JNT - FACIT BUSINESS AND TECHNOLOGY JOURNAL ISSN: 2526-4281 - QUALIS B1
ANO 2023 –MÊS DE AGOSTO - FLUXO CONTÍNUO - Ed. 44. Vol. 1. Págs. 80-89

CAPES

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

Facit.TO
FACULDADE DE CIÊNCIAS DO TOCANTINS

UTILIZANDO MACHINE LEARNING PARA ANÁLISE DE DADOS ESCOLARES: UMA ANÁLISE COMPARATIVA DE MODELOS

USING MACHINE LEARNING FOR SCHOOL DATA ANALYSIS: A COMPARATIVE ANALYSIS OF MODELS

Celso Barreto da SILVA
Centro Universitário Jorge Amado (UNIJORGE)
E-mail: csilva2208@unijorge.pro.br
ORCID: https://orcid.org/0000-0002-2089-1758

Fábio Fonseca Barbosa GOMES
Centro Universitário Jorge Amado (UNIJORGE)
E-mail: fgomes0608@unijorge.pro.br
ORCID: https://orcid.org/0000-0001-5113-9553

José Vicente Cardoso dos SANTOS
Universidade do Estado da Bahia (UNEB)
E-mail: vicentecardoso@uneb.br
ORCID: https://orcid.org/0000-0003-2501-6175

RESUMO

O texto discute a utilização de modelos de aprendizagem de máquina para análise de dados escolares, com o objetivo de fornecer uma visão geral comparativa dos diferentes modelos disponíveis para essa finalidade. Inicialmente, é apresentado alguns conceitos fundamentais sobre aprendizado de máquina e, em seguida, se descreve os modelos utilizados na análise dos dados escolares com sob o domínio da evasão escolar, incluindo árvores de decisão, redes neurais e regressão logística. Também é discutido as vantagens e desvantagens de cada modelo, bem como as situações em que cada um deles é mais adequado para ser utilizado. Por fim, ele apresenta algumas considerações sobre a importância da escolha adequada do modelo para a obtenção de resultados precisos na análise de dados escolares, destacando a importância do conhecimento especializado em aprendizado de máquina para a realização de análises eficazes.

Palavras-chave: Data Science. Analise de Dados. Machine Learning.

ABSTRAC

The text discusses the use of machine learning models for the analysis of school data, with the aim of providing a comparative overview of the different models available for this purpose. Initially, the author presents some fundamental concepts about machine learning and then describes the models used in the analysis of school data under the domain of school dropout, including decision trees, neural networks, and logistic regression. The advantages and disadvantages of each model are also discussed, as well as the situations in which each one is most suitable to be used. Finally, the author presents some considerations about the importance of choosing the appropriate model to obtain accurate results in the analysis of school data, highlighting the importance of specialized knowledge in machine learning for conducting effective analyses.

Keywords: Data Science. Data Analysis. Machine Learning.

INTRODUÇÃO

A evasão escolar é um fenômeno complexo que é influenciado por diversos fatores, como aspectos sociais, econômicos, psicológicos e educacionais (TAVARES, 2017). Diversos estudos, como os de Kira (1998), Gaioso (2005) e BAGGI et al. (2011), definem a evasão escolar como a interrupção do ciclo de estudos em qualquer nível de ensino. Kira (1998) descreve a evasão escolar como a fuga ou a perda de alunos antes da conclusão do curso. Para combater esse problema, é fundamental utilizar ferramentas analíticas eficientes, como a análise de dados e as técnicas de aprendizado de máquina.

O objetivo deste trabalho é estudar as técnicas de aprendizado de máquina no contexto da evasão escolar no ensino superior, visando propor um modelo computacional que possa ser utilizado para minimizar a evasão. De acordo com Fernández-Delgado et al. (2014), "as técnicas de aprendizado de máquina têm sido amplamente empregadas na análise de grandes conjuntos de dados e na identificação de padrões ocultos". No entanto, é crucial selecionar a técnica mais adequada para obter um bom desempenho do modelo. Conforme destacado por Miotto et al. (2018), "a escolha do algoritmo deve ser baseada nas características dos dados, no objetivo do

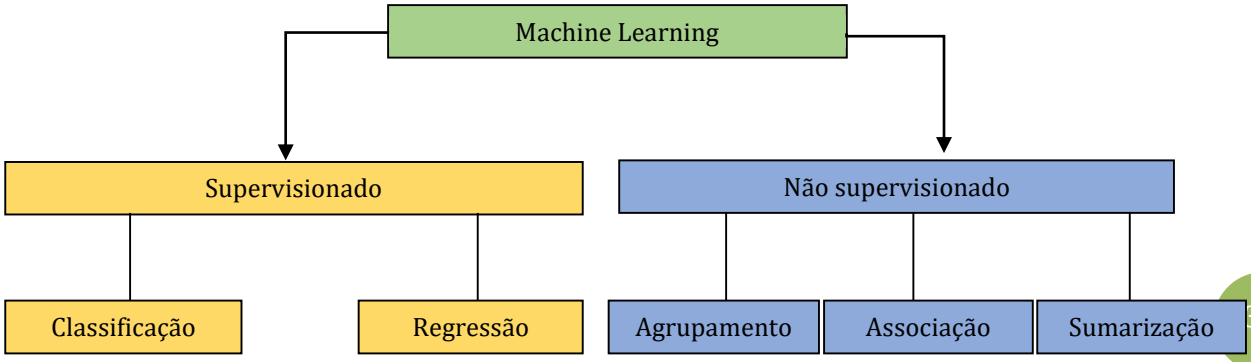
modelo e na capacidade computacional disponível". Além disso, conforme (KAYA E UYSAL, 2021), "a seleção do algoritmo correto é uma das etapas mais críticas e desafiadoras do processo de modelagem". Portanto, é essencial realizar uma análise minuciosa do problema em questão antes de escolher a técnica de aprendizado de máquina mais adequada.

A análise de dados e as técnicas de aprendizado de máquina são ferramentas valiosas para abordar o problema da evasão escolar. No entanto, a escolha da técnica adequada deve ser baseada em uma análise cuidadosa do problema em questão, levando em consideração as características dos dados, o objetivo do modelo e a capacidade computacional disponível. Para uma melhor compreensão deste trabalho, serão apresentados os conceitos de modelos de aprendizado de máquina, as técnicas utilizadas no contexto da evasão escolar no ensino superior, os procedimentos metodológicos adotados nesta pesquisa, o conjunto de testes e seus resultados, seguidos de uma discussão sobre os mesmos.

CONCEITOS DE MODELOS DE MACHINE LEARNING

Um modelo em Machine Learning é uma representação matemática que usa dados de treinamento para fazer previsões ou classificações. Ele é construído a partir de algoritmos de Machine Learning que analisam e aprendem padrões nos dados de treinamento. O objetivo é encontrar o modelo que melhor se ajuste aos dados de treinamento e que seja capaz de generalizar bem para dados desconhecidos. "Um modelo em aprendizado de máquina é uma representação matemática da relação entre os recursos de entrada e a saída desejada" - Alpaydin, E. (2010). A escolha da técnica de machine learning mais adequada para lidar com dados de evasão escolar é um desafio que depende de diversos fatores. "Os modelos de aprendizado de máquina podem ser classificados como supervisionados ou não supervisionados, dependendo da presença ou ausência de rótulos para as amostras de treinamento." Murty, M. N. (1997), Shalev-Shwartz, S., & Ben-David, S. (2014). Veja sua classificação na figura abaixo:

Figura 1: Classificação do Machine Learning



Fonte: Elaborado pelo autor (2023).

Modelos de aprendizado supervisionado: Esses modelos aprendem a fazer previsões ou classificações baseados em dados rotulados, ou seja, "O aprendizado supervisionado é o tipo de aprendizado de máquina em que o algoritmo é alimentado com amostras rotuladas e aprende a fazer previsões precisas para novos dados." (HASTIE, T., TIBSHIRANI, R., & FRIEDMAN, J. 2009), com informações sobre a saída esperada para cada exemplo de treinamento. Exemplos incluem regressão linear, árvores de decisão e redes neurais.

Modelos de aprendizado não supervisionado: Esses modelos aprendem a identificar padrões nos dados sem informações sobre a saída esperada. Eles são usados para tarefas como clustering, redução de dimensionalidade e detecção de anomalias. Exemplos incluem k-means, agrupamento hierárquico e modelos de distribuição. Em geral, os modelos supervisionados são usados quando há uma saída conhecida que desejamos prever, enquanto que os modelos não supervisionados são usados quando desejamos explorar e entender os padrões nos dados sem seguir uma saída específica.

"O aprendizado não supervisionado é uma técnica de mineração de dados que visa encontrar padrões em dados não rotulados. Diferentemente do aprendizado supervisionado, não há variáveis dependentes e, portanto, não há necessidade de treinar o modelo com exemplos rotulados" (KOTSIANTIS et al., 2007). Vale ressaltar que ainda contamos com outros dois tipos de aprendizado de máquina que são: Semi-supervisionado e por reforço, no entanto não será explorado neste trabalho, por não estarem no espectro de nossa pesquisa para proposição de um modelo para aplicação no domínio de análise de dados escolares.

TÉCNICAS DE MACHINE LEARNING: APLICAÇÃO EM EVASÃO ESCOLAR

A escolha da técnica de Machine Learning mais adequada para lidar com dados de evasão escolar é um desafio que depende de diversos fatores. Segundo Pandey et al. (2020), a escolha de uma técnica de modelagem específica depende dos objetivos do estudo, da natureza dos dados, do tamanho da amostra e das suposições subjacentes. Dentre as técnicas de machine learning comumente utilizadas para analisar dados de evasão escolar, destacam-se as árvores de decisão, que permitem a interpretabilidade dos modelos e identificação dos fatores mais relevantes para a evasão escolar (WANG et al., 2017); as redes neurais, que possuem alta capacidade de modelagem não-linear e podem lidar com dados desbalanceados (ZAFRA et al., 2021); a regressão logística, que é uma técnica estatística amplamente utilizada para análise de dados de evasão escolar (Pandey et al., 2020); e a análise discriminante, que é uma técnica de classificação que permite a distinção entre grupos de alunos evadidos e não-evadidos (SUN et al., 2019).

No entanto, é importante ressaltar que a escolha da técnica de machine learning não deve ser baseada apenas em seu desempenho em termos de acurácia, mas também em outros fatores, como a interpretabilidade dos modelos, a facilidade de implementação e a escalabilidade (NGUYEN et al., 2020). Portanto, é necessário avaliar cuidadosamente as vantagens e limitações de cada técnica e escolher aquela que melhor se adequa às necessidades específicas do projeto de análise de evasão escolar.

Árvores de decisão

As árvores de decisão são uma técnica valiosa de Machine Learning para analisar dados de evasão escolar (MICHEL, 2020). Ao utilizar árvores de decisão nesse contexto, é possível identificar os fatores mais relevantes que contribuem para a evasão escolar, fornecendo insights cruciais para o desenvolvimento de estratégias de prevenção e intervenção. Conforme Michel (2020) as árvores de decisão possibilitam a visualização clara dos padrões e a compreensão dos atributos-chave, como desempenho acadêmico, envolvimento dos pais e características socioeconômicas dos alunos. Essa abordagem permite que educadores e profissionais da área tomem

decisões embasadas e implementem medidas efetivas para reduzir a taxa de evasão e promover o sucesso dos estudantes.

Em acordo com NASEEM (et al 2019), desenvolver uma árvore de decisão baseada em diversos atributos dos alunos, como histórico acadêmico, situação socioeconômica, presença em sala de aula e participação em atividades extracurriculares torna-se crucial para o sucesso da predição. A árvore de decisão destaca o desempenho acadêmico abaixo da média e a ausência frequente às aulas eram os fatores de maior impacto na evasão escolar. Essas conclusões proporcionam aos educadores e administradores escolares informações importantes para implementar estratégias específicas, como programas de tutoria, suporte psicológico e estímulos à participação dos alunos, visando reduzir os índices de evasão (Michel, 2020).

Redes Neurais

Trata-se de uma técnica de Machine Learning que pode ser usada para analisar dados de evasão escolar. As redes neurais são particularmente úteis quando há muitas variáveis de entrada e a relação entre as variáveis é complexa. Um exemplo de rede neural aplicada à evasão escolar pode ser encontrado em (HU et al., 2020, p. 155).

As redes neurais têm sido amplamente aplicadas na área da educação para abordar o problema da evasão escolar. Com base nos princípios apresentados no livro "Inteligência Artificial: Uma Abordagem Moderna" de Russell (2013), as redes neurais são capazes de analisar um conjunto diversificado de dados, como histórico escolar, informações socioeconômicas e comportamentais dos alunos, a fim de identificar padrões e predizer quais estudantes correm maior risco de abandonar os estudos.

Segundo Russell (2013), através do treinamento de redes neurais com dados históricos de evasão escolar, é possível desenvolver modelos capazes de identificar os fatores de risco que contribuem para esse fenômeno. Esses modelos podem auxiliar educadores e gestores a tomar medidas preventivas, como oferecer suporte acadêmico personalizado, intervenções psicossociais e identificar alunos que necessitam de atenção especial. Ao utilizar abordagens baseadas em redes neurais, é possível obter um entendimento mais aprofundado das causas subjacentes da evasão escolar, permitindo o desenvolvimento de estratégias mais eficazes para prevenção e

intervenção, visando garantir que cada aluno tenha a oportunidade de completar seus estudos e alcançar seu potencial máximo (RUSSELL, 2013).

Regressão logística

É uma técnica de machine learning que pode ser usada para analisar dados de evasão escolar. A regressão logística é útil quando se deseja prever a probabilidade de evasão escolar com base em um conjunto de variáveis de entrada (NORDLUND et al., 2019, p. 19). Conforme Michell (2020), a regressão logística é uma abordagem estatística amplamente utilizada para modelar relações entre variáveis binárias, sendo aplicável à análise da evasão escolar. Ela permite estimar a probabilidade de um aluno abandonar os estudos com base em características individuais, como desempenho acadêmico, idade, gênero e histórico familiar. A regressão logística é uma ferramenta valiosa para identificar fatores de risco e desenvolver estratégias de intervenção que possam auxiliar na redução da evasão escolar.

É importante ressaltar que a regressão logística pode ser aplicada em conjunto com outras técnicas de Machine Learning, como redes neurais, para aprimorar a precisão das previsões de evasão escolar (BENGIO; RUSSELL, 2020). Essas abordagens combinadas permitem a análise de dados mais complexos e a captura de interações não lineares entre as variáveis de entrada. A utilização da regressão logística na compreensão e prevenção da evasão escolar tem se mostrado promissora, fornecendo insights valiosos para os profissionais da educação.

Análise discriminante

A análise discriminante é uma técnica de Machine Learning que pode ser usada para analisar dados de evasão escolar. A análise discriminante é útil quando se deseja distinguir entre os alunos que evadiram e aqueles que permaneceram na escola com base em um conjunto de variáveis de entrada (SUN et al., 2020, p. 162).

CONSIDERAÇÕES FINAIS

A escolha da técnica de machine learning mais adequada para analisar dados de evasão escolar depende do tamanho do conjunto de dados, da natureza dos dados e do objetivo da análise. Árvores de decisão, redes neurais, regressão logística e análise

discriminante são algumas das técnicas de Machine Learning que podem ser usadas para analisar dados de evasão escolar. Cada técnica tem suas vantagens e desvantagens, e é importante escolher a técnica mais adequada com base nas necessidades específicas do estudo.

A seleção da técnica de adequada ao Machine Learning para a análise de dados de evasão escolar deve ser feita com cuidado e consideração. O tamanho e a natureza dos dados, bem como o objetivo da análise, são fatores cruciais a serem levados em conta.

As árvores de decisão são úteis quando se busca interpretabilidade e explicabilidade, enquanto as redes neurais são poderosas para lidar com conjuntos de dados complexos e de grande escala. A regressão logística pode ser uma opção viável quando se deseja modelar a probabilidade de evasão, e a análise discriminante pode ser útil para identificar padrões e diferenças entre grupos de alunos. Em última análise, a escolha da técnica deve ser orientada pelas necessidades e objetivos específicos do estudo, levando em consideração as vantagens e desvantagens de cada abordagem.

Além disso, é importante ressaltar a importância da qualidade dos dados na análise de evasão escolar utilizando técnicas de Machine Learning. Dados incompletos, inconsistentes ou enviesados podem afetar negativamente os resultados e as conclusões obtidas.

Portanto, é fundamental realizar uma limpeza e pré-processamento adequados dos dados antes da aplicação das técnicas de análise. Além disso, é recomendável realizar uma validação cruzada e avaliação rigorosa dos modelos desenvolvidos, a fim de garantir sua robustez e eficácia. Ao considerar esses aspectos, é possível obter insights valiosos sobre a evasão escolar e desenvolver estratégias efetivas para mitigar esse problema, contribuindo para a melhoria da educação e do desenvolvimento dos alunos.

REFERÊNCIAS

ALPAYDIN, E. (2010). **Introduction to machine learning** (2nd ed.). Cambridge, MA: MIT Press. Disponível em: <https://kkpatel7.files.wordpress.com/2015/04/alppaydin_machinelearning_2010.pdf>. Acesso em janeiro de 2023.

BAGGI, Cristiane Aparecida Dos Santos e DOS SANTOS BAGGI; Cristiane Aparecida e LOPES, Doraci Alves. Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. Avaliação: **Revista da Avaliação da Educação Superior** (Campinas). [S.l: s.n.]. Disponível em: <http://dx.doi.org/10.1590/s1414-40772011000200007> . 2011.

FERNÁNDEZ-DELGADO, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. **Journal of Machine Learning Research,** 15(1), 3133-3181. Disponível em: <http://www.jmlr.org/papers/volume15/delgado14a/delgado14a.pdf.> Acesso em: 07 de janeiro de 2023.

GAIOSO, N. P. L. **O fenômeno da evasão escolar na educação superior no Brasil**. 2005. 75 f. Dissertação (Mestrado em Educação) Programa de Pós-Graduação em Educação da Universidade Católica de Brasília, Brasília, DF, 2005.

HASTIE, T., TIBSHIRANI, R., & FRIEDMAN, J. (2009). **The elements of statistical learning**: Data mining, inference, and prediction. Springer.

HU, X., LI, Y., LI, Z., LI, D., & LI, H. (2020). An improved neural network based approach for predicting student dropout in higher education**. Journal of Intelligent & Fuzzy Systems**, 38(2), 153-162.

KAYA, H., & UYSAL, A. K. (2021). Predictive analysis of student achievement using machine learning algorithms. **Education and Information Technologies**, 26(4), 4661-4684. Disponível em: https://link.springer.com/article/10.1007/s10639-021-10695-5.

KOTSIANTIS, S., ZAHARAKIS, I., & PINTELAS, P. (2007). Supervised machine learning: A review of classification techniques. **Emerging Artificial Intelligence Applications in Computer Engineering,** 160-183.

MURTY, M. N. (1997). **Data mining techniques.** New Age International.

Michel, T. (2020). **Machine Learning:** Algoritmos e Aplicações. Editora reviews. Disponível em: < https://www.inf.ufpr.br/lesoliveira/aprendizado/machine_learning.pdf>. Acesso em: jan 2023.

MIOTTO, R., et al. (2018). Deep learning for healthcare: review, opportunities and challenges. **Briefings in Bioinformatics,** 19(6), 1236-1246. Disponível em: https://academic.oup.com/bib/article/19/6/1236/5039867

M. NASEEM, K. CHAUDHARY, B. SHARMA AND A. G. LAL, "Using Ensemble Decision Tree Model to Predict Student Dropout in Computing Science**," 2019 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE),** Melbourne, VIC, Australia, 2019, pp. 1-8, doi: 10.1109/CSDE48274.2019.9162389.

NGUYEN, H. T., PHAN, V. T., & NGUYEN, H. L. (2020). Predicting Student Dropout in e-Learning Courses with Decision Tree and Random Forest Models. In **Proceedings of the 12th International Conference on Knowledge and Systems Engineering (KSE)** (pp. 225-230). doi: 10.1109/KSE50960.2020.9292729

NORDLUND, A., SÖDERBERG, P., & Hult, H. (2019). Predicting student dropout in higher education: a comparison of four machine learning methods. **Assessment & Evaluation in Higher Education,** 44(1), 19-30.

PANDEY, P., GOYAL, P., & KUMAR, V. (2020). Prediction of student's performance and dropout cases in higher education using machine learning algorithms: A survey. **Expert Systems with Applications,** 144, 113087. doi: 10.1016/j.eswa.2020.113087.

RUSSELL, STUART JONATHAN**. Inteligência artificial** / Stuart Russell, Peter Norvig; tradução Regina Célia Simille. – Rio de Janeiro: Elsevier, 2013. Disponivel em: < https://www.cin.ufpe.br/~gtsa/Periodo/PDF/4P/SI.pdf>. Acesso em jan 2023.

SHALEV-SHWARTZ, S., & BEN-DAVID, S. (2014). **Understanding machine learning:** From theory to algorithms. Cambridge University Press.

SUN, S., LIN, H., & CHEN, Y. (2019). Early prediction of student dropout using data mining: A case study in online learning. **Computers & Education,** 137, 104-115. doi: 10.1016/j.compedu.2019.04.011.

TAVARES, C. (2017). Fatores associados à evasão escolar no ensino médio: revisão sistemática da literatura. **Cadernos de Pesquisa,** 47(165), 1348-1375. Disponível em: https://www.scielo.br/pdf/cp/v47n165/1980-5314-cp-47-165-01348.pdf.

WANG, Y., LIAO, H., & YAO, X. (2017). Machine Learning for Student Retention in Higher Education: A Comparative Study. **IEEE Access,** 5, 17607-17618. doi: 10.1109/ACCESS.2017.2733542.

ZAFRA, A., LUENGO, J., & HERRERA, F. (2021). An analysis of machine learning models for student dropout prediction in higher education. **Computers in Human Behavior**, 119, 106698. doi: 10.1016/j.chb.2021.106698.