EDUCATIONAL DATA MINING: PREDIÇÃO DE EVASÃO ESCOLAR EM COLÉGIOS PÚBLICOS DE SALVADOR

Rogério Magnavita Nogueira Júnior;

Celso Barreto da Silva;

Fábio Fonseca Barbosa Gomes;

José Vicente Cardoso Santos

RESUMO

Educational Data Mining diz respeito à aplicação de técnicas de mineração de dados, capaz de reconhecer padrões nos dados, no campo da educação. Este projeto tem como objetivo geral a construção de um modelo de aprendizagem de máquina capaz de realizar predições e classificar um aluno de acordo com a probabilidade do aluno evadir dos estudos. Os objetivos específicos estão alinhados com a metodologia, divididos na coleta e na limpeza dos dados, no treinamento do modelo e na visualização e análise dos resultados. Os dados foram coletados através de um formulário. Os algoritmos usados foram Random Forest, Naive-bayes e KNN. Random Forest obteve o maior desempenho e foi o escolhido para o modelo final.

Palavras-chave: Mineração de dados educacionais; Aprendizagem de máquina; *Random Forest*; *K-nearest Neighbor*; *Naive-bayes*.

1. INTRODUÇÃO

Levando em consideração a sociedade da informação como um campo de estudo da definição de informação e suas contribuições para a sociedade em diversos campos (RABAN et al., 2011), é possível retornar à frase "dados são o novo petróleo" (HUMBY, 2006, apud PALMER, 2006). São valiosos, mas precisam ser refinados e analisados para demonstrarem seu valor; dão suporte para tomadas de decisões corretas (PALMER, 2006). Ao expandir-se para a área de educação, duas vertentes são notadas. Uma delas é a Mineração de Dados Educacionais (EDM), que já encontrava tração desde 2006 (ROMERO; VENTURA, 2007).

De acordo com Kabathova e Drlik (2021), EDM lida principalmente com dados relacionados ao estudo para entender o comportamento do estudante. Complementando, há a área de Learning Analytics (LA), que lida com a coleta, análise e o relatório de informações sobre o histórico e o contexto do aluno. Entre essas duas áreas, a evasão escolar é um dos temas mais pesquisados.

A aplicação de técnicas de aprendizagem de máquina (do inglês Machine Learning, abreviado ML) é capaz de retirar informações não tão óbvias acerca da evasão escolar, como a correlação entre uma característica do aluno e a sua possibilidade de abandonar os estudos. O presente trabalho utiliza o conceito de evasão em termos gerais de Fialho (2014).

A definição de evasão escolar não é um consenso. Para Fialho (2014), "evasão significa a não conclusão de qualquer unidade educacional que guie o aluno a um conhecimento especializado, em termos gerais." Segundo Evaristo (2017), evasão difere de abandono; seria abandono a infrequência do estudante durante o ano letivo, já a evasão é caracterizada pela não realização da matrícula para continuar seus estudos.

Apesar da evasão ser um tema popular, seu escopo possibilita a justificativa do estudo, pois os níveis de ensino e contextos locais influenciam os indicadores e os resultados. O estudo e aplicação das técnicas em bancos de dados do nível superior é popular (TEODORO; KAPPEL, 2020), (MANHÃES et al. 2012), (NEVES et al. 2021). No nível de Ensino Médio, o estudo pode indicar a raiz do problema e contribuir para intervenções eficazes, resultando no combate às suas implicações (GIL et al. 2020).

Para tais finalidades, são aplicados métodos estatísticos que encontram padrões nos dados. Esse procedimento generalizado define o que visa o campo de Machine Learning. Em nível de aplicação, pode ser exemplificado como um sistema de carro com piloto automático que recebe dados de seus sensores e encontra padrões de "perigo" (CARLEO et al. 2019).

Conforme os conceitos vistos, o estudo propõe a elaboração de um modelo de aprendizagem de máquina, com a finalidade de identificar alunos que sejam prováveis de abandonar seus estudos. Ao fornecer o modelo a uma instituição de ensino, os devidos profissionais podem utilizá-lo para fundamentar intervenções. Os objetivos específicos

são divididos na coleta de dados, no tratamento e filtragem desses dados para que sejam relevantes, na aplicação e treinamento dos algoritmos de ML, e na análise desses resultados.

2. FUNDAMENTAÇÃO TEÓRICA

Para Gil et al. (2020), motivos comuns para fracassar nos estudos estão relacionados a dificuldades financeiras, problemas familiares, questões de saúde, baixo desempenho em notas e o ambiente de aprendizado. Alguns dos indicadores utilizados para definir se o aluno tem possibilidade de abandonar os estudos foram idade, gênero, notas, porcentagem de presença, distância entre o domicílio e a escola, escolaridade dos pais e renda familiar.

A questão de gênero é abordada de maneira mais profunda, no contexto da Índia por Hassan e Mirza (2020). Através de entrevistas com parentes, alunos, professores e ex-estudantes, foram levantados alguns indicadores. A justificativa para o gênero ser relevante está em questões como: crenças sociais que definem estereótipos e papéis de gênero, preocupações com violência, gravidez, menstruação, casamentos arranjados desde a infância e a preferência de focar nos estudos dos filhos homens. Também é levantada a quantidade de irmãos, uma vez que existem casos de abandono por ser necessário cuidar dos irmãos mais novos.

Ainda no estudo de Hassan e Mirza (2020), indicadores relacionados à infraestrutura da escola também são pertinentes, como a existência de rampas de acessibilidade e banheiros bem conservados; em alguns casos, até a presença de professoras mulheres têm impacto, devido ao medo de violência baseada em gênero. Em cima dos dados levantados, foram utilizados os algoritmos de regressão logística, SVM (Support Vector Machine), CART (Classification and Regression Trees), Naive Bayes e Random Forest, através do método de aprendizagem supervisionada. As métricas de performance indicaram que os algoritmos mais eficazes foram regressão logística e Random Forest.

Retornando para o contexto brasileiro, com o intuito de encontrar aplicações e testar técnicas de Machine Learning, Souza (2020) realizou um trabalho que utilizou a metodologia CRISP-DM em sua construção. Tal metodologia possui 6 etapas:

compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implementação.

Seguindo essa metodologia, durante a preparação e modelagem dos dados, o estudo de Souza (2020) reuniu 32 indicadores acadêmicos e socioeconômicos. Entre eles, informações sobre as instituições de ensino onde o aluno estudou anteriormente são um diferencial entre os estudos citados anteriormente. Necessidades especiais também foram levadas em conta. Os melhores algoritmos para a tarefa de acordo com métricas de performance foram Two-class Boosted Decision Tree e Two-class Neural Network.

A aprendizagem de máquina pode ser dividida em supervisionada e não supervisionada. Na supervisionada, o sistema recebe dados de entrada também com os desejados dados de saída durante o treinamento; o objetivo é que o sistema aprenda a exibir o resultado correto dada uma inédita entrada (GHAHRAMANI, 2003). O tipo de saída subdivide a aprendizagem supervisionada em duas categorias: classificação e regressão. Variáveis contínuas são sábados de regressão, enquanto em classificação a variável de sáida corresponde a um rótulo dentre um conjunto finito de possibilidades (SOOFI; AWAN, 2017).

Para alcançar bons resultados, é importante que os algoritmos sejam adequados ao tipo e problema. Entre os algoritmos de classificação encontrados na literatura (KOTSIANTIS et al. 2006) e nos trabalhos relacionados (TEODORO; KAPPEL, 2020) (NEVES et al., 2021) foram selecionados:

Random Forest: um classificador composto por múltiplas árvores de decisão, onde cada árvore individualmente realiza uma predição e o modelo decide pelo resultado mais recorrente entre todo o conjunto de árvores (BREIMAN, 2001).

Uma árvore de decisão, por sua vez, classifica os dados ao realizar uma série de testes acerca dos atributos dos itens estudados. Cada teste é representado por um nó, e seus resultados são representados por um ramo que se dá em outro nó. Com esses elementos, é possível realizar uma hierarquia de testes. Os rótulos de sáida são denominados de ”folha”; são os resultados finais (KINGSFORD; SALZBERG, 2008).

K-Nearest Neighbor (KNN): a partir da hipótese de que instâncias semelhantes estão próximas num espaço com *n* dimensões, o algoritmo utiliza da distância entre dois

pontos para realizar a comparação entre a tupla de teste e a de treinamento (NEVES et al. 2021).

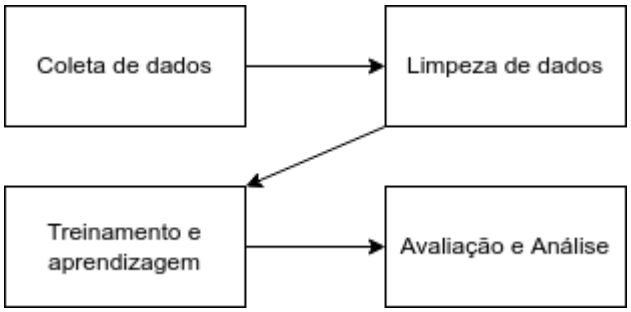
Classificador Naive-Bayes: se baseia na teoria de probabilidade. O algoritmo assume que não há dependência entre seus atributos (HAN et al. 2011) e os confere o mesmo nível de importância (GIL et al. 2020).

Com a utilização de mais de um algoritmo, o objetivo principal é encontrar o mais adequado para o problema. Para isso, são utilizadas métricas de avaliação através da quantificação de performance.

3. METODOLOGIA

O presente projeto é estruturado na coleta e limpeza dos dados, seguido do treinamento dos modelos de aprendizagem de máquina, e finalizando na utilização de métricas de avaliação e na análise dos resultados.

Figura 1. Metodologia do projeto



Fonte: Elaborado pelo autor (2022).

Os processos de coleta e limpeza de dados estão alinhados com o processo de Extração, Transformação e Carga (ETL, do inglês Extract Transform Load), pois os dados utilizados neste projeto foram extraídos de mais de uma fonte (FERREIRA et al. 2010). Nessa seção, serão descritas as etapas do projeto.

3.1. Coleta dos dados

A população desse projeto são estudantes ou ex-estudantes de colégios públicos em Salvador. O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

(INEP) disponibiliza abertamente uma vasta quantidade de dados do Censo Escolar, que foram utilizados na escrita do artigo.

Quadro 1. Descrição do formulário

Pergunta	Atributo	Observação
Você estuda/estudou em um colégio público de Salvador, Bahia?	is_valid	Boolean (verdadeiro ou falso). Determina a resposta como dentro do escopo da pesquisa.
Nome (opcional)	nome	Texto.
Idade	idade	Numérico. A ser utilizado para dizer a distorção idade-série.
Gênero	genero	Múltipla escolha entre feminino, masculino, outro e "prefiro não dizer".
Renda da família	faixa_renda	Múltipla escolha entre faixas salariais:"menor que 1 salário mínimo", "entre 1 até 2 salários mínimos", "maior que 2 salários mínimos até 3 salários mínimos", "maior que 3 salários mínimos até 4 salários mínimos", "maior que 4 salários mínimos".
Acerca de uma instituição de ensino, você está:	status_aluno	Define o status atual de quem está preenchendo o formulário; se está matriculado e frequentando, matriculado e não frequentando, ou não matriculado.
Caso esteja matriculado, você mora perto da escola?	mora_perto	Boolean (verdadeiro ou falso).
Caso você tenha abandonado os estudos, qual foi o motivo? Selecione todas as opções aplicáveis.	motivo_evasao	Múltipla escolha dentre as opções: Financeiro, saúde, familiar, pouco suporte emocional, desinteresse, dificuldades de aprendizagem, distância/dificuldade de locomoção, bullying, gravidez não planejada e a opção "outros", que permite respostas por escrito.
Caso não tenha abandonado os estudos, alguma vez já pensou em abandonar? Se sim, por qual motivo?	possivel_motivo_evasao	Permite as mesmas escolhas da pergunta anterior, porém com a adição da opção "nunca pensei em abandonar os estudos".
Caso tenha abandonado os estudos, em qual ano você esteve matriculado em uma escola por último? (exemplo: 2019).	ano_evasao	Para saber se cabe no escopo da pesquisa e na hora da análise ter o contexto geral da época.
Caso tenha evadido, o que o faria mudar de ideia para voltar aos estudos?	possivel_intervencao	Permite uma resposta por escrito. A resposta pode ser utilizada para entender ainda mais o motivo da evasão.
Em qual série(ano) você está? (ou esteve por último, caso tenha abandonado os estudos).	serie	Múltipla escolha entre o 6º ano do Ensino Fundamental até o 3º ano do Ensino Médio (de acordo com a oferta de ensino dos colégios onde o formulário foi compartilhado) com uma opção "outros" para respostas não esperadas.
Maior escolaridade dos pais ou responsáveis	escolaridade_pais	Múltipla escolha, desde fundamental incompleto até a pós-graduação, também tendo como alternativa a opção "Não sei ou prefiro não dizer".

Fonte: Elaborado pelo autor (2022).

Foi realizado um pedido para o INEP e para a Secretaria de Educação do Estado da Bahia (SEC) disponibilizarem dados classificados que seriam utilizados no treinamento do modelo, como dados a nível individual com notas e frequência, pedido que foi negado. Como método alternativo, foi criado um formulário na plataforma Google Forms, com perguntas estratégicas que definirão os indicadores (quadro 1).

Os dados acerca da infraestrutura de escolas foram retirados dos microdados do Censo Escolar de 2021, disponibilizados no portal do INEP para acesso público, foram

baixados em formato de planilha, para ter uma visão geral das escolas a serem contatadas. Os dados foram filtrados de acordo com o município, obtendo a amostra de dados dos colégios de Salvador. Em seguida, os dados foram organizados de acordo com a quantidade de estudantes matriculados nas escolas, de acordo com o maior número de matrículas para o menor. A partir daí, foram selecionados os colégios para tentativa de contato.

O formulário foi compartilhado no Colégio Estadual Batista Neves, no Instituto Municipal de Educação Professor José Arapiraca e no Colégio Estadual Monsenhor Manoel Barbosa, que das escolas contatadas foram as que responderam e permitiram o compartilhamento. Os alunos que responderam os formulários também foram informados para compartilhá-lo com colegas que evadiram, caso ainda mantivessem contato.

3.2. Limpeza de dados

Nessa etapa, foram exportados os resultados do formulário em uma planilha do Google Sheets para realizar os processos de eliminação de respostas em branco, inválidas e não condizentes com a pesquisa.

A respeito da figura 2, no campo ”idade” e ”ano evasão”, os textos foram removidos e a coluna passou a receber apenas números. Campos com respostas extensas tiveram suas sentenças reduzidas e simplificadas através da ferramenta ”localizar e substituir”, como o campo de renda da família e o campo de status. Respostas inconsistentes, como as de alunos que marcaram que estavam matriculados e frequentando, mas responderam o formulário como se tivessem evadido, foram consideradas como não evasão. Pessoas que marcaram ”Não matriculado” e responderam a pergunta acerca do ano de evasão foram consideradas como ”Evasão”, e as que não responderam foram consideradas como ”Não evasão”. Respostas duplicadas e identificadas pela duplicidade no campo ”nome” tiveram a versão mais recente mantida e a mais antiga deletada, partindo do pressuposto que a resposta recente está corrigindo algum erro enviado na anterior.

3.3. Treinamento e aprendizagem

Nessa etapa foi utilizado o software Weka, versão 3.9.6. Os dados foram carregados e o campo ”nome” foi removido do modelo. Devido a um desbalanço nas classes, onde 22 eram casos de evasão e 65 não eram, foi utilizado o filtro SMOTE (Synthetic Minority Over-sampling Technique) disponibilizado na ferramenta, que adiciona novas instâncias entre uma instância da classe em minoria e seus vizinhos mais próximos. Após a aplicação do filtro, os dados possuem 65 instâncias de não evasão e 44 instâncias de evasão.

Os campos ”motivo evasão” e ”possível motivo evasão” foram removidos do modelo durante o treinamento, partindo do pressuposto que em um caso real, na aquisição de novos dados, o modelo não terá acesso a essas informações.

Os algoritmos selecionados são: Random Forest, *K-Nearest Neighbor* (KNN) e Naive-Bayes.

Para cada algoritmo foram utilizadas duas opções na divisão entre dados de treina- mento e dados de teste: validação cruzada k-fold (assumindo k = 10) e a divisão por porcentagem (reservando 66% do conjunto de dados para treinamento e os outros 34% para testes). A validação cruzada funciona ao dividir o conjunto de dados em *“k”* subgrupos, treinando o algoritmo em selecionados *“k*-1*”* desses grupos. O grupo restante é utilizado como teste. Esse processo é repetido até que todos os subgrupos tenham sido utilizados uma vez como subgrupo de teste.

Foi utilizada a configuração padrão do WEKA para os algoritmos Random Forest e Naive-Bayes. Para o KNN, foram realizados ajustes no valor de K conforme testes consecutivos para o que encontrasse melhor desempenho na validação cruzada (Cross Validation) ao mesmo tempo que evitasse sobreajuste. No WEKA, o algoritmo KNN é chamado de IBk.

3.4. Métricas de avaliação

Para avaliar um modelo de aprendizagem de máquina, algumas métricas são comumente adotadas. As métricas utilizadas serão: ”Acurácia”, ”Precisão” e ”Revocação”, ambas com base na matriz de confusão (figura 2), e a partir delas, o F-measure.

A ”Classe Prevista” corresponde à classe que o modelo aponta nos dados de saída, e a ”Classe Real” corresponde à verdadeira classe dos dados de entrada. Desse modo, o Verdadeiro Negativo (VN) diz respeito a um aluno que não evadiu e que o modelo apontou como não evasão. O Falso Negativo (FN) são os casos onde o modelo apontou como não evasão, mas os dados correspondem a um aluno que evadiu. O Falso Positivo (FP) corresponde a um aluno que o modelo apontou como evasão, mas ele não evadiu. Por fim, o Verdadeiro Positivo (VP) corresponde a um aluno que o modelo apontou como evasão, e o aluno de fato evadiu.

Figura 2. Matriz de confusão

		Classe Prevista	
		Não Evasão	**Evasão**
Classe Real	**Não Evasão**	Verdadeiro Negativo (VN)	Falso Positivo (FP)
	Evasão	Falso Negativo (FN)	Verdadeiro Positivo (VP)

Fonte: Elaborado pelo autor (2022).

A acurácia é uma razão entre o número de predições corretas (VP e VN) e o total de predições (equação 1). A precisão é a razão dos VP sobre os VP e os FP (equação 2), ou seja, é uma medida que está relacionada a quantos do total de valores que o modelo apontou como evasão (VP + FP) são casos de evasão de fato. A revocação (também chamada de recall) é a razão dos VP sobre os reais casos de evasão (VP e FN) (equação 3). O F-measure (também chamado de F-score ou F1) é a média harmônica entre a precisão e o recall (equação 4). O presente trabalho utilizou essas métricas para avaliar o modelo, que são fornecidas automaticamente pela ferramenta WEKA ao testar o modelo.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \quad (1)$$

$$Precisão = \frac{VP}{VP + FP} \quad (2)$$

$$Revocação(Recall) = \frac{VP}{VP + FN} \quad (3)$$

$$F1 = \frac{2 * Precisão * Recall}{Precisão + Recall} \quad (4)$$

4. RESULTADOS E DISCUSSÃO

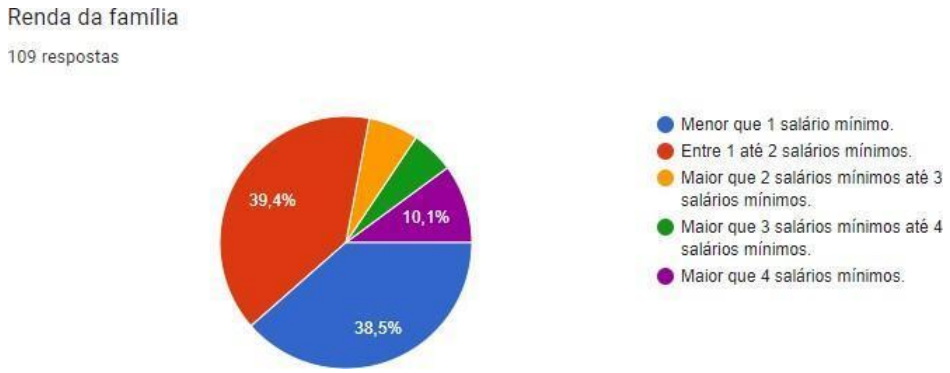
Quadro 2. Desempenho dos algoritmos.

Algoritmo	Opção de testes	Acurácia	Precisão	Recall	F-Measure
Naive-Bayes	Cross Validation 10-fold	0,789	0,714	0,795	0,753
Naive-Bayes	Treino em 66%, Teste em 34%	0,811	0,737	0,875	0,800
KNN (k=11)	Cross Validation 10-fold	0,816	0,740	0,841	0,787
KNN (k=11)	Treino em 66%, Teste em 34%	0,756	0,706	0,750	0,727
Random Forest	Cross Validation 10-fold	0,871	0,841	0,841	0,841
Random Forest	Treino em 66%, Teste em 34%	0,919	0,882	0,938	0,909

Fonte: Elaborado pelo autor (2022).

A tabela da figura 4 mostra as métricas de avaliação dos algoritmos selecionados para ambas as opções de teste. Pode-se observar que na opção de teste de divisão em porcentagem, o KNN encontrou o menor desempenho, apesar de ter alcançado um bom desempenho na validação cruzada. Já o algoritmo Naive-Bayes teve o menor desempenho na validação cruzada. Entretanto, o algoritmo Random Forest encontrou o melhor desempenho em ambas as opções de teste, consolidando-se então como o algoritmo selecionado. Uma acurácia de 0,879 e de 0,919, respectivamente, correspondem a 87,9% e 91,9% de instâncias classificadas corretamente (quadro 2).

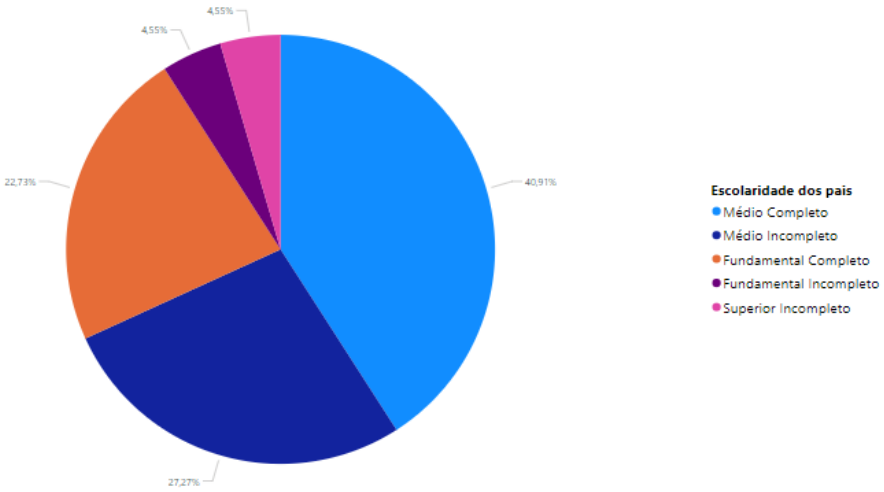
Figura 3. Gráfico do formulário sobre a renda familiar



Fonte: Elaborado automaticamente pelo Google Forms (2022).

Algumas informações de valor que podem ser retiradas das respostas do formulário são relacionadas à renda familiar (figura 3): 77,9% dizem obter renda familiar total de 2 salários mínimos ou menos, ou seja, a grande maioria da amostra. Tal resultado está de acordo com o principal fator marcados pelos alunos que evadiram ser financeiro, correspondendo a 30,14% dos casos de evasão obtidos, seguido de motivos de saúde (20,5%), de desinteresse (15,9%) e de distância e/ou dificuldade de locomoção até a escola (13,6%). Através da pergunta sobre o que faria ou fez os alunos evadidos voltarem a estudar, observou-se que a maioria que marcou “saúde” comentou sobre a pandemia do coronavírus e suas consequências.

Figura 4. Gráfico da escolaridade dos pais nos casos de evasão



Fonte: Elaborado pelo autor (2022).

O gráfico da figura 4 mostra a escolaridade dos pais nos casos de evasão. Observa-se que aproximadamente 55% não chegaram a concluir o Ensino Médio. Ainda assim, a porcentagem que concluiu ainda é significativa. Nota-se uma ausência de casos de Ensino Superior Completo e Pós-graduação, que nas respostas totais do formulário correspondem, juntos, a 8,3% dos casos. Entretanto, ao considerar os casos válidos, é uma amostra pequena para corroborar uma conclusão generalizada.

Entre os casos de evasão obtidos, é notável a discrepância entre os gêneros: aproximadamente 2/3 correspondem ao gênero feminino e 1/3 ao masculino. Entretanto, os motivos de evasão relatados são semelhantes: financeiro em sua maioria, seguido de motivos de saúde. Casos de gravidez são minoria, correspondendo a apenas um (4,5%).

5 CONSIDERAÇÕES FINAIS

O presente trabalho buscou construir um modelo de aprendizagem de máquina capaz de classificar e indicar os alunos com maior potencial de evasão, tal que é uma área de estudo importante para contribuir para o aumento da taxa de retenção dos alunos em colégios públicos.

Para alcançar a construção do modelo, definiu-se alguns objetivos específicos. Tais objetivos foram fundamentais e lógicos para as etapas de construção. Os principais obstáculos encontrados durante o trabalho foram na etapa de coleta de dados, visto que existem diversas diretrizes e políticas para o acesso a dados pessoais e sensíveis. Como o acesso não foi liberado, partiu-se de iniciativa individual a obtenção desses dados através de um formulário, entretanto sob limitações de tipos e de tamanho da amostra. O modelo construído utilizou-se do algoritmo Random Forest, que demonstrou o maior desempenho.

Em pesquisas e trabalhos futuros, pode-se insistir nessa parceria e realizar um trabalho interno com a aprovação da Secretaria de Educação do Estado da Bahia ou da Prefeitura de Salvador. Também é uma opção mudar o escopo para uma instituição privada em específico com o intuito de fornecer o modelo para ela, para que os devidos profissionais possam utilizar para criar intervenções institucionais e pedagógicas para

evitar que alunos com as maiores chances de evasão assim o façam. Tais caminhos possibilitam a inclusão de atributos importantes como a frequência e o desempenho escolar do aluno em conformidade com as diretrizes de proteção de dados.

REFERÊNCIAS

BREIMAN, Leo. Random forests. **Machine learning**, v. 45, n. 1, p. 5-32, 2001.

CARLEO, Giuseppe et al. Machine learning and the physical sciences. **Reviews of Modern Physics**, v. 91, n. 4, p. 045002, 2019.

EVARISTO, Macaé. **Macaé Evaristo - abandono e evasão escolar**. Entrevista concedida ao portal Observatório de Educação. Disponível em: <https://observatoriodeeducacao.institutounibanco.org.br/em-debate/abandono-evasão-escolar>. Acesso em: 15 maio 2022.

FERREIRA, João et al. O processo etl em sistemas data warehouse. In: **INForum**. 2010. p. 757-765.

FIALHO, Marillia Gabriella Duarte et al. A evasão escolar e a gestão universitária: o caso da Universidade Federal da Paraíba. 2014.

GHAHRAMANI, Zoubin. Unsupervised learning. In: **Summer school on machine learning**. Springer, Berlin, Heidelberg, 2004. p. 72-112.

GIL, Jay S.; DELIMA, Allemar Jhone P.; VILCHEZ, Ramcis N. Predicting students' dropout indicators in public school using data mining approaches. **International Journal of Advanced Trends in Computer Science and Engineering**, v. 9, n. 1, p. 774-778, 2020.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. Data Mining: Concepts and Techniques [Internet]. Waltham. 2011.

HASSAN, M.; MIRZA, T. Prediction of school drop outs with the help of machine learning algorithms. **GIS Sci. J**, v. 7, p. 253-263, 2020.

KABATHOVA, Janka; DRLIK, Martin. Towards predicting student's dropout in university courses using different machine learning techniques. **Applied Sciences**, v. 11, n. 7, p. 3130, 2021.

KINGSFORD, Carl; SALZBERG, Steven L. What are decision trees?. **Nature biotechnology**, v. 26, n. 9, p. 1011-1013, 2008.

KOTSIANTIS, Sotiris B.; ZAHARAKIS, Ioannis D.; PINTELAS, Panayiotis E. Machine learning: a review of classification and combining techniques. **Artificial Intelligence Review**, v. 26, n. 3, p. 159-190, 2006.

MANHÃES, Laci Mary Barbosa et al. Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. In: **Brazilian symposium on computers in education (simpósio brasileiro de informática na educação-sbie)**. 2012.

NEVES, Felipe et al. Uso de Aprendizado de Máquina para detecção de risco de evasão no curso de Licenciatura em Computação. In: **Lynx**, v. 1, n. 2, 2021.

PALMER, Michael. Data is the new oil. **ANA marketing maestros**, v. 3, 2006.

RABAN, Daphne Ruth; GORDON, Avishag; GEIFMAN, Dorit. The information society: The development of a scientific specialty. **Information, Communication & Society**, v. 14, n. 3, p. 375-399, 2011.

ROMERO, Cristobal; VENTURA, Sebastian. Educational data mining: A survey from 1995 to 2005. **Expert systems with applications**, v. 33, n. 1, p. 135-146, 2007.

SOOFI, Aized Amin; AWAN, Arshad. Classification techniques in machine learning: applications and issues. **Journal of Basic & Applied Sciences**, v. 13, p. 459-465, 2017.

SOUZA, Alex Marques de. **Machine learning e a evasão escolar: análise preditiva no suporte à tomada de decisão**. 2020. Tese de Doutorado. Mestrado em Sistemas de Informação e Gestão do Conhecimento.

DE ALMEIDA TEODORO, Leonardo; KAPPEL, Marco André Abud. Aplicaçao de técnicas de aprendizado de máquina para prediçao de risco de evasao escolar em instituiçoes públicas de ensino superior no brasil. **Revista Brasileira de Informática na Educação**, v. 28, p. 838-863, 2020.