UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

Taciana dos Santos Vasconcelos

REESTRUTURAÇÃO ÉTICA NA MINERAÇÃO DE DADOS EDUCACIONAIS SUPERIORES: CONFORMIDADE COM A LEI GERAL DE PROTEÇÃO DE DADOS

Recife

Agosto de 2025

Taciana dos Santos Vasconcelos

REESTRUTURAÇÃO ÉTICA NA MINERAÇÃO DE DADOS EDUCACIONAIS SUPERIORES: CONFORMIDADE COM A LEI GERAL DE PROTEÇÃO DE DADOS

Artigo apresentado ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Universidade Federal Rural de Pernambuco – UFRPE
Departamento de Estatística e Informática
Curso de Bacharelado em Sistemas de Informação

Orientadores: Roberta Gouveia; Gabriel Alves

Recife
Agosto de 2025

REESTRUTURAÇÃO ÉTICA NA MINERAÇÃO DE DADOS EDUCACIONAIS SUPERIORES: CONFORMIDADE COM A LEI GERAL DE PROTEÇÃO DE DADOS

ETHICAL RESTRUCTURING IN HIGHER EDUCATION DATA MINING: COMPLIANCE WITH THE GENERAL DATA PROTECTION REGULATION

REESTRUCTURACIÓN ÉTICA EN LA MINERÍA DE DATOS DE EDUCACIÓN SUPERIOR: CUMPLIMIENTO CON LA LEY GENERAL DE PROTECCIÓN DE DATOS

Taciana dos Santos Vasconcelos[1]
Ebony Marques Rodrigues[2]
Roberta Macêdo Marques Gouveia[3]
Gabriel Alvez de Albuquerque Junior[4]
Maria da Conceição Moraes Batista[5]

RESUMO
Este estudo aborda a reestruturação dos dados públicos educacionais do Ensino Superior promovida pelo INEP, alinhada à LGPD. Inspirado pelo estudo conduzido por Rodrigues (2021), que analisou concluintes de graduação, visando à construção de modelos de classificação utilizando fatores socioeconômicos e tempo estimado para conclusão da graduação em IES públicas. Este trabalho examina como as mudanças na configuração dos dados do ENADE e do Censo da Educação Superior afetam a realização de pesquisas

[1] Pós-Graduada em Cloud Computing e Ciência de Dados pela Universidade Anhanguera. Graduada em Gestão da Tecnologia da Informação pela Universidade Estácio de Sá. Graduanda em Sistemas de Informação pela Universidade Federal Rural de Pernambuco (UFRPE) e cursa MBA em Data Science pela Uninassau. Atuou em projeto de Mineração e Análise de Dados Públicos de Inovação da Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE), e em pesquisa aplicada à educação. Atua como residente em TI na Procuradoria Geral do Estado do Rio Grande do Norte (PGE-RN). Lattes: http://lattes.cnpq.br/8691839294756407
[2] Mestre em Informática Aplicada e graduado em Sistemas de Informação (UFRPE). Atuou como analista de sistemas na Empresa Municipal de Informática (Emprel), em Recife. Atualmente, é Analista de TI na UFRPE, com experiência em engenharia e análise de dados. Pesquisa ciência de dados aplicada à educação, com foco em descoberta de conhecimento em dados educacionais abertos. Lattes: http://lattes.cnpq.br/5929185711837204
[3] Doutora em Engenharia pela Universidade Federal Rural da Paraíba (UFPB), mestre e bacharela em Ciência da Computação (UFPB). Professora da UFRPE, com atuação nas áreas de *Data Science*, *Machine Learning* e *Open Data*, com foco em educação e design centrado no usuário. Lattes: http://lattes.cnpq.br/2024317361355224
[4] Doutor e mestre em Ciência da Computação pela Universidade Federal de Pernambuco (UFPE). Professor da UFRPE e coordenador do Observatório de Dados da Graduação. Pesquisa avaliação de performabilidade e *Learning Analytics*. Lattes: http://lattes.cnpq.br/1399502815770584
[5] Doutora e mestre em Ciência da Computação pela Universidade Federal de Pernambuco (UFPE). Professora da UFRPE, com experiência em banco de dados, qualidade e integração da informação. Atuou como gerente de projetos no Tribunal de Contas do Estado de Pernambuco (TCE-PE) e em iniciativas de *Business Inteligence* no setor privado. Lattes: https://lattes.cnpq.br/8167265341219263

científicas. Em resposta às mudanças nos microdados efetuadas pelo INEP, que impossibilitam a reprodução de trabalhos com finalidade de análise individual de discentes, este estudo direcionou seu enfoque para as informações dos cursos e tempo de graduação dos discentes nesses cursos, considerando medidas de tendência central. Consideraram-se os anos de 2016 a 2018 para dados do ENADE e 2018 para Censo da Educação Superior. Utilizou-se o processo de *Knowledge Discovery in Database* (KDD) ao longo do trabalho, desde a seleção até a interpretação de dados. Usando 5.170 registros de cursos, técnicas do Aprendizado de Máquina Supervisionado foram empregadas para construção de modelos de regressão e classificação. Essa abordagem visa superar os desafios éticos e metodológicos da reestruturação dos dados, garantindo a utilidade dos dados para fins de pesquisa científica. Resultados mostram que as mudanças permitiram o uso eficaz de modelos de Aprendizado de Máquina. O estudo destaca a importância da ética nos dados educacionais e na inteligência artificial, garantindo a proteção da privacidade e a responsabilidade na utilização dos dados para tomada de decisões futuras.

PALAVRAS-CHAVE: Educação. Dados educacionais. LGPD. Ética. Inteligência artificial. Decisões.

ABSTRACT

This study addresses the restructuring of public educational data in Higher Education promoted by INEP, aligned with LGPD. Inspired by the study conducted by Rodrigues (2021), which analyzed undergraduate graduates, aiming to construct classification models using socioeconomic factors and estimated time for graduation completion in public HEIs. This work examines how changes in the configuration of ENADE and Higher Education Census data affect the conduct of scientific research. In response to changes in microdata made by INEP, which make it impossible to reproduce works for the purpose of individual analysis of students, this study directed its focus to course information and graduation time of students in these courses, considering measures of central tendency. The years 2016 to 2018 were considered for ENADE data and 2018 for the Higher Education Census. The Knowledge Discovery in Database (KDD) process was used throughout the work, from data selection to interpretation. Using 5,170 course records, techniques of Supervised Machine Learning were employed for the construction of regression and classification models. This approach aims to overcome the ethical and methodological challenges of data restructuring, ensuring the utility of data for scientific research purposes. Results show that the changes allowed the effective use of Machine Learning models. The study highlights the importance of ethics in educational data and artificial intelligence, ensuring the protection of privacy and responsibility in the use of data for future decision-making.

KEYWORDS: Education. Educational data. LGPD. Ethics. Artificial intelligence. Decisions.

RESUMEN

Este estudio aborda la reestructuración de los datos educativos públicos en la Educación Superior promovida por el INEP, alineada con la LGPD. Inspirado en el estudio realizado por Rodrigues (2021), que analizó graduados universitarios, con el objetivo de construir modelos de clasificación utilizando factores socioeconómicos y tiempo estimado para la conclusión de la graduación en instituciones de educación superior públicas. Este trabajo examina cómo los cambios en la configuración de los datos del ENADE y del Censo de la Educación Superior afectan la realización de investigaciones científicas. En respuesta a los cambios en los microdatos realizados por el INEP, que hacen imposible reproducir trabajos con fines de análisis individual de estudiantes, este estudio dirigió su enfoque hacia la información del curso y el tiempo de graduación de los estudiantes en estos cursos, considerando medidas de tendencia central. Se consideraron los años 2016 a 2018 para los datos del ENADE y 2018 para el Censo de la Educación Superior. Se utilizó el proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD) a lo largo del trabajo, desde la selección de datos hasta la interpretación. Utilizando 5,170 registros de cursos, se emplearon técnicas de Aprendizaje Automático Supervisado para la construcción de modelos de regresión y clasificación. Este enfoque tiene como objetivo superar los desafíos éticos y metodológicos de la reestructuración de datos, garantizando la utilidad de los datos para fines de investigación científica. Los resultados muestran que los cambios permitieron el uso efectivo de modelos de Aprendizaje Automático. El estudio resalta la importancia de la ética en los datos educativos y la inteligencia artificial, garantizando la protección de la privacidad y la responsabilidad en el uso de datos para la toma de decisiones futuras.

PALABRAS CLAVE: Educación. Datos educativos. LGPD. Ética. Inteligencia artificial. Decisiones.

INTRODUÇÃO

No âmbito do ensino superior brasileiro, a divulgação de dados públicos desempenha um papel crucial, não apenas para promover a transparência, mas também para avaliar as políticas educacionais em vigor. As mudanças decorrentes da promulgação da Lei Geral de Proteção de Dados Pessoais (LGPD) apresentaram desafios significativos que exigiram adaptações fundamentais no tratamento e compartilhamento de informações sensíveis. Em resposta a essas demandas, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) implementou alterações substanciais na divulgação de microdados educacionais, alinhando-se às diretrizes da LGPD.

Esses arquivos, conhecidos como "microdados educacionais", contêm informações sobre estudantes, cursos, instituições de ensino e outros elementos relevantes para o sistema educacional brasileiro. A LGPD, promulgada em agosto de 2018, estabeleceu diretrizes rigorosas para garantir a integridade dos dados pessoais, com o objetivo de conter acessos não autorizados e o uso indevido de informações sensíveis.

A adaptação dos microdados à LGPD foi motivada pela identificação, em colaboração com a Universidade Federal de Minas Gerais (UFMG), de potenciais riscos de identificação dos estudantes por meio da correlação de identificadores no Censo da Educação Superior de 2018. A conclusão da análise realizada pelo Termo de Execução Descentralizada (TED) 8750 da UFMG indicou que a utilização de três variáveis — dia e ano de nascimento e código do curso — poderia resultar na identificação de até 39% dos alunos. Por sua vez, a combinação de quatro variáveis — dia, mês e ano de nascimento, e código do curso — acarretaria na identificação de até 80% dos indivíduos.

Esse processo de adaptação resultou na temporária suspensão do acesso aos dados do INEP, seguida por uma sequência programada de republicação das bases de dados. O clímax desse processo ocorreu em outubro de 2022, com a republicação completa das edições de 2004 a 2021 do Exame Nacional de Desempenho de Estudantes (ENADE), representando não apenas uma restauração, mas também um esforço complexo de harmonização das informações à nova estrutura preconizada pela LGPD.

Nesse contexto de reestruturação, houve transformações abrangendo aspectos estruturais, de formatação e de apresentação dos dados. Um aspecto notável foi a padronização das variáveis, visando à uniformização das representações previamente divergentes. Entretanto, as alterações mais substanciais foram direcionadas aos dados pessoais dos participantes do Exame Nacional do Ensino Médio (ENEM) e dos estudantes do ensino superior. Devido à natureza sensível dessas informações, medidas foram adotadas para mitigar eventuais riscos de identificação.

A reestruturação dos microdados teve como finalidade dificultar a identificação dos indivíduos, sendo realizada mediante a segmentação dos dados em grupos específicos de

informações, organizados conforme distintas variáveis. Nesse panorama, é relevante destacar que a pesquisa realizada por Rodrigues (2021) estava predominantemente voltada à análise do tempo de graduação dos discentes. Entretanto, diante das alterações na estrutura dos dados e da impossibilidade da vinculação individual dos discentes, essa abordagem não pôde ser empregada diretamente.

Como alternativa, houve a necessidade de observar as informações no âmbito dos cursos, usando o número do curso como identificador, armazenado na variável “CO_CURSO”. Dessa forma, os dados foram sumarizados em relação aos cursos. Nesse contexto, é fundamental esclarecer que a sumarização dos dados foi uma técnica essencial para transformar a visão dos discentes para o curso, possibilitando a análise e a compreensão dos desfechos e do tempo de graduação dos estudante nos cursos.

A metodologia adotada segue o processo de *Knowledge Discovery in Databases* (KDD), compreendendo etapas de análise exploratória e mineração de dados para desvendar informações pertinentes nos conjuntos de dados educacionais. As fases englobam a seleção, o pré-processamento, a transformação, a mineração de dados e a interpretação/avaliação, com adaptações para se adequarem com a nova estrutura dos microdados em consonância com a LGPD.

O objetivo geral desta pesquisa é reproduzir e adaptar o estudo conduzido por Rodrigues (2021), intitulado "Técnicas de Aprendizado de Máquina para Descoberta de Conhecimento sobre Dados Abertos do Ensino Superior Público Brasileiro", considerando a nova estrutura dos microdados do INEP, decorrente das alterações realizadas sobre os dados para que estejam em conformidade com a LGPD. Busca-se compreender a transformação na estrutura dos dados educacionais do ensino superior brasileiro, bem como examinar o impacto dessas mudanças nas análises e previsões realizadas.

Para atender a esse objetivo geral, os seguintes objetivos específicos foram delineados:

- Analisar as modificações inseridas na estrutura dos dados relacionados ao ENADE (2016-2018) e do Censo da Educação Superior (2018) decorrentes das ações do INEP em conformidade com a LGPD;
- Examinar as variáveis empregadas no estudo de Rodrigues (2021) e identificar atributos equivalentes, levando em consideração a nova organização dos microdados do INEP após a adaptação à LGPD;
- Avaliar a viabilidade de reproduzir o estudo de Rodrigues (2021), considerando as alterações nos microdados, e examinar o impacto dessas transformações nas análises realizadas;
- Compreender a base de dados resultante das sumarizações e agregações realizadas, empregando técnicas de análise exploratória para compreender suas características e peculiaridades;
- Investigar como as mudanças na estrutura dos microdados influenciaram as etapas de seleção, pré-processamento, transformação e análise, bem como os resultados obtidos

com o emprego das técnicas de Aprendizado de Máquina Supervisionado;
- Realizar experimentos de regressão e classificação com o objetivo de identificar padrões que influenciem no desempenho dos cursos e de detectar tendências relevantes a partir da nova configuração dos microdados.

Esses objetivos específicos direcionaram o desenvolvimento da pesquisa, abordando tanto as mudanças e limitações nas análises quanto às novas possibilidades de aplicação das técnicas de Aprendizado de Máquina Supervisionado na nova estrutura dos microdados educacionais do ensino superior.

METODOLOGIA

O *Knowledge Discovery in Databases* (KDD) é um processo complexo para a revelação de conhecimento a partir de bases de dados. Ele abrange uma série de etapas que auxiliam a compreensão e extração de informações relevantes. O KDD não se resume à mineração de dados, mas a um conjunto de procedimentos que envolve pré-processamento, pós-processamento e mineração de dados (Han & Kamber, 2006). Para uma abordagem mais específica, o processo KDD pode ser decomposto em cinco etapas distintas: seleção de dados, pré-processamento, transformação, mineração de dados, interpretação e avaliação de resultados. Importante ressaltar que, embora a mineração de dados seja amplamente conhecida, não deve ser considerada como um fim em si mesma, mas como um meio para alcançar a descoberta de conhecimento (Han et al., 2012).

Para conduzir o presente estudo, foram empregadas ferramentas de programação em um ambiente local, com ênfase na linguagem de programação Python, juntamente com uma série de bibliotecas, tais como Numpy, Pandas, Seaborn, Matplotlib e Scikit-Learn. A seleção dessas ferramentas foi guiada pela capacidade do Python de oferecer flexibilidade, eficiência e uma ampla variedade de bibliotecas específicas para manipulação, visualização e análise de dados.

As etapas do KDD forneceram uma estrutura metodológica sólida para abordar as transformações decorrentes da LGPD nos conjuntos de dados do ENADE e do Censo da Educação Superior. A análise foi realizada, desde a seleção dos dados até a interpretação dos resultados, garantindo que as informações obtidas estivessem alinhadas aos objetivos do estudo.

A pesquisa prosseguirá com as próximas etapas do KDD na presente discussão, abordando os processos específicos de seleção de dados, pré-processamento, transformação, mineração de dados e interpretação e avaliação de resultados. Essa abordagem permitirá uma compreensão das abordagens adotadas para enfrentar os desafios apresentados pelas transformações regulatórias e pela complexidade dos dados.

CONSTRUINDO O ALICERCE DO ESTUDO: ABORDAGENS PARA SELEÇÃO, PRÉ-PROCESSAMENTO E TRANSFORMAÇÃO DE DADOS

As três primeiras etapas do processo de KDD incluem a seleção, pré-processamento e transformação dos dados observados, visando à descoberta de conhecimento (Fayyad et al.,1996). Essas etapas são denominadas como tal e, neste estudo, as atividades correspondentes seguem a nomenclatura do KDD.

A etapa inicial do KDD, denominada seleção de dados, é fundamental. Cada base de dados requer a escolha dos atributos a serem empregados na modelagem, originando conjuntos de dados de interesse, também conhecidos como *target data* (Fayyad et al., 1996). Neste estudo, a seleção dos atributos orienta-se por considerações específicas, influenciadas pelo escopo do trabalho de Rodrigues (2021), que empregou 30 atributos do ENADE para classificar a permanência de estudantes universitários em IES.

Todavia, a reorganização resultante das diretrizes da LGPD introduziu alterações na disposição dos atributos. Em vez de estarem dispostos em uma única base de dados, esses atributos passaram a estar distribuídos em 23 arquivos não sequenciais, de um total de 32 fornecidos pelo INEP. Essa distribuição desigual, motivada por questões de privacidade, apresentou desafios para a identificação e organização das variáveis relevantes.

A abordagem segue o caminho do trabalho de Rodrigues (2021), restringindo a análise a cursos de bacharelado e licenciatura, com a seleção de instituições classificadas como faculdades, centros universitários ou universidades estaduais e federais. No estudo do referido autor, realizado antes da LGPD, foi possível realizar a exclusão de registros de estudantes com menos de 2 anos de permanência na IES, permitindo uma análise mais precisa. Contudo, uma mudança significativa surge da impossibilidade de remover registros de estudantes com menos de 2 anos de permanência na IES hoje, devido à inviabilidade de vinculação individual, conforme a LGPD. Essa alteração orienta o foco da análise para as características dos cursos.

A análise mais aprofundada dos arquivos revela um conjunto diversificado de atributos, muitos dos quais podem não ser pertinentes aos objetivos deste estudo. Além disso, o refinamento da seleção se estende aos cursos considerados na análise. Dos conjuntos iniciais de cursos (4.300 em 2016, 10.571 em 2017 e 8.813 em 2018), um número substancialmente menor de cursos (791, 3.625 e 918, respectivamente) atende aos critérios.

Com a conclusão da seleção de atributos, a análise dos dados para identificação de valores ausentes e outras inconsistências diversas é crucial. Esses registros inconsistentes precisam ser corrigidos ou descartados, a fim de preservar a qualidade dos modelos a serem construídos. Essas análises constituem a segunda etapa do KDD, chamada de pré-processamento de dados.

Ao tratar dos dados do ENADE, as operações realizadas na segunda etapa compreendem a padronização de certos atributos, como aqueles que indicam o turno de graduação dos estudantes. Tais atributos mostraram variações na edição de 2016, mas foram padronizados nas edições de 2017 e 2018. Mudanças gerais também foram aplicadas a todos os dados, incluindo a renomeação de atributos para maior clareza. Além disso, um conjunto de análises e verificações específicas foi desenvolvido para identificar inconsistências nos dados, com procedimentos definidos para tratar essas questões.

As análises de inconsistência trouxeram à luz o desafio dos valores ausentes nos conjuntos de dados do ENADE. As taxas de valores ausentes nos anos de 2016 (4,13%), 2017 (14,32%) e 2018 (10,13%) revelam a persistência desse problema. O tratamento requerido para lidar com valores ausentes é um aspecto do pré-processamento, visto que a falta de dados completos pode comprometer a confiabilidade das análises subsequentes.

Neste estudo, a análise detalhada identificou cursos específicos que continham exclusivamente valores ausentes nos questionários socioeconômicos, em todas as edições anuais analisadas. Para resolver esse problema, um processo foi conduzido, arquivo por arquivo, para verificar a correspondência dos códigos de cursos ("CO_CURSO").

Após identificar os cursos afetados, estes foram removidos dos conjuntos de dados. Essa abordagem reduziu as porcentagens de valores ausentes. Por exemplo, na edição de 2016, a porcentagem de valores ausentes foi reduzida para 3,93%. Nas edições de 2017 e 2018, a porcentagem foi ajustada para 14,17% e 10,06%, respectivamente. Vale ressaltar que outros valores ausentes são abordados na fase seguinte de sumarização dos dados por curso.

É importante enfatizar que as razões subjacentes para a ausência desses dados podem ser variadas, derivando de diferentes fontes, como opção do estudante por não responder a certas perguntas ou perda de dados.

A detecção de valores duplicados nos dados também trouxe à tona questões relevantes. Neste estudo, identificou-se a duplicação de cursos de "Serviço Social" em determinadas Instituições de Ensino Superior (IES), tanto em 2016 quanto em 2018, conforme documentado no manual do usuário fornecido pelo INEP.

A solução adotada para lidar com essa questão consistiu na aplicação de um filtro que considerou apenas a primeira ocorrência do "CO_CURSO" para cada curso de "Serviço Social" nas IES específicas. Essa abordagem permitiu a análise subsequente com dados corretos, eliminando as distorções que as duplicações indevidas poderiam causar. Esse procedimento assegura confiabilidade e a representatividade dos dados na análise, bem como a integridade dos resultados.

A conclusão da etapa de pré-processamento de dados direciona o estudo para a fase de transformação de dados, a terceira etapa do processo KDD. Essa etapa compreende operações para moldar os dados preparados na etapa anterior, a fim de atender às demandas do estudo. Por meio de análises, formatação e organização dos dados, a transformação de dados visa

criar um ambiente propício para a exploração de informações relevantes, alinhando-se aos objetivos definidos.

Nesse contexto, a introdução de novas variáveis é um aspecto notável da transformação dos dados. A análise da diferença entre o ano de início da graduação e o ano de conclusão do ensino médio, por exemplo, resultou nas três variáveis alvo centrais desta pesquisa: “mediana_tempo_graduacao”, “media_tempo_graduacao” e “moda_tempo_graduacao”. Essas variáveis segmentam os estudantes em intervalos de anos específicos e desempenham um papel fundamental na análise do desempenho dos cursos, permitindo-nos compreender melhor os fatores que influenciam o tempo de graduação. Além disso, houve uma agregação dos dados para representar faixas etárias e áreas de conhecimento. Para categorizar essas áreas, recorreu-se à CINE Brasil.

A sumarização dos dados é um componente fundamental nessa fase. A sumarização dos dados, que envolveu a agregação dos discentes em cursos, foi uma medida implementada devido à impossibilidade de associar individualmente os discentes, uma vez que as diretrizes da LGPD requerem a proteção da privacidade dos dados pessoais dos discentes. Durante o processo, as instâncias com o mesmo código de curso (“CO_CURSO”) foram combinadas, resultando na criação de um único registro para cada curso.

No caso dos valores ausentes que persistiram após o pré-processamento, a impossibilidade de vinculá-los a indivíduos levou a uma abordagem específica. Esses valores ausentes não foram diretamente considerados na sumarização das variáveis socioeconômicas. No entanto, os dados dos estudantes com valores válidos para outras variáveis não afetadas pela ausência foram integralmente incorporados ao processo de sumarização. Isso indica que, mesmo sem intervenção direta nos dados ausentes, o processo de sumarização já considerava os dados existentes dos estudantes em outras dimensões. O problema foi resolvido, uma vez que o processo de sumarização considera apenas os dados existentes.

A etapa de transformação culminou na criação de um conjunto de dados composto por 121 atributos distintos. Entre esses atributos, três variáveis desempenham um papel de destaque: "moda_tempo_graduacao," "media_tempo_graduacao," e "mediana_tempo_graduacao." Essas variáveis foram geradas a partir da sumarização dos discentes que compartilham o mesmo código de curso. A criação dessas três variáveis estatísticas de centralidade foi motivada pela necessidade de fornecer uma representação resumida e significativa do tempo de graduação dos cursos. Ao avaliar a diferença entre o ano de realização do ENADE e o ano de início da graduação para os discentes de cada curso, as medidas de moda, média e mediana foram selecionadas como indicadores fundamentais.

Para lidar com possíveis desequilíbrios nos dados, foi aplicada a técnica de *undersampling*, a qual é específica para os modelos de classificação deste estudo. O objetivo do *undersampling* é lidar com a diferença entre o número de observações em diferentes classes, visando alcançar uma distribuição mais equilibrada. Isso é alcançado ao reduzir a quantidade de exemplos na classe majoritária, evitando qualquer viés em favor da classe dominante (Chawla et al., 2002).

Com essa transformação, os dados estão prontos para a fase de mineração de dados, na qual os modelos serão aplicados para extrair conhecimento e *insights* relevantes.

EXPERIMENTOS DE APRENDIZADO DE MÁQUINA

A quarta etapa do processo de KDD, conhecida como mineração de dados, constitui um momento na pesquisa em que os dados preparados ao longo das etapas anteriores são submetidos a métodos e técnicas que visam à descoberta efetiva de conhecimento. Essas técnicas exploram os dados para identificar padrões e relações, com o intuito de gerar *insights*. Incorporando técnicas estatísticas, de inteligência artificial e de aprendizado de máquina, essa etapa tem como objetivo principal a extração de conhecimento (Fayyad et al.,1996).

Os experimentos de aprendizado de máquina concentram-se na análise e modelagem do tempo de graduação por curso, utilizando as três medidas centrais — moda, média e mediana — como alvos de predição. Essa abordagem reflete a relevância do contexto educacional.

A escolha de usar moda, média e mediana como alvos de predição é fundamentada em critérios objetivos neste estudo, visando capturar a diversidade temporal dos tempos de graduação por curso. Cada uma dessas medidas de tendência central oferece uma perspectiva única sobre o tempo de graduação dos cursos. A moda representa o valor mais frequente, o que nos ajuda a identificar o tempo de graduação que ocorre com maior regularidade em cada curso. A média oferece um ponto central, proporcionando uma visão geral do tempo médio de graduação. Por fim, a mediana é o valor intermediário que nos dá uma ideia da tendência central, ajudando a entender os tempos de graduação de forma mais equilibrada.

Além disso, a escolha dessas medidas de tendência central é justificada pela sua resistência a valores extremos e variações nos dados. Dado o ambiente educacional complexo e a diversidade de fatores que podem influenciar o tempo de graduação, é relevante empregar medidas de centralidade que possam capturar efetivamente essa amplitude de informações.

Portanto, ao utilizar moda, média e mediana como *targets*, busca-se uma compreensão mais completa e robusta do tempo de graduação dos cursos. Dentre as abordagens exploradas, destacam-se os modelos de regressão e classificação, que representam duas perspectivas analíticas distintas. Os modelos de regressão visam compreender e prever as relações numéricas entre variáveis, permitindo estimativas contínuas do tempo de graduação com base em atributos relevantes. Por outro lado, os modelos de classificação se concentram na categorização dos cursos de acordo com as três tendências centrais, fornecendo uma visão discreta das variações do tempo de graduação.

CENÁRIO DE APRENDIZADO DE MÁQUINA SUPERVISIONADO: REGRESSÃO

Essa abordagem é adotada com o propósito de desenvolver e avaliar modelos capazes de prever valores contínuos com base em atributos preditores. No âmbito do presente estudo, as medidas centrais de tendência, como mediana, média e moda, relacionadas ao tempo de graduação dos estudantes nos cursos, foram escolhidas como alvos de predição. Para isso, construiu-se modelos de previsão utilizando algoritmos de aprendizado de máquina baseados na técnica de *boosting*, incluindo o *Regressor de Gradient Boosting* (GBM), *LightGBM* e *CatBoost*. Esses modelos foram implementados por meio da linguagem de programação Python e das bibliotecas *XGBoost*, *LightGBM* e *CatBoost*.

A seleção desses algoritmos, é respaldada por sua eficácia na manipulação de dados complexos, permitindo a captura de relações não lineares, ou seja, relações que podem existir entre as 118 variáveis independentes (atributos preditores) e as três variáveis dependentes (medidas centrais de tendência do tempo de graduação: mediana, média e moda) que não seguem padrões simples e diretos.

O *Regressor de Gradient Boosting* (GBM) se destaca como uma técnica que aproveita a combinação de múltiplos modelos mais simples, construindo iterativamente modelos sucessivos aprimorados com base nos erros dos modelos anteriores. Dessa forma, é possível criar um modelo robusto ao combinar modelos mais simples. Sua capacidade de lidar eficazmente com dados ruidosos (dados com imprecisões ou erros) e complexos (dados com muitas variáveis conectadas) foi destacada por Friedman (2001), o que o torna uma escolha apropriada para tratar dos desafios apresentados pelos dados de tempo de graduação.

O *LightGBM*, desenvolvido pela Microsoft, é uma estrutura de aprendizado de máquina distribuída eficiente que demonstra um desempenho otimizado e uma velocidade notável, conforme enfatizado por Ke et al. (2017). Esse algoritmo é particularmente hábil por sua capacidade de realizar divisões em níveis profundos das árvores de decisão de maneira eficiente, permitindo uma exploração mais detalhada das relações entre os atributos, o que é relevante para compreender os padrões associados ao tempo de graduação dos estudantes nos cursos.

O *CatBoost*, desenvolvido pela Yandex, também se fundamenta na técnica de *boosting*. Além de ser aplicável tanto em tarefas de regressão quanto de classificação, o *CatBoost* se diferencia por suas estratégias que ajustam automaticamente os hiperparâmetros e simplificam a tratativa de valores ausentes. As características distintivas desse algoritmo, incluindo o emprego de árvores simétricas, foram destacadas por Prokhorenkova et al. (2018). Tais aspectos oferecem uma abordagem diferenciada para lidar com a complexidade inerente aos dados do tempo de graduação.

Os modelos de regressão desenvolvidos passaram por um procedimento uniforme de treinamento e avaliação utilizando o método "*train_test_split*" da biblioteca *Scikit-Learn*.

Essa abordagem foi aplicada de forma contínua a todos os modelos de regressão, garantindo uma avaliação equitativa do desempenho dos modelos.

Diante disso, o conjunto de dados foi dividido em duas partes: o conjunto de treinamento e o conjunto de teste. Tal divisão permitiu que os modelos fossem treinados em uma porção dos dados e posteriormente avaliados em dados que não haviam sido utilizados para o treinamento.

A proporção de 20% foi escolhida para separar o conjunto de teste, visando assegurar que uma quantidade significativa de dados não utilizados durante o treinamento fosse empregada para avaliar o desempenho dos modelos. Além disso, o parâmetro "*random_state*" foi fixado em 42, o que possibilitou repetir a mesma divisão entre conjuntos sempre que o processo fosse executado. Essa abordagem garantiu que a comparação entre diferentes modelos e iterações fosse consistente e reprodutível.

Antes de entrar nos detalhes dos experimentos de regressão, é essencial abordar a última etapa do processo de KDD, dedicada à interpretação e avaliação dos resultados obtidos por meio dos modelos construídos (Fayyad et al., 1996). Nesse contexto, as métricas de avaliação desempenham um papel crucial, uma vez que possibilitam a mensuração e comparação dos resultados. No âmbito deste trabalho, centrado na regressão, quatro métricas foram adotadas, que são Erro Quadrático Médio (MSE), Erro Absoluto Médio (MAE), Raiz Quadrada do Erro Quadrático Médio (RMSE) e Coeficiente de Determinação (R^2), utilizando funções da biblioteca *Scikit-Learn*. As escolhas das técnicas de regressão e das métricas de avaliação foram feitas para que o estudo se alinhasse às particularidades dos dados de tempo de graduação e aos objetivos da pesquisa (Montgomery et al., 2012).

CENÁRIO DE APRENDIZADO DE MÁQUINA SUPERVISIONADO: CLASSIFICAÇÃO

O propósito dessa abordagem foi o desenvolvimento e avaliação de modelos capazes de categorizar os cursos em classes específicas, determinadas pelo tempo de graduação. Os modelos foram construídos utilizando a linguagem de programação *Python*, com o uso de dois algoritmos, *XGBoost* e *CatBoost*.

Essas seleções foram feitas de forma estratégica, considerando as características dos algoritmos e sua adequação aos objetivos do estudo. O *CatBoost* foi escolhido devido ao seu desempenho superior nas métricas de erro utilizadas no cenário de regressão anterior. Sua capacidade de lidar com dados complexos e a implementação de árvores simétricas foram consideradas vantagens significativas para o cenário de classificação. Por outro lado, o *XGBoost* foi adotado com base na sua utilização bem-sucedida no estudo de Rodrigues (2021), mesmo considerando as diferenças na natureza dos dados. Essa escolha foi motivada pela sua comprovada eficácia em problemas de aprendizado de máquina.

O *XGBoost*, é um algoritmo baseado na técnica de *boosting*, que se destaca por sua capacidade de lidar com dados complexos, modelar relações não lineares e oferecer um alto desempenho. A abordagem do *XGBoost* é baseada na construção de uma série de árvores de decisão, cada uma corrigindo os erros da anterior. Esse algoritmo também incorpora estratégias para evitar o *overfitting*, que é uma preocupação comum em modelos complexos (Chen & Guestrin, 2016).

Por sua vez, o *CatBoost* é um algoritmo que também se baseia na técnica de *boosting* e apresenta características distintivas que o tornam uma escolha valiosa para problemas de classificação.

O ponto de partida desta etapa consistiu na categorização dos valores contínuos de tempo de graduação em duas classes distintas: “2 a 4 anos” e “5 ou mais anos”. Essas categorias foram estabelecidas com base nas medidas centrais de tendência — “mediana_tempo_graduacao,” “media_tempo_graduacao,” e “moda_tempo_graduacao” — previamente extraídas dos cursos.

Essa abordagem permitiu a transformação do tempo de graduação, inicialmente contínuo, em uma dimensão discreta. O processo de categorização foi uma etapa importante que vinculou as métricas extraídas das tendências centrais às classes de tempo de graduação, tornando possível a modelagem de classificação baseada nos atributos preditores e nas representações discretas dessas *targets*.

Um passo na preparação dos dados para os modelos de classificação foi a aplicação do *LabelEncoder*. Esse procedimento consiste em transformar as categorias em valores numéricos únicos, permitindo que os algoritmos interpretassem essas categorias como valores ordinais. Assim, as categorias “2 a 4 anos” e “5 ou mais anos” foram codificadas em valores numéricos, sendo representadas respectivamente, como 0 e 1, possibilitando o uso eficiente desses dados nos algoritmos de aprendizado de máquina.

Dois experimentos distintos foram realizados nesta etapa, considerando a natureza das classes de tempo de graduação dos estudantes nos cursos e a distribuição dos dados. No primeiro experimento, utilizou-se um conjunto de dados desbalanceado, com 2.431 amostras da classe 0 e 1.705 amostras da classe 1. No segundo experimento, aplicou-se o balanceamento ao conjunto de dados, resultando em uma distribuição igual de classes, com 1.705 amostras para cada classe. Esse balanceamento foi realizado por meio do método de *'undersampling*' da biblioteca '*Imbalanced-Learn*'.

Assim como nos experimentos de regressão, foi aplicado um procedimento uniforme de treinamento e avaliação aos modelos de classificação, utilizando o método

“train_test_split” da biblioteca Scikit-Learn. Essa abordagem foi aplicada a todos os modelos de classificação, garantindo uma avaliação equitativa do desempenho. A proporção de 20% foi novamente escolhida para separar o conjunto de teste. O parâmetro “random_state” foi fixado em 42, garantindo consistência e reprodutibilidade nas avaliações.

As métricas de avaliação utilizadas para mensurar o desempenho dos modelos de classificação incluem a acurácia, precisão, *recall*, *f1-score* e a matriz de confusão. A acurácia, definida como a proporção de previsões corretas em relação ao total de previsões, é uma métrica amplamente adotada na avaliação de modelos de classificação (James et al., 2013). A precisão, que quantifica a proporção de verdadeiros positivos dentre todas as instâncias classificadas como positivas, é uma medida fundamental para avaliar a qualidade das previsões positivas (Provost & Fawcett, 2013). O *recall*, por sua vez, expressa a proporção de verdadeiros positivos entre todas as instâncias que deveriam ter sido classificadas como positivas, sendo particularmente útil quando o foco está na identificação de casos positivos (Powers, 2011). A matriz de confusão, conforme descrita por Sokolova e Lapalme (2009), permite tabular as previsões do modelo em relação aos resultados reais, facilitando a visualização dos verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. Por fim, o *f1-score*, que é uma média harmônica entre a precisão e o *recall*, oferece uma métrica que considera tanto os falsos positivos quanto os falsos negativos, sendo uma medida útil para avaliar o equilíbrio entre essas duas métricas (Powers, 2011).

O processo de construção e avaliação dos modelos de classificação proporcionou uma compreensão aprofundada da capacidade de predição das categorias de tempo de graduação. Os passos adotados, desde a categorização das classes até o uso de métricas de avaliação específicas, foram fundamentais para alcançar resultados alinhados aos objetivos deste estudo.

RESULTADOS

Na seção anterior, descreveu-se o processo de condução dos experimentos de aprendizado de máquina. Nesta seção, são apresentadas e discutidas as conclusões decorrentes desses experimentos, enfatizando as abordagens empregadas e suas implicações para a compreensão do estudo.

CENÁRIO DE APRENDIZADO DE MÁQUINA SUPERVISIONADO: REGRESSÃO

Nesta seção, a discussão concentra-se nos resultados obtidos por meio da abordagem de aprendizado de máquina supervisionado no contexto de regressão. Como mencionado anteriormente, os modelos desenvolvidos têm o objetivo de prever os valores contínuos das medidas centrais de tendência — moda, média e mediana — relacionadas ao tempo de graduação dos estudantes nos cursos dos cursos. Essa análise visa proporcionar *insights* sobre os fatores que influenciam os períodos de graduação e entender as relações subjacentes.

.

Resultados da Regressão - Mediana

Model	MSE	MAE	RMSE	R^2
GBM Regressor	0.6345	0.602	0.7966	43.56%
LightGBM	0.5762	0.5676	0.759	48.76%
CatBoost	0.5226	0.545	0.7229	53.52%

Figura 1. Resultados dos Modelos de Regressão para a Variável Alvo "tempo_graduacao_mediana"
Fonte: o(s) autor(es).

Resultados da Regressão - Média

Model	MSE	MAE	RMSE	R^2
GBM Regressor	0.5394	0.5396	0.7345	47.53%
LightGBM	0.449	0.4828	0.6701	56.32%
CatBoost	0.4102	0.4586	0.6405	60.09%

Figura 2. Resultados dos Modelos de Regressão para a Variável Alvo "tempo_graduacao_media"
Fonte: o(s) autor(es).

Resultados da Regressão - Moda

Model	MSE	MAE	RMSE	R^2
GBM Regressor	0.9273	0.689	0.963	31.62%
LightGBM	0.8734	0.6752	0.9345	35.60%
CatBoost	0.8415	0.6587	0.9173	37.95%

Figura 3. Resultados dos Modelos de Regressão para a Variável Alvo "tempo_graduacao_moda"
Fonte: o(s) autor(es).

Os resultados dos experimentos de regressão, apresentados nas figuras acima, revelaram padrões distintos nas métricas de avaliação para cada algoritmo empregado. É importante ressaltar que as medidas utilizadas — MSE, MAE, RMSE e R^2 — oferecem uma visão abrangente do desempenho dos modelos em diferentes aspectos da previsão.

Ao examinar a variável dependente "mediana_tempo_graduacao", que tem a medida de tendência mediana, os resultados destacam que o algoritmo *CatBoost* obteve o menor valor de MSE (0.5226), indicando a menor dispersão entre as previsões e os valores reais. Além disso, apresentou o menor valor de RMSE (0.7229), apontando para um menor erro médio nas estimativas.

A métrica MAE revela que o *CatBoost* alcançou um valor de 0.5449, demonstrando a menor diferença média entre as previsões e os valores reais. Por sua vez, o R^2 registrou um valor de 0.5352, o que significa que aproximadamente 53.52% da variabilidade dos dados foi capturada pelo modelo.

O modelo *LightGBM* também demonstrou bom desempenho na previsão da mediana do tempo de graduação, com valores de métricas próximos aos do CatBoost. O *GBM Regressor* obteve valores mais altos nas métricas de erro e menor R^2, indicando que esse modelo teve um desempenho relativamente inferior na previsão dessa medida.

No caso da variável "media_tempo_graduacao", novamente o *CatBoost* se destacou, apresentando o menor MSE (0.4103) e RMSE (0.6405). Registrou também um MAE de 0.4586, indicando a menor diferença média entre as previsões e os valores reais. Em relação ao R^2 obteve um valor de 0.6009, sugerindo que aproximadamente 60.09% da variação nos dados foi explicada pelo modelo.

O *LightGBM* demonstrou desempenho competitivo, com valores de métricas próximos ao *CatBoost*. O *GBM Regressor* obteve métricas mais altas de erro e um R^2 mais baixo em comparação aos outros modelos.

Nos testes com a variável dependente "moda_tempo_graduacao", o *CatBoost* manteve seu padrão de bom desempenho, com menor MSE (0.8415) e RMSE (0.9173). O MAE foi de 0.6587, indicando a menor diferença média entre as previsões e os valores reais. No que diz respeito ao R^2, o modelo alcançou 0.3795, sugerindo que cerca de 37.95% da variabilidade é capturada pelo modelo. O *LightGBM* e o *GBM Regressor* continuaram apresentando desempenho competitivo, embora com valores mais altos em todas as métricas em comparação ao *CatBoost*.

Concluindo a análise dos resultados de regressão, observa-se que o algoritmo *CatBoost* se destacou consistentemente, demonstrando sua capacidade de compreender as complexas relações entre os atributos preditores e as medidas centrais de tendência do tempo de graduação. Especificamente, os modelos construídos tendo como variável dependente a variável "media_tempo_graduacao" apresentaram resultados melhores do que os modelos criados com as outras variáveis dependentes (mediana e moda), o que oferece as melhores previsões no contexto em questão. O *LightGBM* também apresentou resultados competitivos, enquanto o *GBM Regressor* mostrou um desempenho um pouco inferior em relação às métricas de erro e R^2.

A Figura 4 ilustra uma comparação entre os 100 primeiros pontos com valores preditos pelo modelo de melhor desempenho, o CatBoost, e os valores reais da medida

"media_tempo_graduacao". Por meio desta representação gráfica, é possível observar a proximidade entre as previsões geradas pelo modelo e os valores reais. Esse aspecto enfatiza o desempenho do algoritmo em realizar estimativas que se aproximam dos resultados reais, nos dados de tempo de graduação.

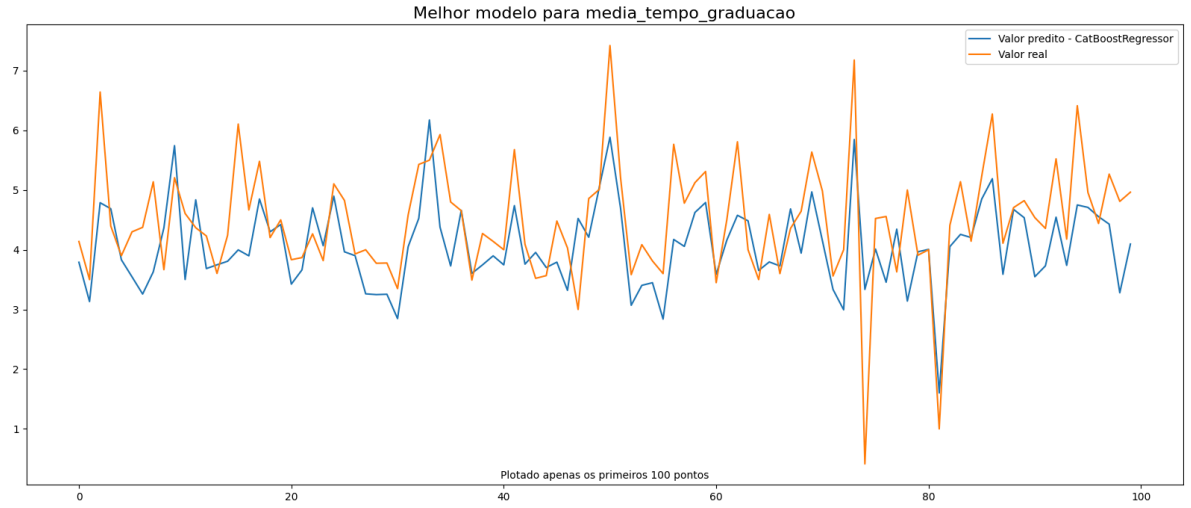

Figura 4. Comparação Entre os Primeiros 100 Pontos dos Valores Preditos pelo Modelo *CatBoost* e Valores Reais de 'média de tempo de graduação'
Fonte: o(s) autor(es).

No conjunto, todos os algoritmos revelaram a habilidade de prever os tempos de graduação dos cursos, contribuindo para a compreensão dos fatores que influenciam esses períodos. Isso sugere a adaptabilidade desses algoritmos à complexidade dos dados educacionais. A escolha estratégica desses algoritmos, baseados na técnica de *boosting*, demonstrou fazer sentido, uma vez que todos se mostraram aptos a lidar com os desafios apresentados pelos dados de tempo de graduação.

CENÁRIO DE APRENDIZADO DE MÁQUINA SUPERVISIONADO: CLASSIFICAÇÃO

Nesta seção, a discussão é centrada nos resultados obtidos por meio da abordagem de aprendizado de máquina supervisionado no contexto de classificação. Como mencionado anteriormente, o objetivo dessa etapa foi categorizar os cursos em grupos específicos com base nas medidas centrais de tendência do tempo de graduação. A análise dos resultados é conduzida com base nas métricas de avaliação, a saber: acurácia, precisão, *recall*, *f1-score* e matriz de confusão.

Durante os experimentos de classificação, empregaram-se dois algoritmos distintos. O primeiro deles, o *XGBoost*, como referido no estudo de Rodrigues (2021), foi escolhido devido à sua aplicabilidade prévia neste campo. O segundo algoritmo escolhido foi o *CatBoost*, que demonstrou um desempenho superior no cenário de regressão do atual estudo.

Além disso, é fundamental destacar a influência do balanceamento dos dados no desempenho dos modelos de classificação. Para uma compreensão mais clara dessa influência na obtenção

dos resultados dos modelos, a Figura 5 ilustra a quantidade de dados em cada classe, antes e depois do processo de balanceamento.

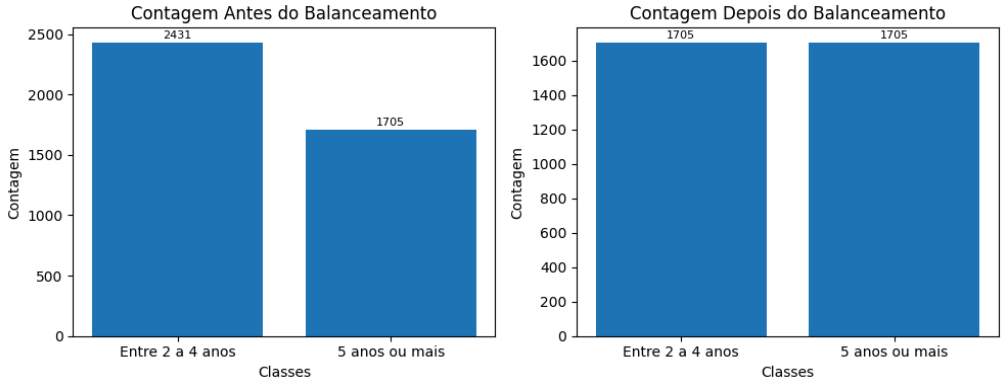

Figura 5. Distribuição de Dados por Classe Antes e Depois do Balanceamento
Fonte: o(s) autor(es).

A seguir, apresentam-se os resultados para o conjunto de dados desbalanceados. A Figura 6 fornece um relatório detalhado da classificação realizada pelo modelo *XGBoost* em relação às variáveis-alvo. Esses relatórios resumem os principais resultados, incluindo as métricas de avaliação mencionadas anteriormente.

Tabela Descritiva - XGBoost

Classe Alvo	Precisão (XGBoost)	Recall (XGBoost)	F1-Score (XGBoost)
Mediana (Entre 2 a 4 anos)	0.78	0.82	0.8
Mediana (5 anos ou mais)	0.73	0.68	0.7
Média (Entre 2 a 4 anos)	0.67	0.56	0.61
Média (5 anos ou mais)	0.83	0.88	0.86
Moda (Entre 2 a 4 anos)	0.79	0.87	0.83
Moda (5 anos ou mais)	0.61	0.48	0.54

Figura 6. Relatório Detalhado da Classificação pelo Modelo *XGBoost* para o Conjunto de Dados Desbalanceados para as três Classes
Fonte: o(s) autor(es).

A Figura 7 apresenta um relatório semelhante, porém para o modelo CatBoost, ainda no contexto do conjunto de dados desbalanceados.

Tabela Descritiva - CatBoost

Classe Alvo	Precisão (CatBoost)	Recall (CatBoost)	F1-Score (CatBoost)
Mediana (Entre 2 a 4 anos)	0.78	0.86	0.82
Mediana (5 anos ou mais)	0.77	0.67	0.72
Média (Entre 2 a 4 anos)	0.69	0.53	0.6
Média (5 anos ou mais)	0.82	0.9	0.86
Moda (Entre 2 a 4 anos)	0.8	0.91	0.85
Moda (5 anos ou mais)	0.71	0.47	0.56

Figura 7. Relatório Detalhado da Classificação pelo Modelo *CatBoost* para o Conjunto de Dados Desbalanceados para as três Classes
Fonte: o(s) autor(es).

Além disso, é efetuada uma análise das matrizes de confusão e dos valores de acurácia para os modelos, considerando suas variáveis-alvo. Essas matrizes fornecem uma visão do desempenho na classificação das classes, auxiliando a compreensão de como os cursos são

corretamente classificados em cada categoria, bem como a identificação de possíveis erros de classificação. A Figura 8 apresenta as matrizes de confusão e acurácia para cada variável-alvo dos modelos *XGBoost*, tendo em vista seu conjunto de dados desbalanceado, oferecendo informações relevantes sobre o desempenho da categorização alcançado por esses modelos.

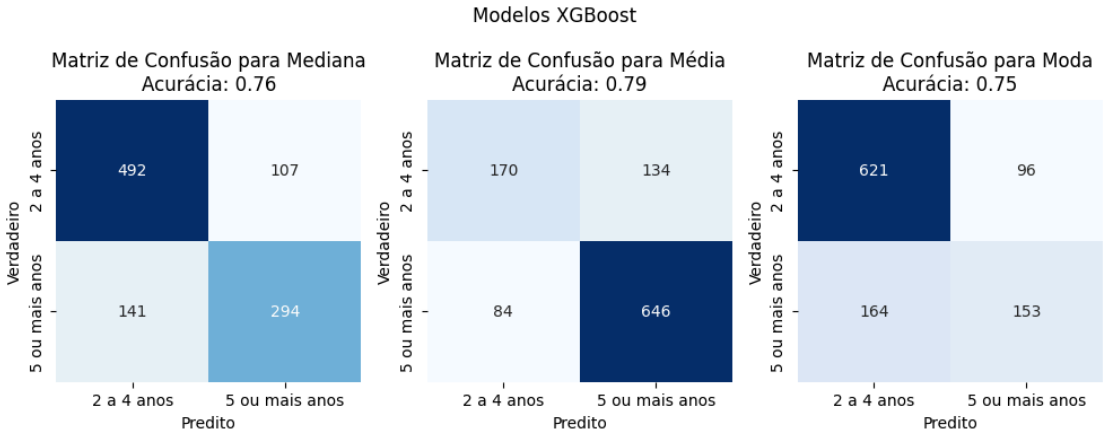

Figura 8. Matrizes de Confusão e Acurácia para Modelos *XGBoost* com Dados Desbalanceado para as três Classes
Fonte: o(s) autor(es).

A Figura 9 complementa essa análise, mostrando a matriz de confusão e acurácia correspondente aos valores das variáveis-alvo dos modelos *CatBoost* no mesmo conjunto de dados desbalanceado.

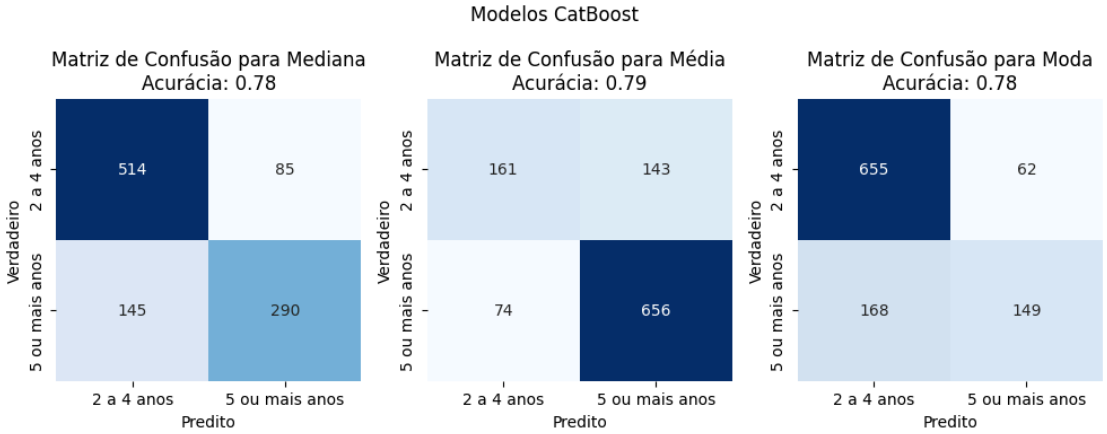

Figura 9. Matrizes de Confusão e Acurácia para Modelos *CatBoost* com Dados Desbalanceado para as três Classes
Fonte: o(s) autor(es).

A próxima etapa da discussão concentra-se nos resultados para o conjunto de dados balanceado.

No cenário balanceado, os algoritmos *XGBoost* e *CatBoost* foram submetidos aos mesmos procedimentos de treinamento e teste. As métricas de avaliação, incluindo acurácia, precisão, *recall*, *f1-score* e matriz de confusão, foram novamente utilizadas para medir o desempenho dos modelos na tarefa de classificação. A análise dos resultados no conjunto de dados balanceado, ilustrada na Figura 10, revela diferenças em relação ao cenário desbalanceado. É

importante destacar que a mitigação do desbalanceamento das classes teve um impacto direto no desempenho dos modelos.

Tabela Descritiva Depois do Balanceamento - XGBoost

Classe Alvo	Precisão (XGBoost)	Recall (XGBoost)	F1-Score (XGBoost)
Mediana (Entre 2 a 4 anos)	0.82	0.77	0.79
Mediana (5 anos ou mais)	0.71	0.77	0.74
Média (Entre 2 a 4 anos)	0.54	0.73	0.62
Média (5 anos ou mais)	0.87	0.75	0.8
Moda (Entre 2 a 4 anos)	0.84	0.73	0.78
Moda (5 anos ou mais)	0.53	0.69	0.6

Figura 10. Relatório Detalhado da Classificação pelo Modelo *XGBoost* com Dados Balanceados para as três Classes.
Fonte: o(s) autor(es).

Em seguida, na Figura 11, apresentam-se os valores das métricas de avaliação referentes ao conjunto de dados balanceado, utilizando o algoritmo *CatBoost*.

Tabela Descritiva Depois do Balanceamento - CatBoost

Classe Alvo	Precisão (CatBoost)	Recall (CatBoost)	F1-Score (CatBoost)
Mediana (Entre 2 a 4 anos)	0.82	0.77	0.79
Mediana (5 anos ou mais)	0.71	0.76	0.73
Média (Entre 2 a 4 anos)	0.53	0.72	0.61
Média (5 anos ou mais)	0.86	0.73	0.79
Moda (Entre 2 a 4 anos)	0.85	0.74	0.79
Moda (5 anos ou mais)	0.55	0.7	0.62

Figura 11. Relatório Detalhado da Classificação pelo Modelo *CatBoost* com Dados Balanceados para as três Classes
Fonte: o(s) autor(es).

A fim de proporcionar uma visão mais abrangente do desempenho, as Figuras 12 e 13 exibem as matrizes de confusão correspondentes a cada modelo, acompanhadas das respectivas acurácias.

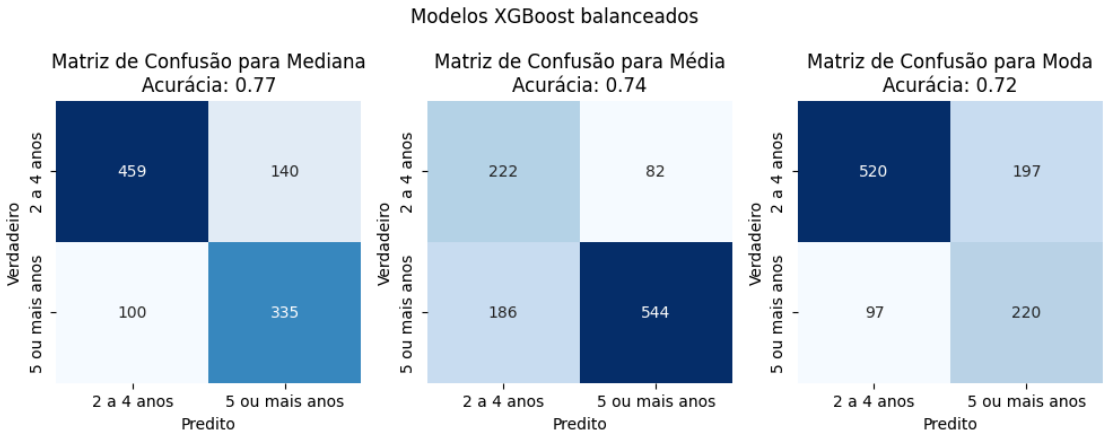

Figura 12. Matrizes de Confusão e Acurácia para Modelos *XGBoost* com Dados Balanceado para as três Classes
Fonte: o(s) autor(es).

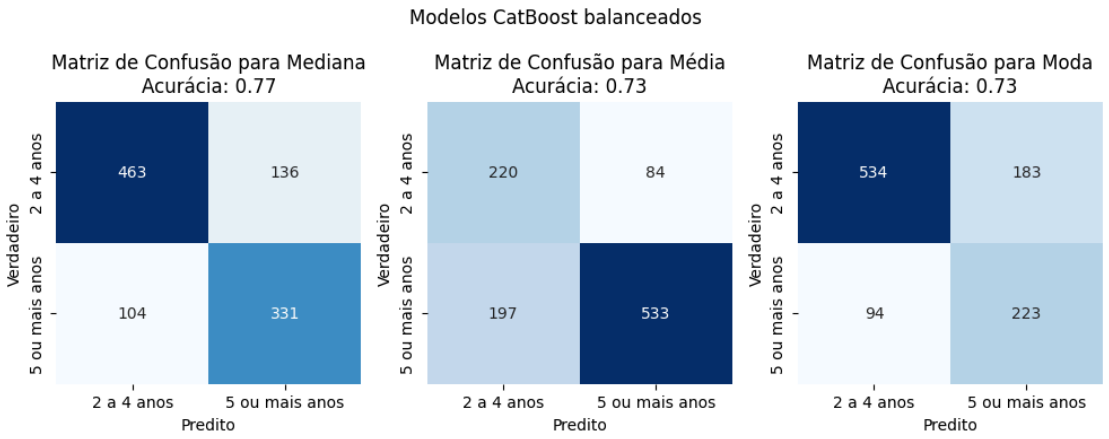

Figura 13. *CatBoost* com Dados Balanceado para as três Classes

Ao analisar os resultados, é evidente que tanto o *XGBoost* quanto o *CatBoost* demonstraram a capacidade de compreender as características dos cursos com base nas medidas de tendência central do tempo de graduação, tanto no cenário desbalanceado quanto no balanceado.

No cenário desbalanceado, ambos os algoritmos tiveram resultados satisfatórios, com o *XGBoost* obtendo uma compreensão adequada das categorias relacionadas a cerca de 2 a 4 anos e 5 anos ou mais de duração do curso. Por outro lado, o *CatBoost* demonstrou consistência, mantendo um desempenho notável nas três tendências centrais: mediana, média e moda.

No cenário balanceado, os modelos continuaram a demonstrar seu potencial. O *CatBoost*, em particular, manteve sua robustez, registrando acurácias consistentes e bons níveis de precisão e recall nas três tendências centrais. O *XGBoost*, embora tenha mostrado uma melhoria modesta em relação ao cenário desbalanceado, ainda alcançou resultados competitivos. Esses resultados encorajam a aplicação de algoritmos de aprendizado de máquina em estudos semelhantes, com o objetivo de descobrir conhecimentos adicionais no contexto da educação e tempo de graduação.

CONSIDERAÇÕES FINAIS

Ao avaliar os resultados deste estudo, constata-se que os objetivos iniciais foram alcançados. A aplicação de algoritmos de aprendizado de máquina para investigar o tempo de graduação em cursos universitários revelou-se uma estratégia eficaz e promissora. A condução de experimentos abrangentes, tanto no cenário de regressão quanto no de classificação, permitiu uma compreensão das dinâmicas que envolvem o tempo de graduação. Isso é relevante, pois esse fenômeno educacional é de suma importância para as instituições de ensino superior e as políticas públicas relacionadas à educação.

A análise das medidas de tendência central (mediana, média e moda) como variáveis-alvo em modelos de regressão oferece *insights* valiosos sobre como diferentes características dos

cursos podem influenciar o tempo necessário para a conclusão da graduação. Isso, por sua vez, pode servir como base para a otimização dos currículos acadêmicos e o desenvolvimento de estratégias para reduzir os períodos de graduação. Considerando os resultados apresentados, é justificável a continuidade de pesquisas com objetivos semelhantes.

Além disso, é importante considerar os aspectos éticos relacionados à aplicação de algoritmos de aprendizado de máquina em contextos educacionais. A utilização de dados sensíveis dos alunos deve ser feita com cautela, garantindo a privacidade e a segurança das informações. É imperativo que as instituições de ensino adotem políticas robustas de proteção de dados e que os pesquisadores estejam cientes e comprometidos com os princípios éticos ao realizar análises com IA.

Estudos futuros podem tratar de análises mais profundas sobre dados educacionais, explorando outras variáveis e abordagens de aprendizado de máquina. Esta pesquisa pode servir como base para o desenvolvimento de políticas públicas e estratégias institucionais que visem melhorar a qualidade do ensino superior e reduzir os índices de retenção e evasão, contribuindo para uma educação mais acessível e eficaz no Brasil.

REFERÊNCIAS

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). **Introduction to linear regression analysis (5th ed.). Wiley.** Disponível em: <https://ocd.lcwu.edu.pk/cfiles/Statistics/Stat-503/IntroductiontoLinearRegressionAnalysisby >. Acesso em: 22 de agosto de 2023.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). **SMOTE: Synthetic minority over-sampling technique.** Journal of Artificial Intelligence Research, 16, 321-357. Disponível em: <https://www.jair.org/index.php/jair/article/view/10302 >. Acesso em: 20 de agosto de 2023.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer. Provost, F., & Fawcett, T. (2013). **Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly Media.** Disponível em: <https://books.google.com.br/books?hl=pt-BR&lr=&id=EZAtAAAAQBAJ&oi=fnd&pg=P1&dq=Provost,+F.,+%26+Fawcett,+T.+(2013)&ots=yl2KVp1Sy_&sig=rN5E0nZGOmIiWIbZyTopkESnQ9Y#v=onepage&q=Provost%2C%20F.%2C%20%26%20Fawcett%2C%20T.%20(2013)&f=false>. Acesso em 18 de agosto de 2023.

Powers, D. M. (2011). **Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation.** Journal of Machine Learning Technologies, 2(1),37-63. Disponível em: <https://arxiv.org/abs/2010.16061>. Acesso em: 20 de agosto de 2023.

Friedman, J. H. (2001). **Greedy function approximation: a gradient boosting machine. Annals of statistics**, 29(5), 1189-1232. Disponível em: <https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full.>. Acesso em: 18 de agosto de 2023.
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). **LightGBM: A highly efficient gradient boosting decision tree. In Advances in neural information processing systems** (pp. 3146-3154). Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.>. Acesso em: 16 de agosto de 2023.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). **CatBoost: unbiased boosting with categorical features. In Advances in neural information processing systems** (pp. 6638-6648). Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf.>. Acesso em: 15 de julho de 2023.

EqualFrequencyDiscretiser. **Feature Engine, Read the Docs, 2020**. Disponível em <https://feature-engine.readthedocs.io/en/latest/discretisation/EqualFrequencyDiscretiser.html>. Acesso em: 15 de julho de 2023.

FRAWLEY, William J.; PIATETSKY-SHAPIRO, Gregory; MATHEUS, Christopher J. **Knowledge Discovery in Databases: An Overview.** AI MAGAZINE, v. 13, nº 3, p. 57, 1992. Disponível em <https://ojs.aaai.org//index.php/aimagazine/article/view/1011.>

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. **From Data Mining to Knowledge Discovery in Databases.** AI MAGAZINE, v. 17, nº 3, p. 37, 1996. Disponível em <https://ojs.aaai.org//index.php/aimagazine/article/view/1230.>

Microdados do Censo da Educação Superior. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, INEP, 2021-7. Disponível em <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior.>. Acesso em 13 de julho de 2023.

BILOGUR, Aleksey. **Undersampling and oversampling imbalanced data**. Kaggle, 2018. Disponível em https://www.kaggle.com/residentmario/undersampling-and-oversampling-imbalanced-data. Acesso em: 13 de julho de 2023.

Sklearn.preprocessing.**OneHotEncoder**. Scikit-Learn, Disponível em https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html. Acesso em: 13 de julho de 2023.

Classification: Accuracy. Machine Learning Crash Course, Google Developers. Disponível em https://developers.google.com/machine-learning/crash-course/classification/accuracy. Acesso em: 13 de julho de 2023.

Classification: Precision and Recall. Machine Learning Crash Course, Google

Developers. Disponível em https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall. Acesso em: 13 de julho de 2023.

Sklearn.metrics.f1_score. Scikit-Learn. Disponível em https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html.

Rodrigues, Ebony Marques. **Técnica de Aprendizado de Máquina para Descoberta de Conhecimento sobre Dados Abertos do Ensino Superior Público Brasileiro.** Trabalho de Conclusão de Curso. Universidade Federal Rural de Pernambuco, 2021.

SCIKITLEARN. Sklearn.preprocessing.**OneHotEncoder**. 2023. ScikitLearn. < https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html >. Acesso em 27 de junho de 2023.

Pandas Development Team. pandas.DataFrame.merge. In: pandas 1.5.3 documentation. 2023. Disponível em: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.merge.html. Acesso em: 27 de junho de 2023.

FOLHA DE SÃO PAULO. **Inep tira do ar informações detalhadas sobre alunos e professores do Censo.** Folha de São Paulo, São Paulo, 9 fev. 2022. Disponível em: https://www1.folha.uol.com.br/educacao/2022/02/inep-tira-do-ar-informacoes-detalhadas-sobre-alunos-e-professores-do-censo.shtml. Acesso em: 1 de maio de 2023.

INEP. Nota técnica nº 21/2022-DAES: **Orientações técnicas para o preenchimento do questionário do estudante referente ao Exame Nacional de Desempenho dos Estudantes (ENADE) 2022. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira**, 2022. Disponível em: https://download.inep.gov.br/microdados/nota_tecnica_21-2022-daes.pdf. Acesso em: 1 de maio de 2023.

GERON, Aurélien**. Aprendizado de Máquina com Scikit-Learn e TensorFlow: Conceitos, Ferramentas e Técnicas para Construir Sistemas Inteligentes.** São Paulo: Novatec, 2017. p. 63.

HAN, J.; KAMBER, M. **Data Mining: Concepts and Techniques.** 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2006.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques.** 3rd ed. Morgan Kaufmann, 2012.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. **Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet).** Diário Oficial da União, Brasília, DF, 15 ago. 2018. Seção 1, p. 1. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm. Acesso em: 29 de abril de 2023.

Questionário do Estudante do ENADE. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, INEP, 2021-5. Disponível em https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enade. Acesso em 29 de março de 2023.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **Parecer nº 00018/2022/PF**. Brasília: INEP, 18 de fevereiro de 2022. Disponível em: https://www.gov.br/inep/pt-br/assuntos/noticias/institucional/nota-de-esclarecimento-divulgacao-dos-microdados. Acesso em: 18 de março de 2023.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **Microdados Enade 2018**. Brasília: Inep, 2022. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enade. Acesso em: 15 de março de 2023.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **Cine Brasil.** Brasília: Inep, 2022. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/cine-brasil/classificacao. Acesso em: 10 de março de 2023.

APÊNDICE

Apêndice 1 - Dicionário de Dados
Apêndice 2 - Dicionário de Dados de atributos utilizado por Rodrigues (2021)