



**UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
UNIVERSIDADE DO ESTADO DO RIO GRANDE DO NORTE
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA
COMPUTAÇÃO**



LEONARDO TORRES MARQUES

**MATEO: UMA ABORDAGEM DE DESCOBERTA DE
CONHECIMENTO PARA DESVENDAR AS CAUSAS
DA EVASÃO ESCOLAR**

Mossoró-RN

2020

LEONARDO TORRES MARQUES

**MATEO: UMA ABORDAGEM DE DESCOBERTA DE
CONHECIMENTO PARA DESVENDAR AS CAUSAS
DA EVASÃO ESCOLAR**

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação - associação ampla entre a Universidade do Estado do Rio Grande do Norte e a Universidade Federal Rural do Semi-Árido, para a obtenção do título de Mestre em Ciência da Computação.

Linha de Pesquisa: Projeto de Sistemas e Circuitos

Orientadora: Prof^ª Dra. Angélica Félix de Castro,

Coorientador: Prof^º Dr. Paulo Gabriel Gadelha Queiroz

Mossoró-RN

2020

© Todos os direitos estão reservados a Universidade Federal Rural do Semi-Árido. O conteúdo desta obra é de inteira responsabilidade do (a) autor (a), sendo o mesmo, passível de sanções administrativas ou penais, caso sejam infringidas as leis que regulamentam a Propriedade Intelectual, respectivamente, Patentes: Lei nº 9.279/1996 e Direitos Autorais: Lei nº 9.610/1998. O conteúdo desta obra tomar-se-á de domínio público após a data de defesa e homologação da sua respectiva ata. A mesma poderá servir de base literária para novas pesquisas, desde que a obra e seu (a) respectivo (a) autor (a) sejam devidamente citados e mencionados os seus créditos bibliográficos.

M357m Marques, Leonardo torres marques.
MATEO: UMA ABORDAGEM DE DESCOBERTA DE
CONHECIMENTO PARA DESVENDAR AS CAUSAS DA EVASÃO
ESCOLAR / Leonardo torres marques Marques. - 2020.
71 f. : il.

Orientadora: Angélica Félix de Castro.
Coorientador: Paulo Gabriel Gadelha Queiroz.
Dissertação (Mestrado) - Universidade Federal
Rural do Semi-árido, Programa de Pós-graduação em
Ciência da Computação, 2020.

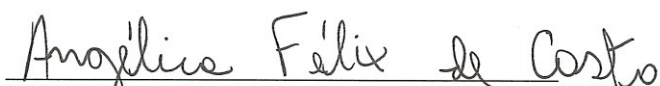
1. Evasão Escolar. 2. Abordagem de Descoberta
de Conhecimento. 3. Aprendizado de Máquina. 4.
Modelos de Classificação. I. de Castro, Angélica
Félix, orient. II. Queiroz, Paulo Gabriel
Gadelha, co-orient. III. Título.

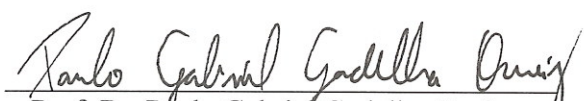
LEONARDO TORRES MARQUES

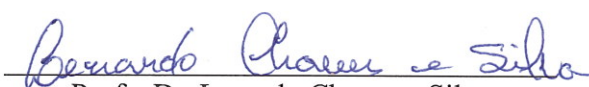
MATEO: UMA ABORDAGEM DE DESCOBERTA DE CONHECIMENTO PARA
DESVENDAR AS CAUSAS DA EVASÃO ESCOLAR

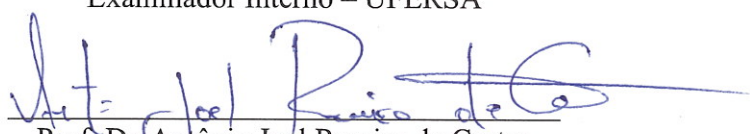
Dissertação apresentada ao Programa de Pós-Graduação
em Ciência da Computação para a obtenção do título de
Mestre em Ciência da Computação.

APROVADA EM: 14 / 02 / 2020


Prof. Dra. Angélica Félix de Castro
Orientadora e Presidente da Banca


Prof. Dr. Paulo Gabriel Gadelha Queiroz
Co-Orientador e Examinador Interno – UFERSA


Profa. Dr. Lenardo Chaves e Silva
Examinador Interno – UFERSA


Prof. Dr. Antônio Joel Ramiro de Castro
Examinador Externo – UFC

Dedico este trabalho especialmente a Deus, aos meus pais, Antônio Lucildo Marques Gonçalves e Hilda Gonçalves Torres Marques, e aos meus irmãos, Bruno Torres Marques e Raísa Carvalho Torres Marques.

Agradecimentos

Gratidão é a melhor definição pra expressar o meu sentimento quando olho pra trás e vejo que não estive sozinho durante essa trajetória. Sou muito grato à todos que doaram um pouco de si, para contribuir com minha formação acadêmica e como cidadão. Todos fizeram parte dessa conquista.

Antes de tudo, agradeço a Deus por ter me concedido a perseverança, fé, sabedoria e serenidade necessária para enfrentar as dificuldades encontradas ao longo dessa jornada, para aproveitar as oportunidades a mim proporcionadas.

À minha família, que sempre acreditou em mim e nunca me deixou desistir dos meus sonhos. Aos meus pais, Antônio Lucildo e Hilda Torres, por todo amor, carinho e dedicação, e aos meus irmãos, Bruno Torres e Kamylla Torres, que me apoiaram desde o princípio. E não poderia esquecer dos meus inesquecíveis avós *in memória*, sinto a presença de vocês a todo instante.

À minha amiga e orientadora Angélica Félix e meu amigo e coorientador Gabriel Gadelha, por toda a atenção e orientação concedida. Obrigado por todos os ensinamentos, compreensão e companheirismo que vocês me proporcionaram. Aprendi muito com vocês.

Aos meus amigos de sempre que se fizeram presentes em minha vida. Em especial aos meus colegas de PPgCC e do LES, que compartilharam comigo momentos de aprendizagens e descontração.

À CAPES pelo apoio financeiro para o desenvolvimento desta pesquisa. Enfim, agradeço a todos que, mesmo sem saber, colaboraram para o meu crescimento acadêmico.

“O insucesso é apenas uma oportunidade para recomeçar de novo com mais inteligência.”

Henry Ford

Resumo

A evasão escolar é um dos problemas mais intrigantes e cruciais da Educação. Esse problema permeia os vários níveis e modalidades de educação e gera perdas sociais, econômicas, políticas, acadêmicas e financeiras para todos os envolvidos no processo educacional. Portanto, é essencial o desenvolvimento de métodos eficientes para prever o risco de abandono dos estudantes, permitindo que as instituições adotem ações proativas para minimizar a situação. Assim, objetiva-se com este trabalho, apresentar uma abordagem de descoberta de conhecimento, em banco de dados, com o objetivo de identificar grupos de estudantes em risco de abandono nos cursos presenciais de ensino superior. A abordagem foi validada, utilizando dados de ex-alunos do curso superior de Ciência da Computação da Universidade Federal Rural do Semi-Árido, Campus Mossoró/RN. Neste trabalho foram utilizados modelos de classificação, uma das técnicas de Inteligência Artificial, que possibilita o aprendizado contínuo. Na validação da abordagem foram empregados cinco modelos de aprendizado de máquina, sendo que dois modelos obtiveram melhores índices de acurácia (*SVM* e *Adaboost*). Em relação aos resultados, observou-se que: alunos que estudaram em escolas privadas no ensino médio, tem uma maior probabilidade de concluir o curso; assim como, os alunos naturais do estado do RN; os alunos naturais de outras cidades que mudaram pra morar em Mossoró/RN; os alunos que moraram na vila acadêmica; aqueles que tem a renda familiar melhor; alunos que ganham bolsa no decorrer do curso; alunos que tem acompanhamento institucional; alunos que concluem o curso acreditam que os profissionais de Tecnologia da Informação são remunerados adequadamente e alunos que tem pai com nível superior.

Palavras-chave: Evasão Escolar; Abordagem de Descoberta de Conhecimento; Aprendizado de Máquina; Modelos de Classificação.

Abstract

Dropout is one of the most intriguing and crucial problems of education. This problem permeates the various levels and modalities of education and generates social, economic, political, academic and financial losses for all involved in the educational process. Therefore, the development of efficient methods to predict the risk of student dropout is essential, allowing institutions to take proactive actions to minimize the situation. Thus, the objective of this work is to present a knowledge discovery approach, in a database, aiming to identify groups of students at risk of dropping out in face-to-face higher education courses. The approach was validated using data from alumni of the Computer Science college at the Federal Rural University of Semi-Arid, Mossoró/RN. In this work, classification models were used, one of the Artificial Intelligence techniques, which enables continuous learning. In the validation of the approach, five machine learning models were used, and two models obtained better accuracy indexes (SVM and Adaboost). Regarding the results, it was observed that: students who studied in private schools in high school, are more likely to complete the course; as well as the natural students of the state of RN; natural students from other cities who moved to live in Mossoró/RN; the students who lived in the academic village; those with the best monthly student income; students who earn scholarships during the course; students with institutional support; Students who complete the course believe that Information Technology professionals are adequately remunerated and students who have a parent with a college degree.

Keywords: School Dropout; Knowledge Discovery Approach; Machine Learning; Classification Models.

Lista de ilustrações

Figura 1 – Resultado das buscas por base	32
Figura 2 – Quantitativo de trabalhos selecionados por ano	34
Figura 3 – Fatores Indutores para à Evasão Escolar.	36
Figura 4 – Abordagem de descoberta de conhecimento para identificar causas da evasão escolar no ensino superior.	42
Figura 5 – Correção de dados da variável “Situação_Ex-Aluno”.	49
Figura 6 – Agregação de dados da variável “Renda_Mensal”.	50
Figura 7 – Normalização de dados da variável “Cidade_Morava”.	50

Lista de tabelas

Tabela 1 – Palavras-chave e Sinônimos	28
Tabela 2 – Lista de fontes	29
Tabela 3 – Ficha de extração de dados	31
Tabela 4 – Resultados dos filtros	33
Tabela 5 – Resultados dos trabalhos encontrados em jornais.	33
Tabela 6 – Resultados dos trabalhos encontrados em conferências.	34
Tabela 7 – Relação dos trabalhos selecionados	35
Tabela 8 – Abordagens de conhecimento utilizadas nos trabalhos	37
Tabela 9 – Ferramentas utilizadas na MD	38
Tabela 10 – Comparativo das Ferramentas	39
Tabela 11 – Técnicas e algoritmos utilizados na MD	39
Tabela 12 – Fatores e variáveis exploradas e respectivas descrições.	48
Tabela 13 – Aplicação da Lei de Entropia para verificação do ganho de informação das variáveis.	51
Tabela 14 – Características obtidas com os modelos <i>Naive Bayes</i>	52
Tabela 15 – Validação dos modelos <i>Naive Bayes</i>	53
Tabela 16 – Características obtidas com o modelo <i>J48</i>	54
Tabela 17 – Validação dos modelos <i>J48</i>	54
Tabela 18 – Características obtidas com o modelo <i>KNN</i>	55
Tabela 19 – Validação dos modelos <i>KNN</i>	55
Tabela 20 – Características obtidas com o modelo <i>SVM</i>	56
Tabela 21 – Validação dos modelos <i>SVM</i>	56
Tabela 22 – Características obtidas com o modelo <i>AdaBoost</i>	57
Tabela 23 – Validação dos modelos <i>SVM</i>	57

Lista de abreviaturas e siglas

BD	Banco de Dados
CC	Ciência da Computação
CE	Critério de Exclusão
CI	Critério de Inclusão
CQ	Critério de Qualidade
HLM	<i>Hierarchical Linear Modeling</i>
IA	Inteligência Artificial
IEEE	<i>Institute of Electrical and Electronics Engineers</i>
IES	Instituições de Ensino Superior
KDD	<i>Knowledge Discovery in Databases</i>
KNN	<i>K-Nearest Neighbors</i>
MDE	Mineração de Dados Educacionais
MD	Mineração de Dados
MLP	<i>Multilayer Perceptron</i>
QC	Questões Conceituais
QGP	Questão Geral de Pesquisa
QP	Questões Práticas
QT	Questões Tecnológicas
RENOTE	Revista Novas Tecnologias na Educação
RN	Rio Grande do Norte
RSL	Revisão Sistemática de Literatura
SAS	<i>Statistical Analysis System</i>
SI	Sistemas de Informação

SPSS	<i>Statistical Package for the Social Science</i>
SQL	<i>Structured Query Language</i>
SVM	<i>Support Vector Machine</i>
TI	Tecnologia da Informação
UFERSA	Universidade Federal Rural do Semi-Árido
Weka	<i>Waikato environment for Knowledge Analysis</i>

Sumário

1	INTRODUÇÃO	15
1.1	Motivação	16
1.2	OBJETIVOS	17
1.2.1	Objetivo Geral	17
1.2.2	Objetivos Específicos	17
1.3	QUESTÕES DE PESQUISA	18
1.4	Contribuições do Trabalho	19
1.5	ORGANIZAÇÃO DO DOCUMENTO	19
2	FUNDAMENTAÇÃO TEÓRICA	21
2.1	Descoberta de Conhecimento em Bases Dados	21
2.1.1	Dificuldades e Desafios da descoberta de conhecimento em bases de dados	22
2.2	Mineração de Dados Educacionais	23
2.3	Determinantes da Evasão no Ensino Superior	24
2.3.1	Fatores Inerentes às Características Individuais do Aluno	25
2.3.2	Fatores Internos às Instituições	25
2.3.3	Fatores Externos às Instituições	25
2.4	Considerações Finais	26
3	REVISÃO SISTEMÁTICA DA LITERATURA	27
3.1	Procedimentos e Métodos	27
3.1.1	Objetivos	27
3.1.2	Questões de Pesquisa	28
3.1.3	Definição das Palavras-chave e <i>String</i> de Busca	28
3.1.4	Intervalo de Tempo para Pesquisa	29
3.1.5	Bases de Dados de Pesquisa	29
3.1.6	CrITÉRIOS de Inclusão e Exclusão	29
3.1.7	Procedimentos para Seleção dos Estudos	30
3.1.8	Extração e Sintetização de Dados	31
3.2	Condução da RSL	31
3.2.1	Fase 1	31
3.2.2	Fase 2	32
3.2.3	Fase 3	33
3.3	Resultados da RSL	35
3.3.1	QP1: Quais fatores indutores têm sido investigados para a descoberta das causas da evasão escolar?	35

3.3.2	QP2: Quais abordagens de descoberta de conhecimento têm sido utilizadas para descoberta das causas da evasão escolar?	36
3.3.3	QP3: Quais ferramentas de MD têm sido utilizadas para descoberta das causas da evasão escolar?	37
3.3.4	QP4: Quais técnicas de MD têm sido utilizadas para descoberta das causas da evasão escolar?	39
3.4	Considerações Finais	40
4	ABORDAGEM DE DESCOBERTA DE CONHECIMENTO	42
4.1	Atividades de Processamento	43
4.2	Atividades de Transformação	44
4.3	Atividades da Construção do Modelo	45
5	VALIDAÇÃO DA ABORDAGEM	47
5.1	Naive Bayes	52
5.2	<i>J48</i>	54
5.3	<i>KNN (K-Nearest Neighbors)</i>	54
5.4	<i>SVM (Support Vector Machine)</i>	55
5.5	<i>AdaBoost</i>	56
5.6	Considerações Finais	57
6	CONCLUSÕES E TRABALHOS FUTUROS	59
6.1	Sumário da Pesquisa	59
6.2	Limitações	60
6.3	Trabalhos Futuros	60
6.4	Publicações	61
	REFERÊNCIAS	62
	APÊNDICES	66
	APÊNDICE A – QUESTIONÁRIO PARA EXTRAÇÃO DE DADOS	67

1 Introdução

Nos últimos anos tem crescido o interesse e a preocupação em muitos países, em especial no Brasil, com o problema da evasão escolar e a determinação de seus principais fatores relacionados. Esse interesse é evidenciado por meio da observação de publicações científicas, recentes, sobre o tema, tais como as de: Márquez-Vera *et al.* (2016), Weybright *et al.* (2017), Adelman *et al.* (2018).

A evasão escolar é um caso social complexo, definido como o encerramento do ciclo de estudos (GAIOSO, 2005). Esse problema aflige as instituições de ensino em geral, sejam públicas ou privadas, uma vez que a saída dos alunos ocasiona externalidades negativas nos âmbitos sociais, acadêmicos, econômicos e ambientais, repercutindo de maneira negativa nas políticas públicas e privadas de investimento e desenvolvimento educacional.

O problema é ainda mais grave em cursos da área de Ciências Exatas. No Brasil, setenta e três por cento (73%) dos alunos de Ciência da Computação (CC) e sessenta e sete vírgula três por cento (67,3%) dos alunos de Sistemas de Informação (SI) desistem do curso (HOED, 2016). Embora esse problema se mostre relevante, ainda não se identificou um delineamento que possa esclarecer todos os fatores envolvidos no risco de abandono do ensino superior, tanto de curso, de instituição ou de sistema educacional (SCALI *et al.*, 2009).

No Brasil, a principal referência em relação ao tema é o relatório elaborado pela Comissão Especial de Estudos da Evasão, em que são aludidas as prováveis causas determinantes da evasão escolar no ensino superior. De acordo com esse relatório, o fenômeno da evasão escolar está relacionado à três (3) dilatadas categorias: às características individuais do estudante; aos fatores internos à instituição; e, aos fatores externos à instituição (BRASIL, 2009).

A identificação dos fatores que influenciam a evasão escolar e a atribuição de uma ordem de importância para esses fatores é um trabalho complexo, que está diretamente ligado à análise do conjunto de alunos (MANHÃES *et al.*, 2011). Vale destacar que cada instituição deve identificar as causas do fenômeno em seu ambiente educacional (NAGAI; CARDOSO, 2017).

Uma solução muito propícia para o estudo das causas da evasão escolar é o uso da descoberta de conhecimento, por meio de técnicas de bancos de dados ou Mineração de Dados (MD), denominado de Mineração de Dados Educacionais (MDE) (ROMERO; VENTURA, 2007). Essa área de pesquisa é responsável pelo desenvolvimento de métodos para entender melhor os alunos e o contexto em que eles aprendem (ROMERO; VENTURA, 2010). É possível aplicar técnicas de MDE para criar modelos que preveem especificamente

o abandono escolar.

Contudo, o processo de descoberta de conhecimento é complexo e composto por diversas etapas, tais como: seleção de dados, pré-processamento de dados, transformação de dados, mineração de dados, interpretação de dados e etc Kovanović *et al.*, (2015) . Algumas dessas etapas não são triviais. Além disso, para facilitar sua aplicação e melhorar a acurácia dos resultados, tanto o processo quanto suas etapas, devem ser adaptados de acordo com os diferentes domínios de aplicação.

A fase de pré-processamento de dados, comumente, requer uma dedicação significativa e pode consumir, em torno, de sessenta por cento (60%) a noventa por cento (90%) dos recursos do processo de descoberta de conhecimento, segundo Kovanović *et al.*, (2015) . Ademais, os autores indicam que a educação tem a complicação adicional de precisar mover dados em diferentes níveis do sistema, indo e vindo entre bancos de dados de sala de aula, escola, distrito e estado.

Além das lacunas de pesquisa no processo de descoberta de conhecimento em MDE (necessidade de adequação do processo/etapas de acordo com o domínio), existem questionamentos sobre os fatores que contribuem para a evasão escolar. Segundo Tontini e Walter (2017), no que ecoa aos fatores indutores à evasão escolar no ensino superior, estudos têm sido profusamente publicados, porém com abordagens e tratamentos diferentes para as causas indutoras. Adicionalmente, observa-se que poucos estudos foram produzidos para compreender a importância das diferentes dimensões e impacto em estudantes das Instituições de Ensino Superior (IES) do Brasil.

Outra ausência na literatura são estudos que tenham explorado métodos que possam auxiliar às IES na determinação do real risco de evasão escolar. Além disso, poucas pesquisas analisaram o quanto a identificação das causas da evasão escolar pode impactar na sua redução (TONTINI; WALTER, 2017).

1.1 Motivação

Nos últimos dez (10) anos houve um crescimento exponencial na Universidade Federal Rural do Semi-Árido (UFERSA), mais especificamente no campus Mossoró/RN, principalmente quanto ao número de cursos e de alunos matriculados em nível de graduação e pós-graduação. Todavia, um ponto de preocupação para o desenvolvimento da universidade é o crescimento do número de alunos que evadiram ou ficaram mais tempo que o previsto para concluírem a graduação.

A inserção de novos cursos a partir de dois mil e quatro (2004), vem aumentando a participação de candidatos nos processos seletivos da UFERSA, associado ao crescente número de cursos, tem-se também uma maior evasão escolar, seja por abandono, transferência

ou mobilidade (BRASIL *et al.*, 2015).

No curso de CC ofertado pela UFERSA, campus Mossoró/RN, o problema da evasão escolar é alarmante. Os registros mostram que desde o semestre de dois e seis ponto dois (2006.2), período que ingressou a primeira turma, se matricularam seiscentos e oitenta e nove (689) alunos até o semestre de dois mil e dezoito ponto um (2018.1), sendo que destes cinquenta e três vírgula um por cento (53,1%) evadiram. Apenas doze vírgula oito por cento (12,8%) concluíram o curso, enquanto os demais alunos são ativos ou estão com o curso trancado, podendo ainda retornarem. Essas estatísticas foram disponibilizadas pela própria instituição.

Atualmente, o processo de identificação das causas da evasão escolar e retenção dos alunos da UFERSA, campus Mossoró/RN, é manual, subjetivo, empírico e sujeito falhas, pois depende primordialmente da experiência acadêmica e do envolvimento, tanto dos docentes, como dos alunos. Geralmente, os docentes desempenham inúmeras atividades tais como orientação de alunos de pós-graduação, atividades administrativas e gerenciamento de projetos de pesquisa, ensino e extensão, além das atividades de sala de aula. Portanto, é difícil acompanhar e reconhecer as necessidades de cada aluno e identificar aqueles alunos que apresentam risco de evasão escolar. Contudo, entende-se que a adoção de mecanismos automatizados podem viabilizar a detecção precoce de grupos de alunos com risco de evasão escolar. Como consequência da detecção precoce, a instituição pode tomar medidas para mitigar os riscos e reduzir os índices de evasão escolar.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

O objetivo deste estudo é definir uma abordagem capaz de identificar alunos, em nível superior, com maior probabilidade de evadir.

1.2.2 Objetivos Específicos

A partir do objetivo geral foram definidos os seguintes objetivos específicos:

1. identificar fatores que influenciam no processo de evasão escolar;
2. definir quais dados relacionados a evasão escolar devem ser utilizados e com extraí-los;
3. definir quais técnicas de descoberta de conhecimento de dados podem ser utilizadas;
4. definir quais processos ou abordagens podem ser utilizados para a descoberta de conhecimento na base de dados;

5. definir quais ferramentas de descoberta de conhecimento podem ser utilizadas;
6. definir com que precisão a nova abordagem de descoberta de conhecimento em bases de dados é capaz de prever precocemente quais alunos podem evadir;
7. elencar as potenciais dimensões que levam os alunos a evadirem na UFERSA;
8. definir qual o impacto de cada dimensão na evasão escolar dos alunos de nível superior.

1.3 QUESTÕES DE PESQUISA

A partir do contexto e das motivações apresentadas anteriormente, este trabalho foi guiado para responder as seguintes questões de pesquisa:

Questão Geral de Pesquisa (QGP) - Quais as características dos alunos, em cursos de nível superior, com potencial para evadir?

Para detalhar a questão de pesquisa e definir questões específicas à abordagem, com as respectivas ferramentas, processos, algoritmos e perspectivas de resultados, são exibidas questões que foram subdivididas em três tipos de questões, são elas: Questões Conceituais (QC), Questões Tecnológicas (QT) e Questões Práticas (QP), respectivamente apresentadas a seguir:

1. **QC - Quais os fatores podem influenciar no processo de evasão escolar?**
 - a) Quais dados devem ser utilizados e como extraí-los para desvendar as causas da evasão escolar?
2. **QT - Quais as técnicas de descoberta de conhecimento de dados podem ser utilizadas?**
 - a) Qual processo ou abordagem de descoberta de conhecimento podem ser utilizados?
 - b) Quais as técnicas de MD/ aprendizagem de máquina podem ser utilizados para a descoberta de conhecimento na base de dados?
 - c) Quais ferramentas de descoberta de conhecimento podem ser utilizadas?
3. **QP - Com que precisão a nova proposta de abordagem de descoberta de conhecimento em bases de dados é capaz de prever precocemente quais alunos podem evadir?**
 - a) A abordagem de descoberta de conhecimento em bases de dados é capaz de determinar com que probabilidade as dimensões que levam os alunos a evadirem?

- b) A abordagem de descoberta de conhecimento em bases de dados é capaz de definir qual o impacto de cada dimensão na evasão escolar dos alunos de nível superior?

1.4 Contribuições do Trabalho

Por meio da Revisão Sistemática de Literatura (RSL) (capítulo 3) foi apresentado neste trabalho uma perspectiva do estado da arte em relação ao MDE. Após coletar e consultar a bibliografia publicada na área de MDE, os estudos mais importantes da área foram selecionados. Em seguida, foi classificado cada estudo não apenas pelo tipo de dados e técnicas de MD utilizadas, mas também, e mais importante, por aqueles que obtiveram os melhores níveis de acurácia.

A partir disso foi proposta uma abordagem de descoberta de conhecimento inovadora para identificar, automaticamente, de forma contínua os alunos considerados no grupo de risco de abandono, em salas de aula do ensino superior. Para validar a abordagem (MATEO), utilizou-se um banco de dados com informações coletadas de ex-alunos do curso de CC da UFERSA, campus Mossoró/RN.

Em geral, em relação aos resultados de classificação obtida no estudo de caso, por meio da utilização da abordagem (MATEO) de conhecimento proposta, têm-se as seguintes inferências:

- Mostrou-se a utilidade da técnica de seleção de atributos quando se tem um grande número de variáveis (lei de entropia). Neste caso especificamente, reduziu-se o número de atributos usados de vinte e três (23) atributos, inicialmente disponíveis, para os vinte e um (21) melhores atributos, obtendo menos regras e condições sem perder o desempenho da classificação;
- Constatou-se que algoritmos de classificação podem ser usados com sucesso para prever o desempenho acadêmico de um aluno e, em particular, para modelar a diferença entre alunos evadidos e concludentes.

Desta forma, conclui-se que é possível utilizar a abordagem para predição de grupos de alunos com tendência a evadir, alcançando, portanto, os objetivos definidos para este trabalho.

1.5 ORGANIZAÇÃO DO DOCUMENTO

O restante deste trabalho está estruturado da seguinte maneira: No Capítulo 2 se apresenta o referencial teórico, em que é abordada a descoberta de conhecimento em bases

de dados, mineração de dados educacionais, e, ainda, é apresentada uma contextualização dos determinantes da evasão escolar no ensino superior. No Capítulo 3 são mostrados os principais trabalhos da literatura que tratam do problema da evasão escolar, trazendo uma perspectiva de quais fatores, abordagens, ferramentas e técnicas são utilizados para desvendar as causas da evasão escolar. No Capítulo 4 é apresentada a abordagem proposta para desvendar o problema da evasão escolar no ensino superior conhecida como (MATEO). No Capítulo 5 são apresentados os resultados da validação da abordagem proposta e, por fim, no Capítulo 6 são apresentadas as considerações finais à cerca do trabalho.

2 Fundamentação Teórica

Neste capítulo são apresentados os conceitos fundamentais para a realização deste trabalho. Primeiramente, é apresentado o conceito de descoberta de conhecimento em bases de dados; em seguida, mineração de dados educacionais; e, por fim, determinantes para evasão escolar.

2.1 Descoberta de Conhecimento em Bases Dados

A tarefa de levantar e armazenar dados de diferentes tipos, formatos e fontes, ultrapassou os limites da capacidade humana de analisar, sumarizar e extrair conhecimento a partir desses dados. Além de fornecer grande capacidade de armazenamento, os bancos de dados tradicionais (relacionais) também disponibilizam ferramentas adequadas e robustas para manipulação desses. Em contrapartida, a assimilação e a análise dos dados requerem a utilização de ferramentas apropriadas, que automatizem o processo de análise dos dados e descoberta de conhecimento (FAYYAD *et al.*, 1996).

Os conceitos concernentes a descoberta de conhecimento em bases de dados flexionam fundamentos provenientes de diversas áreas, tais como a Inteligência Artificial (IA), a aprendizagem automática, o reconhecimento de padrões e a estatística. As aplicações de descoberta de conhecimento em bases de dados incorporam teorias, métodos e algoritmos oriundos dessas diferentes áreas de conhecimento, tendo como finalidade a extração de conhecimento a partir de grandes Banco de Dados (BD) (ÁLVARES, 2003).

A automatização do processo de descoberta de conhecimento em bases de dados ainda é passível de evoluções e adaptações, pois existem muitos problemas que ainda necessitam da intervenção do usuário para serem solucionados (FELDENS; CASTILHO, 2015). Essa área tem evoluído por meio de sistemas projetados e implementados para fins específicos, além de diversas adaptações de processos para fins de melhorias nos que já existem. Além disso, podem ser utilizados em várias bases de dados, mas que o objetivo da descoberta seja semelhante. Em compensação, os sistemas mais genéricos e algumas bases de dados mais complexas dependem da intervenção do usuário, requerendo deste, um conhecimento prévio e significativo sobre os dados, para que algum conhecimento possa, de fato, ser encontrado (ÁLVARES, 2003).

Existem diversos problemas que estão diretamente conexos com as decisões que precisam ser tomadas durante o processo de descoberta de conhecimento como, por exemplo, a representação do conhecimento extraído, a definição das prioridades do conhecimento descoberto e a seleção do método de MD mais adequado, sobretudo na escolha de métodos

para o tratamento de dados na etapa de pré-processamento.

2.1.1 Dificuldades e Desafios da descoberta de conhecimento em bases de dados

Em Fayyad *et al.* (2017) são destacados alguns problemas encontrados para lidar com dados, além das lacunas dos processos de descoberta de conhecimento em bases de dados para solucionar esses desafios. A seguir, são apresentados alguns desses problemas:

1. **Bancos de dados maiores:** bancos de dados com centenas de campos e tabelas e milhões de bases dados com tamanhos na escala de *terabytes* ou até maiores. Os métodos para trabalhar com grandes volumes de dados envolvem algoritmos mais eficientes, amostragem, aproximação e processamento altamente paralelo;
2. **Alta dimensionalidade:** frequentemente, nas bases de dados, existem um grande número de registros, mas também podem haver um número excessivo de campos (atributos, variáveis); nesses casos, a dimensionalidade do problema é alta. Um conjunto de dados alto-dimensional gera problemas em termos de aumentar o tamanho do espaço de busca para a inferência do modelo de forma combinatória. Além disso, aumenta as chances de que um algoritmo de MD não encontre padrões e comportamentos válidos. Abordagens de tratamento para este problema abrangem métodos para reduzir a dimensionalidade do problema e o uso de conhecimento prévio para identificar variáveis irrelevantes;
3. **Dados e conhecimentos alterados:** pequenas mudanças nos dados (não estacionários) podem fazer padrões previamente descobertos inválidos, e ainda, as variáveis medidas numa dada base de dados de aplicação podem ser modificadas, suprimidas, ou aumentadas com novas medições, com as mudanças sofridas pelas bases constantemente. As soluções possíveis incluem métodos incrementais para atualizar os padrões e tratar mudanças como uma oportunidade para a descoberta, usando-o para melhorar a busca de padrões;
4. **Dados em falta e “barulhentos”:** este problema é singularmente comum em bancos de dados. Em alguns casos as taxas de erros ou de dados faltantes chega a vinte por cento (20%) em alguns campos. Atributos importantes podem estar faltando se o banco de dados não foi projetado para posteriormente os dados armazenados serem utilizados para descoberta de conhecimento. As soluções possíveis envolvem estratégias estatísticas mais aprimoradas para identificar variáveis escondidas e dependências;
5. **Integração com outros sistemas:** um sistema de descoberta independente pode não ser muito útil. Problemas de integração particulares compreendem a integração

com um sistema de gerenciamento de banco de dados (por exemplo, através de uma interface de consulta), integração com planilhas e ferramentas de visualização, e que acomodam leituras de sensores em tempo real.

2.2 Mineração de Dados Educacionais

A MDE é um campo de pesquisa interdisciplinar, no qual técnicas de MD se justapõem em dados educacionais. O propósito é entender como os alunos adquirem conhecimento, reconhecem as configurações em que aprendem para melhorar os resultados educacionais e, ainda, verificar como suas respectivas características tendenciam a certos comportamentos, a exemplo, a decisão de evadir ou não. Observa-se que os sistemas educacionais podem armazenar uma enorme quantidade de dados oriundos de várias fontes e em diferentes formatos (VENKATACHALAPATHY *et al.*, 2017).

Logo, estudando os parâmetros/dados que afetam a educação ou comportamentos de estudantes, é possível prever o desempenho do aluno durante o programa acadêmico. Os alunos que se encontram em risco de concluir o programa acadêmico podem ser ajudados, tanto pessoalmente quanto academicamente, pela instituição familiar e acadêmica, para evitar que se afastem (MEHTA; BUCH, 2016).

Cada problema educacional singular diligencia diferentes tipos de técnicas de MD, porque as técnicas tradicionais de MD não podem ser aplicadas diretamente a esses tipos de dados e problemas. Há muitas ferramentas de MD disponíveis, mas elas não são projetadas para trabalhar com dados educacionais. Assim um gestor escolar não pode utilizar essas ferramentas sem o conhecimento de conceitos de MD.

A indisponibilidade do conjunto global de dados se torna um problema para uma pesquisa rica. O pré-processamento de dados ruidosos pode ser realizado usando várias técnicas, porém, de acordo com Papamitsiou e Economides (2014), o pré-processamento de dados educacionais não tem recebido esforços analíticos suficientes. Em contrapartida a etapa é o primeiro passo em qualquer processo de MD, sendo uma das tarefas mais importantes na MDE.

A falta de estudos e exploração dessa etapa da descoberta de conhecimento é surpreendente, uma vez que diversos autores destacam a importância da atividade para o processo, visto que a fase de pré-processamento de dados comumente pleiteia um esforço significativo e pode levar especialmente entre sessenta por cento (60%) e noventa por cento (90%) dos recursos do processo de descoberta do conhecimento (KOVANOVIĆ *et al.*, 2015). Em Mining (2012) é explicitada a dispendiosidade da etapa, pois ela pode consumir em torno de setenta por cento (70%) e muitas vezes chegando de oitenta por cento (80%) à oitenta e cinco por cento (85%) do esforço em análise de dados, dedicados à limpeza, formatação ou alinhamento de dados. Ainda foi destacado, que a educação tem

a complicação adicional de precisar mover dados em diferentes níveis do sistema, indo e vindo entre bancos de dados de sala de aula, escola, distrito e estado.

2.3 Determinantes da Evasão no Ensino Superior

A evasão escolar no ensino superior é um dos problemas mais amplos e fundamentais no campo da educação. A evasão ultrapassa inúmeros graus e peculiaridades de ensino e tem originado prejuízos sociais, econômicos, políticos, acadêmicos e financeiros a todos os envolvidos no processo educativo (MARTINHO *et al.*, 2013). Assim, faz-se necessário desenvolver métodos capazes de auxiliar na predição dos alunos em risco de abandono, possibilitando a adoção de intervenções proativas para minimizar o contexto.

Esse problema é inquestionável, uma vez que os estudos encontrados na RSL Marques *et al.* (2019), apresentam dados sobre a evasão escolar dos cursos de graduação das IES, em nível internacional, em que os estudos sobre a evasão escolar contornam os mais diversos cursos de graduação de uma IES.

Os primeiros relatos identificados de estudos acadêmicos e pesquisas científicas sobre o abandono escolar de alunos no ensino superior começaram na década de mil novecentos e trinta (1930), nos Estados Unidos, de forma modesta e insuficiente. Nos anos mil novecentos e setenta (1970), a preocupação com a situação, a complexidade do caso e a inevitabilidade de encontrar soluções acentuaram-se. Diante disso, Tinto (1975) apresentou o primeiro modelo de fatores contribuintes para o processo de evasão escolar. Esse modelo, apesar de ter passado por alterações, ainda é o mais empregado para se compreender o fenômeno da evasão/permanência de estudantes na universidade.

Nesse sentido, o autor recomendou como variáveis importantes a serem analisadas: as particularidades de ingresso do estudante, as incumbências e objetivos iniciais, as experiências institucionais, a adaptação acadêmica e social, os objetivos e compromissos finais e a decisão de permanência ou evasão escolar. De acordo com ele, os estudantes ingressam na universidade com diferenciadas características e circunstâncias pessoais, familiares e acadêmicas, que amoldam suas disposições iniciais e as intenções de frequentar o curso superior. Esses compromissos são provavelmente alterados e reformulados por meio das interações entre o estudante e os sistemas acadêmicos e sociais da instituição, que permitem a construção de uma condição de integração acadêmica e social no ensino superior.

Todavia, não existe um delineamento nacional ou internacional que possa esclarecer todos os fatores envolvidos no risco de abandono do ensino superior, tanto de curso, de instituição ou de sistema (SCALI *et al.*, 2009). No Brasil, ao que tange sobre o tema, foi identificado um relatório elaborado pela Comissão Especial de Estudos da Evasão, em que são assinaladas as possíveis causas determinantes da evasão no ensino superior, ao apontar

que esse fenômeno está associado à três amplas categorias, inerentes às Características individuais do Aluno; internos às instituições e externos às instituições (BRASIL, 2009):

2.3.1 Fatores Inerentes às Características Individuais do Aluno

São fatores pertinentes à adoção precoce da profissão, formação escolar anterior, insatisfação ou desinteresse dos alunos com cursos elegidos em segunda (2ª) ou terceira (3ª) opção, embaraços pessoais de adequação à vida universitária, falta de informação em relação a natureza dos cursos, descoberta de novos interesses que levam ao cumprimento de novo processo seletivo, dificuldades financeiras do aluno, divergências entre a vida acadêmica e as exigências do mundo do trabalho, fraco desempenho e dificuldades de aprendizagem, constantes reprovações e/ou baixa frequência às aulas, habilidade de estudo, personalidade do estudante, entre outros.

O estudante, ao ingressar em um curso superior, passa a viver uma outra realidade. A contraposição entre seu ambiente anterior (as escolas que frequentou nos níveis fundamental e médio, e no cursinho pré-vestibular) e o contexto que lhe é apresentado, muitas vezes, causa um desapontamento com o curso escolhido, e, por consequência, com a profissão escolhida, estimulando-o a repensar a sua escolha, podendo até acarretar no abandono do curso. Em contrapartida, uma avaliação positiva do curso, também por decorrência, pode reforçar a ideia do acerto da profissão escolhida (SCALI *et al.*, 2009).

2.3.2 Fatores Internos às Instituições

São fatores conectados à questões acadêmicas (por exemplo: currículos ultrapassados e distendidos, cadeia de pré-requisitos inteiriçados, além da falta de clareza sobre o próprio projeto pedagógico do curso); cultura institucional de desmerecimento da docência na graduação; estrutura débeis de apoio ao ensino de graduação (laboratórios, equipamentos de informática, bibliotecas, etc.); didático-pedagógicas (métodos inadequados de avaliação do desempenho do aluno, deficiência de formação pedagógica ou desinteresse do professor, desprovimento ou número reduzido de programas institucionais para o estudante, como iniciação científica, monitoria, entre outros); deficiência do sistema público nacional que proporcione a racionalização da utilização das vagas, suprimindo a possibilidade da matrícula em duas universidades.

2.3.3 Fatores Externos às Instituições

São fatores concernentes ao mercado de trabalho, à qualidade do ensino fundamental e médio associados às condições econômicas, ao reconhecimento social da profissão escolhida. Além disso, ainda se constata a depreciação das profissões, a exemplo, nos cursos de licenciaturas, às dificuldades de modernizar a universidade frente aos avanços tecnológicos,

econômicos e sociais da atualidade; à falta de políticas públicas plausíveis e continuadas direcionadas para o ensino de graduação.

O mercado de trabalho e a notoriedade da profissão escolhida também estão entre os fatores indutores à evasão escolar. A supressão de reconhecimento social da carreira escolhida, os baixos salários e as condições melindrosas de trabalho estão entre os fatores que cooperam para que os estudantes deixem os cursos (SCALI *et al.*, 2009).

2.4 Considerações Finais

Neste capítulo foi apresentado os conceitos fundamentais para construção e entendimento acerca dos principais assuntos abordados neste estudo. Primeiro foi apresentado o conceito de descoberta de conhecimento em bases de dados, em que foi destacada a motivação para o surgimento do conceito, em seguida, as semelhanças e a relação do tema com outras áreas, tais como: IA, aprendizagem automática, reconhecimento de padrões e estatística. Ainda em relação a esse tema, apresentou-se possíveis aplicações e principais dificuldades e desafios da descoberta de conhecimento em bases de dados.

O segundo tema abordado foi a MDE, no qual foi deslindado sobre os conceitos do assunto, as possíveis aplicações e às dificuldades inerentes e ainda foi apresentado um panorama dos recursos utilizados na etapa de pré-processamento, em que foi frisado que essa etapa pode consumir entre sessenta por cento (60%) e noventa por cento (90%) dos recursos.

No último tema abordado, apresentou-se sobre os determinantes da evasão escolar no ensino superior, destacou-se os principais prejuízos derivados do problema evasão escolar, as dimensões e amplitude acerca do problema, os primeiros relatos de pesquisa em busca de soluções e por fim, foi abordado um relatório em que são assinaladas as principais causas da evasão escolar no ensino superior do Brasil, destacando que esses fatores estão associados a três (3) grandes categorias: características individuais do aluno, internos às instituições e externos às instituições. No próximo capítulo, é apresentado uma RSL em que se buscou investigar temas e artefatos para auxiliar no desenvolvimento deste trabalho.

3 Revisão Sistemática da Literatura

Neste capítulo é apresentada uma RSL, com os objetivos de determinar os fatores investigados para a descoberta das causas da evasão escolar, as abordagens de descoberta de conhecimento, as ferramentas e técnicas de MD têm sido utilizadas para descoberta das causas da evasão escolar.

Sendo assim, buscou-se responder algumas questões de pesquisa deste estudo, como: Quais os fatores indutores têm sido investigados para a descoberta das causas da evasão escolar? Quais abordagens de descoberta de conhecimento têm sido utilizadas para descoberta das causas da evasão escolar? Quais ferramentas de MD têm sido investigadas e abordadas para descoberta das causas da evasão escolar? Quais técnicas de MD têm sido utilizadas para descoberta das causas da evasão escolar?

De acordo com Kitchenham (2004), uma RSL é uma pesquisa episódica que usa uma metodologia adequada e confiável para identificar, avaliar e interpretar toda a pesquisa disponível relevante para uma questão de pesquisa em particular, ou área de tópico, ou fenômeno de interesse de uma maneira repetível e imparcial. Tratando-se especificamente deste trabalho, a realização da RSL se dá pela necessidade de elucidar as principais ferramentas, técnicas e demais utensílios para desvendar causas da evasão escolar.

O restante deste capítulo está organizado da seguinte forma: na Seção 3.1, são apresentados os Procedimentos e Métodos; na Seção 3.2, descreve-se a Condução da RSL; na Seção 3.3, são apresentados os Resultados da RSL; e, na Seção 3.4, são discutidas as Conclusões e Considerações Finais deste capítulo.

3.1 Procedimentos e Métodos

Nesta seção são apresentados os detalhes do processo utilizado nesta RSL. Salienta-se, que esta RSL se baseia no protocolo definido por Biolchini *et al.* (2007), em que são aludidos os objetivos da pesquisa e as questões de pesquisa; são definidas as palavras-chave e *string* de busca; delimita-se o intervalo de tempo para pesquisa; são definidas as bases de dados de pesquisa, os critérios de inclusão e exclusão e, os procedimentos para extração e apresentação dos dados.

3.1.1 Objetivos

Diante do contexto supracitado, foram definidos os seguintes objetivos que nortearam as buscas:

- **Objetivo 1:** desvendar quais fatores indutores vêm sendo utilizado(s) para exploração das causas da evasão escolar;
- **Objetivo 2:** identificar quais abordagens de descoberta têm sido utilizadas para descoberta das causas da evasão escolar;
- **Objetivo 3:** conhecer quais as melhores ferramentas de MD são utilizadas para exploração dos dados de evasão escolar;
- **Objetivo 4:** identificar quais técnicas de MD são utilizados para exploração dos dados.

3.1.2 Questões de Pesquisa

De acordo com os objetivos apresentados na Seção 3.1.1, foram definidas as seguintes questões de pesquisa que norteiam esta RSL:

- **Questão Primária (QP1):** Quais os fatores indutores têm sido investigados para a descoberta das causas da evasão escolar?
- **Questão Primária (QP2):** Quais abordagens de descoberta de conhecimento têm sido utilizadas para descoberta das causas da evasão escolar?
- **Questão Primária (QP3):** Quais ferramentas de MD têm sido utilizadas para descoberta das causas da evasão escolar?
- **Questão Primária (QP4):** Quais técnicas de MD têm sido utilizadas para descoberta das causas da evasão escolar?

3.1.3 Definição das Palavras-chave e *String* de Busca

Tabela 1 – Palavras-chave e Sinônimos

Palavra chave	Sinônimos em inglês
Data Mining	<i>Knowledge Discovery, Data Extraction, Predicting, Prediction, Detection e Detecting</i>
School Dropout	<i>School Drop-Out, School Evasion, School Retention, School Truancy and School Failure</i>

A determinação das palavras-chave em RSLs é estabelecida a partir das questões de pesquisa sob investigação. Preliminarmente, foram realizados testes com as *strings* de busca que seriam utilizadas nas buscas. Na Tabela 1, são apresentadas as palavras-chave e seus sinônimos utilizados para montar a *string* de busca.

Após se efetuar as combinações das palavras-chave e sinônimos, foi delineada a seguinte *string* de busca:

(“Data Mining” OR “Knowledge discovery” OR “data extraction” OR predicting OR prediction OR detection OR detecting) AND (“school dropout” OR “school drop-out” OR “school evasion” OR “school retention” OR “school truancy” OR “school failure”)

3.1.4 Intervalo de Tempo para Pesquisa

Desde o início da década de mil novecentos e setenta (1970), já existia preocupação com o problema da evasão escolar, diante da complexidade e a necessidade de encontrar soluções plausíveis. Neste período, Tinto (1975) apresentou o primeiro modelo de fatores contribuintes para o processo de evasão escolar. Entretanto, considerou-se estudos realizados no período dos últimos dez (10) anos, compreendido entre dois mil e oito (2008) e (dezembro de) dois mil e dezoito (2018), por se considerar que, as constantes evoluções, no que se trata de fatores, abordagens, ferramentas e técnicas tem aperfeiçoado a busca das possíveis causas da evasão escolar.

3.1.5 Bases de Dados de Pesquisa

Para a RSL em questão, foram definidas quatro (4) bases de dados para pesquisa conforme se apresenta na Tabela 2. O critério de escolha dessas bases de dados se deve ao fato de serem as bases mais difundidas no campo da tecnologia. Essas bases de dados foram desenvolvidas com o intuito de propiciar, em uma única plataforma, centenas de revistas científicas e seus respectivos trabalhos, assentes por qualidade e originalidade (DANTAS *et al.* 2018).

Tabela 2 – Lista de fontes

Fontes	Endereço Eletrônico (link)
<i>IEEEExplore</i>	http://ieeexplore.ieee.org/Xplore/home.jsp
<i>ACM Digital Library</i>	http://dl.acm.org/
<i>Science Direct</i>	http://www.sciencedirect.com/
<i>Scopus</i>	https://www.scopus.com

3.1.6 Critérios de Inclusão e Exclusão

Os critérios de inclusão e exclusão são definidos para auxiliar na condução de uma RSL com o intuito de apoiar a classificação de relevância dos estudos (BIOLCHINI *et al.*, 2007). Os critérios de inclusão estão intimamente relacionados às questões de pesquisa em análise. Vale ressaltar que para um trabalho ser selecionado é suficiente atender a pelo menos um critério de inclusão. De maneira análoga, para excluir um trabalho é suficiente atender a um critério de exclusão.

Os seguintes critérios de inclusão de trabalhos foram definidos:

- Critério de Inclusão (CI1): o trabalho discute, aplica, propõe ou investiga, fatores ou dados relacionados ao problema da evasão escolar de alunos no contexto educacional;
- Critério de Inclusão (CI1): o trabalho discute, aplica, propõe ou investiga, abordagens de descoberta de conhecimento relacionado ao problema da evasão escolar de alunos no contexto educacional;
- Critério de Inclusão (CI2): o trabalho discute, aplica, propõe ou investiga, ferramentas de MD relacionado ao problema da evasão escolar de alunos no contexto educacional;
- Critério de Inclusão (CI1): o trabalho discute, aplica, propõe ou investiga, técnicas de exploração ou descoberta de conhecimento em bases de dados relacionado ao problema da evasão escolar de alunos no contexto educacional;
- Critério de Inclusão (CI2): o trabalho aborda estudos experimentais para validar os resultados que tratam do problema da evasão escolar.

Os seguintes critérios de exclusão de trabalhos foram definidos:

- **CE1:** trabalhos em andamento, que não tenham sido concluídos, ou que o estudo final não esteja disponível;
- **CE2:** trabalhos que não tratem da utilização de MD na descoberta das causas da evasão escolar, em casos práticos, serão excluídos;
- **CE3:** trabalhos em outros idiomas que não o inglês, devido à baixa ascendência de estudos sobre o tema em outros idiomas;
- **CE4:** caso existam trabalhos duplicados, apenas um deles será aceito (o mais recente/desenvolvido), excluindo-se todos os demais;
- **CE5:** trabalhos que foram publicados em *Journal*, Revista ou Conferência com *Qualis* acima de B5 ou que tenha sido publicado em periódicos com fator de impacto igual ou superior a zero vírgula setecentos (0,700), independente da área de estudo.

3.1.7 Procedimentos para Seleção dos Estudos

Neste trabalho, para realização da seleção de estudos, foram seguidas três (3) fases:

- **Fase 1:** selecionar todos os trabalhos científicos, relacionados à *string* de busca;
- **Fase 2:** realizar a leitura de títulos e resumos das publicações coletadas e eleger aquelas que satisfazem aos critérios de inclusão e excluir os que atendem aos critérios de exclusão;

- **Fase 3:** analisar os trabalhos selecionados na fase anterior, que nesta fase, são lidos integralmente e novamente aplicar os critérios de inclusão e exclusão. Os trabalhos aceitos nesta fase são utilizados para responder as questões de pesquisa.

3.1.8 Extração e Sintetização de Dados

A estratégia de extração de dados compreendeu em elaborar um formulário de forma a sintetizar as análises feitas em cada trabalho, além de gráficos e tabelas com estatísticas de inclusões e exclusões em cada máquina de busca pesquisada. Com o formulário de extração, buscou-se informações a respeito das questões de pesquisa que norteiam este trabalho e informações relevantes aos objetivos da RSL, conforme é apresentado na Tabela 3.

Tabela 3 – Ficha de extração de dados

Informações Gerais				
Título:				
Fonte:	Nº Páginas:	Data da Publicação:	Qualis:	
Idioma:				
Resumo do Trabalho:				
Critérios de Inclusão:				
Observações:				

3.2 Condução da RSL

A RSL foi conduzida em um intervalo de cinco (5) meses (08/2018 à 12/2018), seguindo o planejamento previsto nas seções anteriores. Foram realizados testes (busca piloto) na *IEEEExplore*, com diferentes *strings* de busca e verificada uma quantidade limitada de resultados, com o propósito de escolher aquela *string* que obteve a melhor taxa de resultados aceitáveis. Ao todo, foram retornados cento e cinquenta (150) trabalhos considerando todas as máquinas de busca, que foram submetidos para as fases um (1) e dois (2). Nas subseções a seguir, são apresentados mais detalhes das buscas realizadas em cada base.

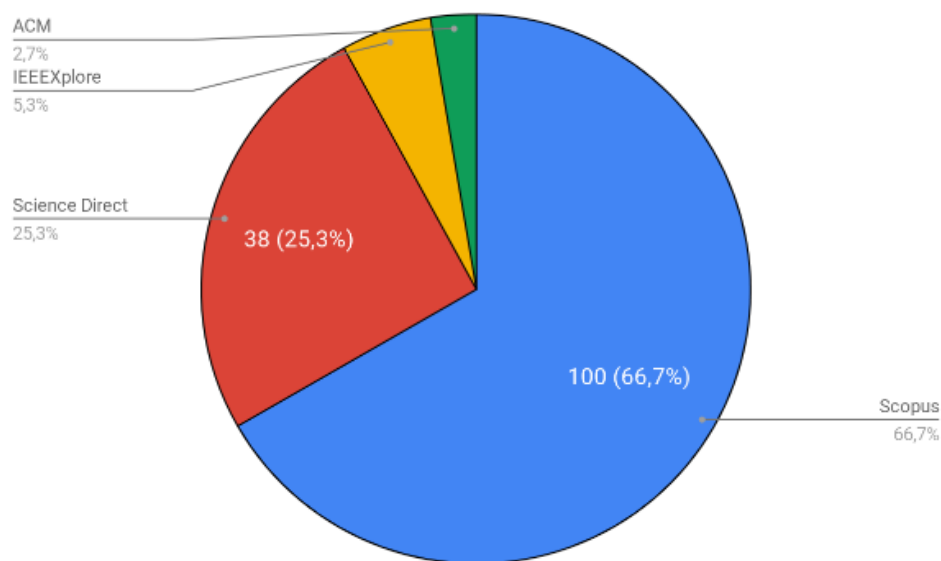
3.2.1 Fase 1

A ordem de pesquisa nas bases, bem como o quantitativo de trabalhos encontrados, são apresentadas, para cada fonte a seguir. Para todas as fontes foi utilizada a mesma *string* de busca definida na subseção 3.1.7, todos os trabalhos aceitos, foram publicados no intervalo de dois mil e oito (2008) até dois mil e dezenove (2019), é importante observar

que mesmo tendo finalizado a pesquisa em dezembro de dois mil e dezoito (2018), alguns trabalhos que seriam publicados em dois mil e dezenove (2019) foram retornados.

- **IEEEExplore:** a primeira (1ª) base pesquisada foi a *IEEEExplore*, que retornou oito (8) trabalhos, utilizando o *Advanced Search*, na aba *Command Search*;
- **ACM Digital Library:** a segunda (2ª) base pesquisada foi a *ACM Digital Library*, que retornou quatro (4) trabalhos, utilizando o *Search*;
- **Science Direct:** a terceira (3ª) base utilizada foi a *Science Direct*, que retornou trinta e oito (38) trabalhos. Observa-se que foi utilizado o comando *Advanced search*, na aba *Expert search*;
- **Scopus:** A quarta (4ª) base utilizada foi a *Scopus*, que retornou cem (100) trabalhos, sendo utilizado o comando *Advanced*.

Figura 1 – Resultado das buscas por base



Fonte: Autoria Própria (2020).

De forma geral, o resultado obtido da execução da busca para cada uma das fontes definidas se encontra representado na Figura 1. Constata-se que foram retornados cento e cinquenta (150) trabalhos, salientando-se o número expressivo de trabalhos retornados na *Scopus* em detrimento das demais bases pesquisadas, com aproximadamente sessenta e sete por cento (67%) dos trabalhos encontrados.

3.2.2 Fase 2

Nesta fase foi realizada a leitura de títulos e resumos dos trabalhos retornados com a *string* de busca. Os resultados inerentes a cada base, após este procedimento, podem ser

visualizados na Tabela 4. Destaca-se que após a fase 1, a máquina de busca em que foram selecionados mais trabalhos foi a *Scopus*, na qual foram selecionados um total de quatorze (14) trabalhos. Em contrapartida, na *ACM Digital Library* apenas um (1) trabalho restou após os filtros. No total, vinte e dois (22) trabalhos foram selecionados para serem avaliados na fase seguinte (Fase 3).

Tabela 4 – Resultados dos filtros

Bases	Exclusão - Fase 2	Exclusão - Fase 3	Selecionados
<i>IEEEExplore</i>	4	2	2
<i>ACM</i>	3	0	1
<i>Science Direct</i>	35	0	3
<i>Scopus</i>	86	6	8

3.2.3 Fase 3

Nesta fase, os trabalhos obtidos depois dos filtros da fase 2, foram lidos completamente e mais uma vez foi verificado se atendiam aos critérios de inclusão mencionados anteriormente. Destaca-se, que foram excluídos seis (6) trabalhos da máquina de busca *Scopus* e dois (2) da *IEEEExplore*, conforme apresentado na Tabela 4.

Sendo assim, ressalta-se que foram selecionados finalmente após todos os filtros, quatorze (14) trabalhos, sendo estes utilizados para responder as questões de pesquisa. A seguir, são apresentadas características relacionadas aos trabalhos selecionados, levando em consideração pontos que julgam-se relevantes:

Apresenta-se na Figura 2, uma visão geral do quantitativo dos trabalhos selecionados entre os anos de dois mil e oito (2008) e dois mil e dezoito (2018). Evidencia-se que os anos de 2013 e 2017, foram retornados um maior número de trabalhos com três (3) trabalhos aceitos. Em compensação os anos de dois mil e oito (2008) e dois mil e onze (2011) não foram aceitos nenhum trabalho.

Na Tabela 5, exibe-se o quantitativo de periódicos em que os trabalhos aceitos foram encontrados. Nota-se, que o *Journal of Adolescence* foi o que retornou o maior número de trabalhos com três (3) e, logo em seguida, o *The Journal of Educational Research* com dois (2) trabalhos. Os demais periódicos retornaram apenas um (1) trabalho.

Tabela 5 – Resultados dos trabalhos encontrados em jornais.

Periódicos	Link	Quantidade
<i>Journal of Adolescence</i>	https://www.journals.elsevier.com/journal-of-adolescence/	3
<i>The Journal of Educational Research</i>	https://www.tandfonline.com/loi/vjer20	2
<i>Plos One</i>	https://journals.plos.org/plosone/	1
<i>School Psychology quarterly</i>	https://www.apa.org/pubs/journals/spq/	1
<i>Education Economics</i>	https://www.tandfonline.com/loi/cede20	1
<i>South African Journal of Education</i>	http://www.sajournalofeducation.co.za/index.php/saje	1
<i>Expert Systems</i>	https://www.expertsystem.com/	1
<i>Journal of Latin-American Learning Technologies</i>	http://iecc-edusociety.org/about/latin-american-learning-technology-rita	1
<i>International Journal of Educational Development</i>	https://www.journals.elsevier.com/international-journal-of-educational-development	1

Figura 2 – Quantitativo de trabalhos selecionados por ano



Fonte: Autoria Própria (2020).

Enquanto isso, observa-se na Tabela 6 que foram retornados trabalhos de duas (2) conferências diferentes, em ambas foi retornado um (1) trabalho.

Tabela 6 – Resultados dos trabalhos encontrados em conferências.

Conferências	Link	Quantidade
<i>International Conference on Learning Analytics and Knowledge</i>	https://solaresearch.org/events/lak/	1
<i>International conference on Tools with Artificial Intelligence</i>	http://www.ictai2019.org/	1

Na Tabela 7 são apresentados os trabalhos selecionados, com os respectivos anos de publicação e o continente em que a pesquisa foi explorada. Constata-se que apesar das oscilações nos últimos dez (10) anos, tem crescido o número de publicações e práticas referentes a busca pelas possíveis causas da evasão escolar e por soluções para o problema.

Em relação a distribuição por continentes, é possível observar que o interesse e pesquisas em relação a evasão escolar, tem sido grande em diversas partes do planeta. Destaca-se o quantitativo na América do Norte, de onde foram retornados metade dos trabalhos. Vale salientar, que os Estados Unidos tem sido o país que mais investiga sobre o problema.

Tabela 7 – Relação dos trabalhos selecionados

Índice	Trabalhos	Ano_Publicação	Continentes
#1	(ARCHAMBAULT <i>et al.</i> , 2009)	2009	América do Norte
#2	(BOWERS, 2010)	2010	América do Norte
#3	(FALL; ROBERTS, 2012)	2012	América do Norte
#4	(SABATES <i>et al.</i> , 2013)	2013	Ásia
#5	(MÁRQUEZ-VERA <i>et al.</i> , 2013)	2013	América do Norte
#6	(MARTINHO <i>et al.</i> , 2013)	2013	América do Sul
#7	(EICHER <i>et al.</i> , 2014)	2014	Europa
#8	(JIMÉNEZ-GÓMEZ <i>et al.</i> , 2015)	2015	Europa
#9	(PARR; BONITZ, 2015)	2015	América do Norte
#10	(MÁRQUEZ-VERA <i>et al.</i> , 2016)	2016	América do Norte
#11	(WEYBRIGHT <i>et al.</i> , 2017)	2017	África
#12	(ROVIRA <i>et al.</i> , 2017)	2017	Europa
#13	(WOOD <i>et al.</i> , 2017)	2017	América do Norte
#14	(ADELMAN <i>et al.</i> , 2018)	2018	América Central

3.3 Resultados da RSL

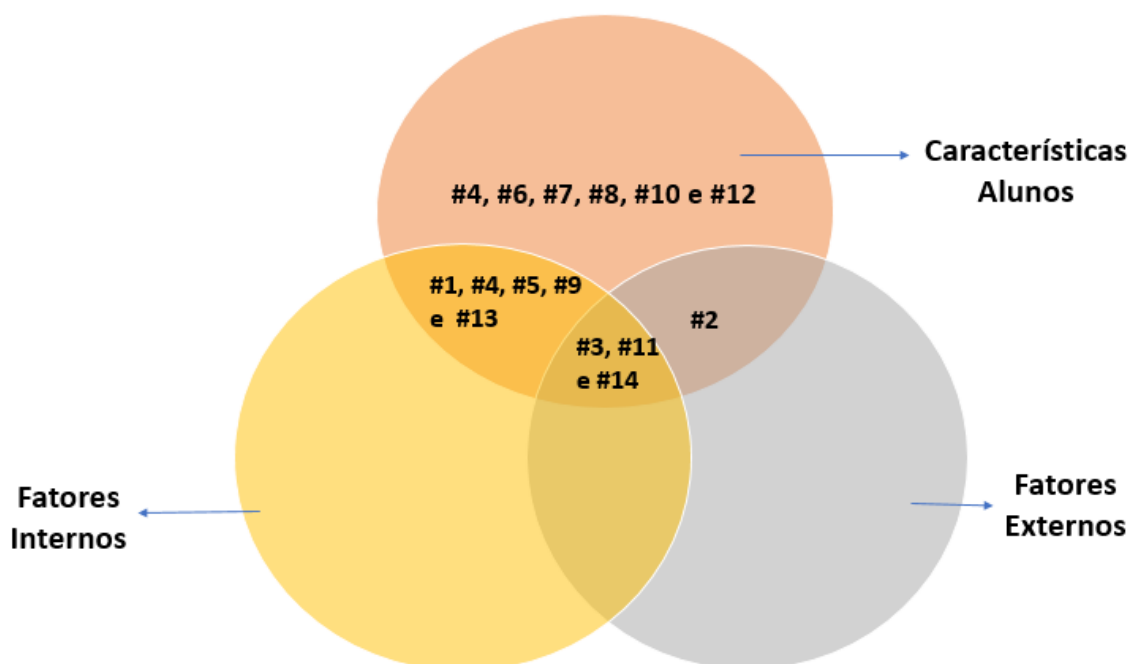
Nesta seção são apresentadas as respostas para as questões de pesquisa, que foram encontradas a partir da análise dos trabalhos selecionados. Por meio de formulários de extração de dados visto na Seção 3.1, buscou-se pontos importantes dos trabalhos que respondessem às questões de pesquisa e atendessem aos objetivos, também destacados na Seção 3.1.

3.3.1 QP1: Quais fatores indutores têm sido investigados para a descoberta das causas da evasão escolar?

Na Figura 3, são exibidos os fatores envolvidos no risco de evasão escolar, com os índices dos respectivos trabalhos retornados na RSL, a relação de índices e trabalhos é apresentada na Tabela 7. Salienta-se que em todos os trabalhos, foram explorados dados inerentes às características individuais do aluno. Destaca-se que apenas nos trabalhos (#3, #11 e #14) foram explorados dados inerentes às características individuais do aluno, fatores internos às instituições de ensino e externos às instituições de ensino.

Os fatores relacionados às características individuais do aluno, nos trabalhos (#1, #2, #4, #7, #8 e #14) foram explorados dados relacionados a desinteresse e insatisfação com a escola ou universidade, descoberta de novos interesses e fraco desempenho. Enquanto isso, nos trabalhos (#5, #6, #11, #12 e #13) foram abordados dados relacionados as dificuldades de aprendizagem, constantes reprovações e baixa frequência nas aulas, e, por fim, nos trabalhos (#3, #9, #10 e #11) foi investigado como causa da evasão escolar a falta de habilidade de estudo dos alunos.

Figura 3 – Fatores Indutores para à Evasão Escolar.



Fonte: Autoria Própria (2020).

Em relação aos fatores associados às instituições, abordou-se nos trabalhos (#3, #4 e #13) dados relativos a currículos ultrapassados e distendidos, estrutura débeis de apoio ao ensino (laboratórios, equipamentos de informática, bibliotecas, etc); enquanto nos trabalhos (#1, #5, #9, #11 e #14) dados sobre didático-pedagógicas (métodos inadequados da avaliação do desempenho do aluno, deficiência de formação pedagógica ou desinteresse do professor, desprovimento ou número reduzido de programas institucionais para o estudante), entre outros.

Tratando-se de fatores externos às instituições nos trabalhos (#2, #3, #11 e #14) foram estudados dados concernentes ao meio em que o aluno está inserido. São exemplos: falta de reconhecimento social com educação, falta de perspectivas empregatícias, falta de políticas públicas plausíveis e continuadas voltadas para educação, entre outras.

3.3.2 QP2: Quais abordagens de descoberta de conhecimento têm sido utilizadas para descoberta das causas da evasão escolar?

Na Tabela 8 são destacadas as abordagens de descoberta de conhecimento utilizados para descoberta de conhecimento em bases de dados nos trabalhos retornados, salienta-se o quantitativo de propostas de abordagens de descoberta de conhecimento totalizando sete (7) trabalhos. Logo em seguida, aparecem as abordagens que foram adaptadas contabilizando um total de seis (6) trabalhos, sendo que todas essas adaptações são constituídas de uma ou mais etapas do *Knowledge Discovery in Databases (KDD)*. Por sua vez, o *KDD* foi utilizado integralmente em um (1) trabalho.

Tabela 8 – Abordagens de conhecimento utilizadas nos trabalhos

Índice	Abordagens	Quantidade de trabalhos
1	Propôs Abordagem	7
2	Adaptou Abordagem	6
3	KDD	1

Nos trabalhos em que foram utilizadas adaptações de outras abordagens, as principais abordagens bases foram: *KDD*, *CRISP-DM* e *SEMMA*. Assim, em cada trabalho em particular foram utilizadas apenas algumas etapas da abordagem base, ou mesmo junções de etapas de duas ou mais abordagens. Foram utilizadas adaptações nos trabalhos de Fall e Roberts (2012), Sabates *et al.* (2013), Parr e Bonitz (2015), (Archambault *et al.* (2009), Bowers (2010), Rovira *et al.* (2017) e Martinho *et al.* (2013).

É importante realçar que as principais modificações/adaptações, foram sobre a etapa de pré-processamento, uma vez que, constata-se que é a etapa mais trabalhosa. Ainda se deve salientar, que os autores levaram em consideração a disposição dos dados.

Nos trabalhos que foram propostas abordagens diferentes das abordagens tradicionais, foram adicionadas novas etapas ou mesmo foram consideradas outras metodologias para descoberta de conhecimento. Na maioria dos trabalhos, foram adicionadas etapas para facilitar a extração de dados e para definir os melhores atributos para utilização nos modelos. Esta abordagem foi utilizada nos seguintes trabalhos: Adelman *et al.* (2018), Márquez-Vera *et al.* (2013), Wood *et al.* (2017), Márquez-Vera *et al.* (2016), Jiménez-Gómez *et al.* (2015) e Eicher *et al.* (2014).

Enquanto isso, o KDD que é constituído de cinco (5) etapas principais para a descoberta de conhecimento, são as etapas de seleção, pré-processamento transformação, mineração de dados e avaliação e interpretação dos dados, foi utilizado no trabalho de Weybright *et al.* (2017).

Nesse caso, ressalta-se a abordagem de descoberta de conhecimento *KDD*, uma vez que, foram utilizadas algumas de suas etapas em todos os trabalhos que fizeram adaptações de abordagens, e, ainda foi utilizada em um (1) trabalho integralmente. Além disso, é o processo com maior destaque na literatura, utilizado em trabalhos que envolvem MD.

3.3.3 QP3: Quais ferramentas de MD têm sido utilizadas para descoberta das causas da evasão escolar?

Na Tabela 9 é apresentado o quantitativo de uso das ferramentas de MD nos trabalhos selecionados. É possível verificar que as ferramentas *Waikato environment for Knowledge Analysis (Weka)* e *Mplus*, destacam-se sendo utilizadas em três (3) trabalhos, as demais ferramentas *Statistical Analysis System (SAS-STAT)*, *Hierarchical Linear Modeling*

(HLM), *Linguagem Python*, *Statistical Package for the Social Science* (SPSS) e *Software R* foram utilizadas em apenas um (1) trabalho.

Tabela 9 – Ferramentas utilizadas na MD

Índice	Ferramentas	Quantidade de trabalhos
1	<i>Mplus</i>	3
2	<i>Weka</i>	3
3	<i>SPSS</i>	1
4	<i>Linguagem Python</i>	1
5	<i>Software R</i>	1
6	<i>SAS-STAT</i>	1
7	<i>HLM</i>	1

A *Weka* foi utilizada nos trabalhos de Márquez-Vera *et al.* (2016) , Márquez-Vera *et al.* (2013) e Jiménez-Gómez *et al.* (2015). A ferramenta é uma coleção de algoritmos de aprendizado de máquina para tarefas de MD, oferece aparatos para preparação de dados, classificação, regressão, agrupamento, mineração de regras de associação e visualização.

O *Mplus* foi utilizado nos trabalhos de Parr e Bonitz (2015), Archambault *et al.* (2009) e Fall e Roberts (2012). A ferramenta é um programa de modelagem estatística para análises de dados, que oferece uma ampla variedade de modelos, estimadores, algoritmos e exibições gráficas de dados e resultados de análises. Essa ferramenta também permite análise de dados transversais e longitudinais, dados de nível único e multinível, dados provenientes de diferentes populações com heterogeneidade observada ou não observada e dados que contêm valores omissos.

O *SAS-STAT*, empregado no trabalho de Weybright *et al.* (2017), permite realizar análises estatísticas, incluindo análise de variância, regressão, análise categórica de dados, análise multivariada, análise de sobrevivência, análise psicométrica, análise de *cluster*, análise não paramétrica, análise de modelos mistos e análise de dados de pesquisa com numerosos exemplos, além de informações sobre sintaxe e uso.

O *HLM* utilizado no trabalho de Wood *et al.*, (2017), é um *software* estatístico para modelagem linear hierárquica. O *Python* foi utilizado para manipular os dados no trabalho de Rovira *et al.*, (2017) com as bibliotecas *pandas* e *scikit-learn*, é uma linguagem de programação de uso geral que pode ser usada para diferentes fins: coleta de dados, engenharia de dados, análise, *Web Scraping*, construção de aplicativos *web* e etc. O *Software R* utilizado no trabalho de Eicher *et al.* (2014), dá suporte em análise de dados e estatística. O *SPSS* manuseado no trabalho de Bowers (2010), é um pacote de *software* usado para análise estatística interativa ou em lote.

Apresenta-se na Tabela 10, um comparativo das ferramentas abordadas com base em três (3) critérios. Observa-se que as ferramentas (*Weka*, *Python* e *Software R*) tem licença livre, todas as ferramentas oferecem documentação e apenas as ferramentas (*Weka*

Tabela 10 – Comparativo das Ferramentas

Ferramentas	Licença	Documentação	Open Source
<i>Weka</i>	Livre	Sim	Sim
<i>Mplus</i>	Paga	Sim	Não
<i>SAS-STAT</i>	Paga	Sim	Não
<i>HLM</i>	Paga	Sim	Não
<i>Python</i>	Livre	Sim	Não
<i>Software R</i>	Livre	Sim	Sim
<i>SPSS</i>	Paga	Sim	Não

e Software R) são *Open Source*.

3.3.4 QP4: Quais técnicas de MD têm sido utilizadas para descoberta das causas da evasão escolar?

Na Tabela 11 são exibidas as técnicas de MD que foram exploradas nos trabalhos encontrados, enfatiza-se as técnicas de classificação que foram utilizadas em quatro (4) trabalhos, em seguida as Equações Estruturais empregadas em três (3) trabalhos.

Tabela 11 – Técnicas e algoritmos utilizados na MD

Índice	Técnicas	Quantidade de trabalhos
1	Classificação	4
2	Equações Estruturais	3
3	Regressão	2
4	Análise de Sobrevivência	2
5	Análise Fatorial	2
6	Redes Neurais	1
7	Análise Multinível	1

A técnica de Classificação foi explorada nos trabalhos de Rovira *et al.* (2017) , Márquez-Vera *et al.* (2016) , Jiménez-Gómez *et al.* (2015) e Sabates *et al.* (2013) . Essa técnica é uma tarefa da MD que associa ou classifica objetos a determinadas classes. Ela busca prever uma classe de um novo dado automaticamente. As Equações Estruturais utilizadas nos trabalhos de Parr e Bonitz (2015), Archambault *et al.* (2009) e Fall e Roberts (2012), são utilizadas para analisar múltiplas variáveis, que normalmente estão associadas a um determinado indivíduo por exemplo.

A técnica de Regressão abordada nos trabalhos de Adelman *et al.* (2018) e Wood *et al.* (2017), é utilizada para se prever um valor real, por exemplo, em que período do curso um aluno irá evadir do curso. A técnica de Análise Sobrevivência, foi explorada nos trabalhos de Weybright *et al.* (2017) e Bowers (2010), é utilizada para se estabelecer o tempo até a ocorrência de determinado evento.

A Análise Fatorial empregada nos trabalhos de Archambault *et al.* (2009) e Fall e Roberts (2012), é uma técnica estatística exploratória, que tem como objetivo reduzir grandes conjuntos variáveis em grupos menores e altamente correlacionadas. A técnica de Rede Neural foi utilizada no trabalho de Martinho *et al.* (2013), essa técnica é inspirada no sistema nervoso central de um animal, para realização de aprendizado de máquina e reconhecimento de padrões. A Análise Multinível manipulada no trabalho de Eicher *et al.* (2014), é uma técnica que permite a regressão diferenciada de dados longitudinais, ou seja, coletados ao longo do tempo.

3.4 Considerações Finais

Neste capítulo foram apresentados o planejamento, a condução e os resultados de uma RSL sobre o problema de evasão escolar, em que se buscou elencar ferramentas, técnicas, abordagens de conhecimento e fatores indutores utilizados para prever as causas do problema nos últimos dez (10) anos. A busca pelos trabalhos resultou na pré-seleção de dezesseis (16) trabalhos, dentre os quais quatorze (14) foram incluídos para a extração de dados.

A partir dos resultados, evidenciam-se que a ferramenta *Weka* se consolida como uma das ferramentas de MD mais utilizadas para auxiliar a desvendar causas da evasão escolar, por meio da aplicação de técnicas e algoritmos sobre conjuntos de dados. Com essa ferramenta, é possível acompanhar processos de MD de pequenas bases de dados, realizando a avaliação dos resultados obtidos e a comparação de algoritmos. Ademais, é possível executar tarefas relacionadas ao pré-processamento de dados como, por exemplo, a seleção e a transformação de atributos.

Foi possível verificar que a técnica de Classificação vem sendo bastante utilizada e tem retornado alta precisão nas previsões de tendência de evasão escolar. Vale realçar que em setenta e cinco por cento (75%) das vezes que essa técnica foi utilizada, utilizou-se a ferramenta *Weka*. Os algoritmos de Classificação mais utilizados nos trabalhos foram: *Naive Bayes (NB)*, *Support-Vector Machine (SMO)*, *Retwork of Radial Basis Function (RBFNetwork)*, *Multilayer Perceptron (MLP)*, *k-Nearest Neigh-Bours (KNN)*, *Jrip*, *OneR*, *J48*, *PART* e *AdaBoostM1*.

Em relação as abordagens para descoberta de conhecimento, pelo menos metade dos trabalhos propuseram novas abordagens, porém destaca-se o *KDD*, que foi usado parcialmente em todos os trabalhos que fizeram adaptações e usado integralmente em outro trabalho. Quanto aos fatores indutores para a evasão escolar, constatou-se que em todos os trabalhos são investigadas possíveis causas relacionadas às características individuais do aluno, analisando dados armazenados nas instituições e/ou levantando dados por meio de questionários, da perspectiva do aluno. Vale ressaltar, que possíveis fatores relacionados à

instituição e externos à instituição, ainda são pouco explorados.

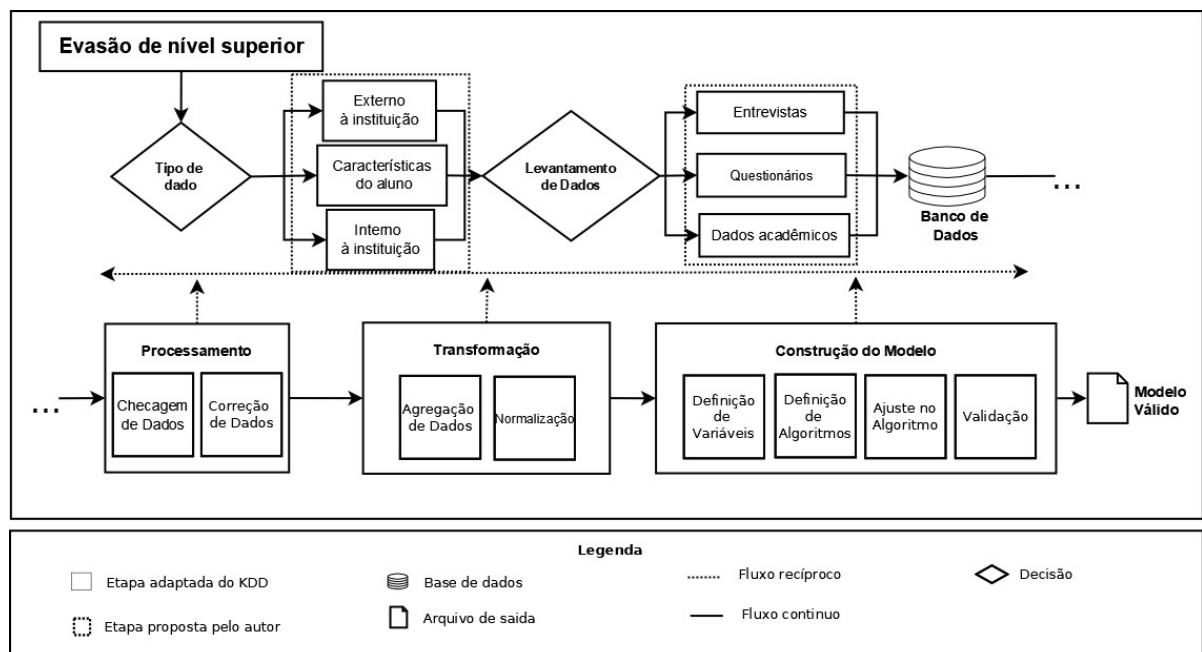
Por fim, é importante enfatizar que qualquer pesquisa desta natureza possui riscos à validade dos seus resultados. Para esta RSL em particular, detectam-se os seguintes pontos de fragilidade: a pré-seleção dos trabalhos, devido ao conjunto limitado de bases consultadas, ao número de trabalhos encontrados, e pelo fato de alguns títulos e resumos não terem refletido de forma adequada o conteúdo de seus textos; as palavras-chave de busca que, mesmo se manuseando alguns sinônimos, podem não contemplar a todos os existentes para os termos da pesquisa, podendo ser insuficientes para capturar todos os trabalhos relevantes; e, o dilatado tempo necessário para a efetuação de uma RSL, podendo conduzir a uma diferente captura de resultados nas bases consultadas, em diferentes datas.

4 Abordagem de Descoberta de Conhecimento

Considerando o contexto supracitado nos capítulos anteriores e visando facilitar a descoberta das possíveis causas da evasão escolar no ensino superior, neste capítulo se apresenta uma abordagem de descoberta de conhecimento como mostrado na Figura 4. Nesta abordagem, uma vez definida os fatores que serão estudados e, posteriormente, adquiridos os dados, estes dados passam por um tratamento programado através de um encadeamento de atividades nas etapas de processamento e transformação, que tenciona aumentar a qualidade do BD.

Logo após, um modelo de predição é construído por meio da execução de um algoritmo de MD. O especialista avalia o modelo desenvolvido, se os resultados se mostrarem relevantes, o mesmo pode implementar o modelo e usá-lo em previsões futuras. Senão, retorna-se a etapas anteriores com o intuito de aumentar a qualidade dos resultados. O conjunto de etapas incorporadas nesta abordagem de conhecimento são expostas adiante.

Figura 4 – Abordagem de descoberta de conhecimento para identificar causas da evasão escolar no ensino superior.



Fonte: Autoria Própria (2020).

A abordagem tem início a partir de uma decisão de qual tipo de fatores indutores para evasão escolar devem ser utilizados para desvendar as causas da evasão escolar, como visto na Seção 2.3, as causas da evasão escolar se enquadram a três (3) grandes categorias: fatores inerentes às características individuais do aluno, fatores internos às instituições

e fatores externos às instituições. Vale salientar, que não necessariamente é obrigatório decidir por apenas uma categoria, o especialista pode decidir por uma mais categorias para utilizar nas buscas das causas da evasão escolar em determinada instituição.

Uma vez que se estabelece que tipos de fatores utilizar, inicia-se a segunda etapa da abordagem, que é decidir como será o levantamento dos dados e de onde extraí-los. Na Subseção 3.3.1 foi possível observar, que em todos os trabalhos selecionados na RSL, empregou-se pelo menos uma das seguintes formas para levantamento de dados: dados acadêmicos, questionários e entrevistas. Assim sendo, nesta etapa se decide por qual dessas formas é possível utilizar e a mais conveniente, assim como na etapa anterior, nesta etapa também é possível decidir por utilizar uma (1) ou mais das opções disponíveis.

Nos casos em que se decide pela utilização de questionários e/ou entrevistas, recomenda-se o emprego de questões objetivas de múltiplas escolhas fechadas essa estratégia evita ou minimiza trabalhos posteriores com processamento e transformação de dados. Uma vez que, a característica fundamental destas questões é evitar o caráter subjetivo das respostas, o oposto disto leva as dificuldades de tratamento de dados.

Deve-se salientar que as duas (2) primeiras etapas foram propostas pelo autor deste trabalho, enquanto as demais etapas (processamento, transformação e construção do modelo) foram acordadas no processo de *KDD*, pois foi a abordagem que apareceu em maior frequência nos trabalhos retornados na RSL, como visto na Subseção 3.3.2. Assim, para cada uma dessas etapas foram definidas atividades ou subetapas, sendo essas definidas com base também na RSL. O objetivo de definir subetapas, foi facilitar a resolução de um problema maior em subproblemas menores. Nas próximas seções, as etapas da abordagem proposta são apresentadas em mais detalhes.

4.1 Atividades de Processamento

Sendo assim, definida a base de dados que será utilizada, principia-se a etapa de processamento, em que se pode aumentar a qualidade do banco de dados, por meio da checagem e da correção dos dados. A checagem dos dados é uma atividade mutável que tem como finalidade revelar os atributos com valores nulos, raros e impossíveis. Recursos visuais como gráficos, implementações de código em uma linguagem de programação e consultas em *Structured Query Language* (SQL) a partir de um banco de dados, dentre outras ferramentas, fornecem aos especialistas subsídios de detectar, de maneira eficiente, falhas em repositórios de dados de grandes volumes.

Recomenda-se como passo inicial nesta atividade (Checagem dos dados), averiguar-se a existência ou não de valores nulos, uma vez que é o procedimento mais trivial pela facilidade de reconhecer quando acontece esse tipo de problema. Em seguida, indica-se verificar a presença ou não de valores raros, esse procedimento pode ser realizado por meio

da utilização da tabela de frequência absoluta, em que o propósito é verificar dados que se repete em menor frequência. E, por fim, realiza-se a certificação da existência de valores impossíveis. Neste ponto é necessário identificar o conjunto de variáveis dependentes e independentes, com as variáveis dependentes, verifica-se a correlação destas com as independentes. De maneira geral, para realização de cada procedimento desta atividade o especialista deve ter total conhecimento do domínio de aplicação dos dados e como está disposta a base de dados a ser tratada.

Na correção dos dados, o especialista toma suas decisões sobre os atributos, de acordo com a situação detectada na atividade anterior. Ele pode resolver desconsiderar o atributo suspeito, corrigir manualmente atributos, formatar casas decimais, definir uma constante global (média ou mediana), coletar informações necessárias para modelar ou estimar ruído (e.g., regressão ou inferência), dentre outros possíveis procedimentos (GARCÍA *et al.*, 2016) .

4.2 Atividades de Transformação

Com os dados processados, inicia-se a etapa de transformação cujo objetivo é aumentar a performance de cada algoritmo. Uma agregação é realizada com o intuito de reduzir o conjunto de dados e aumentar o desempenho na construção de modelos de predição. A redução é definida em concordância com a necessidade do especialista: por exemplo, se for considerado a elaboração de um modelo de previsão para definir causas da evasão escolar, pode-se transformar os dados de renda em faixas salariais, o que indubitavelmente reduziria substancialmente o conjunto de dados e aumentaria a performance na execução dos algoritmos, uma vez que diversos registros seriam unidos em poucas faixas.

Algoritmos de MD, em geral, constroem modelos preditivos de formas diferentes e podem ter um melhor comportamento e exatidão nos resultados, se executados em um conjunto de entrada apropriado: por exemplo, um método pode construir um modelo para predição mais acurado (ou mais rápido), se um atributo salário for representado na forma nominal “mil reais”, em vez da forma numérica mil (1000) (GARCÍA *et al.*, 2016) . Diante do exibido, o especialista pode estabelecer diferentes maneiras para simbolizar alguns dos atributos processados sem afetar a integridade deles e usá-los como entrada para os algoritmos na investida de conseguir melhores resultados.

Antes de iniciar a construção dos modelos, propõe-se normalizar todos os atributos em uma escala de dados para simplificar e acelerar a execução dos algoritmos. A normalização lida com a mudança e padronização da dimensão de escalas dos dados, assim sendo, atributos podem ser normalizados com casas decimais deliberadas após a vírgula ou mecanismos que definem uma única escala para todos os atributos (e.g., de 0 a 1)

(GARCÍA *et al.*, 2016) . Em ambos os casos, a normalização pode aumentar o desempenho dos modelos, posto que permite ao algoritmo executar de forma mais simples omitindo cálculos extensos.

4.3 Atividades da Construção do Modelo

A partir dos dados transformados, pode-se executar os algoritmos automáticos (ou semiautomáticos) e congruentes para predição da possibilidade de um aluno evadir. Os padrões de evasão escolar são diversos tornando difícil definir um único método para uma predição. Por isso, profusos algoritmos devem ser examinados no desenvolvimento de modelos de predição. Desta forma, dispõe-se quatro (4) atividades essenciais para auferir resultados plausíveis: análise dos atributos de entrada, definição do algoritmo, ajuste do algoritmo e validação.

Na análise de atributos se indica detectar uma considerável coleção de entrada para os algoritmos. Atributos em demasia, insignificantes na entrada do modelo afligem os seus resultados de acerto e performance. Então, pode-se detectar nessa análise que um determinado algoritmo poderia executar em poucos segundos e conseguir melhores resultados de predição se fosse combinado em sua entrada somente seis (6) atributos ao invés de doze (12), por exemplo. Efetivamente, precisa-se analisar e delinear um relevante conjunto de entrada para cada algoritmo, de tal modo a atingir uma verticalidade entre desempenho e acurácia nos resultados.

A definição do algoritmo consiste em estabelecer um algoritmo adequado ou o mais acurado possível para um conjunto de dados previamente definido. É importante salientar que não existe um algoritmo ideal para todos os problemas, sendo assim, é possível que nessa atividade seja necessário uma série de testes para decidir qual se adequará melhor. Como passo inicial para decidir por qual algoritmo utilizar, indica-se analisar na base de dados a dimensão e a quantidade de dados, visto que o desempenho dos algoritmos está condicionado as características das bases de dados. Por exemplo, um determinado algoritmo pode se comportar melhor com uma base que tem uma quantidade demasiada de variáveis (SVM), enquanto isso, outro algoritmo pode funcionar melhor com uma quantidade de variáveis reduzida (MLP), por exemplo.

Quanto aos testes dos algoritmos, orienta-se dividir o conjunto global de dados em dois (2) conjuntos, o primeiro conjunto se trata dos dados de treino, em que o propósito é treinar o modelo preditor. Enquanto isso, o segundo conjunto é destinado aos testes propriamente dito, em que se averigua a precisão de acerto dos modelos. Ambos os conjuntos são definidos de forma aleatório, empregando-se uma função responsável por tal procedimento (SILVA; SPATTI; FLAUZINO, 2010).

Definida a coleção de dados, pode-se ajustar os parâmetros de cada algoritmo e

analisar os resultados. O ajuste dos algoritmos depende de uma série de fatores, como o tamanho da amostra de dados, projeções dos sinais e a quantidade de variáveis. Compete ao especialista executar uma série de testes com o propósito de detectar um plano adequado de configuração nas medidas para cada um dos algoritmos. Uma possibilidade de teste é examinar a acurácia do modelo com todos os atributos iniciais e, logo em seguida, fazer o mesmo processo anterior com os melhores atributos, esses atributos podem ser definidos, por exemplo, com a Lei de Entropia (GARCÍA *et al.*, 2016) .

Posteriormente à execução dos algoritmos, um modelo de predição é concebido e deve ser validado por um especialista, o modelo de predição mencionado anteriormente é o resultado de todo o processo utilizado para desvendar as causas da evasão escolar, ou seja, desde à formulação de questões até a validação por meio das técnicas estatísticas. Os resultados da predição de evasão escolar podem ser avaliados por muitas métricas estatísticas (*recall*, *f1-score*, precisão e etc.) que especificam uma representação numérica para o erro. Destaca-se na literatura, sobre critérios mais convenientes para validar modelos de predição: Willmott e Matsuura (2005); Chai e Draxler (2014); Wang e Bovik (2009); Goodwin e Lawton (1999); Tayman e Swanson (1999). Conquanto, ainda não existe uma única abordagem global para avaliar modelos de predição, em virtude que cada métrica municiona uma diferente visão do erro para o especialista.

5 Validação da Abordagem

Com o intuito de validar a abordagem de descoberta de conhecimento proposta no Capítulo 4, foi realizado um estudo de caso com dados de ex-alunos (concludentes ou evadidos) do curso de CC da UFERSA Campus Mossoró/RN. Salienta-se que foram contempladas todas as etapas da abordagem, desde o levantamento das variáveis a partir de questões exploradas no questionário (Apêndice A) até a validação dos modelos de MD empregados. O detalhamento completo das atividades realizadas e os resultados obtidos com a aplicação dos modelos são apresentados a seguir.

Como definido na abordagem apresentada na Figura 4, a primeira etapa da abordagem consiste em definir quais fatores indutores para evasão escolar devem ser abordados na pesquisa, sendo assim, foram definidas variáveis que abrangessem fatores das três (3) categorias possíveis: características do aluno, internos à instituição e externos à instituição. Na Tabela 12 são apresentados os fatores explorados neste estudo, em que foram agrupados de acordo com a categoria à qual pertencem. Logo em seguida se apresenta uma descrição de cada fator estudado.

Na segunda etapa, definiu-se a forma de coleta ou levantamento dos dados, decidiu-se pelo questionário, uma vez que foi o meio mais fácil de levantamento para o contexto do estudo de caso. O questionário foi aplicado de forma *online* no período que compreendeu 01/04/2019 à 30/04/2019, enfatiza-se que a instituição disponibilizou os *emails* dos ex-alunos para que fosse feito o contato. Sendo assim, destaca-se que foram obtidas respostas de cem (100) ex-alunos, entre as quais cinquenta e duas (52) foram de evadidos e quarenta e oito (48) de concludentes.

Uma vez que os dados foram levantados, iniciou-se as atividades (checagem dos dados e correção dos dados) da terceira etapa da abordagem, a etapa de processamento. Vale destacar, que todas as atividades realizadas neste estudo de caso estão disponíveis no *GitHub* ¹ deste trabalho, evidenciando que foi utilizada a linguagem *Python* e suas bibliotecas (*pandas*, *matplotlib* e *sklearn*) para a realização das atividades, pois, como visto no Capítulo 3 foi uma das ferramentas resultantes da RSL e que de acordo com os critérios de avaliação o autor deste trabalho juntamente com outros especialistas (orientadores), julgaram a mais adequada para este trabalho.

Na atividade de checagem dos dados, tratou-se de buscar quais variáveis apresentavam valores nulos, raros ou impossíveis. Constatou-se, que oito (8) variáveis apresentavam pelo menos um dos problemas supracitado (*situação_ex-aluno*, *ano_ingresso*, *ano_saída*, *experiência_prévia*, *instituição_estudava_ensino_médio*, *tipo_instituição_estudava_en-*

¹ <https://github.com/AnaliseDeDadosUfersa/AnalisedeDados>

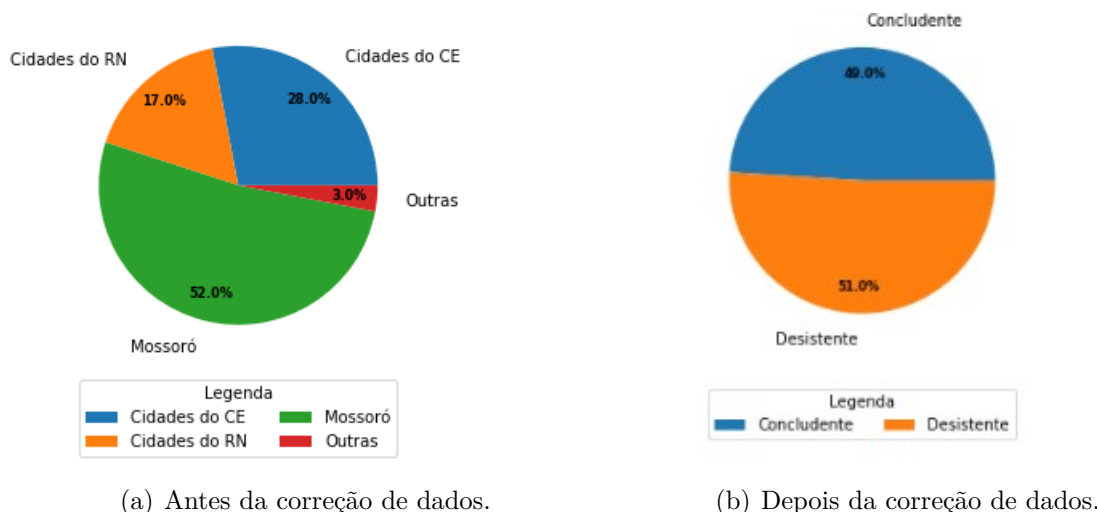
Tabela 12 – Fatores e variáveis exploradas e respectivas descrições.

Fatores	Variáveis	Descrição
Características individuais do aluno	situação_ex-aluno	a situação do ex-aluno (concludente ou evadido), no momento em que respondeu o questionário;
Características individuais do aluno	ano_ingresso	o ano de ingresso do aluno no curso;
Características individuais do aluno	ano_saída	o ano que o aluno concluiu ou evadiu do curso;
Características individuais do aluno	experiência_prévia	se o aluno tinham experiências na área do curso anterior ao ingresso no curso;
Características individuais do aluno	instituição_estudava_ensino_médio	a instituição que o aluno estudava no ensino médio (Pública ou Privada);
Características individuais do aluno	tipo_instituição_estudava_ensino_médio	o tipo de escola pública que os alunos estudavam, ex: escola militar, institutos federais, etc;
Características individuais do aluno	cidade_morava	cidade que o aluno morava antes de ingressar no curso;
Características individuais do aluno	mudou_cidade_instituição	se o aluno se mudou ou não para a cidade do curso;
Características individuais do aluno	Renda_Mensal_Familiar	renda familiar do aluno ao entrar na universidade;
Características individuais do aluno	pertence_qual_raça	a raça que o aluno pertence, ex: branco, negro, pardo, etc;
Características individuais do aluno	aproveitamento_acadêmico	ao aproveitamento acadêmico do aluno ao final do primeiro semestre;
Características individuais do aluno	relação_pais	o estado civil dos pais dos alunos durante a graduação (casados ou separados);
Características individuais do aluno	grau_instrução_mãe	o grau de instrução da mãe do aluno;
Características individuais do aluno	grau_instrução_pai	o grau de instrução do pai do aluno;
Externos à instituição	morou_vila_acadêmica	se o aluno chegou à morar ou não na vila acadêmica durante a graduação;
Externos à instituição	obteve_bolsa	o aluno ganhou ou não bolsa durante a graduação;
Externos à instituição	acompanhamento_instituição	o aluno teve ou não o devido acompanhamento da instituição, ex: coordenação de curso, assistência social, etc;
Externos à instituição	grade_curricular	a adequação ou não da grade curricular do curso;
Externos à instituição	cadeia_pré_requisitos	se a cadeia de pré-requisitos das disciplinas são adequados;
Externos à instituição	critérios_avaliação	os critérios de avaliação utilizados pelos professores;
Externos à instituição	infraestrutura_instituição	a adequação ou não da infraestrutura da instituição para a realização das atividades do curso;
Externos à instituição	qualidade_professores	a qualidade dos professores do curso;
Internos à instituição	carreiras_relacionadas_curso	as carreiras relacionadas ao curso, são socialmente reconhecidas;
Internos à instituição	remuneração_profissionais	os profissionais formados no curso, são pagos adequadamente no mercado de trabalho.

sino_médio, pertence_qual_raça e relação_pais); essas variáveis tiveram que ser corrigidas na atividade de correção dos dados. Dentre outras manipulações, um exemplo, apresenta-se na Figura 5 com a variável “Situação_Ex-Aluno”, salientando que na Figura 5(a) são

apresentados os dados antes da correção e na Figura 5(b) o resultado após a correção de dados.

Figura 5 – Correção de dados da variável “Situação_Ex-Aluno”.



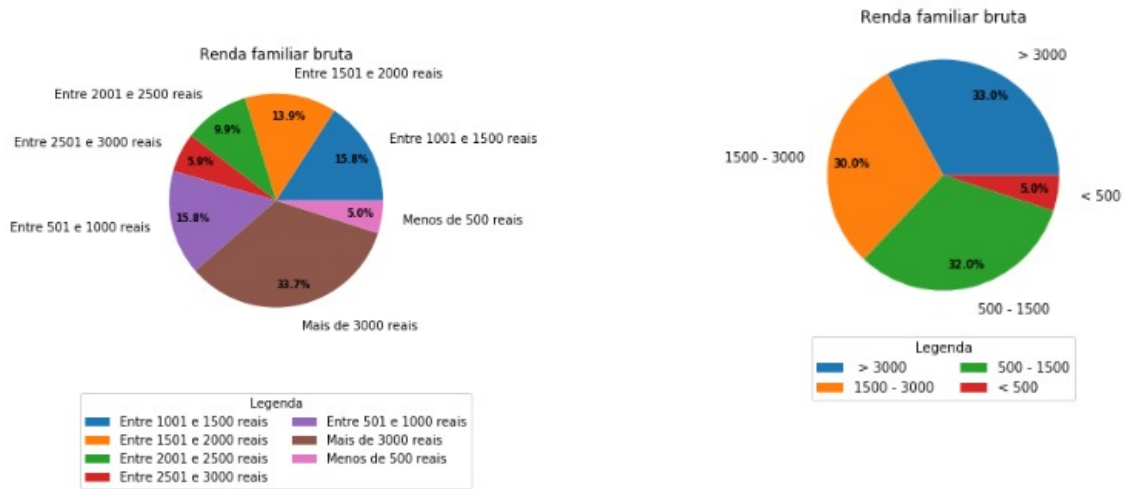
Fonte: Autoria Própria (2020).

Dado que as variáveis em que se observou problemas foram corrigidas, iniciou-se as atividades (agregação de dados e normalização) da quarta etapa do processo, a etapa de transformação. Na atividade de agregação de dados, percebeu-se que era conveniente agregar dados em oito (8) variáveis (Ano_Ingresso, Ano_Saída, Tipo_Instituição_Estudava_Ensino_Médio, Cidade_Morava, Renda_Mensal_Familiar, Aproveitamento_Acadêmico, Grau_Instrução_Mãe e Grau_Instrução_Pai); pois, essas variáveis tinham muitos dados dispersos, que diminui substancialmente a eficiência dos algoritmos. Na Figura 6, apresenta-se um exemplo de agregação com o fator “Renda_Mensal”, em que é mostrado um panorama dos dados brutos na Figura 6(a), enquanto na Figura 6(b) se apresenta o resultado após a agregação de dados.

Após o tratamento adequado dos dados na atividade de agregação de dados, a última atividade da etapa de transformação, consistiu em normalizar todos as variáveis em uma escala de dados para simplificar e acelerar a execução dos algoritmos. Dessa forma, na Figura 7 é possível observar um exemplo das manipulações que foram realizadas nesta atividade. Mostra-se na Figura 7(a) a variável “Cidade_Morava” antes da normalização e na Figura 7(b) o resultado após à normalização de dados.

A partir das manipulações realizadas até então, a etapa seguinte consistiu na construção do modelo preditivo. Neste caso, a primeira atividade foi definir uma coleção de variáveis de entrada para os algoritmos. Assim, utilizou-se a Lei da Entropia para definir os melhores atributos, esta lei significa que toda a energia em um sistema eremítico se move do estado ordenado para o estado desordenado (RIFKIN, 1980). Dessa forma, o estado mais ordenado, em que a concentração é maior, é visto o estado da entropia mínima.

Figura 6 – Agregação de dados da variável “Renda_Mensal”.

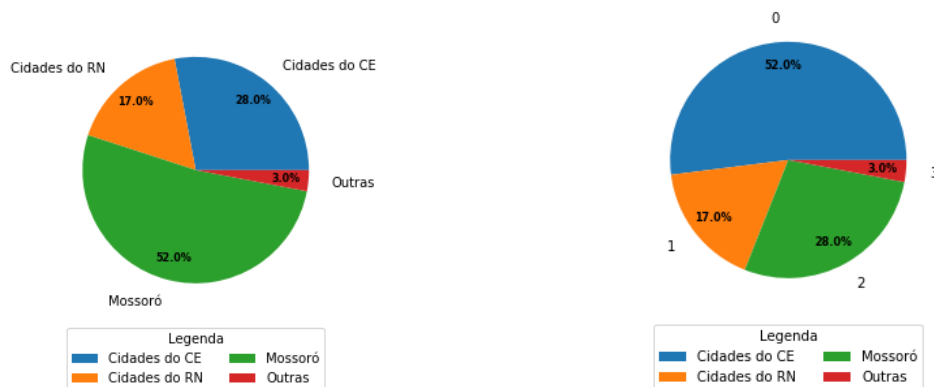


(a) Antes da agregação de dados.

(b) Depois da agregação de dados.

Fonte: Autoria Própria (2020).

Figura 7 – Normalização de dados da variável “Cidade_Morava”.



(a) Antes da normalização de dados.

(b) Depois da normalização de dados.

Fonte: Autoria Própria (2020).

Em contrapartida, o estado mais desordenado é considerado o estado da entropia máxima (RIFKIN, 1980). Assim, a noção de entropia está ligada ao grau de desorganização existente na fonte (podendo ser chamado na estatística de dispersão, dados dispersos/espalhados).

Apesar da Lei de Entropia ser utilizada para medir o grau de desordem de um sistema, neste trabalho a utilização dessa lei foi para determinar o ganho de informação de determinada variável em relação a variável *target*. Sendo assim, o resultado da Tabela 13 não é referente ao grau de entropia de cada variável, mas sim do ganho substancial de informação para inferência da desordem da variável *target*. De forma resumida representa o ganho de informação de uma variável para encontrar o valor real da *target*, isso significa que os sistemas de maior grau de ganho de informação são mais constituintes nos dois (2) grupos de concludentes e evadidos.

Tabela 13 – Aplicação da Lei de Entropia para verificação do ganho de informação das variáveis.

Índice	Variáveis	Ganho de Informação
1	ano_ingresso	0.417533
2	ano_saída	0.379427
3	experiência_prévia	0.005532
4	instituição_estudava_ensino_médio	0.108663
5	tipo_instituição_estudava_ensino_médio	0.205925
6	cidade_morava	0.149580
7	mudou_cidade_instituição	0.023541
8	morou_vila_acadêmica	0.023422
9	renda_mensal_familiar	0.402667
10	pertence_qual_raça	0.106036
11	obteve_bolsa	0.210807
12	aproveitamento_acadêmico	0.436557
13	acompanhamento_instituição	0.071961
14	grade_curricular	0.026584
15	cadeia_pré-requisitos	0.060191
16	critérios_avaliação	0.042411
17	infraestrutura_instituição	0.029666
18	qualidade_professores	0.053293
19	carreiras_relacionadas_curso	0.017665
20	Remuneração_Profissionais	0.034771
21	relação_pais	0.055591
22	grau_instrução_mãe	0.197992
23	grau_instrução_pai	0.179309

O resultado da aplicação da Lei da Entropia pra cada variável deste estudo pode ser verificado na Tabela 13. Considerou-se que as melhores variáveis são aqueles que obtiveram ganho de informação igual ou superior a zero vírgula zero dois (0,02), nesse caso se entende que as variáveis (Experiência_Prévia e Carreiras_Relacionadas_Curso) devem ser desconsideradas para o estudo, pois, ambas contribuem minimamente para o modelo e podendo até diminuir a acurácia do modelo.

O passo seguinte foi definir os algoritmos, como visto no Capítulo 3 a técnica mais utilizada no estudo das causas de evasão escolar, foi a classificação. Dessa forma, neste trabalho os algoritmos de classificação utilizados, foram: *Naive Bayes*, *J48*, *KNN*, *SVM* e *Adaboost*; tais algoritmos foram escolhidos com base nos resultados da RSL. Em seguida, tratou-se de ajustar as variáveis resultantes da atividade “definição de variáveis” em cada um dos algoritmos. Nas subseções a seguir, são apresentados os resultados obtidos em cada um dos algoritmos escolhidos.

5.1 Naive Bayes

Nesse primeiro caso foram utilizadas quatro (4) variações da coleção de classificadores *Naive Bayes* baseados no Teorema de *Bayes* (NARAYANAN *et al.*, 2013). Vale ressaltar, que todos compartilham um princípio comum, ou seja, cada par de características sendo classificadas é independente uma da outra. Neste trabalho, considerou-se as variações que estão disponíveis na biblioteca *sklearn* do *Python* (*GaussianNB*, *BernoulliNB*, *MultinomialNB* e *ComplementNB*):

- O *GaussianNB* é um tipo especial de algoritmo *naive bayes*, empregado especificamente quando os recursos possuem valores contínuos. Também se presume que todos os recursos estão seguindo uma distribuição Gaussiana, isto é, distribuição normal (ACOSTA, 2019);
- O *BernoulliNB* é utilizado para dados distribuídos de acordo com as distribuições multivariadas de *Bernoulli*; isto é, pode haver vários recursos, mas cada um é assumido como sendo uma variável de valor binário (booleano) (ACOSTA, 2019);
- O *MultinomialNB* é uma das duas variantes clássicas ingênuas de *Bayes* usadas na classificação de texto, onde os dados são tipicamente representados como contagens vetoriais de palavras (ACOSTA, 2019);
- O *ComplementNB* é uma adaptação do algoritmo padrão *MultinomialNB* que é particularmente adequado para conjuntos de dados desequilibrados. Especificamente, o *ComplementNB* usa estatísticas do complemento de cada classe para calcular os pesos do modelo (ACOSTA, 2019).

Tabela 14 – Características obtidas com os modelos *Naive Bayes*.

Variáveis	GaussianNB Características dos evadidos	GaussianNB Características dos concludentes	BernoulliNB Características dos evadidos	BernoulliNB Características dos concludentes	MultinomialNB Características dos evadidos	MultinomialNB Características dos concludentes	ComplementNB Características dos evadidos	ComplementNB Características dos concludentes
ano_ingresso	2006 - 2007	2014 - 2015	2006 - 2007	2014 - 2015	2006 - 2007	2014 - 2015	2006 - 2007	2014 - 2015
ano_saida	2010 - 2011	2006 - 2007	2014 - 2015	2006 - 2007	2010 - 2011	2006 - 2007	2010 - 2011	2006 - 2007
instituicao_estudava_ensino_medio	Pública	Privada	Pública	Privada	Pública	Privada	Pública	Privada
tipo_instituicao_estudava_ensino_medio	Pública	Privada	Pública	Privada	Pública	Privada	Pública	Privada
cidade_morava	Cidades do CE	Cidades do RN	Cidades do CE	Outras	Cidades do CE	Mossoró	Cidades do CE	Mossoró
mudou_cidade_instituicao	Sim	Não	Sim	Não	Sim	Não	Sim	Não
morou_vila_academica	Não	Sim	Sim	Não	Não	Sim	Não	Sim
renda_mensal_familiar	Menos que 500	Mais que 3000	Menos que 500	Mais que 3000	Menos que 500	Mais que 3000	Menos que 500	Mais que 3000
pertence_qual_raça	Outros	branco	Pardo ou Negro	branco	Outros	branco	Pardo ou Negro	branco
obteve_bolsa	Sim	Não	Não	Sim	Não	Sim	Não	Sim
aproveitamento_academico	Reprovado em até duas disciplinas	Aprovado em todas	Reprovado em até duas disciplinas	Aprovado em todas	Reprovado em até duas disciplinas	Aprovado em todas	Reprovado em até duas disciplinas	Aprovado em todas
acompanhamento_instituicao	Não	Sim	Não	Sim	Não	Sim	Não	Sim
grade_curricular	Não	Sim	Não	Sim	Não	Sim	Não	Sim
cadeia_pre-requisitos	Não	Sim	Não	Sim	Não	Sim	Não	Sim
critérios_avaliação	Não	Sim	Não	Sim	Não	Sim	Não	Sim
infraestrutura_instituicao	Não	Sim	Não	Sim	Não	Sim	Não	Sim
qualidade_professores	Sim	Não	Sim	Não	Sim	Não	Sim	Não
remuneração_profissionais	Não	Sim	Não	Sim	Não	Sim	Não	Sim
relação_pais	Não	Sim	Não	Sim	Não	Sim	Não	Sim
grau_instrução_mãe	Nível Fundamental	Nível Superior	Nível Fundamental	Nível Superior	Nível Fundamental	Nível Superior	Nível Fundamental	Nível Superior
grau_instrução_pai	Outros	Nível Superior	Outros	Nível Superior	Outras	Nível Superior	Outras	Nível Superior

Os dados foram ajustados em cada uma das variações supracitadas, na Tabela 14 é apresentado um panorama completo dos resultados obtidos para cada algoritmo, sendo enfatizado os perfis dos ex-alunos (evadidos e concludentes). Enquanto isso, apresenta-se

na Tabela 15 o desempenho obtido para cada algoritmo. Salienta-se, que foram utilizadas três (3) técnicas estatísticas disponíveis na biblioteca *sklearn*:

- *precisão*: refere-se a como estimativas próximas de amostras diferentes são para a outra. Por exemplo, o erro padrão é uma medida de precisão. Quando o erro padrão é pequeno, as estimativas de diferentes amostras terão um valor próximo; e vice-versa (BEAVIS, 1998);
- *recall*: refere-se ao número de verdadeiros positivos dividido pelo número de verdadeiros positivos mais o número de falsos negativos. Verdadeiros positivos são pontos de dados classificados como positivos pelo modelo que são realmente positivos (o que significa que estão corretos) e falsos negativos são pontos de dados que o modelo identifica como negativos que são realmente positivos (incorretos) (BEAVIS, 1998);
- *f1-score*: na análise estatística da classificação binária, o *f1-score* (também *f-score* ou *f-measure*) é uma medida da precisão de um teste. A pontuação *f1* pode ser interpretada como uma média ponderada da precisão e recuperação, onde uma pontuação *f1* atinge seu melhor valor em um (1) e pior em zero (0) (BEAVIS, 1998).

Neste trabalho, essas técnicas foram utilizadas para medir a precisão de acerto dos algoritmos escolhidos. A estratégia utilizada para medir quão preciso é um algoritmo, foi realizar a média aritmética das técnicas para cada algoritmo.

Tabela 15 – Validação dos modelos *Naive Bayes*.

Algoritmo	Concludente			Evadido			Média
	precisão	recall	f1-score	precisão	recall	f1-score	
GaussianNB	0.85	0.86	0.85	0.85	0.84	0.85	0.85
BernoulliNB	0.80	0.80	0.80	0.80	0.80	0.80	0.80
MultinomialNB	0.89	0.92	0.90	0.91	0.88	0.90	0.90
ComplementNB	0.88	0.88	0.88	0.88	0.88	0.88	0.88

Salienta-se que o conjunto global de dados foi dividido em dois (2) conjuntos de dados, conjunto de treino e conjunto de teste. No conjunto de treino foi colocado setenta por cento (70%) dos dados e no conjunto de teste trinta por cento (30%) dos dados. Observa-se na Tabela 15 que o *MultinomialNB* foi o algoritmo que teve o melhor desempenho com os dados de teste, obtendo noventa por cento (90%) de acertos, enquanto isso, o *BernoulliNB* acertou apenas em oitenta por cento (80%) dos casos. Quanto aos dados de treino, o *MultinomialNB* foi também o que obteve melhor desempenho acertando em torno de oitenta por cento (80%) e de maneira análoga o *BernoulliNB* foi o que obteve o pior desempenho nos testes acertando em apenas cinquenta e cinco por cento (55%) dos casos.

5.2 J48

O modelo *J48* é uma implementação do algoritmo *C4.5* desenvolvido por (QUINLAN, 2014). O *J48* se baseia na criação de uma árvore de decisão a partir do cálculo de entropia, determinando os nós mais acima (níveis superiores) pelo maior ganho de informação em relação aos nós mais abaixo (níveis inferiores) na árvore. Na Tabela 16 é apresentada as características de concludentes e evadidos obtidas a partir da aplicação do modelo *J48*.

Tabela 16 – Características obtidas com o modelo *J48*.

Variáveis	Características evadidos	Características concludentes
ano_ingresso	2014 - 2015	2006 - 2007
ano_saida	2006 - 2007	2010 - 2011
instituição_estudava_ensino_médio	Pública	Privada
tipo_instituição_estudava_ensino_médio	Pública	Privada
cidade_morava	Cidades do CE	Outras
mudou_cidade_instituição	Não	Sim
morou_vila_acadêmica	Não	Sim
renda_mensa_familiar	Menos de 500	Mais de 3000
pertence_qual_raça	Pardo ou Negro	Branco
obteve_bolsa	Não	Sim
aproveitamento_acadêmico	Reprovado em até duas disciplinas	Aprovado em todas
acompanhamento_instituição	Não	Sim
grade_curricular	Não	Não
cadeia_pré-requisitos	Não	Não
critérios_avaliação	Não	Não
infraestrutura_instituição	Não	Não
qualidade_professores	Não	Não
remuneração_profissionais	Não	Não
relação_pais	Não	Não
grau_instrução_mãe	Nível Médio	Nível Superior
grau_instrução_pai	Outros	Nível Superior

Tabela 17 – Validação dos modelos *J48*.

Algoritmo	Concludente			Evadido			Média
	precisão	recall	f1-score	precisão	recall	f1-score	
J48	0.96	0.90	0.93	0.91	0.96	0.93	0.93

Enquanto isso, na Tabela 17 é apresentado o resultado da validação do modelo *J48*, destaca-se que o desempenho do modelo ficou em torno de noventa e três por cento (93%) com os dados de teste. Salienta-se, que não houve uma diferença considerável entre o resultado de treino e teste, uma vez que, no treino o resultado foi por volta de noventa e três vírgula três por cento (93,3%).

5.3 KNN (*K-Nearest Neighbors*)

O modelo *KNN* se baseia em um cálculo que determina a proximidade (distância) de um determinado ponto em relação aos seus vizinhos mais próximos. Esse cálculo define

Tabela 18 – Características obtidas com o modelo *KNN*.

Variáveis	Características evadidos	Características concludentes
ano_ingresso	2014 - 2015	2006 - 2007
ano_saída	2006 - 2007	2010 - 2011
instituição_estudava_ensino_médio	Pública	Privada
tipo_instituição_estudava_ensino_médio	Pública	Privada
cidade_morava	Cidades do CE	Mossoró
mudou_cidade_instituição	Não	Sim
morou_vila_acadêmica	Não	Sim
renda_mensal_familiar	Menos que 500	Mais que 3000
pertence_qual_raça	Pardo ou Negro	Branco
obteve_bolsa	Não	Sim
aproveitamento_acadêmico	Reprovado em mais de cinco disciplinas	Aprovado em todas
acompanhamento_instituição	Não	Sim
grade_curricular	Não	Sim
cadeia_pré-requisitos	Não	Sim
critérios_avaliação	Não	Sim
infraestrutura_instituição	Não	Sim
qualidade_professores	Não	Sim
remuneração_profissionais	Não	Sim
relação_pais	Não	Sim
grau_instrução_mãe	Nível fundamental	Nível superior
grau_instrução_pai	outros	Nível superior

uma relação estatística estabelecendo a probabilidade do determinado ponto pertencer ao mesmo grupo que os seus vizinhos mais próximos (ZHANG *et al.*, 2017). Na Tabela 18 são apresentadas as características de concludentes e evadidos respectivamente, encontradas a partir do modelo *KNN*.

Tabela 19 – Validação dos modelos *KNN*.

Algoritmo	Concludente			Evadido			Média
	precisão	recall	f1-score	precisão	recall	f1-score	
KNN	0.92	0.92	0.92	0.92	0.92	0.92	0.92

Verifica-se na Tabela 19 o desempenho obtido pelo modelo *KNN*, destaca-se que o modelo em média acertou em noventa e dois por cento (92%) dos casos. Com isso, constata-se que houve uma variação negativa em relação ao desempenho do modelo no treinamento, uma vez que, no treinamento o desempenho ficou por volta de noventa e sete por cento (97%).

5.4 SVM (Support Vector Machine)

O funcionamento de uma *SVM* consiste no desenvolvimento de um modelo tomando as entradas de treinamento, mapeando-as no espaço multidimensional e utilizando regressão para encontrar um hiperplano (um hiperplano é uma superfície em espaço de n dimensões que o separa em duas metades de espaço) que melhor separa duas classes de entradas. Uma vez que a *SVM* tenha sido treinada, ela é capaz de avaliar novas entradas em relação ao hiperplano divisor e classificá-las em uma entre duas categorias (YE, 2019). Na Tabela

Tabela 20 – Características obtidas com o modelo *SVM*.

Variáveis	Características evadidos	Características concludentes
ano_ingresso	2014 - 2015	2006 - 2007
ano_saída	2006 - 2007	2010 - 2011
instituição_estudava_ensino_médio	Pública	Privada
tipo_instituição_estudava_ensino_médio	Pública	Privada
cidade_morava	Cidades do CE	Outras
mudou_cidade_instituição	Não	Sim
morou_vila_acadêmica	Não	Sim
renda_mensal_familiar	Menos que 500	Mais que 3000
pertence_qual_raça	Pardo ou negro	Branco
obteve_bolsa	Não	Sim
aproveitamento_acadêmico	Reprovado em até duas disciplinas	Aprovado em todas
acompanhamento_instituição	Não	Sim
grade_curricular	Não	Sim
cadeia_pré-requisitos	Não	Sim
critérios_avaliação	Não	Sim
infraestrutura_instituição	Não	Sim
qualidade_professores	Não	Sim
remuneração_profissionais	Não	Sim
relação_pais	Não	Sim
grau_instrução_mãe	Nível fundamental	Nível superior
grau_instrução_pai	Outros	Nível superior

20 se apresenta o detalhamento das características de concludentes e evadidos, obtidas a partir do modelo *SVM*.

Tabela 21 – Validação dos modelos *SVM*.

Algoritmo	Concludente			Evadido			Média
	precisão	recall	f1-score	precisão	recall	f1-score	
SVM	0.98	0.98	0.98	0.98	0.98	0.98	0.98

Na Tabela 21 é apresentada a validação do modelo *SVM*, salienta-se o excelente desempenho do modelo com noventa e oito por cento (98%) de acurácia em relação aos dados de testes. Vale destacar, que não houve uma grande variação negativa do resultado de teste comparado ao de treino, uma vez que, com os dados de treino o modelo conseguiu acertar em cem por cento (100%) dos casos.

5.5 AdaBoost

O *AdaBoost* é um classificador que se baseia em um cálculo similar ao cálculo do MLP. São definidas várias variáveis W_n que determinam os pesos dos modelos fracos contidos no modelo original, esses W_n são atribuídos a partir de predições sucessivas mediante seus erros. Os modelos fracos, como o próprio nome diz, são modelos que não conseguem por si só predizer com eficiência a classificação de um determinado dado. O *AdaBoost* determina um cálculo interno e soma as predições dos classificadores fracos multiplicados pelo seu peso W para determinar a saída Y (HUANG *et al.*, 2019).

Tabela 22 – Características obtidas com o modelo *AdaBoost*.

Variáveis	Características evadidos	Características concludentes
ano_ingresso	2014 - 2015	2006 - 2007
ano_saída	2006 - 2007	2016 - 2017
instituição_estudava_ensino_médio	Pública	Privada
tipo_instituição_estudava_ensino_médio	Pública	Privada
cidade_morava	Cidades do CE	Outras
mudou_cidade_instituição	Não	Sim
morou_vila_acadêmica	Não	Sim
renda_mensal_familiar	Menos que 500	Mais que 3000
pertence_qual_raça	Pardo ou negro	Branco
obteve_bolsa	Não	Sim
aproveitamento_acadêmico	Reprovado em até duas disciplinas	Aprovado em todas
acompanhamento_instituição	Não	Sim
grade_curricular	Não	Sim
cadeia_pré-requisitos	Não	Sim
critérios_avaliação	Não	Sim
infraestrutura_instituição	Não	Sim
qualidade_professores	Não	Sim
remuneração_profissionais	Não	Sim
relação_pais	Não	Sim
grau_instrução_mãe	Nível fundamental	Nível superior
grau_instrução_pai	Outros	Nível superior

Tabela 23 – Validação dos modelos *SVM*.

Algoritmo	Concludente			Evadido			Média
	precisão	recall	f1-score	precisão	recall	f1-score	
Adaboost	0.98	0.98	0.98	0.98	0.98	0.98	0.98

Apresenta-se, na Tabela 22, as características de concludentes e evadidos obtidas a partir do modelo *AdaBoost*. Enquanto isso, na Tabela 23 é apresentada a validação do modelo *AdaBoost* salientando que, assim como o modelo *SVM*, este modelo obteve noventa e oito por cento (98%) de acurácia com os dados de teste. Porém, a variação em relação ao resultado de treino foi positiva, uma vez que, no treino, o desempenho foi em torno de noventa e três por cento (93%).

5.6 Considerações Finais

A partir dos modelos *SVM* e *AdaBoost*, considerando que os mesmos se mostraram como os melhores na classificação de alunos que evadem e concluem, as principais inferências são:

- alunos que estudaram em escolas privadas no ensino médio, têm uma maior probabilidade de concluir o curso de CC da UFERSA Campus Mossoró/RN;
- alunos naturais do estado do RN, têm uma maior probabilidade de concluir o curso de CC da UFERSA Campus Mossoró/RN;

- alunos naturais de outras cidades que mudaram pra morar em Mossoró/RN, têm uma maior probabilidade de concluir o curso de CC da UFERSA Campus Mossoró/RN;
- alunos que moraram na vila acadêmica, têm uma maior probabilidade de concluir o curso de CC da UFERSA Campus Mossoró/RN;
- a renda mensal de alunos que concluíram o curso de CC da UFERSA Campus Mossoró/RN, é melhor em detrimento aos que evadiram;
- alunos que ganharam bolsa no decorrer do curso, têm uma maior probabilidade de concluir o curso de CC da UFERSA Campus Mossoró/RN;
- alunos que têm acompanhamento institucional, têm uma maior probabilidade de concluir o curso de CC da UFERSA Campus Mossoró/RN;
- alunos que concluíram o curso acreditam que os profissionais de Tecnologia da Informação (TI) são remunerados adequadamente;
- alunos que têm pais com nível superior, têm uma maior probabilidade de concluir o curso de CC da UFERSA Campus Mossoró/RN.

Sendo assim é possível traçar o seguinte perfil dos alunos que desistem do curso de CC da UFERSA Campus Mossoró/RN: são alunos que predominantemente estudaram em escolas públicas no ensino médio, naturais de outros estados, não mudaram pra morar em Mossoró/RN, não moraram na vila acadêmica, renda mensal da família baixa, não ganharam bolsa no decorrer do curso, não tiveram acompanhamento institucional, os mesmos acreditam que os profissionais não são remunerados adequadamente e geralmente alunos que tem pais com baixa escolaridade são mais propensos a evadir. Quanto aqueles que concluem, o perfil é o contrário daqueles que evadem.

6 Conclusões e Trabalhos Futuros

Visando uma maior compreensão do processo de evasão no ensino superior, no presente estudo foi proposto uma abordagem de descoberta de conhecimento que visa identificar os determinantes de evasão de alunos de cursos superiores. No que se refere aos determinantes responsáveis pela decisão de não permanência no curso, evidenciou-se que há um conjunto amplo de motivos associados à evasão, indicando que são variados os aspectos que atuam sobre os estudantes.

Com base na literatura e nas análises realizadas neste estudo, constatou-se que não existe um único e melhor algoritmo global para previsão da evasão escolar que possa ser aplicado em qualquer situação para fornecer os melhores resultados, devido ao fato dos padrões de evasão serem influenciados por muitos fatores. No entanto, existem algoritmos robustos aplicados na abordagem proposta que podem ser utilizadas em uma dada situação para fornecer resultados relevantes, tais como o *SVM* e o *AdaBoost*.

6.1 Sumário da Pesquisa

Em busca de conhecimentos relevantes, na problemática deste trabalho foram definidas questões de pesquisa que enriqueceram o seu desenvolvimento. As questões foram respondidas ao decorrer do trabalho, proporcionando alternativas de soluções para a problemática abordada através da aplicação de novos conhecimentos científicos, identificando o que realmente é eficaz para desvendar as causas da evasão escolar. De forma sintetizada, as respostas referentes às questões de pesquisa definidas no capítulo 1 foram:

Quais fatores podem influenciar no processo de evasão escolar?

A partir dos resultados obtidos com a RSL, desvendou-se que as causas da evasão estão associadas a três (3) grandes categorias: fatores relacionados às características individuais do aluno, fatores associados às instituições e fatores externos às instituições;

Quais técnicas de descoberta de conhecimento de dados podem ser utilizadas?

Buscou-se especificamente com esta questão, as ferramentas, algoritmos e abordagens de MD que até então vinham sendo utilizadas para desvendar as causas da evasão no ensino superior. Com a RSL, constatou-se que das ferramentas comumente utilizadas (*Weka*, *Mplus*, *SAS-STAT*, *HLM*, *Python*, *Software R* e *SPSS*), sendo *Weka* e o *Software R*, as ferramentas que obtiveram melhores avaliações, pois, ambas têm licença livre, oferecem documentação e são *Open Sources*.

Quanto aos algoritmos, destacam-se os principais: classificação, equações estruturais, regressão, análise de sobrevivência, análise fatorial, rede neural e análise multinível. É importante salientar que neste trabalho a técnica explorada foi classificação. E, por fim, a abordagem com maior destaque foi o KDD, uma vez que foi utilizada integralmente em um trabalho e parcialmente em diversos outros trabalhos.

Com que precisão a nova abordagem de descoberta de conhecimento em bases de dados é capaz de prever precocemente quais alunos podem evadir?

A abordagem foi testada com cinco (5) modelos diferentes, destacando-se que em dois (2) deles se alcançou a excelente acurácia de noventa e oito por cento (98%) de acertos.

6.2 Limitações

Este estudo trata da previsão de ações resultantes de resoluções e decisões de seres humanos. Assim, reconhece as limitações da metodologia e as possíveis falhas, dado que as previsões fora do determinismo completo sobre a evasão escolar podem ser resultado de um processo estocástico. Outra limitação é a quantidade de dados utilizados para definir os classificadores, portanto, considerando os resultados, é notório que a abordagem é inovadora para prever alunos no grupo de risco de abandono escolar no ensino superior, contribuindo com uma lacuna identificada nas produções da comunidade científica Nacional e Internacional.

6.3 Trabalhos Futuros

A partir do modelo gerado pelos algoritmos de MD, um sistema para alertar coordenadores de curso, docentes e os demais departamentos da instituição sobre os alunos que estão potencialmente em risco de falhar ou desistir pode ser implementado. Como exemplo de ação possível, propõe-se que, uma vez que os alunos sejam considerados de risco a evadir, eles sejam designados para um tutor, a fim de lhes fornecer apoio acadêmico e orientação para motivar e tentar evitar o insucesso escolar.

Finalmente, como próximo passo desta pesquisa, pretende-se realizar mais experimentos usando uma base de dados maior de alunos de Computação e também, utilizando dados de alunos de outros cursos, para testar se os mesmos resultados de acurácia dos algoritmos são obtidos com diferentes bases de dados. Como trabalho futuro, destaca-se o seguinte:

- adicionar novas variáveis que possibilite prever o fracasso do aluno o mais rápido possível, pois, quanto mais cedo melhor;
- propor estratégias para ajudar os alunos identificados dentro do grupo de risco.

6.4 Publicações

Apresenta-se a seguir, uma publicação fruto deste trabalho na Revista Novas Tecnologias na Educação (RENOTE):

Marques, Leonardo Torres, Angélica Félix De Castro, Bruno Torres Marques, Jesaias Carvalho Pereira Silva and Paulo Gabriel Gadelha Queiroz. "Mineração de dados auxiliando na descoberta das causas da evasão escolar: Um Mapeamento Sistemático da Literatura." **RENOTE-Revista Novas Tecnologias na Educação** 17, no. 3 (2019): 194-203.

Referências

- ACOSTA, F. A. Z. *Construcción de un dataset de noticias para el entrenamiento y evaluación de clasificadores automatizados*. Tese (Doutorado) — Telecomunicacion, 2019. Citado na página 52.
- ADELMAN, M. *et al.* Predicting school dropout with administrative data: new evidence from guatemala and honduras. *Education Economics*, Taylor & Francis, v. 26, n. 4, p. 356–372, 2018. Citado na página 15.
- ÁLVARES, L. O. C. *Algoritmos e Ferramentas de Descoberta de Conhecimento em Bancos de Dados Geográficos*. Tese (Doutorado) — Dissertation, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2003. Citado na página 21.
- ARCHAMBAULT, I. *et al.* Student engagement and its relationship with early high school dropout. *Journal of adolescence*, Elsevier, v. 32, n. 3, p. 651–670, 2009. Nenhuma citação no texto.
- BEAVIS, W. D. Qtl analyses: power, precision, and accuracy. *Molecular dissection of complex traits*, v. 1998, p. 145–162, 1998. Citado na página 53.
- BIOLCHINI, J. C. de A. *et al.* Scientific research ontology to support systematic review in software engineering. *Adv. Eng. Inform.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 21, n. 2, p. 133–151, abr. 2007. ISSN 1474-0346. Disponível em: <<http://dx.doi.org/10.1016/j.aei.2006.11.006>>. Nenhuma citação no texto.
- BOWERS, A. J. Grades and graduation: A longitudinal risk perspective to identify student dropouts. *The Journal of Educational Research*, Taylor & Francis, v. 103, n. 3, p. 191–207, 2010. Citado 3 vezes nas páginas 35, 37 e 39.
- BRASIL. *Diretrizes Curriculares Nacionais Gerais para a organização e o funcionamento dos cursos superiores de tecnologia*. 2009. Disponível em: <<http://portal.mec.gov.br/cne/arquivos/pdf/cp29.pdf>>. Acesso em: 30 Novembro 2018. Citado 2 vezes nas páginas 15 e 25.
- BRASIL, R. F. Avaliação da evasão em uma instituição de ensino superior do rio grande do norte. Universidade Federal Rural do Semi-Árido, 2015. Nenhuma citação no texto.
- CHAI, T.; DRAXLER, R. Arguments against avoiding rmse in the literature. *Geosci. Model Dev*, v. 7, 2014. Citado na página 46.
- DANTAS, Á. M. C. *et al.* Internet das coisas e aprendizagem colaborativa: Revisão sistemática da literatura. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2018. v. 29, n. 1, p. 278. Nenhuma citação no texto.
- EICHER, V.; STAERKLÉ, C.; CLÉMENCE, A. I want to quit education: A longitudinal study of stress and optimism as predictors of school dropout intention. *Journal of adolescence*, Elsevier, v. 37, n. 7, p. 1021–1030, 2014. Nenhuma citação no texto.

- FALL, A.-M.; ROBERTS, G. High school dropouts: Interactions between social context, self-perceptions, school engagement, and student dropout. *Journal of adolescence*, Elsevier, v. 35, n. 4, p. 787–798, 2012. Citado 4 vezes nas páginas 35, 37, 39 e 40.
- FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37, 2017. Nenhuma citação no texto.
- FELDENS, M. A.; CASTILHO, J. M. V. d. Engenharia da descoberta de conhecimento em bases de dados: estudo e aplicação na área de saúde. *Porto Alegre*, 2015. Citado na página 21.
- GAIOSO, N. d. L. O fenômeno da evasão escolar na educação superior no brasil. *Brasília, DF: Universidade Católica de Brasília*, 2005. Citado na página 15.
- GARCÍA, S.; LUENGO, J.; HERRERA, F. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge-Based Systems*, Elsevier, v. 98, p. 1–29, 2016. Nenhuma citação no texto.
- GOODWIN, P.; LAWTON, R. On the asymmetry of the symmetric mape. *International journal of forecasting*, Elsevier, v. 15, n. 4, p. 405–408, 1999. Citado na página 46.
- HOED, R. M. Análise da evasão em cursos superiores: o caso da evasão em cursos superiores da área de computação. *Brasília, DF: Universidade de Brasília*, 2016. Citado na página 15.
- HUANG, Q. *et al.* On combining biclustering mining and adaboost for breast tumor classification. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, 2019. Citado na página 56.
- JIMÉNEZ-GÓMEZ, M. Á. *et al.* Discovering clues to avoid middle school failure at early stages. In: ACM. *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. [S.l.], 2015. p. 300–304. Nenhuma citação no texto.
- KITCHENHAM, B. *Procedures for Performing Systematic Reviews*. Department of Computer Science, Keele University, UK, 2004. Citado na página 27.
- KOVANOVIĆ, V. *et al.* Penetrating the black box of time-on-task estimation. In: ACM. *Proceedings of the fifth international conference on learning analytics and knowledge*. [S.l.], 2015. p. 184–193. Citado na página 23.
- MANHÃES, L. M. B. *et al.* Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2011. v. 1, n. 1. Nenhuma citação no texto.
- MARQUES, L. T. *et al.* Mineração de dados auxiliando na descoberta das causas da evasão escolar: Um mapeamento sistemático da literatura. *RENTE-Revista Novas Tecnologias na Educação*, v. 17, n. 3, p. 194–203, 2019. Citado na página 24.
- MÁRQUEZ-VERA, C. *et al.* Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, Wiley Online Library, v. 33, n. 1, p. 107–124, 2016. Citado na página 15.

- MÁRQUEZ-VERA, C.; MORALES, C. R.; SOTO, S. V. Predicting school failure and dropout by using data mining techniques. *IEEE Revista Iberoamericana de Tecnologías del Aprendizaje*, IEEE, v. 8, n. 1, p. 7–14, 2013. Nenhuma citação no texto.
- MARTINHO, V. R. D. C.; NUNES, C.; MINUSSI, C. R. An intelligent system for prediction of school dropout risk group in higher education classroom based on artificial neural networks. In: IEEE. *Tools with Artificial Intelligence (ICTAI), 2013 IEEE 25th International Conference on*. [S.l.], 2013. p. 159–166. Nenhuma citação no texto.
- MEHTA, A. A.; BUCH, N. J. Depth and breadth of educational data mining: Researchers' point of view. In: IEEE. *Intelligent Systems and Control (ISCO), 2016 10th International Conference on*. [S.l.], 2016. p. 1–6. Citado na página 23.
- MINING, T. E. D. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. In: *Proceedings of conference on advanced technology for education*. [S.l.: s.n.], 2012. Citado na página 23.
- NAGAI, N. P.; CARDOSO, A. L. J. A evasão universitária: Uma análise além dos números. *Revista Estudo & Debate*, v. 24, n. 1, 2017. Citado na página 15.
- NARAYANAN, V.; ARORA, I.; BHATIA, A. Fast and accurate sentiment classification using an enhanced naive bayes model. In: SPRINGER. *International Conference on Intelligent Data Engineering and Automated Learning*. [S.l.], 2013. p. 194–201. Nenhuma citação no texto.
- PAPAMITSIOU, Z.; ECONOMIDES, A. A. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society*, JSTOR, v. 17, n. 4, 2014. Citado na página 23.
- PARR, A. K.; BONITZ, V. S. Role of family background, student behaviors, and school-related beliefs in predicting high school dropout. *The Journal of Educational Research*, Taylor & Francis, v. 108, n. 6, p. 504–514, 2015. Citado 3 vezes nas páginas 35, 37 e 39.
- QUINLAN, J. R. *C4. 5: programs for machine learning*. [S.l.]: Elsevier, 2014. Citado na página 54.
- RIFKIN, J. Entropy: a new world view.[social and political implications of the second law of thermodynamics]. Viking Press, New York, NY, 1980. Citado 2 vezes nas páginas 49 e 50.
- ROMERO, C.; VENTURA, S. Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, Elsevier, v. 33, n. 1, p. 135–146, 2007. Citado na página 15.
- ROMERO, C.; VENTURA, S. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Ieee, v. 40, n. 6, p. 601–618, 2010. Citado na página 15.
- ROVIRA, S.; PUERTAS, E.; IGUAL, L. Data-driven system to predict academic grades and dropout. *PLoS one*, Public Library of Science, v. 12, n. 2, p. e0171207, 2017. Nenhuma citação no texto.

- SABATES, R.; HOSSAIN, A.; LEWIN, K. M. School drop out in bangladesh: insights using panel data. *International Journal of Educational Development*, Elsevier, v. 33, n. 3, p. 225–232, 2013. Nenhuma citação no texto.
- SCALI, D. F. *et al.* Evasão nos cursos superiores de tecnologia: a percepção dos estudantes sobre seus determinantes. [sn], 2009. Nenhuma citação no texto.
- SILVA, I. N. D.; SPATTI, D. H.; FLAUZINO, R. A. Redes neurais artificiais para engenharia e ciências aplicadas curso prático. *São Paulo: Artliber*, 2010. Citado na página 45.
- TAYMAN, J.; SWANSON, D. A. On the validity of mape as a measure of population forecast accuracy. *Population Research and Policy Review*, Springer, v. 18, n. 4, p. 299–322, 1999. Citado na página 46.
- TINTO, V. Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, Sage Publications Sage CA: Thousand Oaks, CA, v. 45, n. 1, p. 89–125, 1975. Citado na página 24.
- TONTINI, G.; WALTER, S. A. Pode-se identificar a propensão e reduzir a evasão de alunos? ações estratégicas e resultados táticos para instituições de ensino superior. *Avaliação: Revista da Avaliação da Educação Superior*, SciELO Brasil, v. 19, n. 1, 2017. Citado na página 16.
- VENKATACHALAPATHY, K.; VIJAYALAKSHMI, V.; OHMPRAKASH, V. Educational data mining tools: A survey from 2001 to 2016. In: IEEE. *Recent Trends and Challenges in Computational Models (ICRTCCM), 2017 Second International Conference on*. [S.l.], 2017. p. 67–72. Nenhuma citação no texto.
- WANG, Z.; BOVIK, A. C. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, IEEE, v. 26, n. 1, p. 98–117, 2009. Citado na página 46.
- WEYBRIGHT, E. H. *et al.* Predicting secondary school dropout among south african adolescents: A survival analysis approach. *South African Journal of Education*, Education Association of South Africa, v. 37, n. 2, 2017. Citado na página 15.
- WILLMOTT, C. J.; MATSUURA, K. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, v. 30, n. 1, p. 79–82, 2005. Citado na página 46.
- WOOD, L. *et al.* Predicting dropout using student-and school-level factors: An ecological perspective. *School Psychology Quarterly*, Educational Publishing Foundation, v. 32, n. 1, p. 35, 2017. Citado na página 35.
- YE, K. Key feature recognition algorithm of network intrusion signal based on neural network and support vector machine. *Symmetry*, Multidisciplinary Digital Publishing Institute, v. 11, n. 3, p. 380, 2019. Citado na página 55.
- ZHANG, S. *et al.* Learning k for knn classification. *ACM Transactions on Intelligent Systems and Technology (TIST)*, ACM, v. 8, n. 3, p. 43, 2017. Citado na página 55.

Apêndices

APÊNDICE A – Questionário para extração de dados

1 - Qual a sua situação atual em relação ao curso de CC?

A. ☐ Concludente/Egresso

B. ☐ Evadido/Desistente

2 - Em qual ano e semestre você ingressou no curso de CC? Exemplo 2007.1

A. ☐ 2006.2

B. ☐ 2007.1

B. ☐ 2007.2

C. ☐ 2008.1

D. ☐ 2008.2

F. ☐ Outros, Qual?

3 - Em qual semestre você concluiu, desistiu ou encontra-se atualmente? Exemplo 2007.1

A. ☐ 2006.2

B. ☐ 2007.1

C. ☐ 2007.2

D. ☐ 2008.1

E. ☐ 2008.2

F. ☐ Outros, Qual?

4 - Antes de ingressar no curso você já tinha alguma experiência com a área?

A. ☐ Sim

B. ☐ Não

5 - Antes de ingressar na universidade, em que tipo de escola você estudava?

A. ☐ Pública

B. ☐ Privada

6 - Caso tenha estudado em escola pública, qual foi o tipo de instituição?

A. ☐ Instituto Federal

B. ☐ Escola Militar

C. ☐ Escola Estadual

D. ☐ Outros, Qual?

7 - Antes de entrar na universidade em qual cidade/estado você morava?
Exemplo Mombaça/CE

A. ☐ Mossoró/RN

B. ☐ Natal/RN

C. ☐ Russas/CE

D. ☐ Limoeiro do Norte/CE

E. ☐ Outros, Qual?

8 - Ao ingressar na universidade você se mudou para morar em Mossoró/RN?

A. ☐ Sim

B. ☐ Não

C. ☐ Eu já morava em Mossoró/RN

9 - Você chegou a morar na vila acadêmica?

A. ☐ Sim

B. ☐ Não

10 - Qual a renda mensal da sua família ao ingressar na universidade?

A. ☐ Menos de 500 reais

B. ☐ Entre 501 e 1000 reais

- C. ☐ Entre 1001 e 1500 reais
- D. ☐ Entre 1501 e 2000 reais
- E. ☐ Entre 2001 e 2500 reais
- F. ☐ Entre 2501 e 3000 reais
- G. ☐ Mais de 3000 reais

11 - A qual raça você se considera pertencer?

- A. ☐ Amarelo
- B. ☐ Branco
- C. ☐ Negro
- D. ☐ Pardo
- E. ☐ Prefiro não informar

12 - Você ganhou alguma bolsa durante o período que esteve no curso?

- A. ☐ Sim
- B. ☐ Não

13 - Em relação ao seu aproveitamento acadêmico, até o segundo semestre do curso você foi?

- A. ☐ Aprovado em todas as disciplinas
- B. ☐ Reprovado em uma disciplina
- C. ☐ Reprovado em duas disciplinas
- D. ☐ Reprovado em três disciplinas
- E. ☐ Reprovado em quatro disciplinas
- F. ☐ Reprovado em mais de cinco disciplinas

14 - Você considera que durante o período que esteve no curso teve o devido acompanhamento da universidade? Exemplo: PROAE ou Coordenação do curso.

- A. ☐ Sim
- B. ☐ Não

15 - Você considera que a grade curricular do curso de CC é adequado ao que se exige no mercado de trabalho?

A. ☐ Sim

B. ☐ Não

16 - Em relação a cadeia de pré-requisitos das disciplinas do curso, você considera adequado?

A. ☐ Sim

B. ☐ Não

17 - Em relação aos critérios de avaliação do desempenho do aluno, você considera adequado?

A. ☐ Sim

B. ☐ Não

18 - Em relação a infraestrutura da instituição como laboratórios, bibliotecas, equipamentos de informática e etc, você considera adequada?

A. ☐ Sim

B. ☐ Não

19 - Em relação aos professores, você considera que os mesmos apresentam deficiências pedagógicas, falta de interesse ou outros problemas?

A. ☐ Sim

B. ☐ Não

20 - Você considera que as carreiras relacionadas ao curso de CC são socialmente reconhecidas?

A. ☐ Sim

B. ☐ Não

21 - Em relação ao mercado, você considera que o profissional formado em CC é pago adequadamente?

A. ☐ Sim

B. ☐ Não

22 - Em relação aos seus pais, durante o período que você estava no curso, eles eram casados?

A. ☐ Sim

B. ☐ Não

23 - Em relação a sua mãe, qual o nível de escolaridade dela?

A. ☐ Ensino fundamental incompleto

B. ☐ Ensino fundamental Completo

C. ☐ Ensino médio incompleto

D. ☐ Ensino médio completo

E. ☐ Ensino superior incompleto

F. ☐ Ensino superior Completo

G. ☐ Pós-graduação incompleta

H. ☐ Pós-graduação completa

I. ☐ Ensino primário incompleto

24 - Em relação a seu pai, qual o nível de escolaridade dele?

A. ☐ Ensino fundamental incompleto

B. ☐ Ensino fundamental Completo

C. ☐ Ensino médio incompleto

D. ☐ Ensino médio completo

E. ☐ Ensino superior incompleto

F. ☐ Ensino superior Completo

G. ☐ Pós-graduação incompleta

H. ☐ Pós-graduação completa

I. ☐ Ensino primário incompleto