Matos Bezerra. II. Souza, Jonas Otaviano Praça de. III. Souza, Leandro Carlos de. IV. Título.

UFPB/BC CDU 004.8(043)

Elaborado por Walqueline da Silva Araújo - CRB-15/514

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Ata da Sessão Pública de Defesa de Dissertação de Mestrado de **DANIEL BAPTISTA VIO**, candidato ao título de Mestre em Informática na área de Ciências da Computação, realizada em 26 de fevereiro de 2025.

Aos vinte e seis dias do mês de fevereiro do ano de dois mil e vinte e cinco, às 9h, no Centro de Informática, reuniram-se os membros da Banca Examinadora constituída para julgar o Trabalho Final do discente **DANIEL BAPTISTA VIO**, vinculado a Universidade Federal da Paraíba sob a matrícula nº 20231003810, candidato ao grau de Mestre em Informática, na área de "Ciências da Computação", na linha de pesquisa "Metodologia e Técnicas de Computação" do Programa de Pós-Graduação em Informática. A comissão examinadora foi composta pelos professores Gustavo Henrique Matos Bezerra Motta, orientador e presidente da banca; Jonas Otaviano Praca de Souza, segundo orientador; Leandro Carlos de Souza, coorientador; Clauirton de Albuquerque Siebra, examinador interno ao programa; e Osmar Abílio de Carvalho Junior, examinador externo à Instituição. Dando início aos trabalhos, o Presidente da Banca cumprimentou os presentes, comunicou a finalidade da reunião e passou a palavra ao candidato para que fizesse, oralmente, a exposição do trabalho de dissertação intitulado "**Aprendizagem de Máquina para Detecção de Áreas de Aluvião: Um Estudo no Semiárido Brasileiro"**. Concluída a exposição, o candidato foi arguido pela Banca Examinadora, que emitiu o seguinte parecer: "***aprovado***". Do ocorrido, eu, Gean Paulo P. M. de Barros, secretário, lavrei a presente ata que vai assinada por mim e pelos membros da Banca Examinadora. João Pessoa, 26 de fevereiro de 2025.

Gean Paulo P. M. de Barros
Secretário - SIAPE 2326476

Prof. Dr. Gustavo Henrique Matos Bezerra Motta
Orientador (PPGI)

Documento assinado digitalmente
gov.br
GUSTAVO HENRIQUE MATOS BEZERRA MOTTA
Data: 26/02/2025 15:16:29-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Jonas Otaviano Praca de Souza
Segundo orientador (UFPB)

Documento assinado digitalmente
gov.br
JONAS OTAVIANO PRACA DE SOUZA
Data: 27/02/2025 11:20:50-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Leandro Carlos de Souza
Coorientador (PPGI)

Documento assinado digitalmente
gov.br
LEANDRO CARLOS DE SOUZA
Data: 26/02/2025 18:30:01-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Clauirton de Albuquerque Siebra
Examinador interno ao Programa (PPGI)

Documento assinado digitalmente
gov.br
CLAUIRTON DE ALBUQUERQUE SIEBRA
Data: 26/02/2025 19:29:47-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Osmar Abílio de Carvalho Junior
Examinador externo à Instituição (UnB)

Documento assinado digitalmente
gov.br
OSMAR ABILIO DE CARVALHO JUNIOR
Data: 26/02/2025 21:13:02-0300
Verifique em https://validar.iti.gov.br

ABSTRACT

In semi-arid regions, where water resources are scarce, the development of precise methods for their identification and mapping is pertinent. This study aimed to investigate the use of artificial intelligence techniques to map alluvial deposits in the Riacho Grande basin, Pernambuco. To this end, an extensive, high-resolution, georeferenced database was employed, comprising eleven geomorphological and hydrological dimensions for the characterisation of alluvial areas. Initially, the database, containing over 315 million points distributed across the Basin, underwent an optimisation process using the extended Geometric Markovian Diffusion (e-GMD) method, with the aim of reducing the data volume at different levels without compromising its representativeness. Subsequently, supervised machine learning models (KNN, CART/DT, and RF) were constructed and evaluated for the detection of alluvial areas. Each model was trained and validated using 10-fold cross-validation, and its hyperparameters were tuned to optimise performance. The results demonstrated that the Random Forest (RF) algorithm excelled, achieving an F1-score of 89.8% using only 5% of the original data, outperforming KNN (86.1%) and CART/DT (86.3%). This study demonstrates the potential of employing artificial intelligence techniques for the effective mapping of alluvial deposits in semi-arid regions.

Keywords: Random Forest, alluvial deposits, machine learning, semi-arid regions, data reduction.

RESUMO

Em regiões semiáridas, onde os recursos hídricos são escassos, é relevante o desenvolvimento de métodos precisos para sua identificação e mapeamento. Este estudo teve como objetivo investigar o uso de técnicas de inteligência artificial para mapear depósitos aluvionares na bacia do Riacho Grande, em Pernambuco. Para isso, utilizou-se um extenso banco de dados georreferenciado de alta resolução, composto por onze dimensões geomorfológicas e hidrológicas para a caracterização de áreas aluvionares. Inicialmente, o banco de dados, com mais de 315 milhões de pontos distribuídos pela Bacia, passou por um processo de otimização utilizando o método Difusão Geométrica Markoviana estendido (DGM-e), visando reduzir o volume de dados, em diferentes patamares, sem comprometer sua representatividade. Posteriormente, foram construídos e avaliados modelos de aprendizado de máquina supervisionado (KNN, CART/DT e RF) para detecção de áreas aluvionares. Cada modelo foi treinado e validado utilizando a técnica de validação cruzada com 10 folds, e seus hiperparâmetros foram ajustados para otimizar o desempenho. Os resultados demonstraram que o algoritmo Random Forest (RF) se destacou, alcançando um F1-score de 89,8% utilizando apenas 5% dos dados originais, superando o KNN (86,1%) e o CART/DT (86,3%). Este estudo aponta o potencial do uso de técnicas de inteligência artificial para o mapeamento eficaz de depósitos aluvionares em regiões semiáridas.

Palavras-chave: Floresta Aleatória, depósitos aluvionares, aprendizado de máquina, regiões semiáridas, redução de dados.

AGRADECIMENTOS

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro que viabilizou esta pesquisa.

Meus sinceros agradecimentos aos meus orientadores, Prof. Dr. Gustavo Henrique Matos Bezerra Motta, Prof. Dr. Jonas Otaviano Praça de Souza e Prof. Dr. Leandro Carlos de Souza, pela dedicação, orientação valiosa e amizade durante todo este percurso.

Agradeço também aos professores Prof. Dr. Clauirton de Albuquerque Siebra (CI/UFPB) e Prof. Dr. Osmar Abílio de Carvalho Junior (UnB) pelas valiosas contribuições, críticas e sugestões durante a banca de qualificação e de defesa, que enriqueceram este trabalho.

Às professoras e professores, bem como aos colegas do Programa de Pós-Graduação em Informática, meu reconhecimento pelo compartilhamento de conhecimentos.

Aos pesquisadores do Laboratório de Inteligência Artificial Aplicada (LIAA) do Centro de Informática da UFPB, pela colaboração e suporte.

Por fim, mas não menos importante, à equipe do Laboratório de Estudos Fluviais (CCEN) e ao Grupo de Estudos de Ambientes Fluviais Semiáridos (GEAFS/CCEN), pelo companheirismo, incentivo e contribuições fundamentais ao longo desta jornada.

SUMÁRIO

ABSTRACT ..... **v**

RESUMO ..... **vi**

AGRADECIMENTOS ..... **vii**

LISTA DE FIGURAS ..... **ix**

LISTA DE QUADROS ..... **x**

1. INTRODUÇÃO **12**
1.1 Objetivos 14
1.2 Contribuições 15
1.3 Estrutura do Trabalho 15

2. REFERENCIAL TEÓRICO **17**
2.1 Detecção de Áreas Aluviais 17
2.2 Técnicas de Inteligência Artificial 20
2.2.1 CART - DT 22
2.2.2 Floresta Aleatória - RF 23
2.2.3 K-Vizinhos Mais Próximos - KNN 24
2.3 DGM-e para redução de instâncias 26
2.3.1 Estruturação das Instâncias 27
2.3.2 Cálculo da Representatividade 28
2.3.3 O processo de Simplificação 29

3 MATERIAL E MÉTODOS **31**
3.1 Área objeto da pesquisa 31
3.2 Obtenção dos Dados 33
3.3 Redução das instâncias e sua avaliação 36
3.4 Construção dos Modelos e Detecção de Áreas de Aluvião 38

4 RESULTADO E DISCUSSÃO **41**

5 CONCLUSÕES E TRABALHOS FUTUROS **49**

REFERÊNCIAS **51**

LISTA DE FIGURAS

Figura 1 - Paradigmas do Aprendizado de Máquina 21
Figura 2 - Estrutura da Árvore de Decisão 22
Figura 3 - A variação do k no KNN 25
Figura 4 - Representação das Etapas do método DGM-e 26
Figura 5 - Identificação das áreas de aluvião na Bacia do Riacho Grande 32
Figura 6 - Localização e Drenagem da Bacia do Riacho Grande 32
Figura 7 - Etapas de trabalho para obtenção dos dados 34
Figura 8 - Etapas para implementação da redução a partir do DGM-e 37
Figura 9 - Etapas de trabalho para aplicação das técnicas de aprendizagem 39
Figura 10 - Resultado da aplicação do KNN 42
Figura 11 - Resultados da aplicação do CART/DT 42
Figura 12 - Resultados obtidos com o RF 43
Figura 13 - Importância das variáveis utilizadas na predição 44
Figura 14 - Matriz de confusão RF (Banco reduzido para 5% do total) 45
Figura 15 – Curva Precision-Recall (Banco reduzido para 5% do total) 46
Figura 16 – Validação cruzada resultados (Banco reduzido para 5% do total) 46
Figura 17- Instâncias aluviais restantes nos diferentes patamares de redução 47

LISTA DE QUADROS

Quadro 1 – Descrição das variáveis utilizadas na modelagem..... 36
Quadro 2 – Percentuais aplicados em cada etapa de redução..... 38
Quadro 3 – Resultado obtido por cada uma das técnicas de AM..... 41

1. INTRODUÇÃO

A preocupação global com a preservação dos recursos hídricos tem se intensificado nas últimas décadas, diante das crescentes crises causadas pelas mudanças climáticas (TAYER et al., 2023). Estudos recentes, como o conduzido por Jasechko et al. (2024), analisaram milhares de medições em poços e aquíferos localizados em regiões áridas e semiáridas, evidenciando um declínio acelerado das reservas de água subterrânea em escala global. Esse cenário reforça a importância de estratégias para a gestão sustentável dos recursos hídricos, sobretudo em áreas particularmente vulneráveis às transformações climáticas.

Nesse contexto, os Objetivos de Desenvolvimento Sustentável (ODS), estabelecidos pela Organização das Nações Unidas (ONU), reforçam a necessidade de ações globais coordenadas para enfrentar os desafios relacionados à água e ao clima. O ODS número 6 prioriza a garantia de acesso universal à água potável e ao saneamento, enquanto o ODS número 13 destaca a importância de implementar medidas eficazes para mitigar os impactos das mudanças climáticas. Essas metas ressaltam a relevância de iniciativas que aliem conservação ambiental e gestão sustentável, sobretudo em regiões onde a falta de água ameaça diretamente a vida da população.

No semiárido da Região Nordeste do Brasil, a irregularidade dos regimes pluviométricos agrava os desafios relacionados à disponibilidade hídrica. O déficit de chuvas nesta região contribui para a predominância de rios efêmeros ou intermitentes, cujo fluxo ocorre apenas durante a estação chuvosa (MCLEOD et al., 2024). Além disso, eventos de seca recorrentes impactam o desenvolvimento socioeconômico, provocando o esgotamento de reservatórios superficiais, como lagos e açudes (BÚRQUEZ et al., 2024).

Nesse cenário, os aquíferos aluviais, definidos como formações geológicas porosas e permeáveis compostas por sedimentos não consolidados (cascalho, areia, silte e argila) depositados por corpos d'água, que armazenam e transmitem água subterrânea (FREEZE; CHERRY, 1979), destacam-se como fontes estratégicas de água subterrânea, devido à sua elevada capacidade de infiltração e proteção contra a evaporação, oferecendo uma alternativa viável para mitigar os efeitos das secas prolongadas (SILVA; SOUZA, 2023).

Apesar da importância dos aquíferos aluviais como fonte de água sua identificação e mapeamento precisos representam um desafio significativo. Os métodos tradicionais,

baseados em descrições visuais e análises manuais de amostras de solo (AMIT et al., 1996), são processos demorados, onerosos e de alcance limitado, dificultando a obtenção de uma visão abrangente em grandes áreas. Embora o sensoriamento remoto multiespectral e o radar de abertura sintética (SAR) tenham permitido análises em larga escala, a interpretação dos dados ainda pode ser complexa e a resolução espacial nem sempre é suficiente para identificar feições aluvionares com a precisão necessária. A introdução do LiDAR melhorou a análise topográfica, mas a identificação da natureza do material sedimentar e a distinção entre áreas aluvionares com potencial aquífero e outras formações geológicas superficiais permanecem dependentes de análise e interpretação especializadas, muitas vezes com limitações de escala e custo para o mapeamento extensivo necessário na região.

É importante destacar, que áreas aluvionares, regiões superficiais com depósitos de sedimentos transportados por rios, nem sempre correspondem a aquíferos produtivos, embora sejam zonas prioritárias para investigação devido ao seu potencial hídrico. No entanto, identificar e mapear esses depósitos representa um grande desafio também devido à sua distribuição irregular e à ausência de uma forma geométrica definida (CERVI; TAZIOLI, 2021). Consolidar estratégias de mapeamento eficiente, integrando tecnologias modernas e baixo custo, como as técnicas de inteligência artificial abordadas neste estudo, pode ampliar significativamente as possibilidades de gestão hídrica na região.

Os avanços nas técnicas de inteligência artificial, possibilitam a identificação desses depósitos em maior escala. Métodos de aprendizado de máquina têm mostrado grande potencial para modelar sistemas hídricos complexos, eliminando a necessidade de estabelecer previamente relações matemáticas fixas entre as variáveis de entrada (ZARESEFAT; DERAKHSHANI, 2023).

Apesar da grande quantidade de pesquisa que aplica técnicas de aprendizado de máquina em estudos de água subterrânea (DÍAZ-ALCAIDE; MARTÍNEZ-SANTOS, 2019; ALI et al., 2023; SEIFU et al., 2023), e embora alguns estudos tenham abordado a identificação de feições geomorfológicas relacionadas a ambientes fluviais (Pipaud; Lehmkuhl, 2017; Babič et al., 2021; Rabanaque et al., 2022), uma lacuna específica persiste na aplicação dessas técnicas para a identificação direta e em larga escala de áreas aluvionares com potencial aquífero no contexto do semiárido brasileiro. Os estudos existentes frequentemente utilizam dados com menor resolução espacial (como SRTM de

30 metros), ou têm como foco principal a delimitação de trechos de uma bacia hidrográfica.

Para preencher essa lacuna este trabalho desenvolveu e testou modelos preditivos para identificação de áreas aluvionares utilizando dados da bacia hidrográfica do Riacho Grande, localizada no sertão do Estado de Pernambuco. A bacia, que se estende por aproximadamente 315 Km^2, foi mapeada com aproximadamente 315 milhões de pontos, um ponto por metro quadrado. Para cada uma dessas instancias, foram capturadas dez variáveis de entrada e uma variável de saída que indica se a área é de aluvião ou não.

Um dos principais desafios do trabalho foi realizar a redução de instâncias, etapa essencial para viabilizar a aplicação eficiente de técnicas de aprendizagem de máquina. O crescimento exponencial na geração de dados em diversas áreas do conhecimento traz significativos desafios para o processamento e análise de informações. Nesse contexto, a redução de dados apresenta-se como uma solução para criar representações mais compactas, preservando, na medida do possível, as características originais dos dados.

Este trabalho aplicou o método Difusão Geométrica Markoviana estendido (DGM-e) (SILVA, 2012) como metodologia para redução de dados. A escolha se justificou pela necessidade de um método capaz de preservar a representatividade dos dados enquanto reduzia substancialmente seu volume. Baseado no método DGM original, proposto por Souza (2011), o DGM-e foi adaptado para tratar dados genéricos, utilizando uma análise probabilística fundamentada na propagação de calor entre elementos de um grafo.

Essa abordagem foi capaz de identificar e preservar as instâncias mais representativas entre os pontos originais. Após a redução dos dados, foram aplicadas diferentes técnicas de aprendizado de máquina, incluindo Árvore de Decisão do tipo CART (CART/DT), Florestas Aleatórias (RF) e K-Vizinhos Mais Próximos (KNN).

1.1 Objetivos

O objetivo geral desta dissertação é investigar o uso de técnicas de aprendizado de máquina para a identificação de depósitos aluvionares em regiões semiáridas, promovendo avanços metodológicos no mapeamento do fenômeno e contribuindo, de forma indireta, com subsídios técnicos para o planejamento territorial. Além dele, a dissertação apresenta os seguintes objetivos específicos:

Objetivo 1: Construir e comparar modelos preditivos com base em diferentes algoritmos de aprendizado de máquina, avaliando o desempenho por meio das métricas F1-score, recall e precisão, a fim de identificar quais abordagens oferecem maior eficácia na detecção de depósitos aluvionares.

Objetivo 2: Produzir modelos e resultados que possam ser aproveitados por iniciativas públicas voltadas ao mapeamento de áreas com potencial hídrico e uso estratégico do solo, especialmente em regiões semiáridas com demandas por desenvolvimento territorial sustentável.

1.2 Contribuições

Atendidos os objetivos apresentados na seção anterior, os resultados do trabalho eventualmente podem contribuir para formular políticas públicas mais eficazes na identificação e exploração de depósitos aluvionares, particularmente em regiões semiáridas. A modelagem resultante poderá subsidiar decisões estratégicas de gestão de recursos hídricos, atuando como uma ferramenta relevante para a proteção e uso sustentável das fontes de água subterrânea. Essa contribuição é particularmente relevante no semiárido nordestino, onde os aquíferos aluviais representam uma solução vital para mitigar os impactos das secas recorrentes e atender às necessidades de abastecimento hídrico de comunidades rurais isoladas.

1.3 Estrutura do Trabalho

Os capítulos subsequentes que compõe este trabalho estão organizados da seguinte forma:

- **Capítulo 2:** Neste capítulo, são apresentados os conceitos fundamentais que embasam o desenvolvimento deste trabalho. Inicialmente, será feita uma breve caracterização sobre como as detecções de áreas de aluvião evoluíram ao longo do tempo, desde os métodos tradicionais baseados em trabalho de campo até as abordagens modernas que utilizam inteligência artificial. Em seguida, serão detalhadas as técnicas de aprendizado de máquina empregadas na presente pesquisa, CART/DT, RF e KNN. Por fim, a última parte do capítulo será dedicada à apresentação detalhada do método DGM-e, explorando suas principais etapas e

características. Inicialmente, será descrito o processo de estruturação das instâncias, etapa essencial para organizar e preparar os dados de forma eficiente. Em seguida, será explicado o cálculo da representatividade de cada instância, destacando como o método identifica os elementos mais relevantes dentro do conjunto de dados. Por fim, será apresentado o processo de simplificação, que visa preservar as características estruturais do conjunto original.

- **Capítulo 3:** Este capítulo apresenta a metodologia utilizada na pesquisa. Inicialmente, será descrita a área de estudo, a bacia hidrográfica do Riacho Grande, localizada no sertão de Pernambuco, destacando suas características geográficas e sua importância no contexto hidrológico de regiões semiáridas. Em seguida, será detalhado como, utilizando o software ArcGIS Pro versão 10.3, foram obtidas as dez variáveis de entrada e a variável de saída (classificação de aluvião), extraídas de um conjunto de aproximadamente 315 milhões de pontos espalhados pela bacia, com densidade de um ponto por metro quadrado. Posteriormente, será apresentada a aplicação do método DGM-e para a redução de instâncias, destacando os quatro níveis de redução implementados (10%, 5%, 1% e 0,1%). Por fim, serão apresentados elementos relevantes utilizados para a aplicação das técnicas de inteligência artificial.

- **Capítulo 4:** O capítulo apresenta os resultados obtidos na redução dos dados e pela aplicação das técnicas de aprendizado de máquina implementadas, incluindo CART/DT, RF e KNN. Indicadores de desempenho, como F1-Score, Recall e Precisão, são comparados entre os diferentes modelos, identificando aquele que apresentou o melhor desempenho em cada nível de redução (10%, 5%, 1% e 0,1%). Por fim, é realizada uma discussão sobre os resultados obtidos na identificação de áreas aluvionares, destacando as limitações e implicações dos modelos desenvolvidos no contexto do estudo.

- **Capítulo 5:** Este capítulo apresenta as conclusões do trabalho, sintetizando os principais resultados alcançados na identificação de áreas aluvionares em regiões semiáridas. Além disso, discute as motivações para futuros desdobramentos, sugerindo caminhos para a continuidade da pesquisa.

2 REFERENCIAL TEÓRICO

Este capítulo, estruturado em três seções, apresenta os fundamentos que sustentam a metodologia desta pesquisa. A primeira seção traz uma revisão sobre a evolução das abordagens para detecção de áreas aluvionares, desde os métodos tradicionais até contemporâneos baseadas em inteligência artificial. A segunda seção é dedicada à descrição das técnicas de aprendizado de máquina empregadas, explorando seus conceitos. Por fim, a terceira seção apresenta o método DGM-e, destacando sua concepção, fundamentos e aplicabilidade no contexto investigado.

2.1 Detecção de Áreas Aluviais

A formação de aquíferos aluviais resulta da fragmentação de rochas e do transporte de sedimentos pela chuva. Esses depósitos sedimentares desempenham um papel crucial na retenção de água em regiões áridas, atuando como reservatórios naturais que contribuem para a sustentabilidade hídrica durante períodos de seca (BRAGA, 2016). No entanto, a identificação dessas formações sempre representou um desafio. Inicialmente, os aluviões eram reconhecidos com base em descrições visuais e análises manuais de amostras de solo, um processo demorado e oneroso que exigia inspeções extensivas e levantamentos topográficos (AMIT et al., 1996).

Com o avanço da tecnologia, as limitações desses métodos tradicionais impulsionaram a busca por abordagens mais eficientes. Nesse contexto, o sensoriamento remoto multiespectral emergiu como uma alternativa promissora, permitindo análises em larga escala com maior precisão e reduzindo a necessidade de trabalhos de campo exaustivos. Estudos pioneiros demonstraram o impacto dessa abordagem na identificação de aluviões. Gillespie; Kahle; Palluconi (1984), por exemplo utilizaram um escâner multiespectral de infravermelho térmico para mapear leques aluviais no Vale da Morte, Califórnia, evidenciando a eficácia do sensoriamento térmico na discriminação de diferentes sedimentos. Da mesma forma, Crouvi et al. (2006) aplicaram espectrometria de campo e sensoriamento remoto hiperespectral no deserto do Negev, Israel, identificando assinaturas espectrais específicas associadas a depósitos aluvionares.

Apesar dos avanços proporcionados pelo sensoriamento remoto multiespectral, a necessidade de mapear áreas ainda mais extensas e obter dados detalhados em condições desafiadoras levou à adoção do radar de abertura sintética (SAR). Essa tecnologia

ampliou significativamente as possibilidades de análise, sendo particularmente útil para o estudo de superfícies aluvionares em regiões com vegetação densa ou de difícil acesso. Farr; Chadwick (1996) demonstraram sua aplicabilidade ao utilizar dados de SAR para mapear leques aluviais nas Montanhas Kun Lun, na China, possibilitando uma análise detalhada da morfologia e dos processos geomórficos dessas formações. Estudos posteriores reforçaram esse avanço, como os de Hetz et al., (2016) e Gaber; Koch; El-Baz (2010), que empregaram SAR na análise de superfícies aluvionares em desertos, ampliando o conhecimento sobre a dinâmica desses depósitos sedimentares.

A introdução do LiDAR (Light Detection and Ranging), com sua capacidade de gerar modelos digitais de elevação altamente precisos, transformaram os modos de se classificar e identificar áreas de aluvião. Esses modelos facilitam a análise detalhada da topografia e da estrutura de aluviões, possibilitando a identificação de características geomorfológicas com uma precisão sem precedentes. Conforme destacado por Hohenthal et al., (2011) e Cavalli et al., (2008), o LiDAR se estabeleceu como uma ferramenta eficiente para obter informações topográficas detalhadas, mesmo em áreas montanhosas e densamente florestadas. Estudos pioneiros de Staley; Wasklewicz; Blaszczynski (2006) e Frankel; Dolan (2007) demonstraram a eficácia do LiDAR na análise de padrões de deposição e na caracterização da rugosidade de superfícies de leques aluviais. Esses estudos revelaram zonas de deposição distintas e permitiram diferenciar superfícies de leques aluviais de diferentes idades. Pesquisas subsequentes, como as realizadas por Cavalli; Marchi (2008) e Regmi, McDonald; Bacon (2014), também utilizaram LiDAR para identificar e classificar leques aluviais.

Mais recentemente, o uso de técnicas de aprendizado de máquina potencializou a análise de grandes volumes de dados geoespaciais. Essas abordagens, aliadas às tecnologias anteriores, têm ampliado a capacidade de detecção e classificação de áreas aluvionares. Conforme destacado por Muñoz-Carpena et al., (2023), a integração dessas abordagens pode reduzir a incerteza dos modelos hidrológicos e melhorar a precisão das previsões, especialmente em grandes sistemas integrados.

Dois tipos de estudos sobre água subterrânea utilizando técnicas de aprendizado de máquina se destacam na literatura. O primeiro foca na detecção de água subterrânea, com exemplos incluindo os seguintes trabalhos: trabalhos de (DÍAZ-ALCAIDE; MARTÍNEZ-SANTOS, 2019), (ALI et al., 2023), (SEIFU et al., 2023), (MARTÍNEZ-SANTOS; RENARD, 2020) e (NGUYEN et al., 2020). O segundo tipo envolve o desenvolvimento de modelos específicos para a previsão do nível de água subterrânea,

como os seguintes estudos: (ARDABILI et al., 2019), (TAO et al., 2022), (UC-CASTILLO et al., 2023), (GHOLAMI et al., 2023), (KAYHOMAYOON et al., 2022), (VADIATI et al., 2022), (SHAKYA; BHATTACHARJYA; DADHICH, 2022), (SRIVASTAVA; SHUKLA; JEMNI, 2022), (GAFFOOR et al., 2022), (EL BILALI; TALEB; BROUZIYNE, 2021) e (LUIZ, 2022).

Embora o estudo da água subterrânea tenha recebido ampla atenção na aplicação de técnicas de aprendizado de máquina, a identificação e classificação específicas de formações aluvionares têm sido pouco investigadas. O levantamento bibliográfico realizado revelou que poucas pesquisas se dedicaram especificamente a essa temática.

Pipaud; Lehmkuhl (2017) realizaram uma pesquisa que apresentou um método para a delimitação e classificação de leques aluviais utilizando MDEs, combinados com a técnica de clusterização *mean-shift* e uma máquina de vetores de suporte (SVM). As variáveis de entrada, utilizadas na segmentação, incluíram parâmetros morfométricos como declividade, curvatura transversal, curvatura longitudinal, assimetria dos valores de altitude e o desvio do gradiente em relação ao ápice do leque. O estudo foi conduzido utilizando dados SRTM com uma resolução de 30 metros. A segmentação *mean-shift* foi aplicada repetidamente com diferentes parâmetros para capturar a variabilidade dos leques aluviais. Posteriormente, uma SVM foi utilizada para a classificação dos objetos já agrupados. Os resultados mostraram que essa abordagem, denominada Análise Morfométrica Baseada em Objetos (OBMA), alcançou bons resultados, medidos a partir da utilização de valores de pertinência *fuzzy* derivados da classificação SVM para a seleção da segmentação mais apropriada para cada leque aluvial identificado.

Babič et al., (2021) também modelaram e classificaram leques aluviais utilizando Modelos Digitais de Elevação e diversas técnicas de aprendizagem de máquina. O estudo se concentrou nos leques aluviais da Eslovênia, identificando sete principais parâmetros geomorfométricos: inclinação média da hinterlândia, inclinação média da torrente, número de rugosidade de Melton da bacia de captação, razão de relevo, relação entre a área do leque e a área da hinterlândia, número de Melton do leque aluvial e inclinação média do leque. Através da comparação de cinco métodos de aprendizagem, incluindo Random Forest, Programação Genética, SVM, Rede Neural e um método híbrido de grafo de Euler, os pesquisadores demonstraram a eficácia dessas abordagens na classificação automática de leques aluviais propensos a fluxos de detritos. O estudo utilizou dados de várias fontes de imagens de satélite, como ASTER, GeoEye, Ikonos, WorldView, ALOS e SPOT Image, com resoluções variando de 2 metros (WorldView) a 30 metros (ASTER

e SRTM). Os resultados, validados com dados empíricos, mostraram que a Programação Genética apresentou o melhor desempenho na classificação.

Rabanaque et al., (2022) realizaram uma análise hidromorfológica em larga escala de riachos efêmeros utilizando algoritmos de aprendizado de máquina, especificamente a SVM e RF, para segmentar e classificar canais fluviais e formas fluviais associadas. As variáveis de entrada incluíram largura do canal ativo, largura do fundo do vale, gradiente de declive, distância de rota e potência específica do fluxo, além de dados de sensoriamento remoto das bandas espectrais do Sentinel-2 (RGB, NIR1, SWIR1, SWIR2) e índices espectrais como NDVI, GRVI e NDWI. Os dados LiDAR do Projeto PNOA-LiDAR foram utilizados para criar um modelo digital de elevação (DEM) reamostrado para maior resolução, permitindo análises mais detalhadas. A precisão dos modelos foi avaliada usando a Matriz de Confusão, Precisão e o Índice de Kappa de Cohen, com o SVM obtendo uma precisão média de 0,87 e Kappa de 0,84, enquanto o RF obteve uma precisão média de 0,85 e Kappa de 0,81.

A evolução das técnicas de detecção de áreas aluvionares reflete o avanço tecnológico e computacional, mas também a crescente necessidade de compreender os processos geomorfológicos de maneira detalhada para adoção de políticas públicas mais eficientes.

2.2 Técnicas de Inteligência Artificial

O aumento significativo do poder computacional nas últimas décadas elevou consideravelmente a relevância do aprendizado de máquina. Um dos desafios centrais enfrentados pelos algoritmos dessa área é maximizar sua capacidade de generalização, ou seja, a habilidade de fornecer respostas eficientes a situações inéditas, não encontradas durante o processo de treinamento. Essa característica é crucial para garantir que os sistemas possam tomar decisões adequadas em uma variedade de cenários (MITCHELL, 1997).

O aprendizado de máquina é tradicionalmente dividido em quatro paradigmas principais: aprendizado supervisionado, aprendizado não supervisionado, aprendizado por reforço e aprendizado semi-supervisionado. Os dois primeiros (ver Figura 1) são os mais amplamente utilizados.

No aprendizado supervisionado, o modelo é treinado com dados rotulados, ou seja, cada entrada no conjunto de treinamento está associada a uma saída conhecida. Isso

permite que o algoritmo aprenda a mapear corretamente as entradas para as saídas esperadas. Esse paradigma é amplamente empregado em tarefas preditivas, como classificação e regressão.

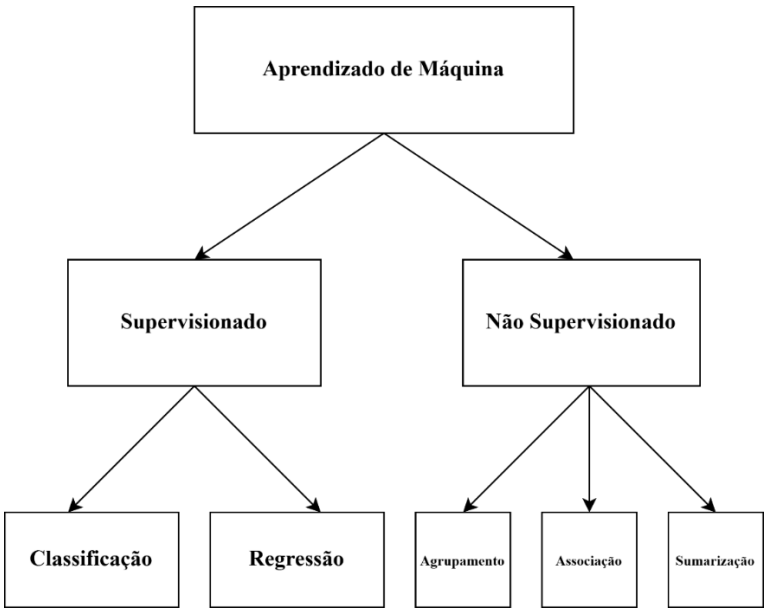


Figura 1 - Paradigmas do Aprendizado de Máquina

Por outro lado, o aprendizado não supervisionado busca identificar padrões ou relações em dados sem rótulos associados. É aplicado principalmente em tarefas descritivas, como agrupamento e associação. O aprendizado por reforço, por sua vez, envolve a interação contínua de um agente com o ambiente, em que este recebe recompensas ou penalidades para aprimorar suas decisões ao longo do tempo. Finalmente, o aprendizado semi-supervisionado combina dados rotulados e não rotulados, oferecendo uma abordagem híbrida entre os paradigmas supervisionado e não supervisionado (FACELLI et al., 2021).

No contexto desta pesquisa, o enfoque está no aprendizado supervisionado, dado o uso de dados rotulados. Esses algoritmos são projetados para aprender uma função que mapeie com precisão entradas para saídas específicas, com base em um conjunto de dados de treinamento previamente estabelecido (Russel, 2022).

Com o objetivo de identificar áreas de aluvião, foram desenvolvidos modelos supervisionados como CART/DT, K-Vizinhos Mais Próximos (KNN) e Floresta Aleatória (RF). Cada um desses algoritmos apresenta características específicas que os tornam adequados para diferentes aspectos do problema investigado. Para contextualizar

a aplicação desses modelos, este capítulo apresenta uma breve descrição acerca dos principais das características de cada técnica utilizada nesta pesquisa.

2.2.1 CART - DT

As Árvores de Decisão (DT) são uma forma eficaz de representar o conhecimento adquirido a partir de conjuntos de dados, organizando as informações como uma combinação de restrições nos valores dos atributos das instâncias. Cada caminho da raiz até uma folha da árvore representa uma sequência de testes nos atributos. Os nós internos correspondem a pontos de decisão, enquanto os ramos indicam os diferentes resultados possíveis dessas decisões.

Durante a construção de uma DT (Figura 2), a seleção dos atributos para divisão é determinada pela pureza de cada nó. O processo se inicia no nó raiz e busca criar uma estrutura de classificação que seja compacta e eficiente (Mitchell, 1997). Isso é feito através da seleção recursiva dos atributos mais informativos para formar os nós internos e os ramos correspondentes a cada valor possível do atributo escolhido. A construção prossegue até que os exemplos em um nó sejam homogêneos, formando nós folha, onde a classificação final é realizada.

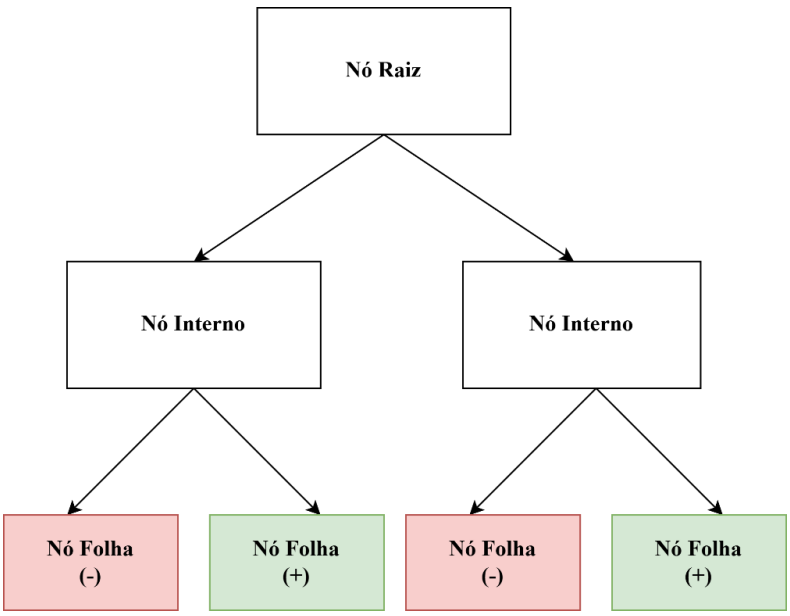


Figura 2 - Estrutura da Árvore de Decisão

A escolha do atributo para dividir os dados em cada nó é fundamentada em medidas de pureza, com o objetivo de criar a DT mais precisa e menor possível. Nesta pesquisa, foi utilizado o algoritmo CART (Classificação e Regressão de Árvores), que é

adequado tanto para problemas de classificação quanto de regressão. Na classificação, o algoritmo CART utiliza como métrica de decisão o índice de Gini (Marsland, 2015). Esse índice é calculado como:

$$Gini = \sum_{i=1}^{n} p_i^2 \quad (1)$$

Onde *n* corresponde ao número de classes, e *pi* representa a proporção da classe *i* no nó. Um índice de Gini próximo a 0 indica alta pureza no nó, enquanto valores mais próximos de 0,5 indicam maior impureza. Essa métrica garante que cada nó seja o mais homogêneo possível, ajudando a evitar inconsistências na classificação.

Para validar os modelos de DT, esta pesquisa utilizou o método de validação cruzada *k-fold* (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Nesse método, o conjunto de dados é dividido em k grupos ou dobras. Em cada iteração, uma das dobras é utilizada como dados de validação, enquanto as dobras restantes são usadas para o treinamento do modelo. Esse processo é repetido k vezes, garantindo que cada grupo seja usado uma vez como conjunto de validação. Ao final, o erro médio de todas as iterações é calculado, fornecendo uma estimativa mais precisa do desempenho do modelo e reduzindo o risco de superajuste.

2.2.2 Floresta Aleatória - RF

As Florestas Aleatórias (RF) são uma evolução das DT, desenvolvidas para aumentar a precisão e a robustez dos modelos preditivos. Essa técnica consiste na construção de um conjunto de n árvores de decisão independentes, cada uma treinada com um subconjunto aleatório dos dados e das variáveis. Esse procedimento, conhecido como *bagging* (ou *bootstrap aggregating*), ajuda a reduzir o risco de sobreajuste (*overfitting*) ao combinar as previsões de diversas árvores. A previsão final é obtida por meio da média, no caso de regressão, ou pelo voto majoritário, no caso de classificação (BREIMAN, 2001).

Cada árvore na floresta é gerada utilizando amostras aleatórias com reposição a partir do conjunto de dados original. Durante o processo de divisão dos nós, apenas um subconjunto aleatório de variáveis é analisado, o que introduz maior diversidade entre as árvores. Essa característica torna as RF potencialmente eficazes para capturar padrões

complexos em dados e especialmente vantajosas em cenários com muitas dimensões ou ainda com variáveis altamente correlacionadas.

A abordagem que combina várias árvores confere às RF maior estabilidade e precisão, minimizando os impactos de instâncias ruidosas ou de outliers no desempenho geral do modelo. Outra vantagem significativa das Florestas Aleatórias é sua capacidade de medir a importância relativa das variáveis, auxiliando na identificação dos atributos mais relevantes para o problema em estudo. Esse recurso é particularmente útil para análises exploratórias e para a construção de modelos mais interpretáveis.

2.2.3 K-Vizinhos Mais Próximos - KNN

O algoritmo K-Vizinhos Mais Próximos (KNN) é reconhecido por sua eficiência e relativa simplicidade no momento da implementação. Este método pertence à categoria de algoritmos supervisionados baseados em instâncias, sendo que ele não realiza uma etapa de treinamento explícita antes de produzir previsões. Em vez disso, o KNN utiliza todo o conjunto de treinamento armazenado para prever a saída de uma nova amostra com base na proximidade entre essa amostra e as instâncias existentes no conjunto (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

O princípio do KNN é medir a similaridade entre uma amostra desconhecida e as instâncias do conjunto de treinamento, com base em métricas de distância, como a Euclidiana, Manhattan ou Minkowski. Após calcular essas distâncias, são selecionados os k vizinhos mais próximos, cuja classe ou valor será usado para determinar a saída. Para tarefas de classificação, o resultado é baseado na classe predominante entre os vizinhos; em problemas de regressão, a saída é calculada como a média ponderada dos valores das instâncias selecionadas.

Embora sua implementação seja direta, a eficiência do KNN depende fortemente do valor de k, que define o número de vizinhos considerados. Valores muito pequenos para k podem tornar o modelo mais suscetível a ruídos nos dados, enquanto valores maiores podem diluir a relevância das instâncias mais próximas, reduzindo a precisão do modelo.

A Figura 3 ilustra o impacto da escolha do parâmetro k no comportamento do modelo K-Vizinhos Mais Próximos (KNN), evidenciando como diferentes valores podem influenciar a classificação de uma mesma amostra. No exemplo apresentado na figura, observa-se que, para um valor menor de k, o modelo tende a classificar a amostra com

base em vizinhos mais próximos, o que pode resultar em decisões altamente sensíveis a variações locais e à presença de ruídos. Por outro lado, à medida que *k* aumenta, o modelo passa a considerar um conjunto maior de vizinhos na tomada de decisão, o que pode suavizar a classificação, reduzindo a influência de outliers, mas, ao mesmo tempo, aumentando o risco de incluir instâncias irrelevantes ou até mesmo pertencentes a outras classes. Dessa forma, a Figura 3 destaca visualmente a necessidade de um ajuste criterioso do parâmetro *k* para equilibrar precisão e generalização no modelo.

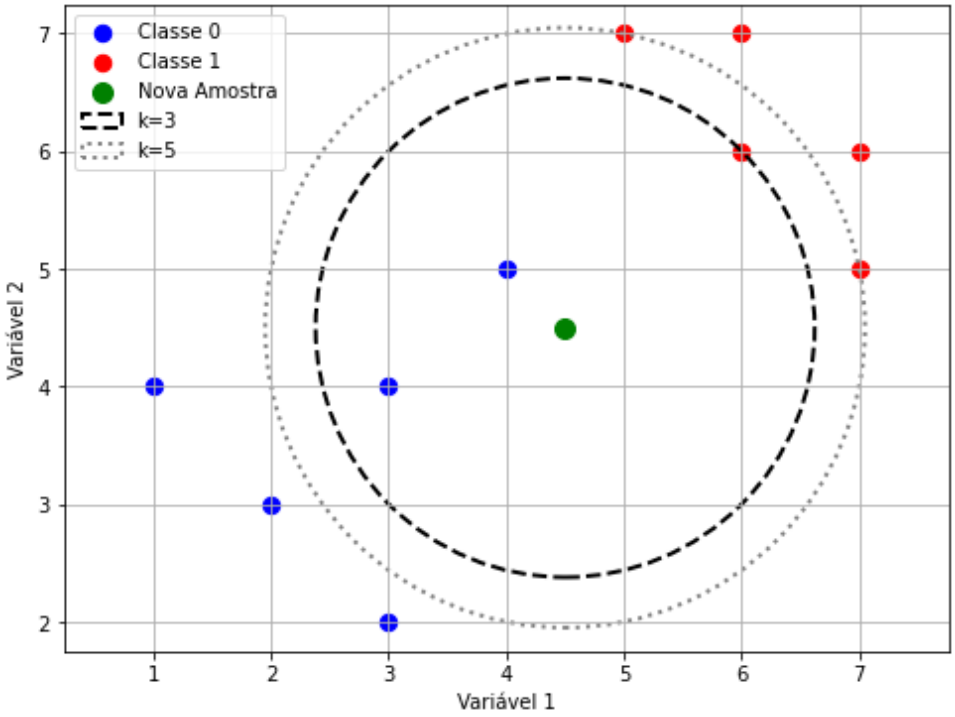


Figura 3 - A variação do k no KNN

O KNN possui algumas limitações importantes. Por exemplo, o custo computacional pode ser alto para grandes volumes de dados, já que cada previsão exige o cálculo de distâncias para todas as instâncias do conjunto de treinamento. Além disso, o desempenho do KNN pode ser prejudicado em cenários de alta dimensionalidade, que reduz a eficácia das métricas de distância. Para lidar com essas questões, técnicas como redução de dimensionalidade, métodos de ponderação por distância e a seleção adaptativa do valor de *k* são amplamente investigadas na literatura.

O KNN tem sido amplamente aplicado em diversas áreas. Sua estrutura simples o torna uma ferramenta adequada para experimentação inicial, enquanto variações mais

avançadas buscam melhorar sua eficácia e reduzir suas limitações em cenários mais complexos.

2.3 DGM-e para redução de instâncias

O aumento no volume de dados gerados atualmente exige a adoção de estratégias eficientes para seu manejo, sendo a redução de instâncias uma abordagem possível. Os métodos de redução podem ser classificados em duas categorias principais: aqueles baseados em técnicas de clusterização e os supervisionados. Este trabalho estendeu o método chamado Difusão Geométrica Markoviana - DGM-e (SILVA, 2012), a partir do método original, desenvolvido por SOUZA (2011).

O DGM é baseado na equação diferencial do calor entre partículas, onde se observa que após certo tempo partículas com menor interação com partículas vizinhas retém menos calor, mostrando-se relativamente diferentes do conjunto de partículas próximas, por outro lado, elementos que podem ser considerados como mais importantes no sistema são os que apresentam maior capacidade de reter calor.

Na sua forma original, o método foi concebido para realizar a redução em malhas triangulares, preservando os vértices que apresentam características específicas de interesse durante o processo de simplificação. A extensão demonstrou a capacidade do método de ser aplicado a outros dados para além dos originalmente testados, para tanto a adaptação precisou criar uma estrutura geométrica para que dados genéricos fossem suportados.

O DGM-e redunda em um método de redução de instancias que pode ser representado no fluxograma abaixo:

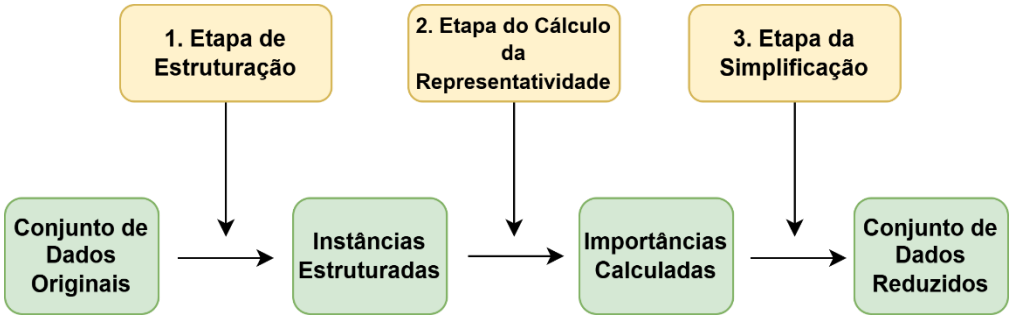


Figura 4 - Representação das Etapas do método DGM-e

A seção segue com uma breve apresentação de cada uma das etapas estruturantes do método DGM-e.

2.3.1 Estruturação das Instâncias

A estruturação das instâncias é o que permite aos elementos assumirem a estrutura de um grafo, ou seja, após a passagem dos elementos pelo algoritmo de estruturação, os pontos assumem a estrutura de um grafo o que permite a aplicação do DGM-e. O primeiro passo para a estruturação das instâncias é a realização de escalonamento dos dados. Os elementos a serem processados podem apresentar grandes variações de escala. Calcular a distância entre instâncias, sem essa transformação prévia, não é adequado.

Assim para a instancia x_{ij} correspondente a j^{th} dimensão de x_i, a equação utilizada para o escalonamento foi:

$$\hat{x}_{ij} = \frac{x_{ij} - m_j}{M - m} \tag{2}$$

Onde m_j é o valor mínimo da dimensão j_{th}, m e M equivalem ao mínimo e máximo globais. Com o escalonamento acima as instâncias assumem valores entre 0 e 1 e fica garantida a manutenção da relação de distância entre as instâncias.

A segunda etapa de estruturação dos dados consiste na partição do espaço. O DGM-e utiliza a técnica KD-Tree (MOORE, 1991), que funciona de modo semelhante a uma generalização da d binária de busca. O algoritmo escolhe a mediana da partição atual e a divide em duas partes. Dessa forma, para duas dimensões, a divisão será uma linha e cada divisão um plano. No caso de três dimensões teremos cubos, generalizando a estrutura para mais dimensões teremos hipercubos.

A segmentação é parametrizada por:

$$r = (n/k)^{\frac{1}{k}} \tag{3}$$

Onde r é o número de segmentos, n o número de instâncias e k o número de dimensões. Após a divisão cria-se um espaço com r^k hipercubos, tal criação ocorre através da fórmula abaixo:

$$h_{ij=} \lfloor \hat{x}_{ij} \cdot r \rfloor \quad (4)$$

Onde h_{ij} representa a coordenada inteira da instância i ao longo da $j-ésima$ dimensão do hipercubo ao qual ela será alocada dentro da estrutura particionada. O termo $\hat{x}_{ij}$ denota o valor normalizado da instância i na $j-ésima$ dimensão do espço de dados antes da sua discretização nos hipercubos. A função $\lfloor \cdot \rfloor$ representa a função floor, responsável por arredondar o valor resultante da multiplicação.

Para evitar uma explosão combinatória, o método DGM-e propõe uma otimização que consiste em gerar apenas os hipercubos populados ao invés de se gerar todos para depois efetuar a população.

Sucede a criação dos hipercubos o último passo para estruturação dos dados, a construção da vizinhança. O DGM-e utiliza a própria estrutura dos hipercubos como base inicial para construção da relação de vizinhança entre as instâncias. O método adota a estratégia de determinar a vizinhança de uma instância a partir do método de amostragem Monte Carlo. Assim, são selecionadas ***p*** instâncias pertencentes ao mesmo hipercubo da instância. A vantagem de tal abordagem é se mostrar efetiva em conjuntos de dados de alta densidade.

O resultado da estruturação é o surgimento de uma tupla de vizinhança de cada instância. Deve-se destacar que em tal estruturação a simetria de ligação entre as instâncias fica garantida, ou seja, se uma instância "a" está ligada a "b", deve-se garantir que "b" também esteja ligada a "a".

2.3.2 Cálculo da Representatividade

O DGM-e propõe a construção de um núcleo para cálculo da similaridade entre as instâncias. Esse núcleo deve ser uma função que atenda as seguintes relações:

$$k(x_i, x_j) = k(x_j, x_i) \quad (5)$$

$$(x_i, x_j) \geq 0 \; se \; i \neq j \quad (6)$$

$$k(x_i, x_j) > 0 \tag{7}$$

A simetria fica garantida em (5) removendo a influência de direção de ligação entra as instâncias. São as inequações subsequentes (6) e (7) que permitem associar os valores obtidos por *k* como medidas de probabilidade.

Adotou-se no DGM-e a distância euclidiana para comparar relação entre os elementos. Assim, considerando-se o núcleo Gaussiano $k(x_i, x_j) = \exp\ (-|x_i - x_j|)^2$, temos $|x_i - x_j|^2$ como a distância euclidiana ao quadrado. Para efetuar a comparação dos elementos pode-se generalizar a distância para qualquer função de distância.

Os valores do núcleo são utilizados para construção de uma matriz simétrica sem valores negativos. Uma normalização ($d(x_i)$) é aplicada, resultando em uma matriz de Transição (P). Assim temos:

$$P(x_i, x_j) = \frac{k(x_j, x_i)}{d(x_i)} \tag{8}$$

Onde $d(x_i)$ realiza a normalização conforme relação que segue:

$$d(x_i) = \sum_{x_j \in X} k(x_i, x_j) \tag{9}$$

.

A matriz de transição (P) incorpora a estrutura geométrica do elemento com seus vizinhos imediatos. A matriz, construída a partir da função núcleo, tem em sua diagonal justamente o valor da similaridade entre os vizinhos. A partir de tal valor de similaridade é que ocorre a redução de elementos.

2.3.3 O processo de Simplificação

Os valores da diagonal da matriz de Transição (P) refletem a relação de similaridade entre a vizinhança de cada instância. No entanto, o valor encontrado na diagonal, e representativo de cada ponto (vértice), ainda sofre a influência da conectividade que possui em virtude da normalização aplicada (7). Essa influência pode comprometer a avaliação da representatividade de um elemento, uma vez que diferentes

pontos podem possuir diferentes densidades de conexões. Diante disso, o DGM-e propõe um cálculo de importância V que leve em conta o número de conexões que um elemento possui. O cálculo de importância de um elemento no método DGM-e ocorre assim:

$$V(x_i) = \frac{Deg(x_i)}{P(x_i,x_i)} \tag{10}$$

onde $Deg(x_i)$ é o número de conexões que um elemento x_i possui na estrutura geométrica e $P(x_i, x_i)$ é o valor do ponto na matriz de transição.

A importância de um ponto mede a propensão daquela região ser removida. No DGM-e a simplificação atua nos elementos com baixos valores. Para valores mais altos de importância opta-se pela manutenção, pois trata-se de regiões de maior relevância para a estrutura dos dados. Além disso, o DGM-e usa como critério de parada o percentual de redução de instâncias que o usuário decidir.

De maneira geral o processo de simplificação ocorre da seguinte forma: inicialmente, calcula-se a importância $V(x_i)$ de cada elemento da base de dados, considerando a diagonal da matriz de transição (P) e no grau $d(x_i)$. As instâncias são ordenadas, de acordo com sua importância, e as menos representativas (com menores valores de $V(x_i)$) são removidas de maneira iterativa. Após cada remoção, a matriz de transição é recalculada ajustando os valores de importância para as instâncias restantes. Esse processo permanece até que o ponto de parada seja atingido. Essa abordagem preserva as características essenciais do conjunto, de modo a manter os elementos mais relevantes.

3 MATERIAL E MÉTODOS

Este Capítulo se inicia com a apresentação da área objeto de estudo, no caso a bacia do Riacho Grande, localizada no semiárido da Região Nordeste do Brasil. Coube a segunda seção do capítulo descrever a obtenção e manipulação inicial dos dados até a formação de um banco de dados. Por fim, as duas últimas seções do capítulo apresentam respectivamente o método empregado na redução de instâncias do banco e a metodologia para construção de modelos de aprendizagem de máquina para detecção de áreas aluvionares.

3.1 Área objeto da pesquisa

A área de estudo desta pesquisa é a bacia hidrográfica do Riacho Grande, situada no sertão de Pernambuco e abrangendo os municípios de Serra Talhada, Calumbi, Flores e Betânia. Para melhor contextualizar a região, suas características serão apresentadas na seguinte sequência: clima, geologia, geomorfologia e hidrografia.

A região da bacia do Riacho Grande possui um clima tropical semiárido, marcado por elevadas temperaturas médias, que variam entre 25°C e 30°C, e baixos índices pluviométricos anuais, com precipitação média entre 450 e 700 mm. A ocorrência de chuvas concentra-se principalmente nos meses de novembro a abril.

A bacia encontra-se integralmente localizada no Planalto da Borborema, uma formação geológica caracterizada pela predominância de rochas cristalinas. Essa composição rochosa influencia diretamente o relevo da região.

O relevo da bacia do Riacho Grande apresenta variações altimétricas significativas, com altitudes que oscilam entre aproximadamente 430 e 870 metros, diminuindo no sentido leste para oeste (a jusante). As maiores declividades são observadas nas áreas de cabeceira, enquanto a bacia, de modo geral, caracteriza-se por áreas de baixa declividade. As regiões de maior altitude atuam como zonas de produção de sedimentos, que são posteriormente transportados pela água da chuva para as áreas mais baixas, onde ocorre a deposição, processo fundamental para a formação das áreas de aluvião (Figura 5).

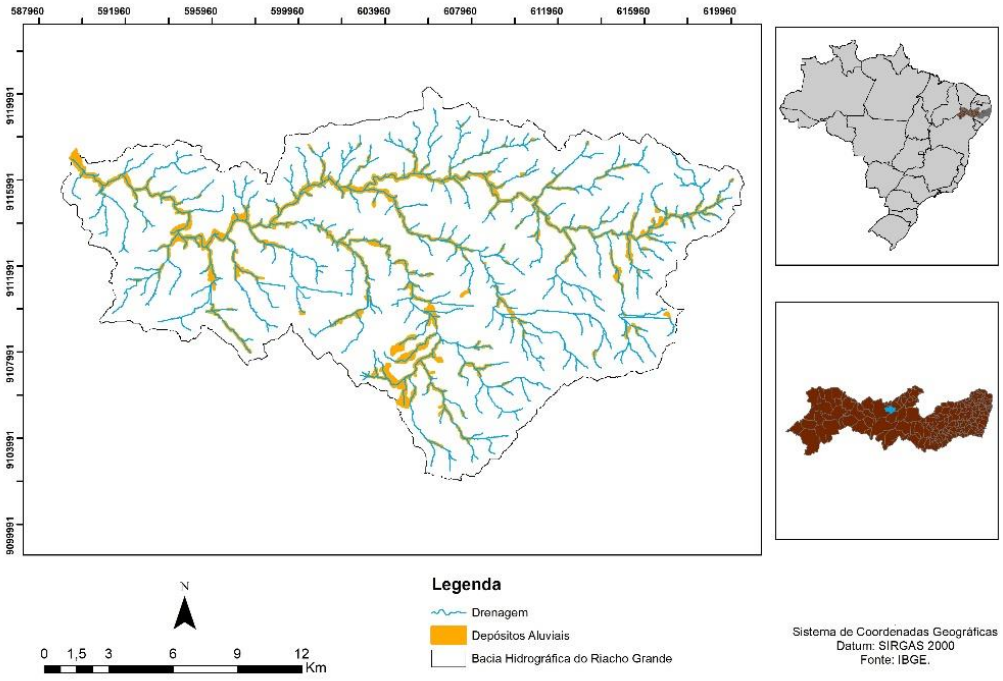


Figura 5 - Identificação das áreas de aluvião na Bacia do Riacho Grande

A rede hidrográfica da bacia do Riacho Grande é composta por rios efêmeros, cujo fluxo de água é intermitente, ocorrendo principalmente durante a estação chuvosa (Figura 6). O principal curso d'água é o Riacho Grande, que, apesar de sua relevância para a bacia, é uma sub-bacia do Rio Pajeú.

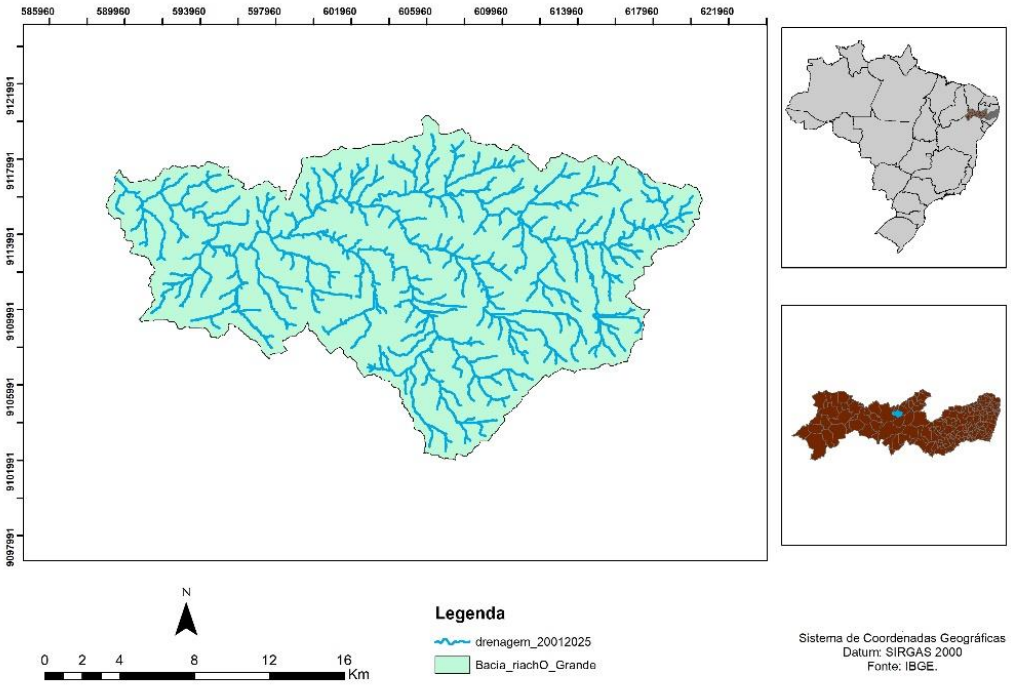


Figura 6 - Localização e Drenagem da Bacia do Riacho Grande

Com a área de estudo contextualizada em termos climáticos, geológicos, geomorfológicos e hidrográficos, a próxima etapa metodológica consistiu na obtenção e no tratamento dos dados geoespaciais que permitiram a caracterização detalhada da bacia do Riacho Grande. A definição e o processo de coleta dessas informações são detalhados na seção seguinte.

3.2 Obtenção dos Dados

A definição das onze dimensões que compõem o banco de dados baseou-se na relevância dessas variáveis para a caracterização geomorfológica e hidrológica das áreas aluvionares, considerando fatores essenciais para a formação e identificação desses depósitos. Foram utilizadas as variáveis disponíveis nos dados extraídos da bacia hidrográfica do Riacho Grande, priorizando aquelas que apresentavam potencial para influenciar a ocorrência de aluviões. A seleção foi fundamentada tanto em estudos anteriores sobre predição de aluviões quanto no conhecimento consolidado sobre os processos de transporte e deposição de sedimentos em bacias hidrográficas.

A definição da variável de saída (Figura 6), que indica a presença ou ausência de áreas aluvionares, foi realizada por meio de trabalhos de campo e análise de imagens de satélite. Inicialmente, foram realizadas vistorias em trechos da bacia do Riacho Grande considerados críticos para a ocorrência de aluviões, com o objetivo de identificar e delimitar a extensão dessas áreas. Posteriormente, essas informações foram complementadas e refinadas através da análise visual de imagens de alta resolução do Google Earth. A precisão dos polígonos criados a partir do trabalho foi cuidadosamente avaliada para garantir a representatividade da variável de saída utilizada na modelagem.

A obtenção das onze dimensões (Quadro 1) que compõem o banco de dados ocorreu por meio do uso do software ArcMap, versão 10.8, uma ferramenta de Sistema de Informação Geográfica (SIG). Através dessa ferramenta, foram realizadas análises espaciais e geográficas fundamentais para coleta das variáveis.

O primeiro passo do trabalho foi a obtenção dos Modelos Digitais de Elevação junto ao Programa Pernambuco Tridimensional (PE3D), que utiliza tecnologias de aerofotogrametria e o LiDAR (Light Detection and Ranging) para realizar o mapeamento do Estado de Pernambuco. Os modelos obtidos possuem uma resolução espacial de 1 metro quadrado (1x1m), o que representa uma precisão significativamente superior aos modelos gratuitos disponíveis globalmente, como o Shuttle Radar Topography Mission

(SRTM) e o ALOS PALSAR, que possuem resolução de 30x30 metros. Essa alta resolução permitiu a extração detalhada de variáveis topográficas e hidrológicas, aumentando a precisão da análise das áreas aluvionares.

Com os modelos processados, foram realizadas as etapas de tratamento das imagens e extração de atributos relevantes para o estudo. O fluxograma abaixo ilustra essas etapas, apresentando desde a aquisição dos dados brutos até a geração do conjunto final de informações utilizadas para a modelagem. O fluxograma destaca as principais operações realizadas, como a interpolação dos dados de elevação, o cálculo de atributos geomorfológicos e hidrológicos e a integração das variáveis em uma base unificada. Essas etapas foram fundamentais para garantir a representatividade dos dados e viabilizar a aplicação das técnicas de inteligência artificial para a identificação de áreas aluvionares.

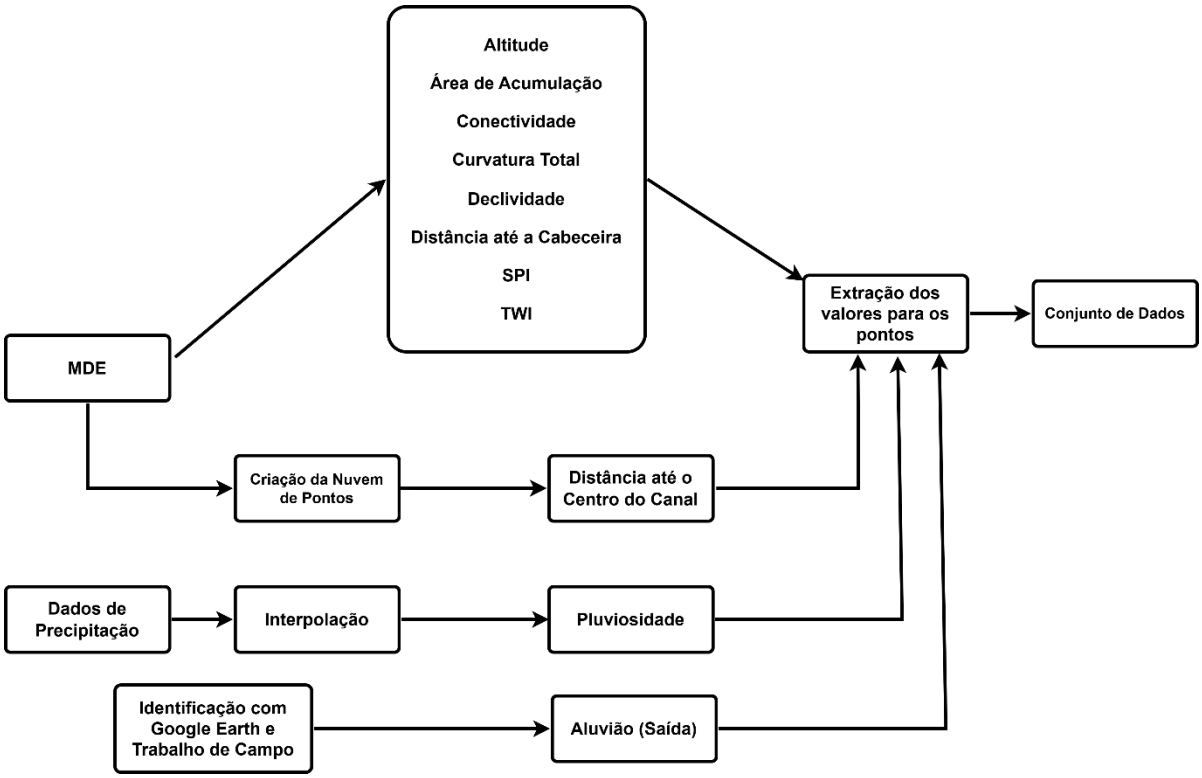


Figura 7 - Etapas de trabalho para obtenção dos dados

A Figura 7 ilustra o fluxo de trabalho para a obtenção do conjunto de dados utilizado nesta pesquisa. O processo inicia-se com a aquisição de Modelos Digitais de Elevação (MDE) provenientes do Programa Pernambuco Tridimensional (PE3D) e de dados de precipitação fornecidos pela Agência Pernambucana de Águas e Clima (APAC). Paralelamente, a identificação das áreas de aluvião (Saída) é realizada através da análise de imagens do Google Earth e de trabalhos de campo.

A partir do MDE, são derivadas variáveis geomorfológicas e hidrológicas, incluindo Altitude, Área de Acumulação, Conectividade, Curvatura Total, Declividade, Distância até a Cabeceira, SPI (Stream Power Index) e TWI (Topographic Wetness Index). Para a variável de Distância até o Centro do Canal, utiliza-se o MDE para a criação da Nuvem de Pontos e posterior cálculo da distância. Os dados de Precipitação passam por um processo de interpolação para serem espacializados em toda a bacia hidrográfica, resultando na variável de Pluviosidade.

Os valores de todas as onze dimensões são extraídos para cada ponto da bacia, integrando-se para formar o Conjunto de Dados final, que servirá de base para a modelagem de aprendizado de máquina

Ao longo de toda a bacia do Riacho Grande foram espalhados pontos a cada metro quadrado, resultando em um banco de dados com 11 dimensões e pouco mais de 315 milhões de instancias. O quadro 1 caracteriza cada um dos atributos utilizados na modelagem.

Atributo	**Descrição**
Altitude	Refere-se à elevação do terreno em relação ao nível do mar, medida em metros, para cada ponto analisado.
Área de Acumulação	Indica a área acumulada de drenagem desde a cabeceira até o ponto em análise, medida em quilômetros quadrados.
Conectividade	Representa o índice que mensura, em escala de pixel, a conectividade de um ponto com outros trechos da bacia hidrográfica. Este índice varia no intervalo [-∞, +∞], sendo utilizado para avaliar a continuidade do fluxo hídrico.
Curvatura Total	Combina as curvaturas plana e de perfil do terreno, fornecendo uma representação integrada da curvatura do relevo.
Declividade	Mede a inclinação do terreno em graus, indicando a relação angular em relação à horizontal.
Distância até o Centro do Rio	Mensura, em metros, a distância de cada ponto até o centro do canal principal de drenagem.
Distância até a Cabeceira	Refere-se à distância acumulada do ponto analisado até a cabeceira do curso d'água. Este atributo mede a extensão total percorrida pela água ao longo do fluxo, a partir do ponto mais elevado da bacia hidrográfica até o ponto em questão.

Precipitação	Reflete a quantidade média de chuva, em milímetros, estimada para cada um dos pontos com base em uma interpolação realizada a partir de dados de cinco estações pluviométricas (Betânia, Calumbi, Flores, Serra Talhada e Custódia). As informações foram fornecidas pela Agência Pernambucana de Águas e Clima (APAC) e correspondem à média de precipitação dos últimos 30 anos.
SPI (Stream Power Index)	Indica o potencial de erosão exercido pela água corrente sobre o terreno, considerando a declividade e o fluxo acumulado.
TWI (Topographic Wetness Index)	Representa a umidade topográfica potencial, calculada com base na área de contribuição e na inclinação do terreno. É utilizado para avaliar a capacidade de retenção de água em uma área específica.
Aluvião (Saída)	Variável de saída utilizada no modelo. Os pontos identificados como áreas aluvionares foram atribuídos o valor 1, enquanto os pontos não aluvionares receberam o valor 0.

Quadro 1 – Descrição das variáveis utilizadas na modelagem

Dado o extenso volume do banco de dados resultante da etapa de obtenção, a próxima seção detalha o método aplicado para a redução do número de instâncias, visando otimizar o processamento para a modelagem de aprendizado de máquina.

3.3 Redução das instâncias e sua avaliação

Devido ao grande volume de dados, foi necessário aplicar um método de redução de instâncias para viabilizar o processamento e a aplicação de técnicas de aprendizado de máquina. O método escolhido foi o Difusão Geométrica Markoviana estendido - DGM-e (SILVA, 2012), uma abordagem validada na literatura (SILVA; SOUZA; MOTTA, 2016) para a redução de grandes conjuntos de dados, preservando a representatividade das informações originais. O DGM-e foi aplicado para obtenção de diferentes níveis de redução (10%, 5%, 1% e 0,1% do total de instâncias), permitindo a criação de bancos de dados menores, mas ainda representativos.

Para a realização do processamento, foi utilizado um computador com processador Intel Core i7 de 13ª geração, 64 GB de memória RAM e uma placa gráfica dedicada com 8 GB de memória. Além disso, quatro outros computadores de suporte foram empregados, todos equipados com processadores Intel Core i7-4770 CPU @ 3.4 GHz (8 núcleos), 16 GB de memória RAM e gráficos integrados Intel HD Graphics 4600.

Essa configuração conjunta acelerou significativamente o trabalho de redução das instâncias.

O conjunto de dados original possuía aproximadamente 40 GB, que foi dividido em 16 partes. Essa divisão foi essencial para a efetivação da redução. O processo de trabalho para redução está retratado na Figura 8.

Cada uma das 16 partes passou por uma primeira etapa de redução. Subsequentemente, as instâncias foram reunidas em quatro grupos e submetidas à segunda etapa de redução. A etapa final consistiu na unificação dos quatro conjuntos e na realização da terceira e última fase de redução.

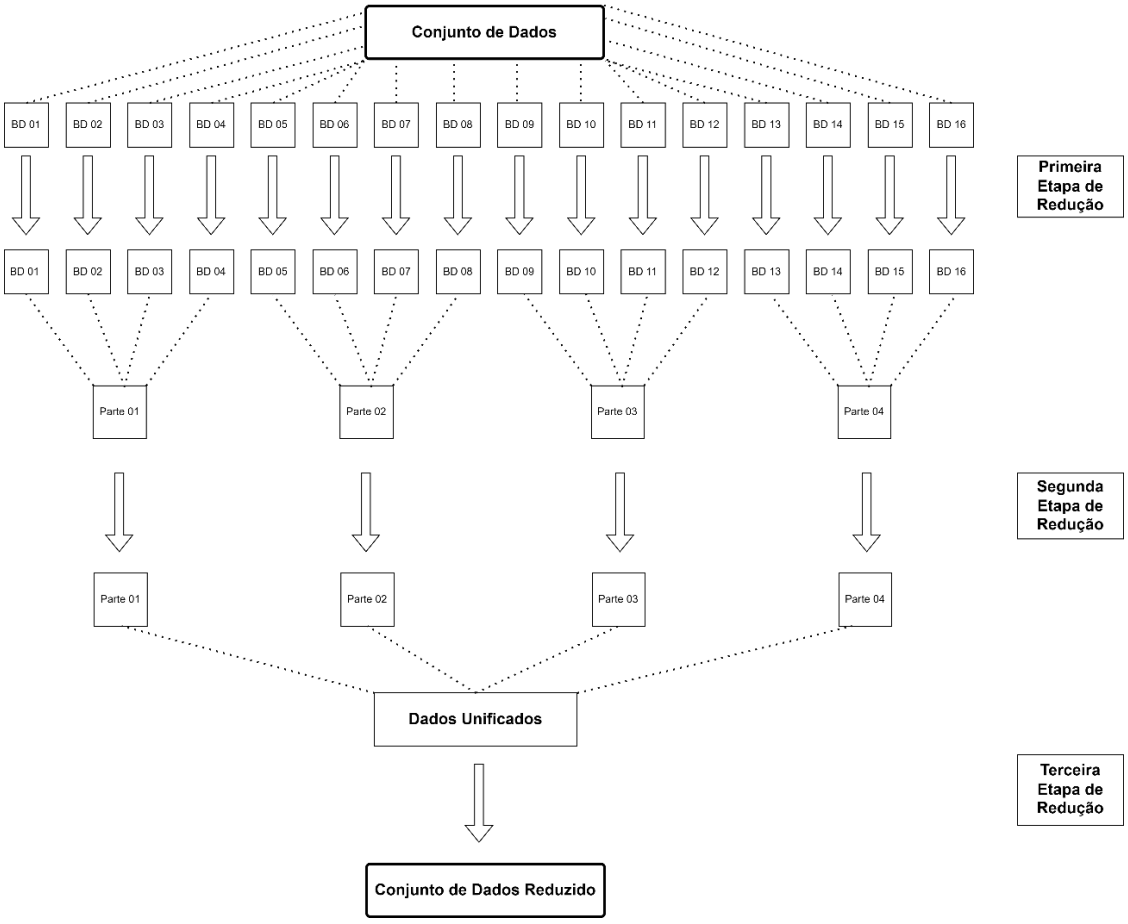


Figura 8 - Etapas para implementação da redução a partir do DGM-e

Foram realizadas reduções no conjunto de dados original, aplicando percentuais totais de 10%, 5%, 1% e 0,01%. Os resultados obtidos serão detalhados nos experimentos descritos no próximo capítulo. O conjunto de dados original possui aproximadamente 315 milhões e 870 mil instancias, o Quadro 2 apresenta o percentual reduzido em cada uma das etapas.

Patamares de Redução (% final alcançado)	% Retido Após a 1ª Etapa	Instâncias Após a 1ª Etapa (Milhões)	% Retido Após a 2ª Etapa	Instâncias Após a 2ª Etapa (Milhões)	% Retido Após a 3ª Etapa	Instâncias Após a 3ª Etapa - Final (Milhões)
0,1%	10,1%	31,9	1,1%	3,5	0,1%	0,3
1%	11%	34,7	2%	6,3	1%	3,2
5%	15%	47,4	6%	19	5%	15,8
10%	20%	63,2	11%	34,7	10%	31,6

Quadro 2 – Percentuais aplicados em cada etapa de redução

Resta esclarecer que, durante nossa aplicação do processo de redução, os dados foram escalonados para o intervalo entre -10 e 10, da seguinte maneira:

$$\hat{x}_{ij} = -10 + \frac{x_{ij-}\, m_j}{M-m} * 20 \qquad (11)$$

onde x_{ij} representa o valor original, $m_j e\ M$ correspondem respectivamente, ao menor e maior valor da variável em questão. Esse escalonamento foi adotado em substituição ao intervalo originalmente proposto (2), de 0 a 1, com o objetivo de ampliar a representatividade dos valores e permitir maior representatividade no momento da construção da vizinhança.

Dado o grande volume do conjunto de dados original, foi a redução que possibilitou a realização do treinamento e dos testes para os testes. O método apresentado buscou, dentro dos recursos computacionais disponíveis, otimizar o tempo necessário para a conclusão de todas as etapas de redução.

3.4 Construção dos Modelos e Detecção de Áreas de Aluvião

Os experimentos para a detecção de áreas de aluvião foram realizados utilizando a linguagem Python, versão 3.13.1, em conjunto com o ambiente de desenvolvimento Spyder 6. A proposta do experimento, cujos resultados serão apresentados no próximo capítulo, consistiu em comparar o desempenho de diferentes técnicas de aprendizado de máquina na identificação de aluviões.

Foram aplicadas as técnicas de CART/DT, K-Vizinhos Mais Próximos e Floresta Aleatória para a detecção de aluviões. Devido ao alto volume de dados, não foi possível testar essas abordagens no banco de dados completo, tornando essencial a aplicação do método de redução. Assim, as análises foram conduzidas em bases reduzidas para 10%, 5%, 1% e 0,1% do total de instâncias, permitindo a comparação do desempenho de cada técnica. O critério principal de avaliação foi a eficácia na classificação das áreas aluvionares, considerando métricas como precisão, recall e F1-score.

A Figura 9 apresenta as etapas metodológicas empregadas na construção e validação dos modelos. O processo iniciou-se com a aplicação de diferentes níveis de redução ao banco de dados original, que continha aproximadamente 315 milhões de pontos. Foram gerados quatro conjuntos de dados reduzidos: 31,5 milhões de pontos (redução para 10%), 15,7 milhões de pontos (redução para 5% do total), 3,15 milhões de pontos (redução para 1%) e 315 mil pontos (redução para 0,1% do total), correspondendo a sucessivas reduções no tamanho do banco.

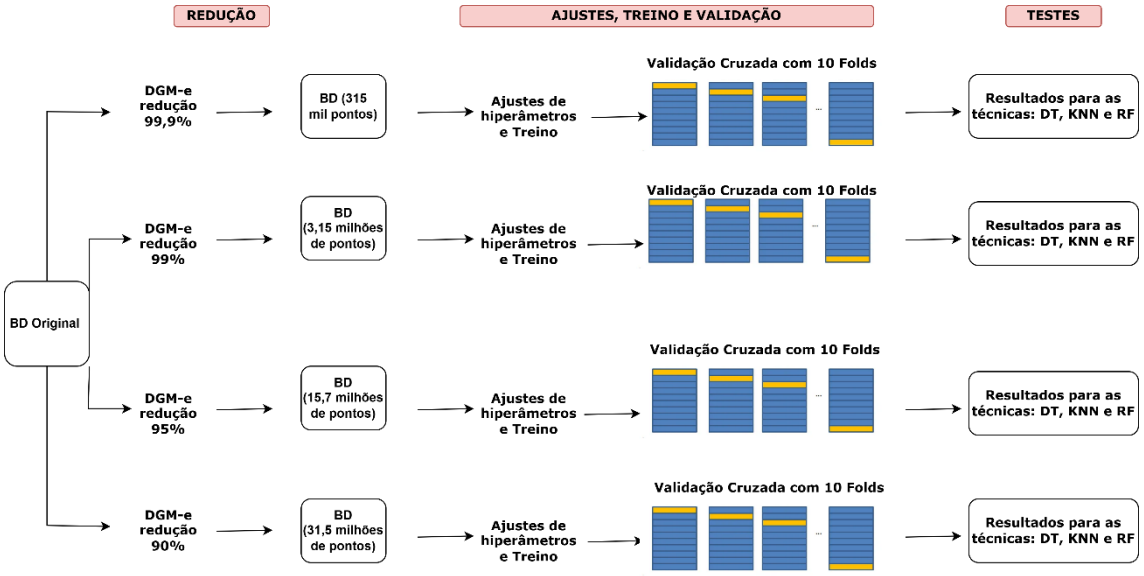


Figura 9 - Etapas de trabalho para aplicação das técnicas de aprendizagem

Cada conjunto reduzido foi submetido a etapas de ajustes, treinamento e validação, empregando validação cruzada com 10 *folds* para garantir consistência na avaliação dos modelos. Para otimizar o desempenho das técnicas utilizadas, CART/DT, KNN e RF, os melhores hiperparâmetros foram selecionados por meio de *grid search* combinado com validação cruzada.

Para o algoritmo de KNN, o grid search identificou que o número ideal de vizinhos (n_neighbors) foi de 7, utilizando a ponderação por distância (weights='distance') e a métrica de distância Manhattan (metric='manhattan'). Esses hiperparâmetros se mostraram consistentes para todos os níveis de redução do conjunto de dados.

No caso do CART/DT, os hiperparâmetros ótimos encontrados através do grid search foram: profundidade máxima da árvore (max_depth) de 20, número mínimo de amostras para dividir um nó (min_samples_split) igual a 5, número mínimo de amostras nas folhas (min_samples_leaf) de 2 e um peso para a classe positiva (class_weight) de 6. Assim como no KNN, essa configuração se manteve a melhor para todas as bases de dados reduzidas.

Finalmente, para o algoritmo RF, o grid search determinou os seguintes hiperparâmetros como os mais eficazes: profundidade máxima da árvore (max_depth) de 25, número mínimo de amostras para dividir um nó (min_samples_split) de 2, número mínimo de amostras nas folhas (min_samples_leaf) de 2, peso para a classe positiva (class_weight) de 5 e o número de árvores na floresta (n_estimators) igual a 50. Esses hiperparâmetros também foram consistentes entre os diferentes níveis de redução dos dados.

Esse processo permitiu identificar as configurações mais adequadas para cada técnica, maximizando a eficácia preditiva dos modelos. Por fim, os resultados de cada etapa foram registrados e comparados, considerando o impacto do nível de redução no desempenho das técnicas.

4 RESULTADO E DISCUSSÃO

O Quadro 3 apresenta os resultados obtidos com a aplicação de três técnicas de classificação (KNN, CART/DT e Random Forest) nos quatro bancos gerados após o processo de redução dos dados.

Não foi possível aplicar as técnicas de aprendizado de máquina ao conjunto total de dados, que continha aproximadamente 315 milhões de pontos, devido ao alto custo computacional envolvido. Dessa forma, a redução dos dados tornou-se essencial para viabilizar a modelagem, permitindo que os experimentos fossem conduzidos dentro do limite da capacidade dos recursos disponíveis. Após alguns testes, o percentual de 10% do total de instancias representou o maior volume de dados que foi possível processar sem comprometer a execução dos modelos, garantindo uma análise consistente sem exceder os limites operacionais.

Técnica Aplicada	**Total de Instâncias**	**F1-Score**	**Precisão**	**Recall**
KNN	0,1%	72,7%	75,1%	70,4%
KNN	1%	66,3%	69,5%	63,5%
KNN	5%	86,1%	87,3%	84,9%
KNN	10%	85,5%	86,7%	84,3%
CART/DT	0,1%	70,3%	73,2%	67,9%
CART/DT	1%	74,4%	76,1%	72,9%
CART/DT	5%	86,3%	88,2%	84,5%
CART/DT	10%	85,3%	87,1%	83,7%
RF	0,1%	75%	78,5%	71,7%
RF	1%	77,4%	80,2%	74,8%
RF	5%	89,8%	91,0%	88,7%
RF	10%	89,0%	90,2%	87,9%

Quadro 3 – Resultado obtido por cada uma das técnicas de AM

Os dados analisados representam quatro diferentes níveis de redução do conjunto original, com proporções de 0,1%, 1%, 5% e 10% do total, permitindo avaliar o impacto da redução na performance dos classificadores. A acurácia não foi utilizada como critério

de avaliação, uma vez que o extremo desbalanceamento dos dados tornaria essa métrica pouco representativa para a qualidade real dos modelos.

Entre as técnicas avaliadas, o KNN (Figura 10) apresentou desempenho inferior na maioria dos casos, especialmente nos bancos mais reduzidos, evidenciando uma sensibilidade maior à disponibilidade de dados.

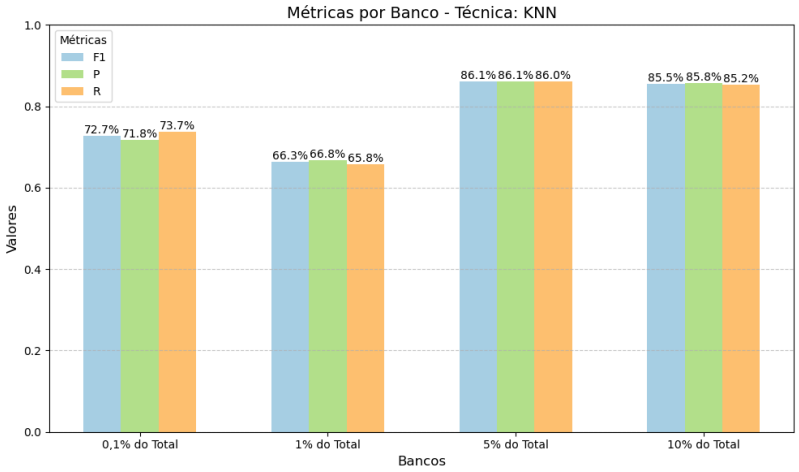


Figura 10 - Resultado da aplicação do KNN

O CART/DT mostrou resultados intermediários, com melhorias progressivas à medida que a proporção de dados aumentava (Figura 11). Isso sugere que a técnica é menos sensível à redução de dados do que o KNN, mas ainda depende de um maior volume de informações para obter resultados mais consistentes.

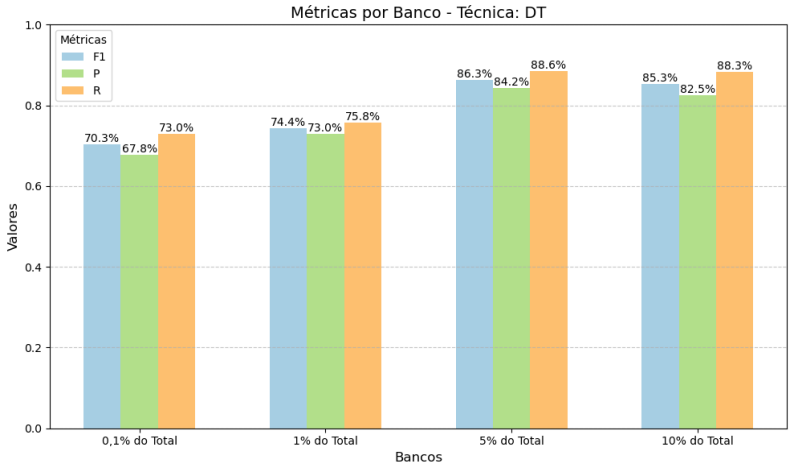


Figura 11 - Resultados da aplicação do CART/DT

Por outro lado, o RF (Figura 12) destacou-se como o melhor classificador em todos os bancos, apresentando maior estabilidade e métricas relativamente elevadas. Os resultados obtidos com RF nos bancos reduzidos para 5% e 10% do total foram próximos, com F1-Score e Precisão superiores a 89% e Recall próximo desse patamar.

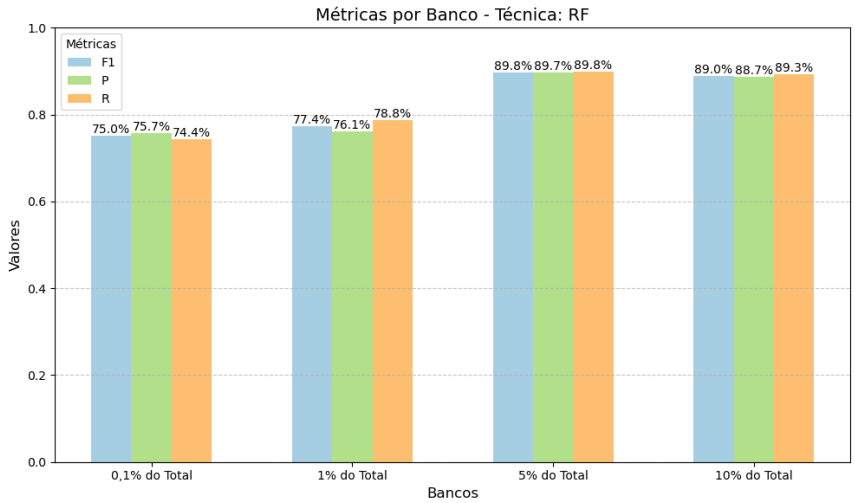


Figura 12 - Resultados obtidos com o RF

A análise dos modelos treinados com 5% e 10% dos dados apontou uma leve vantagem para o modelo de 5%, com pequenas diferenças observadas nas métricas de desempenho. Para verificar essa diferença, foram calculados os intervalos de confiança de 95% para o F1-Score no conjunto de teste final.

Os resultados demonstraram que o modelo treinado com 5% dos dados obteve um F1-Score de 0.8975, com um intervalo de confiança de (0.8972, 0.8978), enquanto o modelo treinado com 10% dos dados apresentou um F1-Score de 0.8897, com intervalo de (0.8895, 0.8899). Como esses intervalos não se sobrepõem, há evidências de que o modelo treinado com 5% dos dados manteve um desempenho levemente superior ao modelo treinado com 10%, reforçando que a redução da quantidade de dados, neste caso, não comprometeu a qualidade preditiva.

Dessa forma, as análises que serão apresentadas a partir deste ponto terão como referência o modelo treinado com 5% dos dados.

Ao comparar os resultados obtidos neste estudo com as métricas apresentadas na literatura, observa-se que as abordagens utilizadas, como Random Forest, alcançam níveis de precisão e F1-Score comparáveis aos relatados em estudos similares, como o de

Rabanaque et al. (2022). Esse estudo obteve uma acurácia média de 85% e um coeficiente Kappa de 0,81 na classificação de elementos fluviais, incluindo áreas aluvionares.

Os resultados do presente estudo, onde o Random Forest atingiu um F1-Score de 89,8%, além de precisão e recall próximos desse patamar indicam desempenhos similares, considerando as diferenças metodológicas e os contextos de aplicação. Isso reforça a aplicabilidade do aprendizado de máquina na identificação de áreas aluvionares, indicando que modelos como o Random Forest podem obter resultados consistentes mesmo com diferentes bases de dados e metodologias.

Com base nos resultados obtidos utilizando o Random Forest no banco com 5% do total, podemos destacar que as variáveis de maior importância no modelo foram a distância ao canal, a altitude e a precipitação. Esses fatores apresentaram os maiores pesos na classificação das áreas aluvionares. Estudos como os de Pipaud e Lehmkuhl (2017) e Babič et al. (2021) destacam a relevância de variáveis geomorfométricas, como declividade e curvatura, que, embora tenham sido consideradas neste estudo, apresentaram menor impacto na modelagem, com valores de importância inferiores aos das variáveis de maior relevância. A Figura 13 ilustra a importância relativa das variáveis utilizadas, evidenciando a hierarquia dos fatores mais influentes na classificação, bem como a menor contribuição de atributos geomorfométricos como declividade e curvatura.

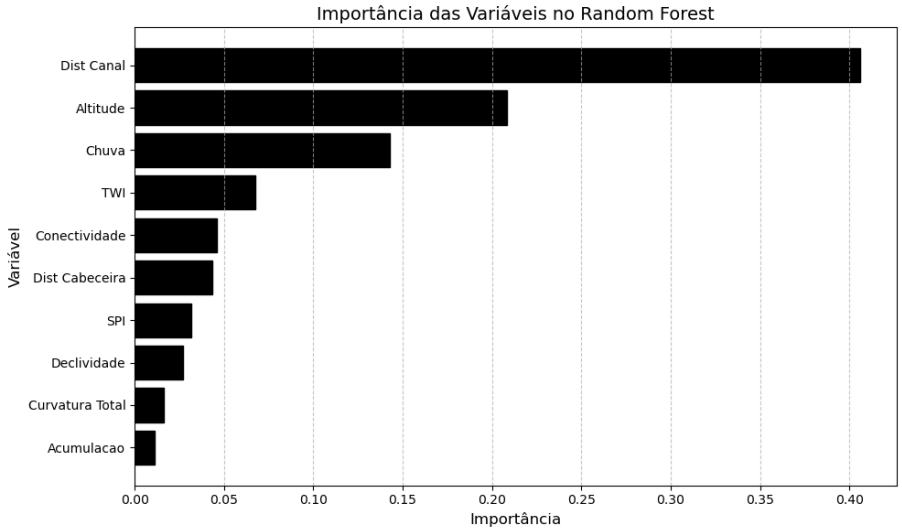


Figura 13 - Importância das variáveis utilizadas na predição

A matriz de confusão gerada para o Random Forest no banco de 5% do total é apresentada na Figura 14. Ela aponta que o modelo foi capaz de classificar corretamente a maior parte das amostras, com 2.898.662 verdadeiros negativos (áreas corretamente

identificadas como não sendo de aluvião) e 211.549 verdadeiros positivos (áreas de aluvião corretamente classificadas). Os erros foram relativamente baixos, com 24.308 falsos positivos (áreas classificadas erroneamente como aluvião) e 24.037 falsos negativos (áreas de aluvião não identificadas pelo modelo).

É importante destacar que o conjunto de dados apresenta um cenário desbalanceado, com uma predominância expressiva de áreas não aluvionares. No total, aproximadamente 92,5% das amostras pertencem à classe 0 (áreas sem aluvião), enquanto apenas 7,5% correspondem à classe 1 (áreas de aluvião). Ainda assim, o Random Forest conseguiu manter uma boa taxa de classificação para ambas as classes, identificando corretamente a maioria das áreas aluvionares e não aluvionares.

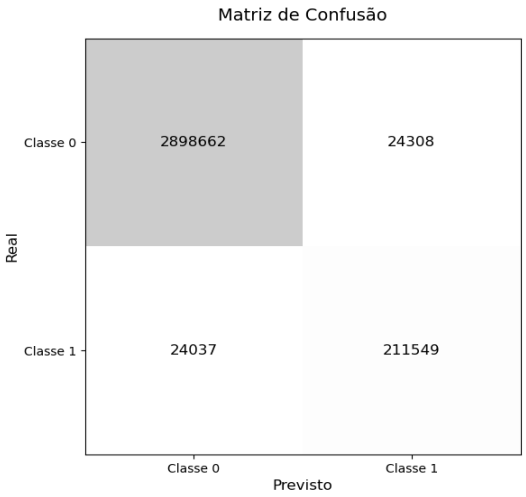


Figura 14 - Matriz de confusão RF (Banco reduzido para 5% do total)

Além disso, a análise da curva Precision-Recall (Figura 15) revela um adequado desempenho do modelo na identificação das áreas de aluvião. A precisão média ponderada pelo recall (Average Precision - AP) obtida foi de 0.97, indicando que o modelo mantém uma alta precisão mesmo cobrindo uma proporção significativa das áreas de aluvião reais.

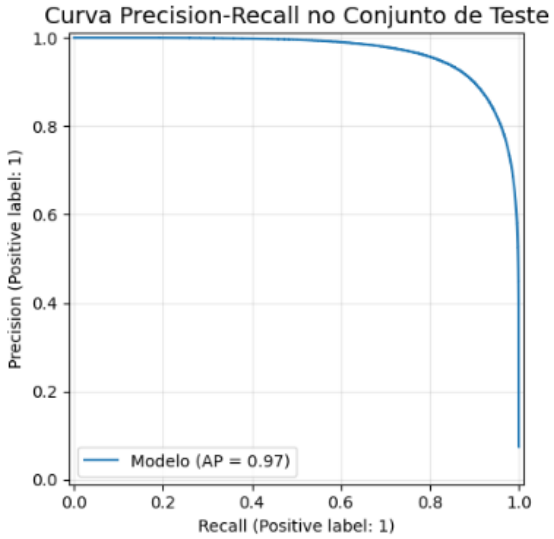


Figura 15 – Curva Precision-Recall (Banco reduzido para 5% do total)

Os resultados alcançados durante a validação cruzada reforçam a estabilidade do modelo Random Forest na tarefa de identificar áreas aluvionares. A Figura 16 apresenta a distribuição das métricas F1-Score, Precisão e Recall ao longo das divisões realizadas na validação cruzada (com k = 10). Observa-se que as métricas mantêm valores consistentes entre os diferentes folds, com baixa variabilidade e medianas elevadas, próximas das médias calculadas.

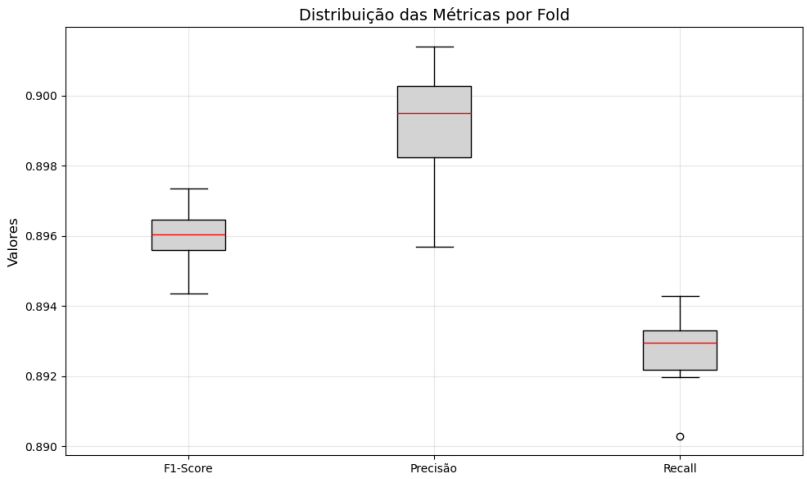
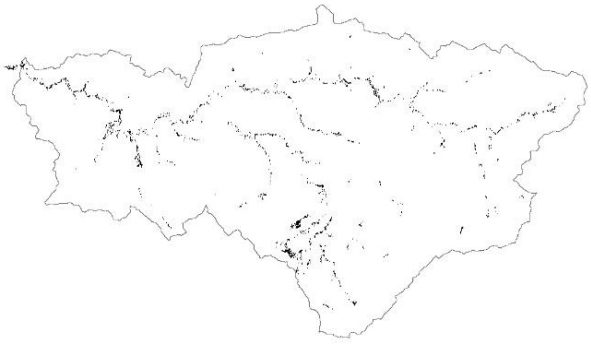


Figura 16 – Validação cruzada resultados (Banco reduzido para 5% do total)

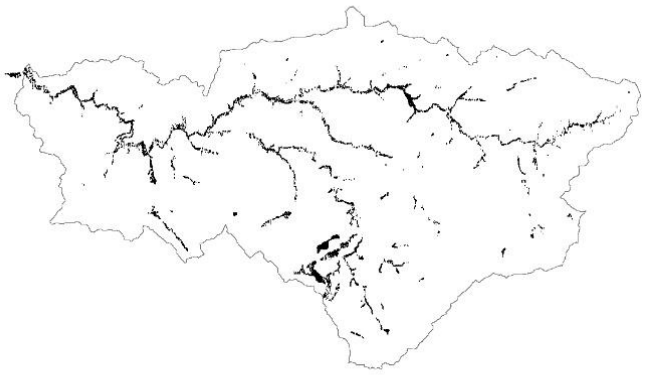
Esses resultados apontam que o modelo é confiável e apresenta alta capacidade de generalização, mesmo ao lidar com diferentes amostras. Assim, o desempenho demonstrado pelo Random Forest valida seu uso como uma ferramenta eficiente para a identificação de áreas aluvionares, com potencial aplicação em estudos geomorfológicos e no planejamento territorial.

Os mapas apresentados na Figura 17 mostram o resultado das diferentes reduções aplicadas às áreas aluvionares, partindo de 0,1% do total (canto superior esquerdo) até 10% do total (canto inferior direito). A figura permite observar a distribuição espacial das áreas aluvionares à medida que diferentes níveis de redução são aplicados, destacando a concentração em torno dos principais canais fluviais.

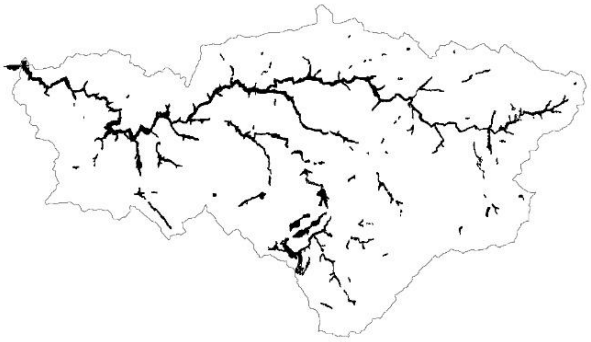
Redução para 0,1% do Total



Redução para 1% do Total



Redução para 5% do Total



Redução para 10% do Total

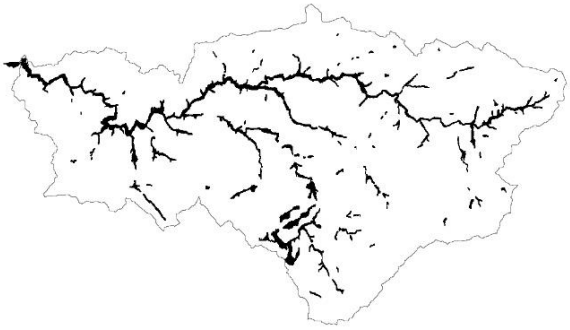


Figura 17- Instâncias aluviais restantes nos diferentes patamares de redução

Nas reduções mais extremas (correspondente a 0,1% e 1%), as áreas aluvionares ficam altamente restritas. Por outro lado, com percentuais de 5% e 10%, observa-se uma expansão gradativa dessas áreas, abrangendo não apenas os rios principais, mas também porções significativas de seus afluentes. Essa variação evidencia a hierarquia fluvial e o papel das áreas aluvionares em trechos de diferentes relevâncias na bacia hidrográfica.

O processo de redução progressiva demonstrado nos mapas é útil para diversos objetivos. Em cenários com alta redução (ex.: 0,1%), o foco é apenas nos trechos mais críticos ou de maior relevância, enquanto as menores reduções (ex.: 10%) permitem uma representação mais abrangente das áreas aluvionares, preservando trechos com menor densidade de drenagem.

A escolha do percentual adequado depende diretamente das necessidades da análise, sejam elas voltadas ao planejamento ambiental, à modelagem hidrológica ou à preservação de habitats associados aos rios.

5 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho realizou uma análise sobre a aplicação de diferentes técnicas de aprendizado de máquina na classificação de áreas aluvionares, considerando conjuntos de dados reduzidos em diferentes proporções. As técnicas utilizadas foram o KNN, CART/DT e RF, aplicadas a bancos com 0,1%, 1%, 5% e 10% dos dados originais. A análise permitiu identificar o impacto da redução no desempenho dos classificadores, destacando que o Random Forest apresentou resultados mais consistentes e robustos em todos os cenários testados.

Os resultados mais expressivos foram obtidos com o Random Forest, onde as métricas de F1-Score e Precisão superaram 89%, e o Recall apresentou um desempenho próximo desse patamar. Isso evidencia a capacidade do modelo em lidar com a complexidade e variabilidade das características geomorfológicas, mesmo em bases reduzidas. Entre as variáveis analisadas, destacaram-se como mais relevantes a distância ao centro do canal, a altitude e a precipitação.

Além disso, a comparação entre os quatro patamares de redução mostrou que a redução do volume de dados não comprometeu significativamente a eficiência dos classificadores. Sugere-se que o DGM-e foi eficaz, oportunizando seu uso em análises futuras, com menor custo computacional e maior agilidade no processamento.

Embora os três objetivos deste estudo tenham sido alcançados, ainda existem possibilidades de aprofundamento. Um dos próximos passos será submeter os bancos de dados a um conjunto mais diversificado de técnicas de aprendizado de máquina, incluindo métodos como Gradient Boosting, Extreme Gradient Boosting (XGBoost) e redes neurais. A aplicação dessas técnicas poderia oferecer uma análise complementar do desempenho em relação às abordagens já testadas, permitindo explorar a capacidade de modelagem em diferentes cenários e identificar potenciais melhorias no processo de classificação.

Outro ponto central para futuros desenvolvimentos é a comparação da redução realizada pelo método DGM-e com outras abordagens para redução de instâncias. Essa análise permitirá avaliar a eficiência do método proposto em relação a alternativas existentes, identificando suas vantagens e limitações em contextos variados. Tal comparação será crucial para consolidar o DGM-e como uma ferramenta útil e confiável em estudos geomorfológicos e aplicações de aprendizado de máquina.

Por fim, este trabalho reforça a importância do uso de técnicas modernas de aprendizado de máquina para a classificação de áreas aluvionares. Além disso, aponta caminhos para otimizações futuras, que poderão ampliar o alcance e a aplicabilidade das metodologias apresentadas, contribuindo para o avanço no entendimento de processos geomorfológicos e no planejamento territorial.

REFERÊNCIAS

ALI, Rayees; SAJJAD, Haroon; SAHA, Tamal Kanti; ROSHANI; MASROOR, Md; RAHAMAN, Md Hibjur. Effectiveness of machine learning ensemble models in assessing groundwater potential in Lidder watershed, India. Acta Geophysica, [S. l.], 2023. DOI: 10.1007/s11600-023-01237-8.

AMIT, R.; HARRISON, J. B. J.; ENZEL ', Y.; PORAT, N. Soils as a tool for estimating ages of Quaternary fault scarps in a hyperarid environment-the southern Arava valley, the Dead Sea Rift, IsraelCATENA ELSEVIER Catena. [s.l: s.n.].

ARDABILI, Sina; MOSAVI, Amir; DEHGHANI, Majid; VARKONYI-KOCZY, Annamaria R. Deep learning and machine learning in hy-drological processes climate change and earth systems a systematic review. [S. l.], 2019. DOI: 10.20944/preprints201908.0166.v1. Disponível em: www.preprints.org.

BABIČ, Matej; PETROVIČ, Dušan; SODNIK, Jošt; SOLDO, Božo; KOMAC, Marko; CHERNIEVA, Olena; KOVAČIČ, Miha; MIKOŠ, Matjaž; CALÌ, Michele. Modeling and classification of alluvial fans with dems and machine learning methods: A case study of Slovenian torrential fans. Remote Sensing, [S. l.], v. 13, n. 9, 2021. DOI: 10.3390/rs13091711.

BRAGA, Ricardo. Águas de Areia. [s.l: s.n.].

BREIMAN, Leo. Random Forests. Machine Learning, [S. l.], v. 45, n. 1, p. 5–32, 2001. DOI: 10.1023/A:1010933404324.

BÚRQUEZ, Alberto; BOJÓRQUEZ OCHOA, Mirsa; MARTÍNEZ-YRÍZAR, Angelina; DE SOUZA, Jonas Otaviano Praça. Human-made small reservoirs alter dryland hydrological connectivity. Science of The Total Environment, [S. l.], v. 947, p. 174673, 2024. DOI: 10.1016/j.scitotenv.2024.174673.

CAVALLI, M.; MARCHI, L. Characterisation of the surface morphology of an alpine alluvial fan using airborne LiDAR. Hazards Earth Syst. Sci. [s.l: s.n.]. Disponível em: www.nat-hazards-earth-syst-sci.net/8/323/2008/.

CAVALLI, Marco; TAROLLI, Paolo; MARCHI, Lorenzo; DALLA FONTANA, Giancarlo. The effectiveness of airborne LiDAR data in the recognition of channel-bed morphology. Catena, [S. l.], v. 73, n. 3, p. 249–260, 2008. DOI: 10.1016/j.catena.2007.11.001.

CERVI, Federico; TAZIOLI, Alberto. Quantifying Streambed Dispersion in an Alluvial Fan Facing the Northern Italian Apennines: Implications for Groundwater Management of Vulnerable Aquifers. Hydrology, [S. l.], v. 8, n. 3, p. 118, 2021. DOI: 10.3390/hydrology8030118.

CROUVI, Onn; BEN-DOR, Eyal; BEYTH, Michael; AVIGAD, Dov; AMIT, Rivka. Quantitative mapping of arid alluvial fan surfaces using field spectrometer and hyperspectral remote sensing. Remote Sensing of Environment, [S. l.], v. 104, n. 1, p. 103–117, 2006. DOI: 10.1016/j.rse.2006.05.004.

DÍAZ-ALCAIDE, S.; MARTÍNEZ-SANTOS, P. Review: Advances in groundwater potential mapping. Hydrogeology JournalSpringer Verlag, , 2019. DOI: 10.1007/s10040-019-02001-3.

EL BILALI, Ali; TALEB, Abdeslam; BROUZIYNE, Youssef. Comparing four machine learning model performances in forecasting the alluvial aquifer level in a semi-arid region. Journal of African Earth Sciences, [S. l.], v. 181, 2021. DOI: 10.1016/j.jafrearsci.2021.104244.

FACELLI, Katti; LORENA, Ana Carolina; GAMA, João; ALMEIDA, Tiago Agostinho De; CARVALHO, André Carlos Ferreira. Inteligência Artificial: Uma abordagem de Aprendizado de Máquina. 2. ed. Rio de Janeiro: LTC, 2021.

FARR, Tom G.; CHADWICK, Oliver A. Geomorphic processes and remote sensing signatures of alluvial fans in the Kun Lun Mountains, China. Journal of Geophysical Research: Planets, [S. l.], v. 101, n. E10, p. 23091–23100, 1996. DOI: 10.1029/96JE01603.

FRANKEL, Kurl L.; DOLAN, James F. Characterizing arid region alluvial fan surface roughness with airborne laser swath mapping digital topographic data. Journal of Geophysical Research: Earth Surface, [S. l.], v. 112, n. 2, 2007. DOI: 10.1029/2006JF000644.

FREEZE, R. A.; CHERRY, J. A. Groundwater. Englewood Cliffs: Prentice-Hall, 1979.

GABER, Ahmed; KOCH, Magaly; EL-BAZ, Farouk. Textural and compositional characterization of Wadi Feiran deposits, Sinai Peninsula, Egypt, using Radarsat-1, PALSAR, SRTM and ETM+ data. Remote Sensing, [S. l.], v. 2, n. 1, p. 52–75, 2010. DOI: 10.3390/rs2010052.

GHOLAMI, V.; KHALEGHI, M. R.; TEIMOURI, M.; SAHOUR, H. Prediction of annual groundwater depletion: An investigation of natural and anthropogenic influences. Journal of Earth System Science, [S. l.], v. 132, n. 4, 2023. DOI: 10.1007/s12040-023-02184-0.

GILLESPIE, Alan R.; KAHLE, Anne B.; PALLUCONI, Frank D. MAPPING ALLUVIAL FANS IN DEATH VALLEY, CALIFORNIA, USING MULTICHANNEL THERMAL INFRARED IMAGES. [s.l: s.n.].

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. The Elements of Statistical Learning. New York, NY: Springer New York, 2009. DOI: 10.1007/978-0-387-84858-7.

HETZ, Guy; MUSHKIN, Amit; BLUMBERG, Dan G.; BAER, Gidon; GINAT, Hanan. Estimating the age of desert alluvial surfaces with spaceborne radar data. Remote Sensing of Environment, [S. l.], v. 184, p. 288–301, 2016. DOI: 10.1016/j.rse.2016.07.006.

HOHENTHAL, Johanna; ALHO, Petteri; HYYPPÄ, Juha; HYYPPÄ, Hannu. Laser scanning applications in fluvial studies. Progress in Physical Geography, [S. l.], v. 35, n. 6, p. 782–809, 2011. DOI: 10.1177/0309133311414605.

JASECHKO, Scott; SEYBOLD, Hansjörg; PERRONE, Debra; FAN, Ying; SHAMSUDDUHA, Mohammad; TAYLOR, Richard G.; FALLATAH, Othman; KIRCHNER, James W. Rapid groundwater decline and some cases of recovery in aquifers globally. Nature, [S. l.], v. 625, n. 7996, p. 715–721, 2024. DOI: 10.1038/s41586-023-06879-8.

KAYHOMAYOON, Zahra; GHORDOYEE-MILAN, Sami; JAAFARI, Abolfazl; ARYA-AZAR, Naser; MELESSE, Assefa M.; KARDAN MOGHADDAM, Hamid. How does a combination of numerical modeling, clustering, artificial intelligence, and evolutionary algorithms perform to predict regional groundwater levels? Computers and Electronics in Agriculture, [S. l.], v. 203, 2022. DOI: 10.1016/j.compag.2022.107482.

LUIZ, Thiago Boeno Patricio. Previsão de Dados de Água Subterrânea utilizando modelos baseados em Aprendizado de Máquina. [s.l: s.n.].

MARSLAND, Stephen. Machine Learning: an algorithmic Perspective. Boca Raton: Taylor & Francis Group, LLC, 2015.

MARTÍNEZ-SANTOS, P.; RENARD, P. Mapping Groundwater Potential Through an Ensemble of Big Data Methods. Groundwater, [S. l.], v. 58, n. 4, p. 583–597, 2020. DOI: 10.1111/gwat.12939.

MCLEOD, Jonah S. et al. Landscapes on the edge: River intermittency in a warming world. Geology, [S. l.], v. 52, n. 7, p. 512–516, 2024. DOI: 10.1130/G52043.1.

MITCHELL, Tom M. (Tom Michael). Machine Learning. [s.l: s.n.].

MOORE, Andrew W. An intoductory tutorial on kd-trees. 1991. Tese de Doutorado - Universith of Cambridge, [S. l.], 1991.

MUÑOZ-CARPENA, Rafael; CARMONA-CABRERO, Alvaro; YU, Ziwen; FOX, Garey; BATELAAN, Okke. Convergence of mechanistic modeling and artificial intelligence in hydrologic science and engineering. PLOS Water, [S. l.], v. 2, n. 8, p. e0000059, 2023. DOI: 10.1371/journal.pwat.0000059.

NGUYEN, Phong Tung et al. Improvement of credal decision trees using ensemble frameworks for groundwater potential modeling. Sustainability (Switzerland), [S. l.], v. 12, n. 7, 2020. DOI: 10.3390/su12072622.

PIPAUD, Isabel; LEHMKUHL, Frank. Object-based delineation and classification of alluvial fans by application of mean-shift segmentation and support vector machines. Geomorphology, [S. l.], v. 293, p. 178–200, 2017. DOI: https://doi.org/10.1016/j.geomorph.2017.05.013. Disponível em: https://www.sciencedirect.com/science/article/pii/S0169555X16310364.

RABANAQUE, Maria Pilar; MARTÍNEZ-FERNÁNDEZ, Vanesa; CALLE, Mikel; BENITO, Gerardo. Basin-wide hydromorphological analysis of ephemeral streams using machine learning algorithms‡. Earth Surface Processes and Landforms, [S. l.], v. 47, n. 1, p. 328–344, 2022. DOI: 10.1002/esp.5250.

REGMI, Netra R.; MCDONALD, Eric V.; BACON, Steven N. Mapping Quaternary alluvial fans in the southwestern United States based on multiparameter surface roughness of lidar topographic data. Journal of Geophysical Research: Earth Surface, [S. l.], v. 119, n. 1, p. 12–27, 2014. DOI: 10.1002/2012JF002711.

RITCHIE, Hannah; EISMA, Jessica A.; PARKER, Alison. Sand Dams as a Potential Solution to Rural Water Security in Drylands: Existing Research and Future Opportunities. Frontiers in Water, [S. l.], v. 3, 2021. DOI: 10.3389/frwa.2021.651954.

RUSSELL, Stuart J. (Stuart Jonathan); NORVIG, Peter; DAVIS, Ernest. Artificial intelligence : a modern approach. [s.l: s.n.].

SEIFU, Tesema Kebede; ESHETU, Kidist Demessie; WOLDESENBET, Tekalegn Ayele; ALEMAYEHU, Taye; AYENEW, Tenalem. Application of advanced machine learning algorithms and geospatial techniques for groundwater potential zone mapping in Gambela Plain, Ethiopia. Hydrology Research, [S. l.], v. 54, n. 10, p. 1246–1266, 2023. DOI: 10.2166/nh.2023.083.

SHAKYA, Chandra Mohan; BHATTACHARJYA, Rajib Kumar; DADHICH, Sharad. Groundwater level prediction with machine learning for the Vidisha district, a semi-arid region of Central India. Groundwater for Sustainable Development, [S. l.], v. 19, 2022. DOI: 10.1016/j.gsd.2022.100825.

SILVA, Adonai Felipe Pereira De Lima; SOUZA, Jonas Praça. Análise da qualidade da agua nos aquíferos aluviais da bacia Riacho do Tigre-PB: Uma abordagem hidrológica em ambientes fluviais semiáridos no Brasil Water Quality Analysis of Alluvial Aquifers in the Riacho do Tigre Basin-PB: A Hydrological Approach in a Semi-Arid Fluvial Environment in Brazil. [s.l: s.n.].

SILVA, Duílio Assis Nobre. Um Método Geral Para Redução de Grandes Conjuntos de Dados Baseados em Difusão Geométrica Markoviana. 2012. Dissertação de Mestrado - UFPB, João Pessoa, 2012.

SILVA, Duílio Assis Nobre; SOUZA, Leandro Carlos; MOTTA, Gustavo Henrique Matos Bezerra. An instance selection method for large datasets based on Markov geometric diffusion. Data & Knowledge Engineering, [S. l.], v. 101, p. 24-41, 2016.

SOUZA, Leandro Carlos De. Simplificação de Malhas com Preservação de Feições Baseada em Difusão Geométrica Markoviana. 2011. Dissertação de Mestrado - PUC-Rio, Rio de Janeiro, 2011.

SRIVASTAVA, Devesh Kumar; SHUKLA, Aditi; JEMNI, Divyans. Prediction of Ground Water Level in Rajasthan State Using Machine Learning. Em: PROCEDIA COMPUTER SCIENCE 2022, Anais [...]. : Elsevier B.V., 2022. p. 1702–1711. DOI: 10.1016/j.procs.2023.01.148.

STALEY, Dennis M.; WASKLEWICZ, Thad A.; BLASZCZYNSKI, Jacek S. Surficial patterns of debris flow deposition on alluvial fans in Death Valley, CA using airborne laser swath mapping data. Geomorphology, [S. l.], v. 74, n. 1–4, p. 152–163, 2006. DOI: 10.1016/j.geomorph.2005.07.014.

TAO, Hai et al. Groundwater level prediction using machine learning models: A comprehensive review. NeurocomputingElsevier B.V., , 2022. DOI: 10.1016/j.neucom.2022.03.014.

TAYER, Thiaggo C.; BEESLEY, Leah S.; DOUGLAS, Michael M.; BOURKE, Sarah A.; MEREDITH, Karina; MCFARLANE, Don. Identifying intermittent river sections with similar hydrology using remotely sensed metrics. Journal of Hydrology, [S. l.], v. 626, 2023. DOI: 10.1016/j.jhydrol.2023.130266.

UC-CASTILLO, José Luis; MARÍN-CELESTINO, Ana Elizabeth; MARTÍNEZ-CRUZ, Diego Armando; TUXPAN-VARGAS, José; RAMOS-LEAL, José Alfredo. A systematic review and meta-analysis of groundwater level forecasting with machine learning techniques: Current status and future directions. Environmental Modelling and SoftwareElsevier Ltd, , 2023. DOI: 10.1016/j.envsoft.2023.105788.

VADIATI, Meysam; RAJABI YAMI, Zahra; ESKANDARI, Effat; NAKHAEI, Mohammad; KISI, Ozgur. Application of artificial intelligence models for prediction of groundwater level fluctuations: case study (Tehran-Karaj alluvial aquifer). Environmental Monitoring and Assessment, [S. l.], v. 194, n. 9, 2022. DOI: 10.1007/s10661-022-10277-4.

ZARESEFAT, Mojtaba; DERAKHSHANI, Reza. Revolutionizing Groundwater Management with Hybrid AI Models: A Practical Review. Water, [S. l.], v. 15, n. 9, p. 1750, 2023. DOI: 10.3390/w15091750.