

Evasão no Ensino Básico da Rede Pública Municipal de Juiz de Fora: uma Análise com Mineração de Dados*

**Frederico de Oliveira Sales¹, Yan Mendes Ferreira¹, Francisco H C Ferreira¹,
Bruno José Dembogurski², Gustavo Silva Semaan³, Edelberto Franco Silva¹**

¹Laboratório NetLab – PPGCC – Departamento de Ciência da Computação
Universidade Federal de Juiz de Fora (UFJF) - Juiz de Fora - MG

²Universidade Federal Rural do Rio de Janeiro (UFRRJ) - Nova Iguaçu - RJ

³Instituto do Noroeste Fluminense de Educação Superior (INFES)
Universidade Federal Fluminense (UFF) - Santo Antônio de Pádua - RJ

***Abstract.** Inep's data shows that only 53,8% of students who join Elementary school concludes it. These kids are usually socially disadvantaged, therefore they end up increasing the city bills. This work is the first research which analyses the drop out in public primary education in Brazilian southeast based on predictive computational methods. The research accomplished shows the efficiency in using machine learning algorithms to school evasion prediction during 2017 in Juiz de Fora city. In these case, approximately 1448 students from the public education left their schools. The proposed method is applied using real data and presenting the most relevant attributes. It also achieves 76% accuracy and covers 97% in the best case to identify the students who tend to drop out.*

***Resumo.** Dados do INEP revelam que apenas 53,8% dos ingressantes concluem o Ensino Fundamental no Brasil. Trata-se de um relevante prejuízo aos cofres públicos e também um problema para a sociedade, uma vez que o jovem torna-se socioeconomicamente mais vulnerável, e o problema direta ou indiretamente afeta também setores públicos como segurança e saúde. Este trabalho é a primeira pesquisa para análise da evasão no ensino básico da rede pública do Sudeste do Brasil de acordo com a revisão sistemática da literatura. Utilizando algoritmos de aprendizado de máquina para a previsão da evasão escolar no ano letivo de 2017 na cidade de Juiz de Fora/MG, onde cerca de 1448 alunos da rede pública deixaram a sala de aula, o trabalho destaca os atributos de maior relevância, alcança uma precisão de 76% e uma cobertura de 98% no melhor caso de identificação daqueles alunos mais propensos a evadir.*

1. Introdução

Dados do INEP revelam que apenas 53,8% dos ingressantes concluem o Ensino Fundamental (EF) no Brasil (UNESCO 2015). Isso representa um prejuízo aos cofres públicos de cerca de R\$100 Bilhões e atinge de maneira direta e indireta a sociedade. A evasão escolar ocorre por diversas causas, desde aquelas diretamente relacionadas à instituição, como a qualidade do ensino e o ambiente escolar a vida social, como questões familiares e/ou meio em que o(a) aluno(a) vive. Em (Neri et al. 2015) são apresentados alguns

*Contato: edelberto@ice.ufjf.br

motivos da evasão e do abandono diretamente relacionados a problemas na vida social do aluno, como a gravidez na adolescência, a necessidade de trabalhar para auxiliar na renda familiar ou a idade avançada decorrente de reprovações (desmotivação do aluno). Destaca-se, entretanto, que abandono e evasão, embora correlacionados, são conceitos diferentes. No abandono o(a) aluno(a) deixa de frequentar as aulas, o que compromete o ano letivo, mas retorna em um momento posterior. Já na evasão o(a) aluno(a) deixa de frequentar as aulas e não retorna para dar continuidade (Neri et al. 2015).

Pesquisas realizadas para identificar e mensurar a evasão escolar são, na maioria das vezes, baseadas em regiões específicas devido a características particulares. Assim, além da importância do tema evasão que possui impacto direto na sociedade, a Prefeitura Municipal de Juiz de Fora, no estado de Minas Gerais, tem como demanda identificar e classificar o perfil do(a) aluno(a) evasor(a). Nesse sentido, com base em suas características, por meio do processo de descoberta de conhecimento em bases de dados e de Inteligência Computacional, deve-se selecionar os candidatos mais propensos a evadir e, assim, apresentar soluções para minimizar os casos.

O objetivo geral do trabalho é abordar o problema de evasão escolar no ensino fundamental do município de Juiz de Fora, e propor soluções que apoiem a prefeitura em tomadas de decisão. Nesse sentido, como objetivos específicos foram considerados: (i) Apresentar uma contextualização com Estatísticas Iniciais e dados dos alunos do município; (ii) Realizar e apresentar uma revisão sistemática da literatura em relação aos termos *evasão* e *abandono*. (iii) Introduzir o tema Descoberta de Conhecimento em Bases de Dados e detalhar as etapas operacionais e atividades realizadas no presente trabalho, sobretudo a Classificação de Dados na etapa Mineração de Dados;

Este trabalho está organizado da seguinte maneira: a seção 2 apresenta dados recentes da rede municipal de ensino de Juiz de Fora; na seção 3 são apresentados os principais trabalhos relacionados e uma breve revisão da literatura; já na seção 4 apresenta proposta que aborda o processo de conhecimento de bases de dados (KDD, do inglês *Knowledge Discovery in Databases*) e suas etapas operacionais de pré-processamento, mineração de dados e pós-processamento; na seção 5 são apresentados os resultados obtidos e, por fim, a seção 6 apresenta as conclusões e trabalhos futuros.

2. Contextualização

Como uma visão geral da Rede Municipal de Ensino Básico do município de Juiz de Fora¹, esta seção apresenta, de maneira objetiva, dados importantes para compreender os resultados desta pesquisa.

A Figura 1 apresenta o gráfico Alunos por ciclo no EF, em que é possível identificar uma maior concentração nos dois primeiros anos deste ciclo, havendo uma queda de aproximadamente 3 mil alunos entre o primeiro e o quarto ano. Destaca-se ainda a queda a partir do quinto ano, com redução total de quase 300% com relação ao primeiro ano.

O gráfico da Figura 2 apresenta o Quantitativo de alunos por turno, em que pode-se observar a concentração de alunos nos turnos da manhã (18.787 alunos) e da tarde (17.483 alunos). Assim, o turno da noite e o regime integral, juntos, correspondem a

¹Mais informações sobre a cidade de Juiz de Fora podem ser encontrados em: <https://www.pjf.mg.gov.br/cidade/>

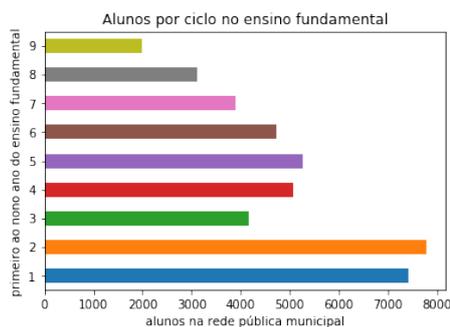


Figura 1. Alunos por ciclo (EF).

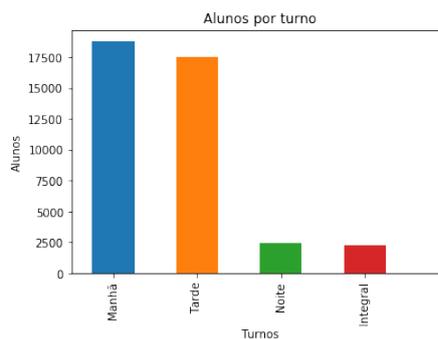


Figura 2. Quantitativo de alunos por turno.

cerca de 17% do total de alunos (7.402 alunos).

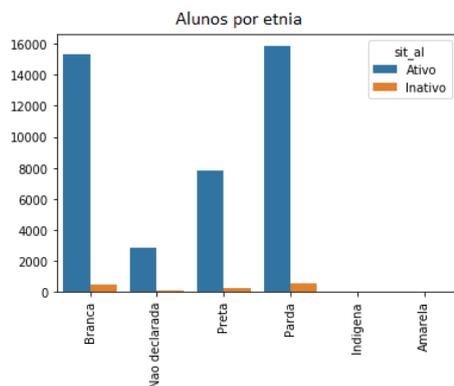


Figura 3. Alunos ativos e inativos por etnia.

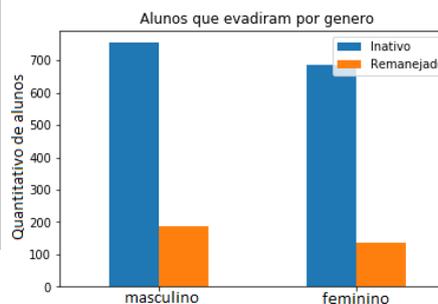


Figura 4. Alunos ativos e inativos por gênero.

A Figura 3 mostra o Gráfico Alunos por Etnia, em que a maior quantidade de alunos ativos está concentrada nas etnias declaradas branca e parda. Entretanto, o índice de evasão da etnia negra é proporcionalmente maior. Já o gráfico da Figura 4, Evasão por Gênero, observa-se que a evasão em ambos os gêneros é proporcional. Com base nos dados apresentados, busca-se a formação de padrões para ser possível classificar (identificar) candidatos a evasão. Para isso será considerado e apresentado na seção 4 o processo de KDD e, em especial, a tarefa Classificação na etapa de mineração de dados.

3. Trabalhos Relacionados

O presente trabalho não tem como objetivo realizar uma Revisão Sistemática da Literatura, embora seja de grande importância mapear trabalhos relacionados em acervos consolidados e consagrados.

Conforme relatado na seção 1, embora correlacionados, *abandono* e *evasão* são conceitos diferentes. Além disso, para a realização de pesquisas relacionadas a evasão escolar, é de fundamental importância considerar características regionais. Assim, na pesquisa do presente trabalho optou-se por desconsiderar importantes bases como *ACM Digital Library*, *IEEE Xplore Digital Library* e *ScienceDirect*, e foi considerado apenas

o acervo com as publicações dos últimos 18 anos do SBIE (Simpósio Brasileiro de Informática na Educação), o principal evento científico da área Informática na Educação do Brasil, que conta com 2.147 artigos publicados por mais de 4.200 (co)autores, e está disponível de maneira simples e organizada em seu portal online.

Uma pesquisa exploratória realizada em fevereiro de 2019 buscou nos títulos dos trabalhos dos anais do SBIE os termos *evasão* e *abandono*. Os critérios adotados para a revisão foram: analisar as publicações dos últimos 10 anos, no período de 2009 a 2018, escritos em português com ao menos um dos termos no título. A busca resultou em apenas 13 artigos com termo *evasão*, e nenhum com o termo *abandono*. Além disso, 12 dos 13 artigos localizados foram publicados nos últimos 4 anos, no período entre 2015 e 2018, o que indica que o tema além de importante está em debate pela comunidade. Os 13 artigos selecionados possuem 50 (co)autores distintos, com apenas 5 autores com participação em dois trabalhos e nenhum autor atuou em três ou mais trabalhos com os termos pesquisados.

Sendo assim, existem diversos trabalhos relacionados à evasão escolar que consideram métodos computacionais, porém destaca-se que nenhum atuou em unidades de ensino da rede pública municipal de educação de um estado da região sudeste. Nesse sentido, devido a relevância das características regionais na investigação, é de grande importância destacar que a pesquisa é inovadora no âmbito nacional.

A presente seção relata oito trabalhos selecionados referentes a publicações mais recentes (2017 e 2018), em que quatro trabalhos tratam da evasão em ambientes de educação à distância, três trabalho relacionado à evasão em cursos superiores presenciais e um trabalho que trata da evasão em escolas da educação básica, que mais se assemelha a presente pesquisa.

Um dos desafios das instituições de ensino é reduzir os altos índices de evasão em seus cursos superiores. Em (Ieza Damasceno 2018) e (Máriele Lanes 2018) foram considerados dados reais obtidos dos sistemas acadêmicos de suas instituições para tratar o assunto. Em (Ieza Damasceno 2018) foi apresentada uma metodologia para estabelecer e analisar o panorama da evasão de discentes em cursos de computação da Universidade Federal de Uberlândia (UFU). A busca por fatores que podem estar relacionados ocorre por meio de uso de medidas de evasão. (Máriele Lanes 2018) considerou mineração de dados educacionais, com o uso do algoritmo de classificação J48, para identificar estudantes que apresentam risco de evasão a partir do seu primeiro ano no curso de graduação.

(Digiampietri et al. 2016) apresenta um estudo acerca da evasão escolar do curso de graduação em Sistemas de Informação da Universidade de São Paulo (USP). A pesquisa teve o objetivo de classificar os discentes com risco de desistência, em que o abandono é caracterizado pelo bacharel que não conclui o curso. De acordo com o trabalho, a base de dados considerou apenas variáveis relacionadas às notas obtidas em toda a graduação, desprezando variáveis como características dos docentes, estrutura da universidade e variáveis pessoais relacionadas aos discentes. A classificação é realizada pelo algoritmo *Rotation Forest*.

Em (Queiroga et al. 2017) é utilizada apenas a contagem de interações dos estudantes dentro do AVA, além de atributos derivados dessas contagens. A premissa inicial é que essa estratégia permite uma maior generalização em diferentes plataformas e AVA,

uma vez que não utiliza diferenciações entre as maneiras de interação e desconsidera informações de outra ordem encontradas fora do AVA (dados demográficos, exames, questionários, etc). Modelos de predição foram testados e treinados com dados de 4 diferentes cursos técnicos EAD em dois cenários diferentes.

Já em (Ramos et al. 2017) é aplicado o uso dos construtos da Teoria da Distância Transacional como preditores da evasão de alunos na EAD. Foi utilizado um conjunto de variáveis representativas desses construtos, obtido de uma base de dados com dois cursos superiores ofertados por EAD, por meio de um processo de análise multivariada de dados. Em seguida, foram utilizados algoritmos de aprendizagem supervisionadas para definição de modelos preditivos da evasão de alunos, sendo o modelo por regressão logística o que apresentou melhores resultados, com índices de acerto superiores a 89%.

O trabalho (Azevedo et al. 2017) propõe uma técnica de análise de sentimento que auxilie o professor na detecção da motivação do aluno a partir de postagens de fóruns educacionais, prevendo um possível caso de evasão. Os experimentos mostram que a proposta alcançou uma taxa de acerto de 82% na classificação da motivação dos estudantes. Já em (Jorge Luis Cavalcanti Ramos 2018) apresenta comparativo de classificadores na previsão da evasão de alunos em EAD, em que com o uso de mineração e aprendizagem de máquina em contextos educacionais busca-se refinar a predição de alunos com risco de evasão.

Por fim, em (Calixto et al. 2017) foi realizado um estudo para identificar as variáveis relativas à evasão escolar na educação básica nos Estados do Ceará e Sergipe com base nos dados do INEP de 2014 a 2016. Foi considerada a metodologia CRISP-DM para entender, preparar e modelar os dados e técnicas de indução de regras e regressão logística apresentaram alta acurácia. As principais variáveis identificadas: a idade, etapa de ensino, modalidade de ensino, existência de laboratórios e localização da escola. Trata-se do trabalho de maior semelhança à pesquisa do presente trabalho.

4. Processo de KDD

Os temas Descoberta de Conhecimento e Bases de Dados e Mineração de Dados embora estejam mais relacionados a áreas da Ciência da Computação, como Banco de Dados, Inteligência Computacional e Aprendizado de Máquina, já são bem conhecidos e abordados em áreas como Informática na Educação. No acervo de anais do SBIE, por exemplo, ao pesquisar a palavra *mineração* retornam 63 artigos publicados.

O processo de KDD foi definido por (Fayyad 1996) como “*processo não trivial, formado por várias etapas, interativo e iterativo, para identificação de padrões válidos, novos, potencialmente úteis e compreensíveis, em grandes conjuntos de dados*”. Ele é considerado interativo por necessitar de especialistas em sua condução, e iterativo por ser possível retornar ou avançar etapas. Basicamente, as etapas operacionais do processo de KDD são organizadas em três grandes grupos: pré-processamento, mineração de dados e pós-processamento. Dentre as tarefas de mineração de dados existentes na literatura, a Classificação consiste em associar objetos (registros como alunos) a uma classe (como candidato (ou não) a evasão), tendo por base uma coleção de classes predefinidas (um histórico ou conjunto de dados de treinamento)(Goldschmidt 2005)(Tan 2009).

Nesta seção é apresentado o processo de KDD para a solução do problema abordado. Para isso serão apresentados os passos do pré-processamento com o tratamento de

dados na subseção 4.1 e etapa de mineração de dados na subseção 4.2, em que é detalhado o modelo matemático e parâmetros do classificador, bem como a justificativa de sua escolha. Os resultados obtidos serão apresentados na seção 5.

4.1. Pré-processamento

O conjunto de dados reais disponibilizados pela Prefeitura de Juiz de Fora possui 43.672 instâncias, correspondente aos dados dos alunos nos anos de 2017 e 2018, com cerca de 96% de alunos ativos. Destaca-se que dados como nome e filiação não foram fornecidos devido a questões de privacidade. Entretanto, pelo mesmo motivo, dados que os autores acreditam ser de grande potencial (ou decisivos) para o estudo, como o indicativo sócio-econômico, as notas e as frequências dos alunos também não foram disponibilizados. Os dados considerados no trabalho foram:

- *Nível*: nível de ensino ao qual o aluno se encontra (*e.g.* fundamental, básico);
- *Turno*: manhã, tarde, noite ou regime integral;
- *Etnia*: a etnia declarada pelo aluno (branca, parda, preta, indígena, amarela ou não declarada);
- *Repetente*: se o aluno é repetente;
- *Responsável*: se o aluno possui responsável. Em geral um parente responde pelo aluno, mas há casos em que é um vizinho ou o conselho tutelar;
- *Necessidade Especial*: se o aluno possui algum tipo de necessidade especial;
- *Transporte*: se o aluno necessita de transporte público entre sua casa e a escola;
- *Genero*: masculino ou feminino;
- *situação atual*: indica se o aluno está ativo na rede pública, se mudou de escola ou evadiu.

Conforme citado na seção 1, a presente pesquisa surgiu com base em uma demanda existente na Prefeitura de Juiz de Fora. Em uma pesquisa realizada no início de 2018 com os diretores de escolas municipais, cerca de 93% deles afirmaram ter interesse em uma solução para identificação de alunos candidatos a evasão.

O primeiro da etapa de pré-processamento para o tratamento dos dados foi a realização de reduções nos dados. A redução horizontal consiste em filtrar registros conforme os critérios estabelecidos. No presente trabalho foram ignorados registros que não possuem os dados necessários. A redução vertical consiste em remover os atributos indesejados. Em outras palavras, trata-se da redução da dimensionalidade do problema. Em geral, são removidos os atributos que não contribuem no processo, como a matrícula. Uma vez que a maioria dos atributos considerados são categóricas, a aplicação de técnicas como matriz de correlação para melhor refinamento da dimensão torna-se inviável.

4.2. Mineração de Dados

A mineração de dados é a etapa do KDD que consiste em aplicar algoritmos para análise e descoberta de padrões ou modelos de dados. Conforme o objetivo a ser alcançado, deve-se utilizar a tarefa mais adequada, e diversos algoritmos e técnicas podem ser considerados, como redes neurais, classificadores baseados em regras e algoritmos que atuam em árvores de decisão. Como foco do presente trabalho, a classificação de dados consiste em identificar um objeto em uma classe com base em um conjunto de objetos previamente classificados [KUMAR].

A arquitetura de classificação deste trabalho foi a Floresta Randômica Ponderada (Gislason et al. 2006). Essa escolha foi feita por permitir um grau de interpretabilidade maior do que outras abordagens, como redes neurais ou máquinas de vetor suporte, já que as decisões tomadas para a classificação podem ser recuperadas. Além disso, evita o enviesamento por ter um considerável desbalanceamento das classes, uma vez que a abordagem considera a frequência de cada classe no processo de treinamento (Chen et al. 2004). Floresta Randômica Ponderada (do inglês *Weighted Random Forest* - WRF) é uma floresta de Árvores de Decisão (do inglês *Decision Tree* - DT), e uma DT é uma árvore que classifica dados, com base em condições binárias (verdadeiras ou falsas) sobre as variáveis avaliadas. Nesta representação, as arestas representam os caminhos tomados mediante as indagações feitas nos nós, como visto na Figura 4.2.

Ainda com base na Árvore de Decisão da Figura 5, A, B, C e D são classes, e X, Y, Z e W são atributos. Assim, um registro que tem os valores $X = 11$ e $Y = -3$, independente do valor do atributo Z, é alocado a classe B. A Figura 6 ilustra o processo de classificação de um registro que possui os valores $X = 8$, $Y = -1$ e $Z = 4$, alocado a classe D.

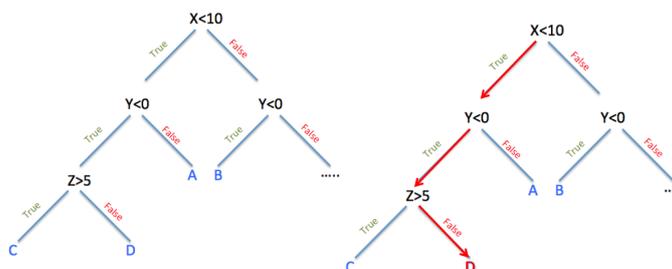


Figura 5. Exemplo de árvore de decisão.

Figura 6. Classificação.

No contexto de aprendizado de máquina, o processo de construção de uma DT consiste nas escolhas das divisões que melhor se ajuste aos dados e classifique-os de maneira mais precisas. No presente trabalho, essa divisão é avaliada através do coeficiente de Gini. Já uma WRF trata-se de um conjunto de DTs, e seu objetivo é construir distintas árvores para classificar um conjunto de dados, realizando uma votação entre estes diferentes classificadores para eleger classe mais popular (Breiman 2001).

4.2.1. Treinamento, Teste e Avaliação

O conjunto de dados foi fracionado em 66% para compor uma base de treinamento e os restantes 34% para a base de teste. Por se tratar de uma massa de dados altamente desbalanceada, a fim de garantir a representatividade da classe minoritária (alunos que evadiram) no conjunto de testes, o processo de separação se deu de forma estratificada, *i.e.* a mesma proporção de desbalanceamento é encontrada nos dois conjuntos. Além disso, por se tratar de uma abordagem estocástica, para cada configuração de parâmetros avaliada, 30 modelos distintos foram considerados para o mesmo conjunto (de treinamento e teste).

Como métricas de avaliação foram utilizados Precisão, Cobertura e Medida-f (média harmônica entre precisão e cobertura), definidas nas Equações 1, 2 e 3 respectiva-

mente. Cabe ressaltar que todos os resultados expostos na seção 5 são exclusivos da classe minoritária, *i.e.* a precisão e a cobertura de classificação de um aluno propenso à evasão.

$$precisao = \frac{|\{\text{documentos_relevantes}\} \cap |\{\text{documentos_recuperados}\}|}{|\{\text{documentos_recuperados}\}|} \quad (1)$$

$$cobertura = \frac{|\{\text{documentos_relevantes}\} \cup |\{\text{documentos_recuperados}\}|}{|\{\text{documentos_relevantes}\}|} \quad (2)$$

$$F_{\beta} = (1 + \beta^2) \cdot \frac{precisao \cdot cobertura}{\beta^2 \cdot precisao + cobertura} \quad (3)$$

A precisão exprime a quantidade de classificações que foram feitas corretamente, e verifica entre os "documentos" recuperados quantos são relevantes (estão corretos). A cobertura (também conhecida como *recall*) visa identificar a proporção entre os dados que foram classificados corretamente e o conjunto de todos os dados que deveriam ter sido classificados, *i.e.* de todos os alunos que evadiram, quantos deles foram identificados pelo classificador de forma correta.

Pelo domínio em questão, a precisão não é a principal métrica de avaliação, apesar de ter sua importância. Nesse sentido, a cobertura possibilita ao gestor de educação ter acesso a uma informação mais abrangente sobre quais alunos são potenciais evasores. Já a Equação 3 combina a precisão e a cobertura e, a partir da utilização do parâmetro β , é possível calibrar a relação de importância na média entre as duas métricas. Assim, este trabalho atribui $\beta = 2$ favorecendo a cobertura em relação à precisão no momento da avaliação.

5. Resultados

Nesta seção são apresentados os resultados para a execução da proposta apresentada na seção 4. A partir dos dados de mais de 40 mil alunos da rede pública de ensino básico do município da cidade de Juiz de Fora/MG. Para a execução do modelo proposto sobre os dados foi utilizada a implementação dos métodos disponível pelo *scikit-learn* (*Machine Learning in Python*) (Pedregosa et al. 2011).

Em um primeiro momento uma análise foi realizada com o objetivo de identificar os atributos mais relevantes do aluno em relação a classificá-lo como candidato a evasão. A Figura 7 apresenta um gráfico com a relevância de todos os atributos considerados no presente trabalho. É possível verificar que os três atributos mais importantes conforme a classificação são: a etnia do aluno, se o aluno possui responsável e o turno em que ele estuda.

Para o cálculo da precisão e da cobertura, o modelo proposto foi submetido ao algoritmo cinco vezes considerando a mesma base de dados. Na Tabela 1 são apresentados os resultados para cada uma das execuções realizadas, em que a melhor precisão foi de 76% de precisão e a melhor cobertura obtida foi de 98%. Além disso, em média, os resultados de precisão e cobertura ficam em 70% e 97%, respectivamente, e comprovam que o uso da abordagem proposta é satisfatória e apresenta-se como um caminho promissor para a resolução de problemas semelhantes.

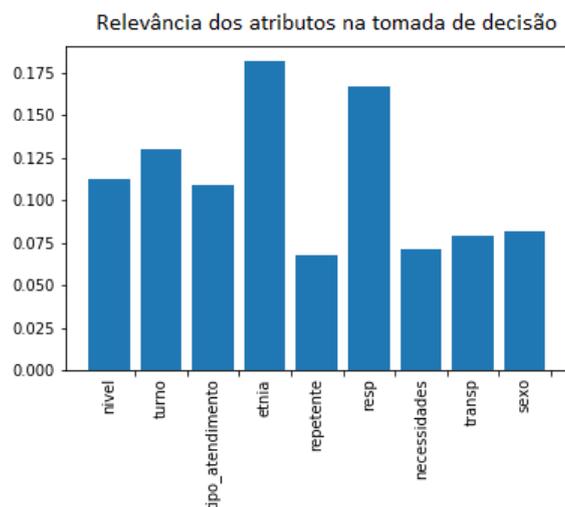


Figura 7. Relevância dos atributos na tomada de decisão.

Execução	Precisão	Cobertura	F_{β}
#1	0.76	0.97	0.85
#2	0.71	0.97	0.82
#3	0.65	0.97	0.78
#4	0.71	0.97	0.82
#5	0.71	0.98	0.85
Média	0.70	0.97	0.82

Tabela 1. Resultados sobre precisão e cobertura.

6. Conclusões

Esse trabalho apresentou, com base na revisão sistemática da literatura apresentada, a primeira pesquisa para análise da evasão no ensino básico da rede pública do Sudeste do Brasil. Para a identificação de alunos candidatos a evasão foram considerados o processo de KDD e métodos computacionais preditivos. A pesquisa apresentada neste estudo mostrou a eficácia do uso de algoritmos de aprendizado de máquina para a previsão da evasão escolar no ano letivo de 2017 na cidade de Juiz de Fora, onde cerca de 1448 alunos da rede pública abandonaram as salas de aula. O modelo de análise preditiva proposto foi aplicado em dados reais e, apesar das limitações devido a questões de privacidade, foram obtidos valores satisfatórios de precisão (76%) e de cobertura (98%). Além disso, em uma análise preliminar os atributos mais relevantes foram mensurados e destacados.

Como trabalhos futuros deve-se incluir outros atributos relevantes, como o Índice de Desenvolvimento para a Educação Básica (IDEB), índices sociais e de criminalidade da região bem como dados escolares como notas e frequências. Pode-se, também comparar do modelo proposto com outros modelos de classificação existentes na literatura.

Referências

- [Azevedo et al. 2017] Azevedo, D., Ferreira, R., Mendonca, V., and Miranda, P. (2017). Aplicação de análise de sentimento em fóruns educacionais para prevenir evasão. In *SBIE*, volume 28, page 1097.

- [Breiman 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- [Calixto et al. 2017] Calixto, K., Segundo, C., and de Gusmão, R. P. (2017). Mineração de dados aplicada a educação: um estudo comparativo acerca das características que influenciam a evasão escolar. In *SBIE*, volume 28, page 1447.
- [Chen et al. 2004] Chen, C., Liaw, A., and Breiman, L. (2004). Using random forest to learn imbalanced data. *University of California, Berkeley*, 110:1–12.
- [Digiampietri et al. 2016] Digiampietri, L. A., Nakano, F., and de Souza Lauretto, M. (2016). Mineração de dados para identificação de alunos com alto risco de evasão: Um estudo de caso. *Revista de Graduação USP*, 1(1):17–23.
- [Fayyad 1996] Fayyad, U. M.; Piatetsky–shapiro, G. S. P. (1996). Knowledge discovery and data mining: Towards a unifying framework. In *International Conference on Knowledge Discovery and Data Mining (KDD-96)*.
- [Gislason et al. 2006] Gislason, P. O., Benediktsson, J. A., and Sveinsson, J. R. (2006). Random forests for land cover classification. *Pattern Recognition Letters*, 27(4):294–300.
- [Goldschmidt 2005] Goldschmidt, R.; Passos, E. (2005). *Data Mining: Um Guia Prático*. Elsevier.
- [Ieza Damasceno 2018] Ieza Damasceno, M. C. (2018). Panorama da evasão no curso de sistemas de informação da universidade federal de Uberlândia: Um estudo preliminar. In *SBIE*.
- [Jorge Luis Cavalcanti Ramos 2018] Jorge Luis Cavalcanti Ramos, João Silva, L. P. A. G. R. R. (2018). Um estudo comparativo de classificadores na previsão da evasão de alunos em ead. In *SBIE*.
- [Máriele Lanes 2018] Máriele Lanes, C. A. (2018). Predição de alunos com risco de evasão: estudo de caso usando mineração de dados. In *SBIE*.
- [Neri et al. 2015] Neri, M. et al. (2015). Motivos da evasão escolar.
- [Pedregosa et al. 2011] 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.
- [Queiroga et al. 2017] Queiroga, E., Cechinel, C., and Araújo, R. (2017). Predição de estudantes com risco de evasão em cursos técnicos a distância. In *SBIE*, volume 28, page 1547.
- [Ramos et al. 2017] Ramos, J. L. C., Gomes, A. S., Rodrigues, R., Silva, J., de Souza, F. d. F., de Gouveia Zambom, E., and Prado, L. (2017). Um modelo preditivo da evasão dos alunos na ead a partir dos construtos da teoria da distância transacional. In *SBIE*, volume 28, page 1227.
- [Tan 2009] Tan, P.; Steinbach, M. K. V. (2009). *Introdução ao data mining, Mineração de Dados*. Ciência Moderna.
- [UNESCO 2015] UNESCO (2015). Relatório educação para todos no Brasil.