

Seleção de Atributos para Classificadores de Evasão Escolar com Dados da Plataforma Nilo Peçanha

Isleimar S. Oliveira¹, Francisco Petrônio A. Medeiros¹, Fabio G. Andrade²

¹ Programa de Pós-Graduação em Tecnologia da Informação (PPGTI) - Instituto Federal de Educação Ciência e Tecnologia de Paraíba (IFPB)
Av. Primeiro de Maio, nº 720, Jaguaribe - CEP: 58.015-435
João Pessoa/PB

² Instituto Federal de Educação Ciência e Tecnologia de Paraíba, Campus (IFPB),
Rua José Antônio da Silva, 300, Bairro Jardim Oásis - CEP: 58.900-000
Cajazeiras/PB

oliveira.isleimar@academico.ifpb.edu.br, {petronio, fabio}@ifpb.edu.br

Abstract. *School dropout is a problem that has affected students, communities, and educational institutions around the world. To Struggle against school dropouts, preventive actions can be carried out aiming at the permanence of these students. For this, it is necessary to identify these students with a potential risk of dropping out. This work presents the results of a study for the application of a predictive algorithm capable of identifying students with the potential to drop out of school, using a dataset available on the Nilo Peçanha Platform for technical and higher education courses. The Random Forest algorithm was used for feature selection and prediction, presenting satisfactory results for predicting school dropouts.*

Resumo. *A evasão escolar é um problema que tem afetado estudantes, comunidades e instituições educacionais em todo o mundo. Para combatê-la, ações preventivas podem ser realizadas visando a permanência desses alunos. Para isso, é necessário identificar os alunos com potencial risco de evasão o mais cedo possível. Este trabalho apresenta os resultados de um estudo para aplicação de um algoritmo preditivo capaz de identificar alunos com potencial de evasão escolar utilizando um conjunto de dados disponível na Plataforma Nilo Peçanha para cursos técnicos e superiores. O algoritmo Random Forest foi utilizado para seleção e predição de características, apresentando resultados satisfatórios para prever a evasão escolar.*

1. Introdução

A educação tem um papel fundamental no processo de transformação da sociedade, propiciando igualdade social, desenvolvimento, melhorando os aspectos moral, intelectual e material tanto para o indivíduo quanto para o ambiente que o cerca [Oliveira et al. 2013]. A alta taxa de evasão escolar é um problema que afeta instituições públicas e privadas de ensino no Brasil e no mundo. Pesquisas realizadas por Siebra et. al (2020) e Pal e Pal (2013) identificaram que causas pessoais e institucionais da evasão escolar contribuem de forma relevante para que alunos

decidam por abandonar seus estudos. Uma forma de combater a evasão escolar e, conseqüentemente, os problemas causados por ela, consiste em identificar de forma antecipada os perfis de alunos com risco potencial de evadir [Romero e Ventura, 2020].

Segundo dados fornecidos pelo Instituto Nacional de Estudo e Pesquisa Educacional Anísio Teixeira (INEP) em sua Sinopse Estatística da Educação Superior, do total de 8.603.824 matrículas de alunos vinculados aos cursos de graduação presencial e a distância no Brasil, excetuando os 1.229 alunos falecidos, um total de 3.744.522 alunos evadiram de seus cursos, o que representa aproximadamente 43,5% dos alunos matriculados. Nos últimos anos, vários estudos têm sido desenvolvidos na área de tecnologia da informação para reconhecer características que contribuem para o problema da evasão escolar [Kaur et al. 2015], tendo a Mineração de Dados (em inglês *Data Mining*, DM) contribuído com resultados promissores no reconhecimento dessas características [Rigo et al., 2017]. A Mineração de Dados tem como objetivo realizar a extração de conhecimento a partir da análise de um conjunto de dados volumosos, mediante técnicas estatísticas e algoritmos que buscam por padrões presentes nos dados. A Mineração de Dados no contexto da Educação é conhecida como Mineração de Dados Educacionais (MDE) (em inglês *Educational Data Mining*) [Rodrigues et al., 2013].

2. Trabalhos Relacionados

A utilização de EDM como forma de prever a possibilidade de evasão escolar tem sido um tema estudado por diversos pesquisadores no Brasil e no mundo. A escolha dos atributos utilizados para prever o risco de evasão variam de acordo com os cursos, as instituições e os ambientes sociais nos quais as pesquisas foram aplicadas. Manhães et al. (2014) utilizaram o algoritmo Naive Bayes para identificar alunos com risco de evasão. O modelo desenvolvido neste trabalho foi treinado e avaliado com base em dados acadêmicos de estudantes da UFRJ e foram restritos aos semestres letivos entre 2003 e 2004. Segundo os autores, o modelo gerado ainda estava em fase de estudos devido às limitações de acesso aos dados e pela falta de referências sobre quais atributos eram mais adequados para analisar o problema da evasão escolar, mas apresentou precisão em torno de 80% sobre a previsão da evasão escolar.

Silva (2019) identificou os alunos que apresentaram risco de evasão escolar por meio de um processo de EDM que agrupou dados gerais, acadêmicos, socioeconômicos, demográficos, programas de auxílio estudantil e participação em projetos de pesquisa e extensão. O autor descobriu padrões e correlações entre os atributos utilizados, sendo levantadas as características que implicam em maior chance de evasão por parte dos alunos. Brito et al. (2015) realizaram um trabalho semelhante a partir de doze anos de dados históricos das notas dos alunos do primeiro período do curso de Bacharelado em Ciência da Computação da UFPB. Nesse estudo, os dados foram analisados utilizando o algoritmo Naive Bayes. e o modelo gerado neste estudo alcançou uma acurácia de até 86,9%. Queiroga et al. (2020) aplicaram técnicas de mineração de dados utilizando Algoritmos Genéticos com o objetivo de identificar alunos em risco de abandono em cursos técnicos à distância do IFRGS. Os resultados preditivos apontaram melhorias significativas em relação ao algoritmo de aprendizado de máquina sem otimização.

Ao analisar esses trabalhos, constatou-se que a utilização de algoritmos de mineração de dados para a predição do risco de evasão escolar é um campo muito promissor [de Andrade et al., 2021]. Entretanto, apesar de relevantes, os trabalhos analisados são voltados para conjuntos de dados muito específicos, o que limita a sua aplicação em outros conjuntos de dados. Para Siebra et. al (2020), apesar de existirem atributos pré-definidos para modelos de classificação de evasão escolar, como idade, gênero, fator socioeconômico, notas em disciplinas específicas, entre outros, não é possível afirmar que a utilização desse conjunto de características garanta sempre os melhores resultados para qualquer conjunto de dados. O trabalho proposto neste artigo oferece um estudo de caso da aplicação das técnicas de seleção de atributos para classificadores de predição da evasão escolar em cursos de graduação e cursos técnicos disponibilizados na Plataforma Nilo Peçanha.

3. Referencial teórico

A Descoberta de Conhecimento em Banco Dados (em inglês *Knowledge Discovery in Databases*, KDD) é em geral utilizada em processos predição, pois tem como propósito extrair informação em grandes conjuntos de dados buscando padrões explicáveis nos mesmos, o que permite a interpretação e extrapolação para eventos futuros [Fayyad et al., 1996]. Para Baker et al. (2011), o processo de KDD extrai conhecimento implícito a partir de conjuntos de dados volumosos ensejando a tomadas de decisões.

Os dados utilizados no KDD são oriundos de diversas fontes, sendo necessário integrar esses dados através de um processo chamado de *Extração, Transformação e Carga* (em inglês *Extract, Transform and Load*, ETL). O ETL é dividido em três etapas básicas. A primeira etapa é a extração dos dados, que pode ocorrer a partir de diversas fontes. A transformação é a segunda etapa, na qual os dados são modificados aplicando técnicas de higienização, padronização, filtragem e dicionário de dados. Por fim, na etapa de carga os dados são persistidos em uma base de dados [Moscoso-Zea e Luján-Mora, 2016]. A principal etapa do processo de KDD é a *Mineração de Dados*, na qual é realizada a análise de grandes quantidades de dados para se obter conhecimento, agregar significado e prover utilidade aos dados [Castro e Ferrari, 2016]. A MDE torna-se cada vez mais reconhecida como uma ferramenta no desenvolvimento de métodos para explorar os dados provenientes do cenário educacional [Romero e Ventura, 2020].

Atualmente, os dados analisados são cada vez mais volumosos e heterogêneos, em virtude das diversas fontes de origem. A solução para analisar esses grandes volumes de dados de forma mais automatizada é através da utilização de técnicas de *Aprendizagem de Máquina* (em inglês *Machine Learning*, ML). As técnicas de ML permitem que programas de computador modifiquem seu comportamento sem a necessidade de serem explicitamente programados, usando como base resultados alcançados anteriormente. Tais comportamento são guiados pela análise amostral dos dados, o que possibilita flexibilidade nas tomadas de decisão sem a necessidade de se seguir regras e instruções predefinidas. Essa maleabilidade permite que as soluções desenvolvidas por meio dessas técnicas sejam constantemente revisadas e adequadas às novas configurações, quando necessário [Géron, 2019].

Muitas vezes, o processo de ML trabalha com conjuntos extremamente dimensionados. Esse número excessivo de dimensões no conjunto de dados leva ao aumento da complexidade, dificultando a interpretação, aumentando o tempo do processo de treinamento e/ou impactando na precisão das previsões [Liu e Motoda, 2007]. Técnicas de seleção de características (em inglês *Feature Selection*) eliminam recursos irrelevantes, reduzindo a complexidade do modelo resultante. Para esses autores, o objetivo final é ter um modelo enxuto que seja mais rápido de processar e não prejudique a qualidade preditiva. Assim, a seleção de recursos não significa necessariamente reduzir o tempo de treinamento, tendo visto que algumas técnicas podem inclusive aumentar o tempo total do treinamento, mas sim reduzir o tempo de execução do modelo.

Para avaliar os modelos de predição é necessário definir alguma métrica de desempenho. Hoje, existem diversas métricas para avaliar o desempenho dos modelos de classificação. Mas, em síntese, um modelo apresenta as melhores medidas de desempenho quando é capaz de produzir uma maior quantidade de resultados corretos [Géron, 2019]. Neste trabalho foram utilizadas as métricas de desempenho matriz de confusão, acurácia, precisão, cobertura, A-ROC e F1.

4. Método da Pesquisa

As etapas do arcabouço metodológico, que serão descritos nesta seção, são métodos de extração, processamento e persistência dos dados, construção das ferramentas para análise dos dados, construção e mensuração do modelo preditivo de evasão escolar.

4.1. Obtenção dos dados para a pesquisa

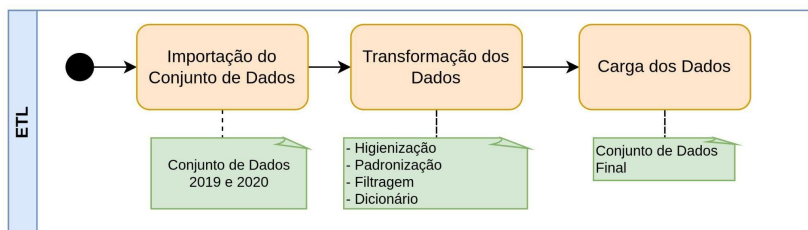
Para execução do projeto foi utilizada a base de dados aberta da Plataforma Nilo Peçanha (PNP) do governo federal, onde são disponibilizados microdados referentes às instituições de ensino que compõem a rede federal. A PNP publica anualmente, desde 2018, microdados sobre as instituições de ensino que compõem os Institutos e as Universidades Federais. Dentre os dados disponibilizados pela plataforma, foram utilizados nesta pesquisa os “*Microdados Matrícula*” que possuem informações relacionadas às matrículas de alunos, cursos e instituições. Para cada aluno, são disponibilizadas informações como sexo, etnia, faixa de renda, idade, tipo de curso, eixo tecnológico, modalidade de ensino, tipo do curso, vagas ofertadas pelo curso, turno, instituição de ensino, município, unidade federativa e situação do aluno no curso. Os microdados disponibilizados em cada ano são referentes à situação dos alunos no ano anterior. Durante a execução deste trabalho, foram usados os microdados da PNP disponibilizados nos anos de 2019 e 2020, visto que os dados de 2018 apresentaram limitações em relação à quantidade de informações disponíveis.

4.2. Implementação da ferramenta ETL

Uma vez definida a origem dos dados, foi possível implementar os processos da ferramenta ETL, responsável por ler os dados obtidos na plataforma PNP, transformá-los para o esquema lógico proposto e carregá-los localmente. Esse processo foi composto por três etapas, ilustradas na Figura 1. A primeira etapa do processo de ETL corresponde à importação do conjunto de dados. Nela, os dados da PNP foram

extraídos, descompactados e armazenados localmente com a utilização de um arquivo de script escrito na linguagem de programação Python. Todos os dados utilizados neste trabalho foram disponibilizados pela plataforma no formato CSV.

Figura 1 - Diagrama de atividades



Os dados importados foram transformados e convertidos para o esquema lógico definido para o armazenamento. Esta é a etapa mais dispendiosa do processo de ETL devido ao grande número de registros existentes em cada arquivo e ao esforço para analisar a integridade dos dados. Na etapa de higienização, os dados foram examinados para identificar inconsistências nos registros, pois durante a análise exploratória dos dados notou-se a existência de algumas imprecisões, como a presença de valores nulos ou fora do alcance razoável para o tipo de atributo. Na etapa de padronização, os registros que indicavam situações iguais, porém com formatos diferentes, foram convertidos para um mesmo formato. Deste modo, para realizar a integração das duas bases de dados fez-se necessário transformar os valores destes campos para um mesmo formato visando à padronização da base de dados final.

Na etapa de filtragem foram aplicados filtros de acordo com as regras de negócio. O campo “categoria_situacao” objetivou selecionar apenas os registros que apresentavam valores “Concluintes” e “Evadidos”, desse modo não foram incluídos os registros cujo campo “categoria_situacao” apresentem o valor “Em Curso”, para fins do treinamento dos algoritmos classificadores. Por último, o filtro do campo “tipo_curso” restringiu os valores a “Bacharelado”, “Tecnologia” e “Licenciatura” para delimitar o foco dos dados aos cursos superiores. Para os cursos técnicos, o valor do campo era “Técnico”. O campo “ano” foi incluído, derivando sua origem do período da base de dados onde foi extraído. Na etapa de criação dos dicionários, campos de valores relacionados que apresentaram nomes diferentes foram modificados para que todas as instâncias pudessem ser acessadas utilizando o mesmo rótulo, desta forma o dicionário foi construído como forma de padronizar os nomes dos atributos para serem utilizados na integração das duas bases de dados selecionadas.

4.3. Análise Preditiva

Na etapa de preditiva dos dados da PNP foi utilizado o algoritmo de classificação Random Forest disponibilizado na biblioteca do scikit-learn¹ versão 1.1.0, da linguagem de programação Python versão 3.8.10. Na seleção de atributos foi utilizado a solução de *Feature Selection* incorporada ao Random Forest.

A escolha do algoritmo Random Forest para a classificação e seleção de atributos ocorreu em virtude dos melhores resultados obtidos, em comparação às médias das métricas entre outros algoritmos identificados na pesquisa bibliográfica. Quando utilizado o conjunto de dados nacionais a média das métricas do Random Forest foi

de 0.89323, do Multilayer Perceptron foi de 0.81690, da Árvore de Decisão foi de 0.80093 e do Naive Bayes foi de 0.67700.

5. Resultados e discussões

Os dados oriundos da Plataforma Nilo Peçanha possibilitam uma análise, em diversas perspectivas, para os dados oriundos dos cursos de instituições de ensino federais no Brasil. A utilização desses dados possibilitou neste trabalho tanto a pesquisa sobre a evasão escolar quanto a aplicação de técnicas de *Feature Selection* para a identificação das características mais relevantes para serem utilizadas em algoritmos de classificação.

5.1. Seleção de atributos

Utilizando o subconjunto de dados extraídos da PNP, foram selecionados os 15 (quinze) principais atributos segundo o algoritmo de seleção de atributos *Random Forest*. Para os dados provenientes dos cursos técnicos no Brasil, foram retornados os seguintes atributos: *carga_horaria*, *codigo_municipio_dv*, *codigo_unidade_ensino_sistec*, *cor_raca*, *data_matricula*, *faixa_etaria*, *fim_ciclo*, *idade*, *inicio_ciclo*, ***instituicao***, *mes_ocorrendia*, *municipio*, ***total_inscritos***, ***unidade_ensino*** e *vagas_ofertadas*. Quando utilizado os dados dos cursos técnicos do IFPB os seguintes atributos foram retornados: *carga_horaria*, *codigo_municipio_dv*, *codigo_unidade_ensino_sistec*, *cor_raca*, *data_matricula*, *faixa_etaria*, *fim_ciclo*, *idade*, *inicio_ciclo*, *mes_ocorrendia*, *municipio*, *nome_curso*, ***renda_familiar***, ***total_inscritos*** e *vagas_ofertadas*.

Os atributos em destaque *instituicao*, *total_inscritos* e *unidade_ensino* foram selecionados quando utilizados os dados nacionais, não possuindo a mesma relevância para o algoritmo quando utilizado nos conjuntos de dados do IFPB. Já os atributos selecionados apenas quando utilizados o conjunto de dados do IFPB foram *nome_curso*, *renda_familiar* e *total_inscritos*. Quando utilizado os dados nacionais dos cursos superiores, os atributos selecionados foram: *carga_horaria*, ***codigo_municipio_dv***, *codigo_unidade_ensino_sistec*, *cor_raca*, *data_matricula*, *faixa_etaria*, *fim_ciclo*, *idade*, *inicio_ciclo*, *mes_ocorrendia*, *municipio*, *renda_familiar*, *total_inscritos*, *unidade_ensino* e *vagas_ofertadas*.

Para os dados dos cursos superiores do IFPB, os atributos selecionados foram: *carga_horaria*, *codigo_unidade_ensino_sistec*, *cor_raca*, *data_matricula*, *faixa_etaria*, *fim_ciclo*, *idade*, *inicio_ciclo*, *mes_ocorrendia*, *municipio*, ***nome_curso***, *renda_familiar*, *total_inscritos*, *unidade_ensino* e *vagas_ofertadas*. Para os dados nacionais, apenas o atributo *codigo_municipio_dv* não está presente entre os atributos selecionados, enquanto para os dados do IFPB, e apenas o atributo *nome_curso* aparece nos dados do IFPB de maneira exclusiva.

5.2. Algoritmo de Classificação

Para a classificação dos alunos, foi utilizado o algoritmo de classificação *Random Forest*, utilizando o atributo *categoria_situacao* como alvo, para identificar os alunos como Concluintes ou Evadidos. A Figura 2 apresenta os resultados da classificação dos cursostécnicos, comparando os resultados entre os dados nacionais e o do IFPB.

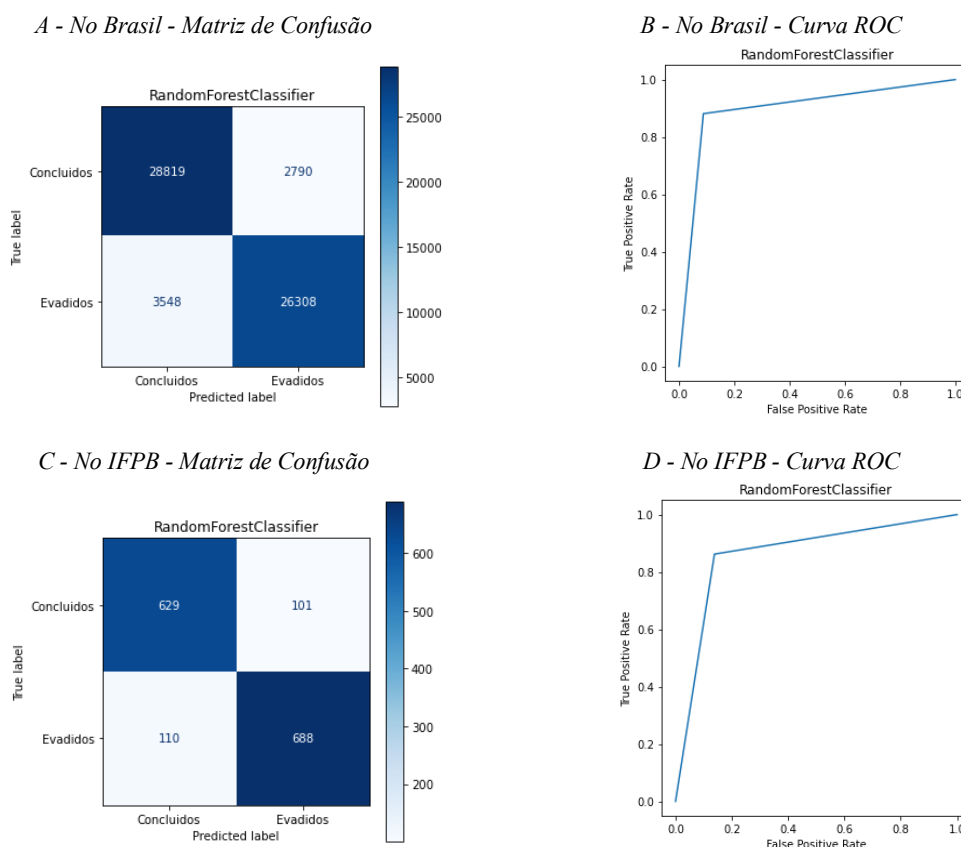


Figura 2 - Resultado da predição dos Cursos Técnicos

A classificação para os conjuntos de dados dos cursos técnicos nacionais e do IFPB são apresentados no quadro 1, onde é possível identificar que tanto para os dados nacionais quanto para os dados do IFPB a precisão apresenta resultados melhores que a cobertura, demonstrando ser mais eficientes na identificação de Falsos Positivos (os casos em que os alunos foram classificados como evadidos, porém não são).

Quadro 1 - Resultados das Métricas dos Cursos Técnicos

Dados	Métricas					
	Acurácia	Precisão	Cobertura	A-ROC	F1	Média
Nacionais	0.896884	0.904117	0.881163	0.896448	0.892492	0.894221
IFPB	0.861911	0.871990	0.862155	0.861900	0.867045	0.865002

A figura 3 apresenta os resultados da classificação dos cursos superiores, comparando os resultados entre os dados nacionais e o do IFPB.

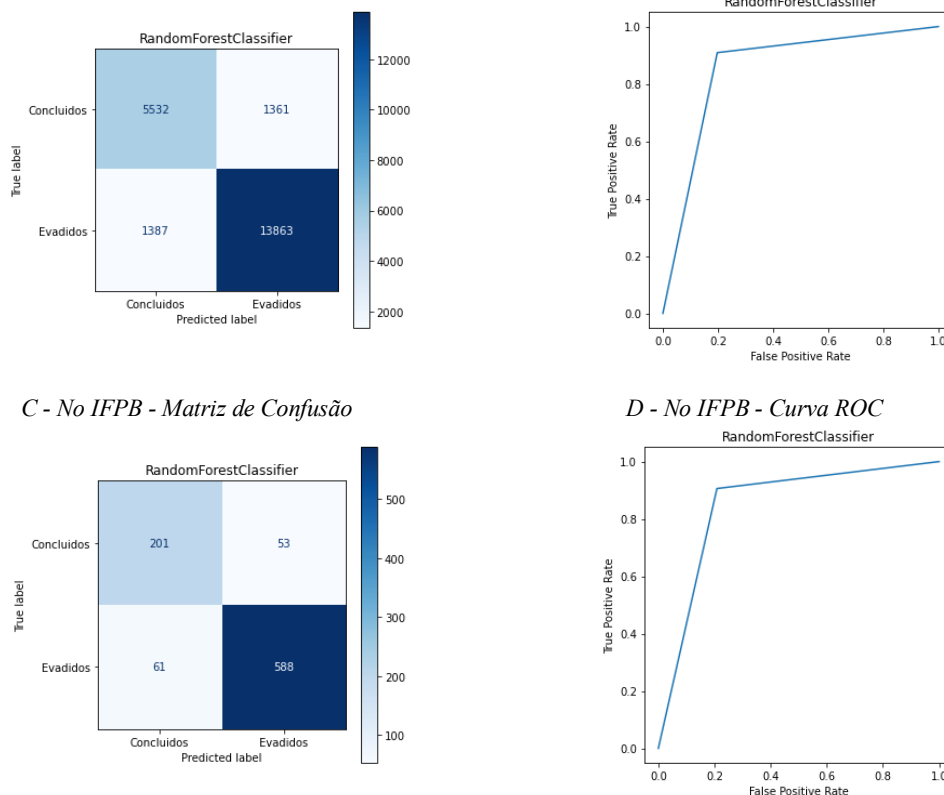


Figura 3 - Resultado da predição dos Cursos Superiores

A classificação para os dados dos cursos superiores nacionais e do IFPB são apresentados no quadro 2. De forma semelhante aos Cursos Técnicos, a precisão dos dados apresentou melhores resultados, tornando a identificação de Falsos Positivos mais efetiva. O valor curva ROC para os Cursos Superiores do IFPB foi o único que apresentou valor abaixo de 0.85, mesmo que com uma diferença inferior a 0.002.

Quadro 2 - Resultados das Métricas dos Cursos Superiores

Dados	Métricas					
	Acurácia	Precisão	Cobertura	A-ROC	F1	Média
Nacionais	0.875898	0.910602	0.909049	0.855801	0.909825	0.892235
IFPB	0.873754	0.917317	0.906009	0.848674	0.911628	0.891476

A análise dos dados demonstra que os atributos selecionados apresentaram relevância na utilização dos algoritmos para classificar os alunos quanto à persistência nos cursos. Os atributos que apresentaram maior relevância na seleção de atributos presentes em todos os grupos foram: carga horária, cor/raça, faixa etária e município. Os resultados da etapa de predição retornaram valores satisfatórios variando entre 0.92 e 0.85, tendo a médias das métricas apresentando valores superiores a 0.86 tanto referente aos dados nacionais quanto aos dados do IFPB para os cursos técnicos e superiores.

6. Considerações finais

Os resultados deste trabalho demonstram que a utilização da MDE e métodos de seleção de atributos na Plataforma Nilo Peçanha permitem identificar atributos relevantes relacionados à evasão escolar. Nesse conjunto de dados, os atributos carga horária, cor/raça, faixa etária e município apresentaram as maiores relevâncias tanto para cursos técnicos quanto para cursos superiores quando se comparou um extrato dos dados nacionais e no contexto da instituição IFPB.

A utilização das características mais relevantes, extraídas da etapa de seleção de atributos, apresentou resultados satisfatórios quando utilizadas para a predição da evasão escolar utilizando o algoritmo de classificação *Random Forest*, com valores médios superiores a 0.86 nas métricas aplicadas, em todos os conjuntos de dados utilizados. Como próximas etapas desta pesquisa e, a partir dos resultados obtidos neste trabalho, é possível expandir a análise para alunos em outros níveis de ensino como o fundamental emédio, realizando comparações entre os atributos selecionados, bem como a aplicação da análise em conjuntos de dados oriundo de sistemas de controle acadêmico.

Referências

- Baker, R.; Isotani, S.; Carvalho, A. Mineração de dados educacionais: Oportunidades para o Brasil. *Revista Brasileira de Informática na Educação*, v. 19, n. 02, p. 03, 2011.
- Brito, D. M. De et al. Identificação de estudantes do primeiro semestre com risco de evasão através de técnicas de Data Mining. *Nuevas Ideas en Informática Educativa TISE*. v. 11, p. 459-463, 2015.
- Castro, Leandro N. de, Ferrari, Daniel G. *Introdução à Mineração de Dados: conceitos básicos, algoritmos e aplicações*. 1ª edição. São Paulo: Saraiva, 2016.
- de Andrade, T. L., de Almeida, C. M. M., Barbosa, J. L. V., & Rigo, S. J. Metodologias Ativas integradas a um Sistema de Recomendação e Mineração de Dados Educacionais para a mitigação de evasão em EaD. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação* (pp. 824-835). SBC, 2021.
- Fayyad, Usama; Piatetsky-Shapiro, Gregory e Smyth, Padhraic. From data mining to knowledge discovery in databases. *AI magazine* 17 , no. 3, 37-54. 1996.
- Moscoso-Zea, O., & Luján-Mora, S. Datawarehouse design for educational data mining. In *2016 15th International Conference on Information Technology Based Higher Education and Training (ITHET)* (pp. 1-6). IEEE, 2016.
- Géron, Aurélien. *Mãos à Obra Aprendizado de Máquina com Scikit-Learn & TensorFlow: Conceitos, Ferramentas e Técnicas Para a Construção de Sistemas Inteligentes*. Rio de Janeiro. Alta Books, 2019.
- Grus, Joel. *Data Science do zero*. Traduzido por Welington Nascimento. - Rio de Janeiro:Alta Books, 2016.
- INEP, Sinopse Estatística da Educação Superior. INEP - Instituto Nacional de Estudo e Pesquisa Educacional Anísio Teixeira. Disponível em <<https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados>>.

Acesso em: 22 nov. 2020.

- Kaur, P.; Singh, M.; Josan, G. S. Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector. *Procedia Computer Science*, v. 57, p. 500–508, 2015. ISSN 18770509. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1877050915019018?via%3Dihub>>. Acesso em: 17 nov. 2021.
- Liu, Huan; Motoda, Hiroshi. *Computational Methods of Feature Selection*. Chapman & HALL/CRC. 2007.
- Manhães, L. M. B.; Cruz, S. M. S.; Zimbrao, G. Wave: an architecture for predicting dropout in undergraduate courses using EDM .In: *Proceedings of the 29th annual acmsymposium on applied computing*. v. 1, p. 243-247, 2014.
- Oliveira, Terezinha; VIANA, Ana Paula dos Santos; BOVETO, Lais; SARACHE, Mariana Vieira. Escola, Conhecimento e Formação de Pessoas: Considerações Históricas. 2013. *Políticas Educativas*, Porto Alegre, v. 6, n. 2, p. 145-160.
- Pal, A., & Pal, S. (2013). Data Mining Techniques in EDM for Predicting the Performance of Students. *International Journal of Computer and Information Technology*, 2(6), 1110–1116.
- Rigo, F; Cazella, S. Identificando o perfil de evasão de alunos de graduação através da mineração de dados educacionais: um estudo de caso de uma universidade comunitária, VI Congresso Brasileiro de Informática na Educação, Nov. 2017.
- Rodrigues, R. L., De Medeiros, F. P., & Gomes, A. S. Modelo de Regressão Linear aplicado à previsão de desempenho de estudantes em ambiente de aprendizagem. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)* (Vol. 24, No. 1, p. 607), 2013.
- Romero, C., & Ventura, S. Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and KnowledgeDiscovery*, 10(3), e1355, 2020.
- Siebra, Claurton Albuquerque. Santos, Ramon N. Lino, Natasha C. Q. A Self-Adjusting Approach for Temporal Dropout Prediction of E-Learning Students. 2020. *International Journal of Distance Education Technologies*. v. 18, Issue 2.
- Silva, E. M. C, Um Modelo Descritivo para Auxiliar o Acompanhamento da Evasão Escolar nos Cursos Técnicos e Superiores no Instituto Federal do Rio Grande do Norte – Campus São Gonçalo do Amarante. Natal/RN:Universidade Federal do Rio Grandedo Norte, 2019.