

Detecção Automática de Anomalias Baseada em Difusão Geométrica Markoviana

Erikson Carlos Ramos



CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

João Pessoa, 2025

Erikson Carlos Ramos

Detecção Automática de Anomalias Baseada em Difusão Geométrica Markoviana

Dissertação apresentada ao
Programa de Pós-Graduação em Informática - PPGI
do Centro de Informática, da Universidade Federal da Paraíba,
como requisito parcial para a obtenção do grau de Mestre em Informática

Orientador: Leandro Carlos de Souza

Coorientador: Gustavo Henrique Matos Bezerra Motta

Dezembro de 2025

Catálogo na publicação
Seção de Catalogação e Classificação

R175d Ramos, Erikson Carlos.

Detecção automática de Anomalias baseada em Difusão Geométrica Markoviana / Erikson Carlos Ramos. - João Pessoa, 2025.

93 f. : il.

Orientação: Leandro Carlos de Souza.

Coorientação: Gustavo Matos Bezerra Motta.

Dissertação (Mestrado) - UFPB/CI.

1. Informática. 2. Difusão Geométrica Markoviana. 3. Detecção automática Anomalias. I. Souza, Leandro Carlos de. II. Motta, Gustavo Matos Bezerra. III. Título.

UFPB/BC

CDU 004(043)



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



Ata da Sessão Pública de Defesa de Dissertação de Mestrado de **ERIKSON CARLOS RAMOS**, candidato ao título de Mestre em Informática na área de Ciências da Computação, realizada em 29 de agosto de 2025.

1 Aos vinte e nove dias do mês de agosto do ano de dois mil e vinte e cinco, às 11h, via
2 videoconferência, reuniram-se os membros da Banca Examinadora constituída para julgar o
3 Trabalho Final do discente **ERIKSON CARLOS RAMOS**, vinculado a Universidade Federal da
4 Paraíba sob a matrícula nº 20231003776, candidato ao grau de Mestre em Informática, na
5 área de “Ciências da Computação”, na linha de pesquisa “Metodologia e Técnicas de
6 Computação” do Programa de Pós-Graduação em Informática. A comissão examinadora foi
7 composta pelos professores Leandro Carlos de Souza, orientador e presidente da banca;
8 Claurton de Albuquerque Siebra, examinador interno ao programa; Gustavo Henrique Matos
9 Bezerra Motta, examinador interno ao programa; Bruno Cesar Barreto de Figueiredo,
10 examinador externo à Instituição; e Leandro Aparecido Simal Moreira, examinador externo à
11 Instituição. Dando início aos trabalhos, o presidente cumprimentou os presentes, comunicou
12 a finalidade da reunião e passou a palavra ao candidato para que fizesse, oralmente, a
13 exposição do trabalho de dissertação intitulado “**Deteção Automática de Anomalias**
14 **Baseada em Difusão Geométrica Markoviana**”. Concluída a exposição, o candidato foi
15 arguido pela Banca Examinadora, que emitiu o seguinte parecer: “**aprovado**”. Do ocorrido,
16 eu, Gean Paulo P. M. de Barros, secretário, lavrei a presente ata que vai assinada por mim e
17 pelos membros da Banca Examinadora. João Pessoa, 29 de agosto de 2025.

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

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

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

Prof. Dr. Gustavo Henrique Matos Bezerra Motta
Examinador interno ao Programa (PPGI)

Prof. Dr. Bruno Cesar Barreto de Figueiredo
Examinador externo à Instituição (UERR)

Leandro Aparecido Simal Moreira
Examinador externo à Instituição (CASNAV)

Documento assinado digitalmente



LEANDRO CARLOS DE SOUZA
Data: 02/09/2025 13:22:29-0300
Verifique em <https://validar.iti.gov.br>

Documento assinado digitalmente



CLAURTON DE ALBUQUERQUE SIEBRA
Data: 02/09/2025 13:32:52-0300
Verifique em <https://validar.iti.gov.br>

Documento assinado digitalmente



GUSTAVO HENRIQUE MATOS BEZERRA MOTTA
Data: 03/09/2025 14:55:54-0300
Verifique em <https://validar.iti.gov.br>

Documento assinado digitalmente



BRUNO CESAR BARRETO DE FIGUEIREDO
Data: 04/09/2025 08:45:40-0300
Verifique em <https://validar.iti.gov.br>

Documento assinado digitalmente



LEANDRO APARECIDO SIMAL MOREIRA
Data: 04/09/2025 10:36:39-0300
Verifique em <https://validar.iti.gov.br>

A mente que se abre a uma nova ideia jamais voltará ao seu tamanho original.

Albert Einstein

DEDICATÓRIA

Dedico este trabalho à minha esposa, Damaris, e à minha filha, Antonella, cujo amor e apoio incondicionais iluminaram os dias mais desafiadores desta jornada acadêmica. Aos meus pais, Lenubia e Dikson, e aos meus irmãos, pela paciência e empatia; à minha sogra, Nadir, e a todos os demais familiares e amigos, pela força constante ao longo da minha vida. Vocês são meu alicerce e minha inspiração diária, responsáveis pela construção de um lar cheio de amor e acolhimento. Esta conquista acadêmica é uma homenagem ao vínculo extraordinário que compartilhamos. Agradeço profundamente por toda dedicação, compreensão e incentivo que tornaram possível este trabalho. Amo vocês do fundo do meu coração e dedico esta realização à nossa trajetória conjunta e aos valores que moldam nossa existência.

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus pela saúde, força e pela oportunidade de estudar e desenvolver este trabalho.

À minha esposa, Damaris, sou profundamente grato pela compreensão, paciência, apoio incondicional e motivação constante ao longo desta jornada acadêmica e da vida.

À minha filha, Antonella, minha pequena inspiração diária, que com apenas 1 ano e 8 meses enche meus dias de alegria, especialmente nos momentos em que, com ternura, vem ao meu encontro enquanto estou no computador.

Ao professor e orientador, Leandro Carlos de Souza, agradeço profundamente pela paciência, confiança, dedicação, pelas reuniões esclarecedoras e pelo apoio fundamental no desenvolvimento desta pesquisa.

Ao professor e coorientador, Gustavo Motta, agradeço pela disponibilidade, pelas reuniões produtivas e pelas valiosas contribuições durante as revisões necessárias à conclusão deste trabalho.

Aos professores Claurton Siebra e Leandro Moreira, expresso minha sincera gratidão pelas valiosas contribuições e comentários construtivos oferecidos tanto no exame de qualificação quanto na banca de defesa desta dissertação de mestrado. Ao professor Bruno Figueirêdo, agradeço pelos comentários pertinentes e sugestões enriquecedoras durante a banca de defesa.

Ao amigo Rafael Spíndola, agradeço pelas contribuições e pelo suporte que impulsionaram as fases iniciais desta pesquisa.

Aos colegas do mestrado, agradeço pelas conversas enriquecedoras e pelo apoio mútuo ao longo de toda a pós-graduação.

Ao Tribunal de Contas do Estado da Paraíba, agradeço pelo apoio institucional e pela flexibilização concedida nos momentos necessários para a realização deste mestrado. Minha gratidão especial ao chefe Adjalton Souza, pela compreensão e suporte, e aos colegas de trabalho pelo companheirismo e incentivo constante, com destaque para José Eronildo (Eron) e os demais colegas de sala, cujo apoio diário foi fundamental nesta jornada.

À Universidade Federal da Paraíba e a todos os professores do Programa de Pós-Graduação em Informática, sou grato pela significativa contribuição para minha formação acadêmica e profissional.

Estes agradecimentos representam minha mais sincera gratidão a todos que contribuíram para o êxito deste trabalho e para o meu desenvolvimento profissional, pessoal e acadêmico.

RESUMO

A detecção automática de anomalias é essencial em domínios como saúde, finanças e cibersegurança, nos quais pequenos desvios podem indicar fraude, falha ou risco iminente. Este estudo apresenta um método não supervisionado denominado Detecção Automática de Anomalias baseada em Difusão Geométrica Markoviana (DA-DGM), aplicável a conjuntos de dados unidimensionais e multidimensionais. A técnica usa difusão geométrica para identificar relações não lineares entre instâncias; em dados multidimensionais, o parâmetro de escala é ajustado por entropia de Shannon mediante busca global seguida de refinamento local, garantindo adaptabilidade ao contexto dos dados. A principal contribuição consiste em integrar a sensibilidade da Difusão Geométrica Markoviana ao ajuste automático do parâmetro de escala baseado em entropia de Shannon, visando eliminar a dependência de ajuste manual de hiperparâmetros e fornecer uma métrica de anomalia de interpretação direta baseada na probabilidade de auto-transição. Em avaliações com dados sintéticos e reais, o DA-DGM iguala ou supera métodos clássicos como KNN (*k-Nearest Neighbors*) e LOF (*Local Outlier Factor*), além de apresentar desempenho competitivo frente ao IForest (*Isolation Forest*) em termos de área sob a curva ROC (AUC-ROC), com destaque para cenários de geometria complexa e baixa prevalência de anomalias. Esses resultados evidenciam o potencial do DA-DGM em aplicações críticas e reforçam a utilidade de abordagens difusivas na mineração de dados.

Palavras-chave: detecção automática de anomalias; difusão geométrica Markoviana; aprendizado não supervisionado.

ABSTRACT

Automatic anomaly detection is essential in domains such as healthcare, finance, and cybersecurity, where small deviations may indicate fraud, failure, or imminent risk. This study introduces an unsupervised method called Anomaly Detection based on Markovian Geometric Diffusion (DA-DGM), applicable to one-dimensional and multi-dimensional datasets. The technique employs geometric diffusion to identify nonlinear relationships among instances; in multi-dimensional data, the scale parameter is tuned via Shannon entropy using a global search followed by local refinement, ensuring adaptability to the data context. The main contribution lies in integrating the sensitivity of Markovian geometric diffusion with automatic scale-parameter tuning based on Shannon entropy, thereby eliminating reliance on manual hyperparameter adjustment and providing a directly interpretable anomaly metric grounded in self-transition probability. In evaluations on synthetic and real datasets, DA-DGM matches or surpasses classical methods such as KNN (k -Nearest Neighbors) and LOF (Local Outlier Factor), and achieves competitive performance against IForest (Isolation Forest) in terms of area under the ROC curve (AUC-ROC), with particular advantages in scenarios of complex geometry and low anomaly prevalence. These results highlight DA-DGM’s potential for critical applications and reinforce the usefulness of diffusion-based approaches in data mining.

Keywords: automatic anomaly detection; Markovian geometric diffusion; unsupervised learning.

LISTA DE FIGURAS

1	Exemplo de grafo simples não orientado.	14
2	Exemplo de grafo simples não orientado e ponderado.	15
3	Exemplo de grafo orientado e não ponderado.	15
4	Espectros dos autovalores da matriz de transição construída com núcleo gaussiano para $\varepsilon = 10^{-1}, 1, 10$ e 100 (leitura: esquerda para a direita, de cima para baixo). O eixo horizontal indica o índice dos autovalores ordenados por magnitude, enquanto o eixo vertical apresenta a respectiva magnitude λ	22
5	Representação visual do modelo <i>bunny</i> com pontos coloridos conforme sua importância.	25
6	Processo de Redução de Dados.	26
7	Resultados para as distribuições de dados unidimensionais <i>Gaussiana</i> , <i>Beta</i> e <i>Gama</i> . Linha 1: histogramas das amostras, em que barras azuis representam instâncias normais e vermelhas indicam anomalias. Linha 2: <i>boxplots</i> dos percentis do ranqueamento da probabilidade de auto-transição obtidos pelo DA-DGM (azul = normal, vermelho = anômalo). Linha 3: curvas <i>ROC</i> com as áreas sob a curva (AUC) dos algoritmos avaliados, permitindo comparar seu desempenho discriminatório.	45
8	Resultados para os conjuntos de dados bidimensionais <i>Anel</i> , <i>Cantos</i> e <i>Curvas</i> . Linha 1: curvas $\varepsilon \times$ entropia de Shannon utilizadas para selecionar o parâmetro de escala. Linha 2: gráficos de dispersão das instâncias classificadas pelo DA-DGM (azul = normal, vermelho = anômalas). Linha 3: <i>boxplots</i> dos percentis do ranqueamento da probabilidade de auto-transição. Linha 4: curvas <i>ROC</i> com as áreas sob a curva (AUC) dos algoritmos avaliados, permitindo comparar seu desempenho discriminatório.	47
9	Resultados para os conjuntos de dados bidimensionais <i>Espiral</i> , <i>Gaussiano</i> e <i>Ondas</i> . Linha 1: curvas $\varepsilon \times$ entropia de Shannon utilizadas para selecionar o parâmetro de escala. Linha 2: gráficos de dispersão das instâncias classificadas pelo DA-DGM (azul = normal, vermelho = anômalas). Linha 3: <i>boxplots</i> dos percentis do ranqueamento da probabilidade de auto-transição. Linha 4: curvas <i>ROC</i> com as áreas sob a curva (AUC) dos algoritmos avaliados, permitindo comparar seu desempenho discriminatório.	48

10	Resultados para os conjuntos de dados tridimensionais <i>Cantos 3D</i> , <i>Espiral 3D</i> e <i>Curva Helicoidal</i> . Linha 1: curvas $\varepsilon \times$ entropia de Shannon utilizadas para selecionar o parâmetro de escala. Linha 2: gráficos de dispersão tridimensionais das instâncias classificadas pelo DA-DGM (azul = normal, vermelho = anômalas). Linha 3: <i>boxplots</i> dos percentis do ranqueamento da probabilidade de auto-transição. Linha 4: curvas <i>ROC</i> com as áreas sob a curva (AUC) dos algoritmos avaliados, permitindo comparar o desempenho discriminatório em cada cenário.	50
11	Resultados para os conjuntos de dados reais <i>Ionosphere</i> , <i>Letter</i> e <i>MNIST</i> . Linha 1: curvas $\varepsilon \times$ entropia de Shannon usadas para selecionar o parâmetro de escala. Linha 2: <i>boxplots</i> dos percentis do ranqueamento da probabilidade de auto-transição (DA-DGM). Linha 3: curvas <i>ROC</i> com as áreas sob a curva (AUC) dos algoritmos avaliados.	54
12	Resultados para os conjuntos de dados reais <i>Pima</i> , <i>Vowels</i> e <i>FMSJP 2016</i> . Linha 1: curvas $\varepsilon \times$ entropia de Shannon usadas para selecionar o parâmetro de escala. Linha 2: <i>boxplots</i> dos percentis do ranqueamento da probabilidade de auto-transição (DA-DGM). Linha 3: curvas <i>ROC</i> com as áreas sob a curva (AUC) dos algoritmos avaliados.	55
13	Resultados para os conjuntos de dados reais <i>FMSJP 2017</i> e <i>FMSJP 2018</i> . Linha 1: curvas $\varepsilon \times$ entropia de Shannon usadas para selecionar o parâmetro de escala. Linha 2: <i>boxplots</i> dos percentis do ranqueamento da probabilidade de auto-transição (DA-DGM). Linha 3: curvas <i>ROC</i> com as áreas sob a curva (AUC) dos algoritmos avaliados.	56
14	Resultados para os conjuntos de dados reais <i>FMSJP 2019</i> e <i>FMSJP 2020</i> . Linha 1: curvas $\varepsilon \times$ entropia de Shannon usadas para selecionar o parâmetro de escala. Linha 2: <i>boxplots</i> dos percentis do ranqueamento da probabilidade de auto-transição (DA-DGM). Linha 3: curvas <i>ROC</i> com as áreas sob a curva (AUC) dos algoritmos avaliados.	57
15	Painel de empenhos sinalizados como anômalos pelo DA-DGM (FMSJP, exercícios de 2016 a 2018). Exemplos: Linha 1 (2016): 392733 e 390551; Linha 2 (2017): 393598, 392484, 391311 e 390062; Linha 3 (2018): 392588, 391460 e 390819.	63
16	Painel de empenhos sinalizados como anômalos pelo DA-DGM (FMSJP, exercícios de 2019 e 2020). Exemplos: Linha 1 (2019): 393919, 392459, 391208, 390061; Linha 2 (2020): 393391, 393000, 392169, 391682, 390061. . .	64

17	Painel de empenhos sinalizados como anômalos pelo DA-DGM (FMSJP, exercícios de 2017 a 2020). Exemplos: Linha 1 (2017): 393569, 392458, 391189, 390001; Linha 2 (2018): 393834, 392549, 391457, 390821; Linha 3 (2019): 393478, 392469, 391210, 390429; Linha 4 (2020): 392704, 392161, 391919, 391679, 390055.	65
----	--	----

LISTA DE TABELAS

1	Comparação metodológica e de complexidade entre algoritmos de detecção de anomalias.	41
2	Resultados consolidados para os conjuntos de dados unidimensionais. . . .	44
3	Resultados consolidados para os conjuntos de dados bidimensionais. . . .	46
4	Resultados consolidados para os conjuntos de dados tridimensionais. . . .	49
5	Resultados consolidados para os conjuntos de dados reais.	52

LISTA DE ABREVIATURAS

1D - Unidimensional

2D - Bidimensional

3D - Tridimensional

ABOD - *Angle-Based Outlier Detection*

ADBench - *Anomaly Detection Benchmark*

AEKNN - *AutoEncoder KNN*

AUC-ROC - *Area Under the Curve - Receiver Operating Characteristic*

Ball-Tree - *Ball tree*

CM - Cadeias de Markov

COF - *Connectivity-Based Outlier Factor*

DA-DGM - Detecção Automática de Anomalias baseada em Difusão Geométrica Markoviana

DGM - Difusão Geométrica Markoviana

DGM-e - Difusão Geométrica Markoviana Estendida

EIF - *Extended Isolation Forest*

e.g. - *exempli gratia* (“por exemplo”)

FastABOD - *Fast Angle-Based Outlier Detection*

FMSJP - Fundo Municipal de Saúde de João Pessoa

FN - *False Negative*

FP - *False Positive*

FPR - *False Positive Rate*

HBOS - *Histogram-Based Outlier Score*

IForest - *Isolation Forest*

INFLO - *Influence Function-based LOF*

IoT - *Internet of Things*

KD - *k-dimensional* (usado em “árvore KD”/KD-Tree; estrutura de índice espacial).

KD-Tree - árvore k-d; estrutura de partição espacial para busca de vizinhos.

KDD - *Knowledge Discovery in Databases*

k-NN - *k*-Nearest Neighbors

kth-NN - *k*-th Nearest Neighbor

LDOF - *Local Distance-Based Outlier Factor*

LOCI - *Local Correlation Integral*

LDOF - *Local Distance-Based Outlier Factor*

LOF - *Local Outlier Factor*

LoOP - *Local Outlier Probability*

MNIST - *Modified National Institute of Standards and Technology*

ODDS - *Outlier Detection DataSets*

OOSE - *Out-of-Sample Extension*

OpenML - *Open Machine Learning*

PyOD - *Python Outlier Detection*

SAGRES - Sistema de Acompanhamento da Gestão dos Recursos da Sociedade (plataforma do TCE-PB)

SECODA - *Segmentation- and Combination-based Detection of Anomalies*

SimplifiedLOF - *Simplified Local Outlier Factor*

TCE-PB - Tribunal de Contas do Estado da Paraíba

TN - *True Negative*

TP - *True Positive*

TPR - *True Positive Rate*

UCI - *University of California, Irvine*

Sumário

LISTA DE FIGURAS	v
LISTA DE TABELAS	viii
LISTA DE ABREVIATURAS	ix
1 INTRODUÇÃO	1
1.1 Objetivos - Gerais e Específicos	2
1.2 Hipótese	3
1.3 Contribuição	4
1.4 Trabalhos Relacionados	5
1.5 Estrutura da Dissertação	11
2 FUNDAMENTAÇÃO TEÓRICA	13
2.1 Grafos	13
2.2 Cadeias de Markov	16
2.3 Cadeias de Markov Geométricas	18
2.4 Propriedades Espectrais da Matriz de Markov	20
2.5 Método DGM para Simplificação de Malhas	22
2.6 Método estendido DGM-e para Redução de Grandes Conjuntos de Dados .	25
2.7 k -Vizinhos Mais Próximos	28
2.8 Entropia de Shannon	30
3 DETECÇÃO AUTOMÁTICA DE ANOMALIAS BASEADA EM DI- FUSÃO GEOMÉTRICA MARKOVIANA	32
3.1 Pré-Processamento	32
3.2 Construção do Grafo Local via K-Vizinhos Mais Próximos	33
3.3 Seleção do Parâmetro de Escala via Entropia de Shannon	34
3.4 Detecção de Anomalias	37
3.5 Análise de Complexidade	40

4	AVALIAÇÃO EXPERIMENTAL	42
4.1	Métrica de Avaliação	42
4.2	Protocolo de Comparação e Visualizações	43
4.3	Conjuntos de Dados Sintéticos	43
4.3.1	Dados Unidimensionais	44
4.3.2	Dados Bidimensionais	45
4.3.3	Dados Tridimensionais	48
4.4	Dados Reais	51
4.5	Discussão	57
4.5.1	Dados Sintéticos	59
4.5.2	Dados Reais	61
4.6	Limitações e Trabalhos Futuros	66
5	CONCLUSÃO	68
	REFERÊNCIAS	70

1 INTRODUÇÃO

Dados são descritos por Kock, McQueen e Baker (1996) como elementos brutos e desprovidos de contexto que, após serem processados, podem ser convertidos em informações úteis. Esses elementos representam registros do estado do mundo em um momento específico, refletindo características associadas a entidades, objetos ou eventos (Gomes; Pimenta; Schneider, 2019). A transformação desses registros em conhecimento significativo exige análises rigorosas e a aplicação de técnicas especializadas (Souza, 2021). Diante do crescimento exponencial na geração de dados, é imperativo desenvolver métodos analíticos robustos, capazes de converter dados brutos em informações de valor. Esse processo, denominado Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* – KDD), visa identificar padrões relevantes e detectar desvios significativos, comumente denominados anomalias ou *outliers* (Souza, 2021; Fayyad; Piatetsky-Shapiro; Smyth, 1996). Em diferentes domínios de aplicação, incluindo operações industriais, auditorias financeiras e sistemas de saúde, a detecção precisa de tais anomalias é fundamental, uma vez que eventos atípicos, como falhas operacionais, fraudes financeiras e erros instrumentais, podem ocasionar impactos severos em áreas que abrangem desde auditorias contábeis até diagnósticos médicos (Chandola; Banerjee; Kumar, 2007; Bouman; Bukhsh; Heskes, 2024).

Embora métodos tradicionais, tais como os baseados em estatística, distância, densidade, isolamento e agrupamento, sejam eficazes em diversas aplicações, a dependência de parâmetros ajustados manualmente compromete frequentemente a robustez e a capacidade de generalização, especialmente em ambientes ruidosos ou estruturalmente heterogêneos (Samariya; Thakkar, 2023; Bouman; Bukhsh; Heskes, 2024). Conforme Bouman, Bukhsh e Heskes (2024), o desempenho de algoritmos não supervisionados em dados tabulares do mundo real varia com o tipo de anomalia (global ou local), muitos dependem de seleção cuidadosa de hiperparâmetros e, em vários casos, um pequeno subconjunto, como KNN (*k-Nearest Neighbors*) e EIF (*Extended Isolation Forest*), é suficiente graças às suas forças complementares e eficiência computacional. Estudos sobre Sistemas Biológicos de Alerta Precoce (*Biological Early Warning Systems*), por exemplo, que aplicam técnicas não supervisionadas como LOF (*Local Outlier Factor*) e IForest (*Isolation Forest*), confirmam a necessidade de calibração cuidadosa para evitar falsos alarmes (Grekov et al., 2023). Ademais, apesar de sua eficiência, o *Isolation Forest* (iForest) ainda apresenta desvantagens, como forte aleatoriedade, baixa generalização e estabilidade insuficiente (Chen et al., 2023).

A robustez é fundamental em cenários práticos. Em tarefas de Monitoramento da Integridade Estrutural (SHM - *Structural health monitoring*), a detecção de eventos extremos, como sísmicos, é complexa devido à interferência de dados defeituosos e à presença

de múltiplos tipos de anomalias (por exemplo, dados ausentes, pequenos desvios ou *drift*) (Tang et al., 2025; Qian et al., 2025). Além disso, a perda não uniforme de dados é um desafio realista que afeta tanto a avaliação estrutural normal quanto a detecção de mudanças incomuns (Liu et al., 2025). Esses desafios somam-se à complexidade das anomalias industriais, que frequentemente exibem variações significativas de escala e demandam redes robustas (Yang et al., 2024). Por outro lado, modelos de difusão têm emergido como alternativas promissoras graças à capacidade de modelar distribuições complexas por meio de processos iterativos de remoção de ruído (Bhosale et al., 2024). Essa promessa é reforçada pelo uso de modelos de difusão em grafos em redes neurais hierárquicas para elevar a qualidade dos dados e mitigar esparsidade ou baixa qualidade em redes de sensores, como em centros de usinagem (Huang; Yang, 2023). Contudo, a aplicação desses modelos à detecção de anomalias permanece relativamente pouco explorada.

Nesse contexto, a Teoria dos Mapas de Difusão (TMD) oferece vantagens ao capturar a geometria intrínseca dos dados, apresentando aplicações promissoras em áreas como restauração de imagens médicas e controle de qualidade industrial (Bercea et al., 2024; Neto; Souza; de Magalhães, 2019). Todavia, o desafio central persiste: desenvolver métodos que preservem a sensibilidade geométrica sem ajustes manuais e que produzam resultados estatisticamente consistentes.

Os avanços recentes aqui revisados expõem uma lacuna persistente de pesquisa. Embora modelos baseados em difusão se destaquem em visão e dados sequenciais, ainda não há uma solução não supervisionada que trate conjuntos tabulares heterogêneos, elimine o ajuste manual de hiperparâmetros e ofereça uma métrica de anomalia facilmente interpretável. A relevância desse foco não supervisionado é enfatizada pela prática em domínios críticos, como detecção de fraudes com cartões de crédito, frequentemente reformulada como um problema não supervisionado de detecção de anomalias com redes avançadas, como *Autoencoders* e GANs (*Generative Adversarial Networks*) (Jiang et al., 2023). Nesse contexto, objetiva-se reduzir essa lacuna ao introduzir a Detecção Automática de Anomalias baseada em Difusão Geométrica Markoviana (DA-DGM), que combina difusão geométrica com seleção automática do parâmetro de escala (ε).

1.1 Objetivos - Gerais e Específicos

Este trabalho tem como objetivo geral desenvolver um método não supervisionado de detecção de anomalias baseado em Difusão Geométrica Markoviana (DGM), capaz de identificar instâncias atípicas em conjuntos de dados unidimensionais e multidimensionais. O método busca preservar a geometria intrínseca dos dados, eliminar a necessidade de ajuste manual de hiperparâmetros e produzir uma métrica de anomalia de fácil interpretação.

Os objetivos específicos são:

- **Objetivo 1:** Fundamentar teoricamente o método DA-DGM, descrevendo em detalhe as Cadeias de Markov Geométricas, as matrizes de transição e a maximização da entropia de Shannon, e demonstrar como esses elementos subsidiam a métrica ordenada de anomalia.
- **Objetivo 2:** Projetar e implementar o DA-DGM, enfatizando a construção da matriz de transição e a geração do vetor de probabilidades ordenado das instâncias do conjunto de dados.
- **Objetivo 3:** Desenvolver um procedimento de otimização automatizado para selecionar o parâmetro de escala (ε) em dados multidimensionais, utilizando uma busca global logarítmica seguida de refinamento local que maximiza a entropia de Shannon.
- **Objetivo 4:** Avaliar o desempenho do DA-DGM em conjuntos de dados sintéticos e reais, comparando-o com detectores clássicos (KNN, LOF, IForest, entre outros) por meio da Área Sob a Curva ROC (AUC-ROC).
- **Objetivo 5:** Garantir a interpretabilidade do vetor de probabilidades gerado, destacando as instâncias mais atípicas no topo do *ranking* e facilitando a tomada de decisão em aplicações práticas.

1.2 Hipótese

As hipóteses formuladas nesta pesquisa visam direcionar o processo investigativo e fundamentar as análises e interpretações dos resultados experimentais. Elas são desenvolvidas com base nas lacunas identificadas na literatura, nomeadamente a necessidade de um método não supervisionado para detecção de anomalias em dados tabulares heterogêneos que reduza substancialmente a dependência de ajuste manual de hiperparâmetros e mantenha a sensibilidade à geometria intrínseca dos dados.

Dessa forma, a hipótese principal testada neste estudo, que guia o desenvolvimento e a avaliação do método DA-DGM, é a seguinte:

- O método DA-DGM, que integra Difusão Geométrica Markoviana com a seleção automática do parâmetro de escala (ε) por maximização da entropia de Shannon, alcança desempenho igual ou superior em capacidade discriminativa à de detectores clássicos de anomalias (como KNN, LOF e IForest) em conjuntos de dados sintéticos e reais com geometrias variadas.

Formalmente, essa hipótese pode ser representada em termos da métrica de avaliação AUC-ROC, que é utilizada para quantificar a capacidade de discriminação dos algoritmos:

- Hipótese nula (H_0): O desempenho do método DA-DGM, medido pela AUC-ROC, não é significativamente superior ao desempenho dos detectores clássicos (KNN, LOF, IForest, entre outros) em conjuntos de dados sintéticos e reais.
- Hipótese alternativa (H_1): O método DA-DGM alcança AUC-ROC superior à de detectores clássicos (KNN, LOF, IForest, entre outros) em *benchmarks* sintéticos e reais.

O detalhamento do experimento, incluindo o protocolo de comparação, a seleção dos 22 conjuntos (sintéticos 1D (unidimensionais), 2D (bidimensionais) e 3D (tridimensionais) e bases reais ODDS (*Outlier Detection DataSets*) e FMSJP (Fundo Municipal de Saúde de João Pessoa)) e a avaliação dessa hipótese, é apresentado no Capítulo 4.

1.3 Contribuição

Neste estudo, apresenta-se o DA-DGM, método não supervisionado de detecção de anomalias que estende a DGM (Souza, 2011) e sua variante multivariada, DGM-e (Silva, 2012). A estrutura dos dados é modelada por grafos de proximidade e por uma matriz de transição de Markov, enquanto o parâmetro de escala (ε) é escolhido automaticamente pelo máximo da entropia de Shannon (Shannon, 1948) calculada sobre seus elementos diagonais. O vetor de auto-transições (P_{ii}) fornece escores interpretáveis e um ranqueamento de instâncias adequado à triagem em detecção não supervisionada (Bouman; Bukhsh; Heskes, 2024).

Para validar essa abordagem, foi realizada uma avaliação experimental extensiva em 22 conjuntos de dados distintos, englobando distribuições unidimensionais sintéticas, estruturas bidimensionais complexas, cenários tridimensionais não lineares e dez bases de dados reais, sendo cinco amplamente reconhecidas. As análises qualitativas revelaram clara separação visual entre amostras normais e anômalas, enquanto, quantitativamente, o método alcançou 100% de AUC-ROC em todos os cenários sintéticos analisados. Em bases reais, o DA-DGM mostrou desempenho robusto e estável. Nos conjuntos ODDS (*Outlier Detection DataSets*) (Rayana; Akoglu, 2016), liderou três de cinco bases (*Ionosphere*, *Letter* e *MNIST*), enquanto o KNN prevaleceu em *Vowels* e *Pima*. Nos dados do FMSJP, o IForest foi consistentemente superior, e o DA-DGM compôs um segundo grupo competitivo.

As principais contribuições deste trabalho incluem: o delineamento de um método não supervisionado que integra Difusão Geométrica Markoviana com o ajuste automático

do parâmetro de escala (ε) baseado em entropia de Shannon e a demonstração experimental de sua robustez e capacidade de generalização em múltiplos contextos de aplicação.

1.4 Trabalhos Relacionados

Nesta seção, serão apresentados os principais trabalhos relacionados à detecção de anomalias, tema recorrente na literatura de mineração de dados, sendo aplicada em diversas áreas devido a sua importância.

Pesquisadores dedicam esforços consideráveis a aprimorar métodos existentes e a desenvolver novas abordagens capazes de identificar padrões anômalos que indiquem comportamentos irregulares ou falhas de sistema. Como ponto de partida, distinguem-se as três principais metodologias destinadas à detecção de anomalias: supervisionada, não supervisionada e semissupervisionada (Chandola; Banerjee; Kumar, 2007).

Na abordagem supervisionada, um modelo de classificação é treinado com base em instâncias rotuladas de treinamento de ambas as classes, normal e anômala, a fim de prever rótulos de classe para dados de teste, que consistem em uma combinação de dados normais e anomalias. Em muitas aplicações do mundo real, a falta de disponibilidade de amostras de treinamento rotuladas da classe anômala pode ser um obstáculo (Chandola; Banerjee; Kumar, 2007).

No que diz respeito à abordagem não supervisionada, as instâncias em um banco de dados são classificadas diretamente com base em alguma pontuação de anomalia. Durante o treinamento, um modelo de pontuação é aprendido sem o uso de rótulos, sendo empregado para calcular a pontuação de anomalia dos dados de teste e classificá-los na fase de teste. Essa abordagem pressupõe que as anomalias sejam poucas e diversas. No entanto, essa premissa pode não ser válida em situações em que há um grande número de anomalias (Chandola; Banerjee; Kumar, 2007; Breiman, 2001).

Quanto à abordagem semi-supervisionada, um perfil de dados normais é construído exclusivamente usando dados de treinamento da classe normal, sem a utilização das informações de rótulo durante o processo de aprendizado. Nesse contexto, ela difere da abordagem de aprendizado semi-supervisionado discutida em outros problemas de aprendizado de máquina, em que um modelo é treinado na fase de treinamento com um conjunto parcialmente rotulado (apenas um subconjunto dos dados de treinamento possui rótulos). Na fase de teste, as instâncias são classificadas com base em seu ajuste ao perfil de dados normais aprendido. Essa abordagem não faz suposições sobre anomalias nem demanda amostras de treinamento da classe anômala (Chandola; Banerjee; Kumar, 2007; Breiman, 2001).

O presente trabalho concentra-se em métodos não supervisionados, cuja relevância é ressaltada por estudos como Goldstein e Uchida (2016), Domingues et al. (2018), Cam-

pos et al. (2016), Pang et al. (2021) e Darban et al. (2024). Esses trabalhos identificam três desafios recorrentes: escalabilidade para grandes conjuntos de dados de alta dimensionalidade, robustez ao ruído e representação de estruturas não lineares complexas (Pang et al., 2021; Darban et al., 2024).

A seguir são apresentados os principais estudos relacionados à detecção de anomalias, com ênfase nas abordagens não supervisionadas. Esse foco reflete cenários em que rótulos confiáveis para eventos raros são onerosos ou inviáveis, os perfis de anomalia evoluem ao longo do tempo e métodos adaptativos sem intervenção humana são preferíveis. O método proposto (DA-DGM) é intrinsecamente não supervisionado, construindo uma representação difusiva do espaço de dados sem informação de classe. Estudos comparativos também destacam a importância e os desafios dessa linha de pesquisa, particularmente a escalabilidade em ambientes de alta dimensionalidade, a robustez ao ruído e a representação de estruturas não lineares complexas (Goldstein; Uchida, 2016; Domingues et al., 2018; Campos et al., 2016; Pang et al., 2021; Darban et al., 2024).

Dentro desse panorama mais amplo, a robustez desponta como um requisito central, especialmente em aplicações práticas em que o comportamento dos dados é irregular ou degradado. No Monitoramento da Integridade Estrutural, por exemplo, detectar eventos sísmicos é particularmente desafiador devido à interferência de dados defeituosos e à coexistência de múltiplos tipos de anomalia (como dados ausentes, pequenos desvios ou *drift*) (Tang et al., 2025; Qian et al., 2025). Além disso, a perda não uniforme de dados é um desafio realista que afeta tanto a avaliação estrutural quanto a detecção de mudanças incomuns (Liu et al., 2025). No domínio financeiro, a detecção de fraudes com cartão de crédito é recorrentemente reformulada como um problema não supervisionado utilizando modelos como *Autoencoders* ou GANs (*Generative Adversarial Networks*) (Jiang et al., 2023).

Nesse contexto, o uso do termo “difusão” merece esclarecimento, pois ele se aplica a duas famílias conceitualmente distintas. A primeira refere-se aos Mapas de Difusão, uma técnica de aprendizado de variedades que constrói um grafo de afinidade e aplica análise espectral para obter uma incorporação de baixa dimensionalidade que preserva a conectividade intrínseca (Coifman; Lafon, 2006; Belkin; Niyogi, 2003). Nesse espaço de difusão, *outliers* usualmente aparecem como pontos isolados em relação à estrutura geométrica principal da variedade (Sipola; Juvonen; Lehtonen, 2011).

Aplicações de Mapas de Difusão demonstraram utilidade em domínios como registros de intrusão em redes (Sipola; Juvonen; Lehtonen, 2011), perfis de expressão gênica (Xu et al., 2007), sinais biomédicos (Shirotori; Sugitani, 2024) e processos industriais (Li; Zhang, 2014; Li; Liu; Zang, 2018). Sua principal vantagem reside em capturar geometrias não lineares. Não obstante, persistem desafios importantes, incluindo custo computacional, sensibilidade à escolha do parâmetro de escala (ε) e dificuldades

com a extensão fora da amostra (OOSE - *Out-of-Sample Extension*). O OOSE consiste no procedimento que projeta novos dados, não incluídos na decomposição espectral original, para o espaço de difusão aprendido. Essa projeção permite mapear pontos inéditos para a variedade sem recomputar toda a estrutura espectral. Entretanto, o método pode falhar quando anomalias estão ausentes do conjunto de referência utilizado na construção da variedade (Coifman; Lafon, 2006; Olson; Judd; Nichols, 2017; Mishne et al., 2017).

Esses desafios tornam-se ainda mais críticos em grandes conjuntos de dados, nos quais a aplicação direta de Mapas de Difusão é inviável e a OOSE passa a depender de amostragem para construir a variedade de referência (Mishne; Cohen, 2017; Mishne; Cohen, 2015; Mishne; Cohen, 2013; Mishne; Cohen, 2014; Juvonen; Hämmäläinen, 2014). Nesses cenários, a qualidade da extensão fora da amostra está condicionada à representatividade do subconjunto utilizado na decomposição espectral, de modo que anomalias não observadas tendem a ser mapeadas para regiões associadas ao fundo (*background*), entendido como o conjunto de observações que compõem o comportamento predominante e esperado dos dados. Isso ocorre porque a OOSE realiza interpolação em relação à estrutura amostrada, não extrapolando padrões inexistentes (Mishne; Cohen, 2017; Olson; Judd; Nichols, 2017; Sipola; Juvonen; Lehtonen, 2011). Soma-se a isso a sensibilidade ao parâmetro de escala (ε), cuja escolha inadequada pode aumentar excessivamente a conectividade no grafo, tornando anomalias indistinguíveis do *background* e reduzindo sua separabilidade no espaço de difusão (Sipola; Juvonen; Lehtonen, 2011; Coifman; Lafon, 2006; Mishne; Cohen, 2013).

A escolha do valor do parâmetro de escala (ε) não é trivial e tem sido objeto de estudo intenso, dado o seu impacto na robustez e na interpretabilidade dos resultados (Belkin; Niyogi, 2006). Teoricamente, o parâmetro de escala ε deve variar desde o limite de conectividade (a menor distância entre pares de pontos) até o diâmetro do conjunto de dados (a maior distância) (Oliveira, 2022).

Apesar da existência de métodos heurísticos amplamente utilizados na prática (como aqueles baseados no limite de conectividade, na distância mínima $\varepsilon_{\min}$, ou na média das distâncias mínimas $\varepsilon_{\text{mean}}$) (Coifman; Lafon, 2006), essas abordagens frequentemente falham em fornecer um critério unificado que seja aplicável a diferentes conjuntos de dados (Kushner; Yin, 2003).

Para contornar essas limitações, a literatura tem explorado abordagens sofisticadas de otimização. Trabalhos mais recentes, como o de Sule, Evans e Cameron (2023), avançaram ao fornecer estimativas de erro mais precisas para a escolha do parâmetro de escala (ε), denominado no referido estudo como largura de banda, especialmente em aplicações que envolvem medidas alvo ou problemas de *committor* em sistemas dinâmicos. Esses estudos revelaram que a largura de banda ótima (ε^*) pode ser determinada considerando a complexidade do *manifold* subjacente e o número de amostras disponíveis,

propondo uma formulação que minimiza tanto o erro de viés quanto o de variância de forma adaptativa. Essa abordagem é crucial para garantir que o mapeamento preserve tanto as estruturas locais quanto as globais, permitindo uma representação fidedigna da distribuição dos dados no espaço de difusão. Além disso, Sule, Evans e Cameron (2023) validam a aplicação do teste *Ksum* como uma heurística prática para a seleção de ε^* , demonstrando sua eficácia especialmente quando combinada com amostragem quase uniforme para encontrar parâmetros próximos do ótimo teórico.

Tais abordagens buscam um equilíbrio mais eficaz entre o erro de variância e o viés. Adicionalmente, a escolha do parâmetro de escala ε tem consequências diretas nas propriedades espectrais: valores pequenos podem resultar em autovalores próximos de 1, limitando o impacto do parâmetro temporal t na técnica de Mapas de Difusão, ao passo que valores muito grandes causam decaimento acentuado dos autovalores (Oliveira, 2022). Essa complexidade reforça a necessidade de um critério de seleção automatizado.

A segunda família inclui Modelos Generativos de Difusão, que aprendem a reverter um processo de adição de ruído (Ho; Jain; Abbeel, 2020; Liu et al., 2024). Esses modelos têm sido aplicados para sintetizar pseudo-anomalias ou para definir escores de desvio (Ho; Jain; Abbeel, 2020; Liu et al., 2024; Zhang et al., 2025; He et al., 2023). Em centros de usinagem, o uso de um módulo de *graph diffusion augmentation* auxilia no enfrentamento da esparsidade e da baixa qualidade em conjuntos de sensores (Huang; Yang, 2023). Entretanto, modelos generativos demandam poder computacional substancial e oferecem interpretabilidade limitada (Arrieta et al., 2020), o que motiva o emprego de técnicas de Inteligência Artificial Explicável (Arrieta et al., 2020). Trabalhos futuros sobre reconstrução de dados em SHM (*Structural health monitoring*) consideram introduzir Modelos de Difusão ou GANs (*Generative Adversarial Networks*) para gerar dados sintéticos de alta fidelidade como *augmentation* (Liu et al., 2025). Como o DA-DGM se ancora na primeira família, aprofunda-se a seguir a evidência empírica e as limitações dos Mapas de Difusão em detecção não supervisionada.

A dificuldade em estimar a dimensionalidade intrínseca ótima (Li; Zhang, 2014) e o fato de que a detecção de anomalias pode ser vista como um problema de agrupamento altamente desbalanceado (Mishne; Cohen, 2017) também representam limitações. Apesar desses desafios, a contínua pesquisa, como a introdução de *autoencoders* baseados em mapas de difusão (*Diffusion Nets*) para OOSE, detecção de *outliers* e *denoising* (Mishne et al., 2017), demonstra o potencial e a evolução da metodologia.

No universo dos métodos clássicos, Samariya e Thakkar (2023) classificam 52 algoritmos em sete categorias: estatística, densidade, distância, agrupamento, isolamento, *ensemble* e subespaço. O estudo de Bouman, Bukhsh e Heskes (2024) comparou 33 algoritmos não supervisionados em 52 bases tabulares multivariadas do mundo real, com tamanhos de 19 a 619.326 amostras e prevalência de anomalias de 0,03% a 34,99 %. Os

resultados indicam que o EIF (*Extended Isolation Forest*) supera a maioria dos métodos em anomalias globais, presentes em regiões de baixa densidade, enquanto o KNN apresenta melhor desempenho em anomalias locais, próximas a aglomerados densos. Cerca de um terço dos conjuntos analisados contém anomalias locais, favorecendo métodos como LOF e COF (*Connectivity-Based Outlier Factor*), enquanto os demais são dominados por anomalias globais, nas quais EIF e *Isolation Forest* se destacam (Bouman; Bukhsh; Heskes, 2024).

O HBOS (*Histogram-Based Outlier Score*) sobressai entre os métodos estatísticos não supervisionados pela simplicidade e escalabilidade. O algoritmo constrói, para cada atributo, histogramas que estimam a densidade marginal; em seguida, essas densidades são normalizadas e ponderadas conforme a importância relativa de cada atributo, resultando em um escore de anomalia para cada instância. A intuição subjacente é que observações em regiões de baixa probabilidade indicam anomalias, ao passo que aquelas em regiões densas refletem comportamento normal. Essa abordagem oferece bons resultados quando os atributos são aproximadamente independentes, mas sua eficácia pode ser menor em conjuntos de dados com forte correlação entre variáveis (Goldstein; Dengel, 2012; Samariya; Thakkar, 2023).

Na família de métodos não supervisionados baseados em densidade, destacam-se LOF (*Local Outlier Factor*) (Breunig et al., 2000), COF (*Connectivity-Based Outlier Factor*) (Tang et al., 2002), LoOP (*Local Outlier Probability*) (Kriegel et al., 2009), LOCI (*Local Correlation Integral*) (Papadimitriou et al., 2003) e suas variantes INFLO (*Influence Function-based LOF*) (Jin et al., 2006) e SimplifiedLOF (*Simplified Local Outlier Factor*) (Schubert; Zimek; Kriegel, 2014); todos comparam a densidade local de um objeto à de seus vizinhos mais próximos, identificando discrepâncias significativas como indício de anomalia, sendo que o LOF é o mais difundido, o COF ajusta a forma de estimar a densidade, o LoOP utiliza estimadores de densidade local mais robustos, o LOCI apresenta um fator de desvio multigranularidade, o INFLO é outra variante do LOF que usa o relacionamento de vizinhança simétrica para identificar anomalias e o SimplifiedLOF substitui a distância de alcance local pela distância de KNN na medida de pontuação de anomalia do LOF; apesar da eficácia em baixa dimensionalidade, o desempenho coletivo desses algoritmos é menor em conjuntos esparsos ou de alta dimensão, em que a estimativa confiável de densidade é mais difícil. A intuição por trás desses algoritmos de detecção de anomalias é que a densidade do objeto anormal é significativamente diferente da instância normal (Samariya; Thakkar, 2023).

Entre os métodos não supervisionados baseados em distância, destacam-se KNN (Knox; Ng, 1998; Knorr; Ng; Tucakov, 2000), kth-NN (Ramaswamy; Rastogi; Shim, 2000), ABOD (*Angle-Based Outlier Detection*) e sua variante eficiente FastABOD (Kriegel; Schubert; Zimek, 2008), bem como LDOF (*Local Distance-Based Outlier Factor*)

(Zhang; Hutter; Jin, 2009); todos classificam como anômalas as instâncias que estão significativamente afastadas de seus vizinhos, seja pela soma das distâncias aos k vizinhos mais próximos no KNN, pela distância ao k -ésimo vizinho no k th-NN, pela variância ponderada dos ângulos formados com os demais pontos no ABOD ou pela razão entre distâncias locais no LDOF. Embora o k th-NN se destaque pela eficiência computacional e o FastABOD reduza a complexidade temporal para $O(n^2 + nk^2)$, o desempenho coletivo dessa família de algoritmos tende a ser menor em conjuntos de alta dimensionalidade ou com estruturas não lineares, cenários nos quais a concentração das distâncias dificulta a diferenciação entre padrões normais e anômalos (Samariya; Thakkar, 2023).

O *Isolation Forest* (IForest) destaca-se entre os métodos não supervisionados baseados em isolamento pela eficiência computacional: o algoritmo gera diversas árvores isoladas (*Isolation Trees*) que particionam aleatoriamente o espaço de atributos e isolam instâncias, atribuindo-lhes uma pontuação proporcional ao comprimento médio do caminho necessário para separá-las; como pressupõe que anomalias são raras e exibem características que facilitam sua rápida segregação, pontos anômalos tendem a ser isolados em níveis mais rasos do que os exemplos normais (Liu; Ting; Zhou, 2008; Souza, 2021); embora apresente desempenho robusto quando a contaminação é baixa e dispense estimativas de densidade ou distância, o método preserva apenas parcialmente estruturas geométricas complexas, o que pode reduzir sua sensibilidade em domínios de alta complexidade espacial.

Estratégias híbridas de baixo custo, como AEKNN (*AutoEncoder KNN*) (Guo et al., 2018) e SECODA (*Segmentation-and Combination-based Detection of Anomalies*) (Foorhuis, 2017), mitigam algumas limitações dos métodos puramente baseados em distância ou densidade, mas continuam dissociadas dos avanços na teoria de Mapas de Difusão (MDs). Embora MDs preservem a geometria dos dados e gerem amostras sintéticas que reforçam padrões normais, sua aplicação ainda se limita majoritariamente a visão computacional e séries temporais, com pouca investigação em dados tabulares heterogêneos (Liu et al., 2025; Zhang et al., 2025; He et al., 2023). Além disso, demandam alto poder de processamento e oferecem interpretabilidade limitada, o que limita seu uso em ambientes com recursos escassos ou que exijam transparência.

Todavia, nenhuma investigação abrangente avaliou a DGM, derivada dessa teoria, em conjuntos de dados tabulares heterogêneos, lacuna que este estudo busca suprir. Além disso, as limitações dos métodos tradicionais e dos baseados em redes neurais, como a dependência de ajustes manuais de hiperparâmetros, o desempenho variável conforme o tipo de anomalia (global ou local) e a reduzida interpretabilidade (Bouman; Bukhsh; Heskes, 2024), reforçam a necessidade de uma abordagem alternativa.

A abordagem proposta generaliza o uso de conceitos de Mapas de Difusão, especialmente a Difusão Geométrica Markoviana, para a detecção de anomalias em uma

variedade mais ampla de conjuntos de dados multidimensionais. Embora as aplicações anteriores desses conceitos tenham sido eficientes e específicas, este trabalho se distingue ao explorar a adaptabilidade e a aplicabilidade desta técnica em um espectro mais amplo de cenários. Ao validar a versatilidade da Difusão Geométrica Markoviana como uma ferramenta valiosa na análise de dados, o estudo atual abre novas possibilidades para sua aplicação em áreas ainda não exploradas. Assim, posiciona-se como uma técnica complementar às existentes, introduzindo eficácia na detecção de anomalias e destacando o potencial ainda inexplorado da DGM.

1.5 Estrutura da Dissertação

Esta dissertação se divide em cinco capítulos principais e mais as referências bibliográficas, detalhando desde os objetivos até as conclusões finais e sugestões para pesquisas futuras.

O **Capítulo 1** apresenta o contexto e a motivação do estudo: introduz o ciclo dados–informação no âmbito do KDD, formula o problema de detecção de anomalias e sua relevância prática, discute limitações de abordagens tradicionais e o potencial dos Mapas de Difusão/DGM, explicita a lacuna investigada, enuncia os objetivos (geral e específicos), destaca as contribuições do trabalho e descreve a organização da dissertação.

No **Capítulo 2**, apresenta-se a base conceitual do trabalho: inicia-se com grafos e suas representações matriciais; introduzem-se as cadeias de Markov e as cadeias de Markov geométricas; discutem-se as propriedades espectrais da matriz de transição; revisitam-se a DGM e sua extensão a dados genéricos (DGM-e); e, por fim, resumem-se elementos de k -vizinhos mais próximos (KNN) e de entropia de Shannon, que serão empregados na modelagem e na seleção de parâmetros.

No **Capítulo 3**, detalha-se o método proposto (DA-DGM): construção do grafo KNN e da matriz de transição, seleção automática do parâmetro de escala ε via entropia de Shannon e ranqueamento de anomalias a partir das auto-probabilidades, com descrição do pré-processamento, da busca global e do refinamento local, bem como do algoritmo completo e dos modos de uso.

O **Capítulo 4** descreve a avaliação experimental do DA-DGM. Inicia-se com experimentos em dados sintéticos (1D, 2D e 3D) e, na sequência, em bases reais (ODDS e FMSJP). Em seguida, apresentam-se as principais métricas e uma discussão integrada dos resultados. O capítulo encerra com a análise de limitações e direções para trabalhos futuros.

O **Capítulo 5** apresenta a síntese dos resultados obtidos, relaciona-os aos objetivos propostos, discute o desempenho do DA-DGM em dados sintéticos e reais, analisa limitações e custo computacional e propõe perspectivas de continuidade.

Por fim, as referências bibliográficas reúnem as fontes consultadas e citadas ao longo da dissertação, oferecendo suporte teórico e metodológico ao desenvolvimento do trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

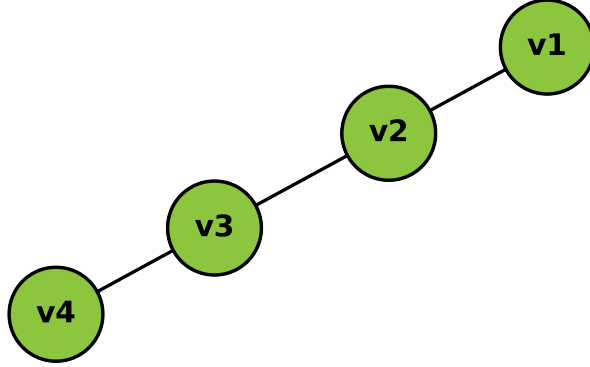
Este capítulo estabelece a base conceitual da modelagem utilizada nesta dissertação. Inicia-se pela teoria de grafos e por sua representação matricial; em seguida, apresentam-se as cadeias de Markov como passeios aleatórios sobre grafos; por fim, descrevem-se a DGM e sua extensão para dados genéricos, o método DGM-estendido (DGM-e), ambas fundamentadas na teoria dos mapas de difusão. Demonstra-se que os elementos centrais do modelo, como relações de vizinhança, probabilidades de transição e medidas de relevância, podem ser formulados de modo unificado por meio de estruturas matriciais derivadas de um mesmo objeto fundamental: o grafo ponderado.

2.1 Grafos

Esta seção apresenta uma visão geral dos elementos de teoria dos grafos empregados na DGM, com foco na ponderação de arestas e no uso de núcleos de similaridade para induzir dinâmicas estocásticas coerentes com a geometria dos dados. São introduzidas as estruturas fundamentais, como vértices, arestas e matrizes de adjacência nas versões ponderadas ou não e orientadas ou não, bem como medidas locais como grau, emissão e recepção. Também são discutidos os conceitos de caminho e conectividade e seu papel na construção de cadeias de Markov que sustentam o processo difusivo utilizado na detecção de anomalias. As definições formais e os exemplos ilustrativos são desenvolvidos nos parágrafos subsequentes.

Em termos formais, define-se um grafo como $G = (V, E, w)$, em que $V = \{v_1, \dots, v_n\}$ é o conjunto enumerável de vértices (ou nós), $E = \{e_{ij} = (v_i, v_j) \mid v_i, v_j \in V\}$ é o conjunto de arestas (*edges*) que ligam pares de vértices e $w : E \rightarrow W \subset \mathbb{R}$ é uma função que atribui a cada aresta um peso numérico (*weights*), interpretável como intensidade, custo ou distância da ligação. Embora essa definição seja abstrata, um exemplo elementar com $V = \{1, 2, 3, 4\}$ e $E = \{(1, 2), (2, 3), (3, 4)\}$, mostrado na Figura 1, já mostra a versatilidade dos grafos para modelar relações em contextos variados, como redes sociais, rotas de transporte ou fluxos de informação.

Figura 1: Exemplo de grafo simples não orientado.



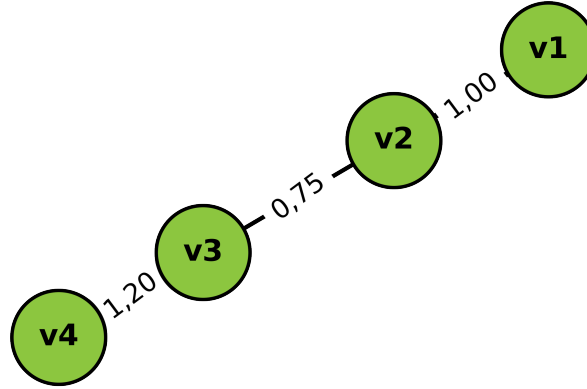
Fonte: O autor, 2025.

A representação algébrica dessa estrutura surge naturalmente pela matriz de adjacência $A \in \mathbb{R}^{n \times n}$, na qual cada elemento a_{ij} indica a existência (ou intensidade) da ligação entre v_i e v_j . Em grafos não ponderados, como o mostrado na Figura 1, a_{ij} assume valores binários; em grafos ponderados, como o ilustrado na Figura 2, incorpora o próprio peso da aresta (Moreira, 2010). Essa matriz é simétrica em grafos não orientados e reflete a conectividade entre os vértices. Tal codificação viabiliza o uso de técnicas de álgebra linear para inferir propriedades globais do sistema, permitindo, por exemplo, a detecção de comunidades ou o cálculo de medidas de centralidade e, assim, identificar os atores mais relevantes no cenário modelado, quantificando a relevância e ranqueando os nós centrais da rede.

A Figura 2 exhibe o grafo $G = (V, E, w)$ com $V = \{1, 2, 3, 4\}$, $E = \{(1, 2), (2, 3), (3, 4)\}$ e $w = \{(1, 2) \mapsto 1.00, (2, 3) \mapsto 0.75, (3, 4) \mapsto 1.20\}$. A respectiva matriz de adjacência é

$$A = \begin{bmatrix} 0 & 1.00 & 0 & 0 \\ 1.00 & 0 & 0.75 & 0 \\ 0 & 0.75 & 0 & 1.20 \\ 0 & 0 & 1.20 & 0 \end{bmatrix}.$$

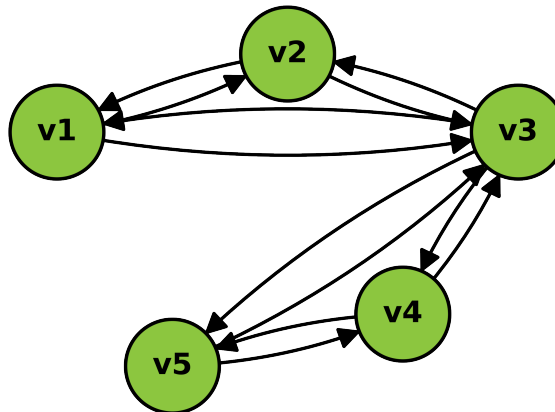
Figura 2: Exemplo de grafo simples não orientado e ponderado.



Fonte: O autor, 2025.

Essa mesma estrutura admite diversas classificações que afetam diretamente a escolha de algoritmos de análise. Quando $W = \{0,1\}$, tem-se um grafo não ponderado; quando $W \subset \mathbb{R}^+$, alcança-se a generalidade dos grafos ponderados, em que pesos representam distâncias, custos ou probabilidades (Moreira, 2010). A orientação das arestas acrescenta outra dimensão: se $w_{ij} = w_{ji}$, o grafo é não orientado, como nas Figuras 1 e 2; caso contrário, assume-se um grafo orientado, como o ilustrado na Figura 3, no qual a assimetria modela fluxos direcionais, como tráfego ou relações de precedência (Moreira, 2010).

Figura 3: Exemplo de grafo orientado e não ponderado.



Fonte: O autor, 2025.

A centralidade de grau (*degree centrality*) mede a importância local de cada vértice por meio do seu grau $d(v_i)$ (Figueirêdo, 2021). Em grafos não orientados e sem pesos, como o ilustrado na Figura 1, o grau de v_i corresponde ao número de vizinhos; em grafos ponderados, como o ilustrado na Figura 2, utiliza-se a soma dos pesos incidentes, refletindo a intensidade da conectividade local (Moreira, 2010). Em grafos orientados, distinguem-se o grau de emissão e o grau de recepção, que contabilizam, respectivamente, as arestas que partem e as que chegam a um vértice (Silva, 2012). Esses indicadores revelam quais vértices exercem maior influência ou concentram maior dependência no sistema.

A noção de caminho conecta essas ideias: uma sequência de vértices adjacentes que une v_i a v_j define um caminho cujo comprimento é dado pelo número de arestas percorridas (Moreira, 2010). A existência de caminhos entre todos os pares de vértices caracteriza um grafo conexo, como os ilustrados nas Figuras 1, 2 e 3, atributo crucial quando se pretende analisar a difusão de informação ou o espalhamento de perturbações na rede.

Todas as construções anteriores têm um elo natural em cadeias de Markov, pois a matriz de transição de um passeio aleatório sobre o grafo nada mais é que uma normalização estocástica da matriz de adjacência (Moreira, 2010). Essa correspondência permite interpretar pesos como probabilidades de deslocamento entre estados, utilizar resultados clássicos de processos estocásticos e, sobretudo, estabelecer a base para a Difusão Geométrica Markoviana. A DGM recorre a núcleos de similaridade para construir cadeias de Markov cujas probabilidades de transição respeitam a geometria intrínseca dos dados; a partir da evolução dessa cadeia, mede-se a importância de cada elemento de modo análogo à retenção de calor em um processo difusivo (Silva, 2012). Assim, conceitos de densidade, agrupamento ou detecção de anomalias emergem como consequências diretas do fluxo probabilístico sobre o grafo.

Em síntese, a progressão dos grafos para as cadeias de Markov e destas para a DGM revela um *continuum* conceitual: começa-se representando relações discretas, passa-se a interpretar essas relações como passos de um processo estocástico e, finalmente, explora-se a propagação difusiva para extrair padrões e anomalias em conjuntos de dados de alta complexidade.

2.2 Cadeias de Markov

Esta seção introduz as ferramentas probabilísticas que sustentarão os algoritmos de difusão usados nas seções seguintes. Inicia-se pela formulação clássica das cadeias de Markov, discute-se como as probabilidades de transição podem ser organizadas de forma matricial e avança-se até variações geométricas que integram noções de similaridade entre dados.

Um conceito fundamental em Cadeias de Markov (CM) é o de variável aleatória.

É uma função que mapeia elementos do espaço amostral para valores numéricos, isto é, $X : \Omega \rightarrow A$, em que $A \subseteq \mathbb{R}$, ou seja, um subconjunto dos números reais (Ross, 2000).

Um processo estocástico discreto $X(t), t \in T$ é definido como uma coleção de variáveis aleatórias indexadas por t . Em outras palavras, para cada $t \in T$, $X(t)$ corresponde a uma variável aleatória específica. O parâmetro t frequentemente representa o tempo, e $X(t)$ o estado do processo nesse instante específico (Ross, 2000).

Pode-se então definir uma Cadeia de Markov. Considere um processo estocástico discreto $X(n), n = 0, 1, 2, \dots, n$, que possui um conjunto finito e enumerável de estados. Assumindo que o processo se encontra no estado i , seja P_{ij} a probabilidade (fixa) de transição para o estado j no instante seguinte (Ross, 2000). Formalmente,

$$P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = P_{ij}. \quad (1)$$

Em termos simples, a probabilidade de o processo entrar em qualquer estado subsequente depende unicamente do estado atual. Um processo estocástico discreto com essa característica é denominado Cadeia de Markov. A probabilidade de transição entre os estados n e $n+1$, $P[X(n+1) = x_{n+1} \mid X(n) = x_n]$, constitui uma medida essencial em tal sistema. Com P_{ij} definidos para $0 \leq i \leq n$ e $0 \leq j \leq n$, em que todas as probabilidades são positivas, a seguinte condição é satisfeita (Ross, 2000):

$$\sum_{j=0}^n P_{ij} = 1. \quad (2)$$

No exemplo elementar $P = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$ a interpretação é direta: partindo do estado 1, a probabilidade de permanecer é 0,7 e de migrar para o estado 2 é 0,3, quantificando de forma compacta o comportamento do sistema.

Uma aplicação prática e desafiadora das Cadeias de Markov, que ilustra o conceito de transição de estados em um sistema complexo, é o algoritmo *PageRank* (Page et al., 1999) do motor de busca Google. Nesse modelo, a Web é interpretada como um grafo direcionado (*Web Graph*), em que cada página *web* é um estado e os *hyperlinks* são as arestas. O algoritmo *PageRank* é modelado com base no comportamento de um navegador *web* aleatório (*randomized Web surfer*). A cada passo, o navegador move-se de uma página (estado) para outra ao escolher aleatoriamente um *link* de saída. A transição de um estado i para um estado j ocorre com uma probabilidade p_{ij} que depende apenas do estado atual, e não de estados anteriores. A matriz de transição de Markov representa essas probabilidades de movimento. A probabilidade limite (estacionária) de um navegador aleatório infinitamente dedicado visitar uma determinada página é o seu *PageRank*, o que demonstra como a CM é usada para prever o comportamento de um

sistema que viaja de um estado para outro, considerando apenas a condição atual.

Para que a cadeia explore todo o espaço de estados e exiba convergência adequada em horizontes longos, costumam-se exigir irreducibilidade e aperiodicidade. Uma Cadeia de Markov é classificada como irreducível se, para cada par de estados x_i e x_j , a probabilidade de transição entre eles for estritamente positiva, assegurando que todos os estados são acessíveis entre si (Ross, 2000), ou seja, qualquer estado pode ser alcançado a partir de qualquer outro, o que equivale a dizer que o grafo associado à matriz de transição é conexo (Moreira, 2010).

A Cadeia de Markov é considerada aperiódica quando, além de ser irreducível, existe pelo menos um $n \geq 0$ e um estado x_j tal que (Ross, 2000):

$$P(X_n = x_j \mid X_0 = x_j) > 0 \text{ e} \quad (3)$$

$$P(X_{n+1} = x_j \mid X_0 = x_j) > 0 \quad (4)$$

A aperiodicidade impede padrões cíclicos estritos de retorno, eliminando periodicidades que limitariam a evolução temporal (Moreira, 2010).

Quando essas condições se verificam, o Teorema de *Perron-Frobenius* garante que o maior autovalor da matriz é 1 e possui multiplicidade algébrica 1 (Moreira, 2010; Meyer, 2023), resultado fundamental para provar existência e unicidade da distribuição estacionária.

2.3 Cadeias de Markov Geométricas

Esta seção introduz as Cadeias de Markov Geométricas, nas quais a estrutura espacial dos dados orienta a dinâmica estocástica por meio de uma função núcleo $k(x_i, x_j)$ simétrica e não negativa. Definem-se $k(x_i, x_j)$ e as condições para sua interpretação probabilística, procede-se à normalização que gera a matriz de transição P , e discute-se como essa construção incorpora a vizinhança local. Por fim, apresentam-se os aspectos espectrais de P relevantes à difusão, que fundamentam procedimentos de análise e detecção de anomalias sem exigir o cálculo explícito de autovetores.

A teoria de Cadeias de Markov estende-se às *Cadeias de Markov Geométricas* (Souza, 2011), que vinculam a configuração geométrica de um conjunto de dados multidimensional à dinâmica estocástica da cadeia. Para isso, define-se uma função de ponderação (núcleo) que atribui pesos às interações entre estados, isto é, arestas que representam ligações multidimensionais, em que cada estado corresponde a uma instância do conjunto. É, portanto, essencial especificar com precisão essa função para realçar a relevância das interações e favorecer a detecção efetiva de anomalias.

Tomando X como o conjunto de n estados possíveis, representados pelos elementos do conjunto de dados $X = \{x_1, x_2, \dots, x_n\}$, a função núcleo k é estabelecida tendo como referência o conceito do núcleo de calor, tal como abordado por Silva, Souza e Motta (2016), sendo definida como $k : X \times X \rightarrow \mathbb{R}^+$. Para assegurar propriedades probabilísticas à função núcleo, devem ser satisfeitas algumas condições (Burdea; Coiffet, 2003; Biyikoğlu; Leydold; Stadler, 2007; Collado et al., 2004). Portanto, para $x_i, x_j \in X$, exige-se que:

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

$$k(x_i, x_j) \geq 0 \text{ se } i \neq j \quad (6)$$

$$k(x_i, x_i) > 0 \quad (7)$$

A neutralização do impacto da direcionalidade nas conexões entre os elementos ocorre pela equivalência na Igualdade (5), que não favorece nenhuma direção específica nas conexões. As condições expressas pelas desigualdades (6) e (7) facilitam a interpretação dos valores gerados pela função núcleo k como probabilidades (Silva; Souza; Motta, 2016; Collado et al., 2004; Franc; Skala, 2001).

Utilizando os valores gerados pelo núcleo, é possível compor uma matriz simétrica livre de valores negativos. A normalização transforma esta matriz em uma que exhibe propriedades probabilísticas, originando uma matriz de probabilidade de transição, denominada matriz P , que se associa à cadeia de Markov definida sobre o conjunto de estados X . A formulação dessa normalização é especificada como (Silva; Souza; Motta, 2016; Biyikoğlu; Leydold; Stadler, 2007; Collado et al., 2004):

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

com o fator de normalização especificado por:

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

em que $d(x_i)$ é um fator de normalização especificado por:

A matriz de transição elaborada incorpora a configuração geométrica local de um elemento e suas interações com os vizinhos mais próximos. A maneira como as informações são extraídas e interpretadas a partir de cada configuração local é influenciada pelas características definidoras da função núcleo aplicada ao conjunto de dados (Souza, 2011). Uma cadeia de Markov que emerge de um fundamento geométrico é denominada cadeia geométrica de Markov.

A disposição em ordem decrescente dos autovalores absolutos nas matrizes de transição, iniciando com o maior valor fixado em um e os subsequentes em declínio, destaca a natureza espectral dessas matrizes. Essa propensão e a convergência acelerada dos autovalores em direção ao zero são validadas pelo teorema da convergência espectral. Adicionalmente, o traço da matriz, sendo a soma dos elementos ao longo de sua diagonal principal, tem uma relação direta com a soma de seus autovalores, um conceito fundamentado pelo teorema do traço. O teorema de *Hoffman-Wielandt* (Elsner, 1993) fornece a base para minimizar distorções espectrais ao remover instâncias do conjunto de dados com base em valores diagonais, favorecendo o uso da Teoria dos Mapas de Difusão. Essa abordagem permite a estimação dos autovalores a partir dos elementos diagonais, eliminando a necessidade de calcular diretamente os autovalores e autovetores e facilitando a identificação das instâncias a serem mantidas, como apontado nos estudos de Coifman e Lafon (2006); Kriegel, Schubert e Zimek (2008); Silva, Souza e Motta (2016) e (Strang, 1988), quando a técnica é utilizada com o objetivo de simplificação.

2.4 Propriedades Espectrais da Matriz de Markov

Esta seção apresenta as propriedades espectrais da matriz de transição P em cadeias de Markov discretas: define a decomposição em autovalores λ_l e autovetores e mostra como o parâmetro de escala ε do núcleo gaussiano regula a conectividade do grafo e o decaimento espectral. Por fim, conecta esses elementos aos mecanismos de difusão do DA-DGM e à detecção de anomalias.

As propriedades espectrais da matriz de transição P descrevem o comportamento da cadeia em termos de autovalores (λ_l) e autovetores; no DA-DGM, como P é construída sobre um conjunto finito de instâncias, a análise é intrinsecamente discreta. Um autovalor é um escalar λ tal que, para um vetor não-nulo u (o autovetor), a transformação linear da matriz P resulta em $Pu = \lambda u$. Na decomposição espectral da matriz de transição discreta, os termos $\psi_l(\cdot)$ e $\varphi_l(\cdot)$ devem ser interpretados como autovetores à direita e à esquerda, respectivamente, associados aos autovalores λ_l , e não como funções próprias, que são típicas de formulações contínuas do operador Laplaciano.

A decomposição espectral pode ser expressa como:

$$p(x_i, x_j) = \sum_{l \geq 0} \lambda_l \psi_l(x_i) \varphi_l(x_j), \quad (10)$$

em que $\psi_l(\cdot)$ e $\varphi_l(\cdot)$ são os autovetores à direita e à esquerda associados aos autovalores λ_l . Um resultado crucial da análise espectral é que os autovalores λ_i da matriz de transição P satisfazem a seguinte relação (Moreira, 2010):

$$\lambda_0 = 1 > |\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n|, \quad (11)$$

indicando que o maior autovalor é $\lambda_0 = 1$, enquanto os demais autovalores decrescem em magnitude. Além disso, à medida que o número de iterações n aumenta, o menor autovalor se aproxima de zero (Meyer, 2023):

$$\lim_{n \rightarrow \infty} \lambda_n = 0. \quad (12)$$

A matriz de transição P é construída a partir da normalização do núcleo gaussiano (ou *heat kernel*), uma função que estabelece a similaridade entre as instâncias x_i e x_j (estados) com base em sua distância euclidiana, conforme definido detalhadamente na Seção 3.2. O parâmetro de escala ε atua como o termo de decaimento nesse núcleo, regulando a conectividade local do grafo e, conseqüentemente, o espectro da matriz. A função núcleo é dada por:

$$k(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{\varepsilon}} \quad (13)$$

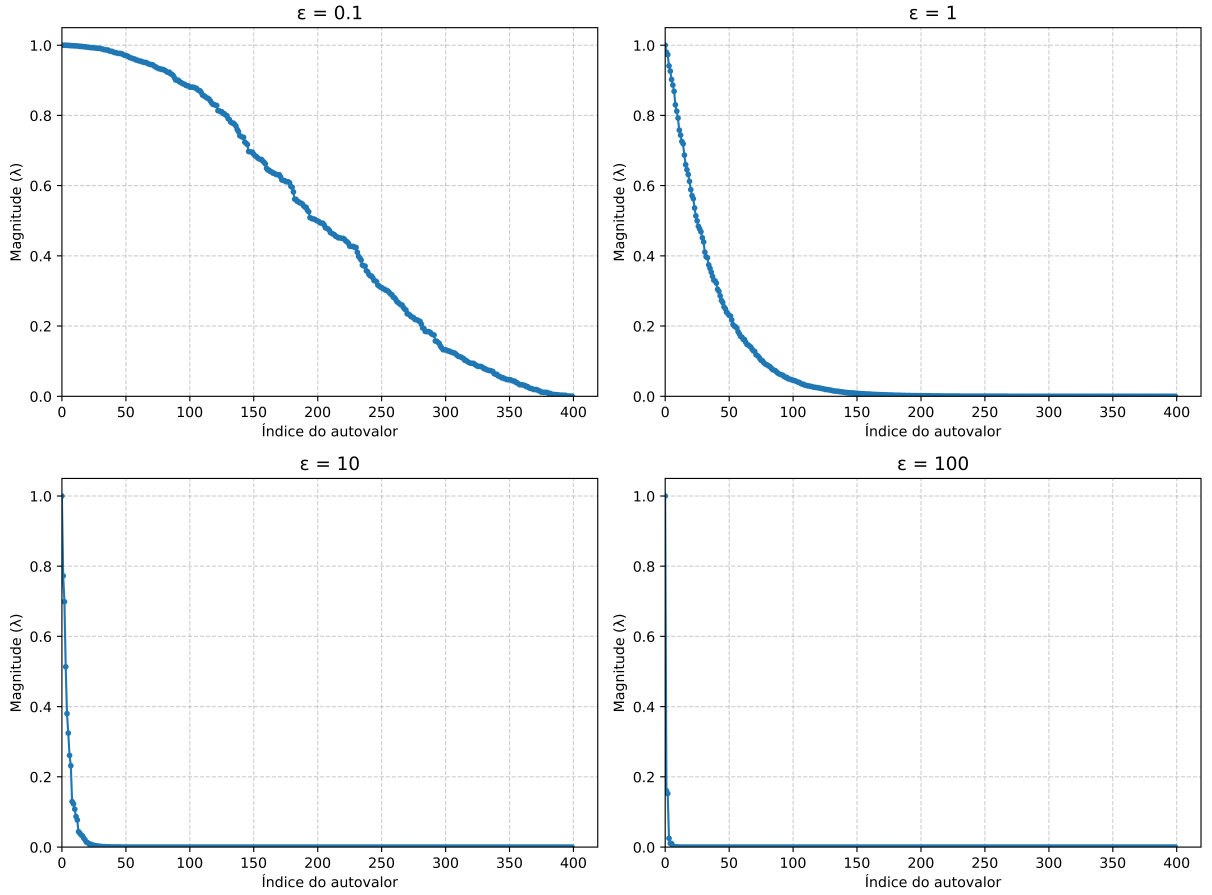
O escopo efetivo do núcleo é definido pelo parâmetro $\varepsilon > 0$. Quando ε assume valores reduzidos, privilegiam-se a formação de conexões essencialmente locais, resultando em uma matriz P mais esparsa e sensível a nuances geométricas finas. Por outro lado, valores elevados tendem a produzir conectividade quase homogênea entre os pontos, suavizando variações e podendo, no limite, diluir estruturas relevantes (Singer; Wu, 2011; Coifman; Lafon, 2006). Desse modo, ε atua como um controlador de escala, equilibrando fidelidade local e mistura global, com influência direta sobre o espectro de P , inclusive sobre o ritmo de decaimento de seus autovalores.

Essas características espectrais desempenham um papel fundamental na compreensão da convergência e estabilidade de cadeias de Markov, sendo essenciais em diversas aplicações práticas. Os espectros dos autovalores da matriz de transição $P \in \mathbb{R}^{400 \times 400}$, com valores positivos na diagonal e não-negativos nas outras posições, calculados para quatro valores do parâmetro de escala ε , estão apresentados na Figura 4. Observa-se que, à medida que o parâmetro de escala aumenta, o decaimento espectral se torna mais abrupto: para $\varepsilon = 10^{-1}$ os autovalores diminuem gradualmente, enquanto para $\varepsilon = 100$ a queda ocorre nos primeiros índices. Este comportamento é previsto pelo teorema de convergência espectral (Meyer, 2023)

A influência do parâmetro de escala $\varepsilon > 0$ também se estende ao estudo dos autovalores da matriz de difusão, com implicações diretas no parâmetro temporal t (da técnica de Mapas de Difusão). Como discutido na tese Oliveira (2022), a escolha de um valor pequeno do parâmetro de escala pode resultar em autovalores próximos de 1, li-

mitando o impacto do parâmetro temporal t na técnica de Mapas de Difusão, uma vez que a potência desses autovalores permanece praticamente inalterada. Por outro lado, valores muito grandes de ε podem causar um decaimento acentuado dos autovalores, o que também pode não ser ideal dependendo da aplicação. Essa dependência reforça a importância da seleção automática de ε no DA-DGM, contribuindo para que o mapeamento preserve as características geométricas relevantes do conjunto de dados.

Figura 4: Espectros dos autovalores da matriz de transição construída com núcleo gaussiano para $\varepsilon = 10^{-1}, 1, 10$ e 100 (leitura: esquerda para a direita, de cima para baixo). O eixo horizontal indica o índice dos autovalores ordenados por magnitude, enquanto o eixo vertical apresenta a respectiva magnitude λ .



Fonte: O autor, 2025.

2.5 Método DGM para Simplificação de Malhas

Esta seção apresenta o método DGM aplicado à simplificação de malhas triangulares e organiza o conteúdo em quatro passos: construção da matriz de transição a partir de características geométricas locais, cálculo de uma métrica de relevância para priorização de vértices, definição de um limiar automático com critério de parada e execução do ciclo de remoção até a convergência, com visualização das regiões a preservar.

A técnica de simplificação de malhas triangulares desenvolvida por Souza (2011) utiliza o conceito de difusão de informação para construir uma matriz de transição que representa variações locais nas propriedades geométricas de um conjunto de pontos 3D. Essa matriz é gerada com base nos vetores normais aos vértices, que são mapeados por uma função núcleo gaussiana, cuja diagonal indica a similaridade entre as normais de vértices adjacentes, refletindo a congruência em uma vizinhança. A curvatura direcional, determinada pela diferença entre vetores normais de vértices conectados por uma aresta, é uma propriedade central nesse processo, conforme descrito por Sun, Ovsjanikov e Guibas (2009). Assim, a DGM permite estimar a concentração de informação de cada elemento, associando o resultado da redução às características selecionadas, com o núcleo gaussiano garantindo uma representação multiescala concisa e informativa.

Com o auxílio de uma função núcleo adequada, forma-se uma matriz de transição, cujos elementos diagonais refletem a congruência entre os vetores normais em uma determinada vizinhança. A escolha recai sobre o núcleo gaussiano, conforme descrito por Silva, Souza e Motta (2016):

$$k(x_i, x_j) = \begin{cases} e^{-\|N_i - N_j\|^2}, & \text{se } x_i \text{ e } x_j \text{ são conectados,} \\ 0, & \text{caso contrário.} \end{cases} \quad (14)$$

em que N_i e N_j representam vetores normais nos vértices x_i e x_j , respectivamente.

Os vértices em uma malha podem apresentar variados níveis de conectividade, complicando a avaliação comparativa de sua relevância. Para abordar essa questão, Souza (2011) estabeleceu o valor de importância V de um vértice como:

$$V(X_i) = P(X_i, x_i) \cdot \text{Deg}(x_i) \quad (15)$$

em que $\text{Deg}(x_i)$ e P_{ii} são, respectivamente, o grau do ponto x_i e o seu valor correspondente na diagonal da matriz de transição.

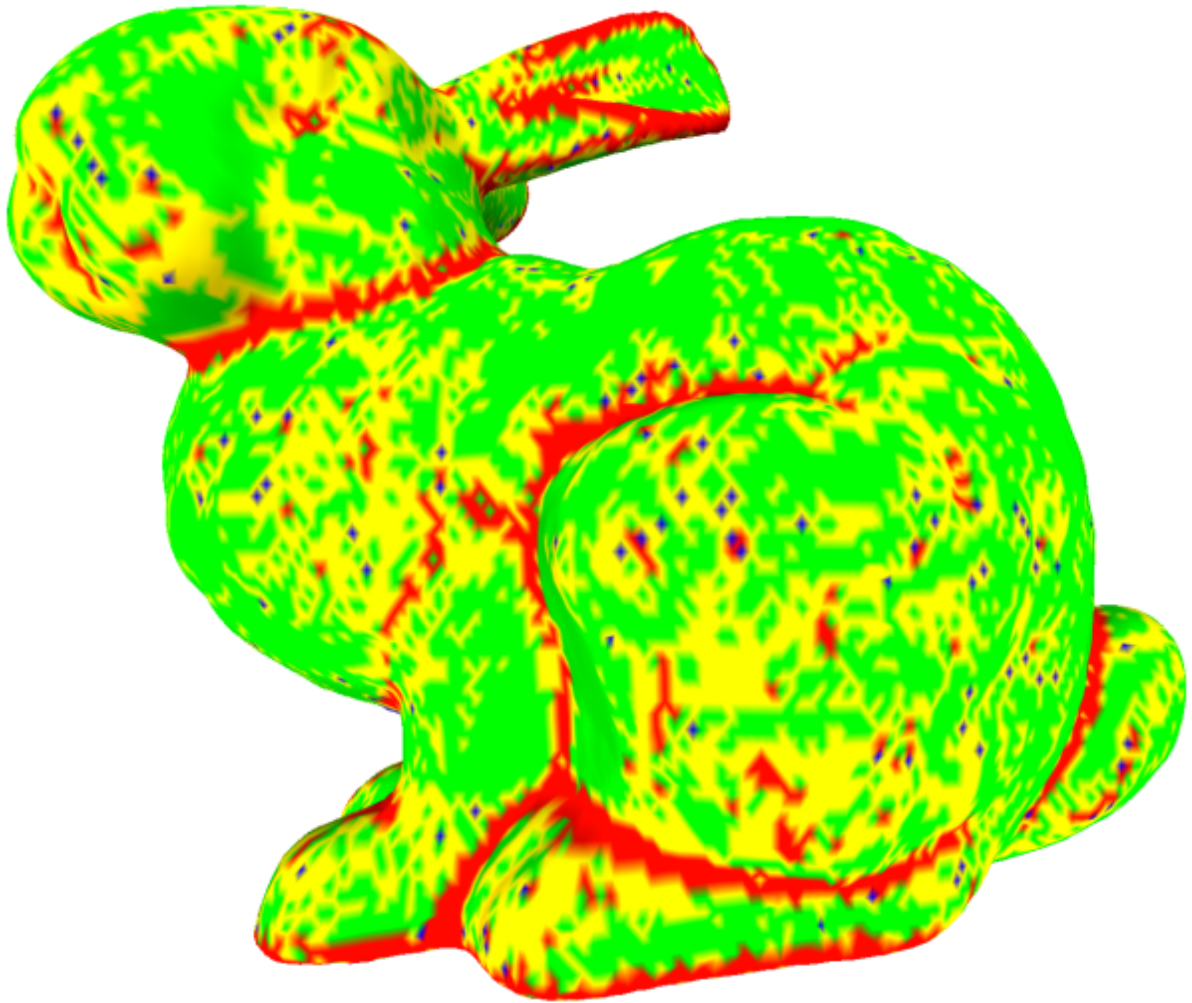
Inicialmente, determinar um limiar preciso para exclusão de elementos pode ser desafiador. Uma abordagem para automatizar esse processo envolve o uso do modelo gaussiano sobre os valores de importância V . Elementos de baixa representatividade são identificados no intervalo $[-\infty, l - js]$ e aqueles de alta representatividade em $[l + js, +\infty[$, sendo l a média e s o desvio padrão do modelo gaussiano, calculados a partir dos valores de $V(x_i)$. O parâmetro $j \geq 0$ serve como um ajuste para o tamanho do intervalo e pode ser estabelecido através de uma heurística relacionada ao volume do conjunto de dados (Silva; Souza; Motta, 2016).

Ao longo das iterações do método, a relevância dos elementos remanescentes é reexaminada. De acordo com Souza (2011), analisou-se as alterações nos valores médios

e na variância após diversas iterações de exclusão de vértices, observando uma tendência de decréscimo no valor médio. Assim, a taxa de redução da média foi adotada como critério para cessar as iterações do método, atingindo a convergência quando a diferença entre as médias de iterações consecutivas se mostra reduzida. Observou-se também uma diminuição no número de elementos excluídos a cada iteração (Silva; Souza; Motta, 2016).

Como ilustrado na Figura 5, o modelo *bunny* é exibido com uma codificação cromática proporcional à importância de cada vértice, calculada pela função de representatividade $V(x_i)$. Essa codificação decorre de um modelo gaussiano ajustado aos valores $V(x_i)$ e do parâmetro κ , que atua como margem de segurança na definição do intervalo de confiança usado na seleção de vértices candidatos à remoção, evitando a exclusão de pontos representativos. No caso específico $\kappa = 1$, os intervalos que orientam as cores são definidos em torno da média μ e do desvio-padrão σ : azul para $V(x_i) \in (0, \mu - \sigma]$, verde para $V(x_i) \in (\mu - \sigma, \mu]$, amarelo para $V(x_i) \in (\mu, \mu + \sigma]$ e vermelho para $V(x_i) \in (\mu + \sigma, +\infty)$. Conforme observado na Figura 5, as feições da malha, regiões de maior relevância que devem ser preservadas, aparecem em vermelho, enquanto valores mais baixos tendem ao azul e indicam áreas potencialmente passíveis de simplificação.

Figura 5: Representação visual do modelo *bunny* com pontos coloridos conforme sua importância.



Fonte: Souza (2011).

2.6 Método estendido DGM-e para Redução de Grandes Conjuntos de Dados

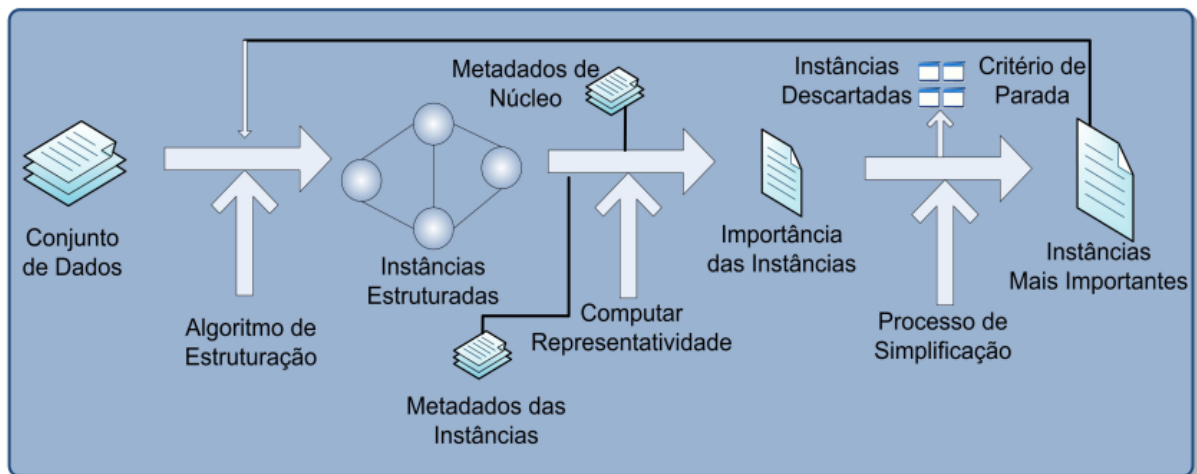
Esta seção apresenta o DGM-e e a forma como o conteúdo é organizado, incluindo a visão geral do método, o fluxo operacional, o algoritmo de construção de vizinhanças, o critério de parada e a análise de custo e escalabilidade.

O DGM-e é uma extensão do DGM, projetada para processar diferentes tipos de dados, e não apenas para a simplificação de malhas triangulares. Esta versão incorpora duas modificações significativas: a geração de uma estrutura geométrica a partir de dados não estruturados e a capacidade de comparar distintos atributos, seja nominal ou contínuo. O DGM-e engloba três fases cruciais: estruturação dos dados, avaliação da representatividade e simplificação das instâncias. Ações como escalonamento dos dados, divisão do espaço de dados e formação de vizinhanças são executadas para converter con-

juntos não estruturados em um grafo, viabilizando a aplicação do DGM. O algoritmo adota métricas de distância apropriadas para tratar atributos nominais e contínuos, e a escolha das instâncias segue um processo de simplificação guiado pelos valores de relevância. O critério de parada baseia-se na porcentagem de instâncias a serem removidas. O procedimento do DGM-e em mineração de dados envolve calcular a representatividade das instâncias e simplificá-las, repetindo esse ciclo até satisfazer o critério de parada. A análise da complexidade do método enfatiza sua eficiência escalável para grandes conjuntos de dados, apresentando uma complexidade linear $O((k+3q+3)n)$. Essa complexidade é alcançada considerando que o número de dimensões (k) e o limite para a busca aleatória de vizinhos (p) são muito menores que o número de instâncias (n) (k e $p \ll n$). Nesta expressão: n representa o número de instâncias no conjunto de dados, k representa o número de dimensões e p representa o limite para a busca aleatória de vizinhos (Silva, 2012).

O procedimento do modelo DGM-e para a simplificação de dados genéricos é detalhado na Figura 6 (Silva, 2012). Inicialmente, as instâncias são organizadas geometricamente através de um algoritmo de estruturação. Após esta fase, o próximo passo envolve o cálculo da representatividade de cada instância, o que define seus valores de importância. Esta etapa inclui a análise da representatividade com base nas informações provenientes dos arquivos de metadados, que detalham o tratamento dos atributos e as configurações do núcleo. Finalmente, o processo culmina na eliminação de instâncias, realizada conforme sua relevância e a metodologia de remoção adotada.

Figura 6: Processo de Redução de Dados.



Fonte: Silva (2012).

O Algoritmo 1 sintetiza o processo de determinação da vizinhança em um conjunto de dados S com n instâncias em k dimensões. Primeiro, aplica-se uma normalização Min-Max (Jayalakshmi; Santhakumaran, 2011) para $[0, 1]$ com faixa global, dada pela

Eq. (16); em seguida, identifica-se o hipercubo $H_{i,j}$ ao qual cada instância i pertence em cada dimensão j (Eq. 17), usando o número de segmentos r definido em 18. Depois, selecionam-se aleatoriamente p vizinhos por instância dentro do seu hipercubo. As instruções finais garantem reciprocidade nas ligações: se a instância u é conectada à instância v , então v também é conectada a u .

$$\hat{x}_{ij} = \frac{x_{ij} - m_j}{M - m}, \quad \text{com } m_j = \min_{1 \leq z \leq n} x_{zj}, \quad m = \min_{\substack{1 \leq z \leq n \\ 1 \leq w \leq k}} x_{zw}, \quad M = \max_{\substack{1 \leq z \leq n \\ 1 \leq w \leq k}} x_{zw}. \quad (16)$$

Algoritmo 1 Construção de Vizinhança

```

1: function BUILDNEIGHBORHOOD( $S$ : conjunto de dados normalizado)
2:   Atribuir o número de segmentos ( $r$ ) usando a equação (18)
3:   for cada instância  $i$  no conjunto  $S$  do
4:     Encontrar todas as  $j$  ( $j \leftarrow 1, k$ ) dimensões para o hipercubo  $H_{ij}$  usando a
       equação (17)
5:   end for
6:   for cada instância  $i$  no conjunto  $S$  do
7:     for  $j \leftarrow 1$  até  $p$  do
8:       Selecionar aleatoriamente dentro do hipercubo da instância uma instância
       vizinha  $e$ 
9:       Atribuir a instância  $e$  na lista de vizinhos de  $i$ 
10:      Atribuir a instância  $i$  na lista de vizinhos de  $e$ 
11:    end for
12:  end for
13:  return:  $S$  ▷ Agora cada instância em  $S$  possui seus vizinhos
14: end function

```

A construção de vizinhança em dados multidimensionais envolve estruturar o espaço de dados em segmentos uniformes. Inspirado pela *KD-Tree* (Moore, 1991), que generaliza árvores binárias para dimensões superiores, o espaço é dividido em hipercubos através de hiperplanos (Silva, 2012).

A equação (17), em que H_{ij} representa a coordenada j do hipercubo para a instância x_i , permite alocar cada instância a um hipercubo específico com complexidade linear $O(kn)$, otimizando o uso de memória ao gerar apenas os hipercubos que contêm instâncias. O número de segmentos r é adaptativamente determinado pela equação (18), considerando o número de instâncias n e a dimensão k , o que reflete a expectativa de uma redução no número de instâncias por iteração do método, mantendo a escalabilidade.

$$H_{ij} = \text{floor}(\hat{x}_{ij} * r) \quad (17)$$

$$r = \left(\frac{n}{k}\right)^{\frac{1}{k}} \quad (18)$$

No pré-processamento de dados utilizando o método descrito, a definição da vizinhança das instâncias é um passo crucial. Tradicionalmente, essa vizinhança poderia ser determinada pelas instâncias dentro do mesmo hipercubo ou estendendo-se aos hipercubos adjacentes. No entanto, para lidar com variações de densidade de dados e evitar vizinhanças de tamanho excessivo que prejudicam a performance, optou-se pela amostragem de Monte Carlo (Silva, 2012). Este método seleciona eficientemente um número fixo de vizinhos aleatórios de cada hipercubo, mantendo a complexidade do processamento em uma ordem linear $O(pn)$, ideal para grandes volumes de dados e regiões de alta densidade.

2.7 k -Vizinhos Mais Próximos

Esta seção introduz o algoritmo dos k -vizinhos mais próximos (KNN, k -nearest neighbours), técnica não paramétrica amplamente empregada em regressão e classificação, cuja ideia central é que instâncias semelhantes tendem a apresentar características ou rótulos de resposta similares (Izbicki; Santos, 2020).

Do ponto de vista formal, seja n o tamanho de uma amostra formada por pares independentes e identicamente distribuídos $(X_1, Y_1), \dots, (X_n, Y_n) \sim (X, Y)$, em que Y indica a variável resposta e $X = (x_1, \dots, x_d)$ reúne as covariáveis (ou *features*); para cada observação i escreve-se $x_i = (x_{i,1}, \dots, x_{i,d})$. Dada uma métrica $d(\cdot, \cdot)$, define-se para um ponto de consulta x o conjunto dos k vizinhos mais próximos

$$\mathcal{N}_x = \left\{ i \in \{1, \dots, n\} : d(x_i, x) \leq d_k^x \right\}, \quad (19)$$

em que d_k^x é a distância entre x e o seu k -ésimo vizinho. A forma mais usual utiliza a métrica Euclidiana, mas alternativas como Minkowski, Manhattan ou Mahalanobis podem ser adotadas conforme a natureza dos dados e do problema (Altman, 1992; Izbicki; Santos, 2020). No contexto da regressão, o KNN prevê o valor de uma nova observação calculando a média (ou mediana) das respostas de seus k vizinhos mais próximos no conjunto de treinamento (Izbicki; Santos, 2020).

$$g(x) = \frac{1}{k} \sum_{i \in \mathcal{N}_x} y_i, \quad (20)$$

enquanto, em classificação, a predição é determinada pela classe majoritária entre esses k vizinhos, funcionando como um voto majoritário (Izbicki; Santos, 2020).

Em síntese, o algoritmo KNN se inicia fixando o hiperparâmetro k e a métrica $d(\cdot, \cdot)$; em seguida, constrói-se (ou atualiza-se) uma estrutura de índices eficiente, como

KD-Tree, Ball-Tree ou grafos de vizinhança aproximada, para acelerar a busca; por fim, para cada nova observação x^* calculam-se as distâncias aos pontos de treino, selecionam-se os k menores valores que compõem $\mathcal{N}_{x^*}$ e obtém-se a predição $\hat{y}$ pela média das respostas em regressão ou pela classe modal em classificação (Cover; Hart, 1967; Altman, 1992; Izbicki; Santos, 2020).

A determinação do hiperparâmetro k é central para o desempenho do KNN, pois regula o compromisso entre viés e variância: valores pequenos reduzem o viés, mas aumentam a variância e a propensão ao sobreajuste, enquanto valores grandes suavizam excessivamente a função estimada, elevando o viés e favorecendo o subajuste (Izbicki; Santos, 2020; Fukunaga; Hostetler, 1973). O ponto ótimo de k geralmente é obtido por técnicas de validação, especialmente validação cruzada ou *data splitting*, que avaliam o erro preditivo em conjuntos de teste não utilizados no treinamento, garantindo poder de generalização adequado ao modelo (Izbicki; Santos, 2020).

A forma mais comum no KNN é a média uniforme, em que todos os k vizinhos contribuem igualmente para a predição. Alternativamente, adotam-se ponderações baseadas na distância, dando mais influência aos vizinhos mais próximos. No estimador de Nadaraya–Watson (Nadaraya, 1964; Watson, 1964), seja x_i o vetor de atributos da i -ésima observação de treino e x o ponto de consulta; a predição é uma média ponderada na qual os pesos $w_i(x)$ são definidos por uma função *kernel* de similaridade entre x_i e x e normalizados para somar 1. Assim, observações mais próximas recebem maior peso na estimativa (Izbicki; Santos, 2020).

A complexidade do KNN deriva da necessidade de calcular, para cada consulta, a distância entre a nova observação e todos os n pontos do conjunto de treinamento, operação que custa $O(nd)$ em tempo quando se consideram d covariáveis (ou *features*/atributos), além de exigir $O(nd)$ em memória para armazenar todo o conjunto de treino (Izbicki; Santos, 2020). Estruturas de partição espacial, como KD-Tree ou Ball-Trees, podem reduzir o tempo médio de busca para $O(\log n)$ em baixas dimensões, mas perdem eficiência à medida que d cresce, fenômeno conhecido como maldição da dimensionalidade; nestes casos, técnicas aproximadas são indispensáveis para preservar a escalabilidade do método.

Em comparação com métodos paramétricos ou baseados em modelo, o KNN mostra vantagens e desvantagens distintas. Uma de suas principais vantagens é ser um método não paramétrico e flexível, que não impõe suposições sobre a forma funcional da relação entre as covariáveis e a variável resposta, o que o torna capaz de capturar relações complexas nos dados. Além disso, sua simplicidade conceitual e facilidade de implementação são pontos fortes. Contudo, o KNN sofre significativamente com a maldição da dimensionalidade (*curse of dimensionality*), em que seu desempenho preditivo diminui drasticamente em espaços de alta dimensão, a menos que haja suposições adicionais como esparsidade

ou redundância nas covariáveis. Outra desvantagem é o custo computacional elevado para grandes conjuntos de dados, tanto em tempo de predição quanto em requisitos de memória. Adicionalmente, o KNN é sensível à escala das covariáveis e à presença de variáveis irrelevantes, que podem distorcer a noção de proximidade e impactar negativamente a qualidade das vizinhanças formadas (Izbicki; Santos, 2020).

2.8 Entropia de Shannon

Esta seção introduz a entropia de Shannon como medida da incerteza média de uma variável aleatória no escopo da Teoria da Informação, apresentando sua formulação $H(p) = -\sum_i p_i \log_b p_i$ e as unidades associadas à base b (*bits*, *nats* ou *dits*).

A noção de entropia foi introduzida por Clausius, em 1865, para representar a parcela de energia interna de um sistema que não pode ser convertida em trabalho útil. Décadas depois, em 1948, Claude Shannon aplicou o termo em um contexto totalmente distinto ao desenvolver a Teoria da Informação, definindo a entropia como a incerteza média, ou quantidade média de informação, associada ao resultado de uma variável aleatória discreta (Shannon, 1948; Cover; Thomas, 2005). Hoje, essa medida perpassa a Mecânica Estatística, a Probabilidade, os Sistemas Dinâmicos e, em especial, a Ciência de Dados e o Aprendizado de Máquina, servindo como fundamento para quantificar incerteza e otimizar processos de codificação (Sepúlveda-Fontaine; Amigó, 2024).

Para uma distribuição de probabilidade discreta $p = \{p_1, \dots, p_W\}$, a entropia de Shannon é formalizada por

$$H(p) = -\sum_{i=1}^W p_i \log_b p_i, \quad (21)$$

em que b define a unidade (*bits* para base 2, *nats* para base e ou *dits* para base 10) e, por convenção, o termo $0 \log_b 0$ é nulo. Essa expressão é a única, até um fator de escala positivo, que satisfaz os axiomas formulados por Shannon e Khinchin. Em primeiro lugar, a entropia nunca assume valores negativos e só é zero quando toda a probabilidade se concentra em um único evento. Ela varia continuamente com a distribuição de probabilidades e permanece inalterada se forem acrescentados eventos de probabilidade nula. Além disso, atinge o valor máximo quando todos os eventos são equiprováveis, refletindo o estado de maior incerteza possível. Por fim, apresenta a chamada aditividade forte: a entropia conjunta de duas variáveis pode ser decomposta na entropia de uma delas mais a entropia condicional da outra; caso sejam independentes, a soma das entropias individuais reproduz a entropia do sistema composto (Cover; Thomas, 2005; Amigó; Keller; Unakafova, 2015).

A partir desse núcleo conceitual emergem outras quantidades úteis. A informação

mútua expressa, em termos qualitativos, o quanto o conhecimento de uma variável reduz a incerteza sobre outra, sendo essencial em tarefas de seleção de atributos. A divergência de Kullback–Leibler avalia o custo, em média, de representar dados de uma distribuição usando outra, o que fornece um critério para treinamento de modelos probabilísticos. O princípio da máxima entropia, proposto por Jaynes, recomenda escolher, entre todas as distribuições que obedecem às restrições conhecidas, aquela com maior entropia, garantindo a inferência mais imparcial possível (Jaynes, 1957).

Esses conceitos alimentam um leque amplo de aplicações. Na seleção de atributos e na modelagem de linguagem, a informação mútua ajuda a identificar variáveis realmente informativas (Brown et al., 2012). Na detecção de anomalias em redes de sensores e ambientes IoT (*Internet of Things*), a entropia atua como termômetro de irregularidades em sequências de dados (DeMedeiros; Hendawi; Alvarez, 2023). Em métodos de Análise Envoltória de Dados e processos de decisão multicritério, ela fornece pesos objetivos para critérios, melhorando o poder discriminatório de avaliações de eficiência (Peykani et al., 2022). Também surge em análise de sinais biomédicos, diagnóstico de falhas mecânicas, regularização de estratégias em controle ótimo e inúmeras outras situações que exigem medir quantitativamente a incerteza ou a dispersão de informações. Por sua clareza formal e versatilidade prática, a entropia de Shannon continua sendo um dos fundamentos mais sólidos para o estudo e o tratamento da informação em múltiplas disciplinas.

3 DETECÇÃO AUTOMÁTICA DE ANOMALIAS BASEADA EM DIFUSÃO GEOMÉTRICA MARKOVIANA

Este capítulo descreve o algoritmo de *Detecção Automática de Anomalias baseada em Difusão Geométrica Markoviana* (DA-DGM), proposto como uma extensão dos métodos DGM e DGM-e para identificação não supervisionada de anomalias. O algoritmo baseia-se na difusão estocástica local sobre grafos de similaridade construídos a partir da estrutura de vizinhança dos dados.

A detecção ocorre por meio do cálculo das entradas diagonais da matriz de transição de Markov P , conforme definido na Equação (24). Essas auto-probabilidades P_{ii} indicam a persistência local da difusão em cada instância. Ao ranquear as amostras com base nas auto-probabilidades, de forma decrescente, obtém-se uma ordenação de anormalidade: os pontos com maior probabilidade de permanecer em si mesmos são considerados mais isolados e, portanto, mais anômalos.

O procedimento completo compreende quatro fases principais: pré-processamento dos dados; construção do grafo local via k vizinhos mais próximos; estimação do parâmetro de escala ε por meio da maximização da entropia de Shannon das auto-probabilidades; e detecção e ranqueamento dos escores de anomalia. O Algoritmo 2 resume o fluxo geral, e as seções seguintes detalham cada etapa, incluindo uma análise de complexidade e escalabilidade.

3.1 Pré-Processamento

Nesta etapa, todos os atributos são normalizados para o intervalo $[0, 1]$ por meio da transformação Min-Max (Jayalakshmi; Santhakumaran, 2011), garantindo comparabilidade entre as dimensões e estabilidade numérica durante o processo de difusão. A normalização é dada por:

$$x'_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)}, \quad (22)$$

em que x'_{ij} é o valor normalizado do atributo j da instância x_i , e os operadores $\min(x_j)$, $\max(x_j)$ referem-se aos extremos do atributo j no conjunto de dados.

Esse procedimento reduz vieses induzidos por variáveis de alta magnitude, melhora a sensibilidade do cálculo de similaridade e favorece uma difusão mais fiel à geometria local dos dados.

3.2 Construção do Grafo Local via K-Vizinhos Mais Próximos

Nesta seção, serão apresentados os fundamentos da construção do grafo local via k vizinhos mais próximos (KNN), incluindo o núcleo gaussiano de conectividade, o procedimento de normalização probabilística, os critérios de ajuste do parâmetro k , e as estratégias de implementação que garantem escalabilidade e reprodutibilidade.

Para modelar a estrutura geométrica dos dados, o DA-DGM constrói um grafo de vizinhança baseado em KNN. Para cada instância x_i , identifica-se um conjunto $\mathcal{N}_k(x_i) \subseteq X$ com as k instâncias mais próximas, segundo a métrica euclidiana.

A conectividade local é então quantificada por um núcleo gaussiano (*heat kernel*) sobre as distâncias, Equação (13), aplicado apenas para $j \in \mathcal{N}_k(x_i)$. O valor de k controla o alcance local da difusão: valores menores de k destacam microestruturas e anomalias pontuais, enquanto valores maiores suavizam irregularidades, reforçando estruturas globais.

Para garantir interpretabilidade probabilística, os pesos de cada instância são normalizados de forma que a soma de cada linha da matriz de transição seja 1. O termo de normalização é dado por:

$$d(x_i) = 1 + \sum_{j \in \mathcal{N}_k(x_i) \setminus \{x_i\}} e^{-\frac{\|x_i - x_j\|^2}{\epsilon}} \quad (23)$$

em que o termo 1 representa o auto-laço. Devido à construção via k vizinhos mais próximos, a matriz de transição $P = [P_{ij}]$ é esparsa e localmente adaptada, ou seja, os elementos P_{ij} são não nulos apenas para as transições entre uma instância x_i e seu conjunto de k vizinhos mais próximos ($j \in \mathcal{N}_k(x_i)$), e incorpora a configuração geométrica local e as interações com esses vizinhos.

A inclusão da etapa de k vizinhos mais próximos é necessária para garantir a viabilidade e a eficácia do DA-DGM, uma vez que ela estabelece a estrutura de vizinhança local sobre a qual a difusão opera. Se essa restrição não fosse imposta, a matriz de transição seria totalmente conectada (densa), o que resultaria em um custo computacional e de memória da ordem de $\Theta(n^2d)$ (no caso da construção de todas as distâncias), limitando a escalabilidade do método para grandes volumes de dados. Ao restringir o cálculo do núcleo gaussiano apenas aos k vizinhos mais próximos, o método assegura que a matriz P seja esparsa e localmente adaptada, o que permite ao algoritmo alcançar um crescimento em tempo médio sub-quadrático (ver Seção 3.5) para conjuntos multidimensionais e manter a sensibilidade geométrica local crucial para a identificação precisa de microestruturas e anomalias pontuais.

Por fim, adotou-se $k \geq 5$ como valor mínimo para garantir uma conectividade local

mínima e estimativas estáveis das probabilidades de transição. O parâmetro k foi ajustado automaticamente para cada base de dados, mantendo-o pequeno em relação ao tamanho da amostra n , de modo a equilibrar sensibilidade local e robustez global. É importante distinguir o ajuste de k da otimização do parâmetro de escala ε . Embora k seja ajustado automaticamente (no sentido de ser adaptativo ao conjunto de dados), ele é determinado por um procedimento de busca empírica iterativa (detalhado no Algoritmo 2). Nesse procedimento, k é inicializado e incrementado em passos fixos (Δk) até que se atinja um critério de parada, como o limite superior $k_{\max}$ ou a estabilidade do parâmetro ε (isto é, $\varepsilon \leq 1$). Assim, o ajuste de k visa garantir a estabilidade do grafo local e manter a esparsidade, sendo guiado por hiperparâmetros de busca ($k_0, \Delta k, k_{\max}$), e não pela maximização direta de um critério informacional, como ocorre com a Entropia de Shannon para a seleção de ε .

A formulação da normalização na Equação (23) constitui uma adaptação local e esparsa da definição geral de Cadeias de Markov Geométricas, Equações (8) e (9), embora ambas sirvam ao mesmo propósito probabilístico de garantir que a soma das probabilidades de transição de uma linha seja unitária. A principal distinção é dupla: primeiro, o fator $d(x_i)$ na Equação (23) restringe explicitamente a soma dos pesos apenas ao conjunto limitado dos k vizinhos mais próximos, $N_k(x_i) \setminus \{x_i\}$, uma estratégia essencial para garantir a escalabilidade do DA-DGM e manter a matriz P esparsa, mitigando o custo quadrático, $O(N^2)$, de matrizes densas. Segundo, o termo de afinidade do auto-laço, $k(x_i, x_i)$, é explicitamente fixado em 1 e somado separadamente, diferindo da inclusão implícita do auto-laço na soma total da Equação (9) sobre todo o conjunto X .

As consultas de KNN foram realizadas por meio de uma *KD-Tree* (Moore, 1991) em bases multidimensionais, aproveitando a indexação espacial para otimizar a velocidade das buscas. Nos casos em que $d = 1$, utilizou-se o método de força bruta, o qual preserva resultados exatos com complexidade computacional total de $O(n^2 d)$ para a construção do grafo. Ressalta-se que, para ordenar as distâncias do KNN, adotou-se o método estável *mergesort*, que resolve empates de forma determinística e mantém o próprio ponto na primeira posição da lista de vizinhos (distância zero). Desconsidera-se essa posição inicial no cálculo dos pesos (utilizam-se apenas $(j \geq 1)$, o que preserva a definição de P_{ii} e assegura reprodutibilidade dos resultados, com custo adicional desprezível por ocorrer em vetores de tamanho k por linha.

3.3 Seleção do Parâmetro de Escala via Entropia de Shannon

Esta seção define o parâmetro de escala (ε) e explica como ele é estimado maximizando-se a entropia de Shannon por meio de uma busca global em grade logarítmica seguida de refinamento local.

O parâmetro de escala, também denominado parâmetro de difusão ε em vários contextos, desempenha um papel fundamental em métodos que envolvem a análise de dados por difusão, como os Mapas de Difusão. Este parâmetro é frequentemente associado à noção de similaridade entre os dados em um espaço métrico, influenciando diretamente a conectividade do grafo que modela essas relações. Em termos mais práticos, o parâmetro de escala define o “tamanho” da vizinhança considerada ao avaliar a similaridade entre pontos de dados, afetando a matriz de similaridade que é central para muitos algoritmos de aprendizado de máquina e análise de dados (Silva, 2012; Breiman, 2001; Singer et al., 2009). Valores pequenos de ε privilegiam conexões locais, produzindo matrizes de transição esparsas que capturam detalhes geométricos finos, ao passo que valores grandes promovem conectividade quase uniforme e, conseqüentemente, borram variações relevantes (Singer; Wu, 2011; Coifman; Lafon, 2006).

A seleção de hiperparâmetros em modelos estocásticos, como a escolha do parâmetros de escala, é crucial, mas heurísticas *ad hoc* ainda são comuns porque não existe um valor universalmente ótimo. A escolha do valor de ε depende do conjunto de dados em questão. Neste trabalho, adota-se um critério de natureza informacional: escolher o valor do parâmetro de escala que maximiza a entropia de Shannon das entradas diagonais da matriz de transição, associadas às probabilidades de auto-transição. Até o momento, não foram encontrados registros desse critério na literatura.

Do ponto de vista estatístico, a entropia de Shannon quantifica a incerteza ou desordem de um sistema: entropias maiores indicam maior variabilidade na distribuição de probabilidades e, portanto, aumento de imprevisibilidade. Explorando esse princípio, a estratégia proposta usa a entropia da diagonal de P como regularizador dinâmico. Maximizar essa quantidade desencoraja persistências extremas, isto é, estados excessivamente estáveis ou puramente transitórios, e produz uma distribuição mais homogênea de probabilidades de auto-transição. O modelo resultante torna-se menos suscetível a sobreajustes a dinâmicas frágeis ou ruidosas, ampliando sua capacidade de generalização. Diferentemente de métodos que requerem dados externos de treinamento, esta abordagem baseia-se apenas em propriedades intrínsecas de P . A formulação matemática correspondente é apresentada a seguir.

Seja $P = [P_{ij}]$ a matriz de transição introduzida na Equação (8), em que $k(x_i, x_j)$, Equação (13), é o núcleo gaussiano e o termo de normalização, $d(x_i)$, é dado na Equação (23).

Esse núcleo gaussiano, Equação (13), representa uma generalização do método DGM original, que empregava vetores normais para simplificação de malhas, para o contexto de conjuntos genéricos de dados (DGM-e). Essa mudança substitui a dependência de vetores normais pela distância Euclidiana ($\|x_i - x_j\|^2$) no espaço de atributos, o que é essencial para aplicar o método a dados tabulares multivariados.

Para determinar o parâmetro de escala automaticamente, adota-se como critério a entropia de Shannon das auto-probabilidades P_{ii} , que medem a permanência da difusão em cada ponto (entradas diagonais):

$$P_{ii} = \frac{1}{d(x_i)}. \quad (24)$$

A entropia de Shannon associada a essas probabilidades é:

$$H(P) = - \sum_{i=1}^n P_{ii} \ln(P_{ii}), \quad (25)$$

e o parâmetro de escala selecionado sob o critério de Shannon é:

$$\varepsilon = \arg \max_{\varepsilon} H(P). \quad (26)$$

A otimização automática é viável porque o P_{ii} de cada instância é diretamente dependente do fator de normalização $d(x_i)$, Equação (24), que por sua vez é uma função do parâmetro de escala ε , Equação (23)). O ε regula geometricamente a conectividade local no grafo; assim, ao variar ε , varia-se o vetor de auto-probabilidades. O critério da maximização da Entropia de Shannon é, portanto, aplicado como uma função objetivo para encontrar o ε que gera o maior contraste (ou variabilidade) na distribuição dos valores de P_{ii} . Esse ε ideal equilibra a alta probabilidade de auto-transição em anomalias isoladas (altos P_{ii}) com a baixa probabilidade de auto-transição em instâncias normais (baixos P_{ii}), tornando a distribuição dos escores mais discriminativa e eliminando a dependência do ajuste manual de hiperparâmetros.

A entropia favorece um equilíbrio entre pontos densos (pequenos P_{ii}) e isolados (altos P_{ii}), amplificando o contraste entre instâncias típicas (normais) e aquelas que se desviam do padrão (anômalas).

O parâmetro de escala ε é crítico e sua faixa de valores ideal depende intrinsecamente da distribuição das distâncias no conjunto de dados. Essa faixa define o espaço em que ε deve ser buscado para capturar adequadamente a geometria local e global dos dados. A estratégia de maximização da entropia de Shannon, Equação (26), é aplicada para localizar o ponto ótimo dentro dessa faixa relevante.

A estimação do parâmetro de escala ocorre em duas etapas: uma busca global sobre malha logarítmica e um refinamento local em torno do máximo identificado. Essa abordagem hierárquica garante tanto a robustez quanto a precisão do parâmetro. Embora o limite teórico superior de ε seja o diâmetro do conjunto de dados, e o inferior, o limite de conectividade, a busca é implementada sobre uma malha logarítmica. Esta escolha é

justificada pela necessidade de cobrir eficientemente múltiplas ordens de magnitude, visto que ε impacta a conectividade de forma exponencial. Os limites experimentais ajustáveis (como $\varepsilon \in [10^{-15}, 10^0]$) são definidos para serem conservadores, abrangendo de forma robusta os limites de distância teóricos em conjuntos de dados que foram previamente normalizados para o intervalo 0-1. A normalização tende a concentrar o diâmetro da maioria dos conjuntos em uma ordem de magnitude próxima de 1, garantindo que a busca entre 10^{-15} e 10^0 encontre o ótimo de entropia mesmo em conjuntos com distâncias mínimas extremamente pequenas ou que exijam suavização máxima.

O refinamento local define um intervalo I centrado no valor ótimo encontrado na busca global (ε_g). A amplitude desse intervalo é controlada pelo parâmetro largura relativa (δ), que estabelece $I = [(1 - \delta)\varepsilon_g, (1 + \delta)\varepsilon_g]$. A largura relativa é um dos parâmetros de entrada listados no Algoritmo 2 e garante que a otimização local ocorra em um raio limitado do valor de escala global.

Para conjuntos de dados unidimensionais previamente normalizados, utiliza-se diretamente $\varepsilon = 1$, conforme evidências experimentais que apontam para um desempenho ótimo sem a necessidade de otimização adicional. Esse valor é consistente com a formulação original da DGM e de sua extensão, a DGM-e, que implicitamente adota o mesmo parâmetro, garantindo robustez e simplicidade na aplicação.

3.4 Detecção de Anomalias

Esta seção apresenta o mecanismo de detecção de anomalias do DA-DGM, detalhando como o valor de P_{ii} é utilizado como medida de isolamento e convertido em um *ranking* de anormalidade. Descreve-se o procedimento de ordenação das instâncias, o uso opcional da taxa de contaminação ν para rotulagem direta e o funcionamento do método quando tal parâmetro é desconhecido. Por fim, introduz-se o algoritmo que reúne todas as etapas do processo até a geração da lista final de anomalias.

Uma vez definido o parâmetro de escala e construída a matriz P , o método prossegue para a etapa de detecção. A lógica é simples: quanto maior for P_{ii} , maior o isolamento da instância x_i , e, portanto, maior a chance de ser uma anomalias.

As instâncias são ranqueadas por:

$$r = \text{argsort}(P_{ii}, \text{decrecente}), \quad (27)$$

de forma que $r[1]$ representa a instância mais anômala.

O algoritmo oferece dois modos de classificação: quando a taxa de contaminação ν , que representa a proporção ou porcentagem esperada de anomalias em um determinado conjunto de dados, é conhecida, o método rotula como anômalas as $\lceil \nu n \rceil$ primeiras

posições de r ; caso contrário, retorna-se o vetor r como um *ranking* contínuo, permitindo análise interativa ou definição posterior do limiar de corte com base em critérios estatísticos ou conhecimento de domínio.

O Algoritmo 2 formaliza todas as etapas, desde o pré-processamento até a geração do *ranking* final de anomalias.

Algoritmo 2 DA-DGM: Detecção Automática de Anomalias por Difusão Geométrica Markoviana

Require: Dados normalizados $X \in \mathbb{R}^{n \times d}$.

Require: Malha logarítmica $\mathcal{E} = \{\varepsilon_j\}_{j=1}^m$, definida conforme a base:

- 1D: $\varepsilon = 1$ (fixo);
- 2D/3D: $\varepsilon \in [10^{-8}, 10^{-1}]$;
- Multivariadas com atributos específicos (e.g., financeiros): $\varepsilon \in [10^{-15}, 10^0]$;
- Demais multidimensionais: $\varepsilon \in [10^{-4}, 10^3]$.

(Limites ajustáveis para garantir contraste adequado em P_{ii} .)

Require: Parâmetros: k_0 (início), Δk (passo), $k_{\max}$ (limite superior), N_{refine} (padrão = 20), δ (largura relativa; padrão = 0,20).

Refinamento local em $I = [(1 - \delta)\varepsilon_g, (1 + \delta)\varepsilon_g]$ com N_{refine} amostras uniformes.

Ensure: Parâmetro de escala ε , número final de vizinhos k_{eff} , vetor de auto-probabilidades P_{ii}^* , *ranking* $\mathbf{r}$

```

1: if  $d = 1$  then                                ▷ Valor fixo para dados unidimensionais
2:   Definir  $k_{\text{eff}} \leftarrow n, \varepsilon^* \leftarrow 1$ 
3:   Ir para CÁLCULO FINAL
4: end if
5: Inicializar  $k \leftarrow k_0$ 
6: repeat
7:   Computar distâncias quadradas  $D \in \mathbb{R}^{n \times k}$  via KNN-DISTÂNCIAS( $X, k$ ) ▷ KNN-
   DISTÂNCIAS retorna, para cada ponto, as  $k$  menores distâncias quadradas (incluindo
   a si mesmo)
8:   for all  $\varepsilon \in \mathcal{E}$  do
9:     Calcular  $P_{ii}(\varepsilon)$  com AUTOPROBS( $D, \varepsilon$ )                                ▷  $P_{ii} = \frac{1}{1 + \sum_{j=2}^k e^{-\|x_i - x_{i_j}\|^2/\varepsilon}}$ 
10:    Calcular  $H(\varepsilon) = -\sum_i P_{ii} \log P_{ii}$ 
11:  end for
12:  Determinar  $\varepsilon_g \leftarrow \arg \max_{\varepsilon \in \mathcal{E}} H(\varepsilon)$ 
13:  Definir  $I = [(1 - \delta)\varepsilon_g, (1 + \delta)\varepsilon_g]$ ; amostrar  $N_{\text{refine}}$  valores  $\{\tilde{\varepsilon}_\ell\} \subset I$ 
14:  for all  $\tilde{\varepsilon}_\ell$  do                                ▷ Refinamento local
15:    Repetir o cálculo de  $P_{ii}$  e  $H(\tilde{\varepsilon}_\ell)$ 
16:  end for
17:  Selecionar  $\varepsilon \leftarrow \arg \max H$  entre todos os valores testados
18:  if  $\varepsilon \leq 1$  ou  $k \geq k_{\max}$  then
19:    Definir  $k_{\text{eff}} \leftarrow k$  e sair do laço
20:  else
21:    Atualizar  $k \leftarrow k + \Delta k$ 
22:  end if
23: until condição de parada
Cálculo Final:
24: Recalcular  $D \leftarrow \text{KNN-DISTÂNCIAS}(X, k_{\text{eff}})$ 
25: Obter  $P_{ii} \leftarrow \text{AUTOPROBS}(D, \varepsilon)$ 
26: Gerar  $\mathbf{r} \leftarrow \text{ARGSORT}(-P_{ii})$ 
27: return  $\varepsilon, k_{\text{eff}}, P_{ii}, \mathbf{r}$ 

```

3.5 Análise de Complexidade

Esta seção apresenta a complexidade computacional do DA-DGM, relacionando o custo temporal e o uso de memória aos parâmetros do algoritmo. Diferenciam-se os casos unidimensionais e multidimensionais, ressaltando o impacto da varredura exaustiva ou do uso de *KD-Tree* na busca de vizinhos. Ao final, descrevem-se a expressão geral do tempo de execução, o pico de memória associado e uma tabela comparativa com outros algoritmos.

Para compreender o comportamento algorítmico em diferentes cenários, é necessário primeiro estabelecer a notação e identificar os fatores que determinam o custo computacional em cada regime dimensional.

Sejam n o número de instâncias, d a dimensionalidade, k o número de vizinhos por difusão, m a cardinalidade da malha global de ε , r a cardinalidade do refinamento local, $c = m + r$ o total de valores de ε avaliados por tentativa, e $t = \lceil (k_{\max} - k_{\text{start}})/k_{\text{step}} \rceil + 1$ o número máximo de iterações do laço adaptativo de k . No Algoritmo 2, a cada tentativa de k ocorre um pré-cálculo de vizinhos seguido da avaliação de c valores de ε sobre o mesmo conjunto de distâncias.

O mecanismo de busca de vizinhança adota varredura exaustiva quando $d = 1$ e *KD-Tree* quando $d \geq 2$. Para $d = 1$, impõe-se $k = n$ e realiza-se uma comparação todos-contra-todos entre os pontos; como não há etapa relevante de indexação, o custo temporal reduz-se ao cálculo de todas as distâncias, isto é, $\Theta(n^2 d) = \Theta(n^2)$. As distâncias podem ser mantidas em uma matriz $n \times n$, o que implica uso de memória $\Theta(n^2)$. Para $d \geq 2$, com *KD-Tree*, o custo médio do pré-cálculo de k -NN por tentativa é $\mathcal{O}(dn \log n + d n k)$, enquanto a memória para armazenar as k distâncias por ponto é $\Theta(nk)$; em alta dimensionalidade pode haver degradação para $\mathcal{O}(n^2 d)$.

Dado o bloco de distâncias $D \in \mathbb{R}^{n \times k}$, cada valor de ε é avaliado em $\Theta(nk)$ para computar os pesos gaussianos e P_{ii} . Em cada tentativa de k , a busca global com m valores e o refinamento local com r valores totalizam $T_\varepsilon = \Theta(c n k)$. Durante a busca global, o Algoritmo 2 mantém transitoriamente uma lista com m vetores P_{ii} de tamanho n , o que gera pico adicional de memória $\Theta(mn)$ além de $\Theta(nk)$; esse pico não persiste fora da função.

Somando as parcelas por tentativa de k e considerando até t iterações, o tempo total para $d \geq 2$ é

$$T_{\text{total}} = \mathcal{O}\left(t[d n \log n + d n k + c n k]\right), \quad (28)$$

com pico de memória

$$M_{\text{pico}} = \Theta(nk + mn). \quad (29)$$

Tabela 1: Comparação metodológica e de complexidade entre algoritmos de detecção de anomalias.

Algoritmo	Complex.	Escala	Sensibilidade a parâmetros	Base conceitual
DA-DGM ($d \geq 2$)	$\mathcal{O}(t[dN \log N + dNk + cNk])$	Subquadrática (média)	Baixa (ε automatizado por entropia)	Difusão geom. Markoviana
DA-DGM ($d = 1$)	$\Theta(N^2)$	Quadrática (alta)	Baixa (ε fixo em 1)	Difusão geom. Markoviana
KNN	$\mathcal{O}(N^2d)$	Quadrática (alta)	Alta (ajuste manual de k)	Distância / proximidade
LOF / COF	$\mathcal{O}(N^2d)$	Quadrática (alta)	Alta (ajuste manual de k)	Densidade local relativa
ABOD (fast)	$\mathcal{O}(N^2 + Nk^2)$	Quadrática (alta)	Alta (variância angular)	Ângulo / distância
HBOS	$\mathcal{O}(N)$	Linear (baixa)	Moderada (tamanho do <i>bin</i>)	Estatística (histograma)
IForest (teste)	$\mathcal{O}(N t \log \psi)$	Quase linear (baixa)	Moderada (t , subamostra ψ)	Particionamento de isolamento
Notas: N instâncias; d dimensionalidade; k vizinhos; t iterações/árvores; ψ subamostra; c nº de valores de ε .				

Para $d = 1$, como o Algoritmo 2 fixa $k = n$ e avalia um único ε , valem $T_{\text{total}} = \Theta(n^2)$ e $M_{\text{pico}} = \Theta(n^2)$. Na prática, com t, k, c moderados, o termo dominante em n para $d \geq 2$ é tipicamente $n \log n$ do *KD-Tree*, resultando em crescimento sub-quadrático; a etapa específica do DA-DGM após o pré-cálculo escala linearmente em cnk .

A tabela a seguir resume a complexidade temporal e o contraste metodológico entre os algoritmos em que N é o número de instâncias, d é a dimensionalidade, k é o número de vizinhos, t é o número de iterações/árvores, ψ é o tamanho da subamostra e c é o total de valores ε avaliados por tentativa.

O contraste metodológico entre o DA-DGM e os algoritmos de referência revela importantes *trade-offs* entre escalabilidade e robustez geométrica. Em termos de complexidade algorítmica, o DA-DGM, ao utilizar indexação espacial (*KD-Tree*) e reutilização de vizinhanças (custo médio sub-quadrático para $d \geq 2$), é mais escalável que KNN, LOF, COF e ABOD (tipicamente $\mathcal{O}(N^2)$), embora seja menos eficiente que IForest e HBOS (Quase-Linear/Linear). A principal distinção reside na sensibilidade a parâmetros: enquanto métodos de vizinhança como KNN e LOF dependem criticamente do ajuste manual de k , o DA-DGM elimina a dependência do ajuste manual do parâmetro de escala (ε) por meio da maximização automática da Entropia de Shannon. Quanto à interpretabilidade, o DA-DGM ancora-se em conceitos geométricos (probabilidade de auto-transição P_{ii}), sendo capaz de capturar relações em variedades curvas, ao passo que o IForest, embora eficiente, baseia-se em particionamentos retilíneos, preservando apenas parcialmente estruturas geométricas complexas.

4 AVALIAÇÃO EXPERIMENTAL

Este capítulo descreve os experimentos realizados com o DA-DGM, conduzidos em duas etapas complementares. Inicialmente, foram utilizados conjuntos de dados sintéticos, nos quais as instâncias normais formam estruturas geométricas complexas e as anomalias estão dispersas em diferentes regiões. Esses conjuntos permitem controle rigoroso sobre a geometria, o nível de contaminação e as escalas, possibilitando a validação precisa do ranqueamento baseado nas autoprobabilidades de transição e no parâmetro de escala ε , selecionado por meio da maximização da entropia de Shannon. Essa abordagem testou a robustez do algoritmo em diversas distribuições. Posteriormente, o método foi avaliado em dez bases de dados reais com diferentes dimensionalidades e proporções de anomalias, para verificar seu desempenho em cenários práticos.

4.1 Métrica de Avaliação

Esta seção apresenta a métrica utilizada para avaliar o desempenho dos métodos, adotando a Área Sob a Curva ROC (AUC-ROC) como medida principal. Resume-se o funcionamento da curva ROC por meio das taxas TPR (*True Positive Rate*) e FPR (*False Positive Rate*), bem como a interpretação probabilística da AUC e seus intervalos de qualidade, justificando sua escolha por ser independente de limiar, robusta a transformações monótonas e adequada à comparação de algoritmos de detecção de anomalias.

A métrica adotada é a AUC-ROC, por oferecer medida global, independente de limiar e estável a transformações monótonas dos escores (Bradley, 1997; Fawcett, 2006). A curva ROC plota a Taxa de Verdadeiros Positivos (TPR) no eixo y em função da Taxa de Falsos Positivos (FPR) no eixo x (Fawcett, 2006), com

$$\text{TPR} = \frac{TP}{TP + FN} \quad \text{e} \quad \text{FPR} = \frac{FP}{FP + TN}.$$

em que TP (*True Positive*) e FN (*False Negative*) correspondem, respectivamente, a anomalias corretamente e incorretamente identificadas, enquanto TN (*True Negative*) e FP (*False Positive*) denotam normais corretamente e incorretamente identificados.

A AUC quantifica a área sob a ROC e admite interpretação probabilística: a probabilidade de o classificador ranquear uma instância positiva aleatória acima de uma negativa aleatória (Bradley, 1997; Hand; Till, 2001). Seus valores variam de 0 a 1 (**com 0,5 indicando desempenho aleatório e 1,0 separação perfeita**) (Fawcett, 2006). Para referência interpretativa, (Hosmer David W.; Lemeshow; Sturdivant, 2013) sugerem: $0,7 \leq \text{AUC} < 0,8$ (**aceitável**), $0,8 \leq \text{AUC} < 0,9$ (**excelente**) e $\text{AUC} \geq 0,9$ (**excepcional**). A AUC-ROC é menos sensível do que a acurácia às variações de prevalência e a custos desiguais de erro e não exige a definição prévia de um limiar (Bradley, 1997;

Fawcett, 2006; Hand; Till, 2001). Assim, ela é apropriada para comparar métodos em detecção de anomalias quando o ponto de operação não é previamente fixado (Samariya; Thakkar, 2023).

4.2 Protocolo de Comparação e Visualizações

Esta seção apresenta o protocolo de comparação e visualizações. Para suporte visual, apresentam-se *boxplots* das auto-probabilidades P_{ii} e projeções uni, bi e tridimensionais que auxiliam na interpretação do ranqueamento e na distinção entre instâncias normais e anômalas.

Esta seção descreve o protocolo experimental empregado para comparar o DA-DGM com métodos clássicos de detecção de anomalias, bem como as visualizações utilizadas na análise dos resultados. Diante da variedade de técnicas disponíveis, a escolha de um método adequado não é trivial; por isso foram selecionados seis algoritmos representativos de diferentes famílias metodológicas (Samariya; Thakkar, 2023): HBOS (Goldstein; Dengel, 2012) (estatístico), LOF (Breunig et al., 2000) e COF (Tang et al., 2002) (baseados em densidade), KNN (Knorr; Ng; Tucakov, 2000) e ABOD (Kriegel; Schubert; Zimek, 2008) (baseados em distância) e *Isolation Forest* (Liu; Ting; Zhou, 2008) (baseado em isolamento). Todas as implementações utilizaram a biblioteca PyOD (Zhao; Nasrullah; Li, 2019) e foram configuradas apenas com a taxa de contaminação nominal de cada conjunto, sem ajustes adicionais, o que espelha condições de uso real e garante comparabilidade justa entre métodos. Embora Bouman, Bukhsh e Heskes (2024) relatem que o *Extended Isolation Forest* se destaca em anomalias globais e o KNN em anomalias locais, o *Isolation Forest* foi mantido aqui como representante canônico da família de isolamento em virtude de sua maturidade e ampla adoção.

Como apoio interpretativo, apresentam-se *boxplots* das auto-probabilidades P_{ii} , que evidenciam posição, dispersão, assimetria e anomalias; e projeções uni, bi e tridimensionais para contraste visual entre instâncias normais e anômalas. No ranqueamento do DA-DGM (ordenação decrescente de P_{ii}), as anomalias tendem a ocupar os percentis mais baixos (primeiras posições), refletindo maiores auto-probabilidades em regiões de baixa conectividade difusiva.

4.3 Conjuntos de Dados Sintéticos

Os experimentos com dados sintéticos foram divididos em três grupos: (i) distribuições unidimensionais *Gaussiana*, *Beta* e *Gama*; (ii) seis conjuntos bidimensionais com geometrias distintas (*Anel*, *Cantos*, *Curvas*, *Espiral*, *Gaussiano* e *Ondas*); e (iii) três conjuntos tridimensionais mais complexos (Cantos 3D, Curva Helicoidal e Espiral 3D). Para cada

grupo especificam-se o processo de geração, o critério de rotulagem das anomalias, os parâmetros do DA-DGM e as métricas usadas na comparação com os métodos de referência.

4.3.1 Dados Unidimensionais

Geraram-se três conjuntos unidimensionais de 10.000 instâncias a partir de distribuições estatísticas diferentes: *Gaussiana*, *Beta* ($\alpha = 2$, $\beta = 5$) e *Gama* ($\alpha = 5$, $\beta = 1$). As anomalias foram identificadas pela regra do *Z-score* ($|z| > 3\sigma$), resultando em 23, 34 e 91 anomalias, respectivamente.

A Tabela 2 apresenta, para essas distribuições, o número de anomalias detectadas corretamente por cada método e a AUC correspondente.

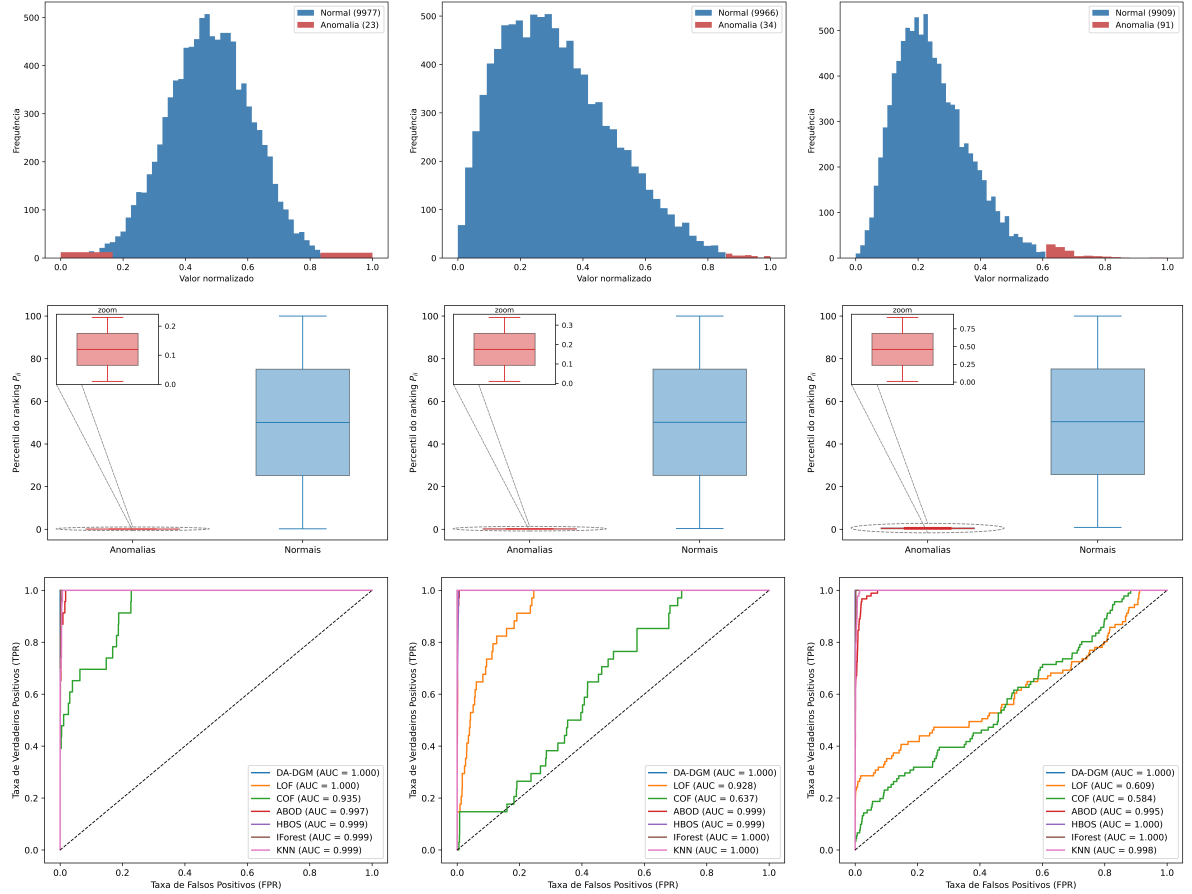
Tabela 2: Resultados consolidados para os conjuntos de dados unidimensionais.

Distribuição	Métrica	DA-DGM	KNN	HBOS	LOF	COF	ABOD	IForest
<i>Gaussiana</i>	Anomalias	23	16	7	17	9	14	15
	AUC (%)	100,00	99,85	99,89	99,96	93,45	99,74	99,94
<i>Beta</i>	Anomalias	34	30	20	5	0	21	34
	AUC (%)	100,00	99,99	99,90	92,81	63,69	99,88	100,00
<i>Gama</i>	Anomalias	91	71	22	22	6	60	78
	AUC (%)	100,00	99,82	99,98	60,91	58,35	99,46	99,96

AUC (%) = área sob a curva ROC; valores em negrito indicam o melhor resultado em cada linha.

A Figura 7 resume os resultados obtidos para as distribuições *Gaussiana*, *Beta* e *Gama*. Nos histogramas, observa-se que as anomalias se concentram nas regiões de cauda das distribuições normais, em que a densidade de pontos é menor. Os diagramas de caixa (*boxplots*), que representam o *ranking* do modelo DA-DGM, mostram que as instâncias normais se distribuem por toda a faixa de percentis (0-100), com mediana em 50%. Em contrapartida, para todas as três distribuições, as anomalias são mapeadas para percentis de *ranking* extremamente baixos, com suas distribuições inteiramente contidas abaixo do percentil 1. Por fim, as curvas ROC indicam o desempenho dos sete algoritmos: o DA-DGM alcançou 100% de AUC em todos os cenários, enquanto IForest, HBOS, KNN e ABOD mantiveram AUCs consistentemente acima de 99%. Os modelos LOF e COF registraram os menores valores, com quedas acentuadas na distribuição *Gama*, em que suas AUCs foram de 60,91% e 58,35%, respectivamente.

Figura 7: Resultados para as distribuições de dados unidimensionais *Gaussiana*, *Beta* e *Gama*. Linha 1: histogramas das amostras, em que barras azuis representam instâncias normais e vermelhas indicam anomalias. Linha 2: *boxplots* dos percentis do ranqueamento da probabilidade de auto-transição obtidos pelo DA-DGM (azul = normal, vermelho = anômalo). Linha 3: curvas *ROC* com as áreas sob a curva (AUC) dos algoritmos avaliados, permitindo comparar seu desempenho discriminatório.



Fonte: O autor, 2025.

4.3.2 Dados Bidimensionais

Inspirado no trabalho de Aryal, Santosh e Dazeley (2021), construíram-se seis conjuntos de dados bidimensionais denominados *Anel*, *Cantos*, *Curvas*, *Espiral*, *Gaussiano* e *Ondas*, cada um contendo pouco mais de 2.000 instâncias e projetado para reproduzir diferentes configurações geométricas, com taxas de anomalia previamente controladas.

A Tabela 3 resume, para cada conjunto, o total de anomalias recuperadas corretamente e a AUC dos métodos avaliados.

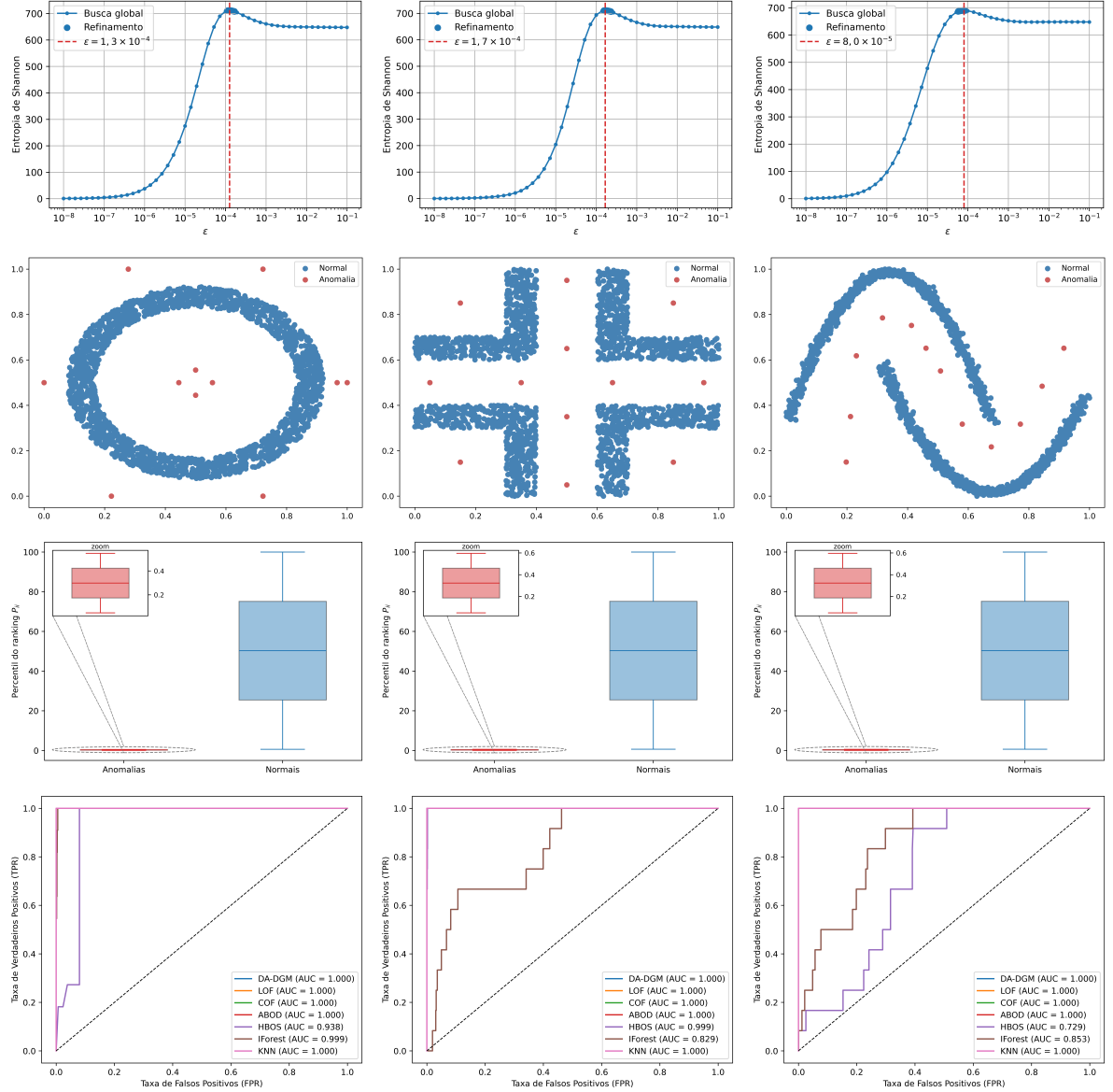
Tabela 3: Resultados consolidados para os conjuntos de dados bidimensionais.

Conjunto de dados	Métrica	DA-DGM	KNN	HBOS	LOF	COF	ABOD	IForest
<i>Anel</i>	Anomalias	11	11	0	11	11	11	7
	AUC (%)	100,00	100,00	93,84	100,00	100,00	100,00	99,86
<i>Cantos</i>	Anomalias	12	12	9	12	12	12	0
	AUC (%)	100,00	100,00	99,95	100,00	100,00	100,00	82,93
<i>Curvas</i>	Anomalias	12	12	1	12	12	12	1
	AUC (%)	100,00	100,00	72,93	100,00	100,00	100,00	85,34
<i>Espiral</i>	Anomalias	15	15	1	15	15	15	3
	AUC (%)	100,00	100,00	67,61	100,00	100,00	100,00	68,90
<i>Gaussiano</i>	Anomalias	20	20	16	20	16	20	17
	AUC (%)	100,00	100,00	99,76	100,00	99,82	100,00	99,94
<i>Ondas</i>	Anomalias	12	12	0	12	12	12	2
	AUC (%)	100,00	100,00	65,42	100,00	100,00	100,00	79,34

AUC (%) = área sob a curva ROC; valores em negrito indicam o melhor resultado em cada linha.

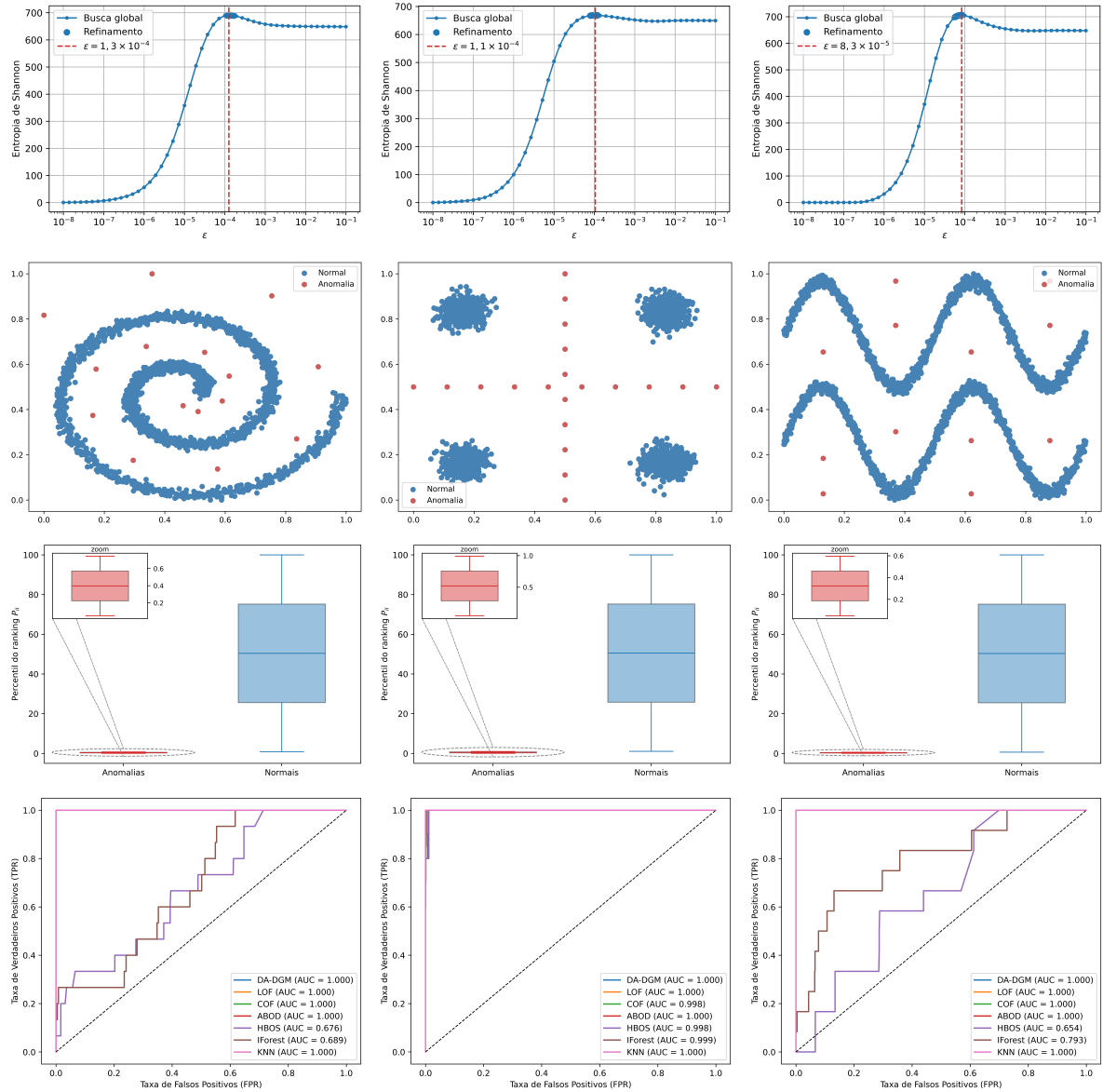
As Figura 8 e 9 apresentam os resultados para os seis conjuntos de dados bidimensionais analisados. As curvas $\varepsilon \times$ entropia de Shannon mostram a presença de picos bem definidos, utilizados para a seleção do parâmetro de escala em cada cenário. Os diagramas de dispersão revelam que os conjuntos possuem geometrias complexas e não-lineares, com anomalias distribuídas tanto de forma global quanto local. Os *boxplots*, correspondentes ao ranqueamento da probabilidade de auto-transição obtido pelo DA-DGM, mostram que as instâncias normais se distribuem ao longo de toda a faixa de percentis (0–100), com mediana próxima de 50%, enquanto as anomalias são sistematicamente mapeadas para valores próximos de zero, com distribuições concentradas abaixo do percentil 1. Por fim, as curvas ROC confirmam um desempenho agrupado em dois blocos: DA-DGM, KNN, LOF e ABOD atingiram AUC de 100% em todos os testes, enquanto IForest e HBOS apresentaram resultados variáveis, com quedas acentuadas nos cenários mais complexos, registrando AUC entre 65,42% e 85,34%.

Figura 8: Resultados para os conjuntos de dados bidimensionais *Anel*, *Cantos* e *Curvas*. Linha 1: curvas $\varepsilon \times$ entropia de Shannon utilizadas para selecionar o parâmetro de escala. Linha 2: gráficos de dispersão das instâncias classificadas pelo DA-DGM (azul = normal, vermelho = anômalas). Linha 3: *boxplots* dos percentis do ranqueamento da probabilidade de auto-transição. Linha 4: curvas *ROC* com as áreas sob a curva (AUC) dos algoritmos avaliados, permitindo comparar seu desempenho discriminatório.



Fonte: O autor, 2025.

Figura 9: Resultados para os conjuntos de dados bidimensionais *Espiral*, *Gaussiano* e *Ondas*. Linha 1: curvas $\varepsilon \times$ entropia de Shannon utilizadas para selecionar o parâmetro de escala. Linha 2: gráficos de dispersão das instâncias classificadas pelo DA-DGM (azul = normal, vermelho = anômalas). Linha 3: *boxplots* dos percentis do ranqueamento da probabilidade de auto-transição. Linha 4: curvas *ROC* com as áreas sob a curva (AUC) dos algoritmos avaliados, permitindo comparar seu desempenho discriminatório.



Fonte: O autor, 2025.

4.3.3 Dados Tridimensionais

Para simular cenários de maior complexidade espacial, geraram-se três conjuntos tridimensionais: *Cantos 3D*, *Curva Helicoidal* e *Espiral 3D*, cada qual com pouco mais de 2.000 instâncias e proporções conhecidas de anomalias.

A Tabela 4 lista o total de anomalias identificadas corretamente e a AUC correspondente.

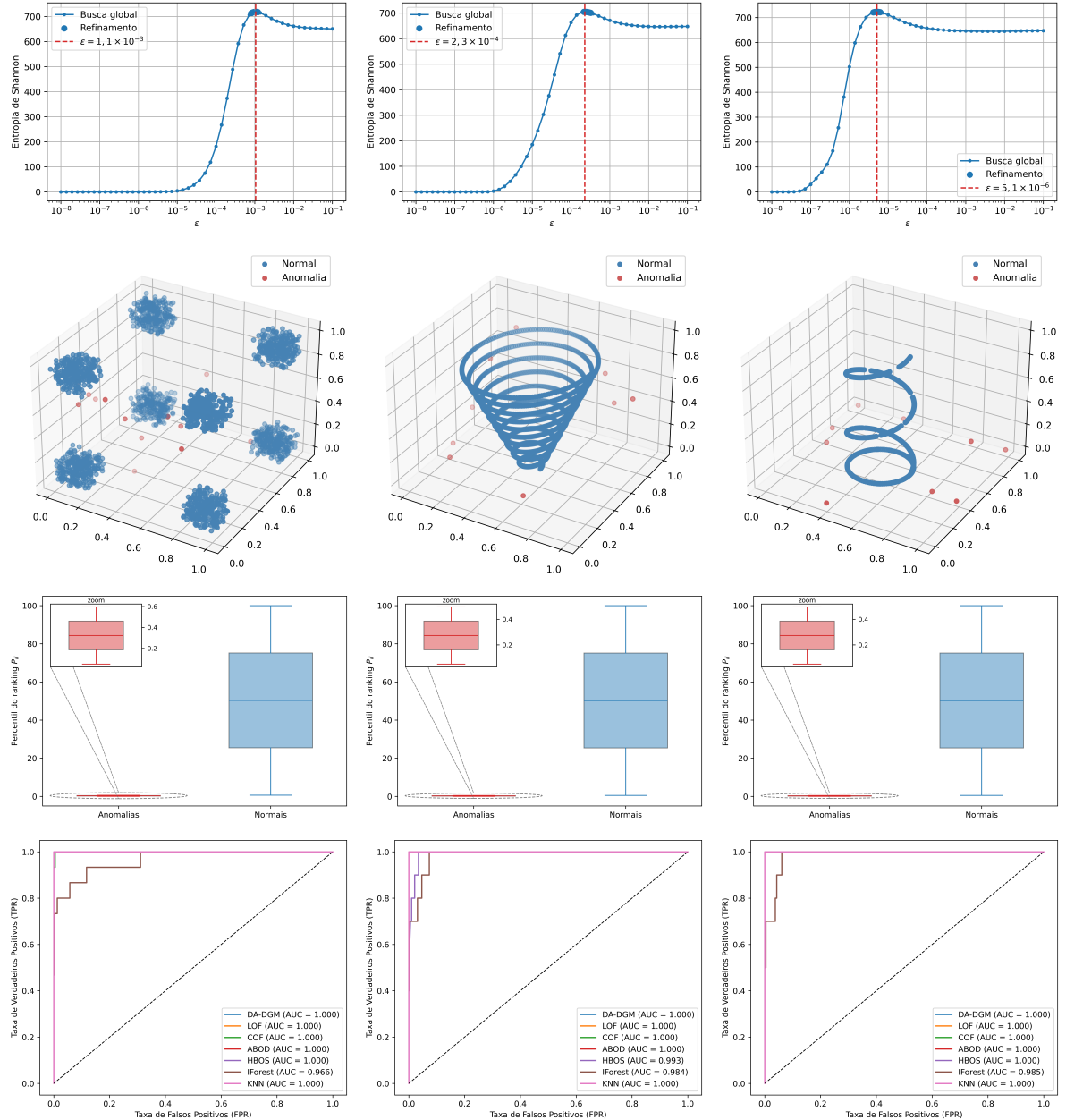
Tabela 4: Resultados consolidados para os conjuntos de dados tridimensionais.

Conjunto de dados	Métrica	DA-DGM	KNN	HBOS	LOF	COF	ABOD	IForest
<i>Cantos 3D</i>	Anomalias	15	15	15	15	14	15	9
	AUC (%)	100,00	100,00	100,00	100,00	99,97	100,00	96,62
<i>Curva Helicoidal</i>	Anomalias	10	10	10	10	10	10	5
	AUC (%)	100,00	100,00	100,00	100,00	100,00	100,00	98,50
<i>Espiral 3D</i>	Anomalias	10	10	5	10	10	10	6
	AUC (%)	100,00	100,00	99,27	100,00	100,00	100,00	98,42

AUC (%) = área sob a curva ROC; valores em negrito destacam o melhor resultado em cada linha.

A Figura 10 consolida os resultados para os três conjuntos de dados tridimensionais: Cantos 3D, Curva Helicoidal e Espiral 3D. Os diagramas de dispersão revelam estruturas de alta complexidade, incluindo múltiplos *clusters* esféricos, uma hélice e uma espiral cônica, com anomalias distribuídas globalmente no espaço. Os *boxplots* do modelo DA-DGM mostram um resultado consistente em todos os três casos: as instâncias normais ocupam toda a faixa de percentis de *ranking* (0-100), enquanto as anomalias são separadas de forma perfeita, com seus *rankings* invariavelmente concentrados abaixo do percentil 1. As curvas ROC indicam um desempenho excepcional de quase todos os algoritmos. Os modelos DA-DGM, KNN, LOF, COF e ABOD atingiram um AUC de 100% em todos os cenários. HBOS e IForest também obtiveram resultados excelentes, com seus piores desempenhos registrando AUCs de 99,27% e 96,62%, respectivamente.

Figura 10: Resultados para os conjuntos de dados tridimensionais *Cantos 3D*, *Espiral 3D* e *Curva Helicoidal*. Linha 1: curvas $\varepsilon \times$ entropia de Shannon utilizadas para selecionar o parâmetro de escala. Linha 2: gráficos de dispersão tridimensionais das instâncias classificadas pelo DA-DGM (azul = normal, vermelho = anômalas). Linha 3: *boxplots* dos percentis do ranqueamento da probabilidade de auto-transição. Linha 4: curvas *ROC* com as áreas sob a curva (AUC) dos algoritmos avaliados, permitindo comparar o desempenho discriminatório em cada cenário.



Fonte: O autor, 2025.

4.4 Dados Reais

Nesta seção avalia-se o desempenho do DA-DGM em conjuntos de dados reais selecionados para cobrir amplo espectro de tamanhos, dimensionalidades e taxas de contaminação, verificando se a robustez observada nos dados sintéticos se mantém em cenários práticos e heterogêneos. Utilizaram-se cinco conjuntos da coleção ODDS (Rayana; Akoglu, 2016): *Ionosphere* (351 instâncias, 33 atributos, 35,90% de anomalias) (Sigillito et al., 1989); *Letter* (1.600 instâncias, 32 atributos, 6,25%) (Slate, 1991; Micenková; McWilliams; Assent, 2014; Rayana; Akoglu, 2016); *MNIST* (7.603 instâncias, 100 atributos, 9,21%) (Lecun et al., 1998; Bandaragoda et al., 2014); *Pima Indians Diabetes* (768 instâncias, 8 atributos, 34,90%) (Kahn, ; Smith et al., 1988) e *Vowels* (1.456 instâncias, 12 atributos, 3,43%) (Kudo; Toyama; Shimbo, 1999; Aggarwal; Sathe, 2015; Sathe; Aggarwal, 2016). A maioria desses conjuntos tem origem no repositório de aprendizado de máquina da UCI (Kelly; Longjohn; Nottingham, 2025), exceto o *MNIST*.

Embora o repositório ODDS original (<https://odds.cs.stonybrook.edu/>), acessado em 3 de setembro de 2024) esteja atualmente indisponível, a autora (Rayana; Akoglu, 2016) mantém uma página pessoal com os *datasets* usados neste trabalho. O endereço é: <https://shebuti.com/outlier-detection-datasets-odds/>, acessado em 22 de agosto de 2025. Em paralelo, a comunidade de pesquisa em detecção de anomalias tem adotado crescentemente as suítes PyOD (Zhao; Nasrullah; Li, 2019) e ADBench (Han et al., 2022), que disponibilizam versões padronizadas e pré-processadas de *datasets* da UCI, OpenML e coleções adicionais voltadas para visão computacional, processamento de texto e séries temporais.

Além das bases apresentadas anteriormente, utilizaram-se cinco conjuntos reais referentes às despesas do Fundo Municipal de Saúde de João Pessoa (FMSJP) de 2016 a 2020, obtidos através do SAGRES (TCE-PB, 2023) e previamente analisados por Souza (2021). A base original, disponibilizada como um único arquivo cobrindo 2016–2020, foi adaptada aos experimentos por meio de desmembramento por exercício, preservando metadados e integridade dos registros. A rotulagem de anomalias tomou por referência as ocorrências validadas por especialistas e consolidadas na Tabela 9 do estudo de Souza (2021) sob o cenário de pior caso, entendido aqui como o critério mais restritivo, no qual uma instância é considerada *outlier* verdadeiro apenas quando há concordância unânime dos três especialistas; nesse critério, a base completa reúne 895 *outliers*. Essas ocorrências foram reconciliadas com o ano de ocorrência e distribuídas entre os exercícios, formando cinco bases anuais com rótulos validados e as seguintes características: FMSJP 2016, 4.208 instâncias, 23 atributos e 146 anomalias (3,47%); FMSJP 2017, 4.372 instâncias, 23 atributos e 159 anomalias (3,64%); FMSJP 2018, 4.435 instâncias, 23 atributos e 201 anomalias (4,53%); FMSJP 2019, 4.131 instâncias, 23 atributos e 185 anomalias (4,48%); e FMSJP 2020, 3.243 instâncias, 23 atributos e 204 anomalias (6,29%). A inclusão desses

dados permite avaliar o desempenho do DA-DGM em informação financeira e administrativa real, verificando a manutenção da robustez observada em cenários sintéticos e repositórios padronizados diante de um contexto heterogêneo com anotações validadas por especialistas.

Neste estudo, empregaram-se as taxas de contaminação conhecidas para facilitar a apresentação dos resultados e validar a correta detecção de anomalias, assegurando que o protocolo experimental permaneça comparável aos relatados na literatura. Os resultados obtidos pelo DA-DGM são apresentados a seguir.

A Tabela 5 consolida, para cada conjunto, o número de anomalias recuperadas corretamente e a AUC obtida pelo DA-DGM e pelos métodos comparativos.

Tabela 5: Resultados consolidados para os conjuntos de dados reais.

Conjunto de dados	Métrica	DA-DGM	KNN	HBOS	LOF	COF	ABOD	IForest
<i>Ionosphere</i>	Anomalias	110	110	61	97	97	104	81
	AUC (%)	94,11	93,16	65,46	89,03	86,13	91,89	85,04
<i>Letter</i>	Anomalias	48	39	7	44	43	39	11
	AUC (%)	92,28	89,50	59,03	88,72	87,34	89,94	62,96
<i>MNIST</i>	Anomalias	288	283	101	169	148	205	232
	AUC (%)	83,23	80,41	61,40	64,98	62,18	71,89	82,41
<i>Pima</i>	Anomalias	150	148	145	118	113	142	128
	AUC (%)	70,97	71,37	69,56	59,78	58,22	67,05	66,13
<i>Vowels</i>	Anomalias	33	27	6	18	31	36	6
	AUC (%)	97,10	97,97	68,07	94,43	94,93	97,05	71,63
FMSJP 2016	Anomalias	67	70	19	20	16	64	77
	AUC (%)	94,06	94,43	54,86	74,95	68,08	92,83	96,04
FMSJP 2017	Anomalias	88	89	53	37	29	84	97
	AUC (%)	94,67	96,04	67,40	75,47	66,79	91,03	97,77
FMSJP 2018	Anomalias	100	104	70	38	48	84	128
	AUC (%)	87,85	87,69	76,21	79,96	70,42	84,84	97,47
FMSJP 2019	Anomalias	101	111	87	18	23	83	132
	AUC (%)	90,87	92,16	79,41	78,69	67,32	86,77	98,29
FMSJP 2020	Anomalias	94	96	108	33	31	83	137
	AUC (%)	87,75	87,74	82,03	68,72	64,96	86,04	96,91

AUC (%) = área sob a curva ROC. Os valores em negrito indicam simultaneamente o melhor desempenho em cada linha e os resultados obtidos pelo DA-DGM.

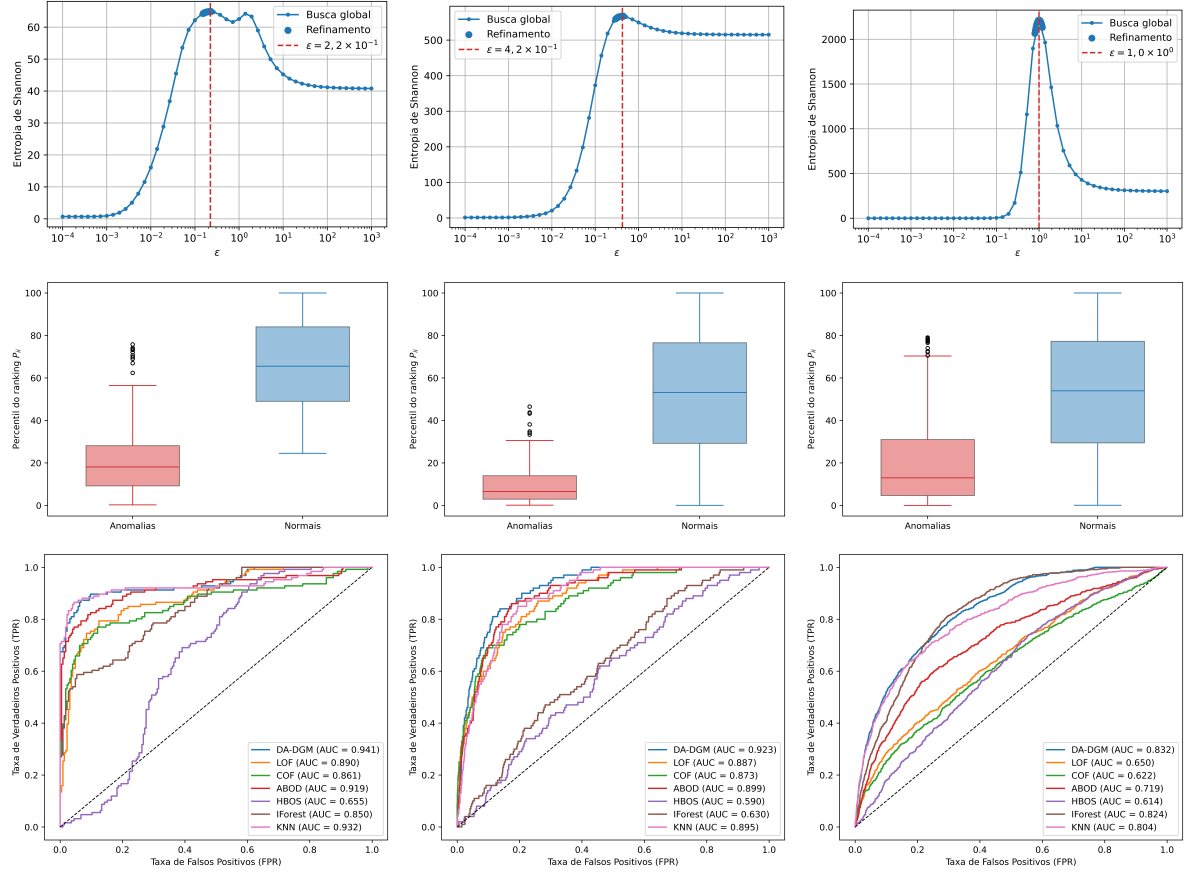
Inicialmente, é preciso registrar que em todos os painéis, cada coluna de figuras apresenta, de cima para baixo: (a) a curva $\varepsilon \times$ entropia de Shannon, cujo máximo define

o parâmetro de escala; (b) o *boxplot* das posições no ranqueamento da probabilidade de auto-transição, mostrando o deslocamento sistemático dos escores anômalos para percentis inferiores; e (c) a curva *ROC* com a área sob a curva (AUC), permitindo comparar a capacidade discriminativa dos diferentes algoritmos.

As Figuras 11 a 14 reúnem os resultados do DA-DGM para todos os conjuntos de dados reais avaliados. As duas primeiras Figuras (11 e 12) abrangem as bases *Ionosphere*, *Letter*, *MNIST*, *Pima* e *Vowels*, além do conjunto *FMSJP 2016*. As áreas sob a curva ROC revelam um cenário desafiador, no qual nenhum algoritmo atingiu desempenho perfeito e há variação considerável entre as bases. Os modelos DA-DGM e KNN destacaram-se com os maiores valores de AUC na maioria dos casos; por exemplo, DA-DGM obteve 94,11% na base *Ionosphere* e KNN alcançou 97,97% na base *Vowels*. Em contrapartida, HBOS e IForest estiveram, em geral, entre os de menor desempenho. A base *Pima* mostrou-se a mais difícil, com AUCs variando entre 58,22% e 71,37%, achado corroborado pelos *boxplots* do DA-DGM, que exibem significativa sobreposição entre as distribuições de escores de instâncias normais e anômalas, indicando separação incompleta no espaço de *ranking* do modelo.

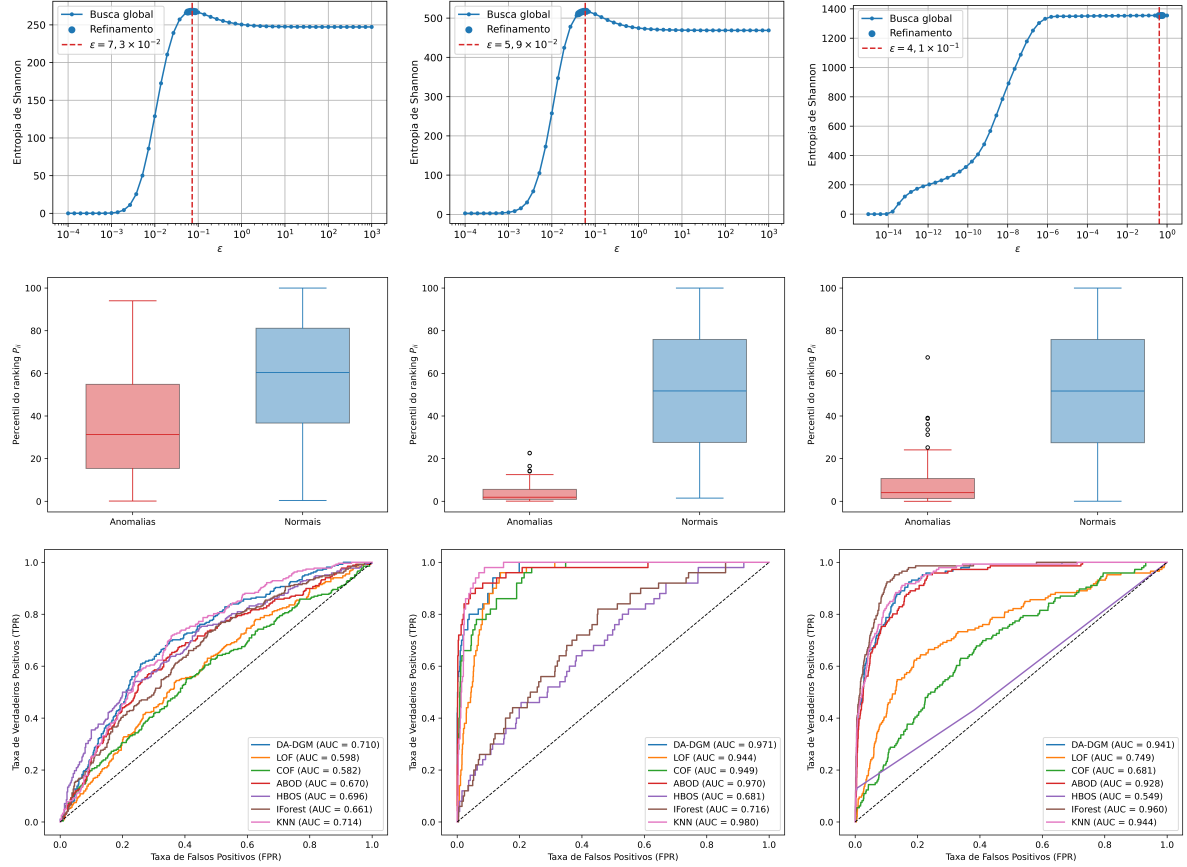
As Figuras seguintes (13 e 14) apresentam os resultados para as bases do *Fundo Municipal de Saúde de João Pessoa* (FMSJP) de 2017 a 2020. Nesse grupo, o algoritmo IForest obteve desempenho consistentemente superior, com AUC sempre acima de 96%. Um segundo bloco, formado por DA-DGM, KNN e ABOD, mostrou desempenho bom e estável, com AUCs tipicamente entre 85% e 96%. Por outro lado, LOF, COF e HBOS registraram os menores resultados, com este último frequentemente exibindo a menor capacidade discriminativa. Assim como nas demais bases, os *boxplots* do DA-DGM indicam considerável sobreposição entre os percentis de ranking de normais e anômalos, reforçando a complexidade inerente à separação dessas classes.

Figura 11: Resultados para os conjuntos de dados reais *Ionosphere*, *Letter* e *MNIST*. Linha 1: curvas $\varepsilon \times$ entropia de Shannon usadas para selecionar o parâmetro de escala. Linha 2: *boxplots* dos percentis do ranqueamento da probabilidade de auto-transição (DA-DGM). Linha 3: curvas *ROC* com as áreas sob a curva (AUC) dos algoritmos avaliados.



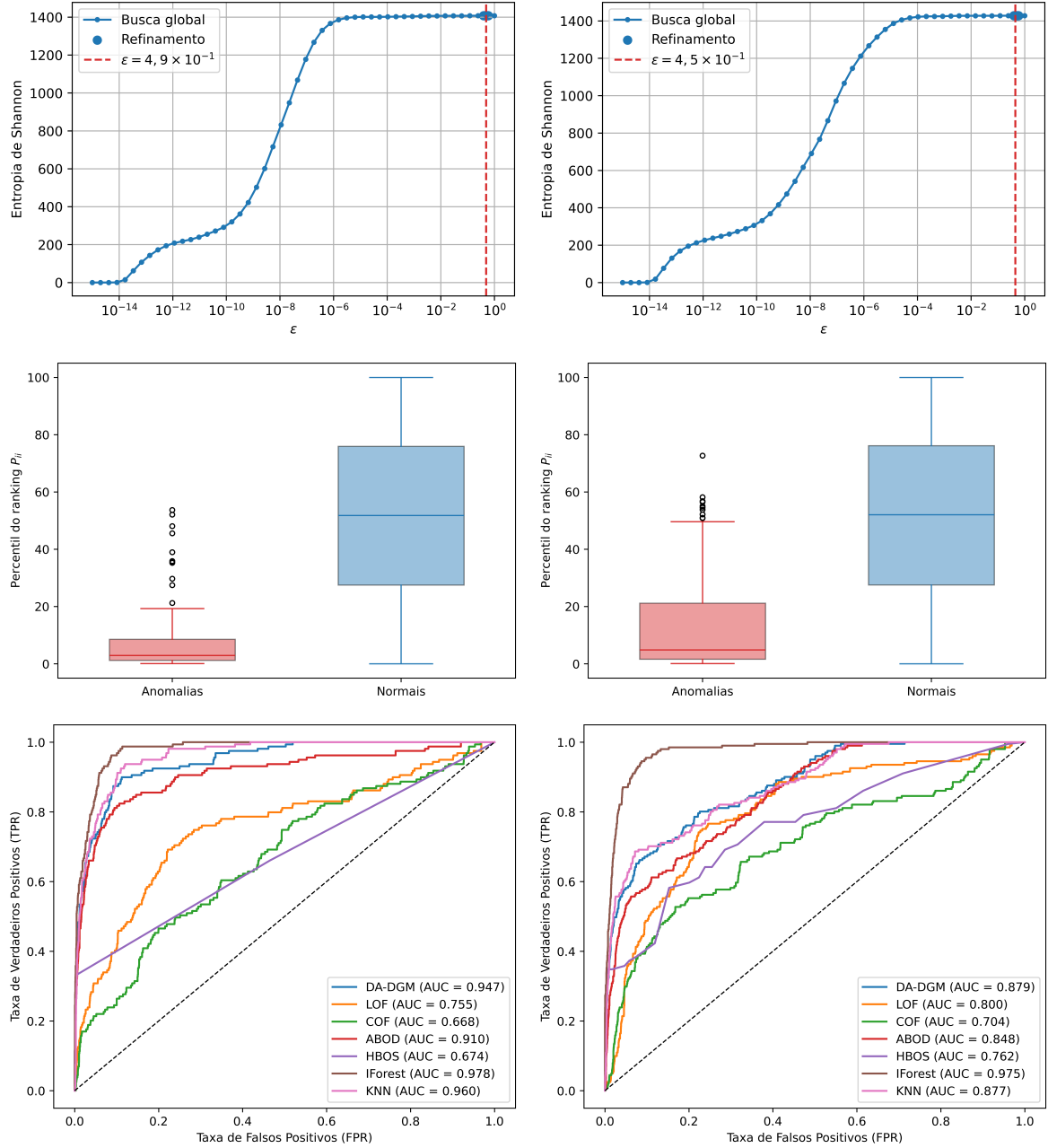
Fonte: O autor, 2025.

Figura 12: Resultados para os conjuntos de dados reais *Pima*, *Vowels* e *FMSJP 2016*. Linha 1: curvas $\varepsilon \times$ entropia de Shannon usadas para selecionar o parâmetro de escala. Linha 2: *boxplots* dos percentis do ranqueamento da probabilidade de auto-transição (DA-DGM). Linha 3: curvas *ROC* com as áreas sob a curva (AUC) dos algoritmos avaliados.



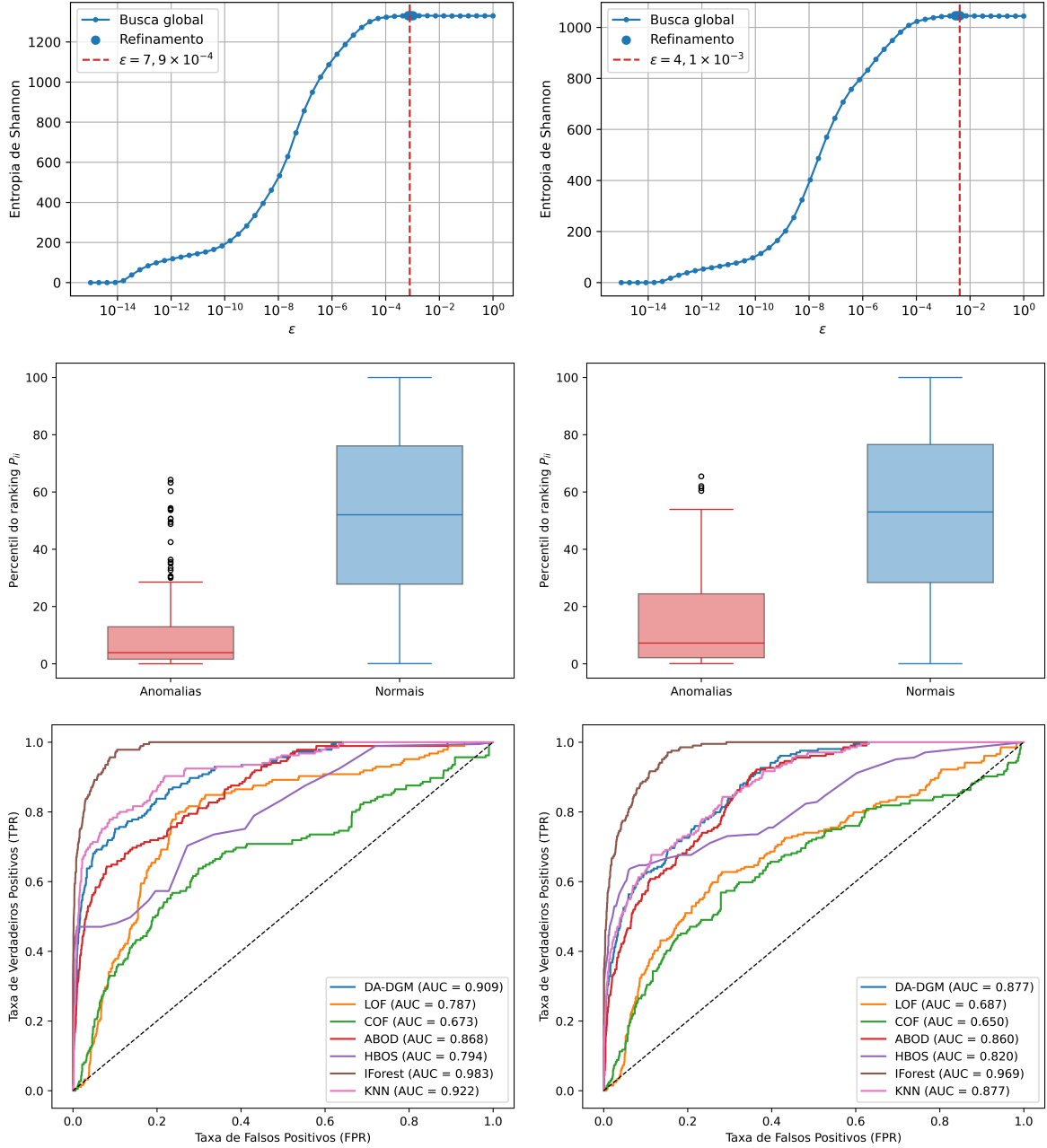
Fonte: O autor, 2025.

Figura 13: Resultados para os conjuntos de dados reais *FMSJP 2017* e *FMSJP 2018*. Linha 1: curvas $\varepsilon \times$ entropia de Shannon usadas para selecionar o parâmetro de escala. Linha 2: *boxplots* dos percentis do ranqueamento da probabilidade de auto-transição (DA-DGM). Linha 3: curvas *ROC* com as áreas sob a curva (AUC) dos algoritmos avaliados.



Fonte: O autor, 2025.

Figura 14: Resultados para os conjuntos de dados reais *FMSJP 2019* e *FMSJP 2020*. Linha 1: curvas $\varepsilon \times$ entropia de Shannon usadas para selecionar o parâmetro de escala. Linha 2: *boxplots* dos percentis do ranqueamento da probabilidade de auto-transição (DA-DGM). Linha 3: curvas *ROC* com as áreas sob a curva (AUC) dos algoritmos avaliados.



Fonte: O autor, 2025.

4.5 Discussão

Esta seção consolida e interpreta os achados experimentais, comparando o DA-DGM com métodos de referência por meio da área sob a curva ROC (AUC-ROC). Primeiro, examinam-se cenários sintéticos uni, bi e tridimensionais, destacando a influência da ge-

ometria na separabilidade. Em seguida, discutem-se os testes em conjuntos reais, enfatizando a contribuição da seleção automática do parâmetro de escala via maximização da entropia de Shannon. Por fim, analisam-se limitações observadas em alta dimensionalidade e em contextos com forte sobreposição entre classes; a Subseção 4.6 delinea frentes para ampliar eficiência, adaptabilidade e interpretabilidade.

Em dados sintéticos, o DA-DGM alcançou $AUC = 100\%$ em todas as distribuições unidimensionais, bidimensionais e tridimensionais avaliadas, indicando separação perfeita do *ranking* derivado de P_{ii} . Nos cenários bidimensionais, o desempenho ideal também foi observado para KNN, LOF e ABOD, ao passo que COF, HBOS e IForest apresentaram quedas sob geometrias mais desafiadoras (como *Cantos*, *Curvas* e *Ondas*). Em 3D, a separação permaneceu essencialmente perfeita para DA-DGM, KNN, LOF e ABOD, com reduções moderadas em IForest (AUC entre $\approx 96,6\%$ e $98,5\%$) e em HBOS (mínimo de $\approx 99,3\%$). Em todos os painéis, as curvas $\varepsilon \times$ entropia exibiram picos nítidos que sustentaram a seleção do parâmetro de escala, e os *boxplots* mostraram anomalias sistematicamente concentradas abaixo do percentil 1 do *ranking*.

Nos conjuntos reais, o comportamento foi heterogêneo, como esperado para bases de maior variabilidade. O DA-DGM liderou em *Ionosphere* ($AUC = \mathbf{94,11\%}$) e *Letter* ($AUC = \mathbf{92,28\%}$), superando todos os comparativos. Em *MNIST*, de maior dimensionalidade, manteve o melhor resultado ($AUC = \mathbf{83,23\%}$) frente a KNN (80,41%) e demais métodos. Em *Vowels*, o DA-DGM obteve $AUC = \mathbf{97,10\%}$, enquanto o melhor geral foi KNN (97,97%). Já em *Pima*, cenário notoriamente desafiador pela sobreposição entre classes, o DA-DGM ($AUC = \mathbf{70,97\%}$) ficou muito próximo de KNN (71,37%) e acima de LOF/COF.

Nos dados financeiros reais do FMSJP (2016–2020), observou-se um padrão distinto: o IForest foi consistentemente superior ($AUC \geq 96\%$ nos quatro últimos anos e 96,91% em 2020), enquanto o DA-DGM manteve desempenho estável e competitivo (AUC entre $\approx 87,75\%$ e $94,67\%$), usualmente no segundo bloco de melhores resultados ao lado de KNN e ABOD. Os *boxplots* do DA-DGM nesses conjuntos evidenciaram maior sobreposição entre normais e anômalos, sinalizando separação incompleta em contextos de alta heterogeneidade.

Os resultados indicam que o DA-DGM: iguala ou supera métodos clássicos em cenários sintéticos e em parte relevante dos conjuntos reais; mostra robustez particular em geometrias não lineares, nas quais a difusão capta conectividade intrínseca; e preserva competitividade em bases reais desafiadoras, ainda que sem dominar universalmente (como ilustrado pelos resultados do IForest no FMSJP). As principais limitações emergem na alta dimensionalidade com forte sobreposição (e.g., *Pima*) e no custo de construção de vizinhanças. Ainda assim, a implementação otimizada evita formar a matriz completa de distâncias $n \times n$, reutiliza KNN ao longo da malha de ε e reduz o custo para $O(c \cdot n \cdot k)$

em tempo e $O(n \cdot k)$ em memória, mitigando o gargalo clássico de $O(n^2)$.

Em síntese, a seleção automática de ε por entropia de Shannon, combinada ao *ranking* por P_{ii} , explica o excelente desempenho em geometrias complexas e a boa competitividade em dados reais. Permanecem oportunidades de ganho em bases com sobreposição acentuada e alta dimensionalidade (e.g., *Pima*), bem como em cenários administrativos/financeiros como FMSJP, em que *ensembles* ou hibridizações com métodos de isolamento podem capturar padrões residuais complementares.

4.5.1 Dados Sintéticos

Esta subseção discute criticamente os achados obtidos nos conjuntos sintéticos uni, bi e tridimensionais, articulando a geometria dos dados com o comportamento das métricas e com os mecanismos internos dos algoritmos avaliados. A ênfase recai sobre a estabilidade do ranqueamento produzido pelo DA-DGM, a sensibilidade dos métodos a diferentes estruturas e a interpretação dos casos em que há saturação de AUC.

Em relação aos resultados dos experimentos com conjuntos sintéticos unidimensionais, os resultados da Tabela 2 e da Figura 7 mostram que o DA-DGM atingiu AUC de 100% nas três distribuições (*Gaussiana*, *Beta* e *Gama*), recuperando integralmente as anomalias rotuladas. O comportamento dos *boxplots* é consistente com esse desempenho, pois as anomalias são mapeadas para percentis extremamente baixos do ranqueamento (concentradas abaixo do percentil 1), enquanto as instâncias normais ocupam toda a faixa de 0 a 100, com mediana próxima de 50%.

Em contraste, LOF e COF mostraram forte queda nas distribuições assimétricas *Beta* e, sobretudo, *Gama*, refletindo a dificuldade de métodos estritamente baseados em densidade local quando a cauda da distribuição normal apresenta gradientes suaves de densidade que reduzem o contraste entre normalidade e anormalidade. Já KNN, HBOS, ABOD e IForest mantiveram AUCs próximas de 100% no conjunto do painel, embora a contagem de anomalias “recuperadas” corretamente varie entre métodos, indicando diferenças de calibração do limiar na conversão do ranqueamento contínuo em decisão binária. Em síntese, em 1D, a separação produzida pelo DA-DGM no espaço de ranqueamento explica diretamente a AUC perfeita e mostra robustez a assimetrias e caudas pesadas.

No que tange aos resultados dos experimentos com conjuntos sintéticos bidimensionais, os resultados das Tabelas 3 e Figuras 8 e 9 indicam um agrupamento claro de desempenho. DA-DGM, KNN, LOF e ABOD atingiram AUC igual a 100% nos conjuntos *Anel*, *Cantos*, *Curvas*, *Espiral*, *Gaussiano* e *Ondas*. IForest e HBOS, por sua vez, tiveram quedas pronunciadas nas geometrias não lineares e em variedades curvas, com AUCs entre 65,42% e 85,34% em *Curvas*, *Espiral* e *Ondas*. Esse padrão é compatível com as hipóteses estruturais de cada técnica. Modelos que particionam o espaço por cor-

tes essencialmente retilíneos tendem a fragmentar de forma inadequada conjuntos sobre variedades curvadas, prejudicando a separação de anomalias que se organizam ao longo de estruturas contínuas. Em contrapartida, abordagens baseadas em vizinhança e grafos, como o DA-DGM, preservam relações locais independentemente da curvatura global e, portanto, mantêm a separação no ranqueamento. As curvas $\varepsilon \times$ entropia de Shannon mostram picos bem definidos em todos os conjuntos, sustentando a seleção de escala e a estabilidade do ranqueamento subsequente. Assim, em 2D, a combinação entre seleção de escala informada por entropia e difusão sobre grafo explica a separação consistente observada nos *boxplots* e a saturação de AUC.

Embora a saturação da AUC em 100% por DA-DGM, KNN, LOF e ABOD nas bases bidimensionais e tridimensionais sugira uma equivalência de capacidade discriminativa, a vantagem do DA-DGM reside primariamente em sua arquitetura geométrica e em sua automação. Diferentemente dos métodos clássicos, cujo desempenho frequentemente depende de ajustes manuais de hiperparâmetros, o DA-DGM alcança esse ranqueamento perfeito por meio da seleção automática do parâmetro de escala (ε) guiada pela Entropia de Shannon. Essa otimização baseada em critérios informacionais elimina a dependência de ajuste manual, garantindo que o método seja adaptável à geometria intrínseca dos dados sem intervenção humana, o que o torna uma solução intrinsecamente mais robusta e estável para a detecção de anomalias em conjuntos com estruturas variadas.

No que diz respeito aos resultados dos experimentos com conjuntos sintéticos tridimensionais, a Tabela 4 e a Figura 10 sugerem um ponto de inflexão favorável à maioria dos métodos. KNN, LOF e ABOD atingiram AUC de 100% em *Cantos 3D*, *Curva Helicoidal* e *Espiral 3D*. HBOS teve desempenhos muito próximos à perfeição, com o pior resultado em 99,27%. IForest permaneceu alto, porém abaixo do teto, entre 96,62% e 98,50%. A melhora substancial dos métodos baseados em densidade, em especial quando comparada à sua queda em 1D, resulta do maior isolamento relativo das anomalias globais no espaço tridimensional, o que aumenta o contraste geométrico entre regiões normais e regiões anômalas. O DA-DGM manteve AUC de 100% em todos os conjuntos e *boxplots* com separação perfeita, reforçando que a difusão em grafos captura com naturalidade estruturas curvas em dimensão moderada e projeta as anomalias para a cauda inferior do ranqueamento.

Diante do exposto, quatro conclusões surgem de forma consistente. Primeiro, a representação por grafo do DA-DGM, combinada à seleção de ε guiada por entropia, gera um ranqueamento com zona de exclusão para anomalias próxima de zero, o que se traduz em AUCs máximas e em decisões mais estáveis entre distribuições e geometrias distintas. Segundo, métodos baseados em densidade são sensíveis à forma da distribuição em baixa dimensão, mas tendem a ganhar contraste quando a dimensionalidade permite isolar melhor as anomalias globais. Terceiro, técnicas que dependem de partições retilíneas po-

dem apresentar limitações em variedades não lineares, como mostram os desempenhos de HBOS e IForest nas bases com curvatura pronunciada. Quarto, a saturação de AUC observada em diversos cenários sugere que, para discriminar finamente entre métodos de alto desempenho, é recomendável complementar a avaliação com métricas sensíveis ao topo do ranqueamento, análise de estabilidade do ranqueamento sob perturbações e análise de calibração do limiar para converter escores em decisões binárias. Em conjunto, os resultados indicam que o DA-DGM oferece uma solução robusta e geometricamente coerente para detecção de anomalias em dados com estruturas variadas, mantendo desempenho competitivo mesmo quando outros métodos já atingem o teto de AUC.

4.5.2 Dados Reais

Esta subseção discute criticamente os resultados obtidos pelo DA-DGM e pelos métodos de referência em bases reais, incluindo conjuntos consagrados na literatura (ODDS) e séries históricas de despesas do Fundo Municipal de Saúde de João Pessoa (FMSJP). A análise considera métricas de desempenho, com ênfase na área sob a curva ROC (AUC-ROC), bem como a distribuição dos escores de anomalia no ranqueamento do DA-DGM e o impacto da definição do parâmetro de escala ε a partir do máximo da entropia de Shannon. O objetivo é contrastar padrões de desempenho, limitações e vantagens de cada abordagem em cenários padronizados e em dados aplicados de elevada complexidade.

Nos conjuntos ODDS, observaram-se ganhos expressivos do DA-DGM em três de cinco bases: *Ionosphere* (AUC = 94,11%), *Letter* (92,28%) e *MNIST* (83,23%). O KNN prevaleceu em *Vowels* (97,97%) e teve o melhor resultado em *Pima* (71,37%), coerente com cenários em que a estrutura local é bem definida e as anomalias manifestam-se como desvios de vizinhança. Em contrapartida, HBOS e IForest mostraram limitações em bases com geometria não linear e distribuição multimodal, como *Letter* (59,03% e 62,96%, respectivamente). A base *Pima* permaneceu desafiadora a todos os métodos (AUCs entre 58,22% e 71,37%), com *boxplots* do DA-DGM mostrando sobreposição substancial entre percentis de instâncias normais e anômalas, o que indica separabilidade intrinsecamente limitada.

No FMSJP, o padrão se inverteu: o IForest superou os demais em todos os cinco anos avaliados, com AUCs superiores a 96% e elevada detecção absoluta, a exemplo de 2020 (137 instâncias). Esse desempenho sugere forte compatibilidade entre a natureza dos dados financeiros-administrativos do FMSJP e a estratégia de particionamento recursivo do IForest. O DA-DGM, ao lado de KNN e ABOD, formou um segundo grupo de bom desempenho (AUCs tipicamente entre 85% e 96%, por exemplo, 94,06% em 2016 e 94,67% em 2017), indicando que a difusão geométrica capta padrões de anomalia quando há continuidade estrutural e relações proximais relevantes entre atributos contábeis e temporais.

O desempenho contrastante em bases reais decorre da compatibilidade entre a geometria dos dados e a estratégia de modelagem de cada algoritmo. Nas bases ODDS, as curvas entropia- ε exibem picos nítidos, o que favorece a seleção de escala e a separação dos escores do DA-DGM. Já no FMSJP, são comuns patamares largos da entropia em função de redundâncias e correlações entre variáveis financeiras; nesses casos, pequenas variações de ε preservam a ordem relativa de P_{ii} (ranqueamento estável), mas reduzem a nitidez dos intervalos interquartílicos nos *boxplots* e aumentam a sobreposição entre normais e anômalas. Em termos práticos, o **DA-DGM produz um ranqueamento útil para priorização, ainda que a separação global não seja perfeita.**

A utilidade aplicada do método no FMSJP é corroborada por exemplos em que o DA-DGM sinalizou empenhos anteriormente validados por Auditores de Controle Externo do TCE-PB durante o estudo de Souza (2021). Tais achados reforçam o papel do DA-DGM como filtro de alto *recall* operacional, especialmente quando combinado a IForest em esquema de *ensembles* ou de interseção/união de listas para triagem.

Por fim, destaca-se que detecção automática de anomalias não implica, por si, irregularidade. O propósito é priorizar instâncias com alta propensão a atipicidade para análise humana aprofundada. Em auditoria, recomenda-se ordenar o *backlog* pelo ranqueamento do DA-DGM (complementado por IForest) e aplicar verificações documentais e contextuais. Ainda que as instâncias ilustradas nas Figuras 15, 16 e 17 tenham sido classificadas como anômalas, nem este estudo nem o de Souza (2021) confirmaram, por si só, a existência de irregularidades; tal juízo exige avaliação técnico-jurídica subsequente.

As figuras a seguir ilustram exemplos de instâncias corretamente classificadas como anômalas pelo DA-DGM, validadas por Auditores de Controle Externo do Tribunal de Contas da Paraíba durante o desenvolvimento da dissertação de Souza (2021). Essas validações corroboram a precisão do método na identificação de anomalias, consolidando sua relevância prática em auditorias de dados financeiros.

Figura 15: Painel de empenhos sinalizados como anômalos pelo DA-DGM (FMSJP, exercícios de 2016 a 2018). Exemplos: Linha 1 (2016): 392733 e 390551; Linha 2 (2017): 393598, 392484, 391311 e 390062; Linha 3 (2018): 392588, 391460 e 390819.

		Início		Municipal		Sobre		Ajuda		Exercício 2016		João Pessoa		Fundo Municipal de Saúde de João Pessoa			
Empenhos																	
Arraste colunas aqui para agrupá-las																	
Classificação institucional		Dados principais				Valores				Natureza da Despesa							
Unidade Gestora	Nº do Empenho	Data	Mês	CPF/CNPJ	Fornecedor	Valor Empenhado	Valor Liquidado	Valor Pago	Elemento								
		dd/mm			laure												
>	Fundo Municipal de Saúde de João Pessoa	0390100	04/01/2016	01-Janeiro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 25.500,00	R\$ 25.500,00	R\$ 25.500,00	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0394519	05/12/2016	12-Dezembro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 30.000,00	R\$ 0,00	R\$ 0,00	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0394054	30/09/2016	09-Setembro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 30.753,00	R\$ 4.572,00	R\$ 4.572,00	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0390098	04/01/2016	01-Janeiro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 39.852,00	R\$ 39.852,00	R\$ 39.852,00	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0394056	30/09/2016	09-Setembro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 58.218,00	R\$ 51.000,00	R\$ 51.000,00	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0392734	12/07/2016	07-Julho	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 87.327,00	R\$ 76.500,00	R\$ 76.500,00	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0390525	02/02/2016	02-Fevereiro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 89.748,80	R\$ 89.748,80	R\$ 89.748,80	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0393119	29/07/2016	07-Julho	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 119.430,00	R\$ 111.510,00	R\$ 111.510,00	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0390526	02/02/2016	02-Fevereiro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 127.500,00	R\$ 127.500,00	R\$ 127.500,00	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0390338	18/01/2016	01-Janeiro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 1.813.170,28	R\$ 1.813.170,28	R\$ 1.813.170,28	92 - Despesas de Exercícios Anteriores							
>	Fundo Municipal de Saúde de João Pessoa	0394520	05/12/2016	12-Dezembro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 2.952.875,00	R\$ 0,00	R\$ 0,00	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0390099	04/01/2016	01-Janeiro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 3.185.129,99	R\$ 3.185.129,99	R\$ 3.185.129,99	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0394055	30/09/2016	09-Setembro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 6.114.926,52	R\$ 3.842.660,54	R\$ 3.842.660,54	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0392733	12/07/2016	07-Julho	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 11.058.783,49	R\$ 10.239.237,31	R\$ 10.239.237,31	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0390551	02/02/2016	02-Fevereiro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 17.486.924,86	R\$ 17.486.924,86	R\$ 17.486.924,86	39 - Outros Serviços de Terceiros - Pes							

		Início		Municipal		Sobre		Ajuda		Exercício 2017		João Pessoa		Fundo Municipal de Saúde de João Pessoa			
Empenhos																	
Arraste colunas aqui para agrupá-las																	
Classificação institucional		Dados principais				Valores				Natureza da Despesa							
Unidade Gestora	Nº do Empenho	Data	Mês	CPF/CNPJ	Fornecedor	Valor Empenhado	Valor Liquidado	Valor Pago	Elemento								
		dd/mm			laureano												
>	Fundo Municipal de Saúde de João Pessoa	0392483	17/07/2017	07-Julho	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 0,00	R\$ 0,00	R\$ 0,00	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0392485	17/07/2017	07-Julho	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 86.827,50	R\$ 51.000,00	R\$ 51.000,00	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0391312	20/04/2017	04-Abril	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 86.827,50	R\$ 73.627,50	R\$ 73.627,50	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0390022	02/01/2017	01-Janeiro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 86.827,50	R\$ 86.827,50	R\$ 86.827,50	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0391309	20/04/2017	04-Abril	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 89.479,50	R\$ 89.479,50	R\$ 89.479,50	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0392482	17/07/2017	07-Julho	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 89.572,50	R\$ 50.526,00	R\$ 50.526,00	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0390021	02/01/2017	01-Janeiro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 89.572,50	R\$ 89.572,50	R\$ 89.572,50	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0393598	10/10/2017	10-Outubro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 6.500.000,00	R\$ 1.715.008,45	R\$ 1.715.008,45	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0392484	17/07/2017	07-Julho	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 11.106.715,00	R\$ 9.247.400,46	R\$ 9.247.400,46	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0391311	20/04/2017	04-Abril	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 11.106.715,00	R\$ 10.670.377,56	R\$ 10.670.377,56	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0390062	02/01/2017	01-Janeiro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 11.106.715,00	R\$ 11.106.715,00	R\$ 11.106.715,00	39 - Outros Serviços de Terceiros - Pes							

		Início		Municipal		Sobre		Ajuda		Exercício 2018		João Pessoa		Fundo Municipal de Saúde de João Pessoa			
Empenhos																	
Arraste colunas aqui para agrupá-las																	
Classificação institucional		Dados principais				Valores				Natureza da Despesa							
Unidade Gestora	Nº do Empenho	Data	Mês	CPF/CNPJ	Fornecedor	Valor Empenhado	Valor Liquidado	Valor Pago	Elemento								
		dd/mm			laureano												
>	Fundo Municipal de Saúde de João Pessoa	0394755	27/12/2018	12-Dezembro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 0,00	R\$ 0,00	R\$ 0,00	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0391162	26/03/2018	03-Março	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 25.500,00	R\$ 25.500,00	R\$ 25.500,00	92 - Despesas de Exercícios Anteriores							
>	Fundo Municipal de Saúde de João Pessoa	0390920	06/03/2018	03-Março	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 34.578,00	R\$ 34.578,00	R\$ 34.578,00	92 - Despesas de Exercícios Anteriores							
>	Fundo Municipal de Saúde de João Pessoa	0394472	07/12/2018	12-Dezembro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 50.000,00	R\$ 0,00	R\$ 0,00	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0390818	27/02/2018	02-Fevereiro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 82.887,90	R\$ 0,00	R\$ 0,00	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0393944	04/10/2018	10-Outubro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 112.000,00	R\$ 0,00	R\$ 0,00	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0390820	27/02/2018	02-Fevereiro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 127.500,00	R\$ 127.500,00	R\$ 25.500,00	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0392721	23/07/2018	07-Julho	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 225.000,00	R\$ 193.950,00	R\$ 0,00	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0394768	28/12/2018	12-Dezembro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 545.989,95	R\$ 545.989,95	R\$ 0,00	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0393491	10/09/2018	09-Setembro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 1.886.135,00	R\$ 1.886.135,00	R\$ 1.886.135,00	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0393866	05/10/2018	10-Outubro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 3.775.000,00	R\$ 732.557,29	R\$ 732.557,29	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0394114	07/11/2018	11-Novembro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 4.000.000,00	R\$ 0,00	R\$ 0,00	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0391460	11/04/2018	04-Abril	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 9.000.000,00	R\$ 9.000.000,00	R\$ 9.000.000,00	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0392588	05/07/2018	07-Julho	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 10.000.000,00	R\$ 10.000.000,00	R\$ 10.000.000,00	39 - Outros Serviços de Terceiros - Pes							
>	Fundo Municipal de Saúde de João Pessoa	0390819	27/02/2018	02-Fevereiro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO E OUTROS	R\$ 11.168.692,77	R\$ 11.168.692,77	R\$ 11.168.692,77	39 - Outros Serviços de Terceiros - Pes							

Fonte: SAGRES, módulo “Empenhos” (TCE-PB, 2023).

Figura 16: Painel de empenhos sinalizados como anômalos pelo DA-DGM (FMSJP, exercícios de 2019 e 2020). Exemplos: Linha 1 (2019): 393919, 392459, 391208, 390061; Linha 2 (2020): 393391, 393000, 392169, 391682, 390061.

 Início Municipal Sobre Ajuda Exercício 2019 João Pessoa Fundo Municipal de Saúde de João Pessoa										
Empenhos Detalhes de empenho										
Arraste colunas aqui para agrupá-las										
Classificação institucional	Dados principais					Valores			Natureza da Despesa	
Unidade Gestora	Nº do Empenho	Data	Mês	CPF/CNPJ	Fornecedor	Valor Empenhado	Valor Liquidado	Valor Pago	Elemento	
> Fundo Municipal de Saúde de João Pessoa	0390062	22/01/2019	01-Janeiro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 50.000,00	R\$ 50.000,00	R\$ 50.000,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	
> Fundo Municipal de Saúde de João Pessoa	0393924	22/11/2019	11-Novembro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 61.250,00	R\$ 0,00	R\$ 0,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	
> Fundo Municipal de Saúde de João Pessoa	0390060	22/01/2019	01-Janeiro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 82.887,90	R\$ 0,00	R\$ 0,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	
> Fundo Municipal de Saúde de João Pessoa	0393927	22/11/2019	11-Novembro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 100.000,00	R\$ 25.500,00	R\$ 0,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	
> Fundo Municipal de Saúde de João Pessoa	0392072	06/06/2019	06-Junho	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 270.000,00	R\$ 221.650,00	R\$ 126.850,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	
> Fundo Municipal de Saúde de João Pessoa	393919	21/11/2019	11-Novembro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 450.000,00	R\$ 450.000,00	R\$ 450.000,00	43 - Subvenções Sociais	
> Fundo Municipal de Saúde de João Pessoa	0394067	03/12/2019	12-Dezembro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 1.300.000,00	R\$ 1.300.000,00	R\$ 0,00	43 - Subvenções Sociais	
> Fundo Municipal de Saúde de João Pessoa	0394364	26/12/2019	12-Dezembro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 1.900.000,00	R\$ 1.900.000,00	R\$ 1.900.000,00	43 - Subvenções Sociais	
> Fundo Municipal de Saúde de João Pessoa	0393489	07/10/2019	10-Outubro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 3.578.000,00	R\$ 1.479.891,86	R\$ 171.073,80	39 - Outros Serviços de Terceiros - Pessoa Ju...	
> Fundo Municipal de Saúde de João Pessoa	392459	03/07/2019	07-Julho	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 9.500.000,00	R\$ 6.715.135,33	R\$ 6.715.135,33	39 - Outros Serviços de Terceiros - Pessoa Ju...	
> Fundo Municipal de Saúde de João Pessoa	391208	03/04/2019	04-Abril	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 10.521.585,69	R\$ 10.521.585,69	R\$ 10.521.585,69	39 - Outros Serviços de Terceiros - Pessoa Ju...	
> Fundo Municipal de Saúde de João Pessoa	390061	22/01/2019	01-Janeiro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 10.905.512,83	R\$ 10.905.512,83	R\$ 10.905.512,83	39 - Outros Serviços de Terceiros - Pessoa Ju...	

 Início Municipal Sobre Ajuda Exercício 2020 João Pessoa Fundo Municipal de Saúde de João Pessoa										
Empenhos Detalhes de empenho										
Arraste colunas aqui para agrupá-las										
Classificação institucional	Dados principais					Valores			Natureza da Despesa	
Unidade Gestora	Nº do Empenho	Data	Mês	CPF/CNPJ	Fornecedor	Valor Empenhado	Valor Liquidado	Valor Pago	Elemento	
> Fundo Municipal de Saúde de João Pessoa	0391303	15/04/2020	04-Abril	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 60.000,00	R\$ 60.000,00	R\$ 60.000,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	
> Fundo Municipal de Saúde de João Pessoa	0390076	24/01/2020	01-Janeiro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 60.000,00	R\$ 60.000,00	R\$ 60.000,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	
> Fundo Municipal de Saúde de João Pessoa	0390060	22/01/2020	01-Janeiro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 75.000,00	R\$ 0,00	R\$ 0,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	
> Fundo Municipal de Saúde de João Pessoa	0391949	17/06/2020	06-Junho	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 100.000,00	R\$ 100.000,00	R\$ 100.000,00	43 - Subvenções Sociais	
> Fundo Municipal de Saúde de João Pessoa	0393805	15/12/2020	12-Dezembro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 148.322,00	R\$ 148.322,00	R\$ 148.322,00	43 - Subvenções Sociais	
> Fundo Municipal de Saúde de João Pessoa	0393001	09/10/2020	10-Outubro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 150.000,00	R\$ 103.350,00	R\$ 103.350,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	
> Fundo Municipal de Saúde de João Pessoa	0391921	12/06/2020	06-Junho	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 189.381,76	R\$ 189.381,76	R\$ 189.381,76	43 - Subvenções Sociais	
> Fundo Municipal de Saúde de João Pessoa	0391860	02/06/2020	06-Junho	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 500.000,00	R\$ 500.000,00	R\$ 500.000,00	43 - Subvenções Sociais	
> Fundo Municipal de Saúde de João Pessoa	0390747	16/03/2020	03-Março	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 1.000.000,00	R\$ 1.000.000,00	R\$ 1.000.000,00	43 - Subvenções Sociais	
> Fundo Municipal de Saúde de João Pessoa	0392158	21/10/2020	10-Outubro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 1.425.454,00	R\$ 1.425.454,00	R\$ 1.425.454,00	43 - Subvenções Sociais	
> Fundo Municipal de Saúde de João Pessoa	0394012	28/12/2020	12-Dezembro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 1.950.000,00	R\$ 1.950.000,00	R\$ 1.950.000,00	43 - Subvenções Sociais	
> Fundo Municipal de Saúde de João Pessoa	0391922	12/06/2020	06-Junho	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 2.591.254,04	R\$ 2.591.254,04	R\$ 2.591.254,04	43 - Subvenções Sociais	
> Fundo Municipal de Saúde de João Pessoa	0393750	11/12/2020	12-Dezembro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 2.647.809,00	R\$ 2.647.809,00	R\$ 2.647.809,00	43 - Subvenções Sociais	
> Fundo Municipal de Saúde de João Pessoa	0391302	15/04/2020	04-Abril	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 3.500.000,00	R\$ 3.500.000,00	R\$ 3.183.880,30	39 - Outros Serviços de Terceiros - Pessoa Ju...	
> Fundo Municipal de Saúde de João Pessoa	0393391	04/11/2020	11-Novembro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 5.000.000,00	R\$ 0,00	R\$ 0,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	
> Fundo Municipal de Saúde de João Pessoa	0393000	09/10/2020	10-Outubro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 6.000.000,00	R\$ 2.781.179,02	R\$ 2.781.179,02	39 - Outros Serviços de Terceiros - Pessoa Ju...	
> Fundo Municipal de Saúde de João Pessoa	0392169	08/07/2020	07-Julho	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 7.000.000,00	R\$ 7.000.000,00	R\$ 7.000.000,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	
> Fundo Municipal de Saúde de João Pessoa	0391682	18/05/2020	05-Maio	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 7.445.795,18	R\$ 7.445.795,18	R\$ 7.445.795,18	39 - Outros Serviços de Terceiros - Pessoa Ju...	
> Fundo Municipal de Saúde de João Pessoa	0390061	22/01/2020	01-Janeiro	09.112.236/0001-94	HOSPITAL NAPOLEÃO LAUREANO	R\$ 10.000.000,00	R\$ 10.000.000,00	R\$ 10.000.000,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	

Fonte: SAGRES, módulo “Empenhos” (TCE-PB, 2023).

Figura 17: Painel de empenhos sinalizados como anômalos pelo DA-DGM (FMSJP, exercícios de 2017 a 2020). Exemplos: Linha 1 (2017): 393569, 392458, 391189, 390001; Linha 2 (2018): 393834, 392549, 391457, 390821; Linha 3 (2019): 393478, 392469, 391210, 390429; Linha 4 (2020): 392704, 392161, 391919, 391679, 390055.

Início Municipal Sobre Ajuda Exercício 2017 João Pessoa Fundo Municipal de Saúde de João Pessoa										
Empenhos Amarrar colunas aqui para agrupá-las Classificação institucional Dados principais Valores Natureza da Despesa Dados Gerais										
Unidade Gestora	Nº do Empenho	Data	Mês	CPF/CNPJ	Fornecedor	Valor Empenhado	Valor Liquidado	Valor Pago	Elemento	Nº Licitação
Fundo Municipal de Saúde de João Pessoa	0394291	06/12/2017	12-Dezembro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 150.000,00	R\$ 0,00	R\$ 0,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0393569	09/10/2017	10-Outubro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 1.617.145,50	R\$ 1.168.114,10	R\$ 1.168.114,10	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0392458	14/07/2017	07-Julho	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 1.617.145,50	R\$ 1.617.145,50	R\$ 1.617.145,50	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0391189	12/04/2017	04-Abril	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 1.617.145,50	R\$ 1.617.145,50	R\$ 1.617.145,50	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0390001	02/01/2017	01-Janeiro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 1.617.145,50	R\$ 1.617.145,50	R\$ 1.617.145,50	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0390644	09/02/2018	02-Fevereiro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 1.114,46	R\$ 1.114,46	R\$ 1.114,46	92 - Despesas de Exercícios Anteriores	000000000
Fundo Municipal de Saúde de João Pessoa	0391168	23/08/2018	08-Agosto	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 12.000,00	R\$ 0,00	R\$ 0,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0394496	11/12/2018	12-Dezembro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 100.000,00	R\$ 0,00	R\$ 0,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0394754	27/12/2018	12-Dezembro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 199.899,55	R\$ 199.899,55	R\$ 0,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0391457	11/04/2018	04-Abril	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 1.245.000,00	R\$ 1.245.000,00	R\$ 1.245.000,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0393834	04/10/2018	10-Outubro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 1.455.800,00	R\$ 894.159,65	R\$ 894.159,65	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0390821	27/02/2018	02-Fevereiro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 1.617.145,47	R\$ 1.617.145,47	R\$ 1.617.145,47	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0392549	04/07/2018	07-Julho	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 1.617.145,53	R\$ 1.617.145,53	R\$ 1.617.145,53	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0392586	18/07/2019	07-Julho	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 0,00	R\$ 0,00	R\$ 0,00	43 - Subvenções Sociais	000000000
Fundo Municipal de Saúde de João Pessoa	0390084	22/01/2019	01-Janeiro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 0,00	R\$ 0,00	R\$ 0,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0390083	22/01/2019	01-Janeiro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 0,00	R\$ 0,00	R\$ 0,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0391109	28/03/2019	03-Março	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 3.400,57	R\$ 3.400,57	R\$ 3.400,57	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0390428	31/01/2019	01-Janeiro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 40.890,54	R\$ 40.890,54	R\$ 40.890,54	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0391209	03/04/2019	04-Abril	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 47.114,40	R\$ 0,00	R\$ 0,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0394385	27/12/2019	12-Dezembro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 100.000,00	R\$ 0,00	R\$ 0,00	43 - Subvenções Sociais	000000000
Fundo Municipal de Saúde de João Pessoa	0393930	25/11/2019	11-Novembro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 200.000,00	R\$ 200.000,00	R\$ 200.000,00	43 - Subvenções Sociais	000000000
Fundo Municipal de Saúde de João Pessoa	0394386	27/12/2019	12-Dezembro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 290.000,00	R\$ 0,00	R\$ 0,00	43 - Subvenções Sociais	000000000
Fundo Municipal de Saúde de João Pessoa	0390460	08/02/2019	02-Fevereiro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 297.322,92	R\$ 297.322,92	R\$ 297.322,92	92 - Despesas de Exercícios Anteriores	000000000
Fundo Municipal de Saúde de João Pessoa	0394363	26/12/2019	12-Dezembro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 1.036.335,15	R\$ 1.036.335,15	R\$ 0,00	42 - Auxílios	000000000
Fundo Municipal de Saúde de João Pessoa	0392469	03/07/2019	07-Julho	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 1.932.000,00	R\$ 1.932.000,00	R\$ 1.932.000,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0390429	31/01/2019	01-Janeiro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 2.068.689,69	R\$ 2.068.689,69	R\$ 2.068.689,69	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0391210	03/03/2019	04-Abril	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 2.268.671,55	R\$ 2.268.671,55	R\$ 2.268.671,55	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0393478	07/10/2019	10-Outubro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 2.350.000,00	R\$ 2.084.154,41	R\$ 2.084.154,41	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0391678	18/05/2020	05-Maio	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 45.000,00	R\$ 15.704,80	R\$ 15.704,80	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0391304	15/04/2020	04-Abril	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 45.000,00	R\$ 27.400,82	R\$ 27.400,82	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0390054	22/01/2020	01-Janeiro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 45.000,00	R\$ 45.000,00	R\$ 45.000,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0392953	02/10/2020	10-Outubro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 100.400,00	R\$ 100.400,00	R\$ 100.400,00	43 - Subvenções Sociais	000000000
Fundo Municipal de Saúde de João Pessoa	0391926	12/06/2020	06-Junho	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 125.846,00	R\$ 125.846,00	R\$ 125.846,00	43 - Subvenções Sociais	000000000
Fundo Municipal de Saúde de João Pessoa	0391928	12/06/2020	06-Junho	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 148.000,00	R\$ 148.000,00	R\$ 148.000,00	43 - Subvenções Sociais	000000000
Fundo Municipal de Saúde de João Pessoa	0391920	12/06/2020	06-Junho	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 206.257,37	R\$ 206.257,37	R\$ 206.257,37	43 - Subvenções Sociais	000000000
Fundo Municipal de Saúde de João Pessoa	0392954	02/10/2020	10-Outubro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 300.000,00	R\$ 300.000,00	R\$ 300.000,00	43 - Subvenções Sociais	000000000
Fundo Municipal de Saúde de João Pessoa	0391929	12/06/2020	06-Junho	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 400.000,00	R\$ 400.000,00	R\$ 400.000,00	43 - Subvenções Sociais	000000000
Fundo Municipal de Saúde de João Pessoa	0391927	12/06/2020	06-Junho	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 480.000,00	R\$ 480.000,00	R\$ 480.000,00	43 - Subvenções Sociais	000000000
Fundo Municipal de Saúde de João Pessoa	0393387	04/11/2020	11-Novembro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 750.000,00	R\$ 667.924,58	R\$ 667.924,58	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0391305	15/04/2020	04-Abril	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 750.000,00	R\$ 750.000,00	R\$ 750.000,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0392704	08/09/2020	09-Setembro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 1.400.000,00	R\$ 1.400.000,00	R\$ 1.400.000,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0392161	08/07/2020	07-Julho	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 1.400.000,00	R\$ 1.400.000,00	R\$ 1.400.000,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0391679	18/05/2020	05-Maio	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 1.500.000,00	R\$ 1.466.921,84	R\$ 1.466.921,84	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000
Fundo Municipal de Saúde de João Pessoa	0391919	12/06/2020	06-Junho	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 1.977.367,99	R\$ 1.977.367,99	R\$ 1.977.367,99	43 - Subvenções Sociais	000000000
Fundo Municipal de Saúde de João Pessoa	0390055	22/01/2020	01-Janeiro	08.667.206/0001-81	HOSPITAL PADRE ZÉ	R\$ 2.200.000,00	R\$ 2.200.000,00	R\$ 2.200.000,00	39 - Outros Serviços de Terceiros - Pessoa Ju...	000000000

Fonte: SAGRES, módulo “Empenhos”(TCE-PB, 2023).

4.6 Limitações e Trabalhos Futuros

Esta seção resume as limitações do DA-DGM, com quedas em alta dimensionalidade, sobreposição entre classes, patamares na curva entropia- ε sobretudo no FMSJP e custo das vizinhanças KNN. Como rumos, sugerem-se novos núcleos e *ensembles*, redução de dimensionalidade com esparsificação e vizinhança adaptativa, processamento em fluxo, além de paralelismo, GPU (*Graphics Processing Unit*) e execução distribuída, acompanhados de avanços em interpretabilidade e teoria envolvendo a relação da probabilidade de auto-transição com centralidades, a consistência do estimador de ε e a estabilidade do ranqueamento.

Os resultados indicam que o DA-DGM apresenta desempenho consistente em dados sintéticos e competitivo em diversos cenários reais. No entanto, algumas limitações foram mostradas. Observou-se queda de desempenho em contextos de alta dimensionalidade e forte sobreposição entre instâncias normais e anômalas, como nas bases *MNIST* e *Pima*, em que a separação no espaço de ranqueamento não foi completa. Nessas situações, o critério atual de seleção do parâmetro de escala via maximização da entropia de Shannon pode não capturar adequadamente estruturas relevantes para discriminação, especialmente quando a densidade local das instâncias é relativamente homogênea.

A observação dos patamares largos na curva entropia- ε (especialmente no FMSJP) sugere que a redundância e correlação entre atributos financeiros impactam a densidade de dados no *manifold*, limitando a capacidade do critério de maximização de ε em gerar contraste discriminativo sobre um amplo intervalo de escala. Embora o DA-DGM em sua formulação atual utilize exclusivamente o núcleo gaussiano (*heat kernel*), a investigação de diferentes famílias de difusão (*kernels*) não foi realizada. Contudo, essa é uma via reconhecida como promissora para trabalhos futuros, visando a criação de *ensembles* de difusão que combinem múltiplos núcleos para maior adaptabilidade. Adicionalmente, a análise formal das condições que dão origem a esses patamares e de sua relação com a densidade local está entre as recomendações de pesquisa teórica futura.

No aspecto computacional, o custo de tempo e memória associado à construção das vizinhanças por KNN permanece um fator limitante. A versão otimizada do DA-DGM evita formar a matriz $n \times n$ e reutiliza distâncias KNN para todos os valores de ε , reduzindo a avaliação do parâmetro de escala, sobre uma vizinhança fixa, para $O(c \cdot n \cdot k)$ em tempo e $O(n \cdot k)$ em memória. Esse ganho, porém, depende do número de iterações do laço adaptativo de k . A obtenção inicial das vizinhanças com *backend* KD-Tree tem custo médio $O(dn \log n + dnk)$, com possível degradação para $O(n^2d)$ em alta dimensão. No caso unidimensional, esta implementação força $k = n$, o que implica custos $O(n^2)$ em tempo e memória. Durante a busca global ainda ocorre um pico transitório de memória $O(mn)$ devido ao armazenamento de vetores de P_{ii} por ε . Uma estratégia de processamento em

fluxo, com manutenção apenas do melhor valor corrente, reduz esse pico para $O(n)$ sem alterar o resultado do critério de máximo.

Direções futuras de pesquisa incluem investigar técnicas de esparsificação adaptativa da vizinhança, explorando índices espaciais; integrar métodos não lineares de redução de dimensionalidade para mitigar ruído e redundância, melhorando a separação em espaços de alta dimensão; estender o DA-DGM para ambientes de *streaming*, com atualização incremental de ε e das vizinhanças, viabilizando detecção em tempo real; desenvolver esquemas de meta-aprendizado para seleção automática do critério de informação mais adequado a cada contexto; e criar *ensembles* de difusão que combinem múltiplos núcleos e incorporem módulos de interpretabilidade, como escores de importância em nível de vértice e visualizações de grafos, aumentando a transparência e a confiabilidade em aplicações críticas.

No plano computacional, propõe-se explorar paralelismo e distribuição para acelerar etapas dominantes: paralelização por *multithreading* das consultas de vizinhança e das avaliações de ε , dado seu baixo acoplamento; uso de GPU para acelerar o cálculo de distâncias e operações matriciais associadas às transições de Markov; e execução distribuída por particionamento de dados e/ou *sharding* (ou fragmentação) do grafo, com estratégias de *approximate* KNN para reduzir latência em coleções muito grandes, preservando acurácia prática.

A comparação do ranqueamento de anomalias por auto-probabilidade P_{ii} com as medidas clássicas de centralidade em grafos, embora seja um exercício crucial para a interpretabilidade, não foi objeto da avaliação experimental. Contudo, o DA-DGM estabelece uma relação teórica inversa com uma forma de centralidade de grau ponderado, visto que P_{ii} é inversamente proporcional ao fator de normalização $d(x_i)$, que representa a soma dos pesos de afinidade local. Para fortalecer a interpretabilidade, recomenda-se, como trabalho futuro, realizar uma análise comparativa formal entre o ranqueamento obtido por P_{ii} e medidas de importância topológica, como as centralidades de proximidade ou grau. Esse exercício, alinhado à agenda de estudo de vínculos com a teoria espectral de grafos e módulos de interpretabilidade, permitiria demonstrar como o isolamento probabilístico capturado pela difusão se correlaciona, ou se distingue, da importância estrutural dos vértices nas redes complexas.

No plano teórico, recomenda-se uma análise formal das condições que dão origem a patamares na curva entropia- ε e de sua relação com separabilidade, ruído e densidade local. Essa agenda inclui: caracterização da consistência do estimador de ε (sob regimes assintóticos e de alta dimensão), estudo da estabilidade do ranqueamento induzido por P_{ii} frente a perturbações do grafo e do ruído nas distâncias, e vínculos com resultados de teoria espectral de grafos e perturbação de autovalores. Essas garantias teóricas tendem a subsidiar a definição de ε de forma mais robusta, bem como a formulação de critérios

híbridos capazes de manter desempenho confiável mesmo em contextos adversos.

5 CONCLUSÃO

Este trabalho apresentou o DA-DGM, método não supervisionado de detecção de anomalias que estende a Difusão Geométrica Markoviana, incorporando a seleção automática do parâmetro de escala ε por maximização da entropia de Shannon. O protocolo experimental abrangeu dados sintéticos com geometrias variadas, bases reais da coleção ODDS e cinco séries históricas de despesas do FMSJP, permitindo avaliar robustez, generalidade e aplicabilidade prática. Assim, foram alcançados os Objetivos 1–5: (1) fundamentação teórica; (2) projeto e implementação do algoritmo; (3) automação da escolha do parâmetro de escala; (4) avaliação comparativa; e (5) garantia de interpretabilidade, conforme detalhado nos Capítulos 2, 3 e 4.

O método herda duas características centrais do processo difusivo: capta estruturas não lineares por meio de um grafo de proximidade e produz escores de anomalia interpretáveis a partir das auto-transições. A escolha do parâmetro de escala guiada por informação reduziu a dependência de ajuste manual e estabilizou o ranqueamento das instâncias, favorecendo o uso operacional em triagem.

A hipótese principal alternativa (H_1) postula que o DA-DGM alcança AUC-ROC superior à de detectores clássicos (KNN, LOF, IForest, entre outros) em *benchmarks* sintéticos e reais, enquanto a hipótese nula (H_0) sustenta que o desempenho do DA-DGM não é significativamente superior ao dos concorrentes. Os resultados experimentais fornecem forte evidência para a rejeição de H_0 em cenários controlados com geometria complexa e sustentam a aceitação parcial de H_1 em contextos específicos. O DA-DGM alcançou AUC = 100% em todos os conjuntos sintéticos unidimensionais, bidimensionais e tridimensionais, superando ou igualando os demais algoritmos, especialmente em geometrias não lineares e variedades curvas em que IForest e HBOS apresentaram quedas pronunciadas. Esse desempenho é atribuído à integração entre difusão geométrica e ajuste automático do parâmetro de escala por entropia de Shannon, que induz uma zona de exclusão próxima de zero para anomalias. Em bases reais ODDS, o DA-DGM foi robusto e competitivo, obtendo as melhores AUCs em três de cinco conjuntos (*Ionosphere* 94,11%, *Letter* 92,28% e *MNIST* 83,23%), o que evidencia capacidade de preservar separabilidade em diferentes contextos.

Em dados financeiros-administrativos do FMSJP, a análise de H_1 exige leitura mais cuidadosa: o IForest foi consistentemente superior (AUC $\geq 96\%$ nos cinco anos), o que sugere alta compatibilidade de sua estratégia de particionamento com a natureza desses dados. Nesses cenários, o DA-DGM compôs, com KNN e ABOD, um segundo grupo de bom desempenho (AUCs tipicamente entre 84% e 96%), sem superioridade estatística clara sobre todos os clássicos. As limitações do DA-DGM em bases que apresentam maior mistura entre classes (por exemplo, *Pima* e *MNIST*) manifestaram-se em sobreposição

substancial dos *boxplots* de ranqueamento. A queda de desempenho em alta dimensionalidade e a presença de patamares amplos na curva de entropia em função de ε (especialmente no FMSJP) indicam que, nesses casos, o critério de seleção pode não capturar estruturas ideais para discriminação completa. Em síntese, os resultados apoiam H_1 no sentido de que o DA-DGM constitui alternativa robusta e geometricamente informada que supera a maioria dos clássicos em diversas condições, ainda que detectores como o IForest possam prevalecer em domínios específicos. Esse é um contexto em que o DA-DGM é particularmente útil em arranjos de *ensemble* e triagem.

Do ponto de vista computacional, a reutilização de vizinhanças permitiu avaliar c valores do parâmetro de escala ε com custo linear em nk por tentativa, em que n é o número de instâncias, k é o número de vizinhos por difusão, d é a dimensionalidade dos dados, m é a cardinalidade da malha global de ε usada na busca inicial, r é a cardinalidade do refinamento local e $c = m + r$ é o total de valores de ε avaliados por tentativa. A análise de complexidade (Seção 3.5) indica crescimento médio subquadrático para $d \geq 2$, com tempo $\mathcal{O}(n \log n + nk(d + c))$ e pico de memória $\Theta(nk + mn)$, em que o termo mn decorre do armazenamento transitório, durante a busca global, de vetores P_{ii} para m valores de ε ao longo de n instâncias. Em $d = 1$, quando se força $k = n$, os custos tornam-se quadráticos. Tais propriedades sustentam a viabilidade do método em conjuntos de tamanho moderado a grande, desde que k , c e o número de iterações do laço adaptativo permaneçam em faixas moderadas.

As limitações identificadas concentram-se no custo de construção de vizinhanças e nos cenários de alta dimensionalidade com forte sobreposição entre normais e anômalas. Nesses contextos, patamares na curva entropia- ε preservam o ranqueamento, porém reduzem a nitidez da separação global, o que ajuda a explicar a vantagem do IForest no FMSJP. Ainda assim, o conjunto de evidências empíricas e a análise de complexidade indicam que o DA-DGM oferece boa relação custo-benefício para dados com estruturas complexas e potencial de impacto prático em auditoria pública e outros domínios intensivos em dados.

REFERÊNCIAS

- AGGARWAL, C. C.; SATHE, S. Theoretical Foundations and Algorithms for Outlier Ensembles. *SIGKDD Explor. Newsl.*, v. 17, n. 1, p. 24–47, set. 2015. ISSN 1931-0145.
- ALTMAN, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, Taylor & Francis, v. 46, n. 3, p. 175–185, 1992.
- AMIGÓ, J. M.; KELLER, K.; UNAKAFOVA, V. A. On entropy, entropy-like quantities, and applications. *Discrete and Continuous Dynamical Systems - B*, American Institute of Mathematical Sciences (AIMS), v. 20, n. 10, p. 3301–3343, 2015. ISSN 1531-3492. Disponível em: <http://dx.doi.org/10.3934/dcdsb.2015.20.3301>.
- ARRIETA, A. B. et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, Elsevier, v. 58, p. 82–115, 2020.
- ARYAL, S.; SANTOSH, K.; DAZELEY, R. usfAD: A robust anomaly detector based on unsupervised stochastic forest. *Int. J. Mach. Learn. & Cyber.*, v. 12, n. 4, p. 1137–1150, abr. 2021. ISSN 1868-808X.
- BANDARAGODA, T. R. et al. Efficient Anomaly Detection by Isolation Using Nearest Neighbour Ensemble. In: *2014 IEEE International Conference on Data Mining Workshop*. [S.l.: s.n.], 2014. p. 698–705. ISSN 2375-9259.
- BELKIN, M.; NIYOGI, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, MIT Press, v. 15, n. 6, p. 1373–1396, 2003. Disponível em: <https://doi.org/10.1162/089976603321780317>.
- BELKIN, M.; NIYOGI, P. Convergence of laplacian eigenmaps. In: *Advances in Neural Information Processing Systems*. [S.l.]: MIT Press, 2006. v. 19.
- BERCEA, C. I. et al. Diffusion Models with Implicit Guidance for Medical Anomaly Detection. In: LINGURARU, M. G. et al. (Ed.). *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Cham: Springer Nature Switzerland, 2024. p. 211–220. ISBN 978-3-031-72120-5.
- BHOSALE, A. et al. *Anomaly Detection Using Diffusion-based Methods*. 2024.
- BIYIKOĞU, T.; LEYDOLD, J.; STADLER, P. F. *Laplacian eigenvectors of graphs: Perron-Frobenius and Faber-Krahn type theorems*. [S.l.]: Springer, 2007.
- BOUMAN, R.; BUKHSH, Z.; HESKES, T. Unsupervised Anomaly Detection Algorithms on Real-world Data: How Many Do We Need? *Journal of Machine Learning Research*, v. 25, n. 105, p. 1–34, 2024. ISSN 1533-7928. Disponível em: <http://jmlr.org/papers/v25/23-0570.html>.
- BRADLEY, A. P. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, Elsevier, v. 30, n. 7, p. 1145–1159, 1997.
- BREIMAN, L. Random Forests. *Machine Learning*, v. 45, n. 1, p. 5–32, out. 2001. ISSN 1573-0565.

- BREUNIG, M. M. et al. LOF: Identifying density-based local outliers. In: *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. [s.n.], 2000. p. 93–104. Disponível em: <https://doi.org/10.1145/335191.335388>.
- BROWN, G. et al. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *The journal of machine learning research*, JMLR. org, v. 13, n. 1, p. 27–66, 2012.
- BURDEA, G. C.; COIFFET, P. *Virtual reality technology*. [S.l.]: John Wiley & Sons, 2003.
- CAMPOS, G. O. et al. On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study. *Data Min Knowl Disc*, v. 30, n. 4, p. 891–927, jul. 2016. ISSN 1573-756X. Disponível em: <https://doi.org/10.1007/s10618-015-0444-8>.
- CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Outlier detection: A survey. *ACM Computing Surveys*, v. 14, p. 15, 2007. Disponível em: https://www.researchgate.net/profile/Vipin-Kumar-54/publication/242403027_Outlier_Detection_A_Survey/links/0deec52d83600bb86c000000/Outlier-Detection-A-Survey.pdf.
- CHEN, J. et al. An anomaly detection method for wireless sensor networks based on the improved isolation forest. *Applied Sciences*, MDPI, v. 13, n. 2, p. 702, 2023.
- COIFMAN, R. R.; LAFON, S. Diffusion maps. *Applied and computational harmonic analysis*, v. 21, n. 1, p. 5–30, 2006. ISSN 1063-5203. Disponível em: <https://doi.org/10.1016/j.acha.2006.04.006>.
- COLLADO, J. F. et al. *New methods for triangulation-based shape acquisition using laser scanners*. [S.l.]: Universitat de Girona, 2004.
- COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE transactions on information theory*, IEEE, v. 13, n. 1, p. 21–27, 1967.
- COVER, T. M.; THOMAS, J. A. Data compression. *Elements of Information Theory*, John Wiley & Sons, Ltd, p. 103–158, 2005.
- DARBAN, Z. Z. et al. Deep Learning for Time Series Anomaly Detection: A Survey. *ACM Comput. Surv.*, v. 57, n. 1, p. 15:1–15:42, out. 2024. ISSN 0360-0300.
- DEMEDEIROS, K.; HENDAWI, A.; ALVAREZ, M. A survey of ai-based anomaly detection in iot and sensor networks. *Sensors*, MDPI, v. 23, n. 3, p. 1352, 2023.
- DOMINGUES, R. et al. A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognition*, v. 74, p. 406–421, fev. 2018. ISSN 0031-3203. Disponível em: <https://doi.org/10.1016/j.patcog.2017.09.037>.
- ELSNER, L. A note on the hoffman-wielandt theorem. *Linear Algebra and its applications*, Elsevier, v. 182, p. 235–237, 1993.
- FAWCETT, T. An introduction to ROC analysis. *Pattern Recognition Letters*, v. 27, n. 8, p. 861–874, jun. 2006. ISSN 0167-8655.
- FAYYAD, U.; Piatetsky-Shapiro, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, v. 17, n. 3, p. 37–37, mar. 1996. ISSN 2371-9621.

FIGUEIRÊDO, B. C. B. d. *Uma nova centralidade para Redes Multiplex não direcionadas*. Dissertação (Mestrado) — Universidade Federal do Amazonas, Manaus, Brasil, 2021.

FOORTHUIS, R. SECODA: Segmentation- and Combination-Based Detection of Anomalies. In: *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. [s.n.], 2017. p. 755–764. Disponível em: <https://doi.org/10.1109/DSAA.2017.35>.

FRANC, M.; SKALA, V. Parallel triangular mesh decimation without sorting. In: *IEEE. Proceedings Spring Conference on Computer Graphics*. [S.l.], 2001. p. 22–29.

FUKUNAGA, K.; HOSTETLER, L. Optimization of k nearest neighbor density estimates. *IEEE Transactions on information theory*, IEEE, v. 19, n. 3, p. 320–326, 1973.

GOLDSTEIN, M.; DENGEL, A. Histogram-based Outlier Score (HBOS): A fast Unsupervised Anomaly Detection Algorithm. In: . [s.n.], 2012. Disponível em: <https://www.goldiges.de/publications/HBOS-KI-2012.pdf>.

GOLDSTEIN, M.; UCHIDA, S. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one*, v. 11, n. 4, p. e0152173, 2016. ISSN 1932-6203. Disponível em: <https://doi.org/10.1371/journal.pone.0152173>.

GOMES; PIMENTA; SCHNEIDER. Data mining in information science research: Challenges and opportunities. 2019. Disponível em: <https://zenodo.org/records/3521038>.

GREKOV, A. N. et al. Anomaly detection in biological early warning systems using unsupervised machine learning. *Sensors*, MDPI, v. 23, n. 5, p. 2687, 2023.

GUO, J. et al. An Anomaly Detection Framework Based on Autoencoder and Nearest Neighbor. In: *2018 15th International Conference on Service Systems and Service Management (ICSSSM)*. [s.n.], 2018. p. 1–6. ISSN 2161-1904. Disponível em: <https://doi.org/10.1109/ICSSSM.2018.8464983>.

HAN, S. et al. ADBench: Anomaly Detection Benchmark. *Advances in Neural Information Processing Systems*, v. 35, p. 32142–32159, dez. 2022.

HAND, D. J.; TILL, R. J. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine learning*, Springer, v. 45, n. 2, p. 171–186, 2001.

HE, H. et al. *DiAD: A Diffusion-based Framework for Multi-class Anomaly Detection*. [S.l.]: arXiv, 2023.

HO, J.; JAIN, A.; ABBEEL, P. Denoising diffusion probabilistic models. In: *Advances in Neural Information Processing Systems*. [s.n.], 2020. v. 33, p. 6840–6851. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

HOSMER DAVID W., J.; LEMESHOW, S.; STURDIVANT, R. X. *Applied Logistic Regression*. 3. ed. Hoboken, NJ: John Wiley & Sons, 2013. Disponível em: <https://doi.org/10.1002/9781118548387>.

- HUANG, J.; YANG, Y. Anomaly detection in machining centers based on graph diffusion-hierarchical neighbor aggregation networks. *Applied Sciences*, MDPI, v. 13, n. 23, p. 12914, 2023.
- IZBICKI, R.; SANTOS, T. M. dos. *Aprendizado de máquina: uma abordagem estatística*. [S.l.]: Rafael Izbicki, 2020.
- JAYALAKSHMI, T.; SANTHAKUMARAN, A. Statistical normalization and back propagation for classification. *International Journal of Computer Theory and Engineering*, v. 3, n. 1, p. 1793–8201, 2011.
- JAYNES, E. T. Information theory and statistical mechanics. *Physical review*, APS, v. 106, n. 4, p. 620, 1957.
- JIANG, S. et al. Credit card fraud detection based on unsupervised attentional anomaly detection network. *Systems*, MDPI, v. 11, n. 6, p. 305, 2023.
- JIN, W. et al. Ranking outliers using symmetric neighborhood relationship. In: SPRINGER. *Pacific-Asia conference on knowledge discovery and data mining*. [S.l.], 2006. p. 577–593.
- JUVONEN, A.; HÄMÄLÄINEN, T. An efficient network log anomaly detection system using random projection dimensionality reduction. In: *New Technologies, Mobility and Security (NTMS), 2014 6th International Conference on*. [S.l.: s.n.], 2014. p. 1–5.
- KAHN, M. *Diabetes*. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5T59G>.
- KELLY, M.; LONGJOHN, R.; NOTTINGHAM, K. *The UCI Machine Learning Repository*. 2025.
- KNORR, E. M.; NG, R. T.; TUCAKOV, V. Distance-based outliers: Algorithms and applications. *The VLDB Journal*, v. 8, n. 3, p. 237–253, fev. 2000. ISSN 0949-877X. Disponível em: <https://doi.org/10.1007/s007780050006>.
- KNOX, E. M.; NG, R. T. Algorithms for Mining Distance-Based Outliers in Large Datasets. In: CITESEER. *Proceedings of the international conference on very large data bases*. 1998. p. 392–403. Disponível em: <https://www.vldb.org/conf/1998/p392.pdf>.
- KOCK, N. F. J.; MCQUEEN, R. J.; BAKER, M. Learning and process improvement in knowledge organizations: a critical analysis of four contemporary myths. *The Learning Organization*, MCB UP Ltd, v. 3, n. 1, p. 31–41, 1996.
- KRIEGEL, H.-P. et al. LoOP: Local outlier probabilities. In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. New York, NY, USA: Association for Computing Machinery, 2009. (CIKM '09), p. 1649–1652. ISBN 978-1-60558-512-3.
- KRIEGEL, H.-P.; SCHUBERT, M.; ZIMEK, A. Angle-based outlier detection in high-dimensional data. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2008. (KDD '08), p. 444–452. ISBN 978-1-60558-193-4. Disponível em: <https://doi.org/10.1145/1401890.1401946>.

- KUDO, M.; TOYAMA, J.; SHIMBO, M. *Japanese Vowels*. [S.l.]: UCI Machine Learning Repository, 1999.
- KUSHNER, H. J.; YIN, G. G. *Stochastic Approximation and Recursive Algorithms and Applications*. 2. ed. New York, NY: Springer, 2003. v. 35. (Stochastic Modelling and Applied Probability, v. 35).
- LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, v. 86, n. 11, p. 2278–2324, nov. 1998. ISSN 1558-2256.
- LI, Y.; LIU, Y.; ZANG, C. Discriminant diffusion maps based k-nearest-neighbour for batch process fault detection. *Chemical Engineering Research and Design*, 2018.
- LI, Y.; ZHANG, X. Diffusion maps based k-nearest-neighbor rule technique for semiconductor manufacturing process fault detection. *Chemometrics and Intelligent Laboratory Systems*, v. 136, p. 47, 2014. Disponível em: <https://doi.org/10.1016/j.chemolab.2014.05.003>.
- LIU, F. T.; TING, K. M.; ZHOU, Z.-H. Isolation Forest. In: *2008 Eighth IEEE International Conference on Data Mining*. [s.n.], 2008. p. 413–422. ISSN 2374-8486. Disponível em: <https://doi.org/10.1109/ICDM.2008.17>.
- LIU, J. et al. *A Survey on Diffusion Models for Anomaly Detection*. [S.l.]: arXiv, 2025.
- LIU, K. et al. Graph diffusion models for anomaly detection. In: *Proceedings of the ACM Web Conference 2024*. [S.l.: s.n.], 2024.
- LIU, P. et al. Nonuniform data loss reconstruction based on time-series-specialized neural networks for structural health monitoring. *Structural Health Monitoring*, SAGE Publications Sage UK: London, England, p. 14759217251321760, 2025.
- MEYER, C. D. *Matrix analysis and applied linear algebra*. [S.l.]: SIAM, 2023.
- MICENKOVÁ, B.; MCWILLIAMS, B.; ASSENT, I. Learning Outlier Ensembles: ACM SIGKDD 2014 Workshop ODD2. In: . [S.l.: s.n.], 2014.
- MISHNE, G.; COHEN, I. Multiscale anomaly detection using diffusion maps. *IEEE Journal of Selected Topics in Signal Processing*, v. 7, n. 1, p. 111–123, 2013.
- MISHNE, G.; COHEN, I. Multiscale anomaly detection using diffusion maps and saliency score. In: *Proceedings of the 39th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2014.
- MISHNE, G.; COHEN, I. Multi-channel wafer defect detection using diffusion maps. In: *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. [S.l.: s.n.], 2015. p. 1740–1744.
- MISHNE, G.; COHEN, I. Iterative diffusion-based anomaly detection. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2017. p. 2337–2341.
- MISHNE, G. et al. Diffusion nets. *Applied and Computational Harmonic Analysis*, 2017.

- MOORE, A. W. An introductory tutorial on kd-trees. *Computer Laboratory, University of Cambridge, Tech. Rep*, v. 209, 1991.
- MOREIRA, L. A. S. *Funções de Difusão e Regiões de Zero de Campos de Vetores Planares: Mapeamentos de Zonas de Singularidades*. Dissertação (Mestrado) — Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brasil, 2010.
- NADARAYA, E. A. On estimating regression. *Theory of Probability & Its Applications*, SIAM, v. 9, n. 1, p. 141–142, 1964.
- NETO, F. M.; SOUZA, P.; de Magalhães, M. S. Determining baseline profile by diffusion maps. *Eur. J. Oper. Res.*, v. 279, n. 1, p. 107–123, 2019. ISSN 0377-2217. Disponível em: <https://doi.org/10.1016/j.ejor.2019.05.032>.
- OLIVEIRA, C. A. L. d. S. d. *Análise de sinais eletroquímicos por mapas de difusão*. 167 p. Tese (Doutorado) — Universidade do Estado do Rio de Janeiro, Instituto Politécnico, Nova Friburgo, Brasil, 2022.
- OLSON, C. C.; JUDD, K. P.; NICHOLS, J. M. Manifold learning techniques for unsupervised anomaly detection. *Expert Systems with Applications*, v. 00, p. 1–22, 2017.
- PAGE, L. et al. *The PageRank citation ranking: Bringing order to the web*. [S.l.], 1999.
- PANG, G. et al. Deep Learning for Anomaly Detection: A Review. *ACM Comput. Surv.*, v. 54, n. 2, p. 38:1–38:38, mar. 2021. ISSN 0360-0300.
- PAPADIMITRIOU, S. et al. LOCI: Fast outlier detection using the local correlation integral. In: *Proceedings 19th International Conference on Data Engineering (Cat. No.03CH37405)*. [S.l.: s.n.], 2003. p. 315–326.
- PEYKANI, P. et al. Input/output variables selection in data envelopment analysis: A shannon entropy approach. *Machine Learning and Knowledge Extraction*, MDPI, v. 4, n. 3, p. 688–699, 2022.
- QIAN, G. et al. Consistent seismic event detection using multi-input end-to-end neural networks for structural health monitoring. *Structural Control and Health Monitoring*, Wiley Online Library, v. 2025, n. 1, p. 9966359, 2025.
- RAMASWAMY, S.; RASTOGI, R.; SHIM, K. Efficient algorithms for mining outliers from large data sets. In: *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. New York, NY, USA: Association for Computing Machinery, 2000. (SIGMOD '00), p. 427–438. ISBN 978-1-58113-217-5.
- RAYANA, S.; AKOGLU, L. Less is More: Building Selective Anomaly Ensembles. *ACM Trans. Knowl. Discov. Data*, v. 10, n. 4, p. 42:1–42:33, maio 2016. ISSN 1556-4681.
- ROSS, S. M. *Introduction to probability models*. 7. ed. [S.l.]: Academic Press, 2000.
- SAMARIYA, D.; THAKKAR, A. A comprehensive survey of anomaly detection algorithms. *Annals of Data Science*, Springer, v. 10, n. 3, p. 829–850, 2023. ISSN 2198-5804. Disponível em: <https://doi.org/10.1007/s40745-021-00362-9>.

- SATHE, S.; AGGARWAL, C. LODS: Local Density Meets Spectral Outlier Detection. In: *Proceedings of the 2016 SIAM International Conference on Data Mining*. [S.l.]: Society for Industrial and Applied Mathematics, 2016. p. 171–179. ISBN 978-1-61197-434-8.
- SCHUBERT, E.; ZIMEK, A.; KRIEGEL, H.-P. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. *Data mining and knowledge discovery*, Springer, v. 28, n. 1, p. 190–237, 2014.
- SEPÚLVEDA-FONTAINE, S. A.; AMIGÓ, J. M. Applications of entropy in data analysis and machine learning: a review. *Entropy*, MDPI, v. 26, n. 12, p. 1126, 2024.
- SHANNON, C. E. A mathematical theory of communication. *The Bell system technical journal*, v. 27, n. 3, p. 379–423, 1948. ISSN 0005-8580.
- SHIROTORI, M.; SUGITANI, Y. Change-point detection using diffusion maps for sleep apnea monitoring with contact-free sensors. *PLoS ONE*, v. 19, n. 6, p. e0306139, 2024. Disponível em: <https://doi.org/10.1371/journal.pone.0306139>.
- SIGILLITO, V. et al. *Ionosphere*. [S.l.]: UCI Machine Learning Repository, 1989.
- SILVA, D. A. N. S. *Um Método Geral para Redução de Grandes Conjuntos de Dados Baseado em Difusão Geométrica Markoviana*. Dissertação (Mestrado) — Universidade Federal da Paraíba, João Pessoa, Brasil, 2012.
- SILVA, D. A. N. S.; SOUZA, L. C.; MOTTA, G. H. M. B. An instance selection method for large datasets based on Markov Geometric Diffusion. *Data & Knowledge Engineering*, v. 101, p. 24–41, jan. 2016. ISSN 0169-023X.
- SINGER, A. et al. Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps. *Proceedings of the National Academy of Sciences*, National Academy of Sciences, v. 106, n. 38, p. 16090–16095, 2009. Disponível em: <https://www.pnas.org/doi/abs/10.1073/pnas.0905547106>.
- SINGER, A.; WU, H.-t. Orientability and diffusion maps. *Applied and Computational Harmonic Analysis*, v. 31, n. 1, p. 44–58, jul. 2011. ISSN 1063-5203. Disponível em: <https://doi.org/10.1016/j.acha.2010.10.001>.
- SIPOLA, T.; JUVONEN, A.; LEHTONEN, J. Anomaly detection from network logs using diffusion maps. In: *Engineering Applications of Neural Networks*. [s.n.], 2011. v. 363, p. 172–181. Disponível em: https://doi.org/10.1007/978-3-642-23957-1_20.
- SLATE, D. *Letter Recognition*. [S.l.]: UCI Machine Learning Repository, 1991.
- SMITH, J. W. et al. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In: *Proceedings of the annual symposium on computer application in medical care*. [S.l.: s.n.], 1988. p. 261.
- SOUZA, L. C. d. *Simplificação de Malhas com Preservação de Feições baseada em Difusão Geométrica Markoviana*. Dissertação (Mestrado) — Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brasil, 2011.

- SOUZA, R. A. S. d. *Um Sistema de Apoio à Detecção de Anomalias em Dados Governamentais usando Múltiplos Classificadores*. Dissertação (Mestrado) — Universidade Federal da Paraíba (UFPB), Paraíba, Brasil, abr. 2021.
- STRANG, G. *Linear algebra and its applications*. 3. ed. [S.l.]: Brooks Cole, 1988.
- SULE, S.; EVANS, L.; CAMERON, M. Sharp error estimates for target measure diffusion maps with applications to the committor problem. *arXiv preprint arXiv:2312.14418*, 2023. Disponível em: <https://doi.org/10.48550/arXiv.2312.14418>.
- SUN, J.; OVSJANIKOV, M.; GUIBAS, L. A concise and provably informative multi-scale signature based on heat diffusion. In: WILEY ONLINE LIBRARY. *Computer graphics forum*. 2009. v. 28, n. 5, p. 1383–1392. Disponível em: <https://doi.org/10.1111/j.1467-8659.2009.01515.x>.
- TANG, J. et al. Enhancing Effectiveness of Outlier Detections for Low Density Patterns. In: CHEN, M.-S.; YU, P. S.; LIU, B. (Ed.). *Advances in Knowledge Discovery and Data Mining*. Berlin, Heidelberg: Springer, 2002. p. 535–548. ISBN 978-3-540-47887-4. Disponível em: https://doi.org/10.1007/3-540-47887-6_53.
- TANG, Z. et al. Automated seismic event detection considering faulty data interference using deep learning and bayesian fusion. *Computer-Aided Civil and Infrastructure Engineering*, Wiley Online Library, v. 40, n. 14, p. 1910–1931, 2025.
- TCE-PB. *SAGRES - Sistema de Acompanhamento da Gestão dos Recursos da Sociedade*. 2023. Disponível em: <https://sagresonline.tce.pb.gov.br>. Acesso em: 17 set. 2023.
- WATSON, G. S. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, JSTOR, p. 359–372, 1964.
- XU, R. et al. Applications of diffusion maps in gene expression data-based cancer diagnosis analysis. In: *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. [s.n.], 2007. p. 2200–2203. Disponível em: <https://doi.org/10.1109/IEMBS.2007.4353367>.
- YANG, Y. et al. Mst: Multiscale flow-based student–teacher network for unsupervised anomaly detection. *Electronics*, MDPI, v. 13, n. 16, p. 3224, 2024.
- ZHANG, H. et al. DiffusionAD: Norm-Guided One-Step Denoising Diffusion for Anomaly Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1–13, 2025. ISSN 1939-3539. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.
- ZHANG, K.; HUTTER, M.; JIN, H. A New Local Distance-Based Outlier Detection Approach for Scattered Real-World Data. In: THEERAMUNKONG, T. et al. (Ed.). *Advances in Knowledge Discovery and Data Mining*. Berlin, Heidelberg: Springer, 2009. p. 813–822. ISBN 978-3-642-01307-2.
- ZHAO, Y.; NASRULLAH, Z.; LI, Z. PyOD: A Python Toolbox for Scalable Outlier Detection. *Journal of Machine Learning Research*, v. 20, n. 96, p. 1–7, 2019. ISSN 1533-7928. Disponível em: <http://jmlr.org/papers/v20/19-011.html>.