UNIVERSIDADE
FEDERAL RURAL
DE PERNAMBUCO

Rodolfo André Barbosa Ferreira

Análise da Evasão no Ensino Superior: Predição e Prevenção por Meio da Mineração de Dados Educacionais

Recife

2024

Rodolfo André Barbosa Ferreira

Análise da Evasão no Ensino Superior: Predição e Prevenção por Meio da Mineração de Dados Educacionais

Monografia apresentada ao Curso de Bacharelado em Ciências da Computação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Ciências da Computação.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Computação

Curso de Bacharelado em Ciência da Computação

Orientador: Rafael Ferreira de Mello

Recife

2024

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

F383a Ferreira, Rodolfo André Barbosa
Análise da Evasão no Ensino Superior: Predição e Prevenção por Meio da Mineração de Dados Educacionais Recife 2024 / Rodolfo André Barbosa Ferreira. - 2024.
27 f.

Orientador: Rafael Ferreira de Mello.
Inclui referências.

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Ciência da Computação, Recife, 2024.

1. Mineração de Dados Educacionais. 2. Aprendizado de máquina. 3. Predições. 4. Evasão. I. Mello, Rafael Ferreira de, orient. II. Título

CDD 004

MINISTÉRIO DA EDUCAÇÃO E DO DESPORTO
UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO (UFRPE)
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

http://www.bcc.ufrpe.br

Ficha de Aprovação do Trabalho de Conclusão de Curso

Trabalho defendido por Rodolfo André Barbosa Ferreira às 15h do dia 05/03/2024, no Recife, como requisito para conclusão do curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco, intitulado "Análise da Evasão no Ensino Superior: Predição e Prevenção por Meio da Mineração de Dados Educacionais", orientado por Rafael Ferreira de Mello e aprovado pela seguinte banca examinadora:

ORIENTADOR
DC/UFRPE

AVALIADOR
DC/UFRPE

À Este trabalho é dedicado à minha mãe, que sempre se sacrificou para me proporcionar uma educação de qualidade. Seu apoio incondicional, incentivo e força foram a luz que me guiou durante toda minha graduação. Suas palavras de encorajamento e amor foram o combustível que me impulsionou a alcançar meus objetivos. Esta conquista é dedicada a você, minha inspiração e guia.

Agradecimentos

Agradeço profundamente ao meu orientador, Rafael Ferreira Leite de Mello, cuja orientação, apoio e expertise foram fundamentais para o desenvolvimento deste estudo. Sua dedicação incansável, paciência e incentivo foram essenciais ao longo deste processo. Agradeço sinceramente por sua orientação cuidadosa, insights valiosos e por sempre acreditar no meu potencial (mesmo quando até eu duvidei). Este trabalho não teria sido possível sem sua orientação exemplar. Rafael, esta conquista também é sua.

Agradeço à minha parceira desde o primeiro período, Juliana, que esteve ao meu lado em todos os desafios acadêmicos, compartilhando alegrias, dificuldades e sonhos. Sua presença constante e apoio inabalável foram fundamentais para minha jornada acadêmica.

Aos meus amigos Caio, Francisco e Victor, que me acompanharam durante toda essa jornada, oferecendo suporte, compreensão e momentos de descontração que tornaram os desafios mais leves e as vitórias mais significativas. Sua amizade foi um presente que valorizarei para sempre.

Dedico este trabalho ao professor Robson, cuja paixão pela programação e habilidade em transmitir conhecimento foram tão inspiradoras que despertaram em mim uma verdadeira paixão por esta área. Agradeço por abrir os meus olhos para as possibilidades emocionantes da programação e por me guiar com tanta dedicação e maestria. Sua influência foi fundamental para o meu crescimento acadêmico e pessoal.

Quero expressar minha mais profunda gratidão a todos os professores que contribuíram para minha jornada acadêmica. À Valmir por sua contribuição fundamental na minha jornada acadêmica, especialmente ao apresentar a disciplina de IA (Inteligência Artificial). Além disso, agradeço imensamente por sempre estar disponível para ajudar e esclarecer minhas dúvidas. Nunca esquecerei das tardes que passei na sala dos professores, tentando implementar um MLP (Perceptron Multicamadas), e da sua paciência e orientação constante durante esse processo, Ao professor André Camara, que tornou a matemática discreta II uma experiência empolgante, especialmente com o projeto envolvente sobre criptografia.

Agradeço também a todos os outros professores que, de maneira direta ou indireta, contribuíram para minha formação acadêmica. Cada um de vocês deixou uma marca indelével em minha jornada, e seus ensinamentos estarão presentes em cada linha de código e em cada passo que eu der na minha vida profissional. Mais uma vez, meu sincero agradecimento por seu compromisso, dedicação e impacto duradouro em

minha educação e desenvolvimento pessoal.

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

Resumo

Tendo em vista que a evasão ocorre por abandono, transferência ou desistência do curso; quando o aluno se desliga da instituição que está matriculado ou quando o aluno abandona definitivamente ou não o curso superior, este artigo busca identificar métodos e técnicas automáticas para auxiliar os gestores a prevenir casos de evasão por meio das predições. Para realizar o estudo foi utilizada a Mineração de Dados Educacionais (MDE), que aplica técnicas de mineração de dados, tais como banco de dados, estatísticas e aprendizado de máquina nas áreas da educação. Foram empregues dados de 5144 alunos com características relacionadas ao curso, semestre e demografia constantes no banco de dados fornecido pelo Sistema de Informações e Gestão Acadêmica (SIGA) da Universidade Federal Rural de Pernambuco (UFRPE) para os cursos de Zootecnia, Engenharia de Pesca e Agronomia. Os dados, exceto aqueles que são informações pessoais, restritas e sensíveis, foram separados em Características Acadêmicas por Semestre, Acadêmicas Gerais, dos Cursos, Demográficas e Característica alvo. O estudo usa o algoritmo de aprendizado de máquina LSTM e os otimizadores SGD e Adam, explorando diferentes valores para os parâmetros de taxa de aprendizagem, momentum, tamanho de lotes e número de épocas.

Palavras-chave: Mineração de Dados Educacionais, Aprendizado de máquina, Predições, Curso superior, Evasão.

Abstract

Considering that dropout occurs due to abandonment, transfer, or withdrawal from the course; when the student disengages from the institution they are enrolled in or when the student definitively abandons or does not complete higher education, this article seeks to identify methods and automated techniques to assist managers in preventing dropout cases through predictions. To conduct the study, Educational Data Mining (EDM) was used, which applies data mining techniques such as database, statistics, and machine learning in education. Data from 5144 students with characteristics related to course, semester, and demographics were used from the database provided by the Academic Information and Management System (SIGA) of the Federal Rural University of Pernambuco (UFRPE) for the courses of Animal Science, Fisheries Engineering, and Agronomy. The data, except for those containing personal, restricted, and sensitive information, were separated into Academic Characteristics per Semester, General Academic Characteristics, Course-related, Demographic, and Target Characteristics. The study employs the LSTM machine learning algorithm and the SGD and Adam optimizers, exploring different values for the parameters of learning rate, momentum, batch size, and number of epochs.

Keywords: Educational Data Mining, Machine Learning, Predictions, Higher education, Dropout .

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

UFRPE	Universidade Federal de Pernambuco
SIGA	Sistema Integrado de Gestão Acadêmica
LSTM	Long-short term memory
INEP	Instituto Nacional de Estudo e Pesquisas Educacionais
MDE	Mineração de dados educacionais

Sumário

1 Introdução

A evasão de cursos de nível superior é um problema que atinge várias universidades pelo Brasil e pelo mundo. De acordo com (COUTINHO et al., 2018) existem três tipos de evasão: curso, quando ocorre um abandono, transferência ou desistência do curso; institucional, quando o aluno se desliga da instituição em que está matriculado; e do sistema, que ocorre quando o aluno abandona definitivamente ou não o curso superior.

De acordo com os dados apresentados por (VIANA; SANTANA; RABêLO, 2022) a taxa de evasão dos cursos de graduação na rede pública foi de cerca de 40% em 2022. E, segundo dados do INEP em instituições federais de nível superior, de todos os alunos que ingressaram em 2011 apenas 43% se formaram e 55% evadiram [BRASIL 2020]. Levando em consideração esses dados é preciso encontrar formas de avaliar o risco de evasão dos alunos para, assim, poder atuar na permanência dos mesmos. Se torna imprescindível, portanto, criar métodos e técnicas automáticas para auxiliar os gestores a prevenir casos de evasão.

Uma das metodologias que podem ser aplicadas para a automação e análise desse processo é a Mineração de Dados Educacionais (MDE), que aplica técnicas de mineração de dados tais como banco de dados, estatísticas e aprendizado de máquina nas áreas da educação. Com essa metodologia podemos extrair informações importantes a partir de um grande volume de dados educacionais, acompanhando, analisando e avaliando os dados. A área de Mineração de dados educacionais pode auxiliar na formação de políticas públicas para permanência dos estudantes nas instituições de ensino superior. (VIANA; SANTANA; RABêLO, 2022)

Neste trabalho utilizaremos o banco de dados provido em (CARVALHO et al., 2019). fornecido pelo Sistema de Informações e Gestão Acadêmica (SIGA) da Universidade Federal Rural de Pernambuco (UFRPE) para os cursos de Zootecnia, Engenharia de Pesca e Agronomia. Todos os dados sensíveis como nome, CPF e endereço foram removidos. Os dados foram separados em 5 tipos de características: Características Acadêmicas por Semestre, Características Acadêmicas Gerais, Características dos Cursos, Características Demográficas e Característica Alvo.

Um dos algoritmos de aprendizado de máquinas é o Long Short-Term Memory (LSTM), uma das redes neurais mais promissoras para aplicação de dados temporais. Ela aplica o conceito de células de memória substituindo os neurônios das redes neurais tradicionais. E, assim como aplicado em (CHEN; ZHOU; DAI, 2015), apresenta um bom algoritmo com alta capacidade de previsão.

Neste estudo exploraremos uma aplicação da Mineração de Dados Educacionais na análise da evasão em cursos de ensino superior. Faremos uso dos dados fornecidos pelo Sistema de Informações e Gestão Acadêmica da Universidade Federal Rural de Pernambuco supracitado. Utilizando o algoritmo Long Short-Term Memory (LSTM), examinaremos a eficácia e o desempenho de diferentes parâmetros e otimizadores na previsão da evasão dos alunos com base nos dados disponibilizados.

2 Trabalhos relacionados

(COUTINHO et al., 2018) Propôs uma adaptação métrica do índice de cálculo empregado pelo [MEC, 1997]. Os dados foram extraídos do sistema de gestão acadêmica e com isso foram aplicados à métrica. Então, com os resultados obtidos, foram encontrados alguns pontos como desorganização do curso e incertezas sobre o mercado de trabalho como fatores que levam à evasão do curso.

Em (VIANA; SANTANA; RABêLO, 2022) foi feito um estudo utilizando o Knowledge Discovery in Databases(KDD), o processo de descoberta de conhecimento em banco de dados. Com 5 fases, onde é feita a coleta, pré-processamento, transformação, mineração e avaliação dos resultados. Foi utilizado um banco de dados com cerca de 130 atributos sociais e, com ajuda de analista, foram aplicados 12 atributos. Além do banco de dados com atributos sociais, outros bancos com dados de período e curso foram adotados. Após a coleta dos dados, foi usado um algoritmo de seleção de dados bastante indicado na literatura, o Random Forest (RF). Após isso, os atributos selecionados foram processados nos seguintes algoritmos: Radom Forest, Decision Tree, Extra Trees e Multilayer Perceptron, Support Vector Machine, K-Nearest Neighbors e Gaussian Naive Bayes. Como resultado do processamento, os algoritmos: Random Forest, Extra Trees, Multilayer Perceptron and Support Vector Machine apresentaram melhores resultados.

Nos estudos de (CARVALHO et al., 2019) foi utilizado um questionário sobre a situação socio-econômica e demográfica(CSED) com aproximadamente 40 perguntas que foram disponibilizadas no portal do aluno AVA, uma adaptação da plataforma Moodle, no primeiro período do curso. Primeiramente foi feita uma análise sobre os dados do final do primeiro período, como as notas e os coeficientes de rendimento, no qual foi classificado se o aluno era desistente ou formado. Em seguida, esses dados foram associados à pesquisa CSED. Como resultado foi possível analisar que o coeficiente de rendimento do aluno no primeiro período está associado com a situação final do estudante (formado ou desistente).

Já em (CHEN; ZHOU; DAI, 2015) foi utilizado o LSTM para predição no mercado de ações. Como atributos foram utilizados sucessivamente os preços e as negociações das ações em N dias. O modelo consiste em uma única camada de entrada com o mesmo número de células de memória que os recursos de aprendizado de sequência que podem também acomodar multiplas camadas LSTM, uma camada densa e por fim, uma única camada de saída com o mesmo número de células de memória que as categorias que definem o desempenho da sequência. Apesar de ter uma acurácia

em torno de 24%, pelo fato do mercado de ações da China ser muito imprevisível, foi um bom resultado comparado com outros trabalhos. O estudo mostra o quão eficaz o LSTM pode ser em aprendizado sequencial.

Dessa forma, neste projeto, propomos a aplicação do algoritmo LSTM, amplamente empregado na literatura para lidar com problemas temporais, na questão da evasão escolar. Nosso objetivo é prever se um aluno apresenta probabilidade de evadir do curso. Os dados necessários serão fornecidos pelo SIGA da UFRPE, garantindo total anonimato.

3 Materiais e métodos

3.1 Base de dados

Através do Sistema de Informações e Gestão Acadêmica (SIGA) foi criado um banco de dados com os dados de estudantes que ingressaram nos cursos de ciências agrárias (Agronomia, Engenharia de Pesca e Zootecnia) entre os anos 2019 e 2021. Durante a coleta dos dados foi levada em consideração a privacidade e o anonimato dos alunos ignorando campos identificáveis como Nome, CPF, Endereço e outros campos sensíveis. No total são dados de 5144 alunos com características relacionadas ao curso, semestre e demografia. As características a serem utilizadas nesse estudo estão apresentadas nas tabelas, nas quais incluem características acadêmicas gerais, características dos alunos por semestre, características dos cursos e características demográficas. Como objetivo temos o atributo alvo que é a conclusão ou evasão do aluno em relação ao curso.

Tabela 1 – Características Acadêmicas Gerais

Características	Descrição	Tipo
ano_conc_ensm	Ano de Conclusão do Ensino Médio	Numérico (1970 a 2019)
nota_enem	Nota Obtida na Prova do Enem	Numérico (0 a 100)
categoria_ensm	Categoria do Ensino Médio	Nominal
ano_admis	Ano de Admissão	Numérico (2009 a 2021)
semtr_admis	Semestre de Admissão	Dicotômica (1 e 2)
semtr_conc	Semestre de Conclusão	Dicotômica (1 e 2)
tp_ingrs	Tipo de Ingresso na Universidade	Nominal
ano_conc_prev	Ano de Conclusão Previsto	Numérico (2013 a 2026)
perd_conc_prev	Período de Conclusão Previsto	Dicotômica (1 e 2)
tp_admis	Tipo de Admissão na Universidade	Nominal
qtd_trancmt_acum	Quantidade de Trancamentos	Numérico (0 a 6)
sem_perd_letivo	Semestre Período Letivo	Dicotômico (1 e 2)
ano_perd_letivo	Ano Período Letivo	Numérico (2009 a 2021)

Tabela 2 – Características Acadêmicas por Semestre

Características	Descrição	Tipo
media_geral	Média Geral	Numérico (0 a 10)
nu_ranking	Média Geral por Ranking	Numérico (0 a 1000)
ds_tp_sit_vinc	Tipo de Situação do Vínculo	Nominal
chtotal_aprovado	Carga Horária Total de Aprovações	Numérico (0 a 1000)
chtotal_naoaprov	Carga Horária Total de Reprovações	Numérico (0 a 1000)
ch_aprovadoacum	Carga Horária Aprovações Acumuladas	Numérico (0 a 100)
ch_reprovadoacum	Carga Horária Reprovações Acumuladas	Numérico (0 a 100)
retido_fim_perd	Retido no Final do Período	Dicotômica (T ou F)

Tabela 3 – Características dos Cursos

Características	Descrição	Tipo
cd_progr_form	Código do Programa de Formação	Numérico (0 a 100)
cd_turno	Código do Turno	Numérico (1 e 2)
nm_progr_form	Nome do Programa de Formação	Nominal
nm_campus	Nome do Campus	Nominal
cd_perf	Código do Perfil dos Cursos	Numérico
duracao_curso	Duração do Curso por Semestres	Dicotômico (10 e 11)
ch_total	Carga Horária Total do Curso	Numérico (3000 a 5000)
ch_und_cur	Carga Horária Unidade do Curso	Numérico (3000 a 5000)
median_ch_perd	Mediana da Carga Horária do Período	Numérico (300 a 400)
perd_est_min_fim	Período Estimado Mínimo Final	Numérico (0 a 10)
perd_est_ma_fim	Período Estimado Máximo Final	Numérico (0 a 11)
perd_est_me_fim	Período Estimado Médio Final	Numérico (0 a 6)

Tabela 4 – Características Demográficas

Características	Descrição	Tipo
dt_nasc	Data de Nascimento	Numérico
raca	Cor da Pele do Estudante	Nominal
est_civil	Estado civil	Nominal
nm_sexo	Sexo do Estudante	Nominal
idade_admis_aprox	Idade Aproximada de Admissão no Curso	Numérico
sigl_uf_rg	Sigla do Estado em que Nasceu	Nominal

Tabela 5 – Categoria Alvo

Características	Descrição	Tipo
vinculo	Indica se o aluno é não evadido ou evadido	Dicotômica

Tabela 6 – Mapeamento ordinal da unidade federativa

Valor Nominal	Valor Numeral
Pernambuco	1
Paraíba	2
Minas Gerais	3
Distrito Federal	4

3.2 Metodologia

Nesta seção passamos a descrever o processo para conseguir classificar se um aluno tem chances ou não de evadir do curso como um problema de solução binária. Para isso, utilizamos técnicas de aprendizado de máquina para criar um sistema capaz de identificar - de acordo com as características relacionadas ao curso, semestre e demografia - as chances de um aluno perder o vínculo ou não com o curso e, por conseguinte, com a instituição de ensino. Utilizando o banco de dados provido pelo Sistema de Informações e Gestão Acadêmica fizemos o teste e validação para identificar a qualidade das predições.

Destaques-se que para desenvolver um sistema computacional capaz de auxiliar gestores academicos a indetificar se o aluno pode evadir ou não do curso é preciso da utilização de técnicas de análise de dados, com isso podemos a partir de dados relacionados a determinado aluno possa ser classificado e dado como resposta.

3.2.1 Normalização de dados

Na fase inicial do projeto foi feita a análise dos dados para identificar aqueles que requerem normalização. As colunas referentes ao semestre, unidade federativa, situação de vínculo, estado civil, cor/raça, sexo e data de nascimento foram identificadas como exigindo representação numérica. Foi utilizado o método de normalização ordinal, onde cada valor único em cada coluna mencionada é mapeado para um número distinto. Por exemplo:

Após realizar o mapeamento de todos os valores, os valores nominais foram substituídos por valores numéricos em cada coluna da tabela de dados. Essa transformação permite que os dados sejam compatíveis com o algoritmo LSTM e empregados de forma eficaz nas análises.

3.2.2 Algoritmo de aprendizagem de máquina

Em seguida, utilizamos o algoritmo de aprendizado de máquina para criar o modelo para ser apresentado. Para isso, foi utilizado o algoritmo LSTM que demonstra uma boa eficiência em classificação de dados (BELAGOUNE et al., 2021), (CHEN; ZHOU; DAI, 2015) e (YADAV et al., 2020).Os otimizadores SGD (Stochastic Gradient Descent) e Adam foram empregados. Ambos são frequentemente utilizados na literatura, conforme exemplificado em (SAURABH, 2020) e (YADAV et al., 2020), sendo o Adam aplicado principalmente em problemas de predição. (BELAGOUNE et al., 2021), (CHANDRIAH; NARAGANAHALLI, 2021). Foi empregada a função de perda logarítmica BinaryCrossEntropy, recomendada para problemas binários. Esta escolha se justifica pelo fato de lidarmos com uma situação binária, como a evasão ou não evasão do aluno, tornando-a adequada para o contexto desse trabalho.

3.2.3 Parâmetros

No algoritmo LSTM a configuração dos parâmetros é uma etapa crucial que impacta diretamente na performance e nos resultados alcançados. Foram ajustados os seguintes parâmetros: taxa de aprendizagem, momentum, tamanho do lote e número de épocas. A partir desses parâmetros conduziram-se experimentos visando atingir a máxima eficácia, levando em conta o tempo de execução do treinamento do algoritmo.

(A) A taxa de aprendizagem é fundamental para determinar a magnitude das atualizações nos pesos da rede neural durante o treinamento. Ajustar esse parâmetro corretamente pode acelerar a convergência do modelo para reduzir a perda. Alterar a taxa de aprendizado pode ter um impacto significativo no desempenho do treinamento e na velocidade com que o modelo aprende a partir dos dados. (YU et al., 2020)

(B) O momentum é uma técnica que visa acelerar o processo de descida do gradiente, acumulando uma velocidade vetor na direção de uma redução constante do objetivo ao longo das iterações. (SUTSKEVER et al., 2013)

(C) O tamanho do lote refere-se à quantidade de pontos de dados em um mini lote, o qual é uma amostra representativa dos dados de treinamento. A partir desse mini lote, o gradiente é calculado em cada etapa do otimizador de descida de gradiente estocástico (SGD) ou em suas variantes. (NEISHI et al., 2017)

(D) O número de épocas representa a quantidade de repetições do processo de aprendizado executado pelo LSTM. Quanto maior o número de épocas, mais tempo o algoritmo leva para processar os dados. No entanto, é importante destacar que um aumento no número de épocas não necessariamente resulta em uma

melhoria na acurácia do modelo. Idealmente, esperamos alcançar bons resultados com um número reduzido de épocas. (HASTOMO et al., 2021)

3.3 Experimentos e Resultados

Nesta seção, será fornecida uma descrição detalhada dos experimentos realizados para validar a proposta apresentada nas seções anteriores e avaliar a eficácia e qualidade do sistema de auxílio proposto para identificar alunos em risco de evasão. Utilizamos os dados obtidos pelo Sistema de Gestão Acadêmica para teste e validação. Durante esse processo empregamos o mesmo conjunto de dados de teste e validação para avaliar diferentes conjuntos de parâmetros, possibilitando uma análise comparativa dos resultados. Todos os experimentos foram conduzidos utilizando a plataforma do Google Colab.

Os testes foram conduzidos com variações nos seguintes parâmetros: taxa de aprendizado, momentum, tamanho do lote e número de épocas. Para avaliar possíveis impactos significativos nos resultados, empregamos valores com uma ampla disparidade entre si durante os testes. O impacto do momentum na velocidade do algoritmo foi investigado através dos valores (0, 0.5 e 1), abrangendo a gama de 0 (indicando ausência de momentum) até 1. A taxa de aprendizagem desempenha um papel fundamental, já que valores baixos podem retardar o progresso do algoritmo, enquanto valores altos podem resultar em oscilações excessivas. Portanto, foram testados os seguintes valores: (10^{-3}, 10^{-2}, 10^{-1}). O número de épocas representa quantas vezes o conjunto de dados é processado pela rede neural durante o treinamento, permitindo que os pesos da rede sejam ajustados. Neste contexto, foram empregados valores que variam de 20 a 80 . Em determinados algoritmos que demonstraram melhorias com o aumento do número de épocas foram aplicados valores de 100 e 200. O tamanho do lote se refere ao número de entradas a serem processadas simultaneamente durante o treinamento. Esse parâmetro tem um impacto direto tanto na eficiência quanto na estabilidade do treinamento do algoritmo. Para esse propósito, foram adotados valores de 50 e 10.

3.4 Otimizadores

Na fase de processamento o otimizador desempenha um papel crucial ao ajustar os parâmetros da rede LSTM. Neste estudo foram utilizados dois otimizadores, SGD e Adam, explorando diferentes valores para os parâmetros de taxa de aprendizagem, momentum, tamanho de lotes e número de épocas.

Tabela 7 – Resultados - SGD

AUC (%)	Tempo de exec. (s)	Épocas	Tam. de lotes	Taxa de Aprend.	Momentum
59.81	450.81	40	50	10^{-3}	0.5
59.64	109.08	10	50	10^{-3}	0.5
59.53	196.21	20	50	10^{-3}	0
59.28	293.41	40	100	10^{-3}	0.5
59.12	328.37	30	50	10^{-3}	0
57.29	266.49	30	100	10^{-3}	0.5
57.18	84.79	10	100	10^{-3}	0
56.94	472.18	40	50	10^{-1}	0.2
56.04	87.54	10	100	10^{-3}	0.5
55.9	151.44	20	100	10^{-3}	0.5
55.41	148.62	20	100	10^{-3}	0
55.31	87.6	10	100	10^{-2}	0.9
55.24	267.21	30	100	10^{-3}	0
54.95	206.61	20	50	10^{-3}	0.5
54.81	328.77	40	100	10^{-3}	0
54.03	447.90	40	50	10^{-3}	0
52.8	327.05	30	50	10^{-3}	0.5
52.67	152.18	10	50	10^{-3}	0
51.95	446.62	40	50	10^{-2}	0.9
48.75	327.52	30	50	10^{-1}	0.2
48.41	326.45	30	50	10^{-2}	0.9
44.6	327.23	40	100	10^{-2}	0.9
42.27	152.93	10	50	10^{-2}	0.9
41.63	327.31	40	100	10^{-1}	0.2
39.77	194.41	20	50	10^{-1}	0.2
38.87	210.52	30	100	10^{-1}	0.2
38.71	86.91	10	100	10^{-1}	0.2
31.74	267.32	20	50	10^{-2}	0.9
30.58	146.35	20	100	10^{-1}	0.2
30.55	208.67	20	100	10^{-2}	0.9
28.21	266.44	30	100	10^{-2}	0.9
25.08	147.31	10	50	10^{-1}	0.2

3.4.1 Otimizador SGD

O SGD (Stochastic Gradient Descent), também conhecido como Método do Gradiente Estocástico, é um otimizador amplamente empregado em problemas de classificação binária e multiclasses. Sua eficácia é notável em conjuntos de dados volumosos, pois permite que o modelo seja atualizado com base em amostras individuais. No entanto, é importante notar que essa abordagem pode levar a uma convergência mais ruidosa em comparação com métodos de otimização mais determinísticos.

3.4.2 Otimizador Adam

Outro otimizador muito utilizado é o Adam (Estimativa de Momento Adaptativo). O algoritmo ADAM apresenta uma convergência mais rápida em comparação com o descenso de gradiente estocástico convencional (SGD), sendo igualmente eficiente em termos computacionais e adequado para otimizar modelos com um grande número de parâmetros. Em contraste com a abordagem tradicional de atualização dos pesos com uma taxa de aprendizado constante, o ADAM baseia-se na correção das estimativas devido à distorção provocada pela média móvel do gradiente e do gradiente quadrático. (CHANDRIAH; NARAGANAHALLI, 2021) Ao contrário do SGD, o algoritmo Adam não possui um parâmetro explícito de momentum.

Tabela 8 – Resultados - Adam

AUC (%)	Tempo de exec. (s)	Épocas	Tam. de lotes	Taxa de Aprend.
77.32	923.42	80	50	10^{-3}
77.27	448.25	40	50	10^{-2}
75.51	567.05	60	50	10^{-3}
74.47	987.53	90	50	10^{-3}
73.86	627.2	60	50	10^{-2}
73.36	507.07	50	50	10^{-3}
72.26	687.60	70	50	10^{-3}
72.15	171.53	20	100	10^{-1}
72.1	568.42	50	50	10^{-1}
71.61	268.09	30	100	10^{-3}
71.50	101.69	10	50	10^{-3}
70.45	568.1	50	50	10^{-2}
69.27	207.01	20	50	10^{-3}
68.41	287.06	40	100	10^{-2}
66.35	988.68	90	50	10^{-2}
66.9	447.08	60	100	10^{-2}
66.08	827.93	80	50	10^{-2}
65.93	268.99	20	50	10^{-1}
65.67	2067.95	200	50	10^{-3}
65.40	328.21	40	100	10^{-3}
65.40	389.91	30	50	10^{-3}
64.50	447.00	40	50	10^{-3}
64.22	328.57	30	50	10^{-2}
63.73	447.89	50	100	10^{-2}
62.38	1048.32	100	50	10^{-3}
62.25	148.36	20	100	10^{-3}
61.54	214.25	30	100	10^{-2}
62.41	611.38	80	100	10^{-2}
60.7	773.25	90	100	10^{-2}
58.15	748.96	70	50	10^{-2}
57.81	507.65	70	100	10^{-2}
57.37	77.77	10	100	10^{-2}
54.3	164.3	20	100	10^{-2}
53.47	267.45	20	50	10^{-2}
49.45	147.76	10	50	10^{-2}
51.45	76.71	10	100	10^{-3}
50.15	487.26	40	50	10^{-1}
50.0	267.35	30	100	10^{-1}
50.0	323.65	30	50	10^{-1}
50.0	385.89	40	100	10^{-1}
50.0	508.59	60	100	10^{-1}
50.0	628.55	70	100	10^{-1}

4 Discussão

Conforme os estudos realizados, obtivemos nosso melhor resultado com a métrica AUC de 77,32%, ao utilizar o otimizador Adam com os seguintes parâmetros: 80 épocas de treinamento, tamanho de lote igual a 50 e uma taxa de aprendizagem de 10-3. O tempo total de treinamento para este modelo foi de aproximadamente 15 minutos. Por outro lado, o melhor resultado alcançado com o otimizador SGD foi de 59,81%, utilizando 50 épocas de treinamento, uma taxa de aprendizagem de 10-3 e um momentum de 0.5. O tempo de execução do algoritmo para esta configuração foi em torno de 7 minutos.

Com base nestes dados podemos concluir que o otimizador Adam demonstra um desempenho superior ao otimizador SGD para problemas temporais como o apresentado neste estudo.

Uma análise dos 10 melhores resultados revela que o algoritmo apresenta um desempenho otimizado com tamanhos de lote de 50, com apenas 2 exceções utilizando 100. O número de épocas ideal varia entre 20 e 80, enquanto valores acima (90, 100, 200) não apresentaram melhoras significativas na AUC, apenas aumentando o tempo de execução. Por exemplo, um experimento com 200 épocas levou 35 minutos para ser executado, com um AUC de 65,67%, mais que o dobro do tempo do melhor resultado. A taxa de aprendizagem ideal varia entre 10^{-3} e 10^{-2}.

O segundo melhor resultado, com AUC de 77,27%, foi obtido com uma taxa de aprendizagem de 10^{-2} e 40 épocas, o que resultou em um tempo de execução de 7,5 minutos, metade do tempo do melhor resultado (15 minutos). Essa observação indica que é possível ajustar a taxa de aprendizagem para otimizar o tempo de execução sem sacrificar significativamente a performance do algoritmo.

O estudo demonstra que o algoritmo apresenta um bom desempenho na tarefa de predição. O otimizador Adam se mostrou superior ao SGD para este problema, enquanto o tamanho de lote ideal foi de 50. O número de épocas ideal depende do conjunto de dados e dos recursos computacionais, enquanto a taxa de aprendizagem pode ser ajustada para otimizar o tempo de execução sem sacrificar a performance. O modelo com 77,32% de AUC pode ser aplicado ao problema de evasão escolar por ter o melhor custo-benefício em termos de tempo de execução.

5 Conclusão

Este trabalho abordou a problemática da predição de evasão acadêmica no ensino superior utilizando o algoritmo de aprendizado de máquina LSTM. Inicialmente, um banco de dados do Sistema de Gestão Acadêmica foi analisado e utilizado como conjunto de teste e avaliação para o algoritmo. Em seguida, diversos otimizadores e parâmetros foram aplicados com o objetivo de otimizar o desempenho do modelo.

O melhor resultado obtido foi uma taxa de acerto de 77,32% utilizando o otimizador Adam, 80 épocas e um tamanho de lote de 50. Este resultado demonstra o potencial do algoritmo LSTM para a predição de evasão acadêmica, fornecendo uma ferramenta valiosa para auxiliar nas ações de retenção de alunos no ensino superior.

Aumentar o banco de dados com informações de outros cursos e períodos pode ser crucial para aprimorar a AUC do algoritmo de predição de evasão. A inclusão de mais dados expandiria o conhecimento do modelo e permitiria uma aplicação mais específica aos diferentes cursos. Através da segmentação em grupos e da criação de modelos distintos, seria possível realizar comparações entre os modelos específicos e o modelo genérico, possibilitando uma análise mais aprofundada e insights valiosos sobre os fatores que influenciam a evasão em cada contexto.

Em trabalhos futuros podemos aplicar outros classificadores de aprendizagem de máquina como o MLP e o Random Forest Tree a fim de comparação de AUC e tempo de execução do algoritmo. Assim como podemos aplicar seleção de atributos e filtrar características que podem atrapalhar o algoritmo. Outra possibilidade é utilizar os tipos de dados (dados demográficos, do curso, do semestre) e aplicar aos algoritmos de aprendizagem de máquina separadamente e comparar ao genérico e avaliar a utilização de pesos.

Referências

BELAGOUNE, S. et al. Deep learning through lstm classification and regression for transmission line fault detection, diagnosis and location in large-scale multi-machine power systems. *Measurement*, v. 177, p. 109330, Jun 2021. Citado na página 19.

CARVALHO, L. et al. Detecção precoce de evasão em cursos de graduação presencial em computação: um estudo preliminar. In: *Anais do XXVII Workshop sobre Educação em Computação*. Porto Alegre, RS, Brasil: SBC, 2019. p. 233–243. ISSN 2595-6175. Disponível em: <https://sol.sbc.org.br/index.php/wei/article/view/6632>. Citado 2 vezes nas páginas 12 e 14.

CHANDRIAH, K. K.; NARAGANAHALLI, R. V. Rnn / lstm with modified adam optimizer in deep learning approach for automobile spare parts demand forecasting. *Multimedia Tools and Applications*, Apr 2021. Citado 2 vezes nas páginas 19 e 22.

CHEN, K.; ZHOU, Y.; DAI, F. A lstm-based method for stock returns prediction: A case study of china stock market. In: LUO, F. et al. (Ed.). *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2015. (Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015), p. 2823–2824. Funding Information: The work is partially supported by the National Natural Science Foundation of China (Grant No. 61221001), and Shanghai Science and Technology Committees of Scientic Research Project (Grant No. 14DZ1101200). Publisher Copyright: © 2015 IEEE.; 3rd IEEE International Conference on Big Data, IEEE Big Data 2015 ; Conference date: 29-10-2015 Through 01-11-2015. Citado 3 vezes nas páginas 12, 14 e 19.

COUTINHO, E. et al. Uma análise da evasão em cursos de graduação apoiado por métricas e visualização de dados. In: *Anais do XXIV Workshop de Informática na Escola*. Porto Alegre, RS, Brasil: SBC, 2018. p. 31–40. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/wie/article/view/14314>. Citado 2 vezes nas páginas 12 e 14.

HASTOMO, W. et al. Characteristic parameters of epoch deep learning to predict covid-19 data in indonesia. *Journal of Physics: Conference Series*, IOP Publishing, v. 1933, n. 1, p. 012050, jun 2021. Disponível em: <https://dx.doi.org/10.1088/1742-6596/1933/1/012050>. Citado na página 20.

NEISHI, M. et al. A bag of useful tricks for practical neural machine translation: Embedding layer initialization and large batch size. In: NAKAZAWA, T.; GOTO, I. (Ed.). *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, 2017. p. 99–109. Disponível em: <https://aclanthology.org/W17-5708>. Citado na página 19.

SAURABH, N. Lstm-rnn model to predict future stock prices using an efficient optimizer. *International Research Journal of Engineering and Technology (IRJET)*, v. 7, n. 11, p. 672, 2020. Citado na página 19.

SUTSKEVER, I. et al. On the importance of initialization and momentum in deep learning. In: DASGUPTA, S.; MCALLESTER, D. (Ed.). *Proceedings of the 30th International Conference on Machine Learning*. Atlanta, Georgia, USA: PMLR, 2013. (Proceedings of Machine Learning Research, 3), p. 1139–1147. Disponível em: <https://proceedings.mlr.press/v28/sutskever13.html>. Citado na página 19.

VIANA, F.; SANTANA, A.; RABêLO, R. Avaliação de classificadores para predição de evasão no ensino superior utilizando janela semestral. In: *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*. Porto Alegre, RS, Brasil: SBC, 2022. p. 908–919. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/sbie/article/view/22469>. Citado 2 vezes nas páginas 12 e 14.

YADAV, K. et al. Bi-lstm and ensemble based bilingual sentiment analysis for a code-mixed hindi-english social media text. *2020 IEEE 17th India Council International Conference (INDICON)*, Dec 2020. Citado na página 19.

YU, C. et al. Llr: Learning learning rates by lstm for training neural networks. *Neurocomputing*, v. 394, p. 41–50, Jun 2020. Citado na página 19.