

Identificando o perfil de evasão de alunos de graduação através da Mineração de dados Educacionais: um estudo de caso de uma Universidade Comunitária

Fábio Josende Paz^{1,2}, Silvio César Cazella^{1,3}

¹Pós Graduação em Informática na Educação - Universidade Federal do Rio Grande do Sul (UFRGS)

²Universidade da Região da Campanha (URCAMP)

³Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA)

Fabiopaz@urcamp.edu.br, silvio.cazella@gmail.com

Abstract: This article presents a case study that sought to identify profiles of students with potential for evasion in a community university through the application of the Knowledge Discovery in Database (DCBD) process. The data were collected and refer to undergraduate students enrolled in the second semester of 2016. Data were mined applying the data classification task, using the J48 algorithm and the decision tree technique. The results obtained showed an accuracy of 90%, and it is possible to identify profiles of students evading the Higher Education Institution (HEI) in their campuses and courses.

Resumo: Este artigo apresenta um estudo de caso que buscou identificar perfis de alunos com potencial de evasão em uma universidade comunitária através da aplicação do processo de Descoberta de Conhecimento em Base de Dados (DCBD). Os dados analisados foram coletados e referem-se a alunos de graduação regularmente matriculados no segundo semestre de 2016. Os dados foram minerados aplicando a tarefa de classificação de dados, utilizando a técnica de árvores de decisão através do algoritmo J48. Os resultados alcançados apresentaram acurácia de 90%, sendo possível identificar perfis de alunos evasores da Instituição de Nível Superior (IES) em seus campi e cursos.

1 Introdução

A evasão escolar é um tema complexo e que está interferindo na gestão universitária por todo o Brasil, principalmente em Instituições de Nível Superior (IES) privadas e comunitárias (RIBEIRO, 2015). Corroborando Costa *et al.* (2015), afirmam que um grave problema enfrentado pelas universidades brasileiras é o aumento da Evasão escolar. Nesse sentido, este artigo aborda esta temática que segundo diversos estudos (JOHANN, 2012; COSTA *et al.*, 2015; RIBEIRO, 2015; FIGUEIREDO E SALLES, 2017), é um grande desafio a ser superado na Educação Universitária.

Para auxiliar as IES a superar este desafio, técnicas de mineração de dados Educacionais foram utilizadas em vários estudos (MANHÃES *et al.*, 2011; COSTA *et al.*, 2015; PRABHA, SHANAVAS, 2014) permitindo identificar níveis e grupos de alunos, além de perfis de alunos evasores, gerando informação para apoio para gestores destas instituições (PRABHA, SHANAVAS, 2014).

Portanto, o objetivo principal deste artigo é apresentar o resultado de um estudo que buscou identificar perfis de alunos com potencial de evasão em uma universidade comunitária através da aplicação do processo de Descoberta de Conhecimento em Base de Dados e da mineração de dados educacionais, procurando contribuir com os gestores universitários no planejamento de ações efetivas para a retenção de alunos desta IES. Para auxiliar na pesquisa, foram elaboradas algumas hipóteses: -Hipótese 1: Alunos de semestres iniciais possuem maior tendência a evasão ; -Hipótese 2: Alunos que residem em municípios fora do campus onde estudam tem tendência a evasão; -Hipótese 3: Os incentivos fornecidos (bolsas) têm relação com a evasão.

Este artigo está estruturado em seis seções, incluídas a introdução e Conclusão. Na seção 2, estão os estudos teóricos que embasaram este trabalho, na sequência estão apresentados na seção 3 os materiais e métodos. A seção 4 dedica-se a explicar a aplicação do Processo de Descoberta de Conhecimento em Base de Dados (DCBD), e a seção 5 apresenta a análise dos resultados e discussão.

2 Fundamentação Teórica

Esta seção apresenta os principais conceitos referente aos temas evasão universitária, descoberta de conhecimento em base de dados, mineração de dados educacionais, apresentando trabalhos relacionados.

2.1 Evasão Universitária

É considerada evasão de curso quando o aluno rompe o vínculo jurídico estabelecido, não renovando seu compromisso de continuar no estabelecimento de ensino, uma vez que ao não renovar a matrícula rompe-se o vínculo existente entre aluno e instituição (JOHANN, 2012). Segundo Lobo (2011) há tempos o problema da evasão é sério no Brasil, em pesquisa realizada pelo Instituto Lobo para o Desenvolvimento da educação, da ciência e da Tecnologia, as perdas financeiras com a evasão no ensino superior em 2009, giram em torno de R\$ 9 bilhões. Ainda conforme Costa *et al.* (2015), a evasão escolar é um problema social, acadêmico e econômico (FIGUEIREDO E SALLES, 2017).

Neste contexto, também estão às universidades comunitárias que constituem um segmento de IES que tem fins educacionais, serviços sociais e á comunidade (BRASIL, 2013), são comumente confundidas com as particulares, devido à origem de seus recursos serem oriundas de mensalidades, porém não tem fins lucrativos (VEIGA *et al.*, 2014). No semestre 2016/2 da Universidade comunitária estudada teve uma evasão de 9,13%, que representa 420 alunos, utilizando o cálculo de que o aluno em média por ano custa R\$ 12.000,00 e se multiplicarmos por média de três anos (a evasão geralmente acontece nos primeiros semestres) chegamos a mais de R\$ 15 milhões que deixaram de ser arrecadados pela instituição.

2.2 Mineração de dados educacionais

Segundo Baker *et al.* (2011) a mineração de dados educacionais (MDE) tem como principal objetivo o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais, e que esta apesar de ser uma área recente de pesquisa, possui um grande potencial para melhorar a qualidade do ensino. Ainda a mineração de dados educacionais pelo seu enorme potencial de transformação, pode ser usada para aprender, prever e compreender o comportamento da aprendizagem, além de ajudar a desenvolver tecnologias melhores e mais inteligentes para apoiar alunos e professores (BAKER, 2014).

Com a utilização da MDE é possível de forma mais eficaz que a manual, compreender os alunos, como eles aprendem, como ocorre a aprendizagem, além de outros fatores que influenciam na aprendizagem. (BAKER *et al.*, 2011)

Para analisar a evasão será aplicado o processo de Descoberta de Conhecimento em Base de Dados (DCDB), que busca identificar padrões válidos e úteis de comportamento, onde a Mineração de Dados constitui-se em uma das suas etapas. O DCDB que procura extrair conhecimento de grandes volumes de dados para apoiar a análise de perfis de alunos para geração de informações e alertas que possam apoiar atividades preventivas pelos professores, embasando e agilizando as tomadas de decisões [COSTA *et al.* 2015]. A Figura 1 apresenta resumidamente as 05 etapas envolvidas no processo DCDB.

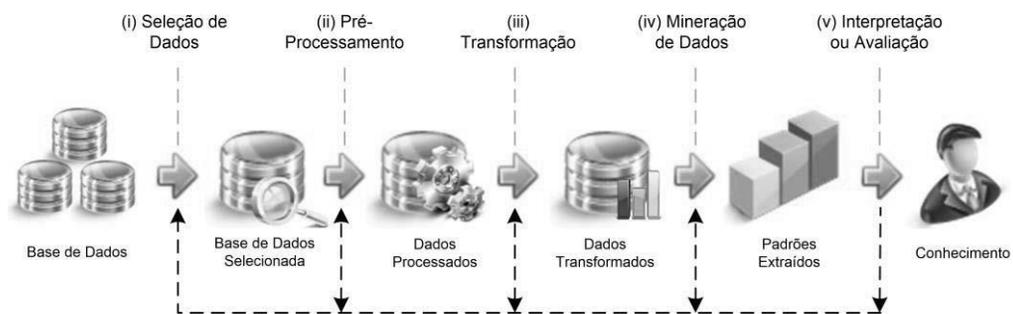


Figura 1: Etapas do processo de DCBD.

Fonte: Rigo *et al.*, 2014

Na sequência, apresenta-se uma breve descrição sobre cada etapa do processo de DCBD, já apresentado na Figura 1.

i) Seleção de Dados: Envolve a compreensão do domínio e dos objetivos da tarefa, juntamente com a criação do conjunto de dados envolvendo as variáveis necessárias.

ii) Pré-Processamento: Os dados disponíveis, geralmente não estão prontos para análise, eles podem apresentar inconformidades como: falta de dados em alguns campos, duplicidade, entre outros, por isso é necessário uma limpeza, tratamento e redução do volume de dados.

iii) Transformação: Os dados já processados serão transformados em um padrão que possam ser utilizados nos algoritmos de extração de padrões.

iv) Mineração de Dados: é a fase que transforma dados puros em informação útil, ou seja, caracteriza-se pela descoberta de conhecimento na aplicação do algoritmo de mineração de dados.

v) Interpretação ou Avaliação: essa etapa é onde os conhecimentos encontrados são interpretados e poderão ser usados no suporte ao processo de tomada de decisão na área de domínio da aplicação.

2.3 Trabalhos Relacionados

A utilização de técnicas de mineração de dados sobre dados educacionais é relativamente recente conforme destaca Baker *et al.* (2009, 2011), porém recentemente alguns trabalhos sobre a evasão em universidades tem obtido resultados satisfatórios, na sequencia apresenta-se alguns desses trabalhos.

Manhães *et al.* (2011) aplicaram 10 algoritmos de classificação que são técnicas de mineração de dados para identificar precocemente perfis de evasão sobre uma base de dados de alunos de graduação do curso de Engenharia Civil da Escola Politécnica da UFRJ, obtendo êxito em seu estudo com taxas de até 80% de acurácia.

Rigo *et al.* (2014) em seu estudo durante 18 meses em cursos de graduação na modalidade de Educação a Distância, buscou entender os fatores envolvidos na evasão escolar com a inclusão de elementos associados ao papel exercido pelos docentes, como resultados positivos cita-se a identificação de perfis de alunos evasores com taxa de acurácia de 87%, nesse trabalho também foram realizadas intervenções pedagógicas, com obtenção de taxas médias de 11% na redução da evasão.

O presente trabalho distingue-se destes, pois busca compreender a evasão de uma Universidade Comunitária e não apenas cursos isolados, buscando compreender se o incentivo que o aluno recebe como bolsas e auxílios tem influência na evasão.

3 Materiais e Métodos

Esta pesquisa caracteriza-se por ser uma pesquisa exploratória de acordo com seu objetivo, de natureza aplicada, com abordagem quali-quantitativa uma vez que métricas são assumidas para avaliar os modelos obtidos. O método de pesquisa aplicado foi estudo de caso único. Os materiais utilizados na pesquisa foram provenientes dos cursos de graduação de uma Universidade Comunitária do Rio Grande do Sul disponibilizados pelo setor de Tecnologia de Informação (TI) da respectiva IES.

4 Aplicando o Processo Descoberta de Conhecimento em Base de dados

A seguir será descrita a aplicação detalhada do processo de Descoberta de Conhecimento em Base de Dados (DCBD) nos dados objetos da análise.

4.1 Seleção da base de dados

A próxima etapa do processo DCDB é a seleção dos dados, nesse estudo foram utilizados os dados amostrais de 4697 alunos matriculados em cursos de graduação no semestre 2016/2 de todos os campi da Universidade objeto de estudo, os alunos não foram identificados nestes dados, garantindo assim a privacidade dos mesmos.

Os dados foram coletados de doze tabelas, contendo 322 atributos, o fluxo da seleção de dados aplicados nessa pesquisa envolveu os dados referentes a possíveis influências na evasão dos alunos.

4.2 Pré-Processamento

Para escolha da base de dados não houve aleatorização, e sim escolha da amostra com maior percentual de não *missing data*. A base de dados foi analisada por especialistas educacionais para identificar quais informações poderiam ser relevantes para a evasão, e no primeiro momento optou-se por usar os seguintes dados de cada aluno (atributos a serem analisados): a) Currículo do aluno; b) Campus; c) Incentivo; d) Data de Nascimento; e) Semestre atual; f) Município onde reside; g) Evasão. Foram retiradas instâncias com dados incompletos, finalizando a base em 4601 instâncias. Em seguida para facilitar o entendimento e análise dos dados optou-se por transformar o atributo Data de Nascimento para Idade, também foram agrupados incentivos do mesmo tipo, porém com percentuais diferentes. Adaptou-se os dados da Evasão para que este fosse o atributo objetivo da mineração (os demais configuraram os atributos preditivos). Ainda para que fosse possível encontrar regras significativas os semestres foram agrupados em semestres iniciais 1º e 2º (1) e demais semestres. Importante explicar a regra do atributo Evasão, no qual são considerados alunos evasores aqueles que não se matricularam no semestre 2017/1 e não se formaram em 2016/2.

Os dados selecionados foram convertidos em arquivo CSV para serem operados na ferramenta de mineração de dados denominada WEKA (*Waikato Environment for Knowledge Analysis*) (HALL *et al.*, 2009) a qual foi escolhida para condução deste estudo.

4.3 Mineração de Dados

Na etapa de mineração dos dados foi utilizada a tarefa de classificação e o algoritmo escolhido para realizar tal tarefa foi o J48, que permite visualizar o modelo preditivo, em formato de árvore de decisão. Como não se pretendia comparar resultados de performance de algoritmos e sim obter um modelo simples e inteligível ao especialista, optou-se pelo algoritmo J48. Foram utilizados diversos conjuntos de treinamento de acordo com o objetivo de cada mineração as quais serão explicados posteriormente na seção análise de Resultados, no entanto, para todos os treinamentos foi utilizado o método de teste denominado *cross-validation*, assumindo o valor de 10 *folds*.

4.4 Pós-Processamento

Ao final das minerações de dados, obtiveram-se modelos preditivos que foram analisados através das métricas oferecidas pela ferramenta WEKA.

5. Análise de Resultados

Para melhor apresentação e entendimento dos resultados optou-se por apresentá-los de acordo com as hipóteses apresentadas no início deste artigo.

-Hipótese 1: Alunos de semestres iniciais possuem maior tendência a evasão ;
-Hipótese 2: Alunos que residem em municípios fora do campus onde estudam tem tendência a evasão; -Hipótese 3: Os incentivos fornecidos (bolsas) têm relação com a evasão.

Com o intuito de responder as hipóteses já mencionadas, e também aplicar as tarefas de mineração identificadas para o domínio de dados criado, foram realizados dois experimentos, listados a seguir:

- A) Analisar a evasão nos cursos com mais alunos da instituição;
- B) Analisar a evasão nos dois campi com mais alunos da instituição;

Devido à quantidade de currículos da Universidade em todos os campi optou-se por utilizar seis (06) currículos que representam 73% dos alunos da instituição, o treinamento deste modelo foi realizado com um conjunto de 3360 instâncias. O modelo preditivo classificou corretamente 3066 (91.25%) instâncias e incorretamente 294 (8.75%), analisando-se a Matriz de Confusão (*Confusion Matrix*) observa-se que 3066 alunos foram identificados corretamente pelo algoritmo J48 e 294 alunos foram incorretamente classificados gerando falsos positivos (24 alunos incluídos na classe não evasão). A Figura 2 apresenta o resultado obtido confirmando a percepção inicial de que o incentivo e o currículo do aluno estão relacionados à tendência de evasão.

The screenshot shows the Weka Explorer interface with the following data:

Classifier: J48-S-C 0.25-M 2

Test options: Cross-validation Folds: 10

Classifier output:

```

time taken to build model: 0.01 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 3066 91.25 %
Incorrectly Classified Instances 294 8.75 %
Kappa statistic 0.2077
Mean absolute error 0.1543
Root mean squared error 0.2801
Relative absolute error 90.6601 %
Root relative squared error 96.1122 %
Total Number of Instances 3360
  
```

Detailed Accuracy By Class:

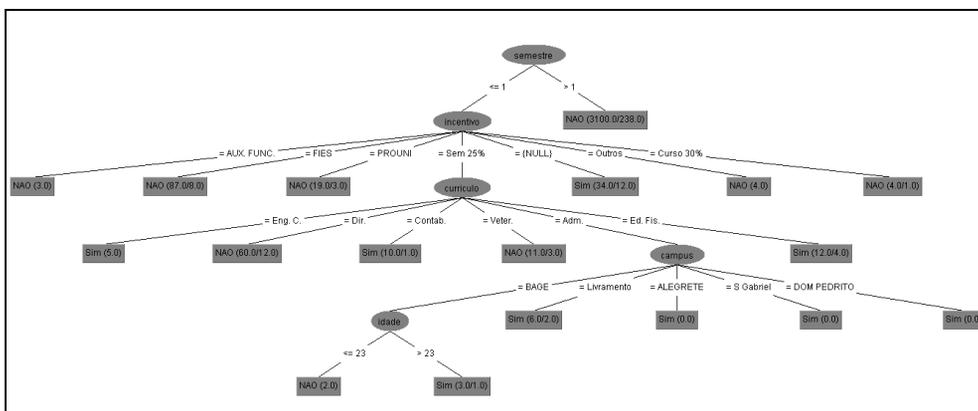
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Weighted Avg.	0,992	0,008	0,918	0,992	0,954	0,277	0,548	0,210	Não
	0,143	0,008	0,652	0,143	0,234	0,277	0,548	0,210	Sim

Confusion Matrix:

	a	b	← classified as
3021	24		a = Não
270	45		b = Sim

Figura 2: Resultado do experimento A**Fonte: WEKA. Adaptado pelo Autor**

A figura 3 apresenta a árvore de decisão gerada pela mineração dos dados. Onde o atributo semestre representa o primeiro nodo da árvore gerada, demonstrando que alunos que estão em semestres avançados não evadem, porém alunos de semestres iniciais e que não possuem incentivo (NULL), tem uma alta tendência à evasão. Ainda pode-se visualizar que alunos de semestres iniciais com incentivo semestre cheio 25% (Sem 25%) dos currículos Engenharia Civil (Eng. C), Ciências Contábeis (Contab.), Educação Física (Ed. Fis.) e Administração (Adm.) com exceção de alunos com menos de 24 anos do campus Bagé possuem tendência à evasão.

**Figura 3: Árvore de decisão do experimento A****Fonte: WEKA. Adaptado pelo Autor**

O Experimento B diz respeito aos campi com mais alunos da instituição os quais são Bagé com 2437 alunos e Alegrete com 839, devido à quantidade de currículos existentes no campus Bagé optou-se por apresentar os 06 cursos com mais alunos e agrupar o restante dos currículos em Outros. Portanto, o treinamento do modelo Bagé foi realizado com um conjunto de 2437 instâncias, o modelo preditivo classificou corretamente 2228 (91,42%) instâncias e incorretamente 209 (8,58%).

A árvore de decisão gerada pela mineração dos dados apresenta-se na Figura 4. Onde o atributo semestre representa o primeiro nodo da árvore gerada, demonstrando que alunos que estão não estão em semestres iniciais não tem tendência à evasão, no entanto a evasão novamente está nos semestres iniciais onde algumas regras de tendência à evasão foram formadas:

- Semestres iniciais + cursos com desconto de 30% (Cursos 30%);
- Sem incentivo (NULL) + Município de Bagé ou não informada;

- Incentivo Semestre Cheio 25% (Sem 25%) + Currículos de Eng. C; Ed. Fis. Ou Outros;
- Incentivo Semestre Cheio 25% (Sem 25%) + Currículo Adm. + idade acima de 23 anos.
- Incentivo Semestre Cheio 25% (Sem 25%) + Currículo Veter. (Medicina Veterinária) + idade menor que 19 anos.

Importante salientar que não se comprovou a hipótese inicial de que alunos que são de outros Municípios (Não Alegrete) teriam tendência à evasão, devido a viagens diárias para frequentar a Universidade.

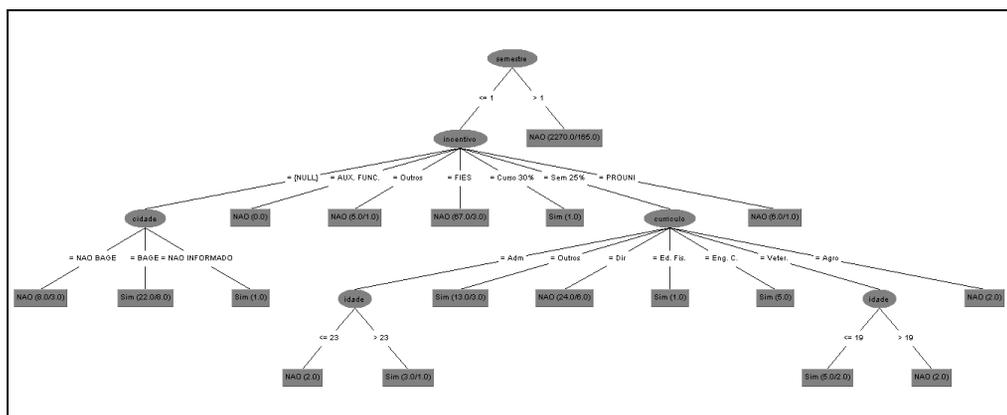


Figura 4: Árvore de Decisão do experimento B - Bagé

Fonte: WEKA. Adaptado pelo Autor

O treinamento do modelo Alegrete no experimento B foi realizado com um conjunto de 839 instâncias, o modelo preditivo classificou corretamente 772 (92,01%) instâncias e incorretamente 67 (7,99%), a árvore de decisão gerada apresenta-se na Figura 5, onde mais uma vez os semestres iniciais apresentam tendência à evasão nos currículos de Ciências Contábeis (Contábeis) e na Educação Física (Ed. Fis.) dos alunos que moram em Alegrete ou não informaram o Município de sua residência.

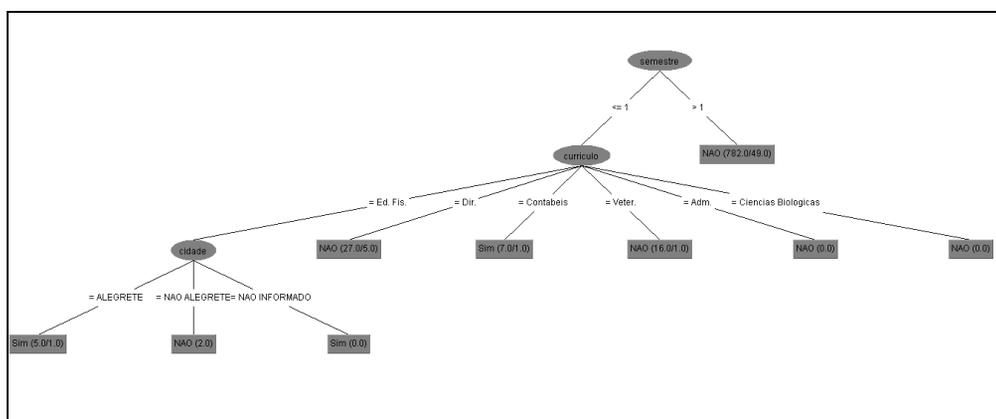


Figura 5: Árvore de Decisão do experimento B - Alegrete

Fonte: WEKA. Adaptado pelo Autor

5 Considerações finais

A evasão em cursos de graduação é um problema complexo, tanto econômico quanto social, na Universidade comunitária estudada não é diferente, este estudo através da mineração de dados possibilitou evidenciar perfis de alunos evasores na Universidade, em seus campi e cursos, gerando conhecimentos valiosos para auxiliar no processo de tomada de decisão. Os experimentos realizados tiveram um alto grau de precisão atingindo acurácia superiores a 90%.

Dentre os principais resultados evidencia-se a confirmação da percepção inicial de que o incentivo e o currículo dos alunos estão diretamente relacionados à tendência de evasão, e que estas ocorrem geralmente nos semestre iniciais dos cursos com alunos sem incentivo. No entanto, não se comprovou a hipótese de evasão em alunos que não moram no Município do campus. Como limitação deste trabalho a análise de apenas um semestre da IES e para trabalhos futuros destaca-se a ampliação da base de dados para mais semestres, a inclusão de novos atributos para novas análises como: renda per capita, se o aluno trabalha e frequência na biblioteca. Como estudo de caso único os resultados desta pesquisa não são generalizáveis.

Referências

- BAKER, R. and YACEF K. (2009), “The State of Educational Data Mining in 2009: A Review and Future Visions.” Pages 3-17. JEDM -Journal of Educational Data Mining, 2009, Volume 1, Issue 1, October.
- BAKER, R., ISOTANI, S., CARVALHO, A. (2011), Mineração de Dados Educacionais: Oportunidades para o Brasil. Revista Brasileira de Informática na Educação, 19(2), p. 3-13.
- BAKER, R. S. J. (2014), Educational Data Mining: An Advance for Intelligent Systems in Education. IEEE Intelligent Systems, 29 (3), p. 78-82.
- BRASIL. (2017) Lei No 12.881, de 12 de novembro de 2013. Da definição, qualificação, prerrogativas e finalidades das instituições comunitárias de educação superior. Disponível em http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2013/Lei/L12881.htm. Acesso 10 de julho de 2017.
- COSTA, Susane Santos da; CAZELLA, Silvio; RIGO, Sandro José. (2015), Minerando dados sobre o desempenho de alunos de cursos de educação permanente em modalidade EAD: Um estudo de caso sobre evasão escolar na UNA-SUS. RENOTE, v. 12, n. 2.
- FIGUEIREDO, Natália Gomes da Silva; SALLES, Denise Medeiros Ribeiro. (2017), Educação Profissional e evasão escolar em contexto:

motivos e reflexões. Ensaio: aval. pol. públ. Educ., Rio de Janeiro, p. 1-37.

HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P. and WITTEN. I.H. (2009), “The WEKA Data Mining Software: An Update” SIGKDD Explorations, Volume 11, Issue 1.

JOHANN, C. C. (2012), Evasão escolar no Instituto Federal Sul-Rio-Grandense: um estudo de caso no campus Passo Fundo. Dissertação (Mestrado em Educação) – Universidade de Passo Fundo, Passo Fundo.

LOBO, Maria Beatriz de Carvalho Melo. (2011), Esclarecimentos metodológicos sobre os cálculos de evasão. Instituto Lobo para o Desenvolvimento da Educação, da Ciência e da Tecnologia. Mogi das Cruzes, SP:. Disponível em: <http://www.institutolobo.org.br/imagens/pdf/artigos/art_087.pdf> Acesso em: 02 agosto de 2017.

MANHÃES, Laci Mary Barbosa; DA CRUZ, Sérgio Manuel Serra; COSTA, Raimundo J. Macário; ZAVALA, Jorge; ZIMBRÃO, Geraldo. (2011), Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. Anais do XXII SBIE - XVII WIE Aracaju, p. 150-159, 21 a 25 de novembro de 2011.

PRABHA, S. Lakshmi; SHANAVAS, A.R. Mohamed. (2014), EDUCATIONAL DATA MINING APPLICATIONS. Operations Research and Applications: An International Journal (ORAJ). Vol. 1, No. 1, p. 1- Agosto.

RIBEIRO, M. A. – (2005), O Projeto Profissional Familiar como Determinante da Evasão Universitária– Um Estudo Preliminar – Revista Brasileira de Orientação Profissional, 6(2), p. 55 – 70.

RIGO, Sandro J.; CAMBRUZZI, Wagner; BARBOSA, Jorge L. V.; CAZELLA, Sílvio C. (2014), Aplicações de Mineração de Dados Educacionais e *Learning Analytics* com foco na evasão escolar: oportunidades e desafios. Revista Brasileira de Informática na Educação, Volume 22, Número 1, p. 132-146.

VEIGA, Luciane da; DREHMER, Clarice Luiza; URNAU, Jeferson Rafael; SILVA, Talita da; LIZOTE, Suzete Antonieta; TERRES, José Carlos. (2014), O QUE É UMA UNIVERSIDADE COMUNITÁRIA? Um estudo sobre o grau de conhecimento dos estudantes de uma Instituição de Ensino Superior. *XII Coloquio Internacional de Gestión Universitaria*, p. 1-15.