



UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de São José do Rio Preto

Bruno Cesar Silva Rodrigues

Análise de Padrões na Educação via Redes Neurais e Aprendizado de Máquina

São José do Rio Preto

2024

Bruno Cesar Silva Rodrigues

Análise de Padrões na Educação via Redes Neurais e Aprendizado de Máquina

Trabalho de Conclusão de Curso (TCC) apresentado como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação, junto ao Conselho de Curso de Bacharelado em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de São José do Rio Preto.

UNESP - Universidade Estadual Paulista

IBILCE - Instituto de Biociências, Letras e Ciências Exatas

Orientador: Prof. Dr. Wallace Correa de Oliveira Casaca

São José do Rio Preto

2024

R696a

Rodrigues, Bruno Cesar Silva

Análise de padrões na educação via redes neurais e aprendizado de máquina / Bruno Cesar Silva Rodrigues. -- São José do Rio Preto, 2024

56 p. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (UNESP), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto

Orientador: Wallace Correa de Oliveira Casaca

1. Redes neurais (Computação). 2. Representações dos grafos. 3. Inteligência artificial. 4. Censo escolar. I. Título.

BRUNO CESAR SILVA RODRIGUES

Análise de Padrões na Educação via Redes Neurais e Aprendizado de Máquina

Trabalho de Conclusão de Curso (TCC) apresentado como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação, junto ao Conselho de Curso de Bacharelado em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Câmpus de São José do Rio Preto.

Banca Examinadora:

Prof. Dr. Wallace Correa de Oliveira
Casaca(Orientador)
UNESP – Câmpus de São José do Rio Preto

Prof. Dr. Carlos Roberto Valêncio
UNESP – Câmpus de São José do Rio Preto

Prof. Dr. Lucas Correia Ribas
UNESP – Câmpus de São José do Rio Preto

São José do Rio Preto

2024

Para minha mãe Maria, ao meu pai Paulo e à minha vó Luzia

Agradecimentos

Gostaria de começar agradecendo aos meus pais, Paulo Cesar Rodrigues e Maria dos Reis Silva Rodrigues, e à minha avó Luzia do Carmo Silva, a quem devo todas as minhas conquistas durante a graduação e além.

Agradeço também aos excelentes professores que nos acompanharam ao longo dessa jornada, inspirando-nos com seu conhecimento e apoio, com destaque especial ao meu orientador, Wallace Correa de Oliveira Casaca, por sua dedicação no desenvolvimento dos meus projetos.

Por fim, expresso minha gratidão aos amigos e colegas que tive o privilégio de conhecer durante a graduação.

*"But things we take for granted...
They're really the most essential things.
And they're precious."
(The Partner - Mystery Dungeon)*

Resumo

O Censo Escolar do INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) é uma fonte indispensável de dados educacionais no Brasil, servindo como base para o planejamento de ações e melhorias na educação. No entanto, o grande volume e a complexidade desses dados impõem desafios significativos para sua análise e interpretação. Portanto, este trabalho propõe a aplicação de Redes Neurais de Grafos (GNNs) como uma abordagem inovadora para enfrentar os desafios associados à alta dimensionalidade, à falta de padronização e às dificuldades de interpretação dos dados do mais recente Censo Escolar de 2023. Através deste trabalho, foi demonstrado como as GNNs podem ser efetivas em capturar relações complexas entre os dados, facilitando a extração de informações relevantes, bem como a geração de insights educacionais significativos.

Palavras-chave: Redes neurais artificiais, Representações de grafos, Inteligência artificial, Censo escolar, GNN.

Abstract

The INEP School Census (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) is an indispensable source of educational data in Brazil, serving as the foundation for planning actions and improving education. However, the large volume and complexity of this data collection pose significant challenges for exploration and interpretation. Therefore, this study proposes the application of Graph Neural Networks (GNNs) as an innovative tool to address the challenges associated with high dimensionality, lack of standardization, and difficulties in interpreting the data from the most recent 2023 School Census. This work demonstrates how GNNs can effectively capture complex relationships within the data, facilitating the extraction of relevant information and generating meaningful educational insights.

Keywords: Artificial neural networks, Representations of graphs, Artificial intelligence, School census, GNN.

Lista de figuras

Figura 2.3.1–Convolução 2-D versus convolução em grafos. (a) Convolução 2-D: Cada pixel em uma imagem é considerado um nó em que os vizinhos são determinados pelo tamanho do filtro. A convolução 2-D considera a média ponderada dos valores de pixel do nó vermelho junto com seus vizinhos. Os vizinhos de um nó são ordenados e têm um tamanho fixo. (b) Convolução em grafos: Para obter a representação oculta do nó vermelho, uma solução simples que a operação de convolução em grafos toma, é de obter o valor médio das características do nó vermelho junto com seus vizinhos. Diferentemente dos dados da imagem, os vizinhos de um nó não são ordenados e têm tamanho variável. Fonte: (WU et al., 2020)	20
Figura 2.3.2–Diferentes modelos GNN criados com camadas de convolução em grafo. Gconv representa uma camada de convolução em grafos, MLP um perceptron multicamadas e CNN uma camada convolucional tradicional. Fonte: (WU et al., 2020)	21
Figura 2.3.3–Diferenças entre GCN e GAT: (a) O GCN atribui explicitamente um peso não paramétrico $a_{ij} = \frac{1}{\sqrt{\deg(v_i)\deg(v_j)}}$ ao vizinho v_j de v_i durante o processo de agregação. (b) O GAT captura implicitamente o peso a_{ij} por meio de uma arquitetura neural end-to-end, permitindo que nós mais importantes recebam pesos maiores. Fonte: (WU et al., 2020)	22
Figura 3.0.1–Pipeline da metodologia do projeto	27
Figura 3.2.1–Subgrafo da UF SP com Dependência Administrativa Privada	30
Figura 3.3.1–Histograma dos Erros de Reconstrução	32
Figura 4.3.1–Concentração de Anomalias por Região	48

Lista de tabelas

Tabela 1 – Comparação das Características dos Nós - Global	33
Tabela 2 – Comparação das Características dos Nós - Região Norte	35
Tabela 3 – Comparação das Características dos Nós - Região Nordeste	36
Tabela 4 – Comparação das Características dos Nós - Região Sudeste	37
Tabela 5 – Comparação das Características dos Nós - Região Sul	38
Tabela 6 – Comparação das Características dos Nós - Região Centro-Oeste	39
Tabela 7 – Comparação das Características dos Nós - Global (Sobreajuste)	40
Tabela 8 – Comparação das Características dos Nós - Região Norte (Sobreajuste) . . .	41
Tabela 9 – Comparação das Características dos Nós - Região Nordeste (Sobreajuste) .	43
Tabela 10 – Comparação das Características dos Nós - Região Sudeste (Sobreajuste) .	44
Tabela 11 – Comparação das Características dos Nós - Região Sul (Sobreajuste)	45
Tabela 12 – Comparação das Características dos Nós - Região Centro-Oeste (Sobreajuste)	46

Sumário

1	INTRODUÇÃO	13
1.1	Motivação	13
1.2	Objetivo	14
1.3	Estrutura do Trabalho	14
2	REVISÃO BIBLIOGRÁFICA	16
2.1	IA	16
2.1.1	Aprendizado de Máquina	16
2.1.2	Modelos Generativos (LLMs)	18
2.2	Mineração de Dados	19
2.2.1	Definição	19
2.2.2	Aplicações	19
2.3	GNNs	20
2.3.1	Definição	20
2.3.2	Aplicações	23
2.4	Trabalhos Relacionados	24
2.4.1	Predição de Links na Educação	24
2.4.2	Rastreamento de Conhecimento na Educação usando Grafos	25
2.4.3	Aprendizagem Adaptativa utilizando GNNs e Aprendizado por Reforço	26
2.4.4	Relação com este projeto	26
3	METODOLOGIA	27
3.1	Etapas de Pré-Processamento, Limpeza e Seleção dos Dados	27
3.2	Criação do Grafo	29
3.3	Criação do Modelo	31
3.3.1	Criação do Modelo Sobreajustado	32
4	RESULTADOS	33
4.1	Resultados dos Dados Originais	33
4.1.1	Resultados do Grafo Global	33
4.1.2	Resultados da Região Norte	35
4.1.3	Resultados da Região Nordeste	36
4.1.4	Resultados da Região Sudeste	37
4.1.5	Resultados da Região Sul	38
4.1.6	Resultados da Região Centro-Oeste	39
4.2	Resultados com Dados com Sobreajuste	40

4.2.1	Resultados do Grafo Global (Sobreajuste)	40
4.2.2	Resultados da Região Norte (Sobreajuste)	41
4.2.3	Resultados da Região Nordeste (Sobreajuste)	43
4.2.4	Resultados da Região Sudeste (Sobreajuste)	44
4.2.5	Resultados da Região Sul (Sobreajuste)	45
4.2.6	Resultados da Região Centro-Oeste (Sobreajuste)	46
4.3	Outras Análises	48
5	CONCLUSÃO	49
5.1	Grafo Global (Original)	49
5.2	Regiões Específicas (Originais)	49
5.3	Grafo Global (Sobreajuste)	50
5.4	Regiões Específicas (Sobreajuste)	50
5.5	Limitações	51
5.6	Melhorias Futuras	52
	REFERÊNCIAS	53

1 Introdução

O Censo Escolar Nacional (CEN), conduzido pelo INEP, é uma fonte imprescindível de informações sobre a educação brasileira, uma vez que reúne dados de diferentes modalidades e natureza, como dados demográficos, estatísticas sobre matrículas, informações sobre a infraestrutura escolar, e até mesmo resultados de avaliações educacionais. Tais dados estão dispostos de forma a representar toda a realidade educacional brasileira (DINIZ, 1999), possibilitando assim a gestão e implementação de políticas públicas em escopo nacional. Porém, essa grande quantidade de informações é compilada na forma de arquivos CSV (Comma Separated Values), que apresentam centenas de colunas e milhares de linhas, o que torna a busca por informações relevantes uma árdua tarefa.

A educação é um pilar fundamental para o desenvolvimento de qualquer país, e a análise eficiente dos dados educacionais pode oferecer *insights* valiosos para a melhoria contínua do sistema educacional. Entretanto, a complexidade e o volume dos dados do CEN representam um desafio significativo para gestores e pesquisadores, que muitas vezes carecem de ferramentas e técnicas adequadas para extrair informações úteis de forma eficaz.

Entre os principais desafios na análise dos dados do CEN estão a integração de diferentes tipos de dados, a complexidade das relações entre variáveis e a necessidade de identificar padrões relevantes em um mar de informações. Além disso, a heterogeneidade dos dados, que variam amplamente em termos de qualidade e completude, adiciona outra camada de dificuldade para análises precisas e confiáveis.

1.1 Motivação

O CEN possui uma enorme quantidade de informações que pode auxiliar significativamente na tomada de decisões tanto em âmbito municipal (especialmente em municípios rurais ou menores) quanto em âmbito regional (capitais e grandes cidades). No entanto, devido ao escopo abrangente da pesquisa e à quantidade excessiva de colunas, os dados não são facilmente interpretáveis.

A falta de ferramentas adequadas para interpretar esses dados pode levar a decisões mal informadas e, conseqüentemente, a políticas públicas ineficazes. Por exemplo, gestores educacionais em municípios pequenos podem não ter acesso aos recursos analíticos necessários para identificar deficiências específicas em suas escolas, enquanto gestores de grandes cidades podem enfrentar dificuldades para priorizar investimentos em infraestrutura escolar de maneira eficiente.

Considerando o escopo deste trabalho, o objetivo central desta pesquisa é estudar,

explorar e otimizar Redes Neurais de Grafos de forma a contornar as principais dificuldades e *gaps* existentes nos processos de extração da informação em meio aos dados do CEN-INEP. Algumas das dificuldades encontradas são sumarizadas a seguir:

- **Alta dimensionalidade dos dados:** Como dito anteriormente os dados estão distribuídos em centenas de colunas, o que dificulta consideravelmente a análise e interpretação dos mesmos por parte de usuários não familiarizados com ferramentas da Tecnologia da Informação.
- **Dificuldade para encontrar informações relevantes:** Devido ao grande volume de dados disponibilizados, encontrar informações relevantes sobre um determinado tópico de interesse torna-se uma tarefa de difícil realização. A verificação da existência de padrões, bem como comportamentos e correlações específicas acerca dos dados também constitui um desafio a ser superado neste contexto.
- **Ausência de padronização dos dados:** Os dados foram coletados e apresentados de forma não padronizada entre cada versão do CEN, o que dificulta tecer comparações e a definição de hipóteses de estudo.
- **Dificuldade na interpretação dos resultados:** Mesmo após a análise dos dados, pode ser um desafio interpretar e representar as conclusões de maneira acessível as partes interessadas.

1.2 Objetivo

O objetivo deste trabalho é tornar os dados do CEN mais facilmente interpretáveis para gestores públicos e privados, de modo que se tornem uma ferramenta útil na tomada de decisões. Para alcançar esse objetivo, serão utilizadas técnicas de Redes Neurais e Aprendizado de Máquina, especificamente Redes Neurais de Grafos (do inglês, GNNs - *Graph Neural Networks*)(WU et al., 2020; CHADHA; JAIN, 2020), e uma de suas variantes, as Redes Neurais de Grafos Convolucionais (do inglês, GCNs - *Graph Convolutional Networks*). Por meio da representação dos dados em formato de grafos, buscamos oferecer uma visão mais clara das características-chave e suas inter-relações, além de identificar as principais falhas presentes na área educacional de cada região brasileira.

1.3 Estrutura do Trabalho

O restante deste trabalho está organizado da seguinte maneira. Inicialmente, a revisão bibliográfica detalhando as principais técnicas computacionais utilizadas na área da educação. Em seguida, será explicada a metodologia por trás da análise dos dados e a transformação dos

arquivos, do formato CSV para representação via grafos. Na próxima seção, será detalhado os passos da implementação da GNN a partir dos dados sob análise. Após isso, os resultados obtidos serão explorados e discutidos, incluindo uma breve descrição dos principais *insights* educacionais encontrados. Por fim, a última seção do trabalho será dedicada à conclusão da pesquisa.

2 Revisão Bibliográfica

A pesquisa sobre Inteligência Artificial (IA) e suas aplicações educacionais tem crescido exponencialmente nos últimos anos. Esta seção revisa os principais avanços em Aprendizado de Máquina, Modelos Generativos e Mineração de Dados, destacando suas contribuições e desafios na educação.

2.1 IA

A área de IA (Inteligência Artificial) dedica-se ao estudo e desenvolvimento de máquinas e programas computacionais capazes de reproduzir o comportamento humano na tomada de decisões e na realização de tarefas, desde as mais simples até as mais complexas. IA abrange diversas sub-áreas, que têm uma grande importância não só na educação (AYANWALE et al., 2022; BERNIUS; KRUSCHE; BRUEGGE, 2022), mas em diversas outras aplicações. Algumas dessas sub-áreas e suas aplicações gerais e educacionais serão descritas a seguir.

2.1.1 Aprendizado de Máquina

2.1.1.1 Definição

Aprendizado de Máquina (MITCHELL; MITCHELL, 1997; ALPAYDIN, 2020) é um subcampo da IA que se concentra no desenvolvimento de algoritmos e técnicas que permitem aos computadores aprender a partir de dados e fazer previsões ou tomar decisões sem serem explicitamente programados para cada tarefa. Em vez disso, esses algoritmos utilizam dados históricos para construir modelos matemáticos que identificam padrões e *insights*. Existem 3 tipos de treinamento de modelos de aprendizado de máquina:

- **Supervisionado:** O modelo é treinado com dados rotulados, ou seja, o algoritmo aprende a partir de entradas e saídas conhecidas.
- **Não Supervisionado:** O modelo é treinado com dados não rotulados e o objetivo é encontrar estruturas ocultas ou agrupamentos nos dados.
- **Reforço:** O modelo aprende por meio de interações com o ambiente, recebendo recompensas ou penalidades com base nas ações que toma.

2.1.1.2 Aplicações

O aprendizado de máquina tem sido utilizado em várias áreas (KUČAK; JURIČIĆ; DAMBIĆ, 2018) para resolver problemas complexos e melhorar a eficiência. Algumas aplicações

incluem:

- **Processamento de Linguagem Natural (PLN):** Ferramentas como o Google Tradutor utilizam algoritmos de aprendizado de máquina que se atualizam com base na entrada dos usuários, incorporando novas palavras e sintaxes. Assistentes virtuais como Siri, Alexa e Google Assistente dependem do PLN para reconhecer e sintetizar fala, permitindo-lhes entender ou pronunciar palavras novas.
- **Sistemas de Recomendação:** Serviços como Netflix, Amazon e Google utilizam o aprendizado de máquina para oferecer recomendações personalizadas com base na atividade de busca dos usuários. Essas recomendações melhoram significativamente as experiências online, com algoritmos que podem prever com alta certeza o que você comprará e quando. A Amazon, por exemplo, possui um sistema de "envio antecipado", que envia produtos para o armazém mais próximo antes mesmo de serem encomendados, para agilizar a entrega.
- **Trading Algorítmico:** O trading algorítmico (ou Algo Trading) envolve comportamento aleatório, dados em constante mudança e vários fatores. Enquanto os analistas financeiros não conseguem prever todo esse comportamento, os algoritmos de aprendizado de máquina podem fazê-lo e responder às mudanças no mercado muito mais rapidamente do que um humano.

Na área educacional, algumas aplicações notáveis são:

- **Previsão de Desempenho Estudantil:** O aprendizado de máquina pode prever o desempenho dos estudantes. Ao "aprender" sobre cada aluno, o modelo pode identificar fraquezas e sugerir formas de melhoria, como aulas adicionais ou leituras complementares (ZAMBRANO et al., 2021).
- **Melhoria na Retenção de Alunos:** Técnicas de aprendizado de máquina, como a análise de aprendizado, ajudam a melhorar as taxas de retenção. Identificando alunos "em risco", as escolas então podem tomar medidas para impedir que esse aluno cometa evasão escolar ou se torne "atrasado" quando comparado aos colegas (AL-SHABANDAR et al., 2019).
- **Correção de Provas:** Algoritmos baseados em aprendizado de máquina podem ajudar na classificação de provas manuscritas, aliviando a carga de trabalho dos professores e aumentando a eficiência do processo de correção (CHEN; HE, 2013; CELAR et al., 2015).

2.1.2 Modelos Generativos (LLMs)

2.1.2.1 Definição

Modelos generativos são uma classe de algoritmos de aprendizado de máquina que têm a capacidade de gerar novos dados a partir de padrões aprendidos a partir de dados de treinamento. Esses modelos não apenas aprendem a classificar ou prever, mas também a criar novas amostras que se assemelham aos dados de entrada. *Large Language Models* (LLMs), como ChatGPT, são uma subcategoria específica de modelos generativos que se especializam no processamento e geração de linguagem natural. Eles são treinados em grandes corpora de texto e utilizam técnicas avançadas de redes neurais, como transformadores, para gerar texto coerente e contextualmente relevante.

2.1.2.2 Aplicações

Modelos generativos têm uma ampla gama de aplicações em diversos campos, incluindo, mas não se limitando a:

- **Criação de Conteúdo:** Modelos generativos podem criar diversos tipos de conteúdo, como histórias, scripts, código de software e até mesmo arte.
- **Saúde:** Na área médica, modelos generativos podem ser utilizados na geração de imagens médicas sintéticas para treinamento de modelos ou na descoberta de novos medicamentos através de modelagem molecular.

Na área educacional, modelos generativos podem ser usados para:

- **Modelos de Ensino Personalizados:** Como discutido em (KASNECI et al., 2023) modelos generativos como o ChatGPT podem ser utilizados para auxiliar a educação, criando conteúdo educacional, melhorando o engajamento e a interação dos alunos, e criando experiências de ensino personalizadas. O artigo também comenta as dificuldades que isso pode trazer ao ensino, com a necessidade de um preparo dos professores e dos alunos antes de sua implementação mais geral.
- **Provas e Testes Automatizados:** Como proposto em (QU; JIA; WU, 2021) modelos generativos podem criar questões e suas respostas automaticamente para uso em provas ou testes de múltipla escolha para os alunos, além disso, esses modelos podem ser fornecidos diretamente a alunos com mais necessidade para auxiliá-los a acompanhar seus colegas. Os autores de (SARSA et al., 2022) também propuseram uma versão interessante mais focada na área computacional.

2.2 Mineração de Dados

2.2.1 Definição

Mineração de Dados (do inglês, Data Mining), é o processo de descobrir padrões, correlações, tendências e informações úteis a partir de grandes conjuntos de dados utilizando métodos estatísticos, algoritmos de aprendizado de máquina e técnicas de análise de dados. O objetivo principal do data mining é transformar dados brutos em informações significativas e úteis, permitindo a tomada de decisões informadas em diversas áreas (MINING, 2006).

2.2.2 Aplicações

Para pessoas normais ou empresas, o Data Mining pode ser aplicado em diversas técnicas como:

- **Marketing e Vendas:** Segmentação de clientes e vendas baseada em diferentes grupos sociais. Um dos exemplos mais clássicos de Data Mining foi quando o Walmart encontrou uma correlação entre pais comprando fraldas e bebidas alcoólicas juntos.
- **Finanças:** A técnicas pode ser utilizada para tentar detectar fraudes em transações financeiras, realizar uma análise de risco de crédito, modelagem de scores de crédito ou até mesmo para tentar prever as tendências do mercado financeiro.
- **Redes Sociais:** Algumas das aplicações de Data Mining em redes sociais são análise de sentimento em postagens e comentários, recomendação de conteúdo personalizado baseado nos gostos do usuário e nas páginas que frequenta e detecção de spam e comportamento malicioso em redes sociais.

Na área educacional, diversos trabalhos e pesquisas já foram realizados com base nos dados do CEN, como por exemplo:

- **Análise de Desempenho Escolar:** No trabalho de (FRENEDA et al., 2020) foi detalhado como a mineração de dados aplicada ao CEN e às planilhas do ENEM foi utilizada para analisar o desempenho escolar dos alunos com o passar do tempo.
- **Análises Estatísticas:** Já no trabalho de (VIZZOTTO, 2020) foram utilizadas técnicas de mineração de dados e estatística para avaliar a inclusividade da educação brasileira ao longo do tempo.
- **Análise de Insights:** Por fim, O trabalho de (COLPANI, 2018) mostra como técnicas de mineração de dados podem ser usadas para procurar *insights* educacionais que contribuam para a melhoria da educação no país.

Foi observado que a mineração de dados tem sido a principal técnica empregada na extração de *insights* e na análise de dados em grande parte dessas pesquisas. Embora tenha proporcionado resultados positivos, é possível que esteja alcançando seu limite de eficácia devido à saturação. Diante disso, considerando a necessidade de inovação e avanços na área, este trabalho propõe uma abordagem alternativa para a análise dos dados do Censo Escolar, fundamentada no emprego e exploração de técnicas modernas de Inteligência Artificial (IA), sobretudo, de algoritmos baseados em Redes Neurais de Grafos.

2.3 GNNs

2.3.1 Definição

Graph Neural Networks (GNNs) são uma classe de redes neurais projetadas para trabalhar diretamente com dados estruturados como grafos. Diferente das redes neurais tradicionais que operam em dados euclidianos (como imagens ou texto), as GNNs são capazes de capturar dependências estruturais complexas entre os nós do grafo e suas conexões (arestas). As GNNs utilizam operações de agregação e atualização que permitem que a informação seja propagada e combinada através dos nós do grafo, tornando-as adequadas para tarefas como classificação de nós, previsão de links e geração de grafos (WU et al., 2020; SANCHEZ-LENGELING et al., 2021).

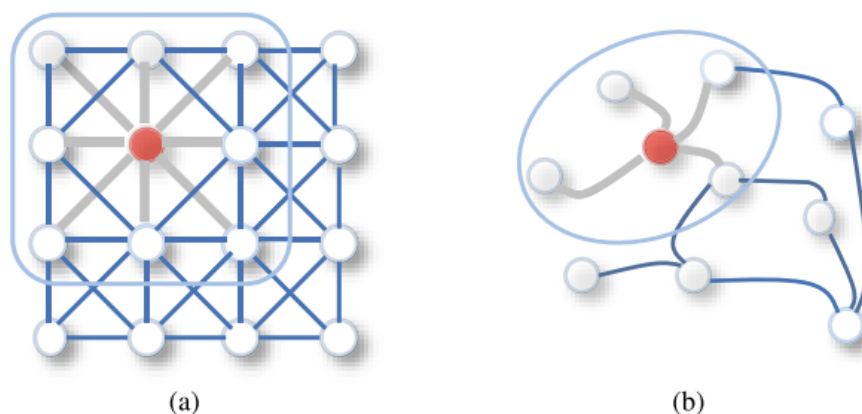


Figura 2.3.1 – Convolução 2-D versus convolução em grafos. (a) Convolução 2-D: Cada pixel em uma imagem é considerado um nó em que os vizinhos são determinados pelo tamanho do filtro. A convolução 2-D considera a média ponderada dos valores de pixel do nó vermelho junto com seus vizinhos. Os vizinhos de um nó são ordenados e têm um tamanho fixo. (b) Convolução em grafos: Para obter a representação oculta do nó vermelho, uma solução simples que a operação de convolução em grafos toma, é de obter o valor médio das características do nó vermelho junto com seus vizinhos. Diferentemente dos dados da imagem, os vizinhos de um nó não são ordenados e têm tamanho variável. **Fonte: (WU et al., 2020)**

A operação de convolução nas GNNs é um dos elementos centrais que as tornam especialmente poderosas para lidar com dados estruturados como grafos. Diferentemente da convolução em dados euclidianos, como imagens, onde os vizinhos de um pixel (nó) são fixos e ordenados pelo tamanho do filtro, a convolução em grafos opera em vizinhos que não são ordenados e podem variar em tamanho. Essa característica reflete a flexibilidade dos grafos em modelar relações complexas e heterogêneas entre entidades.

Enquanto na convolução 2-D os valores dos pixels vizinhos são combinados com base em uma média ponderada definida pelo filtro, nas GNNs, uma abordagem comum é calcular a média ou soma das características dos nós vizinhos, ajustadas por pesos aprendidos. Essa abordagem permite que informações contextuais se propaguem por todo o grafo, enriquecendo a representação dos nós e das arestas.

A Figura 2.3.1 ilustra a diferença entre a convolução tradicional em 2-D e a convolução em grafos, destacando as principais características e diferenças operacionais entre ambas.

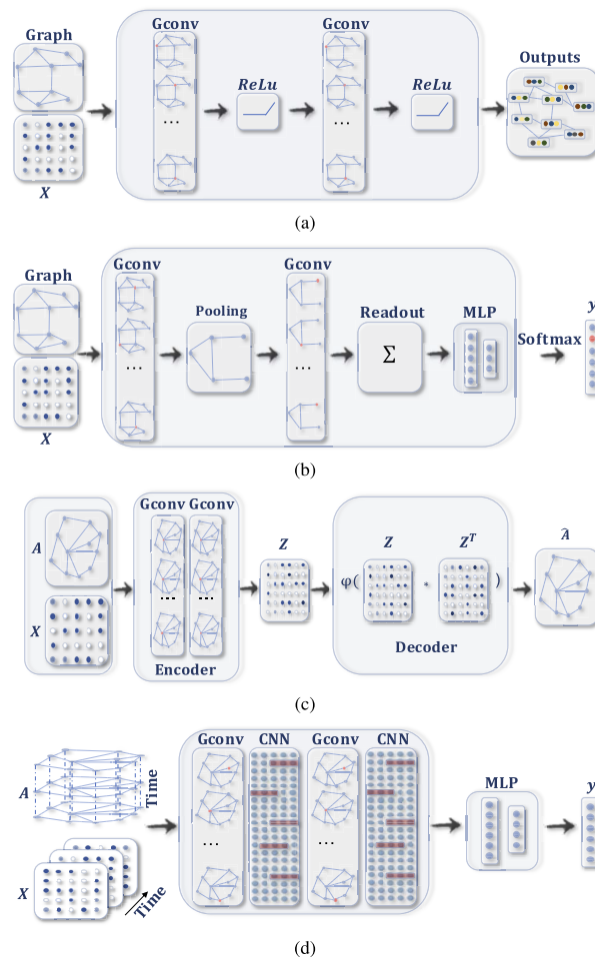


Figura 2.3.2 – Diferentes modelos GNN criados com camadas de convolução em grafo. Gconv representa uma camada de convolução em grafos, MLP um perceptron multi-camadas e CNN uma camada convolucional tradicional. **Fonte: (WU et al., 2020)**

A Figura 2.3.2 ilustra quatro diferentes modelos de GNN com camada de convolução em grafo, sendo eles respectivamente: (a) ConvGNN com múltiplas camadas de convolução em grafos, permitindo a agregação de informações de vizinhos mais distantes; (b) ConvGNN com camadas de pooling e readout para classificação de grafos, onde subgrafos representam níveis mais altos de abstração; (c) GAE (Autoencoder de grafos) para embedding de redes, que reconstrói a matriz de adjacência minimizando diferenças entre os dados reais e os preditos; (d) STGNN para previsão espaço-temporal, combinando convolução em grafos e CNN 1D para capturar dependências espaciais e temporais.

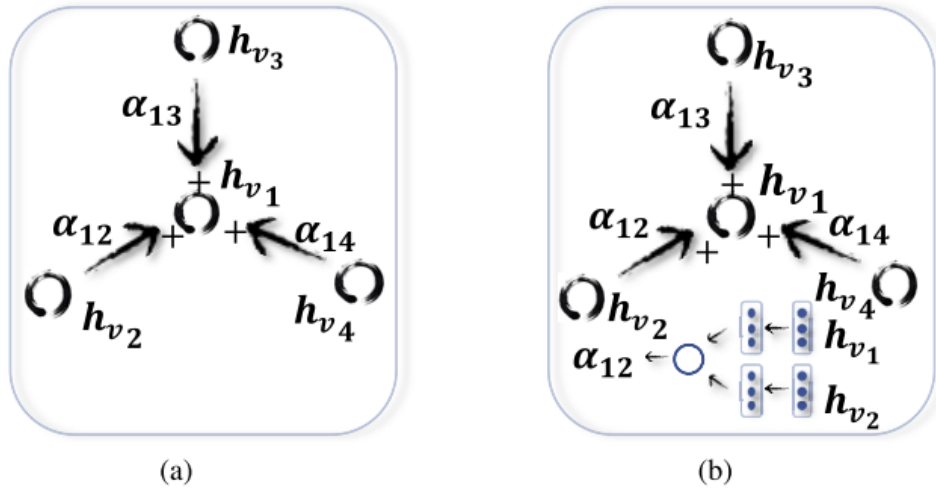


Figura 2.3.3 – Diferenças entre GCN e GAT: (a) O GCN atribui explicitamente um peso não paramétrico $a_{ij} = \frac{1}{\sqrt{\deg(v_i)\deg(v_j)}}$ ao vizinho v_j de v_i durante o processo de agregação. (b) O GAT captura implicitamente o peso a_{ij} por meio de uma arquitetura neural end-to-end, permitindo que nós mais importantes recebam pesos maiores. **Fonte: (WU et al., 2020)**

Por último temos a Figura 2.3.3 onde fica demonstrado a diferença entre uma GCN (*Graph Convolutional Network*) e uma GAT (*Graph Attention Network*).

2.3.1.1 GCN

As Graph Convolutional Networks (GCNs) aplicam uma operação de convolução sobre grafos, onde a informação de um nó é combinada com a de seus vizinhos em um processo iterativo. A convolução em GCNs utiliza uma matriz de adjacência normalizada para calcular os pesos entre os nós e seus vizinhos. Especificamente, um nó v_i é atualizado pela seguinte fórmula:

$$h_i^{(k+1)} = \sigma \left(\sum_{j \in \mathcal{N}(i) \cup \{i\}} \frac{1}{\sqrt{\deg(v_i)\deg(v_j)}} W^{(k)} h_j^{(k)} \right), \quad (2.3.1)$$

onde:

- $h_i^{(k)}$ é a representação do nó v_i na camada k ;
- $W^{(k)}$ são os pesos aprendidos pela rede;
- σ é uma função de ativação, como ReLU;
- $\deg(v_i)$ é o grau do nó v_i .

2.3.1.2 GAT

As Graph Attention Networks (GATs) introduzem um mecanismo de atenção, permitindo que os pesos atribuídos aos vizinhos sejam aprendidos dinamicamente durante o treinamento. Em vez de usar pesos fixos derivados da estrutura do grafo (como nas GCNs), as GATs calculam um peso de atenção a_{ij} para cada par de nós conectados. O peso é obtido através de uma função de similaridade:

$$a_{ij} = \text{softmax} \left(\text{LeakyReLU} \left(\vec{a}^T [Wh_i || Wh_j] \right) \right), \quad (2.3.2)$$

onde:

- $\vec{a}$ é um vetor de parâmetros aprendidos;
- W é a matriz de transformação linear;
- $||$ representa a concatenação de vetores;
- *softmax* normaliza os pesos para que somem 1.

Esse mecanismo permite que a rede foque nos vizinhos mais relevantes, ajustando automaticamente a importância de cada conexão. Como resultado, as GATs tendem a ser mais robustas em grafos com densidade irregular ou conexões ruidosas.

Uma vantagem clara das GATs é sua capacidade de capturar relações heterogêneas entre nós, onde a relevância dos vizinhos pode variar amplamente. No entanto, a maior complexidade computacional do mecanismo de atenção pode ser um desafio em grafos muito grandes.

2.3.2 Aplicações

As GNNs têm uma ampla gama de aplicações em diversos campos devido à sua capacidade de modelar relacionamentos complexos em dados estruturados como grafos:

- **Redes Sociais:** GNNs são utilizadas para prever a propagação de informações, detecção de comunidades e recomendação de amigos baseadas na estrutura das redes sociais (WU et al., 2020).

- **Biologia e Química:** GGNNs (uma variante de GNNs, significa Gated Graph Neural Networks e fazem uso do mecanismo GRU) como são usadas para prever interações entre proteínas, propriedades de moléculas e para a descoberta de novos medicamentos (GILMER et al., 2017).
- **Recomendação de Conteúdo:** Plataformas como Pinterest utilizam GNNs para melhorar a personalização de recomendações, aproveitando as interações e conexões entre usuários e itens (YING et al., 2018; WU et al., 2022).

No contexto educacional, as GNNs podem ser aplicadas em várias áreas para melhorar a qualidade do ensino e a gestão educacional:

- **Análise de Desempenho Estudantil:** GNNs podem ser usadas para prever o desempenho dos alunos com base em redes de interação entre estudantes, professores e recursos educacionais. Por exemplo, a interação entre estudantes em fóruns online e grupos de estudo pode ser modelada como um grafo, permitindo a análise de influência social no desempenho acadêmico (LI et al., 2022).
- **Recomendações Personalizadas de Recursos:** GNNs podem fornecer recomendações personalizadas de materiais de estudo, atividades e cursos com base nas interações e preferências dos alunos. Isso ajuda a criar um ambiente de aprendizado adaptativo que responde às necessidades individuais dos alunos (LIANG et al., 2023; WU et al., 2022).
- **Deteção de Anomalias em Redes Educacionais:** GNNs podem ser utilizadas para detectar comportamentos anômalos ou fraudulentos em redes educacionais, como plágio ou fraude em exames, através da análise de padrões de interação atípicos (AKOGLU; TONG; KOUTRA, 2015).

2.4 Trabalhos Relacionados

2.4.1 Predição de Links na Educação

No trabalho realizado por (SUSANTI et al., 2023), são abordados desafios enfrentados pelas instituições de ensino superior na Indonésia, como altas taxas de evasão e atrasos na graduação. Em resposta, foram realizadas pesquisas para começar a adotar a transformação digital para melhorar os processos educacionais e a tomada de decisão, o que resultou em um grande volume de dados educacionais. Com isso, há uma necessidade crescente de métodos avançados para analisar esses dados.

O artigo propõe o uso de GCNs para prever quais disciplinas optativas os estudantes provavelmente escolherão, com base em seu desempenho anterior. A escolha de GCNs pelo autor vem do fato que já existem diversos modelos de análise de dados baseados em modelos de aprendizado de máquina clássicos e também devido ao fato que as GCNs são capazes de modelar tanto as entidades (estudantes e cursos) quanto as relações entre elas, tornando-as eficazes para prever escolhas futuras.

Com 52 datasets de alunos como entrada, o modelo atingiu uma acurácia de 62,5% ao prever as escolhas de disciplinas optativas. Por fim, o autor concluiu que previsão de disciplinas optativas é essencial para personalizar o plano de estudo dos alunos e ele também sugere que com um melhor *finetuning* dos hiper-parâmetros e um maior número de datasets seria possível aumentar consideravelmente a acurácia do modelo.

2.4.2 Rastreamento de Conhecimento na Educação usando Grafos

O artigo de (NAKAGAWA; IWASAWA; MATSUO, 2019) propõe uma nova abordagem para Rastreamento de Conhecimento (*Knowledge Tracing*) utilizando GNNs. Com o objetivo de prever o desempenho dos alunos ao longo do tempo nos exercícios realizados durante o curso. Diferente dos modelos tradicionais como *Deep Knowledge Tracing* (DKT), que não consideravam a estrutura de grafo subjacente ao conhecimento, a abordagem proposta, chamada *Graph-Based Knowledge Tracing* (GKT), reformula o problema de rastreamento de conhecimento como uma tarefa de classificação de séries temporais em nível de nó, dentro de uma estrutura de grafo.

O modelo GKT estrutura o conhecimento como um grafo, onde os nós representam conceitos e as arestas representam as relações entre esses conceitos. A técnica utiliza GNNs para capturar a evolução do estado de conhecimento dos alunos ao longo do tempo, atualizando o estado de conhecimento de conceitos vizinhos sempre que um conceito é avaliado em um exercício.

Os autores também propõem métodos para implementar a estrutura de grafo quando esta não é fornecida explicitamente. Eles sugerem abordagens baseadas em estatísticas e aprendizado, como o uso de mecanismos de atenção (como das GATs) e *variational autoencoders* (VAE) para inferir as relações entre os conceitos.

Experimentos realizados com dois conjuntos de dados educacionais mostraram que o GKT superou os métodos anteriores em termos de precisão na previsão do desempenho dos alunos. Além disso, o modelo forneceu previsões mais interpretáveis, ajudando a visualizar como o estado de conhecimento dos alunos evolui ao longo do tempo de maneira mais clara e específica para cada conceito.

2.4.3 Aprendizagem Adaptativa utilizando GNNs e Aprendizado por Reforço

A aprendizagem adaptativa visa fornecer experiências de aprendizado personalizadas para atender às necessidades únicas de cada aluno. A personalização de caminhos de aprendizagem é um subcampo importante, focado em recomendar sequências de atividades educacionais para maximizar os resultados dos alunos. Recentemente, como visto em (VAS-SOYAN; VIE; LEMBERGER, 2023) a combinação de GNNs e Aprendizado por Reforço (RL) tem mostrado potencial significativo para abordar problemas complexos de personalização e otimização em diversos contextos.

A integração de GNNs com RL se destaca pela sua capacidade de modelar a estrutura relacional dos dados educacionais, otimizando de forma dinâmica as atividades recomendadas para cada aluno. As GNNs são eficazes em capturar as relações entre conceitos educacionais e o progresso do aluno ao longo do tempo, enquanto o aprendizado por reforço permite que o sistema adapte continuamente as recomendações com base nas interações do aluno e seus resultados.

Essa combinação mostrou ser promissora, mesmo em ambientes de dados pequenos, proporcionando uma solução escalável para sistemas educacionais que desejam aplicar inteligência adaptativa. Além disso, essa abordagem permite a generalização para uma variedade de tarefas, como controle de tráfego e otimização de redes, evidenciando a versatilidade e o potencial dos modelos de GNN com RL na personalização educacional.

2.4.4 Relação com este projeto

Os trabalhos mencionados acima demonstram a eficácia das GNNs em diferentes contextos educacionais, reforçando o papel dessas técnicas na análise e predição de dados complexos, como o comportamento estudantil e o desempenho acadêmico. Os estudos evidenciam as vantagens dos modelos gráficos em comparação com abordagens tradicionais, destacando a capacidade de capturar tanto as entidades quanto as relações entre elas, proporcionando *insights* mais detalhados e previsões mais precisas.

Neste projeto, busca-se explorar essas mesmas vantagens das GNNs, adaptando-as ao contexto do Censo Escolar do INEP 2023, um conjunto de dados com estrutura altamente relacional. Pretende-se aplicar as lições aprendidas nos estudos relacionados, como o uso de técnicas para estruturar dados como grafos e a implementação de GNNs para modelar essas relações, com o objetivo de aprimorar a análise educacional, detectar casos anômalos e fornecer *insights* mais robustos para a formulação de políticas educacionais.

3 Metodologia

Nesta seção, é descrito o processo utilizado para converter os dados do Censo Escolar Nacional em um formato adequado para análise por meio de uma Rede Neural de Grafos. Como mencionado previamente, este trabalho busca encontrar relações complexas e extrair informações relevantes do Censo Escolar Nacional. Contudo, como o mecanismo a ser utilizado para realizar essa análise e interpretação será uma Rede Neural de Grafos, é necessário, antes de tudo, que o alvo da pesquisa seja transformado em um grafo.

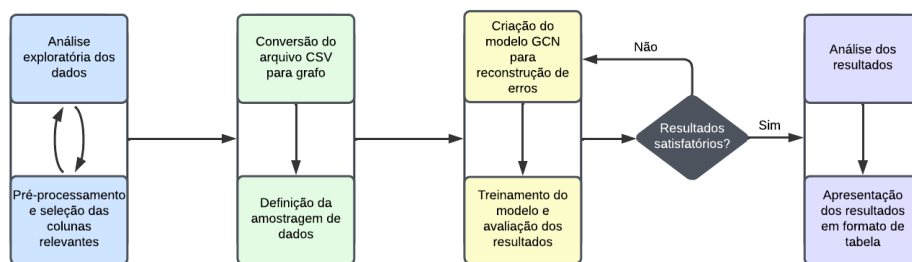


Figura 3.0.1 – Pipeline da metodologia do projeto

3.1 Etapa de Pré-Processamento, Limpeza e Seleção dos Dados

Antes de descrever as etapas de conversão do Censo Escolar, de um arquivo CSV para o formato de um grafo, é importante abordar sobre a estrutura do documento do Censo e quais as transformações realizadas nele. Para tal, foi utilizada a última versão publicamente disponível do CEN em sua modalidade de ensino de educação básica (Microdados do Censo da Educação Básica 2023), o qual possui 359 colunas-variáveis contendo informações de 217.625 entidades escolares distintas. Esses dados foram obtidos diretamente do site do INEP¹. É importante mencionar que, mesmo o arquivo tratando-se de um censo sobre a educação básica, este também abrange dados de ensino infantil, fundamental, e médio, EJA (Educação de Jovens e Adultos), e educação profissionalizante.

Dentre as 359 colunas disponíveis, muitas já haviam sido descontinuadas nesta nova versão de apresentação dos dados e, portanto, não foram preenchidas por grande parte das entidades. Posto isso, elas foram removidas do arquivo de forma a garantir que a análise seja a mais representativa possível com relação ao cenário real nacional. Além das colunas deprecadas, também havia uma quantidade excessiva de informações presentes para uma análise inicial e, portanto, viu-se necessário a realização de uma avaliação exploratória para verificar quais

¹ <https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados/censo-escolar>

as colunas apresentam maior importância para a extração e representação do conhecimento ali contido. Desta forma, os dados considerados mais relevantes para a análise foram:

- **Informações Geográficas:** Dados de alta relevância para uma análise de busca de padrões, ter conhecimento de qual região geográfica brasileira a entidade pertence, além de qual unidade federativa e se sua cidade é um território considerado rural ou urbano.
- **Dependência Administrativa:** Também corresponde a um dado vital, a fim de conhecer se a entidade estudada é uma escola de cunho público (Federal, Estadual ou Municipal) ou privado para fins de comparação justa.
- **Situação de Funcionamento:** Outra informação relevante é saber se a escola está em pleno funcionamento, paralisada ou se foi extinta nos últimos anos.
- **Alunos Matriculados:** É de sumo interesse saber se há alunos matriculados em cada entidade e quais os níveis escolares são ofertados pela escola, garantindo que as relações estabelecidas façam sentido.
- **Características Diversas:** É importante incluir diversas características sobre cada entidade para diferenciá-las, como informações sobre o cumprimento de necessidades básicas como água, luz e esgoto; informações sobre quais tipos de instalação física estão presentes no local, se há ou não exame de seleção, entre outras.
- **Colunas Adequadas:** É importante selecionar colunas que possam ser efetivamente processadas pelo modelo de GNN. Isso significa remover colunas com valores numéricos de alta variação, como a quantidade de alunos por turma, que podem introduzir ruído e dificultar a convergência do modelo.

Após a seleção dessas colunas, o arquivo foi carregado em seu formato CSV no site da plataforma Google Colab², onde por meio da biblioteca Pandas³, foram realizadas diversas tarefas de pré-processamento necessárias. Dentre essas tarefas, destacamos: a verificação dos dados de forma a encontrar valores duplicados, faltantes ou não numéricos(VANDERPLAS, 2016). Também foi realizada a consolidação de colunas, como foi o caso da criação da coluna "NECESSIDADES_BASICAS". A qual foi criada com a junção das colunas: "IN_AGUA.INEXISTENTE", "IN_ENERGIA.INEXISTENTE", "IN_ESGOTO.INEXISTENTE" e "IN_BANHEIRO". A nova coluna recebe valor 1 caso as colunas com sufixo inexistente possuam valor 0 e a coluna "IN_BANHEIRO" possua valor 1. Caso contrário, a nova coluna terá o valor 0. Isso foi feito para reduzir a complexidade na criação do grafo sem perda de informação.

² <https://colab.google/>

³ <https://pandas.pydata.org/>

Além disso, foi realizado o preenchimento ou remoção das linhas encontradas previamente dependendo de sua relação com outras colunas. Por exemplo, no caso da coluna "TP_SITUACAO_FUNCIONAMENTO" possuir valores iguais a 2, 3 ou 4 (indicando que a entidade está paralisada ou foi extinta no ano do censo ou anterior a ele), as linhas correspondentes foram removidas, pois o estudo não encontraria informações relevantes em entidades que não estão mais em funcionamento.

Outro ponto importante no pré-processamento foi o tratamento da coluna 'UF', que contém 27 valores distintos, correspondentes às Unidades Federativas do Brasil. Como essa coluna poderia gerar ruído no modelo devido à grande variedade de categorias, foi aplicada uma codificação (encoding) para transformá-la em uma representação numérica apropriada. No entanto, essa abordagem aumentou significativamente o tempo de execução do modelo, sendo, portanto, excluída na versão final do modelo com sobreajuste devido ao custo computacional. Essa decisão foi tomada a fim de otimizar o desempenho, sem comprometer drasticamente a qualidade dos resultados.

3.2 Criação do Grafo

Após a sanitização dos dados, pôde-se iniciar a modelagem do problema em termos de um grafo, definindo assim os nós, suas características e as arestas que estarão presentes no grafo-base. Mais precisamente, modelos dos dados no grafo a partir das seguintes representações de instâncias:

- **Nós:** Para os nós, foi utilizada a coluna "CO_ENTIDADE", que contém o valor do código único daquela entidade escolar definido pelo próprio INEP.
- **Características:** Para as características dos nós, incluiu-se uma ampla variedade de colunas, abrangendo informações sobre o cumprimento de necessidades básicas, tipos de instalação física presentes no local, a presença de exame de seleção, refeição para os alunos, equipamentos técnicos, presença de profissionais extras como de serviço geral e saúde, existência de material escolar, acesso à internet, projetores multimídia, entre outras informações relevantes. Essas características permitem mapear as inter-relações entre elas, bem como a localização da entidade e os níveis de escolarização oferecidos pela escola (infantil, fundamental, médio, profissionalizante ou EJA), dentre outras possíveis conexões.
- **Arestas:** Por fim, para as arestas, foi utilizado um sistema de pesagem onde nós que possuem pelo menos 3 características iguais entre as colunas referentes à localização geográfica, unidade federativa (UF), localidade (rural ou urbana) e dependência administrativa são conectados. O objetivo foi ligar nós de escolas com características geográficas e administrativas similares, mantendo comparações justas.

3.3 Criação do Modelo

Após o termino da criação do grafo usando o NetworkX, foi utilizada a biblioteca PyTorchGeometric⁵ para a criação e manipulação da GNN. Para tal propósito, o grafo NetworkX foi convertido em PyTorch Geometric Data. O modelo utilizado para analisar os dados do grafo foi um GCN (Graph Convolutional Network) (KIPF; WELLING, 2017), com o intuito de realizar a detecção de anomalias (AN; CHO, 2015) nos dados ali presentes.

A detecção de anomalias ocorrerá da seguinte forma:

1. Primeiramente, as características dos nós são carregadas e armazenadas em um tensor, que será utilizado como entrada para o modelo de detecção de anomalias.
2. Em seguida, o modelo é construído utilizando uma Graph Convolutional Network (GCN) com três camadas convolucionais. A primeira camada recebe as características dos nós e as transforma em 26 características intermediárias (igual ao total de características presentes por nó). Em seguida, uma segunda camada reduz para 13 características, enquanto a terceira camada reconstrói as características de volta para o número original de características por nó. Cada camada convolucional é seguida por uma função de ativação ReLU.
3. O modelo foi treinado por 50 épocas, garantindo que o processo de ajuste dos pesos do modelo fosse suficientemente longo para capturar as características relevantes de cada nó sem causar sobreajuste.
4. O limiar para definir se um nó é anômalo é calculado como a média dos erros de reconstrução somada a duas vezes o desvio padrão. Este critério permite identificar nós cujas características foram mal reconstruídas, sinalizando a presença de anomalias.
5. A função de perda que foi utilizada é a MSE (Mean Squared Error Loss), que mede a diferença entre as características reconstruídas e as características reais dos nós. O objetivo é minimizar essa diferença, garantindo que as características dos nós sejam reconstruídas de forma fiel, exceto em casos de anomalias.
6. O Treinamento é realizado utilizando o otimizador Adam (KINGMA; BA, 2017) para ajustar os pesos do modelo. Os ajustes são feitos de forma a minimizar a perda de reconstrução. A cada época, a perda é recalculada, e os gradientes são atualizados para melhorar a precisão do modelo.

Abaixo temos uma figura representando os erros detectados em escala logarítmica (para melhor representação)

⁵ <https://pytorch-geometric.readthedocs.io/en/latest/>

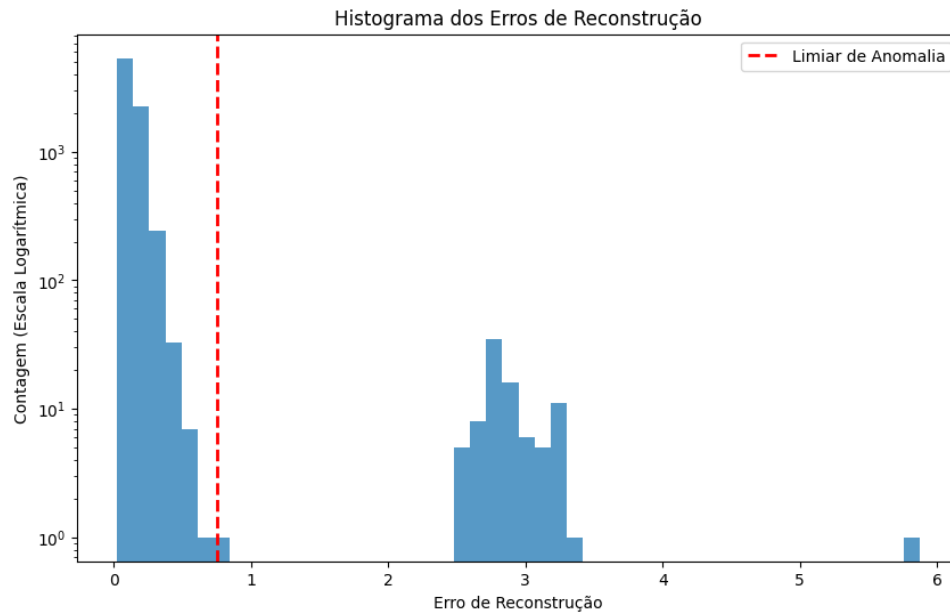


Figura 3.3.1 – Histograma dos Erros de Reconstrução

Após o treinamento, o modelo reconstrói as características dos nós. O erro de reconstrução é calculado para cada nó. Nós com erros acima de um limiar (calculado como a média do erro mais duas vezes o desvio padrão) são considerados anômalos.

3.3.1 Criação do Modelo Sobreajustado

Além desse primeiro modelo, será criado um segundo modelo treinado com dados sobreajustados. Para atingir esse objetivo, o modelo será treinado por 200 épocas (o quádruplo do número de épocas do primeiro modelo), e as camadas convolucionais terão o dobro de unidades sem as reduzir na segunda camada, utilizando 52 unidades em vez de 26 e 13. Essa abordagem visa permitir que o modelo capture ainda mais padrões nos dados, possibilitando uma análise mais aprofundada das anomalias.

4 Resultados

Nesta seção, serão apresentados os resultados obtidos com a implementação do modelo descrito na seção anterior. O modelo foi aplicado ao grafo global e aos subgrafos regionais, correspondentes a cada região geográfica brasileira.

Primeiramente, os resultados da aplicação no grafo global com um total de 8000 nós escolhidos aleatoriamente são apresentados a seguir:

4.1 Resultados dos Dados Originais

4.1.1 Resultados do Grafo Global

Tabela 1 – Comparação das Características dos Nós - Global

Características	Não Anômalo (Moda)	Anômalo (Moda)	Não Anômalo (Mediana)	Anômalo (Mediana)
Região	Nordeste (2)	Sudeste (3)	Sudeste (3)	Sul (4)
UF	São Paulo (35)	Santa Catarina (42)	Minas Gerais (31)	Santa Catarina (42)
Dependência	Municipal (3)	Municipal (3)	Municipal (3)	Municipal (3)
Localização	Urbana (1)	Rural (2)	Urbana (1)	Rural (2)
Necessidades Básicas	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Tratamento de Lixo	Não (0)	Não (0)	Não (0)	Não (0)
Laboratório Ciências	Não (0)	Não (0)	Não (0)	Não (0)
Laboratório Informática	Não (0)	Não (0)	Não (0)	Não (0)
Quadra Esportes	Não (0)	Não (0)	Não (0)	Não (0)
Refeitório	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Acessibilidade	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Equipamento Multimídia	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Internet	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Profissional Serviços Gerais	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Profissional Saúde	Não (0)	Não (0)	Não (0)	Não (0)
Alimentação	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Material Pedagógico	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Exame de Seleção	Não (0)	Não (0)	Não (0)	Não (0)
Regular	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Diurno	Sim (1)	Sim (1)	Sim (1)	Sim (1)
EAD	Não (0)	Não (0)	Não (0)	Não (0)
Alunos Matriculados	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Infantil	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Fundamental	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Médio	Não (0)	Não (0)	Não (0)	Não (0)
Ensino Profissionalizante	Não (0)	Não (0)	Não (0)	Não (0)
Educação de Jovens e Adultos	Não (0)	Não (0)	Não (0)	Não (0)

Analisando esses dados, podemos observar que a maior parte dos nós parecem vir da região Sudeste, com a UF mais comum sendo São Paulo, a dependência administrativa do tipo municipal e a localização geográfica urbana. Também podemos observar que quase todas as entidades analisadas cumprem as necessidades básicas, mas não realizam tratamento de lixo e não possuem instalações adicionais como laboratórios, embora possuem refeitório.

Em geral, a maioria das escolas oferece o mínimo de acessibilidade, tem acesso a projetor, internet, profissionais de serviço geral e alimentação. Como esperado, a maioria das

escolas oferece um ensino regular no período diurno, sem exame de seleção para ingresso e sem modalidade EAD. As escolas filtradas possuem alunos regularmente matriculados, majoritariamente no Ensino Infantil ou Fundamental.

Os dados revelam uma grande concentração de escolas na região Sudeste, especialmente no estado de São Paulo, com administração municipal e localização urbana. Isso vai ao encontro da distribuição populacional brasileira, pois, de acordo com o IBGE (2017), a densidade demográfica do Brasil revela que a região Sudeste contém 41% da população do País, sendo seguida pela região Nordeste (28%), Sul (14%), Norte (9%) e Centro-Oeste (8%).

Embora a maioria das escolas apresentem condições decentes para os alunos, muitas características benéficas ao desenvolvimento dos alunos ainda não são comuns. A ausência de Ensino Médio por grande parte dos dados é estranha, por um lado pode ser causada pela maior amostragem de somente 8000 nós, por outro lado pode ser consequência do fato que essa edição do Censo abrangeu uma quantidade muito menor de escolas graças a pandemia.

Outro fato interessante é a ausência de escolas na modalidade EAD. Após a ocorrência da pandemia, era de se esperar que pelo menos uma parte optasse por manter tal modalidade de formação, o que não foi o caso observado nos dados analisados.

Em seguida, serão apresentados os resultados regionais começando pela região Norte.

4.1.2 Resultados da Região Norte

Tabela 2 – Comparação das Características dos Nós - Região Norte

Características	Não Anômalo (Moda)	Não Anômalo (Mediana)
Região	Norte (1)	Norte (1)
UF	Pará (15)	Pará (15)
Dependência	Municipal (3)	Municipal (3)
Localização	Rural (2)	Rural (2)
Necessidades Básicas	Sim (1)	Sim (1)
Tratamento de Lixo	Não (0)	Não (0)
Laboratório Ciências	Não (0)	Não (0)
Laboratório Informática	Não (0)	Não (0)
Quadra Esportes	Não (0)	Não (0)
Refeitório	Não (0)	Não (0)
Acessibilidade	Sim (1)	Sim (1)
Equipamento Multimídia	Não (0)	Não (0)
Internet	Sim (1)	Sim (1)
Profissional Serviços Gerais	Sim (1)	Sim (1)
Profissional Saúde	Não (0)	Não (0)
Alimentação	Sim (1)	Sim (1)
Material Pedagógico	Sim (1)	Sim (1)
Exame de Seleção	Não (0)	Não (0)
Regular	Sim (1)	Sim (1)
Diurno	Sim (1)	Sim (1)
EAD	Não (0)	Não (0)
Alunos Matriculados	Sim (1)	Sim (1)
Ensino Infantil	Sim (1)	Sim (1)
Ensino Fundamental	Sim (1)	Sim (1)
Ensino Médio	Não (0)	Não (0)
Ensino Profissionalizante	Não (0)	Não (0)
Educação de Jovens e Adultos	Não (0)	Não (0)

Essa região foi um caso à parte, sendo a única região que, independentemente da seed ou amostragem utilizada na geração do grafo, nunca apresentou anomalias. Isso pode indicar um viés no modelo ou no grafo gerado, algo que pode ser melhorado no futuro. Também pode sugerir que a região Norte é mais homogênea que as outras regiões do país, sem grande anomalias positivas ou negativas.

Analisando os dados, é visível que o Pará é o estado com mais escolas, que são primariamente rurais. Comparadas aos dados globais, existe a ausência de refeitórios e equipamentos multimídia. Indicando que a região é geralmente mais carente de recursos comparada às outras.

4.1.3 Resultados da Região Nordeste

Tabela 3 – Comparação das Características dos Nós - Região Nordeste

Características	Não Anômalo (Moda)	Anômalo (Moda)	Não Anômalo (Mediana)	Anômalo (Mediana)
Região	Nordeste (2)	Nordeste (2)	Nordeste (2)	Nordeste (2)
UF	Bahia (29)	Bahia (29)	Paraíba (25)	Pernambuco (26)
Dependência	Municipal (3)	Estadual (2)	Municipal (3)	Estadual (2)
Localização	Urbana (1)	Rural (2)	Urbana (1)	Rural (2)
Necessidades Básicas	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Tratamento de Lixo	Não (0)	Não (0)	Não (0)	Não (0)
Laboratório Ciências	Não (0)	Não (0)	Não (0)	Não (0)
Laboratório Informática	Não (0)	Não (0)	Não (0)	Não (0)
Quadra Esportes	Não (0)	Não (0)	Não (0)	Não (0)
Refeitório	Não (0)	Não (0)	Não (0)	Não (0)
Acessibilidade	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Equipamento Multimídia	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Internet	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Profissional Serviços Gerais	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Profissional Saúde	Não (0)	Não (0)	Não (0)	Não (0)
Alimentação	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Material Pedagógico	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Exame de Seleção	Não (0)	Não (0)	Não (0)	Não (0)
Regular	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Diurno	Sim (1)	Sim (1)	Sim (1)	Sim (1)
EAD	Não (0)	Não (0)	Não (0)	Não (0)
Alunos Matriculados	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Infantil	Sim (1)	Não (0)	Sim (1)	Não (0)
Ensino Fundamental	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Médio	Não (0)	Sim (1)	Não (0)	Sim (1)
Ensino Profissionalizante	Não (0)	Não (0)	Não (0)	Não (0)
Educação de Jovens e Adultos	Não (0)	Não (0)	Não (0)	Não (0)

A região Nordeste seguiu bem o padrão global, com algumas mudanças notáveis. Primeiramente, os nós anômalos tendem a ter como dependência o estado, enquanto os nós não anômalos são predominantemente de dependência municipal. Além disso, a ausência de refeitórios foi uma característica comum tanto em nós anômalos quanto em nós não anômalos, destacando um ponto crítico relacionado à infraestrutura escolar quando comparada ao padrão global.

Outro aspecto significativo foi a discrepância na oferta de ensino. Enquanto os nós anômalos tendem a ofertar Ensino Fundamental e Médio, há uma ausência marcante de Ensino Infantil entre os nós anômalos. Isso contrasta com o padrão global, onde o Ensino Infantil está presente em ambos os nós anômalos e não anômalos. Essa diferença pode indicar que as escolas anômalas abriram mão do Ensino Infantil, que já está presente na maior parte das escolas, para enfatizarem o Ensino Fundamental e Médio.

4.1.4 Resultados da Região Sudeste

Tabela 4 – Comparação das Características dos Nós - Região Sudeste

Características	Não Anômalo (Moda)	Anômalo (Moda)	Não Anômalo (Mediana)	Anômalo (Mediana)
Região	Sudeste (3)	Sudeste (3)	Sudeste (3)	Sudeste (3)
UF	São Paulo (35)	Rio de Janeiro (33)	São Paulo (35)	Rio de Janeiro (33)
Dependência	Municipal (3)	Municipal (3)	Municipal (3)	Municipal (3)
Localização	Urbana (1)	Rural (2)	Urbana (1)	Rural (2)
Necessidades Básicas	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Tratamento de Lixo	Não (0)	Não (0)	Não (0)	Não (0)
Laboratório Ciências	Não (0)	Não (0)	Não (0)	Não (0)
Laboratório Informática	Não (0)	Não (0)	Não (0)	Não (0)
Quadra Esportes	Sim (1)	Não (0)	Sim (1)	Não (0)
Refeitório	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Acessibilidade	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Equipamento Multimídia	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Internet	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Profissional Serviços Gerais	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Profissional Saúde	Não (0)	Não (0)	Não (0)	Não (0)
Alimentação	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Material Pedagógico	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Exame de Seleção	Não (0)	Não (0)	Não (0)	Não (0)
Regular	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Diurno	Sim (1)	Sim (1)	Sim (1)	Sim (1)
EAD	Não (0)	Não (0)	Não (0)	Não (0)
Alunos Matriculados	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Infantil	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Fundamental	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Médio	Não (0)	Não (0)	Não (0)	Não (0)
Ensino Profissionalizante	Não (0)	Não (0)	Não (0)	Não (0)
Educação de Jovens e Adultos	Não (0)	Não (0)	Não (0)	Não (0)

A região Sudeste seguiu o padrão global, como esperado, com São Paulo e Rio de Janeiro aparecendo como as principais UF's, e todas as escolas analisadas sob dependência municipal. Entretanto, algumas diferenças importantes foram observadas entre os nós anômalos e não anômalos.

Um ponto de destaque é a presença de quadras de esportes nas escolas não anômalas, enquanto as escolas anômalas não possuem essa infraestrutura. Esse dado pode indicar uma correlação entre a ausência de instalações esportivas e a classificação das escolas como anômalas, possivelmente refletindo uma menor qualidade de infraestrutura geral.

Outro aspecto relevante é a localização. Enquanto as escolas não anômalas estão situadas em áreas predominantemente urbanas, uma proporção significativa das escolas anômalas está em áreas rurais, corroborando a ideia inicial de que as escolas anômalas dessa região podem ser escolas com menos recursos.

No entanto, em termos de oferta educacional e acesso a recursos como internet e equipamentos multimídia, tanto as escolas anômalas quanto as não anômalas apresentaram padrões semelhantes, indicando que talvez a causa das anomalias não seja a falta de recursos, e sim algum outro fator que não foi perceptível com as características utilizadas para treinar o modelo nesta versão.

4.1.5 Resultados da Região Sul

Tabela 5 – Comparação das Características dos Nós - Região Sul

Características	Não Anômalo (Moda)	Anômalo (Moda)	Não Anômalo (Mediana)	Anômalo (Mediana)
Região	Sul (4)	Sul (4)	Sul (4)	Sul (4)
UF	Paraná (41)	Santa Catarina (42)	Santa Catarina (42)	Santa Catarina (42)
Dependência	Municipal (3)	Estadual (2)	Municipal (3)	Estadual (2)
Localização	Urbana (1)	Rural (2)	Urbana (1)	Rural (2)
Necessidades Básicas	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Tratamento de Lixo	Não (0)	Não (0)	Não (0)	Não (0)
Laboratório Ciências	Não (0)	Não (0)	Não (0)	Não (0)
Laboratório Informática	Não (0)	Não (0)	Não (0)	Não (0)
Quadra Esportes	Não (0)	Sim (1)	Não (0)	Sim (1)
Refeitório	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Acessibilidade	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Equipamento Multimídia	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Internet	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Profissional Serviços Gerais	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Profissional Saúde	Não (0)	Não (0)	Não (0)	Não (0)
Alimentação	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Material Pedagógico	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Exame de Seleção	Não (0)	Não (0)	Não (0)	Não (0)
Regular	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Diurno	Sim (1)	Sim (1)	Sim (1)	Sim (1)
EAD	Não (0)	Não (0)	Não (0)	Não (0)
Alunos Matriculados	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Infantil	Sim (1)	Não (0)	Sim (1)	Não (0)
Ensino Fundamental	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Médio	Não (0)	Não (0)	Não (0)	Não (0)
Ensino Profissionalizante	Não (0)	Não (0)	Não (0)	Não (0)
Educação de Jovens e Adultos	Não (0)	Não (0)	Não (0)	Não (0)

A região Sul tem como principal UF Santa Catarina, seguida pelo Paraná. A dependência das escolas é majoritariamente municipal para os casos não anômalos e estadual para os anômalos, o que pode indicar uma diferença de gestão que impacta a ocorrência de anomalias. As escolas não anômalas estão situadas em áreas urbanas, enquanto as anômalas estão em áreas predominantemente rurais, o que pode sugerir que a localização geográfica influencia na qualidade dos recursos e serviços oferecidos.

Um ponto interessante na região Sul é que, apesar de as escolas anômalas estarem em áreas rurais, são as únicas que possuem quadras de esporte, contrastando com o padrão esperado, onde escolas em localidades urbanas tendem a ter mais recursos que escolas rurais.

Outro dado relevante é a ausência do Ensino Infantil em escolas anômalas, enquanto ele está presente nas não anômalas. A oferta do Ensino Fundamental é comum em ambas as categorias, o que demonstra uma consistência na oferta educacional básica. Porém, diferentemente do Nordeste, onde essa falta foi compensada com a adição do Ensino Médio, aqui não houve nenhum acréscimo positivo para compensar sua ausência.

No geral, a região Sul segue o padrão global em termos de infraestrutura e acesso a recursos como internet, alimentação e materiais pedagógicos. As diferenças mais notáveis estão na dependência administrativa das escolas, na localização e na carência de Ensino Infantil.

4.1.6 Resultados da Região Centro-Oeste

Tabela 6 – Comparação das Características dos Nós - Região Centro-Oeste

Características	Não Anômalo (Moda)	Anômalo (Moda)	Não Anômalo (Mediana)	Anômalo (Mediana)
Região	Centro-Oeste (5)	Centro-Oeste (5)	Centro-Oeste (5)	Centro-Oeste (5)
UF	Goiás (52)	Distrito Federal (53)	Goiás (52)	Goiás (52)
Dependência	Municipal (3)	Municipal (3)	Municipal (3)	Municipal (3)
Localização	Urbana (1)	Rural (2)	Urbana (1)	Rural (2)
Necessidades Básicas	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Tratamento de Lixo	Não (0)	Não (0)	Não (0)	Não (0)
Laboratório Ciências	Não (0)	Não (0)	Não (0)	Não (0)
Laboratório Informática	Não (0)	Não (0)	Não (0)	Não (0)
Quadra Esportes	Sim (1)	Não (0)	Sim (1)	Sim (1)
Refeitório	Sim (1)	Não (0)	Sim (1)	Não (0)
Acessibilidade	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Equipamento Multimídia	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Internet	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Profissional Serviços Gerais	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Profissional Saúde	Não (0)	Não (0)	Não (0)	Não (0)
Alimentação	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Material Pedagógico	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Exame de Seleção	Não (0)	Não (0)	Não (0)	Não (0)
Regular	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Diurno	Sim (1)	Sim (1)	Sim (1)	Sim (1)
EAD	Não (0)	Não (0)	Não (0)	Não (0)
Alunos Matriculados	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Infantil	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Fundamental	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Médio	Não (0)	Não (0)	Não (0)	Não (0)
Ensino Profissionalizante	Não (0)	Não (0)	Não (0)	Não (0)
Educação de Jovens e Adultos	Não (0)	Não (0)	Não (0)	Não (0)

A região Centro-Oeste se destaca por apresentar a menor quantidade de dados no Censo 2023, o que já gera um ponto de atenção, sugerindo uma possível sub-representação ou falhas na coleta de dados nessa área. A UF predominante é Goiás, com alguns casos anômalos ocorrendo no Distrito Federal, o que pode refletir uma disparidade regional.

Em termos de dependência administrativa, todas as escolas analisadas, tanto anômalas quanto não anômalas, são de dependência municipal, conforme o padrão global. As localizações seguem o observado em outras regiões: as escolas não anômalas estão majoritariamente em áreas urbanas, enquanto as escolas anômalas estão situadas em áreas rurais, indicando que a localização geográfica pode influenciar a identificação de anomalias.

Um dado interessante é a ausência de quadras de esportes nas escolas anômalas, enquanto essa infraestrutura está presente nas escolas não anômalas. Além disso, a ausência de refeitórios nas escolas anômalas é notável, destacando uma carência em aspectos relacionados à alimentação escolar e à infraestrutura esportiva.

Por outro lado, características como o acesso à internet, equipamentos multimídia e materiais pedagógicos estão presentes tanto nas escolas anômalas quanto nas não anômalas, em conformidade com o padrão global.

No geral, a região Centro-Oeste segue o padrão global em termos de recursos educacionais, mas apresenta diferenças marcantes na infraestrutura das escolas anômalas, princi-

palmente na falta de quadras de esportes e refeitórios, fatores que podem ser cruciais para a qualidade de vida dos alunos.

4.2 Resultados com Dados com Sobreajuste

Após realizar o treino normal do modelo, foi investigada a viabilidade de treinar o modelo com dados propositalmente sobreajustados. A intenção desta abordagem é aprimorar o processo de detecção de anomalias, verificando se o excesso de informações para um modelo desse tipo seria prejudicial ou benéfico.

O sobreajuste, geralmente visto como um problema a ser solucionado para a criação de modelos de aprendizado de máquina, pode ser utilizado de forma estratégica para lidar com dados onde buscamos justamente ter certeza que o modelo acerte para aquele conjunto de dados. A ideia central é que, ao treinar o modelo para sobreajustar-se a uma versão com excesso de dados, ele se torna excessivamente sensível a pequenas variações, que podem destacar de forma mais clara as anomalias nos dados. Essa abordagem nos permite testar se a detecção de anomalias pode ser aprimorada através do treinamento de um modelo com sobreajuste.

4.2.1 Resultados do Grafo Global (Sobreajuste)

Tabela 7 – Comparação das Características dos Nós - Global (Sobreajuste)

Características	Não Anômalo (Moda)	Anômalo (Moda)	Não Anômalo (Mediana)	Anômalo (Mediana)
Região	Sudeste (3)	Nordeste (2)	Sudeste (3)	Nordeste (2)
Dependência	Municipal (3)	Estadual (2)	Municipal (3)	Municipal (3)
Localização	Urbana (1)	Urbana (1)	Urbana (1)	Urbana (1)
Necessidades Básicas	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Tratamento de Lixo	Não (0)	Não Informado (9)	Não (0)	Não Informado (9)
Laboratório Ciências	Não (0)	Não (0)	Não (0)	Não (0)
Laboratório Informática	Não (0)	Não (0)	Não (0)	Não (0)
Quadra Esportes	Não (0)	Não (0)	Não (0)	Não (0)
Refeitório	Sim (1)	Não (0)	Sim (1)	Não (0)
Acessibilidade	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Equipamento Multimídia	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Internet	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Profissional Serviços Gerais	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Profissional Saúde	Não (0)	Não (0)	Não (0)	Não (0)
Alimentação	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Material Pedagógico	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Exame de Seleção	Não (0)	Não (0)	Não (0)	Não (0)
Regular	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Diurno	Sim (1)	Sim (1)	Sim (1)	Sim (1)
EAD	Não (0)	Não (0)	Não (0)	Não (0)
Alunos Matriculados	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Infantil	Sim (1)	Não (0)	Sim (1)	Não (0)
Ensino Fundamental	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Médio	Não (0)	Não (0)	Não (0)	Não (0)
Ensino Profissionalizante	Não (0)	Não (0)	Não (0)	Não (0)
Educação de Jovens e Adultos	Não (0)	Não (0)	Não (0)	Não (0)

Analisando esses novos dados vemos algumas diferenças bastante interessantes, a primeira que chama atenção é o fato que a região Sudeste se manteve como a mais comum para os dados não anômalos, mas a região Nordeste se tornou a região mais comum entre os dados anômalos. Isso está de acordo com o que foi testado ao colocarmos menos características nos nós, e será comentado mais a seguir na seção "Outras Análises".

As dependências administrativas se mantêm bem similares às originais com a única diferença sendo que para a moda dos dados anômalos temos a dependência do tipo estadual e não municipal. A localização se tornou majoritariamente urbana e grande parte das escolas anômalas nem mesmo informaram sobre a sua situação de tratamento de lixo.

Necessidades básicas, laboratórios e quadras se mantiveram com o valor 0. Já refeitórios agora estão ausentes dos casos anômalos em grande parte. De resto o último dado com uma divergência de interesse é o ensino infantil, que está ausente em muitos dos casos anômalos, o que é uma estatística preocupante. Pois a ausência do ensino infantil pode vir a afetar as novas gerações desses locais anômalos, que vão ter de procurar escolas em outros locais.

4.2.2 Resultados da Região Norte (Sobreajuste)

Tabela 8 – Comparação das Características dos Nós - Região Norte (Sobreajuste)

Características	Não Anômalo (Moda)	Anômalo (Moda)	Não Anômalo (Mediana)	Anômalo (Mediana)
Região	Norte (1)	Norte (1)	Norte (1)	Norte (1)
Dependência	Municipal (3)	Estadual (2)	Municipal (3)	Municipal (3)
Localização	Rural (2)	Rural (2)	Rural (2)	Rural (2)
Necessidades Básicas	Sim (1)	Não (0)	Sim (1)	Não (0)
Tratamento de Lixo	Não (0)	Não Informado (9)	Não (0)	Não Informado (9)
Laboratório Ciências	Não (0)	Não (0)	Não (0)	Não (0)
Laboratório Informática	Não (0)	Não (0)	Não (0)	Não (0)
Quadra Esportes	Não (0)	Não (0)	Não (0)	Não (0)
Refeitório	Não (0)	Não (0)	Não (0)	Não (0)
Acessibilidade	Sim (1)	Não (0)	Sim (1)	Não (0)
Equipamento Multimídia	Não (0)	Não (0)	Não (0)	Não (0)
Internet	Sim (1)	Não (0)	Sim (1)	Não (0)
Profissional Serviços Gerais	Sim (1)	Não (0)	Sim (1)	Não (0)
Profissional Saúde	Não (0)	Não (0)	Não (0)	Não (0)
Alimentação	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Material Pedagógico	Sim (1)	Não (0)	Sim (1)	Não (0)
Exame de Seleção	Não (0)	Não (0)	Não (0)	Não (0)
Regular	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Diurno	Sim (1)	Sim (1)	Sim (1)	Sim (1)
EAD	Não (0)	Não (0)	Não (0)	Não (0)
Alunos Matriculados	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Infantil	Sim (1)	Não (0)	Sim (1)	Não (0)
Ensino Fundamental	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Médio	Não (0)	Não (0)	Não (0)	Não (0)
Ensino Profissionalizante	Não (0)	Não (0)	Não (0)	Não (0)
Educação de Jovens e Adultos	Não (0)	Não (0)	Não (0)	Não (0)

A região norte foi a região que sofreu a maior quantidade de mudanças com o uso de dados com sobreajuste. Principalmente porque agora foram encontrados casos anômalos, o que é um bom sinal da eficácia dessa abordagem distinta. Além disso, sobre os dados de

dependência administrativa temos agora que a moda dos casos anômalos é Estadual, o que difere de todos os outros casos nessa região.

Uma das diferenças mais interessantes entre os nós anômalos e não anômalos é a presença de necessidades básicas atendidas. Enquanto as escolas não anômalas têm suas necessidades garantidas, as anômalas apresentam lacunas significativas, indicando possíveis dificuldades logísticas ou falta de investimentos em certas áreas remotas da região (considerando que são predominantemente rurais). Além disso, os casos anômalos mostram que muitas escolas nem sequer informaram o tratamento de lixo, um indicativo de falhas na coleta de dados ou na própria infraestrutura para esse tipo de serviço.

Acessibilidade também é um ponto crítico. Enquanto as escolas não anômalas mostram o mínimo de acessibilidade, os nós anômalos indicam uma ausência total desse recurso, o que é preocupante, pois escolas em áreas rurais já enfrentam dificuldades de acesso físico, e a falta de infraestrutura acessível agrava o problema para alunos com deficiência.

Outro ponto que merece ser comentado é a falta de internet nas escolas anômalas. A conectividade, essencial para o ensino moderno, é quase inexistente nos nós anômalos, contrastando com as escolas não anômalas, onde o acesso à internet está presente. A falta de conectividade limita não apenas o acesso a recursos pedagógicos online, mas também a comunicação com órgãos administrativos e a modernização dos processos educacionais, o que pode explicar a falta de alguns dados no censo escolar.

Outro dado interessante é que, embora a alimentação seja oferecida tanto nas escolas anômalas quanto nas não anômalas, as anômalas não fornecem material pedagógico adequadamente. Isso pode indicar problemas na distribuição de recursos ou na gestão educacional dessas instituições.

Por fim, a ausência do Ensino Infantil nas escolas anômalas chama atenção. Embora o Ensino Fundamental seja oferecido amplamente em todas as escolas, a falta do ensino voltado para a primeira infância nas escolas anômalas pode refletir disparidades no atendimento educacional básico.

A análise da Região Norte revela uma das maiores disparidade entre os nós anômalos e os nós não anômalos, visto que muitas características essenciais que são cumpridas nos casos usuais não são pelos casos anômalos. Em geral, a região parece necessitar de mais recursos ou políticas públicas e possivelmente uma investigação mais a fundo.

4.2.3 Resultados da Região Nordeste (Sobreajuste)

Tabela 9 – Comparação das Características dos Nós - Região Nordeste (Sobreajuste)

Características	Não Anômalo (Moda)	Anômalo (Moda)	Não Anômalo (Mediana)	Anômalo (Mediana)
Região	Nordeste (2)	Nordeste (2)	Nordeste (2)	Nordeste (2)
Dependência	Municipal (3)	Estadual (2)	Municipal (3)	Estadual (2)
Localização	Urbana (1)	Urbana (1)	Urbana (1)	Urbana (1)
Necessidades Básicas	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Tratamento de Lixo	Não (0)	Não Informado (9)	Não (0)	Não Informado (9)
Laboratório Ciências	Não (0)	Não (0)	Não (0)	Não (0)
Laboratório Informática	Não (0)	Não (0)	Não (0)	Não (0)
Quadra Esportes	Não (0)	Não (0)	Não (0)	Sim (1)
Refeitório	Não (0)	Não (0)	Não (0)	Não (0)
Acessibilidade	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Equipamento Multimídia	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Internet	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Profissional Serviços Gerais	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Profissional Saúde	Não (0)	Não (0)	Não (0)	Não (0)
Alimentação	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Material Pedagógico	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Exame de Seleção	Não (0)	Não (0)	Não (0)	Não (0)
Regular	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Diurno	Sim (1)	Sim (1)	Sim (1)	Sim (1)
EAD	Não (0)	Não (0)	Não (0)	Não (0)
Alunos Matriculados	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Infantil	Sim (1)	Não (0)	Sim (1)	Não (0)
Ensino Fundamental	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Médio	Não (0)	Sim (1)	Não (0)	Sim (1)
Ensino Profissionalizante	Não (0)	Sim (1)	Não (0)	Sim (1)
Educação de Jovens e Adultos	Não (0)	Não (0)	Não (0)	Não (0)

A região Nordeste com sobreajuste destaca padrões que, em geral, se mantiveram semelhantes ao modelo sem sobreajuste, mas com algumas mudanças significativas que valem a análise mais aprofundada. Embora os nós anômalos e não anômalos compartilhem muitas características, as diferenças em relação à oferta de ensino e alguns outros atributos sugerem mudanças estruturais que podem ser interessantes.

Assim como no modelo sem sobreajuste, as escolas da região Nordeste continuam majoritariamente localizadas em áreas urbanas. Esse padrão reflete a concentração populacional e a infraestrutura educacional voltada para centros urbanos.

Além disso, características como necessidades básicas, acessibilidade, e equipamentos multimídia permanecem uniformes entre as duas categorias, indicando que, de maneira geral, as escolas possuem acesso mínimo a essas estruturas. Contudo, a falta de quadras esportivas nas escolas não anômalas, enquanto uma minoria de escolas anômalas apresenta quadra, é um dado interessante. Podendo indicar que essas escolas anômalas estejam em regiões mais privilegiadas em termos de infraestrutura, mas que, devido a outros fatores, acabam destoando do padrão geral.

A mudança mais significativa nos nós anômalos ocorre nas características relacionadas à oferta de ensino. Enquanto as escolas não anômalas mantêm um foco predominante em Ensino Infantil e Fundamental, as escolas anômalas se diferenciam por oferecer Ensino Médio e Ensino Profissionalizante. Essa diferença pode indicar que as anomalias são geradas em

função do público atendido. Ao focar em um público mais velho, essas escolas se destacam no modelo como "anômalas" quando comparadas às escolas voltadas para a educação infantil.

Essa mudança de perfil pode refletir a presença de escolas técnicas ou instituições que preparam os alunos para o mercado de trabalho, o que é menos comum no restante da amostra. Esse ponto de divergência ressalta como o sobreajuste ajudou a identificar variações importantes dentro da região, que passariam despercebidas caso ele não fosse realizado.

4.2.4 Resultados da Região Sudeste (Sobreajuste)

Tabela 10 – Comparação das Características dos Nós - Região Sudeste (Sobreajuste)

Características	Não Anômalo (Moda)	Anômalo (Moda)	Não Anômalo (Mediana)	Anômalo (Mediana)
Região	Sudeste (3)	Sudeste (3)	Sudeste (3)	Sudeste (3)
Dependência	Municipal (3)	Privada (4)	Municipal (3)	Privada (4)
Localização	Urbana (1)	Urbana (1)	Urbana (1)	Urbana (1)
Necessidades Básicas	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Tratamento de Lixo	Não (0)	Não Informado (9)	Não (0)	Não Informado (9)
Laboratório Ciências	Não (0)	Não (0)	Não (0)	Não (0)
Laboratório Informática	Não (0)	Não (0)	Não (0)	Não (0)
Quadra Esportes	Não (0)	Não (0)	Não (0)	Não (0)
Refeitório	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Acessibilidade	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Equipamento Multimídia	Sim (1)	Não (0)	Sim (1)	Não (0)
Internet	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Profissional Serviços Gerais	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Profissional Saúde	Não (0)	Não (0)	Não (0)	Não (0)
Alimentação	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Material Pedagógico	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Exame de Seleção	Não (0)	Não (0)	Não (0)	Não (0)
Regular	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Diurno	Sim (1)	Sim (1)	Sim (1)	Sim (1)
EAD	Não (0)	Não (0)	Não (0)	Não (0)
Alunos Matriculados	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Infantil	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Fundamental	Sim (1)	Não (0)	Sim (1)	Não (0)
Ensino Médio	Não (0)	Não (0)	Não (0)	Não (0)
Ensino Profissionalizante	Não (0)	Não (0)	Não (0)	Não (0)
Educação de Jovens e Adultos	Não (0)	Não (0)	Não (0)	Não (0)

Uma das diferenças mais marcantes entre os nós anômalos e não anômalos na região Sudeste está na dependência administrativa. Enquanto as escolas não anômalas são predominantemente municipais, as escolas anômalas se destacam por serem privadas. Esse achado pode estar associado ao fato de que as escolas privadas, embora presentes em menor número, possuem características diferentes em termos de gestão e acesso a recursos, o que as distancia do padrão observado na maioria das escolas públicas da região.

Em termos de infraestrutura, muitas das características essenciais — como a presença de acessibilidade, alimentação e serviços gerais — se mantêm semelhantes entre os nós anômalos e não anômalos, indicando que, de maneira geral, as escolas da região Sudeste apresentam uma base estrutural relativamente sólida.

No entanto, há algumas diferenças que merecem destaque, como a presença de equipamentos multimídia. Enquanto as escolas não anômalas geralmente possuem esse tipo de

recurso, muitas escolas anômalas não possuem. Esse achado é curioso, já que as escolas privadas, em tese, teriam mais acesso a esse tipo de equipamento.

Outro ponto interessante vem do tratamento de lixo. Tanto nas escolas anômalas quanto nas não anômalas, a maioria não possui um tratamento adequado ou essa informação simplesmente não é reportada.

Diferentemente do que foi observado na região Nordeste, a oferta de ensino na região Sudeste é bem similar à do caso original. Tanto as escolas anômalas quanto as não anômalas oferecem Ensino Infantil, porém as escolas anômalas carecem do Ensino Fundamental. O fato que continua peculiar é a ausência de ensino médio, principalmente se considerarmos o fator da maior presença de escolas privadas na região.

4.2.5 Resultados da Região Sul (Sobreajuste)

Tabela 11 – Comparação das Características dos Nós - Região Sul (Sobreajuste)

Características	Não Anômalo (Moda)	Anômalo (Moda)	Não Anômalo (Mediana)	Anômalo (Mediana)
Região	Sul (4)	Sul (4)	Sul (4)	Sul (4)
Dependência	Municipal (3)	Privada (4)	Municipal (3)	Privada (4)
Localização	Urbana (1)	Urbana (1)	Urbana (1)	Urbana (1)
Necessidades Básicas	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Tratamento de Lixo	Sim (1)	Não Informada (0)	Sim (1)	Não (0)
Laboratório Ciências	Não (0)	Não (0)	Não (0)	Não (0)
Laboratório Informática	Não (0)	Não (0)	Não (0)	Não (0)
Quadra Esportes	Não (0)	Não (0)	Não (0)	Não (0)
Refeitório	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Acessibilidade	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Equipamento Multimídia	Sim (1)	Não (0)	Sim (1)	Não (0)
Internet	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Profissional Serviços Gerais	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Profissional Saúde	Não (0)	Não (0)	Não (0)	Não (0)
Alimentação	Sim (1)	Não (0)	Sim (1)	Não (0)
Material Pedagógico	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Exame de Seleção	Não (0)	Não (0)	Não (0)	Não (0)
Regular	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Diurno	Sim (1)	Sim (1)	Sim (1)	Sim (1)
EAD	Não (0)	Não (0)	Não (0)	Não (0)
Alunos Matriculados	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Infantil	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Fundamental	Sim (1)	Não (0)	Sim (1)	Não (0)
Ensino Médio	Não (0)	Não (0)	Não (0)	Não (0)
Ensino Profissionalizante	Não (0)	Não (0)	Não (0)	Não (0)
Educação de Jovens e Adultos	Não (0)	Não (0)	Não (0)	Não (0)

Assim como em outras regiões, a dependência administrativa é uma das características que mais difere entre as escolas anômalas e não anômalas na Região Sul. As escolas não anômalas são predominantemente municipais, enquanto as anômalas tendem a ser privadas. Esse padrão já observado no Sudeste reforça a ideia de que escolas privadas, mesmo com recursos diferenciados, se distanciam do padrão educacional majoritário.

Ao observar as características relacionadas à infraestrutura, percebe-se que, tanto nos nós anômalos quanto nos não anômalos, a maioria das escolas possui acessibilidade e internet. No entanto, uma diferença importante surge no quesito alimentação: enquanto as escolas

não anômalas garantem esse recurso, as escolas anômalas frequentemente não possuem oferta de alimentação. Esse dado pode ser interpretado como uma falha no atendimento de uma necessidade básica em escolas privadas, que podem priorizar outros tipos de infraestrutura ou confiar que os alunos providenciam sua própria alimentação.

Em termos de equipamento multimídia, as escolas anômalas também se destacam por não possuírem esse recurso, enquanto as não anômalas geralmente o possuem. O que novamente é uma contradição interessante, já que se espera que escolas privadas tenham esse tipo de equipamento com uma frequência maior do que as municipais.

Novamente o tratamento de lixo é mais frequentemente registrado em escolas não anômalas, enquanto nas escolas anômalas essa informação geralmente está ausente ou não é considerada.

Por fim, na oferta de ensino, observa-se que as escolas não anômalas geralmente oferecem Ensino Infantil e Ensino Fundamental, enquanto as escolas anômalas frequentemente não oferecem o Ensino Fundamental. Isso pode ser um indicativo de que as escolas anômalas da região Sul são voltadas para fases específicas da educação, como o Ensino Infantil, e não se dedicam ao Ensino Fundamental.

4.2.6 Resultados da Região Centro-Oeste (Sobreajuste)

Tabela 12 – Comparação das Características dos Nós - Região Centro-Oeste (Sobreajuste)

Características	Não Anômalo (Moda)	Anômalo (Moda)	Não Anômalo (Mediana)	Anômalo (Mediana)
Região	Centro-Oeste (5)	Centro-Oeste (5)	Centro-Oeste (5)	Centro-Oeste (5)
Dependência	Municipal (3)	Municipal (3)	Municipal (3)	Municipal (3)
Localização	Urbana (1)	Urbana (1)	Urbana (1)	Urbana (1)
Necessidades Básicas	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Tratamento de Lixo	Não (0)	Sim (1)	Não (0)	Sim (1)
Laboratório Ciências	Não (0)	Não (0)	Não (0)	Não (0)
Laboratório Informática	Não (0)	Não (0)	Não (0)	Não (0)
Quadra Esportes	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Refeitório	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Acessibilidade	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Equipamento Multimídia	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Internet	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Profissional Serviços Gerais	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Profissional Saúde	Não (0)	Não (0)	Não (0)	Não (0)
Alimentação	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Material Pedagógico	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Exame de Seleção	Não (0)	Não Informado (9)	Não (0)	Não Informado (9)
Regular	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Diurno	Sim (1)	Sim (1)	Sim (1)	Sim (1)
EAD	Não (0)	Não (0)	Não (0)	Não (0)
Alunos Matriculados	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Infantil	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Fundamental	Sim (1)	Sim (1)	Sim (1)	Sim (1)
Ensino Médio	Não (0)	Não (0)	Não (0)	Não (0)
Ensino Profissionalizante	Não (0)	Não (0)	Não (0)	Não (0)
Educação de Jovens e Adultos	Não (0)	Não (0)	Não (0)	Não (0)

Ao contrário de outras regiões onde as escolas anômalas tendem a ser privadas ou estaduais, no Centro-Oeste, tanto as escolas não anômalas quanto as anômalas são predo-

minantemente municipais. Essa dependência compartilhada não é suficiente para explicar as anomalias no modelo, sugerindo que os fatores que levam ao comportamento anômalo estão mais relacionados às condições de infraestrutura e recursos.

Nas características de infraestrutura, as escolas anômalas apresentam um nível de recursos semelhante às não anômalas em muitos aspectos. Ambas possuem alta disponibilidade de refeitórios, acessibilidade, equipamentos multimídia, e internet, o que demonstra uma atenção básica à infraestrutura escolar.

No entanto, algumas diferenças surgem, como o tratamento de lixo, onde as escolas anômalas têm mais chances de reportar que possuem essa infraestrutura, enquanto as não anômalas muitas vezes não a possuem. O que é um dado interessante pois não havia ocorrido em nenhuma das outras regiões, nem mesmo nos dados originais sem sobreajuste.

Um ponto interessante é a falta de informações consistentes sobre exames de seleção nas escolas anômalas. Enquanto nas escolas não anômalas essa variável é mais claramente reportada como "não", nas anômalas há uma lacuna de dados, com "não informado" sendo a moda. Isso pode simbolizar uma certa confusão ou descaso na hora de preencher os dados do censo, o que contribui para essa classificação fora do padrão das outras regiões.

Tanto as escolas anômalas quanto as não anômalas têm uma presença quase idêntica na oferta de ensino. A Educação Infantil e o Ensino Fundamental são amplamente oferecidos, enquanto o Ensino Médio, Profissionalizante, e Educação de Jovens e Adultos são praticamente ausentes em ambas. Isso reflete um padrão regional em que a educação básica é amplamente atendida pelas instituições municipais, mas os níveis mais avançados de ensino devem ser cobertos pelos outros tipos de dependência administrativa.

4.3 Outras Análises

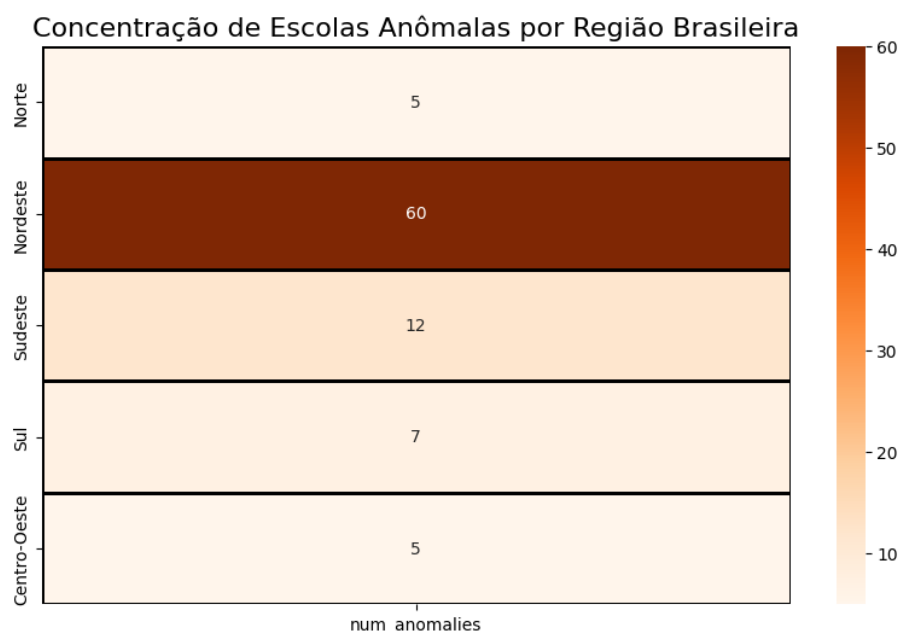


Figura 4.3.1 – Concentração de Anomalias por Região

A análise dos dados da concentração de escolas anômalas por região brasileira mostra uma distribuição bastante desigual de anomalias entre as regiões, com algumas observações importantes:

Nordeste: O Nordeste se sobressai com uma quantidade expressivamente maior de anomalias (60) em comparação com outras regiões. Isso pode indicar que as escolas desta região apresentam mais desafios ou inconsistências nos dados, como infraestrutura inadequada, falta de recursos, ou outros fatores que afetam a normalidade dos dados observados.

Sudeste e Sul: O Sudeste, embora seja uma região mais populosa e com maior concentração de escolas, apresenta 12 anomalias, o que é relativamente baixo comparado ao Nordeste, mas ainda acima das regiões Norte e Centro-Oeste. O Sul segue com 7 anomalias, sugerindo uma leve concentração de irregularidades, embora menor em relação ao Sudeste.

Norte e Centro-Oeste: Ambas as regiões Norte e Centro-Oeste apresentam apenas 5 anomalias cada uma. Esse número pode indicar uma melhor adequação dos dados ou uma possível sub notificação, dependendo de como os dados foram coletados e analisados.

5 Conclusão

Neste trabalho, foi analisada a aplicação do modelo de detecção de anomalias por reconstrução de erros em grafos por meio de uma rede neural convolucional de grafos. Convertendo uma parte de uma grande base de dados para o formato de grafo com um conjunto de até 8000 nós aleatórios, foram investigadas características globais e regionais das escolas para identificar padrões e encontrar possíveis *insights*. Alguns desses *insights* e padrões foram:

5.1 Grafo Global (Original)

- A maioria dos nós não anômalos estão localizados na região Sudeste, com a UF mais comum sendo São Paulo.
- As escolas apresentam condições básicas adequadas, porém, a ausência de tratamento de lixo e instalações adicionais como laboratórios de ciências e informática é notável.
- A presença de Ensino Médio, Técnico (Escolas Federais), Profissionalizante e EJA é bem menor que a do Ensino Infantil ou Fundamental.
- A modalidade de Ensino a Distância (EAD) não foi observada em peso, o que é surpreendente dado o contexto pós-pandemia.

5.2 Regiões Específicas (Originais)

- **Norte:** Notável pela ausência de anomalias, indicando uma possível homogeneidade nas condições escolares.
- **Nordeste:** Diferenças significativas entre nós anômalos e não anômalos em termos de dependência administrativa (estadual versus municipal) e oferta de níveis de ensino, além de ser a segunda região com mais escolas nos dados observados.
- **Sudeste:** Segue o padrão global (principalmente por ser a região mais populosa) com algumas variações, especialmente na presença de quadras de esporte.
- **Sul:** Se diferencia por meio da dependência e localização entre nós anômalos e não anômalos, com anômalos possuindo mais quadras de esporte.
- **Centro-Oeste:** Similaridades com o padrão global, mas com variações na presença de quadras de esporte e refeitórios.

Além da versão original dos dados, utilizando um modelo tradicional, foi criado um modelo alternativo utilizando dados sobreajustados e mais focados para encontrar com maior precisão as anomalias em cada região. Graças a esse modelo foram feitas as seguintes descobertas e análises:

5.3 Grafo Global (Sobreajuste)

- A maioria dos nós anômalos estão localizados na região Nordeste, seguido do Sudeste, Sul, Norte e Centro-Oeste. Além disso, nesse novo modelo a maior parte das escolas se encontra na localidade urbana.
- Assim como no modelo original, em geral as escolas cumprem as necessidades básicas esperadas, porém os casos anômalos do modelo sobreajustado não possuem refeitório e não informaram sobre a situação de tratamento de lixo.
- O Ensino Fundamental está presente na maior parte dos dados, mas o Ensino Infantil foi mais proeminente em dados não anômalos. Também é de interesse comentar que novamente houve uma grande ausência de Ensino Médio, Profissionalizante e EJA.
- Em ambos os modelos, a modalidade de Ensino a Distância (EAD) foi virtualmente inexistente, um achado surpreendente dado o contexto pós-pandemia. Algo que levanta questões sobre o preparo da educação brasileira para essa nova realidade.

5.4 Regiões Específicas (Sobreajuste)

- **Norte:** Diferente da versão original agora foram encontradas anomalias, também houve uma mudança na dependência administrativa, uma falha no atendimento de necessidades básicas nos casos anômalos, falta de acessibilidade, internet e até mesmo material pedagógico. Todos esses fatores somados mostram um cenário que necessita de melhoras o quanto antes.
- **Nordeste:** Em geral se manteve bem similar ao modelo original, mas com algumas mudanças interessantes. Dentre elas destacamos: A presença de Ensino Médio e Profissionalizante nos dados anômalos é algo único e só foi observado nessa região.
- **Sudeste:** Se diferencia da original pela presença de dados anômalos com a dependência administrativa do tipo Privada. Algo que é peculiar visto que os dados anômalos da região Sudeste indicam uma certa precariedade ou falta de recursos, visto pela ausência de projetores multimídia, ausência de ensino médio e ausência de tratamento de lixo.

- **Sul:** Novamente pode-se perceber a aparição do tipo administrativo Privado, mas esse não é o foco dessa região. O detalhe de mais interesse que a difere das outras é a presença de tratamento de lixo nas escolas não anômalas e a ausência de alimentação nas anômalas, algo que foi comentado em mais detalhes na seção de resultados da região.
- **Centro-Oeste:** Finalizando com a região com a menor quantidade de mudanças. A região Centro-Oeste em geral se manteve muito similar a versão original com algumas mudanças em foco: Presença de tratamento de lixo nas anômalas, presença de quadra de esportes em geral, falha da informação sobre exame de seleção e uma homogeneidade na parte de ofertas de ensino.

Os resultados deste estudo oferecem uma nova perspectiva sobre a distribuição e características das escolas brasileiras, ressaltando as disparidades regionais. As conclusões podem ser úteis para políticas educacionais e alocação de recursos, especialmente em regiões mais carentes de infraestrutura adequada.

5.5 Limitações

- **Amostragem:** A análise foi limitada a 8000 nós, o que, embora represente uma amostra significativa, não cobre a totalidade das escolas brasileiras. Essa amostragem pode não refletir adequadamente a diversidade regional e estrutural de todas as instituições.
- **Dados Limitados pela Pandemia:** A coleta de dados ocorreu em um período marcado pela pandemia de COVID-19, o que pode ter afetado não apenas a disponibilidade de recursos nas escolas, mas também a precisão dos dados. Fatores como fechamento temporário de escolas e mudanças no regime de ensino (presencial para remoto) podem ter gerado distorções.
- **Quantidade de Colunas Analisadas:** Muitas colunas foram cortadas da análise devido à análise exploratória e ao tempo de execução elevado ou à sua similaridade com outras colunas. Isso pode ter deixado de revelar padrões adicionais ou fatores determinantes para o comportamento anômalo em certas regiões.
- **Falta de Estudos Direcionados:** Devido à falta de algumas colunas, não foi realizado um estudo direcionado a cidades ou municípios específicos, o que poderia ter proporcionado *insights* bastante interessantes para a pesquisa.

5.6 Melhorias Futuras

- **Expansão da Amostra:** Expandir a amostra analisada permitirá capturar uma visão mais detalhada e abrangente da realidade educacional brasileira, incluindo escolas de diferentes perfis e regiões menos representadas.
- **Expansão das Características Analisadas:** Incluir um número maior de características de nós para uma análise com mais informações de interesse. O que trará uma análise mais robusta e permitirá identificar com maior precisão os fatores que contribuem para as anomalias.
- **Refinamento do Modelo:** Ajustes no modelo de detecção de anomalias, como o uso de algoritmos mais avançados e ajustes nos hiperparâmetros, podem reduzir vieses e melhorar a sensibilidade do modelo, especialmente nas regiões menos detectadas por anomalias ou nas regiões com muitas inconsistências.
- **Investigações Regionais Detalhadas:** Realizar análises mais detalhadas por região permitirão uma melhor compreensão das disparidades regionais e identificarão causas específicas para as anomalias.

Por fim, podemos concluir que, apesar das limitações, este estudo fornece uma base sólida para entender as condições das escolas brasileiras e identificar áreas que necessitam de atenção especial. As disparidades regionais destacadas pelo estudo reforçam a importância de políticas educacionais adaptadas às necessidades específicas de cada região.

Referências

- AKOGLU, L.; TONG, H.; KOUTRA, D. Graph based anomaly detection and description: a survey. *Data mining and knowledge discovery*, Springer, v. 29, p. 626–688, 2015.
- AL-SHABANDAR, R. et al. Detecting at-risk students with early interventions using machine learning techniques. *IEEE Access*, IEEE, v. 7, p. 149464–149478, 2019.
- ALPAYDIN, E. *Introduction to machine learning*. [S.l.]: MIT press, 2020.
- AN, J.; CHO, S. Variational autoencoder based anomaly detection using reconstruction probability. *Special lecture on IE*, v. 2, n. 1, p. 1–18, 2015.
- AYANWALE, M. A. et al. Teachers' readiness and intention to teach artificial intelligence in schools. *Computers and Education: Artificial Intelligence*, v. 3, p. 100099, 2022. ISSN 2666-920X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666920X22000546>.
- BERNIUS, J. P.; KRUSCHE, S.; BRUEGGE, B. Machine learning based feedback on textual student answers in large courses. *Computers and Education: Artificial Intelligence*, Elsevier, v. 3, p. 100081, 2022.
- BLONDEL, V. D. et al. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, IOP Publishing, v. 2008, n. 10, p. P10008, 2008.
- CELAR, S. et al. Classification of test documents based on handwritten student id's characteristics. *Procedia Engineering*, Elsevier, v. 100, p. 782–790, 2015.
- CHADHA, A.; JAIN, V. Graph neural networks. *Distilled AI*, 2020. <https://aman.ai>.
- CHEN, H.; HE, B. Automated essay scoring by maximizing human-machine agreement. In: *Proceedings of the 2013 conference on empirical methods in natural language processing*. [S.l.: s.n.], 2013. p. 1741–1752.
- COLPANI, R. Mineração de dados educacionais: um estudo da evasão no ensino médio com base nos indicadores do censo escolar. *Informática na educação: teoria & prática*, v. 21, n. 3, 2018.
- DINIZ, E. O censo escolar. *Revista Brasileira de Estudos Pedagógicos*, v. 80, n. 194, 1999.
- FRENEDA, F. C. B. et al. Múltiplos fatores do desempenho escolar: uma análise dos microdados do inep sobre a educação no distrito federal. Universidade Católica de Brasília, 2020.
- GILMER, J. et al. Neural message passing for quantum chemistry. In: PMLR. *International conference on machine learning*. [S.l.], 2017. p. 1263–1272.
- KASNECI, E. et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, Elsevier, v. 103, p. 102274, 2023.
- KINGMA, D. P.; BA, J. *Adam: A Method for Stochastic Optimization*. 2017.

- KIPF, T. N.; WELLING, M. *Semi-Supervised Classification with Graph Convolutional Networks*. 2017.
- KUČAK, D.; JURIČIĆ, V.; ĐAMBIĆ, G. Machine learning in education-a survey of current research trends. *Annals of DAAAM & Proceedings*, v. 29, 2018.
- LI, M. et al. Study-gnn: a novel pipeline for student performance prediction based on multi-topology graph neural networks. *Sustainability*, MDPI, v. 14, n. 13, p. 7965, 2022.
- LIANG, Z. et al. Graph path fusion and reinforcement reasoning for recommendation in moocs. *Education and Information Technologies*, Springer, v. 28, n. 1, p. 525–545, 2023.
- MINING, W. I. D. *Introduction to data mining*. [S.l.]: Springer, 2006.
- MITCHELL, T. M.; MITCHELL, T. M. *Machine learning*. [S.l.]: McGraw-hill New York, 1997. v. 1.
- NAKAGAWA, H.; IWASAWA, Y.; MATSUO, Y. Graph-based knowledge tracing: Modeling student proficiency using graph neural network. In: *IEEE/WIC/ACM International Conference on Web Intelligence*. New York, NY, USA: Association for Computing Machinery, 2019. (WI '19), p. 156–163. ISBN 9781450369343. Disponível em: <https://doi.org/10.1145/3350546.3352513>.
- QU, F.; JIA, X.; WU, Y. Asking questions like educational experts: Automatically generating question-answer pairs on real-world examination data. *arXiv preprint arXiv:2109.05179*, 2021.
- SANCHEZ-LENGELING, B. et al. A gentle introduction to graph neural networks. *Distill*, 2021. <https://distill.pub/2021/gnn-intro>.
- SARSA, S. et al. Automatic generation of programming exercises and code explanations using large language models. In: *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1*. [S.l.: s.n.], 2022. p. 27–43.
- SUSANTI, M. N. I. et al. Link prediction in educational graph data to predict elective course using graph convolutional network model. In: *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*. [S.l.: s.n.], 2023. p. 714–719.
- VANDERPLAS, J. *Python data science handbook: Essential tools for working with data*. [S.l.]: "O'Reilly Media, Inc.", 2016.
- VASSOYAN, J.; VIE, J.-J.; LEMBERGER, P. Towards scalable adaptive learning with graph neural networks and reinforcement learning. *ArXiv*, abs/2305.06398, 2023. Disponível em: <https://api.semanticscholar.org/CorpusID:258615287>.
- VIZZOTTO, P. A. Inclusão na educação básica brasileira: Análise do censo escolar por meio dos microdados do inep. *Ensaio Pedagógicos*, v. 4, n. 1, p. 102–112, 2020.
- WU, S. et al. Graph neural networks in recommender systems: a survey. *ACM Computing Surveys*, ACM New York, NY, v. 55, n. 5, p. 1–37, 2022.
- WU, Z. et al. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, IEEE, v. 32, n. 1, p. 4–24, 2020.

YING, R. et al. Graph convolutional neural networks for web-scale recommender systems. In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. [S.l.: s.n.], 2018. p. 974–983.

ZAMBRANO, J. L. et al. Early prediction of student learning performance through data mining: A systematic review. *Psicothema*, 2021.