Predição de Evasão em Cursos *Online* utilizando *Machine Learning*

Maria Luísa Clemente dos Santos Tanini Vidal
Instituto Federal de Minas Gerais (IFMG) - Campus Ibirité R. Mato Grosso, 02 - Bairro - Vista Alegre, Ibirité - MG – Brasil
luisatanini@gmail.com

Vinícius Caputo Resende de Oliveira
Instituto Federal de Minas Gerais (IFMG) - Campus Ibirité R. Mato Grosso, 02 – Bairro Vista Alegre, Ibirité - MG - Brasil
viniciuscaputoresende@outlool.com.br

Thalita Vieira Sales
Instituto Federal de Minas Gerais (IFMG) - Campus Ibirité R. Mato Grosso, 02 - Bairro – Vista Alegre, Ibirité - MG - Brasil
thalitav.sales25@gmail.com

Thiago Henrique Barbosa de Carvalho Tavares
Instituto Federal de Minas Gerais (IFMG) - Campus Ibirité R. Mato Grosso, 02 - Bairro – Vista Alegre, Ibirité - MG - Brasil
thiago.tavares@ifmg.edu.br

ABSTRACT

This study investigates the application of machine learning techniques to prevent student dropout in online courses. In this context, the objective of this investigation was to implement and compare the predictive models K-Nearest Neighbors (K-NN), Naive Bayes, Support Vector Machines (SVM) and Decision Tree, in the Python programming language, in order to discover, based on the accuracy of the models, which one is most suitable for the application. The models were trained with a database taken from Kaggle, a Google data science platform.

Keywords

School Evasion, Predictive Models, Machine Learning.

RESUMO

Este estudo investiga a aplicação de técnicas de *machine learning* para prever a evasão dos alunos em cursos *online*. Nesse contexto, o objetivo dessa investigação foi implementar e comparar os modelos preditivos *K-Nearest Neighbors* (K-NN)*, Naive Bayes*, *Support Vector Machines* (SVM) e *Decision Tree,* na linguagem de programação *python*, para poder descobrir, com base na acurácia dos modelos, qual deles é o mais adequado para a aplicação. Os modelos foram treinados com um banco de dados retirado da *Kaggle*, uma plataforma de ciência de dados da Google.

Palavras-chaves

Evasão Escolar, Modelos Preditivos, Machine Learning.

1. INTRODUÇÃO

A crescente popularidade dos cursos *online* tem democratizado o acesso à educação, permitindo que um número maior de pessoas busque conhecimento e desenvolvimento profissional. Segundo o Censo da Educação Superior 2021, o ensino a distância cresceu 474% em uma década (INEP, 2022). No entanto, a diversificação das plataformas de ensino à distância trouxe novos desafios. Entre eles, a avaliação e a participação efetiva dos alunos. Logo, identificar e certificar que os estudantes permaneçam motivados é importante na eficiência e na manutenção do ensino, consequentemente, influenciando na conclusão de suas formações. De acordo com o censo de 2022, a taxa acumulada de evasão dos cursos de Educação à Distância (EAD) foi de 59% (Censup, 2022). Nesse contexto, as técnicas de *machine learning* aplicadas servem como ferramentas para analisar padrões de comportamento e prever a evasão dos alunos.

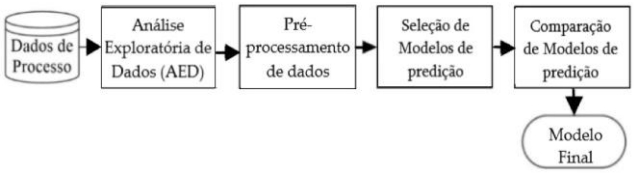
Machine learning é um conjunto de técnicas aplicadas no aprendizado de máquina por meio de treinos e testes implementados por código em grandes massas de dados (Russo et al., 2016). Para lidar com o desafio do engajamento e da evasão em cursos *online*, os modelos preditivos podem auxiliar na tomada de decisão. Isto implica que devem ser capazes de prever quais alunos estão mais propensos a evadir, com base nos resultados obtidos com a predição em comparação com os resultados esperados.

Nesse contexto, o conjunto de dados rotulados contido no banco de dados gera organização das estruturas, bem como permite realizar previsões. Para isto, podemos considerar o algoritmo de *machine learning* supervisionado, conforme mencionado por Lopez (2017).

A Figura 1 ilustra de maneira generalizada os passos utilizados para realizar a seleção do melhor modelo preditivo para determinada aplicação.

A Análise Exploratória de Dados (AED) é uma etapa introdutória que permite desenvolver familiaridade com o conjunto de dados, identificando padrões e tendências que fornecem informações importantes para a análise (Faceli et al. 2022). Já as Técnicas de Pré-Processamento de dados melhoraram a qualidade do *dataset* por meio da eliminação de valores incorretos, inconsistentes, duplicados ou ausentes, minimizando essas ocorrências (Faceli et al. 2022).

Figura 1: Processo de seleção do modelo preditivo.



Fonte: Adaptado de Cunha e Almeida, 2021.

Para lidar com o desafio do engajamento e da evasão em cursos *online*, os modelos preditivos podem auxiliar na tomada de decisão. Esses métodos são projetados para investigar os dados e identificar padrões que possam prever o comportamento dos alunos, possibilitando melhorias eficazes. A utilização de algoritmos permite a construção de modelos que podem prever o risco de evasão e ajudar a traçar estratégias de ensino com o intuito de aumentar a retenção e o sucesso dos discentes. Esses modelos também fornecem uma visão sistêmica dos dados, bem como, a praticidade na otimização da experiência educacional. (SMITH; BROWN, 2023).

Dessa forma, o objetivo central deste estudo é prever a evasão em cursos *online* utilizando os modelos preditivos *K-Nearest Neighbors* (K-NN)*, Naive Bayes*, *Support Vector Machines* (SVM) e *Decision Tree*. Além disso, também é proposto a comparação entre os modelos utilizados a fim de identificar aquele que apresentou melhor desempenho (acurácia) para o problema proposto.

2. REVISÃO BIBLIOGRÁFICA

O cenário educacional contemporâneo caracteriza-se por uma crescente adesão aos cursos *online*. Essa modalidade de ensino, impulsionada pela flexibilidade e acessibilidade que oferece, atrai um público cada vez mais amplo e diversificado. No entanto, essa expansão também traz consigo desafios, principalmente no que diz respeito ao engajamento dos alunos. (Anderson, 2014).

A análise preditiva, que utiliza técnicas de *machine learning*, tem se destacado como uma ferramenta inovadora (Baker & Inventado, 2014). Os modelos citados anteriormente são frequentemente utilizados para analisar grandes volumes de dados educacionais e prever comportamentos dos estudantes (Bishop, 2006).

O modelo *K-Nearest Neighbors* (K-NN) é um dos algoritmos de aprendizado supervisionado eficaz para problemas de classificação e regressão. Sua eficiência em diversos cenários é documentadas na literatura (Chen et al., 2020). Em ambientes de aprendizado *online*, K-NN pode ser utilizado para prever o engajamento dos alunos baseando-se em métricas como tempo de acesso, frequência de participação em fóruns e desempenho em avaliações, entre outros. Um estudo de Chen et al. (2020) demonstrou que o K-NN pode alcançar uma precisão significativa ao prever a desistência de alunos em cursos *online*, usando dados históricos de interação dos alunos com a plataforma.

O *Naive Bayes* é outro modelo preditivo, que é geralmente utilizado em contextos de classificação de texto e filtragem de *spam* (Russell & Norvig, 2016). Este modelo baseia-se no Teorema de Bayes, assumindo independência entre os atributos, o que facilita seu uso em grandes conjuntos de dados educacionais (Kaya et al., 2019). Em um estudo sobre engajamento em cursos *online*, Kaya et al. (2019) utilizaram o *Naive Bayes* para prever a participação dos alunos em atividades *online*, com resultados que mostraram uma alta taxa de acurácia.

O método SVM é amplamente utilizado em tarefas de classificação e regressão devido à sua capacidade de encontrar um hiperplano que maximiza a margem entre diferentes classes de dados. Na análise de engajamento de cursos *online*, pode ser aplicado para classificar alunos com base em seu comportamento, identificando aqueles que estão mais propensos a se engajar ou a desistir do curso. (Shah, Razzak, & Imran, 2020).

As árvores de decisão, também conhecidas como *Decision Tree*, são ferramentas amplamente utilizadas em aprendizado de máquina e estatística para tarefas de classificação e regressão. Elas operam dividindo um conjunto de dados em subconjuntos menores e mais homogêneos com base em uma série de testes de decisão realizados em diferentes atributos dos dados (Murthy et al., 2023).

A implementação desses modelos preditivos em *python* é facilitada pelas vastas bibliotecas especializadas disponíveis. A flexibilidade dessas ferramentas permite a construção de modelos robustos de maneira eficiente (Pedregosa et al., 2011). Estudos recentes têm mostrado que a combinação de diferentes modelos preditivos pode melhorar a precisão das previsões de engajamento (Khalil & Ebner, 2016). A capacidade de integrar e analisar dados de diversas fontes em tempo real torna essas técnicas úteis para o desenvolvimento de ambientes de aprendizado.

3. METODOLOGIA

O desenvolvimento deste estudo foi realizado no ambiente virtual do Google Colab, com o apoio das bibliotecas *numpy*, *pandas*, *sklearn* e *matplotlib*.

3.1. Análise exploratória dos dados

O banco de dados utilizado para o treinamento dos modelos preditivos é um arquivo CSV retirado da plataforma de ciência de dados Kaggle[1]. O arquivo é intitulado de "*Predict Online*

[1] https://www.kaggle.com/datasets/rabieelkharoua/predict-online-course-engagement-dataset/data

Course Engagement Dataset", e inclui dados demográficos do usuário, dados específicos do curso e métricas de engajamento.

Essa base de dados contém 9.000 linhas registradas distribuídas em 9 colunas, sendo essas:

- ID do Usuário: Essencial para a identificação dos alunos, mas desnecessária para a aplicação desenvolvida neste artigo.
- Categoria do Curso: Tipo de curso que cada usuário está matriculado.
- Tempo Gasto no Curso: Representa o tempo gasto, em horas, pelos usuários em seus respectivos cursos.
- Número de Vídeos Assistidos: Número total de vídeos assistidos pelos usuários em seus respectivos cursos..
- Número de Testes Realizados: Número total de testes realizados pelos usuários em seus respectivos cursos.
- Notas dos Testes: Média das notas alcançadas por cada usuário ao longo do curso.
- Taxa de Conclusão: Percentual de conteúdo do curso concluído pelo usuário.
- Tipo de Dispositivo: Tipo de dispositivo utilizado pelo usuário para realizar o curso (0 para computador e 1 para dispositivo móvel).
- Conclusão do Curso: Estado de conclusão do curso (0 para não concluído e 1 para concluído). Esta é a variável alvo para a predição dos modelos.

Foram utilizadas funções das bibliotecas citadas para conferir se a base de dados possui valores nulos ou ausentes. Além disso, foi contabilizada a distribuição da variável alvo e montada uma matriz de correlação para observar a interação entre as variáveis.

3.2. Pré-processamento de dados

Antes de prosseguir com os modelos preditivos, foram realizadas algumas alterações no *dataset*. Uma dessas alterações foi a remoção da coluna "ID do Usuário", por se tratar apenas de uma variável de identificação, não interferindo na eficácia das predições.

Identificou-se que a coluna "Categoria do Curso" se trata de uma variável categórica, sendo assim, foi aplicado o método de transformação *One-Hot Encoding* para representá-la de forma binária, em colunas exclusivas para cada curso. Essa abordagem preserva a natureza categórica dos dados e evita a introdução de uma ordenação artificial, permitindo que os modelos preditivos processem as informações de forma mais precisa e eficaz.

Em seguida, as características foram normalizadas utilizando a função *StandardScaler* da biblioteca *sklearn*, o que é indispensável para algoritmos como K-NN e SVM, que são sensíveis à escala dos dados.

3.2. Modelos preditivos

A seleção dos modelos preditivos utilizados neste artigo levou em consideração a facilidade de entendimento e interpretação dos mesmos, além da natureza dos dados.

A implementação dos modelos foi realizada com as funções da biblioteca *sklearn*. Além disso, para utilizar os modelos, os dados foram separados em 70% de treinamento e 30% de teste.

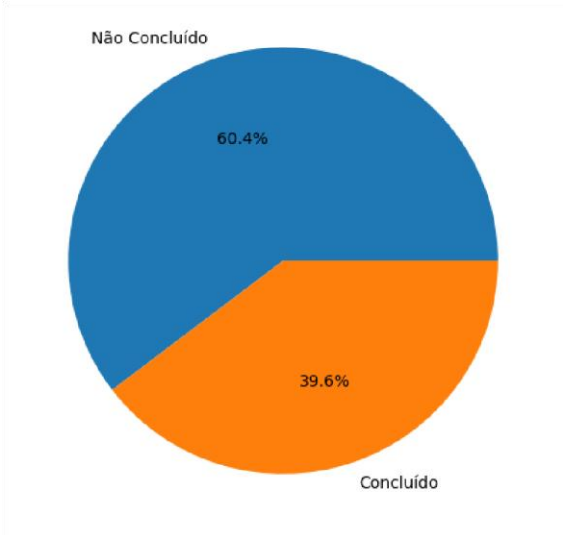
Foram realizadas, em *loop,* 50 iterações para cada um dos modelos implementados, variando os dados de treinamento e de teste de forma aleatória O cálculo da acurácia de cada modelo, em cada iteração, foi armazenado em um novo *dataframe* e, ao fim das iterações, foi realizado o cálculo da acurácia média, desvio padrão e maior acurácia para cada modelo, a fim de realizar as comparações finais e determinar qual modelo obteve o melhor desempenho.

Posteriormente, para confirmar melhorias ou mudanças na avaliação final, esse processo foi repetido, porém considerando apenas as variáveis com maior correlação com a variável alvo. Para isso, foi realizada uma breve análise da matriz de correlação e foi determinado que as características com correlação menor que 0,1 poderiam ser retiradas, sendo elas "Tipo de Dispositivo" e as colunas provenientes da coluna original de "Categoria de Curso".

4. RESULTADOS

A Figura 2 ilustra a distribuição da variável alvo, tendo em base o *dataset* sem qualquer tipo de alteração. É possível ver que o percentual de alunos que não concluíram seus cursos (60,4%) é maior que daqueles que concluíram (39,6%).

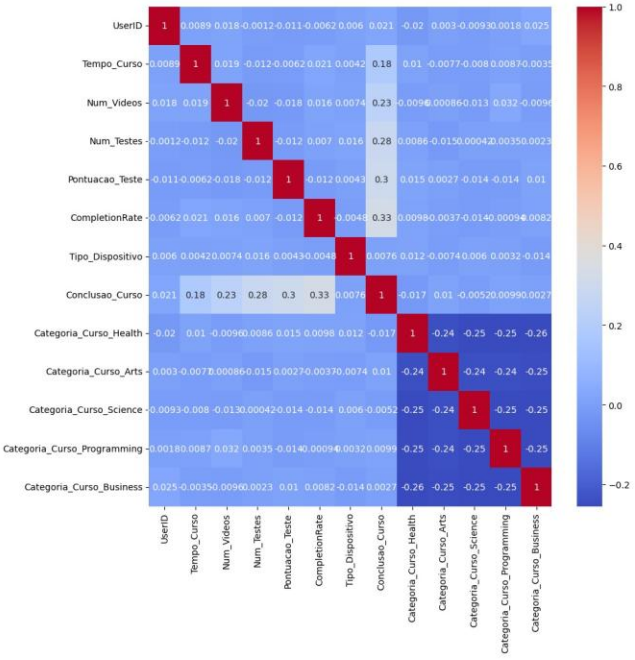
Figura 2: Distribuição da variável alvo.



Fonte: Elaborada pelos autores.

A matriz de correlação, apresentada na Figura 3, demonstra que não existem correlações muito fortes entre as variáveis do banco de dados. Em relação a variável alvo, destacam-se aquelas que possuem correlação maior que 0,1, sendo elas "Tempo Gasto no Curso", "Número de Vídeos Assistidos", "Número de Testes Realizados", "Notas dos Testes" e "Taxa de Conclusão".

Figura 3: Matriz de correlação.



Fonte: Elaborada pelos autores.

A Figura 4 apresenta as cinco primeiras linhas do *dataframe* geradas a partir das iterações em *loop*, antes de realizar a simplificação com base na matriz de correlação. Os valores exibidos são a acurácia calculada em cada iteração para cada modelo preditivo.

Figura 4: Porcentagem de conclusão do curso.

Execution	KNN	SVM	Naive Bayes	Decision Tree
1	0.857407	0.862963	0.827037	0.921481
2	0.847407	0.854444	0.821111	0.927037
3	0.853704	0.851852	0.819259	0.918148
4	0.854815	0.855185	0.811481	0.919259
5	0.850000	0.861852	0.814815	0.921111

Fonte: Elaborada pelos autores.

A Figura 5 expõe a acurácia média, o desvio padrão e a acurácia máxima alcançadas por cada modelo ao final das 50 iterações. É notável que o *Decision Tree,* em geral, alcançou os melhores valores, com 92,2%, seguido pelos modelos SVM, K-NN e Naive Bayes com 86,02%, 85,52% e 82,34% de acurácia média, respectivamente. Além disso, *Decision Tree* apresentou o menor desvio padrão com 0,005279 e 93,3% de acurácia máxima.

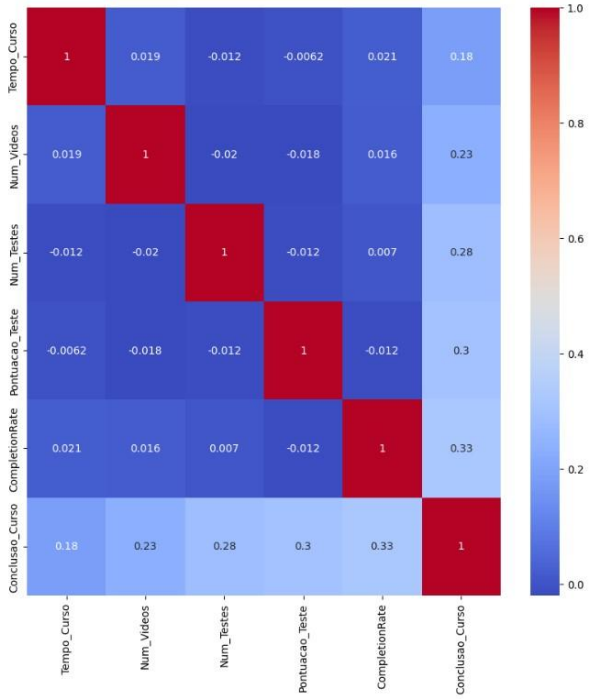
Figura 5: Desempenho dos modelos.

Modelo	Acurácia Média	Desvio Padrão	Acurácia Máxima
KNN	0.855207	0.006177	0.865926
SVM	0.860193	0.006388	0.874074
Naive Bayes	0.823430	0.006330	0.843333
Decision Tree	0.921933	0.005279	0.932963

Fonte: Elaborada pelos autores.

Já na Figura 6, é possível visualizar a matriz de correlação do banco de dados após a remoção das colunas que apresentaram menor relação com a variável alvo. Note que as novas colunas que serão utilizadas possuem relação superior a 0,1.

Figura 6: Matriz de correlação simplificada.



Fonte: Elaborada pelos autores.

Por fim, na Figura 7, é esboçado os novos resultados após executar os modelos novamente com 50 iterações, mas agora apenas com as colunas esboçadas na Figura 6.

Figura 7: Desempenho após simplificação.

Modelo	Acurácia Média	Desvio Padrão	Acurácia Máxima
KNN	0.855207	0.006177	0.865926
SVM	0.860193	0.006388	0.874074
Naive Bayes	0.823430	0.006330	0.843333
Decision Tree	0.921993	0.005585	0.934444

Fonte: Elaborada pelos autores.

É possível observar que, mesmo com as mudanças realizadas no *dataset*, os modelos continuaram apresentando os mesmos valores de acurácia média. O *Decision Tree* também continuou apresentando o menor desvio padrão, com 0,005585, e a maior acurácia, de 93,44%.

Mesmo que as diferenças entre as métricas do primeiro e do segundo teste tenham sido pequenas, isso demonstra que é possível alcançar os mesmos resultados reduzindo tempo de processamento e custo computacional, uma vez que foi possível encontrar resultados semelhantes e satisfatórios utilizando menos dados (colunas).

5. CONCLUSÃO

Este estudo explorou a aplicação de modelos preditivos de *machine learning* para prever a evasão em cursos *online*, utilizando um conjunto de dados retirado da Kaggle. A análise dos dados incluiu a geração de uma matriz de correlação, que permitiu identificar as correlações entre a variável alvo "Conclusão do Curso" e as demais variáveis do banco de dados.

Os resultados indicaram que, independentemente do banco de dados estar simplificado apenas com as variáveis de maior correlação ou não, o modelo *Deep Tree* apresentou a melhor acurácia média, com 93,34%, destacando-se entre os demais modelos utilizados. Seguido por ele, encontram-se os modelos SVM, K-NN e Naive Bayes com 86,02%, 85,52% e 82,34% de acurácia média, respectivamente. Com relação a acurácia máxima encontrada por cada modelo, o *ranking* ainda se mantém o mesmo.

Para trabalhos futuros, a exploração de redes neurais profundas e técnicas avançadas de *machine learning*, podem complementar este estudo. Tais técnicas, juntamente com a personalização dos modelos preditivos, podem proporcionar uma compreensão mais profunda dos fatores que contribuem para a evasão de cursos *online*.

6. REFERÊNCIAS

[1] Bowman, M., Debray, S. K., and Peterson, L. L. 1993. ABADI, M. et al. Tensorflow: Um sistema para aprendizado de máquina em larga escala. Em: OSDI, v. 16, p. 265-283, 2016.

[2] ANDERSON, T. The Theory and Practice of Online Learning. AU Press, 2014.

[3] BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Ensino a distância cresce 47,4% em uma década. Disponível em: https://www.gov.br/inep/pt-br/assuntos/noticias/censo-da-educacao-superior/ensino-a-distancia-cresce-474-em-uma-decada. Acesso em: 22 jul. 2024.

[4] CHEN, X.; ZOU, D.; CHENG, G.; XIE, H. Detecting and Preventing Students' Dropout with K-Nearest Neighbors and Decision Tree. Journal of Educational Technology & Society, v. 23, n. 4, p. 33-46, 2020.

[5] COVER, T.; HART, P. Classificação de padrões de vizinhos mais próximos. Transações IEEE sobre teoria da informação, v. 13, n. 1, p. 21-27, 1967.

[6] CUNHA, P. M. C.; ALMEIDA, G. M. Sensor virtual para classificação de emissões de SO2 baseado em k-NN (k-Nearest Neighbors). Revista Ibero Americana de Ciências Ambientais, v. 12, n. 9, p. 278-292, 2021.

[7] DUDANI, S. A. A regra do k-vizinho mais próximo ponderada pela distância. Transações IEEE sobre sistemas, homem e cibernética, v. 1, n. 4, p. 325-327, 1976.

[8] FACELI, K. [et al.]. Inteligência artificial: uma abordagem de aprendizado de máquina. 2. ed. - [Reimpr.]. Rio de Janeiro: LTC, 2022.

[9] KAYA, M.; SENEL, S.; DELEN, D. Predicting student performance in online learning environments using data mining techniques. Journal of Information, 2019.

[10] LOPEZ, M. A.; LOBATO, A.; MATTOS, D.; ALVARENGA, I. B.; DUARTE, O. C., et al. Um Algoritmo Não Supervisionado e Rápido para Seleção de Características em Classificação de Tráfego, 2017.

[11] MURTHY, K.; KUMAR, P.; SINGH, R. Adaptive pruning techniques for decision trees to enhance generalization. *IEEE Transactions on Neural Networks and Learning Systems*, v. 34, n. 2, p. 345-358, 2023.

[12] PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, v. 12, p. 2825-2830, 2011.