INFORMÁTICA NA EDUCAÇÃO: teoria & prática Porto Alegre, v.21, n.3, set./dez. 2018. ISSN digital 1982-1654
ISSN impresso 1516-084X

# Mineração de Dados Educacionais: um Estudo da Evasão no Ensino Médio com Base nos Indicadores do Censo Escolar

# Educational Data Mining: a Study of Withdrawal in High School Based on the Indicators of the School Census

ROGÉRIO COLPANI
Centro Universitário da Fundação Educacional Guaxupé - Unifeg

**Resumo:** O emprego de técnicas de Mineração de Dados proporciona obter conhecimentos úteis com o intuito de melhorar a proposta educacional. Com o uso dos indicadores educacionais disponibilizados pelo INEP e utilizando a metodologia CRISP-DM, foram aplicadas as técnicas de correlação e regressão linear a fim de analisar as variáveis que se relacionam com a evasão escolar. Como resultados, foi possível verificar que a taxa de distorção idade-série apresentou a maior correlação positiva, sendo $r = 0{,}59$. Em seguida, foi obtido o modelo preditivo no qual o resultado da Raiz do Erro Médio Quadrático foi de 5.06. Contudo, pode-se verificar que a Mineração de Dados Educacionais pode ser aplicada para exploração dos dados e gerar conhecimento de modo a servir como auxílio para soluções de problemas e suporte a mecanismos em apoio ao ensino.

**Palavras-chave:** Mineração de Dados. Evasão. Descoberta de Conhecimento em Base de Dados.

**Abstract:** The use of Data Mining techniques provides useful knowledge in order to improve the educational proposal. Using the educational indicators provided by INEP and using the CRISP-DM methodology, the correlation and linear regression techniques were applied in order to analyze the variables that are related to school dropout. As results, it was possible to verify that the age-series distortion rate presented the highest positive correlation, where r = 0.59. Then, the predictive model was obtained, in which the root mean square error was 5.06. However, it can be verified that Educational Data Mining can be applied to data mining and generate knowledge to serve as an aid to problem solving and support mechanisms in support of teaching.

**Keywords:** Data Mining. Withdrawal. Knowledge Discovery in Database.

COLPANI, Rogério. Mineração de Dados Educacionais: um Estudo da Evasão no Ensino Médio com Base nos Indicadores do Censo Escolar. *Informática na Educação*: teoria & prática, Porto Alegre, v. 21, n. 3, p. 109-123, set./dez. 2018.

## 1 Introdução

A educação, nos últimos anos, vem passando por muitas mudanças que se configuram como melhorias estruturais para as instituições de ensino, possibilitando a escolarização de milhões de pessoas na educação básica (CALIXTO; SEGUNDO; GUSMÃO, 2017).

De acordo com a Lei de Diretrizes e Bases da Educação Nacional (LDB)[1], "a educação básica tem por finalidade desenvolver o educando, assegurar-lhe a formação comum indispensável para o exercício da cidadania e fornecer-lhe meios para progredir no trabalho e em estudos posteriores". Entretanto, diversos problemas vêm sendo enfrentados pelas instituições de ensino, entre eles, a evasão escolar (SILVIA; NUNES, 2015).

A evasão escolar se define em alunos que não completam os cursos, podendo também ser considerados evadidos aqueles que se matriculam e não iniciam o curso (MAIA; MEIRELLES; PELA, 2004). Porém, nos trabalhos de Colpani (2018) e Rigo *et al.* (2014), os autores relatam que o termo evasão é apresentado sob diversas formas e vista sob várias interpretações.

Segundo levantamentos realizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), entre 2014 e 2015 a taxa de evasão na primeira série do ensino médio foi de 12,7%, seguida por 12,1% dos alunos matriculados na segunda série. Já a terceira série obteve 6,7% de evasão, chegando a 11% do total de alunos nessa etapa de ensino (INEP, 2017a). Ainda sob os relatos do mesmo instituto, observa-se que a evasão é maior nas escolas rurais, em todas as etapas de ensino, ficando com o estado do Pará a maior taxa de evasão, chegando a 16% no ensino médio (INEP, 2017b).

A problemática da evasão escolar atinge diversas instituições, estando presentes em todos os níveis de ensino, e é considerada uma grande preocupação para empresários, diretores, pesquisadores, pais e alunos (LOBO, 2012; BITTENCOURT; MERCADO, 2014). As perdas ocasionadas pela evasão podem acarretar prejuízos às instituições de ensino, sendo, para o setor público, os recursos investidos sem o devido retorno; para o setor privado, importante perda de receita; para ambos os setores, fonte de ociosidade de professores, funcionários, equipamentos e espaço físico (COLPANI, 2018).

Calixto, Segundo e Gusmão (2017), apontam que a evasão escolar ocorre por diversas causas, dentre elas, estão relacionadas a qualidade da educação oferecida pela instituição de ensino, o ambiente escolar, a relação familiar, o meio social em que o estudante vive, motivos pessoais do próprio aluno, necessidade de trabalhar para auxiliar na renda familiar, idade avançada decorrente de reprovações que acabam instigando o aluno a se sentir deslocado, entre outros.

Diante de tal problema, pesquisas vêm sendo realizadas na área de evasão escolar, dentre elas, aquelas que fazem uso de Mineração de Dados Educacionais (MDE) para fins de tomada de decisão (MACHADO et *al*., 2015). Esta se classifica como uma área de pesquisa que lida com o desenvolvimento de métodos para explorar conjuntos de dados de contextos educacionais (BRITO et *al*., 2015). A MDE consiste na aplicação de técnicas de Mineração de Dados (MD) e aplicação de algoritmos com o intuito de extrair padrões em bases de dados, cujo objetivo é descobrir conhecimento útil nestes dados (BRITO et *al*., 2015).

Os métodos de Classificação, Associação e *Clustering* são os mais utilizados em pesquisas de MDE, sendo os principais objetivos: estimativa ou modelagem de desempenho de estudantes, modelagem de grupos ou aprendizagem colaborativa, mediação ou recomendação pedagógica e apoio ao estudante e *feedback* (RODRIGUES et *al*., 2014). Em adição, Machado et

[1] Lei de Diretrizes e Bases da Educação Nacional. Disponível em http://www.planalto.gov.br/ccivil_03/leis/L9394.htm. Acesso em 15 out. 2018.

*al*. (2015), apontam que Árvores de Decisão, Redes Neurais, Regressão Logística, *Clustering* e Regras de Associação, são os principais métodos de MDE aplicados a problemática de evasão escolar.

Com base no exposto, o presente trabalho tem como objetivos (1) realizar uma análise correlacional dos indicadores educacionais do Censo Escolar 2017, fornecidos publicamente pelo INEP, com o intuito de determinar quais variáveis educacionais estão relacionadas com a evasão escolar, no ensino médio; (2) propor um modelo preditivo, usando o método de regressão linear, para determinar a equação que modela os dados correlacionados.

Este trabalho está estruturado em cinco seções. Na primeira seção foi realizada a introdução ao tema, expondo os problemas de evasão escolar e a necessidade de explorar técnicas de MDE. A segunda seção apresenta uma breve contextualização sobre MD e MDE, bem como os trabalhos relacionados e as contribuições desta pesquisa. A terceira seção aborda a metodologia empregada no trabalho. A quarta seção mostra os resultados obtidos. Por fim, a quinta seção apresenta as considerações finais da pesquisa.

## 2 Mineração de Dados Educacionais e a Evasão Escolar

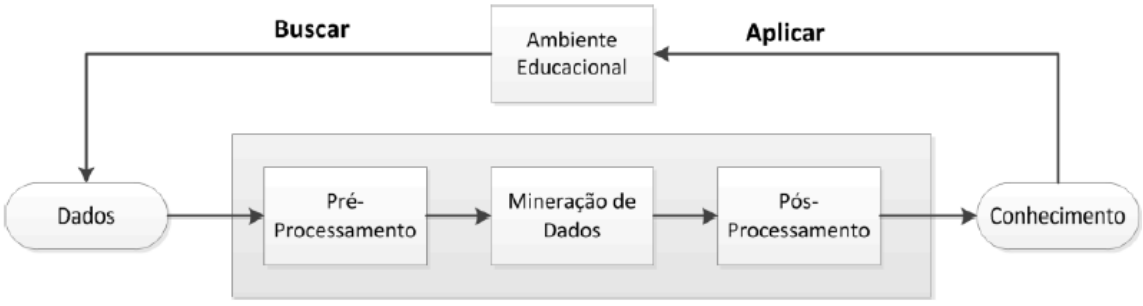
O avanço das Tecnologias de Informação e Comunicação e da *internet*, com a massiva difusão do uso de sistemas informatizados nas instituições de ensino, possibilita o crescimento a cada dia de uma considerável quantidade de dados gerados e disponibilizados (RIGO et *al*, 2014).

Este grande volume de dados tem instigado o interesse na sua utilização, junto de técnicas de MD, a fim de obter respostas para diversas perguntas específicas na área da Educação (BAKER; ISOTANI; CARVALHO, 2011; MANHÃES et *al.*, 2011).

O que tange a área de MD, Witten e Frank (2000) mencionam que esta pode ser considerada como uma das etapas do processo de Descoberta de Conhecimento em Base de Dados (DCBD), no qual se faz uso de algoritmos para extração de padrões de bases de dados (REZENDE, 2005).

O processo de DCBD é definido, segundo Goldschmidt, Passos e Bezerra (2015, p.03), como um "processo de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados". A Figura 1 ilustra, de modo sucinto, as etapas operacionais do processo de DCBD:

Figura 1 - Etapas operacionais do processo de DCBD aplicado ao ambiente educacional.




Fonte: Rodrigues et *al*. (2014).

**i.** *Pré-processamento:* o primeiro passo é entender o domínio do problema e realizar a seleção do conjunto de dados com base nos objetivos a serem alcançados. Em

seguida, são realizadas as funções de organização, tratamento e preparação dos dados para a próxima etapa, a MD (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

**ii.** *Mineração de Dados:* principal etapa do processo de DCBD, ocorre a busca efetiva por conhecimentos novos e úteis a partir da aplicação de algoritmos sobre os dados (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

**iii.** *Pós-processamento:* envolve a visualização, análise e a interpretação do modelo de conhecimento gerado pela etapa de MD (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

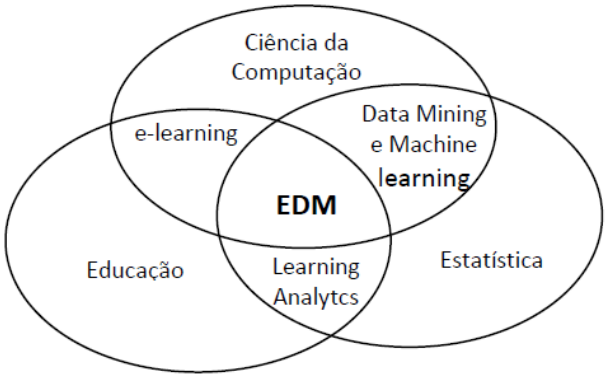
Além das etapas operacionais supracitada, Goldschmidt, Passos e Bezerra (2015) apresentam as tarefas envolvidas no processo de DCBD agrupadas em: tarefas de associação, classificação, regressão, sumarização e *clustering*. Em relação aos objetivos da tarefa de MD, estes são classificados em:

- *Predição:* busca um modelo de conhecimento que permita prever valores de determinados atributos em novas situações (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).
- *Descrição:* busca um modelo que descreva o comportamento dos dados (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

A MDE, de acordo com Romero e Ventura (2013), é a combinação de três principais áreas de conhecimento: Computação, Educação e Estatística. A intersecção destas áreas gera três subáreas: *E-learning*, *Data Mining* e *Machine Learning*, conforme ilustrado na Figura 2.

Segundo Baker, Isotani e Carvalho (2011) e Rigo et *al*. (2014), a MDE é definida como a área de pesquisa que tem como principal foco a aplicação das técnicas de MD para explorar conjunto de dados coletados em ambientes educacionais.

Figura 2 - Principais áreas relacionadas a MDE.




Fonte: Rodrigues et *al*. (2014)

Diversos trabalhos vêm sendo realizados com o objetivo de identificar as tendências de evasão. Brito et *al*. (2015), apresentam uma estratégia de apoio aos docentes para detectar de forma precoce alunos no curso de Ciência da Computação, com risco de evasão, baseada em suas notas obtidas nas matérias do vestibular e disciplinas cursadas no primeiro semestre do curso. Os algoritmos de *NaiveBayes*, J48, *AdaBoostM1*, *SimplesLogistic*, IBK, *DecisionTable* e VFI foram empregados como métodos de MD a fim de comparar os resultados obtidos. A acurácia da classificação dos alunos se manteve superior a 84% em todos os algoritmos, sendo o *DecisionTable* o que obteve o melhor resultando, com 86,9%.

Rigo et *al*. (2014), descrevem a aplicação desenvolvida e os resultados obtidos no uso do sistema de MDE e *Learning Analytics* em uma experiência prática de aplicação de recursos tecnológicos em busca do tratamento do problema de evasão escolar. Fazendo uso da Rede Neural Artificial *Mutilayer Perceptron* com o algoritmo *Backpropagation* verificou-se que a taxa de acerto da predição de evasão ficou entre 68% e 83% nas disciplinas analisadas.

Oliveira Júnior, Noronha e Kaestner (2014), aplicam técnicas de MD em duas bases de dados. A primeira base, refere-se a um Sistema de Controle Acadêmico, já a segunda base de um Sistema de Controle de Biblioteca a fim de verificar se existe alguma correlação entre o empréstimo de livros em biblioteca com a evasão no curso. Com o uso dos métodos de MD J48, Rede Neural Artificial *Multilayer Perceptron*, SMO, IBK e *Naive Bayes*, verificou-se que o algoritmo J48 apresentou a melhor acurácia entre os demais métodos empregados, com um valor próximo de 80%. Além disso, destacou-se que 92,60% dos alunos com idade entre 17 e 18 anos desistiram do curso. Também foi observado que 80,6% dos alunos, com mais de 18 anos, que emprestavam livros na biblioteca continuam frequentando o curso o que apresenta um indício de correlação entre estes dois atributos.

Paz e Cazella (2017) buscaram identificar perfis de alunos com potencial de evasão em uma universidade comunitária, procurando contribuir com os gestores universitários no planejamento de ações efetivas para a retenção de alunos desta instituição de ensino. Fazendo uso do método J48, os experimentos realizados tiveram alto grau de precisão atingindo acurácia superiores a 90%. Dentre os principais resultados ficou evidenciado que o incentivo e o currículo dos alunos estão diretamente relacionados à tendência de evasão, e que estas ocorrem geralmente nos semestres iniciais dos cursos com alunos sem incentivo.

Oliveira Júnior, Noronha e Kaestner (2017), empregaram um método de seleção de atributos aplicados na previsão da evasão escolar, utilizando classificação, criação e seleção de atributos, com a finalidade de auxiliar a análise da evasão de alunos, aplicado em cursos presenciais de graduação. Os métodos empregados para MD foram JRip, J48, *Naive Bayes*. Como resultado, verificou-se que aproximadamente 80% da evasão ocorre até o 3° período. Nas bases de dados utilizadas, a abordagem *wrapper* obteve a melhor acurácia, com resultados entre 83% e 87%. Dentre os atributos aplicados na previsão da evasão escolar, o atributo 'dificuldade média das disciplinas cursadas pelo aluno' revelou-se uma boa medida de prognóstico de desempenho do aluno.

Por fim, Nascimento, Cruz Júnior e Fagundes (2018) utilizam bases de dados educacionais fornecidas pelo INEP e aplicaram técnicas de MD com finalidade de melhor explicar indicadores como a evasão e reprovação escolar no ensino fundamental. Após análises correlacionais das variáveis, modelos de regressão linear e de regressão robusta foram desenvolvidos a fim de comparar o desempenho e fornecer um modelo que minimize o erro de previsão. Os resultados obtidos indicam que a regressão robusta obteve melhores resultados na estimação das variáveis elencadas no estudo. A dispersão idade-série dos alunos, o esforço dos docentes para exercer a profissão, a quantidade de alunos por turma e a formação dos docentes, mostram-se variáveis mais associadas com estes indicadores.

Em suma, diante das pesquisas realizadas, é possível verificar que as aplicações foram empregadas no ensino superior e fundamental, sendo a maioria dos trabalhos fazendo uso da linguagem de programação R[2] e do *software* Weka[3].

Com base no exposto, a realização do presente trabalho visa contribuir com o uso de um modelo de predição diferente dos apresentados por Machado et *al.* (2015). Além disso, propõe-se uma análise correlacional entre os indicadores de evasão e reprovação, o que não foi realizado nos trabalhos supracitados. O estudo também se difere ao se concentrar na exploração dos indicadores do ensino médio, fornecidos pelo INEP em 2017, e fazer uso da

[2] Linguagem R. Disponível em https://www.r-project.org/. Acesso em 20 nov. 2018.

[3] *Software* Weka. Disponível em https://www.cs.waikato.ac.nz/ml/weka/. Acesso em 20 nov. 2018.

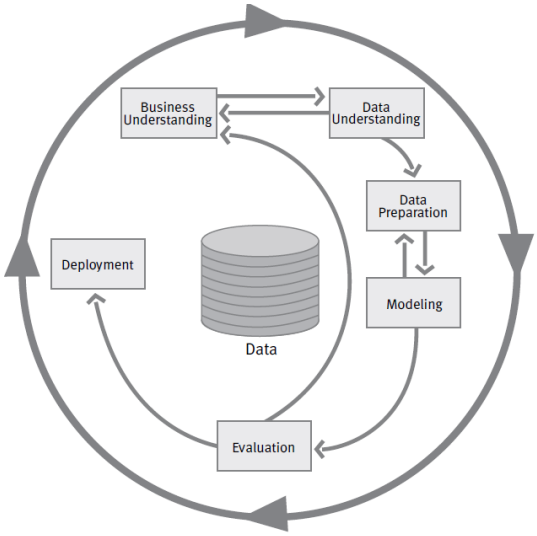
linguagem de programação Python na implementação de todas as etapas. A justificativa para o uso da linguagem Python se dá por ser uma das linguagens mais utilizadas no mercado[4] atualmente e por não ter sido encontrado em nenhum trabalho, conforme supracitado.

## 3 Metodologia para descoberta de conhecimento em base de dados

O presente trabalho adotou a metodologia *Cross-Industry Standard Process for Data Mining* (CRISP-DM) para nortear o processo de DCBD. Este modelo foi escolhido por ser um dos mais utilizados, aceitos e bem-sucedidos em várias aplicações de DCBD no mercado mundial, além de apresentar um conjunto de fases muito bem definidas (GOLDSHMIDT; PASSOS; BEZERRA, 2015).

A metodologia define uma sequência não rígida de seis fases - (a) Compreensão do negócio; (b) Compreensão dos dados; (c) Preparação dos dados; (d) Modelagem; (e) Avaliação; (f) Desenvolvimento – que permite a construção e implementação de um modelo de DCBD em um ambiente real, auxiliando nas decisões de negócios (GOLDSHMIDT; PASSOS; BEZERRA, 2015). Assim, o presente trabalho segue as fases do modelo em questão, ilustrado na Figura 3, e descritas a seguir.

Figura 3 - Fases do Modelo CRISP-DM.




Fonte: Chapman et *al.* (2000).

### 3.1 Compreensão do negócio

Segundo Goldshmidt, Passos e Bezerra (2015), a primeira etapa do CRISP-DM consiste em conhecer o contexto em que o processo de DCBD será realizado e converter o conhecimento adquirido em uma definição de problema de MD (NASCIMENTO; CRUZ JÚNIOR; FAGUNDES, 2018). Assim, para a presente pesquisa, foram analisados os trabalhos relacionados de MDE aplicados à problemática de evasão escolar. Em seguida, os objetivos a serem alcançados foram formulados, sendo eles:

[4] *The 2018 Top Programming Languages*. Disponível em https://spectrum.ieee.org/at-work/innovation/the-2018-top-programming-languages. Acesso em 20 nov. 2018.

i. Realizar uma análise correlacional dos indicadores (‘média de alunos por turma’, ‘percentual de docentes com curso superior’, ‘reprovação’, ‘média de horas-aula diária por escola’ e ‘taxa de distorção idade-série por escola’) educacionais do Censo Escolar 2017, fornecidos publicamente pelo INEP, com o intuito de determinar quais variáveis educacionais estão relacionadas com o indicador ‘evasão escolar’, no ensino médio (INEP, 2017c).
ii. Propor um modelo preditivo, usando o método de regressão linear, para determinar a equação que modela os dados correlacionados.

### 3.2 Compreensão dos Dados

A segunda etapa do CRISP-DM requer um estudo detalhado das informações disponíveis (GOLDSHMIDT; PASSOS; BEZERRA, 2015).

A pesquisa fez uso dos indicadores educacionais derivados do Censo Escolar de 2017, disponíveis abertamente ao público (INEP, 2017c). De acordo com o próprio INEP (2017d), o “Censo Escolar é o principal instrumento de coleta de informações da educação básica e o mais importante levantamento estatístico educacional brasileiro”. Ainda de acordo com os relatos do INEP (2017d), os indicadores educacionais proporcionam valores estatísticos que permitem conhecer o desempenho dos alunos, o contexto econômico, social e as condições em que ocorrem os processos de ensino e de aprendizagem, além de monitorar o acesso, a permanência e a aprendizagem de todos os alunos.

O presente estudo utiliza as bases de dados no nível ‘escolas’, referente ao estado do Pará. A justificativa para a escolha desse estado se dá com base ao relato do INEP (2017b), por ser o que apresentou maior taxa de evasão no ensino médio. Os indicadores educacionais utilizados na pesquisa são apresentados no Quadro 1 com as respectivas abreviaturas empregadas no restante deste trabalho e os campos em comum em todas as bases de dados.

Quadro 1 – Indicadores utilizados na pesquisa.

| Sigla | Indicador | Dados contidos nas Base de Dados |
|---|---|---|
| MAT | Média de Alunos por Turma | • Ano,<br>• Região,<br>• UF,<br>• Código do Município,<br>• Código da Escola,<br>• Nome da Escola,<br>• Localização,<br>• Dependência,<br>• Nível de Escolaridade (Infantil, Fundamental e Médio) com o campo Total. |
| MHAD | Média de Horas-Aula Diária | |
| PDCS | Percentual de Docentes com Curso Superior | |
| TDIS | Taxa de Distorção Idade-Série | |
| TE | Taxa de Evasão | |
| TR | Taxa de Reprovação | |

Fonte: Desenvolvido pelo Autor.

### 3.3 Preparação dos Dados

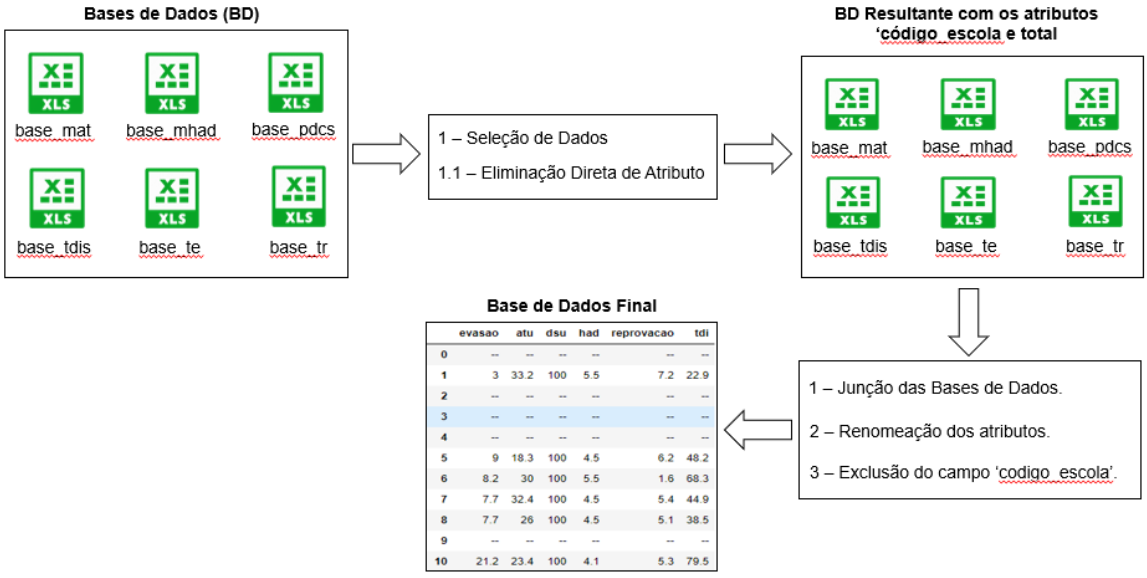
A terceira etapa constitui-se de funções relacionadas à captura, organização, tratamento e preparação dos dados para a etapa de modelagem (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

A primeira função utilizada foi a seleção de dados, que compreende na identificação das informações, dentre as bases de dados existentes, que serão consideradas durante o processo de DCBD (GOLDSHMIDT; PASSOS; BEZERRA, 2015). Após análise, foi verificado que os campos pertinentes para o presente trabalho, em um primeiro momento, são 'codigo_escola' e 'total' (referente ao nível de escolaridade do ensino médio). Assim, todos os demais campos foram removidos por meio da técnica de eliminação direta de atributos (GOLDSHMIDT; PASSOS; BEZERRA, 2015).

Com o auxílio da linguagem de programação Python, foi realizado a junção de todas as bases de dados, resultantes do procedimento anterior 'seleção de dados', por meio do campo 'codigo_escola'. Por fim, os atributos foram renomeados e a exclusão do campo 'codigo_escola' foi empregado. Após realizar as tarefas presentes, foi criada uma única base de dados com 6 variáveis e 6369 instâncias. A Figura 4 ilustra os procedimentos realizados.

Após obter a Base de Dados Final, foi possível observar valores ausentes e instâncias que possuíam valor 0 (zero). Para resolver tais problemas, foi adotada a limpeza de informações, por meio da técnica de exclusão de casos (GOLDSHMIDT; PASSOS; BEZERRA, 2015). Desse modo, foram removidas todas as instâncias que continham atributo com valores ausentes ou zero. Após esta etapa, a BD a ser utilizada na etapa de Modelagem resultou em 610 instâncias.

Figura 4 - Procedimento de Seleção de Dados.




| | evasao | atu | dsu | had | reprovacao | tdi |
|---|---|---|---|---|---|---|
| 0 | -- | -- | -- | -- | -- | -- |
| 1 | 3 | 33.2 | 100 | 5.5 | 7.2 | 22.9 |
| 2 | -- | -- | -- | -- | -- | -- |
| 3 | -- | -- | -- | -- | -- | -- |
| 4 | -- | -- | -- | -- | -- | -- |
| 5 | 9 | 18.3 | 100 | 4.5 | 6.2 | 48.2 |
| 6 | 8.2 | 30 | 100 | 5.5 | 1.6 | 68.3 |
| 7 | 7.7 | 32.4 | 100 | 4.5 | 5.4 | 44.9 |
| 8 | 7.7 | 26 | 100 | 4.5 | 5.1 | 38.5 |
| 9 | -- | -- | -- | -- | -- | -- |
| 10 | 21.2 | 23.4 | 100 | 4.1 | 5.3 | 79.5 |

Fonte: Desenvolvido pelo autor.

## 4. Modelagem

A etapa de modelagem representa o desenvolvimento dos modelos para o problema, com base nos dados pré-processados. Na presente pesquisa, foi empregado o modelo de análise de correlação e a regressão linear. A Correlação "é uma relação entre duas variáveis. Os dados podem ser representados por pares ordenados $(x, y)$, onde $x$ é a variável independente e $y$ é a variável dependente" (LARSON; FARBER, 2010, p.395). Neste trabalho, o indicador de 'evasão' é adotado como variável dependente e os demais indicadores como as variáveis independentes.

A fim de medir a força e a direção da relação entre as variáveis foi adotado o coeficiente de correlação de *Pearson*, ou $r$. O coeficiente $r$ pode assumir valores entre -1 e 1, sendo $r = 1$ uma correlação perfeita positiva e $r = -1$ uma correlação perfeita negativa, caso contrário quanto mais próximo de 0 (zero) o valor de $r$ a correlação se torna mais fraca (LARSON; FARBER, 2010). O coeficiente $r$ é calculado como apresentado na Equação 1:

$$r = \frac{n\sum xy - (\sum x)(y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} \quad (1)$$

O Modelo de Regressão Linear é uma técnica que permite prever os valores de uma variável independente (resposta) $y$ para uma dada variável dependente (explicativa) $x$ (LARSON; FARBER, 2010). A equação da reta de regressão é dada pela Equação 2.

$$\hat{y} = mx + b \quad (2)$$

A inclinação $m$ e a intersecção $b$ são dadas pelas Equações 3 e 4, respectivamente:

$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} \quad (3) \text{ e } b = \frac{\sum y}{n} - m\frac{\sum x}{n} \quad (4)$$

A entrada de dados para o modelo é dada pelas variáveis dependentes $x$ com maior valor de correlação. A configuração final do modelo preditivo é dada por 75% dos dados para a etapa de treino e 25% para a etapa de teste. A divisão das bases foi realizada por meio do método *holdout* (GOLDSHMIDT; PASSOS; BEZERRA, 2015).

### 3.5 Avaliação

A etapa de avaliação consiste na análise do modelo desenvolvido. Para a pesquisa, foram adotados o coeficiente de determinação e a Raiz do Erro Médio Quadrático (REMQ).

O coeficiente de determinação, foi utilizado para mostrar a relação da variação explicada com a variação total (LARSON; FARBER, 2010), por meio da Equação 5.

$$r^2 = \frac{Variação\,explicada}{Variação\,total} \quad (5)$$

A $variação\,explicada$ é dada pela "soma dos quadrados das diferenças entre o valor $y$ de cada par pedido e cada valor $y$ previsto corretamente", conforme ilustrado na Equação 6. (LARSON; FARBER, 2010, p.420). Já a $variação\,total$, "sobre uma linha de regressão, é a soma dos quadrados das diferenças entre o valor $y$ e cada par pedido e a média de $y$" (LARSON; FARBER, 2010, p.419), como apresentado na Equação 7.

$$Variação\,explicada = \sum (\hat{y}_i - \bar{y})^2 \quad (6)$$

$$Variação\,total = (y_i - \bar{y})^2 \quad (7)$$

A REMQ, apresentado na Equação 8, foi utilizada para verificar a acurácia da interpolação. O REMQ é uma medida de magnitude da divergência entre o valor predito no modelo de regressão linear, $\hat{y}_i$, e o conjunto de dados $y_i$. Como resultado, valores próximos de 0 (zero) possuem maior qualidade na estimativa (LARSON; FARBER, 2010).

$$REMQ = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (8)$$

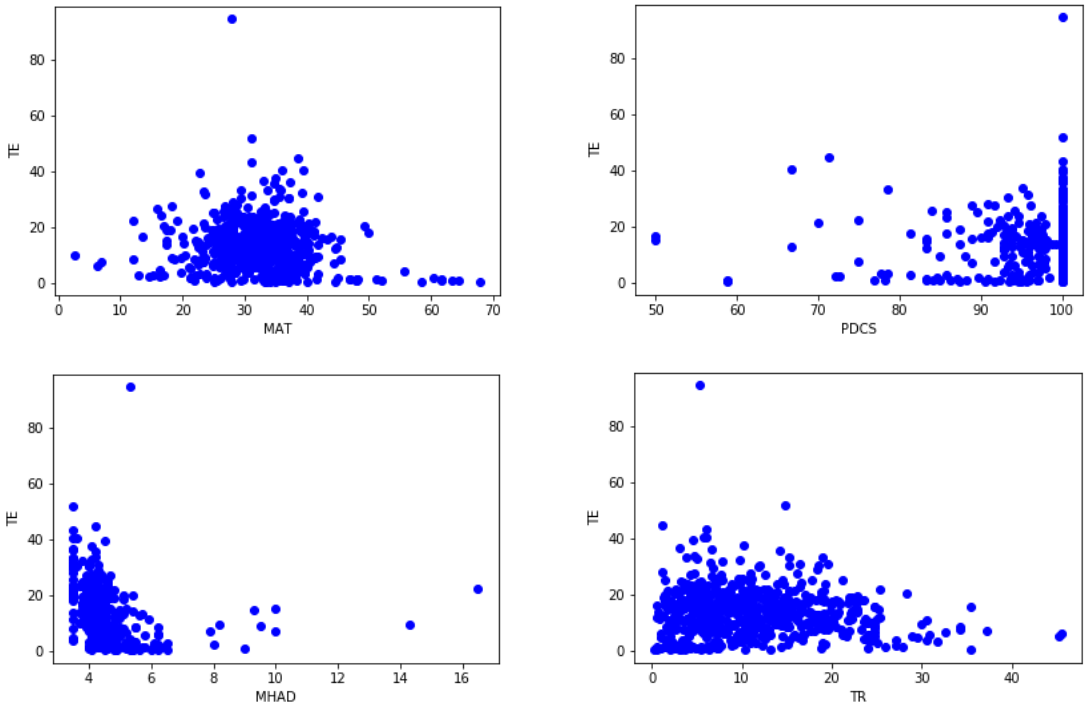
## 4 Resultados

Os gráficos de dispersão[5] ilustrados na Figura 5 e os índices de correlação apresentados na **Erro! Fonte de referência não encontrada.** mostram o relacionamento entre as variáveis independentes e a variável dependente (evasão). Com base nessas informações é possível verificar que a variável TDIS possui a maior correlação positiva, sendo $r = 0{,}59$.

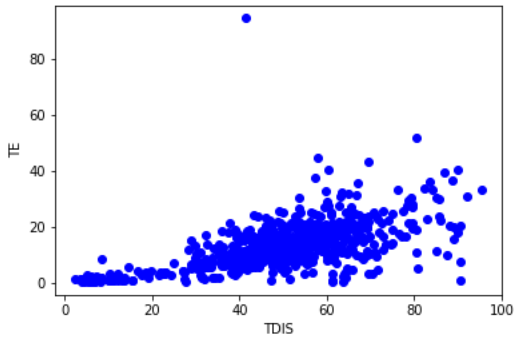
Diante o exposto, é possível inferir que a medida que a taxa de distorção idade-série (TDIS) aumenta, maior é a taxa de evasão (TE).

Para as variáveis MAT e MHAD é possível observar uma fraca correlação negativa. Já para as variáveis PDCS e TR não há correlação com a variável dependente.

Figura 5 - Gráfico de dispersão da análise correlacional entre o indicador evasão e os demais indicadores da base de dados.




[5] Um gráfico de dispersão, segundo Larson e Farber (2010), é utilizado para apresentar a relação entre duas variáveis quantitativas, por meio de pares ordenados, que são representados como pontos em um plano coordenado.




Fonte: Desenvolvido pelo autor.

Tabela 1 - Análise de Correlação entre a variável dependente e as variáveis independentes.

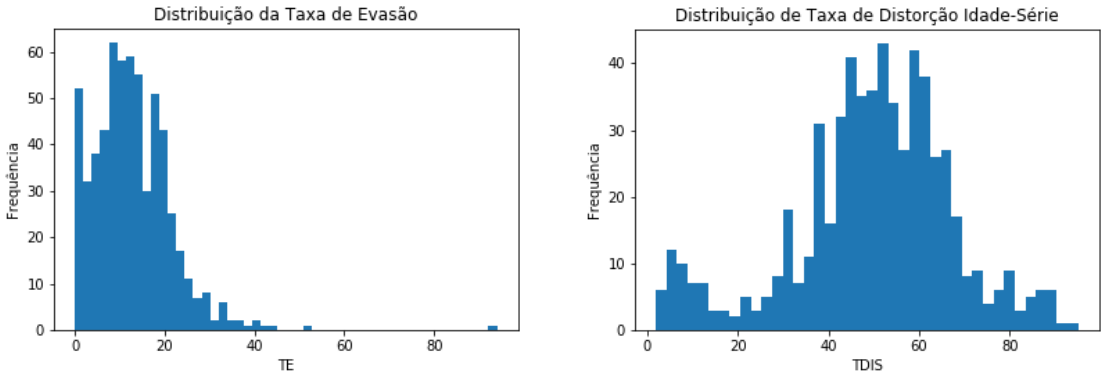
| Variável Dependente | Valor Correlacional *r* com as Variáveis Independentes | | | | |
|---|---|---|---|---|---|
| TE | MAT | PDCS | MHAD | TR | TDIS |
| | -0,10 | 0,045 | -0,26 | -0,06 | 0,59 |

Fonte: Desenvolvido pelo autor.

Para o modelo preditivo, inicialmente, foi realizada uma análise exploratória dos dados da variável independente TDIS, por ser aquela que apresentou maior correlação, e da variável dependente TE, com o intuito de verificar a disposição dos dados e a necessidade de processamento antes de aplicar a regressão linear.

Os histogramas, apresentados na Figura 6, mostram a distribuição de frequência[6] do conjunto de dados das variáveis TE e TDIS. Ao analisar o histograma da Figura 6a, é possível verificar que a maior parte das escolas apresentam uma taxa de evasão de até 20%. Já o histograma da Figura 6b, aponta que a maioria das escolas possuem uma taxa de 38% a 65% de alunos com idade superior à idade recomendada.

Figura 6 - Histogramas de frequência das variáveis TE e TDIS.




[6] A distribuição de frequência é representada por um histograma de frequência, um diagrama de barras que representa a quantidade de dados de um determinado grupo (LARSON; FARBER, 2010).

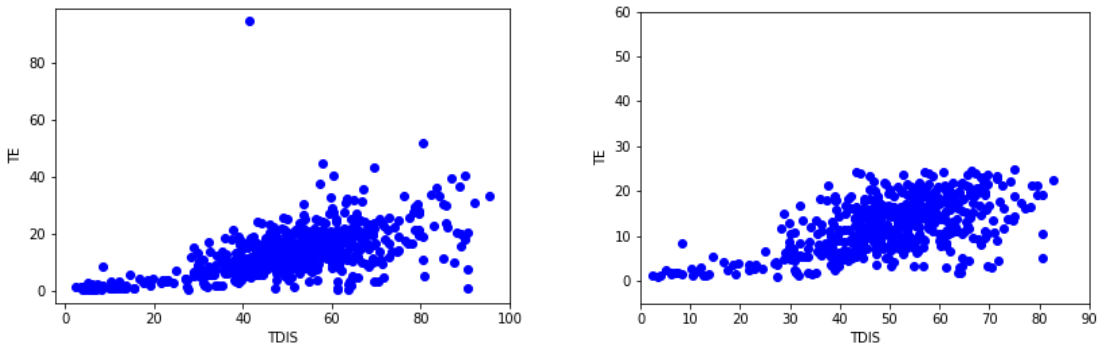
(a) (b)

Fonte: Desenvolvido pelo autor.

Após a análise das classes com alto número de frequência nos histogramas da Figura 6, foram observadas as classes com baixa frequência e, junto à análise do gráfico de dispersão apresentado na Figura 7a foi realizado um pré-processamento na base de dados com o intuito de minimizar os valores discrepantes. Assim, para a variável TDIS foram excluídas as instâncias com valores maiores ou iguais a 85.0. Já para a variável TE, foram removidas as instâncias com valores superiores ou iguais a 25.0 e inferiores a 1.0 (sendo, para estes, a justificativa de estarem próximos de 0).

Após realizar o pré-processamento, é possível verificar, na Figura 7b, uma maior concentração dos dados.

Figura 7 - Gráfico de dispersão TDIS x TE.




(a) (b)

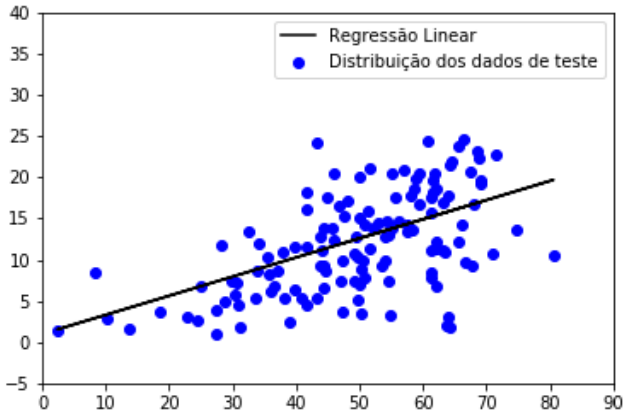
Fonte: Desenvolvido pelo autor.

O modelo de regressão tratou de encontrar os parâmetros que fornecessem o ajuste entre as variáveis observadas. Os valores obtidos para as variáveis $a$ e $b$ após o treinamento do modelo foram de 0.23 e 0.99, respectivamente. A Equação 9 apresenta o resultado da equação da reta.

$$y = 0.23x + 0.99 \quad (9)$$

A **Erro! Fonte de referência não encontrada.** mostra a dispersão dos dados de predição em relação à equação da reta. Como resultado de validação do modelo, o coeficiente de determinação foi de 0.33. Isso significa que 33% da variação da TE pode ser explicado pela variação da TDIS; e cerca de 69% restante da variação é não explicada, devido a outros fatores. Por fim, o valor da REMQ da TE para uma TDIS é cerca de 5.06. A alta magnitude pode ser justificada pelas presenças de valores discrepantes, ainda contidos no conjunto de dados da amostra.

Figura 8 - Equação da reta e a dispersão dos dados de teste.




Fonte: Desenvolvido pelo autor.

## 5 Considerações finais

O presente trabalho buscou investigar as relações dos indicadores educacionais, disponibilizados pelo Censo Escolar de 2017, relacionados com a evasão escolar, no ensino médio, referente ao estado do Pará.

O uso da metodologia CRISP-DM permitiu nortear o processo de DCBD e melhor entender o relacionamento entre as variáveis da base de dados. A etapa de pré-processamento foi importante para compreensão sobre a temática de evasão escolar e para a preparação dos dados para aplicação da MD.

A análise de correlação apontou a TDIS como a variável mais associada com a evasão escolar, tendo a maior concentração das escolas valores entre 38% e 65% de alunos com idade acima da recomendada.

Já a aplicação do modelo de regressão linear, mostrou uma variação alta na estimativa dos valores. Contudo, a justificativa para o resultado obtido é dada pelo fato de a regressão linear ser sensível a valores discrepantes, ainda contidos na base de dados. Embora exista um pré-processamento para tratar os valores discrepantes entre a TE e TDIS, essa tratativa não se mostrou suficiente.

Com base no exposto, a pesquisa apresentou que a utilização da MDE pode ser aplicada a fim de gerar conhecimento e servir como auxílio para soluções de problemas e suporte para o desenvolvimento de mecanismos em apoio ao ensino.

Dessa forma, como trabalhos futuros, pretende-se utilizar métodos robustos com o intuito de analisar os comportamentos dos dados discrepantes e explorar outras granularidades de escolaridade, como regional e nacional, além de realizar uma análise histórica desses dados.

## Referências

BAKER, R. S. J.; ISOTANI, S.; CARVALHO, A. M. J. B. Mineração de Dados Educacionais: Oportunidades para o Brasil. *Revista Brasileira de Informática na Educação*, v. 19, n. 2, 2011.

BITTENCOURT, I. M.; MERCADO, L. P. Evasão nos cursos na modalidade de educação a distância: estudo de caso do Curso Piloto de Administração da UFAL/UAB. *Ensaio: Avaliação e Políticas Públicas em Educação*, v. 22, n. 83, Rio de Janeiro, apr-jun, 2014.

BRITO D. M. et *al*. Identificação de estudantes do primeiro semestre com risco de evasão através de técnicas de Data Mining. *Nuevas Ideas en Informática Educativa*, p. 459 – 463, 2015.

CALIXTO, K. E. A.; SEGUNDO, C. V. N.; GUSMÃO, R. P. Mineração de dados aplicado a educação: um estudo comparativo acerca das características que influenciam a evasão escolar. *In VI Congresso Brasileiro de Informática na Educação*, p. 1447 - 1456. Recife, PE, 2017.

CHAPMAN, P. et *al*. *CRISP-DM 1.0: step-by-step.* (2000). Disponível em https://www.the-modeling-agency.com/crisp-dm.pdf. Acesso em 18 out. 2018.

COLPANI, R. Educação a Distância: Identificação dos Fatores que contribuíram para a Evasão dos Alunos no Curso de Gestão Empresarial da Faculdade de Tecnologia de Mococa. EAD EM FOCO, [S.l.], v. 8, n. 1, ago. 2018. ISSN 2177-8310. Disponível em: <http://eademfoco.cecierj.edu.br/index.php/Revista/article/view/688>. Acesso em: 14 out. 2018. doi:https://doi.org/10.18264/eadf.v8i1.688.

GOLDSHMIDT, R.; PASSOS, E.; BEZERRA, E. *Data Mining: conceitos, técnicas, algoritmos, orientações e aplicações*. Rio de Janeiro: Elsevier, 2015.

INEP. Evasão no ensino médio supera 12%, revela pesquisa inédita. 2017a. Disponível em http://portal.mec.gov.br/ultimas-noticias/211-218175739/50411-evasao-no-ensino-medio-supera-12-revela-pesquisa-inedita. Acesso em 14 out. 2018.

INEP. Inep divulga dados inéditos sobre fluxo escolar na educação básica. 2017b. Disponível em http://portal.inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/inep-divulga-dados-ineditos-sobre-fluxo-escolar-na-educacao-basica/21206. Acesso em 14 out. 2018.

INEP. Indicadores Educacionais. 2017c. Disponível em http://portal.inep.gov.br/web/guest/indicadores-educacionais. Acesso em 20 out. 2018.

INEP. Indicadores Educacionais do Censo Escolar 2017 estão disponíveis para consulta. 2017d. Disponível em http://portal.inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/indicadores-educacionais-do-censo-escolar-2017-estao-disponiveis-para-consulta/21206. Acesso em 20 out. 2018.

LARSON, F.; FARBER, B. *Estatística aplicada*. 4ed. São Paulo: Pearson Prentice Hall, 2010.

LOBO, M. B. de C. M. Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. *ABMES Cadernos*, n. 25, 2012.

MACHADO, R. D. et *al*. (2015). Estudo bibliométrico em mineração de dados e evasão escolar. *In XI Congresso Nacional de Excelência em Gestão*. Disponível em http://www.inovarse.org/node/4003. Acesso em 14 out. 2018.

MAIA, M. C.; MEIRELLES, F. S.; PELA, S. K. (2004). Análise dos índices de evasão nos cursos superiores a distância do Brasil. Disponível em http://www.abed.org.br. Acesso em 14 out. 2018.

MANHÃES, L. M. B.; CRUZ, S. M. S.; COSTA, R. J. M.; ZAVALETA, J.; ZIMBRÃO, G. Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. *In XVII Simpósio Brasileiro de Informática na Educação*, 2011.

NASCIMENTO, R. L. S.; CRUZ JÚNIOR, G. G.; FAGUNDES, R. A. A. Mineração de Dados Educacionais: Um Estudo Sobre Indicadores da Educação em Bases de Dados do INEP. *Novas Tecnologias na Educação*, v. 16, n. 1, 2018.

OLIVEIRA JÚNIOR, J. G.; NORONHA, R. V.; KAESTNER, C. A. A. Análise de Correlação de Evasão de Cursos de Graduação com o Empréstimo de Livros em Biblioteca. *In 3° Congresso Brasileiro de Informática na Educação*, 2014.

OLIVEIRA JÚNIOR, J. G.; NORONHA, R. V.; KAESTNER, C. A. A. Método de Seleção de Atributos Aplicados na Previsão da Evasão de Cursos de Graduação. *Revista de Informática Aplicada*, v. 13, n. 2, p. 54 – 67, 2017.

PAZ, F. J.; CAZELLA, S. C. Identificando o perfil de evasão de alunos de graduação através da Mineração de Dados Educacionais: um estudo de caso de uma Universidade Comunitária. *In VI Congresso Brasileiro de Informática na Educação*, 2017.

REZENDE, S. Mineração de Dados. *In XXV Congresso da Sociedade Brasileira de Computação*. Minicurso, Universidade do Vale do Rio dos Sinos – Unisinos, São Leopoldo, Julho de 2005.

RIGO, S. J. *et al.* Aplicações de Mineração de Dados Educacionais e Learning Analytics com foco na evasão escolar: oportunidades e desafios. *Revista Brasileira de Informática na Educação*, v. 22, n.1, 2014.

RODRIGUES, R. L. et *al*. A literatura brasileira sobre mineração de dados educacionais. *In 3° Congresso Brasileiro de Informática na Educação*. 2014.

ROMERO, C.; VENTURA, S. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 3, n.1, p. 12-27, 2013.

SILVIA, J. L. D.; NUNES, I. D. Mineração de Dados Educacionais como apoio para a classificação de alunos do Ensino Médio. *In Anais do XXVI Simpósio Brasileiro de Informática na Educação*, p. 1112 – 1121, 2015.

WITTEN, I. H.; FRANK, E. *Data mining: practical machine learning tools and techniques with java implementations*. San Francisco: Morgan Kaufmann, p. 369, 2000.

*Recebido em novembro de 2018*
*Aprovado para publicação em dezembro de 2018*

**Rogério Colpani**
Centro Universitário da Fundação Educacional Guaxupé – Unifeg
rocolpani@gmail.com