Instituto Federal de Educação, Ciência e Tecnologia Fluminense

Programa de Pós-graduação em Sistemas Aplicados à Engenharia e Gestão

IDENTIFICAÇÃO DO COMPORTAMENTO DOS ESTUDANTES EVADIDOS DE CURSOS TÉCNICOS UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

RENATA GOMES CORDEIRO

2017

Instituto Federação de Educação, Ciência e Tecnologia Fluminense
Programa de Pós-graduação em Sistemas Aplicados à Engenharia e Gestão

IDENTIFICAÇÃO DO COMPORTAMENTO DOS ESTUDANTES EVADIDOS DE CURSOS TÉCNICOS UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

RENATA GOMES CORDEIRO

Henrique Rego Monteiro da Hora
(Orientador)

Dissertação submetida como requisito parcial para obtenção do grau de **Mestre** no Programa de Pós-graduação em Sistemas Aplicados à Engenharia e Gestão, Área de Concentração em Sistemas Computacionais.

Campos dos Goytacazes, RJ
Novembro de 2017

Biblioteca Anton Dakitsch
CIP - Catalogação na Publicação

C794i

Cordeiro, Renata Gomes
Identificação do comportamento dos estudantes evadidos de cursos técnicos utilizando técnicas de mineração de dados. / Renata Gomes Cordeiro - 2017.
79 f.: il. color.

Orientador: Henrique Rego Monteiro da Hora

Dissertação (mestrado) -- Instituto Federal de Educação, Ciência e Tecnologia Fluminense, Campus Campos Centro, Curso de Mestrado Profissional em Sistemas Aplicados à Engenharia e Gestão, Campos dos Goytacazes, RJ, 2017.
Referências: f. 79 a 79.

1. Mineração de dados. 2. Evasão. 3. Ensino Técnico. 4. Educação. I. Hora, Henrique Rego Monteiro da, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da Biblioteca Anton Dakitsch do IFF com os dados fornecidos pelo(a) autor(a).

Instituto Federação de Educação, Ciência e Tecnologia Fluminense
Programa de Pós-graduação em Sistemas Aplicados à Engenharia e Gestão

RENATA GOMES CORDEIRO

Dissertação submetida como requisito parcial para obtenção do grau de **Mestre** no Programa de Pós-graduação em Sistemas Aplicados à Engenharia e Gestão, Área de Concentração em Sistemas Computacionais.

APRESENTADO EM 17/11/2017

Angellyne Moço Rangel
Doutora em Sociologia Política – UENF

Breno Fabrício Terra Azevedo
Doutor em Informática na Educação – UFRGS

Carlos Artur de Carvalho Arêas
Mestre em Administração - UFSC

Henrique Rego Monteiro da Hora
Doutor em Engenharia de Produção – UFF
(Orientador)

AGRADECIMENTOS

À minha família por se sentir realizada através das minhas conquistas e por estarem sempre me apoiando.

A Munir, pelas contribuições, parceria e por me transmitir a calma necessária em todos os momentos.

Aos amigos da Diretoria de Gestão de Tecnologia da Informação do Instituto Federal Fluminense pelo apoio necessário durante essa caminhada. Em especial a Natanael por toda ajuda.

Ao meu orientador Henrique da Hora pelos ensinamentos e apoio durante o mestrado.

Ao Instituto Federal Fluminense por fornecer os dados necessários para o cumprimento dessa pesquisa.

LISTA DE FIGURAS

LISTA DE QUADROS

SUMÁRIO

1. APRESENTAÇÃO

A partir do momento em que a informatização começou a ocorrer na educação, tornou-se possível o armazenamento de dados de alunos durante toda a trajetória escolar, desde informações acadêmicas à informações socioeconômicas. Porém, para Machado, Benitez, Corleta e Augusto (2015), bases de dados tornam-se pouco úteis sem utilização de ferramentas para interpretação e análise. A mineração de dados utilizada como ferramenta para análise desses dados torna-se útil e de grande valor pela possibilidade de extrair informações importantes a partir de grandes volumes de dados.

A mineração de dados educacionais é uma área emergente para a descoberta de conhecimento em grandes volumes de dados. É um campo aberto para pesquisa no intuito de tornar o processo de ensino-aprendizagem mais planejado e eficaz para os alunos e para a sociedade (Mehta & Buch, 2016)**.** Nessa área, a abordagem da evasão escolar tem tido o intuito de propiciar a identificação precoce de alunos com possibilidade de evadir.

Considerando o impacto da educação na vida de cada indivíduo e da sociedade, é importante que sejam analisados não só os alunos que ingressam em novos cursos, mas também aqueles que não concluem e abandonam o curso caracterizando evasão.

De acordo com Márquez-Vera *et al.* (2016), quanto mais cedo forem identificados os estudantes propensos a evadir o curso, maiores são as chances de sucesso da política de permanência escolar. Hoed (2016) afirma que é uma necessidade da instituição conter a evasão, por estar diretamente associada à perda de recursos financeiros. A evasão de um aluno em qualquer nível da educação pode custar não só a ele e sua família, também pode deixar marcas no futuro da sociedade e no crescimento da instituição. E estudos que permitam detectar alunos propensos a evadir o curso tornam possíveis a elaboração de políticas mais focadas que o incentivem a permanecer.

Cunha, Moura e Analide (2016) apontam que a evasão e a reprovação estão relacionadas às áreas de conhecimento do aluno, ao nível de educação e às metodologias de ensino e aprendizado. Métodos e ferramentas que tornem possíveis a análise de vários fatores em conjunto proporcionam uma visão ampla sobre as causas de problemas.

Considerando a problemática da evasão e o grande volume de dados armazenados que podem propiciar informação valiosas para as instituições de ensino, este trabalho busca explorar os dados na área educacional através da utilização de técnicas de mineração de dados e

identificar o comportamento de alunos evadidos em cursos de nível técnico através de um estudo de caso.

Sendo a questão desta pesquisa definida da seguinte forma:

Qual o comportamento dos alunos que evadem os cursos técnicos nas modalidades concomitante e subsequente no Instituto Federal Fluminense?

A presente pesquisa está estruturada em cinco capítulos. Este primeiro capítulo traz uma contextualização do tema, bem como a apresentação e definição da questão de pesquisa. No segundo capítulo é apresentado um estudo bibliométrico de trabalhos que abordam a evasão escolar utilizando mineração de dados. A partir dos trabalhos encontrados no primeiro artigo foi realizada uma análise sistemática, com o intuito de consolidar os dados, métodos e metodologias mais utilizadas para identificação de alunos evadidos utilizando mineração de dados, resultado que está apresentado no terceiro capítulo. No quarto capítulo, é realizado um estudo de caso no Instituto Federal Fluminense e são identificados os comportamentos de alunos evadidos nas modalidades concomitante e subsequente do ensino técnico. Finalizando, o quinto capítulo traz as considerações finais deste estudo, seguido pela lista de referências utilizadas na pesquisa.

2. ARTIGO 1 - UM ESTUDO BIBLIOMÉTRICO SOBRE MINERAÇÃO DE DADOS EDUCACIONAIS COM FOCO NA EVASÃO NO ENSINO TÉCNICO

2.1. Resumo

Contexto: O grande volume de dados armazenados na área educacional requer a utilização de técnicas que propiciem a interpretação e análise desses dados com objetivo de agregar valor à gestão e aos educadores. A mineração de dados educacionais é uma área emergente para a descoberta de conhecimento em grandes volumes de dados.
Objetivo: O objetivo desta pesquisa é analisar os trabalhos publicados na área da mineração de dados educacionais com foco na evasão escolar.
Metodologia: É realizado um estudo bibliométrico na base de dados *Scopus* a partir da definição de conceitos chave com o intuito de analisar a produção acadêmica na área pesquisada.
Resultados: A partir das buscas combinando as quatro palavras chave definidas, são realizadas análises sobre principais autores, trabalhos mais citados, periódicos e conferências além de uma análise cronológica de produção
Conclusões: É verificada a ausência de trabalhos que abordem a utilização de mineração de dados para análise da evasão no ensino técnico. Mesmo para as outras buscas realizadas foi verificado que é uma área com potencial a ser explorado em que a maioria das publicações são posteriores ao ano 2000.

Palavras-chave: Mineração de dados; Educação; Evasão; Ensino Técnico.

2.2. Abstract

Context: *The large volume of data stored in the educational area requires the use of techniques that allow the interpretation and analysis of such data in order to add value to the management and to educators. Educational data mining is an emerging area for knowledge discovery in large volumes of educational data*
Objective: *The objective of this research is to analyze the published works in the field of educational data mining with focus on truancy.*
Methodology: *A Bibliometric study on Scopus database from the definition of key terms in order to analyze the academic production in the area searched.*
Results: *From the search combining the four keywords defined, are performed analyses on major authors, most cited papers, journals and conferences as well as a chronological analysis of production*
Conclusions: *It is verified the absence of works that address the use of data mining to analysis of avoidance in technical education. Same for the other searches checked which is an area with potential to be explored when most publications are post-year 2000.*

Keywords: Data mining; education; Dropout; Technical education.

2.3. Introdução

A educação tem impacto em aspectos sociais, econômicos e culturais na sociedade. Rigo, Cazella e Cambruzzi (2012) e Cunha, Moura e Analide (2016) apontam a evasão escolar como um desafio a ser superado na área da educação. Sendo de grande importância estudos que abordem o tema principalmente com análises e propostas que contribuam para um tratamento.

Com o desenvolvimento da área de tecnologia da informação torna-se possível o armazenamento de dados de alunos durante toda a trajetória escolar, desde informações acadêmicas à informações socioeconômicas. Dado o grande volume de dados, torna-se necessária a utilização de técnicas que contribuam para a análise dos mesmos. Para Machado, Benitez, Corleta e Augusto (2015), bases de dados tornam-se pouco úteis sem a utilização de ferramentas de interpretação e análise. A mineração de dados propicia técnicas que podem ser utilizadas para análises que contribuam para o trabalho de gestores.

A partir daí surge a área de pesquisa mineração de dados educacionais. É uma área emergente para a descoberta de conhecimento em grandes volumes de dados educacionais. No mundo, em diferentes países, pesquisadores têm se esforçado para encontrar padrões e fatores que possam contribuir para o aperfeiçoamento da educação (Mehta & Buch, 2016). Na mineração de dados educacionais a evasão escolar tem sido abordada com o intuito de propiciar a identificação precoce de alunos com possibilidade de evadir.

O objetivo desta pesquisa é analisar os trabalhos publicados na área da mineração de dados educacionais com foco na evasão escolar. Buscando verificar os autores mais citados, periódicos e os anos em que houveram mais publicações. Além disso busca-se verificar se existem trabalhos utilizando mineração de dados com foco em evasão escolar no ensino técnico. Acredita-se que através desta pesquisa seja possível obter um panorama sobre as pesquisas na área que une mineração de dados, educação e evasão. Para Araújo Júnior, Perucchi e Lopes (2013), a análise quantitativa da produção de uma área pode determinar tendências, aeo conhecimento sobre essa produção aumenta a possibilidade de serem feitas inferências significativas para o desenvolvimento futuro.

2.4. Metodologia

Com o intuito de identificar trabalhos que abordem a evasão escolar no ensino técnico utilizando técnicas de mineração de dados é realizado um estudo bibliométrico. Para a

elaboração desta pesquisa é utilizada como base a metodologia proposta por Costa (2010). São executadas 6 etapas descritas a seguir:

1. Definição da amostra da pesquisa;
2. Pesquisa na amostra com uso de palavras-chave e seus tesauros;
3. Identificação dos periódicos e conferências com maior número de artigos publicados;
4. Identificação dos autores com maior número de publicações;
5. Artigos com maior número de citações;
6. Levantamento da cronologia da produção;

Na primeira etapa, é definida como amostra os artigos indexados na base de conhecimento *Scopus*. Esta escolha se deve à representatividade e abrangência da base de dados que contem artigos de conferência, periódicos, anais, entre outros. Sendo que para compor o conjunto de trabalhos analisados, foram selecionados os artigos de periódicos e conferências. Em relação ao recorte temporal, a pesquisa contempla todos os trabalhos publicados até 2016.

Para a segunda etapa, são definidas as palavras-chave, os tesauros e termos correspondentes para a realização da pesquisa na base Scopus. A partir da seleção das palavras-chave: *data mining*, *education*, *dop out* e *trade*, é realizada uma consulta no site Thesaurus (Thesaurus, 2013) para a definição dos tesauros.

Na terceira e quarta etapas, são identificados os periódicos e conferências além dos autores das publicações, adotando como recorte, os periódicos e os autores com duas ou mais publicações.

Na quinta etapa, são identificados os autores e os trabalhos com mais citações. São considerados os cinco trabalhos mais citados e são apresentados o título, os autores, o ano de publicação, o periódico ou conferência em que foi publicado e o número de citações.

Na última etapa da metodologia, é realizado um estudo cronológico dos trabalhos produzidos. É apresentada a quantidade de publicações por ano em formato gráfico.

2.5. Análise Bibliométrica

2.5.1. Pesquisa na amostra com uso de palavras-chave e seus tesauros

Este trabalho compõe uma pesquisa que objetiva utilizar técnicas de mineração de dados na problemática da evasão escolar, mais especificamente, no ensino técnico. Para a realização da segunda etapa, foram definidos, primeiramente, os conceitos que definem o método, o objeto de pesquisa e o objetivo: mineração de dados (método), educação (objeto de pesquisa), evasão (objeto de pesquisa) e ensino técnico (objetivo). A partir daí foram definidas as palavras-chave: *data mining, education, dropout* e *trade.* Além das palavras-chaves foram definidos seus tesauros e termos que também as representam na literatura, apresentados na Figura 2.1.

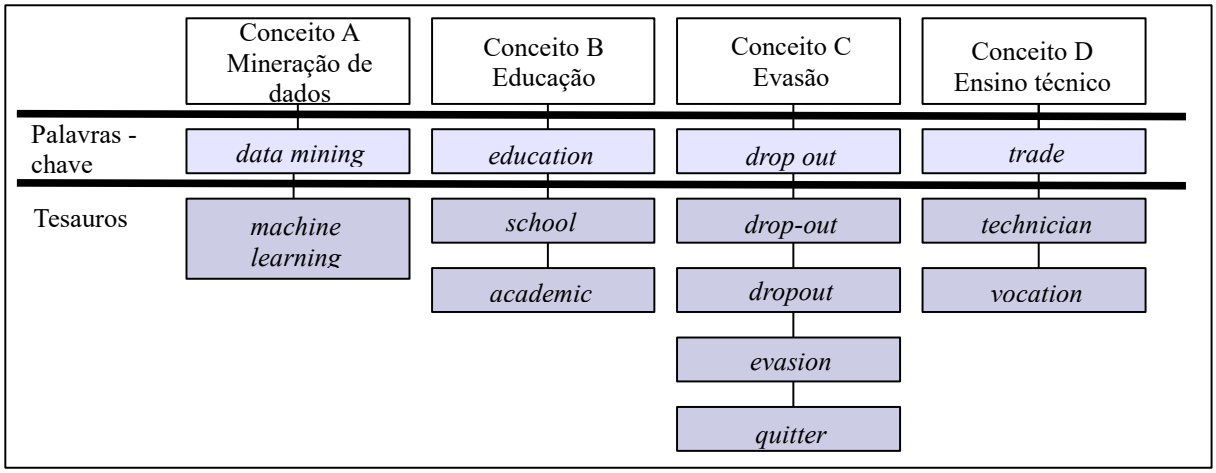

Figura 2.1 - Palavras-chave, tesauros e termos correspondentes. Fonte: Elaboração própria.

Como demonstrado na Figura 2.1, foi feita também uma organização classificando cada conceito como A, B, C e D. Essa classificação continuará a ser utilizada no decorrer deste trabalho.

Após a definição dos conceitos a serem utilizados foi formulada e realizada a pesquisa na base *Scopus*. A pesquisa inclui o corte de tipo de trabalho, considerando apenas os artigos de periódicos e de conferências. E o corte temporal, excluindo os trabalhos publicados a partir de 2017, por este ano não estar consolidado, o que prejudicaria a série temporal. A pesquisa está apresentada no Quadro 2.1.

Quadro 2.1 - Pesquisa em base de conhecimento.

```
( TITLE-ABS-KEY
   ( "data mining"  OR  "machine* Learning" )                          #Tesauros de A
 AND  TITLE-ABS-KEY
   ( education*  OR  school*  OR  academic* )                          #Tesauros de B
 AND  TITLE-ABS-KEY
   ( "drop out"  OR  drop-out  OR  dropout  OR  "evasion*"  OR  "Quitter" )   #Tesauros de C
AND  TITLE-ABS-KEY
   ( trade*  OR  technician*  OR  vocation* ) )                        #Tesauros de D
 AND
   ( LIMIT-TO ( DOCTYPE ,  "cp " )                                     #Corte de tipo
      OR  LIMIT-TO ( DOCTYPE ,  " ar " ) )
AND
   ( EXCLUDE ( PUBYEAR ,  2017 ) )                                     #Corte temporal
```

Fonte: Elaboração própria.

Além da pesquisa apresentada no Quadro 2.1 também foram realizadas outras buscas na base *Scopus*, combinando 2 ou 3 conceitos. No total foram formuladas e executadas 11 pesquisas. O diagrama de *Venn* apresentado na Figura 2.2 apresenta a quantidade de trabalhos encontrados em todas as combinações possíveis com os quatro conceitos.

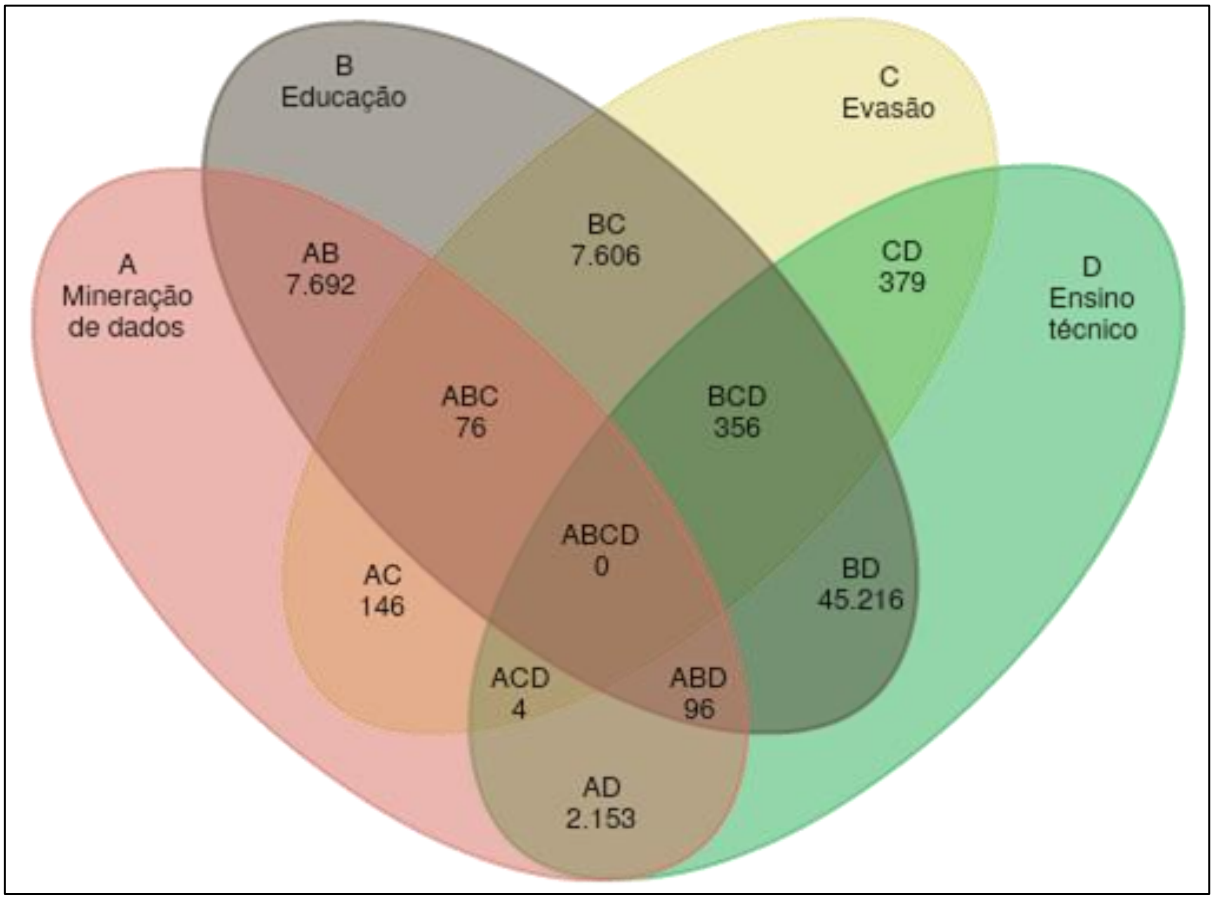

Figura 2.2 - Diagrama de Venn com a quantidade de trabalhos encontrados na base Scopus. Fonte: Elaboração própria.

A partir do diagrama é possível notar, primeiramente, que não foram encontrados trabalhos que reúnam os quatro conceitos. Na pesquisa que reuniu os conceitos mineração de dados, educação e evasão (ABC) foram encontrados 76 trabalhos, considerado um número baixo se comparado à outras pesquisas, como, por exemplo, nas buscas que utilizaram apenas dois conceitos.

Uma quantidade ainda menor de trabalhos foi retornada na pesquisa que reuniu mineração de dados, evasão e ensino técnico, com o retorno de apenas 4 trabalhos demonstrando que esta é uma área ainda pouco explorada. Para a pesquisa que excluiu o conceito evasão e utilizou mineração, educação e ensino técnico foram retornados 96 trabalhos.

Como não foram encontrados trabalhos na pesquisa que reúnam todos os conceitos, as etapas seguintes foram realizadas paras as pesquisas reunindo no mínimo três das palavras-chave definidas, onde umas dessas três é mineração de dados. Visto que esta que representa o método, é imprescindível na pesquisa.

2.5.2. Identificação dos periódicos com maior número de publicações

Nesta seção são apresentados os periódicos com mais publicações em cada uma das três pesquisas utilizando três dos conceitos definidos. Para a pesquisa com os conceitos ABC foram retornados 21 periódicos e 38 anais de conferência. Na busca com os conceitos ABD o total de periódicos foi 32 e de anais de conferência 47. Já a pesquisa com os conceitos ACD resultou em 4 anais de conferência. O Quadro 2.2 apresenta os periódicos com no mínimo duas publicações.

Quadro 2.2 - Quantidade de artigos por periódicos ou anais de conferências.

	Mineração de dados, educação e evasão (ABC)	**Mineração de dados, educação e ensino técnico (ABD)**	**Mineração de dados, evasão, ensino técnico (ACD)**
ACM International Conference Proceeding Series / ISBN: 9781450347709, 9781450341905, 9781450340205, 9781450335515, 978145033417	6	-	-
Anthropologist / ISSN: 0972-0073	-	2	-
CEUR Workshop Proceedings / ISSN: 16130073	4	-	-
Data Mining And Knowledge Discovery / ISSN: 1384-5810	-	2	-
Expert Systems With Applications / ISSN:0957-4174	-	3	-

Indian Journal Of Science And Technology / ISSN: 0974-6846	2	-	-
Lecture Notes In Computer Science Including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics / ISSN: 0302-9743	7	8	2
Lecture Notes In Electrical Engineering / ISSN: 1876-1100	-	3	-
Machine Learning / ISSN: 0885-6125	-	5	-
Proceedings Frontiers In Education Conference Fie / ISSN: 0190-5848	2	-	-
Proceedings Of The ACM Symposium On Applied Computing	2	-	-
Revista Iberoamericana De Tecnologias Del Aprendizaje / ISSN: 1932-8540	2	-	-

Fonte: Elaboração própria.

No total, o Quadro 2.2 lista 10 veículos de comunicação científica, sendo que o periódico *Lecture Notes In Computer Science Including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics* aparece nas três buscas analisadas. Significando ser um periódico que deve ser monitorado por ter uma atuação maior na área de pesquisa analisada.

Na Figura 2.3 é demonstrada a relação entre a quantidade de periódicos ou anais de eventos e a quantidade de publicações.

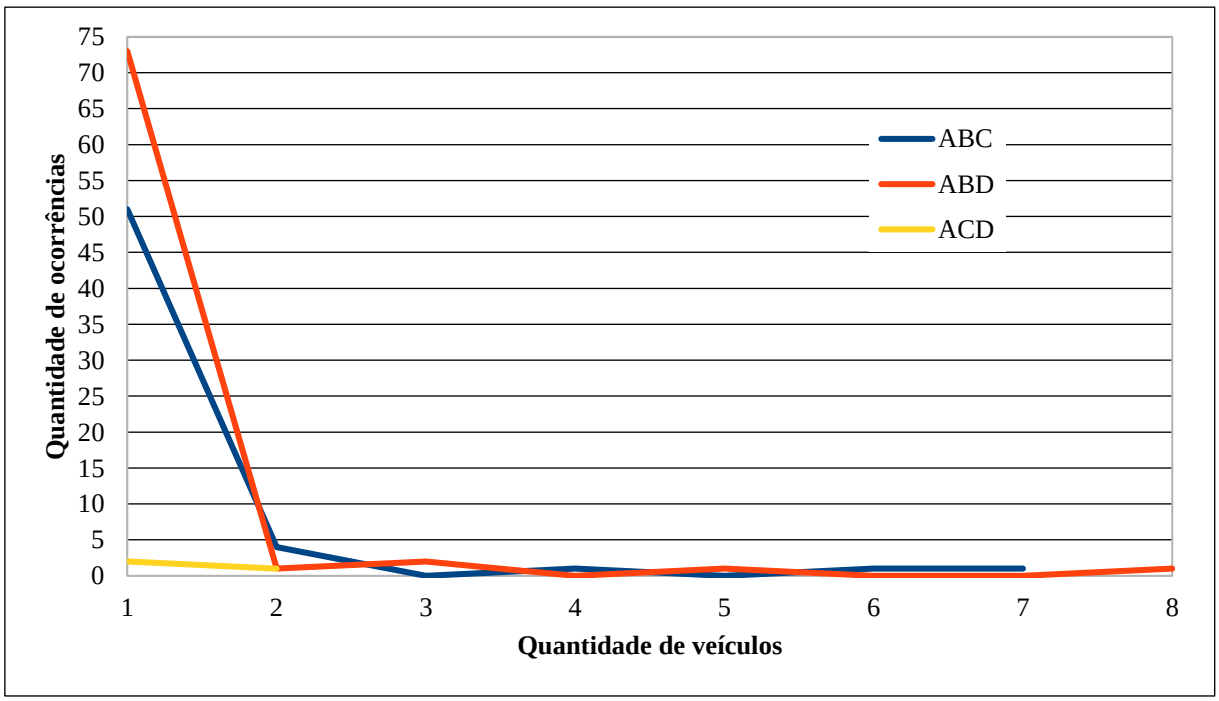

Figura 2.3 - Quantidade de ocorrências X Quantidade de veículos. Fonte: Elaboração própria.

Através da Figura 2.3 é possível verificar que a maioria dos veículos retornados não possuem mais que dois trabalhos publicados. Para as pesquisas com os conceitos ABC e com os conceitos ABD foram retornados, respectivamente, mais de 50 e mais de 70 veículos com apenas uma publicação.

2.5.3. Identificação dos autores com maior número de publicações

Na terceira etapa da pesquisa são apresentados os autores com duas ou mais publicações.

Quadro 2.3 - Quantidade de publicações por autores e coautores

Autor	Número de publicações
Mineração de dados, educação e evasão (ABC)	
Ventura S.	3
Da Cruz S.M.S.	2
García-Saiz D.	2
Li C.	2
Liang J.	2
Luna J.M.	2
Manhães L.M.B.	2
Márquez-Vera C.	2
Palazuelos C.	2
Reategui E.	2
Romero C.	2
Zheng L.	2
Zimbrão G.	2
Zorrilla M.	2
Mineração de dados, educação e ensino técnico (ABD)	
Knauf, R.	4
Tsuruta, S.	4
Buldu, A.	2
Chang, K.L.	2
Kaski, S.	2
Omar, E.B.	2
Peltonen, J.	2
Sakurai, Y.	2
Sakurai, Y.	2
Takada, K.	2
Yu, J.	2
Üçgün, K.	2

Mineração de dados, evasão, ensino técnico (ACD)	
-	-

Fonte: Elaboração própria.

No Quadro 2.3 são apresentados 26 autores. Sendo que para a busca que reuniu os conceitos mineração de dados, evasão e ensino técnico não foi identificado nenhum autor com mais de uma publicação.

2.5.4. Artigos com maior número de citações

Nesta etapa foram identificados os artigos mais citados em cada uma das três buscas analisadas. Para compor o Quadro 2.4 foram considerados os cinco artigos mais citados, com exceção da busca com os conceitos ACD, que retornou apenas quatro artigos.

Quadro 2.4 - Artigos com maior número de citações.

Título do artigo	Autores	Ano de publicação	Título do periódico ou conferência	Número de citações
Mineração de dados, educação e evasão (ABC)				
Predicting students drop out: A case study	Dekker, G.W., Pechenizkiy, M., Vleeshouwers, J.M.	2009	EDM'09 - Educational Data Mining 2009: 2nd International Conference on Educational Data Mining	74
Dropout prediction in e-learning courses through the combination of machine learning techniques	Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., Loumos, V.	2009	Computers and Education	57
Data mining model for higher education system	Ayesha, S., Mustafa, T., Raza Sattar, A., Khan, M.I.	2010	European Journal of Scientific Research	23
Preventing student dropout in distance learning using machine learning techniques	Kotsiantis, S.B., Pierrakeas, C.J., Pintelas, P.E.	2003	Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)	20
Predicting who will drop out of nursing courses: A machine learning exercise	Moseley, L.G., Mead, D.M.	2008	Nurse Education Today	19
Mineração de dados, educação e ensino técnico (ABD)				
Very Simple Classification Rules Perform Well on Most Commonly Used Datasets	Holte, R.C.	1993	Machine Learning	843
A Tree Projection Algorithm for Generation of Frequent Item Sets	Agarwal, R.C., Aggarwal, C.C., Prasad, V.V.V.	2001	Journal of Parallel and Distributed Computing	255

Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining	Novak, P.K., Lavrač, N., Webb, G.I.	2009	Journal of Machine Learning Research	159
Empirical support for winnow and weighted-majority algorithms: Results on a calendar scheduling domain	Blum, A.	1997	Machine Learning	94
Cost-Sensitive Learning of Classification Knowledge and Its Applications in Robotics	Tan, M.	1993	Machine Learning	87
Mineração de dados, evasão, ensino técnico (ACD)				
Feature cross-substitution in adversarial classification	Li, B., Vorobeychik, Y.	2014	Advances in Neural Information Processing Systems 3(January)	8
Using data mining techniques in fiscal fraud detection	Bonchi, F., Giannotti, F., Mainetto, G., Pedreschi, D.	1999	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	3
Detecting stepping-stone connection using association rule mining	Kuo, Y.-W., Huang, S.-H.S.	2009	Proceedings - International Conference on Availability, Reliability and Security, ARES 2009	1

Fonte: Elaboração própria.

Analisando o Quadro 2.4 é possível identificar que os artigos mais citados estão concentrados na busca com os conceitos ABD, ou seja, a busca que não inclui o conceito evasão. Sendo esta também, a que retornou entre os mais citados, os artigos mais antigos.

Em seguida, com artigos mais citados está a busca com os conceitos ABC, sendo todos os cinco artigos publicados a partir de 2003. A busca que inclui os conceitos ACD, retornou como artigo mais citado um trabalho publicado em 2014, o mais recente entre todos apresentados no Quadro 2.4.

A Figura 2.4 apresenta o gráfico de frequência acumulada das citações dos trabalhos relacionados a cada conceito pesquisado.

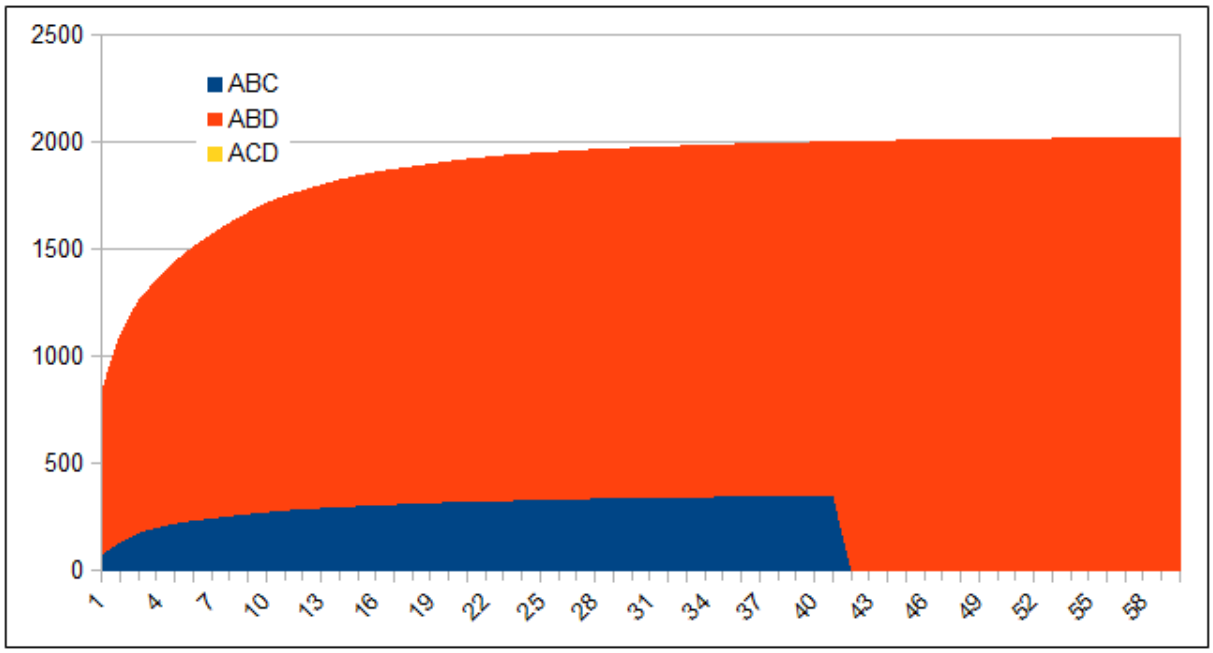

Figura 2.4 - Gráfico de frequência acumulada das citações. Fonte: Elaboração própria.

Analisando a Figura 2.4 nota-se que a área englobando os conceitos ABD (mineração de dados, educação e ensino técnico) é mais difundida em relação às outras duas buscas. Principalmente em relação à área que engloba mineração de dados, evasão e ensino técnico (ACD), que no gráfico não ficou visível devido aos valores irrisórios em relação às demais.

2.5.5. Levantamento da cronologia da produção

Para a análise da produção de trabalhos foram consideradas as pesquisas realizadas no *Scopus* que combinaram três dos quatro conceitos definidos. O corte temporal foi realizado a partir de 1990, devido ao fato de apenas uma pesquisa ter retornado trabalhos publicados anteriores à década de 90. A Figura 2.5 demonstra a cronologia da produção de 1990 a 2016.

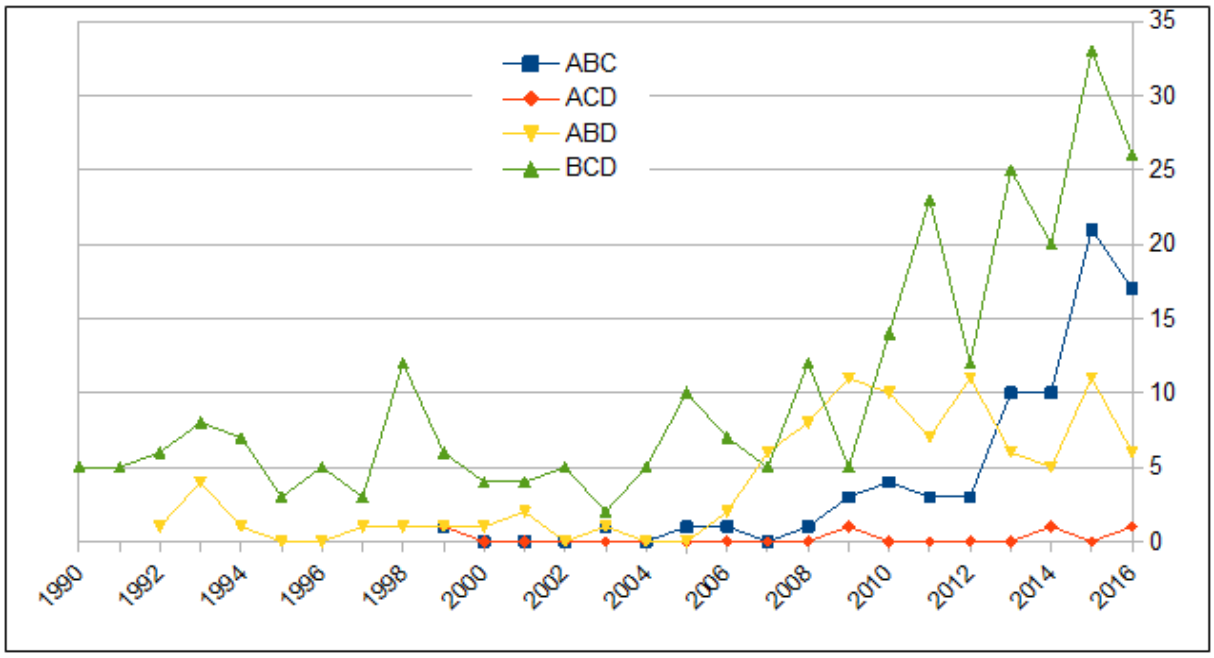

Figura 2.5 - Quantidade de artigos por ano de publicação. Fonte: Elaboração própria.

De acordo com o gráfico, os artigos encontrados a partir da pesquisa ABC que utilizam os conceitos mineração de dados, educação e evasão, começaram a ser publicados em 1999 e começaram a ter um aumento em 2013. Recentemente, em 2015, ocorreu a maior quantidade de publicações.

Os trabalhos que exploram mineração de dados, evasão e ensino técnico (ACD) mais antigos datam de 1999, apesar de ter tido um pico em 2014, em nenhum ano a quantidade de trabalhos foi superior a 5. Este conjunto foi o único que apresentou um aumento no ano de 2016.

Os artigos pertencentes ao conjunto ABD, que reúnem mineração de dados, educação e ensino técnico começaram a ser publicados no início da década de 90 e obtiveram um aumento na quantidade de publicação a partir de 2007, mantendo até 2016 uma quantidade superior a 5.

A partir da Figura 2.5 é possível observar também que, o conjunto BCD que reúne educação, evasão e ensino técnico é aquele com publicações mais antigas. Havendo, inclusive, trabalhos anteriores a década de 90 mas que não estão indicados na figura. Esse grupo de trabalhos também foi o que reuniu maior quantidade de trabalhos a partir de 2010.

De forma geral é possível observar que as publicações em sua maioria são posteriores a 2000, com aumento mais significativo a partir de 2008. O gráfico reforça a ideia de que na área de pesquisa que reúne mineração de dados, educação, evasão e ensino técnico ainda há muito o que explorar.

2.6. Considerações Finais

Em todas as análises realizadas é possível notar que as pesquisas envolvendo mineração de dados na área educacional com foco em evasão são recentes, com a maioria dos trabalhos publicados partir de 2010. Sendo uma área em ascendência, através do estudo bibliométrico, é notável que ainda é uma área pertinente e com muito a explorar.

A busca com maior quantidade de trabalhos publicados foi a que reuniu os conceitos mineração de dados, educação e ensino técnico, e, ainda assim, foi inferior a 100. Destaca-se que essa busca exclui o conceito que representa o objetivo da pesquisa: evasão.

Considerando a pesquisa que objetiva utilizar técnicas de mineração de dados na problemática da evasão escolar, mais especificamente, no ensino técnico, é interessante ressaltar primeiro que a busca envolvendo os quatro conceitos definidos na etapa 2 não retornou nenhum trabalho. Essa ausência de trabalhos indica uma área inédita a ser explorada. Em segundo lugar, ressalta-se o fato da busca que uniu os conceitos mineração de dados, evasão e ensino técnico retornar apenas 4 trabalhos.

Em relação à análise sobre os autores com mais trabalhos publicados foi possível observar que nenhum deles possui uma quantidade considerável, sendo Knauf, R. o autor com maior número de publicações e este número igual a quatro.

Entre os artigos mais citados estão os encontrados na busca que reuniu os conceitos mineração de dados, educação e ensino técnico. Entre as buscas que utilizaram o conceito evasão, os artigos mais citados estão naquela que utilizou mineração de dados, educação e evasão. E por último, estão os artigos resultantes da pesquisa que inclui os conceitos mineração de dados, evasão e ensino técnico.

Além de trazer um conhecimento sobre os trabalhos mais citados e autores com mais artigos publicados, este trabalho demonstra que a utilização de métodos de mineração de dados em dados educacionais, principalmente com foco em evasão, é uma área ainda em exploração. Este fato pode ser notado principalmente através da análise cronológica onde é visível que a atenção a essa área começou a aumentar apenas a partir de 2010.

Como trabalhos futuros propõe-se a utilização de outras bases de conhecimento, o que provavelmente aumentaria a amostra da pesquisa. Além disso, podem ser realizadas outras análises além das apresentadas neste trabalho.

Referências

Araújo Júnior, R. H. de, Perucchi, V. & Lopes, P. R. D. (2013). Analise bibliométrica dos temas inteligência competitiva, gestão do conhecimento e conhecimento organizacional no repositório institucional da universidade de Brasília. *Perspectivas em Ciência da Informação*, *18*(4), 54–69.

Costa, H. G. (2010). Modelo para webibliomining: proposta e caso de aplicação. *Revista da FAE, 13*, 115–126.

Cunha, J. A., Moura, E. & Analide, C. (2016). Data mining in academic databases to detect behaviors of students related to school dropout and disapproval. *Advances in Intelligent Systems and Computing*, *445*, 189–198. https://doi.org/10.1007/978-3-319-31307-8_19

Machado, R. D., Benitez, E. O., Corleta, J. N. & Augusto, G. (2015). Estudo Bibliométrico em mineração de dados e evasão escolar. Apresentado na XI CONGRESSO NACIONAL DE EXCELÊNCIA EM GESTÃO, Rio de Janeiro, RJ.

Mehta, A. A. & Buch, N. J. (2016). Depth and breadth of educational data mining: Researchers' point of view. Em *Proceedings of the 10th International Conference on Intelligent Systems and Control, ISCO 2016*. Coimbatore, India.

Rigo, S. J., Cazella, S. C. & Cambruzzi, W. (2012). Minerando Dados Educacionais com foco na evasão escolar: oportunidades, desafios e necessidades. *Anais do Workshop de Desafios da Computação Aplicada à Educação*, *0*(0), 168–177.

Thesaurus. (2013). Thesaurus. Obtido 3 de Fevereiro de 2017, de http://www.thesaurus.com

3. ARTIGO 2 - MINERAÇÃO DE DADOS EDUCACIONAIS COM FOCO NA EVASÃO: UMA REVISÃO SISTEMÁTICA SOBRE ATRIBUTOS E TÉCNICAS

3.1. Resumo

Contexto: O grande volume de dados armazenados na área educacional requer a utilização de técnicas que propiciem a interpretação e análise desses dados com objetivo de agregar valor à gestão e aos educadores. A mineração de dados educacionais é uma área emergente em que muitos trabalhos buscam a análise de dados com foco na evasão.
Objetivo: O objetivo deste trabalho é identificar as etapas e os métodos mais utilizados na área de mineração de dados para identificação de alunos propensos a evadir.
Metodologia: Compondo a metodologia é realizada a pesquisa bibliográfica, buscando trabalhos relacionados e uma visão sobre os estudos na área de mineração de dados educacionais.
Resultados: A partir da análise dos trabalhos mais relevantes foi possível verificar os atributos e métodos mais utilizados. Além das principais etapas que compõem os trabalhos analisados: definição dos atributos, tratamento dos dados, higienização dos dados e mineração de dados, etapa em que de fato são identificados os perfis dos alunos evadidos.
Conclusão: Acredita-se que os resultados dessa pesquisa tragam contribuições para as pesquisas que busquem a identificação do comportamento dos alunos evadidos.

Palavras-chave: Mineração de Dados; Educação; Evasão.

3.2. Abstract

Context: *The large volume of data stored in the educational area requires the use of techniques that allow the interpretation and analysis of data in order to add value to management and educators. An educational data mining is an emerging area in which many papers seek data analysis with a focus on evasion.*
Objective: *The objective of this work is to identify the steps and methods most used in the area of data mining to identify students likely to evade.*
Methodology: *Composing the methodology is carried out the bibliographic research, searching for related works and a view on studies in the area of educational data mining.*
Results: *From the analysis of the most relevant works it was possible to verify the attributes and methods most used. Besides the main steps that make up the analyzed works: definition of the attributes, data treatment, data hygiene and data mining, in which stage the profiles of the evaded students are in fact identified.*
Conclusions: *It is believed that the results of this research bring contributions to the researches that seek to identify the behavior of the evaded students.*

Keywords: Data mining; Education; Dropout.

Palavras-chave: Mineração de Dados, Educação, Evasão.

3.3. Introdução

A área de pesquisa Mineração de Dados Educacionais é uma área emergente para a descoberta de conhecimento em grandes volumes de dados armazenados por instituições de ensino. Nessa área, a abordagem da evasão escolar tem tido o intuito de propiciar a identificação precoce de alunos com possibilidade de evadir.

A partir do momento em que as instituições de ensino caminham para a informatização torna-se possível o armazenamento de dados com grande potencial para o planejamento de políticas da gestão escolar, torna-se atrativo buscar técnicas para obter informações a partir dos dados educacionais.

Para Machado, Benitez, Corleta e Augusto (2015), bases de dados tornam-se pouco úteis sem utilização de ferramentas para interpretação e análise. A mineração de dados utilizada como ferramenta para análise desses dados torna-se útil e de grande valor pela possibilidade de extrair informações importantes a partir de grandes volumes de dados.

O presente trabalho visa identificar os métodos e etapas mais utilizados no processo de identificação do comportamento de alunos evadidos. Para isso, realiza uma revisão sistemática em um conjunto de trabalhos.

De acordo com Mehta e Bunch (2016), a educação e a qualidade da educação dependem de vários parâmetros como social, econômico, cultural, geográfico, ambiente institucional, entre outros. Tais fatores podem contribuir para um desempenho acadêmico excelente ou para uma desestimulação. A evasão de um aluno em qualquer nível da educação pode custar não só a ele e sua família, também pode deixar marcas no futuro da sociedade e no crescimento da instituição. Portanto é importante a realização de estudos para identificar os estudantes com possíveis chances de evasão. Além disso, com um estudo que permita detectar alunos propensos a evadir o curso é possível traçar políticas mais focadas que os incentivem a permanecer.

De acordo com Márquez-Vera *et al.* (2016), quanto mais cedo forem identificados os estudantes propensos a evadir o curso, maiores são as chances de sucesso da política de permanência escolar. Estudantes cujo risco de evasão é conhecido podem ser ajudados de forma mais personalizada tanto pela família quanto pela instituição de ensino.

Em seu trabalho, Hoed (2016) afirma que, além de zelar pela aprendizagem e bem-estar dos discentes, também é obrigação e uma necessidade da instituição conter a evasão, por estar diretamente associada à perda de recursos financeiros.

Sendo assim, o presente trabalho justifica-se por explorar a área de mineração de dados educacionais focando na evasão, com intuito de contribuir para a realização de políticas institucionais que visem diminuir os índices de alunos evadidos.

3.4. Descoberta de conhecimento em base de dados

Dado o grande volume de dados armazenados em uma instituição de ensino é necessária uma metodologia e métodos que permitam extrair as informações relevantes. Além da mineração de dados, é necessário que se realize o tratamento prévio dos mesmos. Considera-se então o *Knowledge Discovery in Databases* (KDD) por referir-se ao processo de descoberta de conhecimento útil de um banco de dados, enquanto a mineração de dados está inserida como um dos principais passos necessários para a realização do processo de KDD (Cardoso & Machado, 2008; Fayyad, Piatetsky-Shapiro & Smyth, 1996).

Fayyad *et al.* (1996) definem o KDD como "o processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e, em última instância, compreensíveis em dados". O processo pode ser definido como a descoberta de padrões, relacionamentos e tendências significativas por meio de uma cuidadosa análise dos grandes conjuntos de dados armazenados.

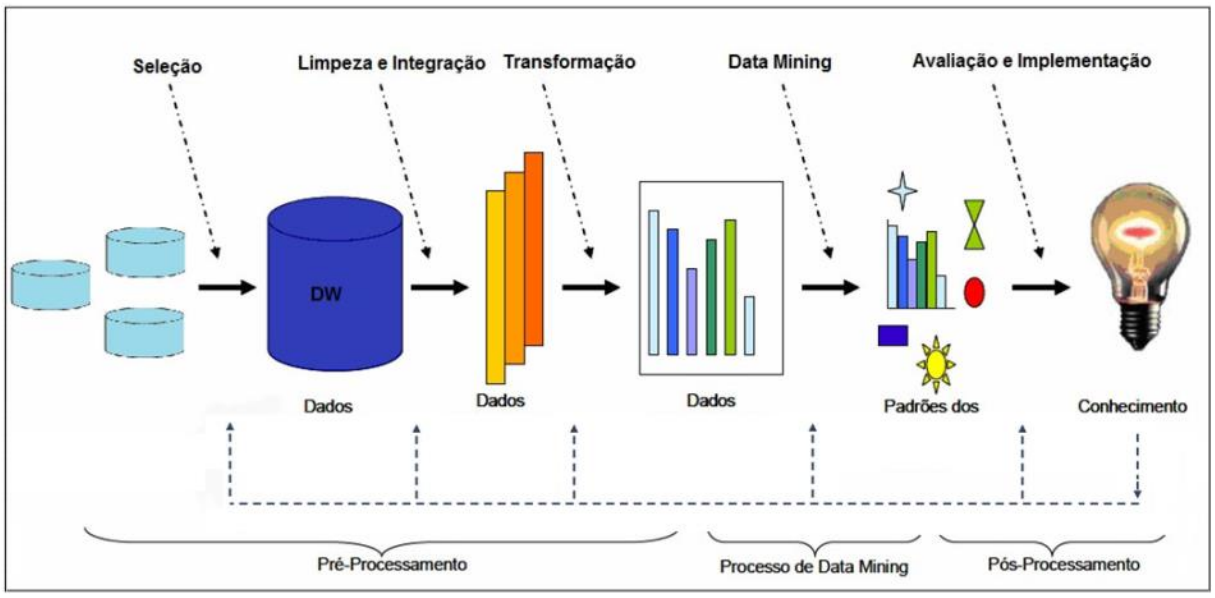

Figura 3.1 - Etapas do Processo de Descoberta de Conhecimento em Banco de Dados. Fonte: (R. M. da Silva, Gomes, Shimoda & Santos, 2010; adaptado de Fayyad *et al.*, 1996).

As etapas do processo de KDD apresentadas na Figura 3.1 são descritas por Goldschmidt e Passos (2005):

a) **Seleção de dados**: identificar e selecionar quais informações, dentre as bases de dados existentes, devem ser efetivamente consideradas durante o processo;

b) **Limpeza de dados**: Realizar tratamento sobre os dados, a fim de assegurar a qualidade relacionada à completude, veracidade e integridade, ou seja, dados inconsistentes ou fora dos padrões são removidos;

c) **Integração de dados**: Reunir várias fontes de dados, mantendo a consistência e a coerência dos dados integrados;

d) **Transformação de dados**: Codificar os dados para o formato apropriado para a próxima etapa. Esta fase é realizada dependendo do algoritmo que será aplicado na mineração de dados;

e) ***Data Mining*** **(Mineração de Dados)**: Aplicar métodos com o propósito de extrair os padrões de interesse;

f) **Avaliação de padrões**: Identificar os padrões de interesse de acordo com algum critério do usuário;

g) **Apresentação de conhecimento**: Tornar o conhecimento extraído compreensível ao homem através de gráficos, diagramas ou relatórios demonstrativos.

Mineração de dados é o conjunto de técnicas que permitem extrair, analisar e explorar conhecimento de uma massa de dados em busca de padrões, erros e associações que, de outra maneira, permaneceriam escondidos nas grandes bases (Amaral, 2016).

De modo geral os trabalhos que utilizam mineração de dados para estudos de evasão escolar se baseiam nas etapas definidas pelo KDD (Jiménez-Gómez, Luna, Romero & Ventura, 2015; Mehta & Buch, 2016; Pradeep, Das & Kizhekkethottam, 2015).

3.5. Metodologia

A seguir são apresentados os procedimentos metodológicos realizados neste trabalho. Apresenta-se a classificação da pesquisa e o detalhamento de cada etapa realizada.

3.5.1. Classificação da pesquisa

Do ponto de vista de sua natureza, este trabalho é uma pesquisa aplicada uma vez que de acordo com Silva e Menezes (2005), pesquisas deste tipo "objetivam gerar conhecimentos para aplicação prática e dirigidos à solução de problemas específicos".

Quanto à forma de abordagem, pode ser classificada como qualitativa, uma vez que foi realizada uma análise dos trabalhos publicados na área de mineração de dados educacionais (Gil, 2009; E. L. da Silva & Menezes, 2005).

Em relação aos objetivos, a pesquisa desenvolvida, classifica-se como descritiva. De acordo com Gil (2009), as pesquisas deste tipo visam descrever características de determinada população e estabelecer relações entre variáveis, inclusive por meio de técnicas padronizadas.

Do ponto de vista dos procedimentos técnicos realizados, é classificada como bibliográfica (E. L. da Silva & Menezes, 2005), pelo fato de utilizar como base os trabalhos já publicados na área e, de acordo com Gil (2009), trabalhos que recebem essa classificação são elaborados a partir de material já publicado, e nesse caso, principalmente, a partir de artigos científicos.

3.5.2. Etapas da pesquisa

O desenvolvimento da pesquisa pode ser dividido em cinco etapas, apresentadas na Figura 3.2.

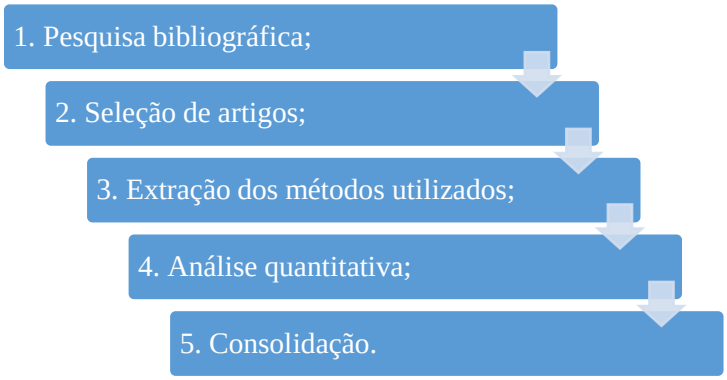

Figura 3.2 - Etapas da metodologia. Fonte: Elaboração própria.

Na primeira etapa é realizada a pesquisa bibliográfica sobre o tema estudado. Para a realização dessa etapa, primeiramente, são definidas as palavras-chaves que descrevem da

melhor forma a proposta do presente trabalho, além de seus tesauros. A pesquisa é realizada na base de conhecimento *Scopus*.

Para a seleção dos artigos, que ocorre na segunda etapa, são considerados os trabalhos retornados na pesquisa que aborda mineração de dados, educação e evasão. Por ser esta a que reúne os conceitos referentes ao método e ao objeto de pesquisa deste trabalho. Dos 76 artigos retornados foram selecionados aqueles com mais de 10 citações e os que apresentaram mais afinidade com a pesquisa proposta.

A etapa três consiste da análise e estudo dos artigos selecionados na etapa anterior. O intuito desta etapa é extrair os métodos utilizados e as etapas executadas nos trabalhos.

Dando continuidade à etapa anterior, na etapa quatro é realizada a análise quantitativa, através da elaboração de quadros onde podem ser visualizados quais trabalhos utilizaram quais métodos e, consequentemente, os métodos mais utilizados.

Na última etapa os resultados obtidos são consolidados reunindo as etapas e os métodos mais adotados nos trabalhos encontrados, com o intuito de obter uma metodologia para identificação de comportamento de alunos evadidos.

3.5.3. Pesquisa bibliográfica

Para a realização da etapa de pesquisa bibliográfica, são definidos, primeiramente, os conceitos que definem o método, o objeto de pesquisa e o objetivo do presente estudo: mineração de dados (método), educação (objeto de pesquisa), evasão (objetivo). A partir daí são definidas as palavras-chave: *data mining*, *education* e *dropout*. Além das palavras-chaves são definidos seus tesauros e conceitos que também as representam na literatura. Para a definição dos tesauros é realizada uma consulta no site Thesaurus (Thesaurus, 2013). A pesquisa é realizada na base de conhecimento *Scopus* (*www.scopus.com*). Esta escolha se deve à representatividade da base de dados que abrange artigos de conferência, periódicos, anais, entre outros. Sendo que para compor o conjunto de trabalhos analisados foram selecionados os artigos de periódicos e conferências. Em relação ao recorte temporal a pesquisa contempla todos os trabalhos publicados até 2016. O Quadro 3.1 apresenta a pesquisa realizada.

Quadro 3.1 - Pesquisa em base de conhecimento.

```
( TITLE-ABS-KEY
   ( "data mining"  OR  "machine* Learning" )                          #Tesauros de A
 AND  TITLE-ABS-KEY
   ( education*  OR  school*  OR  academic* )                          #Tesauros de B
 AND  TITLE-ABS-KEY
   ( "drop out"  OR  drop-out  OR  dropout  OR  "evasion*"  OR  "Quitter" )   #Tesauros de C
AND
   ( LIMIT-TO ( DOCTYPE ,  "cp " )                                     #Corte de tipo
      OR  LIMIT-TO ( DOCTYPE ,  " ar " ) )
AND
   ( EXCLUDE ( PUBYEAR ,  2017 ) )                                     #Corte temporal
```

Fonte: Elaboração própria.

Além da pesquisa apresentada no Quadro 3.1 também são realizadas outras buscas na base *Scopus*, combinando 2 ou 3 palavras-chave. No total são formuladas e executadas 4 pesquisas. O diagrama de *Venn* apresentado na Figura 3.3 apresenta a quantidade de trabalhos encontrados em todas as combinações possíveis com os três conceitos.

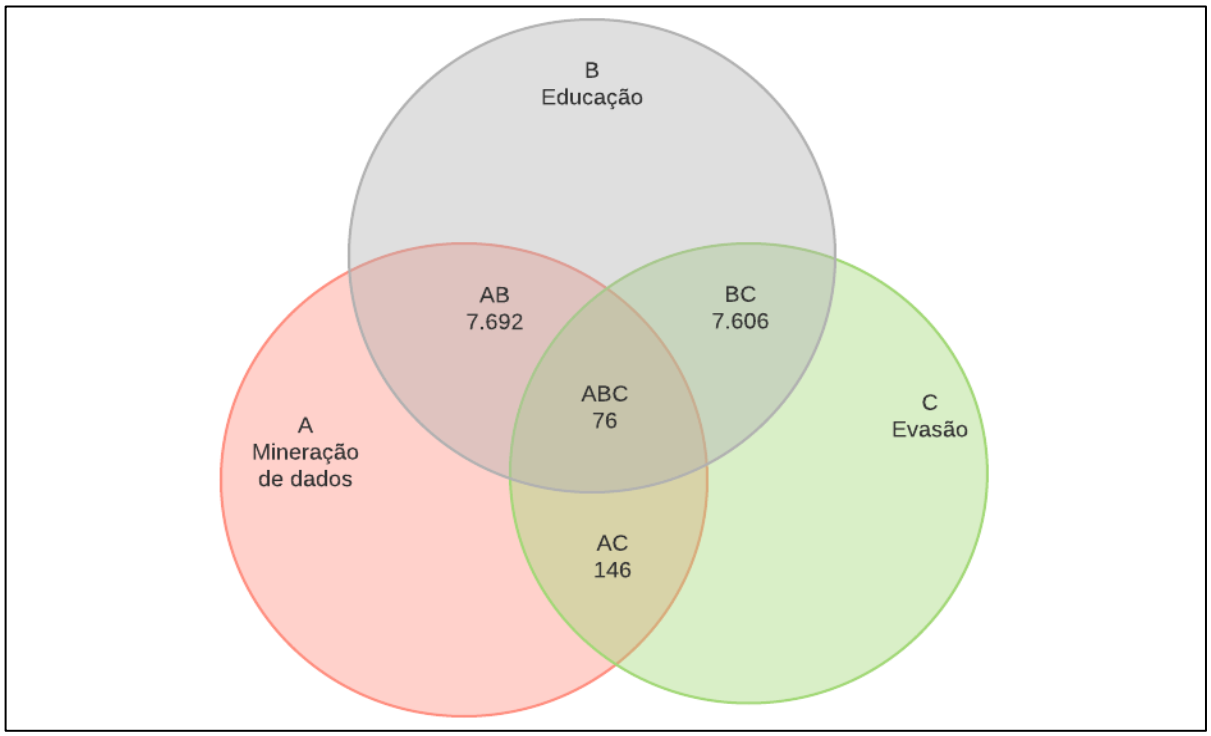

Figura 3.3 - Diagrama de *Venn* com a quantidade de trabalhos encontrados na base *Scopus*. Fonte: Elaboração própria.

A partir do diagrama é possível notar, primeiramente, que na pesquisa que reúne os conceitos mineração de dados, educação e evasão foram encontrados 76 trabalhos, considerado

um número baixo se comparado à outras pesquisas, como por exemplo nas buscas que utilizam apenas dois conceitos.

Em relação as pesquisas que reuniram dois conceitos, é possível notar que aquela que retornou menos trabalhos foi a pesquisa que reúne os conceitos mineração de dados e evasão. Essa pesquisa retornou 146 trabalhos, demonstrando que essa ainda é uma área que não é tão explorada. As outras duas pesquisas, que combinaram mineração de dados e educação; e educação e evasão retornaram mais de sete mil trabalhos. Isso demonstra que a mineração de dados aliada à educação assim como o estudo da evasão na educação são temas bem discutidos.

3.5.4. Mineração de dados na área educacional

Diante dos avanços na área de tecnologia da informação, é crescente a utilização de sistemas também na área educacional. Isso proporciona uma quantidade cada vez maior de dados armazenados, e esses dados, quando associados a ferramentas e métodos capazes de interpretá-los e analisá-los, podem transformar-se em informações valiosas. Essas informações podem contribuir para o tratamento de problemas como a evasão, para o processo de ensino-aprendizagem e elaboração de políticas visando o aprimoramento do processo educacional.

Os estudos na área de mineração de dados educacionais abordam a utilização de técnicas de mineração de dados no contexto educacional. O que a difere é a natureza dos dados, que é mais diversa do que os tradicionalmente utilizados, necessitando de adaptações (Rigo *et al.*, 2012). De acordo com Romero e Ventura (2010), a mineração de dados educacionais difere-se ainda por englobar métodos como de regressão, correlação, entre outros que não são considerados como pertencentes à mineração de dados.

Cunha *et al.* (2016) utilizam mineração de dados em base de dados para detectar comportamentos relacionados à evasão escolar e à reprovação. Os autores apontam que é sabido que a evasão e a reprovação estão relacionadas as áreas de conhecimento do aluno, ao nível de educação e as metodologias de ensino e aprendizado. Mas, neste trabalho, os autores buscam fatores relacionados à evasão e reprovação através de dados disponíveis em uma base de dados acadêmica e formulários anuais aplicados a professores e alunos. Na análise dos resultados, foi identificado que o grupo de alunos evadidos possuem algumas características: originário de escolas estaduais, com renda familiar de até um salário mínimo, vivendo com os pais e com baixo coeficiente de rendimento em algumas disciplinas voltadas para a lógica e matemática.

No trabalho de Márquez-Vera *et al.* (2016), o objetivo é prever a evasão de alunos no nível escolar anterior à universidade. Com o intuito de obter alertas sobre estudantes com possível risco de evasão o mais cedo possível, os autores propõem um algoritmo de classificação. A partir do pressuposto que os resultados dependem a qual etapa do curso os dados correspondem, o método tenta definir em qual etapa deve-se extrair os dados para obter uma previsão de evasão confiável, considerando que quanto mais cedo for possível obter uma previsão de evasão, melhores serão os efeitos das políticas que a impeçam. Os autores, utilizando os dados de 419 estudantes e dados coletados em 6 etapas ao longo do primeiro semestre, obtiveram boa precisão de evasão utilizando o algoritmo classificatório proposto.

Em seu trabalho, Mehta e Buch (2016) realizam uma análise de estudos com foco em mineração de dados educacionais. Os autores citam algumas arquiteturas utilizadas na mineração de alguns trabalhos, em geral, arquiteturas compostas por: obtenção dos dados, pré-processamento, mineração e interpretação dos resultados. Analisam também as variáveis utilizadas, além de relatar etapas de pré-processamento. É destacada a importância da etapa do pré-processamento ao afirmarem que decisões de qualidade necessitam de dados com qualidade sendo importante retirar dados inconsistentes, incompletos e ruídos. Em relação às variáveis, em geral, são utilizados dados demográficos, como gênero e escolaridade dos pais, geográficos, como área onde residem e dados relacionados ao desempenho no curso como notas e frequência. O autor concluiu que a disponibilidades de dados é primordial para um resultado de qualidade, e que a mineração de dados educacionais é um campo aberto para pesquisa no intuito de tornar o processo de educação mais planejado e eficaz para os alunos e para a sociedade.

No trabalho de Xing, Chen, Stein e Marcinkowski (2016) é realizado um estudo de evasão em cursos *online*. Os cursos analisados são aqueles caracterizados como MOOCs (*Massive open online courses*), em que não há limite de inscritos, geralmente são sem custo e o conteúdo é ensinado através de uma plataforma web. Considerando que são turmas com uma grande quantidade de alunos, fica mais difícil para um tutor oferecer a atenção necessária a todos a ponto de identificar os alunos com risco de evadir. Sendo assim, utilizando mineração de dados, analisando dados como por exemplo, interação em fórum de discussões, os autores propõem um modelo com intuito de auxiliar os tutores a identificar os alunos com risco de evasão para que possam ser tomadas atitudes que estimulem estes alunos a permanecerem no curso.

Em Jiménez-Gómez *et al.* (2015) é objetivado encontrar qual o estágio mais cedo do curso em que os dados possam ser coletados para prever, com boa precisão, os alunos com chance de evasão. O estudo se concentra em estudantes na fase escolar anterior ao vestibular e considera dados pessoais dos estudantes, notas e ausências. O banco de dados foi alimentado de forma incremental a cada estágio do curso de forma que ficasse perceptível em qual estágio foi atingida uma boa precisão na previsão. Sendo concluído que para o curso de 4 anos analisado, ao final do primeiro, já é possível realizar uma boa análise de previsão.

Utilizando diversas técnicas de mineração de dados para analisar a melhor a previsão de estudantes com risco de evasão, Pradeep *et al.* (2015) analisaram dados de 670 estudantes com 57 atributos entre os anos de 2011 e 2013. Após a seleção de atributos, foram selecionados 12. Foram realizados três testes: com a base de dados com os 57 atributos, com a base de dados apenas com os 12 atributos selecionados e com a base de dados com os 12 atributos e após a execução do algoritmo de balanceamento. O algoritmo de balanceamento tem o objetivo de balancear os dados pertencentes às duas classes: concluintes e evadidos, considerando que a quantidade de dados de alunos que concluíram é muito superior à dos alunos que evadiram. Algoritmos de árvore de decisão e regras de indução foram utilizados. Como resultado, os autores concluíram que ambos os tipos de algoritmos obtiveram boa performance. A redução de atributos não reduziu a precisão dos algoritmos e gerou regras mais simples de serem interpretadas.

Tamhane, Ikbal, Sengupta, Duggirala e Appleton (2014) utilizam técnicas de mineração de dados para prever estudantes que terão baixo rendimento. Os autores afirmma que uma das consequências desse baixo rendimento é a evasão. O estudo é aplicado na educação primária e secundária, sendo estes níveis da educação determinantes para um futuro de sucesso na vida acadêmica do aluno. Para os autores, a identificação de alunos com risco de baixo rendimento é realizada de acordo com a intuição do professor, tornando algo subjetivo e dependente da experiência dos profissionais. E um método quantitativo capaz de realizar este tipo de previsão pode tornar a identificação mais rápida, permitindo atitudes o mais breve possível para reverter a situação de baixo rendimento. Os autores verificaram que é possível realizar a previsão da performance dos alunos, diferenciando entre estudantes em risco ou não. Além disso, verificaram que as notas obtidas nas avaliações e os dados demográficos dos alunos foram cruciais para uma boa acurácia do método.

O estudo de Márquez-Vera, Romero e Ventura (2013) propõe a utilização de técnicas de mineração de dados para prever estudantes na educação média com risco de evasão. Foram utilizados dados referentes ao primeiro ano de *high school* de estudantes entre 15 e 18 anos. Os dados utilizados foram dados pessoais, dados constantes em um formulário preenchido pelos estudantes ao ingressarem no curso e notas obtidas nas disciplinas. Ao total, os dados utilizados continham informações de 670 alunos e 77 atributos. Após a seleção de atributos foram selecionados 15. No pré-processamento também foi realizado o balanceamento dos dados, igualando a quantidade de dados referentes a alunos que evadiram com a de alunos que concluíram. Neste trabalho, os autores consideraram o custo dos erros, atribuindo pesos mais altos para os alunos que evadiram classificados erroneamente. Comparando os resultados obtidos antes de atribuir custo aos erros e depois, foi verificada uma melhora na precisão para alguns algoritmos e piora para outro. Sendo assim, o autor considerou que, no geral, não houve uma melhora na classificação.

No trabalho desenvolvido por Adamopoulos (2013), a partir da análise de que os cursos *online* abertos, em que há uma quantidade massiva de alunos e há também alta taxa de evasão, o autor propõe um método que seja capaz de expor as causas da evasão e a partir de então as instituições possam elaborar políticas com intuito de reduzir essa taxa. Com o objetivo de realizar uma análise holística sobre esta problemática, no trabalho são utilizadas técnicas de mineração de texto, mineração de opinião e análises econométricas. Em um primeiro momento são analisados fatores que influenciam na permanência do aluno e em seguida é verificada possibilidade de prever os alunos com propensão a evadir. A partir do estudo foi identificado que algumas características do curso, como a disponibilidade de um professor, influenciam de forma positiva a permanência do aluno no curso, e outras, como cursos sem calendário definido, afetam de forma negativa. As características relacionadas ao aluno não influenciam na previsão de propensões à evasão.

Segundo Dekker, Pechenizkiy e Vleeshouwers (2009) é importante o monitoramento dos estudantes assim que ingressam na faculdade com o intuito de prevenir a evasão. Os autores destacam que encontrar fatores preditores do sucesso acadêmico pode auxiliar professores e gestores a tomar medidas adequadas com o intuito de reduzir a evasão. O estudo foi aplicado no departamento de Engenharia Elétrica da Universidade Tecnológica de Eindhoven. Foram três conjuntos de dados: o primeiro com dados acadêmicos do período escolar pré-universitário, o segundo com dados do período universitário e o terceiro unindo os dois conjuntos de dados.

Os autores constataram que a utilização de técnicas de seleção de atributos não trouxe melhoras significativas. Para o estudo com o conjunto de dados que uniu informações pré-universitárias e universitárias foi constatado que os resultados ficaram mais próximos dos encontrados no teste com dados universitários, sendo possível concluir que os dados pré-universitários não contribuem com informações que possam melhorar significativamente os resultados. Os autores concluíram que foram obtidas boas taxas de acerto mesmo com métodos de classificação não muito sofisticados.

No estudo de Lykourentzou, Giannoukos, Nikolopoulos, Mpardis e Loumos (2009), é proposto um método para prever evasão em cursos *online*. Para os autores, os cursos *online* oferecem facilidades e flexibilidades principalmente para quem precisa equilibrar as demandas de estudo e trabalho, porém, esses cursos possuem taxas de evasão muito superiores aos cursos presenciais. A taxa de permanência está entre os indicadores das universidades e é utilizado para a análise da qualidade de uma instituição de ensino. Os autores afirmam que a ênfase crescente em permanência combinada com a alta taxa de evasão em cursos *online* torna a busca pela diminuição da evasão fundamental para o sucesso desse tipo de curso. Foi verificado que ocorriam casos em que o aluno apresentava um bom desempenho nos testes e projetos do curso mas, ainda assim, evadia. O fato de ser um curso *online* e, por isso, ter todas as informações registradas em sistema permite uma previsão com mais precisão. Como resultado, foi verificado que a combinação de três técnicas de *machine learning* trouxe resultados com precisão acima de 90%. Foi verificado também que os dados obtidos dos questionários respondidos pelos alunos, antes de iniciar o curso, com atributos invariantes como sexo e local de residência não trouxe grandes contribuições para o resultado.

Moseley e Mead (2008) abordam a evasão no curso superior de enfermagem. A partir da análise de que houve um aumento no número de alunos nos cursos de enfermagem e também de evasão, foi detectado que este é um problema a ser investigado e prevenido. Os autores analisaram que outros trabalhos com o mesmo objetivo abordaram apenas os estudantes que evadiram sem considerar os que concluíram o curso e comparar os dois grupos. Foi utilizado um método de regra de indução e verificado que é possível realizar boas previsões, porém o método requer dados com qualidade, consistentes.

No trabalho de Kotsiants, Pierrakeas e Pintelas (2003) é investigada a evasão em cursos universitários à distância, afirmando que, nessa modalidade, a taxa de evasão é superior, tendo causas profissionais, acadêmicas, relacionadas à saúde, família e razões pessoais. Os autores

buscam o melhor método entre as técnicas de *machine learning* para prever os estudantes com risco de evasão para que esse método, inserido no ambiente virtual utilizado pelo aluno, possa alertar os tutores sobre a probabilidade de evasão. O estudo foi aplicado em cursos na área de informática e, como resultado, foi verificado que o método *Naive Bayes* foi o mais apropriado, com mais precisão, entre os métodos avaliados. Os autores concluíram que algoritmos de aprendizagem são uma ferramenta que, de fato, pode auxiliar na previsão e redução das taxas de evasão.

3.6. Métodos e etapas na mineração de dados educacionais

Dos 76 trabalhos, identificados na etapa de pesquisa bibliográfica, foram selecionados aqueles com mais de 10 citações e os que apresentaram mais afinidade com a pesquisa proposta. Esta seção analisa as técnicas e etapas de mineração de dados para análise de dados educacionais com foco na evasão utilizadas nos 12 trabalhos selecionados. O Quadro 3.2 apresenta os atributos utilizados, a última linha apresenta quantas vezes cada categoria de atributos foi utilizada.

Quadro 3.2 - Atributos utilizados em cada pesquisa

	Assiduidade	**Desempenho educacional**	**Dados Pessoais**	**Dados socioeconômicos**	**Outros**
(Cunha et al., 2016)	x	x	x	x	
(Márquez-Vera et al., 2016)		x	x		x
(Xing et al., 2016)		x			
(Jiménez-Gómez et al., 2015)		x	x		x
(Pradeep et al., 2015)		x	x	x	
(Tamhane et al., 2014)	x	x	x		
(Adamopoulos, 2013)		x	x		x
(Márquez-Vera et al., 2013)		x	x	x	
(Dekker et al., 2009)		x			x
(Lykourentzou et al., 2009)		x		x	
(Moseley & Mead, 2008)	x	x	x		
(Kotsiantis et al., 2003)	x	x		x	
Total	4	12	8	5	4

Fonte: Elaboração própria.

A partir da análise dos trabalhos elencados no Quadro 3.2, os atributos utilizados nesses trabalhos foram organizados em cinco categorias: assiduidade, desempenho educacional, dados

pessoais, dados socioeconômicos e outros. A categoria assiduidade inclui informações referente a presenças e ausências. A categoria desempenho educacional engloba dados como notas obtidas, tempo para conclusão e reprovações. No trabalho de Xing *et al.* (2016), que analisa evasão em cursos *online*, são utilizados como medidas de desempenho educacional dados como número de *posts* e interações em fóruns no ambiente virtual. A categoria dados pessoais inclui informações como sexo, idade e tempo de experiência profissional. Em dados socioeconômicos são englobados dados referentes à quantidade de filhos, local de moradia, renda familiar e grau de escolaridade. Por fim, a categoria outros engloba os dados menos utilizados nos trabalhos analisados como, por exemplo, dados da fase pré-universitária (Dekker *et al.*, 2009; Márquez-Vera *et al.*, 2016), atributos referentes ao curso, como duração e avaliações (Adamopoulos, 2013), habilidade físicas (Márquez-Vera *et al.*, 2016) e avaliações comportamentais (Jiménez-Gómez *et al.*, 2015).

Analisando o Quadro 3.2 é possível perceber que todos os trabalhos analisados utilizam dados referentes ao desempenho educacional e a maioria utiliza dados pessoais ou socioeconômicos e em alguns casos, ambos. Informações sobre assiduidade foram consideradas em quatro dos doze trabalhos, e quatro trabalhos também utilizaram dados classificados como outros.

Alguns trabalhos realizam a seleção de atributos, identificando aqueles que mais influenciam para a previsão dos alunos com risco de evasão. No trabalho de Cunha *et al.* (2016) foi utilizado um método de árvore de decisão. Já em Márquez-Vera *et al.* (2013) e em Pradeep *et al.* (2015) foi selecionado um conjunto de algoritmos e os atributos que foram selecionados por dois ou mais algoritmos foram os escolhidos. Porém, nos trabalhos de Márquez-Vera e*t al*. (2016), Pradeep *et al.* (2015) e Dekker *et al.* (2009), os autores verificaram que não houve ganho significativo na etapa de mineração de dados após a seleção de atributos.

Com base nos trabalhos pesquisados também foi possível identificar os métodos mais utilizados na mineração dos dados. Como no trabalho de Machado *et al.* (2015), para uma melhor organização, optou-se por apresentar um quadro com os métodos e sua abreviação e em seguida um outro quadro com os autores e os métodos utilizados.

Quadro 3.3 - Métodos mais utilizados e suas abreviações.

Método	Abreviação
Árvore de decisão	AD
Clusterização	C

Feed-forward neural networks	FFNN
Interpretable Classification Rule Mining	ICRM
K-nearest neighbours	KN
Multilayer Perceptron	MP
Naive Bayes	NB
Neural network of radial basis function	RBFN
Probabilistic ensemble simplified fuzzy ARTMAP	PSF
Regra de Indução	RI
Regressão Logística	RL
Support Vector Machine	SVM

Fonte: Elaboração própria

O Quadro 3.4 apresenta os autores dos trabalhos analisados e os métodos utilizados pelos mesmos. Os nomes dos métodos podem ser obtidos no Quadro 3.3.

Quadro 3.4 - Métodos e trabalhos nos quais foram utilizados.

	AD	**C**	**FFNN**	**ICRM**	**KN**	**MP**	**NB**	**RBFN**	**PSF**	**RI**	**RL**	**SVM**
(Cunha et al., 2016)	x	x										
(Márquez-Vera et al., 2016)				x								
(Xing et al., 2016)	x											
(Jiménez-Gómez et al., 2015)	x				x	x	x	x		x		x
(Pradeep et al., 2015)	x									x		
(Tamhane et al., 2014)	x						x				x	
(Adamopoulos, 2013)	x											
(Márquez-Vera et al., 2013)	x									x		
(Dekker et al., 2009)	x									x		
(Lykourentzou et al., 2009)			x						x			x
(Moseley & Mead, 2008)										x		
(Kotsiantis et al., 2003)	x						x					
Total	9	1	1	1	1	1	3	1	1	5	1	2

Fonte: Elaboração própria

Através do Quadro 3.4 pode ser observado que em sua maioria os trabalhos utilizam mais de um método de mineração de dados com o intuito de realizar comparações e detectar o método mais preciso. Entre os mais utilizados estão os métodos de árvore de decisão e os de regra de indução. Entre os trabalhos analisados apenas o trabalho de Cunha *et al.* (2016) utilizou um método de clusterização com o intuito de gerar grupos com os diversos perfis de alunos com risco de evasão.

Para a realização da etapa de mineração de dados, os trabalhos analisados utilizaram o *software* Weka (University of Waikato, 2016). É um software de código aberto, com uma coleção de algoritmos para mineração de dados. Contém recursos para pré-processamento, classificação, regressão, clusterização e regras de associação. Foi utilizado por Márquez-Vera *et al.* (2013), Pradeep *et al.* (2015) e Dekker *et al.* (2009) em seus trabalhos. Em Tamhane *et al.* (2014) além do Weka foi utilizado também o software SPSS Modeler, assim como por Moseley e Mead (2008). No trabalho de Cunha *et al.* (2016) foi utilizado o Analysis Services *software*.

De forma geral os trabalhos analisados seguem as etapas do KDD demonstradas na Figura 3.1. Inicialmente os dados são obtidos, em alguns casos é realizada uma seleção dos atributos mais relevantes. Em seguida os dados são tratados, através de discretização e balanceamento da base de dados, por exemplo. Na etapa de mineração de dados é utilizado mais de um método e por fim é realizada a comparação e análise dos resultados.

3.7. Consolidação dos trabalhos relatados

A partir das etapas do KDD e dos trabalhos pesquisados foram identificadas as macroetapas comumente adotadas em trabalhos na área de mineração de dados educacionais, apresentadas da Figura 3.4.

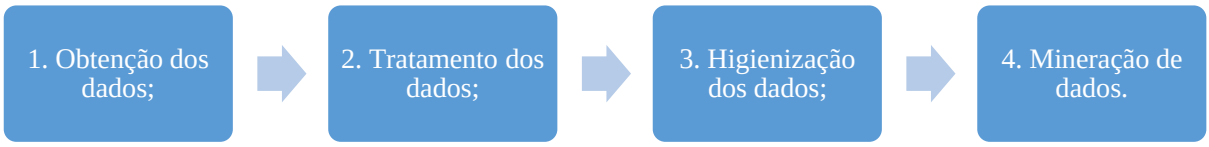

Figura 3.4 - Macroetapas. Fonte: Elaboração própria.

Na primeira etapa, ocorre a definição dos atributos utilizados e também a obtenção dos dados. Como verificado no Quadro 3.2 não ocorre grande variedade de categorias de dados, diante disso, são selecionados dados pertencentes a cada uma das categorias e apresentadas no Quadro 3.5.

Quadro 3.5 – Atributos identificados para cada categoria.

Categoria	Atributo
Assiduidade	Ausências contabilizadas
Dados educacionais	Média final das disciplinas, situação da matrícula, turno
Dados pessoais	Idade, sexo, ocupação, grau de escolaridade, estado civil
Dados Socioeconômicos	Renda familiar, grau de escolaridade dos pais, local de moradia, quantidade de filhos, acesso a computador, tipo de escola que cursou o nível escolar anterior

Fonte: Elaboração própria

No Quadro 3.5 são apresentados os atributos identificados nos trabalhos pesquisados divididos em categorias. Na categoria assiduidade está contido o dado referente à frequência do aluno nas aulas. A categoria dados educacionais é composta por dados que descrevem o desempenho educacional do aluno no curso. Os dados pessoais trazem informações como idade e estado civil. Os dados socioeconômicos são mais abrangentes, por conter, além de dados do aluno, dados relacionados a família e ao seu histórico.

A segunda etapa corresponde ao tratamento dos dados, a importância desta etapa está em tratar alguns dados de forma que os tornem mais propícios à mineração e à análise dos resultados. Um dos tratamentos é a discretização de alguns dados. No caso das médias finais as mesmas são convertidas em categorias de desempenho: excelente, bom, regular, ruim, péssimo. Nos trabalhos de Márquez-Vera *et al.* (2013) e Pradeep *et al.* (2015), foi realizado este tipo de tratamento, com o intuito de tornar os dados mais compreensíveis. Além disso, o atributo relativo à assiduidade é convertido em porcentagem de presença. Em relação ao atributo idade, a maioria das bases de dados possuem a informação de data de nascimento, sendo preciso realizar o cálculo necessário para obter a idade.

A terceira etapa inclui a higienização dos dados. Nesta última etapa de preparação da base de dados é necessário excluir os registros com dados faltantes. É importante ainda analisar os valores contidos na base para a exclusão de registros com valores discrepantes. Em geral, os dados são inseridos nos sistemas por funcionários do registro acadêmico ou pelos próprios alunos, sendo passíveis de erros de digitação. Um exemplo é em relação ao dado idade, pode ser que sejam encontrados registros com o dado idade não condizente com a realidade, sendo necessário excluí-lo.

Na última etapa ocorre a mineração dos dados. Conforme afirmado no trabalho de Márquez-Vera *et al.* (2016), o algoritmo de classificação é a técnica de mineração de dados mais amplamente utilizada para prever evasão escolar. Esta afirmação corrobora os resultados do presente trabalho, onde nove dos doze trabalhos analisados adotaram a árvore de decisão. Além disso, como afirmado por Márquez-Vera *et al.* (2013), estes são métodos caixa branca por fornecerem uma explicação, através de regras para a classificação dos dados, sendo possível extrair quais características ou atributos levaram à classificação dos alunos como evadidos.

3.8. Conclusão

Considerando o grande volume de dados armazenados por uma instituição de ensino, torna-se vantajosa e necessária a utilização de técnicas que propiciem a estratificação de informações relevantes, a partir das quais podem ser traçadas políticas com intuito de trazer melhorias para o ensino. Focando no problema da evasão escolar, foi verificada, através da revisão sistemática a viabilidade de utilização de técnicas de mineração de dados como forma de identificar os alunos propensos a evadir.

Através desta pesquisa, foi possível identificar as principais etapas realizadas e métodos utilizados por trabalhos da área de mineração de dados educacionais com foco na evasão. Foi possível perceber que as etapas realizadas e os métodos adotados são comuns a trabalhos que utilizam como amostra alunos de diversos níveis e modalidades de educação. Porém, como demonstrado neste trabalho, a utilização da mineração de dados na área educacional para o estudo da evasão ainda é um tema pouco explorado.

A partir da revisão sistemática, foi identificado que o método de mineração de dados mais utilizado é o método de classificação de árvore de decisão, utilizado em nove dos doze trabalhos analisados. Em relação à metodologia utilizada pelos autores, foram consolidadas as quatro etapas: obtenção, tratamento, higienização e mineração dos dados.

Foi verificado, ainda, que, na maioria dos trabalhos analisados, o objetivo dos autores é encontrar as melhores taxas de acerto na classificação de alunos evadidos, sem explorar como os atributos que compõem as bases de dados contribuem para a evasão. Entre os trabalhos analisados, apenas no trabalho de Cunha *et al.* (2016) esta etapa foi realizada. Porém, os métodos mais utilizados na etapa de mineração oferecem como resultado, além da taxa de acerto, como os atributos contribuem para a classificação de um aluno como evadido, permitindo que sejam realizadas análises desse tipo.

Acredita-se que através dos resultados obtidos, e sobretudo a partir da aplicação das macroetapas apresentadas, as instituições possam realizar pesquisas a fim de traçar políticas com o intuito de reduzir a evasão, sendo este um problema que afeta não somente o aluno no nível pessoal, mas também a sociedade e a instituição.

Referências

Adamopoulos, P. (2013). What makes a great MOOC? An interdisciplinary analysis of student retention in online courses (Vol. 5, pp. 4720–4740). Apresentado na International Conference on Information Systems (ICIS 2013): Reshaping Society Through Information Systems Design.

Amaral, F. (2016). *Aprenda Mineração de Dados: Teoria e prática*. Rio de Janeiro: Alta Books Editora.

Cardoso, O. N. P. & Machado, R. T. M. (2008). Gestão do conhecimento usando data mining estudo de caso na Universidade Federal de Lavras. Obtido de http://repositorio.ufla.br/jspui/handle/1/184

Cunha, J. A., Moura, E. & Analide, C. (2016). Data mining in academic databases to detect behaviors of students related to school dropout and disapproval. *Advances in Intelligent Systems and Computing*, *445*, 189–198. https://doi.org/10.1007/978-3-319-31307-8_19

Dekker, G. W., Pechenizkiy, M. & Vleeshouwers, J. M. (2009). Predicting students drop out: A case study (pp. 41–50). Apresentado na EDM'09 - Educational Data Mining 2009: 2nd International Conference on Educational Data Mining.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Gil, A. C. (2009). *Como elaborar projetos de pesquisa*. São Paulo: Atlas.

Hoed, R. M. (2016). *Análise da evasão em cursos superiores: o caso da evasão em cursos superiores da área de Computação*. Universidade de Brasília, Brasília, DF. Obtido de http://repositorio.unb.br/bitstream/10482/22575/1/2016_RaphaelMagalh%C3%A3esHoed.pdf

Jiménez-Gómez, M. A., Luna, J. M., Romero, C. & Ventura, S. (2015). Discovering clues to avoid middle school failure at early stages (Vol. 16-20-NaN-2015, pp. 300–304). Apresentado na ACM International Conference Proceeding Series. https://doi.org/10.1145/2723576.2723597

Kotsiantis, S. B., Pierrakeas, C. J. & Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques (Vol. 2774 PART 2, pp. 267–274). Apresentado na Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science).

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G. & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers and Education*, *53*(3), 950–965. https://doi.org/10.1016/j.compedu.2009.05.010

Machado, R. D., Benitez, E. O., Corleta, J. N. & Augusto, G. (2015). Estudo Bibliométrico em mineração de dados e evasão escolar. Apresentado na XI CONGRESSO NACIONAL DE EXCELÊNCIA EM GESTÃO, Rio de Janeiro, RJ.

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa, F. & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, *33*(1), 107–124. https://doi.org/10.1111/exsy.12135

Márquez-Vera, C., Romero, M. & Ventura, S. (2013). Predicting school failure and dropout by using data mining techniques. *Revista Iberoamericana de Tecnologias Del Aprendizaje*, *8*(1), 7–14. https://doi.org/10.1109/RITA.2013.2244695

Mehta, A. A. & Buch, N. J. (2016). Depth and breadth of educational data mining: Researchers' point of view. Em *Proceedings of the 10th International Conference on Intelligent Systems and Control, ISCO 2016*. Coimbatore, India.

Moseley, L. G. & Mead, D. M. (2008). Predicting who will drop out of nursing courses: A machine learning exercise. *Nurse Education Today*, *28*(4), 469–475. https://doi.org/10.1016/j.nedt.2007.07.012

Pradeep, A., Das, S. & Kizhekkethottam, J. J. (2015). Students dropout factor prediction using EDM techniques. Apresentado na Proceedings of the IEEE International Conference on Soft-Computing and Network Security, ICSNS 2015. https://doi.org/10.1109/ICSNS.2015.7292372

Rigo, S. J., Cazella, S. C. & Cambruzzi, W. (2012). Minerando Dados Educacionais com foco na evasão escolar: oportunidades, desafios e necessidades. *Anais do Workshop de Desafios da Computação Aplicada à Educação*, *0*(0), 168–177.

Romero, C. & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Silva, E. L. da & Menezes, E. M. (2005). *Metodologia da Pesquisa e Elaboração de Dissertação* (4º). Florianópolis.

Silva, R. M. da, Gomes, G. R. R., Shimoda, E. & Santos, T. de A. (2010). Percepção dos discentes em relação aos docentes através da aplicação de técnicas e métodos de mineração de dados. Apresentado na XXXVIIII Congresso Brasileiro de educação em Engenharia, Fortaleza, CE, Brasil. Obtido de http://www.abenge.org.br/CobengeAnteriores/2010/artigos/549.doc

Tamhane, A., Ikbal, S., Sengupta, B., Duggirala, M. & Appleton, J. (2014). Predicting student risks through longitudinal analysis (pp. 1544–1552). Apresentado na Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2623330.2623355

Thesaurus. (2013). Thesaurus. Obtido 3 de Fevereiro de 2017, de http://www.thesaurus.com

University of Waikato. (2016). Weka (Versão 3.8). Hamilton, Nova Zelândia.

Xing, W., Chen, X., Stein, J. & Marcinkowski, M. (2016). Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. *Computers in Human Behavior*, *58*, 119–129. https://doi.org/10.1016/j.chb.2015.12.007

4. ARTIGO 3 - COMPORTAMENTO DE ESTUDANTES EVADIDOS DE CURSOS TÉCNICOS: UM ESTUDO UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

4.1. Resumo

Contexto: Dada a problemática da evasão e a quantidade de dados armazenados pelas instituições de ensino, através da utilização da mineração de dados torna-se possível extrair características que descrevem os alunos.
Objetivo: O objetivo do presente trabalho é aplicar a mineração de dados para identificação de comportamento dos alunos evadidos em cursos de nível técnico.
Metodologia: Para o cumprimento do objetivo é realizado um estudo de caso no Instituto Federal Fluminense. São executadas as etapas de obtenção, tratamento, higienização e mineração dos dados. Na última etapa é utilizado o método de classificação J48.
Resultados: A partir das árvores geradas pelo método foi possível identificar que em sua maioria, os alunos evadidos são descritos pelo desempenho ou frequência nas disciplinas dos primeiros módulos do curso.
Conclusão: Os resultados são considerados satisfatórios por terem permitido realizar análises que de fato descrevem comportamentos de alunos que evadiram na amostra analisada.

Palavras-chave: Mineração de dados; Evasão; Ensino Técnico.

4.2. Abstract

Context: *Given the problem of evasion and the amount of data stored by educational institutions, through the use of data mining in the databases of these institutions it is possible to extract characteristics that describe the students.*
Objective: *The objective of the present work is to apply the data mining to identify the behavior of the students evaded in courses of technical level.*
Methodology: *To accomplish the objective, a case study is carried out at the Federal Fluminense Institute. The steps of obtaining, treating, sanitizing and mining the data are performed. In the last step the classification method J48 is used.*
Results: *From the trees generated by the method it was possible to identify that the majority of the students of the evaded profiles are described by the performance or frequency in the subjects of the first modules of the course.*
Conclusions: *The results are considered satisfactory because they allowed to carry out analyzes that in fact describe profiles of students that evaded in the sample analyzed.*

Keywords: Data mining; Dropout; Technical education.

4.3. Introdução

A informatização na área da educação, assim como em diversas outras áreas, deve ser um aliado não só na eficiência dos serviços, mas, também, na eficácia de planos e políticas que buscam aprimoramento e melhores resultados. Na educação, além do seu objetivo final de levar conhecimento e ensinar, busca-se cada vez mais o desenvolvimento de políticas focadas na

redução da evasão escolar. Sendo os dados armazenados sobre toda a trajetória escolar dos alunos uma fonte de informação relevante para o desenvolvimento da gestão escolar.

De acordo com Souza (2014), a evasão é um tema discutido em todo sistema educacional brasileiro, sendo necessário propor estratégias que garantam uma educação inclusiva e que busquem fatores propícios a conclusão do curso por parte do aluno. Cruz (2013) afirma que entre os diversos aspectos apresentados para a evasão dos estudantes, estão fatores que têm ligação com a complexidade da vida pessoal, familiar e financeira. Além disso, existem as instituições responsáveis pela educação e as políticas sociais que nem sempre atendem às necessidades dos estudantes.

Dada a problemática da evasão e a quantidade de dados armazenados pelas instituições de ensino, através da utilização da mineração de dados nas bases de dados dessas instituições, torna-se possível extrair características que descrevem os alunos. E apesar da questão da evasão ser multifatorial, a identificação dessas características é capaz de contribuir para o direcionamento da gestão escolar.

Figueiredo e Salles (2017) ressaltam que indicadores de permanência e evasão podem fornecer informações relevantes às pesquisas que visam avaliar a eficiência e a eficácia de programas governamentais que buscam a ampliação da oferta de cursos técnicos. Porém os dados estatísticos não são suficientes, é necessário extrair informações qualitativas que tragam um mapeamento dos alunos e uma realidade social mais complexa. Os autores afirmam que a evasão é um problema enfrentado pela maioria das instituições de ensino, oriundo de diversos fatores e que quanto mais precoce for identificado o risco da evasão, maiores são as possibilidades de sucesso das políticas de permanência escolar. E para a identificação do risco de evasão, a exposição dos fatores que podem levar a isso torna-se uma forte aliada.

A partir das transformações econômicas e, consequentemente, no mercado de trabalho, a Educação Profissional tem estado cada vez mais em evidência. Tendo governos e empresas constituindo alianças no sentido de criar e manter cursos que, de alguma maneira, possam suprir a demanda de mão de obra para o desenvolvimento do país, contribuindo, igualmente, para a elevação do nível de escolarização dos trabalhadores (Figueiredo & Salles, 2017).

O ensino técnico é uma alternativa, principalmente aos jovens, com objetivo de inserir-se no mercado de trabalho como mão de obra qualificada de forma mais rápida. Porém, assim como cresce a demanda por cursos técnico, cresce também o índice de evasão. Souza (2014) acredita que ao evadir, o aluno está desistindo do que por alguma razão não atendeu às suas

expectativas, e que isso tem um significado forte na vida do educando e consequentemente, da escola.

De acordo com Araújo e Santos (2012), a baixa formação qualificada e falta de habilitação profissional que atualmente existe no mercado de trabalho se deve significativamente ao problema do acesso e da permanência do cidadão em instituições que proporcionem formação qualificada e isso constitui um problema de ordem democrática. As autoras afirmam ainda que por consistir em um desestímulo aos estudos, para compreender a evasão é necessário um estudo profundo em várias perspectivas: aluno, escola e sociedade.

Em seu trabalho Figueiredo e Salles (2017) afirmam que estudos têm revelado a insuficiência de esforços, oriundos das mais diversas esferas de atuação das escolas, no sentido de pensar projetos e desenvolver ações que favoreçam a permanência dos estudantes nos cursos. Mesmo sabendo que será impossível impedir a evasão na sua totalidade, sendo esta com origem na própria estrutura do sistema econômico em que estamos inseridos.

A evasão no ensino técnico tem se tornado um tema em evidência a partir da percepção de que a avaliação estatística de oferta e demanda por esses cursos não são suficientes para analisar o seu resultado para a sociedade. É necessário investigar a proporção de alunos que se forma e as causas da não formação. Como afirmado por Veiga e Bergiante (2016), o desperdício gerado pela evasão escolar não é apenas de ordem financeira, mas também social, por levar o indivíduo a um profundo processo de exclusão.

O objetivo do presente trabalho é aplicar a mineração de dados para identificação do comportamento dos alunos evadidos em cursos de nível técnico. Sua importância se dá no fato de explicitar características comuns aos alunos que evadem, permitindo, por parte da instituição de ensino a criação de políticas focadas em alunos que possuam as características identificadas.

4.4. Metodologia da Pesquisa

4.4.1. População e amostra

A pesquisa é realizada no Instituto Federal de Educação, Ciência e Tecnologia Fluminense (IFFluminense). De acordo com o portal institucional (*iff.edu.br*), o IFFluminense está presente em 11 municípios, é composto por 12 *campi*, um Polo de Inovação, um Centro de Referência em Tecnologia, Informação e Comunicação na Educação e a Reitoria. Conta ainda com polos de Educação a Distância em 4 municípios.

Atualmente, o IFFluminense oferta cursos técnicos nas modalidades: subsequente, para os que já concluíram o ensino médio, concomitante, para os que estão cursando o ensino médio e o integrado com o ensino médio.

São considerados como amostra os cursos técnicos na modalidade subsequente e concomitante. São utilizados os dados dos alunos que estão matriculados, concluíram ou evadiram o curso e que se matricularam a partir de 2014, pois nesse ano foi iniciada a utilização do sistema de inscrição *online*. Nesse sistema, qualquer cidadão que queira participar de algum processo seletivo para estudar no IFFluminense deve realizar o cadastro no sistema de inscrições, informando dados pessoais e respondendo ao questionário socioeconômico, sendo um cadastro único por pessoa que pode ser atualizado a qualquer momento.

4.4.2. Procedimentos técnicos

A metodologia da presente pesquisa baseia-se nas macroetapas consolidadas no trabalho de Cordeiro, Mussa e Hora (2017). Para o cumprimento do objetivo desse trabalho são seguidas as etapas apresentadas na Figura 4.1.

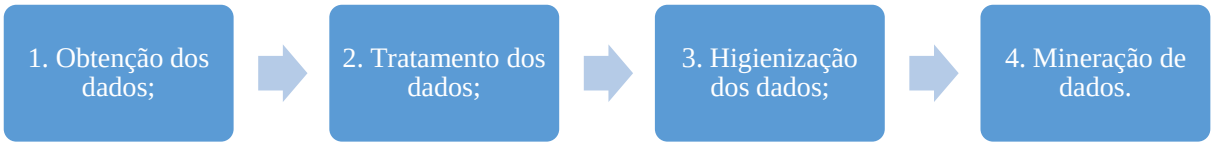

Figura 4.1 - Etapas da pesquisa (Cordeiro *et al.*, 2017).

1. Obtenção de Dados: Na primeira etapa é realizada obtenção dos dados. Nesse trabalho os dados são oriundos de duas bases distintas: a base do sistema acadêmico e a base do sistema de inscrições *online*. O Quadro 4.1 apresenta os dados retirados de cada uma das bases de dados utilizadas.

Quadro 4.1 - Dados retirados da base de dados do sistema acadêmico e do sistema de inscrições.

Dados do sistema acadêmico	Dados do sistema de inscrições
• Matrícula • Data da matrícula • Situação da matrícula • CPF • Data de nascimento • Grau de instrução • Descrição do curso • Modalidade do curso	• CPF • Sexo • PCD (pessoas com deficiência visual, auditiva, física ou intelectual) • Estado civil • Cor • Tipo de escola que cursou o nível escolar anterior

• Turno • *Campus* • Nota • Quantidade de faltas • Disciplina • Quantidade de aulas	• Turno em que cursou o nível escolar anterior • Situação do curso superior • Motivo de escolha do curso • Exercício de atividade remunerada • Renda mensal familiar • Participação na economia familiar • Utilização de computador • Data de preenchimento do formulário • Data de atualização do formulário

Fonte: Elaboração própria.

Como apresentado no Quadro 4.1, foram extraídos 14 atributos da base de dados do sistema acadêmico e 15 da base do sistema de inscrições. Da base de dados do sistema acadêmico, foram retirados dados pessoais e relacionados ao desempenho acadêmico. Do sistema de inscrições, foram retirados dados pessoais que não constavam para todos os registros retirados do sistema acadêmico e também aqueles referentes ao questionário socioeconômico que os alunos respondem ao se cadastrarem no sistema de inscrições. Após obter os dados das duas bases, foi necessário uni-los em uma base única, e para isso, o atributo CPF foi utilizado como referência. As etapas seguintes são realizadas nessa nova base.

2. Tratamento de Dados: Na segunda etapa, são realizados tratamentos na base de dados. Os tratamentos foram:

- Cálculo da proporção de ausência em cada disciplina, dividindo o número de faltas pelo número de aulas ministradas;
- Cálculo da idade através da data de nascimento;
- Cálculo do intervalo entre a data de matrícula e data da última atualização do questionário socioeconômico.

Neste último item são excluídos os registros que possuem um intervalo maior que dois anos. Como o questionário reflete dados variáveis da vida pessoal do aluno, como por exemplo: renda mensal e participação na economia familiar, avaliou-se que restringindo esse intervalo a dois anos haveria menos discrepância da realidade do aluno entre o momento em que atualizou o questionário pela última vez e o momento da matrícula.

Nessa etapa a base de dados também é dividida em outras bases, separando os dados por *campus* e pela modalidade do curso. A Figura 4.2 apresenta as bases formadas após a divisão.

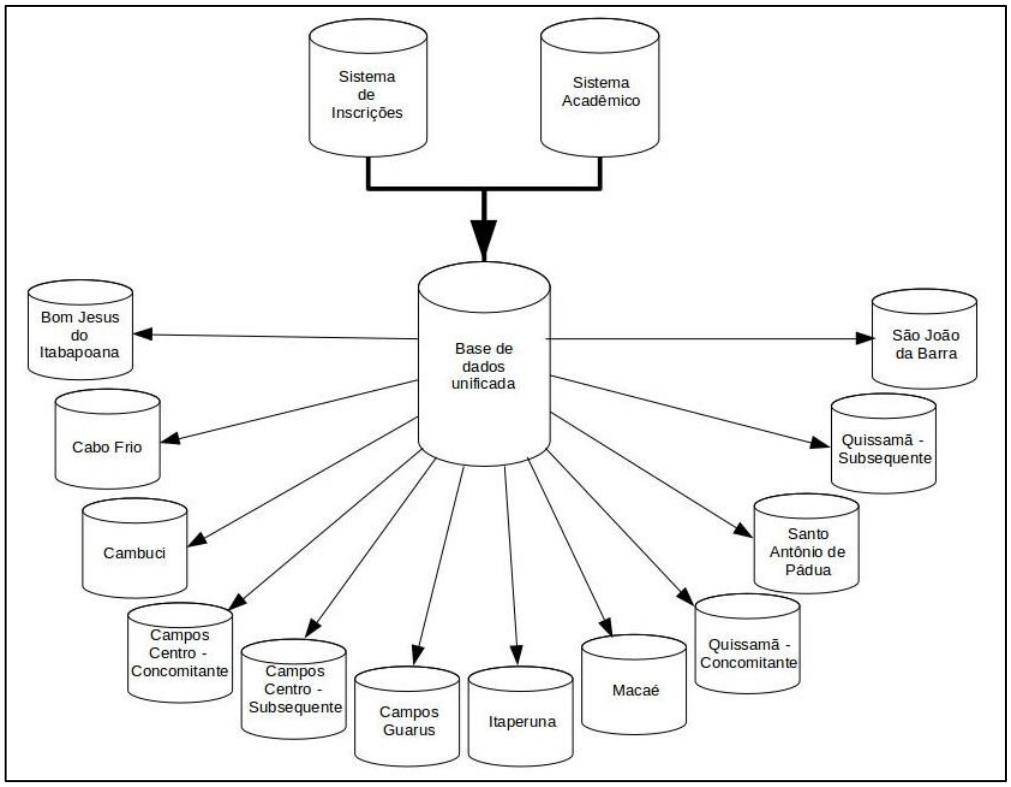

Figura 4.2 - Composição do banco de dados do estudo proposto. Fonte: Elaboração própria.

Como apresentado na Figura 4.2, a partir dos dados disponíveis, foram formadas bases de dados para 10 *campi,* sendo que para os *campi* Campos Centro e Quissamã foram formadas duas bases uma com dados dos alunos dos cursos concomitantes e outra com dados dos alunos dos cursos subsequentes.

Outro importante tratamento realizado nesta etapa é a normalização dos dados que descrevem as disciplinas cursadas e a proporção de ausência em cada disciplina. A nota de cada disciplina é transformada em um atributo assim como a proporção de ausência em cada disciplina.

Ao final da segunda etapa, o banco de dados resultante contém os dezessete atributos apresentados no Quadro 4.3 do Apêndice A, entre esses dados está o dado classificador que é aquele que descreve a situação do aluno, podendo ser matriculado, concluído ou evadido. Além dos dados apresentados no Quadro 4.3 do Apêndice A, nesta etapa, a base ainda contém a modalidade do curso: concomitante ou subsequente; a nota e a proporção de ausência obtida em cada disciplina. Sendo assim, a quantidade de atributos para cada *campus* variou de acordo com a quantidade de disciplinas.

3. Higienização dos Dados: Na terceira etapa ocorre o pré-processamento dos dados que inclui a higienização dos mesmos. A higienizações realizadas foram:

- Exclusão dos registros com dados faltantes
- Exclusão de dados discrepantes:
  - idades menores que 15;
  - grau de escolaridade igual a ensino fundamental completo, ensino fundamental incompleto, não declarado ou alfabetizado.

4. Mineração dos Dados: Na quarta etapa também foram realizadas adaptações nos dados, transformando alguns dados nominais em numéricos, as transformações estão demonstradas no Quadro 4.4 do Apêndice A. Essas alterações foram feitas para três atributos: grau de instrução, renda mensal familiar e exercício de atividade remunerada. Visto que as opções de valores para esses atributos são graduais, a transformação para numérico permite que o método classificador da etapa de mineração gere resultados traçando valores mínimos ou máximos para esses atributos nas classificações.

Na etapa de mineração de dados é utilizado o método de classificação J48. Sendo esse um método de árvore de decisão implementado com base no algoritmo C4.5 proposto por Quinlan (1987). Nesse método, as folhas são as classes e os nós são os atributos. Os ramos da árvore são gerados de acordo com os valores possíveis de cada atributo. Dessa forma, a classificação de uma instância é realizada a partir da folha alcançada percorrendo a árvore de acordo com os valores de seus atributos. Os nós determinam o atributo que será avaliado e o percurso no nível seguinte da árvore é definido a partir do valor do atributo do nó na instância analisada.

Para análise de confiabilidade dos resultados, a etapa de classificação é realizada duas vezes, utilizando métodos diferentes em cada uma delas. Além do método J48 é utilizado para o reteste o método JRip, baseado no algoritmo *Ripper* (*Repeated Incremental Pruning to Produce Error Reduction* ou Poda Incremental Repetida para Produzir Redução de Erro) proposto por Cohen (1995). O método JRip foi escolhido por também ser amplamente utilizado em trabalhos que investigam a evasão a partir da mineração de dados educacionais (Dekker *et al.*, 2009; Jiménez-Gómez *et al.*, 2015; Márquez-Vera *et al.*, 2016; Pradeep *et al.*, 2015).

Cordeiro *et al.* (2017) apontam os métodos J48 (árvore de decisão) e JRip (regras de indução) como os mais utilizados na mineração de dados educacionais, então para garantir a confiabilidade dos dados, é adotada a técnica de teste-reteste (Guttman, 1945), onde o mesmo objeto é submetido à dois tratamentos e seus resultados são comparados. O JRip nesta pesquisa não tem outra função senão apenas aferir a confiabilidade da base de dados, uma vez que o J48 é adotado como algoritmo para gerar os resultados a serem analisados.

Para a execução do método é utilizado o método *k-fold cross validation. N*esse método, os dados são divididos em k conjuntos mutuamente exclusivos. A validação é realizada k vezes, de forma que a cada validação um conjunto diferente é utilizado como teste e os outros k-1 conjuntos são utilizados para treinamento. O resultado final é obtido a partir da média de todas as k validações.

4.5. Resultados

4.5.1. A evasão no IFFluminense

As Figuras 4.3 e 4.4 mostram o quantitativo de alunos concluintes e evadidos que ingressaram a partir de 2014 no IFFluminense nos cursos concomitantes e subsequentes, respectivamente. Não foram considerados os dados do ano de 2017 por não estarem disponíveis. Os dados foram retirados do Sistema Unificado de Administração Pública (SUAP), utilizado no IFFluminense. Nas Figuras 4.3 e 4.4, no eixo das abscissas são apresentados os cursos e o quantitativo de alunos evadidos e concluintes.

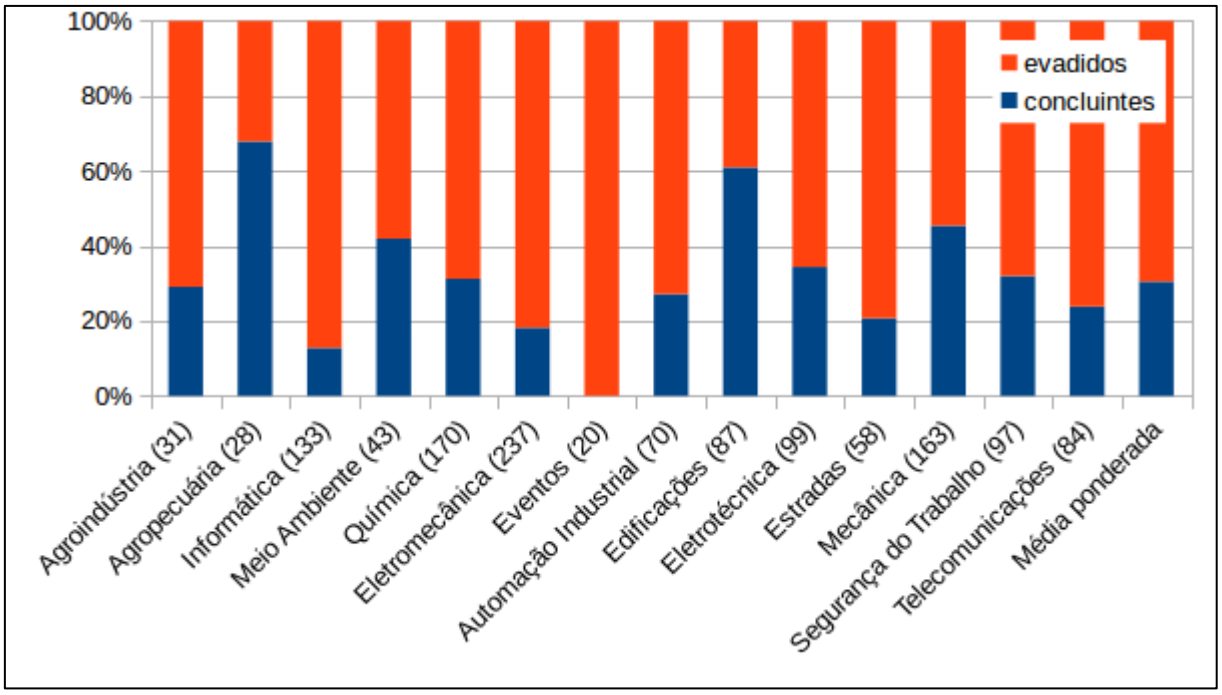

Figura 4.3 - Quantitativo de alunos concluintes e evadidos entre 2014 e 2016 nos cursos concomitantes. Fonte: Elaboração própria.

Na Figura 4.3, é possível observar que apenas nos cursos técnicos de Agropecuária e Edificações o número de concluintes foi superior ao de evadidos. O curso de Eventos foi o curso com o maior número de evasões, seguido pelos cursos de Eletromecânica e Informática.

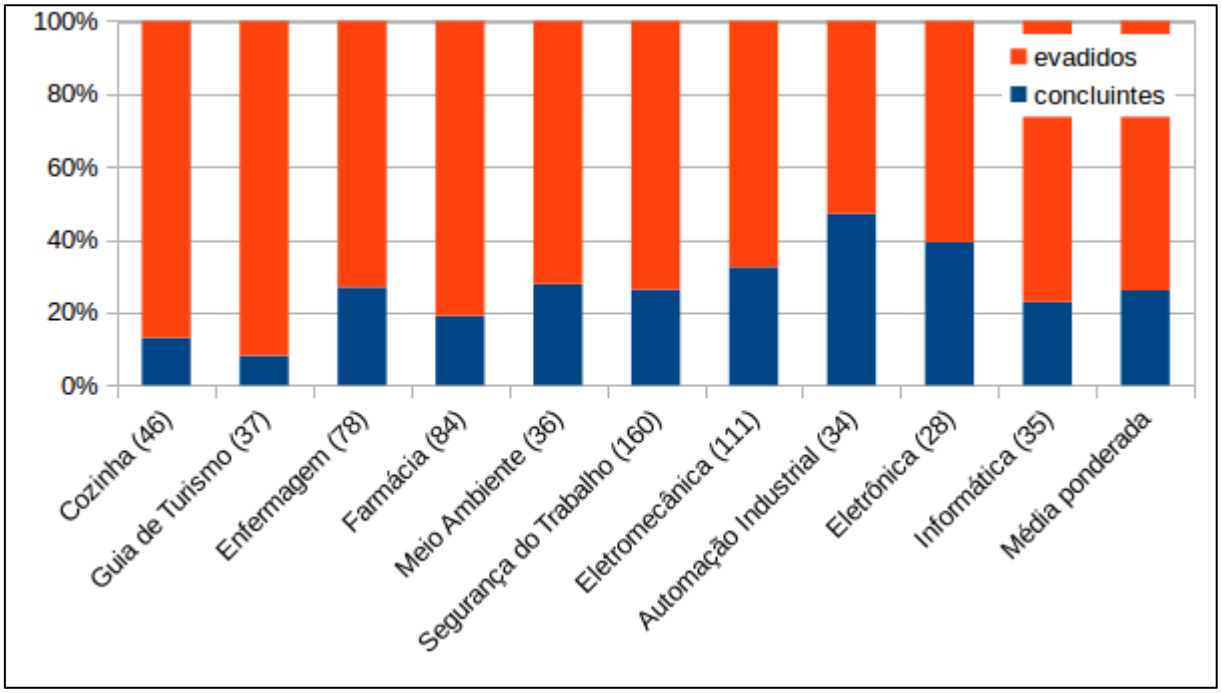

Figura 4.4 - Quantitativo de alunos concluintes e evadidos entre 2014 e 2016 nos cursos subsequentes. Fonte: Elaboração própria.

Na Figura 4.4, relacionada aos cursos da modalidade subsequente, observa-se que nenhum curso possui o quantitativo de concluintes superior ao de evadidos. Sendo o curso de Segurança do Trabalho o com maior proporção de evadidos, e o curso de Automação com menor proporção de evadidos.

Em ambas as Figuras é possível observar a discrepância entre a quantidade de concluintes e evadidos, demonstrando que este é um problema enfrentado pela instituição.

4.5.2. Mineração de dados na identificação do comportamento de alunos evadidos

Nesta seção serão apresentados os resultados obtidos após a execução da etapa de mineração de dados da metodologia apresentada na seção 4.4.2.

O Quadro 4.2 demonstra as taxas de acertos obtidas em cada uma das bases de dados para o método de classificação J48.

Quadro 4.2 - Taxas de acerto para as bases de dados utilizando o método J48

Campus	**Modalidade**	**Taxa de acerto**
Bom Jesus do Itabapoana	Concomitante	73,01 %
Cabo Frio	Concomitante	57,85%
Cambuci	Concomitante	61,11%
Campos Centro	Concomitante	65,96%
	Subsequente	47,06%
Campos Guarus	Subsequente	63,16%
Itaperuna	Concomitante	74,26%
Macaé	Subsequente	61,39%
Santo Antônio de Pádua	Concomitante	77,78%
Quissamã	Concomitante	66,24%
	Subsequente	58,82%
São João da Barra	Concomitante	66,67%

Fonte: Elaboração própria.

Para a obtenção dos resultados apresentados no Quadro 4.2, o método J48 foi executado no *software Weka*. Sendo o valor para o parâmetro de número mínimo igual a 2, que significa o número mínimo de instâncias para geração de um ramo na árvore. Também foi realizado o teste para o valor do parâmetro igual a 10% da quantidade de instâncias da classe mais rara. Porém, em todos os casos, o melhor resultado foi para o valor igual a 2.

Como demonstrado no Quadro 4.2, o melhor resultado obtido foi 77,78% para o *campus* Santo Antônio de Pádua. E o pior resultado foi 47,06% para o *campus* Campos Centro na modalidade subsequente. Os demais resultados variaram entre 74,26% e 57,85%.

A seguir serão apresentadas, para algumas bases, os desfechos de evadido nas árvores obtidas a partir do método J48. Serão analisadas as árvores cuja taxa de acerto é maior que 65%, o nível máximo, ou profundidade da árvore, é 50, e em que houve taxa de acerto para a classe de alunos evadidos.

4.5.2.1. *Campus* Bom Jesus do Itabapoana

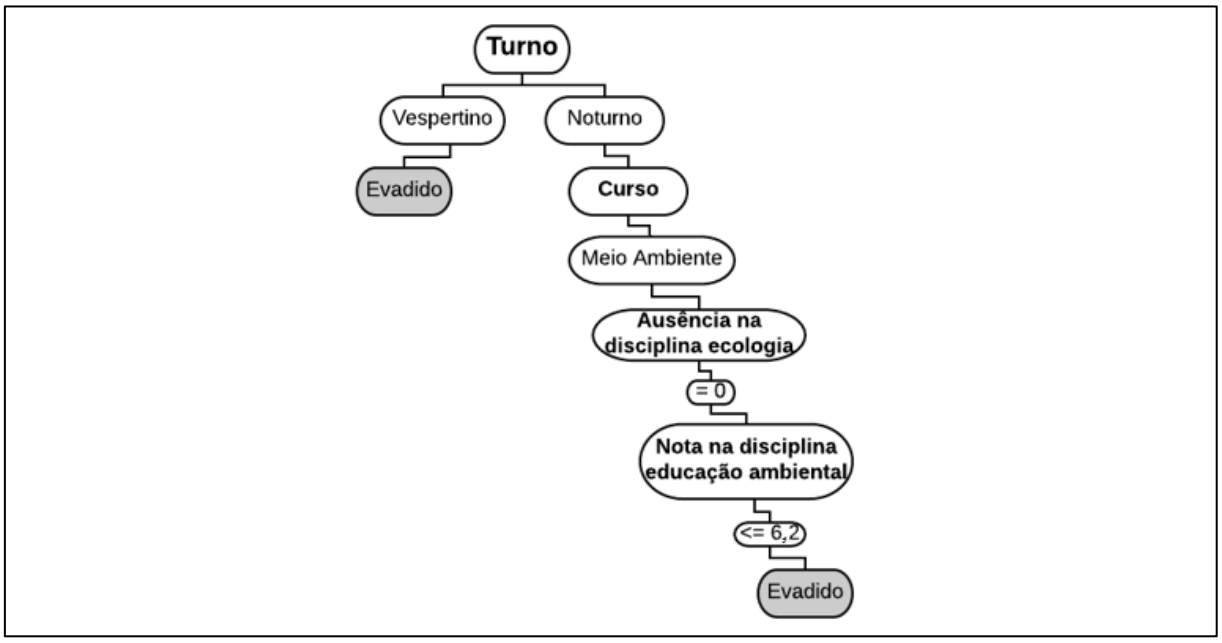

Figura 4.5 - Árvore com desfecho evadido do campus Bom Jesus do Itabapoana. Fonte: Elaboração própria.

Como demonstrado na Figura 4.5, no *campus* Bom Jesus do Itabapoana, onde o atributo principal foi o turno, houve ocorrência de evasão entre os alunos que estudavam no período vespertino. Casos de evasão também foram registrados entre os que estudavam no período noturno, eram do curso técnico em Meio Ambiente, não registraram ausência na disciplina de ecologia mas tenham obtiveram uma nota inferior a 6,2 na disciplina de educação ambiental. Ambas as disciplinas pertencem ao 1º módulo do curso.

4.5.2.2. *Campus* Itaperuna

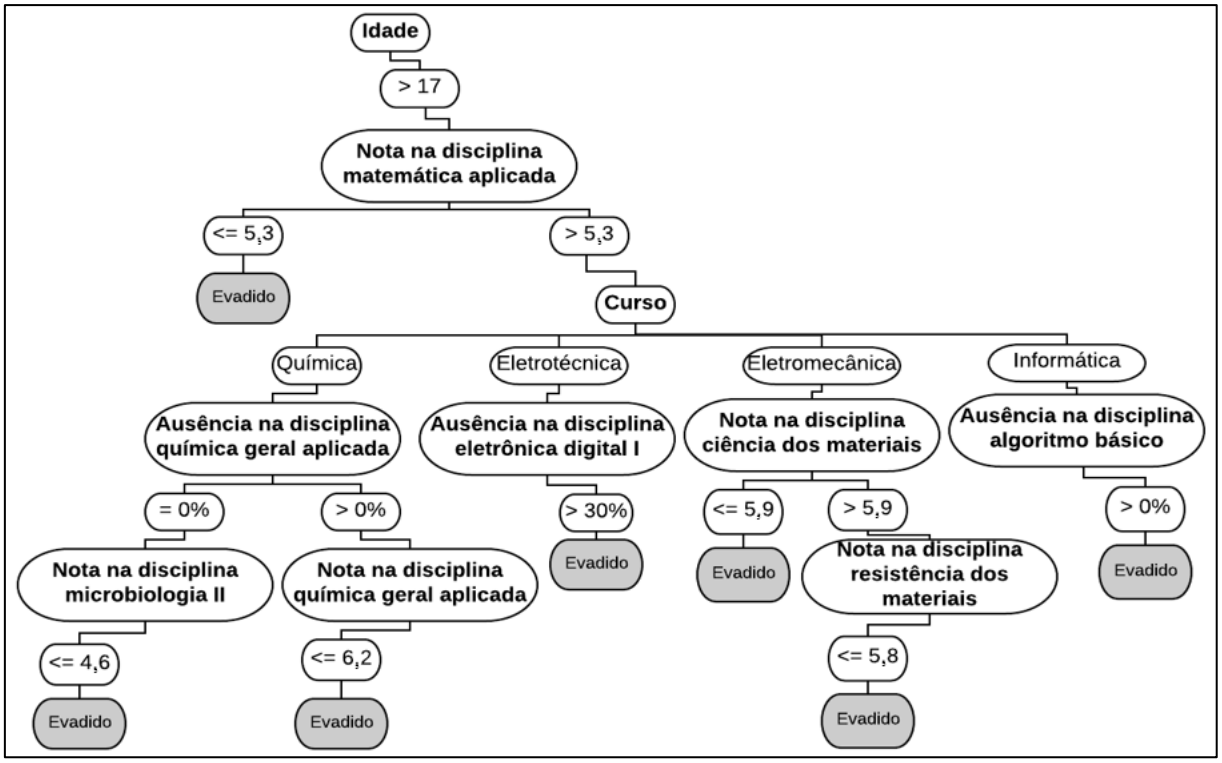

Figura 4.6 - Árvore com desfecho evadido do campus Itaperuna. Fonte: Elaboração própria.

Na árvore do *campus* Itaperuna, demonstrada na Figura 4.6, o primeiro atributo foi a idade seguido pelo desempenho na disciplina matemática aplicada. Estão entre os evadidos os que possuem mais de 17 anos e apresentaram um desempenho igual ou inferior a 5,3 na disciplina matemática aplicado, essa disciplina está presente no 1º módulo dos cursos Eletromecânica, Mecânica, Eletrotécnica e Química. Para os que obtiveram nota superior a 5,3 em matemática aplicada, o próximo atributo é o curso. No curso Técnico em Química, evadiram os que não tiveram ausência nas aulas de química geral aplicada, que compõe o 1º módulo, mas obtiveram nota igual ou inferior a 4,6 em microbiologia II, que pertence ao 2º módulo. Evadiram também os que não apresentaram 100% de presença na disciplina química geral aplicada e apresentaram rendimento igual ou inferior a 6,2 nessa mesma disciplina. No curso Técnico em Eletrotécnica houve evadidos entre os que apresentaram mais de 30% de ausência na disciplina eletrônica digital I no 1º módulo do curso. No curso Técnico em Eletromecânica, evadiram alunos que obtiveram nota igual ou inferior a 5,9 em ciência dos materiais, sendo essa uma disciplina do 1º módulo, ou obtiveram nota maior que 5,9 nessa mesma disciplina, mas em resistência dos materiais, que pertence ao 2º módulo, obtiveram uma nota igual ou inferior a 5,8. Já no Curso Técnico em Informática houve evasão entre os que não foram a todas as aulas da disciplina de algoritmo básico.

4.5.2.3. *Campus* Santo Antônio de Pádua

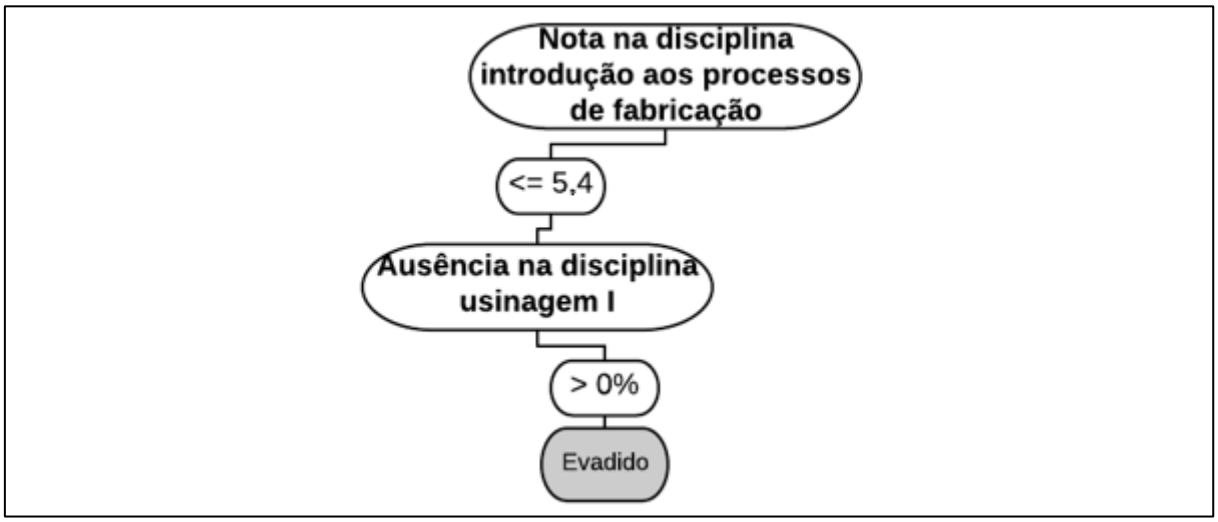

Figura 4.7 - Árvore com desfecho evadido do campus Santo Antônio de Pádua. Fonte: Elaboração própria.

Para o *campus* Santo Antônio de Pádua, as duas disciplinas que compõem a árvore apresentada na Figura 4.7, pertencem ao 1º módulo do curso Técnico em Mecânica. Entre os evadidos estão aqueles que obtiveram nota inferior ou igual a 5,4 na disciplina introdução aos processos de fabricação e faltaram pelo menos a uma aula de usinagem I.

4.5.2.4. *Campus* Quissamã

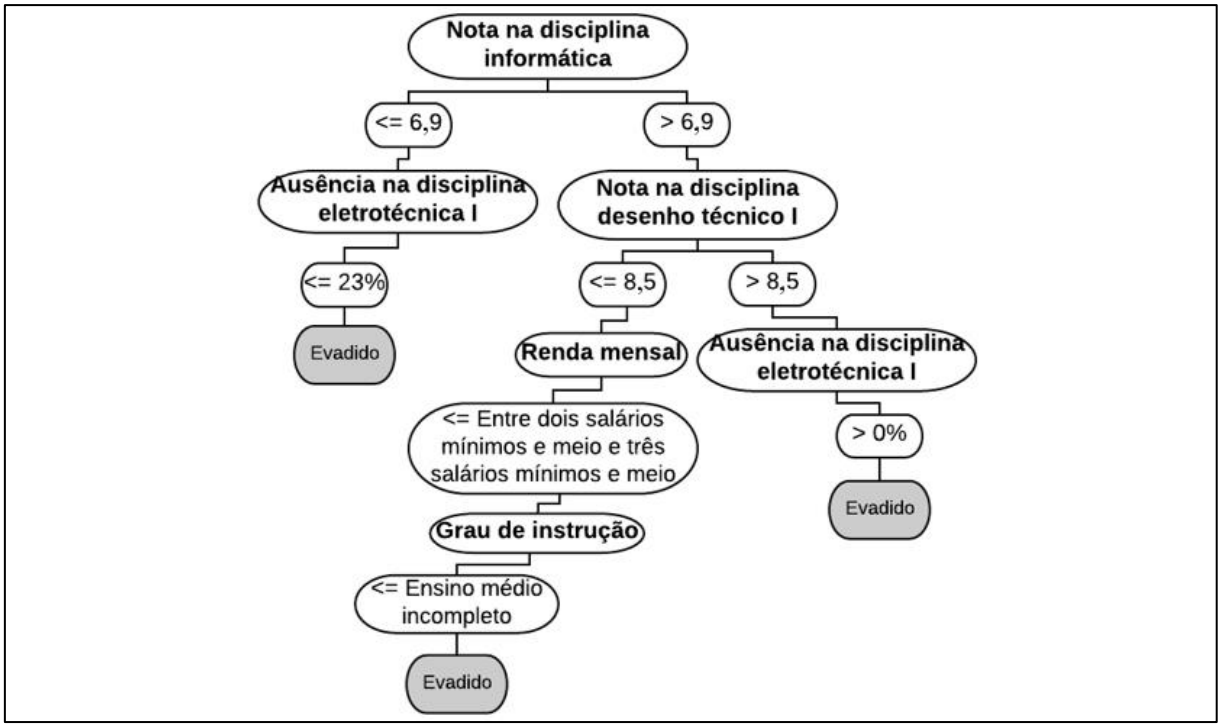

Figura 4.8 - Árvore com desfecho evadido do campus Quissamã na modalidade concomitante. Fonte: Elaboração própria.

No *campus* Quissamã, em que a base de dados da modalidade concomitante foi composta apenas por alunos do curso Técnico em Eletromecânica, o primeiro atributo que aparece na árvore, apresentada na Figura 4.8, é o desempenho na disciplina informática. Entre os alunos evadidos, estão aqueles que obtiveram nota igual ou inferior a 6,9 em informática e também um índice máximo de ausência de 23% nas aulas de eletrotécnica I. Também houve evadidos entre os que obtiveram nota superior a 6,9 na disciplina informática, sendo que parte desses obtiveram nota igual ou inferior a 8,5 na disciplina desenho técnico I, renda mensal máxima de três salários mínimos e meio e ensino médio incompleto. Os evadidos que obtiveram nota superior a 6,9 na disciplina informática e superior a 8,5 na disciplina desenho técnico I também apresentaram alguma ausência nas aulas de eletrotécnica I. Todas as disciplinas que aparecem na árvore compõem o 1º módulo do curso.

Analisando as quatro árvores apresentadas, de início é possível observar que em todas elas o atributo nota foi considerado em alguma disciplina. Em 4 delas foi considerada também a porcentagem de ausência nas aulas. Outro ponto interessante é que todas as disciplinas que aparecem nas árvores pertencem ao 1º ou 2º módulo dos cursos. Os outros atributos considerados foram: turno, renda mensal familiar, grau de instrução, curso e idade.

Para as bases de dados compostas por cursos da modalidade subsequente, como apresentado no Quadro 4.2, foram obtidas taxas de acerto mais baixas. Será analisada a árvore cuja taxa de acerto foi superior a 60% e em que houve taxa de acerto para a classe de alunos evadidos.

4.5.2.5. *Campus* Campos Guarus

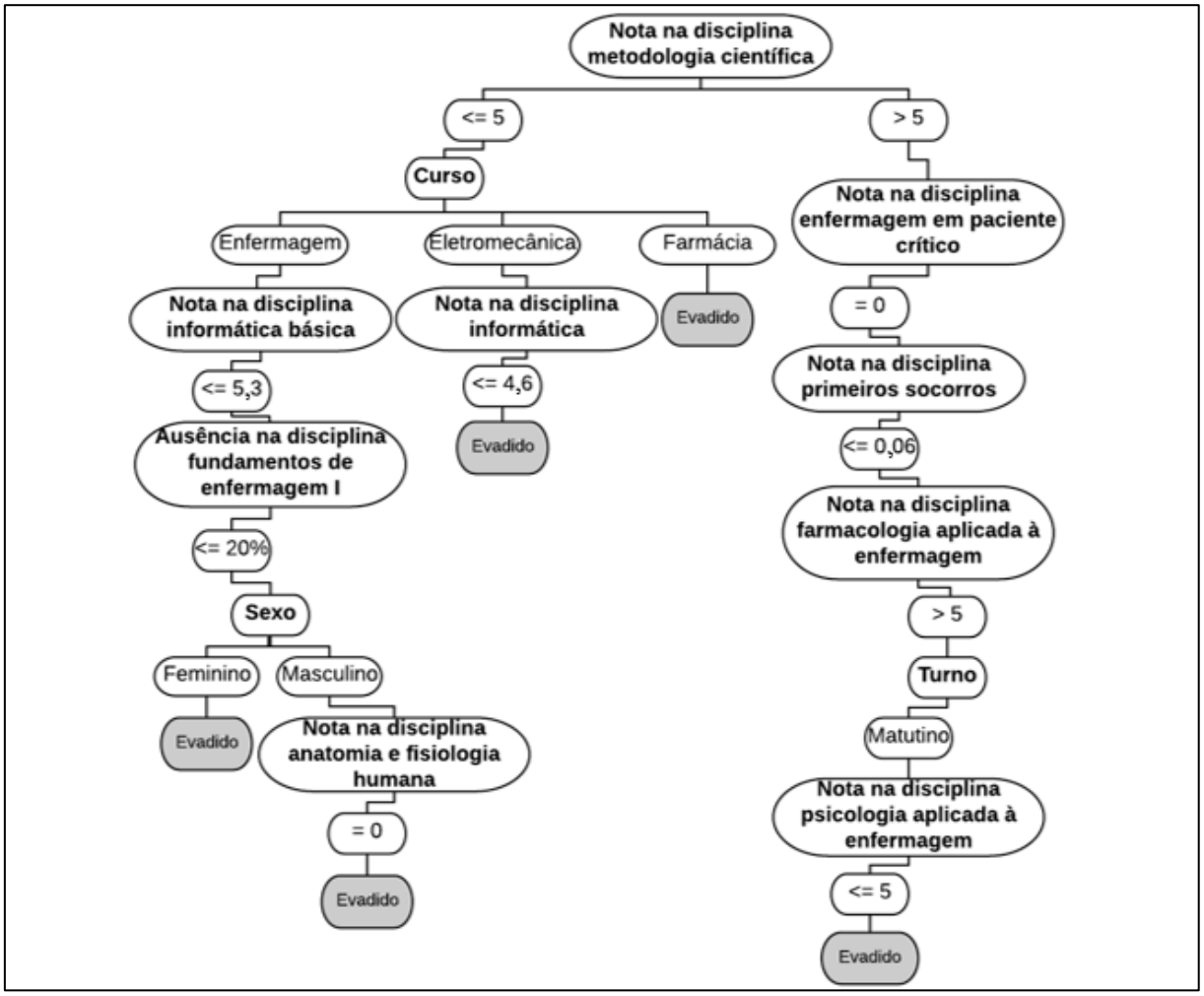

Figura 4.9 - Árvore com desfecho evadido do campus Campos Guarus. Fonte: Elaboração própria.

Na árvore da Figura 4.9, gerada para a base de dados do *campus* Guarus ocorreram cinco desfechos para a classe evadido. O principal atributo é a nota na disciplina metodologia científica. Ocorreu evasão entre os que obtiveram nessa disciplina nota inferior ou igual a 5 e eram do curso de Enfermagem, Eletromecânica ou Farmácia. Sendo que, no curso Técnico em Enfermagem, onde ocorreu o primeiro desfecho, houve evasão entre os que obtiveram nota igual ou inferior a 5,3 na disciplina informática básica mesmo apresentando uma ausência igual ou inferior a 20% das aulas de fundamentos de enfermagem I. Desses, houve evasão entre os

alunos do sexo feminino, e do sexo masculino para os que obtiveram nota igual a 0 na disciplina anatomia e fisiologia humana. Todas as disciplinas citadas compõem o 1º módulo do curso. No curso Técnico de Eletromecânica, foi observada evasão entre os que obtiveram nota igual ou inferior a 4,6 na disciplina de informática, que pertence ao 1º módulo. Já no curso Técnico de Farmácia apenas a nota inferior ou igual a 5 na disciplina de metodologia científica foi decisiva para a evasão. Já entre os alunos evadidos que obtiveram nota superior a 5 em metodologia científica, o comportamento identificado foi daqueles que obtiveram nota igual 0 na disciplina enfermagem em paciente crítico, nota inferior ou igual a 0,06 em primeiros socorros, nota superior a 5 na disciplina farmacologia aplicada à enfermagem, cursavam no turno matutino e obtiveram nota igual ou inferior a 5 em enfermagem. Para esse último desfecho, todas as disciplinas são do curso de enfermagem, sendo que a disciplina enfermagem com paciente critico pertence ao 3º módulo, primeiros socorros e farmacologia aplicada à enfermagem pertencem ao 2º módulo e psicologia aplicada à enfermagem ao 1º módulo do curso.

A partir da árvore analisada para a base de dados composta por cursos na modalidade subsequente, é possível perceber que, diferenciado das árvores para os cursos na modalidade concomitante, além de disciplinas dos 1º e 2º módulo dos cursos, esteve presente também o atributo de nota em uma disciplina do 3º módulo.

O Apêndice B apresenta todas as árvores geradas e completas. Além disso, é apresentada a matriz de confusão para cada uma das bases e o quantitativo de registros por curso para cada classe: matriculado, evadido e concluído.

4.6. Discussão de Resultados

A partir dos resultados obtidos analisa-se que em geral os atributos que mais descrevem os comportamentos dos alunos evadidos no ensino técnico estão relacionados à ausência e às notas obtidas nas disciplinas do curso. Corroborando com o estudo de Jiménez-Gómez *et al.* (2015), foi vislumbrado que os comportamentos podem ser identificados a partir de características detectáveis já nos módulos iniciais do curso.

No estudo realizado por Souza (2014), busca-se a identificação de fatores que levam a evasão em cursos técnicos subsequentes no Instituto Federal do Rio Grande do Norte através da aplicação de questionários e realização de entrevistas. A autora constatou que os principais fatores destacados pelos alunos estão relacionados ao currículo, disciplinas de difícil

compreensão e metodologia dos professores. Assim como na presente pesquisa, em que, em sua maioria as regras incluem o desempenho ou frequência em alguma disciplina.

Bastos e Gomes (2017), que investigaram fatores que levam a evasão no ensino técnico, identificaram que a maioria dos alunos evade no segundo módulo do curso, e que em geral as disciplinas que determinam o comportamento dos alunos evadidos ocorrem no primeiro módulo. Um ponto em comum com os resultados obtidos nesta pesquisa, onde a maioria dos atributos que compõem as árvores estão relacionados à disciplinas do 1º ou 2º módulo dos cursos.

Cruz (2013) investiga as principais causas de evasão de cursos profissionalizantes gratuitos. Em relação à análise dos alunos evadidos, o autor constata que a maioria são jovens entre 18 e 27 anos e renda mensal de até R$ 2040,00. O resultado relacionado ao *campus* Itaperuna apresentou como principal atributo a idade, onde os alunos que evadiram possuíam no mínimo 17 anos, englobando alunos jovens assim como no trabalho de Cruz (2013). Em relação a renda mensal, no *campus* Quissamã também foi obtido resultado compatível.

Bastos e Gomes (2017), Cruz (2013) e Veiga e Bergiante (2016) identificaram como determinante para a evasão o fato do aluno ter que conciliar os estudos com o trabalho. O presente estudo não observou a mesma realidade, já que não foram obtidos resultados onde o atributo decisivo foi aquele que descreve se o aluno exerce atividade remunerada.

No trabalho de Cunha *et al.* (2016), além de fatores relacionados ao perfil socioeconômico do aluno, foi identificado que o rendimento dos alunos em determinadas disciplinas também foi um fator em comum entre os alunos evadidos. Os autores verificaram que os alunos evadidos demonstraram dificuldades em disciplinas relacionadas à lógica e à matemática assim como disciplinas da área técnica. Na presente pesquisa os atributos que mais descreveram o comportamento dos alunos evadidos também estão relacionados às disciplinas técnicas, e no *campus* Itaperuna a primeira disciplina da árvore é matemática aplicada, corroborando com os resultados de Cunha *et al.* (2016).

4.7. Conclusão

A partir de dados armazenados em sistemas utilizados pelas instituições de ensino e de técnicas de mineração de dados que propiciem a extração de informações relevantes desses dados é possível explorar o problema de evasão escolar. Esta exploração se dá no sentido de

oferecer informações que contribuem para a elaboração de políticas estudantis que busquem reduzir a evasão.

De acordo com Cruz (2013), pode-se concluir que a análise da evasão no ensino é uma tarefa desafiadora, por buscar compreender que o modo como o estudante enfrenta e reage às novas tarefas, às mudanças e às vivências escolares pode influenciar no sucesso do processo de ensino-aprendizagem, diplomação, permanência e evasão escolar.

Através deste trabalho buscou-se aplicar as etapas e métodos mais utilizados, de acordo com Cordeiro *et al.* (2017), para extração de características de alunos evadidos no ensino técnico. O estudo de caso foi realizado no Instituto Federal Fluminense, mais especificamente, utilizando dados de alunos de cursos técnicos nas modalidades concomitante e subsequente.

O fracasso escolar produz marcas no aluno, favorecendo a baixa estima e o desenvolvimento de um processo depreciativo que promove a desmotivação com os estudos e com a escola de modo geral. Porém, as marcas no desenvolvimento humano são mais profundas, uma vez que o aluno passa a ter comprometidas suas potencialidades e habilidades (Araújo & Santos, 2012). Isso reforça a necessidade de estudos no sentido de entender a evasão e acredita-se que este trabalho contribui para o alcance desses objetivos.

A partir dos resultados obtidos, conclui-se que, de forma geral, não há diferença entre os comportamentos dos alunos evadidos de cursos técnicos na modalidade concomitante e na modalidade subsequente. Em ambos os casos, as disciplinas dos primeiros módulos que traçam os comportamentos. Em sua maioria, são disciplinas técnicas e os fatores determinantes são o desempenho e a ausência nas aulas.

Apesar das taxas de acertos obtidas não terem ultrapassado 80%, ainda assim, os resultados são considerados satisfatórios por terem permitido realizar análises que de fato descrevem os comportamentos de alunos que evadiram na amostra analisada. Ainda que não representem um comportamento considerado genérico, no universo dos cursos técnicos são resultados pertinentes ao estudo de caso realizado.

Acredita-se que a principal contribuição dessa pesquisa é explicitação dos fatores que descrevem os comportamentos de alunos evadidos no ensino técnico. Além da utilização de técnicas de mineração em dados educacionais, área identificada como emergente, foram identificados os atributos que descrevem o comportamento dos alunos evadidos no ensino técnico, sendo este um nível de ensino pouco explorado na mineração de dados educacionais.

Entre as dificuldades encontradas nesse trabalho está o fato do sistema acadêmico não ser constantemente atualizado em todos os *campi*. Isso resulta em dados faltantes e consequentemente a exclusão de alunos da base de dados analisada.

Como trabalhos futuros propõe-se aplicar a seleção de atributos previamente a fase de mineração de dados, com o intuito de verificar se houve melhora nas taxas de acerto obtidas. Além disso, com o mesmo intuito, poderá ser investigada a utilização de outros métodos de classificação. Futuramente, sugere-se ainda analisar o comportamento dos alunos concluintes e aplicar a mesma metodologia em bases de dados contendo apenas dados socioeconômicos com o intuito de observar como esses dados descrevem o comportamento do aluno evadido.

Referências

Araújo, C. F. de & Santos, R. A. dos. (2012). A EDUCAÇÃO PROFISSIONAL DE NÍVEL MÉDIO E OS FATORES INTERNOS/ EXTERNOS ÀS INSTITUIÇÕES QUE CAUSAM A EVASÃO ESCOLAR. Em *The 4ŧħ international congress university industry cooperation*. Taubate, SP.

Bastos, O. G. A. & Gomes, C. F. S. (2017). A evasão escolar no Ensino Técnico: entendendo e enfrentando as dificuldades - Um estudo de caso do CEFET-RJ. Apresentado no Congresso Nacional em excelência em gestão.

Cohen, W. W. (1995). Fast Effective Rule Induction. Em *In Proceedings of the Twelfth International Conference on Machine Learning* (pp. 115–123). Morgan Kaufmann.

Cordeiro, R. G., Mussa, M. de S. & Hora, H. M. R. da. (2017). Mineração de dados educacionais com foco na evasão: Uma revisão sistemática. In Anais do VII Encontro Fluminense de Engenharia de Produção. Nova Iguaçu, RJ: ENFEPro.

Cruz, A. P. da. (2013). *Evasão nos cursos técnicos profissionalizantes: uma análise das principais causas e identificação de perfil dos alunos evadidos do Senac Sete Lagoas*. Fundação Pedro Leopoldo, Pedro Leopoldo.

Cunha, J. A., Moura, E. & Analide, C. (2016). Data mining in academic databases to detect behaviors of students related to school dropout and disapproval. *Advances in Intelligent Systems and Computing*, *445*, 189–198. https://doi.org/10.1007/978-3-319-31307-8_19

Dekker, G. W., Pechenizkiy, M. & Vleeshouwers, J. M. (2009). Predicting students drop out: A case study (pp. 41–50). Apresentado na EDM'09 - Educational Data Mining 2009: 2nd International Conference on Educational Data Mining.

Figueiredo, N. G. da S. & Salles, D. M. R. (2017). Educação Profissional e evasão escolar em contexto: motivos e reflexões. *Ensaio: Avaliação e Políticas Públicas em Educação*, *25*(95), 356–392. https://doi.org/10.1590/s0104-40362017002500397

Jiménez-Gómez, M. A., Luna, J. M., Romero, C. & Ventura, S. (2015). Discovering clues to avoid middle school failure at early stages (Vol. 16-20-NaN-2015, pp. 300–304). Apresentado na ACM International Conference Proceeding Series. https://doi.org/10.1145/2723576.2723597

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa, F. & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, *33*(1), 107–124. https://doi.org/10.1111/exsy.12135

Pradeep, A., Das, S. & Kizhekkethottam, J. J. (2015). Students dropout factor prediction using EDM techniques. Apresentado na Proceedings of the IEEE International Conference on Soft-Computing and Network Security, ICSNS 2015. https://doi.org/10.1109/ICSNS.2015.7292372

Quinlan, J. R. (1987). Simplifying Decision Trees. *Int. J. Man-Mach. Stud.*, *27*(3), 221–234. https://doi.org/10.1016/S0020-7373(87)80053-6

Souza, J. (2014). *PERMANÊNCIA E EVASÃO ESCOLAR: UM ESTUDO DE CASO EM UMA INSTITUIÇÃO DE ENSINO PROFISSIONAL – Mestrado em Gestão e Avaliação da Educação Pública*. Universidade Federal de Juiz de Fora, Juiz de Fora.

Veiga, C. & Bergiante, N. (2016). Fatores predominantes da evasão escolar no ensino médio profissional: uma revisão de literatura. Apresentado na Congresso Nacional de Execelência em Gestão.

Apêndice A

Quadro 4.3 - Atributos utilizados com descrição e tipos de dados

Campo	Descrição	Tipo
Dados educacionais		
sit_matricula	Situação da matrícula • Matriculado • Concluído • Evadido • Concludente	texto
desc_curso	Nome do curso	texto
desc_turno	Turno em que ocorrem as aulas	texto
Dados Pessoais		
idade	Idade	numérico
sexo	Sexo	texto
estado_civil	Estado civil • Solteiro(a) • Casado(a) • Viúvo(a) • Separado(a) Judicialmente • Divorciado(a) • Outros	texto
pcd	Possui deficiência visual, auditiva, física ou intelectual • Sim • Não	booleano
desc_grau_instrucao	Grau de instrução • Ensino médio incompleto • Ensino médio completo • Ensino Técnico • Superior incompleto • Superior completo • Pós-graduação incompleta • Pós-graduação • Doutorado	numérico
Dados do questionário socioeconômico		
cor_raca	Cor ou raça • Branco • Negro • Amarela • Indígena • Outra	texto
tipo_escola_fundamental	Em que tipo de estabelecimento você cursa/cursou o ensino médio (2º grau) e/ou fundamental ? • Somente em estabelecimento público	texto

	• Somente em estabelecimento particular • Parte em pública e parte em particular • No exterior • Nenhuma das alternativas anteriores	
periodo_fundamental	Em que período cursa/cursou o ensino médio (2º grau) e/ou fundamental? • Somente diurno • Somente noturno • Parte diurna e parte noturna • Outro	texto
situacao_curso_superior	Você já fez ou vem fazendo algum curso superior, qual das seguintes alternativas melhor expressa sua situação no referido curso? • Abandonei-o • Já o concluí • Pretendo desistir do curso atual se passar neste processo seletivo • Pretendo frequentar dois cursos ao mesmo tempo • Outro	texto
motivo_escolha_curso	Qual o motivo predominante na escolha do curso para o qual você se inscreveu? • Possibilidade de poder contribuir com a sociedade • Prestígio Social da Profissão • Possibilidade de realização pessoal • Amplas possibilidades salariais • Tradição profissional da família • Mercado de trabalho • Possibilidade de dar continuidade a seus estudos • Baixa concorrência pelas vagas	texto
exerce_atividade_remunerada	Você exerce alguma atividade remunerada ? • Não • Sim, em tempo parcial (cerca de 20 horas semanais) • Sim, em tempo integral (cerca de 30 horas semanais) • Sim, mas se trata de um trabalho eventual	numérico
renda_mensal_familia	Qual a renda mensal da sua família ? • Até meio salário mínimo • Entre meio salário mínimo e uma salário mínimo e meio • Entre um salário mínimo e meio e dois salários mínimos e meio • Acima de três salários mínimos e meio • Entre dois salários mínimos e meio e três salários mínimos e meio	numérico
participacao_economia_familia	Qual a sua participação na vida econômica da família ? • Trabalho, mas recebo ajuda financeira da família ou de outras pessoas	texto

	• Trabalho e sou o principal responsável pelo sustento da família • Trabalho, sou responsável pelo meu próprio sustento e contribuo parcialmente para o sustento da família ou de outra pessoa • Não trabalho e meus gastos são financiados pela família ou por outras pessoas	
utiliza_computador	Você costuma usar microcomputadores ? • No trabalho e em casa • Em casa • Em casa de amigos e parentes • Não tenho acesso a microcomputadores • No trabalho	texto

Fonte: Elaboração própria.

Quadro 4.4 - Conversão dos atributos nominais para numéricos

Valor nominal	Valor numérico
Atributo: Grau de instrução	
Ensino médio incompleto	1
Ensino médio completo	2
Ensino técnico	3
Superior completo	4
Superior incompleto	5
Pós-graduação incompleta	6
Pós-graduado	7
Doutorado	8
Atributo: Renda mensal familiar	
Até meio salário mínimo	1
Entre um salário mínimo e meio e dois salários mínimos e meio	2
Entre dois salários mínimos e meio e três salários mínimos e meio	3
Entre meio salário mínimo e uma salário mínimo e meio	4
Acima de três salários mínimos e meio	5
Atributo: Exercício de atividade remunerada	
Não	0
Sim, mas se trata de um trabalho eventual	1
Sim, em tempo parcial (cerca de 20 horas semanais)	2
Sim, em tempo integral (cerca de 30 horas semanais)	3

Fonte: Elaboração própria.

Apêndice B

Campus Bom Jesus do Itabapoana

Modalidade: Concomitante

Quadro 4.5 - Quantitativo de alunos do *campus* Bom Jesus do Itabapoana por curso e por situação de matrícula.

Curso	Matriculado	Evadido	Concluído	Total
Técnico em Agroindústria	0	17	7	24
Técnico em Agropecuária	18	13	2	33
Técnico em Alimentos	11	0	0	11
Técnico em Informática	65	19	6	90
Técnico em Meio Ambiente	31	21	16	68
Total	125	70	31	226

Fonte: Elaboração própria.

Quadro 4.6 - Matriz de confusão do *campus* Bom Jesus do Itabapoana.

	Classificado como:		
	matriculado	**evadido**	**concluído**
matriculado	120	1	4
evadido	38	29	3
concluído	9	6	16

Fonte: Elaboração própria.

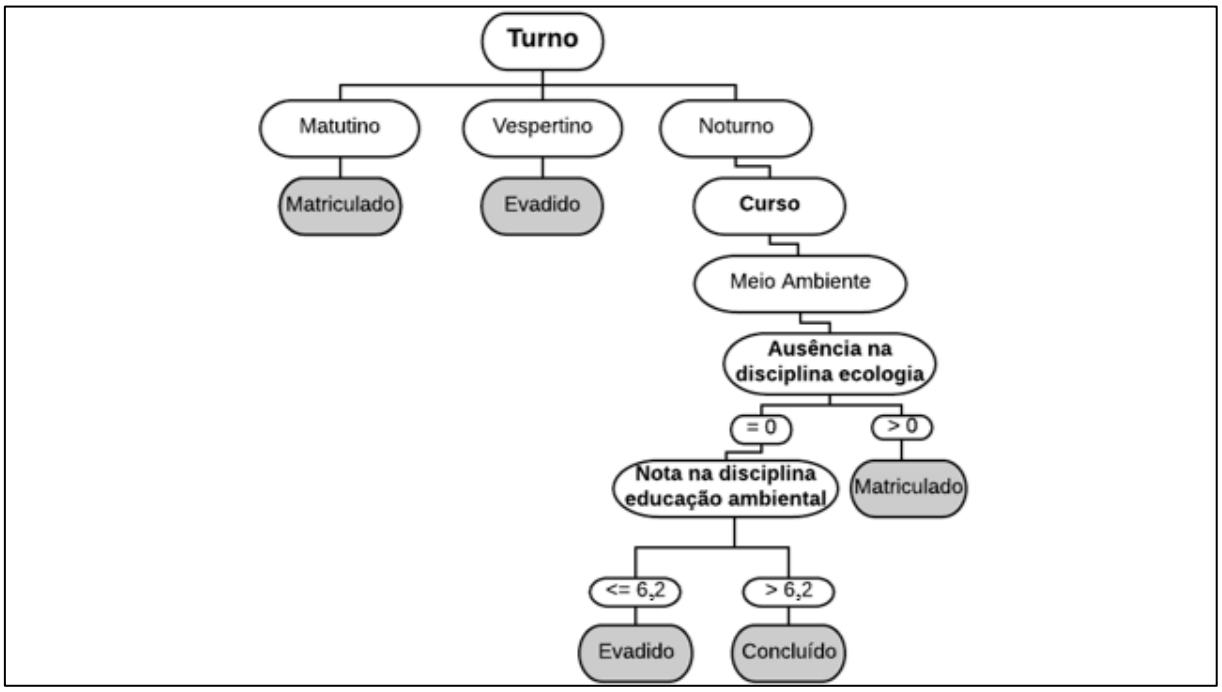

Figura 4.10 - Árvore de classificação do *campus* Bom Jesus do Itabapoana. Fonte: Elaboração própria.

Campus Cabo Frio

Modalidade: Concomitante

Quadro 4.7 - Quantitativo de alunos do *campus* Cabo Frio por curso e por situação de matrícula.

Curso	Matriculado	Evadido	Concluído	Total
Técnico em Cozinha	18	10	11	39
Técnico em Eletromecânica	60	14	37	111
Técnico em Eventos	22	3	0	25
Técnico em Química	34	12	21	67
Total	134	39	69	242

Fonte: Elaboração própria.

Quadro 4.8 - Matriz de confusão do *campus* Cabo Frio.

	Classificado como:		
	matriculado	**evadido**	**concluído**
matriculado	133	1	1
evadido	30	5	4
concluído	67	0	2

Fonte: Elaboração própria.

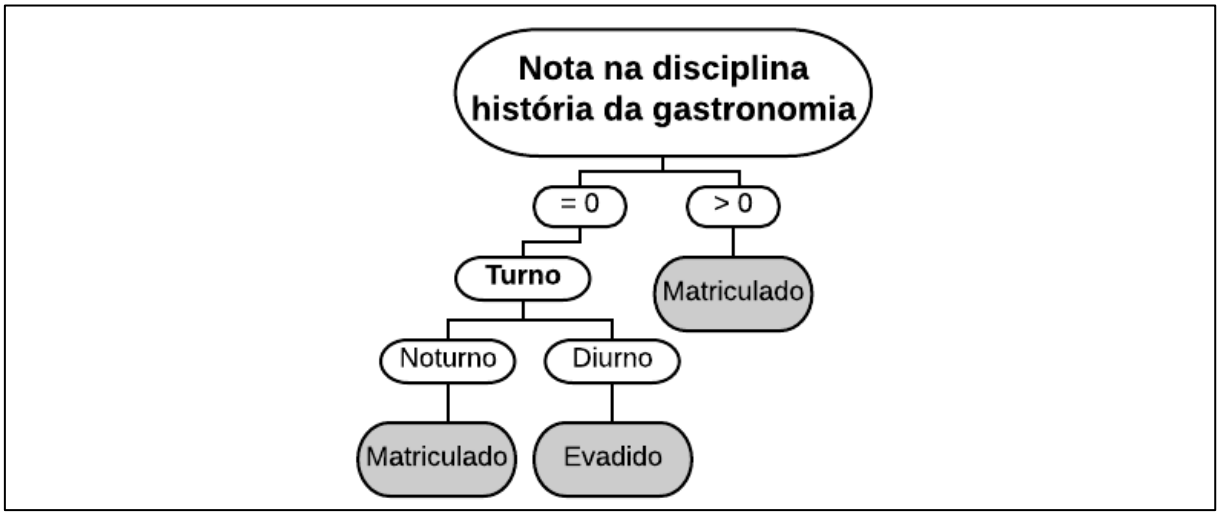

Figura 4.11 - Árvore de classificação do *campus* Cabo Frio. Fonte: Elaboração própria.

Campus Avançado Cambuci

Modalidade: Concomitante

Quadro 4.9 - Quantitativo de alunos do *campus* Cambuci por curso e por situação de matrícula

Curso	Matriculado	Evadido	Concluído	Total
Técnico em Agropecuária	24	8	4	36
Total	24	8	4	36

Fonte: Elaboração própria.

Quadro 4.10 - Matriz de confusão do *campus* Cambuci.

	Classificado como:		
	matriculado	**evadido**	**concluído**
matriculado	21	2	1
evadido	7	1	0
concluído	4	0	0

Fonte: Elaboração própria.

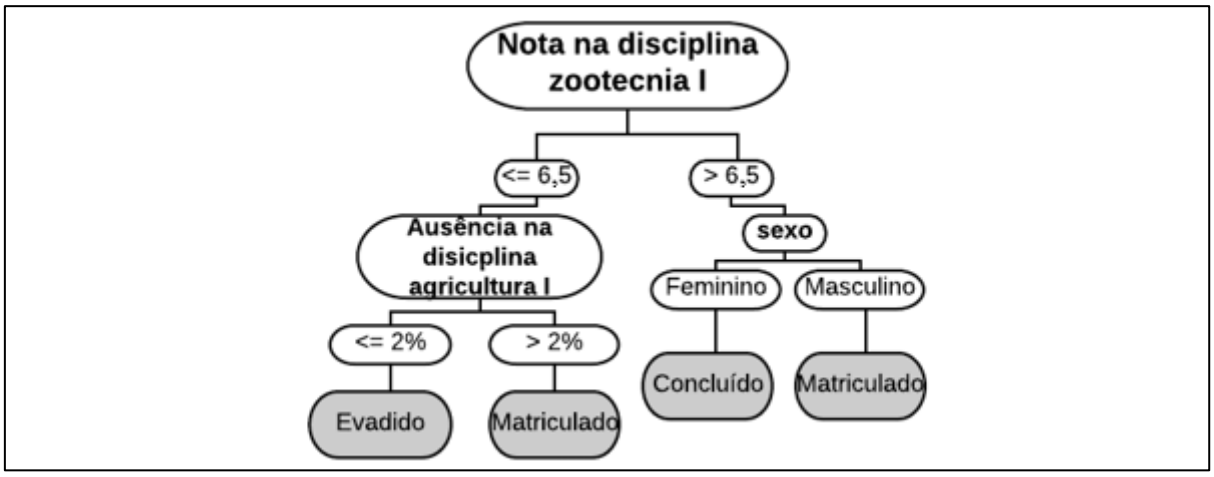

Figura 4.12 - Árvore de decisão do *campus* Cambuci. Fonte: Elaboração própria.

Campus Campos Centro

Modalidade: Concomitante

Quadro 4.11 - Quantitativo de alunos do *campus* Campos Centro na modalidade concomitante por curso e por situação de matrícula.

Curso	Matriculado	Evadido	Concluído	Total
Técnico em Automação Industrial	92	39	20	151
Técnico em Edificações	70	10	58	138
Técnico em Eletrotécnica	52	27	12	91
Técnico em Estradas	45	11	10	66
Técnico em Informática	70	64	14	148
Técnico em Mecânica	149	66	91	306
Técnico em Química	63	1	8	72
Técnico em Telecomunicações	63	75	27	165
Total	604	293	240	1137

Fonte: Elaboração própria.

Quadro 4.12 - Matriz de confusão do *campus* Campos Centro na modalidade concomitante.

	Classificado como:

	matriculado	evadido	concluído
matriculado	506	26	72
evadido	147	123	23
concluído	116	3	121

Fonte: Elaboração própria.

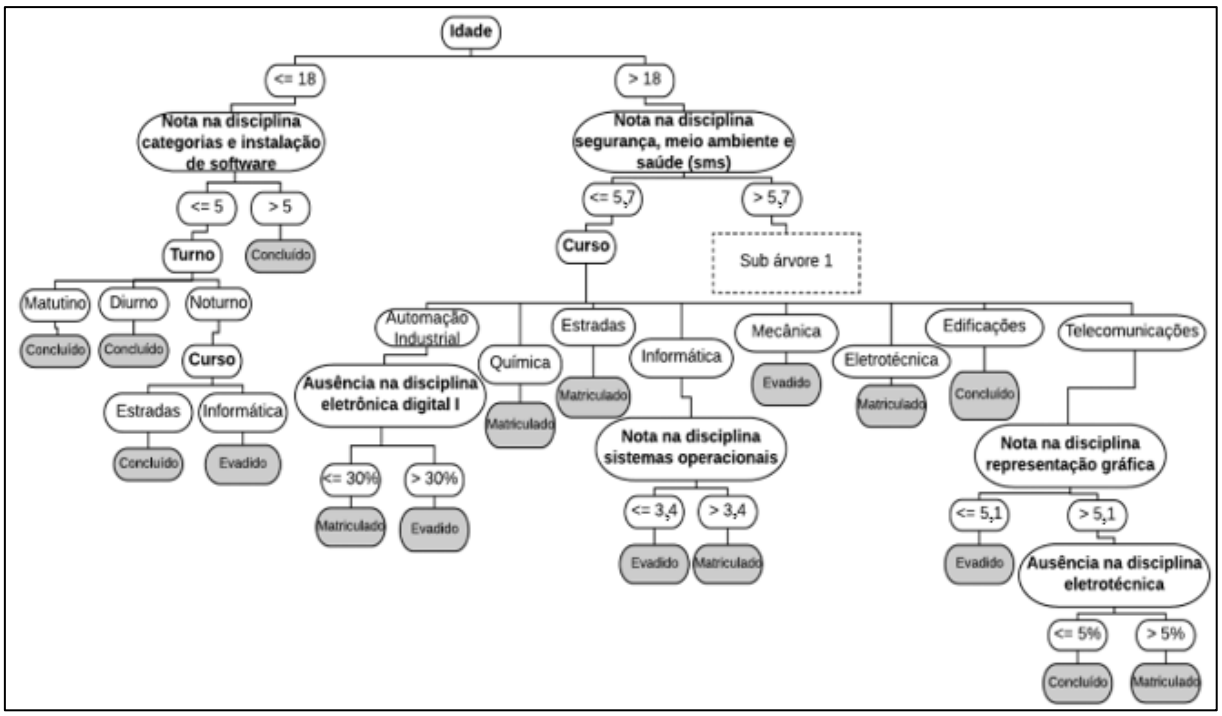

Figura 4.13 - Primeira parte da árvore de decisão do *campus* Campos Centro na modalidade concomitante.
Fonte: Elaboração própria.

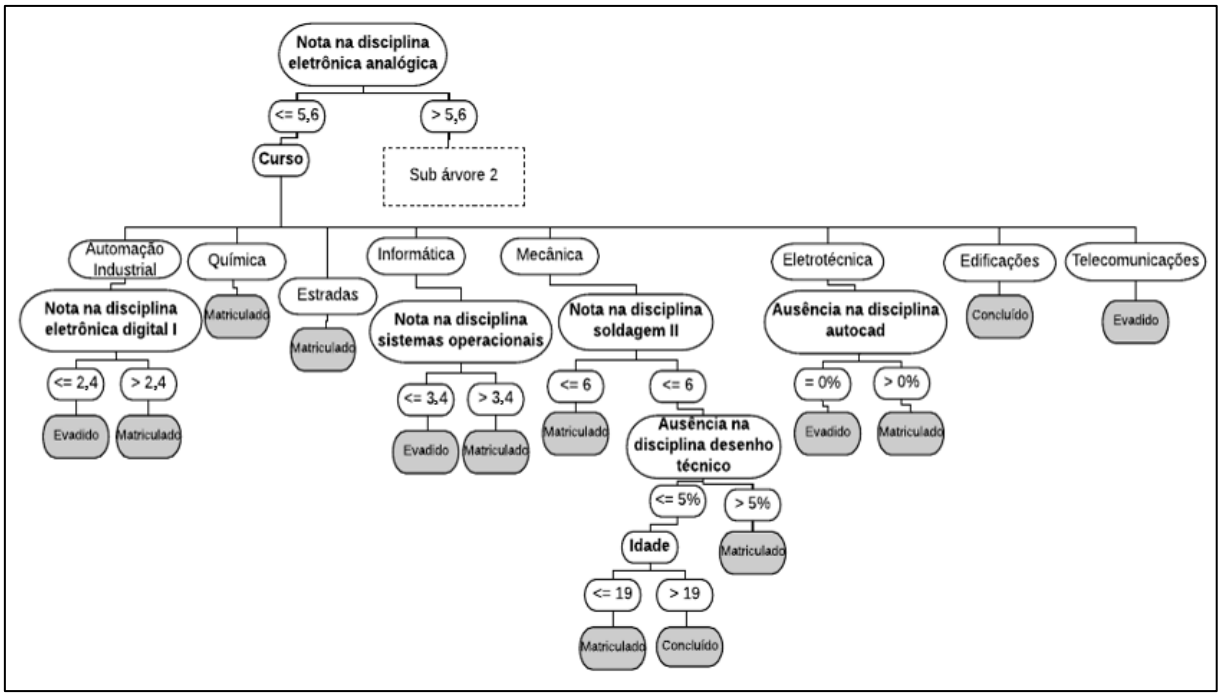

Figura 4.14 - Sub árvore 1 da árvore de decisão do *campus* Campos Centro na modalidade concomitante. Fonte: Elaboração própria.

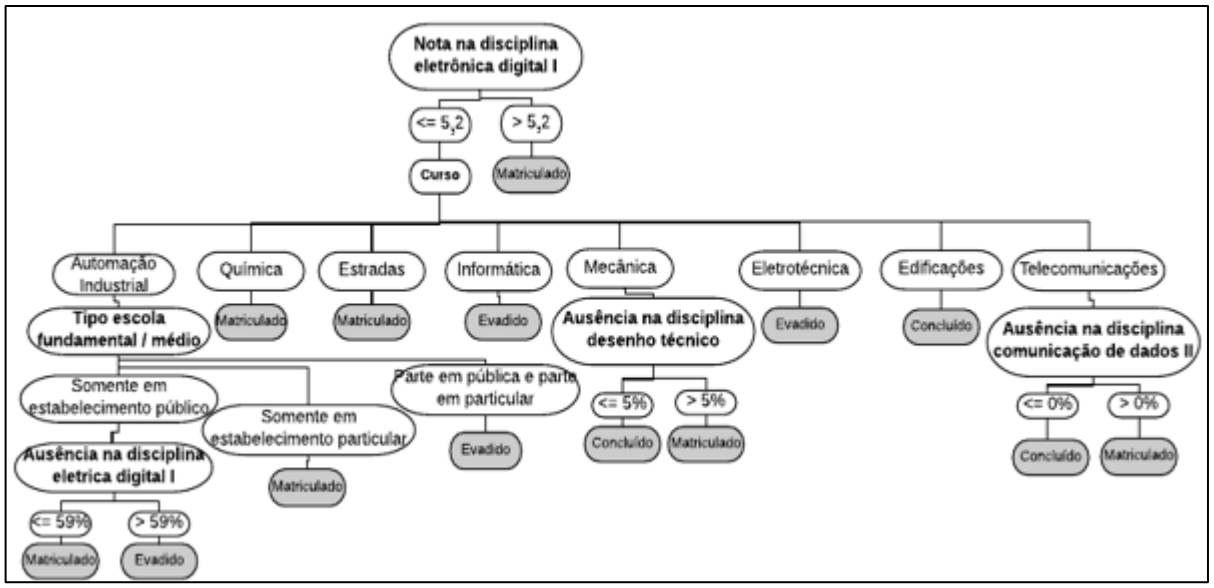

Figura 4.15 - Sub árvore 2 da árvore de decisão do *campus* Campos Centro na modalidade concomitante. Fonte: Elaboração própria.

Modalidade: Subsequente

Quadro 4.13 - Quantitativo de alunos do *campus* Campos Centro na modalidade subsequente por curso e por situação de matrícula.

Curso	Matriculado	Evadido	Concluído	Total
Técnico em Segurança do Trabalho	36	15	17	68
Total	36	15	17	68

Fonte: Elaboração própria.

Quadro 4.14 - Matriz de confusão do *campus* Campos Centro na modalidade subsequente.

	Classificado como:		
	matriculado	**evadido**	**concluído**
matriculado	32	4	0
evadido	15	0	0
concluído	17	0	0

Fonte: Elaboração própria.

Não foi gerada árvore pelo desfecho demonstrado na matriz de confusão, em que não houve classificação de evadido e concluído.

Campus Guarus

Modalidade: Subsequente

Quadro 4.15 - Quantitativo de alunos do *campus* Guarus por curso e por situação de matrícula.

Curso	**Matriculado**	**Evadido**	**Concluído**	**Total**
Técnico em Enfermagem	59	33	23	115
Técnico em Eletromecânica	49	23	27	99
Técnico em Farmácia	44	40	14	98
Técnico em Meio Ambiente	24	19	6	49
Total	176	115	70	361

Fonte: Elaboração própria.

Quadro 4.16 - Matriz de confusão do *campus* Guarus.

	Classificado como:		
	matriculado	**evadido**	**concluído**
matriculado	155	16	5
evadido	58	55	2
concluído	48	4	18

Fonte: Elaboração própria.

Figura 4.16 - Árvore de decisão do *campus* Guarus. Fonte: Elaboração própria.

Campus Itaperuna

Modalidade: Concomitante

Quadro 4.17 - Quantitativo de alunos do *campus* Itaperuna por curso e por situação de matrícula.

Curso	Matriculado	Evadido	Concluído	Total
Técnico em Eletromecânica	1	48	14	63
Técnico em Eletrotécnica	41	33	31	105
Técnico em Informática	9	36	0	45
Técnico em Mecânica	44	20	5	69
Técnico em Química	41	28	22	91
Total	136	165	72	373

Fonte: Elaboração própria.

Quadro 4.18 - Matriz de confusão do *campus* Itaperuna.

	Classificado como:		
	matriculado	**evadido**	**concluído**
matriculado	88	33	15
evadido	19	134	12
concluído	13	4	55

Fonte: Elaboração própria.

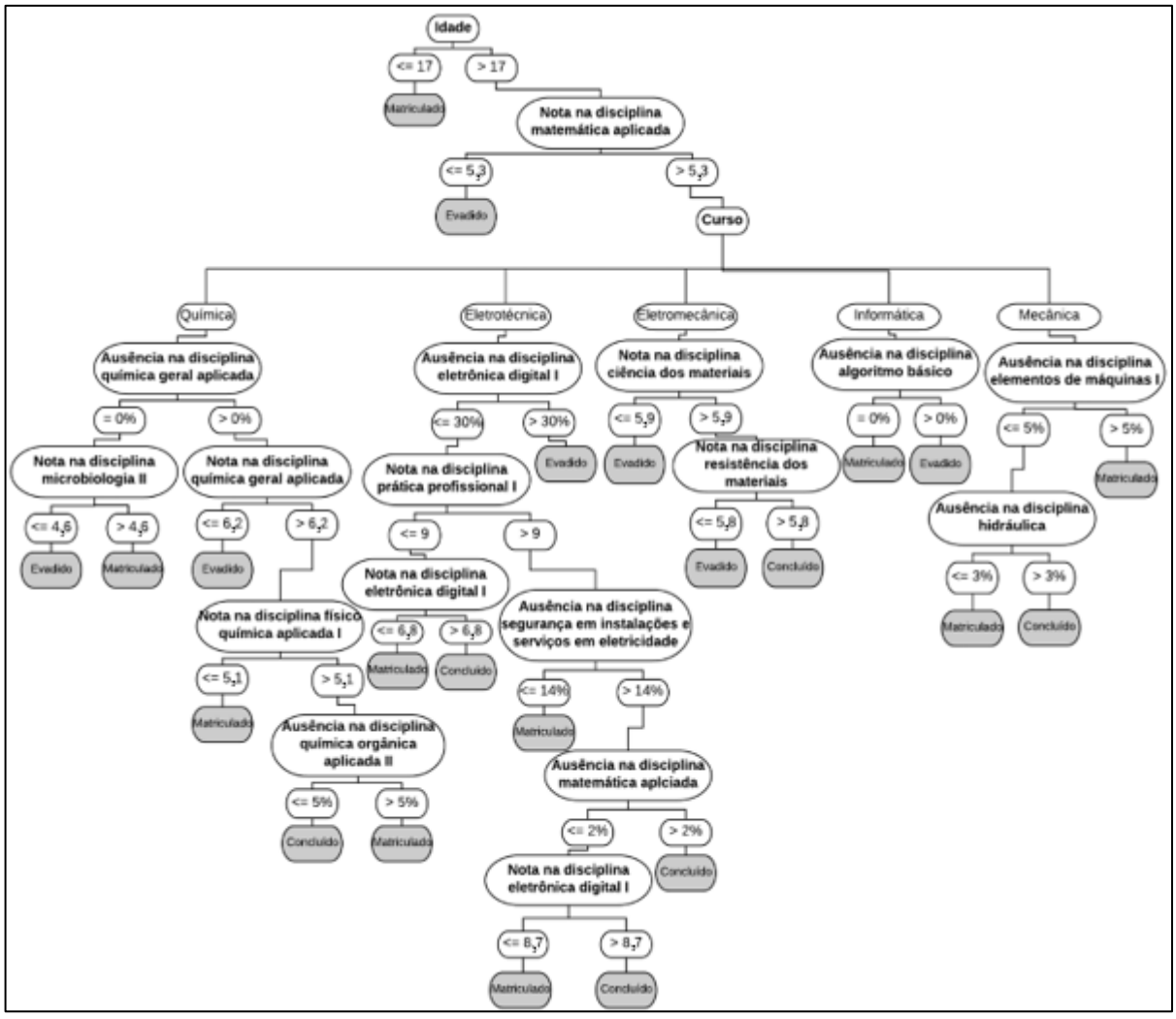

Figura 4.17 - Árvore de decisão do *campus* Itaperuna. Fonte: Elaboração própria.

Campus Macaé

Modalidade: Subsequente

Quadro 4.19 - Quantitativo de alunos do *campus* Macaé por curso e por situação de matrícula.

Curso	**Matriculado**	**Evadido**	**Concluído**	**Total**

Técnico em Automação Industrial	35	8	19	62
Técnico em Eletromecânica	36	0	19	55
Técnico em Eletrônica	2	1	14	17
Técnico em Informática	16	3	8	27
Técnico em Meio Ambiente	31	1	0	32
Técnico em Segurança do Trabalho	34	10	22	66
Total	154	23	82	259

Fonte: Elaboração própria.

Quadro 4.20 - Matriz de confusão do *campus* Macaé.

	Classificado como:		
	matriculado	**evadido**	**concluído**
matriculado	149	0	5
evadido	21	0	2
concluído	72	0	10

Fonte: Elaboração própria.

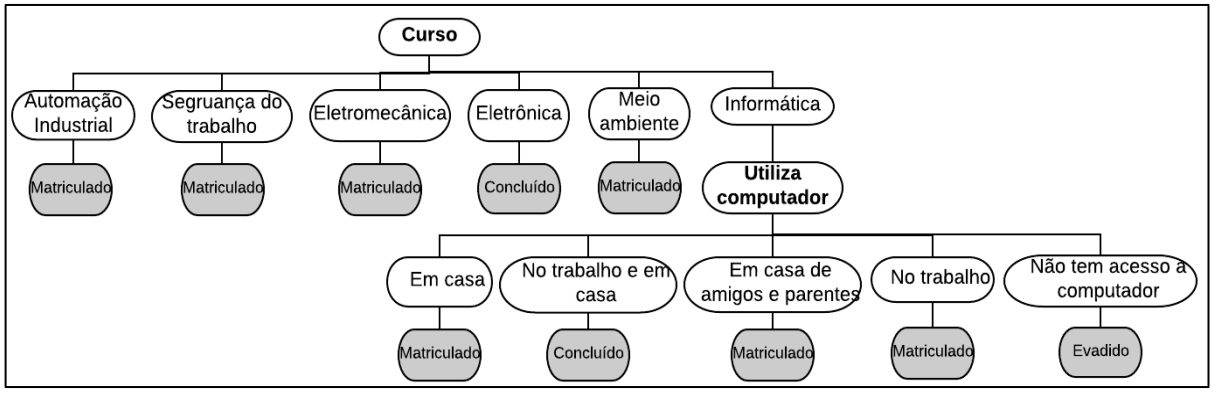

Figura 4.18 - Árvore de decisão do *campus* Macaé. Fonte: Elaboração própria.

Campus Santo Antônio de Pádua

Modalidade: Concomitante

Quadro 4.21 - Quantitativo de alunos do *campus* Santo Antônio de Pádua por curso e por situação de matrícula.

Curso	Matriculado	Evadido	Concluído	Total
Técnico em Mecânica	35	28	-	63
Total	35	28	-	63

Fonte: Elaboração própria.

Quadro 4.22 - Matriz de confusão do *campus* Santo Antônio de Pádua.

	Classificado como:		
	matriculado	**evadido**	**concluído**
matriculado	31	4	-

evadido	10	18	-
concluído	-	-	-

Fonte: Elaboração própria.

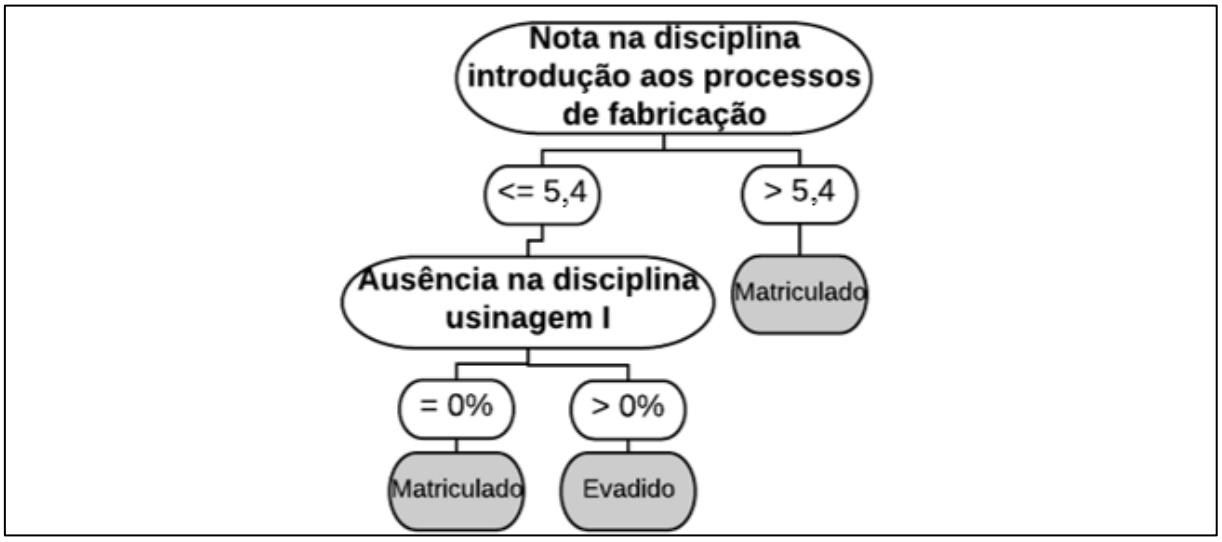

Figura 4.19 - Árvore de decisão do *campus* Santo Antônio de Pádua. Fonte: Elaboração própria.

Campus Quissamã

Modalidade: Concomitante

Quadro 4.23 - Quantitativo de alunos do *campus* Quissamã na modalidade concomitante por curso e por situação de matrícula.

Curso	Matriculado	Evadido	Concluído	Total
Técnico em Eletromecânica	25	43	9	77
Total	25	43	9	77

Fonte: Elaboração própria.

Quadro 4.24 - Matriz de confusão do *campus* Quissamã na modalidade concomitante.

	Classificado como:		
	matriculado	**evadido**	**concluído**
matriculado	17	6	2
evadido	8	33	2
concluído	4	4	1

Fonte: Elaboração própria.

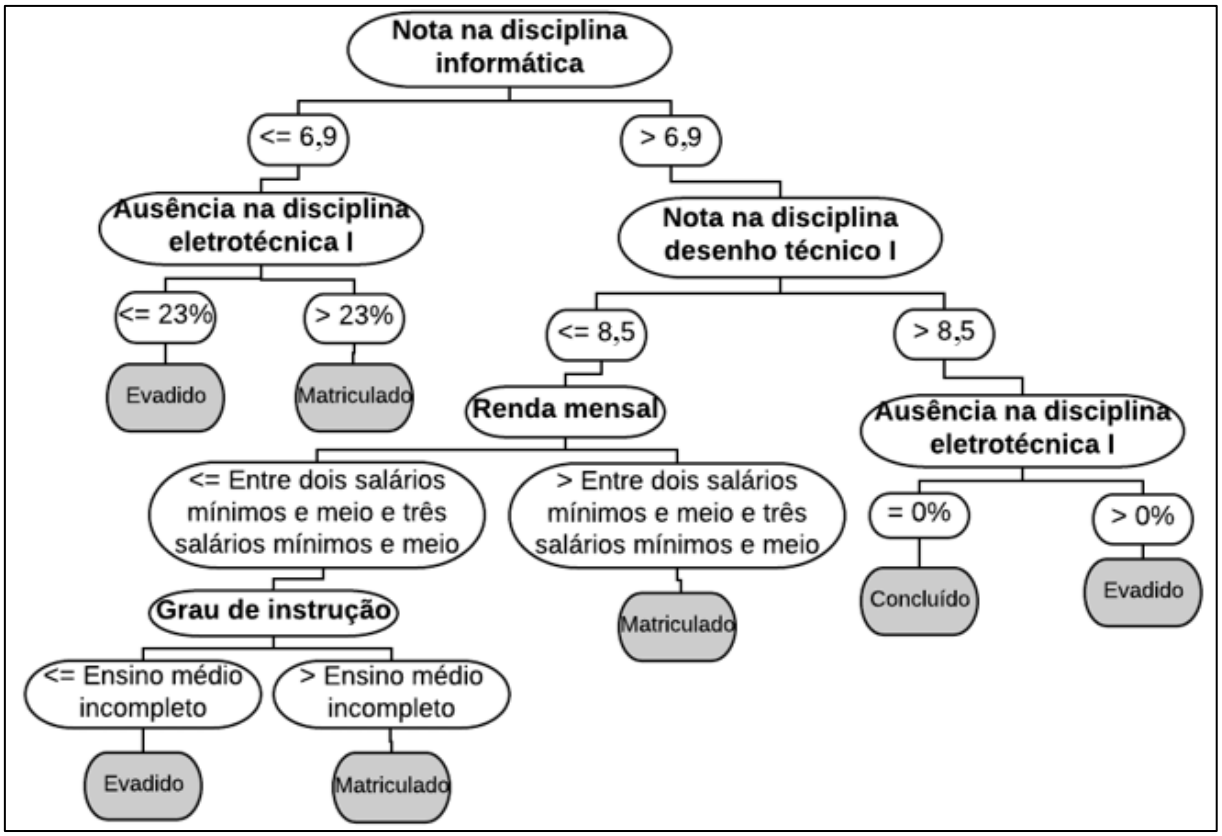

Figura 4.20 - Árvore de decisão do *campus* Quissamã na modalidade concomitante. Fonte: Elaboração própria.

Modalidade: Subsequente

Quadro 4.25 - Quantitativo de alunos do *campus* Quissamã na modalidade subsequente por curso e por situação de matrícula.

Curso	Matriculado	Evadido	Concluído	Total
Técnico em Segurança do Trabalho	9	11	14	34
Total	9	11	14	34

Fonte: Elaboração própria.

Quadro 4.26 - Matriz de confusão do *campus* Quissamã na modalidade subsequente.

	Classificado como:		
	matriculado	**evadido**	**concluído**
matriculado	5	1	3
evadido	2	7	2
concluído	4	2	8

Fonte: Elaboração própria.

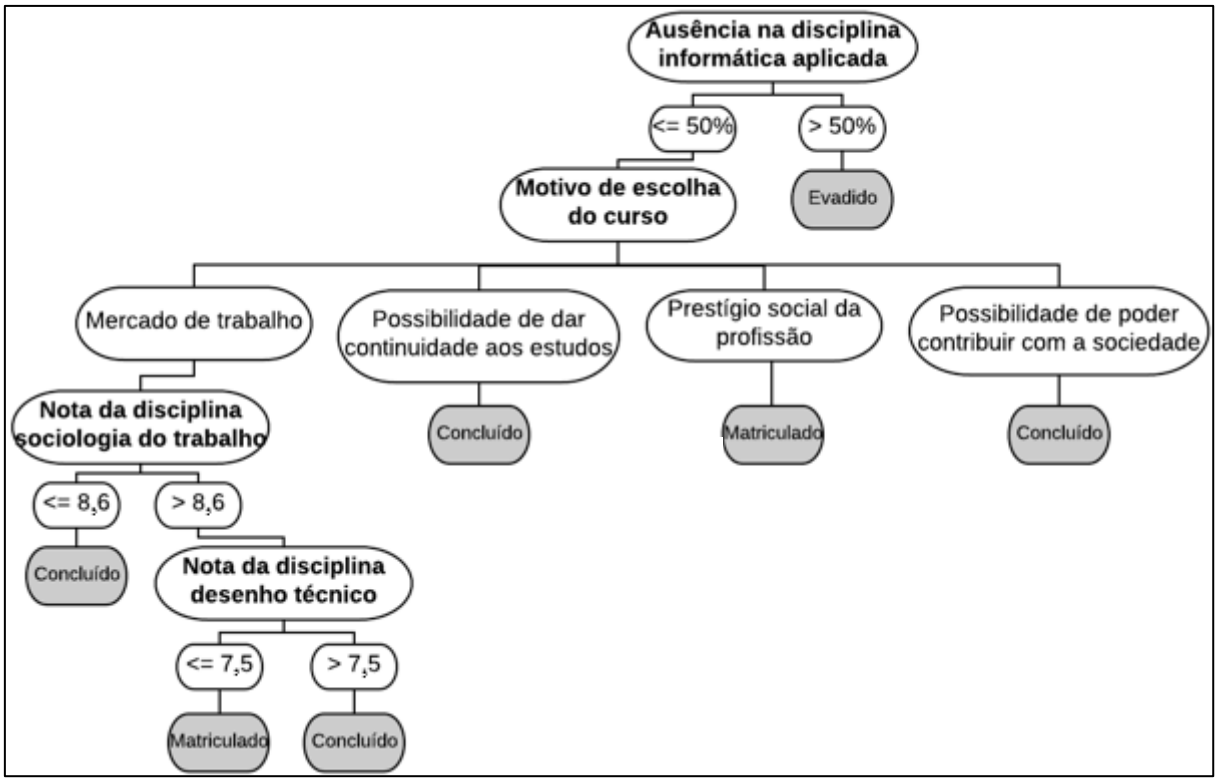

Figura 4.21 - Árvore de decisão do *campus* Quissamã na modalidade subsequente. Fonte: Elaboração própria.

Campus São João da Barra

Modalidade: Concomitante

Quadro 4.27 - Quantitativo de alunos do *campus* São João da Barra por curso e por situação de matrícula.

Curso	Matriculado	Evadido	Concluído	Total
Técnico em Eletromecânica	22	3	2	27
Total	22	3	2	27

Fonte: Elaboração própria.

Quadro 4.28 - Matriz de confusão do *campus* São João da Barra.

	Classificado como:		
	matriculado	**evadido**	**concluído**
matriculado	21	0	1
evadido	3	0	0
concluído	2	0	0

Fonte: Elaboração própria.

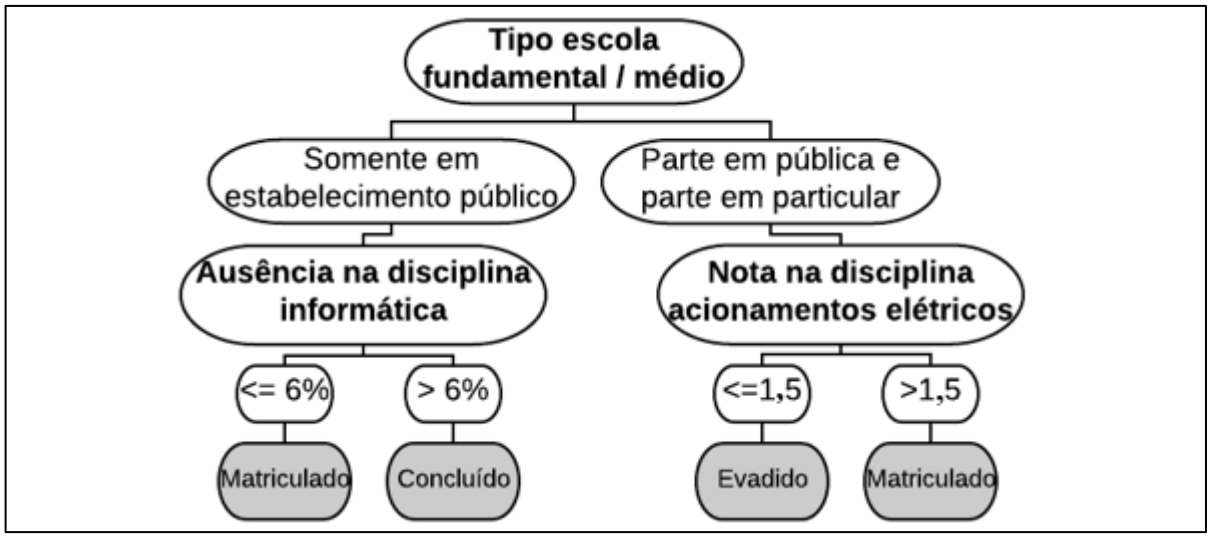

Figura 4.22 - Árvore de decisão do *campus* São João da Barra. Fonte: Elaboração própria.

Apêndice C

TERMO DE COMPROMISSO

TERMO DE COMPROMISSO, que fazem entre si, Renata Gomes Cordeiro, brasileira, portador do documento de identidade número 22.288.906-5, expedido pelo DETRAN, residente e domiciliado nesta cidade, estudante do Programa de Pós-graduação em Sistemas Aplicados à Engenharia e Gestão, doravante denominada de mestranda, e Instituto Federal Fluminense Campus - Reitoria (IFF), sediado nesta cidade, representado pelo Pró Reitor de Ensino Carlos Artur de Carvalho Arêas e pelo Diretor de Gestão Acadêmica e Políticas de Acesso Marcelo Peçanha Sarmento, assumem de comum acordo os seguintes compromissos:

Do objeto

1. Fica instituído que o projeto de pesquisa, com orientação do Professor Henrique Rego Monteiro da Hora, intitulado "Identificação do comportamento dos estudantes evadidos de cursos técnicos utilizando técnicas de mineração de dados", será desenvolvido em parceria entre a mestranda e o IFF.

Do projeto

2. O projeto de pesquisa terá como objetivo estudar o aplicação da mineração de dados para análise do comportamento dos alunos evadidos em cursos de nível técnico.

Das disposições gerais

3. São obrigações da mestranda:
    a. manter sigilosas as informações individuais dos alunos, bem como nomes de pessoas físicas e jurídicas envolvidas no processo;
4. São direitos da mestranda:
    a. divulgar os resultados da pesquisa, apresentando somente as informações macro, sem ferir o item 3.a. do presente termo;
    b. interromper os trabalhos de pesquisa, desde que devidamente justificados.
5. São direitos do IFF:
    a. utilizar como melhor lhe convier os dados levantados durante a pesquisa;
    b. interromper o fornecimento de informações à mestranda , se comprovada a falha no sigilo ditado no item 3.a.

Das disposições gerais

6. Toda responsabilidade sobre as divulgações ficarão a cargo de seus autores.
7. O vazamento de informações citadas neste termo será de inteira responsabilidade da mestranda, desde que comprovada culpa ou dolo.

Assim sendo, assinam o presente TERMO DE COMPROMISSO, a mestranda e o IFF .

Campos dos Goytacazes, 24 de outubro de 2017.

Renata Gomes Cordeiro

RENATA GOMES CORDEIRO – Mestranda

CARLOS ARTUR DE CARVALHO ARÊAS - Pró Reitor de Ensino

MARCELO PEÇANHA SARMENTO - Diretor de Gestão Acadêmica e Políticas de Acesso

Apêndice D

Assunto	**Re: Fwd: Dissertação de mestrado - autorização de acesso a dados**
De	Renata Gomes Cordeiro <renata.cordeiro@iff.edu.br>
Para	<msarmento@iff.edu.br>
Data	2017-05-30 07:14

Bom dia, Marcelo.

Obrigada pela autorização.

Att.,

Renata Gomes Cordeiro

Analista de Tecnologia da Informação
Instituto Federal Fluminense - Reitoria

Em 2017-05-29 23:11, Marcelo Peçanha Sarmento escreveu:

> Renata,
>
> conversei com Christiane e com Henrique sobre a questão. Consideramos muito relevante a pesquisa e nesse sentido o acesso e utilização dos dados estão autorizados. Solicitamos que os dados pessoais não sejam divulgados ou individualizados.
>
> Ótimo trabalho!
>
> Qualquer coisa, estou à disposição!
>
> Atenciosamente,
>
> ---
>
> ***Marcelo Peçanha Sarmento***
> Diretor de Gestão Acadêmica e Políticas de Acesso
> INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA FLUMINENSE
> Pró-Reitoria de Ensino
> Professor do IFFluminense
> (22) 988437177/ (22) 999735527
>
> Em 2017-05-05 19:13, Christiane Menezes Rodrigues escreveu:
>
> > Marcelo,
> >
> > O que vc acha da disponibilização destes dados? Fico temerosa quanto ao fornecimento de informações pessoais dos estudantes.
> > O que vc pensa?
> > Bjs
> > ---------- Mensagem encaminhada ----------
> > De: "Renata Gomes Cordeiro" <renata.cordeiro@iff.edu.br>
> > Data: 5 de mai de 2017 11:29 AM
> > Assunto: Dissertação de mestrado - autorização de acesso a dados
> > Para: "Christiane Menezes Rodrigues" <cmrodrigues@iff.edu.br>
> > Cc: "Henrique Rego Monteiro da Hora" <henrique.dahora@iff.edu.br>
> >
> > > Bom dia, Christiane.
> > >
> > > Conforme conversamos, envio este e-mail para solicitar autorização para acesso à dados do sistemas de inscrições e do QAcadêmico.
> > >
> > > Sou servidora do IFF e trabalho na DGTI. Como aluna do curso de mestrado SAEG e orientada pelo professor Henrique da Hora, pretendo desenvolver minha dissertação objetivando aplicar métodos de mineração de dados em dados educacionais a fim de obter o perfil dos alunos evadidos no cursos técnicos.
> > >
> > > Anexo ao e-mail está o projeto. Para o cumprimento da pesquisa vou precisar dos seguintes dados:
> > >
> > > Do sistemas de inscrições:
> > > - questionários socioeconômico dos inscritos no processos seletivos para cursos de nivel técnico desde 2014.
> > >
> > > Do QAcadêmico:
> > > - frequência
> > > - notas
> > > - dados pessoais
> > >
> > > Aguardo seu retorno.
> > >
> > > Att.,
> > >
> > > --
> > > Renata Gomes Cordeiro
> > >
> > > Analista de Tecnologia da Informação
> > > Instituto Federal Fluminense - Reitoria

5. CONSIDERAÇÕES FINAIS

Considerando o grande volume de dados armazenados por uma instituição de ensino torna-se vantajosa e necessária a utilização de técnicas que propiciem a estratificação de informações relevantes, a partir das quais podem ser traçadas políticas com o intuito de trazer melhorias para o ensino.

O trabalho iniciou com uma pesquisa bibliométrica sobre estudos que abordam evasão no ensino técnico utilizando técnicas de mineração de dados, em que foi identificado que está é uma área ainda a ser explorada. Para a definição da metodologia de extração de características de alunos evadidos no ensino técnico utilizando mineração de dados, foi analisado um conjunto de trabalhos. A análise sistemática demonstrou que em geral as pesquisas possuem etapas em comum assim como os dados utilizados. A aplicação da metodologia foi realizada através de um estudo de caso, onde foram identificados os comportamentos de alunos evadidos em duas modalidades do ensino técnico.

A partir da pesquisa bibliométrica é possível notar que as pesquisas envolvendo mineração de dados na área educacional com foco em evasão são recentes, sendo notável que ainda é uma área pertinente e com muito a explorar.

Através da análise sistemática, foi verificada a viabilidade de utilização de técnicas de mineração de dados como forma de identificar os alunos propensos a evadir. A partir daí foi possível analisar que os trabalhos na área de mineração de dados educacionais, em geral, estão focados na proposta de metodologias que aprimorem os métodos já existentes e consequentemente o resultado final. Neste trabalho buscou-se a análise dos resultados com apresentação de comportamentos de alunos evadidos.

Na realização do estudo de caso no Instituto Federal Fluminense, foram analisados os alunos de cursos técnicos na modalidade concomitante e subsequente. A principal conclusão é que, em geral, os comportamentos dos alunos evadidos são descritos pelo desempenho ou frequência nas disciplinas dos primeiros módulos dos cursos.

Acredita-se que os resultados obtidos sejam do interesse de docentes e gestores. As características identificadas nos alunos evadidos podem sugerir pontos de atenção tanto aos gestores que elaboram políticas de combate a evasão, quanto aos docentes que lidam diariamente com os estudantes na sala de aula.

De modo geral, a pesquisa explorou a área de mineração de dados na educação através da análise da produção de trabalhos e de metodologias utilizadas. E contribuiu através da realização do estudo de caso e análise dos resultados com foco não só nos dados quantitativos, mas também analisando de que forma os atributos utilizados contribuíram para os resultados obtidos.

Como trabalhos futuros propõe-se aplicar a seleção de atributos previamente a fase de mineração de dados, com o intuito de verificar se houve melhora nas taxas de acerto obtidas. Além disso, com o mesmo intuito, poderá ser investigada a utilização de outros métodos de classificação. Futuramente, sugere-se ainda analisar o comportamento dos alunos concluintes e aplicar a mesma metodologia em bases de dados contendo apenas dados socioeconômicos com o intuito de observar como esses dados descrevem o comportamento do aluno evadido.

REFERÊNCIAS BIBLIOGRÁFICAS

Cunha, J. A., Moura, E. & Analide, C. (2016). Data mining in academic databases to detect behaviors of students related to school dropout and disapproval. *Advances in Intelligent Systems and Computing*, *445*, 189–198. https://doi.org/10.1007/978-3-319-31307-8_19

Hoed, R. M. (2016). *Análise da evasão em cursos superiores: o caso da evasão em cursos superiores da área de Computação*. Universidade de Brasília, Brasília, DF. Obtido de http://repositorio.unb.br/bitstream/10482/22575/1/2016_RaphaelMagalh%C3%A3esHoed.pdf

Machado, R. D., Benitez, E. O., Corleta, J. N. & Augusto, G. (2015). Estudo Bibliométrico em mineração de dados e evasão escolar. Apresentado na XI CONGRESSO NACIONAL DE EXCELÊNCIA EM GESTÃO, Rio de Janeiro, RJ.

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa, F. & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, *33*(1), 107–124. https://doi.org/10.1111/exsy.12135

Mehta, A. A. & Buch, N. J. (2016). Depth and breadth of educational data mining: Researchers' point of view. Em *Proceedings of the 10th International Conference on Intelligent Systems and Control, ISCO 2016*. Coimbatore, India.