UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS QUIXADÁ

CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

ANTONIO ALVES DE LIMA JÚNIOR

MINERAÇÃO DE DADOS PARA IDENTIFICAÇÃO DO PERFIL DE EVASÃO DE ALUNOS DA UFC - CAMPUS QUIXADÁ.

QUIXADÁ

2019

ANTONIO ALVES DE LIMA JÚNIOR

MINERAÇÃO DE DADOS PARA IDENTIFICAÇÃO DO PERFIL DE EVASÃO DE ALUNOS DA UFC - CAMPUS QUIXADÁ.

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Sistemas de Informação do Campus Quixadá da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Regis Pires Magalhães

Coorientador: Prof. Tercio Jorge da Silva

QUIXADÁ

2019

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

L696m Lima Júnior, Antonio Alves de.
Mineração de dados para identificação do perfil de evasão de alunos da UFC - Campus Quixadá / Antonio Alves de Lima Júnior. – 2019.
37 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Sistemas de Informação, Quixadá, 2019.
Orientação: Prof. Dr. Regis Pires Magalhães.
Coorientação: Prof. Tercio Jorge da Silva.

1. Mineração de dados (Computação). 2. Evasão escolar. 3. Classificação. 4. Conhecimento. 5. Educação. I. Título.

CDD 005

ANTONIO ALVES DE LIMA JÚNIOR

MINERAÇÃO DE DADOS PARA IDENTIFICAÇÃO DO PERFIL DE EVASÃO DE ALUNOS DA UFC - CAMPUS QUIXADÁ.

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Sistemas de Informação do Campus Quixadá da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Sistemas de Informação.

Aprovada em: ____/____/_____

BANCA EXAMINADORA

Prof. Dr. Regis Pires Magalhães (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Tercio Jorge da Silva (Coorientador)
Universidade Federal do Ceará (UFC)

Profª. M.ª Lívia Almada Cruz Rafael
Universidade Federal do Ceará (UFC)

À minha família, por sua capacidade de acreditar em mim e investir em mim. Mãe, seu cuidado e dedicação foi que deram, em alguns momentos, a esperança para seguir. Pai, seu apoio significou segurança e certeza de que não estou sozinho nessa caminhada.

AGRADECIMENTOS

Ao Prof. Regis Pires Magalhães e ao meu coorientador Tércio Jorge da Silva, deixo um agradecimento especial pelo incentivo e pela dedicação do seu escasso tempo ao meu projeto de pesquisa.

Ao senhor Leonardo Torres Marques, por todo suporte e auxílio durante todas as fases do TCC.

Aos meus pais, Nilda e Antônio, que nos momentos de minha ausência dedicado aos estudos, sempre fizeram entender que o futuro é feito a partir da dedicação no presente, e sempre me deram forças para continuar, agradeço à minha namorada e companheira, Yasmim, que sempre esteve ao meu lado durante o meu percurso acadêmico.

Aos amigos de turma, Lucas Icety, João Paulo, Lucas Rodrigues e Nathan, por estarem junto comigo nessa jornada.

Agradeço a todos os professores, que me deram todo o suporte com suas correções e incentivos. Ao pessoal da biblioteca, meu muito obrigado.

E à Universidade Federal do Ceará (UFC), Campus de Quixadá, por todo suporte durante esses anos.

"Viver é estar continuamente motivado. O significado da vida não é simplesmente existir, sobreviver, mas sim crescer, alcançar e conquistar."

(Arnold Schwarzenegger)

RESUMO

Os altos índices de evasão nos cursos de graduação têm se tornado cada vez mais preocupantes no Brasil, esse problema tem gerado prejuízos tanto para o país, como para alunos e universidades. Nesse contexto, objetivou-se identificar os alunos com tendência a evasão escolar da UFC - Campus Quixadá/CE, por meio de técnicas de mineração dados e utilizando dados históricos de alunos, no qual foram realizados experimentos com dois cenários distintos, o primeiro cenário possuindo o número total de registros com a divisão dos registros por classes desbalanceadas e o segundo cenário contendo uma amostra dos registros com a divisão entre as classes balanceadas. Os resultados obtidos mostram que os potenciais alunos a evadir podem ser identificados com taxas de acerto de até 99% no primeiro cenário e no segundo cenário de até 95,5%. Por intermédio destes resultados, pretende-se auxiliar os gestores da instituição na tomada de decisão e na elaboração de políticas para mitigar à evasão escolar.

Palavras-chave: Mineração de dados educacionais. Evasão escolar. Classificação. Descoberta do conhecimento.

ABSTRACT

The high dropout rates in undergraduate courses have become increasingly worrying in Brazil, this problem has generated losses both for the country and for students and universities. In this context, the objective of this study is to identify students with a tendency to dropout from the UFC - Campus Quixadá/CE, by means of data mining techniques and using historical data of students, in which experiments were conducted with two distinct scenarios, the first scenario having the total number of records with the division of records by unbalanced class and the second scenario containing a sample of records with the division between balanced classes. The results obtained show that the potential students to escape can be identified with hit rates of up to 99% in the first scenario and in the second scenario of up to 95.5%. By means of these results, it is intended to help the managers of the institution in the decision making and in the elaboration of policies to mitigate school dropout.

Keywords: Educational Data Mining. School dropout. Classification. Knowledge Discovery.

LISTA DE FIGURAS

Figura 1 – Processo de KDD 18
Figura 2 – Árvore do algoritmo Tree Cenário 1 37
Figura 3 – Árvore do algoritmo Decision Tree Cenário 2 37

LISTA DE TABELAS

Tabela 1 – Divisão de registros por classe. 27
Tabela 2 – Acurácia, Precisão e Revocação dos Classificadores. 31
Tabela 3 – Acurácia, Precisão e Revocação dos Classificadores. 32

LISTA DE QUADROS

Quadro 1 – Comparação entre os trabalhos relacionados e o proposto. 26

LISTA DE ABREVIATURAS E SIGLAS

CC	Ciência da Computação
EDM	*Educational Data Mining*
ES	Engenharia de Software
GESMA	Grupo de Estudos em Engenharia de Software e Sistemas Multiagente
IES	Instituição de Ensino Superior
IRA	Índice de Rendimento Acadêmico
MD	Mineração de Dados
MDE	Mineração de Dados Educacionais
SGBD	Sistema de Gerenciamento de Banco de Dados
SI	Sistemas de Informação
SMA	Sistema Multiagente
TCC	Trabalho de Conclusão de Curso
TI	Tecnologia da Informação
UECE	Universidade Estadual do Ceará
UFC	Universidade Federal do Ceará
UFRJ	Universidade Federal do Rio de Janeiro

SUMÁRIO

1 INTRODUÇÃO 15
1.1 Objetivo 16
1.1.1 Objetivo Geral 16
1.1.2 Objetivos Específicos 16
1.2 Organização 16
2 FUNDAMENTAÇÃO TEÓRICA 17
2.1 Evasão Escolar 17
2.2 Descoberta de Conhecimento (KDD – Knowledge Discovery in Databases) 18
2.2.1 Seleção 19
2.2.2 Pré-Processamento e Limpeza 19
2.2.3 Transformação de Dados 20
2.2.4 Mineração dos Dados 20
2.2.5 Avaliação dos Dados 21
2.3 Mineração de Dados Educacionais 21
3 TRABALHOS RELACIONADOS 24
3.1 Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados 24
3.2 Auxiliando o Desempenho de Alunos com Tendência a Evasão na Educação a Distância Utilizando Técnicas de Mineração de Dados e Sistemas Multiagentes 24
3.3 Análise Estatística da Relação Entre Evasão e as Respostas do Questionário para Ingressantes da UFC-Quixadá 25
4 PROCEDIMENTOS METODOLÓGICOS 27
4.1 Levantamento de dados 27
4.2 Preparação dos dados 28
4.3 Aplicação dos algoritmos de MD 29
5 RESULTADOS 31
5.1 Resultados do Cenário 1 31
5.2 Resultados do Cenário 2 32
5.3 Perfis de alunos encontrados 33

6 CONCLUSÕES E TRABALHOS FUTUROS 34
REFERÊNCIAS 35
APÊNDICE A – ÁRVORES ANALISADAS PARA A IDENTIFICAÇÃO DOS PERFIS DOS ALUNOS 37

1 INTRODUÇÃO

O elevado índice de evasão escolar nas universidades é um problema que vem se tornando cada vez mais crítico no Brasil, atingindo inúmeras instituições, sejam de ensino público ou privado, gerando desperdício de recursos além de perdas sociais e acadêmicas. Segundo dados do Censo da Educação Superior, constatou-se que 49% dos alunos que ingressaram em 2010 no nível superior no Brasil desistiram no decorrer do curso. Nas instituições privadas, a evasão escolar atingiu 53%, e nas instituições públicas, chegou a 47% nas municipais, 38% nas estaduais e 43% nas federais (OLIVEIRA *et al.*, 2019). Diante desse contexto, é importante conhecer os fatores que influenciam os altos índices de evasão escolar, para que seja possível a tomada de medidas a fim de amenizá-la.

O problema é ainda mais crítico nos cursos da área da Tecnologia da Informação (TI). Segundo dados da secretaria acadêmica da Universidade Federal do Ceará (UFC) – Campus Quixadá/CE que é um polo voltado inteiramente a cursos da área de TI, constatou-se que entre o período de 2013.1 e 2019.1, 108 alunos saíram do curso de Ciência da Computação (CC), salientando que aproximadamente 80,6% são evadidos e apenas cerca de 19,4% são concludentes. No curso de Sistemas de Informação (SI), entre o período de 2007.2 e 2019.1, 358 alunos saíram do curso, enfatiza-se que aproximadamente 63,4% são evadidos e cerca de 36,6% são concluintes e no curso de Engenharia de Software (ES), entre o período de 2010.1 e 2019.1, 275 alunos saíram do curso de ES, dos quais aproximadamente 63,3% são evadidos e cerca de 36,7% são concluintes.

Existem vários fatores que podem motivar o aluno a evadir, segundo Barroso e Falcão (2004) as causas da evasão escolar são classificadas em três agrupamentos, que são: o abandono por inadequação aos métodos de estudo ou insucesso nas disciplinas (evasão institucional), impedimento de manutenção do vínculo por questões socioeconômicas (evasão econômica) e a compreensão de uma escolha de curso inapropriada aos interesses do aluno (evasão vocacional). Para mitigar a evasão escolar as instituições devem identificar as causas do fenômeno em seu ambiente educacional (NAGAI; CARDOSO, 2017).

Um método utilizado para o estudo das causas da evasão escolar, é o uso da descoberta de conhecimento em bases de dados, por meio de técnicas de Mineração de Dados (MD), denominado de *Educational Data Mining* (EDM) (ROMERO; VENTURA, 2007). A descoberta do conhecimento em dados educacionais será adotada neste trabalho, utilizando dados dos alunos do curso de CC da UFC – Campus Quixadá/CE, que é o curso com maior índice de evasão da

universidade. A técnica escolhida na etapa de MD é a classificação que busca reconhecer por meio do aprendizado de uma função a qual classe um determinado registro faz parte (CAMILO; SILVA, 2009).

Sendo assim, objetiva-se com este estudo identificar o perfil dos alunos que evadem dos cursos da UFC - Campus Quixadá/CE, por meio do uso de técnicas de MD. Será utilizada uma base de dados histórica para a descoberta de conhecimento. Salienta-se que os dados foram disponibilizados pela coordenação do curso de CC. Com isso, busca-se auxiliar os gestores da instituição na tomada de decisão e na elaboração de políticas para mitigar à evasão.

1.1 Objetivo

Nesta seção, são apresentados os objetivos deste trabalho.

1.1.1 Objetivo Geral

O objetivo geral desse trabalho foi identificar precocemente alunos com tendência a evadir dos cursos presencias da UFC Campus Quixadá, o intuito é auxiliar os gestores da instituição na tomada de decisão e na elaboração de políticas para mitigar à evasão escolar.

1.1.2 Objetivos Específicos

- Realizar a coleta de dados, por meio de reuniões com os coordenadores dos cursos;
- Executar a limpeza nos dados, removendo registros com campos nulos que impossibilitem a analise;
- Transformar os dados;
- Aplicar o algoritmo de classificação para auxiliar na descoberta de conhecimento;
- Interpretar os resultados obtidos.

1.2 Organização

O restante deste trabalho está organizado da seguinte forma: no Capítulo 2 é abordada a fundamentação teórica, no Capítulo 3 apresentam-se os trabalhos relacionados, no Capítulo 4 são apresentados os procedimentos metodológicos, no Capítulo 5 são detalhados os resultados e, por fim, no Capítulo 6 são mostradas as conclusões e trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo aborda os principais conceitos relacionados a este trabalho e qual a contribuição de cada conceito para o desenvolvimento do trabalho. Na Seção 2.1 são apresentados os conceitos da evasão escolar. Na Seção 2.2 é apresentada a definição de KDD (*Knowledge Discovery in Databases* ou Descoberta de Conhecimento nas Bases de Dados) e as etapas do processo. E por fim na Seção 2.3 são apresentados os conceitos fundamentais da MDE.

2.1 Evasão Escolar

A evasão escolar é um fenômeno complexo e comum às instituições universitárias no mundo contemporâneo. Ao longo dos anos, em razão de sua complexidade e abrangência, tem se tornado objeto de análises e estudos, especialmente nos países desenvolvidos, e tem mostrado tanto a universalidade do fenômeno como a relativa semelhança de comportamento em determinadas áreas do conhecimento, apesar das distinções entre as instituições de ensino e das particularidades socioeconômicos e culturais de cada país (VELOSO; ALMEIDA, 2013). A evasão escolar ocorre em vários graus e modalidades de ensino e tem causado prejuízos acadêmicos, econômicos, políticos, financeiros e sociais a todos os envolvidos no processo educacional, desde o aluno até os órgãos governamentais. (MARTINHO, 2014).

Neste contexto, em busca da compreensão desse fenômeno, vários pesquisadores tem desenvolvido modelos procurando explicar evasão escolar, no qual o conceito de evasão escolar é tratado de diferentes perspectivas.

Em Cardoso (2008), tem-se o conceito de evasão aparente, que seria a mobilidade do aluno, caracterizada por mudança de curso realizada dentro da própria Instituição de Ensino Superior (IES) ou mudança para outra IES, e a evasão real que seria pelo abandono definitivo do sistema de ensino por parte do estudante, e que pode ser gerado por motivos: financeiro, acadêmicos ou sociais.

De acordo com a Comissão Especial de Estudos Sobre Ministério (1997), criada pelo Ministério da Educação, a evasão escolar é definida nas seguintes classificações:

- **evasão de curso**: quando o aluno se desvincula do curso superior por várias situações tais como: abandono (deixar de se matricular), exclusão por norma institucional, reopção ou transferência (mudança de curso), desistência (oficial).
- **evasão da instituição**: quando o aluno rompe o vínculo com a instituição em que está

matriculado.

- **evasão do sistema**: quanto o aluno deixa de maneira definitiva ou temporária o ensino superior.

Em Silva *et al.* (2007) tem-se o conceito de evasão escolar compreendida sob dois aspectos semelhantes, porém não idênticos:

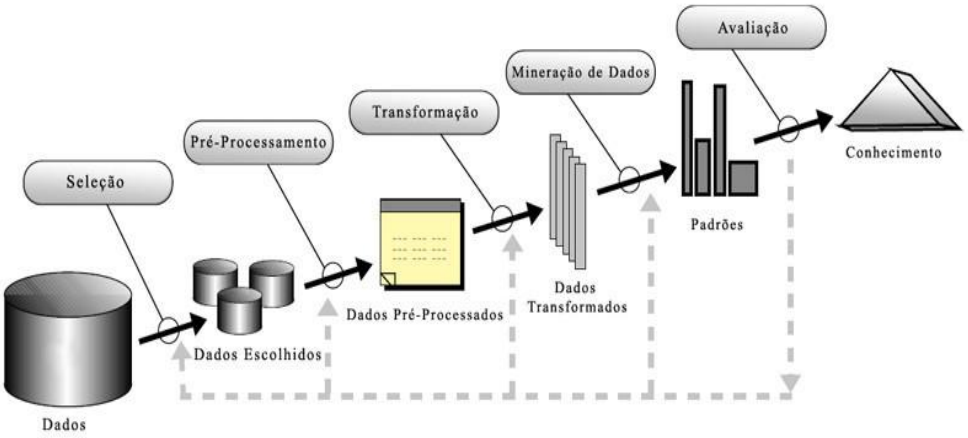
- evasão anual média que é o percentual de estudantes matriculados em uma IES, em um curso, ou em um sistema de ensino, que não se formaram e também não realizaram a matricula no ano ou no semestre seguinte.
- evasão total é dada pela quantidade de estudantes que ingressaram em um determinado curso, IES ou sistema de ensino e não obtiveram o diploma ao fim de uma determinada quantidade de tempo.

Neste estudo, considera-se evasão escolar a saída do aluno de uma IES ou de um de seus cursos de forma definitiva por algum dos motivos seguintes: desistência, transferência de curso ou de IES, deixar de se matricular e exclusão por norma institucional.

2.2 Descoberta de Conhecimento (KDD – Knowledge Discovery in Databases)

O termo Descoberta de Conhecimento em Banco de Dados (KDD – *Knowledge Discovery in Databases*) consiste em um processo não trivial de várias etapas, para identificação de padrões compreensíveis, validos e úteis em grandes conjuntos de dados (FAYYAD *et al.*, 1996). Na Figura 1 podemos ver as fases do processo de descoberta de conhecimento.

Figura 1 – Processo de KDD



Fonte: Fayyad *et al.* (1996)

O processo de KDD é caracterizado por ser interativo e iterativo, no qual envolve uma sequência de fases que devem ser executadas sequencialmente, de forma que, ao final de cada fase, o resultado atingido colabore para a fase seguinte, podendo haver repetições das fases anteriores quantas vezes forem necessárias. As fases do KDD são: seleção, pré-processamento e limpeza, transformação de dados, mineração de dados, interpretação e avaliação dos dados. A seguir cada subseção irá explorar sucintamente as fases do processo de KDD.

2.2.1 Seleção

A seleção dos dados é a fase inicial do processo de descobrimento de conhecimento. Nesta fase é escolhido o conjunto de dados, relacionado a um domínio, incluindo os possíveis atributos (também denominados de variáveis ou características) e registros (também denominados de observações ou casos) que vão fazer parte da análise (PRASS *et al.*, 2004).

Segundo Matos *et al.* (2017), o processo de seleção é considerado complexo, visto que os dados podem vir de várias fontes distintas (planilhas, sistemas legados, data warehouses) e podem possuir uma série de formatos (CSV, ARFF, TXT). Esta fase possui um impacto significante sobre a qualidade do resultado do processo (DUNKEL *et al.*, 1997).

2.2.2 Pré-Processamento e Limpeza

De acordo com Castanheira (2008), a limpeza dos dados inclui uma verificação da consistência das informações e o preenchimento ou a exclusão de valores repetidos e nulos. Nessa fase são identificados e removidos os dados duplicados e corrompidos. Uma boa limpeza dos dados é essencial para obter informações mais coerentes e diminuir o tempo de processamento.

Segundo Oliveira (2000), a fase de pré-processamento e limpeza envolve os seguintes aspectos:

- Padronização dos valores dos atributos: uma vez que um conjunto de dados pode ser composto por diversas fontes, é possível que dados que representam atributos com o mesmo significado possuam tipos diferentes. Por exemplo, que o "gênero"tenha diferentes valores e tipos com significados iguais como "feminino", "f"ou "2". Logo é preciso padronizar esses valores para um tipo em comum.
- Remoção de Registros Duplicados: após a integração dos dados pode ocorrer de diferentes registros armazenarem dados com a mesma informação, mas de formas diferentes. Por exemplo, um determinado registro aparece sem abreviações e em outro ele aparece

abreviado.

- Tratamento e Eliminação de Ruídos: frequentemente os dados coletados podem conter erros gerados por várias causas durante a recuperação de diversas fontes, como por exemplo tipos de dados não suportados pelo gerenciador de base de dados. Os campos que contém ruídos devem ser tratados atribuindo os valores corretos aos dados ou devem ser eliminados da base de dados caso não seja possível tratá-los.
- Tratamento de Valores Ausentes: Encontrar registros que possuem campos de valores nulos é bastante comum. Isto pode ocorrer devido a erros na entrada dos dados, por exemplo, ao preencher um cadastro e sendo esquecido ou ignorado alguns campos. Deste modo, é preciso estabelecer critérios para a correção de atributos ausentes, definindo se os dados serão preenchidos com o valor correspondente ou se serão ignorados.

2.2.3 Transformação de Dados

A Transformação do Dados é a fase do KDD precedente a MD. Após a seleção, pré-processamento e limpeza, os dados necessitam ser devidamente armazenados e formatados para que os algoritmos sejam aplicados. Geralmente os algoritmos usados na MD precisam que os dados se apresentem em um formato adequado, sendo necessária a execução de operações de transformação destes dados (NEVES, 2003).

2.2.4 Mineração dos Dados

Segundo Witten *et al.* (2016), a MD é o nome dado ao conjunto de técnicas que proporcionam a aprendizagem em um sentido prático de padrões a partir de dados, possibilitando encontrar explicações sobre a natureza destes dados e previsões a partir dos padrões identificados.

Em termos gerais, segundo Elmasri (2002), a MD compreende os seguintes objetivos:

- Previsão: a MD pode apresentar como certos atributos dos dados reagirão no futuro;
- Identificação: os padrões de dados podem ser utilizados para identificar a presença de um evento, atividade ou item;
- Classificação: é capaz de dividir os dados de maneira que classes ou categorias distintas consigam ser identificadas a partir da combinação de parâmetros;
- Otimização: a MD pode otimizar a utilização de recursos limitados, como espaço, tempo, dinheiro ou materiais e maximizar variáveis de saída como vendas ou lucros diante de um determinado conjunto de restrições.

2.2.5 Avaliação dos Dados

A fase de avaliação dos dados é a última etapa do processo de descoberta de conhecimento em banco de dados. Após finalizar a aplicação das técnicas de MD, ocorre a interpretação dos resultados obtidos. Nesta etapa, deverá ser realizado um estudo e uma avaliação dos resultados, identificando claramente quais padrões ou prognósticos podem ser usados, sempre baseados em sua expressividade estatística (CÔRTES *et al.*, 2002).

2.3 Mineração de Dados Educacionais

A Mineração de Dados Educacionais (MDE) é um campo de pesquisa interdisciplinar, em que as técnicas de MD são aplicadas em dados educacionais. Seu objetivo é entender melhor como os alunos adquirem conhecimento e reconhecem as configurações em que aprendem para melhorar os resultados educacionais. Os sistemas educacionais podem armazenar uma enorme quantidade de dados oriundos de várias fontes em diferentes formatos (ROMERO; VENTURA, 2013).

Em razão de possuir um enorme potencial de transformação, a MDE pode ser utilizada para descobrir como os alunos aprendem, predizer a aprendizagem e compreender o comportamento de aprendizagem, além de auxiliar no desenvolvimento de tecnologias melhores e mais inteligentes para apoiar alunos e professores (BAKER, 2014).

Segundo Calders e Pechenizkiy (2012), a MDE ajuda a resolver problemas relacionados a diferentes fases do processo de aprendizagem, que pode ser tanto formal (provas, testes) ou informal (jogos educativos), quanto intencional (tutoria) ou inesperada (utilizando mídias sociais).

Cada problema educacional em particular requer diferentes tipos de técnicas de MD, pois as técnicas tradicionais de MD não podem ser aplicadas diretamente a esses tipos de dados e problemas. Existem muitos instrumentos de MD, em geral disponíveis, mas esses não são projetados para tratar dados educativos e um gestor não pode usar esses instrumentos sem conhecimento de conceitos de MD (VENKATACHALAPATHY *et al.*, 2017).

Segundo Baker *et al.* (2011), na maioria dos casos, as técnicas utilizadas no campo de pesquisa da MDE são derivadas diretamente do campo da MD. As técnicas podem ser agrupadas diante da taxonomia das principais subáreas de pesquisa em MDE, apresentadas a seguir:

- Predição: busca prever o valor de um determinado atributo.

  - Classificação: visa identificar por meio do aprendizado de uma função a qual classe um determinado registro faz parte.
  - Regressão: procura predizer um atributo de valor numérico dado um conjunto de dados específico.
- Agrupamento: possui o objetivo de identificar e aproximar os registros semelhantes.
- Mineração de Relações: tem o propósito de descobrir relações entre as variáveis em um conjunto de dados.
  - Mineração de Regras de Associação: visa encontrar regras de conhecimento, relativas ao relacionamento entre atributos em um conjunto de dados. Apresentam a forma: SE atributo X ENTÃO atributo Y.
  - Mineração de Correlações: busca identificar correlações lineares entre variáveis na análise de um conjunto de dados.
  - Mineração de Padrões Sequenciais: visa encontrar a sequência temporal entre eventos e a influência desses eventos em uma variável.
  - Mineração de Causas: possui o objetivo de averiguar se um evento causa outro evento, por meio da análise dos padrões de covariância.
- Destilação de Dados: apresentam-se dados complexos de maneira a facilitar a compreensão dos usuários.
- Descobrimento com modelos: emprega-se um modelo existente (desenvolvido com técnicas de clusterização, classificação, regressão, etc.) como um componente em outra análise.

Segundo Romero e Ventura (2007) em sistemas educacionais, a extração de conhecimento pode ser útil tanto para os educadores responsáveis pela construção, planejamento, concepção e manutenção dos sistemas, como para os estudantes que utilizam e interagem com a aplicação. Assim, a aplicação da MD pode ser orientada a diferentes atores conforme o ponto de vista particular de cada um:

- Orientada para alunos: possui os objetivos de recomendar aos estudantes: tarefas, recursos e atividades que beneficiem e melhorem a sua aprendizagem, sugerir boas experiências de aprendizagem para os alunos, indicar atalhos de percursos ou simplesmente por meio de sugestões de *links*, com base em experiências dos alunos e em tarefas realizadas por outros alunos semelhantes;
- Orientada para educadores: detêm os objetivos de fornecer um *feedback* aos instrutores,

avaliar a estrutura do conteúdo do curso e sua eficácia sobre o processo de aprendizagem, classificar os estudantes em grupos com base em suas necessidades de orientação, descobrir padrões regulares e irregulares de aprendizagem, encontrar os erros mais frequentemente cometidos, descobrir atividades que são mais eficazes, encontrar informações para melhorar a adaptação e personalização dos cursos, reestruturar *sites* para melhor personalizar o curso, organizar os conteúdos de forma eficiente para o progresso do aluno, construir adaptativamente planos de instrução, etc;

- Orientada para gestores educacionais: o propósito é possuir medidas sobre como adaptar o sistema ao comportamento de seus usuários e melhorar sua eficiência, ter parâmetros sobre como melhor organizar os recursos institucionais (materiais e humanos), melhorar a oferta de programas educacionais e determinar a eficiência da nova abordagem de ensino à distância mediada por computador.

3 TRABALHOS RELACIONADOS

Este capítulo apresenta alguns trabalhos relacionados destacando as semelhanças e diferenças com este trabalho.

3.1 Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados

Em Manhães *et al.* (2011), realizou-se um estudo com o propósito de identificar precocemente o subconjunto dos estudantes com tendência a evasão escolar do curso de graduação de Engenharia Civil da Universidade Federal do Rio de Janeiro (UFRJ). O autor utilizou dados históricos de estudantes concluintes e não concluintes do curso de Engenharia Civil com relação às turmas que ingressaram no período de 1994 à 2005, salienta-se que foram utilizadas notas das disciplinas cursadas no primeiro semestre do curso. Foram empregadas um conjunto de técnicas de MD para verificar a sua situação final no curso, identificando se o mesmo possui risco de evasão. Foram utilizados 10 algoritmos de classificação disponíveis na ferramenta Weka, na etapa de MD, que foram avaliados em três experimentos, objetivando-se comparar o desempenho dos algoritmos aplicados ao domínio do problema. Os experimentos mostraram que a acurácia média obtida entre eles era semelhante, variando entre 75% a 80%.

O trabalho de Manhães *et al.* (2011) assemelha-se a este por utilizar técnicas de MD de aprendizagem supervisionada para identificar o perfil dos alunos com tendência a evasão escolar, por meio de algoritmos de classificação. Este estudo se diferencia por analisar dados de um curso da área de TI, além de não restringir o estudo apenas às disciplinas do primeiro semestre.

3.2 Auxiliando o Desempenho de Alunos com Tendência a Evasão na Educação a Distância Utilizando Técnicas de Mineração de Dados e Sistemas Multiagentes

No trabalho de Lira *et al.* (2016) é criado um módulo para predição de alunos que apresentam risco de evasão escolar, no qual foi integrado a um Sistema Multiagente (SMA), desenvolvido pelo Grupo de Estudos em Engenharia de Software e Sistemas Multiagente (GESMA) da UFC – Campus Quixadá/CE. O modulo é responsável pela identificação prévia de características que representam comportamentos que podem levar o aluno a evadir do curso. Para que isso fosse possível, foram analisados dados históricos dos alunos na plataforma *Moodle* da

Universidade Estadual do Ceará (UECE). Os dados passaram por um processo de clusterização, para dividir os perfis dos alunos, e por meio da técnica de classificação foi possível prever o desempenho dos alunos para que fossem auxiliados pelo SMA. Com esses valores do desempenho identificados, as informações são repassadas aos agentes do sistema para que decisões sejam tomadas em grupo. Para avaliar a predição do módulo foram realizados dois testes: o primeiro foi feito por meio da Técnica de *Cross-validation*, no qual com o algoritmo *Random Forest* foi obtido uma acurácia de 98,27%, o segundo foi fornecida pela UECE, uma planilha com a relação de alunos divididos entre alunos graduados, desistentes, que abandonaram e que foram transferidos do curso, no qual foi possível identificar em 92.5% os alunos que poderiam evadir.

O trabalho de Lira *et al.* (2016) assemelha-se a este por buscar identificar o perfil de evasão dos estudantes com tendência a evasão escolar, utilizando dados históricos dos estudantes por meio de técnicas de MD, em contraposição diferencia-se por não possuir integração com nenhum sistema, além de utilizar técnicas de aprendizagem supervisionada na etapa da MD, no qual serão utilizados algoritmos de classificação para identificar se os estudantes possuem risco de evasão escolar.

3.3 Análise Estatística da Relação Entre Evasão e as Respostas do Questionário para Ingressantes da UFC-Quixadá

Em Macêdo *et al.* (2016) foi realizado um estudo com o propósito de analisar os fatores mais relevantes na distinção entre estudantes evadidos e não evadidos, nos três primeiros semestres dos cursos da UFC – Campus Quixadá/CE. Para atingir o objetivo do trabalho, foram analisadas as respostas de um questionário requerido aos alunos ingressantes de 2015. Com os dados obtidos foram realizadas a análises de perfis, em que se comprovou que os estudantes evadidos e não evadidos respondem de maneira distinta o questionário. Com a análise fatorial, foi permitido sumarizar os dados, encontrando-se sete fatores que representam os dados e apresentam 36,3% da variância total. Foi empregado um Teste de Hipótese, para verificar quais perguntas os grupos de estudantes respondem de maneira distinta. Os resultados mostraram um total de sete perguntas que diferenciavam os perfis de alunos e um fator que se destacou agrupando a maior quantidade de perguntas significantes e que se demonstrou mais relevante para tentar prever o estudante que irá evadir.

O trabalho de Macêdo *et al.* (2016) assemelha-se a este por buscar identificar o perfil

de evasão dos alunos que apresentam risco de evasão escolar e levantar possíveis fatores que motivam os alunos a evadir. Contudo, neste estudo se diferencia por aplicar algoritmos de MD para diferenciar o perfil do aluno que evade.

O Quadro 1 apresenta as principais semelhanças e diferenças entre este trabalho e os trabalhos analisados.

Quadro 1 – Comparação entre os trabalhos relacionados e o proposto.

Autores	**Foco do Trabalho**	**Técnica**	**Ferramenta**
Manhães *et al.* (2011)	Predição do subconjunto dos alunos com tendência a evasão escolar.	Classificação	WEKA
Lira *et al.* (2016)	Desenvolvimento de um módulo para predição de alunos que apresentam risco de evasão escolar.	Clusterização e Classificação	WEKA
Macêdo *et al.* (2016)	Analisar os fatores mais relevantes na distinção entre estudantes evadidos e não evadidos.	Análise Fatorial	SPSS
Este trabalho	Identificar o perfil dos alunos que evadem.	Classificação	Orange

Fonte: Elaborado pelo autor

4 PROCEDIMENTOS METODOLÓGICOS

Neste capítulo, são apresentados os procedimentos metodológicos que foram realizados para a conclusão deste trabalho. Na seção 4.1 é apresentado a etapa de levantamento dos dados da UFC - Campus Quixadá/CE. Na seção 4.2 é apresentado a etapa de preparação dos dados necessárias para aplicação dos algoritmos de MD. Na seção 4.3 é apresentada a aplicação da técnica de classificação com o auxílio da ferramenta *Orange* nos dados.

4.1 Levantamento de dados

Neste estudo, foram utilizados dados históricos de estudantes do curso de Bacharelado em CC, da UFC - Campus Quixadá/CE. Os dados foram disponibilizados pela coordenação do curso de CC da universidade e contêm registros dos históricos de estudantes do período de 2013 a 2018. Ressalta-se que os alunos tiveram a sua identificação preservada. Os registros apresentam dados de alunos com matrículas inativas e ativas, ou seja, que concluíram, abandonaram ou que estão ativos no curso. A base de dados é composta por 1914 registros, na Tabela 1 é apresentada a divisão desses registros, conforme a sua situação no curso (classe).

Tabela 1 – Divisão de registros por classe.

Classe	**Número de tuplas**
Ativo	1150
Concluinte	132
Evadido	632

Fonte: Elaborado pelo autor

O aluno é considerado ativo se possuir algum vínculo com a universidade, estando matriculado em pelo menos uma disciplina ou estiver matriculado no estágio ou trabalho de conclusão de curso Trabalho de Conclusão de Curso (TCC). Considera-se um aluno concluinte caso o mesmo tenha completado todos os requisitos para aprovação e conclusão do curso. O aluno é classificado como evadido caso não realize matrícula em disciplinas, nem efetue trancamento total do curso, no decorrer dos períodos letivos seguintes, ou tenha excedido o prazo máximo estabelecido para concluir o curso.

O procedimento de seleção dos dados foi baseado na informação de que o maior número de evasões ocorre no início do curso, portanto decidiu-se utilizar dados acadêmicos do primeiro ano letivo, referentes as disciplinas do primeiro e segundo semestre (SARAIVA; MAS-

SON, 2003). Por meio da análise do banco de dados, foi possível selecionar as informações que teriam maior significância para os experimentos que seriam realizados. Os atributos escolhidos para pesquisa foram:

- Disciplina: identifica as disciplinas do primeiro ano letivo da grade curricular do curso de CC. Por se tratar de apenas um identificador da disciplina este atributo não será considerado na análise pelos algoritmos classificadores para a identificação do perfil dos alunos, dessa forma, constasse que o mesmo é um meta-atributo;
- IRA individual: trata-se do Índice de Rendimento Acadêmico (IRA) individual do aluno, um indicador de desempenho calculado no fim de cada semestre, no qual este atributo representa o IRA calculado no último período letivo do aluno.
- Horas integralizadas: carga horária total cumprida pelo aluno durante o curso;
- Status: refere-se à situação acadêmica final do discente. Este representa o identificador das três classes de alunos analisadas (ativo, concluinte e evadido);
- Reprovações por nota: número de reprovações por nota do aluno na disciplina;
- Reprovações por falta: número de reprovações por falta do aluno na disciplina;
- Trancamentos: número de trancamentos do aluno na disciplina.

4.2 Preparação dos dados

A partir da obtenção dos dados, os mesmos foram armazenados no Sistema de Gerenciamento de Banco de Dados (SGBD) PostgreSQL em duas tabelas, a primeira tabela possuía seis atributos, acerca dos dados do histórico de notas dos alunos compostos por: período letivo, componente curricular, frequência, nota, situação e matrícula. A segunda tabela possuía onze atributos com relação aos dados da situação acadêmica e moradia do aluno formada pela matrícula, cidade/estado nascimento, bairro, cidade atual, estado atual, período inicial, IRA individual, IRA geral, *status*, tipo de saída e horas integralizadas. Com as tabelas criadas e o relacionamento entre elas estabelecido, foram realizadas consultas para gerar uma nova tabela, que foi utilizada para realizar os experimentos com os classificadores, no qual foram criadas visões no banco para fazer a contagem das reprovações por nota, reprovações por falta e trancamentos de cada aluno para todas disciplinas do primeiro ano letivo do curso de CC, juntamente com adição dos atributos mencionados na seção 4.1.

Durante este procedimento, foram encontrados alunos que não possuíam os dados dos históricos das notas, que impossibilitaria a reprodução dos mesmos na nova tabela, devido

a esse fato, todas os alunos que não possuíam essas informações foram removidos na análise. Outra modificação foi realizada no atributo status que identifica as classes analisadas pelos algoritmos, no qual o mesmo era composto pelos valores cancelado, trancado, ativo e concluído, porém após a visualização e análise dos dados, observou-se que o valor trancado era irrelevante para o estudo pois possuía apenas onze registros, que não seria suficiente para que os algoritmos aprendessem a diferenciar o trancamento dos demais tipos, portanto o valor trancamento também foi removido da análise.

4.3 Aplicação dos algoritmos de MD

Nesta etapa foram realizados testes de classificação na base de dados apresentada na seção 4.1, no qual foram utilizados algoritmos populares de aprendizagem de máquina, na ferramenta *Orange*, que fornece um grande conjunto de instrumentos, que permitem que atividades de pré-processamento, mineração e visualização de dados sejam efetuadas de forma simples, além de facilitar que usuários comparem diferentes métodos e identifiquem os mais adequados ao problema.

Foram selecionados oito algoritmos de classificação disponíveis no *Orange*, a escolha se deve ao fato da vasta utilização dos algoritmos em vários contextos. Os algoritmos selecionados foram: *Decision Tree*, *Naive Bayes*, *Support Vector Machine* (SVM), *Logistic Regression*, *Neural Network*, *Random Forest*, *AdaBoost*, *K-Nearest Neighbors* (KNN). Por intermédio do instrumento *test & score* existente no *Orange* é permitido submeter os dados aos algoritmos classificadores aplicando técnicas distintas de amostragem. Nesta fase os dados são divididos, sendo um grupo enviado para teste de predição fornecidas pelo algoritmo classificador, e outro grupo destinado para treinamento do algoritmo em questão. A técnica de amostragem dos dados escolhida foi a validação cruzada (*cross validation*) pelo método *k-folder*, que consiste em dividir o conjunto total de dados em k subconjuntos iguais e com base nisto, um subconjunto é utilizado para teste e os k-1 subconjuntos restantes são utilizados para estimação dos parâmetros e calcula-se a acurácia do modelo. Este processo é repetido k vezes usando um subconjunto de teste distinto em cada iteração (KOHAVI *et al.*, 1995). O valor de k assumido na amostragem foi de 10 *folds*, valor definido por permitir a melhor divisão dos conjuntos com relação ao tamanho dos doados utilizados, em testes executando o parâmetro k com valores 2, 3, 5, 10 e 20 *folds* na ferramenta *Orange*.

Para medir o desempenho dos algoritmos na fase de classificação, utilizou-se o total

de instâncias classificadas corretamente (acurácia), a proporção daqueles que foram classificados como corretos, quando efetivamente eram (precisão) e a taxa de frequência em que o classificador encontra os exemplos de uma classe (revocação). Foram realizados experimentos com dois cenários, o primeiro cenário possuindo o número total de registros com a divisão dos registros por classes desbalanceada. E o segundo cenário contendo uma amostra dos registros com a divisão entre as classes balanceadas, visando validar o desempenho dos algoritmos de mineração de dados aplicados ao domínio do problema. Além da análise dos dados para identificação do perfil dos alunos que evadem.

5 RESULTADOS

Neste capítulo, estão todos os resultados encontrados das análises feitas sobre os dados históricos do curso de CC da UFC com a aplicação da técnica de classificação. Na Seção 5.1 são mostrados os resultados referentes ao primeiro cenário com a aplicação dos algoritmos sobre os dados desbalanceados. Na Seção 5.2 é mostrado os resultados relativos ao segundo cenário com uma amostra dos dados possuindo a divisão dos registros por classes balanceadas. Na Seção 5.3 é mostrado os perfis encontrados após a análise dos dados.

5.1 Resultados do Cenário 1

No primeiro cenário os 8 algoritmos classificadores listados na Seção 4.3 foram executados na ferramenta *Orange* utilizando os dados com a divisão dos registros por classes desbalanceadas. A base de dados utilizada possuía 1912 registros composta por 1150 ativos, 632 abandonados e 132 concluintes, no qual os dados foram divididos em 10 conjuntos aplicando a técnica de validação cruzada utilizando o método *k-fold*, descrita na Seção 2.2, que foi atribuído o valor de 10 *folds*. Na Tabela 2 mostra-se a média da acurácia, precisão e revocação obtidas para cada algoritmo de classificação utilizando as três classes de alunos.

Tabela 2 – Acurácia, Precisão e Revocação dos Classificadores.

Classificadores	Decision Tree	Naive Bayes	SVM	Logistic Regression	Neural Network	Random Forest	KNN	AdaBoost
Acurácia	98,9%	76,5%	75,0%	84,5%	86,6%	98,0%	95,4%	99,0%
Precisão	98,9%	80,1%	74,4%	84,5%	86,5%	98,0%	95,4%	99,0%
Revocação	98,9%	76,5%	75,0%	84,5%	86,6%	98,0%	95,4%	99,0%

Fonte: Elaborado pelo autor

Constata-se na Tabela 1, que a acurácia da classificação dos estudantes se manteve superior a 84 % , com exceção dos algoritmos *SVM* e *Naive Bayes* que obtiveram os menores desempenhos, os resultados mostram que o algoritmo *AdaBoost* alcançou o melhor resultado com 99 % de acurácia, precisão e revocação. Possivelmente, entre os algoritmos testados é o mais adequado quanto a utilização da abordagem proposta para o problema. É válido ressaltar, que o algoritmo *Decision Tree* alcançou resultados igualmente bons com 98,9 % de acurácia, precisão e revocação.

5.2 Resultados do Cenário 2

No segundo cenário também foram executados os 8 algoritmos classificadores listados na seção 4.3 e aplicada a técnica de amostragem dos dados validação cruzada utilizando o método *k-fold*, explicado na Seção 4.3, que foi assumido com o valor de 10 *folds*. Porém foi usada uma amostra dos dados com a divisão entre as classes dos registros balanceadas. A base utilizada possuía 396 registros, formada por 132 ativos, 132 abandonados e 132 concluintes. Na Tabela 3 mostra-se a acurácia, precisão e revocação obtidas para cada algoritmo de classificação utilizando as três classes de alunos.

Tabela 3 – Acurácia, Precisão e Revocação dos Classificadores.

Classificadores	Decision Tree	Naive Bayes	SVM	Logistic Regression	Neural Network	Random Forest	KNN	AdaBoost
Acurácia	95,2%	89,1%	88,1%	86,4%	88,6%	92,2%	89,6%	95,5%
Precisão	95,3%	89,1%	88,1%	86,1%	88,6%	92,2%	89,8%	95,5%
Revocação	95,2%	89,1%	88,1%	86,4%	88,6%	92,2%	89,6%	95,5%

Fonte: Elaborado pelo autor

Constata-se na Tabela 3, que a acurácia da classificação dos estudantes se manteve superior a 86 %, em todos os algoritmos, no qual o algoritmo *AdaBoost* também obteve o melhor resultado com 95,5 % de acurácia, precisão e revocação no cenário dois. Seguido novamente pelo algoritmo *Decision Tree* que alcançou um desempenho bem próximo com 95,2 % de acurácia, precisão e revocação. Neste cenário os algoritmos avaliados tiveram um resultado mais próximos variando a taxa de acerto de acerto em no máximo 9,1%.

O objetivo da realização do experimento com o segundo cenário era verificar se alteração do conjunto de dados desbalanceados afetaria significativamente o desempenho dos algoritmos classificadores avaliados, uma vez que a classe dos alunos ativos representava cerca de 60% dos dados, e esse fator poderia impactar negativamente a analise pois existiria a possibilidade dos classificadores tenderem classificar corretamente as classes de maior proporção e errarem os exemplos das classes com menor proporção (CASTRO; BRAGA, 2011).

Comparando os resultados dos cenários 1 e 2, verificam-se que as duas formas de divisão da base de dados não são significativamente distintas. Relacionando os valores de cada classificador disponível nas Tabelas 2 e 3, observa-se que os resultados são similares na maioria dos classificadores seguindo até uma ordem semelhantes de desempenho, com exceção dos

algoritmos *Naive Bayes* e *SVM* que obtiveram uma boa melhoria de performance do primeiro para o segundo cenário. Os desempenhos satisfatórios dos classificadores nos dois cenários é um forte índicio de que os atributos utilizados são suficientes para realizar a previsão dos estudantes com risco de evasão escolar.

5.3 Perfis de alunos encontrados

Durante a análise usando os algoritmos de árvore de decisão foram encontrados alguns perfis que caracterizam alunos concluintes e alunos com tendência a evasão nos dados do curso de CC, da UFC - Campus Quixadá/CE, as árvores utilizadas para a extração dos perfis estão disponíveis no Apêndice A, os perfis identificados são:

- Alunos que possuem um IRA individual maior que 1.5 e menor ou igual 3.4, três reprovações por media e o número de horas integralizadas menor ou igual a 224 tendem a evadir do curso;
- Alunos que possuem um IRA individual menor ou igual a 3.5, uma reprovação por falta e o número de reprovações por media maior ou igual a três e inferior a cinco tendem a evadir do curso;
- Alunos que possuem um IRA individual maior que 1.5 e menor ou igual 2.2, duas reprovações por media e uma reprovações por falta tendem a evadir do curso;
- Alunos que possuem um IRA individual maior que 3.4, zero reprovações por falta, zero trancamentos e horas integralizadas maior que 3136 tem mais probabilidade de concluir o curso;
- Alunos que possuem um IRA individual maior que 3.4 e menor que 4.6, quantidade de reprovações por media maior ou igual a um e menor que três e número de horas integralizadas menor ou igual a 384 tendem a evadir do curso.

6 CONCLUSÕES E TRABALHOS FUTUROS

Os resultados obtidos mostram que a identificação dos alunos que apresentam risco de evasão escolar, por meio da utilização técnicas de MD é viável. Neste estudo, propôs-se a aplicação de 8 algoritmos de classificação na identificação de estudantes com risco de evasão, com base nos dados históricos de alunos do curso de CC, da UFC - Campus Quixadá/CE, no qual foram efetuados experimentos com dois cenários distintos, o primeiro cenário possuindo o número total de registros com a divisão dos registros por classes desbalanceadas. E o segundo cenário contendo uma amostra dos registros com a divisão entre as classes balanceadas. O desempenho dos algoritmos possuem taxas de acerto de até 99 % no primeiro cenário e no segundo cenário de até 95,5% que indicam que os atributos selecionados provaram ser adequados quanto a utilização da abordagem proposta para o problema. Desse modo, espera-se fornecer um estudo que auxilie os docentes na identificação de estudantes que necessitam de apoio, contribuindo para a tomada de decisão e na elaboração de políticas para mitigar à evasão escolar.

As limitações encontradas neste trabalho, se deu por conta da indisponibilidade atual dos dados referentes a outros cursos existentes da universidade, no qual restringiu a análise dos dados de apenas um curso da UFC - Campus Quixadá/CE, impossibilitando a comparação de resultados com outros subconjuntos de alunos.

Como trabalhos futuros, sugere-se estender este estudo para as disciplinas dos semestres seguintes do curso de CC, verificando a eficiência na predição da evasão escolar dos estudantes a partir do desempenho nas disciplinas destes períodos. Considerasse que seria possível também a aplicação de procedimentos semelhantes para outros cursos da universidade, verificando se os resultados adquiridos se repetem para outros grupos de alunos de graduação da UFC - Campus Quixadá/CE.

REFERÊNCIAS

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de dados educacionais: Oportunidades para o brasil. **Brazilian Journal of Computers in Education**, [S.l.], v. 19, n. 02, p. 03–13, 2011.

BAKER, R. S. Educational data mining: An advance for intelligent systems in education. **IEEE Intelligent systems**, [S.l.], v. 29, n. 3, p. 78–82, 2014.

BARROSO, M. F.; FALCÃO, E. B. Evasão universitária: o caso do instituto de física da ufrj. **IX Encontro Nacional de Pesquisa em Ensino de Física**, [S.l.], v. 9, p. 1–14, 2004.

CALDERS, T.; PECHENIZKIY, M. Introduction to the special section on educational data mining. **Acm Sigkdd Explorations Newsletter**, [S.l.], v. 13, n. 2, p. 3–6, 2012.

CAMILO, C. O.; SILVA, J. C. d. **Mineração de dados:** conceitos, tarefas, métodos e ferramentas. Goiás: Universidade Federal de Goiás (UFG), p. 1–29, 2009.

CARDOSO, C. B. **Efeitos da política de cotas na Universidade de Brasília:** uma análise do rendimento e da evasão. Brasília: Universidade de Brasília (UnB), 2008.

CASTANHEIRA, L. G. **Aplicação de técnicas de mineração de dados em problemas de classificação de padrões**. Belo Horizonte: Universidade Federal de Minas Gerais (UFMG), 2008.

CASTRO, C. d.; BRAGA, A. P. Aprendizado supervisionado com conjuntos de dados desbalanceados. **Rev. Controle Autom**, [S.l.], v. 22, n. 5, p. 441–466, 2011.

CÔRTES, S. da C.; PORCARO, R. M.; LIFSCHITZ, S. **Mineração de dados-funcionalidades, técnicas e abordagens**. [S.l.]: PUC, 2002.

DUNKEL, B.; SOPARKAR, N.; SZARO, J.; UTHURUSAMY, R. Systems for kdd: From concepts to practice. **Future Generation Computer Systems**, [S.l.], v. 13, n. 2-3, p. 231–242, 1997.

ELMASRI, R. **Sistemas de banco de dados**. [S.l.]: LTC, 2002.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, [S.l.], v. 17, n. 3, p. 37–37, 1996.

KOHAVI, R. *et al.* A study of cross-validation and bootstrap for accuracy estimation and model selection. In: MONTREAL, CANADA. **Ijcai**. [S.l.], 1995. v. 14, n. 2, p. 1137–1145.

LIRA, K. C.; OLIVEIRA, M. A. de; MAGALHÃES, R. P.; GONÇALVES, E. J. T. **Auxiliando o desempenho de alunos com tendência a evasão na educação a distância utilizando técnicas de mineração de dados e sistemas multiagentes**. Ceará: Universidade Federal do Ceará (UFC), 2016.

MACÊDO, C. S.; SOUZA, C. P. de; FREITAS, L. I. B. **Análise estatística da relação entre evasão e as respostas do qestionário para ingressantes da UFC-Quixadá**. Ceará: Universidade Federal do Ceará (UFC), 2016.

MANHÃES, L. M. B.; CRUZ, S. M. S. D.; COSTA, R. J. M.; ZAVALETA, J.; ZIMBRÃO, G. 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.

MARTINHO, V. R. d. C. **Sistema inteligente para a predição de grupo de risco de evasão discente**. São Paulo: Universidade Estadual Paulista (UNESP), 2014.

MATOS, Y. C. C. *et al.* **Detecção de fraudes no consumo de energia elétrica usando árvores de decisão**. Belém: Universidade Federal do Pará (UFP), 2017.

MINISTÉRIO, E. da. **Diplomação, retenção e evasão nos cursos de graduação em Instituições de Ensino Superior Públicas**. [S.l.]: MEC Brasília, DF, 1997.

NAGAI, N. P.; CARDOSO, A. L. J. A evasão universitária: uma análise além dos números. **Revista Estudo & Debate**, [S.l.], v. 24, n. 1, 2017.

NEVES, R. d. C. D. d. **Pré-processamento no processo de descoberta de conhecimento em banco de dados**. Porto Alegre: Universidade Federal do Rio Grande do Sul (UFRGS), 2003.

OLIVEIRA, C. H. M. de; SANTOS, F. R. T.; LEITINHO, J. L.; FARIAS, L. G. A. T. Busca dos fatores associados à evasão: um estudo de caso no campus universitário da ufc em crateús. **Revista Internacional de Educação Superior**, v. 5, p. 019006, 2019.

OLIVEIRA, R. B. T. d. **O processo de extração de conhecimento de base de dados apoiado por agentes de software.** Tese (Doutorado) — Universidade de São Paulo (USP), 2000.

PRASS, F. S. *et al.* **Estudo comparativo entre algoritmos de análise de agrupamentos em data mining**. Florianópolis: Universidade Federal de Santa Catarina (UFSC), 2004.

ROMERO, C.; VENTURA, S. Educational data mining: A survey from 1995 to 2005. **Expert systems with applications**, [S.l.], v. 33, n. 1, p. 135–146, 2007.

ROMERO, C.; VENTURA, S. Data mining in education. **Wiley Interdisciplinary Reviews**, Wiley Online Library, v. 3, n. 1, p. 12–27, 2013.

SARAIVA, S.; MASSON, M. Evasão e permanência em uma instituição de tradição: um estudo sobre o processo de evasão de estudantes em cursos de engenharia na escola politécnica da ufrj. **Relatório de Pesquisa**, [S.l.], 2003.

SILVA, R. L. L. F.; MOTEJUNAS, P. R.; HIPÓLITO, O.; LOBO, M. B. C. M. A evasão no ensino superior brasileiro. **SciELO Brasil**, [S.l.], v. 37, n. 132, p. 641–659, 2007.

VELOSO, T. C. M.; ALMEIDA, E. P. de. **Evasão nos cursos de graduação da Universidade Federal de Mato grosso, Campus Universitário de Cuiabá–um processo de exclusão**. Série-Estudos-Periódico do Programa de Pós-Graduação em Educação da UCDB, [S.l.], n. 13, 2013.

VENKATACHALAPATHY, K.; VIJAYALAKSHMI, V.; OHMPRAKASH, V. Educational data mining tools: A survey from 2001 to 2016. In: IEEE. **2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM)**. [S.l.], 2017. p. 67–72.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. **Data Mining**. [S.l.]: Morgan Kaufmann, 2016.

APÊNDICE A – ÁRVORES ANALISADAS PARA A IDENTIFICAÇÃO DOS PERFIS DOS ALUNOS

Essas são as árvores utilizadas para a identificação dos perfis dos alunos, apresentados na Seção 5.3 por meio do algoritmo de árvore de decisão, *Decision Tree* no qual, na Figura 2 é possível visualizar os nove níveis da árvore no cenário 1 e na Figura 3 mostra-se os nove níveis da árvore no cenário 2.

Figura 2 – Árvore do algoritmo Tree Cenário 1

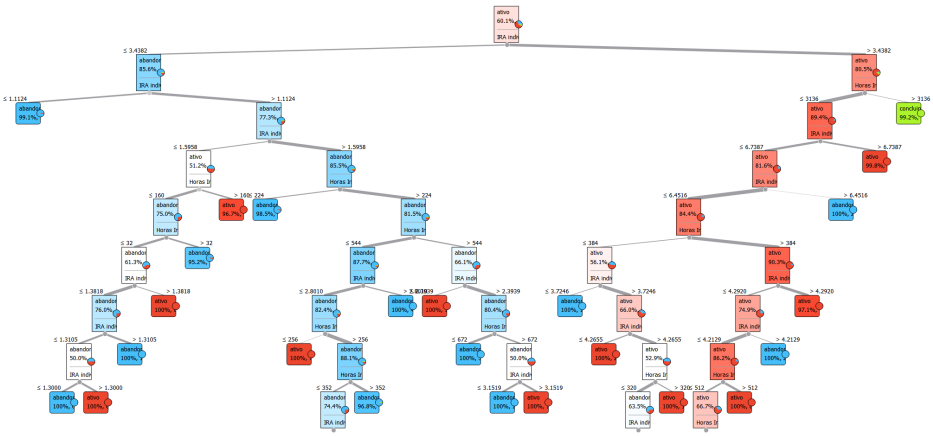


Figura 3 – Árvore do algoritmo Decision Tree Cenário 2