Análise e predição de evasão dos alunos do curso de Sistemas de Informação da Universidade Federal de Santa Maria *Campus* Frederico Westphalen por meio da mineração de dados educacionais

Edson Noetzold[1], **Solange Pertile**[2]

[1]Curso de Bacharelado em Sistemas de Informação
[2]Departamento de Tecnologia da Informação (DTecInf)

Universidade Federal de Santa Maria (UFSM) - Campus Frederico Westphalen - Linha 7 de Setembro, s/n, CEP: 98400-000, BR 386 Km 40- Frederico Westphalen - RS

edsonversusnoetzold@gmail.com, solangepertile@gmail.com

Abstract. *The high school dropout rates are a reality present in several higher education courses offered in Brazil, which highlights the need to investigate this issue, which is responsible for economic losses in institutions and impacts on the global education scenario. This paper aims to develop a study on school dropout patterns in higher education, based on the analysis of data provided by the Information Systems course at the Federal University of Santa Maria (UFSM). These data went through a data processing system, in order to point out indicators related to factors that classify possible evasions. Results were generated through decision trees, which pointed to data regarding the student and his academic performance as important factors for school dropout in higher education.*

Keywords: *School Dropout. Data Mining. Higher Education.*

Resumo. *Os elevados índices de evasão escolar constituem uma realidade presente em diversos cursos superiores ofertados no Brasil, o que evidencia a necessidade de investigação dessa problemática, responsável por perdas econômicas nas instituições e impactos no cenário global da educação. Este artigo tem como proposta desenvolver um estudo sobre os padrões da evasão escolar no ensino superior, com base na análise de dados fornecidos pelo curso de Sistemas de Informação da Universidade Federal de Santa Maria (UFSM). Esses dados passaram por uma sistemática de tratamento de dados, a fim de apontar indicadores relacionados a fatores que classifiquem possíveis evasões. Foram gerados resultados por meio de árvores de decisão, que apontaram dados referentes ao aluno e seu desempenho acadêmico como fatores importantes para a evasão escolar no ensino superior.*

Palavras-Chave: *Evasão Escolar. Mineração de Dados. Ensino Superior.*

1. Introdução

De acordo com Favero (2006), denomina-se a evasão escolar como o processo de desistência do ensino pelo discente de determinado curso, indiferentemente à porcentagem de participação do aluno nas aulas. Já Almeida & Kappel (2020) apontam que tal problemática encontra-se emergente no cenário atual, como observado nas preocupantes taxas de evasão atuais. Segundo Silva Filho et al. (2007), esse impasse vem

impactando até mesmo o cenário internacional, afetando expressivamente os resultados dos sistemas educacionais e causando perdas nas instituições públicas e privadas.

Assim, a motivação para a delimitação da temática do presente Trabalho de Graduação em Sistemas de Informação reside em analisar o problema apresentado, bem como identificar o perfil de alunos propensos a evadir no curso de Sistemas de Informação da Universidade Federal de Santa Maria, utilizando a mineração de dados como ferramenta. A escolha do curso a ser analisado se deu devido a uma análise dos dados dos alunos fornecidos pela coordenação do curso de Sistemas de Informação, a qual constatou uma taxa de evasão no ensino de 41,6% no período entre 2010 (início do curso) e 2019. A partir dos resultados obtidos, espera-se auxiliar os gestores na predição desses alunos, de modo a permitir a adoção de medidas que possam minimizar esse problema.

A estruturação do presente artigo inicia descrevendo o referencial teórico na seção 2, em que são abordados à evasão escolar, as causas do impasse da evasão no ensino superior, a mineração de dados e a apresentação do Software Waikato Environment for Knowledge Analysis (WEKA), que foi utilizado como ferramenta de mineração de dados. Em seguida, a seção 3 apresenta trabalhos cuja proposta é análoga ao objetivo do presente projeto, bem como a contribuição desses projetos para o estado da arte, ou seja, quais pontos dos trabalhos apontados são relevantes para o tema. Em sequência, a seção 4 explicita a solução proposta para o problema abordado, além das etapas que constituem o processo de mineração a ser utilizado. Por fim, a seção 5 aponta as conclusões referentes ao presente TGSI, bem como perspectivas para futuros trabalhos.

2. Referencial Teórico

Esta seção apresenta, primeiramente, algumas definições a respeito da temática da evasão escolar. Em seguida, apresenta algumas causas relacionadas à evasão discente no ensino superior. Após, apresentam-se definições a respeito da mineração de dados, que é a base para o desenvolvimento do estudo. Por fim, inclui a apresentação das ferramentas para a mineração, no caso, o software WEKA.

2.1 Definição de evasão escolar

Define-se como evasão escolar a situação na qual o aluno, por qualquer motivo, rompe o vínculo jurídico estabelecido com a instituição de ensino, não renovando sua matrícula no curso matriculado (Johann, 2012).

Para compreender melhor a questão da evasão, é importante observá-la sob dois panoramas diferentes: a evasão anual média e a evasão total. A evasão anual média visa medir o percentual de estudantes matriculados em um determinado curso de uma instituição de ensino superior que, ainda não formado, não efetivou sua matrícula para o semestre/ano subsequente. O cálculo dessa porcentagem utiliza o total de alunos vinculados ao curso, ou seja, o objetivo é descobrir quantos alunos evadiram em certo período em relação à quantidade total de estudantes (Silva Filho et al., 2007).

Já a evasão total refere-se ao número de alunos que, após sua entrada na instituição de ensino, não concluiu a formação em um número de anos. Essa avaliação é comumente relacionada ao índice de titulação – que mede a porcentagem de formandos na totalidade de ingressantes de um determinado ano. Portanto, se cinquenta alunos ingressaram no ano "x" e vinte e cinco se formaram, tem-se um índice de titulação de 50%, com um índice de evasão também de 50% por consequência (Silva Filho et al., 2007).

De acordo com Silva (2007), é imprescindível a investigação das causas de evasão nas instituições de ensino superior. Isso porque, embora já existam pesquisas apontando os agentes causadores da evasão escolar, é perceptível a ausência de homogeneidade nos graus de evasão para os diferentes cursos, impossibilitando a construção de um padrão universal.

2.2 Dados e causas da evasão escolar no ensino superior

A Plataforma Nilo Peçanha (PNP) é um ambiente virtual de cunho federal iniciado em 2017, desenvolvido com o intuito de unificar e disseminar as estatísticas oficiais da Rede Federal de Educação Profissional, Científica e Tecnológica (Rede Federal). Nesse contexto, o ambiente virtual disponibiliza ao usuário dados sobre as unidades que fazem parte da Rede Federal, como seus cursos ofertados, corpo docente, discente e técnico-administrativo, além de relatórios anuais de análise dos indicadores (Ministério da Educação, 2018).

A PNP publicada no ano de 2020, a qual utilizou 2019 como ano base, apontou índices de evasão escolar no ensino superior das Instituições Federais de aproximadamente 12%, correspondentes principalmente à evasão por abandono (6,51%), seguido de desligamento (4,69%), cancelamento (0,23%), transferência externa (0,22%), reprovação (0,10%) e transferência interna (0,02%) (Plataforma Nilo Peçanha, 2020). Tal percentual soma mais de trinta mil matrículas evadidas, evidenciando os elevados índices de evasão presentes na Rede Federal. O PNP anterior, publicado com base nos dados do ano de 2018, indicou cerca de 29000 matrículas evadidas, permitindo concluir que a evasão no ensino superior está aumentando nas instituições federais (Plataforma Nilo Peçanha, 2019).

Ao se abordar a questão das causas da evasão, é importante citar que se constitui de uma problemática de diversas causas, as quais podem ser classificadas como causas psicológicas, sociológicas, organizacionais, interacionais e econômicas (Schargel e Smink, 2002). A categoria psicológica abrange, de maneira geral, o conjunto de causas ligadas ao comportamento do indivíduo, que por sua vez, interferirá nos seus resultados acadêmicos. Nesse contexto, é possível citar primeiramente a influência das reprovações excessivas na motivação do estudante, que tende a diminuir conforme elas ocorrem (Schargel e Smink, 2002); (Gaioso, 2005). Para Negrine (1994), a motivação é um requisito indispensável ao processo de aprendizagem, visto que ela resulta em um comprometimento verdadeiro com o compromisso acadêmico.

Pereira-Silva e Dessen (2003) também frisam a importância da família ao afirmarem que as interações estabelecidas no microssistema família têm o impacto mais expressivo no desenvolvimento individual, embora outros fatores sociais – como a escola e os outros círculos sociais – também exerçam certa influência. Por fim, a imaturidade e rebeldia constituem fatores influentes no desempenho do aluno (Gaioso, 2005).

No campo da sociologia, é possível citar a falta de orientação vocacional – recurso importante no que tange o autoconhecimento do aluno e que possibilita a escolha profissional mais assertiva (Pimenta, 1981). Além disso, para Kirby et al. (2004), os estudos acabam surgindo como uma fonte de estresse no seio familiar, especialmente se tratando de estudantes mais velhos que já construíram suas próprias famílias, tornando possível apontar a família como um fator causador da evasão. Por fim, as dificuldades enfrentadas no ensino superior como consequência da educação básica deficiente também atuam como aspectos causadores da evasão no ensino válidos de serem declarados (Schargel e Smink, 2002); (Gaioso, 2005).

Já a categoria organizacional, a qual reflete a influência da instituição de ensino sobre o discente, cita como causa para o problema em questão o desconhecimento do aluno para com a metodologia implementada na instituição onde estuda (Schargel e Smink, 2002). Conforme Gaioso (2005), a concorrência entre as diferentes instituições de ensino também interferem negativamente nesse problema.

Outra possibilidade para a evasão é referente a estrutura oferecida pela instituição, cuja qualidade pode ser questionada pelo aluno. Nessa perspectiva, Dalrymple e Parsons (2003) salientaram a importância da gestão da qualidade de serviços das instituições, que devem considerar todos os recursos e suportes necessários ao aluno para o bom rendimento acadêmico – como bibliotecas, secretarias, laboratórios, entre outros.

Além disso, outra questão agravante do problema é a recorrência da exclusão social, o que segundo Mantoan (2003), se manifesta dos modos mais diversos e perversos, e coloca em risco o saber do aluno. Mormente, em consonância com Lemos (2007), a presença do *bullying* no ambiente escolar resulta no comprometimento dos processos de aprendizado, levando à perda da motivação e, consequentemente, à evasão.

Os aspectos econômico-financeiros também possuem certo grau de ligação para com a evasão discente - como as divergências no concílio entre os horários de trabalho com a carga horária do curso e a ausência de vantagens imediatas promovidas pela titulação (Jacob, 2000). Para Gomes (1998), o interesse em cursar o ensino superior está diretamente vinculado à possibilidade de ascensão econômica e social que, quando não ocorre rapidamente, pode levar à frustração do aluno e, consequentemente, à evasão no ensino.

A universidade, diante das transformações econômicas, políticas e culturais recentes que influenciam a educação, enfrenta a necessidade de repensar e transformar seus vínculos com a sociedade. É preciso corrigir alguns fatores e eliminar outros para que os acadêmicos possam ter, além do acesso à universidade, a garantia da conclusão do curso (Amaral, 2013).

Em um cenário de constantes mudanças econômicas, políticas e culturais que afetam direta ou indiretamente a educação, é de suma importância que cada instituição de ensino pesquise e desenvolva estratégias de transformação de seus vínculos com a sociedade. É necessário oferecer recursos para que o estudante possa não apenas ingressar no ensino superior, como também concluir seu curso escolhido, e deve-se combater os fatores divergentes a essa perspectiva (Amaral, 2013).

2.3 Expansão do acesso ao ensino superior nas últimas décadas

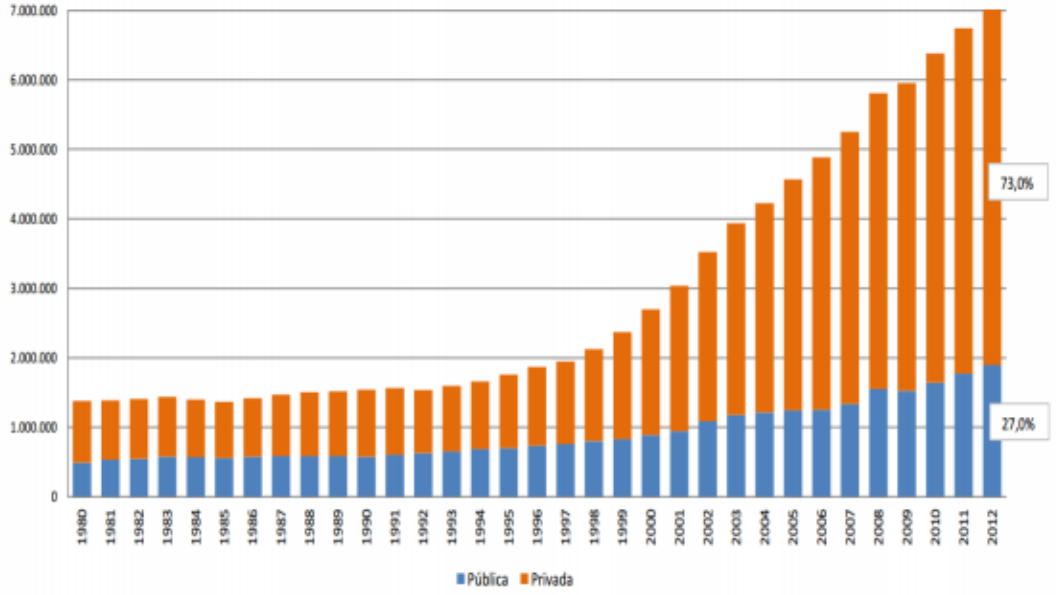
De acordo com Nogueira (2004), a importância da universidade consiste na produção do conhecimento, na formação de cidadãos com pensamento crítico e de profissionais capazes de articular saberes e ascender como líderes intelectuais. Atualmente, o acesso ao ensino configura um recurso imprescindível para suprir o constante aumento das exigências no mercado de trabalho (Amaral e Oliveira, 2011).

Em 1961, foi promulgada a Lei de Diretrizes e Bases da Educação Nacional (LDB) (Brasil, 1961). Tal decreto instituiu a liberdade de ensino e igualdade no envio de recursos para a educação, além da igualdade de atribuições ao Ministério da Educação. Outros pontos importantes da lei foram a instituição da Universidade de Brasília (UnB) e a criação do Conselho Federal de Educação, que viria a ser instaurado de fato em 1962 (Zoccoli, 2009). Entretanto, os anos sessenta não representam o ápice da ascensão da

população ao ensino. Atualmente, a LDB vigente no país corresponde à sancionada em 1996.

O fato é que, desde a instauração da primeira universidade do Brasil no século XIX até os anos finais do século XX, o acesso ao ensino superior no país se manteve acessível apenas a uma pequena parte da população. Foi somente a partir de 1998, com a aplicação de políticas públicas promotoras da educação superior, que alterações perceptíveis na dinâmica de acesso às instituições de ensino superior se tornaram perceptíveis (Dourado, 2002). A evolução no número de matrículas por ano manteve um crescimento discreto até o final dos anos 90, quando o crescimento dos índices começou a crescer acentuadamente, como ilustrado na figura 1.

Figura 1 – Gráfico de evolução das matrículas em instituições de ensino superior públicas e privadas no Brasil entre 1980 e 2012.



Fonte: Ministério da educação, 2013.

As iniciativas de apoio ao ingresso no ensino superior começaram com as instituições de ensino privadas, como é o caso do Fundo de Financiamento Estudantil (Fies) em 1999, e do Programa "Universidade para Todos" (Prouni) em 2004. Pelo Fies, o estudante de baixa renda obtinha acesso a financiamentos de até 100% da mensalidade para o acesso a instituições privadas pela Caixa Econômica Federal. Já o Prouni foi criado com o propósito de garantir o ingresso de discentes de baixa renda a partir de bolsas de estudo integrais ou parciais de 50% - ambas destinadas a não portadores de diploma de ensino superior (Aprile, 2018).

Outro marco importante para a desenvoltura do acesso ao ensino superior reside na criação, em 2005, do Sistema Universidade Aberta do Brasil (UAB), que objetiva o oferecimento de cursos superiores na modalidade a distância (EaD). Esse programa promoveu a expansão e interiorização do ensino superior no Brasil, possibilitando aos residentes das áreas mais remotas do Brasil o acesso ao ensino pela EaD.

Mais adiante, o governo federal criou o Sistema de Seleção Unificada (Sisu) em 2010, que vem atuando até hodiernamente como forma de processo seletivo das instituições públicas de ensino superior. Nele, é utilizada a nota do Exame Nacional do Ensino Médio (Enem) para classificar os estudantes, que escolhem uma primeira e uma

segunda opção de cursos para concorrerem. Assim, os estudantes concorrem para as duas opções de curso solicitadas, e os melhores classificados entre os que optaram por cada curso são selecionados para as vagas disponíveis (Nogueira, 2017). A partir da divulgação dos resultados, o discente classificado é convocado para efetivar sua matrícula na instituição em que foi aprovado.

O Exame Nacional do Ensino Médio (Enem), criado em 1998 como uma ferramenta de avaliação da escolaridade dos estudantes ao final do ensino médio (Andriola, 2011), passou a desempenhar uma ampla importância a partir da criação das iniciativas federais de acesso ao ensino superior, uma vez que passou a ser requisitada a sua realização pelo aluno que almejasse ingressar nas universidades com o auxílio dos programas do governo (Junior et al., 2017). A aplicação do Enem como ferramenta de seleção unificada visou a democratização das oportunidades de acesso às vagas federais de ensino superior (Andriola, 2011).

É importante mencionar que, apesar do intuito dos programas federais objetivarem a superação de obstáculos financeiros para a ascensão educacional da população, alguns grupos permaneceram distantes dessa realidade, levando à criação da Lei N° 12.711, promulgada com o intuito de possibilitar o acesso à educação superior por meio de cotas sociais e raciais (Brasil, 2012). A nova lei permitiu a inclusão dos grupos afastados do ensino e avançou no que tange a diminuição da desigualdade social fomentada pela pobreza e por preconceitos de raça (Marques, 2013).

Entretanto, ainda seguem as dificuldades enfrentadas pelas minorias na sua inclusão efetiva no ensino superior, em especial nos cursos de alta demanda e nos que conferem maior mobilidade social, o que urge o surgimento de políticas de permanência nos cursos superiores para essa parcela da população (Paula, 2017). A autora afirma que é imprescindível incluir as minorias no ensino superior, para assim, tornar eficiente o combate às desigualdades sociais presentes no Brasil.

2.4 Definição de mineração de dados

Devido à multidisciplinaridade da mineração de dados, as definições de mineração variam de acordo com os diferentes autores. Para Hand et al. (2001), a mineração de dados é a análise de grandes conjuntos de dados, a qual procura detectar relações entre essas informações a fim de oferecer maior utilidade para o detentor dos dados. Já Cabena et al. (1998) aponta a mineração como um conjunto composto por "técnicas de máquinas de conhecimentos, reconhecimento de padrões, estatísticas, banco de dados e visualização, para conseguir extrair informações de grandes bases de dados".

De acordo com Dantas (2008), a importância de se utilizar a mineração de dados reside na dificuldade de se interpretar as grandes quantidades de dados armazenadas, o que leva a uma situação de excesso de dados desorganizados e sem utilidade. Nesse contexto, tem-se o que é chamado de *Big Data,* o que de acordo com Silva et al. (2013) se define como o extenso volume de dados de tipo heterogêneo que são produzidos diariamente por fontes descentralizadas e que possuem relações complexas entre si.

É nessa dificuldade de interpretação motivada pela grande quantidade de dados que surge a Descoberta de Conhecimento em Bases de Dados ("*knowledge-discovery in databases"* ou KDD), definida por Dantas (2008) como "o processo de descoberta de padrões e tendências por análise de grandes conjuntos de dados, tendo como principal etapa o processo de mineração." Esse processo se divide em cinco etapas, sendo a primeira a etapa de seleção dos dados de onde as informações são extraídas. A seguir, passa-se para a etapa de pré-processamento, responsável por reduzir a quantidade de

dados - mantendo-se a representatividade - integrar as diferentes fontes de dados obtidas e garantir a qualidade dos dados extraídos, removendo os considerados inconsistentes, por exemplo. A terceira fase, de transformação, diz respeito à adequação dos dados ao software escolhido para a mineração, que corresponde à quarta fase. Por fim, na última fase, interpretam-se os resultados providos pela etapa anterior e avalia-se a qualidade desses resultados (Dantas, 2008).

Um ponto importante da KDD é a capacidade dos softwares de gerar, por meio da mineração de dados, diferentes ferramentas de tomada de decisões, que permitem a visualização de um panorama completo da problemática investigada, e uma delas é a Árvore de Decisão (Welgacz, 2007). Tal ferramenta parte do princípio de decomposição de um grande problema em subproblemas menores, e é construída a partir da seleção dos atributos considerados de maior relevância para o autor da mineração.

Nesse contexto, ficam estabelecidos os fatores de maior influência na questão trabalhada, sendo o fator de maior importância apresentado como o primeiro nó da árvore, enquanto os restantes são alocados nos nós subsequentes. Por meio dessa hierarquização da importância de cada atributo, e ainda por sua facilidade de compreensão, as árvores de decisão podem ser consideradas excelentes ferramentas de tomada de decisão (Crepaldi et al. 2011).

A aplicação das técnicas de mineração de dados tem se expandido expressivamente nos últimos anos, inclusive em modelos de negócios nos quais o armazenamento digital de dados não costumava ser utilizado. Hodiernamente, as organizações vêm produzindo enormes quantidades de dados obtidos a partir de suas operações e pesquisas, e nesse contexto, a mineração de dados está se tornando uma alternativa cada vez mais popular no que tange a descoberta de informações (Da Costa Cortes et al. 2002).

Além da vasta utilização das ferramentas de mineração em marketing, como no caso da sugestão de produtos ao usuário a partir da análise de compras anteriores, a mineração de dados está se tornando aplicável em campos da medicina, educação, finanças e outros (Amaral, 2016). Seus inúmeros benefícios incluem a tomada de decisões automatizada a partir da análise contínua dos dados, e a redução de custos, motivada pelas previsões precisas fornecidas pelo sistema.

Entretanto, é imprescindível mencionar que, embora as ferramentas de mineração de dados atuais sejam altamente automatizadas e façam a maior parte da descoberta do conhecimento pelo usuário do software, essa forma de manipulação de dados ainda depende da análise humana dos resultados obtidos. Assim, não se deve pensar na mineração de dados como um processo totalmente automático e fácil de ser realizado (Camilo e Silva, 2009).

Ainda, existem algumas limitações ligadas à mineração de dados que devem ser avaliadas. Como citado por Wang et al., (2008) um dos impasses na utilização da técnica de mineração inclui o alto nível de exigência na definição das relações entre os dados. Isso porque quando essas relações não são feitas de forma eficiente, criam-se relações errôneas entre os dados, o que leva, por sua vez, as informações inconsistentes com a realidade. O autor também cita a grande quantidade de dados como um obstáculo a ser superado, uma vez que se torna difícil ao executor da mineração selecionar, converter e revisar um volume exacerbado de dados, bem como interpretar os resultados decorrentes da mineração desse grande volume.

Outro conceito importante a ser mencionado é a mineração de dados educacionais, o que de acordo com Costa et al. (2013), constitui-se de uma área emergente na mineração de dados, com enfoque no desenvolvimento ou adaptação de métodos e algoritmos já

existentes, visando compreender os dados no contexto educacional. Assim, por meio do processo de mineração, é possível ampliar os horizontes da educação, contribuindo para a melhora na aprendizagem do estudante por meio da identificação dos entraves a essa melhora.

O processo de mineração de dados educacionais tem como princípio a conversão de conjuntos de dados provenientes de sistemas educacionais e ambientes virtuais de aprendizagem utilizados nas instituições de ensino, como por exemplo o Moodle – um ambiente virtual de aprendizagem utilizado pela UFSM, composto por um conjunto de interfaces, ferramentas e estruturas decisivas para a construção da interatividade e da aprendizagem (Silva, 2006). No caso do Moodle, essa potencialização da aprendizagem se dá pela disponibilidade de diversos recursos que permitem a interação entre o aluno e o professor e entre alunos, como o chat e o fórum, por exemplo (Rostas et al., 2009).

A prática da mineração de dados educacionais abrange o uso de diferentes tarefas, ou seja, técnicas de mineração selecionadas e aplicadas de acordo com os objetivos definidos para cada projeto. Nesse contexto, as tarefas de mineração de maior destaque incluem a predição, o agrupamento, a mineração de relações, a destilação de dados para facilitar decisões humanas e as descobertas com modelos (Baker, Isotani e Carvalho, 2011).

Na tarefa de predição, o objetivo central é desenvolver modelos que permitam deduzir informações sobre os aspectos dos dados disponíveis - as chamadas variáveis preditivas - partindo da análise e associação dos vários aspectos descobertos nos dados, denominados variáveis preditoras. Essa técnica pode ser utilizada para auxiliar no desenvolvimento de atividades instrucionais (Costa et al., 2013) e tem como vantagem a capacidade de predizer os dados de maior importância para um modelo (Baker, Isotani e Carvalho, 2011). Para a mineração de dados educacionais utilizando a tarefa de predição, as técnicas mais frequentemente utilizadas são a classificação, em que as variáveis preditivas são binárias ou categóricas, e a regressão, em que a variável preditiva é contínua (Costa et al., 2013).

A segunda tarefa de mineração passível de citação é o agrupamento, que consiste em dividir o conjunto de dados em grupos menores de acordo com o grau de semelhança entre eles. Uma característica importante do agrupamento é que esse método se constitui de uma forma de aprendizagem não-supervisionada, ou seja, as informações acerca dos dados não são conhecidas previamente, e sim descobertas por meio da associação das semelhanças entre os dados (Costa et al., 2013).

Uma terceira tarefa amplamente utilizada é a mineração de relações, a qual se concentra na descoberta de relações entre as variáveis pertencentes a um banco de dados composto por inúmeras dessas variáveis. Para tanto, escolhe-se uma variável de interesse e investigam-se as variáveis de relação mais expressivas com a escolhida. Existem diversas técnicas contidas no método de mineração de relações, sendo a principal delas a mineração de regras de associação, cuja premissa é formar regras de conhecimento, visando encontrar características ou tendências frequentes entre os conjuntos de dados. Esse princípio pode ser expresso pela proposição “Se X, então Y”, ou seja, aquilo que possui relação com X, também possui relação com Y (Costa et al., 2013).

Indo adiante, a destilação de dados para facilitar decisões humanas constitui-se de uma tarefa com enfoque na apresentação dos dados para a tomada de decisões, tendo como critérios a legibilidade e visualização dos dados e visando a ampla compreensão do usuário (Costa et al., 2013). De acordo com Baker (2010), o método de destilação dos dados possui dois propósitos centrais, sendo o primeiro deles a identificação, ou seja, a

apresentação dos dados de maneira que o usuário possa identificar os padrões com facilidade, e a classificação, que implica em utilizar a destilação com vista em construir modelos de predição.

Por fim, a última das tarefas mais comumente utilizadas é a descoberta de modelos, a qual parte de um modelo previamente desenvolvido por um método de predição ou agrupamento ou até mesmo construído manualmente. A partir disso, esse mesmo modelo é utilizado em uma segunda tarefa de mineração, como a predição ou mineração de relações (Costa et al. 2013).

2.5 *Software Waikato Environment for Knowledge Analysis (WEKA)*

WEKA é um software livre do tipo *open source*, ou seja, passível de execução, acesso e modificação pelo usuário. Ele foi desenvolvido na linguagem de programação JAVA por pesquisadores da Universidade de Waikato (Nova Zelândia) (Gonçalves, 2011). Essa ferramenta possui, implementados em seu código, inúmeros métodos de associação, classificação e agrupamento de dados. Ela ainda conta com ferramentas de adição, visualização e remoção de dados, além da visualização das informações geradas após a mineração (Goldschmidt e Passos, 2005). Outra vantagem do software é que sua interface permite o uso de seus algoritmos de aprendizagem e ferramentas para transformação de bases de dados sem que haja a necessidade de escrever códigos (Costa et al., 2013).

Nativamente, o WEKA utiliza seu próprio formato de dados como requisito para a mineração, o formato ARFF (do inglês "*Attribute-Relation File Format*"), caracterizado pelo uso do símbolo "@" para identificar os atributos presentes no banco de dados. Entretanto, existem outros formatos suportados pelo software, como é o caso do formato CSV (do inglês "*Character-separated Values*") (Morate, 2008). Inicialmente, o banco de dados obtido para o presente projeto foi convertido no formato CSV, possibilitando a conversão para o formato ARFF dentro do próprio WEKA posteriormente. Este último foi o formato utilizado para desenvolver a mineração de dados.

A partir da identificação das características propícias à evasão escolar, será possibilitado aos responsáveis dos cursos de graduação da Universidade Federal de Santa Maria a tomada de medidas preventivas e corretivas quanto à problemática em questão. Desenvolvido, o trabalho usar-se-á pelo curso de Sistemas de Informação do campus de Frederico Westphalen.

3. Trabalhos Relacionados

Nesta seção são apresentados alguns trabalhos com propostas semelhantes às do presente projeto. Primeiramente, é apresentado o trabalho de Paz e Cazzella (2017), que desenvolveram um estudo de caso de uma Universidade Comunitária utilizando a mineração de dados. Em seguida, é apresentado o trabalho de Manhães et al. (2012), que procuraram identificar algumas causas da evasão escolar. Por fim, apresenta-se o trabalho de Gonçalves, Da Silva e Cortes (2018), que identificaram alunos propensos à evasão escolar no ensino superior do Instituto Federal do Maranhão (IFMA) por meio da técnica de mineração de dados.

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

Paz e Cazella (2017) procuraram desenvolver um estudo de caso de cunho exploratório acerca da identificação do perfil de alunos propensos à evasão escolar. O estudo teve

enfoque em uma Universidade Comunitária, de onde foram coletados dados referentes aos alunos regularmente matriculados no segundo semestre de 2016.

Para o desenvolvimento do estudo de caso, foram utilizadas técnicas de mineração de dados. Especificamente, a mineração ocorreu por meio da tarefa de classificação dos dados na ferramenta WEKA, juntamente com a técnica de construção de árvores de decisão a partir do algoritmo J48. Os materiais utilizados na mineração foram cedidos pelos cursos de graduação de uma Universidade Comunitária do Rio Grande do Sul providos pelo setor de Tecnologia da Informação (TI).

De acordo com os resultados da mineração, foi possível aos autores concluir que alunos de semestres iniciais sem auxílio (financeiro, por exemplo) possuem uma alta incidência de evasão. A hipótese levantada inicialmente pelos autores de que morar fora da cidade onde se situa o campus levava à evasão foi descartada. Por último, foi contatado que há uma tendência geral de evasão nos semestres iniciais.

3.2 Identificação dos Fatores que Influenciam a Evasão em Cursos de Graduação Através de Sistemas Baseados em Mineração de Dados: Uma Abordagem Quantitativa

Manhães et al. (2012) investigaram os problemas relacionados a alunos evadidos em uma Universidade Federal brasileira, o que foi desenvolvido a partir de diversas técnicas de mineração de dados. A realização do projeto teve como base um estudo de caso envolvendo alunos de graduação da Universidade Federal do Rio de Janeiro (UFRJ), cujas informações foram mineradas por técnicas de mineração, visando localizar relações que direcionassem o desempenho acadêmico do discente.

As técnicas utilizadas foram avaliadas a fim de medir sua acurácia quando aplicadas a dados estudantis. Nesse contexto, o algoritmo *Naive Bayes* foi selecionado, devido a seu modelo interpretável e à capacidade de conversão gráfica de seus resultados numéricos, o que permite a análise quantitativa dos resultados posteriormente. A execução do algoritmo selecionado se deu no software WEKA, ferramenta escolhida para a realização da mineração devido às suas diversas versões de algoritmos implementados, sua modalidade *open source* e as opções de recursos estatísticos para a análise de dados.

Com os resultados da mineração, foi possível concluir que alunos evadidos apresentavam no mínimo uma disciplina reprovada por falta e média e pelo menos uma disciplina reprovada por média no primeiro semestre, redução no número de disciplinas cursadas e aprovações por semestre e média inferior às dos demais alunos ao final do primeiro semestre.

Quanto aos alunos que vieram a concluir o curso, foi mostrado que costumam manter o número de disciplinas cursadas por semestre, possuem alto índice de aprovações, suas médias semestrais se mantêm próximas do coeficiente de rendimento acadêmico acumulado até em torno do oitavo semestre e seu número de disciplinas reprovadas por média se mantém próximo de zero.

Em suma, foi observado que os alunos que concluíram o curso apresentaram um comportamento homogêneo entre os semestres que se sucederam, com uma tendência geral de aumento de reprovações por média ao final do curso – possivelmente em função de estágios curriculares ou outras atividades.

3.3 Técnicas de mineração de dados: um estudo de caso da evasão no ensino superior do Instituto Federal do Maranhão

Gonçalves, Da Silva e Cortes (2018) utilizaram a mineração de dados como recurso para identificar alunos propensos à evasão escolar no ensino superior do Instituto Federal do Maranhão (IFMA). O projeto foi desenvolvido no formato de estudo de caso, o qual foi composto pela identificação do problema, escolhas das técnicas de mineração para a adequação dos dados, escolha dos algoritmos e escolha dos parâmetros para validação.

Para a identificação do problema, foram construídos levantamentos teóricos acerca da evasão escolar no ensino superior e da aplicação da mineração de dados no contexto educacional, o que permitiu um discernimento maior por parte do grupo quanto à problemática. Em seguida, a segunda etapa se deu com o pré-processamento dos dados fornecidos à equipe, que foram adaptados às exigências dos algoritmos utilizados. Para isso, foram escolhidas as técnicas de pré-processamento *Information Gain* e *Correlation Based Feature Selection*. Ainda, foi definida uma terceira técnica manual para pré-processar os dados em consonância com as técnicas automatizadas.

Por fim, foram escolhidos três algoritmos para a etapa de mineração de dados: o *Naive Bayes*, *Support Vector Machine* e *J48*. A mineração de dados foi feita a partir do software WEKA e, por fim, foram escolhidas as métricas para comparar os resultados, o que ocorreu posteriormente com a utilização da interface *Experimenter* do WEKA. Os resultados foram avaliados na etapa final, buscando validar a utilidade das informações obtidas por meio da mineração.

Os resultados apontaram para uma tendência de evasão entre acadêmicos com tempo de curso inferior a três semestres. As taxas de acertos observadas alcançaram 94% para o *Naive Bayes,* 96% para o *Support Vector Machine* e 97% para o J48. O coeficiente de rendimento também recebeu destaque, sendo indicado pelo algoritmo J48 a propensão à evasão por alunos com rendimento menor ou igual a 5.0, exceto quando esse aluno permanece no curso por mais de nove semestres. O algoritmo classificado como mais eficiente foi o J48.

4. Solução Proposta

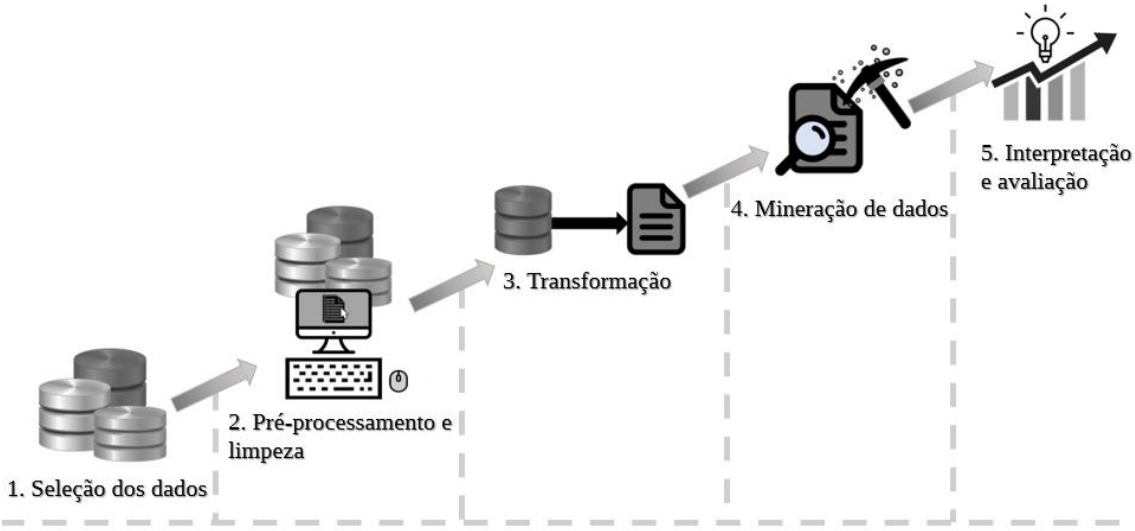
Para entender melhor a problemática da evasão escolar no curso de Sistemas de Informação da Universidade Federal de Santa Maria (UFSM), é preciso antes de tudo identificar os padrões dessa ocorrência para o curso em específico, reiterando a fala de Silva (2007) quando este afirma a diversidade de causas relacionadas a cada curso, o que impossibilita um padrão universal de causas e soluções.

Assim, tem-se como proposta para o presente projeto desenvolver uma metodologia utilizando mineração de dados para predizer alunos que tendem a evadir, visando auxiliar o gestor na tomada de decisão antes que o aluno abandone o curso.

O procedimento de descoberta do conhecimento ocorreu com base nos dados de alunos do curso de Sistemas de Informação da UFSM – Frederico Westphalen, e foi dividido em cinco etapas principais: seleção, pré-processamento e limpeza, transformação, mineração e interpretação dos resultados.

Na Figura 2 é possível visualizar as cinco etapas que constituíram esse procedimento. É importante ressaltar que, embora esse sistema seja composto por uma série de passos subsequentes, o fluxo do procedimento não é unidirecional, logo, é possível voltar a uma das etapas anteriores se preciso (Olson et al., 2008).

Figura 2. Definição das etapas da Descoberta do Conhecimento em Banco de Dados



Fonte: adaptado de Fayyad et al., 1996.

5 Etapas da descoberta do conhecimento

O conjunto de passos que constituiu a descoberta do conhecimento para o presente TGSI se deu da seguinte forma:

5.1 Seleção dos Dados

Primeiramente, o conjunto de dados foi obtido a partir das informações dos alunos disponíveis no Sistema de Informações de Ensino (SIE). Especificamente, foram selecionados os dados de alunos do curso de Sistemas de Informação da Universidade Federal de Santa Maria *campus* Frederico Westphalen. Não foi utilizado qualquer dado pessoal dos alunos que pudesse vir a identificá-los.

A seleção de dados abrangeu quatro planilhas contendo os dados referentes a todos os alunos que mantiveram vínculo com a instituição entre o segundo semestre de 2010 (início do curso) e o segundo semestre de 2019, fornecendo um total de 479 registros (alunos). Visando garantir a privacidade dos estudantes, seus códigos de identificação (ID), contidos em todas as planilhas utilizadas, foram substituídos por um código fornecido pela coordenação do curso. Esse código foi inserido em todas as planilhas utilizadas nas etapas seguintes.

A primeira planilha, denominada “ANUNOS_ELEMENTOS”, possuía dados referentes à manutenção do vínculo entre aluno e instituição, como o ano, semestre e forma de ingresso no curso (“ANO_INGRESSO”, “SEMESTRE_INGRESSO” e “INGRESSO”, respectivamente), situação do aluno ao final do semestre (“SITUACAO”), além do ano, semestre e forma de evasão (“ANO_EVASAO” e “SEMESTRE_EVASAO”). Ainda, apresentou alguns dados gerais sobre os alunos, como é o caso do gênero (“SEXO”), data de nascimento (“DATA_NASCIMENTO”), cidade e estado de naturalidade (“NATURALIDADE” e “UNIDADE_FEDERAL_NATURALIDADE”) e endereço atual de moradia (“ENDERECO_CIDADE”).

Já a segunda planilha, intitulada “ALUNOS_DISCIPLINAS”, contabilizava o desempenho dos estudantes em cada disciplina cursada, apresentando o ano e semestre

de curso da referida matéria (“ANO” e “SEMESTRE”, respectivamente), o código e nome do curso (“CODIGO_CURSO” e “NOME_CURSO”, referentes ao curso de Sistemas de Informação), o código e nome da disciplina (“CODIGO_DISCIPLINA” e “NOME_DISCIPLINA”) e a situação final do aluno ao concluir o semestre (“SITUACAO”).

Na sequência, a terceira planilha ou “MEDIA_NOTA” informou as médias semestrais dos alunos considerados, indicando os atributos referentes à média semestral (“MEDIA”), semestre e ano em que a média fora obtida (“SEMESTRE” e “ANO”), bem como o curso frequentado e seu código e a situação do aluno no referido semestre.

Por fim, a última planilha, nomeada “ALUNOS_FREQUENCIAS”, apresentou os dados de participação do estudante durante o período em que decorreu a disciplina, tais como o código e nome do curso, código, nome e carga horária da referida disciplina (“CODIGO_DISCIPLINA”, “NOME_DISCIPLINA” e “CARGA_HORARIA_TOTAL”) e os créditos referentes a mesma (“CREDITOS”). Por fim, também trouxe o código da turma em que o aluno estava inserido (CODIGO_TURMA), o número de aulas realizadas (AULAS) e o número de presenças e faltas do estudante (“PRESENÇAS” e “FALTAS”).

5.2 Pré-processamento e Limpeza dos Dados

A seguinte etapa teve por objetivo eliminar os dados considerados incompletos ou irrelevantes para o estudo, além de formar novos dados de importância com base na relação entre os já presentes. Foram considerados como evadidos os alunos que evadiram o curso por meio do abandono, transferência e cancelamento. O procedimento de pré-processamento e limpeza foi efetuado em uma planilha eletrônica e foi dividido em três etapas menores: limpeza dos dados, integração dos dados e redução dos dados (Han et al. 2011).

Visando a redução de dados, definiu-se como requisito de utilização dos registros a presença de pelo menos uma matrícula em qualquer disciplina do curso de Sistemas de Informação no referido semestre, de modo que os alunos aprovados na seleção que não tivessem se matriculado fossem excluídos. Como resultado, obteve-se uma redução significativa do conjunto de dados utilizado, que dos 479 dados iniciais, teve 409 registros obedecendo a regra, sendo 67 registros referentes a alunos formados, 207 a evadidos e 135 a alunos regulares, totalizando 70 registros excluídos por falta de matrícula.

No contexto de integração, os dados de interesse para a proposta da mineração foram integrados em três novas planilhas, cada qual criada com o objetivo de avaliar um dos três períodos definidos (2010 a 2019, 2010 a 2014 e 2015 a 2019). A primeira planilha recebeu o nome de “ELEMENTOS_2010_2019”, a qual seria utilizada para analisar os dados do período de 2010 a 2019.

Já a segunda planilha, intitulada “ELEMENTOS_2010_2014”, recebeu os dados de 2010 a 2014 para posterior análise. Por fim, a terceira planilha, denominada “ELEMENTOS_2015_2019” armazenou os dados obtidos entre 2015 e 2019. Inicialmente, as três planilhas de integração receberam dados referentes à idade (“IDADE”), gênero (“SEXO”), forma de ingresso (“INGRESSO”) e situação (“SITUACAO”) dos alunos, de acordo com os períodos definidos para cada uma. O objetivo dessa separação foi investigar uma possível diminuição da evasão a partir de 2015, ano em que a instituição estudada passou por uma significativa melhora em sua infraestrutura.

As informações referentes à situação do aluno em cada semestre foram padronizadas de modo que houvesse apenas três possibilidades para o aluno: regular, formado e abandono. Dessa forma, o status de cancelamento de matrícula, transferência e abandono foram todos classificados como abandono.

Uma das suspeitas levantadas durante o estudo acerca da evasão escolar no ensino superior foi a possível relação entre a evasão escolar no curso de Sistemas de Informação com a distância entre o *campus* e a moradia do estudante. Nesse contexto, tomou-se por objetivo calcular essa distância aproximada em quilômetros, o que foi feito por meio de uma interface de programação de aplicações (API) do Google Maps, cujos resultados foram inseridos no atributo "DISTANCIA", encontrado na planilha "ALUNOS_ELEMENTOS". Para tanto, optou-se por utilizar a seguinte fórmula, desenvolvida e inserida na planilha eletrônica:

```
=SUBSTITUIR(FILTROXML(SERVIÇOWEB("https://
maps.googleapis.com/maps/api/distancematrix/xml?
origins=cidade+campus&destinations=cidade+aluno&k
ey=YOUR_API_KEY");"//distance/text");" km";"")
```

A fórmula criada com o intuito de calcular as distâncias foi desenvolvida na planilha eletrônica LibreOffice Calc, utilizando a função "SERVIÇOWEB", pertencente à planilha eletrônica utilizada, com o objetivo de retornar um resultado do Google Maps (referente à distância entre a cidade residente do aluno e Frederico Westphalen) e alocar esse resultado em uma célula da planilha. Para filtrar o dado de interesse entre as informações devolvidas pela função "SERVIÇOWEB", foram utilizadas as funções "FILTROXML" e "SUBSTITUIR", todas pertencentes à planilha em uso.

A função "FILTROXML" extraiu especificamente a quilometragem do conjunto de informações devolvido pela API, enquanto a função "SUBSTITUIR" removeu outros dados deixados pela função "FILTROXML", restando apenas a distância em números na célula. Um entrave para a realização do cálculo das distâncias foi a semelhança entre os nomes de algumas cidades, o que tornou necessário o cálculo manual dessas distâncias, como foi o caso de Caiçara (localizada no Rio Grande do Sul) e Caiçara do Norte (pertencente ao Rio Grande do Norte).

Outra etapa importante do pré-processamento foi a discretização dos dados em conceitos, com o objetivo de padronizar dados muito diversos para sua utilização durante a mineração de dados. De maneira geral, as informações foram classificadas nos conceitos "MUITO BAIXO", "BAIXO", "MEDIO", "ALTO" e "MUITO ALTO", de acordo com a fórmula abaixo, responsável por estabelecer a amplitude dos intervalos que definiriam cada conceito.

$$\text{média} = (\text{máximo} - \text{mínimo})/\text{número de conceitos}$$

O primeiro conjunto de dados discretizado foi o referente à distância entre a cidade de moradia do aluno e a cidade do *campus*. Nessa perspectiva, o conceito da distância foi considerado "MUITO BAIXO" caso o aluno residisse até 20 km de distância do campus e "BAIXO" se essa distância fosse de 21 km a 40 km. Alunos residentes entre 41 km e 60 km de distância tiveram a distância discretizada como "MEDIO", enquanto a distância

de 61km a 80km foi classificada como “ALTO” e a distância de 81 km a 100 km como “MUITO ALTO”.

As fórmulas utilizadas na discretização dessas distâncias estão contidas no quadro 1, as quais também foram utilizadas na conceituação de outros atributos. Os conceitos obtidos foram adicionados ao atributo “DISTANCIA_CONCEITO”, adicionado às planilhas de integração.

Quadro 1. Fórmulas utilizadas para conceituar o atributo x

Conceito	Fórmula utilizada
MUITO BAIXO	mínimo >= x <= (mínimo + média)
BAIXO	(mínimo + média) > x <= (mínimo + 2 * média)
MEDIO	posição mediana
ALTO	(máximo - média * 2) > x <= (máximo - média)
MUITO ALTO	(máximo – média) > x <= (máximo)

Fonte: elaborado pelo autor (2020).

Nas fórmulas utilizadas, “x” pode assumir o valor referente à distância entre o aluno e o *campus;* a porcentagem de aprovações e reprovações e a porcentagem de presenças e ausências semestrais.

Em seguida, médias semestrais contidas na planilha “MEDIA_NOTAS” foram convertidas em conceitos representados pelas letras do alfabeto latino “A”, “B”, “C” e “D”, utilizando como base os critérios de conversão definidos pela UFSM. Nesse contexto, os alunos que obtiveram média semestral maior ou igual a 9,0 receberam conceito A, enquanto os alunos cuja média semestral manteve-se entre 7,0 e 8,9 receberam conceito B. As médias semestrais situadas entre 5,0 e 6,9 receberam conceito C, e por fim, as médias abaixo de 5 foram convertidas para o conceito D. Calculados os conceitos, estes foram colocados em uma nova coluna, intitulada “MEDIA_CONCEITO”, a qual foi adicionada nas três planilhas de integração. É possível observar os intervalos utilizados para discretizar as médias dos alunos no quadro 2.

Quadro 2. Intervalos definidos para conceituar as médias dos alunos

Conceito	Intervalo
A	média >= 9
B	7 <= média <= 8,9
C	5 <= média <= 6,9
D	média <= 4,9

Fonte: elaborado pelo autor (2020).

Em seguida, definiram-se os conceitos referentes ao número de aprovações obtidas em cada semestre, tomando como base a porcentagem de aprovações em função do número de disciplinas matriculadas por semestre. Dessa maneira, recebeu o conceito “MUITO BAIXO” quem obteve até 20% de aprovação, “BAIXO” quem recebeu entre 21 e 40% de aprovação, “MEDIO” os que obtiveram 41 a 60% de aprovação, “ALTO” para 61 a 80% de aprovação e “MUITO ALTO” para 81 a 100% de aprovação. A conceituação foi definida de acordo com as fórmulas apresentadas no Quadro 1, e os resultados foram contidos na coluna “APROVACAO_CONCEITO”, nas planilhas de integração.

O número semestral de reprovações, tal como o de aprovações, foi convertido em conceitos de acordo com a porcentagem de reprovação semestral e as fórmulas contidas no Quadro 1. Assim, os alunos com zero a 20% de reprovação receberam o conceito

“MUITO BAIXO”, enquanto estudantes com reprovação entre 21 e 40% obtiveram o conceito “BAIXO”. Adiante, alunos com 41 a 60% de reprovação contabilizada adquiriram o conceito “MEDIO”, ao passo que 61 a 80% de reprovação foram classificadas como “ALTO” e 81 a 100% reprovações como “MUITO ALTO”.

Outro conjunto de dados padronizado durante o pré-processamento foi o referente ao número de trancamentos parciais do estudante, ou seja, o trancamento de determinadas matérias ao decorrer de cada semestre. Assim, definiram-se critérios de conversão para os valores contabilizados por cada estudante, sendo considerado “MUITO BAIXO” para os alunos que não trancaram nenhuma matéria, “BAIXO” para um trancamento, “MEDIO” para dois trancamentos, “ALTO” para três trancamentos e “MUITO ALTO” para quatro trancamentos. Os conceitos obtidos foram organizados em uma nova coluna, intitulada “TRANCAMENTO_PARCIAL_CONCEITO”, copiada para as planilhas de integração. Os conceitos definidos de acordo com o número parcial de trancamentos pode ser observado no quadro seguinte.

Quadro 3. Intervalos definidos para conceituar o número parcial de trancamentos

Conceito	Intervalo
MUITO BAIXO	Nenhum trancamento
BAIXO	Um trancamento
MEDIO	Dois trancamentos
ALTO	Três trancamentos
MUITO ALTO	Quatro trancamentos

Fonte: elaborado pelo autor (2020).

Outra coluna, chamada “TRANCAMENTO_TOTAL_CONCEITO”, classificou os dados referentes ao trancamento total do aluno, em que ocorre a suspensão completa das atividades acadêmicas do estudante por período determinado, em “SIM”, para quem efetuou o trancamento total, ou “NÃO”, para os que não o fizeram em cada semestre. Os resultados alcançados, como todos até então, foram alocados nas planilhas de integração.

Por fim, os últimos conjuntos de dados conceituados foram os referentes às porcentagens de presença e ausência semestrais do aluno nas aulas. Nesse aspecto, tanto os conceitos referentes à presença quanto a ausência receberam os mesmos intervalos, sendo considerado “MUITO BAIXO” para zero a 20% de presença/ausência, “BAIXO” para o intervalo de 21 a 40% de presença/ausência, "MEDIO" estando entre 41 a 60% de presença/ausência, “ALTO” para 61 a 80% de presença/ausência e “MUITO ALTO” para 81 a 100% de presença/ausência. Ao final, os conceitos de presença foram inseridos na coluna “PRESENCA_CONCEITO”, enquanto os conceitos referentes à ausência foram colocados na colunas “AUSENCIA_CONCEITO”, ambas inseridas nas tabelas de integração. As fórmulas utilizadas na conceituação de presenças e ausências estão contidas no quadro 1.

Na fase de limpeza de dados, objetivou-se resolver as possíveis inconsistências encontradas no banco de dados obtido, como é o caso de campos vazios ou com dados desatualizados. Nessa perspectiva, outro problema encontrado durante o cálculo das distâncias entre as cidades e o *campus* foi a presença de informações de moradia desatualizadas, como cidades localizadas demasiadamente distantes para possibilitarem o deslocamento diário do estudante até a instituição. Portanto, as distâncias acima de 100 km referentes a alunos residentes no Rio Grande do Sul, foram convertidas à média das distâncias, recebendo o valor cinquenta, enquanto os valores acima de 100 km referentes a outros estados receberam o valor zero.

Do mesmo modo, a distância referente a moradores da cidade onde se situa o *campus* (Frederico Westphalen) foi transformada em zero. Outra ação relacionada à limpeza de dados foi a exclusão das informações de alunos que obtiveram classificação entre as vagas ofertadas pelo curso, mas não efetuaram sua matrícula, bem como de alunos de outros cursos matriculados em apenas uma disciplina do curso de Sistemas de Informação.

Ao final da etapa de pré-processamento e limpeza, as três planilhas de integração, que seriam utilizadas na mineração propriamente, continham os atributos "IDADE", "DISTANCIA_CONCEITO", "MEDIA_CONCEITO", "APROVACAO_CONCEITO", "TRANCAMENTO_TOTAL_CONCEITO", "SEXO", "REPROVACAO_CONCEITO", "TRANCAMENTO_PARCIAL_CONCEITO", "PRESENCA_CONCEITO", "AUSENCIA_CONCEITO" "INGRESSO" e "SITUACAO". A descrição de cada atributo pode ser observada no quadro 4.

Quadro 4. Atributos selecionados para a mineração de dados

Atributos	Descrição
IDADE	Idade do aluno
SEXO	Gênero do aluno
DISTANCIA_CONCEITO	Conceito referente à distância entre a moradia do estudante e o *campus.*
APROVACAO_CONCEITO	Conceito referente ao percentual de aprovação do estudante por semestre
REPROVACAO_CONCEITO	Conceito referente ao percentual de reprovação do estudante por semestre
PRESENCA_CONCEITO	Conceito referente ao percentual de presenças do estudante durante o semestre
AUSENCIA_CONCEITO	Conceito referente ao percentual de ausências do estudante durante o semestre
TRANCAMENTO_TOTAL_CONCEITO	Conceito referente à opção de trancamento total
TRANCAMENTO_PARCIAL_CONCEITO	Conceito referente ao número de trancamentos parciais por semestre
INGRESSO	Forma de ingresso na instituição
SITUACAO	Situação do aluno em cada semestre

Fonte: elaborado pelo autor (2020).

5.3 Transformação dos Dados

O terceiro passo da descoberta do conhecimento consistiu na transformação dos dados pré-processados. Em virtude de a segunda etapa ter sido desenvolvida na planilha eletrônica LibreOffice Calc, o volume de dados foi convertido inicialmente no formato "ODS" (Open Document Spreadsheet), que é nativo do software em questão. Todavia, o prosseguimento do processo de mineração de dados tornou necessária uma nova

conversão dos dados, visando adequá-los ao software a ser utilizado na fase seguinte, o WEKA. Nessa circunstância, o conjunto de dados foi exportado no formato CSV, que possui compatibilidade com o WEKA.

5.4 Mineração de Dados

Terminada a seleção, pré-processamento, limpeza e transformação dos dados, estes foram inseridos na ferramenta WEKA com vistas em iniciar a etapa de mineração de dados, a qual foi dividida em três subetapas distintas. A primeira etapa menor consistiu em minerar os dados referentes ao período de 2010 até 2014, enquanto a segunda avaliou os dados do intervalo entre 2015 e 2019. Ambas as etapas foram definidas de acordo com um interesse em comum, avaliar o impacto da melhora na infraestrutura da instituição nos índices de evasão escolares observadas.

Por fim, a terceira etapa abrangeu a análise do conjunto total de dados, ou seja, todo o decorrer de tempo entre 2010 e 2019, com o objetivo de obter uma visão geral da problemática trabalhada. O algoritmo selecionado para efetuar as três etapas da mineração de dados foi o J48, apontado em trabalhos relacionados como uma eficiente ferramenta para a proposta em questão (Gonçalves, Da Silva e Cortes, 2018).

Com o objetivo de balancear a quantidade de classes presentes no banco de dados exportado, optou-se por aplicar o filtro *Spread Subsample* da própria ferramenta Weka. Essa ação foi importante porque cada estudante recebia um status por semestre cursado, recebendo pelo menos oito status de "REGULAR" antes de receber um "FORMADO". Nessa perspectiva, o filtro produziu uma amostra aleatória dos dados entre a classe mais rara e a mais comum, alterando a proporção de registros dos alunos de 1:1, podendo assim, contornar a discrepância entre os dados a serem minerados.

Para avaliar o grau de acurácia dos resultados obtidos, utilizou-se a chamada estatística Kappa, que se baseia no número de respostas concordantes entre si, considerando a chance dessa concordância ter ocorrido ao acaso (Landis e Koch, 1977). Os valores utilizados pela estatística apresentam-se no intervalo entre zero e um, como é possível observar no quadro 5.

Quadro 5. Interpretação da estatística Kappa

K	Interpretação
Menor que 0	Nenhuma concordância
Entre 0 e 0,2	Leve concordância
Entre 0,21 e 0,4	Concordância regular
Entre 0,41 e 0,6	Concordância moderada
Entre 0,61 e 0,8	Concordância substancial
Entre 0,81 e 1	Concordância quase perfeita

Fonte: Landis e Koch. 1977.

O primeiro conjunto de dados minerado pelo WEKA foi o referente ao período entre 2010 e 2014, composto por 669 registros de alunos com situação "REGULAR", 64 registros como "ABANDONO" e 7 registros com a situação "FORMADO". Após a aplicação do filtro *Spread Subsample*, os dados foram limitados a 7 alunos regulares, 7 formados e 7 em situação de abandono. Utilizando as 21 instâncias contidas em 12 atributos disponíveis, o algoritmo J48 gerou uma árvore de decisão composta por 10 nós e 7 folhas, além de classificar o atributo "APROVACAO_CONCEITO" como nó raiz da árvore.

Em seguida, foi efetuada a etapa de mineração do conjunto de dados referente ao período de 2015 a 2019, que inicialmente contava com 1230 registros de alunos em situação “REGULAR”, 143 como “ABANDONO” e 60 como “FORMADO”. Com a aplicação do filtro *Spread Subsample*, o número de registros diminuiu para 60 alunos em cada uma das situações possíveis. O conjunto final abrangeu 180 instâncias, contidas em 12 atributos, e a aplicação do algoritmo J48 gerou uma árvore de decisão composta por 19 nós e 14 folhas, com o atributo “MEDIA_CONCEITO” sendo utilizado como nó raiz.

A terceira e última etapa da mineração foi a etapa relacionada aos dados pertencentes ao período de 2010 a 2019, composta inicialmente por 1899 registros de alunos com situação “REGULAR”, 207 com “ABANDONO” e 67 com “FORMADO”. As três situações diminuíram para 67 pela aplicação do filtro *Spread Subsample*, com o conjunto de dados final sendo formado por 201 instâncias e 12 atributos, os quais geraram, por meio do algoritmo J48, uma árvore de decisão composta por 19 nós e 14 folhas. O atributo “APROVACAO_CONCEITO” constou como nó raiz.

5.5 Interpretação dos resultados

É destacável a baixa incidência de alunos formados no período entre 2010 e 2014, fato esse que pode ser explicado pelo pouco tempo de existência do curso até aquele momento, o que tornaria necessário um desempenho acadêmico sem quaisquer reprovações ou trancamentos. Isso é importante porque o período entre o segundo semestre de 2010 e o segundo semestre de 2014 corresponde a 9 semestres, tempo necessário para a conclusão do curso.

Em função disso, a estatística Kappa obtida na primeira etapa da mineração foi de 0,1429, indicativa de leve concordância entre os resultados obtidos, o que pode ser explicado pela baixa quantidade de dados utilizada nessa etapa. A interpretação dos resultados obtidos pode ser observada na figura 3.

Figura 3. Regras de classificação geradas pela mineração de dados - período 2010 a 2014

Se APROVACAO_CONCEITO for igual a MUITO ALTO e IDADE for menor ou igual a 21, então o aluno é classificado com a situação REGULAR (3.0).
Se APROVACAO_CONCEITO for igual a MUITO ALTO e IDADE for maior que 21 e IDADE for menor ou igual a 27, então o aluno é classificado com a situação FORMADO (6.0).
Se APROVACAO_CONCEITO for igual a MUITO ALTO e IDADE for maior que 21 e IDADE for maior que 27, então o aluno é classificado com situação REGULAR (3.0/1.0).
Se APROVACAO_CONCEITO for igual a MUITO BAIXO, então o aluno é classificado com a situação ABANDONO (6.0/1.0).
Se APROVACAO_CONCEITO for igual a MEDIO, então o aluno é classificado com a situação REGULAR (0.0).
Se APROVACAO_CONCEITO for igual a ALTO, então o aluno é classificado com a situação REGULAR (2.0/1.0).
Se APROVACAO_CONCEITO for igual a BAIXO, então o aluno é classificado com a situação ABANDONO (1.0).

Fonte: elaborado pelo autor (2020).

No segundo período, entre 2015 a 2019, observa-se um aumento expressivo tanto no número de formados como de evadidos, o que pode ser esperado considerando o maior tempo de existência do curso, o qual possibilitou a conclusão do curso por alunos que não haviam se formado no período estudado anteriormente, bem como a conclusão por novos alunos. Assim, em relação à acurácia, a estatística Kappa obtida na segunda etapa foi de 0,5583, o que classifica a acurácia da mineração desse período como sendo de concordância moderada. As regras geradas nesse período estão disponíveis na Figura 4.

Figura 4. Regras de classificação geradas pela mineração de dados - período 2015 a 2019

Se MEDIA_CONCEITO for igual a D, então o aluno é classificado com a situação ABANDONO (69.0/16.0).
Se MEDIA_CONCEITO for igual a B e IDADE for menor ou igual a 21, então o aluno é classificado com a situação REGULAR (18.0/6.0).
Se MEDIA_CONCEITO for igual a B e IDADE for maior que 21 e PRESENCA_CONCEITO for igual a MUITO BAIXO, então o aluno é classificado com a situação FORMADO (15.0).
Se MEDIA_CONCEITO for igual a B e IDADE for maior que 21 e PRESENCA_CONCEITO for igual a MUITO ALTO e APROVACAO_CONCEITO for igual a MUITO BAIXO, então o aluno é classificado com a situação FORMADO (0.0).
Se MEDIA_CONCEITO for igual a B e IDADE for maior que 21 e PRESENCA_CONCEITO for igual a MUITO ALTO e APROVACAO_CONCEITO for igual a MUITO ALTO, então o aluno é classificado com a situação FORMADO (32.0/7.0).
Se MEDIA_CONCEITO for igual a B e IDADE for maior que 21 e PRESENCA_CONCEITO for igual a MUITO ALTO e APROVACAO_CONCEITO for igual a ALTO, então o aluno é classificado com a situação REGULAR (3.0).
Se MEDIA_CONCEITO for igual a B e IDADE for maior que 21 e PRESENCA_CONCEITO for igual a MUITO ALTO e APROVACAO_CONCEITO for igual a BAIXO, então o aluno é classificado com a situação FORMADO (0.0).
Se MEDIA_CONCEITO for igual a B e IDADE for maior que 21 e PRESENCA_CONCEITO for igual a MUITO ALTO e APROVACAO_CONCEITO for igual a MEDIO e IDADE for menor ou igual a 24, então o aluno é classificado com a situação FORMADO (2.0).
Se MEDIA_CONCEITO for igual a B e IDADE for maior que 21 e PRESENCA_CONCEITO for igual a MUITO ALTO e APROVACAO_CONCEITO for igual a MEDIO e IDADE for maior a 24, então o aluno é classificado com a situação REGULAR (2.0).
Se MEDIA_CONCEITO for igual a B e IDADE for maior que 21 e PRESENCA_CONCEITO for igual a ALTO, então o aluno é classificado com a situação REGULAR (1.0).
Se MEDIA_CONCEITO for igual a B e IDADE for maior que 21 e PRESENCA_CONCEITO for igual a BAIXO, então o aluno é classificado com a situação REGULAR (1.0).
Se MEDIA_CONCEITO for igual a B e IDADE for maior que 21 e PRESENCA_CONCEITO for igual a MEDIO, então o aluno é classificado com a situação FORMADO (1.0).
Se MEDIA_CONCEITO for igual a C, então o aluno é classificado com a situação REGULAR (20.0/3.0).
Se MEDIA_CONCEITO for igual a A, então o aluno é classificado com a situação FORMADO (16.0/4.0).

Fonte: elaborado pelo autor (2020).

Por fim, o período de 2010 a 2019 apresentou a totalidade dos dados registrados desde a criação do curso até o final de 2019. Dessa forma, a última etapa da mineração, de acordo com a estatística Kappa, apresentou concordância moderada, assumindo o valor 0,5224. As regras obtidas nesse período constam na Figura 5.

Figura 5. Regras de classificação geradas pela mineração de dados - período 2010 a 2019

Se APROVACAO_CONCEITO for igual a MUITO ALTO e IDADE for menor ou igual a 19 e IDADE for menor ou igual a 18, então o aluno é classificado com a situação REGULAR (2.0).
Se APROVACAO_CONCEITO for igual a MUITO ALTO e IDADE for menor ou igual a 19 e IDADE for maior que 18, então o aluno é classificado com a situação ABANDONO (2.0).
Se APROVACAO_CONCEITO for igual a MUITO ALTO e IDADE for maior que 19, então o aluno é classificado com a situação FORMADO (75.0/15.0).
Se APROVACAO_CONCEITO for igual a MUITO BAIXO e MEDIA_CONCEITO for igual a B e REPROVACAO_CONCEITO for igual a MUITO BAIXO, então o aluno é classificado com a situação FORMADO (2.0).
Se APROVACAO_CONCEITO for igual a MUITO BAIXO e MEDIA_CONCEITO for igual a B e REPROVACAO_CONCEITO for igual a MUITO ALTO, então o aluno é classificado com a situação ABANDONO (1.0).
Se APROVACAO_CONCEITO for igual a MUITO BAIXO e MEDIA_CONCEITO for igual a B e REPROVACAO_CONCEITO for igual a MEDIO, então o aluno é classificado com a situação REGULAR (0.0).
Se APROVACAO_CONCEITO for igual a MUITO BAIXO e MEDIA_CONCEITO for igual a B e REPROVACAO_CONCEITO for igual a BAIXO, então o aluno é classificado com a situação REGULAR (0.0).
Se APROVACAO_CONCEITO for igual a MUITO BAIXO e MEDIA_CONCEITO for igual a B e REPROVACAO_CONCEITO for igual a ALTO, então o aluno é classificado com a situação REGULAR (2.0).
Se APROVACAO_CONCEITO for igual a MUITO BAIXO e MEDIA_CONCEITO for igual a D, então o aluno é classificado com a situação ABANDONO (75.0/20.0).
Se APROVACAO_CONCEITO for igual a MUITO BAIXO e MEDIA_CONCEITO for igual a C, então o aluno é classificado com a situação REGULAR (2.0/1.0).
Se APROVACAO_CONCEITO for igual a MUITO BAIXO e MEDIA_CONCEITO for igual a A, então o aluno é classificado com a situação REGULAR (1.0).
Se APROVACAO_CONCEITO for igual a MEDIO, então o aluno é classificado com a situação REGULAR (13.0/4.0).
Se APROVACAO_CONCEITO for igual a ALTO, então o aluno é classificado com a situação REGULAR (13.0/5.0).
Se APROVACAO_CONCEITO for igual a BAIXO, então o aluno é classificado com a situação REGULAR (13.0/3.0).

Fonte: elaborado pelo autor (2020).

É destacável que, dentre as três análises realizadas, a que apresentou os melhores resultados de classificação foi a referente ao período 2015-2019, alcançando a classificação correta de 70,56% das instâncias analisadas e 0,56 na estatística Kappa, o que na visão de Landis e Koch (1977) é considerado de concordância moderada. A seguir, o segundo melhor resultado foi o referente a 2010 até 2019, o qual englobou todo o período analisado, apresentando 68,16% de instâncias corretamente classificadas e uma estatística Kappa em 0,52, considerada como de concordância moderada. A diminuição da estatística Kappa, nesse caso, pode ser explicada pelo número maior de registros regulares se comparada ao período de 2015 a 2019.

Por último, o período entre 2010 a 2014 apresentou os piores resultados das três avaliações, com 42,90% de instâncias corretamente classificadas e 0,14 referente à estatística Kappa. Para este conjunto, teve-se um número baixo de instâncias atribuídas à classe "FORMADO", pois a 1ª turma do curso se formou em 2014. Sendo assim, foi necessário balancear as classes, resultando em um número pequeno de instâncias.

Nota-se que, nos três períodos estudados, os atributos mais relevantes para a classificação no algoritmo foram o conceito referente à porcentagem de aprovação e a idade do aluno. Por exemplo, no modelo gerado pelo algoritmo J48 para a análise de 2010 a 2014, a maior probabilidade de evasão está associada ao conceito de aprovação baixo ou muito baixo do aluno no referido semestre.

5. Considerações Finais

O presente Trabalho de Graduação em Sistemas de Informação teve por objetivo investigar as causas da evasão escolar no ensino superior do curso de Sistemas de Informação da Universidade Federal de Santa Maria *campus* Frederico Westphalen, utilizando a mineração de dados como principal ferramenta. Os dados utilizados na mineração foram obtidos no Sistema de Informações de Ensino (SIE), que proveu informações acerca do desempenho semestral de cada um dos estudantes do curso avaliado, bem como dados referentes à idade, sexo e outras informações não limitadas ao desempenho em cada semestre.

Um dos objetivos propostos para esse trabalho foi analisar a relação entre os dados socioeconômicos dos alunos com a tendência à evasão, o que não foi possível devido ao não fornecimento dos dados solicitados pelo órgão responsável. Os resultados obtidos foram considerados satisfatórios, tendo em vista a baixa quantidade de atributos, restritos ao desempenho acadêmico.

Em uma visão geral dos casos, o algoritmo acabou por classificar atributos relacionados a média, aprovação, idade e presença do aluno, ou seja, atributos ligados diretamente ao estudante e seu desempenho acadêmico, que se mostraram fortes indicadores de evasão. O estudo também apontou que a distância entre a moradia do estudante e o *campus* não atuou como um fator relevante para a evasão no caso dos estudantes de Sistemas de Informação.

Para trabalhos futuros, recomenda-se a obtenção de mais atributos que possam ampliar as regras geradas pelo software, bem como a utilização de outros algoritmos que forneçam resultados mais satisfatórios. Essa metodologia também pode ser aplicada em dados de outros cursos da instituição, bem como de outros centros de ensino, com vista em expandir os conhecimentos acerca da evasão escolar no ensino superior.

Referências Bibliográficas

AMARAL, Daniela Patti do; OLIVEIRA, Fátima Bayma de. O Prouni e a conclusão do ensino superior: novas trajetórias pessoais e profissionais dos egressos. **Ensaio: avaliação e políticas públicas em educação**, v. 19, n. 73, p. 861-890, 2011.

AMARAL, Fernando. **Aprenda mineração de dados: teoria e prática.** Alta Books Editora, 2016.

AMARAL, João Batista do. Evasão discente no ensino superior: estudo de caso no Instituto Federal de Educação, Ciência e Tecnologia do Ceará (Campus Sobral). 2013.

ANDRIOLA, Wagner Bandeira. Doze motivos favoráveis à adoção do Exame Nacional do Ensino Médio (ENEM) pelas instituições federais de ensino superior (IFES). 2011.

APRILE, Maria Rita; BARONE, Rosa Elisa Mirra. Educação superior: políticas públicas para inclusão social. **Revista@mbienteeducação**, v. 2, n. 1, p. 39-55, 2018.

BAKER, Ryan; ISOTANI, Seiji; CARVALHO, Adriana. Mineração de dados educacionais: Oportunidades para o brasil. **Revista Brasileira de Informática na Educação**, v. 19, n. 02, p. 03, 2011.

BRASIL, Lei N° 12.711, De 29 De Agosto De 2012. “Lei de Cotas”, Brasília, DF. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/l12711.htm>. Acesso em: 13 jun. 2020.

CABENA, Peter et al. **Discovering data mining: from concept to implementation**. Prentice-Hall, Inc., 1998.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. **Universidade Federal de Goiás (UFC)**, p. 1-29, 2009.

COSTA, Evandro et al. Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. **Jornada de Atualização em Informática na Educação**, v. 1, n. 1, p. 1-29, 2013.

CREPALDI[1], Paola Guarisso et al. Um estudo sobre a árvore de decisão e sua importância na habilidade de aprendizado. 2011.

DA COSTA CÔRTES, Sérgio; PORCARO, Rosa Maria; LIFSCHITZ, Sérgio. **Mineração de dados-funcionalidades, técnicas e abordagens**. PUC, 2002.

DA EDUCAÇÃO SUPERIOR, MEC Censo. Brasília: Ministério da Educação, 2014. 2012.Disponível em: <http://www.andifes.org.br/wp-content/files_flutter/1379600228 mercadante.pdf>. Acesso em: 04 de jun. 2020.

DALRYMPLE, Douglas J.; PARSONS, Leonard J**. Introdução à administração de marketing.** Ltc, 2003.

DANTAS, Eric Rommel G. et al. O uso da descoberta de conhecimento em base de dados para apoiar a tomada de decisões. **V Simpósio de Excelência em Gestão e Tecnologia**, p. 1-10, 2008.

DA SILVA, Adelina Maria Pereira. Processos de ensino-aprendizagem na era digital. 2009.

DA SILVA, Ticiana LC et al. Análise em Big Data e um Estudo de Caso utilizando Ambientes de Computação em Nuvem. 2013.

DA SILVA, Wesley Vieira et al. Avaliação da escolha de um fornecedor sob condição de riscos a partir do método de árvore de decisão. **REGE Revista de Gestão**, v. 15, n. 3, p. 77-94, 2008.

D BAKER, Ryan SJ. Mining data for student models. In: **Advances in intelligent tutoring systems**. Springer, Berlin, Heidelberg, 2010. p. 323-337.

DE ALMEIDA TEODORO, Leonardo; KAPPEL, Marco André Abud. Aplicação de Técnicas de Aprendizado de Máquina para Predição de Risco de Evasão Escolar em Instituições Públicas de Ensino Superior no Brasil. **Revista Brasileira de Informática na Educação**, v. 28, p. 838-863, 2020.

DOURADO, Luiz Fernandes. Reforma do Estado e as políticas para a educação superior no Brasil nos anos 90. **Educação & Sociedade**, v. 23, n. 80, p. 234-252, 2002.

FAVERO, Rute Vera Maria. Dialogar ou evadir: Eis a questão!: um estudo sobre a permanência e a evasão na educação a distância. 2006.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37-37, 1996.

GAIOSO, Natalícia Pacheco de Lacerda. O fenômeno da evasão escolar na educação superior no Brasil. **Brasília, DF: Universidade Católica de Brasília**, p. 20, 2005.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data mining: um guia prático**. Gulf Professional Publishing, 2005.

GOMES, Alberto Albuquerque. Evasão e Evadidos: O discurso dos alunos sobre evasão escolar nos cursos de licenciatura. Nuances: estudos sobre Educação, v. 5, n. 5, 1999.

GONÇALVES, C. E. Data Mining com a ferramenta Weka. **Fórum de Software Livre de Duque de Caxias–2011**, 2011.

GONÇALVES, Tayná Costa; DA SILVA, Josenildo Costa; CORTES, Omar Andres Carmona. Técnicas de mineração de dados: um estudo de caso da evasão no ensino superior do Instituto Federal do Maranhão. **Revista Brasileira de Computação Aplicada**, v. 10, n. 3, p. 11-20, 2018.

HAND, David J.; MANNILA, Heikki; SMYTH, Padhraic. **Principles of data mining (adaptive computation and machine learning)**. MIT Press, 2001.

HAN, Jiawei; PEI, Jian; KAMBER, Micheline. **Data mining: concepts and techniques**. Elsevier, 2011.

JACOB, Celso Alencar Ramos. A evasão escolar e a construção do sujeito/profissional em curso de Ciências Econômicas. 2000.

JOHANN, Cristiane Cabral et al. Evasão escolar no Instituto Federal Sul-Rio-Grandense: um estudo de caso no campus Passo Fundo. 2012.

KIRBY, Peter G. et al. Adults returning to school: The impact on family and work. **The Journal of Psychology**, v. 138, n. 1, p. 65-76, 2004.

LANDIS, J. Richard; KOCH, Gary G. The measurement of observer agreement for categorical data. **biometrics**, p. 159-174, 1977. Disponível em: < https://www.jstor.org/stable/2529310?origin=crossref&seq=1>. Acesso em: 21 mar. 2020.

LEMOS, Anna Carolina Mendonça. Uma visão psicopedagógica do bullying escolar. **Revista Psicopedagogia**, v. 24, n. 73, p. 68-75, 2007.

MANHÃES, Laci Mary Barbosa et al. Identificação dos fatores que influenciam a evasão em cursos de graduação através de sistemas baseados em mineração de dados: Uma abordagem quantitativa. **Anais do VIII Simpósio Brasileiro de Sistemas de Informação, São Paulo**, 2012.

MANTOAN, Maria Teresa Eglér; PRIETO, Rosangela Gavioli. Inclusão escolar: o que é. **Por quê**, p. 12, 2003.

MARQUES, Waldemar. Expansão e oligopolização da educação superior no Brasil. **Avaliação: Revista da Avaliação da Educação Superior**, v. 18, n. 1, p. 69-83, 2013.

MORATE, Diego García. Manual de WEKA. Disponível através do e-mail diego. garcia. morate@ mail. com, 2008.

NEGRINE, Airton. Aprendizagem e desenvolvimento infantil: perspectivas psicopedagógicas. Porto Alegre: Prodil, v. 2, 1994.

NOGUEIRA, Cláudio Marques Martins et al. Promessas e limites: o Sisu e sua implementação na Universidade Federal de Minas Gerais. **Educação em Revista**, v. 33, 2017.

OLSON, David L.; DELEN, **Dursun. Advanced data mining techniques**. Springer Science & Business Media, 2008.

PAULA, Maria de Fátima Costa de. Políticas de democratização da educação superior brasileira: limites e desafios para a próxima década. **Avaliação: Revista da Avaliação da Educação Superior (Campinas)**, v. 22, n. 2, p. 301-315, 2017.

PAZ, Fábio; CAZELLA, Sílvio. Identificando o perfil de evasão de alunos de graduação através da Mineração de dados Educacionais: um estudo de caso de uma Universidade Comunitária. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. 2017. p. 624.

PEÇANHA, Plataforma Nilo. PNP 2019 (ano base 2018). 2019. Disponível em: < http://plataformanilopecanha.mec.gov.br/2019.html>. Acesso em: 11 out. 2020.

PEÇANHA, Plataforma Nilo. PNP 2020 (ano base 2019). 2020. Disponível em: <http://plataformanilopecanha.mec.gov.br/2020.html>. Acesso em: 11 out. 2020.

PIMENTA, Selma Garrido. Orientação e decisão: Estudo crítico da situação do Brasil. 1981.

PRIM, Alexandre Luis; FÁVERO, Jéferson Deleon. Motivos da evasão escolar nos cursos de ensino superior de uma faculdade na cidade de Blumenau. **Revista E-Tech: Tecnologias para Competitividade Industrial-ISSN-1983-1838**, p. 53-72, 2013.

ROSTAS, M. H. S. G.; ROSTAS, Guilherme Ribeiro. O ambiente virtual de aprendizagem (moodle) como ferramenta auxiliar no processo ensino-aprendizagem: uma questão de comunicação. **SOTO, U., MAYRINK, MF, GREGOLIN, IV, orgs. Linguagem, educação e virtualidade [online]. São Paulo: Editora UNESP**, 2009.

SANTOS JUNIOR, José da Silva; REAL, Giselle Cristina Martins. A evasão na educação superior: o estado da arte das pesquisas no Brasil a partir de 1990. Avaliação: Revista da Avaliação da Educação Superior (Campinas), v. 22, n. 2, p. 385-402, 2017.

SCHARGEL, Franklin P.; SMINK, Jay. Estratégias para auxiliar o problema de evasão escolar. **Rio de Janeiro: Dunya**, v. 282, 2002.

SILVA FILHO, Roberto Leal Lobo et al. A evasão no ensino superior brasileiro. **Cadernos de pesquisa**, v. 37, n. 132, p. 641-659, 2007.

SILVA, Nara Liana Pereira; DESSEN, Maria Auxiliadora. Crianças com síndrome de Down e suas interações familiares. **Psicologia: reflexão e crítica**, v. 16, n. 3, p. 503-514, 2003.

SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37-37, 1996.

TIGRINHO, Luiz Maurício Valente. Evasão Escolar nas Instituições de Ensino Superior. Revista Gestão Universitária, v. 173, p. 01-09, 2008. Disponível em: <http://gestaouniversitaria.com.br/artigos/evasao-escolar-nas-instituicoes-de-ensino-superior>. Acesso em: 21 de abr. 2020.

WANG, John; HU, Xiaohua; ZHU, Dan. Minimizing the minus sides of mining data. In: **Data Mining and Knowledge Discovery Technologies**. IGI Global, 2008. p. 254-279.

ZUIN, Antonio AS. Educação a distância ou educação distante? O Programa Universidade Aberta do Brasil, o tutor e o professor virtual. **Educação & Sociedade**, v. 27, n. 96, p. 935-954, 2006.