UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Previsão de evasão de estudantes que preserva a privacidade usando redes de informação heterogêneas

Geraldo Nunes Corrêa

Trabalho de Conclusão de Curso
MBA em Inteligência Artificial e Big Data

ICMC USP
SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Previsão de evasão de estudantes que preserva a privacidade usando redes de informação heterogêneas

Geraldo Nunes Corrêa

USP - São Carlos
2023

Geraldo Nunes Corrêa

Previsão de evasão de estudantes que preserva a privacidade usando redes de informação heterogêneas

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Ricardo M. Marcacini

USP - São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

N972p

Nunes Corrêa, Geraldo
Previsão de abandono de alunos que preserva a privacidade usando redes de informação heterogêneas / Geraldo Nunes Corrêa; orientador Ricardo Marcondes Marcacini. -- São Carlos, .
p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, .

1. MOOC. 2. Evasão de Alunos. 3. LGPD. 4. Redes Complexas. 5. Classificação. I. Marcondes Marcacini, Ricardo , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

DEDICATÓRIA

A minha família pela compreensão, carinho e apoio incansável.

AGRADECIMENTOS

À Coordenação do MBA que me proporcionou uma bolsa de estudos para a realização do curso

À Profa. Dra. Solange Resende Rodrigues, que me fez acreditar que seria possível fazer uma especialização na área de Inteligência Artificial e Big Data.

Ao Prof. Ricardo Marcondes Marcacini, pela inestimável contribuição no desenvolvimento do trabalho.

EPÍGRAFE

“A persistência é o caminho do êxito.”
(Charles Chaplin)

RESUMO

CORRÊA, G. N. **Previsão de evasão de estudantes que preserva a privacidade usando redes de informação heterogêneas**. 2023. 44f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

Os métodos de previsão de evasão escolar têm ganhado importância nas instituições de ensino por permitirem a identificação precoce de estudantes com alto risco de evasão, permitindo intervenções oportunas. Os modelos tradicionais de previsão da evasão escolar baseiam-se frequentemente em informações sensíveis dos estudantes, o que levanta questões de privacidade. O presente projeto propõe o método de Predição de Evasão de Estudantes com Preservação de Privacidade (*Privacy-Preserving Student Dropout Prediction* - P²SDP), que utiliza uma abordagem baseada em grafos para modelar um banco de dados histórico como uma rede heterogênea sem incorporar dados confidenciais. Ao aproveitar a estrutura gráfica e aplicar uma estrutura de regularização, o P²SDP obtém previsões de risco de evasão e as discretiza em níveis de risco baixo, médio e alto. Uma avaliação experimental foi conduzida usando um conjunto de dados reais de uma plataforma de ensino a distância, demonstrando que o P²SDP alcança desempenho competitivo enquanto preserva a privacidade dos estudantes. Além disso, como o P²SDP utiliza apenas registros de ações do usuário, ele pode ser facilmente aplicado a diversas plataformas online.

Palavras-chave: MOOC; Predição de Evasão de Estudantes; Redes Complexas; Classificação

ABSTRACT

CORRÊA, G. N. **Privacy-Preserving Student Dropout Prediction Using Heterogeneous Information Networks.** 2023. 44f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

Student dropout prediction methods have gained importance in educational institutions by enabling early identification of students at high risk of dropping out, allowing timely interventions. Traditional models for dropout prediction often rely on sensitive student information, raising privacy concerns. This project proposes the Privacy-Preserving Student Dropout Prediction (P²SDP) method, which utilizes a graph-based approach to model a historical database as a heterogeneous network without incorporating sensitive data. By leveraging the graph structure and applying a regularization framework, P²SDP obtains dropout risk predictions and discretizes them into low, medium, and high-risk levels. An experimental evaluation was conducted using a real-world dataset from a distance learning platform, demonstrating that P²SDP achieves competitive performance while preserving student privacy. Furthermore, as P²SDP only utilizes user action logs, it can be easily applied to various online platforms.

Keywords: MOOC; Student Dropout Prediction; Complex Networks; Classification

LISTA DE ILUSTRAÇÕES

LISTA DE EQUAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

LGPD	-	Lei Geral de Proteção de Dados
P2SDP	-	Privacy-Preserving Student Dropout Prediction
MOOC	-	Massive Open Online Courses
DPO	-	Data Protection Officer
ANPD	-	Autoridade Nacional de Proteção de Dados ()
GCNs	-	Graph Convolutional Networks
SIGKDD	-	Special Interest Group on Knowledge Discovery and Data Mining
SVM	-	Support Vector Machines

Sumário

1 INTRODUÇÃO

Os métodos de previsão de evasão escolar têm se tornado cada vez mais importantes nas instituições educacionais (Kumar et al. 2017, Krüger et al. 2023). Estes métodos permitem a identificação precoce de estudantes com alto risco de evasão escolar (Oqaidi et al. 2022). Tradicionalmente, os modelos de previsão da evasão escolar têm sido categorizados em dois tipos: modelos exploratórios e modelos preventivos. Os modelos exploratórios analisam dados históricos para identificar estudantes que já abandonaram, com o objetivo de organizar e extrair padrões (Dol e Jawandhiya 2023).

Por outro lado, os modelos preventivos aproveitam dados históricos para classificar os estudantes atuais em diferentes níveis de risco de evasão, proporcionando intervenções oportunas para prevenir a evasão (Márquez-Vera et al. 2016). Embora estes modelos tenham provado ser ferramentas valiosas para as instituições de ensino na melhoria das taxas de retenção de estudantes e na melhoria da experiência global de aprendizagem, a sua formação é desafiadora, pois requer a identificação de variáveis ou características dos estudantes, tais como fatores socioeconômicos e padrões de acesso à plataforma online, que pode estimar o risco de evasão (Chen et al. 2022).

Em ambientes de ensino à distância, inúmeras abordagens têm sido propostas para a previsão da evasão escolar, incluindo algoritmos tradicionais de aprendizagem automática, métodos estatísticos e modelos baseados em grafos (Kumar et al. 2017, Chen et al. 2022). Estes modelos utilizam frequentemente uma combinação de dados socioeconômicos e informações sobre o envolvimento dos estudantes com a plataforma de aprendizagem online para treinar os seus modelos (Mubarak et al, 2021).

Entre essas abordagens, os modelos baseados em grafos ganharam atenção significativa devido à sua interpretabilidade e forte desempenho preditivo (Karimi et al. 2020, Mubarak et al. 2022). Ao aproveitar a estrutura inerente e a conectividade dos dados, os modelos baseados em grafos podem capturar relações e dependências complexas dentro do ambiente de aprendizagem, permitindo previsões precisas dos riscos de evasão (Nitta et al. 2021).

No entanto, a utilização de dados privados ou sensíveis dos estudantes, como informações socioeconômicas, levanta preocupações relativamente à proteção dos dados dos utilizadores (Erickson 2018). A introdução de legislações como o *General Data Protection Regulation* (GDPR), para a comunidade europeia, e a Lei Geral de Proteção de Dados (LGPD), para o Brasil, tendem a tornar impraticável o uso dos atuais modelos preditivos.

Embora possam existir dados históricos contendo informações privadas disponíveis para modelos de formação, os futuros estudantes podem optar por não fornecer os seus dados privados. Portanto, é importante explorar modelos de previsão de evasão escolar que possam funcionar de forma eficaz sem depender de informações privadas dos estudantes (Deho et al, 2022).

Para enfrentar o desafio mencionado acima, este trabalho propõe o método de previsão de evasão de estudantes com preservação de privacidade - *Privacy-Preserving Student Dropout Prediction* (P^2SDP). O método P²SDP consiste em três etapas principais. Primeiro, um banco de dados histórico de estudantes e seus registros de acesso é modelado como uma rede heterogênea. Nessa representação baseada em grafos, estudantes, cursos, matrículas e ações no sistema de ensino a distância são representados como nós, enquanto as relações entre eles são representadas como arestas. Este gráfico captura as interações e dependências complexas dentro do ambiente de aprendizagem.

Na segunda etapa, um subconjunto de estudantes é rotulado com base na sua situação de evasão, levando em consideração os registros históricos. Esta rotulagem inicial serve como um conjunto de treinamento para a tarefa de classificação. Para classificar os estudantes restantes, é aplicada uma estrutura de regularização baseada na propagação de rótulos, aproveitando a topologia do grafo e os estudantes inicialmente rotulados. O processo de propagação de rótulos utiliza a conectividade do grafo para propagar rótulos e estimar riscos de evasão para os estudantes não rotulados.

Na terceira etapa, as pontuações de confiança para a classe de evasão são discretizadas para organizar as previsões em três níveis de risco de evasão: baixo, médio e alto. Isto permite que os gestores da plataforma monitorem e priorizem quais os estudantes que necessitam de apoio adicional para evitar a evasão.

Uma avaliação experimental foi realizada usando um conjunto de dados do mundo real obtido de uma plataforma chinesa de cursos online abertos e massivos (MOOC), que consiste em mais de 4 milhões de ações realizadas por aproximadamente 4.290 estudantes em vários cursos (Feng et al, 2019). Para avaliar o desempenho do método P^2SDP, uma comparação foi feita com um modelo de referência que se baseia em informações privadas dos estudantes.

Os resultados da avaliação demonstraram que o método P²SDP alcançou desempenho competitivo ao mesmo tempo em que garantiu a privacidade dos dados dos estudantes. Além disso, como o método P²SDP utiliza apenas registros de ações do usuário, ele pode ser facilmente aplicado a qualquer plataforma online que registre transações de acesso aos recursos do curso.

2 REVISÃO BIBLIOGRÁFICA

2.1 Lei Geral de Privacidade de Dados

A ascensão exponencial da tecnologia digital trouxe consigo inúmeras conveniências, mas também desafios significativos em relação à privacidade dos dados pessoais. Nesse contexto, a LGPD emerge como um marco normativo crucial, orientado a redefinir as práticas de tratamento de dados no contexto brasileiro (Souza e Teixeira, 2020). Esta seção destaca os principais aspectos e implicações da LGPD, delineando seu papel como guardiã dos direitos individuais e promulgadora de práticas transparentes e éticas no ecossistema digital.

Porém, antes de realizar uma breve descrição dos principais aspectos e implicações da LGPD, é importante apontar que, conforme a linha do tempo de sua criação abaixo, relacionar à sua importância com relação a este trabalho. Como pode ser observado na Figura 1, a LGPD entrou em vigor a partir de agosto de 2021, o que torna organizações, públicas e privadas, vulneráveis a sanções por parte do judiciário brasileiro.

Figura 1 - Linha do tempo da LGPD

Fonte: Softwall (2023).

Princípios sólidos fundamental a LGPD e resguardam os direitos dos titulares dos dados (Pestana, 2020). A transparência, a finalidade específica, a necessidade, a livre acesso, a qualidade dos dados, a segurança, a prevenção, a não discriminação e a responsabilização compõem a base ética do tratamento de dados estabelecida pela lei. Esses princípios não apenas

asseguram a proteção dos dados pessoais, mas também buscam promover a confiança entre os titulares e as entidades que os manipulam.

Uma série de direitos que são conferidos aos titulares representam um avanço significativo na autonomia sobre seus dados pessoais (de Melo, 2022). O direito à confirmação da existência de tratamento, à correção de informações, à eliminação de dados desnecessários e ao acesso facilitado a informações sobre o tratamento são exemplos dessa capacitação do indivíduo. Além disso, o direito de portabilidade e o direito de revogar consentimento conferem um nível inédito de controle sobre o destino e a utilização de informações pessoais.

A LGPD estabelece obrigações claras tanto para os controladores quanto para os operadores de dados. Os controladores são responsáveis por determinar as finalidades e meios do tratamento, enquanto os operadores realizam o processamento em nome dos controladores (Ghisleni, 2022). Ambos devem garantir a segurança dos dados e adotar medidas técnicas e administrativas capazes de proteger a privacidade. A nomeação do Encarregado de Proteção de Dados (*Data Protection Officer* - DPO) é uma exigência significativa, consolidando a responsabilidade e expertise na gestão da privacidade.

Sanções severas estão previstas em caso de violações, incluindo multas substanciais que podem atingir até 2% do faturamento anual da empresa, limitadas a R$ 50 milhões por infração. A Autoridade Nacional de Proteção de Dados (ANPD) é a entidade responsável por fiscalizar e aplicar penalidades, estabelecendo assim um ambiente regulatório que incentiva a conformidade e a adoção de boas práticas de privacidade.

A LGPD surge como um farol orientador em meio às complexidades da era digital. Sua implementação efetiva não apenas cumpre com exigências legais, mas também sinaliza a necessidade de uma cultura de privacidade digital, onde o respeito pelos direitos individuais e a responsabilidade ética tornam-se elementos centrais no desenvolvimento e na utilização de tecnologias que moldam nosso cotidiano. Nesse contexto, o presente trabalho explora o impacto da LGPD e busca contribuir para uma compreensão mais profunda de suas implicações na proteção da privacidade dos dados pessoais.

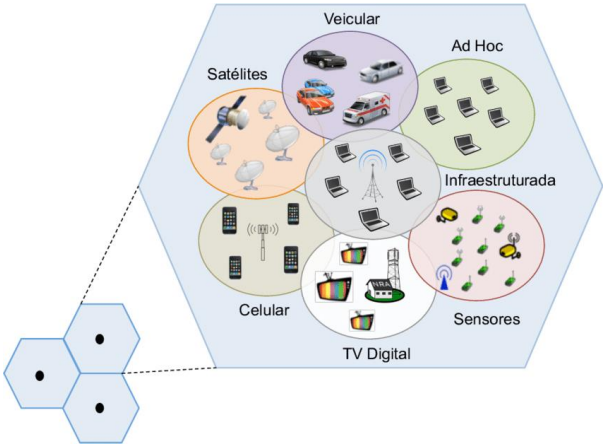
2.2 Redes Heterogêneas

As redes heterogêneas emergem como uma resposta inovadora à crescente complexidade dos dados, especialmente quando estes envolvem uma diversidade estrutural significativa. Nesta seção, serão apresentados os princípios fundamentais, desafios e aplicações

que tornam as redes heterogêneas uma abordagem valiosa na representação de sistemas complexos e multifacetados (Ramesh et al, 2023).

Ao contrário das redes homogêneas, as redes heterogêneas reconhecem e incorporam a heterogeneidade inerente em sistemas do mundo real. Elas permitem a representação de diferentes tipos de entidades (nós) e relações (arestas) em uma única estrutura, proporcionando uma visão mais abrangente e precisa dos sistemas complexos (Wang, 2022). Essa capacidade de modelar diversidade estrutural é essencial em campos que vão desde ciência da computação até biologia e ciências sociais (Figura 3).

Figura 2 - Redes Complexas



Fonte: Researchgate (2023)

Apesar de suas vantagens, as redes heterogêneas apresentam desafios únicos. A heterogeneidade implica em diferentes características e dinâmicas para cada tipo de nó e aresta, tornando a definição de métodos de análise e aprendizado mais complexa. A interpretação e a extração de informações úteis de redes que englobam uma gama diversificada de entidades exigem técnicas avançadas de processamento e modelagem (Alhashimi, 2023).

As redes heterogêneas encontram aplicação em uma variedade de setores. Na área da saúde, por exemplo, elas são utilizadas para modelar interações entre genes, proteínas e compostos em sistemas biológicos complexos. Em redes de transporte, redes heterogêneas são empregadas para representar diferentes modos de transporte e suas interações. Essa versatilidade destaca a capacidade das redes heterogêneas de lidar com a complexidade inerente a sistemas do mundo real.

Uma característica distintiva das redes heterogêneas é sua capacidade de modelagem dinâmica, permitindo a adaptação a mudanças na estrutura da rede ao longo do tempo. Essa flexibilidade é crucial em cenários nos quais as relações entre entidades evoluem, garantindo

que a rede mantenha sua relevância e precisão em face de transformações estruturais (Wang 2022).

À medida que se depara com dados cada vez mais heterogêneos, as redes heterogêneas emergem como uma ferramenta estratégica para modelagem e análise. A capacidade das redes heterogêneas de representar sistemas complexos e dinâmicos fornece uma perspectiva valiosa, capacitando pesquisadores e profissionais a compreender e interagir com a diversidade estrutural presente em diversas disciplinas. A contínua exploração e aprimoramento de técnicas associadas às redes heterogêneas prometem contribuir significativamente para avanços em áreas tão diversas quanto ciência da computação, biologia, e engenharia de sistemas.

2.3 Trabalhos relacionados ao projeto

A previsão da evasão escolar em plataformas de ensino online e à distância atraiu uma atenção significativa da investigação (Chen et al, 2022). Ao longo do tempo, os pesquisadores propuseram várias estratégias e métodos para enfrentar este desafio (Márquez-Vera et al. 2016, Kumar et al. 2017, Feng et al. 2019, Oqaidi et al. 2022, Dol e Jawandhiya 2023). Mais recentemente, as preocupações com a privacidade e a proteção de dados aumentaram, provocando uma ênfase crescente em abordagens de preservação da privacidade para a previsão da evasão escolar (Oneto et al. 2017, Deho et al. 2022). Esta seção oferece uma visão geral do trabalho relacionado neste campo, destacando diversas estratégias e suas implicações para a privacidade.

Os esforços iniciais na previsão da evasão foram baseados em métodos estatísticos e algoritmos tradicionais de aprendizagem automática (Kumar et al, 2017). Esses modelos geralmente dependiam de uma variedade de dados dos estudantes, incluindo dados demográficos, desempenho acadêmico e informações socioeconômicas. Embora estes modelos tenham apresentado resultados promissores na previsão dos riscos de evasão escolar, também levantaram preocupações relativamente à privacidade e segurança dos dados sensíveis dos estudantes (Deho et al, 2022).

À luz das regulamentações de privacidade, os modelos treinados em dados privados podem não ser aplicáveis a novos estudantes que podem estar menos dispostos a fornecer suas informações confidenciais (Erickson, 2018). Portanto, há necessidade de explorar modelos que possam prever efetivamente os riscos de evasão usando apenas logs de acesso aos recursos da plataforma de aprendizagem, uma vez que esses logs podem ser facilmente anonimizados e não contêm informações confidenciais do usuário.

Para abordar questões de privacidade na mineração de dados educacionais, diversas iniciativas foram propostas. A técnica mais popular é a anonimização de dados, que visa preservar a privacidade do estudante durante o pré-processamento e treinamento do modelo (Deho et al, 2022). A ideia geral é remover informações de identificação pessoal, como nomes e números de identificação, dos conjuntos de dados, bem como dados socioeconômicos e outras informações sensíveis.

Outra abordagem que vale a pena mencionar é o uso de métodos diferenciais de privacidade (Boyer et al, 2015), que introduzem ruído ou perturbação durante o treinamento do modelo para evitar a recuperação de dados sensíveis. No entanto, esta estratégia ainda exige que dados sensíveis dos estudantes estejam presentes durante a fase de inferência. Os modelos baseados em grafos têm atraído atenção significativa nos últimos anos devido à sua interpretabilidade e à sua capacidade de capturar relações complexas entre estudantes e recursos de aprendizagem (Nitta et al, 2021).

Contudo, uma limitação desta abordagem é a necessidade de incluir funcionalidades como atributos dos nós, bem como a falta de representação explícita das relações entre os estudantes e das suas ações na plataforma de aprendizagem online. Uma abordagem que pode resolver as limitações acima mencionadas é o uso de redes de informação heterogêneas (do Carmo e Marcacini, 2021).

Redes heterogêneas permitem a representação explícita de diferentes tipos de nós e seus relacionamentos. Por exemplo, ações como acesso a cursos, envio de tarefas e participação em fóruns podem fornecer informações valiosas sobre o envolvimento dos estudantes no processo de aprendizagem. Além disso, redes heterogêneas podem permitir a modelagem de problemas de previsão de evasão usando apenas informações não sensíveis dos estudantes, tornando esta representação potencialmente útil para tais cenários. Na próxima seção, é apresentada a metodologia de desenvolvimento deste trabalho.

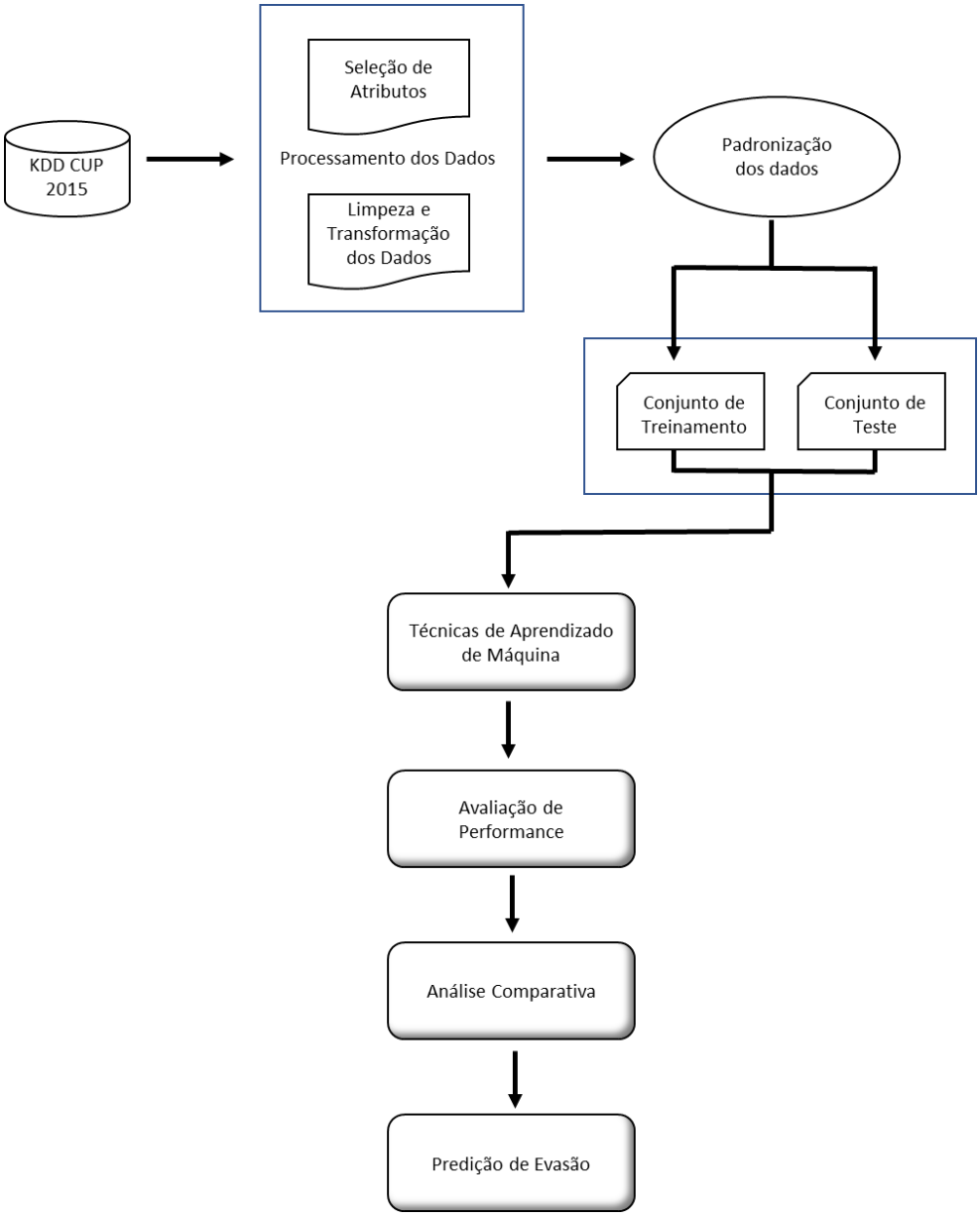
3. Metodologia

A metodologia usada é apresentada conforme mostra a Figura 4, cujas etapas são descritas mais detalhadamente nas subseções seguintes.

O KDD Cup 2015 Dataset foi um conjunto de dados utilizado na competição KDD Cup 2015, um evento anual realizado pela ACM SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining). O objetivo dessa competição era incentivar pesquisadores e

profissionais da área de ciência de dados a desenvolverem soluções inovadoras para um desafio específico.

Figura 3 - Passos da metodologia de pesquisa



Fonte: (Adaptado de NITHYA, 2022)

3.1 KDD Cup 2015 Dataset

O conjunto de dados do KDD Cup 2015 era composto por informações coletadas de um provedor de MOOCs, abrangendo um período de tempo específico. Os dados continham

registros detalhados das atividades dos estudantes, como inscrições em cursos, visualização de aulas, participação em fóruns, realização de tarefas e interações com a plataforma.

O desafio proposto para os participantes da competição era criar modelos preditivos capazes de prever a evasão de estudantes em cursos online. Os competidores deveriam utilizar os dados fornecidos para desenvolver algoritmos de aprendizado de máquina e técnicas de mineração de dados, a fim de identificar padrões e características que pudessem indicar a evasão iminente de um estudante.

O conjunto de atributos do Dataset incluía uma variedade de informações relacionadas às atividades dos estudantes em cursos online. Alguns dos atributos comuns presentes nesse conjunto de dados podem ser:

- **Identificação do estudante**: Um identificador único para cada estudante, permitindo a distinção entre diferentes indivíduos.
- **Informações demográficas:** Dados sobre características demográficas dos estudantes, como idade, gênero, localização geográfica, formação acadêmica, entre outros.
- **Informações do curso:** Detalhes sobre os cursos em que os estudantes estavam matriculados, como nome do curso, área de estudo, duração, nível de dificuldade, entre outros.
- **Atividades do estudante:** Registros das atividades realizadas pelos estudantes na plataforma de ensino online, como visualização de aulas, participação em fóruns, conclusão de tarefas, realização de provas, entre outros.
- **Tempo de interação:** Informações sobre a quantidade de tempo que cada estudante passou interagindo com a plataforma e os cursos, incluindo a duração das sessões de estudo e o intervalo de tempo entre as atividades.
- **Desempenho acadêmico:** Métricas relacionadas ao desempenho dos estudantes, como notas em avaliações, pontuação em testes, taxas de conclusão de tarefas, entre outros.
- **Indicadores de evasão:** Marcadores que indicam se um estudante abandonou ou não o curso, com base em critérios específicos definidos pelo conjunto de dados.

Esses são apenas alguns exemplos de atributos que podem estar presentes no KDD Cup 2015 Dataset. A quantidade e a natureza dos atributos podem variar, dependendo da finalidade específica do conjunto de dados e dos objetivos da competição.

3.2 Processamento dos dados

Conforme mencionado anteriormente, do dataset original foi realizada uma análise e selecionado os seguintes atributos de comportamento e acesso à plataforma MOOC:

- **enroll_id:** Esse atributo representa um identificador único para cada matrícula de estudante em um curso específico. Ele é usado para distinguir diferentes matrículas e rastrear o progresso e as interações dos estudantes ao longo do tempo.
- **username:** O atributo username refere-se ao nome de usuário de cada estudante. Ele é utilizado para identificar e distinguir os estudantes individualmente, permitindo a análise do comportamento e das preferências de cada estudante.
- **course_id:** Esse atributo indica o identificador único do curso em que o estudante está matriculado. Ele é fundamental para agrupar as atividades dos estudantes por curso específico, permitindo análises e modelagens direcionadas a cursos individuais.
- **session_id:** O atributo session_id representa um identificador único para cada sessão ou interação do estudante com a plataforma de ensino online. Ele ajuda a identificar e rastrear o tempo e a sequência das atividades realizadas pelos estudantes.
- **action:** Esse atributo registra a ação específica realizada pelo estudante, como visualizar uma aula, participar de um fórum de discussão, concluir uma tarefa, entre outras possíveis interações com a plataforma.
- **object:** O atributo object descreve o objeto específico com o qual o estudante interagiu em uma determinada ação. Por exemplo, pode indicar qual aula foi visualizada, qual tarefa foi concluída ou qual recurso foi acessado.
- **time:** Esse atributo indica a data e o horário em que a atividade ocorreu. Ele é importante para analisar o padrão de atividades dos estudantes ao longo do tempo e para identificar tendências ou variações nos comportamentos dos estudantes.
- **truth:** O atributo truth é utilizado para indicar se um determinado estudante abandonou ou não o curso. Ele é uma variável binária, onde o valor "1" pode representar o abandono e o valor "0" pode indicar que o estudante não abandonou o curso.

Esses atributos do dataset fornecem informações valiosas sobre o comportamento dos estudantes em MOOCs e são usados para análises, modelagem preditiva e identificação de padrões relacionados ao abandono dos cursos.

O dataset com dados sensíveis contém os seguintes atributos:

- **gender:** define o gênero do(a) estudante, entre homem, mulher e não informado.

- **education:** define o nível de educação do(a) estudante
- **birth:** define a data de nascimento do(a) estudante
- **preferences:** este é um atributo gerado a partir das matrículas dos estudantes nos cursos em que foi aprovado anteriormente.

3.3 Divisão do dataset

Nessa fase, o conjunto de dados disponível foi dividido em duas partes distintas: um conjunto de treinamento e um conjunto de teste.

O conjunto de treinamento é utilizado para treinar o modelo de aprendizado de máquina, permitindo que ele aprenda a partir dos padrões e relações presentes nos dados. Ele é fundamental para a construção do modelo e ajuste dos parâmetros.

Já o conjunto de teste é usado para avaliar o desempenho do modelo em dados não vistos durante o treinamento. Ele é uma medida importante para verificar se o modelo é capaz de generalizar bem para novos dados e se está sofrendo de overfitting (ajuste excessivo) aos dados de treinamento. O conjunto de teste fornece uma avaliação objetiva e imparcial do desempenho do modelo.

A separação entre conjunto de treinamento e conjunto de teste é uma etapa crítica no processo de desenvolvimento de modelos de aprendizado de máquina, garantindo que o modelo seja avaliado corretamente e que suas previsões sejam confiáveis em dados não observados anteriormente.

3.4 Técnicas de Aprendizado de Máquina

Embora diferentes modelos, conforme mencionado na seção 2.3, tenham sido utilizados com sucesso na prospecção de dados educacionais, a sua aplicação à previsão da evasão escolar tem sido relativamente limitada. Notadamente, a tarefa de classificação de nós provou ser valiosa, pois cada estudante é representado como um nó no grafo e as bordas do grafo capturam as relações entre os estudantes. Vários recursos, como dados demográficos, desempenho acadêmico e métricas de engajamento, podem ser incorporados como atributos dos nós. Ao treinar um classificador baseado em grafos, como GCNs (Karimi et al. 2020), o modelo aprende a propagar informações pelo grafo, captando a influência dos estudantes vizinhos nos riscos de evasão.

Desta forma, o uso desta técnica teve como objetivo:

1. **Análise de Centralidade:** As medidas de centralidade, como grau, centralidade de proximidade e centralidade de intermediação, foram aplicadas nas redes complexas construídas a partir do dataset. Essas medidas ajudaram a identificar os estudantes mais influentes, aqueles com maior conexão ou intermediação dentro da rede.
2. **Detecção de Comunidades:** Algoritmos de detecção de comunidades foram aplicados para identificar grupos de estudantes com comportamentos semelhantes em relação ao abandono dos cursos. Esses grupos podem fornecer insights sobre características compartilhadas e padrões de comportamento.
3. **Previsão de Evasão:** As redes complexas foram utilizadas como entrada para modelos de aprendizado de máquina, como Redes Neurais Artificiais ou Algoritmos de Classificação, para prever o abandono de cursos. A estrutura e as características das redes forneceram informações adicionais para melhorar a precisão das previsões.
4. **Análise de Propagação de Influência:** Com base na estrutura da rede complexa, é possível analisar como a influência se propaga entre os estudantes. Isso ajudou a identificar fatores que contribuem para o abandono e entender como o comportamento de um estudante pode influenciar outros.
5. **Identificação de Hubs e Pontes:** As redes complexas permitem identificar hubs (nós com alta conectividade) e pontes (arestas que conectam diferentes grupos). Essas informações foram relevantes para entender a importância de certos estudantes ou conexões na propagação do abandono e na dinâmica da rede.

Após a escolha da técnica, foi desenvolvido, então, a abordagem de Previsão de abandono de estudantes que preserva a privacidade, ou seja, o método P²SDP.

4 Previsão de evasão de estudantes que preserva a privacidade (P²SDP)

O método de previsão de evasão estudantil com preservação da privacidade (P²SDP) visa prever riscos de evasão em plataformas de ensino a distância sem depender de dados confidenciais dos estudantes. O método consiste em três etapas principais:

1. modelagem de redes heterogêneas para previsão de evasão,
2. aprendizado de grafos usando uma estrutura de regularização e
3. discretização em três níveis de risco de evasão.

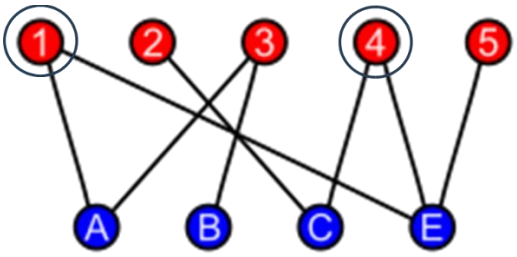
4.1. Modelagem de redes heterogêneas para previsão de evasão

Na primeira etapa, um banco de dados histórico de estudantes e seus registros de acesso é transformado em uma representação de rede heterogênea. A rede heterogênea representa explicitamente diferentes tipos de nós e seus relacionamentos. Por exemplo, os nós podem representar estudantes, cursos, matrículas e ações como acessos a cursos, envios de tarefas e participação em fóruns. As arestas entre os nós capturam os relacionamentos e interações entre eles. Esta abordagem de modelagem permite a incorporação de diversas fontes de informação e fornece uma visão abrangente do envolvimento e envolvimento dos estudantes no processo de aprendizagem.

Seja $G = (V, E)$ denota a rede heterogênea, onde V representa o conjunto de nós e E representa o conjunto de arestas. Os nós em V podem ser categorizados em diferentes tipos, como $V = V_s \cup V_c \cup V_e \cup V_a$, onde V_s representa nós de estudantes, V_c representa nós de curso, V_e representa nós de matrícula e V_a representa nós de ação. As arestas em E capturam os relacionamentos entre os nós, como matrículas em cursos de estudantes ou interações entre estudantes e ações.

Assim, o grafo resultado da modelagem realizada, pode ser analisada na figura 4, onde na camada superior um nó representa os dados de cada estudante e na camada inferior suas interações com a plataforma, tais como quais cursos está matriculado e quais objetos interagiu. Parte dos nós de estudantes foi rotulada com a informação de evasão para viabilizar o treinamento do modelo.

Figura 4 – Vértices representando estudantes



Fonte: (O Autor)

4.2 Aprendizado de grafos usando uma estrutura de regularização

Após a etapa de modelagem de redes heterogêneas, o método P²SDP utiliza uma abordagem de aprendizado de grafos para capturar os padrões de previsão de evasão dentro da rede. Uma estrutura de regularização baseada na propagação de rótulos é aplicada para classificar os estudantes e propagar rótulos através do grafo.

Seja $Y_l = \{ y_i \}$ denota o conjunto de estudantes rotulados com status de evasão conhecido, onde $y_i \in \{0, 1\}$ representa o status de abandono (0 para não evasão e 1 para evasão) do estudante i. O objetivo é inferir a situação de evasão $Y_u = \{ y_j \}$ dos demais estudantes não rotulados.

O processo de propagação de rótulos envolve a atualização iterativa dos rótulos dos estudantes não rotulados com base nos rótulos de seus nós vizinhos no grafo. A equação de propagação do rótulo pode ser definida como apresentada na Equação 1.

Equação 1 – Propagação do Rótulo

$$Q(\mathbf{F}) = \frac{1}{2} \sum_{v_i, v_j \in V} w_{v_i, v_j} \Omega(\mathbf{f}_{v_i}, \mathbf{f}_{v_j}) + \alpha \sum_{v_k \in V_s^L} \Omega(\mathbf{f}_{v_k}, \mathbf{y} v_k)$$

Fonte: (O Autor)

No método P²SDP, a matriz F representa os valores de pertinência de cada nó para cada classe de status de abandono. A função objetivo a ser minimizada, denotada como *Q(F)*, incorpora dois termos e uma função distância **Ω**.

O primeiro termo da função objetivo incentiva a consistência entre os nós vizinhos conectados. Ele garante que os nós vizinhos na rede tenham previsões semelhantes relacionadas ao status de abandono. O peso $W_{Vi,Vj}$ reflete a força da conectividade entre os nós *Vi* e *Vj*, enquanto a função Ω (f_{Vi}, f_{Vj}) mede a similaridade entre as previsões f_{Vi} e f_{Vj}.

O segundo termo da função objetivo, Ω (f_{Vk}, y_{Vk}), quantifica a consistência entre o rótulo previsto f_{Vk} do nó y_{Vk} e seu rótulo observado V_k, para um determinado conjunto de nós estudantes rotulados V_s^L. O parâmetro a controla o trade-off entre a importância dos estudantes inicialmente rotulados. Este termo penaliza o modelo quando há diferenças significativas entre os rótulos previstos e observados.

Minimizando a função $Q(\mathbf{F})$, obtemos os valores de pertinência para o status de abandono de cada nó. No entanto, nosso interesse principal reside nos nós de estudantes não rotulados. Especificamente, o objetivo é discretizar estes valores de adesão em grupos para determinar o risco de abandono, conforme discutido na próxima secção.

4.3. Discretização dos níveis de risco de evasão

Para mapear os valores de adesão contínua para níveis de risco específicos, foi aplicado um processo de discretização à (sub)matriz $\mathbf{F}_{V_s}$ de nós estudantes. Estes valores de adesão refletem o grau de associação entre cada nó de estudante e os diferentes níveis de risco de evasão, sendo que valores mais elevados indicam uma associação mais forte com um determinado nível de risco. Ao considerar esses valores, os estudantes podem ser classificados em categorias de risco distintas.

Um passo crucial no processo de discretização é determinar os limites apropriados. Foram empregadas técnicas tradicionais de *clustering* para identificar esses limites com base nos valores de adesão. Formalmente, o processo de discretização pode ser definido utilizando a distribuição normal para atribuir níveis de risco. Seja $\mathrm{F_{Vi}}$ o valor de adesão para um determinado nó estudante V_i. A equação *Vi* de discretização, diferentes níveis de riscos de evasão (RiskLevel) pode ser expressa da seguinte forma:

$$\mathrm{RiskLevel}(v_i) = \begin{cases} \mathrm{Low}, & \text{if } \mathbf{F}_{v_i} < \mu - \sigma \\ \mathrm{Medium}, & \text{if } \mu - \sigma \leq \mathbf{F}_{v_i} < \mu + \sigma \\ \mathrm{High}, & \text{if } \mathbf{F}_{v_i} \geq \mu + \sigma \end{cases}$$

onde μ representa a média dos valores de pertinência e σ denota o desvio padrão. Assume-se que os níveis de risco seguem uma distribuição normal, o que se mostrou adequado com base nos experimentos conduzidos e discutidos na próxima seção.

Ao discretizar os valores de adesão utilizando a distribuição normal, categorizamos os estudantes em três níveis de risco de abandono: baixo, médio e alto. A eficácia deste processo de discretização será discutida e avaliada mais detalhadamente na próxima seção.

4.4 Avaliação Experimental

Conforme mencionado anteriormente, o conjunto de dados utilizado na avaliação experimental foi coletado da plataforma MoocData, que é baseada no site MOOC XuetangX. Este conjunto de dados fornece uma coleção abrangente de dados que capturam as interações e comportamentos dos estudantes no ambiente de ensino à distância. O (sub)conjunto de dados experimental contém 4.290 usuários que participaram de 20.310 matrículas em 247 cursos diferentes. Esses usuários geraram um total de 4.421.314 transações.

As ações realizadas pelos usuários na plataforma foram registradas e abrangem uma ampla gama de atividades. Essas ações incluem buscar vídeos, pausar e reproduzir vídeos, carregar vídeos, clicar no material didático, fechar o material didático, clicar em informações, parar vídeos, clicar em páginas, monitorar o progresso, acessar fóruns, criar tópicos e comentários, resolver problemas, excluir tópicos e comentários, salvar problemas e fechamento de fóruns. Cada ação está associada a um registro de data/hora, permitindo a análise de padrões temporais no comportamento do usuário.

Para o método P²SDP proposto, três atributos do estudante, nomeadamente sexo, escolaridade do usuário e data de nascimento do usuário, foram excluídos do conjunto de dados. Esses atributos foram retidos e utilizados no modelo de referência para comparação, que serve como benchmark. Além disso, o conjunto de treinamento consistia em 3.003 instâncias, enquanto o conjunto de teste continha 1.287 instâncias. Esta divisão foi baseada na divisão original fornecida pelos autores do conjunto de dados.

Na sequência o objetivo foi avaliar grupos de risco na previsão do abandono, o que é uma tarefa desafiadora, uma vez que o conjunto de dados original não possui inerentemente uma organização em grupos de risco. Para resolver isso, foi adotada uma estratégia extrema de

análise de classificação. Nesse caso, se o modelo tiver um bom desempenho, o grupo de Baixo Risco (Risco L) deverá consistir predominantemente de matrículas de estudantes com status de evasão 0, enquanto o grupo de Alto Risco (Risco H) deverá consistir de matrículas com um status de abandono de 1.

Assim, foi calculada a medida de risco com base na taxa de abandono esperada dentro de cada grupo, normalizada pelo tamanho do grupo. Essa abordagem serve como uma forma de precisão para análise de previsão de risco. A Tabela 1 apresenta uma avaliação abrangente comparando o Modelo de Referência, que utiliza informações confidenciais do usuário, com o modelo P²SDP, que exclui informações confidenciais do usuário do gráfico. O modelo P²SDP proposto alcança medidas de maior precisão em diferentes parâmetros de regularização (∝). Isto indica que a estratégia proposta é competitiva mesmo sem incorporar informações sensíveis que normalmente são relevantes para a previsão da evasão.

Tabela 1 - Comparação das medidas de acurácia do Modelo de Referência e do Modelo P²SDP em diferentes valores dos parâmetros de regularização (∝). As medidas de precisão são calculadas para os grupos de baixo risco (risco L) e de alto risco (risco H)

	Reference Model		P^2SDP Model	
α	L-Risk	H-Risk	L-Risk	H-Risk
0.2	0.529	0.804	0.459	0.818
0.3	0.575	0.817	0.518	0.830
0.5	0.544	0.829	0.528	0.831
0.8	0.567	0.827	0.546	0.830
1.0	0.574	0.828	0.548	0.829

Fonte: O autor.

No entanto, é importante notar que ambas as estratégias encontraram limitações na previsão precisa do grupo de Baixo Risco. Na prática, isso significa que ainda há uma parcela significativa de matrículas de estudantes com situação de evasão 1 presentes no grupo de Baixo Risco. Por outro lado, quando os modelos identificam uma matrícula como pertencente ao grupo de Alto Risco, geralmente acertam mais de 80% das previsões.

Embora ainda haja espaço para melhorias, alcançar resultados competitivos sem depender de informações confidenciais dos usuários é uma conquista significativa. Estas conclusões destacam o potencial dos métodos de preservação da privacidade na prospecção de dados educacionais e fornecem informações valiosas para o desenvolvimento de modelos de previsão de evasão escolar mais robustos e conscientes da privacidade em ambientes de ensino à distância.

5 CONCLUSÃO

Neste trabalho, foi abordado o desafio da previsão da evasão estudantil em plataformas de ensino a distância, com foco na preservação da privacidade. Foi introduzida o P²SDP, que utiliza uma abordagem baseada em grafos e não depende de informações confidenciais dos estudantes.

No P²SDP, foi proposta uma nova abordagem para a previsão da evasão escolar, modelando o ambiente de aprendizagem como uma rede heterogênea. Essa abordagem de modelagem captura as relações entre estudantes, cursos, matrículas e ações de forma abrangente, permitindo previsões precisas dos riscos de evasão.

Em segundo lugar, foi desenvolvida uma estrutura de aprendizagem de grafos baseada na propagação de rótulos, que utiliza a conectividade do grafo para propagar rótulos e estimar riscos de abandono para estudantes não rotulados. Esta estrutura garante a privacidade dos dados dos estudantes, aproveitando apenas informações não confidenciais.

Por último, foi introduzido um processo de discretização para categorizar os estudantes em grupos de baixo, médio e alto risco, fornecendo insights acionáveis para estratégias de intervenção. As direções futuras de pesquisa envolvem a exploração e avaliação do método P²SDP em diversos contextos educacionais e a consideração de diferentes abordagens de modelagem de redes heterogêneas.

Embora método proposto se concentre em um conjunto de dados MOOC específico, é relevante avaliar seu desempenho usando dados de outras plataformas bem conhecidas. A realização de tais avaliações forneceria informações sobre a generalização e aplicabilidade do método P²SDP em diferentes ambientes educacionais.

Além disso, está prevista uma avaliação qualitativa da implementação prática do modelo, com o objetivo de medir como as instituições de ensino podem efetivamente priorizar grupos de risco de abandono escolar em cenários do mundo real. É importante notar que o método atualmente se concentra em estudantes que apresentam um nível mínimo de atividade regular na plataforma. Explorar cenários onde os estudantes apresentam frequência de acesso baixa ou irregular é importante. Por fim, é importante enfatizar que o método P²SDP não é apenas uma contribuição de pesquisa, mas também uma ferramenta computacional.

REFERÊNCIAS

ALHASHIMI, H. F.; HINDIA, M. N.; DIMYATI, K.; HANAFI, E. B.; SAFIE, N.; QAMAR; F.; NGUYEN, Q. N. **A Survey on Resource Management for 6G Heterogeneous Networks: Current Research, Future Trends, and Challenges**. Electronics, 12(3), 647 2023.

BHATI, U. A.; TANG, H.; WU, G.; MARJAN, S.; HUSSAIN, A. **Deep learning with graph convolutional networks: An overview and latest applications in computational intelligence**. International Journal of Intelligent Systems, 2023, 1-28, 2023.

BOYER, S.; GELMAN, B. U.; SCHRECK, B.; VEERAMACHANENI, K. **Data Science foundry for moocs**. In IEEE International Conference on Data Science and Advanced Analytics (DSAA), páginas 1-10. IEEE, 2015.

CAO, P.; ZHU, Z.; WANG, Z.; ZHU, Y.; & NIU; Q. **Applications of graph convolutional networks in computer vision. Neural Computing and Applications**, 34(16), 13387-13405, 2022.

DE MELO, R. O. P.. **Percepção dos Usuários sobre a LGPD: Bases Legais, Princípios e Direitos dos Titulares**. 2022. Tese de Doutorado. UNIVERSIDADE FEDERAL DE PERNAMBUCO.

CHEN, J.; FANG, B.; Zhang, H; Xue, X. **A systematic review for mooc dropout prediction from the perspective of machine learning.** Interactive Learning Environments, páginas 1-14, 2022.

DEHO, O. B.; JOKSIMOVIC, S.; LI, J.; ZHAN, C.; LIU, J; Liu, L. **Should learning analytics models include sensitive attributes? explaining the why.** IEEE Transactions on Learning Technologies. 2022.

DO CARMO, P.; MARCACINI, R. **Embedding propagation over heterogeneous event networks for link prediction.** Conferência Internacional IEEE sobre Big Data (Big Data), páginas 4812-4821. IEEE, 2021.

DOL, S. M. e JAWANDHIYA, P. M. **Classification technique and its combination with clustering and association rule mining in educational data mining - a survey.** Engineering Applications of Artificial Intelligence,122:106071. 2023.

ERICKSON, A. **Comparative analysis of the eu's gdpr and brazil's lgpd: Enforcement challenges with the lgpd**. Ribeiro. J. Internacional L., 44:859, 2018.

FARDOUN, H.; VENTURA, S. **Early dropout prediction using data mining: a case study with high school students.** Expert Systems, 33(1):107–124, 2016.

FENG, W.; TANG, J; LIU, T. X. **Understanding dropouts in moocs**. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, páginas 517-524, 2019.

GHISLENI, J. Z.. **A LGPD E A Risk-Based Approach Da Governança Corporativa: A Primeira Medida Para O Controlador Aplicar Os Princípios**. e3-Revista de Economia, Empresas e Empreendedores na CPLP, v. 8, n. 1, p. 103-126, 2022.

KARIMI, H.; DERR, T.; HUANG, J.; TANG, J. **Online academic course performance prediction using relational graph convolutional neural network**. International Educational Data Mining Society, 2020.

KRÜGER, J. G, C.; DE SOUZA BRITTO JR, A.; BARDDAL, J. P. **An explainable machine learning approach for student dropout prediction.** Expert Systems with Applications,, página 120933, 2023.

KUMAR, M.; SINGH, A; HANDA, D. **Literature survey on educational dropout prediction**. International Journal of Education and Management Engineering, 7(2):8, 2017.
MÁRQUEZ-VERA, C.; CANO, A.; ROMERO, C.; NOAMAN, A. Y. M.; MOUSA

MUBARAK, A. A.; CAO, H.; HEZAM, I. M.; HAO, F. **Modeling students' performance using graph convolutional networks**. Complex & Intelligent Systems,8(3):2183-2201, 2022.
Nitta, I., **Ishizaki**, R., **Shingu**, M., **Nakashima**, S. e **Todoriki**, M. **Graph-based massive open online course (mooc) dropout prediction using clickstream data in virtual learning environment.** In 16th International Conference on Computer Science & Education (ICCSE), páginas 48-52. IEEE, 2021.

ONETO, L.; SIRI, A.; LURIA, G.; ANGUITA, D. **Dropout prediction at university of genoa: a privacy preserving data driven approach**. In ESANN, 2017.

OQAIDI, K.; AOUHASSI, S; MANSOURI, K. **Towards a students' dropout prediction model in higher education institutions using machine learning algorithms.** International Journal of Emerging Technologies in Learning (Online), 17(18):103, 2022.

PESTANA, M. **Os princípios no tratamento de dados na LGPD** (Lei Geral da Proteção de Dados Pessoais). São Paulo: revista Consultor Jurídico. Recuperado de https://www. conjur. com. br/2020-mai-25/marcio-pestana-principios-tratamento-dados-lgpd, 2020.

RAMESH, S.; NIRMALRAJ, S.; MURUGAN, S.; MANIKANDAN, R.; & AL-TURJMAN, F. **Optimization of energy and security in mobile sensor network using classification based signal processing in heterogeneous network**. Journal of Signal Processing Systems, 95(2-3), 153-160, 2023.

RESEARCHGATE, **Redes Heterogêneas**. Nov. 2023. Disponível em: https://www.researchgate.net/figure/Figura-21-Redes-heterogeneas_fig1_342920927

SOFTWALL. **A LGPD já está em vigor, sua empresa está pronta para el**a? Nov. 2023. Disponível em: https://www.softwall.com.br/solucoes/adequacao-lgpd/

SOUZA, B. D.; TEIXEIRA, R. A. de A.. **Breve introdução à LGPD**. Ciência da Informação Express; v. 1 (2020); 1-4, v. 24, n. 2, p. 4-1.

WANG, J.; YAN, Z.; WANG, H.; Li, T.; & Pedrycz, W.; **A survey on trust models in heterogeneous networks**. IEEE Communications Surveys & Tutorials, 2022.