Análise dos principais fatores que influenciam a evasão no ensino superior utilizando técnicas de mineração de dados educacionais

Ronaldo Celso Messias Correia[1], **Harrison Buziquia de Mendonça**[1], **Camila Tolin Santos Da Silva**[1], **Douglas Francisquini Toledo**[1]

[1]Universidade Estadual Paulista - Júlio de Mesquita Filho - UNESP
Presidente Prudente, Brasil

ronaldo.correia@unesp.br, harrison.mendonca@unesp.br,

camila.tolin@unesp.br, douglas.toledo@unesp.br

Abstract. *Dropout rates persist in Brazilian higher education, despite the increase in enrollments. This study proposes an innovative methodology to detect students at risk of dropping out at UNIVESP, using educational data mining and analysis. The technique involves pre-processing, feature selection and machine learning to identify evasion patterns. Anticipating these cases, preventive strategies and personalized support can promote student success. This methodology identifies the main dropout factors, providing insights for educational policies. Preliminary results show 92% accuracy in identifying students at risk, with less than 20% of the data characteristics, demonstrating its effectiveness.*

Resumo. *A evasão persiste no ensino superior brasileiro, apesar do aumento nas matrículas. Este estudo propõe uma metodologia inovadora para detectar alunos em risco de evasão na UNIVESP, utilizando mineração e análise de dados educacionais. A técnica envolve pré-processamento, seleção de características e aprendizado de máquina para identificar padrões de evasão. Antecipando esses casos, estratégias preventivas e apoio personalizado podem promover o sucesso dos estudantes. Esta metodologia identifica os principais fatores de evasão, fornecendo insights para políticas educacionais. Resultados preliminares mostram 92% de acurácia na identificação de alunos em risco, com menos de 20% das características dos dados, evidenciando sua eficácia.*

1. Introdução

O ensino superior à distância no Brasil tem tido um aumento notável, ampliando o acesso à educação superior através de plataformas online e programas de ensino remoto. Para entender o cenário desse tipo de ensino, é fundamental analisar os dados educacionais disponíveis, a fim de aprimorar o sistema e oferecer uma educação superior de qualidade. Estes dados desempenham um papel crucial na avaliação e melhoria do ensino superior à distância no país, fornecendo informações valiosas sobre o desempenho dos alunos, o progresso acadêmico e as áreas a serem aprimoradas. A mineração e a análise destes dados são fundamentais para adaptar os cursos, melhorar os métodos de ensino e avaliar a eficácia das estratégias pedagógicas.

Segundo relatório do Instituto Nacional de Estudos e Pesquisas Educacionais, o INEP, em 2022 foram oferecidas mais de 22,8 milhões de vagas em cursos de graduação, sendo 75,5% vagas novas e 24,4% vagas remanescentes. Do total das vagas em cursos de graduação em 2022, a rede privada ofertou 96,2% do total de vagas . A rede pública correspondeu a 3,8% das vagas ofertadas pelas Instituições de Ensino Superior (IES). Ainda segundo o relatório do INEP, o número de matrículas na modalidade a distância continua crescendo, tendo atingido mais de 4 milhões em 2022, o que já representa uma participação de 45,9% do total de matrículas de graduação. Já o número de concluintes em cursos de graduação presencial teve queda de 4,6% em relação a 2021 – situação semelhante para a modalidade à distância, com redução de 0,3% no mesmo período.

Tão importante quanto o aumento do ingresso de estudantes no ensino superior, é sua permanência. A taxa de evasão é um desafio significativo no contexto do ensino superior à distância no Brasil. Em seu trabalho, [Oliveira and Costa 2021] mostram a análise dos dados do período compreendido entre 2013 e 2017, em que pode-se notar um maior percentual de evasão nos cursos de ensino à distância (EAD) em relação aos presenciais, em todos os anos do período observado.

Este trabalho propõe uma metodologia para a mineração e análise de dados educacionais, assim como a predição de evasão de estudantes EAD da Universidade Virtual do Estado de São Paulo (UNIVESP), visando a descoberta de informações e desenvolvimento de estratégias que auxiliem os alunos a permanecerem na instituição de ensino. Para isso, desenvolvemos uma abordagem inovadora que integra e combina de maneira sinérgica diversas técnicas de mineração de dados educacionais. O objetivo é realizar uma análise abrangente e perspicaz dos principais fatores correlacionados à evasão no ensino superior. Nesse contexto, as contribuições deste estudo são:

- Elaboração de um modelo preditivo iterativo com utilização do SelecKBest para identificação das características mais relevantes na predição dos estudantes com risco de evasão;
- Observação dos fatores e características mais significativos ao analisar a evasão no ensino superior à distância;

O restante deste trabalho está organizado da seguinte forma: na Seção II são apresentados os trabalhos relacionados; na Seção III são apresentados os dados, e explanada a metodologia; a Seção IV descreve os resultados obtidos; a Seção V apresenta a análise e a reprodutibilidade dos resultados; e a Seção VI apresenta as conclusões e os trabalhos futuros.

2. Trabalhos Relacionados

No trabalho relacionado [e Cleber Alcântara 2018] o autor realiza a mineração dos dados educacionais provenientes da Universidade Federal do Rio Grande (FURG) referentes ao período entre 2012 e 2017. O autor utilizou dados do sistema acadêmico da FURG, realizando a engenharia de características para a extração das variáveis faixa etária, média geral no Exame Nacional do Ensino Médio, o ENEM, intervalo de tempo entre a conclusão do ensino médio e o ingresso no curso de graduação, além do último coeficiente de rendimento do aluno. Os cursos foram divididos em quatro categorias (de acordo com a área de cada curso). Também detinham outros atributos como gênero (M ou F), trajetória escolar, estado de origem e se o estudante foi ou não bolsista em algum momento da

graduação. Os testes de predição de evasão foram realizados com o algoritmo J48, uma implementação do algoritmo de árvore de decisão C4.5, desenvolvido por Ross Quinlan e amplamente utilizado para tarefas de classificação, atingindo 90,7% de precisão geral com o modelo.

No trabalho relacionado [Belenke dos Santos 2021], o autor utiliza dados fornecidos pelo Instituto Federal de Santa Catarina (IFSC), campus de Caçador. Após a aquisição dos dados extraídos da base PostgreSQL da instituição, foram realizados os passos de pré-processamento dos dados e engenharia de características, gerando as seguintes variáveis: 'idade', 'status' (número de disciplinas em que o aluno foi aprovado, dividido pelo total de disciplinas que o aluno cursou no semestre), 'penultimo_status' e 'ultimo_status' (equivalentes aos valores de 'status' do penúltimo e último semestres, respectivamente), a média dos 'status', 'diferenca_anos', que contabiliza a diferença entre o ingresso no ensino superior e o término do ensino médio em anos, e a média de faltas no penúltimo e último semestres. Utilizando a técnica *one-hot-encoding* também foram criadas variáveis booleanas correspondentes ao estado civil e à etnia do estudante.

Em seus testes utilizou os algoritmos de Árvore de Decisão e Rede MLP. Com o primeiro conseguiu atingir acurácia de 84%, com taxa de precisão de 87% e de revocação de 76% para a classe dos evadidos. Com a Rede MLP alcançou 82% de acurácia nas predições com taxa de precisão de 78% e taxa de revocação de 82% para a classe dos evadidos. O resumo dos resultados pode ser observado na Tabela 2.

Trabalho Relacionado	Algoritmo	Acurácia Geral
1 - [e Cleber Alcântara 2018]	J48 (Árvore de Decisão)	90,70%
2 - [Belenke dos Santos 2021]	Árvore de Decisão	84%
2 - [Belenke dos Santos 2021]	Rede Neural MLP	82%

Tabela 1. Resultados dos trabalhos relacionados

3. Metodologia de Desenvolvimento

Este trabalho propõe uma abordagem de mineração de dados educacionais para analisar os principais fatores que influenciam a evasão no ensino superior. A metodologia adotada pode ser visualizada no diagrama representado pela Figura 1.

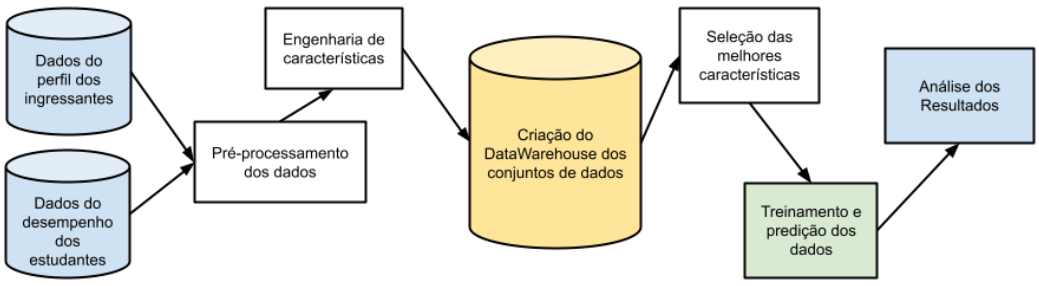

Figura 1. Diagrama da representação dos

A metodologia é composta por cinco etapas fundamentais: Coleta de Dados; Pré-processamento de Dados; Engenharia de características; Data Warehousing; Modelagem

Preditiva; e Análise dos Resultados. Esta metodologia não só busca antecipar a evasão dos estudantes, mas também tem como objetivo fornecer subsídios valiosos para o desenvolvimento de estratégias eficazes de retenção e suporte aos alunos, contribuindo assim para a melhoria do ensino superior a distância. Cada etapa será descrita a seguir.

3.1. Coleta de Dados

Nesta etapa, são obtidos dados históricos dos alunos da Universidade Virtual do Estado de São Paulo (UNIVESP), abrangendo informações demográficas, acadêmicas e de desempenho. Os dados coletados para a pesquisa se referem aos períodos letivos de 2017 a 2022, e foram derivados de duas fontes: os questionários socioeconômicos preenchidos pelos candidatos durante a inscrição do vestibular e os dados dos registros acadêmicos dos estudantes provenientes do SEI - Plataforma Educacional da UNIVESP. Dentre as atividades desenvolvidas nesta etapa, destacam-se: a familiarização com os dados, fundamental para validação dos dados coletados e garantia que sejam precisos, confiáveis e relevantes para o estudo de caso; organização e armazenamento dos dados em formato adequado para facilitar o tratamento e análise; a anonimização dos dados, garantindo que não hajam informações pessoais ou confidenciais expostas.

Os dados brutos dos questionários socioeconômicos foram obtidos no formato XLS, sendo um arquivo para cada respectivo ano, enquanto os dados dos registros acadêmicos foram extraídos diretamente do banco de dados de produção do sistema SEI da UNIVESP, armazenado no Sistema Gerenciador de Banco de Dados PostgreSQL, por meio de consultas na linguagem SQL, e os resultados armazenados no formato XLS.

Esta pesquisa foi realizada de acordo com os princípios éticos da pesquisa científica. Os dados foram anonimizados para garantir a privacidade dos estudantes, as análises foram realizadas de forma a preservar a confidencialidade das informações. Características como nome e CPF dos estudantes foram descartadas no início do processo.

3.1.1. Conjunto de dados 1 - Dados da Inscrição

O primeiro conjunto de dados, contendo 132.763 registros, é composto por dados referentes aos períodos letivos de 2017 a 2022, e foram derivados dos questionários socioeconômicos preenchidos pelos candidatos durante a inscrição do vestibular, incluindo dados acadêmicos, socioeconômicos e comportamentais dos alunos. As variáveis que compõe este conjunto são: código do curso para o qual o estudante se inscreveu, renda familiar e renda individual, ambas mensuradas em salários mínimos, níveis de escolaridade do pai e da mãe, se o estudante possui graduação superior, idade, gênero, código da unidade de ensino, estado civil, etnia, trajetória escolar, participação na renda familiar e situação matricular.

3.1.2. Conjunto de dados 2 - Histórico

O segundo conjunto de dados é composto pelos registros acadêmicos dos estudantes. Contém 430.5611 registros extraídos diretamente do banco de dados de produção do sistema SEI da UNIVESP, armazenado no Sistema Gerenciador de Banco de Dados PostgreSQL, por meio de consultas na linguagem SQL, e os resultados armazenados em for-

mato CSV. Os registros são referentes ao desempenho dos estudantes nas disciplinas cursadas. As variáveis utilizadas são: total de pontos no processo seletivo, código da disciplina cursada, frequência, média final e período.

3.2. Pré-processamento

Nesta etapa, os dados passam por um rigoroso processo de limpeza e transformação para garantir qualidade e integridade, incluindo tratamento de outliers e valores faltantes. O principal problema no Conjunto de Dados 1 foi a falta de padronização nos questionários socioeconômicos, resultando em perguntas similares com respostas diferentes e diferentes padrões de armazenamento. Para resolver isso, os dados foram normalizados para um formato consistente. Também foi feita a limpeza, selecionando apenas os dados dos estudantes matriculados nos anos e cursos correspondentes. Em seguida, aplicou-se *one-hot-encoding* para representar variáveis categóricas como vetores binários, resultando em 507 características a partir das 14 variáveis originais. As amostras foram selecionadas com tamanhos iguais para as classes "Não Evadidos"e "Evadidos", com base na menor classe de "Não Evadidos"que tinha 10.023 registros.

Para a seleção dos evadidos, foram utilizados os registros dos estudantes com situação de matrícula cancelada. Para a seleção dos não evadidos, foram utilizados somente registros de estudantes formados. A última etapa do pré-processamento para este conjunto, antes de utilizá-lo nos testes de predição, foi a eliminação das variáveis constantes, aquelas que têm o mesmo valor em todos as amostras de um conjunto de dados, eliminando 4,5% das variáveis geradas no processo de *one-hot-encoding*.

Antes da utilização do segundo conjunto de dados, estes obtidos da base PostgreSQL que armazena os dados do sistema, fez-se necessária a aplicação dos seguintes passos de pré-processamento: Remoção das disciplinas sem nota avaliativa; Padronização das médias para uma mesma escala (0 - 10), sendo que os valores não estavam dentro de um mesmo intervalo; Padronização dos valores de frequência para uma mesma escala (0 - 100), sendo que os valores não estavam dentro de um mesmo intervalo; Preenchimento das frequências faltantes com a média das frequências; Aplicação da técnica de *one-hot-encoding* para a variável correspondente às disciplinas;

Um dos objetivos deste estudo é prever a evasão de estudantes, destacando a importância de realizar essa previsão o mais cedo possível. Por isso, foram analisados os dados de desempenho do primeiro semestre dos estudantes. Verificou-se que mais de 90% dos alunos cursaram entre seis e oito disciplinas nesse período. Portanto, o número ideal de disciplinas, médias e frequências considerado foi oito. Cada disciplina, média e frequência corresponde a uma coluna na tabela, resultando em oito registros distintos para cada aluno.

Para incluir todas as características nos modelos preditivos, elas devem ser tratadas como pertencentes ao mesmo registro, caso contrário, cada registro seria interpretado como um estudante diferente pelo algoritmo. Os dados do segundo conjunto foram "desnormalizados", resultando em oito colunas para médias e oito para frequências, registrando todos os valores de um estudante em uma mesma linha. Estudantes com menos de oito disciplinas cursadas no primeiro semestre tiveram médias e frequências faltantes preenchidas com a média de suas anteriores. Esse método visou aproveitar ao máximo as informações disponíveis no segundo conjunto.

3.3. Engenharia de Características

Nesta etapa, são geradas novas variáveis a partir das características originais dos dados, buscando informações relevantes que possam afetar a evasão dos estudantes, sendo uma etapa crucial no desenvolvimento de modelos de aprendizado de máquina. Envolve a criação, transformação e seleção de características (também chamadas de variáveis ou atributos) dos dados brutos, a fim de estas características sejam mais representativas para um futuro modelo de aprendizado de máquina [Dong and Liu 2018]. O primeiro conjunto de dados consiste principalmente em variáveis categóricas, com apenas ”idade”como variável discreta. Para manter consistência nas características para os algoritmos preditores, a variável ”idade”foi convertida em ”faixa etária”, utilizando a divisão do Instituto Brasileiro de Geografia e Estatística (IBGE), que agrupa a população a cada quatro anos. Assim, os dados foram classificados em grupos específicos, com o respectivo percentual de estudantes em cada faixa.19 anos ou menos”(5,77%); ”Entre 20 e 24 anos”(13,17%); ”Entre 25 e 29 anos”(18,99%); ”Entre 30 e 34 anos”(20,47%); ”Entre 35 e 39 anos”(17,62%); ”Entre 40 e 44 anos”(11,51%); ”Entre 45 e 49 anos”(6,79%); ”Entre 50 e 54 anos”(3,45%); ”Entre 55 e 59 anos”(1,53%); ”Entre 60 e 64 anos”(0,47%); ”Entre 65 e 69 anos”(0,14%); ”Entre 70 e 74 anos”(0,03%); e ”Mais de 75 anos”(0,009%).

3.4. Data Warehousing

Nesta etapa, os conjuntos foram unidos e os dados pré-processados foram armazenados em um data warehouse. Esses dados passaram por integração e consolidação para garantir qualidade e consistência, gerando um novo conjunto com 566 variáveis, a maioria binária e algumas contínuas. Devido à diversidade de tipos e escalas das variáveis, realizou-se a normalização, utilizando a função MinMaxScaler da biblioteca preprocessing do ScikitLearn. Além disso, as variáveis constantes foram eliminadas, resultando em um conjunto final de 551 variáveis.

3.5. Modelo de Predição para Identificação dos Principais Fatores de Evasão

O modelo proposto para identificar os principais fatores relacionados à evasão combina diversas técnicas avançadas de mineração de dados educacionais. Os seguintes algoritmos de aprendizado de máquina foram utilizados nos testes de predição da evasão: XGBoost, é um algoritmo de aprendizado de máquina baseado em árvores de decisão, otimizado para eficiência e desempenho, utilizando um sistema de gradiente impulsionado [Chen and Guestrin 2016a]; Regressão Logística, considerado componente integral de qualquer análise de dados preocupada em descrever a relação entre uma variável de resposta e uma ou mais variáveis explicativas [Hosmer Jr et al. 2013]; Árvore de Decisão, uma estrutura de dados definida recursivamente como um nó folha que corresponde a uma classe, ou um nó de decisão que contém um teste sobre algum atributo. Para cada resultado do teste existe uma aresta para uma subárvore, que possui a mesma estrutura que a árvore [Monard and Baranauskas 2003]; Floresta Randômica, uma técnica de aprendizado de máquina baseada em *ensemble*, cria várias árvores de decisão durante o treinamento e combina suas previsões para melhorar a precisão e reduzir o *overfitting*.

Além dos algoritmos de aprendizado de máquina, utiliza-se o SelectKBest, método pertencente ao conjunto de ferramentas do SciKit-Learn [Pedregosa et al. 2011]. É uma técnica de seleção de características que escolhe as **k** melhores características com base em testes estatísticos, avaliando a relação entre cada característica e a variável de

saída (rótulo) de forma independente. A função que define o *score*, ou seja, a pontuação das características, deve ser escolhida de acordo com o tipo das variáveis presentes no conjunto de dados analisado. A função utilizada nos testes foi a **f_regression**, que é uma função de seleção de características para problemas de regressão, utilizando a estatística F para avaliar a relação linear entre cada característica numérica e a variável de saída, sendo indicada para características numéricas e problemas de regressão.

4. Resultados

Para analisar os modelos propostos para predição de evasão e identificação de fatores relacionados, foram conduzidos três estudos de caso. O primeiro teve como objetivo testar o desempenho dos algoritmos preditivos utilizando apenas o Conjunto de Dados 1. No segundo estudo de caso, os mesmos algoritmos de aprendizado de máquina foram testados, porém desta vez utilizando ambos os conjuntos de dados disponíveis. Os dois primeiros estudos foram realizados para identificar o algoritmo de aprendizado de máquina com melhor capacidade de classificação para a predição de evasão. Este algoritmo foi então utilizado no terceiro estudo de caso, que aplicou o método SelectKbest para selecionar as melhores características dos conjuntos de dados relacionadas à evasão. Além disso, foi utilizado o algoritmo XGBoost para realizar as predições de evasão, escolhido devido ao seu bom desempenho nos dois primeiros estudos.

Para o conjunto separado para treinamento foram utilizados 75% dos dados, sendo 25% destinados para o conjunto de teste. Os resultados a seguir referem-se ao desempenho dos algoritmos no conjunto de teste. Para facilitar a visualização, serão utilizadas as seguintes abreviaturas: AC para Acurácia (Accuracy), PR para Precisão (Precision), RE para Revocação (Recall), XG para XGBoost, RL para Regressão Logística (Logistic Regression), FR para Floresta Randômica (Random Forest), SVM para Máquina de Vetor de Suporte (Support Vector Machine), AD para Árvore de Decisão (Decision Tree), 0 representa a classe dos não-evadidos e 1 representando a classe dos evadidos.

4.1. Estudo 1

Este estudo foi conduzido utilizando somente os dados do Conjunto de Dados 1. Os resultados demonstram que, com os dados obtidos no momento da inscrição do estudante para o vestibular, pode-se prever os com tendência a evadirem-se com um mínimo de 69% de acurácia, chegando até 74%. Na Tabela 4.1, pode-se observar um desempenho próximo entre os algoritmos XGBoost, SVM e o de Regressão Logística, tendo os dois primeiros classificado corretamente 3.704 dos 5.006 estudantes que compõe o conjunto de dados de teste. O algoritmo de Regressão Logística foi o mais regular dentre todos, atingindo 73% de precisão tanto na acurácia, quanto na precisão e na revocação de ambas as classes, prevendo corretamente 1.825 dos 2.499 evadidos.

.	AC	PR - 0	PR - 1	RE - 0	RE - 1
XG	**74%**	**73%**	75%	**76%**	72%
RL	73%	73%	73%	73%	**73%**
FR	70%	71%	69%	68%	71%
SVM	**74%**	**73%**	74%	75%	72%
AD	69%	69%	69%	69%	68%

Tabela 2. Algoritmos de Aprendizado de Máquina - Conjunto de dados 1

4.2. Estudo 2

Neste foram utilizadas todas as características dos dois conjuntos de dados disponíveis. Com as características de desempenho do estudante, o nível de acurácia em relação à predição dos evadidos, elevou-se para um mínimo de 88%, atingindo 92% em seu melhor resultado. Ao observar na Tabela 4.2 que o algoritmo XGBoost obteve a melhor acurácia, 92%, prevendo corretamente 4.604 dos 5.006 registros selecionados para a base de teste. Os algoritmos SVM, Regressão Logística e Árvore de Decisão também atingiram um desempenho satisfatório, sendo que este último atingiu o melhor desempenho de revocação na classe dos evadidos, 87%, prevendo corretamente 2.174 de um total de 2.499 evadidos.

.	AC	PR - 0	PR - 1	RE - 0	RE - 1
XG	**92%**	**87%**	98%	98%	86%
RL	89%	86%	94%	94%	85%
FR	88%	81%	**99%**	**100%**	77%
SVM	89%	84%	97%	97%	82%
AD	88%	**87%**	89%	89%	**87%**

Tabela 3. Algoritmos de Aprendizado de Máquina - Conjuntos 1 e 2

4.3. Estudo 3

Na condução deste estudo de caso também foram utilizadas as características dos dois conjuntos de dados disponíveis, além do método SelectKBest que pontuou as características através do método 'f_regression' e, utilizando as 10 melhores como ponto de partida, realizou o teste preditivo com o algoritmo XGBoost, sendo adicionada uma característica a cada iteração, de acordo com a classificação fornecida pelo algoritmo. A Figura 2 mostra a evolução do desempenho conforme foram adicionadas as variáveis ao modelo, observando-se que os valores se estabilizam após a adição de aproximadamente 100 características.

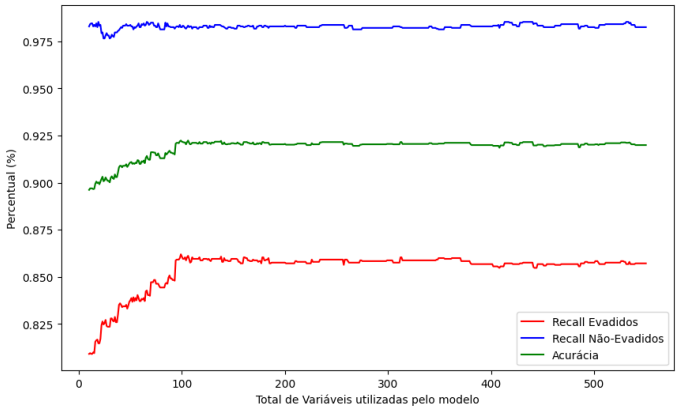

Figura 2. Desempenho do algoritmo XGBoost pontuadas pelo SelectKBest

No processo iterativo de testes das 566 características, o SelectKBest selecionou 99 delas, com 53 do histórico do estudante e 46 do momento da inscrição. Isso resultou em uma acurácia de 92%, com 4.617 de 5.006 estudantes classificados corretamente. A

precisão foi de 88% para os não-evadidos, com 2.463 de 2.808 previstos corretamente, e de 98% para os evadidos, com 2.154 de 2.198 previstos corretamente. A taxa de revocação foi de 98% para os não-evadidos, com 2.463 de 2.507 corretamente classificados, e de 86% para os evadidos, com 2.154 de 2.499 corretamente classificados.

5. Análise dos Resultados e Reprodutividade

As características mais relevantes na antecipação de um estudante com risco de evasão incluem uma combinação dos dados fornecidos pelos dois conjuntos. Os dados do histórico curricular foram particularmente significativos para o algoritmo XGBoost, usado para classificar as características conforme sua importância. Somando as importâncias das características de uma mesma variável, podemos determinar sua relevância global. Na Figura 3, pode-se observar estas características e seu respectivo percentual de importância para o algoritmo classificador.

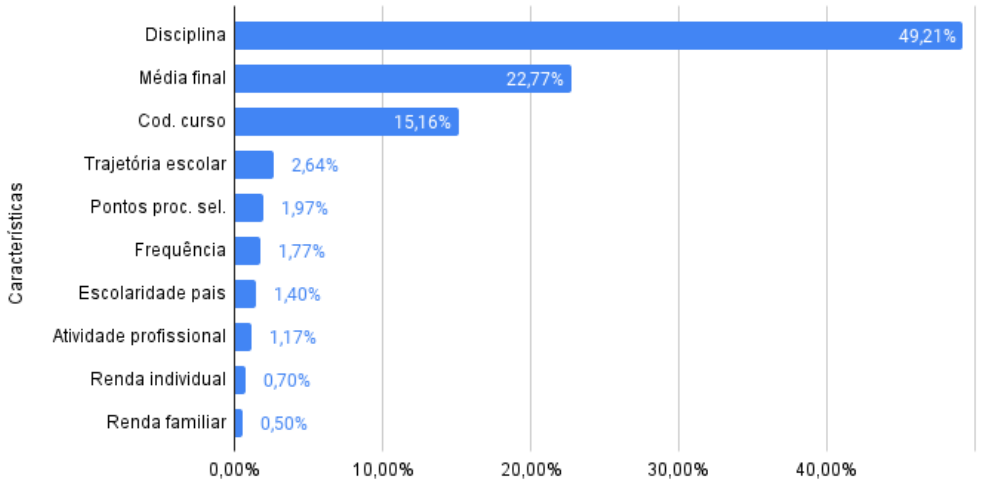

Figura 3. Percentuais de relevância gerados pelo XGBoost

Ao analisar o gráfico, destaca-se que as disciplinas cursadas pelo estudante desempenham um papel significativo na predição da evasão, abrangendo 36% características selecionadas pelo modelo. Uma análise mais detalhada revela que disciplinas como Informática, Cálculo I, Matemática, Física I, Inglês e Informática têm uma taxa de evasão superior a 80%. Esta taxa é calculada dividindo o número total de evadidos que cursaram essas disciplinas pela soma dos totais de formados e evadidos. Disciplinas dos cursos de Engenharia de Produção, Engenharia da Computação, Matemática e Bacharelado em Tecnologia da Informação também estão entre aquelas com uma proporção de evasão superior a 67%. A importância conferida às disciplinas cursadas deve ser considerada com cautela, pois é influenciada pelas taxas de evasão de seus respectivos cursos. O desempenho dos estudantes, medido pelas médias das notas obtidas, também foi considerado um fator relevante pelo algoritmo para prever a probabilidade de evasão.

Já o curso escolhido pelo estudante representou 13% das características selecionadas quando obtido o melhor resultado preditivo, sendo utilizadas todas as geradas a partir desta variável, ou seja, todos os 13 cursos presentes no conjunto de dados foram

utilizados no modelo. A importância desta se comprova ao observarmos a diferença de percentual de evadidos de cada curso. Enquanto Engenharia de Produção, Engenharia da Computação e Matemática possuem taxas de evasão superiores a 45%, os cursos de Bacharelado em Ciências de Dados, Bacharelado em Administração e Tecnólogo em Processos gerenciais ficam abaixo dos 11% de evadidos. O algoritmo também utilizou todos os dados disponíveis acerca da frequência, corroborando com o averiguado no trabalho de [Belenke dos Santos 2021], em que a adição da frequência do estudante melhorou em 8% o desempenho do algoritmo preditivo.

Características relativas a gênero, trajetória escolar, renda individual e à atividade profissional exercida também foram utilizadas, sendo observado em análise que, mulheres evadem menos que homens, 26% a 34%, alunos advindos de escolas técnicas evadem menos que os de escolas públicas não técnicas ou particulares, 14%, 28% e 31% respectivamente, pessoas que ganham até 2 salários mínimos evadem 5% a menos que pessoas com faixa salarial entre 2 e 5 salários, e que estudantes que trabalham na área do curso escolhido evadem menos que os que aqueles que não trabalham, 14% a 35%.

A metodologia apresentada neste trabalho pode ser reproduzida ao seguir as instruções documentadas no endereço do GitHub https://github.com/harrison-bm/mde, estando disponíveis os conjuntos de dados, devidamente anonimizados e o código Python desenvolvido para a realização dos testes. Também pode ser reproduzida em diferentes conjuntos de dados, que tenham características semelhantes, aplicando as técnicas descritas neste trabalho.

6. Conclusões e Trabalhos Futuros

Neste trabalho foi demonstrada a eficácia das técnicas de mineração de dados educacionais e aprendizado de máquina na identificação dos fatores que influenciam a evasão no ensino superior utilizando dados fornecidos pelo estudante no momento da inscrição e posteriormente dados referentes ao desempenho no primeiro semestre de seu curso.

A integração das técnicas sugeridas pela metodologia proposta para mineração e tratamento dos dados e, a partir de sua análise, a seleção das características a serem passadas para os algoritmos de aprendizado de máquina de forma iterativa, mostrou-se eficaz na tarefa de prever estudantes com tendência à evasão, alcançando acurácia geral de 92%, assim como evidenciar as variáveis mais significativas com o uso do SelectKBest, sendo utilizadas menos de 20% das características disponíveis para realizar tal predição. Foi demonstrada também o bom desempenho dd modelo de predição de evasão a partir dos dados do perfil do ingressante, que sem nenhum dado referente ao desempenho do estudante, alcançou bons resultados em prever aqueles com maior probabilidade de evasão, no ato da ingressão do vestibular, atingindo 74% de acurácia.

Para trabalhos futuros recomenda-se o acréscimo de variáveis que representem o desempenho do aluno, inclusive notas do ENEM e dados das interações dos alunos com o ambiente virtual de aprendizagem, que podem fornecer uma visão mais abrangente e refinada dos fatores que influenciam o sucesso acadêmico. Os trabalhos relacionados também mostram bons resultados ao utilizar dados referentes ao período de tempo decorrido entre a formação do estudante no ensino médio e o ingresso no curso superior. A exploração de outros métodos de aprendizagem de máquina, assim como técnicas de aprendizagem profunda, também podem oferecer perspectivas adicionais e complementares.

7. Agradecimentos

O presente trabalho foi parcialmente financiado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES).

8. Referências

Agresti, A. (2012). *Categorical data analysis*, volume 792. John Wiley & Sons.

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish journal of emergency medicine*, 18(3):91–93.

Belenke dos Santos, J. C. (2021). Usando mineração de dados para predição da evasão escolar.

Bittencourt, H. R. (2003). Regressão logística politômica: revisão teórica e aplicações. *Acta Scientiae*, 5(1):77–86.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais (Inep). Censo da Educação Superior. Brasília, DF (2023). INESP. `https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior`.

Chen, T. and Guestrin, C. (2016a). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chen, T. and Guestrin, C. (2016b). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Dong, G. and Liu, H. (2018). *Feature engineering for machine learning and data analytics*. CRC press.

e Cleber Alcântara, M. L. (2018). Predição de alunos com risco de evasão: estudo de caso usando mineração de dados. *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)*, 29(1):1921.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. "O'Reilly Media, Inc.".

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Liu, Y., Wang, Y., and Zhang, J. (2012). New machine learning algorithm: Random forest. In *Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3*, pages 246–252. Springer.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9(1):381–386.

McKinney, W. et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX.

Monard, M. C. and Baranauskas, J. A. (2003). Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, 1(1):32.

Oliveira, L. R. and Costa, S. R. (2021). Fatores que contribuem para a evasão escolar em cursos de nível superior. *Revista Espacios*, 42(11).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Powell, S. (2018). The book of why: The new science of cause and effect. pearl, judea, and dana mackenzie. 2018. hachette uk. *Journal of MultiDisciplinary Evaluation*, 14(31):47–54.

Romero, C., Romero, J. R., and Ventura, S. (2014). A survey on pre-processing educational data. *Educational data mining: applications and trends*, pages 29–64.

Schmitt, J. A. et al. (2018). Identificação de alunos com tendência à evasão nos cursos de graduação à distância por meio de mineração de dados educacionais.