IX Encontro de Iniciação Científica e Tecnológica
IX EnICT
ISSN: 2526-6772
IFSP – *Campus* Araraquara
6 de dezembro de 2025

Felipe de Lucca[1], Miguel Cruciari Kawakami[2], Luiz Henrique Nunes[3]
1 Discente no curso Técnico de Informática Integrado ao Ensino Médio no câmpus Araraquara do IFSP. felipe.lucca@aluno.ifsp.edu.br
[2] Discente no curso Tecnólogo em Análise e Desenvolvimento de Sistemas no câmpus Araraquara do IFSP. m.kawakami@aluno.ifsp.edu.br
3 Docente no câmpus Araraquara do IFSP. lhenriquenunes@ifsp.edu.br

APLICAÇÃO DE LEARNING ANALYTICS E APRENDIZADO DE MÁQUINA PARA IDENTIFICAÇÃO E PREVENÇÃO DE EVASÃO EM CURSOS EAD

Área de conhecimento (Tabela CNPq): 1.03.04.00-2

RESUMO: O e-learning é uma modalidade de Ensino a Distância (EaD) que utiliza a internet e suas tecnologias para facilitar a transmissão de conteúdo e apoiar a instrução dos alunos, embora pesquisas indiquem que a taxa média de conclusão desses cursos seja baixa, em torno de 10%. Nesse contexto, o learning analytics—que consiste na coleta e análise dos dados gerados por estudantes em ambientes digitais—permite compreender comportamentos e atividades relacionadas ao aprendizado, possibilitando o uso de mecanismos automatizados para acompanhar e apoiar o processo educacional. Este trabalho visa aplicar métricas de learning analytics e algoritmos de aprendizado de máquina para identificar alunos com maior probabilidade de evasão nos cursos EaD, avaliando o desempenho dos modelos por meio de métricas como precisão, a fim de propor estratégias eficazes para aumentar a retenção de alunos nesse ambiente.

PALAVRAS-CHAVE: Ambiente digital; Ead; Learning Analytics; Modelos; Probabilidade.

INTRODUÇÃO

Contextualização e Justificativa

A Educação a Distância (EaD) e sua modalidade, o *e-learning*, caracterizada pela separação física entre aluno e professor (ARSHAD e SAEED, 2014; ARIFFIN, RAHMAN, et al., 2014), tem transformado o acesso à educação, permitindo flexibilidade de horário e local. A popularização de Cursos Online Abertos e Massivos (MOOCs) exemplifica essa expansão (RODRIGUES, RAMOS, et al., 2016). No entanto, esta modalidade enfrenta um desafio crítico: a baixa taxa de conclusão. Estudos indicam que a taxa média de conclusão de MOOCs é de aproximadamente 10% (Jordan, 2015; Foley et al., 2019). Um exemplo local no Instituto Federal de São Paulo (IFSP) demonstrou uma taxa de conclusão de apenas 21,5% em uma turma de Licenciatura em Pedagogia EaD, reforçando a urgência em investigar os fatores de evasão.

Neste contexto, o *Learning Analytics* — a coleta e análise de dados gerados por estudantes em Ambientes Virtuais de Aprendizagem (LMS) como o Moodle (LONG e SIEMENS, 2014) — surge como uma ferramenta poderosa. Ele possibilita a compreensão dos comportamentos dos alunos e o desenvolvimento de mecanismos automatizados de apoio. Este trabalho se justifica pela necessidade premente de aumentar a

retenção de alunos no EaD, buscando aplicar métricas de *Learning Analytics* e algoritmos de aprendizado de máquina para identificar precocemente alunos com maior probabilidade de evasão, o que é fundamental para a proposição de intervenções pedagógicas eficazes.

OBJETIVOS

Objetivo Geral

Utilizar métricas de *learning analytics* e algoritmos de aprendizado de máquina para identificar alunos propensos a evadirem dos cursos na modalidade EaD utilizando o LMS Moodle.

Objetivos específicos

- Coletar e explorar dados de atividades dos alunos (logins, tempo de acesso, participação em fóruns, submissões de tarefas e resultados de avaliações).
- Definir indicadores de desempenho e engajamento (frequência de acesso, tempo médio de conclusão de atividades e interação em fóruns).
- Selecionar e implementar algoritmos de classificação (regressão logística e árvores de decisão).
- Treinar e validar os modelos com dados históricos e técnicas como *cross-validation*.
- Medir e comparar a performance dos algoritmos, utilizando métricas como precisão.

FUNDAMENTAÇÃO TEÓRICA

O estudo baseou-se em trabalhos relacionados para estruturar o projeto, focando na aplicação de *Learning Analytics* para prevenir a evasão.

Porto, Dias e Battestin (2023) apresentam os resultados de uma Revisão Sistemática da Literatura sobre as tendências de *Learning Analytics* na plataforma Moodle, com base na análise de 24 estudos selecionados. O principal objetivo foi identificar como a análise de dados de aprendizagem está sendo utilizada nesse ambiente virtual. Como principal destaque, a pesquisa aponta o enorme potencial da análise de dados do Moodle para promover melhorias significativas na educação, como a otimização do desempenho acadêmico dos alunos, a redução das taxas de evasão, a personalização do ensino para atender necessidades individuais, o aprimoramento da interação e o consequente aumento do engajamento estudantil.

Casco et al. (2024) investiga a reconfiguração da atuação docente na era digital, com especial atenção ao contexto do e-learning. Através de uma revisão integrativa da literatura, o artigo tem como objetivo analisar as transformações nas práticas pedagógicas, explorando os desafios e as possibilidades que emergem deste cenário. O principal destaque da pesquisa é a evidência da transição do papel do professor, que assume uma postura de mediador e facilitador de processos de aprendizagem personalizados e colaborativos, superando o modelo de mero transmissor de conteúdo. Aponta-se, contudo, a persistência de desafios relacionados à infraestrutura e formação continuada.

Mônego, Martins e Hartmann (2022) avaliaram as características e potencialidades de quatro plugins de *Learning Analytics* para mapear as interações e a colaboração em um Ambiente Virtual de Ensino-Aprendizagem (AVEA) Moodle. O principal destaque da pesquisa reside na constatação de que as ferramentas analisadas viabilizam o acompanhamento e a gestão das ações dos estudantes. Isso permite ao professor realizar uma avaliação quantitativa e qualitativa das interações discentes, bem como identificar proativamente possíveis dificuldades de colaboração, possibilitando intervenções pedagógicas mais eficazes e direcionadas.

Segundo Silva e Scalabrin (2015) apresentam uma ferramenta de *Learning Analytics*, desenvolvida como um *plugin* para a plataforma Moodle, cujo objetivo é auxiliar os professores na tomada de decisões pedagógicas a partir de dados do ambiente virtual de aprendizagem. O principal destaque da ferramenta é a utilização de gráficos para identificar, de forma ágil, alunos com comportamentos que indicam tendência ao

insucesso, permitindo intervenções e comunicação mais rápidas e eficazes. A finalidade é qualificar as interações entre docentes e discentes, visando aprimorar os resultados de aprendizagem no Moodle.

Zago (2024) investiga métodos para identificar alunos em risco, analisando suas interações no Ambiente Virtual de Aprendizagem (AVA) Moodle. O objetivo principal é desenvolver um modelo de classificação usando técnicas de agrupamento baseadas em dados de interação diária. Um destaque importante do estudo é o uso do algoritmo K-means para agrupar os alunos de acordo com seus padrões de engajamento e, em seguida, analisar a correlação entre esses grupos e as notas finais dos alunos, permitindo a identificação precoce de indivíduos que podem precisar de apoio acadêmico

METODOLOGIA

Nesta seção serão detalhadas as etapas para a aplicação dos métodos de *machine learning*, divididas em preparação (coleta e pré-processamento de dados) e aplicação dos modelos.

Coleta e Exploração de Dados

Os dados foram obtidos da base de dados *Moodle grades and action logs* do Kaggle, divididos em duas tabelas CSV:

- **udk_moodle_log.csv (Moodle_log):** Contém informações sobre as atividades dos alunos, com colunas para , timecreated, eventname, action, target, userid, courseid e other.
- **udk_moodle_all_grads.csv (Moodle_all_grades):** Contém as notas dos módulos e do curso, com colunas para id, timemodified, userid, courseid, finalgrade e itemtype.

Transformação do Dataset

Os dados originais foram processados utilizando a biblioteca Pandas, em linguagem Python, com o objetivo de analisar o engajamento dos alunos em relação às atividades dos módulos. A métrica principal foi definida como o número de vezes que cada ação do ambiente virtual foi executada por um usuário em cada módulo, associando-se também a nota final correspondente.

Durante o pré-processamento, registros com nota igual a -1, bem como aqueles que apresentavam o mesmo valor de timemodified, foram desconsiderados, a fim de eliminar duplicidades e inconsistências. As notas finais foram, então, categorizadas em três grupos de risco de desempenho acadêmico:

- Alto risco, correspondendo a notas de 0 a 4, recodificadas com o valor 0
- Médio risco, correspondendo a notas de 5 a 7, recodificadas com o valor 1
- Baixo risco, correspondendo a notas de 8 a 10, recodificadas com o valor 2

O dataset final resultante contém, para cada registro, o identificador do usuário, do curso e do módulo, o valor final da nota obtida no módulo, o tempo decorrido entre a primeira ação registrada e a obtenção da nota final, além das contagens de cada tipo de ação realizada. Essa estrutura permitiu consolidar informações relevantes sobre o engajamento e o desempenho dos estudantes, servindo de base para a aplicação dos métodos de classificação.

Pré-Processamento

Valores nulos no *dataset* foram preenchidos com 0. Foram selecionadas apenas as métricas de ação que somavam no mínimo 0,5% do total de ações. O atributo da class foi definido como elemento preditivo. Utilizando a biblioteca sklearn, os dados foram divididos em 75% para treinamento e 25% para teste e medição de eficiência.

APLICAÇÃO DOS MÉTODOS DE MACHINE LEARNING

Neste trabalho, foram avaliados quatro métodos supervisionados de Machine Learning para a tarefa de classificação, implementados no ambiente Google Colab com apoio da biblioteca sklearn. Todos os métodos foram treinados utilizando um conjunto de dados composto por 28 métricas extraídas das amostras.

- **Naïve Bayes:** Utilizou-se o classificador GaussianNB, fundamentado no teorema de Bayes sob a suposição de independência entre as variáveis preditoras. Este método destaca-se por sua simplicidade computacional e rapidez, embora apresente limitações quanto à precisão em conjuntos de dados com correlações entre as variáveis.
- **Árvore de Decisão:** Empregou-se o DecisionTreeClassifier, que constrói modelos no formato de árvores binárias, promovendo decisões sucessivas baseadas nas características das amostras. Sua principal vantagem é a interpretabilidade dos resultados; contudo, o algoritmo é suscetível ao overfitting, especialmente em datasets de alta dimensionalidade.
- **Regressão Logística:** Apesar do nome, trata-se de um algoritmo de classificação binária ou multiclasse, capaz de estimar probabilidades de ocorrência de determinadas classes. O método LogisticRegression foi escolhido devido à sua robustez, facilidade de implementação e boa interpretabilidade dos coeficientes.
- **Rede Neural Artificial:** Para explorar padrões complexos não lineares, foi utilizado o classificador MLPClassifier, inspirado na estrutura de redes neurais biológicas. O número de neurônios na camada oculta foi definido como 14, correspondente à metade do número total de métricas. Redes neurais são reconhecidas por sua alta capacidade de modelagem, porém demandam maior poder computacional e ajustes específicos de hiperparâmetros.

RESULTADOS E DISCUSSÃO

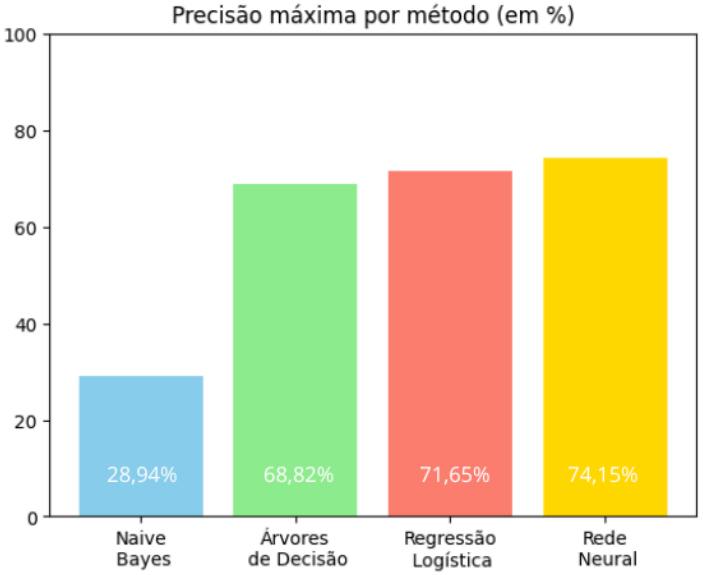
Na Figura 1 é mostrado o gráfico em barra com os testes dos algoritmos de aprendizado de máquina que refletem a precisão obtida na identificação de alunos com probabilidade de evasão.

O método baseado em Rede Neural apresentou o melhor desempenho geral, alcançando uma precisão de 74,15%, evidenciando sua eficácia na identificação de padrões complexos de comportamento estudantil. Em contrapartida, o algoritmo Naïve Bayes apresentou o menor desempenho com 28,94%, provavelmente devido à suposição de independência entre variáveis, pouco adequada ao contexto do conjunto de dados analisado.

Os métodos de Árvore de Decisão e Regressão Logística obtiveram resultados intermediários, com desempenho semelhante, embora a Regressão Logística tenha demonstrado leve vantagem em termos de precisão.

De modo geral, os resultados confirmam o potencial do uso de algoritmos de aprendizado de máquina — especialmente redes neurais artificiais — na análise de logs de interação em ambientes virtuais de aprendizagem. A identificação precoce de estudantes com maior propensão à evasão constitui uma ferramenta relevante para subsidiar ações pedagógicas preventivas, voltadas ao aumento do engajamento, da retenção e da taxa de conclusão dos cursos.

Figura 1. Gráfico da precisão dos métodos no dataset testado



Fonte: Autor

CONCLUSÕES

Este trabalho teve como objetivo principal aplicar métricas de learning analytics e algoritmos de aprendizado de máquina para identificar alunos com maior probabilidade de evasão em cursos na modalidade EaD, especificamente aqueles que utilizam o Ambiente Virtual de Aprendizagem Moodle.

Para alcançar este objetivo, a metodologia foi estruturada em etapas claras, iniciando com a coleta de dados de logs de atividades e notas de alunos (disponibilizados no Kaggle). Estes dados passaram por um processo de transformação e pré-processamento, no qual as notas foram classificadas em três níveis de risco (alto, médio e baixo) e as atividades de interação foram agrupadas em quatro conjuntos distintos de métricas, com base em sua frequência de ocorrência.

Foram implementados e avaliados quatro algoritmos de classificação distintos: Naïve Bayes, Árvore de Decisão, Regressão Logística e Rede Neural. A análise comparativa da precisão revelou diferenças significativas no desempenho de cada método. O classificador Naïve Bayes obteve os resultados mais baixos. Houve uma melhoria progressiva com o uso da Árvore de Decisão e da Regressão Logística.

Em perspectiva de dar continuidade ao trabalho realizado durante este projeto, a aplicação das métricas em outras bases de dados relacionadas ao EaD com objetivo de comparar os resultados obtidos durante este desenvolvimento podem ser realizadas. Ao comparar diferentes bases de dados, será possível identificar quais pontos fortalecem a eficácia dos métodos de Machine Learning.

REFERÊNCIAS

ARIFFIN, N. H. M. et al. **A survey on factors affecting the utilization of a Learning Management System in a Malaysian higher education**. Hawthorn: IEEE. 2014. p. 82–87.

ARSHAD, M.; SAEED, M. N. **Emerging technologies for e-learning and distance learning:** A survey. Dubai: IEEE. 2014. p. 1–6.

CASCO, Silvana T. H. et al. Analítica del aprendizaje para predecir la deserción de estudiantes universitarios. **Cuadernos de Educación**, Asunción, v. 15, n. 1, e8476, 2024.

CASSOL MÔNEGO, Leomar; RODRIGUES MARTINS, Márcio André; HARTMANN, Ângela Maria. Ambiente Moodle na formação inicial de professores: estratégia de mapeamento das interações com o uso de plugins. **RENOTE**, Porto Alegre, v. 20, n. 1, p. 11-20, 2022.

FOLEY, K. A. et al. Massive Open Online Courses (MOOC) Evaluation Methods: Protocol for a Systematic Review. **JMIR Research Protocols**, v. 8, 2019.

JORDAN, K. Massive Open Online Course Completion Rates Revisited: Assessment, Length and Attrition. **The International Review of Research in Open and Distributed Learning**, v. 16, p. 341-358, 2015.

LONG, P.; SIEMENS, G. Penetrating the fog: analytics in learning and education. **Italian Journal of Educational Technology**, v. 22, p. 132–137, 2014.

PORTO, Bruno; DIAS, Diego M.; BATTESTIN, Vania. Tendências de Learning Analytics em Moodle: uma Revisão Sistemática. **EaD em Foco**, Rio de Janeiro, v. 13, n. 1, e2070, 2023.

RODRIGUES, R. L. et al. Discovery engagement patterns MOOCs through cluster analysis. **IEEE Latin America Transactions**, v. 14, p. 4129-4135, 2016.

SILVA, Cristiane D. P. K. S. da; SCALABRIN, Edson E. Predição da Evasão na Educação a Distância Aplicando Técnicas de Mineração de Dados Educacional. *In*: CONGRESO INTERNACIONAL DE INFORMÁTICA EDUCATIVA (TISE), 20., 2015, Santiago. **Anais** [...]. Santiago: TISE, 2015. p. 821-824.

ZAGO, Eduardo. **Análise de clusters de estudantes no Moodle**: comparação entre diferentes combinações de atributos de entrada. 2024. Trabalho de Conclusão de Curso (Graduação em Engenharia de Computação) – Universidade Federal de Santa Catarina, Araranguá, 2024