

# Capítulo 21

MINERAÇÃO DE DADOS PARA DETECTAR EVASÃO ESCOLAR UTILIZANDO ALGORITMOS DE CLASSIFICAÇÃO: UM ESTUDO DE CASO Luciano Bruno Gomes de Medeiros Thereza Patrícia Pereira Padilha DOI 10.22533/at.ed.75319180421

- [RESUMO | ABSTRACT](#)
- [1 | INTRODUÇÃO](#)
- [2 | MINERAÇÃO DE DADOS EDUCACIONAIS](#)
- [3 | TAREFA DE CLASSIFICAÇÃO](#)
- [4 | ESTUDO DE CASO](#)
- [5 | CONSIDERAÇÕES FINAIS](#)
- [REFERÊNCIAS](#)

# RESUMO | ABSTRACT

**RESUMO:** A tecnologia da informação está transformando o mundo de forma muito veloz com a inserção da internet, smartphones, tablets e diversos dispositivos conectados, gerando e armazenando diversos tipos de dados. Na área educacional, há uma infinidade de dados (públicos ou privados) que podem ser explorados para ajudar em processos de tomada de decisão, tais como: notas nas disciplinas, frequência, e disciplinas cursadas, trancadas, reprovadas e evadidas. Diante deste contexto, este trabalho apresenta os resultados obtidos da mineração de dados de uma pesquisa de campo com alunos de uma escola X da rede pública do estado da Paraíba, utilizando algoritmos de classificação da ferramenta Weka com o objetivo de detectar perfis de alunos evadidos para suporte a campanhas e políticas de evasão escolar. Diversos fatores foram identificados para evasão escolar observando diferentes faixas etárias, tais como: trabalho (de 16 a 20 anos) e gravidez (de 21 a 25 anos), por exemplo.

**ABSTRACT:** Information technology is changing the world very fast with the insertion of the internet, smartphones, tablets and several connected devices, generating and storing various data types. In the educational area, there are several data (public or private) that can be explored to assist in decision-making processes, grades obtained in the classes, attendance, and locked, disapproved, and dropout classes. In this context, this paper presents the results from a data mining with students of the public school of the State X using classification algorithms, from Weka tool, with the goal to detect student profiles for support campaigns and school dropout rates policies. Some factors were identified for school dropout for different age groups, such as work (from 16 to 20 years) and pregnancy (from 21 to 25 years), for example.

# 1 | INTRODUÇÃO

A evasão ainda é um grande problema nas escolas do Brasil. São diversas as dificuldades para resolver este problema que, há anos, se alastra em nosso país, principalmente, sobre as causas fundamentais da evasão escolar. Contudo, contribuindo para a problemática, existem limites a serem rompidos, seja por parte do alunado seja por parte da escola em lidar com tais questões. Um levantamento feito pelo movimento “Todos Pela Educação” com base na Pesquisa Nacional por Amostragem Domiciliar de 2013 indica que 45,7% dos jovens brasileiros não conseguem concluir o ensino médio até os 19 anos, 02 anos depois de idade adequada. Converter esse quadro não é tarefa fácil. Variáveis como situação social e dinâmica familiar estão envolvidas, entre outros elementos que vão além dos muros da escola, mas há posturas que podem ser adotadas e que podem melhorar gradativamente a situação.

Segundo Neri (2009 apud CUNHA, 2014) reconhece as causas da evasão escolar a partir de três motivos básicos de motivação, sendo eles: desconhecimento dos gestores da política pública, restringindo a oferta dos serviços educacionais; falta de interesse dos pais e dos alunos sobre a educação ofertada e as restrições de renda e do mercado de crédito que impedem as pessoas de explorar os altos retornos oferecidos pela educação a longo prazo. Para CUNHA (2014), várias causas da evasão escolar são elencadas, e pode-se levar em consideração alguns fatores que determinam essa ocorrência:

- **escola:** não atrativa, autoritária, com professores despreparados, insuficiente e ausência de motivação;
- **aluno:** desinteressado, indisciplinado, problemas de saúde e gravidez;
- **pais/responsáveis:** não cumprimento do pátrio poder e desinteresse em relação ao destino dos filhos;
- **social:** trabalho com incompatibilidade de horário para os estudos, agressão entre os alunos e violência.

Portanto, diversos fatores internos e externos, como permanência na escola, drogas, gravidez, reprovações sucessivas, trabalho, localização da escola, falta de atratividade em sala de aula, dentre outros, podem ser decisivos para o aluno evadir-se da escola. De acordo com o Inep/MEC, baseado no censo escolar de 2016, o índice de evasão escolar entre crianças e jovens é alarmante, conforme mostra a Tabela 1.

Etapa Escolar	Taxa de Reprovação	Taxa de Abandono	Taxa de Aprovação
Fundamental (anos iniciais)	5,9%	0,9%	93,2%
Fundamental (anos finais)	11,4%	3,1%	85,5%
Ensino médio	12,0%	6,6%	81,5%

Tabela 1. Taxa de Rendimento do Ensino Fundamental e Médio - 2016.

Diante deste contexto, este trabalho tem como objetivo mostrar os resultados alcançados a partir da aplicação de quatro algoritmos de mineração utilizando dados de alunos de uma escola da rede estadual de ensino da Paraíba, para compreender os motivos que levaram estudantes a evadirem da escola. Os algoritmos utilizados estão disponíveis em uma ferramenta de mineração chamada Weka.

O presente artigo está estruturado da seguinte forma: na seção 2 serão abordados aspectos da mineração de dados educacionais. A seção 3 mostra o detalhamento da tarefa de classificação de dados, bem como exemplifica uma forma de representação do conhecimento descoberto através desta tarefa (regras). Na seção 4 é apresentado o estudo de caso realizado, sobretudo, os resultados alcançados por cada algoritmo utilizado. Por fim, na seção 5, são descritas as considerações finais e, em seguida, as referências bibliográficas.

# 2 | MINERAÇÃO DE DADOS EDUCACIONAIS

Mineração de dados (MD) é uma área que explora grandes volumes de dados em busca de padrões. Para isso, existem algoritmos de machine Learning, como árvores de decisão, que são capazes de fazer com que o computador aprenda usando dados de eventos passados. Fayyad (1996) propôs um processo para transformar um conjunto de dados em padrões (conhecimento), que é conhecido como processo de descoberta de conhecimento de bases de dados (Knowledge Discovery in Databases), composto de 5 etapas, conforme é ilustrado na Figura 1.

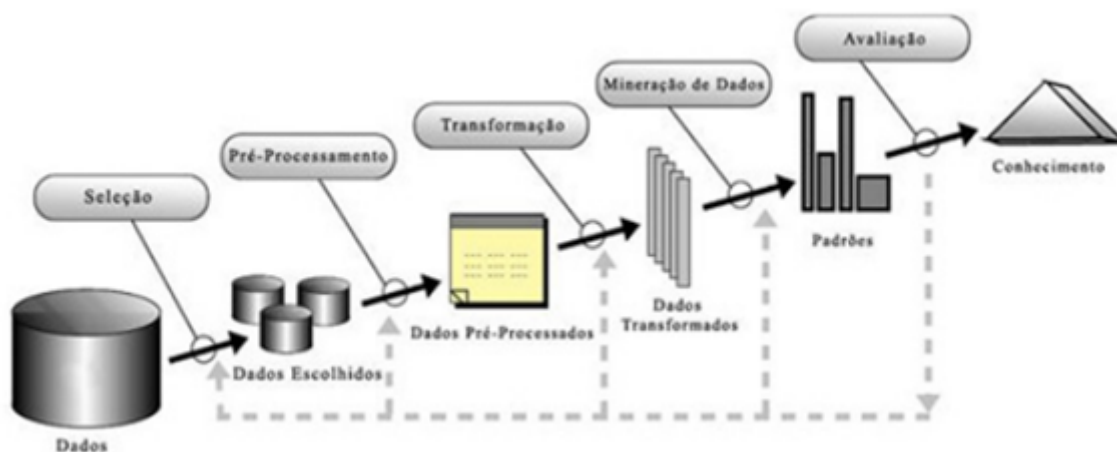


Figura1. Etapas do Processo KDD.

Fonte: (FAYYAD, 1996).

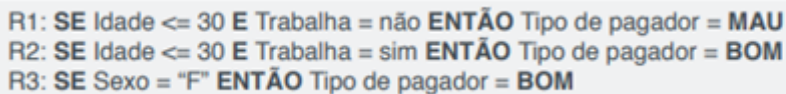
Na etapa de seleção, o objetivo é definir/escolher dados (atributos/características) a serem minerados a partir de dados brutos, podendo ser extraídos de diversas fontes, tais como diários escolares, planilhas eletrônicas, data warehouses, entre outros. A etapa de pré-processamento envolve operações como tratar a falta de dados em alguns atributos, limpeza de dados, redução da quantidade de atributos (características), preenchimento ou eliminação de valores nulos e remoção de dados duplicados. A etapa de transformação, basicamente, se resume em formatar dados para serem interpretados pelos algoritmos de aprendizado, ou seja, alterar o formato do arquivo de dados. A etapa de mineração de dados caracteriza-se pela aplicação de um ou mais algoritmos para extrair conhecimento implícito em padrões. E a etapa de avaliação consiste em interpretar e validar o conhecimento descoberto para processos de tomada de decisão, sendo necessário, o processo pode recomeçar por uma etapa anterior.

Segundo Baker (2009), grande parte dos métodos utilizados em Mineração de Dados Educacionais (Educational Data Mining- EDM) são oriundos da própria mineração de dados, adaptados às necessidades e particularidades da área da Educação. Em Baker (2009), é possível encontrar uma

lista de tarefas de aprendizado que podem ser realizadas com dados educacionais, sendo que cada uma possui um objetivo específico. Neste trabalho, a tarefa de aprendizado escolhida foi classificação porque busca encontrar características de alunos que evadiram analisando os fatores que contribuíram.

# 3 | TAREFA DE CLASSIFICAÇÃO

Na tarefa de classificação, o objetivo é descrever ou prever as características de um atributo especial chamado de “atributo-classe” ou “classe”, podendo ter dois ou mais valores possíveis. Por exemplo, considere um atributo-classe chamado de “Tipo de Pagador”, podendo ter dois valores possíveis: sim (representa um bom pagador) e não (representa um mau pagador). Assim, a partir dos atributos existentes no conjunto de dados, o algoritmo de classificação identificará características que representem pessoas que tenham um perfil de bom e mau pagador. A Figura 2 ilustra um conjunto de 03 regras do tipo Se ... , então (R1, R2 e R3) que identificam situações que diferencia nas duas classes citadas.



R1: SE Idade <= 30 E Trabalha = não ENTÃO Tipo de pagador = MAU  
R2: SE Idade <= 30 E Trabalha = sim ENTÃO Tipo de pagador = BOM  
R3: SE Sexo = "F" ENTÃO Tipo de pagador = BOM

Figura 2. Exemplo de regras geradas pela tarefa de classificação.

Fonte: autoria própria.

Neste exemplo, a classificação serviu para identificar e diferenciar perfil de bons e maus pagadores. A regra R1 informa que se uma pessoa tiver idade menor ou igual a 30 e se não trabalha, então será um mau pagador. Por outro lado, na regra R2, se uma pessoa tiver idade menor ou igual a 30 e trabalhar, então é um bom pagador. A regra R3, por sua vez, informa que pessoas do sexo feminino são boas pagadoras.

A tarefa de classificação tem como objetivo classificar/descrever os grupos existentes observando as características comuns em um conjunto de dados. Na tarefa A tarefa de classificação tem como objetivo classificar/descrever os grupos existentes observando as características comuns em um conjunto de dados. Na tarefa de classificação, a forma de representação do conhecimento (padrões) pode ser com regras do tipo (SE...ENTÃO), (conforme exemplo da Figura 2) ou árvores de decisão (forma gráfica e hierárquica na apresentação das características dos grupos). Cada tarefa pode ainda ter várias implementações através de diferentes algoritmos. Na literatura, há uma infinidade de algoritmos de classificação disponíveis. No caso da ferramenta de mineração de dados Weka, Universidade de Waikato na Nova Zelândia, existem os seguintes algoritmos de classificação: Part, OneR, J48 e randomtree (FRANK et al., 2016).

# 4 | ESTUDO DE CASO

## 4.1 Seleção

Conforme a explicação anterior, a etapa de seleção envolve a compreensão do domínio e dos objetivos da tarefa a ser desenvolvida, bem como a obtenção dos dados (atributos/características). Para esse estudo de caso, os dados foram coletados a partir das respostas de um formulário disponibilizado para alunos que estudavam na escola X. As questões deste formulário continham:

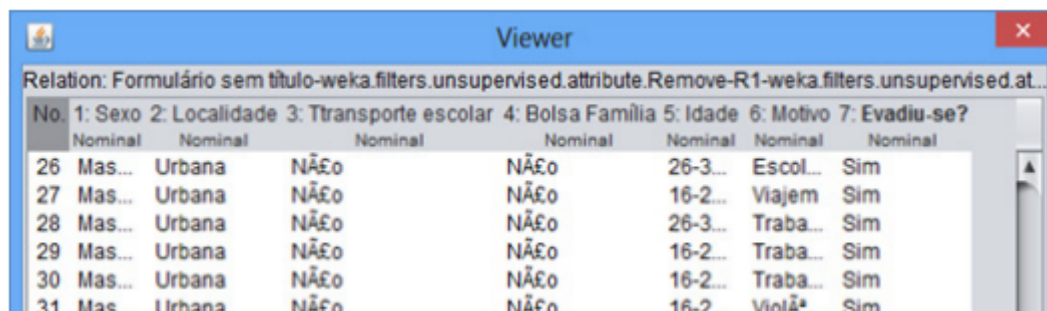
1. Sexo (Masculino, Feminino)
2. Localidade de residência (Rural, Urbana)
3. Utiliza transporte escolar (Sim, Não)
4. Participa do projeto social bolsa família (Sim, Não)
5. Idade (6-10, 11-15, 16-20, 21-25, 26-30)
6. Houve abandono da escola alguma vez (Sim, Não)
7. Qual motivo (os) que ocasionou (ram) o abandono: (falta de perspectiva profissional, casamento, bullying, escola não atrativa, gravidez, trabalho ou desinteresse).
8. A partir dos dados, o objetivo geral foi compreendido a partir do perfil dos alunos que evadiram da escola estadual X alguma vez e os fatores que levaram este abandono.

## 4.2 Pré-processamento

Foram eliminados dados de alunos que não evadiram, pois, o objetivo era justamente compreender os fatores que levaram os alunos a desistirem de continuar estudando em algum momento. No total, o conjunto de dados continha dados de 200 alunos.

## 4.3 Transformação

Nesta etapa, os dados foram formatados para que pudessem ser lidos pela ferramenta Weka. Assim, os dados foram exportados em formato CSV (Comma-separated values), em que cada dado apresenta-se separado por vírgula. A Figura 3 apresenta uma amostra dos dados coletados (06 exemplos) e carregado na ferramenta Weka.



No.	1: Sexo	2: Localidade	3: Transporte escolar	4: Bolsa Família	5: Idade	6: Motivo	7: Evadiu-se?
	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
26	Mas...	Urbana	NÃO	NÃO	26-3...	Escol...	Sim
27	Mas...	Urbana	NÃO	NÃO	16-2...	Viajem	Sim
28	Mas...	Urbana	NÃO	NÃO	26-3...	Traba...	Sim
29	Mas...	Urbana	NÃO	NÃO	16-2...	Traba...	Sim
30	Mas...	Urbana	NÃO	NÃO	16-2...	Traba...	Sim
31	Mas...	Urbana	NÃO	NÃO	16-2...	Violã*	Sim

Figura 3. Amostra do conjunto de dados coletado.



Fonte: Autoria própria.

#### 4.4 Mineração de dados

Nesta etapa foram utilizados os algoritmos de classificação Part, OneR, J48 e Randomtree, disponíveis na ferramenta Weka, para identificação de padrões (conhecimento). As cinco primeiras regras geradas por cada algoritmo são apresentadas nas Figuras 4, 5, 6 e 7.

R1: **SE** Idade 21-25 **ENTÃO** Motivo = Gravidez  
R2: **SE** Idade 16-20 **ENTÃO** Motivo = Trabalho  
R3: **SE** Idade 26-30 **ENTÃO** Motivo = Casamento  
R4: **SE** Idade 6-10 **ENTÃO** Motivo = Trabalho  
R5: **SE** Idade 11-15 **ENTÃO** Motivo = Trabalho

Figura 4. Regras geradas pelo algoritmo OneR

Fonte: autoria própria.

R1: **SE** Idade 26-30 **E** Bolsa Família = Não **E** Sexo = Masculino **ENTÃO** Motivo = Trabalho  
R2: **SE** Idade 21-25 **ENTÃO** Motivo = Gravidez  
R3: **SE** Idade 26-30 **ENTÃO** Motivo = Casamento  
R4: **SE** Sexo = Masculino **ENTÃO** Motivo = Trabalho  
R5: **SE** Bolsa Família = Não **ENTÃO** Motivo = Trabalho

Figura 5. Regras geradas pelo algoritmo Part

Fonte: autoria própria.

R1: **SE** Idade 21-25 **ENTÃO** Motivo = Gravidez  
R2: **SE** Idade 16-20 **ENTÃO** Motivo = Trabalho  
R3: **SE** Idade 26-30 **ENTÃO** Motivo = Casamento  
R4: **SE** Idade 6-10 **ENTÃO** Motivo = Trabalho  
R5: **SE** Idade 11-15 **ENTÃO** Motivo = Trabalho

Figura 6. Regras geradas pelo algoritmo J48

Fonte: autoria própria.

R1: **SE** Idade 21-25 **E** Bolsa Família = Sim **E** Residência = Urbana **ENTÃO** Motivo = Casamento  
R2: **SE** Idade 21-25 **E** Bolsa Família = Sim **E** Residência = Rural **ENTÃO** Motivo = Trabalho  
R3: **SE** Idade 21-25 **E** Bolsa Família = Não **E** Sexo = Masculino **ENTÃO** Motivo = Bullying  
R4: **SE** Idade 21-25 **E** Bolsa Família = Não **E** Sexo = Feminino **E** Transporte Escolar = Não **ENTÃO** Motivo = Gravidez  
R5: **SE** Idade 21-25 **E** Bolsa Família = Não **E** Sexo = Feminino **E** Transporte Escolar = Sim **ENTÃO** Motivo = Casamento

Figura 7. Regras geradas pelo algoritmo Randomtree

Fonte: autoria própria.

#### 4.5 Avaliação

Esta etapa destinou-se a interpretação e avaliação dos resultados gerados na etapa anterior. Pôde-se verificar que as regras geradas pelo algoritmo OneR, conforme Figura 4, idade foi o único atributo utilizado para diferenciar o motivo da evasão. Na maioria dos casos, aponta trabalho como motivo da evasão, exceto para as idades de 21 a 30. Em relação as regras geradas pelo algoritmo Part, apresentadas na Figura 5, observou-se que alunos do sexo masculino e que não recebem bolsa família evadem tendo como motivo o trabalho. O algoritmo J48, por sua vez, também identificou trabalho como sendo o motivo principal para a evasão, exceto para as idades de 21-25 (gravidez) e 26-30 (casamento), conforme mostra a Figura 6. O algoritmo Randomtree gerou regras com o maior nível de detalhe, como pode ser visto na Figura 7. Percebeu-se que para a faixa etária de 21-25, os motivos podem ser variados (casamento, trabalho, bullying ou gravidez).

# 5 | CONSIDERAÇÕES FINAIS

Embora a evasão escolar possa ocorrer por inúmeros motivos socioeconômicos, esta pesquisa foi motivada pela preocupação com o alto número de desistentes das escolas públicas brasileiras e, também, pela tentativa de compreender a correlação do perfil dos alunos e os motivos que influenciaram a evasão NA ESCOLA PESQUISADA. Para isso, aplicou-se as etapas do processo KDD para descobrir algum tipo de conhecimento oculto nos dados visando identificar um padrão do aluno evadido.

Foi analisado um conjunto de dados contendo idade, sexo, tipo de transporte usado para chegar à escola, participação no programa bolsa família e localidade de moradia de 200 alunos de uma escola X do estado da Paraíba. No geral, analisando os resultados gerados dos quatro algoritmos de classificação, constatou-se que os motivos que levam os alunos a evadirem são: A idade de 11 a 20 (trabalho), 21 a 25 (gravidez) e 26 a 30 (casamento). Assim, tais informações podem dar um suporte a gestores escolares e equipes pedagógicas para criarem campanhas focadas em cada faixa etária, fazendo com que a taxa de evasão atenuie nas escolas onde atuam.

# REFERÊNCIAS

BAKER, R. S. J.; YACEF, K. **The state of educational data mining in 2009: a review and future visions**. Journal of Educational Data Mining, 1(1):3-17, 2009. Disponível em: < <https://jedm.educationaldatamining.org/index.php/JEDM/article/download/8/2/>>. Acesso em: 04/11/2017.

CUNHA, V. F. **Evasão escolar e suas causas**. Disponível em: < [http://www.diaadiaeducacao.pr.gov.br/portals/cadernospde/pdebusca/producoes\\_pde/2014/2014\\_uepg\\_cien\\_artigo\\_valdemar\\_fernandes\\_da\\_cunha.pdf](http://www.diaadiaeducacao.pr.gov.br/portals/cadernospde/pdebusca/producoes_pde/2014/2014_uepg_cien_artigo_valdemar_fernandes_da_cunha.pdf)>. Acesso em: 07/11/2017.

FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. **From data mining to knowledge discovery: An overview**. In: Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press, England, 1996, p.1-34.

FRANK, E.; HALL, M.; WITTEN, I. **The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques**, 4<sup>o</sup> ed., 2016.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA - INEP. **Sinopse Estatística da Educação Básica 2016**. Brasília - DF. Disponível em: <<http://portal.inep.gov.br/sinopses-estatisticas-da-educacao-basica>>. Acesso em: 21/10/2017.

QEDU, **Taxas de Rendimento 2016**. Disponível em: <<http://www.qedu.org.br/brasil/taxasrendimento>>. Acesso em: 07/11/2017.