UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS ARARANGUÁ
CENTRO DE CIÊNCIAS, TECNOLOGIAS E SAÚDE (CTS)
TECNOLOGIAS DA INFORMAÇÃO E COMUNICAÇÃO

Laura Giuliani De Pellegrin De Souza

MODELO PREDITIVO PARA IDENTIFICAÇÃO DE ESTUDANTES COM RISCO DE EVASÃO: Análise de Dados Acadêmicos no Moodle

Araranguá
2024

Laura Giuliani de Pellegrin de Souza

MODELO PREDITIVO PARA IDENTIFICAÇÃO DE ESTUDANTES COM RISCO DE EVASÃO: Análise de Dados Acadêmicos no Moodle

Trabalho de Conclusão de Curso submetido ao curso de Tecnologias da Informação e Comunicação do Centro de Ciências, Tecnologia e Saúde do Campus de Araranguá da Universidade Federal de Santa Catarina como requisito para a obtenção do título de Bacharela em Tecnologias da Informação e Comunicação.

Orientador(a): Prof. Marina Carradore Sérgio, Dr.(a)

Araranguá

2024

Giuliani De Pellegrin De Souza, Laura
MODELO PREDITIVO PARA IDENTIFICAÇÃO DE ESTUDANTES COM RISCO DE EVASÃO : Análise de Dados Acadêmicos no Moodle / Laura Giuliani De Pellegrin De Souza ; orientadora, Marina Carradore Sérgio, 2024.
75 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Campus Araranguá, Graduação em Tecnologias da Informação e Comunicação, Araranguá, 2024.

Inclui referências.

1. Tecnologias da Informação e Comunicação. 2. Evasão Escolar. 3. Aprendizado de máquina. 4. Predição da evasão. I. Carradore Sérgio, Marina. II. Universidade Federal de Santa Catarina. Graduação em Tecnologias da Informação e Comunicação. III. Título.

Laura Giuliani de Pellegrin de Souza

Modelo preditivo para identificação de estudantes com risco de evasão: Análise de Dados Acadêmicos no Moodle

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do título de bacharel e aprovado em sua forma final pelo Curso de Tecnologias da Informação e Comunicação.

Araranguá, 17 de dezembro de 2024.

Documento assinado digitalmente
Fabricio Herpich
Data: 21/12/2024 11:38:19-0300
CPF: ***.335.260-**
Verifique as assinaturas em https://v.ufsc.br

Coordenação do Curso

Banca examinadora

Documento assinado digitalmente
Marina Carradore Sergio
Data: 21/12/2024 12:33:30-0300
CPF: ***.746.359-**
Verifique as assinaturas em https://v.ufsc.br

Prof.(a) Marina Carradore Sérgio, Dr.(a)
Orientador(a)

Documento assinado digitalmente
Cristian Cechinel
Data: 21/12/2024 19:58:09-0300
CPF: ***.628.449-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Cristian Cechinel, Dr.
Universidade Federal de Santa Catarina

Documento assinado digitalmente
Thiago da Silva Fialho
Data: 23/12/2024 14:02:06-0300
CPF: ***.391.186-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Thiago da Silva Fialho, Mestrando
Universidade Federal de Santa Catarina

Araranguá, 2024.

AGRADECIMENTOS

Agradeço, primeiramente, à minha família, pelo amor incondicional, apoio e por acreditarem em mim em todos os momentos dessa jornada acadêmica. Cada palavra de incentivo foi fundamental para que eu persistisse.

À minha orientadora, Prof.ª Dr.ª Marina Carradore Sérgio, expresso minha mais profunda gratidão pela paciência, pelas orientações precisas e por compartilhar seu vasto conhecimento, que foram essenciais para a realização deste trabalho.

Aos meus amigos, que estiveram ao meu lado nos momentos mais desafiadores, oferecendo apoio, palavras de conforto e motivação. Vocês tornaram essa caminhada mais leve e significativa.

Ao meu namorado, agradeço o amor, compreensão e por ser meu porto seguro. Sua presença e apoio constante foram indispensáveis para que eu alcançasse mais essa conquista.

RESUMO

Este estudo apresenta o desenvolvimento de um modelo preditivo para identificar estudantes em risco de evasão escolar em cursos de graduação, utilizando dados educacionais extraídos da plataforma Moodle. O objetivo foi analisar indicadores de desempenho e comportamento dos estudantes para prever a probabilidade de evasão e auxiliar instituições de ensino na implementação de estratégias preventivas. A evasão escolar é um fenômeno complexo, influenciado por fatores como reprovação e desengajamento ao longo do semestre. Assim, o modelo desenvolvido pode ser aplicado de forma contínua durante o período letivo, utilizando dados parciais de notas e presença, além das métricas finais. Foram utilizados três algoritmos de aprendizado de máquina: Floresta Aleatória (Random Forest), XGBoost e Redes Neurais Feedforward (FNN). A avaliação dos modelos empregou métricas como acurácia, ROC AUC, precisão, recall e f1-score. Entre os algoritmos, o Random Forest apresentou o melhor desempenho, com uma acurácia de 95,65%, destacando-se pela precisão e estabilidade das previsões. Os resultados evidenciam a eficácia do modelo na identificação de padrões comportamentais que indicam risco de evasão, contribuindo para a gestão educacional e a retenção de estudantes.

Palavras-chave: Evasão Escolar; Aprendizado de Máquina; Predição de Evasão.

ABSTRACT

This study presents the development of a predictive model to identify students at risk of dropping out of undergraduate courses, using educational data extracted from the Moodle platform. The goal was to analyze student performance and behavioral indicators to predict the likelihood of dropout and assist educational institutions in implementing preventive strategies. School dropout is a complex phenomenon, influenced by factors such as academic failure and disengagement throughout the semester. Hence, the developed model can be continuously applied during the academic term, using partial data on grades and attendance as well as final metrics. Three machine learning algorithms were used: Random Forest, XGBoost, and Feedforward Neural Networks (FNN). The models were evaluated using metrics such as accuracy, ROC AUC, precision, recall, and f1-score. Among the algorithms, Random Forest showed the best performance, with an accuracy of 95.65%, standing out for its precision and prediction stability. The results highlight the model's effectiveness in identifying behavioral patterns indicative of dropout risk, contributing to educational management and student retention.

Keywords: Student Dropout; Machine Learning; Dropout Prediction.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

AD *Árvore de Decisão*
AM *Aprendizagem de Máquina*
ANN *Artificial Neural Networks*
FNN *Feedforward Neural Network*
AVA *Ambiente Virtual Aprendizagem*
DATASET *Conjunto de Dados*
DLNN Deep *Learning Neural Network*
DSRM *Design Science Research Methodology*
DP *Deep Learning*
CSV *Comma-separated values*
EDM *Educational Data Mining*
IES *Instituição de Ensino Superior*
INEP *Instituto Nacional de Estudos e Pesquisas*
JSON *JavaScript Object Notation*
KDD *Knowledge Discovery in Databases*
MEC *Ministério da Educação*
ML *Machine Learning*
MLP Multilayer Perceptron
RNA *Redes Neurais Artificiais*
RNP *Rede Neural Profunda*
RNS *Rede Neural Simples*
ROC *Receiver Operating Characteristic*
SNN *Simple Neural Network*
SVM *Support Vector Machine*
SHAP *Shapley Additive exPlanations*
TIC *Tecnologias da Informação e Comunicação*

SUMÁRIO

1 INTRODUÇÃO

A evasão escolar é um fenômeno complexo que afeta instituições de ensino em todo o mundo, sendo objeto de estudo tanto em países desenvolvidos quanto em nações em desenvolvimento. Veloso e Almeida (2013) destacam que esse fenômeno não é exclusivo de determinadas regiões ou áreas do saber, mas sim um desafio global que afeta diversas instituições de ensino, independentemente de suas características socioeconômicas ou culturais. No Brasil, a preocupação com a evasão ganhou destaque na década de 1990, com a realização do Seminário sobre evasão nas universidades brasileiras, que visava identificar as causas desse fenômeno nas universidades públicas do país e sugerir medidas para reduzir sua incidência (Kipnis, 2000).

Ao longo dos anos, diversos programas e planos educacionais, como o Programa de Apoio à Reestruturação e Expansão das Universidades Federais (Reuni) e o Plano Nacional de Educação (PNE), têm buscado implementar estratégias para combater a evasão, promovendo a permanência dos estudantes e a conclusão dos cursos (Fialho; Prestes, 2018). A evasão, de maneira geral, refere-se à interrupção definitiva dos estudos antes da conclusão de uma etapa educacional, como um curso ou uma especialização (Fialho, 2014).

Segundo Tinto (1975, 1997), a evasão escolar resulta de uma interação entre fatores pessoais, externos e o nível de integração do aluno na instituição de ensino superior (IES). Outros estudiosos, como Silva Filho *et al*. (2007), destacam os impactos econômicos e sociais da evasão, afetando tanto o setor público quanto o privado. Nesse sentido, a atuação da gestão universitária é fundamental para controlar a evasão e entender as expectativas dos alunos (Osborne; Jones, 2011).

Com o avanço da tecnologia, a análise dos dados educacionais tem se mostrado uma importante aliada na compreensão e combate à evasão. A Mineração de Dados Educacionais, por exemplo, permite que os registros dos alunos sejam transformados em informações úteis, auxiliando na criação de estratégias para reduzir a evasão (Colpani, 2018).

Diante desse cenário, o presente estudo visa explorar a aplicação de técnicas de *machine learning* para prever a evasão escolar, utilizando como base os dados coletados do Moodle em um dos cursos oferecidos por uma instituição federal. A

análise desses indicadores educacionais tem como objetivo não apenas identificar padrões que possam prever desistências, mas também sugerir medidas preventivas que contribuam para mitigar esse problema.

1.1 OBJETIVO

Analisar os indicadores educacionais coletados no ambiente virtual de aprendizagem Moodle, de um curso ofertado por uma instituição federal, a fim de prever a probabilidade alunos com risco de evasão.

1.1.1 Objetivos Específicos

- Coletar e organizar dados educacionais do Moodle relacionados ao desempenho e comportamento dos alunos;
- Aplicar técnicas de *machine learning*, para identificar alunos com maior risco de evasão;
- Analisar os padrões identificados para compreender os principais fatores que contribuem para a desistência e a reprovação;

1.2 JUSTIFICATIVA

A evasão escolar em instituições de ensino gera impactos tanto no setor público quanto no privado, resultando em desperdícios econômicos, sociais e acadêmicos (Silva Filho *et al*., 2007). Para as instituições, a perda de alunos acarreta diminuição de receitas, podendo comprometer a viabilidade de determinados cursos, especialmente no setor privado. Além disso, a evasão afeta a formação acadêmica e a carreira dos estudantes, que não concluem suas qualificações. Assim, é de grande relevância para as instituições desenvolverem métodos para identificar alunos em risco e implementar ações que possam assegurar a permanência dos mesmos.

O uso de técnicas de *machine learning* tem se mostrado promissor na previsão de eventos complexos, como a evasão escolar, ao identificar padrões que podem passar despercebidos por métodos tradicionais. A aplicação de tais técnicas no contexto educacional oferece uma oportunidade valiosa para otimizar a gestão

acadêmica e melhorar as taxas de retenção dos alunos, além de contribuir com novas abordagens de estudo sobre o tema.

1.3 ESTRUTURA DO TRABALHO

Este trabalho está estruturado em cinco capítulos. O primeiro capítulo apresenta a introdução ao tema, os objetivos, a justificativa e a estrutura do trabalho. O segundo capítulo aborda o referencial teórico, com destaque para o conceito de evasão escolar, suas causas, impactos e a utilização de técnicas de *machine learning* para previsão. O terceiro capítulo detalha a metodologia utilizada para a coleta e análise dos dados educacionais e a aplicação das técnicas de *machine learning*. O quarto capítulo apresenta os resultados obtidos com a análise dos dados e discute os achados do estudo. Por fim, o quinto capítulo oferece as conclusões e recomendações para futuras pesquisas.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 EVASÃO ESCOLAR NO ENSINO SUPERIOR

A evasão no ensino superior é amplamente debatida e definida de diferentes maneiras na literatura. Embora existam múltiplos modelos e definições sobre o tema, a maioria converge na ideia de que a evasão representa a perda do vínculo do estudante com a instituição. Essa perda de vínculo pode ocorrer de várias formas, seja pela saída voluntária do aluno, abandono do curso ou desligamento institucional, podendo incluir tanto decisões do próprio estudante quanto imposições institucionais, como desligamentos por questões acadêmicas ou administrativas (Coimbra; Silva; Costa, 2021).

Tinto (1975), um dos principais teóricos no campo da evasão escolar, propôs o Modelo Longitudinal de Evasão Institucional, que se destaca por sua abordagem processual na análise da relação entre estudantes e instituições de ensino superior. Esse modelo sublinha a necessidade de diferenciar entre as diversas formas de evasão, como a temporária e a definitiva, sendo esta última caracterizada pela ausência do aluno por um período superior a dois anos consecutivos (Barroso *et al*., 2022).

O autor também faz distinções importantes entre a evasão voluntária, na qual o estudante opta por deixar o curso, e a evasão forçada, resultante de desligamento devido ao baixo desempenho acadêmico. Sua abordagem busca entender a evasão como um conceito influenciado tanto por fatores pessoais quanto externos, culminando na integração social e acadêmica do estudante, aspecto central para determinar sua permanência ou saída da instituição (Tinto, 1975). Assim, Tinto (1975) entendia que a evasão do Ensino Superior seria uma atitude voluntária, motivada principalmente por desempenho acadêmico insatisfatório e da não integração social ao novo ambiente (Ambiel, 2015).

Em consonância com as ideias de Tinto, a Comissão Especial de Estudos sobre Evasão (1997) estabeleceu três níveis de desvinculação acadêmica. A evasão de curso ocorre quando o estudante abandona o curso em que está matriculado, seja por desistência oficial, abandono, transferência ou exclusão de acordo com normas institucionais. A evasão da instituição acontece quando o aluno rompe completamente

seu vínculo com a instituição de ensino. Já a evasão do sistema envolve a saída definitiva ou temporária do estudante do sistema de ensino superior, abandonando qualquer instituição ou programa.

Baggi e Lopes (2011) também associam a evasão à interrupção do curso em qualquer momento antes de sua conclusão. Embora o conceito proposto pela Comissão Especial de Estudos sobre Evasão (1997) tenha padronizado o entendimento sobre o tema, é essencial reconhecer suas limitações e considerar outros fatores para uma compreensão mais abrangente. Para Ambiel (2015, p. 42), essa concepção delimitou a evasão “como sendo uma decisão ativa do aluno que decide desligar-se de seu curso superior atual por sua própria responsabilidade”.

Ristoff (1997) argumenta que os atos de desligamento de um curso, instituição ou sistema de ensino superior podem ser motivados por diversos fatores, incluindo a escolha pessoal de não seguir a vida universitária, sem que isso reflita necessariamente uma falha institucional. Da mesma forma, Ristoff (1995) defende que a mobilidade de alunos, como a transferência de curso, não deve ser considerada evasão, pois muitas vezes reflete a busca do estudante por uma melhor adequação às suas aspirações pessoais e profissionais.

De modo geral, a evasão escolar refere-se à saída de estudantes que, por diversas razões, iniciam, mas não concluem seus cursos, sendo um fenômeno complexo vinculado à frustração de expectativas. Está diretamente relacionada a fatores internos e externos às instituições de ensino e deve ser compreendida no contexto mais amplo, incluindo aspectos socioeconômicos, políticos e culturais. Além disso, no campo da gestão educacional, a evasão é um indicador importante do desempenho dos sistemas de ensino, refletindo a trajetória escolar dos estudantes desde a educação básica até o ensino superior (Fritsch; Rocha; Vitelli, 2018).

Os fatores que levam à evasão escolar no ensino superior são múltiplos e complexos. Conforme apontado por Esteves *et al*. (2021), as razões para a evasão são multifatoriais e constituem fenômenos educacionais intrincados. Reconhecendo a diversidade desse fenômeno, Dias, Theóphilo e Lopes (2006) classificaram os fatores que contribuem para a evasão em externos e internos à instituição. Os fatores externos incluem aspectos vocacionais, condições socioeconômicas e problemas pessoais dos estudantes. Já os fatores internos referem-se à infraestrutura da

instituição de ensino superior (IES), ao corpo docente e à assistência socioeducacional oferecida.

Para Prestes e Fialho (2018), Esteves *et al*. (2021), os motivos que levam à evasão estão frequentemente associados a aspectos psicológicos e individuais, que desempenham um papel central na decisão do aluno de abandonar os estudos. A inter-relação desses fatores torna desafiador identificar quais têm maior impacto na decisão de evadir. Concluindo que as causas da evasão estão, em grande parte, relacionadas ao próprio aluno, incluindo a falta de tempo para dedicação aos estudos devido à carga horária de trabalho, falta de identificação com o curso e a carreira profissional, além de impedimentos financeiros e familiares (Prestes; Fialho, 2018).

A evasão escolar no ensino superior causa impactos significativos tanto para as instituições quanto para a sociedade. De acordo com Bittencourt e Mercado (2014), a evasão resulta em perdas que se refletem na ociosidade de recursos pessoais e materiais dentro das instituições de ensino, o que pode até levar ao fechamento de cursos que sofrem uma alta taxa de abandono. Esse fenômeno é observado em todos os níveis educacionais, desde a educação básica até o ensino superior, abrangendo também cursos de pós-graduação lato sensu e stricto sensu. A falta de investimentos em políticas de retenção, como destacado por Silva Filho *et al*. (2007), demonstra que as instituições frequentemente priorizam ações de marketing para captar novos estudantes, em detrimento de medidas voltadas à manutenção dos alunos já matriculados.

Além disso, os prejuízos financeiros decorrentes da evasão afetam tanto o setor público quanto o privado. Conforme Colpani (2018), a evasão escolar implica a perda de recursos que foram investidos sem o retorno esperado, além da subutilização de professores, funcionários, equipamentos e espaços físicos. Para instituições privadas, isso também significa perda de receita, tornando a questão econômica um fator relevante na busca por estratégias para reduzir a evasão.

Outro ponto relevante diz respeito aos impactos para os próprios estudantes. Como observado por Martins, Bertuci e Peniani (2020), a evasão representa a perda de importantes oportunidades de desenvolvimento pessoal e profissional, tais como acesso a melhores empregos, crescimento pessoal e melhoria na renda. Dessa forma, o abandono do ensino superior não só gera efeitos negativos para as instituições de

ensino, mas também compromete o futuro dos alunos, limitando seu potencial de crescimento e as oportunidades disponíveis.

2.2 INDICADORES EDUCACIONAIS E EVASÃO ESCOLAR

A análise dos indicadores educacionais tem se mostrado essencial para entender os fatores que contribuem para a evasão no ensino superior. Esses indicadores permitem não só identificar estudantes em risco, mas também oferecer intervenções direcionadas. Com o crescimento do uso de Ambientes Virtuais de Aprendizagem (AVAs) e a aplicação de *learning analytics*, tornou-se possível acompanhar de maneira mais precisa o engajamento dos alunos e suas dificuldades, oferecendo suporte adequado (Chicon; Paschoal; Frantz, 2021; Digiampetri; Nakano; Lauretto, 2016; Tempelaar, 2019).

Os indicadores de desempenho acadêmico são importantes para monitorar o progresso dos estudantes e identificar dificuldades que possam levar ao abandono. Esses indicadores incluem notas em avaliações, participação em fóruns, envio de tarefas, e frequência nas atividades acadêmicas. Chicon, Paschoal e Frantz (2021, p. 119) destacam que "os dados coletados pelos AVAs [...], como participação em fóruns, envio de tarefas e interações em chat, têm sido utilizados para medir tanto os indicadores comportamentais quanto os de desempenho".

Os indicadores comportamentais fornecem uma visão detalhada das interações dos alunos nos AVAs, como o número de logins, tempo de acesso, participação em fóruns, e interações com colegas e professores. Esses indicadores são particularmente relevantes no ensino a distância, pois ajudam a identificar comportamentos que podem indicar desengajamento, um dos principais preditores de evasão (Chicon; Paschoal; Frantz, 2021).

A crescente utilização de Tecnologias da Informação e Comunicação (TIC) tem transformado os Ambientes Virtuais de Aprendizagem (AVAs) em espaços que possibilitam maior interação e mediação entre professores e alunos. Conforme Maieski e Alonso (2021, p. 1431), "com o advento da cultura digital [...], os AVAs acabaram por constituir espaços de formação, possibilitadores de interação e mediação entre professores e alunos, como sujeitos do processo de

ensino/aprendizagem". Essa interação, mediada pelas TIC, é crucial para o engajamento dos estudantes e, por conseguinte, para a retenção.

O uso de *learning analytics* é destacado por Tempelaar (2019) como uma ferramenta poderosa para apoiar estudantes que enfrentam dificuldades na adaptação ao ambiente universitário. A integração dos dados comportamentais e acadêmicos dos estudantes permite um *feedback* direcionado e oportuno, possibilitando que ações de apoio sejam implementadas de maneira mais eficaz. No contexto dos AVAs, o uso desses dados para intervenções formais e informais pode ajudar a reverter trajetórias de desengajamento, aumentando as chances de retenção (Tempelaar, 2019).

Santos, Jorge e Winkler (2021) reforçam que os AVAs, com tecnologias como videoconferências e videoaulas, ampliaram as possibilidades de interação e forneceram novos dados que podem ser utilizados na análise do comportamento dos estudantes. Além disso, Digiampietri, Nakano e Lauretto (2016) afirmam que os registros de interação nos AVAs são fundamentais para a implementação de estratégias de apoio, permitindo uma ação proativa que visa reduzir os índices de evasão.

No entanto, Agudo-Peregrina *et al*. (2014) e Paschoal e Frantz (2021) apontam que ainda não existe um consenso sobre quais interações específicas devem ser analisadas para prever com maior precisão o risco de evasão. Essa falta de padronização representa um desafio para a aplicação de *learning analytics*, mas não diminui o potencial dessa abordagem na análise do comportamento dos alunos.

Os indicadores educacionais, sejam eles de desempenho acadêmico ou comportamentais, desempenham um papel fundamental na compreensão do envolvimento dos alunos e na identificação do risco de evasão. A utilização de *learning analytics*, em conjunto com os dados dos AVAs, oferece uma base sólida para desenvolver estratégias de intervenção mais eficazes. Ferramentas como os relatórios do Moodle e a integração das TIC no processo educacional promovem um acompanhamento mais próximo e eficaz do desempenho e do comportamento dos alunos, possibilitando intervenções que podem aumentar significativamente a retenção dos estudantes no ensino superior (Tempelaar, 2019; Chicon; Paschoal; Frantz, 2021; Santos; Jorge; Winkler, 2021; Maieski; Alonso, 2021).

2.3 MINERAÇÃO DE DADOS EDUCACIONAIS (*EDUCATIONAL DATA MINING*)

A Mineração de Dados Educacionais (MDE), ou *Educational Data Mining* (EDM), é uma área que aplica técnicas de mineração de dados, aprendizado de máquina e análise estatística para processar grandes volumes de dados gerados em ambientes educacionais. A MDE tem como objetivo descobrir padrões e tendências em dados educacionais, visando compreender melhor o processo de ensino-aprendizagem e aprimorar tanto a experiência dos alunos quanto as estratégias pedagógicas (Romero; Ventura, 2013; Baker; Isotani; Carvalho, 2011).

Com o avanço das Tecnologias de Informação e Comunicação (TIC) e o crescente uso de sistemas informatizados, especialmente nas instituições de ensino, a quantidade de dados gerados aumentou significativamente. De acordo com Rigo *et al*. (2014, p. 136), “o avanço das Tecnologias de Informação e Comunicação possibilitou um aumento substancial na quantidade de dados gerados e disponibilizados”. Esses dados incluem desde informações sobre o comportamento dos estudantes até registros administrativos e demográficos, permitindo uma análise detalhada de diferentes dimensões do contexto educacional (Rabelo *et al*., 2017; Romero; Ventura, 2013).

A MDE é capaz de analisar diversos tipos de dados, como o comportamento de navegação dos estudantes em AVAs, respostas em questionários e exercícios interativos, participação em atividades colaborativas e interações em fóruns. Segundo Romero e Ventura (2013), a MDE pode fazer análise de dados de qualquer sistema de informação da área educacional, tal fato abrange desde dados administrativos até interações individuais e coletivas de alunos. Os dados analisados pela MDE são complexos e envolvem várias camadas de hierarquia, como o nível de disciplina, atividades específicas, além de dados longitudinais coletados ao longo de várias sessões de aprendizado (Romero; Ventura, 2013).

Diante desse cenário, a Mineração de Dados Educacionais tem crescido como uma ferramenta essencial para processar e interpretar grandes volumes de informações em ambientes educacionais. Segundo Rabelo *et al*. (2017, p. 1528), a MDE utiliza modelos computacionais para "descobrir padrões e novas informações sobre os conjuntos de dados acerca dos ambientes de aprendizagem, seus sujeitos e as suas configurações". Essa análise permite entender melhor o comportamento dos

estudantes e facilita a criação de modelos preditivos, que são cruciais para a tomada de decisões pedagógicas e administrativas (Baker; Isotani; Carvalho, 2011).

A utilização de MDE envolve a adaptação de algoritmos de mineração de dados para lidar com as características únicas dos dados educacionais, como a hierarquia dos dados e a dependência estatística entre as interações. Costa *et al*. (2012, p. 4) observa que a MDE “procura desenvolver ou adaptar métodos e algoritmos de mineração existentes, de tal modo que se prestem a compreender melhor os dados em contextos educacionais, produzidos principalmente por estudantes e professores, considerando os ambientes nos quais eles interagem". Assim, a MDE auxilia não só na detecção de padrões de desempenho e comportamento, mas também na construção de modelos preditivos que podem orientar intervenções pedagógicas mais eficientes.

A MDE também desempenha papéis no apoio ao processo de tomada de decisão dentro das instituições educacionais. Conforme Baker, Isotani e Carvalho (2011, p. 4), "as informações sobre a relação entre dados e, posteriormente, a descoberta de novos conhecimentos, podem ser muito úteis para realizar atividades de tomada de decisão". Isso é particularmente importante para a personalização do ensino, permitindo que os educadores ajustem suas práticas de acordo com as necessidades específicas de seus alunos, melhorando a experiência educacional de maneira significativa (Rodrigues *et al*., 2014).

Além disso, a ferramenta se beneficia de sua interdisciplinaridade, combinando ciência da computação, educação e estatística. Romero e Ventura (2013) descrevem a Mineração de Dados Educacionais (MDE) como uma área interdisciplinar que abrange diversas abordagens, incluindo recuperação de informação, sistemas de recomendação, análise visual de dados, análise de redes sociais, além de aspectos ligados à psicopedagogia e psicologia cognitiva. Essa diversidade permite que a MDE se adapte a diferentes contextos educacionais e forneça uma análise abrangente e detalhada dos processos de aprendizagem.

Pimentel e Omar (2006) exploram a aplicação de técnicas de Mineração de Dados Educacionais no contexto de avaliação da aprendizagem. Eles conduziram um estudo de caso com dados coletados em quatro sessões de avaliação de lógica de programação, envolvendo 52 participantes e um total de 3.277 registros. Utilizando a ferramenta Weka para a mineração de dados, os autores aplicaram algoritmos de

classificação, como o J48, e técnicas de associação, como o algoritmo Apriori, para extrair padrões dos dados. Um dos principais resultados identificados foi a correlação entre o desempenho dos alunos em questões específicas e suas características metacognitivas, o que ajudou a identificar padrões de comportamento no processo de aprendizagem.

Rabelo *et al*. (2017) realizaram um estudo utilizando dados de 13 turmas de cursos de graduação da Universidade Federal do Rio Grande do Norte (UFRN), armazenados no ambiente de aprendizagem Moodle. O estudo analisou 64 ações associadas a 24 módulos do Moodle, envolvendo 514 alunos. Para o experimento, foram selecionados oito indicadores de desempenho, incluindo ações como login do usuário, visualização de cursos e recursos, participação em fóruns, envio de tarefas e resposta a questionários. As técnicas de Mineração de Dados aplicadas também incluíram o uso do software Weka, com foco na classificação dos dados por meio de algoritmos de árvores de decisão, como o ID3 e o J48. O melhor desempenho foi obtido com o algoritmo J48, que classificou corretamente 96,5% dos alunos, enquanto o ID3 alcançou uma acurácia de 93,97%.

Essa abordagem permitiu aos autores identificarem padrões de uso da plataforma que estão correlacionados com o sucesso ou insucesso acadêmico dos alunos, possibilitando intervenções mais precisas no processo de ensino a distância, além de sugerir melhorias na gestão de ambientes virtuais de aprendizagem (Rabelo *et al*., 2017).

Kampff, Reategui e Lima (2008) investigaram o uso de técnicas de Mineração de Dados Educacionais (MDE) para apoiar professores no acompanhamento de alunos em cursos à distância. O objetivo da pesquisa foi identificar perfis de alunos com risco de evasão ou reprovação, permitindo que o sistema gere alertas automatizados para os professores. Esses alertas seriam baseados em comportamentos identificados, e utilizou dados de logs armazenados em Ambientes Virtuais de Aprendizagem (AVA), analisando informações como quantidade de acessos, interações em fóruns e atividades submetidas, correlacionando esses dados com o desempenho acadêmico dos estudantes.

2.4 MACHINE LEARNING E EDUCAÇÃO

O aprendizado de máquina (ML), também conhecido como *machine learning*, é um ramo da ciência da computação voltado para o desenvolvimento de algoritmos que capacitam sistemas a aprender a partir de dados e melhorar seu desempenho preditivo de forma contínua (Souza, 2020). Esses algoritmos processam grandes volumes de dados, aplicando técnicas de inferência indutiva para identificar padrões, gerar conhecimento e prever eventos futuros. Contudo, para que as previsões sejam confiáveis, é essencial que os dados utilizados sejam consistentes e atualizados, já que dados de baixa qualidade podem resultar em generalizações inadequadas (Ludermir, 2021).

Dentro do aprendizado de máquina, destaca-se o *deep learning* (aprendizado profundo), um subconjunto que emprega redes neurais artificiais com múltiplas camadas e parâmetros (Shinde; Shan, 2018). Esses algoritmos representam uma evolução mais avançada e complexa dos métodos tradicionais de aprendizado de máquina, sendo inspirados na estrutura do cérebro humano e capazes de processar informações de forma hierárquica, por meio de múltiplas camadas de redes neurais (Aws, 2024).

Enquanto o aprendizado de máquina tradicional depende da intervenção humana para selecionar as características mais relevantes, geralmente exigindo dados estruturados, o *deep learning* automatiza o processamento de dados não estruturados, como imagens, textos e sons. As redes neurais artificiais (ANNs), formadas por camadas de nós, processam informações ao ativar conexões com base em valores de limiar. Já o *deep learning* expande essa estrutura, utilizando diversas camadas para acelerar avanços em áreas como visão computacional e reconhecimento de fala (Ibm, 2024).

Segundo Aws (2024), os algoritmos de aprendizado de máquina podem ser classificados em quatro categorias principais:

- **Aprendizado supervisionado:** Essa categoria envolve algoritmos treinados com dados rotulados, onde as entradas e saídas já são conhecidas. É útil para previsões e classificação de dados.
- **Aprendizado não supervisionado:** Trabalha com dados não rotulados e identifica padrões e agrupamentos automaticamente, sem intervenção humana direta.
- **Aprendizado semi supervisionado:** Combina aprendizado supervisionado e não supervisionado, utilizando uma pequena quantidade de dados rotulados para ajudar a classificar um grande volume de dados não rotulados.

- **Aprendizado por reforço:** Baseado em um sistema de recompensas e penalidades para que o algoritmo aprenda com base nas consequências de suas ações, sendo bastante aplicado em jogos e situações com dados complexos e dinâmicos.

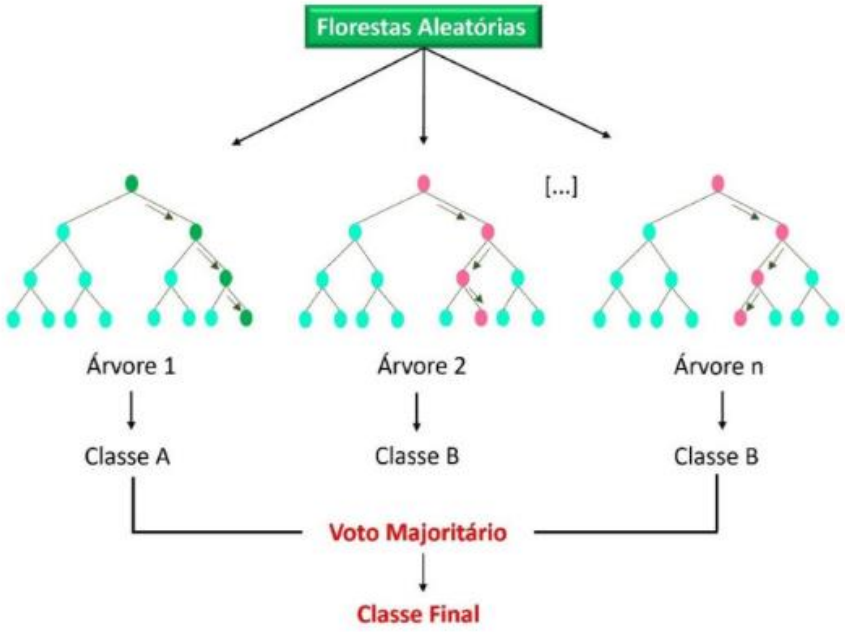
Diversos estudos, como os de Souza (2020), Malerba (2023), Moreira (2020) e Lemos (2021), utilizaram algoritmos supervisionados para realizar a modelagem preditiva da evasão escolar, utilizando técnicas de classificação e regressão para prever o comportamento de alunos com base em variáveis preexistentes. A classificação, dentro do aprendizado supervisionado, tem como objetivo rotular dados de entrada, como prever se um aluno irá ou não evadir. Já a regressão foca em prever valores contínuos, observando as relações entre as variáveis e os resultados (Souza, 2020).

2.4.1 Árvore de decisão e Floresta aleatória

A Árvore de Decisão (AD) é amplamente aplicada em problemas de classificação e regressão, reconhecida por sua simplicidade e visualização clara. Este modelo preditivo organiza os dados em uma estrutura hierárquica, partindo de um nó raiz que se ramifica até os nós folhas, que representam as classificações ou valores finais. Cada nó intermediário executa um teste baseado em atributos específicos, conduzindo a análise por meio de ramos que conectam uma decisão à próxima, até alcançar o nó terminal. O nó folha, também conhecido como nó terminal, representa o resultado após os sucessivos testes realizados desde o nó raiz. Assim estes nós se classificam em nós de probabilidade, decisão e término (Lemos, 2021; Souza, 2020; Moreira, 2020).

As árvores de decisão podem empregar algoritmos variados, como CHAID, CART e C4.5, que diferem principalmente nos critérios de particionamento e na seleção de variáveis, ajustando-se conforme a complexidade do problema e o tipo de dados (Escovedo; Koshiyama, 2020).

Figura 1 - Representação dos algoritmos de árvores de decisão e florestas aleatórias



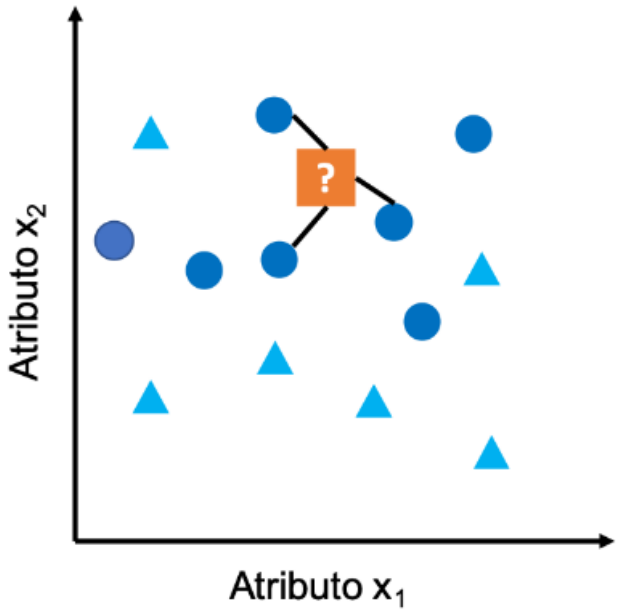
Fonte: Menezes, Scotti e Scotti (2024, p. 7).

Em uma árvore de decisão, cada caminho é ponderado pela probabilidade de sua ocorrência, permitindo avaliar o valor esperado de cada decisão. O caminho com o maior valor esperado é considerado a melhor escolha. A Figura 1 dos autores Menezes, Scotti e Scotti (2024, p. 7) apresenta esta estrutura que possibilita a representação de decisões sequenciais, destacando riscos e possíveis resultados de cada ação. No caso do algoritmo de Floresta Aleatória (*Random Forest*), ele combina várias árvores de decisão (DTs) e utiliza uma análise de consenso para identificar o resultado com maior probabilidade (Menezes; Scotti; Scotti, 2024).

2.4.2 KNN

Conforme descrito por Faria e Monteiro (2015), o algoritmo KNN, conhecido como *k-Nearest Neighbours* ou k-Vizinhos Mais Próximos, atua como um classificador cujo aprendizado é baseado em analogias. O conjunto de treinamento é composto por vetores n-dimensionais, onde cada elemento representa um ponto em um espaço de *n* dimensões. Segundo o autor Oliveira (2016) este algoritmo possui três elementos principais: um conjunto de exemplos com rótulos, uma métrica de distância e o valor de *K*.

Figura 2 - Representação do algoritmo KNN



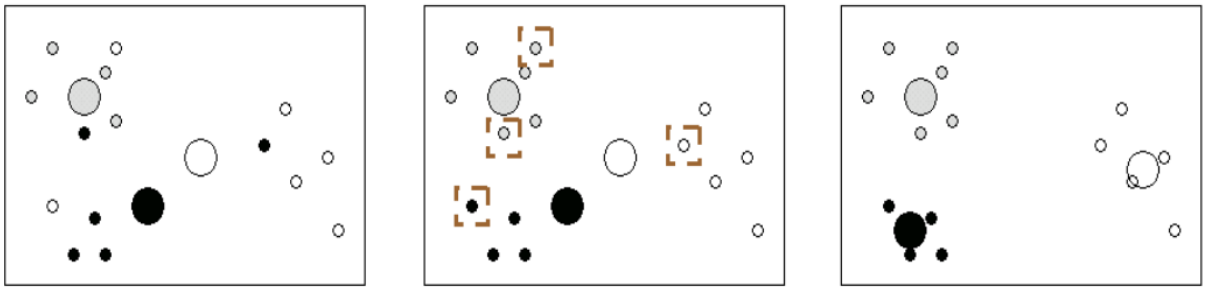
Fonte: Escovedo e Koshiyama (2020).

O algoritmo KNN funciona tanto para Classificação quanto para Regressão. O processo inicia armazenando o conjunto de dados rotulados. Para classificar um novo registro, este é comparado com todos os registros do conjunto de treinamento, visando identificar os *k* vizinhos mais próximos definidos por um parâmetro de entrada utilizando uma métrica de distância específica. A classe atribuída ao novo registro baseia-se nas classes desses vizinhos próximos identificados. Em resumo, considera-se que, nos exemplos vizinhos, a informação que se deseja inserir é similar (Escovedo; Koshiyama, 2020).

2.4.3 K-Means

O algoritmo *K-Means* é um método heurístico de agrupamento não hierárquico que visa minimizar iterativamente a distância entre os elementos e um conjunto de centros ou clusters, representados por $\chi=\{x1,x2,...,xk\}\backslash chi = \backslash\{x_1, x_2, \backslash dots, x_k\backslash\}$χ={x1 ,x2 ,…,xk }, onde kkk é o número de clusters. O processo começa com a seleção aleatória de kkk centros, e cada ponto de dados é atribuído ao centro mais próximo. Em seguida, os centros dos clusters são recalculados com base nos pontos atribuídos a eles. Esse ciclo de atribuição e reajuste continua até que não haja mais mudanças nos grupos, indicando que o algoritmo atingiu a estabilidade (Faria; Monteiro, 2015; Linden, 2009).

Figura 3 - Exemplo de execução do algoritmo de K-Means



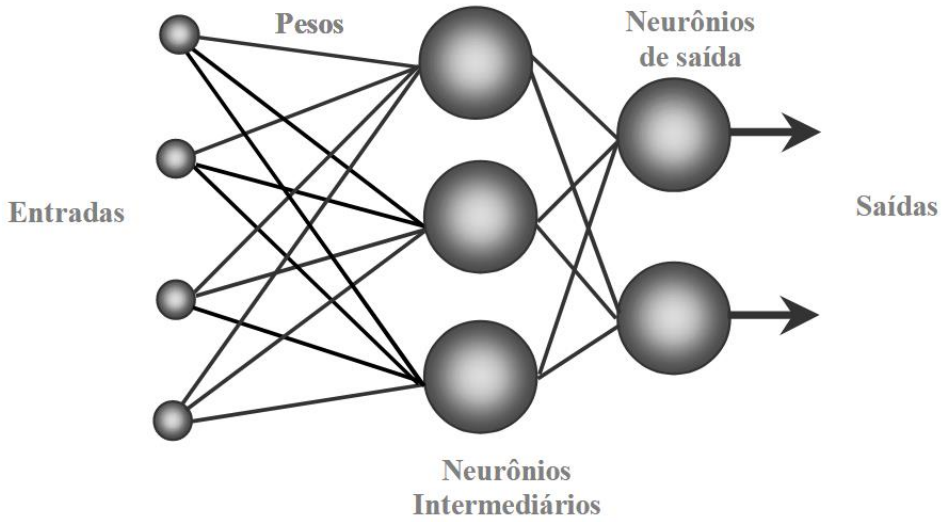
Fonte: Linden (2009, p. 24).

A Figura 3 (Linden, 2009, p. 24) ilustra o funcionamento do algoritmo *K-Means*, onde, inicialmente (a), os objetos são atribuídos aleatoriamente a um dos três grupos, e os centros de cada grupo são calculados. Em seguida (b), cada objeto é realocado para o grupo cujo centro está mais próximo. Por fim (c), os centros dos grupos são recalculados, e os grupos alcançam sua configuração final. Se o equilíbrio ainda não fosse atingido, o processo de realocação e recálculo (passos b e c) seria repetido até a estabilização dos grupos (Linden, 2009).

2.4.4 Redes Neurais Artificiais

As Redes Neurais Artificiais (RNA) são algoritmos inspirados no funcionamento do cérebro humano, compostas por unidades chamadas neurônios interconectados por sinapses com pesos ajustáveis para minimizar o erro na saída da rede, processo no qual ocorre o aprendizado. O neurônio artificial simula a estrutura e função de um neurônio biológico, onde os dendritos são substituídos por entradas conectadas ao corpo celular artificial através de pesos, representando as sinapses. Os estímulos recebidos são processados por uma função de soma, e o limiar de disparo é substituído por uma função de transferência (Malerba, 2023; Souza, 2020).

Figura 4 - Rede Neural Artificial de múltiplas camadas



Fonte: Furtado (2019, p. 11).

Uma das formas mais simples de RNA são os *perceptrons*, que organizam neurônios em camadas: entrada, onde os padrões são apresentados; ocultas (escondidas ou intermediárias), onde ocorre a maior parte do processamento; e saída, onde o resultado é exibido. Conforme observado na Figura 4, as entradas podem conectar-se a vários neurônios, produzindo múltiplas saídas, com as conexões representando sinapses biológicas. Essas conexões transformam o sinal de saída de um neurônio em sinal de entrada para outro ou direcionam o sinal para fora da rede, permitindo diversas arquiteturas de conexão entre as camadas (Furtado, 2019).

2.5 SISTEMAS DE GESTÃO DE APRENDIZAGEM (LMS) E O MOODLE

Os Ambientes Virtuais de Aprendizagem (AVAs) são plataformas digitais projetadas para facilitar a interação entre educadores e alunos, promovendo a criação de comunidades de aprendizado online. Esses ambientes têm como propósito central o compartilhamento de informações e a colaboração, elementos essenciais para o engajamento dos estudantes em seus processos de aprendizagem (Silva, 2009).

De acordo com Soares e Luciano (2004), um AVA deve oferecer mais do que uma simples interface de acesso a conteúdo: ele precisa viabilizar operações pedagógicas que incentivem a interação e a cooperação entre os participantes. Os autores Soares e Luciano (2004) também destacam a relevância de uma interface educacional reside nas operações que ela viabiliza e no plano pedagógico que orienta o processo de aprendizagem, mais do que na própria interface. Em um contexto educacional, esses ambientes são essenciais para promover o envolvimento ativo dos alunos, criando redes de aprendizado que superam as barreiras geográficas.

O Moodle, também conhecido como *Learning Management System* (LMS) ou Sistema de Gerenciamento de Aprendizagem, é amplamente utilizado para a criação de comunidades virtuais voltadas ao aprendizado (Silva, 2009). Inicialmente desenvolvido como uma plataforma de código aberto, o Moodle facilita a interação e a construção colaborativa de conteúdo, fornecendo uma variedade de ferramentas que permite a educadores e estudantes trabalharem juntos em um ambiente dinâmico e adaptável (Magnagnagno; Ramos; Oliveira, 2015).

Entre os recursos disponíveis no Moodle estão videoaulas, podcasts, simulações interativas, ebooks e áudios, além de ferramentas interativas como fóruns de discussão, avaliações online e espaços de colaboração em tempo real. Esses recursos são projetados para enriquecer a experiência dos alunos, proporcionando uma aprendizagem mais completa e integrada (Santos, 2023).

Além disso, o Moodle oferece uma série de relatórios e análises que permitem monitorar o envolvimento dos alunos, como logs de atividades, relatórios de eventos e análise de comprometimento. Esses dados registram informações detalhadas, desde a frequência de logins e duração das sessões até as páginas e recursos visitados, o que permite uma análise minuciosa do comportamento dos estudantes dentro da plataforma. Assim, o sistema facilita não apenas a aprendizagem

colaborativa, mas também o acompanhamento do desempenho acadêmico e da participação dos alunos em atividades (Porto; Dias; Battestin, 2023).

Além das ferramentas de monitoramento, o Moodle oferece opções de personalização que permitem ao professor adaptar o conteúdo às necessidades do curso e às características dos alunos. É possível, por exemplo, configurar fóruns, chats, wikis e questionários, permitindo uma flexibilização que enriquece o aprendizado e promove uma experiência única para cada turma (Magnagnagno; Ramos; Oliveira, 2015).

2.6 MODELOS PREDITIVOS NA PREVENÇÃO DE EVASÃO

A utilização de modelos preditivos no ambiente educacional tem se mostrado uma ferramenta eficaz para identificar estudantes em risco de evasão escolar. Além de descrever o fenômeno e apurar estatísticas descritivas, pesquisadores têm investido no desenvolvimento de modelos preditivos para a evasão, os quais são fórmulas matemáticas criadas a partir de dados históricos que auxiliam na identificação prévia de eventos futuros (Silva, 2021). Esses modelos fornecem a probabilidade de cada estudante evadir do curso e medem o efeito de diferentes variáveis sobre o fenômeno.

A identificação precoce dos alunos em risco permite a adoção de medidas que possam auxiliar na recuperação do estudante e evitar a reprovação na disciplina. Conforme observa Barrozo (2022, p.46):

> Um maior entendimento sobre o efeito que diversos aspectos da vida acadêmica possui sobre os alunos pode trazer diversos benefícios para o aluno. A identificação precoce permite a tomada de uma ação para que o estudante possa se recuperar e evitar uma reprovação na disciplina. As áreas que causam maiores dificuldades podem ser revisadas no planejamento de aula, melhorando a qualidade de ensino em ambos curto e longo prazo.

A utilização de modelos preditivos pode fornecer aos gestores universitários e às políticas públicas educacionais o direcionamento necessário para compreender os principais aspectos que impactam na evasão nos cursos de graduação. Ademais, esses modelos podem apontar, com alto nível de confiança, os alunos com maiores riscos de evasão. Nesse sentido, Silva, Cabral e Pacheco (2020, p. 8) destacam que

"os modelos desenvolvidos contribuem sobremaneira à gestão universitária na medida em que apontam, de forma antecipada e personalizada, a probabilidade de evadir de um dado aluno" (Silva; Cabral; Pacheco, 2020).

Para que os modelos preditivos sejam eficazes, é importante que os dados utilizados sejam corretamente coletados, tratados, armazenados, atualizados e distribuídos. A qualidade dos dados é crucial não apenas para a construção confiável dos modelos, mas também para a aplicação correta das predições aos estudantes em curso. Dados inconsistentes ou desatualizados podem levar a predições equivocadas, resultando em análises e decisões não coerentes (Silva; Cabral; Pacheco, 2020). Como ressalta (Silva, 2021, p. 63):

> Para todos os métodos, a matéria-prima que viabiliza a construção dos modelos preditivos são as bases de dados disponíveis, sejam elas advindas de sistemas institucionais, de bases governamentais, de levantamentos realizados junto ao público-alvo e/ou outros.

Em suma, a modelagem preditiva em ambientes educacionais oferece oportunidades significativas para a gestão universitária e a elaboração de políticas públicas. Ao prever, com certo grau de confiança, a conclusão ou evasão do aluno e identificar os aspectos que mais influenciam na permanência estudantil, as instituições podem desenvolver intervenções mais eficazes e direcionadas. Entretanto, é crucial garantir a validade na utilização desses modelos e assegurar a qualidade das informações utilizadas para seu desenvolvimento (Silva; Cabral; Pacheco, 2020). A identificação precoce e a compreensão dos fatores que levam à evasão possibilitam a implementação de ações que promovam a permanência do aluno antes que ele decida abandonar os estudos.

Diversos estudos têm sido conduzidos para aplicar modelos preditivos na prevenção da evasão escolar. Oliveira (2023) realizou uma pesquisa na Universidade Federal da Paraíba (UFPB) com o objetivo de desenvolver e avaliar modelos preditivos que identificam alunos com maior propensão à evasão. Utilizando técnicas de mineração de dados educacionais baseadas na metodologia CRISP-EDM, o pesquisador aplicou algoritmos de aprendizado de máquina, como Árvore de Decisão, Floresta Aleatória e Máquina de Vetores de Suporte, em dados de autoavaliação institucional. O modelo preditivo alcançou uma acurácia de 87,97%, precisão de 91,72%, recall de 91,67% e medida F de 91,57%, revelando que aproximadamente

59% dos alunos ativos admitidos entre 2017 e 2021 apresentavam maior probabilidade de abandonar seus cursos.

De forma semelhante, Oliveira (2023) propôs uma abordagem preditiva para identificar alunos em risco de evasão, utilizando dados educacionais do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB). Analisando características dos alunos (como gênero, renda familiar, idade, turno e origem escolar) e empregando algoritmos de classificação como Árvore de Decisão, Floresta Aleatória, *Naive Bayes, Multilayer Perceptron* e SVM, o estudo examinou bases de dados do Sistema Unificado de Administração Pública (SUAP) e da Plataforma Nilo Peçanha (PNP). A avaliação dos modelos através da métrica F1-Score mostrou que o conjunto de dados do SUAP obteve resultados superiores (entre 0,84 e 0,98) em comparação ao da PNP (entre 0,58 e 0,94), indicando a necessidade de abordagens distintas conforme o contexto.

Outro exemplo é o estudo de Costa, Cazella e Rigo (2015), que investigaram o problema da evasão escolar nos cursos de educação permanente na modalidade EAD para profissionais da saúde, promovidos pela Universidade Aberta do SUS (UNA-SUS). Utilizando o processo de Descoberta de Conhecimento em Bases de Dados e aplicando a técnica de árvores de decisão para regras de classificação, eles identificaram padrões que explicam esse comportamento. Como resultado, desenvolveram um modelo preditivo que alcançou 97,6% de acerto na classificação do conjunto de treinamento, contribuindo para minimizar o impacto da evasão escolar ao fornecer insights para intervenções mais eficazes.

Barrozo (2022) também contribuiu para esse campo ao propor o desenvolvimento de um método para classificar estudantes em risco de reprovação ou evasão escolar, implementado no *plugin Moodle Analytics Dashboard*. Reconhecendo que altas taxas de reprovação e evasão afetam negativamente a jornada acadêmica dos alunos como perda de motivação e oportunidade. O estudo criou um modelo preditivo baseado apenas nas interações semanais dos estudantes, conforme indicado por pesquisas anteriores. Avaliado em duas disciplinas, o modelo alcançou uma acurácia média de 56,2% nas cinco primeiras semanas, aumentando para 76,5% nas três semanas finais. A menor precisão inicial foi atribuída ao problema de portabilidade dos modelos, já que o treinamento foi realizado com disciplinas não similares.

3 METODOLOGIA

Este capítulo descreve as etapas metodológicas adotadas para a criação de um modelo preditivo de evasão escolar, utilizando dados extraídos da plataforma Moodle. A metodologia é organizada em etapas, que incluem a coleta e preparação dos dados, a construção e treinamento do modelo, e a avaliação de seu desempenho. O objetivo é construir um modelo que possa identificar, os alunos em risco de evasão e auxiliar na tomada de decisões preventivas.

3.1 DEFINIÇÃO DA PESQUISA

A pesquisa científica é uma atividade direcionada à compreensão e explicação de fenômenos, com o propósito de responder questões significativas para o entendimento da realidade. Para alcançar esse objetivo, o pesquisador fundamenta-se no conhecimento acumulado e aplica métodos e técnicas de forma rigorosa, garantindo a obtenção de resultados consistentes que respondam às suas indagações (Dias; Fernandes, 2000).

De acordo com Gil (2008), a pesquisa deve ser conduzida de maneira racional e sistemática, com o objetivo de oferecer soluções para problemas previamente formulados. Esse processo é estruturado em etapas que se iniciam com a definição do problema e se estendem até a apresentação e análise dos resultados obtidos. Complementarmente, Ander-Egg (1978) afirma que a pesquisa é um procedimento reflexivo, sistemático, controlado e crítico, que possibilita a descoberta de novos fatos, dados, relações ou leis em diferentes campos do conhecimento.

3.2 METODOLOGIA CIENTÍFICA E MÉTODO

Segundo Pereira *et al.* (2018), o método científico caracteriza-se por ser um processo sistemático que se baseia na observação organizada dos fatos, na realização de experimentos, na aplicação de deduções lógicas e na validação científica dos resultados obtidos (Pereira *et al*., 2018). Para alguns autores como Almeida (2017) e Tartuce (2006), o método científico pode ser compreendido como a aplicação da lógica à ciência, configurando-se como o caminho estruturado a ser

seguido na busca de respostas para as questões investigadas e na formulação de teorias científicas.

Nesse contexto, a metodologia científica pode ser definida como o estudo dos métodos científicos, com o objetivo de estabelecer padrões e procedimentos que garantam a condução de uma pesquisa de forma rigorosa e confiável. Esse processo visa não apenas a geração de conhecimento, mas também a validação e comprovação de fenômenos ou temas específicos no âmbito da investigação científica (Tartuce, 2006; Almeida, 2017).

3.3 TIPO DE PESQUISA

Este estudo caracteriza-se como uma pesquisa exploratória, sendo seu objetivo principal, proporcionar uma compreensão e aprofundar o conhecimento sobre a aplicação de algoritmos de aprendizado de máquina, na predição de evasão escolar utilizando o Moodle.

Segundo Theodorson e Theodorson (1970), esse tipo de pesquisa permite ao pesquisador se familiarizar com o fenômeno a ser investigado, definindo com maior precisão o problema de pesquisa e formulando hipóteses mais adequadas. Além disso, auxilia na seleção das técnicas mais apropriadas, na identificação das questões que necessitam de maior ênfase e investigação detalhada, bem como na antecipação de possíveis dificuldades e áreas de resistência.

Neste trabalho, a abordagem exploratória visa identificar padrões comportamentais e acadêmicos que possam ser usados para prever o risco de evasão. A análise dos dados coletados possibilita não apenas a compreensão das interações dos estudantes com o AVA, mas também a estruturação de um modelo preditivo eficiente. Assim, a pesquisa busca a aplicação prática de ferramentas preditivas no ambiente educacional, contribuindo para a mitigação do problema da evasão no ensino superior.

3.4 METODOLOGIA DSRM

No presente trabalho, a metodologia adotada é a Metodologia de Pesquisa em Ciência de Design (*Design Science Research Methodology*), que visa projetar e

desenvolver soluções para problemas práticos, especialmente aqueles relacionados a sistemas de informação e inovação. De acordo com Peffers et al. (2007), essa metodologia enfatiza que os pesquisadores não apenas investigam os problemas, mas também contribuem para sua resolução por meio do design de soluções eficazes. A DSRM (*Design Science Research Methodology*), é reconhecida por sua aplicabilidade na criação e avaliação de artefatos tecnológicos com o propósito de resolver problemas identificados e simultaneamente contribuir para o avanço do conhecimento científico.

Segundo Peffers, et al. (2007), em síntese, a DSRM organiza-se em um ciclo composto pelas seguintes etapas:

- Identificação do Problema
- Definição dos Requisitos
- Design da Solução
- Desenvolvimento e Implementação
- Avaliação e Confirmação
- Comunicação dos Resultados
- Refinamento Iterativo (se necessário)

Quadro 1 - Metodologia DSRM

Identificação do problema	**Exploração das causas da evasão escolar no ensino superior, utilizando dados do Moodle, visando compreender os fatores que contribuem para a evasão, bem como identificar a probabilidade de evasões.**
Definição dos requisitos	Definição de parâmetros e variáveis educacionais e comportamentais a serem analisadas, como presença, participação, interação e desempenho dos estudantes, com o objetivo de identificar padrões que possam prever o risco de evasão.
Design da Solução	Proposição de uma solução baseada no uso de técnicas de aprendizado de máquina para modelagem preditiva da evasão. Definição de quais algoritmos seriam empregados (Floresta Aleatória, *XGBoost*, Redes Neurais), além de estabelecer as métricas a serem utilizadas para avaliação dos modelos.

Desenvolvimento e Implementação	Implementação dos modelos preditivos de aprendizado de máquina com os dados pré-processados dos estudantes, incluindo a vetorização dos logs de atividades no Moodle. Utilização de técnicas como *Random Forest*, *XGBoost* e Redes Neurais para treinar e avaliar modelos com base nas variáveis comportamentais e acadêmicas dos estudantes.
Avaliação e Confirmação	Avaliação dos modelos desenvolvidos utilizando métricas como acurácia, ROC AUC, precisão, recall e análise dos fatores mais relevantes para a evasão. Comparação do desempenho dos diferentes algoritmos para confirmar a melhor abordagem no contexto específico da evasão escolar.
Comunicação dos Resultados	Apresentação dos dados obtidos através da análise dos modelos, enfatizando as variáveis mais relevantes para prever a evasão escolar, as tendências identificadas nos grupos de estudantes, e a eficácia dos diferentes algoritmos aplicados na predição.
Refinamento Iterativo	Reciclagem do processo de desenvolvimento dos modelos preditivos, se necessário, com base nos resultados da avaliação. Ajustes nos hiper parâmetros dos algoritmos, coleta de mais dados ou reformulação dos critérios de definição de evasão para melhorar a precisão e relevância da solução.

Fonte: elaborado pelo autor

3.4.1 População e Amostra

Os dados utilizados no presente estudo foram obtidos a partir do Moodle, o Ambiente Virtual de Aprendizagem (AVA) da Universidade Federal de Santa Catarina (UFSC), e referem-se aos registros de uma disciplina ofertada regularmente, de forma semestral, durante o período compreendido entre os anos de 2022 e 2024. A população total foi composta por cinco turmas, abrangendo 116 estudantes.

Com o objetivo de desenvolver e aprimorar o algoritmo de ML empregado neste trabalho, foram considerados três conjuntos de dados distintos (*datasets*), coletados ao longo de 18 semanas, período equivalente à duração média de um semestre letivo. Esses *datasets* possuem informações sobre o comportamento

acadêmico e o desempenho dos estudantes no AVA Moodle, sendo fundamentais para a análise e predição do risco de evasão escolar.

3.4.2 Coleta de Dados e Procedimentos

A coleta de dados para este estudo foi realizada a partir dos relatórios disponibilizados pela plataforma Moodle da UFSC, com exceção do relatório de presença, cujos dados precisaram ser copiados manualmente. Os relatórios utilizados foram organizados nos formatos mais adequados para a análise, abrangendo as seguintes informações:

- Logs de atividades no Moodle
- Notas das atividades avaliativas
- Registros de presença dos estudantes

Os registros das atividades realizadas pelos estudantes foram extraídos no formato JSON e organizados em uma tabela específica, visando facilitar o processamento e a análise subsequente dos dados. As notas das atividades avaliativas foram obtidas de forma independente, armazenadas no formato XLSX, enquanto os registros de presença foram consolidados manualmente em uma tabela complementar, também no formato XLSX.

3.4.3 Análise de dados

Neste trabalho, adotou-se uma abordagem quantitativa, utilizando técnicas de aprendizado de máquina para interpretar os padrões de comportamento dos estudantes no Moodle. As análises têm o objetivo de identificar fatores que contribuem para a evasão escolar, fornecendo subsídios para a criação de um algoritmo preditivo. Os procedimentos a seguir compõem a análise de dados:

- Vetorização dos Logs de Atividade: Os dados de atividades dos estudantes foram processados e transformados em variáveis quantitativas que representam o comportamento dos usuários, como o total de acessos ao AVA, interações por período, dias sem interação, entre outros.

- Criação e Treinamento de Modelos Preditivos: Desenvolveu-se modelos preditivos utilizando algoritmos como *Random Forest* (Floresta aleatória), *XGBoost* e Redes Neurais Feedforward (FNN). Cada um dos modelos foi treinado para prever a probabilidade de evasão com base em variáveis comportamentais e acadêmicas dos estudantes na plataforma Moodle, como porcentagem de presença, sequência de dias sem interação, notas parciais e eventos no sistema.
- Avaliação e Comparação dos Modelos: Os modelos foram avaliados utilizando métricas de acurácia, ROC AUC, precisão e recall. Cada um dos modelos passou por ajustes de hiper parâmetros para maximizar seu desempenho. Além disso, realizaram-se comparações entre os modelos para determinar aquele que apresentou melhor capacidade preditiva no contexto da evasão escolar.
- Visualização dos Resultados e Interpretação dos Padrões Identificados: Utilizaram-se técnicas de visualização de dados para facilitar a compreensão dos resultados obtidos. Foram criadas curvas ROC, gráficos de importância das variáveis e comparações entre os grupos identificados, permitindo uma interpretação clara dos fatores mais relevantes para a evasão.

Essas etapas possibilitaram a melhor compreensão dos comportamentos que contribuem para o risco de evasão e auxiliaram na interpretação do funcionamento dos modelos preditivos e da evasão escolar.

3.4.4 Limitações da pesquisa

Este estudo apresenta algumas limitações que devem ser consideradas ao interpretar os resultados. Primeiramente, a pesquisa se concentra em dados coletados de um único Ambiente Virtual de Aprendizagem (AVA), o Moodle, o que pode limitar a generalização dos resultados para outras plataformas de aprendizagem ou contextos educacionais. Dessa forma, a qualidade dos dados coletados está sujeita a inconsistências relacionadas ao comportamento dos estudantes no ambiente virtual,

como a variabilidade no uso dos recursos e a falta de padronização nas atividades registradas.

Outra limitação diz respeito aos modelos preditivos aplicados: embora técnicas de aprendizado de máquina tenham sido utilizadas, a predição da evasão envolve variáveis complexas e subjetivas que nem sempre são totalmente capturadas pelos dados disponíveis. Assim, os resultados devem ser interpretados como indicativos, e não determinantes, incentivando futuras investigações a explorarem contextos adicionais e a superarem essas limitações metodológicas.

3.4.5 Resultados esperados

Espera-se que esta pesquisa atinja os seguintes resultados:

- Identificação de Estudantes em Risco de Evasão: Com a aplicação de técnicas de aprendizado de máquina, busca-se identificar de forma precisa os estudantes com maior probabilidade de evasão, a partir de variáveis como frequência de acesso ao Moodle, interações no ambiente virtual, desempenho acadêmico e comportamento de uso dos recursos.
- Compreensão dos Padrões de Comportamento Associados à Evasão: Espera-se analisar os padrões comportamentais dos estudantes, de forma a compreender os principais fatores que influenciam a evasão.
- Aprimoramento na Gestão Educacional: Espera-se contribuir para a melhoria na gestão das atividades pedagógicas afim de promover estratégias de intervenção mais assertivas e personalizadas, visando minimizar a evasão escolar.

Além desses resultados diretos, esta pesquisa tem potencial para gerar contribuições acadêmicas, como a publicação dos resultados obtidos.

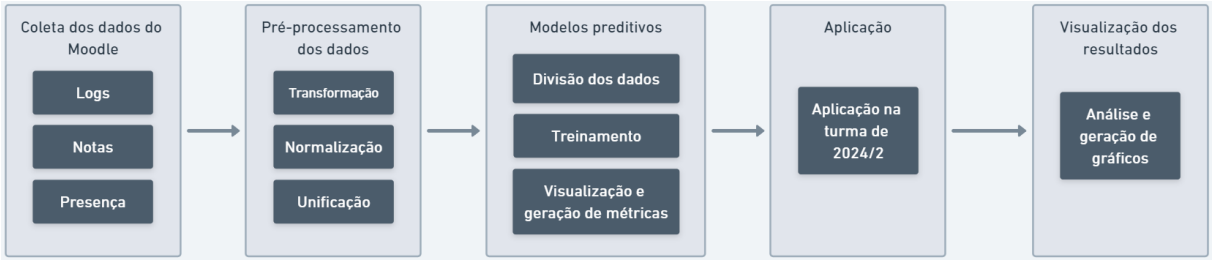
4 APRESENTAÇÃO DO MÉTODO

Neste capítulo é apresentado o método proposto e como foi construído o conjunto de análise. Nesta fase, foram utilizados três conjuntos de dados principais: logs de atividades, notas dos estudantes e registros de presença, todos extraídos do Moodle.

4.1 MÉTODO PROPOSTO

Neste capítulo serão descritos brevemente os métodos para realização do presente estudo.

Figura 5 - Visão geral do método



Fonte: elaborado pelo autor

A primeira etapa envolveu a coleta de dados a partir do ambiente Moodle. Para isso, foram extraídos relatórios de Logs de Atividades, informações de notas e registros de presença dos estudantes, os quais compuseram o conjunto de dados inicial do estudo.

Em seguida, o conjunto de dados foi submetido ao pré-processamento que envolveu a transformação, normalização e unificação dos dados em um *dataset* com as informações relevantes para a predição da evasão. A transformação consistiu na conversão de dados heterogêneos em um formato unificado e no padrão JSON, de modo a facilitar sua manipulação. Ademais, incluiu ainda a eliminação de valores anômalos e *outliers* que poderiam introduzir vieses nos modelos preditivos. A normalização visou ajustar as variáveis em uma escala comum, de forma a evitar que variáveis com amplitude maior tivessem uma influência desproporcional nos

algoritmos de aprendizado de máquina. A unificação, por sua vez, consolidou os dados em um *dataset* homogêneo e adequado para as etapas posteriores.

Para a construção dos modelos preditivos, foi realizada a divisão do *dataset* em subconjuntos de treinamento e teste, sendo utilizado um *split* de 60% dos dados para treinamento e 40% para teste. O treinamento utilizou variáveis comportamentais e acadêmicas dos estudantes, como frequência de presença e número de interações no ambiente. Durante o treinamento, foram calculadas métricas de desempenho, como acurácia, AUC ROC, precisão, recall e f1-score, cada uma fornecendo uma perspectiva diferente sobre a eficácia dos modelos.

Posteriormente, os modelos foram aplicados à turma de 2024/2 para prever o risco de evasão com base nos padrões de comportamento e desempenho acadêmico, validando os algoritmos em um cenário real de uma turma que ainda não finalizou o semestre letivo.

Por fim, os resultados foram visualizados por meio de gráficos e análises aprofundadas, como gráficos de importância de variáveis, curvas ROC e matrizes de confusão, que forneceram resultados interpretativos sobre os fatores que influenciam a evasão escolar e a eficácia comparativa dos modelos preditivos.

4.2 CONJUNTO DE DADOS

Nesta fase, foram utilizados três conjuntos de dados principais: logs de atividades, notas dos estudantes e registros de presença, todos extraídos do Moodle.

4.2.1 Logs de atividades do moodle

O *dataset* Logs de atividades no Moodle contém os registros das interações realizadas pelos estudantes dentro do Moodle. Esses registros são gerados automaticamente pela plataforma e fornecem informações sobre as atividades realizadas, horários de acesso e comportamento dos usuários. Cada entrada no *dataset* inclui os seguintes campos:

- "hora": Indica a data e o horário exatos da ação registrada, permitindo o acompanhamento cronológico das interações dos estudantes.

- "nomecompleto": Apresenta o nome completo do estudante e sua matrícula entre parênteses, garantindo a identificação individual de cada usuário.
- "usurioafetado": Identifica se a ação teve impacto em outro usuário ou se está relacionada apenas ao estudante que realizou a interação.
- "contextodoevento": Especifica o contexto no qual a ação ocorreu, como o curso ou a disciplina correspondente.
- "componente": Indica o módulo ou funcionalidade específica do Moodle que gerou o log, como "Logs" ou outras áreas da plataforma.
- "nomedoevento": Descreve o tipo de evento registrado, como "Relatório de log visto", indicando a ação realizada pelo usuário.
- "descrio": Fornece uma descrição mais detalhada do evento, especificando os IDs de usuários ou cursos envolvidos na interação.
- "origem": Informa a origem da interação, como "web", que indica que o acesso foi realizado por meio de um navegador.
- "endereoip": Registra o endereço IP do dispositivo utilizado para a interação.

4.2.2 Notas

O *dataset* Notas contém informações relacionadas ao desempenho acadêmico dos estudantes em atividades e avaliações realizadas no Moodle. Cada entrada no *dataset* inclui os seguintes campos:

- Nome e Sobrenome: Identificam o estudante de forma única no contexto da disciplina, permitindo cruzamento com outros dados, como logs de atividades e presenças.
- Endereço de e-mail: Informação de contato associada ao estudante, utilizada internamente para comunicações e identificação no sistema.
- Curso: Representa o código ou identificação do curso ao qual a disciplina está vinculada, permitindo a categorização dos dados.

- Nome do Curso: Descrição textual da disciplina ou curso, facilitando a identificação pelo usuário.
- Matrícula: Número de registro único que identifica o estudante no sistema acadêmico.
- Sequência das Notas das Atividades: Lista de notas obtidas pelo estudante em cada atividade avaliativa da disciplina.
- Sem nota total (Real): Soma parcial das notas do estudante que desconsidera atividades ainda não avaliadas ou com peso zero.
- Notas (Real): Total geral das notas obtidas pelo estudante considerando todas as atividades avaliadas.
- Total do Curso (Real): Nota final consolidada do estudante ao término da disciplina.
- Último download realizado neste curso: Indica a data e horário do último download de material realizado pelo estudante na disciplina.

4.2.3 Registro de presença

O *dataset* de presenças foi construído a partir dos registros de frequência dos estudantes ao longo do semestre, fornecidos em formato Excel. Esse conjunto de dados contém informações de presença e ausência para cada aluno em uma série de sessões de aula, identificadas por datas específicas.

Cada linha do *dataset* representa um aluno, identificado pela coluna Nome / Sobrenome cujo inclui a matrícula, permitindo a análise individual da frequência ao longo do curso. As colunas subsequentes, identificadas pelas datas (29/ago, 31/ago, 05/set, etc.), indicam a presença ou ausência dos estudantes em cada sessão de aula ao longo do semestre. Além das datas de presença, o dataset inclui as seguintes colunas adicionais que facilitam a análise quantitativa da frequência:

- P1 e P2: Indicam a presença dos alunos em uma ou ambas as aulas.
- Au: Representa o total de ausências acumuladas pelos estudantes ao longo do semestre.
- Sessões: Indica o número total de sessões realizadas no período considerado.

- Pontos: Relaciona-se aos pontos acumulados pelos estudantes, considerando sua participação e presença nas aulas.
- Porcentagem: Refere-se ao percentual de frequência de cada aluno, calculado com base no total de sessões realizadas e na quantidade de presenças registradas.

4.3 PRÉ-PROCESSAMENTO DOS DADOS

Nesta etapa, os *datasets* obtidos do Moodle foram preparados e ajustados para viabilizar o experimento de construção dos modelos preditivos. O pré-processamento envolveu as etapas sequenciais descritas a seguir:

1. Transformação dos dados: Os dados coletados a partir dos arquivos foram convertidos para um formato estruturado, utilizando scripts específicos para a extração das informações relevantes em formato JSON.
2. Unificação dos *Datasets*: Os dados de diferentes fontes foram combinados para criar um único *dataset* consolidado.

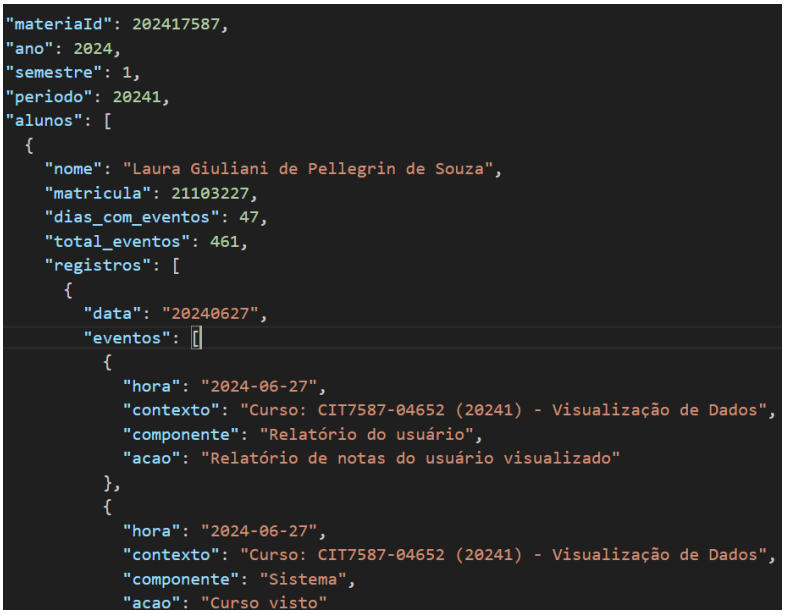
4.3.1 Transformação dos Dados

O pré-processamento dos dados foi realizado através de um *script* em Python desenvolvido especificamente para selecionar e unificar os *datasets* de todos os períodos. Durante essa etapa, os dados foram reorganizados e salvos no formato JSON, garantindo uma estrutura padronizada e de fácil acesso para as etapas subsequentes. Todos os arquivos resultantes foram armazenados em uma pasta chamada "*Outputs*", que contém os *datasets* processados, organizados por "materiaId", "ano", "semestre", "periodo", seguidos pela subdivisão "alunos", onde cada estudante possui suas informações específicas.

No *dataset* de logs representado na Figura 6, a estrutura dos dados de cada aluno inclui os campos "nome", "matricula", "dias_com_eventos", "total_eventos" e "registros". A variável "dias_com_eventos" apresenta a contagem de dias em que o estudante teve interações registradas, enquanto "total_eventos" reflete o número total de eventos realizados pelo usuário. A seção "registros" agrega todos os registros de

interação do aluno no sistema, proporcionando uma visão detalhada das atividades ao longo do tempo. A imagem a seguir ilustra a estrutura final do dataset JSON de Logs, destacando a organização das informações coletadas.

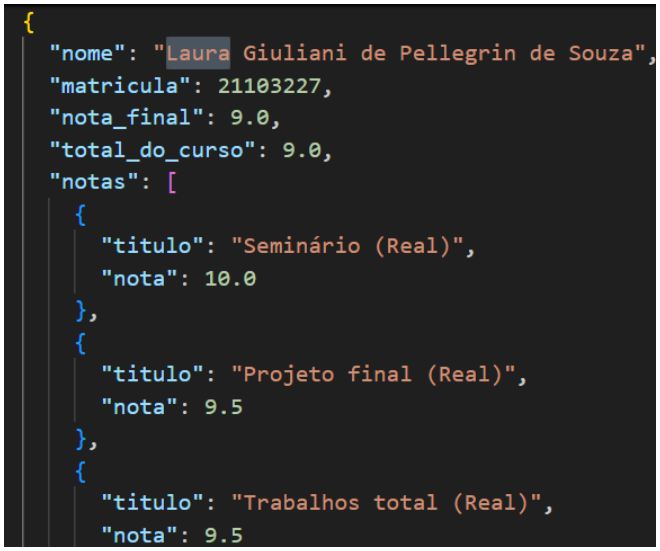
Figura 6 - Estrutura JSON dos Log



Fonte: elaborado pelo autor

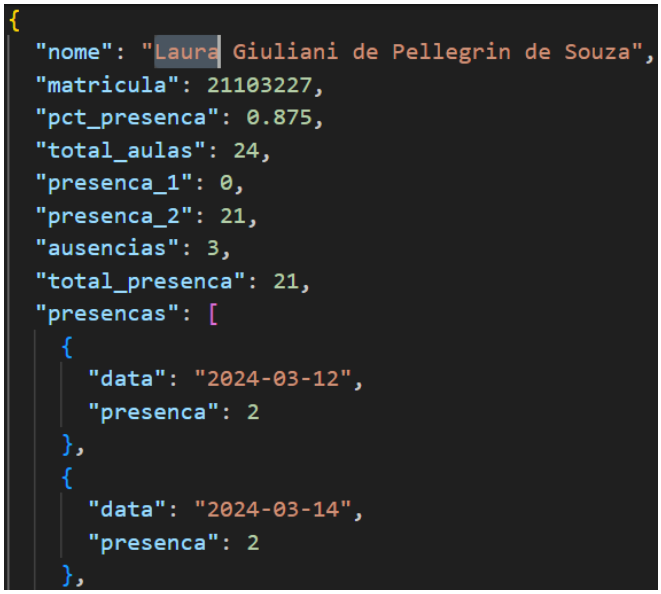
No *dataset* de notas, conforme apresentado na figura 7, a estrutura dos dados de cada aluno inclui os seguintes campos: "nome", "matricula", "nota_final", "total_do_curso" e "notas". A variável "nota_final" representa a nota final obtida pelo estudante ao término do curso, enquanto "total_do_curso" refere-se ao total das notas considerando todas as atividades avaliadas. A seção "notas" contém as notas parciais ao longo do curso, permitindo um acompanhamento detalhado do progresso de cada estudante.

Figura 7 - Estrutura JSON Notas



Fonte: elaborado pelo autor

Figura 8 - Estrutura JSON das Presenças



Fonte: elaborado pelo autor

No que se refere ao *dataset* de presenças representado na figura 8, cada aluno possui a seguinte estrutura de dados: "Nome", "matricula", "pct_presenca", "total_aulas", "presenca_1", "presenca_2", "ausencias", "total_presenca" e "presencas". O campo "pct_presenca" indica a porcentagem de presença do estudante ao longo do curso, enquanto "total_aulas" refere-se ao número total de aulas oferecidas. Os campos "presenca_1" e "presenca_2" fornecem informações

detalhadas sobre a frequência do aluno em períodos específicos, enquanto "ausencias" e "total_presenca" indicam, respectivamente, o número total de ausências e presenças. A seção "presencas" contém registros mais detalhados da frequência dos estudantes ao longo do curso.

4.3.2 Normalização dos Dados

O conjunto de dados analisado apresentou características variadas em termos de distribuição das variáveis, sendo que algumas apresentaram desbalanceamentos relevantes que poderiam impactar os modelos preditivos. Dentre as variáveis mais desbalanceadas, destaca-se a maior sequência de dias sem interações, com média de 64 dias e valores que variam de 4 a 402 dias. A maioria dos estudantes apresenta períodos curtos de inatividade, enquanto uma minoria significativa apresenta longos períodos consecutivos sem interações, evidenciando um desbalanceamento extremo. De forma semelhante, a variável dias sem interações revelou grande dispersão, com uma média de 112 dias e valores que variam de 9 a 443 dias. Enquanto a maior parte dos estudantes acumula até 150 dias sem interações, há casos extremos que ultrapassam 400 dias, refletindo padrões de engajamento muito distintos.

A variável ausências também apresentou desbalanceamento considerável. Enquanto a maioria dos estudantes apresentou de 3 a 8 faltas, correspondendo a 50% dos casos, valores extremos chegaram a 26 ausências, o que pode refletir um indicativo importante de risco de evasão. Já as notas parciais ausentes demonstraram desbalanceamento moderado, com uma média de 5,6 e valores entre 0 e 15, sendo mais frequente a concentração entre 4 e 8 notas ausentes.

Variáveis como porcentagem de presença e soma de interações, embora menos desbalanceadas, também apresentaram padrões que requerem atenção. A porcentagem de presença mostrou concentração acima de 73%, com valores variando de 3,7% a 100%. A soma de interações, por sua vez, apresentou uma média de 418 interações, com valores que variam de 15 a 1.066, sendo que poucos estudantes alcançam níveis muito altos de interação. Variáveis como total de aulas e total de presenças apresentaram distribuição mais homogênea, com menos variação, refletindo uma consistência maior no número de aulas e frequência dos estudantes. Outras variáveis, como total do curso (referente ao progresso dos estudantes) e

período, demonstraram desbalanceamentos menos expressivos. O progresso dos estudantes variou de 0 a 10 semestres, com uma média de 6,4, enquanto o período mostrou pouca variação devido à concentração de dados em um intervalo específico.

Para lidar com os desbalanceamentos identificados, foram empregadas diversas estratégias no processamento dos dados. Inicialmente, foram utilizadas técnicas de oversampling e undersampling para equilibrar a proporção entre as classes minoritárias e majoritárias, especialmente nas variáveis mais desbalanceadas. Nos algoritmos de aprendizado de máquina (*Random Forest,* XGBoost e Redes Neurais) foram ajustados pesos para as classes, de forma a mitigar os impactos do desbalanceamento nos resultados preditivos. Adicionalmente, as variáveis com escalas muito diferentes, como dias sem interação e soma de interações, foram normalizadas e escalonadas para evitar que valores extremos dominassem os cálculos dos algoritmos.

4.3.3 Unificação dos Datasets

Após a estruturação dos dados no formato JSON, procedeu-se à seleção e unificação das informações pertinentes para a utilização no modelo preditivo, resultando em um conjunto de dados exportado em formato CSV. O dataset final unificado abrange as informações extraídas dos logs de atividades, das notas dos estudantes e dos registros de presença. A seguir, apresenta-se uma descrição detalhada dos campos que compõem o dataset:

- finalmatricula: Identificador único do estudante.
- nota_final: Nota final do estudante no curso.
- periodo: Período acadêmico da disciplina (formato "ano/semestre").
- notas_parciais: Lista de notas obtidas em avaliações intermediárias.
- notas_parciais_ausentes: Número de avaliações intermediárias que não foram realizadas.
- pct_presenca: Porcentagem de presença nas aulas.
- total_aulas: Total de aulas realizadas no período.
- ausencias: Número total de ausências do estudante.
- total_presenca: Número total de aulas em que o estudante esteve presente.

- soma_interacoes: Total de interações registradas no Moodle.
- dias_sem_interacoes: Número de dias em que o estudante não realizou nenhuma interação.
- maior_sequencia_dias_sem_interacoes: Maior sequência de dias consecutivos sem qualquer interação no Moodle.

4.4 MODELOS PREDITIVOS

Após o pré-processamento dos dados, foram desenvolvidos modelos preditivos com o objetivo de prever a evasão escolar. A metodologia empregada foi estruturada de forma contínua, abrangendo a escolha dos algoritmos, a divisão dos dados, o treinamento, a avaliação e a interpretação dos resultados.

Neste trabalho foram selecionados três algoritmos principais para a tarefa de previsão da evasão escolar: Floresta Aleatória (*Random Forest*), *XGBoost* e Redes Neurais *Feedforward* (FNN). Cada um desses algoritmos possui características distintas que os tornam particularmente adequados ao problema em questão, como a capacidade de lidar com a complexidade dos dados e captar padrões comportamentais dos estudantes.

A variável alvo (y) utilizada para a construção dos modelos foi definida como um indicador binário de evasão escolar, onde 1 representa estudantes com maior probabilidade de evasão de evasão e 0 representa estudantes que possuem menor probabilidade de evasão. Essa definição foi baseada no desempenho acadêmico dos estudantes, utilizando a coluna total_do_curso do *dataset*, que categorizou os estudantes em risco se o valor fosse inferior a 6, indicando uma reprovação, ou seja, aumentando suas probabilidades de evasão.

As variáveis independentes (X) selecionadas para prever a evasão escolar incluíram uma série de variáveis comportamentais e acadêmicas dos estudantes: pct_presenca (porcentagem de presença), total_presenca (total de presenças), ausencias (número de ausências), soma_interacoes (soma das interações no AVA), dias_sem_interacoes (número de dias sem interação no AVA), maior_sequencia_dias_sem_interacoes (maior sequência de dias consecutivos sem interação) e notas_parciais_ausentes (quantidade de notas parciais ausentes). Essas variáveis foram escolhidas por serem indicativas do nível de engajamento e

desempenho acadêmico dos estudantes, fatores considerados importantes na previsão.

Em seguida, o conjunto de dados previamente preparado foi dividido em dois subconjuntos: 60% para treinamento e 40% para teste. Essa divisão permitiu que os modelos fossem ajustados com uma parte dos dados e, em seguida, avaliados em um conjunto de dados independente, garantindo uma avaliação imparcial da capacidade de generalização dos modelos. Cada modelo foi treinado utilizando as variáveis mencionadas, e os algoritmos foram implementados utilizando bibliotecas amplamente reconhecidas, como *scikit-learn* e *XGBoost*, passando por ajustes de hiper parâmetros para maximizar seu desempenho.

A avaliação dos modelos foi conduzida utilizando métricas como acurácia, ROC AUC, precisão, recall e f1-score, além da matriz de confusão para análise detalhada dos acertos e erros dos modelos. A acurácia mede a proporção de previsões corretas em relação ao total de previsões realizadas, sendo um indicativo global da eficácia do modelo. Uma alta acurácia sugere que o modelo está eficaz em classificar corretamente tanto os casos de evasão quanto os de não evasão.

A ROC AUC (*Receiver Operating Characteristic - Area Under Curve*) representa a capacidade do modelo em distinguir entre as classes positivas (evasão) e negativas (não evasão). Quanto maior o valor da AUC, melhor é o desempenho do modelo em discriminar entre os estudantes em risco e os que não estão. A precisão indica a proporção de previsões positivas que são corretas, refletindo a porcentagem de estudantes identificados como em risco de evasão que de fato estavam em risco. Uma alta precisão implica em menos falsos positivos.

O recall, ou sensibilidade, mede a proporção de casos positivos corretamente identificados pelo modelo, sendo essencial em situações em que é crucial minimizar falsos negativos.

O f1-score é a média harmônica entre a precisão e o recall, sendo especialmente útil em situações em que há um desbalanceamento entre as classes, fornecendo uma medida equilibrada da performance do modelo.

Por fim, a matriz de confusão apresenta a performance do modelo em termos de acertos e erros para cada classe (evasão e não evasão), permitindo uma análise detalhada da quantidade de verdadeiros positivos, verdadeiros negativos, falsos

positivos e falsos negativos, fornecendo uma compreensão mais aprofundada sobre as limitações e pontos fortes de cada abordagem.

As métricas obtidas permitiram a comparação dos modelos e a análise da eficácia de cada um no contexto da previsão da evasão escolar. A análise dos resultados indicou que o modelo de Floresta Aleatória apresentou a melhor acurácia entre os três algoritmos, enquanto o *XGBoost* e a FNN também apresentaram desempenhos competitivos, destacando-se em métricas específicas, como a área sob a curva ROC.

4.5 FERRAMENTAS UTILIZADAS

Para o desenvolvimento deste estudo, foram empregadas ferramentas e bibliotecas foram utilizadas a fim de assegurar o adequado processamento, análise e modelagem dos dados. A seguir, descrevem-se as principais ferramentas empregadas:

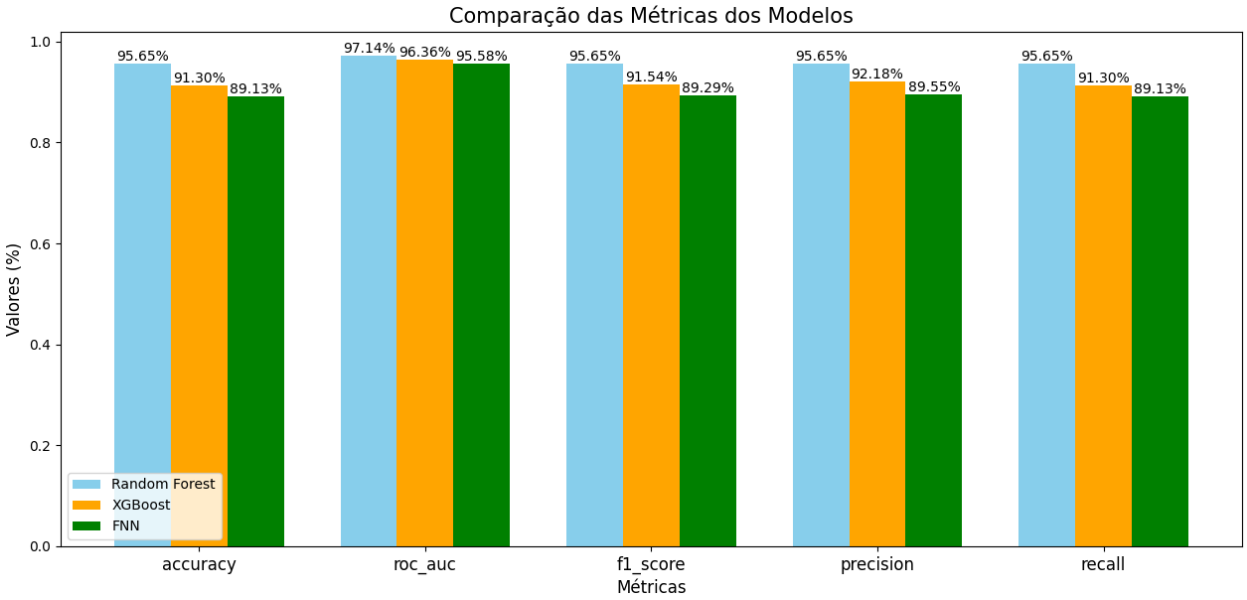
- Python: A linguagem de programação principal utilizada foi o Python, devido à sua ampla aplicabilidade em tarefas de ciência de dados e aprendizado de máquina. Python oferece flexibilidade e uma vasta coleção de bibliotecas para processamento de dados e modelagem.
- Pandas: Utilizada para a manipulação e análise de dados. Essa biblioteca foi fundamental para o carregamento, limpeza e organização dos dados em um formato tabular, facilitando as etapas subsequentes de pré-processamento e análise.
- *NumPy*: Foi utilizada para operações numéricas e manipulação de arrays. *NumPy* é essencial para cálculos matemáticos eficientes e para o suporte ao Pandas e outras bibliotecas de machine learning.
- *Scikit-Learn:* Biblioteca fundamental para a aplicação dos algoritmos de aprendizado de máquina, bem como para a divisão dos dados em conjuntos de treinamento e teste e para a avaliação dos modelos. *Scikit-Learn* foi utilizada para implementar algoritmos como *Random Forest* e Redes Neurais, além de oferecer ferramentas para a normalização dos dados e a avaliação dos modelos.

- *XGBoost*: Um framework avançado de aprendizado de máquina que foi utilizado para a construção de um dos modelos preditivos. O *XGBoost* é conhecido por sua eficiência e alta performance em tarefas de classificação e regressão.
- *Imbalanced-learn* (SMOTE): Utilizado para lidar com o desbalanceamento das classes, através da técnica SMOTE (*Synthetic Minority Over-sampling Technique*). O desbalanceamento dos dados era uma preocupação relevante, visto que a quantidade de alunos que evadem era menor do que a dos que não evadem, e o SMOTE ajudou a balancear o conjunto de dados, melhorando a performance dos modelos.
- Matplotlib: Foi utilizada para a criação de gráficos e visualizações. Esta ferramenta foi essencial para representar os resultados das análises de forma visual, facilitando a interpretação dos padrões encontrados e o desempenho dos modelos preditivos.
- SHAP: Utilizada para análise de interpretabilidade dos modelos. Através do SHAP, foi possível compreender quais variáveis contribuíram mais para as previsões realizadas pelos modelos, oferecendo insights valiosos sobre os fatores de risco de evasão escolar.

5 RESULTADOS E DISCUSSÕES

Após o pré-processamento dos dados, foram desenvolvidos modelos preditivos com o objetivo de prever a probabilidade da evasão escolar em turmas da graduação de uma disciplina presencial da UFSC. A metodologia empregada foi estruturada de forma contínua, abrangendo a escolha dos algoritmos, a divisão dos dados, o treinamento, a avaliação e a interpretação dos resultados.

Figura 9 - Comparação das métricas dos modelos



Fonte: elaborado pelo autor

A Figura 9 apresenta a comparação das métricas de desempenho dos três modelos estudados: *Random Forest*, *XGBoost* e *FNN*. As métricas analisadas incluem acurácia, área sob a curva ROC (AUC), F1-score, precisão e recall. O modelo *Random Forest* apresentou uma acurácia de 95.65%, destacando-se como o mais eficaz em termos gerais, enquanto o *XGBoost* obteve uma acurácia de 91,30%, e o FNN alcançou 89.13%. Sendo assim, o modelo *Random Forest* também apresentou os melhores valores para as demais métricas, indicando um excelente equilíbrio entre a predição de evasão e permanência dos estudantes.

A métrica ROC AUC, também é ligeiramente superior no *Random Forest*, com um valor de 97,14%. O *XGBoost* e a FNN apresentaram valores de 96,36% e 95,58%,

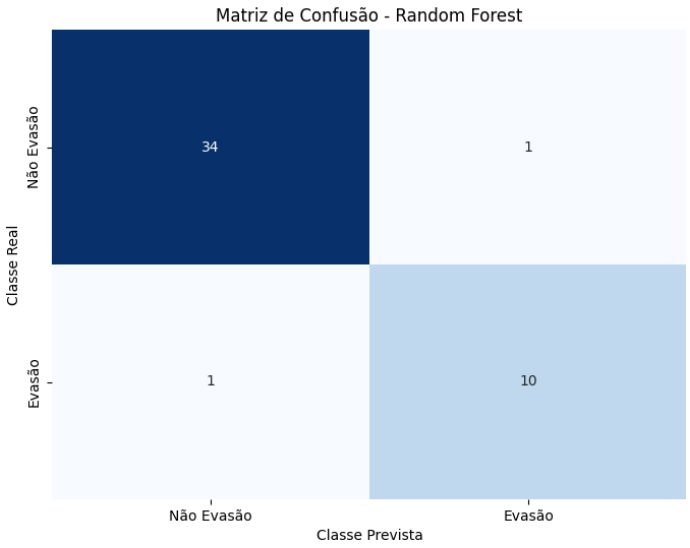
respectivamente, demonstrando que todos os modelos possuem uma boa capacidade de discriminação, embora o *Random Forest* apresente uma vantagem marginal.

O f1-score foi maior para o *Random Forest* (95,65%), enquanto o *XGBoost* e a FNN apresentaram valores de 91,54% e 89,29%, respectivamente, indicando que, embora sejam modelos competentes, não alcançam o mesmo equilíbrio entre minimizar falsos positivos e falsos negativos que o *Random Forest*.

A precisão foi mais alta no *Random Forest* (95,65%), seguida pelo *XGBoost* (92,18%) e pela FNN (89,55%). Isso indica que o *Random Forest* apresenta uma menor taxa de falsos positivos, sendo mais eficaz na correta identificação dos estudantes em risco real de evasão.

O recall também foi superior no *Random Forest* (95,65%), em comparação ao *XGBoost* (91,30%) e à FNN (89,13%). Este resultado sugere que o *Random Forest* possui um melhor desempenho em minimizar os falsos negativos, capturando uma maior proporção de estudantes que realmente estão em risco de evasão.

Figura 10 - Matriz de confusão *Random Forest*

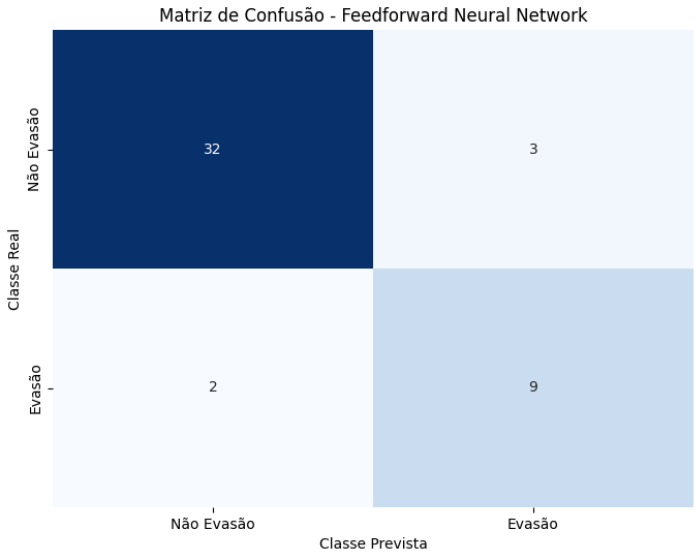


Fonte: elaborado pelo autor

Além da comparação das métricas gerais, as matrizes de confusão foram analisadas para cada modelo, fornecendo uma visão mais detalhada dos acertos e erros de cada abordagem. Na Figura 10, que ilustra a matriz de confusão do *Random Forest*, observa-se que o modelo realizou 34 previsões corretas para estudantes que não evadiram e 10 para aqueles que evadiram, apresentando apenas dois erros. A

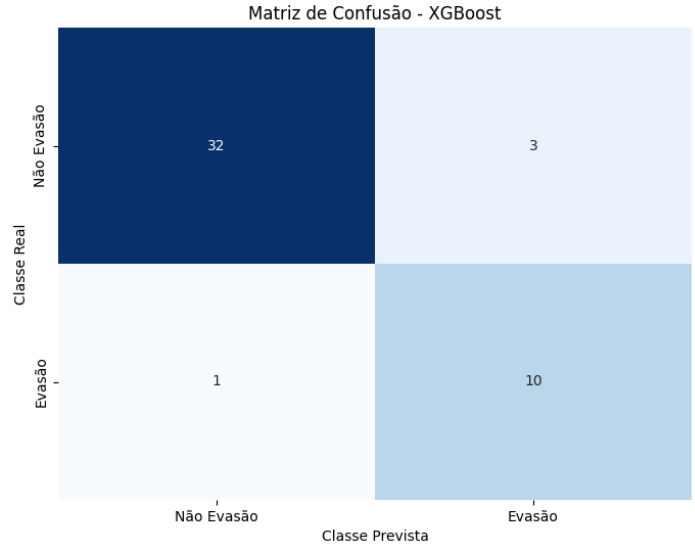
Figura 12, que representa a matriz de confusão do *XGBoost*, mostra um total de quatro erros, sendo três falsos positivos e um falso negativo. Já a Figura 11, que exibe a matriz de confusão do FNN, mostra que o modelo apresentou cinco erros, sendo três falsos positivos e dois falsos negativos.

Figura 11 - Matriz de confusão FNN



Fonte: elaborado pelo autor

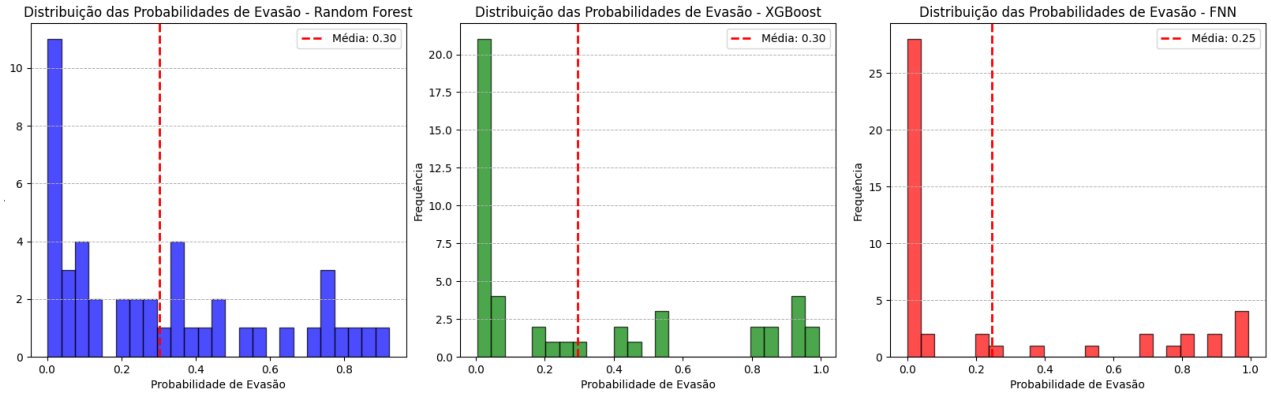
Figura 12 - Matriz de confusão *XGBoost*



Fonte: elaborado pelo autor

A análise da distribuição das probabilidades de evasão previstas pelos modelos também foi realizada, conforme ilustrado na Figura 13. Essa análise possibilitou uma compreensão aprofundada sobre a tendência dos modelos em atribuir diferentes níveis de risco aos estudantes, contribuindo para identificar padrões de previsão e potenciais limitações dos algoritmos aplicados.

Figura 13 - Histograma de distribuição das probabilidades de evasão



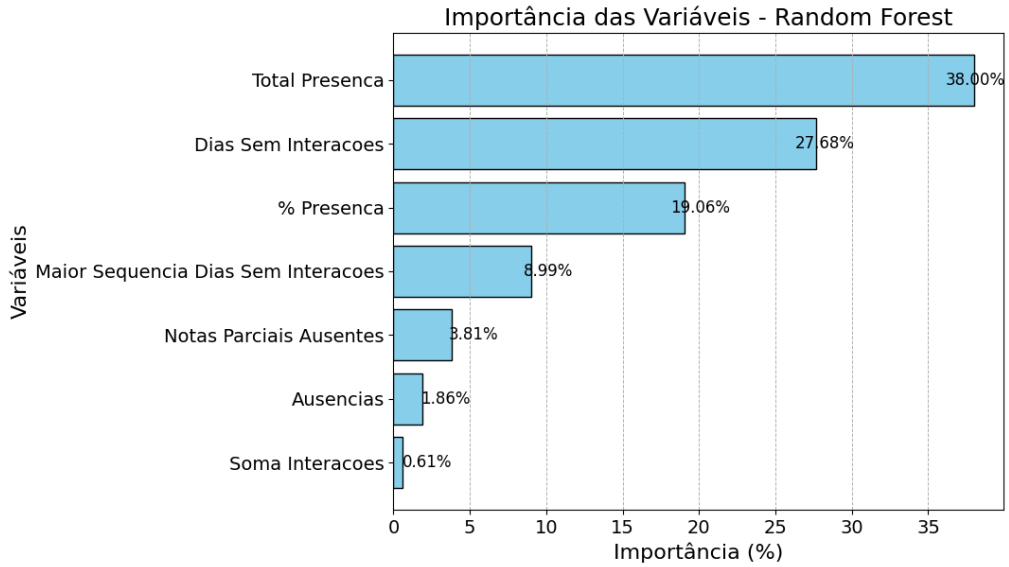
Fonte: elaborado pelo autor

No histograma da RF observa-se uma predominância de valores próximos a zero, indicando que a maioria dos estudantes foi classificada com baixa probabilidade de evasão. A média de probabilidade foi de 0,30, sugerindo uma tendência do modelo em se manter conservador em suas predições, atribuindo menores chances de evasão na maioria dos casos. Isso pode indicar que o modelo está fortemente orientado para reduzir os falsos positivos, ou seja, evita superestimar o risco de evasão, o que pode ser benéfico em termos de intervenções educacionais mais direcionadas.

Por outro lado, o modelo *XGBoost* também apresentou uma distribuição com concentração de probabilidades próximas a zero, mas exibiu uma maior dispersão em valores médios e altos, quando comparado à *Random Forest*. A média de probabilidade de evasão foi igualmente de 0,30. A presença de um número significativo de estudantes com valores próximos de 0,8 e até superiores, sugere que o *XGBoost* pode estar mais propenso a explorar diferentes padrões de risco, sendo mais agressivo em identificar potenciais evadidos. Essa característica pode ser útil quando se pretende identificar uma gama mais ampla de estudantes que possam

estar sob risco, mesmo que em diferentes graus. Por fim, o modelo FNN (*Network Neural Feedforward*) apresentou uma média de 0,25, com uma distribuição concentrada principalmente em valores muito baixos e alguns picos próximos a 1, indicando que o FNN tende a categorizar os estudantes em extremos ou como de baixo risco, ou como de alto risco, com pouca graduação intermediária.

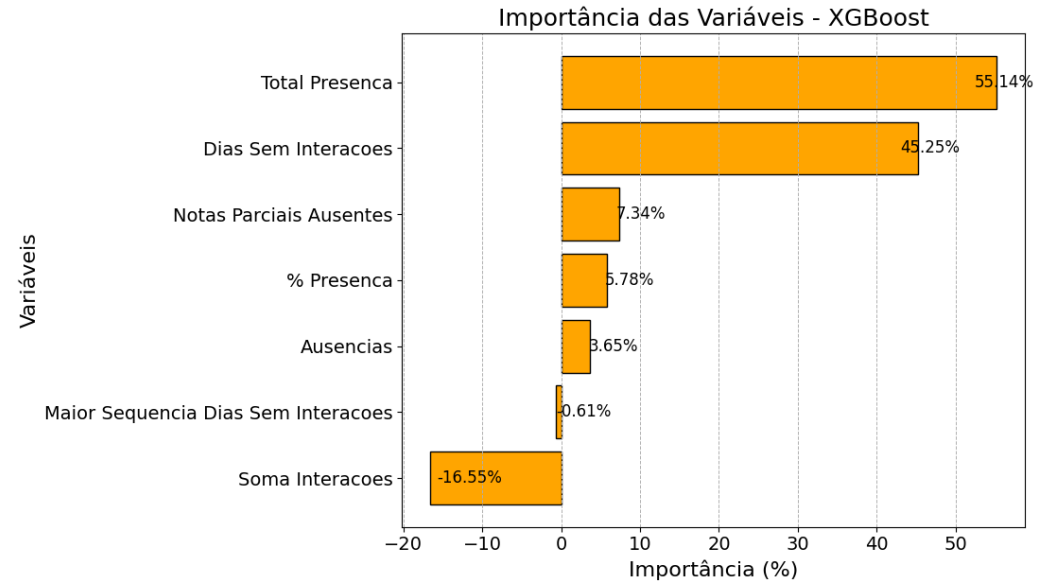
Figura 14 - Importância das variáveis pelo *Random Forest*



Fonte: elaborado pelo autor

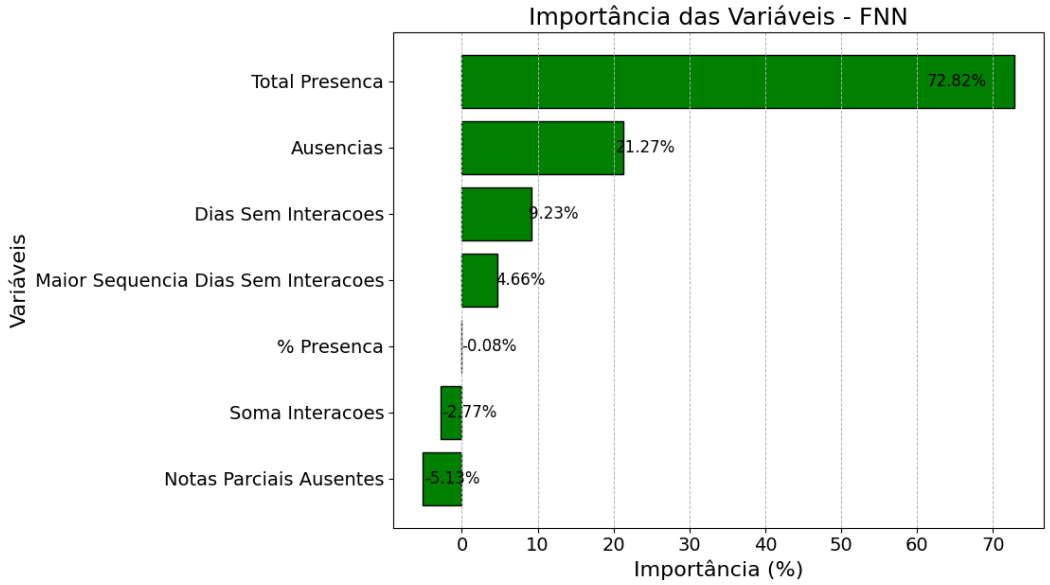
Para compreender melhor os fatores que mais influenciam as previsões dos modelos, foi realizada uma análise da importância das variáveis utilizadas. A Figura 14 apresenta a importância das variáveis no modelo *Random Forest*, destacando que "Total Presença" foi a variável mais relevante, com 38% de contribuição para o modelo, seguida por "Dias Sem Interações" (27,68%) e "% Presença" (19,06%). A Figura 15 ilustra a importância das variáveis no *XGBoost*, evidenciando "Total Presença" como a variável mais importante (55,14%), seguida por "Dias Sem Interações" (45,25%). Por fim, a Figura 16 mostra a importância das variáveis no FNN, com “Total Presença” sendo a variável de maior peso (72,82%), seguida por “Ausências” (21,27%).

Figura 15 – Importância das variáveis pelo *XGBoost*



Fonte: elaborado pelo autor

Figura 16 – Importância das variáveis pelo FNN

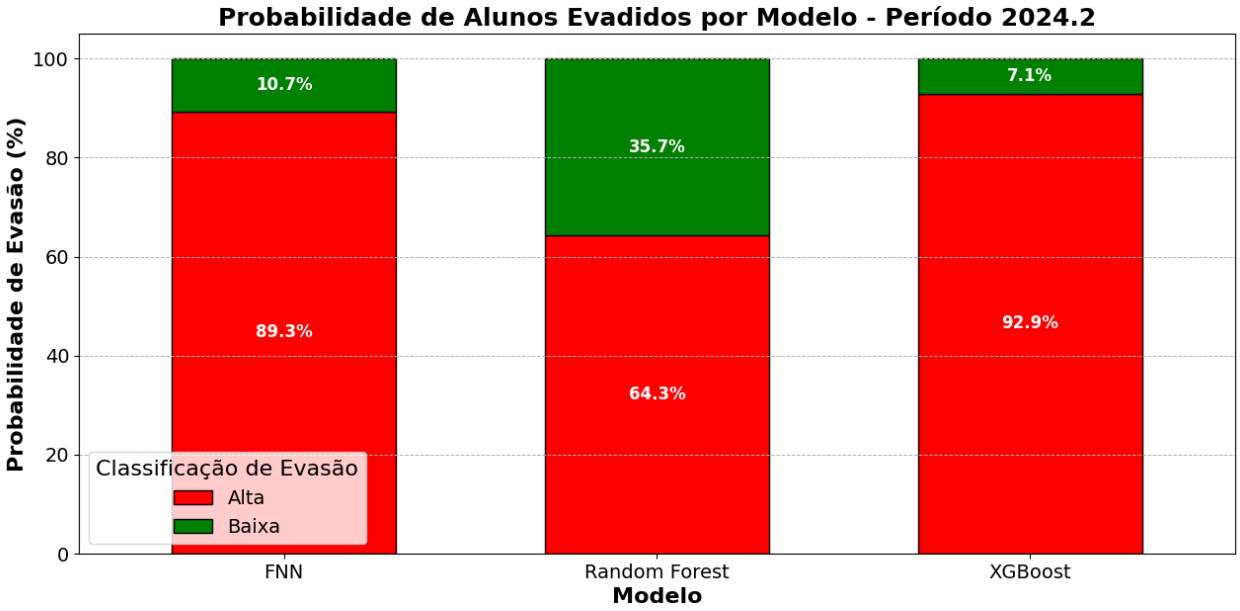


Fonte: elaborado pelo autor

As variáveis com importância negativa indicam que, no contexto específico de alguns modelos, tais variáveis podem não contribuir positivamente para a previsão ou podem sugerir relações indiretas e complexas. Essas relações devem ser interpretadas com cautela, uma vez que a presença de valores negativos pode indicar sobreajuste ou a necessidade de um melhor balanceamento de variáveis. A inclusão

de mais dados ou a utilização de métodos de regularização pode ser uma abordagem promissora para lidar com essas questões em estudos futuros.

Figura 17 - Probabilidade de alunos evadidos por modelo em 2024.2

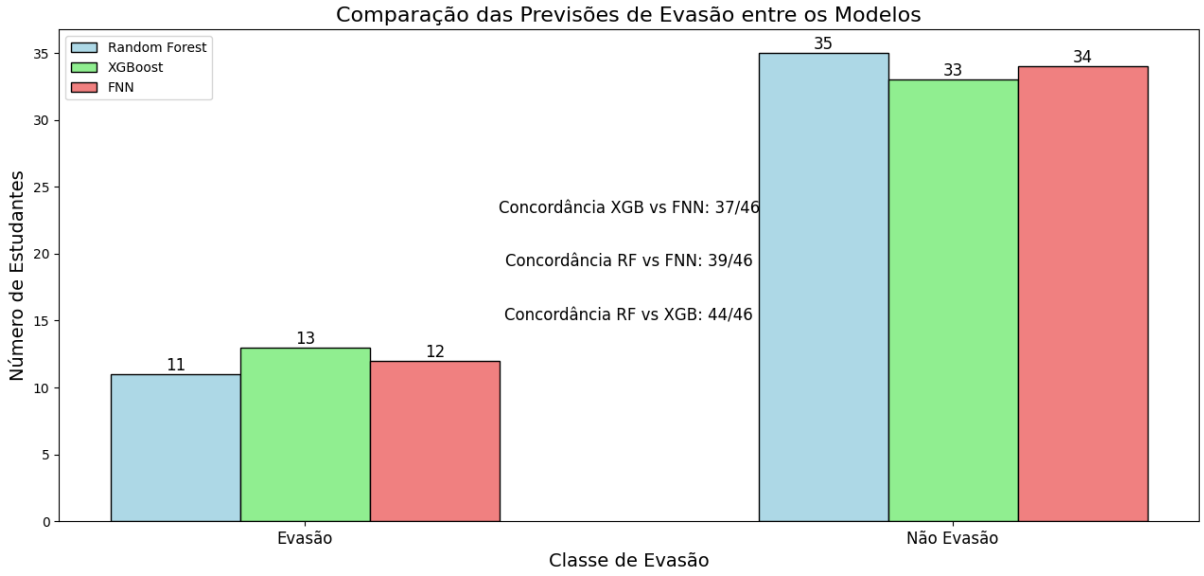


Fonte: elaborado pelo autor

A Figura 17 apresenta a probabilidade de evasão dos alunos por modelo, no período de 2024.2, classificando os alunos como sendo de alta ou baixa probabilidade de evasão. Observa-se que o modelo FNN identificou 78,6% dos alunos como tendo alta probabilidade de evasão, enquanto 21,4% foram classificados como baixa probabilidade. O modelo *Random Forest* apresentou uma distribuição mais equilibrada, com 42,9% dos alunos classificados como alta probabilidade de evasão e 57,1% como baixa probabilidade. Já o *XGBoost* identificou 67,9% dos alunos como alta probabilidade de evasão e 32,1% como baixa probabilidade. Esses resultados indicam que o modelo *Random Forest* foi o mais conservador em termos de classificação de risco, sendo o que apresentou maior proporção de estudantes com baixa probabilidade de evasão, enquanto o FNN foi o mais agressivo, classificando a maior parte dos estudantes como em alto risco. Esses dados informam sobre como diferentes modelos podem impactar a estratégia de mitigação da evasão, sendo crucial compreender o perfil de classificação de cada modelo para melhor adaptação das intervenções pedagógicas e institucionais.

Ademais, é importante mencionar que os alunos incluídos na análise ainda não haviam finalizado o semestre, de modo que suas médias finais eram compostas apenas pelas notas parciais enviadas até o momento. Essa característica dos dados pode ter contribuído para os resultados observados, uma vez que o desempenho parcial pode não refletir com exatidão o desempenho final dos estudantes, introduzindo um grau de incerteza nos modelos de predição. A variável alvo definida dessa forma pode ter influenciado na maior taxa de previsão de risco de evasão em determinados modelos, devido à falta de uma métrica consolidada do desempenho acadêmico ao longo do semestre completo.

Figura 18 - Comparação das previsões de evasão entre os modelos



Fonte: elaborado pelo autor

O gráfico de comparação das previsões de evasão entre os modelos referente a Figura 18, apresenta uma análise das predições realizadas pelos três modelos em relação à classe de evasão (ou não evasão). Observa-se que o modelo *Random Forest* prevê 11 alunos na classe de evasão, enquanto o *XGBoost* prediz 13 e a FNN, 12. Já para a classe de não evasão, os modelos *Random Forest*, *XGBoost* e FNN prevêm 35, 33 e 34 alunos, respectivamente.

Ao relacionar esses resultados com as análises de desempenho das métricas apresentadas anteriormente, podemos observar uma consistência nos resultados do *Random Forest*, que já havia demonstrado a maior acurácia, precisão e recall. A maior

concordância entre o *Random Forest* e o *XGBoost*, com 44 previsões coincidentes em um total de 46 amostras, reforça a robustez observada nas métricas de desempenho desses modelos. A concordância entre o *Random Forest* e a FNN foi de 39/46, enquanto a concordância entre o *XGBoost* e a FNN foi de 37/46.

Esses resultados sugerem que o *Random Forest* e o *XGBoost* possuem uma maior similaridade em suas previsões, o que indica que esses modelos podem compartilhar padrões semelhantes no processo de classificação dos dados. A FNN, embora também tenha uma alta taxa de concordância com os outros dois modelos, se mostra ligeiramente menos consistente nas predições, o que pode ser reflexo de diferenças nos processos de aprendizagem interna e na sensibilidade às variáveis de entrada.

Em termos gerais, a comparação sugere que, embora todos os modelos apresentem um desempenho razoavelmente próximo, o *Random Forest* e o *XGBoost* se destacam pela concordância mais elevada, indicando uma maior estabilidade e possível maior confiabilidade em suas predições. Esses achados reforçam a conclusão de que o *Random Forest*, além de ser o mais eficiente em termos de métricas de avaliação, também se mostra um dos mais consistentes na previsão de evasão escolar, sendo acompanhado de perto pelo *XGBoost*.

6 CONSIDERAÇÕES FINAIS

Este estudo apresentou uma análise dos modelos de predição de evasão escolar aplicados aos dados educacionais de estudantes de graduação, com o objetivo de identificar padrões comportamentais que possam informar estratégias eficazes de retenção acadêmica. A pesquisa percorreu um caminho metodológico, envolvendo a coleta e pré-processamento dos dados e a aplicação de técnicas de aprendizado de máquina, incluindo *Random Forest*, *XGBoost* e *Feedforward Neural Network* (FNN), para a construção e avaliação dos modelos preditivos. Foram utilizadas métricas de desempenho como acurácia, precisão, recall, F1-score e área sob a curva ROC, para mensurar a eficácia dos modelos. O modelo *Random Forest* apresentou uma acurácia de 96%, o *XGBoost* obteve uma acurácia de 91%, e o FNN apresentou uma acurácia de 89%.

O modelo *Random Forest* destacou-se com o melhor desempenho, especialmente em termos de equilíbrio entre precisão e sensibilidade, consolidando-se como a abordagem mais eficaz para a predição da evasão escolar. Embora o modelo *XGBoost* tenha apresentado resultados satisfatórios, observou-se uma tendência a gerar mais falsos positivos, o que pode ocasionar ações preventivas desnecessárias. O modelo FNN, por sua vez, revelou limitações significativas na identificação dos estudantes em risco de evasão, mostrando-se menos robusto em comparação aos demais.

A pesquisa desenvolveu e aplicou modelos preditivos para a evasão escolar no ensino superior, fornecendo informações quantitativas para o monitoramento dos estudantes e a identificação dos padrões de evasão. Através da construção e avaliação de três modelos distintos, uma análise detalhada de suas performances, e a identificação do modelo mais adequado ao contexto investigado.

Este estudo contribui para o campo da análise de dados educacionais, ao demonstrar a aplicabilidade de técnicas de aprendizado de máquina na previsão da evasão escolar e no apoio à tomada de decisões estratégicas por parte das instituições de ensino. As visualizações geradas a partir das análises das matrizes de confusão e das métricas de desempenho facilitaram a interpretação dos resultados, proporcionando uma compreensão dos pontos fortes e limitações de cada modelo.

Entre as limitações deste estudo, destaca-se o tamanho relativamente reduzido do conjunto de dados, bem como a utilização de apenas três modelos de aprendizado de máquina. Futuras pesquisas poderiam ampliar a abrangência dos dados e incorporar técnicas mais sofisticadas, como redes neurais profundas e modelos de aprendizado semi-supervisionado, visando aumentar a precisão e a robustez das previsões. Além disso, a integração de dados de diferentes fontes, tais como indicadores socioeconômicos e acadêmicos, poderia enriquecer a análise e proporcionar um entendimento mais abrangente dos fatores que influenciam a evasão.

Este trabalho focou na análise de evasão escolar utilizando notas finais como variável alvo principal. No entanto, considerando a complexidade do fenômeno da evasão, sugere-se que trabalhos futuros incluam outras variáveis como alvos. Assim, será possível verificar o potencial de capturar aspectos mais dinâmicos do comportamento dos estudantes, permitindo análises contínuas ao longo do semestre. Dessa forma, novos estudos poderiam explorar como a evolução dessas variáveis ao longo do período letivo contribui para identificar com maior antecedência os alunos em risco de evasão.

O uso de técnicas mais avançadas, como *embeddings* e modelos de representação vetorial, também se mostra promissor, pois pode fornecer uma compreensão mais detalhada dos comportamentos dos estudantes e das dinâmicas educacionais. Além disso, a utilização de grafos de conhecimento poderia permitir uma análise das interações entre os fatores que contribuem para a evasão, facilitando a criação de intervenções mais precisas e personalizadas.

REFERÊNCIAS

ALMEIDA, Mauricio B. **Noções básicas sobre metodologia de pesquisa científica**. Belo Horizonte: Universidade Federal de Minas Gerais, [s.d.].

AMBIEL, Rodolfo A. M.. Construção da Escala de Motivos para Evasão do Ensino Superior. **Aval. psicol.**, Itatiba, v. 14, n. 1, p. 41-52, abr. 2015. Disponível em http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1677-04712015000100006&lng=pt&nrm=iso. acessos em 3 dez. 2024.

ANDER-EGG, Ezequiel. **Introducción a las técnicas de investigación social: para trabajadores sociales**. 7. ed. Buenos Aires: Humanitas, 1978.

ASSOCIAÇÃO NACIONAL DOS DIRIGENTES DAS INSTITUIÇÕES FEDERAIS DE ENSINO SUPERIOR. **Diplomação, retenção e evasão nos cursos de graduação em instituições públicas: 1996**. Brasília: ANDIFES, 1997. Disponível em: https://www.andifes.org.br/wp-content/files_flutter/Diplomacao_Retencao_Evasao_Graduacao_em_IES_Publicas-1996.pdf. Acesso em: 20 nov. 2024.

AMAZON WEB SERVICES. **O que é machine learning?**. Disponível em: https://aws.amazon.com/pt/what-is/machine-learning/. Acesso em: 6 set. 2024.

BAGGI, Cristiane Aparecida dos Santos; LOPES, Doraci Alves. Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. **Avaliação: Revista da Avaliação da Educação Superior (Campinas)**, Campinas, v. 16, n. 2, p. 355-374, jul. 2011. FapUNIFESP (SciELO).

BAKER, Ryan; ISOTANI, Seiji; CARVALHO, Adriana. Mineração de Dados Educacionais: oportunidades para o brasil. **Revista Brasileira de Informática na Educação**, Porto Alegre, v. 19, n. 02, p. 3-13, 31 ago. 2011. Sociedade Brasileira de Computação - SB. http://dx.doi.org/10.5753/rbie.2011.19.02.03.

BARROSO, Paula Cristina Freitas; OLIVEIRA, Íris Martins; NORONHA-SOUSA, Dulce; NORONHA, Ana; MATEUS, Cristina Cruz; VÁZQUEZ-JUSTO, Enrique; COSTA-LOBO, Cristina. FATORES DE EVASÃO NO ENSINO SUPERIOR: uma revisão de literatura. **Psicologia Escolar e Educacional**, [S.L.], v. 26, abr. 2022. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/2175-35392022228736.

BARROZO, Caio Blumer. **Módulo preditivo para classificação de estudantes em risco de reprovação**. 2022. 50 f. Trabalho de Conclusão de Curso (Graduação em Engenharia de Computação) – Universidade Federal de Santa Catarina, Araranguá, 2022.

BITTENCOURT, Ibsen Mateus; MERCADO, Luis Paulo Leopoldo. Evasão nos cursos na modalidade de educação a distância: estudo de caso do curso piloto de administração da ufal/uab. **Ensaio**: **Avaliação e Políticas Públicas em Educação**, Rio de Janeiro, v. 22, n. 83, p. 465-504, jun. 2014. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s0104-40362014000200009.

CHICON, Patricia Mariotto Mozzaquatro; PASCHOAL, Leo Natan; FRANTZ, Fabricia Carneiro Roos. Indicadores de Evasão em Ambientes Virtuais de Aprendizagem no

contexto da Educação a Distância: um mapeamento sistemático. **Renote**, Porto Alegre, v. 18, n. 2, p. 111-120, 4 jan. 2021. Universidade Federal do Rio Grande do Sul. http://dx.doi.org/10.22456/1679-1916.110209. Disponível em: https://seer.ufrgs.br/renote/article/view/110209/0. Acesso em: 20 nov. 2024.

COIMBRA, Camila Lima; SILVA, Leonardo Barbosa e; COSTA, Natália Cristina Dreossi. A evasão na educação superior: definições e trajetórias. **Educação e Pesquisa**, São Paulo, v. 47, n. 0, p. 0-0, abr. 2021. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s1678-4634202147228764. Disponível em: https://www.scielo.br/j/ep/a/WRKk9JVNBnJJsnNyNkFfJQj/. Acesso em: 20 nov. 2024.

COLPANI, R. Mineração de Dados Educacionais: um estudo da evasão no ensino médio com base nos indicadores do Censo Escolar. **Informática na educação: teoria & prática,** Porto Alegre, v. 21, n. 3, 2018. DOI: 10.22456/1982-1654.87880. Disponível em: https://seer.ufrgs.br/index.php/InfEducTeoriaPratica/article/view/87880. Acesso em: 20 dez. 2024.

COSTA, Evandro; BAKER, Ryan S.J.D.; AMORIM, Lucas; MAGALHÃES, Jonathas; MARINHO, Tarsis. Mineração de Dados Educacionais: conceitos, técnicas, ferramentas e aplicações. **Jornada de Atualização em Informática na Educação - Jaie.** [S.l], p. 1-29. 2012. Disponível em: http://milanesa.ime.usp.br/rbie/index.php/pie/article/view/2341. Acesso em: 20 nov. 2024.

DIAS, Cláudia; FERNANDES, Denise. **Pesquisa e método científicos**. Brasília, mar. 2000.

DIAS, Ellen Christine Moraes; THEÓPHILO, Carlos R.; LOPES, Maria A. S. Evasão no ensino superior: estudo dos fatores causadores da evasão no curso de Ciências Contábeis da Universidade Estadual de Montes Claros – Unimontes – MG. In: CONGRESSO USP DE INICIAÇÃO CIENTÍFICA EM CONTABILIDADE, 7., 2010, São Paulo. **Anais**. São Paulo: USP, 2010. v. 7. Disponível em: https://congressousp.fipecafi.org/anais/artigos32006/370.pdf. Acesso em: 20 nov. 2024.

DIGIAMPIETRI, Luciano Antonio; NAKANO, Fabio; LAURETTO, Marcelo de Souza. Mineração de Dados para Identificação de Alunos com Alto Risco de Evasão: um estudo de caso. **Revista de Graduação Usp**, São Paulo, v. 1, n. 1, p. 17-8, 18 jul. 2016. Universidade de Sao Paulo, Agência USP de Gestao da Informacao Academica (AGUIA). http://dx.doi.org/10.11606/issn.2525-376x.v1i1p17-23. Disponível em: https://www.revistas.usp.br/gradmais/article/view/117720. Acesso em: 20 nov. 2024.

DIOGO, Maria Fernanda; RAYMUNDO, Luana dos Santos; WILHELM, Fernanda Ax Wilhelm; ANDRADE, Sílvia Patricia Cavalheiro de; LORENZO, Flora Moura; ROST, Flávia Trento; BARDAGI, Marúcia Patta. Percepções de coordenadores de curso superior sobre evasão, reprovações e estratégias preventivas. **Avaliação: Revista da Avaliação da Educação Superior**, Campinas; Sorocaba, SP, v. 21, n. 1, 2016. Disponível em: https://periodicos.uniso.br/avaliacao/article/view/2513. Acesso em: 20 dez. 2024.

ESCOVEDO, Tatiana; KOSHIYAMA, Adriano. **Introdução a data science: algoritmos de machine learning e métodos de análise**. São Paulo: Casa do Código, 2020.

ESTEVES, H. R. C.; DIAS, C. A..; SANTOS, C. M.; HIGUCHI, A. K.. School dropout in Higher Education: a literature review in the years 2014 to 2020. **Research, Society and Development**, *[S. l.]*, v. 10, n. 3, p. e21310313210, 2021. DOI: 10.33448/rsd-v10i3.13210. Disponível em: https://rsdjournal.org/index.php/rsd/article/view/13210. Acesso em: 20 dec. 2024.

FARIA, Mauricio Mendes; MONTEIRO, Ana María. Investigação sobre Técnicas de Detecção de Intrusões em Redes de Computadores com base nos Algoritmos Knn e K-Means. **XI Wcf - Workshop de Computação da Faccamp.** Campo Limpo Paulista, p. 1-5. set. 2015.

FIALHO, Marillia Gabriella Duarte. **A evasão escolar e a gestão universitária: o caso da Universidade Federal da Paraíba.** 2014. 107 f. Dissertação (Mestrado em Gestão de Organizações Aprendentes) - Universidade Federal da Paraí-ba, João Pessoa, 2014.

FRITSCH, R.; ROCHA, C. S. da; VITELLI, R. F. A evasão nos cursos de graduação em uma instituição de ensino superior privada. **Revista Educação em Questão**, [S. l.], v. 52, n. 38, p. 81–108, 2015. DOI: 10.21680/1981-1802.2015v52n38ID7963. Disponível em: https://periodicos.ufrn.br/educacaoemquestao/article/view/7963. Acesso em: 20 dez. 2024.

FURTADO, Maria Inês Vasconcellos. **Redes Neurais Artificiais: Uma Abordagem Para Sala de Aula**. Ponta Grossa: Atena, 2019. 105 p. Disponível em: https://atenaeditora.com.br/catalogo/post/redes-neurais-artificiais-uma-abordagem-para-sala-de-aula. Acesso em: 24 nov. 2024.

GIL, A. C. **Como elaborar projetos de pesquisa**. São Paulo: Atlas, 2008.

Fernandes; TEIXEIRA, Rosenália Ramalho. Evasão escolar no ensino superior público: uma análise de conteúdo nos resumos de artigos que estudaram o tema. **Competência**, Porto Alegre, v. 16, n. 2, dez. 2023.

IBM. **Machine Learning**. Disponível em: https://www.ibm.com/br-pt/topics/machine-learning. Acesso em: 21 nov. 2024.

KAMPFF, A. J. C.; REATEGUI, E. B.; LIMA, J. V. de. Mineração de dados educacionais para a construção de alertas em ambientes virtuais de aprendizagem como apoio à prática docente. **RENOTE**, Porto Alegre, v. 6, n. 1, 2008. DOI: 10.22456/1679-1916.14394. Disponível em: https://seer.ufrgs.br/index.php/renote/article/view/14394. Acesso em: 20 dez. 2024.

KIPNIS, Bernado. A pesquisa institucional e a educação superior brasileira: um estudo de caso longitudinal da evasão. **Linhas Crí-ticas**, [S. l.], v. 6, n. 11, p. 109–130, 2000. DOI: 10.26512/lc.v6i11.2870. Disponível em: https://periodicos.unb.br/index.php/linhascriticas/article/view/2870. Acesso em: 20 dez. 2024.

LEMOS, Ítalo Vinícius do Rego. **Prevendo a evasão escolar em uma instituição de ensino técnico utilizando mineração de dados educacionais**. 2021. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Universidade Federal Rural de Pernambuco, Recife, 2021.

LINDEN, R. Técnicas de agrupamento. **Revista de Sistema da Informação da FSMA**, n. 4, p. 18-36, 2009.

LUDERMIR, Teresa Bernarda. Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências. **Estudos Avançados**, [S.L.], v. 35, n. 101, p. 85-94, abr. 2021. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s0103-4014.2021.35101.007.

MAGNAGNAGNO, Cleber Cicero; RAMOS, Monica Parente; OLIVEIRA, Lucila Maria Pesce de. Estudo sobre o Uso do Moodle em Cursos de Especialização a Distância da Unifesp. **Revista Brasileira de Educação Médica**, [S.L.], v. 39, n. 4, p. 507-516, dez. 2015. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/1981-52712015v39n4e00842014.

MAIESKI, Alessandra; MOROSOV ALONSO, Kátia. Educação a Distância e o uso dos Ambientes Virtuais de Aprendizagem: entre o Ideal e o Possível. **Revista Diálogo Educacional**, [S. l.], v. 21, n. 70, 2021. Pontificia Universidade Catolica do Parana - PUCPR. http://dx.doi.org/10.7213/1981-416x.21.070.ao08.

MALERBA, Adriano. **Previsão de evasão universitária com aprendizado de máquina**. 2023. Dissertação (Mestrado em Engenharia de Produção) – Universidade Federal de Itajubá, Itajubá, 2023.

MARTINS, Ronaldo Rodrigues; BERTUCI, Matheus Henrique; PENIANI, Lucas Polimeno. MINERAÇÃO DE DADOS NO COMBATE À EVASÃO ESCOLAR EM INSTITUIÇÕES DE ENSINO SUPERIOR BRASILEIRAS. **Revista Interface Tecnológica**, [S.L.], v. 17, n. 2, p. 103-115, 18 dez. 2020. Interface Tecnologica. http://dx.doi.org/10.31510/infa.v17i2.885.

MENEZES, Renata de; SCOTTI, Luciana; SCOTTI, Marcus. APRENDIZADO DE MÁQUINA APLICADO A QSAR. **Química Nova**, [S.L.], v. 47, n. 7, p. 1-16, mar. 2024. Sociedade Brasileira de Quimica (SBQ). http://dx.doi.org/10.21577/0100-4042.20240024.

MOREIRA, Fábio Junior Rodrigues. **Aprendizagem de máquina na predição da evasão no ensino superior**. 2020. Monografia (Especialização em Data Science & Big Data) – Universidade Federal do Paraná, Setor de Ciências Exatas, Curitiba, 2020.

OLIVEIRA, Adonias Caetano de. **Máquina de aprendizagem mínima com opção de rejeição**. 2016. 79 f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal do Ceará, Fortaleza, 2016.

OLIVEIRA, Isleimar de Souza. **ANÁLISE DE DADOS APLICADA À EVASÃO ESCOLAR: UM ESTUDO DE CASO DO IFPB**. 2023. 136 f. Dissertação (Mestrado) -

Curso de Tecnologia da Informação, Instituto Federal de Educação, Ciência e Tecnologia da Paraíba – Ifpb, João Pessoa, 2023.

OLIVEIRA, Ronei dos Santos. **Modelo de predição de evasão escolar com base em dados de autoavaliação de cursos de graduação**. 2023. 73 f. Dissertação (Mestrado) - Curso de Tecnologia da Informação, Instituto Federal de Educação, Ciência e Tecnologia da Paraíba - Ifpb, João Pessoa, 2023.

OSBORNE, Jason W.; JONES, Brett D.. Identification with Academics and Motivation to Achieve in School: how the structure of the self influences academic outcomes. **Educational Psychology Review**, [S.L.], v. 23, n. 1, p. 131-158, 16 fev. 2011. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s10648-011-9151-1.

PEFFERS, Ken; TUUNANEN, Tuure; ROTHENBERGER, Marcus A.; CHATTERJEE, Samir. A Design Science Research Methodology for Information Systems Research. **Journal Of Management Information Systems**, [S.L.], v. 24, n. 3, p. 45-77, dez. 2007. Informa UK Limited. http://dx.doi.org/10.2753/mis0742-1222240302.

PEREIRA, Adriana Soares; SHITSUKA, Dorlivete Moreira; PARREIRA, Fabio José; SHITSUKA, Ricardo. **Metodologia da pesquisa científica**. Santa Maria: Núcleo de Tecnologia Educacional da Universidade Federal de Santa Maria, 2018. Disponível em: https://repositorio.ufsm.br/handle/1/15824. Acesso em: 20 nov. 2024.

PIMENTEL, Edson P.; OMAR, Nizam. Descobrindo conhecimentos em dados de avaliação da aprendizagem com técnicas de mineração de dados. **Anais do XXVI Congresso da Sociedade Brasileira de Computação (SBC)**, Campo Grande, MS, 2006.

PORTO, B.; DIAS, D. M.; BATTESTIN, V. Tendências de Learning Analytics em Moodle: uma Revisão Sistemática. **EaD em Foco**, [S. l.], v. 13, n. 1, p. e2070, 2023. Fundacao CECIERJ. http://dx.doi.org/10.18264/eadf.v13i1.2070.

PRESTES, Emília Maria da Trindade; FIALHO, Maríllia Gabriella Duarte. Evasão na educação superior e gestão institucional: o caso da universidade federal da paraíba. **Ensaio: Avaliação e Políticas Públicas em Educação**, [S.L.], v. 26, n. 100, p. 869-889, jul. 2018. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s0104-40362018002601104.

RABELO, Humberto; BURLAMAQUI, Aquiles; VALENTIM, Ricardo; RABELO, Danieli Silva de Souza; MEDEIROS, Soraya. Utilização de técnicas de mineração de dados educacionais para predição de desempenho de alunos de EaD em ambientes virtuais de aprendizagem. **Simpósio Brasileiro de Informática na Educação**, [S.L.], v. 28, n. 1, p. 1527-1536, 27 out. 2017. Brazilian Computer Society (Sociedade Brasileira de Computação - SBC). http://dx.doi.org/10.5753/cbie.sbie.2017.1527.

RIGO, Sandro José; CAMBRUZZI, Wagner; BARBOSA, Jorge L. V.; CAZELLA, Sílvio C.. Aplicações de Mineração de Dados Educacionais e Learning Analytics com foco na evasão escolar: oportunidades e desafios. **Revista Brasileira de Informática na Educação**, [S.L.], v. 22, n. 01, p. 132, 18 maio 2014. Sociedade Brasileira de Computacao - SB. http://dx.doi.org/10.5753/rbie.2014.22.01.132.

RISTOFF, Dilvo Ilvo. **Avaliação institucional: pensando princípios**. In: BALZAN, N. C.; DIAS SOBRINHO, J. (orgs.). Avaliação institucional: teoria e experiências. São Paulo: Cortez, 1995. p. 37-51.

RISTOFF, Dilvo Ilvo. **Considerações sobre a evasão**. In: VASCONCELOS, Silvia Ines Coneglian Carrilho de (org.). Expressão sobre a graduação. Maringá: Universidade Estadual de Maringá, 1997. p. 9-32.

RODRIGUES, Rodrigo Lins; RAMOS, Jorge Luis Cavalcanti; SILVA, João Carlos Sedraz; GOMES, Alex Sandro. A literatura brasileira sobre mineração de dados educacionais. **Workshops do Congresso Brasileiro de Informática na Educação**, [S.L.], v. 1, p. 621, 3 nov. 2014. Sociedade Brasileira de Computação - SBC. http://dx.doi.org/10.5753/cbie.wcbie.2014.621.

ROMERO, Cristobal; VENTURA, Sebastian. Data mining in education. **Wires Data Mining And Knowledge Discovery**, [S.L.], v. 3, n. 1, p. 12-27, 14 dez. 2012. Wiley. http://dx.doi.org/10.1002/widm.1075.

SANTOS DA COSTA, S.; CAZELLA, S.; JOSÉ RIGO, S. Minerando dados sobre o desempenho de alunos de cursos de educação permanente em modalidade EAD: Um estudo de caso sobre evasão escolar na UNA-SUS. **RENOTE**, Porto Alegre, v. 12, n. 2, 2014.

SANTOS, Jorge de Sousa. Recursos multimídias para a educação: o ambiente virtual de aprendizagem na educação superior a distância. **REEDUC - Revista de Estudos em Educação**, v. 9, n. 1, 2023.

SANTOS, Sanval Ebert de Freitas; JORGE, Eduardo Manuel de Freitas; WINKLER, Ingrid. Inteligência artificial e virtualização em ambientes virtuais de ensino e aprendizagem. **Etd - Educação Temática Digital**, [S.L.], v. 23, n. 1, p. 2-19, 17 fev. 2021. Universidade Estadual de Campinas. http://dx.doi.org/10.20396/etd.v23i1.8656150.

SHINDE, Pramila P.; SHAH, Seema. A Review of Machine Learning and Deep Learning Applications. **2018 Fourth International Conference On Computing Communication Control And Automation (Iccubea)**, [S.L.], p. 1-6, ago. 2018. IEEE. http://dx.doi.org/10.1109/iccubea.2018.8697857

SILVA FILHO, Roberto Leal Lobo e; MOTEJUNAS, Paulo Roberto; HIPÓLITO, Oscar; LOBO, Maria Beatriz de Carvalho Melo. A evasão no ensino superior brasileiro. **Cadernos de Pesquisa**, [S.L.], v. 37, n. 132, p. 641-659, dez. 2007. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s0100-15742007000300007.

SILVA, Fernanda Cristina da. **Variáveis para modelos preditivos à evasão na educação superior**. 2021. 259 f. Tese (Doutorado em Administração) – Universidade Federal de Santa Catarina, Florianópolis, 2021.

SILVA, Fernanda Cristina da; CABRAL, Thiago Luiz de Oliveira; PACHECO, Andressa Sasaki Vasques. Evasão ou permanência? Modelos preditivos para a gestão do Ensino Superior. **Education Policy Analysis Archives**, [S.L.], v. 28, n.

149, 19 out. 2020. Mary Lou Fulton Teacher College. http://dx.doi.org/10.14507/epaa.28.5387.

SILVA, Siony da. Ambientes virtuais de aprendizagem e a educação a distância. **Dialogia**, [S.L.], v. 7, n. 2, p. 235-244, 12 nov. 2009. University Nove de Julho. http://dx.doi.org/10.5585/dialogia.v7i2.1285.

SOARES, E. M. S.; LUCIANO, N. A. Formação continuada de professores no contexto das tecnologias digitais. In: CONGRESSO INTERNACIONAL DE EDUCAÇÃO A DISTÂNCIA, 11., 2004 Salvador. **Anais**... Salvador: ABED, 2004.

SOUZA, Alex Marques de. **Machine learning e a evasão escolar - Análise preditiva no suporte à tomada de decisão**. 2020. 133 f. Dissertação (Mestrado) - Curso de Gestão de Sistemas de Informação e do Conhecimento, Universidade Fumec, Belo Horizonte, 2020.

TARTUCE, T. J. A. **Métodos de pesquisa**. Fortaleza: Unice – Ensino Superior, 2006. Apostila.

TEMPELAAR, Dirk. Supporting the less-adaptive student: the role of learning analytics, formative assessment and blended learning. **Assessment & Evaluation In Higher Education**, [S.L.], v. 45, n. 4, p. 579-593, 25 nov. 2019. Informa UK Limited. http://dx.doi.org/10.1080/02602938.2019.1677855.

THEODORSON, G. A. & THEODORSON, A. G. **A modern dictionary of sociology**. London, Methuen, 1970.

TINTO, Vincent. Dropout from Higher Education: a theoretical synthesis of recent research. **Review Of Educational Research**, [S.L.], v. 45, n. 1, p. 89, 1975. SAGE Publications. http://dx.doi.org/10.2307/1170024.

VELOSO, T. C. M. A.; ALMEIDA, E. P. de. Evasão nos cursos de graduação da Universidade Federal de Mato Grosso, campus universitário de Cuiabá – um processo de exclusão. **Série-Estudos - Periódico do Programa de Pós-Graduação em Educação da UCDB**, [S. l.], n. 13, 2013. Disponível em: https://www.serie-estudos.ucdb.br/serie-estudos/article/view/564. Acesso em: 20 dez. 2024.