UNIVERSIDADE FEDERAL DO TOCANTINS
CÂMPUS UNIVERSITÁRIO DE PALMAS
CURSO DE CIÊNCIA DA COMPUTAÇÃO

ANÁLISE DE DADOS EDUCACIONAIS PARA DETERMINAÇÃO DE PERFIS DE EVASÃO UNIVERSITÁRIA

KLESLEY GONÇALVES FRANCISCO

PALMAS (TO)
2021

KLESLEY GONÇALVES FRANCISCO

ANÁLISE DE DADOS EDUCACIONAIS PARA DETERMINAÇÃO DE PERFIS DE EVASÃO UNIVERSITÁRIA

Trabalho de Conclusão de Curso I apresentado à Universidade Federal do Tocantins para obtenção do título de Bacharel em Ciência da Computação, sob a orientação do(a) Prof.(a) Dra. Anna Paula de Sousa Parente Rodrigues.

Orientador: Dra. Anna Paula de Sousa Parente Rodrigues

PALMAS (TO)

2021

KLESLEY GONÇALVES FRANCISCO

ANÁLISE DE DADOS EDUCACIONAIS PARA DETERMINAÇÃO DE PERFIS DE EVASÃO UNIVERSITÁRIA

Trabalho de Conclusão de Curso I apresentado à UFT – Universidade Federal do Tocantins – Câmpus Universitário de Palmas, Curso de Ciência da Computação foi avaliado para a obtenção do título de Bacharel e aprovada em sua forma final pelo Orientador e pela Banca Examinadora.

Data de aprovação: 18 / 8 / 2021

Banca Examinadora:

Profa. Dra. Anna Paula de Sousa Parente Rodrigues

Prof. Dr. Eduardo Ferreira Ribeiro

Prof. Dr. Warley Gramacho da Silva

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da Universidade Federal do Tocantins

F819a Francisco, Klesley Gonçalves.

Análise de dados educacionais para determinação de perfis de evasão universitária. / Klesley Gonçalves Francisco. – Palmas, TO, 2021.

81 f.

Monografia Graduação - Universidade Federal do Tocantins – Câmpus Universitário de Palmas - Curso de Ciências da Computação, 2021.

Orientadora : Dra. Anna Paula de Sousa Parente Rodrigues

1. Evasão Universitária. 2. Censo da Educação Superior Brasileiro. 3. Mineração de dados. 4. Perfis Estudantis. I. Título

CDD 004

Dedico este trabalho aos meus pais, pois é graças aos seus esforços e ensinamentos que hoje posso concluir essa graduação.

AGRADECIMENTOS

Em primeiro lugar, gostaria de agradecer à Professora Dra. Anna Paula de Sousa Parente Rodrigues por toda a atenção, ensinamentos, dedicação e tempo despendido em me ajudar a tornar este trabalho possível.

Minha imensa gratidão a Cassia Gabriela por ajudar a formatar e corrigir este trabalho, e por me encorajar, e às vezes até me obrigar a continuar estudando e desenvolvendo esta pesquisa.

Aos meus pais, Adailde e Clebson, à minha irmã Kethully, e aos meus avós paternos e maternos (mesmo aqueles que não puderam ver este momento), por todo o seu amor, carinho e apoio, que, embora fisicamente distantes, foram de suma importância para completar esta fase da minha vida.

Agradeço também a todos os meus colegas e amigos da Pró-Reitoria de Assuntos Estudantis - Proest, que durante esses anos, mesmo no trabalho, tornaram um pouco mais leve essa difícil jornada. Aos servidores: Inácio, Luciana, Laís, Roselena, Lilian, Eurizane, Valéria, Selma, Luciano, Edisselma, Ana Carolina, Kherlley, Elisama, Victor e Salmo. E para a minha "trupe" de estagiários Thailanny, Raimundo, Pedro, Maelly, Hayanna, Mateus, Irailton, Marina, Joan e João.

E agradeço também aos outros amigos, familiares e professores, que me apoiaram durante e no final desta graduação.

RESUMO

Um dos grandes desafios da atualidade enfrentado por gestores universitários, é manter o estudante na universidade e conduzi-lo à formação acadêmica dentro do prazo estabelecido. Diante disso, um dos principais objetivos deste trabalho é encontrar as diferentes características entre os alunos que formam e os que não formam, e assim criar os perfis estudantis para cada caso. Nesse contexto, este trabalho apresenta o desenvolvimento de modelos preditivos para o acompanhamento do aluno no decorrer dos quatro anos cursados, tendo alcançado valores de acurácia acima dos 60% e obtendo bons índices de confiabilidade dos resultados. Por tanto, conclui-se que este trabalho obteve resultados satisfatórios, onde os objetivos do trabalho são alcançados.

Palavra-chave: Evasão Universitária. Censo da Educação Superior Brasileiro. Mineração de dados. Perfis Estudantis.

ABSTRACT

One of the great challenges of today faced by university managers, is to keep the student at the university and lead him/her to academic training within the established deadline.Therefore, one of the main objectives of this work is to find the different characteristics between students who graduate and those who do not, and thus create student profiles for each case. In this context, this work presents the development of predictive models to monitor the student over the four years attended, having achieved accuracy values above 60% and obtained good reliability rates of the results. Therefore, it is concluded that this work obtained satisfactory results, where the work objectives were achieved.

Keywords: University Dropout. Brazilian Higher Education Census. Data Mining. Student Profiles.

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

Hoje, toda organização ao redor do mundo encara um aumento sem precedentes no volume de dados. Dados que se organizados e analisados podem gerar valor. Porém, para realizar isso, os métodos tradicionais como utilização de planilhas com possibilidade de consulta e geração de gráficos se tornam inviáveis por tal quantidade de dados, e por isso não sendo possível gerar tanto conhecimento sobre os dados dispostos. Por esses fatores, foram desenvolvidas técnicas para realização do que foi nomeado de Descoberta de Conhecimento em Banco de dados (Ou em inglês *Knowledge Discovery in Databases* - KDD). Esse é um processo não trivial para identificação de padrões válidos, não conhecidos, potencialmente úteis e interpretáveis, com o intuito de, basicamente, descobrir conhecimento relevante nos dados armazenados (FAYYD; SHAPIRO; SMYTH, 1996).

O processo de KDD possui várias etapas, que partem desde da seleção da amostra de dados ao conhecimento em si. Dentre essas etapas, existe uma que busca por padrões entre os dados inseridos, chamada de mineração de dados (ou em inglês *Data Mining* - DM). Nesse escopo, o objetivo da mineração de dados é descobrir correlações potencialmente uteis entre uma grande quantidade de dados ou encontrar regras quantitativas relacionadas aos mesmos, possibilitando a previsão de tendências e comportamentos, permitindo um novo processo de tomada de decisão, baseado principalmente no conhecimento acumulado (FAYYD; SHAPIRO; SMYTH, 1996).

Dentre as áreas de aplicação dessas técnicas, uma das mais buscadas nos últimos anos é a educacional. Um dos maiores problemas persiste na evasão do ensino, independente do meio (Ensino a distância ou presencial). Isso ocorre, pois a ineficiência na retenção dos alunos pode provocar significativo desperdício financeiro, uma vez que os recursos, ora investidos na instituição de ensino, não geram o retorno esperado e, ademais, as consequências abrangem os âmbitos pessoal, econômico e social (NAGAI; CARDOSO, 2017). Dessa forma, é percebido que se pode melhor aproveitar esse recurso desperdiçado em outro âmbito educacional ou reinvestindo em outras áreas da sociedade como Saúde, Segurança, Transporte, etc.

A capacidade de prever esses eventos de evasão, com uma antecedência desejável e de forma automática, é de grande interesse para diversas entidades e pesquisadores, pois com essas informações de antemão, administradores do âmbito educacional, por exemplo, teriam seu poder de decisão aprimorado, para desenvolver ações contra tais ocorrências, reduzindo, ou mesmo evitando, a evasão de estudantes.

Por esses motivos, o presente trabalho busca prever tendências de evasão universitária, por meio de algoritmos imbuídos dos conceitos de KDD, por se tratar de uma tecnologia comprovadamente eficaz em situações semelhantes (MACHADO; FRANCE-

LINO, 2020; PAL, 2012; SILVA; ADEODATO, 2012; MARTINS et al., 2017; HEGDE; PRAGEETH, 2018). Para gerar o valor proposto, é utilizada a base de dados do Censo da Educação Superior brasileiro, realizado anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - Inep.

Assim, o planejamento e a tomada de decisões, nas situações que envolvem o tema, tendem ser mais precisos e eficazes, visando evitar a desistência de estudantes e, assim, melhorando o nível intelectual da sociedade, gerando profissionais cada vez mais capacitados.

1.1 Justificativa

Segundo o Ministério da Educação, cerca de 40% dos estudantes que ingressam nas universidades públicas e 30% nas universidades privadas abandonam o curso antes da conclusão (FEDERAL, 2021).

Diante desse cenário, a capacidade de prever situações extremas e de risco relacionadas ao abandono universitário é de suma importância para os responsáveis pela tomada de decisão de uma universidade. Pois com essas informações seria possível direcionar, de forma eficaz, esforços para ações que realmente manteria os estudantes dentro da universidade e com formação no período adequado, podendo ainda melhor aplicar o dinheiro recebido dentro do ambiente universitário.

Portanto, a possibilidade de ajudar a população universitária, em todos os âmbitos, colaborando na tomada de decisão relacionada à melhoria da qualidade do ensino, utilizando de conhecimentos preditivos a respeito do perfil do aluno que serão gerados pela busca de padrões nos dados do censo da educação universitária, seria possível contribuir na realização de um planejamento preventivo bom, ou até mesmo ótimo, com uma eficaz execução deste, evitando situações problemáticas tanto para universidade quanto para o aluno.

1.2 Objetivos

1.2.1 Objetivo Geral

O objetivo principal deste trabalho é realizar a análise dos dados do Censo da Educação Superior Brasileiro dos últimos anos, utilizando técnicas de KDD para compreender indicadores de possíveis evasões universitárias. Isso com o intuito de auxiliar na tomada de decisões, nas ações preventivas e mitigadoras dos impactos relacionados à evasão universitária.

1.2.2 Objetivos Específicos

1. Realizar a coleta de dados do Censo da Educação Superior;

2. Realizar comparativo de performance de algoritmos de mineração dados em bases educacionais;

3. Desenvolver modelos preditivos para colaborar na identificação de indicadores de evasão universitária;

4. Identificar possíveis fatores, caso houver, de evasão universitária, com base nas variáveis inseridas no método utilizado.

5. Elaborar perfis universitários utilizando os dados já pré-processados do Censo da Educação Superior;

1.3 Organização do Trabalho

A dissertação tem a seguinte organização: o primeiro capítulo apresenta a introdução, assim como justificativa, objetivos geral e específicos.

No segundo capítulo é apresentado o levantamento bibliográfico necessário para o embasamento da solução proposta, começando pelo detalhamento do processo de Descoberta de Conhecimento em banco de dados, seguido pela teoria básica de evasão universitária para entender o que leva um estudante deixar a universidade.

No terceiro capitulo é apresentado trabalhos relacionados ao tema de descoberta de conhecimento em bancos de dados educacionais.

O quarto apresenta a metodologia do trabalho no qual será descrito quais instrumentos, fontes de dados e o detalhamento de todas as técnicas que foram utilizadas no trabalho.

Por fim, no último capitulo, são apresentadas as conclusões sobre este trabalho de pesquisa e as sugestões para trabalhos futuros.

2 LEVANTAMENTO BIBLIOGRÁFICO

O desenvolvimento tecnológico tornou possível armazenar facilmente grandes e múltiplos bancos de dados. Com o avanço tecnológico, novos desafios surgem, principalmente na área computacional e, consequentemente, na gestão estratégica. É consenso que os dados são considerados um bem, independentemente da sua natureza, seja comercial, administrativa, governamental, científica, etc.

Porém, não basta ter apenas um grande volume de dados, é preciso organizá-los, processá-los e analisá-los para aproveitá-los ao máximo e, portanto, gerar algum valor. No entanto, com essa quantidade de dados sem precedentes, é impraticável para um ser humano atingir esse objetivo de forma eficaz por métodos tradicionais; portanto, foram desenvolvidas técnicas para realizar o que foi denominado Descoberta de Conhecimento em Bases de Dados (ou em inglês Knowledge Discovery in Databases - KDD).

Neste trabalho, o foco está nos dados educacionais, com o intuito de reduzir o índice de evasão nas instituições brasileiras.

Assim, este capítulo tem como objetivo apresentar o referencial teórico abordado no presente trabalho. A seção 3.1 contém referências ao KDD e suas técnicas. A seção 3.2 trata de questões relacionadas ao abandono universitário no Brasil.

2.1 Descoberta de Conhecimento em Bases de Dados – KDD

Formalizado em 1989, o KDD é um processo que envolve a busca e interpretação de padrões em dados, utilizando algoritmos e análises nos resultados obtidos.

Existem dezenas de definições do que é KDD, porém a mais popular foi a proposta em 1996 por Fayyd, Shapiro e Smyth (1996): “KDD é o processo não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados”.

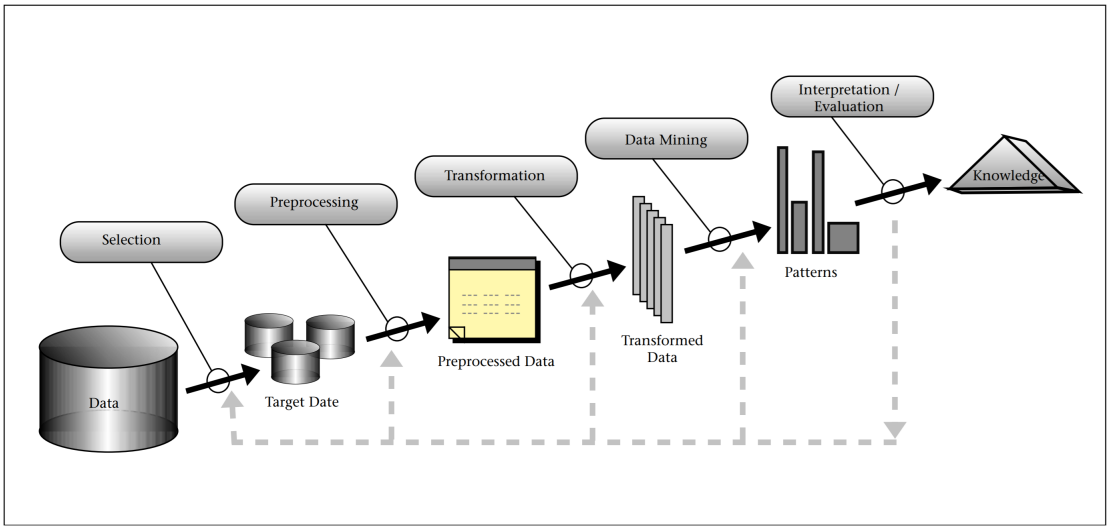
Para entender essa definição de forma mais precisa, é essencial observar alguns dos termos utilizados. “não trivial”, pois será necessário haver alguma inferência ou técnica de busca entre os dados selecionados, “previamente desconhecidas”, uma vez que a informação precisa ser nova para o sistema e de preferência para o usuário. E, “potencialmente úteis”, visto que a informação deve trazer consigo algum benefício para o usuário.

Os autores explicam ainda que o KDD possui duas características relevantes: é interativo e iterativo. Interativo porque o usuário pode decidir como controlar o fluxo entre as fases do processo. E Iterativo porque pode ser repetido quantas vezes forem necessárias na busca por melhores resultados, pois realiza sucessivas tentativas no refinamento de cada etapa para obter o máximo de proveito ao final do processo, portanto o resultado de cada fase depende de quão bem a anterior se sobressai.

Segundo Goldschmidt e Passos (2005), para que a Descoberta de Conhecimento em Bases de Dados alcance resultados úteis, dois tipos de especialistas precisam participar e interagir na execução do processo: Especialista em Domínio e o Especialista em KDD. O primeiro tipo representa a pessoa ou grupo de pessoas que conhece o assunto e o ambiente em que a aplicação de KDD deve ser executada. As informações fornecidas por esse grupo são de fundamental importância no processo, pois influenciam desde a definição dos objetivos até a avaliação dos resultados. E o segundo representa a pessoa ou grupo de pessoas com experiência na execução de processos em KDD. Esse especialista tem a tarefa de interagir com o especialista de domínio e direcionar o processo, definindo o que, como e quando cada ação do processo deve ser executada.

O processo de KDD possui uma série de fases, sendo elas: seleção, pré-processamento, transformação, mineração de dados e interpretação/avaliação. Pode-se perceber que esse processo abrange todo ciclo que o dado percorre para se tornar uma informação útil, ou real conhecimento. A Figura 1 apresenta essas fases.

Figura 1 – Fases do processo de KDD.



Fonte: (FAYYD; SHAPIRO; SMYTH, 1996)

Essas fases são agrupadas por (GOLDSCHMIDT; PASSOS, 2005) em três etapas operacionais: pré-processamento, mineração de dados e pós-processamento. A Figura 2 apresenta essas etapas.

Figura 2 – Etapas Operacionais do Processo de KDD.

Fonte: (GOLDSCHMIDT; PASSOS; BEZERRA, 2015)

As próximas seções têm como objetivo descrever as etapas operacionais do processo de descoberta de conhecimento em bases de dados. Em cada etapa, serão descritas algumas das funções principais e mais utilizadas em aplicações desta natureza.

Em Brachman e Anand (1996), é explicado que antes de iniciar o processo de KDD é necessário realizar dois passos antes. Em primeiro lugar, é necessário realizar uma análise para desenvolver uma compreensão do domínio da aplicação e conhecimento prévio da aplicação (Manuais, artigos que detalham os dados, entrevistas com especialistas, etc.). Em segundo lugar, é necessário identificar o objetivo da realização do processo levando em consideração o ponto de vista do usuário final.

Os autores explicam ainda, que um entendimento claro e objetivo reflete diretamente os objetivos a serem alcançados. Definições errôneas dessas etapas podem tornar os resultados obtidos no processo de mineração de dados totalmente desprezíveis e ineficazes para o que foi proposto.

2.1.1 Pré-Processamento

Esta etapa inclui as funções relacionadas à captura, organização, tratamento e preparação de dados para a próxima etapa: Mineração de Dados.

As bases de dados são dinâmicas, incompletas, redundantes, ruidosas e esparsas, exigindo um pré-processamento para organizá-las e "limpá-las"(NAVEGA, 2002).

O pré-processamento é de fundamental importância para o processo, pois realiza ações para corrigir dados errados e ajustar a formatação dos dados, de forma que algoritmos de mineração de dados sejam utilizados com total eficiência.

De acordo com Mannila (1996), esta etapa pode facilmente perdurar 80% do tempo necessário para todo o processo de descoberta de conhecimento, devido às dificuldades de integração de bases de dados heterogêneas.

Para Goldschmidt e Passos (2005), é importante destacar que a qualidade dos dados tem grande influência na qualidade dos modelos de conhecimento a serem desenvolvidos. Quanto pior a qualidade dos dados reportados ao processo de KDD, pior será a qualidade

dos modelos de conhecimento gerados, por isso a necessidade de realizar esta etapa com extrema cautela.

As próximas subseções apresentam as funções que são realizadas na etapa de pré-processamento.

2.1.1.1 Seleção de Dados

Esta função, também chamada de Redução de Dados, engloba essencialmente a identificação de quais informações das bases de dados existentes devem ser efetivamente consideradas durante o processo.

A seleção varia de acordo com os interesses e resultados esperados. Essas variáveis de interesse podem ser do tipo qualitativa(nominal ou ordinal) ou qualitativa(discreta ou continua)(LOUZADA; DINIZ, 2000)

De acordo com Fayyd, Shapiro e Smyth (1996), esta fase tem um impacto considerável na qualidade do resultado final do processo, uma vez que os dados podem vir de várias fontes e podem ter formatos diferentes, o que torna a fase extremamente complexa. Portanto, se não realizada com cuidado suficiente, a próxima fase, mineração de dados, pode gerar padrões inúteis fazendo com que informações insignificantes sejam descobertas.

Os dados são geralmente organizados em bancos de dados transacionais que são atualizados constantemente ao longo do tempo. Portanto, recomenda-se que seja feita sempre uma cópia dos dados para que o processo de KDD não interfira nas rotinas operacionais relacionadas ao banco de dados. (GOLDSCHMIDT; PASSOS, 2005).

Goldschmidt e Passos (2005) explicam que caso já exista uma estrutura de *Data Warehouse*(depósito de dados orientado a assunto, integrado, não volátil e com variação temporal, para apoiar as decisões de gestão), deve-se considerar a possibilidade de que esses sejam utilizados no processo, pois são conjuntos que já garantem dados confiáveis para o processo. Em outros casos, é comum reunir dados em uma única tabela. Isso se justifica porque a maioria dos algoritmos de mineração de dados pressupõe que os dados são organizados em uma estrutura tabular única, bidimensional, possivelmente muito grande. Portanto, o processo de KDD pode ocorrer independentemente da disponibilidade de *Data Warehouse*, mas a probabilidade de obtenção de resultados úteis será garantida com a utilização destes.

Goldschmidt e Passos (2005) esclarecem ainda que a união entre bases de dados pode ocorrer de duas maneiras: união direta e união orientada. Na direta, todos os atributos e registros do banco de dados transacional são incluídos na nova tabela, sem uma análise crítica sobre quais variáveis e quais casos podem realmente contribuir para o processo de descoberta de conhecimento. Na orientada, o especialista em domínio da aplicação, em colaboração com o especialista em KDD, seleciona atributos e registros com algum potencial de influência em bons resultados e desconsidera atributos que têm uma visão clara de nenhuma melhoria.

Assim, o objetivo da seleção de dados é identificar a fonte dos dados, e extrair um subconjunto dos dados necessários, por meio da seleção das variáveis de interesse para a aplicação da mineração de dados.

2.1.1.2 Limpeza dos Dados

Dados limpos e compreensíveis são requisitos básicos para que a etapa de mineração de dados seja bem-sucedida, como afirmam Louzada e Diniz (2000).

Esta função abrange qualquer tipo de tratamento que assegura a completude, veracidade e integridade dos dados, ou seja, pretende-se garantir a qualidade do conjunto de dados que já foi selecionado na função de pré-processamento anterior. Informações erradas, inconsistentes ou ausentes devem ser corrigidas para não comprometer a qualidade dos modelos de conhecimento que serão extraídos ao final do processo de KDD.

Para Fayyd, Shapiro e Smyth (1996), nesta fase são realizadas operações básicas no conjunto já definido. Essas operações incluem a remoção de ruídos, a coleta das informações necessárias para modelar ou contabilizar o ruído, escolher estratégias para lidar com campos de dados ausentes e contabilizar informações de sequência de tempo e mudanças conhecidas

A função limpeza dos dados envolve verificar a consistência das informações, corrigir possíveis erros, preencher ou excluir valores desconhecidos e redundantes, bem como eliminar valores que não fazem parte do domínio(GOLDSCHMIDT; PASSOS, 2005).

Goldschmidt e Passos (2005) também explicam que a execução desta fase tem por objetivo corrigir a base de dados, eliminando consultas desnecessárias que poderiam ser executadas no futuro pelos algoritmos de mineração de dados, afetando assim o seu desempenho.

2.1.1.3 Codificação dos Dados

Após selecionar e limpar os dados, inicia-se a função que precede a etapa de Mineração de Dados: a codificação dos dados ou também denominada por transformação dos dados.

Os algoritmos possuem padrões que devem ser respeitados, logo esta função de pré-processamento é realizada de acordo com o algoritmo de mineração que será utilizado na tarefa escolhida (CASTANHEIRA, 2008).

Conforme Goldschmidt e Passos (2005), nesta função os dados devem ser codificados para estar em um formato que possa ser usado como entrada para os algoritmos de mineração de dados. Essa codificação pode ser: Numérica - Categórica, que transforma valores reais em categorias ou intervalos, ou Categórica - Numérica, que representa valores de atributos categóricos numericamente.

2.1.1.4 Enriquecimento dos Dados

Uma última função da etapa de pré-processamento é mencionada por alguns autores na literatura: o enriquecimento dos dados já coletados.

Costa et al. (2019) explicam que a fase de enriquecimento adiciona novos dados agregando-os aos existentes, como por exemplo, a informação da cidade e região provenientes da análise dos prefixos dos telefones.

Goldschmidt e Passos (2005) também destacam que, se possível, novos dados devem ser adicionados ao conjunto que já havia sido criado na fase de seleção. Isso, para melhorar o processo de identificação de padrões e, por consequência, informações mais ricas ao final do algoritmo de minerador. Para complementar os dados, pesquisas podem ser realizadas, consultas a bancos de dados fora do domínio, entre várias outras técnicas.

2.1.2 Mineração de Dados

A etapa anterior tem grande relevância no resultado final, porém a etapa que segue recebe maior destaque na literatura, e muitas vezes é vista como sinônimo de todo o processo de KDD: Mineração de dados.

De acordo com Cabena et al. (1998), esta etapa é multidisciplinar, abrangendo principalmente as áreas de Banco de Dados, Aprendizagem de Máquina e Estatística. No trabalho de Zhou (2003), foi feita uma comparação entre essas três perspectivas, a saber:

- Han e Kamber (2001) com uma perspectiva da área de Banco de Dados: "É o processo de descoberta de conhecimento interessante a partir de grandes quantidades de dados armazenados em bancos de dados, armazém de dados ou outros repositórios de informações".

- Witten e Frank (2002) com uma perspectiva da área de Aprendizagem de Máquina: "É a extração de informações implícitas, previamente desconhecidas e potencialmente úteis."

- Hand, Mannila e Smyth (2001) com uma perspectiva da área de Estatística: "É a análise de conjuntos de dados, geralmente grandes, para encontrar relações insuspeitadas e resumir os dados de novas maneiras que sejam compreensíveis e úteis para o proprietário dos dados."

A partir dessas afirmações, pode-se concluir que mineração de dados é o conjunto de técnicas que possibilitam fazer a relação entre dados a fim de buscar padrões não conhecidos, mas existentes e que, consequentemente, proporcionam a extração de conhecimento dentro de uma grande base de dados.

Tan, Steinbach e Kumar (2009) os autores também explicam que as tarefas de *data mining* em geral são separadas em duas categorias:

- Tarefas de previsão: visam prever o valor de um atributo com base nos demais atributos, o atributo a ser previsto é conhecido como atributo alvo, enquanto os demais atributos são conhecidos como variáveis explicativas.

- Tarefas descritivas: visam fornecer padrões de correlações, agrupamentos e tendências, tarefas descritivas são frequentemente utilizadas de forma exploratória, requerendo técnicas de pós-processamento para validação dos dados.

Para Elmasri e Navathe (2011) mineração de dados pode ser utilizada para alcançar quatro classes de objetivos:

- Predição: Pode-se mostrar como determinados atributos dentro dos dados se comportarão no futuro, ou seja, é usada pra definir um provável valor para uma ou mais variáveis.

- Identificação: Padrões de dados podem ser usados para identificar a existência de um item, um evento ou uma atividade.

- Classificação: Pode-se construir um modelo para particionar dados em diferentes classes ou categorias que podem ser identificadas com base em combinações de parâmetros.

- Otimização: É possível otimizar recursos limitados para maximizar as variáveis de saída sob um determinado conjunto de restrições. As técnicas utilizadas neste objetivo são muito semelhantes às utilizadas na área de pesquisa operacional.

Este trabalho foca no objetivo de classificação. Nesta tarefa, os atributos do conjunto de dados são divididos em dois grupos. Um dos grupos contém somente um atributo, que corresponde ao atributo-alvo, ou seja, o atributo que deve ser previsto por um valor. O outro grupo contém os atributos a serem utilizados na predição do valor, denominados atributos previsores ou atributos de predição(GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

Em Louzada e Diniz (2000) é explicado que :

> "Deve-se destacar que cada técnica de *data mining* ou cada implementação especifica de algoritmos que são utilizados para conduzir as operações de *data mining* adapta-se melhor a alguns problemas que a outros, o que impossibilita a existência de um método de *data mining* universalmente melhor. Para cada particular problema tem-se um particular algoritmo. Portanto, o sucesso de uma tarefa de *data mining* está diretamente ligado à experiência e intuição do analista."

Com essa citação, é possível afirmar que não há nenhuma técnica que resolva todos os problemas de mineração de dados, pois diferentes técnicas atendem a propósitos distintos, cada uma contendo vantagens e desvantagens. Ou seja, a escolha do(s) método(s) é realizada com base no problema e no objetivo estratégico que se busca, sendo muito importante que haja pelo menos um profissional com experiência na área.

Existem dezenas de métodos que foram descobertos para desempenhar o papel de minerador de dados, mas algumas técnicas são comumente usadas porque trazem resultados mais consistentes do que outras. Os métodos que serão utilizados neste projeto são abordados na seção seguinte.

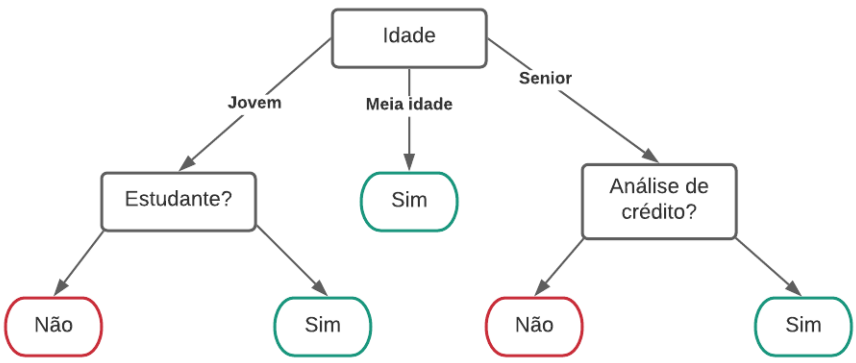
2.1.2.1 Métodos de Mineração de Dados

Os métodos escolhidos para esta pesquisa foram baseados no desenvolvimento do estado da arte, Capítulo 3. Em Kelly et al. (2019), uma comparação de métodos de mineração foi realizada no contexto de dados educacionais, os resultados desse trabalho constataram que o algoritmo *Random Forest* pode ser uma boa escolha em projetos semelhantes.

- *Decision Tree e Random Forest*

Antes de se aprofundar nos algoritmos escolhido, é importante entender o que são e como funcionam às *Decision Tree* (ou Árvore de Decisão). As árvores de decisão são uma forma simples e eficaz de representar o conhecimento. Essa técnica é embasada na abordagem "dividir para conquistar", ou seja sucessivas divisões são feitas no conjunto de dados utilizados para treino, vários subconjuntos, até que cada um destes subconjuntos pertençam a uma mesma classe, ou até uma das classes seja majoritária, não havendo necessidade de novas divisões (GARCIA, 2003).

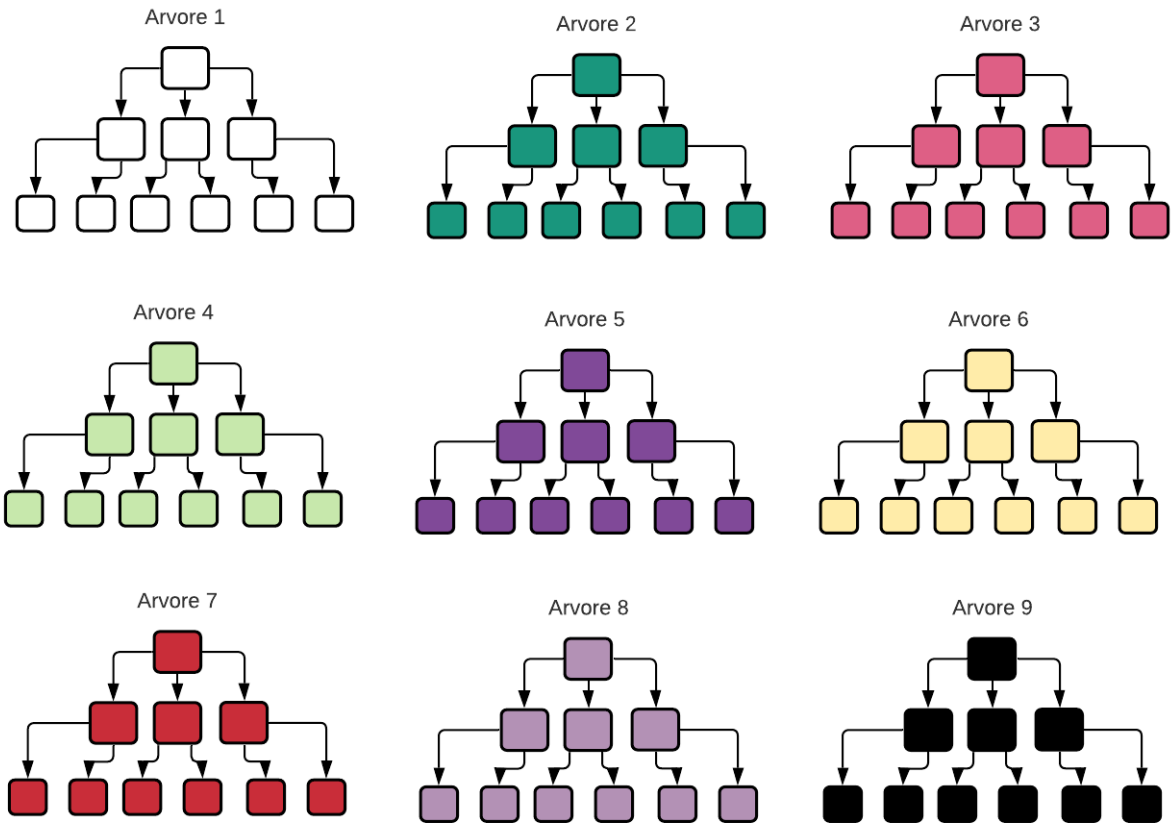
Figura 3 – Exemplo de uma árvore de decisão.



Fonte: (HAN; KAMBER; PEI, 2011)

A *Random Forest* (ou Floresta Aleatória), consiste em um grande número de arvores de decisão individuais que operam em conjunto. Cada árvore de decisão na floresta aleatória exibe uma previsão de classe e a classe com mais votos torna-se a previsão do modelo.

Figura 4 – Exemplo de uma Floresta Aleatória.



Fonte: Próprio Autor.

A vantagem deste algoritmo é o grande número de modelos (arvores) relativamente não correlacionados, que irão operar como um comitê, dessa forma é possível ter um desempenho melhor do que qualquer um dos modelos constituintes individuais.

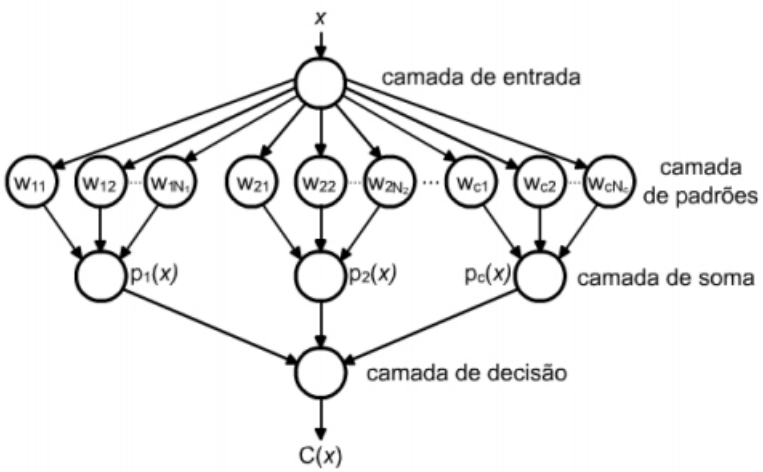
- *Probabilistic Neural Network (PNN)*

A rede neural probabilística (PNN) é uma algoritmo utilizado para problemas de classificação, são redes neurais inspiradas pelos classificadores bayesianos. A PNN possui quatro camadas: A de entrada, padrão, soma e saída. Essas camadas são apresentadas na Figura 5

A função de distribuição de probabilidade pai (PDF) de cada classe é aproximada por uma janela de Parzen e uma função não paramétrica. Então, usando o PDF de cada classe, a probabilidade da classe de um novo dado de entrada é estimada e a regra de

Bayes é então empregada para alocar a classe com maior probabilidade posterior para novos dados de entrada. Por este método, a probabilidade de classificação incorreta é minimizada. Mao, Tan e Ser (2000).

Figura 5 – Exemplo de uma *Probabilistic Neural Network*.



Fonte: (Ciarelli; Patrick; 2018)

2.1.2.2 Mineração de dados Educacionais

Com a informatização acadêmica, grandes quantidades de dados foram acumuladas neste meio; com dados sobre alunos, professores, instituições e demais agentes que fazem parte do processo de aprendizagem aos discentes e sua devida formação profissional. Conforme já explanado neste trabalho, com grandes bases de dados é possível extrair conhecimento para melhorar o processo de tomada de decisão e reflexão quanto às metodologias aplicadas em um determinado contexto. No ambiente acadêmico, isso pode ocorrer de diversas formas: redução de custos dentro das instituições, melhoria na qualidade do ensino, redução da evasão e retenção acadêmica, etc.

Nesse contexto, surgiu a Mineração de Dados Educacionais (do inglês, *Educational Data Mining* - EDM), que é definido como a área de pesquisa cujo foco principal é o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais, desta forma é possível compreender os alunos de forma mais eficaz e adequada, como aprendem, o papel do contexto em que ocorre a aprendizagem, além de outros fatores que influenciam a aprendizagem (PINHEIRO et al., 2014).

Uma definição de EDM pode ser encontrada no *site* da comunidade de *Educational Data Mining* (WWW.EDUCATIONALDATAMINING.ORG, 2021):

> "A mineração de dados educacional é uma disciplina emergente, preocupada em desenvolver métodos para explorar os dados únicos e cada vez mais em grande escala que vêm de ambientes educacionais e usar esses

> métodos para entender melhor os alunos e os ambientes nos quais eles aprendem."

Outra definição é apresentada por Baker, Isotani e Carvalho (2011):

> "A mineração de dados educacionais(EDM) é uma área recente de pesquisa que tem como principal objetivo o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais. Atualmente ela vem se estabelecendo como uma forte e consolidada linha de pesquisa que possui grande potencial para melhorar a qualidade do ensino."

Em Baker, Isotani e Carvalho (2011), citam duas possibilidades de descobertas importantes nos dados do meio acadêmico quando EDM é utilizado:

- Identificar em que situação um tipo de abordagem instrucional(aprendizagem individual ou colaborativa).
- Verificar se um aluno está desmotivado ou confuso e, assim, personalizar o ambiente e os métodos de ensino para oferecer melhores condições de aprendizagem

Em (BAKER; ISOTANI; CARVALHO, 2011) é apresentado a taxinomia das principais sub-áreas do EDM:

- Predição
  - Classificação
  - Regressão
  - Estimação de Densidade
- Agrupamento
- Mineração de relações
  - Mineração de Regras de associação
  - Mineração de Correlações
  - Mineração de Causas
- Destilação de dados para facilitar decisões humanas
- Descobertas com modelos

Em geral, o EDM não difere em muitos aspectos do processo de KDD quando aplicado em outras áreas, pois segue as mesmas etapas operacionais: pré-processamento, mineração de dados e pós-processamento. Porém, muitas vezes, os métodos de mineração de dados precisam ser modificados, pela necessidade de considerar a hierarquia da informação em diversos níveis e por uma falta de independência estatística nos tipos de dados encontrados ao coletar informações em ambientes educacionais (BAKER; ISOTANI; CARVALHO, 2011).

2.1.3 Pós-Processamento

Por fim, é realizado o pós-processamento da etapa de mineração de dados. Esta etapa consiste em visualizar, analisar, interpretar e avaliar os padrões, ou mesmo modelos, que foram indicados pelos algoritmos mineradores da etapa anterior. Segundo Goldschmidt e Passos (2005), nesta fase os especialistas em KDD e no domínio da aplicação avaliam os resultados obtidos e definem novas possibilidades de investigação dos dados.

Nesta etapa, além de ser possível visualizar os padrões extraídos, também é possível identificar alguns fatores que tornaram os resultados da etapa de mineração de dados não tão satisfatórios. Identificar e realizar ajustes, e então retomar para qualquer um dos estágios anteriores, podendo até ser reiniciado, conforme é mostrado na Figura 1.

Em Goldschmidt e Passos (2005), esta etapa é dividida em três passos:

1. Simplificações do Modelo de Conhecimento: Como o próprio nome sugere, consiste em retirar detalhes desse modelo de conhecimento para torná-lo menos complexo, sem perder informações relevantes. A representação do conhecimento por meio de regras é amplamente usada no KDD. No entanto, conjuntos com grandes quantidades de regras são difíceis de interpretar. É muito comum em Mineração de Dados que o usuário estabeleça limites mínimos de precisão e abrangência para as regras, de forma a excluir do modelo de conhecimento gerado todas as regras que não atendam a esses limites.

2. Transformações do Modelo de Conhecimento: Frequentemente, a fim de facilitar a análise de modelos de conhecimento, métodos de transformação podem ser usados nesses modelos. Esses métodos consistem basicamente na conversão da forma de representação do conhecimento de um modelo para outra forma de representação do mesmo modelo.

3. Organização e Apresentação dos Resultados: Os modelos de conhecimento podem ser representados de várias maneiras. Árvores, regras, gráficos em duas ou três dimensões, planilhas, tabelas e cubos de dados são muito úteis na representação do conhecimento. Em geral, as técnicas de visualização de dados estimulam a percepção e inteligência humana, aumentam a capacidade de compreensão e associação de

novos padrões. Portanto, oferecem subsídios para a escolha dos passos a serem seguidas no processo de KDD.

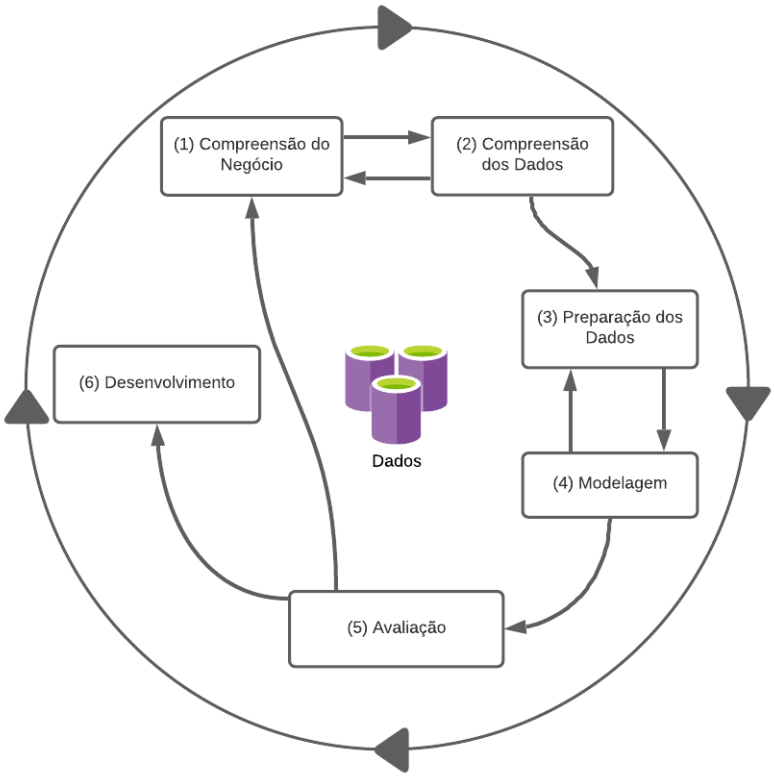
Assim, ao final desta etapa, os dados interpretados podem se tornar informações úteis e compreensíveis pelos usuários, desde que o objetivo de todo o processo tenha sido alcançado: extrair conhecimento.

2.1.4 Modelo de desenvolvimento CRISP - DM

Em projetos imbuídos em KDD, é importante seguir uma modelo de desenvolvimento. O modelo escolhido para este projeto foi o *Cross-industry standard process for data mining* (CRISP - DM); por se tratar de um dos primeiros e mais bem aceitos modelos no desenvolvimento de aplicações desse ramo (SHEARER, 2000).

Esse modelo consiste em seis fases principais: compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e desenvolvimento.

Figura 6 – Fases do modelo de desenvolvimento CRISP - DM.



Fonte: Adaptado de Shearer (2000)

Em Shearer (2000), é explanado sobre as fases apresentadas na Figura 6, a saber:

1. Compreensão do Negócio: Concentra-se em compreender os objetivos do projeto de uma perspectiva de negócios, convertendo este conhecimento em uma definição de

problema de mineração de dados, e então desenvolver um plano preliminar projetado para atingir os objetivos. Para entender quais dados devem ser analisados posteriormente, e como, é vital para os profissionais de mineração de dados compreenderem totalmente o negócio para o qual estão encontrando uma solução.

2. Compreensão dos Dados: Esta fase inicia com uma coleta inicial de dados. O analista então prossegue para aumentar a familiaridade com os dados, para identificar problemas de qualidade de dados, para descobrir *insights* iniciais sobre os dados ou para detectar subconjuntos interessantes para formar hipóteses sobre informações ocultas. A fase de compreensão dos dados envolve quatro etapas, incluindo a coleta dos dados iniciais, a descrição dos dados, a exploração dos dados e a verificação da qualidade dos dados.

3. Preparação dos Dados: Abarca todas as atividades para construir o conjunto de dados final ou os dados que serão alimentados na(s) ferramenta(s) de modelagem a partir dos dados brutos iniciais. As tarefas incluem tabela, registro e seleção de atributos, bem como transformação e limpeza de dados para ferramentas de modelagem. As cinco etapas na preparação de dados são a seleção de dados, a limpeza de dados, a construção de dados, a integração de dados e a formatação de dados.

4. Modelagem: Nesta fase, várias técnicas de modelagem são selecionadas e aplicadas e seus parâmetros são calibrados para valores ideais. Normalmente, existem várias técnicas para o mesmo tipo de problema de mineração de dados. Algumas técnicas possuem requisitos específicos na forma dos dados. Portanto, pode ser necessário voltar à fase de preparação de dados. As etapas de modelagem incluem a seleção da técnica de modelagem, a geração do design de teste, a criação de modelos e a avaliação de modelos

5. Avaliação: Antes de prosseguir para a implantação final do modelo construído pelo analista de dados, é importante avaliar mais detalhadamente o modelo e revisar a construção do modelo para ter certeza de que atende adequadamente os objetivos de negócios estabelecidos. Aqui é fundamental determinar se alguma questão importante de negócios não foi suficientemente considerada. No final desta fase, o líder do projeto deve decidir exatamente como usar os resultados da mineração de dados. As principais etapas aqui são a avaliação dos resultados, a revisão do processo e a determinação das próximas etapas.

6. Desenvolvimento: A criação do modelo geralmente não é o fim do projeto. O conhecimento adquirido deve ser organizado e apresentado de uma forma que o cliente possa usá-lo, o que muitas vezes envolve a aplicação de modelos dentro dos processos

de tomada de decisão de uma organização, como a personalização em tempo real de páginas da Web ou pontuação repetida de bancos de dados de marketing.

Shearer (2000) ainda explica, que um dos motivos do sucesso deste modelo no mercado e na literatura, se deve ao fato de ser possível realimentar e reajustar qualquer parte de seu desenvolvimento e assim melhorar os resultados da aplicação. a um nível satisfatório ou ótimo.

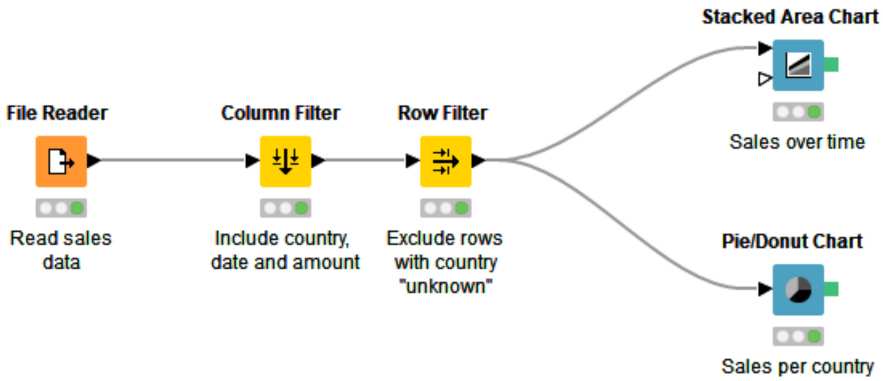
2.1.5 KNIME

Existem inúmeras ferramentas de mineração de dados disponíveis, para uso comercial e acadêmico, que fornecem as mais variadas coleções de algoritmos de pré-processamento dos dados, de mineração de dados, técnicas de visualização de dados, entre outros recursos. Para esse trabalho foi escolhido o *Konstanz Information Miner*(KNIME).

O KNIME é um software de código aberto para a criação de aplicações e serviços de ciência de dados. É intuitivo, código aberto e com integração contínua de novos desenvolvimentos. Essa plataforma torna a compreensão de dados e o projeto de fluxos de trabalho(*workflow*) de ciência de dados com componentes reutilizáveis acessíveis a todos.

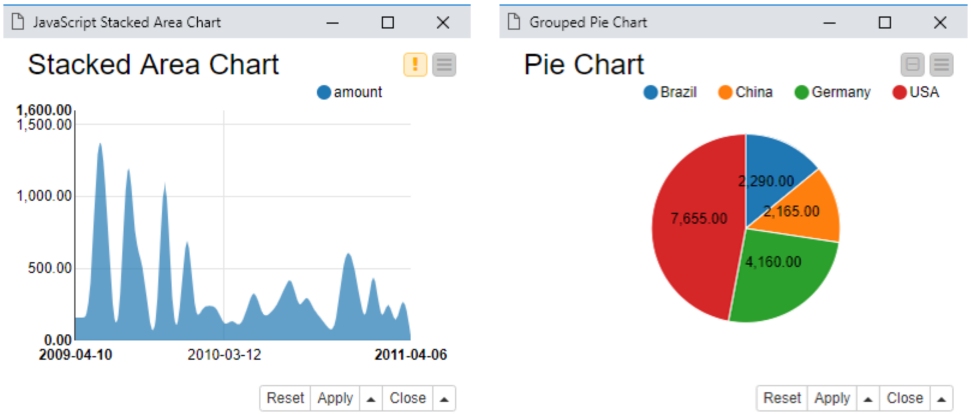
No KNIME, é utilizado de forma intuitiva no qual é possível criar um *workflow* visual, pois conta com uma interface gráfica no estilo arrastar e soltar, sem a necessidade de codificação. Porém, caso o projeto necessite de alguma implementação extra, é possível a integração com as principais linguagens de programação utilizadas para ciência de dados, incluindo integrações com outras plataformas de *Business Intelligence* como *Power BI*, *Tableau*, *Weka* e outras.

Figura 7 – Exemplo de *workflow* na plataforma(KNIME, 2021)



Fonte: (KNIME, 2021)

Figura 8 – Exemplos de visualizações na plataforma.



Fonte: (KNIME, 2021)

2.2 Evasão Universitária

2.2.1 Definição

Existem múltiplos conceitos para o termo Evasão. No contexto universitário, uma das principais definições é apresentada por Mazzetto, Bravo e Carneiro (2002), que conceituam o fenômeno "como toda e qualquer forma de saída do estudante do curso que não tenha sido pela diplomação". Esse conceito também é defendido por Tinto (1975), que se refere à evasão como o discente que se desliga da faculdade e nunca recebe o diploma. Já para os autores Mallada (2011), a evasão é falta de atenção em prol do estudante, resultando no seu justificado abandono. No geral, a evasão universitária é uma interrupção no processo de formação do acadêmico, impossibilitando-o de concluir o curso (PREVE et al., 2017).

De acordo com Lobo (2012), é difícil padronizar e definir tudo aquilo que diz respeito à evasão. Logo, em primeiro lugar, é necessário ter clareza e explicitar o tipo de evasão que está sendo mencionado no estudo, já que há diferentes tipos: evasão do curso, evasão da Instituição de Ensino Superior (IES) e a evasão do sistema; que são apresentadas a seguir (LOBO, 2017).

A Evasão do Curso é aquela em que o aluno deixa um curso por qualquer razão: muda de curso mas permanece na instituição, ou muda para um outro curso de outra IES. Algumas instituições não consideram a mudança de curso uma evasão propriamente dita. Na verdade, estritamente falando, toda vez que um aluno deixa de estudar em um curso, por qualquer razão, o curso teve uma perda, ou seja, houve uma evasão que precisa ser analisada (LOBO, 2017).

A Evasão da Instituição refere-se às evasões onde o aluno deixa a IES, mas não deixa o sistema de ensino superior, ou seja, é aquela onde o aluno troca de instituição. O problema mais comum das evasão da IES, se refere à prática de responsabilizar às

questões de ordem financeira do aluno como sendo a grande e única causa de Evasão da IES. Ou seja, não há uma perspectiva de que a evasão pode ocorrer em razão de problemas acadêmicos e administrativos. Por isso, esse tipo de evasão acaba não sendo tratada como deveria ser: como um problema de gestão institucional (LOBO, 2017).

A Evasão do Sistema trata-se de quando o aluno deixa de estudar e abandona o sistema de ensino, ou seja, não está estudando em nenhuma IES. A Evasão do Sistema é exatamente aquela que exige políticas públicas, que vão além das questões institucionais, acadêmicas ou até das individuais, visto que a evasão é considerada um dos mais sérios problemas de um sistema educacional de qualquer nível, em qualquer lugar do mundo (LOBO, 2017).

2.2.2 Cálculo da evasão

Geralmente, é utilizado regras padrões para realizar o cálculo da evasão em diferentes países, porém, como nem sempre é possível acompanhar individualmente cada aluno, apenas os números gerais são estudados; na maioria das vezes a partir da soma da evasão do conjunto dos diferentes cursos de todas as IES do sistema de ensino superior (LOBO, 2017). Um dos problemas acerca do cálculo da evasão, é que nem sempre é possível saber se o aluno abandonou os estudos de forma temporária ou definitiva. Se o afastamento for temporário, ou seja, se houve apenas um trancamento de estudos, o aluno poderá retornar após um, ou alguns períodos letivos. Com isso, é preciso definir como, e se, os alunos em situação de trancamento vão compor o cálculo da Evasão (LOBO, 2017).

Pelos motivos acima mencionados, muitos países adotam uma fórmula padrão para medir a Taxa de Graduação, que se refere ao número de alunos que se formam anualmente no conjunto de cursos, a partir da qual pode ser calculada a Taxa de Evasão. Sem padronização, nunca seria possível fazer cálculos para comparar diferentes países, ou mesmo estabelecer parâmetros ou propor políticas gerais para melhorar a taxa de titulação e também para combater a evasão (LOBO, 2017).

De acordo com o Instituto Lobo (LOBO, 2017), a fórmula mais correta para realizar os cálculos da evasão nacional, é apresentada por Roberto Lobo, presidente do instituto. O cálculo da fórmula não leva em conta a origem do ingressante: se ele ingressou no curso por meio de processo seletivo, por transferência de curso dentro da mesma instituição ou por transferência de outra instituição.

A fórmula permite que o cálculo de evasão seja feito a partir dos números apresentados nos Censos da Educação Superior Brasileira. Para realizar o cálculo e estimar a taxa de evasão anual do sistema, das IES ou dos cursos, é utilizada a taxa de permanência, ou seja, o número de alunos que permaneceram no curso. A fórmula é apresentada a seguir (LOBO, 2017):

$$P = [M(n) - Ig(n)]/[M(n-1) - Eg(n-1)]$$

Onde:

P = Permanência
n = Ano
M(n) = Matrículas em um determinado ano
M(n-1) = Matrículas do ano anterior
Eg(n-1) = Egressos do ano anterior (ou seja, concluintes)
Ig(n) = novos ingressantes (no ano n)

O índice de evasão é dado pela diferença do resultado da fórmula acima (a taxa de permanência) em relação à 100%, ou seja:

$$\text{Evasão} = (1 - P) * 100$$

A fórmula apresentada pode ser adaptada para realizar outros cálculos sobre a evasão, como a Evasão por IES e do Sistema, retirando-se do cálculo os ingressantes oriundos de transferências entre IES, ou entre cursos na mesma IES (LOBO, 2017).

2.2.3 As principais causas da evasão universitária

A evasão universitária é um fenômeno que vem assumindo grandes proporções e, consequentemente, tem sido objeto de diversos estudos. De acordo com Berger e Lyon (2005), os primeiros estudos e avanços científicos sobre a evasão de alunos tiveram início nos anos 50, nos Estados Unidos, devido ao aumento das instituições de ensino superior e da quantidade de estudantes (ASTIN et al., 2012).

No começo da década de 80 houve uma estagnação no crescimento das matrículas, com isso os esforços das instituições para atrair e manter os estudantes teve um grande crescimento. Foi somente a partir dos anos 90 que os estudos sobre evasão começaram a evoluir de forma rápida, surgindo teorias e modelos sobre o tema, como por exemplo as teorias que explicam os motivos que levam um aluno a abandonar o curso (CISLAGHI et al., 2008).

Os dois modelos mais conhecidos que tratam a evasão de estudantes, foram apresentados por Fishbein e Ajzen (1977) e Ethington (1990). O primeiro modelo, apresentado por Fishbein e Ajzen, diz que “o estudante analisa as interações com o cotidiano, age segundo o sentido que ele lhe dá, e pela aceitação, ou rejeição da ideia de que a formação

superior é significativa para sua vida futura". O segundo modelo é o de Ethington (1990), onde o autor desenvolveu um modelo psicológico em que foram incluídas as metas que os estudantes estabeleciam para si próprios. De acordo com os dois autores, a origem demográfica e as influências pessoais afetam diretamente seus valores, expectativas e aspirações dos estudantes, e influenciam sua decisão de permanecer, ou evadir-se (LOBO, 2017).

O Instituto Lobo, após mais de 12 anos de estudos, pesquisas e consultorias sobre Ensino Superior, conseguiu identificar as causas mais comuns da evasão de estudantes universitários. São elas (LOBO, 2017):

1. A baixa qualidade da educação básica: que é avaliada pelos exames nacionais aplicados e é largamente anunciada e discutida, com ênfase cada vez maior, nos mais diferentes segmentos da sociedade brasileira.

2. A baixa eficiência do ensino médio: que não garante a suficiência de competências do candidato ao Ensino Superior, criando dificuldades de adaptação e acompanhamento do curso.

3. A falta de recursos financeiros: esse aspecto envolve mais os alunos do setor público que, em muitos casos, deixam de estudar por não terem meios financeiros de se manter.

4. A limitação das políticas de financiamento ao estudante: mesmo com a presença dos programas Fundo de Financiamento Estudantil (FIES) e Programa Universidade para Todos (PROUNI), ainda são largamente insuficientes a quantidade de vagas para essas bolsas, não atendendo a todos os estudantes que precisam.

5. A escolha precoce da especialidade profissional: esse aspecto se refere ao aluno que se vê obrigado a realizar, ainda com pouca idade, a escolha de sua profissão, em razão da estrutura e da regulamentação do ensino brasileiro.

6. A dificuldade de mobilidade estudantil: alguns exemplos são a transferência entre as instituições nacionais (em especial para as instituições públicas) e o aproveitamento dos créditos cursados em outra instituição.

7. A dificuldade da autorização/reconhecimento de cursos: inovar os projetos pedagógicos dos cursos é um risco, em especial nas instituições privadas, já que cada Comissão de Autorização e/ou Reconhecimento defende a visão de seus integrantes, nem sempre a mais moderna, ou viável.

8. A legislação sobre a inadimplência no Brasil: o excesso de líderes que educa para o calote, favorece o acúmulo de dívidas pelo aluno e a Evasão das instituições privadas;

9. A enorme quantidade de docentes despreparados para o ensino e para lidar com o aluno real: o que ocorre, entre muitas razões, pela falta de formação didático pedagógica de vários docentes; isso somado à dificuldade de cobrança de desempenho e à pequena valorização do ensino nos planos e promoções de carreira docente.

LOBO (2017) ainda explica, que existem outras várias razões sobre a evasão universitária; os aspectos citados acimas são os mais reconhecidas entre estudiosos e gestores.

2.2.4 Dados da evasão no âmbito nacional

A evasão é um problema internacional, e por isso tornou-se um assunto de grande abrangência (FILHO, 2009). De acordo com os estudos de Rafael e Esteban (2012), os indicadores de evasão no ensino superior internacional variam bastante por dependência administrativa (pública ou privada), região e curso. O estudo apresenta os seguintes indicadores: Espanha (20%), Estados Unidos (35%), Colômbia (45%), Chile (50%) e Itália (60%). Estes indicadores evidenciam que a evasão está presente em todo o globo onde há educação de nível superior (PRESTES; FIALHO; PFEIFER, 2016).

No cenário brasileiro, a evasão atinge tanto as instituições privadas como as públicas. De acordo com os estudos do Instituto Lobo, a evasão teve um crescimento gradual entre 2005 e 2011, onde os indicadores de abandono no ensino superior mostraram um crescimento considerável; no ano de 2005 a taxa de evasão foi de 22% e, no ano de 2011, de 37,9% (PRESTES; FIALHO; PFEIFER, 2016).

De acordo com um estudo realizado em 2016 pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), nas IES públicas, os indicadores permanecem relativamente estáveis nos últimos anos: em torno de 24% para ensino presencial e mais de 30% para os cursos a distância (INEP, 2017). Já nas instituições particulares, o número de evasão vem aumentando; segundo o Panorama do Ensino Superior Privado do Brasil, os indicadores aumentaram 4% entre 2011 e 2016, indo de 19% para 23% (UNIVERSIA, 2019).

Ainda de acordo com o INEP, entre 2017 e 2018, houve um pequeno aumento no índice de evasão das universidades públicas e privadas. Em 2017 a taxa foi de 27,5%, já no ano seguinte subiu para 28,8% (INEP, 2018). Os dados de 2019 e 2020 ainda não foram divulgados, mas de acordo com Capelato, a taxa de evasão de 2020 estima-se em 35%. Esse índice é fundamentado pela quantidade de alunos que podem vir a desistir de estudar por causa da pandemia do Covid-19 (G1, 2020).

Os índices apresentados acima podem não parecer tão alarmantes quando representados como uma porcentagem; mas quando exibido em números reais, outro panorama emerge. De acordo com o censo do INEP, mais de oito milhões de alunos foram matriculados em 2016. Assim, a taxa de evasão para o respectivo ano é equivalente a mais de um milhão de estudantes. Este número é um grande grupo de indivíduos que, por um

conjunto de motivos, optaram por abandonar a graduação (OLIVEIRA, 2018)

Nas instituições privadas, essas taxas de evasão reproduzem uma redução de receita, posto que os evadidos param de contribuir com as mensalidades. No cenário das IES públicas, também é constatado desperdício financeiro, porém, neste caso, o abandono do aluno significa recurso público investido sem o devido retorno, já que os mesmos são designados a professores, funcionários, equipamentos e espaço físico cuja capacidade total não é totalmente usufruída (Silva e Filho et al., 2007). Pesquisas realizadas pelo Instituto Lobo, mostraram que a evasão custa ao ano, em média, 9 bilhões de reais para o Brasil (SLYWITCH; BILAC; SANTOS, 2017).

3 ESTADO DA ARTE

Neste capítulo, serão apresentados trabalhos relacionados ao tema descoberta do conhecimento em bases de dados educacionais, que visam reduzir a evasão universitária.

3.1 Previsão precoce de abandono da faculdade usando mineração de dados

Martins et al. (2017) desenvolveram 321 modelos de predição de evasão estudantil em universidades públicas brasileiras, utilizando-se das técnicas de mineração de dados Aprendizado de máquina (*Deep Learning* - DP), Floresta Aleatória Distribuída (*Distributed Random Forest* - DRF) e Máquina de aumento de gradiente(*Gradient Boosting Machine* - GBM), na qual a proporção de modelos desenvolvidos foram 81, 108 e 132, respectivamente.

Os modelos desenvolvidos foram treinados a partir de dados gerados antes e depois do início do curso universitário escolhido pelo aluno. Dos dados anteriores ao ingresso na universidade, foram coletados os dados do Exame Nacional do Ensino Médio (ENEM), que permite quantificar quanto o aluno do ensino médio foi capaz de absorver em toda a sua trajetória de aprendizagem para que, com essa quantificação, pudesse competir por uma vaga na universidade pública desejada por meio do Sistema de Seleção Unificada (SISU). Quanto aos dados sucessores à admissão, foram coletados, dados pessoais (idade e sexo), forma de ingresso (cota social ou ampla concorrência), dados institucionais (curso, cidade do curso, etc.) e a situação atual do aluno na universidade (cursando, formado e desistente).

A metodologia de desenvolvimento utilizada pelos autores foi CRISP-DM (Cross Industry Standard Process for Data Mining). Essa tecnologia divide o trabalho em seis etapas: compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e desenvolvimento. O CRISP-DM possui a característica de ser um processo cíclico e as suas várias etapas se comunicam, podendo o processo retornar às etapas anteriores para otimizar os resultados.

Após o desenvolvimento de todos os modelos, aqueles com as melhores taxas de acerto foram escolhidos para fazer um comparativo de qual tecnologia obteve um melhor desempenho na previsão de possíveis evasões universitárias. Identificou-se que o DRF foi capaz de gerar uma taxa de 55,4 % de casos corretos de evasão, enquanto o GBM de acertou 63,24 % dos casos, e o DL atingiu 71,1 % dos casos corretos.

Dessa forma, percebeu-se que as técnicas utilizadas para prever possíveis evasões foram bem-sucedidas, com destaque para o Deep Learning. Assim, permitindo que as universidades tenham uma chance maior de implementar ações corretivas já no início do primeiro semestre do aluno.

3.2 Aprendizagem supervisionada no contexto de mineração de dados educacional para evitar o abandono de estudantes universitários

Kelly et al. (2019) realizaram uma comparação entre algoritmos de aprendizagem supervisionada(*Decision Tree, K-nearest Neighbor, Neural Networks, Support Vector Machine , Naive Bayes e Random Forests*) para verificar sua capacidade de fazer previsões de evasão estudantil na Universidade Federal de Sergipe (UFS).

Para a realização do trabalho, foram seguidas cinco etapas, a saber: Aquisição dos dados da UFS, pré-processamento de dados, seleção de recursos com base na contribuição de precisão, aplicação dos algoritmos selecionados e avaliação e análise de resultados. Como ferramenta de desenvolvimento, foi utilizada a biblioteca Scikit-learn. Os recursos da biblioteca foram selecionados com base na melhoria da precisão de todos os algoritmos e uma validação cruzada K-fold com cinco subconjuntos foi aplicada a cada modelo.

Os dados utilizados no experimento foram coletados da base de dados da UFS em um intervalo temporal de 2010 a 2018. Nos registros coletados estavam 23.690 alunos, pertencentes a um curso relacionado à área de computação da universidade, sendo o número de alunos por curso: Ciência da Computação com 12.079 alunos, seguido por Sistemas de Informação com 5.592 alunos e Engenharia da Computação com 5.389 alunos.

Após a execução de todas as etapas, foi possível concluir que os algoritmos Decision Tree e Random Forest alcançaram a melhor precisão na comparação com uma média de 70%. Entre os outros algoritmos treinados, apenas o K-nearest Neighbor atingiu uma precisão acima dos 50%, com uma média de 70%, mas apenas para o curso de Sistema de informação. Percebeu-se também que o algoritmo de Neural Networks estava tendo sua acurácia melhorada à medida que a quantidade de dados aumentava, portanto, com um base de dados mais volumosa, o algoritmo poderia ter se destacado entre aqueles com melhor precisão neste trabalho.

Assim, os resultados deste trabalho indicam que alguns algoritmos podem ser utilizados como ferramentas para apoiar decisões que reduzam o abandono escolar, com destaque para Decision Tree e Random Forest.

3.3 Análise preditiva para identificação de alunos suscetíveis à evasão escolar

Rodrigues et al. (2021) desenvolveram uma ferramenta computacional capaz de identificar alunos com maior chance de evasão escolar. Para a realização desse projeto foi utilizado técnicas de ciências de dados e aprendizado de máquina. Foi aplicado o algoritmo de redes neurais sobre os dados, e a validação dos resultados se deu através de cruzamentos de múltiplas bases de dados de domínio público.

Os dados analisados no projeto são referentes à alunos matriculados no ano de 2019 em cursos de Bacharelado no Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais (IFSEMG). A base de dados coletada contém 575 linhas com

informações sobre idade, sexo e dados socioeconômicos dos alunos.

A metodologia do trabalho foi dividida em quatro etapas, são elas: Seleção dos dados, onde ocorreu a escolha do conjunto de dados; Pré-processamento, a etapa que os dados faltantes são tratados; Aprendizado de máquina, onde ocorre a aplicação do algoritmo escolhido sobre os dados; e por fim a Validação, analisando os resultados obtidos e gerando conhecimento.

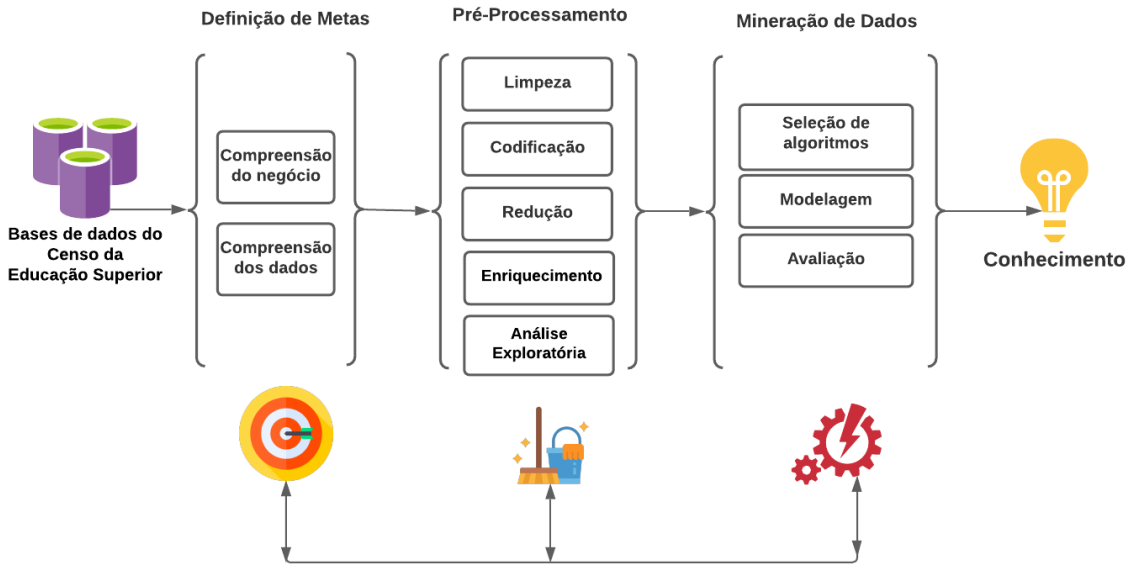
A Rede Neural Artificial desenvolvida teve uma taxa de acurácia de 85,2% durante os testes. Os autores concluíram que a ferramenta desenvolvida é confiável o suficiente para fornecer informações dos alunos com mais probabilidades de abandonarem o curso.

4 METODOLOGIA

Vistos os fundamentos e conceitos mencionados no capítulo 2, este capítulo apresenta a estrutura metodológica, na qual são detalhados todos os métodos adotados para cada uma das etapas de execução do projeto, com destaque para a base de dados, parâmetros de configuração e recursos adotados.

Após a escolha das bases de dados e compreensão do modelo de desenvolvimento CRISP - DM, todas as etapas foram organizadas sequencialmente em um fluxograma como descrito na Figura 9.

Figura 9 – Fluxo do Processo de Desenvolvimento do Projeto.



Fonte: Próprio Autor.

4.1 Processo de desenvolvimento

As etapas de desenvolvimento são organizadas de forma sequencial, conforme mostrado na Figura 9, para a execução das atividades. Essa estrutura auxilia a monitorar quais etapas foram executadas, quais objetivos já foram alcançados e quais etapas precisam ser reformuladas na execução do trabalho. As próximas subseções apresentam o desenvolvimento deste trabalho.

4.1.1 Bases de dados do Censo da Educação Superior

Neste trabalho, é utilizada a base de dados aberta do Censo da Educação Superior. Esses dados são coletados anualmente pelo Instituto Nacional de Estudos e Pesquisas

Educacionais Anísio Teixeira - INEP, com o objetivo de fornecer informações detalhadas sobre a situação e as principais tendências do meio acadêmico universitário.

O censo da educação superior reúne informações sobre as instituições de ensino superior (IES), seus cursos de graduação presencial ou a distância, cursos sequenciais, vagas oferecidas, candidatos, matrículas, ingressantes e concluintes, além de informações sobre docentes, nas diferentes formas de organização acadêmica e categoria administrativa (DADOSABERTOS, 2021).

Por meio de questionário eletrônico, as IES respondem sobre sua estrutura e cursos. Durante o período de preenchimento do questionário, os pesquisadores institucionais podem fazer, a qualquer momento, as alterações ou acréscimos necessários aos dados de suas respectivas instituições. Após este período, o sistema é encerrado para alterações e os dados são disponibilizados às IES, em forma de relatório, para que as informações prestadas possam ser consultadas, validadas ou corrigidas (DADOSABERTOS, 2021).

Após esse período de validação ou correção das informações fornecidas pelas IES, o Inep realiza rotinas de análise na base de dados do censo, para verificar a consistência das informações, e então o censo é finalizado. Os dados são então divulgados e a sinopse estatística é publicada, não podendo haver mais alterações nas informações, pois passam a ser estatísticas oficiais (DADOSABERTOS, 2021).

É importante ressaltar, que o censo realizado no ano vigente se refere ao ano acadêmico anterior, dessa forma o censo da educação superior do ano de 2020, se refere a coleta de dados do ano acadêmico de 2019.

Para essa pesquisa são utilizadas as bases de dados referentes aos Censos de Ensino Superior de 2014 a 2017, especificamente fazendo uso dos dados disponíveis na tabela "DM_ALUNO", que contém dados básicos da IES, do curso, da forma de admissão, o apoio social recebido, as atividades extracurriculares realizadas e as características de cada aluno devidamente matriculado no ano referente à pesquisa.

Cada base de dados selecionada possui um número de colunas com os mesmos significados, porém em algumas bases de dados houve mudanças na nomenclatura dessas variáveis e, portanto, tiveram que ser padronizadas. A forma escolhida para essa operação foi a conservação das nomenclaturas do censo de 2014 aplicada nas demais.

Em relação aos registros utilizados, o trabalho delimitou-se aos discentes dos cursos de tecnologia da informação - TI com duração de 4 anos (Tabela 1), que ingressaram no curso em 2013 (Censo 2014), e assim foi realizado o acompanhamento da trajetória desses alunos ao longo da duração do curso, ou seja, até o ano de 2016 (Censo 2017).

Tabela 1 – Lista com nomes dos cursos selecionados

Nomes dos cursos selecionados
ABI - CIÊNCIA DA COMPUTAÇÃO
ABI - SISTEMAS DE INFORMAÇÃO
ANÁLISE E DESENVOLVIMENTO DE SISTEMAS
CIÊNCIA DA COMPUTAÇÃO
CIÊNCIAS DA COMPUTAÇÃO
CIÊNCIAS DE COMPUTAÇÃO
COMPUTAÇÃO E INFORMÁTICA
COMPUTAÇÃO
REDES DE COMPUTADORES
SISTEMA DE INFORMAÇÃO
SISTEMAS DE COMPUTAÇÃO
SISTEMAS DE INFORMAÇÃO

Esse acompanhamento foi possível através de 4 análises:

1. Análise do primeiro ano cursado - Base composta somente pelos dados do Censo de 2014 - Essa base será chamada de Base 1;

2. Análise dos dois primeiros anos cursados - Base composta somente pelos alunos que não evadiram e se mantiveram no Censo de 2015 - Essa base será chamada de Base 2;

3. Análise dos três anos cursados - Base composta somente pelos alunos que não evadiram e se mantiveram no Censo de 2016 - Essa base será chamada de Base 3;

4. Análise dos quatro anos cursados - Base composta somente pelos alunos que não evadiram e se mantiveram no Censo de 2017 - Essa base será chamada de Base 4;

Com exceção do item 1, todas as bases de dados analisadas são formadas utilizando junções do tipo *inner join* (registros de duas tabelas são usados para gerar dados relacionados de ambas as tabelas).

As junções foram realizadas com a base do censo do ano em análise com o(os) ano(os) antecessor(es), ou seja, cada base é formada pelos estudantes que não evadiram no(os) ano(os) anterior(es) e base do censo em análise, por exemplo a base de dados do item 2, é composta de todos os estudantes que não evadiram no Censo de 2014 e os dados desses alunos disponíveis no Censo de 2015.

Um exemplo disso, é o *inner join* do Censo de 2014 com o de 2015, já a base do item 3 é o *inner join* do censo de 2014 com censo de 2015 e por fim o *inner join* da base resultante com a base do censo de 2016.

Nesta pesquisa as variáveis utilizadas para realizar o *inner join*, citado acima, entre as tabelas são: o código do aluno no ensino superior(CO_ALUNO), código da IES(CO_-IES), código do curso(CO_CURSO) e ano de ingresso do estudante(ANO_INGRESSO).

4.1.2 Definição de metas

Nesta etapa, é realizado a compreensão do negócio e a compreensão dos dados, e em seguida são elaboradas as estratégias de processamento de dados, hipóteses e metas, para atuar de forma mais eficiente no desenvolvimento do processo.

- Compreensão do negócio: Nessa etapa, é realizado todo um levantamento do ambiente educacional que envolve os dados do Censo da Educação Superior. Primeiramente, é realizada uma entrevista com o profissional envolvido no processo de coleta de dados censitários do período escolhido para esta pesquisa, a fim de melhor compreender o processo de coleta e esclarecer dúvidas sobre os dados em análise. Em seguida, são estudados e analisados artigos e resumos técnicos disponibilizados pelo INEP, pois tais dados são utilizados como fonte de dados por diversos públicos, sendo fonte de consulta para gestores de instituições de ensino, gestores de políticas educacionais, órgãos governamentais, estudantes, pesquisadores e demais interessados na educação superior brasileira, a fim de subsidiar processos de análise, pesquisa, planejamento e tomada de decisão.

- Compreensão dos dados: Essa etapa consiste em funções de coleta das bases de dados censitários para compreensão de todas as variáveis neles disponíveis. As bases de dados desta pesquisa possuem como um dos arquivos anexos o dicionário de variáveis(Figura 10), que é uma planilha que contém todas as variáveis que compõem a base e suas especificações como nome, descrição, tipo, tamanho em bytes, categorias quando o campo possui classes e por fim um campo com observações. Essa etapa permite maior familiaridade com os dados, para identificar problemas de qualidade dos dados, construir estratégias para melhorar a qualidade dos bancos de dados censitários e criar hipóteses sobre informações ocultas.

Figura 10 – Primeira página do dicionário de dados da tabela DM_ALUNO

ALUNO (DM_ALUNO)					
N	**Nome da Variável**	**Descrição da Variável**	**Tipo**	**Tam.(1)**	**Categorias**
		DADOS DA IES			
1	CO_IES	Código único de identificação da IES	Num	8	
2	NO_IES	Nome da IES	Char	200	
3	CO_CATEGORIA_ADMINISTRATIVA	Código da Categoria Administrativa	Num	8	1. Pública Federal 2. Pública Estadual 3. Pública Municipal 4. Privada com fins lucrativos 5. Privada sem fins lucrativos 7. Especial
4	DS_CATEGORIA_ADMINISTRATIVA	Nome da Categoria Administrativa	Char	100	
5	CO_ORGANIZACAO_ACADEMICA	Código da Organização Acadêmica	Num	8	1. Universidade 2. Centro Universitário 3. Faculdade 4. Instituto Federal de Educação, Ciência e Tecnologia 5. Centro Federal de Educação Tecnológica
6	DS_ORGANIZACAO_ACADEMICA	Nome da Organização Acadêmica	Char	100	
		DADOS DO CURSO			
7	CO_CURSO	Código único de identificação do curso gerado pelo E-MEC	Num	8	
8	NO_CURSO	Nome do curso	Char	200	
9	CO_CURSO_POLO	Código de identificação do polo vinculado ao curso	Num	8	
10	CO_TURNO_ALUNO	Código do turno do curso ao qual o aluno está vinculado	Num	8	1. Matutino 2. Vespertino 3. Noturno 4. Integral (.) Não aplicável (cursos com modalidade de Ensino a Distância)
11	DS_TURNO_ALUNO	Nome do turno do curso ao qual o aluno está vinculado	Char	13	
12	CO_GRAU_ACADEMICO	Código do grau acadêmico conferido ao diplomado pelo curso	Num	8	1. Bacharelado 2. Licenciatura 3. Tecnológico (.) Não aplicável (cursos com nível acadêmico igual a Sequencial de formação específica ou cursos com Área Básica de Ingresso identificada pela variável TP_ATRIBUTO_INGRESSO)
13	DS_GRAU_ACADEMICO	Nome do grau acadêmico conferido ao diplomado pelo curso	Char	13	
14	CO_MODALIDADE_ENSINO	Código da modalidade de ensino do curso	Num	8	1. Presencial 2. Curso a distância
15	DS_MODALIDADE_ENSINO	Nome da modalidade de ensino do curso	Char	17	
16	CO_NIVEL_ACADEMICO	Código do nível acadêmico do curso	Num	8	1. Graduação 2. Sequencial de formação específica
		DADOS DO ALUNO			
26	CO_ALUNO_CURSO	Código de identificação gerado pelo Inep para o vínculo do aluno ao curso	Num	8	
27	CO_ALUNO_CURSO_ORIGEM	Código de identificação gerado pelo Inep para o vínculo do aluno em seu curso de origem, ou seja, de onde foi transferido	Num	8	
28	CO_ALUNO	Código de identificação gerado pelo Inep para o aluno da Educação Superior	Num	8	
29	CO_COR_RACA_ALUNO	Código da cor/raça do aluno	Num	8	1. Branca 2. Preta 3. Parda 4. Amarela 5. Indígena 6. Não dispõe da informação 0. Aluno não quis declarar cor/raça
30	DS_COR_RACA_ALUNO	Nome da cor/raça do aluno	Char	32	
31	IN_SEXO_ALUNO	Informa o sexo do aluno	Num	8	0. Masculino 1. Feminino
32	DS_SEXO_ALUNO	Nome do sexo do aluno	Char	9	
33	NU_ANO_ALUNO_NASC	Ano de nascimento do aluno	Num	8	Derivadas da variável DT_NASCIMENTO
34	NU_MES_ALUNO_NASC	Mês de nascimento do aluno	Num	8	
35	NU_DIA_ALUNO_NASC	Dia de nascimento do aluno	Num	8	
36	NU_IDADE_ALUNO	Idade que o aluno completa no ano de referência do Censo	Num	8	
37	CO_NACIONALIDADE_ALUNO	Código da nacionalidade do aluno	Num	8	1. Brasileira 2. Brasileira - nascido no exterior ou naturalizado 3. Estrangeira
38	DS_NACIONALIDADE_ALUNO	Nome da nacionalidade do aluno	Char	48	

Fonte: Próprio Autor.

4.1.3 Pré-processamento

Nesta etapa são realizados todos os procedimentos para melhorar a qualidade das bases de dados utilizadas, bem como a remoção de variáveis e registros que não possuem efeito no processo, ou que reduzem a precisão dos modelos em seus desenvolvimentos.

4.1.3.1 Redução de Dados

Esse passo é utilizado para aumentar a eficiência e reduzir custos; suas principais tarefas são: Seleção de um subconjunto de atributos(colunas), reduzindo o número e reduzindo a dimensionalidade(redução vertical). E a redução de registros delimitando assim os dados alvo para esta pesquisa (redução horizontal).

- Redução horizontal

Como já mencionado, nesse trabalho são utilizados somente os registros dos discentes que ingressaram no curso em 2013 (Censo 2014), dos cursos de tecnologia da informação com duração de 4 anos, de nível acadêmico graduação, de modalidade presencial e de grau acadêmico bacharel, logo todos os demais registros que não seguem essas características foram removidos das bases de dados.

Após a remoção de variáveis para a delimitação do tema, foi realizado uma análise nos dados de cada base que resultou na exclusão de alguns registros:

- Remoção de Formados: Foram encontrados estudantes que no mesmo ano de ingresso haviam formado (203 Alunos). Após indagar sobre essa situação, esse questionamento foi feito para o profissional que realizava o Censo, e a partir daí concluiu-se que esses dados não eram erros de digitação ou do sistema utilizado para pesquisa, mas alunos já formados que desejaram melhorar o próprio Coeficiente de Rendimento Acadêmico - CR, pois basta reaproveitar as matérias já cursadas e/ou refazer matérias que havia tirado notas baixas para que o CR tenha um melhora significativa. Os registros desses alunos foram retirados das bases de dados, pois poderiam causar padrões inconsistentes em relação à maioria dos outros registros que eram apenas de alunos que estavam ingressando no curso pela primeira vez.
- Remoção de Falecidos: Foram encontrados estudantes falecidos somente na Base 1, em um total de 2 estudantes, e estes foram removidos pois o estado de falecimento não entra nas condições de evasão.
- Remoção de alunos que realizaram o destrancamento da matrícula: Foi percebido que alunos que realizavam o trancamento da matricula em um semestre anterior, e realizavam o destrancamento na base em análise. Como o objetivo é definir perfis com o intuito de evitar tais trancamentos de matricula e evasões no geral, não é coerente manter esses registros nas análises e desenvolvimento de modelos preditivos.

- Redução vertical:

  As tarefas detalhadas abaixo são realizadas igualmente nas quatro bases analisadas.

  1. Retirada de colunas com 70% ou mais de dados faltantes, algumas delas são: código aluno do curso de origem (CO_ALUNO_CURSO_ORIGEM), código do pais de origem do aluno (CO_PAIS_ORIGEM_ALUNO), código de IES e pais de destino (CO_IES_DESTINO e CO_PAIS_DESTINO) , e outras.
  2. Remoção de colunas que em todos os dados possuem apenas um único valor, como por exemplo todas as variáveis que foram utilizadas para delimitar o tema

na etapa de redução horizontal. A base de dados resultante dessas colunas utilizadas como filtro só possuem um único valor em todos os registros.

3. Seleção de atributos por meio do dicionário de variáveis, no qual foram retiradas colunas sem relevância para a criação de padrões, como nome e código da universidade (NO_IES, CO_IES), nome e código do curso (NO_CURSO, CO_-CURSO), dentre outras.

4. Houve remoção vertical condicionada. Havia dados duplicados, porém representados de maneiras diferentes. Por exemplo, o atributo "Cor do estudante", existe em uma coluna com os valores "Branco"e "Preto" e seu respectivo valor numérico, em outra coluna com os valores "0"e "1". Essas colunas foram divididas para atenderem bases de algoritmos de natureza diferente, ou seja, dependendo do algoritmo de mineração, apenas uma das colunas foi direcionada para a base que seria utilizada. Dentre essas colunas estão turno do curso (DS_TURNO_ALUNO e CO_TURNO_ALUNO), sexo do estudante (DS_SEXO_ALUNO e IN_SEXO_ALUNO), nacionalidade do estudante (DS_-NACIONALIDADE_ALUNO e CO_NACIONALIDADE_ALUNO), cor do estudante (DS_COR_RACA_ALUNO e CO_COR_RACA_ALUNO), e outras.

5. Remoção de variáveis que especificam outras. Há colunas que possuem informações gerais sobre determinada situação e outras que a detalhavam, como por exemplo a coluna que informa se o estudante naquele ano da pesquisa recebeu algum apoio social(IN_APOIO_SOCIAL), esta era seguida por colunas que especificavam qual apoio o estudante recebeu naquele ano da pesquisa, como apoio alimentação (IN_APOIO_ALIMENTACAO), apoio moradia (IN_-APOIO_MORADIA), apoio transporte (IN_APOIO_TRANSPORTE) e outras formas de apoio. Essas colunas que especificam só foram removidas após o desenvolvimento dos modelos na etapa de mineração de dados, pois testes foram realizados utilizando apenas as colunas com dados gerais e assim melhores resultados foram obtidos com maior velocidade para treinamento dos modelos preditivos.

6. Remoção de colunas utilizadas para criação de outras colunas mais específicas (técnica conhecida como *binning*). As Colunas que possuem os dados da situação do aluno (DS_ALUNO_SITUACAO e CO_ALUNO_SITUACAO) foram utilizadas para a geração da coluna com as classes alvo dessa pesquisa(Nomeada de "*targetColumn*"), caso o estudante tivesse como situação os valores "Cursando"ou "Formado"na *targeColumn* o valor seria "Não Evadiu"caso fosse o valor "Matrícula trancada", ou "Desvinculado do curso", ou "Transferido para outro curso da mesma IES"o valor seria "Evadiu". Já a variável idade do aluno(NU_IDADE) teve os valores agregados em intervalos seguindo a pira-

mide etária brasileira (“0-14”,“15-24”,“25-54”,“55-64”e “65+”), essa variável foi nomeada de “CategoriaIdade”. Após tais agregações essas colunas progenitoras foram removidas, com ênfase para a remoção das variáveis de situação do estudante, pois essas tinham um vinculo direto com a *targetColumn* e isso ocasionaria em modelos inúteis.

4.1.3.2 Limpeza dos dados

Esse passo realiza a melhora da qualidade dos dados que são utilizados pelos algoritmos mineradores. Dentre as atividades realizadas estão o preenchimento de dados faltantes e redução de ruídos. As tarefas detalhadas abaixo são realizadas igualmente nas quatro bases analisadas, pois essas possuem problemas em seus dados nas mesmas variáveis.

A primeira situação são as colunas gerais e suas colunas especificadoras(mencionadas no Item 6 da redução vertical). Essas colunas recebem o valor “0”ou “1”, respectivamente, a afirmação ou negação do que significa a variável. Porém, quando as colunas gerais recebem o valor “0”todas as colunas que a especificam recebem o valor “?”(valor vazio). Portanto, todos os registros de variáveis especificadoras no qual o registro de suas variáveis gerais forem o valor “0”, todos os valores “?”são substituídos por “0”, pois representam a negação da variável em questão.

A segunda situação, são as colunas gerais que possuem valores vazios, porém as colunas especificadoras possuem algum valor “1”, desta forma é possível preencher tais campos ausentes apenas realizando uma simples condicional.

A terceira, e última, situação são as colunas com quantidade de dados ausentes inferiores a 10% do seu total. Dentre essas estão as variáveis sexo do aluno(IN_SEXO_ALUNO e DS_SEXO_ALUNO), turno do curso(CO_TURNO_CURSO e DS_TURNO_CURSO). Essas e as demais são preenchidas utilizando a técnica *Most Frequent Value* que verifica qual dos dados dispostos na coluna é o mais frequente e substitui no campo ausente.

4.1.3.3 Codificação

Nesta etapa, é realizada a transformação dos dados já “limpos”em formatos mais adequados para o processo de mineração. Esta etapa geralmente envolve as ações de normalização e discretização dos dados.

O processo de normalização(transformação de todas as variáveis para a mesma ordem de grandeza) dos dados é realizado somente nas variáveis carga horária total(QT_CARGA_HORARIA_TOTAL) e carga horária integralizada(QT_CARGA_HORARIA_INTEG), todas as outras variáveis não possuem valores com uma maior ordem de grandeza, ou essas alternam entre “0”e “1”ou já são valores categóricos que inicia com inteiro

“0”e segue a sequencia até o valor numérico “6”. Portanto, normalizar os dados os dados não relacionado com carga horária, além de não trazer benefícios aos algoritmos de mineração, pode piorar os resultados em algumas situações, o que seria um gasto computacional desnecessário.

Já o processo de discretização é utilizado na variável “CategoriaIdade”, porém somente nos dados de entrada do algoritmo minerador PNN, pois as características do algoritmo precisam de valores numéricos.

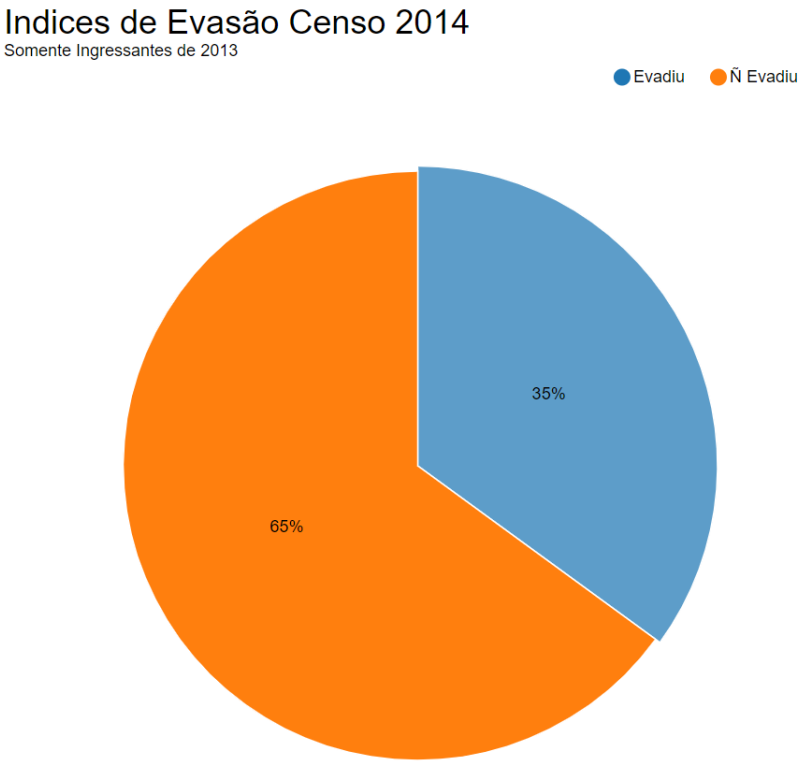
4.1.4 Mineração de dados

Como já mencionado, esse trabalho é desenvolvido com foco na tarefa de classificação. A coluna alvo da base de dados selecionada para realização dessa tarefa é a “targetColumn”, tendo como classes os valores “Não Evadiu”e “Evadiu”. Essa coluna é utilizada como coluna alvo por conter a situação do estudante naquele período da pesquisa do censo, ou seja, é o resultado que possivelmente foi influenciado pelas outras variáveis da base de dados.

Além disso, a partir da descoberta de padrões em função dessas classes, é possível verificar quais fatores levaram um estudante a permanecer ou não no curso. Dessa forma, é viável traçar perfis estudantis e, assim, corroborar no processo de tomada de decisão de gestores universitários, possibilitando a identificação de alunos em situação de evasão, podendo realizar ações mitigadoras para que esses estudantes permaneçam no curso até a formatura e, consequentemente, ter um possível aumento no controle da taxa de evasão.

Com todas as bases de dados já pré-processadas, essas têm seus registros particionados, de forma randômica, em 80% como base de treinamento dos algoritmos e 20% como base de testes, essa no qual os resultados são apresentados no Capítulo 5. Como se pode observar na Figura 11 o número de registros da classe “Não Evadiu”é bem maior em relação a classe “Evadiu”, o que caracteriza as bases de dados como bases desbalanceadas.

Figura 11 – Índices de evasão no Censo de 2014 - Somente ingressantes de 2013



Fonte: Próprio Autor.

Existem dezenas de técnicas para balancear uma base dados, duas das quais são testadas para serem utilizadas no desenvolvimento deste trabalho. A primeira realiza a retirada aleatória de linhas da classe com mais registros até que haja um número aproximado daquela com menos registros. A segunda é a técnica Synthetic Minority Oversampling TEchnique - SMOTE, um algoritmo que sobre-amostra os dados de entrada, ou seja, adiciona linhas artificiais baseadas na base de dados, para enriquecer os dados de treinamento.

Após a realização de testes na Base 1 (Censo 2014), observa-se que a primeira técnica resulta em piores resultados quando comparado à técnica SMOTE, provavelmente devido ao despejo de possíveis dados úteis da classe com maior número de registros realizada na primeira técnica. Assim, a técnica SMOTE é escolhida para balancear as demais bases de dados desta pesquisa.

Os algoritmos mineradores selecionados para essa pesquisa, são Decision Tree - DT, Random Forest - RF e Probabilistic Neural Network - PNN. Esses são treinados, com a base de dados não balanceadas e com as bases de dados balanceadas, no capitulo 5 é apresentado os resultados de ambas as situações.

Logo após a preparação dos dados, foi realizada a modelagem, ou seja, a execução dos algoritmos com as bases já citadas, utilizando a ferramenta KNIME, apresentada na seção 2.1.5.

Figura 12 – Matriz Confusão

		Detectada	
		Sim	Não
Real	Sim	$VerdadeiroPositivo(VP)$	$FalsoNegativo(FN)$
	Não	$FalsoPositivo(FP)$	$VerdadeiroNegativo(VN)$

Como já explicado, o processo de descoberta de conhecimento é um processo iterativo e interativo, o que significa que o processo é executado dezenas de vezes e cada resultado é analisado a fim de aprimorar os modelos gerados até que alcancem precisão e confiabilidade satisfatórias, essa fase é nomeada de avaliação. Os resultados alcançados ao final do processo estão dispostos no Capítulo 5.

Ao realizar a avaliação dos modelos, é gerada uma matriz confusão(é uma tabela que indica os erros e acertos do modelo, comparando com o resultado esperado (rótulos). Um exemplo pode ser observado na Figura 12.

A matriz confusão pode ser interpretada da seguinte forma:

- Verdadeiros Positivos: classificação correta da classe Positivo;
- Falsos Negativos (Erro Tipo 1): erro no qual o modelo previu a classe Negativa quando o valor real era a classe Positiva;
- Falsos Positivos (Erro Tipo 2): erro no qual o modelo previu a classe Positiva quando o valor real era a classe Negativa;
- Verdadeiros Negativos: classificação correta da classe Negativo;

Esse estudo utiliza, de forma predominante três métricas para análise dos resultados:

1. Acurácia: indica um desempenho geral do modelo. Entre todas as classificações, quantas o modelo classificou corretamente.

$$\textbf{Acurácia} = [VP + VN/VP + VN + FP + FN]$$

2. Error: é justamente a diferença do valor da acurácia em relação ao valor 1;
3. Coeficiente Kappa: medida estatística que visa verificar o grau de confiabilidade intermediária.

$$\textbf{Kappa} = [PC - PRC/1 - PRC]$$

Onde:

PC: Probabilidade de concordância;
PRC: Probabilidade de concordância aleatória;

Segundo Simon (2005), uma possível interpretação dos valores seria:

- Deficiente: menor que 0,20;
- Justo: de 0,20 a 0,40;
- Moderado: de 0,40 a 0,60;
- Bom: de 0,60 a 0,80;
- Muito bom: de 0,80 a 1,00.

5 RESULTADOS

Este capítulo é responsável por apresentar os resultados obtidos por meio da etapa de mineração de dados, detalhada no capítulo anterior. É analisada a trajetória dos alunos ingressantes nos cursos de TI no Brasil no ano de 2013. São selecionados para análise os quatro primeiros anos cursados, ou seja, os censos de 2014 a 2017. Os resultados são apresentados da seguinte forma: primeiro o resultado da análise no primeiro ano dos ingressantes (Censo 2014), a seguir o resultado dos dois primeiros anos (Censo 2014 e 2015), depois os três anos cursados(Censo 2014, 2015 e 2016), e por fim os quatro anos de curso(censo 2014, 2015, 2016 e 2017).

Esse estudo utiliza principalmente de duas métricas de avaliação dos dados:

1. *Accuracy*: Precisão do modelo desenvolvido.
2. *Cohen's Kappa*: medida estatística que visa verificar o grau de confiabilidade intermediária de um modelo. Segundo Simon (2005), uma possível interpretação dos valores seria:
    - Deficiente: menor que 0,20;
    - Justo: de 0,20 a 0,40;
    - Moderado: de 0,40 a 0,60;
    - Bom: de 0,60 a 0,80
    - Muito bom: de 0,80 a 1,00.

5.1 Análise do primeiro ano dos ingressantes (Censo 2014)

Os primeiros resultados gerados referem-se à análise dos dados dos alunos ingressantes nos cursos em 2013, o que equivale ao Censo de 2014. Os algoritmos selecionados (*Decision Tree*, *Random Forest* e *Probabilistic Neural Network* (PNN)) são executados nos dados de treinamento balanceados e desbalanceados, os resultados da mineração são apresentados a seguir.

- ***Decision Tree***

A Figura 13 apresenta o resultado da execução do algoritmo Decision Tree sobre a base de dados com balanceamento. Das 5264 instâncias selecionadas para a base de teste, 3325 delas são classificadas corretamente, com uma acurácia de 63,2%; e 1939 estão classificadas incorretamente, tendo uma taxa de erro de 36,8%. De acordo com o coeficiente Kappa, que possui um valor de 0.235, a classificação apresenta um grau de

confiabilidade intermediária, ou “justo” (valores entre 0,20 e 0,40). Na matriz de confusão é possível observar que a classe “Não Evadiu”tem o total de 2072 instâncias classificadas corretamente, e 1386 incorretamente, enquanto a classe “Evadiu”obteve 1253 instancias acertadas e 553 erradas.

O resultado da execução do algoritmo Decision Tree na base sem o balanceamento Figura 14. Nesse resultado uma taxa maior de acurácia (67%) é percebida, logo há uma taxa menor de instâncias classificadas incorretamente (33%); porém o valor do coeficiente Kappa decresceu para 0,23. Com isso, pode-se concluir que o resultado do algoritmo com balanceamento, mesmo tendo uma acurácia menor, é mais confiável do que sem o balanceamento.

Figura 13 – Resultado do algoritmo Decision Tree com balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	2072	1386
Evadiu	553	1253
Classificadas corretamentes: 3,325		
Classificadas incorretamentes: 1,939		
Erro: 36.835%		
Acurácia: 63.165%		
Coeficiente Kappa: 0.264		

Fonte: Próprio Autor.

Figura 14 – Resultado do algoritmo Decision Tree sem balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	2755	699
Evadiu	1036	773
Classificadas corretamentes: 3,528		
Classificadas incorretamentes: 1,735		
Erro: 32.966%		
Acurácia: 67.034%		
Coeficiente Kappa: 0.235		

Fonte: Próprio Autor.

- ***Random Forest***

Os resultados apresentados na Figura 15 e na Figura 16 mostram que o modelo gerado a partir da execução do algoritmo Random Forest na base de dados balanceada obtém uma acurácia de 65,5%, enquanto na base de dados não balanceada alcançou uma taxa de acerto de 68%. Porém, o coeficiente kappa, responsável por definir a confiabilidade do resultado, possui um valor de 0,29 no modelo gerado a partir da base balanceada, e 0,22 na base não balanceada; ou seja, os resultados utilizando a base balanceada têm uma confiabilidade maior.

Figura 15 – Resultado do algoritmo Random Forest com balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	2267	1193
Evadiu	622	1189
Classificadas corretamentes: 3,456		
Classificadas incorretamentes: 1,815		
Erro: 34.434%		
Acurácia: 65.566%		
Coeficiente Kappa: 0.29		

Fonte: Próprio Autor.

Figura 16 – Resultado do algoritmo Random Forest sem balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	2971	489
Evadiu	1190	621
Classificadas corretamentes: 3,592		
Classificadas incorretamentes: 1,679		
Erro: 31.854%		
Acurácia: 68.146%		
Coeficiente Kappa: 0.222		

Fonte: Próprio Autor.

- ***Probabilistic Neural Network (PNN)***

Nos resultados do algoritmo de redes neurais (Figura 17 e Figura 18, são obtidas taxas de acertos bem próximas, 66,9% para a base de treinamento balanceada e 67% para a base não balanceada; porém, assim como os dois primeiros algoritmos, o coeficiente kappa teve um resultado melhor com os dados balanceados. Nesse caso, a diferença nos valores kappa é tão marcante, que é possível dizer que o resultado com dados não balanceados possui grau de confiabilidade deficiente, enquanto aquele com dados balanceados possui confiabilidade justa.

Figura 17 – Resultado do algoritmo PNN com balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	2714	767
Evadiu	978	812
Classificadas corretamentes: 3,526		
Classificadas incorretamentes: 1,745		
Erro: 33.106%		
Acurácia: 66.894%		
Coeficiente Kappa: 0.24		

Fonte: Próprio Autor.

Figura 18 – Resultado do algoritmo PNN sem balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	3369	112
Evadiu	1626	164
Classificadas corretamentes: 3,533		
Classificadas incorretamentes: 1,738		
Erro: 32.973%		
Acurácia: 67.027%		
Coeficiente Kappa: 0.075		

Fonte: Próprio Autor.

5.2 Análise dos dois primeiros anos dos ingressantes (Censo 2014 e 2015)

Para a realização da análise dos dois primeiros anos do curso dos ingressantes em 2013, foi realizado a junção entre as bases do censo de 2014 e 2015. Os resultados dos algoritmos executados sobre os dados balanceados e não balanceados são apresentados a seguir.

- ***Decision Tree***

Nas Figura 19 e Figura 20, é possível observar que foi obtido uma taxa de acurácia bem próxima entre os dados com balanceamento e sem balanceamento, respectivamente, 81,8% e 81,7%. Já o coeficiente kappa possui um maior valor com os dados balanceados(0,41), sendo o primeiro modelo desenvolvido com grau de confiabilidade moderado.

Figura 19 – Resultado do algoritmo Decision Tree com balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	2950	423
Evadiu	325	412
Classificadas corretamentes: 3,362		
Classificadas incorretamentes: 748		
Erro: 18.2%		
Acurácia: 81.8%		
Coeficiente Kappa: 0.412		

Fonte: Próprio Autor.

Figura 20 – Resultado do algoritmo Decision Tree sem balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	2982	399
Evadiu	351	384
Classificadas corretamentes: 3,366		
Classificadas incorretamentes: 750		
Erro: 18.222%		
Acurácia: 81.778%		
Coeficiente Kappa: 0.394		

Fonte: Próprio Autor.

- ***Random Forest***

O resultado da execução do Random Forest dispõe do mesmo padrão dos resultados anteriores, a acurácia obtida no modelo dos dados não balanceados (85,5%) é maior do que a taxa obtida com os dados balanceados (84,6%); porém, a taxa de confiabilidade dos resultados no modelo balanceado é maior.

Figura 21 – Resultado do algoritmo Random Forest com balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	3016	354
Evadiu	279	479
Classificadas corretamentes: 3,495		
Classificadas incorretamentes: 633		
Erro: 15.334%		
Acurácia: 84.666%		
Coeficiente Kappa: 0.507		

Fonte: Próprio Autor.

Figura 22 – Resultado do algoritmo Random Forest sem balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	3230	140
Evadiu	444	314
Classificadas corretamentes: 3,544		
Classificadas incorretamentes: 584		
Erro: 14.147%		
Acurácia: 85.853%		
Coeficiente Kappa: 0.441		

Fonte: Próprio Autor.

- ***Probabilistic Neural Network (PNN)***

A Figura 23 mostra o resultado da execução do algoritmo de redes neurais sobre os dados balanceados, onde 3390 das instâncias são classificadas corretamente, obtendo uma acurácia de 82% e um coeficiente kappa de 0,40. O resultado da execução sobre os dados não balanceados são apresentados na Figura 24, onde mostra que 3525 instâncias estão classificadas corretamente, com uma acurácia de 85,3% e a taxa de confiabilidade de 0,44.

Figura 23 – Resultado do algoritmo PNN com balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	2869	482
Evadiu	526	521
Classificadas corretamentes: 3,390		
Classificadas incorretamentes: 738		
Erro: 17.878%		
Acurácia: 82.122%		
Coeficiente Kappa: 0.474		

Fonte: Próprio Autor.

Figura 24 – Resultado do algoritmo PNN sem balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	3194	157
Evadiu	446	331
Classificadas corretamentes: 3,525		
Classificadas incorretamentes: 603		
Erro: 14.608%		
Acurácia: 85.392%		
Coeficiente Kappa: 0.442		

Fonte: Próprio Autor.

5.3 Análise dos três anos cursados dos ingressantes (Censo 2014, 2015 e 2016)

Os resultados dos algoritmos executados sobre os dados balanceados e não balanceados da junção entre os censo de 2014, 2015 e 2016 são apresentados a seguir.

- ***Decision Tree***

Os resultados da execução do algoritmo Decision Tree sobre os dados da junção dos três primeiros anos é apresentados a seguir. A Figura 25 mostra as taxas obtidas no modelo de dados balanceados, com acurácia de 84.5% e coeficiente kappa de 0.64, sendo categorizado como um resultado bom. Os resultados com os dados não balanceados não estão tão diferentes, detém de uma taxa de acerto de 85.4% e coeficiente kappa de 0.66, também sendo considerado um resultado bom.

Figura 25 – Resultado do algoritmo Decision Tree com balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	1988	256
Evadiu	248	771
Classificadas corretamentes: 2,759		
Classificadas incorretamentes: 504		
Erro: 15.446%		
Acurácia: 84.554%		
Coeficiente Kappa: 0.641		

Fonte: Próprio Autor.

Figura 26 – Resultado do algoritmo Decision Tree sem balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	2016	237
Evadiu	239	778
Classificadas corretamentes: 2,794		
Classificadas incorretamentes: 476		
Erro: 14.557%		
Acurácia: 85.443%		
Coeficiente Kappa: 0.66		

Fonte: Próprio Autor.

- ***Random Forest***

A Figura 27 mostra o resultado do algoritmo Random Forest na base de dados balanceada, possui uma acurácia de 85.7% e coeficiente kappa de 0,66, enquanto a base sem balanceamento (Figura 28) apresenta uma taxa de 86.4% e 0,67.

Figura 27 – Resultado do algoritmo Random Forest com balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	2029	226
Evadiu	242	784
Classificadas corretamentes: 2,813		
Classificadas incorretamentes: 468		
Erro: 14.264%		
Acurácia: 85.736%		
Coeficiente Kappa: 0.667		

Fonte: Próprio Autor.

Figura 28 – Resultado do algoritmo Random Forest sem balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	2063	192
Evadiu	253	773
Classificadas corretamentes: 2,836		
Classificadas incorretamentes: 445		
Erro: 13.563%		
Acurácia: 86.437%		
Coeficiente Kappa: 0.679		

Fonte: Próprio Autor.

- ***Probabilistic Neural Network (PNN)***

No resultado do algoritmo PNN, as taxas do modelo com os dados balanceados tiveram um destaque, pois tanto a taxa de acurácia (88.51%), como a taxa de kappa (0.73) estão maior que as taxas com o modelo não balanceado, 87.23% e 0.68 respectivamente..

Figura 29 – Resultado do algoritmo PNN com balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	2087	219
Evadiu	158	817
Classificadas corretamentes: 2,904		
Classificadas incorretamentes: 377		
Erro: 11.49%		
Acurácia: 88.51%		
Coeficiente Kappa: 0.73		

Fonte: Próprio Autor.

Figura 30 – Resultado do algoritmo PNN sem balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	2134	172
Evadiu	247	728
Classificadas corretamentes: 2,862		
Classificadas incorretamentes: 419		
Erro: 12.77%		
Acurácia: 87.23%		
Coeficiente Kappa: 0.687		

Fonte: Próprio Autor.

5.4 Análise dos quatro anos cursados dos ingressantes (Censo 2014, 2015, 2016 e 2017)

Na execução da análise dos quatro anos de curso, ocorreu a junção entre as bases do censo de 2014, 2015, 2016 e 2017. Os resultados dos algoritmos executados sobre os dados balanceados e não balanceados são apresentados a seguir.

- ***Decision Tree***

Entre os resultados dos três algoritmos na analise dos 4 anos de curso, as taxas do Decision Tree são os que mais diferem entre si, com tudo, ainda não é uma diferença muito marcante. Os resultados possuem taxas de acurácia de 82.4% e confiabilidade 0.57 para o modelo com dados balanceados, e taxa de acerto de 81.6% e kappa de 0.53 para o modelo não balanceado.

Figura 31 – Resultado do algoritmo Decision Tree com balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	476	272
Evadiu	241	1459
Classificadas corretamentes: 1,935		
Classificadas incorretamentes: 413		
Erro: 17.589%		
Acurácia: 82.411%		
Coeficiente Kappa: 0.574		

Fonte: Próprio Autor.

Figura 32 – Resultado do algoritmo Decision Tree sem balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	427	217
Evadiu	214	1485
Classificadas corretamentes: 1,912		
Classificadas incorretamentes: 431		
Erro: 18.395%		
Acurácia: 81.605%		
Coeficiente Kappa: 0.538		

Fonte: Próprio Autor.

- ***Random Forest***

As taxas obtidas pelo Random Forest nos dados com e sem balanceamentos não diferem muito entre si, e possuem respectivamente os valores 86% e 86.1% para a taxa de acurácia, e 0.65 e 0.64 de confiabilidade.

Figura 33 – Resultado do algoritmo Random Forest com balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	498	154
Evadiu	175	1530
Classificadas corretamentes: 2,028		
Classificadas incorretamentes: 329		
Erro: 13.958%		
Acurácia: 86.042%		
Coeficiente Kappa: 0.655		

Fonte: Próprio Autor.

Figura 34 – Resultado do algoritmo Random Forest sem balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	472	180
Evadiu	146	1559
Classificadas corretamentes: 2,031		
Classificadas incorretamentes: 326		
Erro: 13.831%		
Acurácia: 86.169%		
Coeficiente Kappa: 0.649		

Fonte: Próprio Autor.

- ***Probabilistic Neural Network (PNN)***

Observando as Figura 35 e Figura 36, é possível constatar que as taxas de acertos são iguais entre os modelos com dados balanceados e sem balanceamento, ambas possui uma acuraria de 85.9%; enquanto as taxas de confiabilidade dispõe de pouca diferença, 0.64 com balanceamento e 0.62 sem.

Figura 35 – Resultado do algoritmo PNN com balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	490	156
Evadiu	176	1535
Classificadas corretamentes: 2,025		
Classificadas incorretamentes: 332		
Erro: 14.086%		
Acurácia: 85.914%		
Coeficiente Kappa: 0.649		

Fonte: Próprio Autor.

Figura 36 – Resultado do algoritmo PNN sem balanceamento

targetColumn\Previsão	**Não Evadiu**	**Evadiu**
Não Evadiu	418	228
Evadiu	104	1607
Classificadas corretamentes: 2,025		
Classificadas incorretamentes: 332		
Erro: 14.086%		
Acurácia: 85.914%		
Coeficiente Kappa: 0.624		

Fonte: Próprio Autor.

6 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Neste trabalho, as análises foram realizadas em quatro bases de dados, apresentadas em Tabela 4.1.1. As bases de dados contêm informações sobre os alunos que ingressaram em algum curso de tecnologia da informação no ano de 2013 no Brasil. Os resultados das análises serão discutidos nesta seção.

Ao realizar a análise exploratória sobre os dados, foram geradas informações sobre as taxas de evasão no período avaliado, cujos dados são importantes para a discussão dos resultados. As taxas são apresentadas nas Tabela 2. A primeira tabela mostra as taxas totais de evasão para cada ano analisado; a primeira coluna da tabela mostra as bases analisadas e o número de alunos presentes em cada base, e a segunda coluna mostra as taxas de evasão e a quantidade de alunos que abandonaram o curso naquele ano.

Ao analisar as taxas da Tabela 2, é possível verificar que o primeiro ano do curso é o que mais possui evasão, com um total de 34%, ou seja, 8.960 alunos dos 26.355 que ingressaram em 2013. O terceiro ano cursado, possui a segunda maior taxa de evasão, com um valor de 31%, que corresponde a 4.421 dos 14.263 alunos que ingressaram em 2013 e ainda continuaram no curso.

Tabela 2 – Taxas total de evasão por ano

Ano	Total
2014 (26.355 Alunos)	34% (8.960)
2015 (17.394 Alunos)	18% (3.130)
2016 (14.263 Alunos)	31% (4.421)
2017 (9.841 Alunos)	27% (2.657)

Os algoritmos de Mineração de Dados escolhidos (Decision Tree (DT), Random Forest (RF) e Probabilistic Neural Network (PNN)), são executados com a finalidade de analisar os dados dos alunos ingressantes nos cursos de TI em 2013 no Brasil, com o objetivo de encontrar algum padrão entre os perfis dos alunos que evadiram ou não evadiram durante os quatro anos de duração do curso. Com a utilização de três algoritmos para a realização das análises, foi possível também explorar a performance de cada um e determinar qual se saiu melhor na tarefa. A Figura 37 mostra o comparativo entre os resultados dos algoritmos em relação a taxa de acurácia e coeficiente kappa.

No processo da análise, como já explicado anteriormente, foi realizada a aplicação do algoritmo SMOTE (Synthetic Minority Oversampling Technique), ele foi utilizado para tentar equilibrar a base de treinamento com quantidades aproximadas de cada uma das classes alvo. Ou seja, com o intuito de balancear a base de treinamento. Os resultados apresentados na tabela da Figura 37 mostra as taxas dos modelos com balanceamento e

sem balanceamento. A primeira coluna da tabela aponta cada algoritmo, DT - B é a linha referente ao Decision Tree com balanceamento, enquanto a DT é o resultado do algoritmo sem o balanceamento.

Figura 37 – Resultados das analises x algoritmo

	Base 1		Base 2		Base 3		Base 4	
	Acurácia	Kappa	Acurácia	Kappa	Acurácia	Kappa	Acurácia	Kappa
DT - B	63.1%	0.26	81.8%	0.41	84.5%	0.64	82.4%	0.57
DT -	67%	0.23	81.7%	0.39	85.4%	0.66	81.6%	0.53
RF - B	65.5%	0.29	84.6%	0.50	85.7%	0.66	86%	0.65
RF	68%	0.22	85.8%	0.44	86.4%	0.67	86.1%	0.64
PNN-B	66.9%	0.24	82.1%	0.47	88.5%	0.73	85.9%	0.64
PNN	67%	0.075	85.3%	0.44	87.2%	0.68	85.9%	0.62

Fonte: Próprio Autor.

Ao analisar a tabela é possível constatar que os resultados da Base 3 (analise dos três primeiros anos do curso), tiveram um melhor desempenho, tanto em relação às taxas de acurácia como os coeficiente de kappa, que caracteriza os resultados dos modelos da base 3 com confiabilidade intermediária boa. Em seguida encontra-se a Base 4 e Base 2, que possuem as segundas melhores taxas de acerto e de Kappa, com as taxas de confiabilidade definidas entre moderada e boa. Por último, vem a base 1, que tem o menor percentual de acerto e Kappa, isso se deve à falta de dados sobre os ingressantes, pois há poucas informações que façam os algoritmos ”entenderem”quais características se referem a cada classe, possíveis soluções para isso seriam a utilização de dados da escolaridade dos alunos do ensino médio, dados dos vestibulares e do ENEM (Exame Nacional do Ensino Médio) realizados pelos alunos e, assim, enriquecer as bases de dados atuais, com destaque para a base 1, com menor índice de acerto.

Em relação ao resultado de cada algoritmo, temos um destaque para os algoritmos Random Forest e Probabilistic Neural Network, que tiveram um desempenho melhor em todas as bases, em especial o PNN, que obteve a melhor acurácia e melhor taxa de confiabilidade entre todos as análises.

Ao observar a comparação da Figura 37 em relação ao resultado dos algoritmos com e sem balanceamento, nota-se que as taxas de confiabilidade dos resultados obtidos são maiores quando a base de treinamento está balanceada, exceto na Base 3, que

os resultados dos algoritmos DT e DF possuem maior coeficiente Kappa com o modelo não balanceado. Isso ocorre porque com as bases desbalanceadas no treinamento dos algoritmos, estes tendem a categorizar para a classe com mais registros, enquanto com balanceamento a utilização das classes é equilibrada. A Base 3 tem melhores resultados com dados desbalanceados, pois a quantidade de ”Evadiu”e ”Não Evadiu”nessa base são próximos, então balancear uma base já balanceada com a técnica SMOTE causou perda de informações.

6.0.1 Perfis Descobertos

A seguir serão apresentados alguns dos perfis descobertos no decorrer das análises. Para realizar essa tarefa, foram utilizadas as técnicas de entropia(técnica usada para medir a pureza ou impureza de um determinado conjunto de dados, ou seja, quão diferentes ou iguais esses dados são uns dos outros) e de ganho de informação(técnica que seleciona o atributo que gera o melhor ganho de aprendizagem, em relação ao valor de entropia do conjunto), para descobrir quais variáveis possuem uma maior relevância em relação às classes alvo.

É de entendimento geral, que o algoritmo Decision Tree executa essas operações para “decidir”de forma “inteligente”quais dos atributos vêm na raiz(nó pai) e quais são seus subsequentes(nós filhos). Quando algum desses nós serem folhas(nós sem filhos), o caminho para eles será um dos perfis dos estudantes. Assim, os modelos treinados no algoritmo foram utilizados para verificar se perfis foram encontrados. Vale ressaltar que os modelos aqui analisados não eram aqueles com dados balanceados por sobre-amostragem, pois o uso da técnica criaria ramos cada vez mais específicos, o que dificultaria na análise.

Para selecionar tais perfis, foi seguido a seguinte regra: independentemente da classe(“Não Evadiu”e “Evadiu”) relacionada com o atributo em análise, a folha resultante da sequência de atributos que formaria o perfil deveria ter uma disparidade estatística em porcentagem de 70% ou mais e uma quantidade de registros nessa folha de pelo menos 200 registros.

- Perfis do primeiro ano cursado - Base 1

  A seguir será descrito os resultados de cada perfil e o valor percentual da sua classe.

  - “Não Evadiu”

    1. Estudante que recebe financiamento reembolsável do Fundo de Financiamento Estudantil - FIES, independentemente do sexo, cor e apoio social: 84.5%.
    2. Estudante recebe financiamento estudantil reembolsável do FIES, turno do curso é integral ou vespertino, e o aluno fez alguma das atividades extracurriculares disponíveis: 93,8%.

3. Estudante não recebe financiamento estudantil reembolsável do FIES, e o turno do curso é integral ou vespertino, não faz nenhuma das atividades extracurriculares, porém recebe apoio alimentação: 85.5%.
4. Estudante não recebe financiamento estudantil reembolsável do FIES, turno do curso é integral ou vespertino, não faz nenhuma atividade extracurricular, não recebeu apoio alimentação , pertence a faixa etária 15-24: 71.3%.
5. Estudante não recebe financiamento estudantil, o turno do curso é matutino ou noturno , não realiza atividade extracurricular, pertence a faixa etária 25-54, ingressou por vagas remanescentes e sua cor é parda ou amarela: 74.3%.
6. Estudante não recebe financiamento estudantil,o turno do curso é matutino ou noturno , não faz atividade extracurricular, pertence a faixa etária 15-24, ingressou por reserva de vagas: 79.7%.
7. Estudante não recebe financiamento estudantil, o turno do curso é matutino ou noturno, e realizou atividade extracurricular remunerada, do tipo bolsa extensão: 72.6%.

– "Evadiu"

1. Estudante não recebe financiamento estudantil, o turno do curso é noturno, o aluno não realizou qualquer uma das atividade complementares, e possui faixa etária de 55-64: 75%.
2. Estudante não recebe financiamento estudantil, o turno do curso seja noturno, o aluno não realizou qualquer uma das atividade complementares, e possui faixa etária de 25-54, e seu ingresso foi por meio de vagas remanescentes: 82%

- Perfis dos dois primeiros anos cursado - Base 2

  Em 2015, foi adicionada às variáveis de carga horária do curso e carga horaria integralizada pelo aluno, ou seja quantidade de horas cursadas. A seguir será descrito os resultados de cada perfil e o valor percentual da sua classe.

  – "Não Evadiu"

  1. Estudante tem quantidade de horas integralizadas superior a 685.5: 94.2%.
  2. Estudante tem quantidade de horas integralizadas inferior 685.5, porém maior que 284.5, e o turno do curso é integral: 84.4%.
  3. Estudante de curso de turno vespertino ou integral, carga horária integralizada maior que 284.5 e menor ou igual 685.5, não realiza atividades extracurriculares e recebe apoio moradia: 82.5%.

  – "Evadiu"

1. Estudante cursa no turno noturno e as horas integralizadas é menor ou igual a 401, não utiliza de financiamento estudantil reembolsável, e a quantidade horas do curso é superior a 3367: 82.1

- Perfis dos três anos cursados - Base 3

  A seguir será descrito os resultados de cada perfil e o valor percentual da sua classe.

  – "Não Evadiu"

    1. Estudante possui de horas integralizadas um total maior ou igual a 1964: 80%.
    2. Estudante realiza alguma das atividades complementares, remunerada ou não: 86.9%.
    3. Estudante não realize atividade complementar, pertence a faixa etária 15-24 e a forma de ingresso do estudante foi pelo Exame Nacional do Ensino Médio - ENEM: 74.4%.
    4. Estudante não realize atividade complementar, pertence a faixa etária 15-24 e a forma de ingresso do estudante não foi pelo ENEM, e este recebeu apoio alimentação, estes possuem 87%.

  – "Evadiu"

    1. Estudante tem carga integralizada menor ou igual a 657, carga horária total do curso igual a 4160, o turno é matutino ou noturno, não recebeu nenhum apoio social: 88.4%.

- Perfis dos quatro anos cursados - Base 4

  A seguir será descrito os resultados de cada perfil e o valor percentual da sua classe.

  – "Não Evadiu"

    1. Estudante possui de horas integralizadas superior a 1443.5: 88.7%.
    2. Estudante realizou alguma atividade extracurricular remunerada: 72.7%.
    3. Estudante não realizou nenhuma atividade extracurricular, pertence à faixa etária 22-54, o turno do curso é vespertino, recebeu apoio bolsa permanência, apoio bolsa alimentação: 79.1%.

  – "Evadiu"

    1. Estudante possui de horas integralizadas um valor inferior que 556.5, cor preta ou branca e não recebeu apoio bolsa permanência: 79.9%.

7 CONSIDERAÇÕES FINAIS

Os principais objetivos deste trabalho foram o desenvolvimento de modelos preditivos que possam ser utilizados para alcançar o discente que pretende evadir antes mesmo de tomar essa decisão, e buscar definir quais fatores levaram um aluno a permanecer ou não no curso (Subseção 6.0.1), possibilitando assim a identificação dos alunos em situação de evasão, e dessa forma fornecendo informações úteis para auxiliar universidades no processo de tomada de decisão. Para realizar esses objetivos, foram utilizadas técnicas de descoberta de conhecimento sobre os dados do censo da educação superior brasileira para a geração de informações.

A mineração de dados executada forneceu os resultados esperados, e assim foi possível criar alguns perfis de evasão e não evasão dos alunos (Subseção 6.0.1). Além disto, foi possível também analisar as performances dos algoritmos aplicados. Com tudo, pode-se concluir que, com as análises executadas, a avaliação dos algoritmos e os resultados obtidos atendem ao objetivo geral e os objetivos específicos do trabalho.

Durante a execução do projeto foram encontradas alguns desafios. O principal deles é relacionado a base de dados do censo; todo ano ocorre o censo da educação superior e as metodologias de coleta podem mudar, e assim é gerada uma base de dados diferente, alguns anos possuem algumas variáveis e em outros anos não. Como por exemplo, a variável de carga horário, que não está presente no censo de 2014, mas está presente no censo de 2015 em diante. Assim é criado uma dificuldade ao padronizar as bases de dados para análise.

Com o desenvolvimento do presente trabalho, foi possível destacar algumas contribuições principais, tanto para a área de tecnologia quanto para a educação. Para fins tecnológicos, a principal contribuição deu-se pela comparação de algoritmos de classificação, permitindo, assim, analisar quais os que tinham melhor desempenho com bases de dados educacionais. Na área da educação, as principais contribuições foram os modelos preditivos que permitem identificar com um grau de confiabilidade aceitável a evasão em cursos de tecnologia da informação; e a criação de perfis estudantis baseadas nas variáveis que possuem o maior ganho de informação. Dessa forma, são disponibilizados dados aos gestores universitários, o que permitirá uma melhoria no processo de tomada de decisão para a criação ou alteração de projetos contra a evasão universitária.

Após apresentar todo o desenvolvimento, dificuldades e contribuições do projeto, é especificada algumas recomendações para trabalhos futuros:

- Estender esse estudo com o uso de dados do histórico escolar dos estudantes.
- Agregar um *Data Warehouse* ao projeto, permitindo a análise de vários departamentos presente no censo.

- Ampliar o estudo em mais cursos, não apenas cursos da área de tecnologia.
- Utilizar dados que antecedem o ensino superior com o intuito de melhorar a análise e as características dos perfis descobertos, como os dados ensino médio e ENEM.
- Relacionar os dados do censo com os dados disponibilizados pelo Enade.

Em suma, este trabalho obteve resultados satisfatórios, conseguindo atingir os objetivos gerais e específicos definidos. Espera-se que os resultados obtidos neste trabalho possam auxiliar na tomada de decisões dos gestores universitários nas criações de ações preventivas e mitigadoras relacionadas a evasão.

REFERÊNCIAS

ASTIN, A. W. et al. **College student retention: Formula for student success**. [S.l.]: Rowman & Littlefield Publishers, 2012.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineraçao de dados educacionais: Oportunidades para o brasil. **Revista Brasileira de Informática na Educação**, v. 19, n. 02, p. 03, 2011.

BRACHMAN, R. J.; ANAND, T. The process of knowledge discovery in databases. In: **Advances in knowledge discovery and data mining**. [S.l.: s.n.], 1996. p. 37–57.

CABENA, P. et al. **Discovering data mining: from concept to implementation**. [S.l.]: Prentice-Hall, Inc., 1998.

CASTANHEIRA, L. G. Aplicação de técnicas de mineração de dados em problemas de classificação de padrões. Universidade Federal de Minas Gerais, 2008.

CISLAGHI, R. et al. Um modelo de sistema de gestão do conhecimento em um framework para a promoção da permanência discente no ensino de graduação. Florianópolis, SC, 2008.

COSTA, C. N. et al. Descoberta de conhecimento em bases de dados. **Revista Eletrônica: Faculdade Santos Dumont**, v. 2, 2019.

DADOSABERTOS, P. B. de. **Microdados do Censo da Educação Superior**. 2021. Disponível em: <https://dados.gov.br/dataset/microdados-do-censo-da-educacao-superior>.

ELMASRI, R.; NAVATHE, S. B. **Database systems**. [S.l.]: Pearson Education Boston, MA, 2011. v. 9.

ETHINGTON, C. A. A psychological model of student persistence. **Research in higher education**, Springer, v. 31, n. 3, p. 279–293, 1990.

FAYYD, U. M.; SHAPIRO, G. P.; SMYTH, P. From data mining to knowledge discovery: an overview. AAAI Press/The MIT Press, 1996.

FEDERAL, S. **Despreparo de alunos leva a evasão nos cursos superiores, alerta Cristovam**. 2021. Disponível em: <https://www12.senado.leg.br/noticias/materias/2014/09/10/cristovam-buarque-evasao-no-ensino-superior>.

FILHO, J. P. d. S. As reprovações em disciplinas nos cursos de graduação da universidade federal do ceará (ufc) no período de 2000 a 2008 e suas implicações na evasão discente. http://www. teses. ufc., 2009.

FISHBEIN, M.; AJZEN, I. Belief, attitude, intention, and behavior: An introduction to theory and research. 1977.

G1. **Nº de alunos que abandonam faculdade deve subir após a pandemia, e setores poderão enfrentar falta de mão de obra**. 2020. Disponível em: <https://g1.globo.com/educacao/noticia/2020/09/13/no-de-alunos-que-abandonam-faculdade-deve-subir-apos-a-pandemia-e-setores-poderao-enfrentar-fa ghtml>.

GARCIA, S. C. O uso de árvores de decisão na descoberta de conhecimento na área da saúde. 2003.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining: um guia prático**. [S.l.]: Gulf Professional Publishing, 2005.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. **Data Mining**. [S.l.]: Elsevier Brasil, 2015.

HAN, J.; KAMBER, M. Data mining concepts and techniques san francisco moraga kaufman. 2001.

HAN, J.; KAMBER, M.; PEI, J. Data mining concepts and techniques third edition. **The Morgan Kaufmann Series in Data Management Systems**, v. 5, n. 4, p. 83–124, 2011.

HAND, D. J.; MANNILA, H.; SMYTH, P. **Principles of data mining (adaptive computation and machine learning)**. [S.l.]: MIT Press, 2001.

HEGDE, V.; PRAGEETH, P. Higher education student dropout prediction and analysis through educational data mining. In: IEEE. **2018 2nd International Conference on Inventive Systems and Control (ICISC)**. [S.l.], 2018. p. 694–699.

INEP. Censo da educação superior 2016. **Ministério da Educação-Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira**, 2017.

INEP. Censo da educação superior 2017. **Notas Estatísticas**, 2018.

KELLY, J. d. O. et al. Supervised learning in the context of educational data mining to avoid university students dropout. In: IEEE. **2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)**. [S.l.], 2019. v. 2161, p. 207–208.

KNIME. **KNIME Quickstart Guide**. 2021. Disponível em: <https://docs.knime.com/2020-12/analytics_platform_quickstart_guide/index.html>.

LOBO, M. Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. **Associação Brasileira de Mantenedoras de Ensino Superior. Cadernos**, v. 25, 2012.

LOBO, M. B. d. C. M. **Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. 2012**. 2017.

LOUZADA, F.; DINIZ, C. **Data mining: uma introdução**. [S.l.: s.n.], 2000.

MACHADO, L. d. S.; FRANCELINO, W. L. Mineração de dados nos microdados enade computação. **Ciência da Computação-Tubarão**, 2020.

MALLADA, F. J. R. La gestión del absentismo escolar. **Anuario Jurídico y Económico Escurialense**, n. 44, p. 579–596, 2011.

MANNILA, H. Data mining: machine learning, statistics, and databases. In: IEEE. **Proceedings of 8th International Conference on Scientific and Statistical Data Base Management**. [S.l.], 1996. p. 2–9.

MAO, K. Z.; TAN, K.-C.; SER, W. Probabilistic neural-network structure determination for pattern classification. **IEEE Transactions on neural networks**, IEEE, v. 11, n. 4, p. 1009–1016, 2000.

MARTINS, L. C. B. et al. Early prediction of college attrition using data mining. In: IEEE. **2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)**. [S.l.], 2017. p. 1075–1078.

MAZZETTO, S. E.; BRAVO, C. C.; CARNEIRO, S. Licenciatura em química da ufc: perfil sócio-econômico, evasão e desempenho dos alunos. **Química Nova**, SciELO Brasil, v. 25, n. 6B, p. 1204–1210, 2002.

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

NAVEGA, S. Princípios essenciais do data mining. **Anais do Infoimagem**, 2002.

OLIVEIRA, D. d. **Evasão universitária: uma visão sobre o problema**. 2018. Disponível em: <https://blog.lyceum.com.br/evasao-universitaria/#As_taxas_de_evasao_universitaria_no_Brasil>.

PAL, S. Mining educational data using classification to decrease dropout rate of students. **arXiv preprint arXiv:1206.3078**, 2012.

PINHEIRO, M. F. et al. Identificaçao de grupos de alunos em ambiente virtual de aprendizagem: Uma estratégia de análise de log baseada em clusterizaçao. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. [S.l.: s.n.], 2014. v. 3, n. 1, p. 582.

PRESTES, E. M. d. T.; FIALHO, M. G. D.; PFEIFER, D. A evasão no ensino superior globalizado e suas repercussões na gestão universitária. **Paraíba. Acesso em**, v. 13, 2016.

PREVE, W. F. et al. Evasão no curso de licenciatura em educação física da ufsc. Florianópolis, SC, 2017.

RODRIGUES, L. M. et al. Análise preditiva para identificação de alunos suscetíveis à evasão escolar. **Brazilian Journal of Development**, v. 7, n. 7, p. 71631–71643, 2021.

SHEARER, C. The crisp-dm model: the new blueprint for data mining. **Journal of data warehousing**, THE DATA WAREHOUSE INSTITUTE, v. 5, n. 4, p. 13–22, 2000.

SILVA, H. R. B. da; ADEODATO, P. J. L. A data mining approach for preventing undergraduate students retention. In: IEEE. **The 2012 International Joint Conference on Neural Networks (IJCNN)**. [S.l.], 2012. p. 1–8.

SLYWITCH, E. F. V.; BILAC, D. B. N.; SANTOS, A. L. B. dos. Evasão no ensino superior: Estudo de caso com os alunos do curso de ciêncais contábeis da faculdade itop. **Humanidades & Inovação**, v. 4, n. 5, 2017.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao datamining: mineração de dados**. [S.l.]: Ciência Moderna, 2009.

TINTO, V. Dropout from higher education: A theoretical synthesis of recent research. **Review of educational research**, Sage Publications Sage CA: Thousand Oaks, CA, v. 45, n. 1, p. 89–125, 1975.

UNIVERSIA. **Evasão universitária no Brasil: causas e possíveis soluções**. 2019. Disponível em: <https://www.universia.net/br/actualidad/orientacao-academica/evaso-universitaria-brasil-causas-e-possiveis-soluces-1165821.html>.

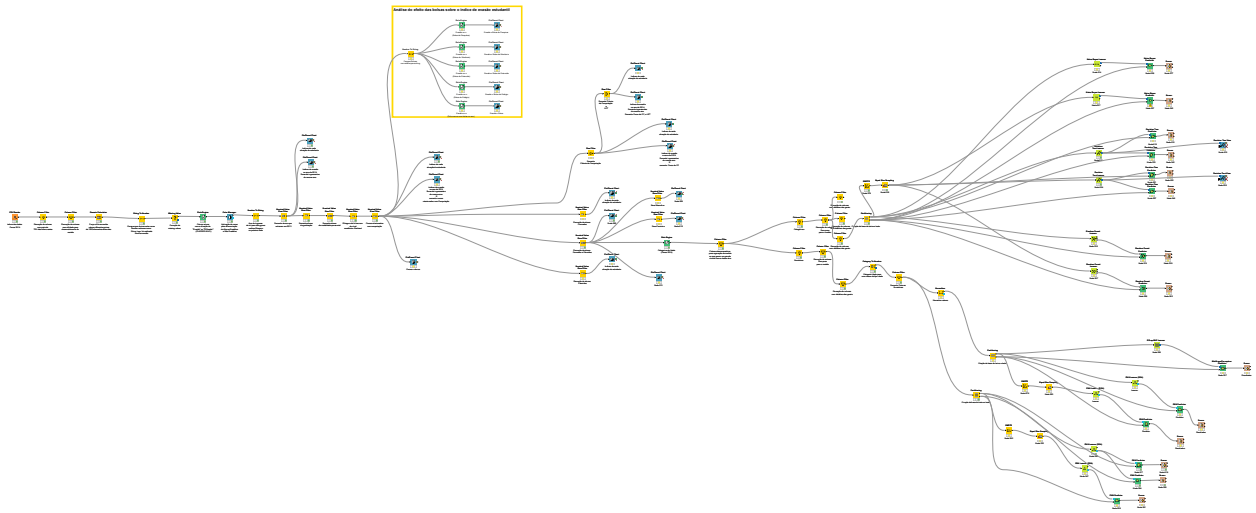
WITTEN, I. H.; FRANK, E. Data mining: practical machine learning tools and techniques with java implementations. **Acm Sigmod Record**, ACM New York, NY, USA, v. 31, n. 1, p. 76–77, 2002.

WWW.EDUCATIONALDATAMINING.ORG. **educationaldatamining.org**. 2021. Disponível em: <https://educationaldatamining.org/>.

ZHOU, Z.-H. Three perspectives of data mining. **Artificial Intelligence**, v. 143, n. 1, p. 139–146, 2003. ISSN 0004-3702. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0004370202003570>.

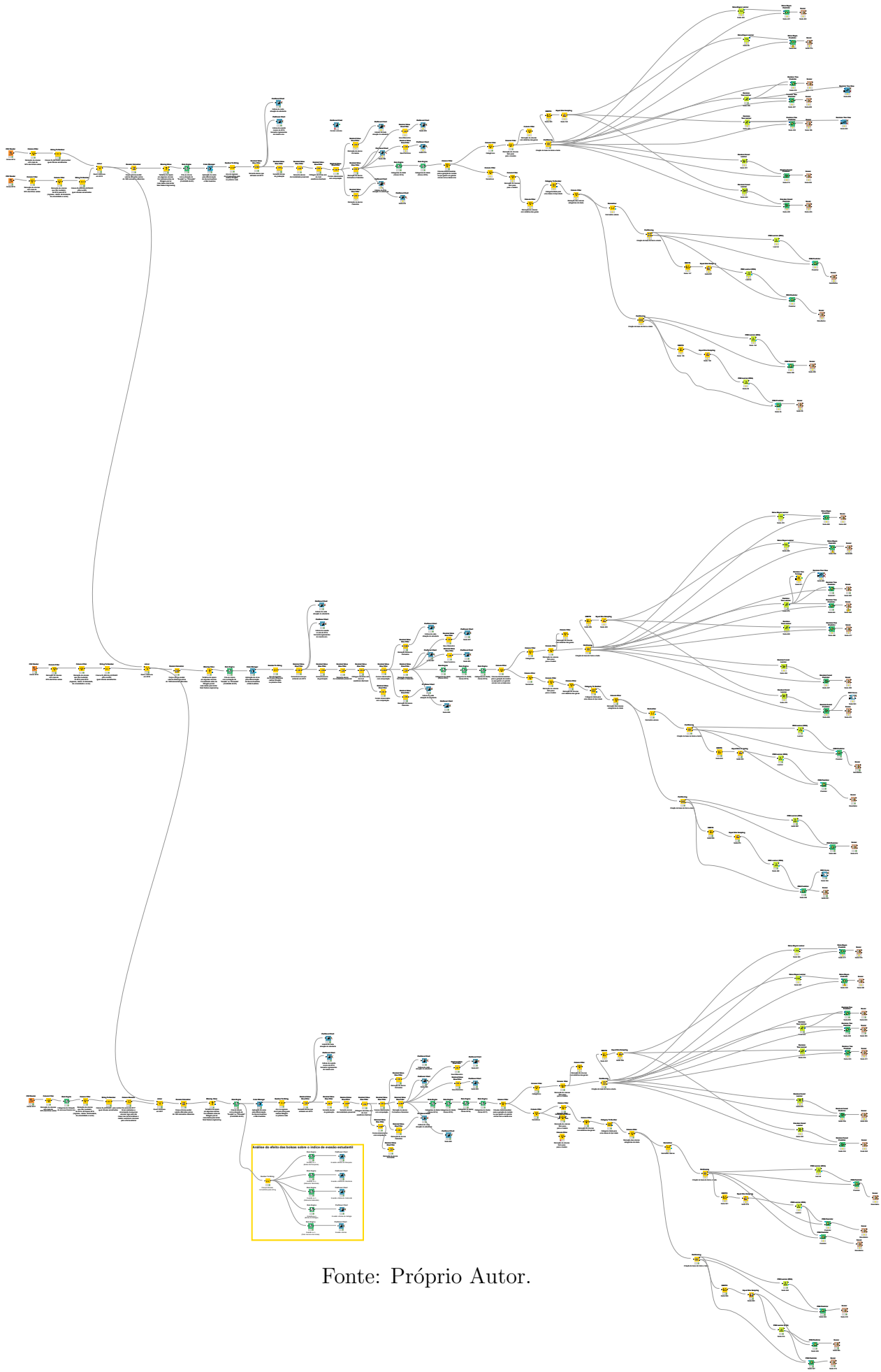
APÊNDICE A - KNIME WORKFLOWS DESENVOLVIDOS

Figura 38 – Knime workflow desenvolvido para análise da base 1.



Fonte: Próprio Autor.

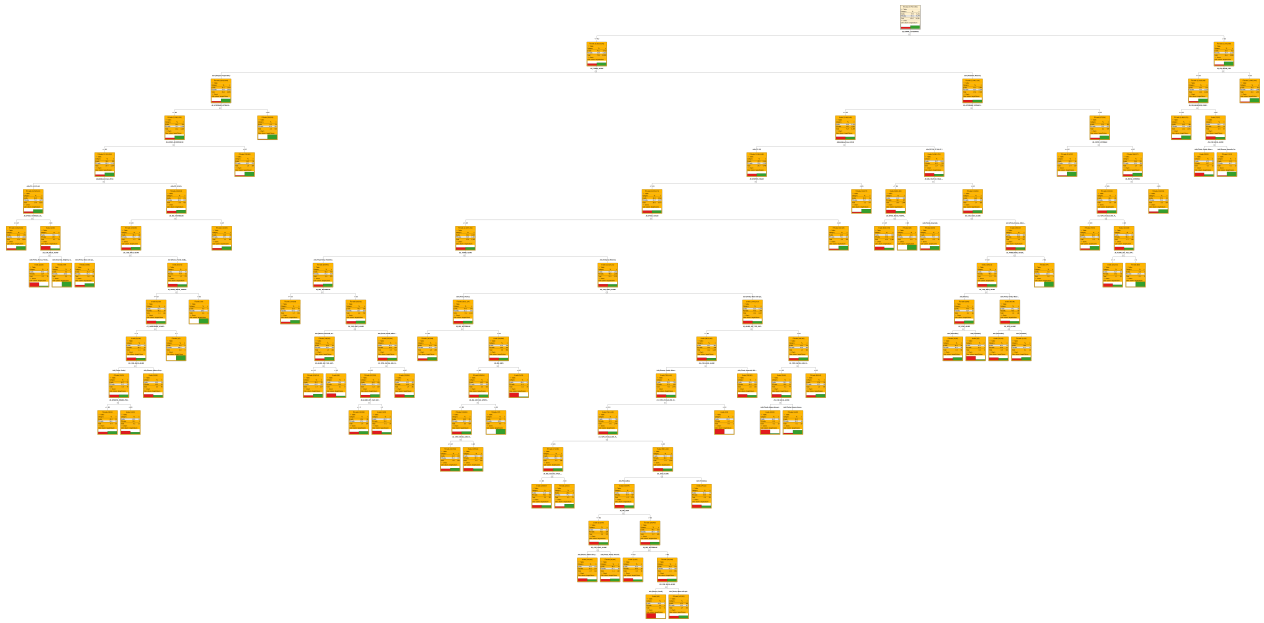
Figura 39 – Knime workflows desenvolvidos para análise das bases 2,3 e 4.



Fonte: Próprio Autor.

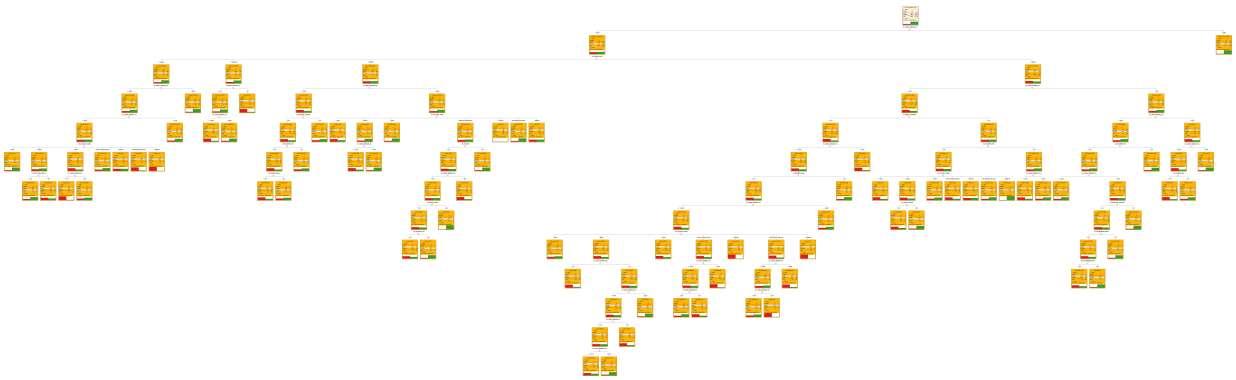
APÊNDICE B - DECISION TREES UTILIZADAS PARA DESCOBERTA DE PERFIS

Figura 40 – Decision Tree utilizada para descobrir perfis estudantis no primeiro ano cursados.



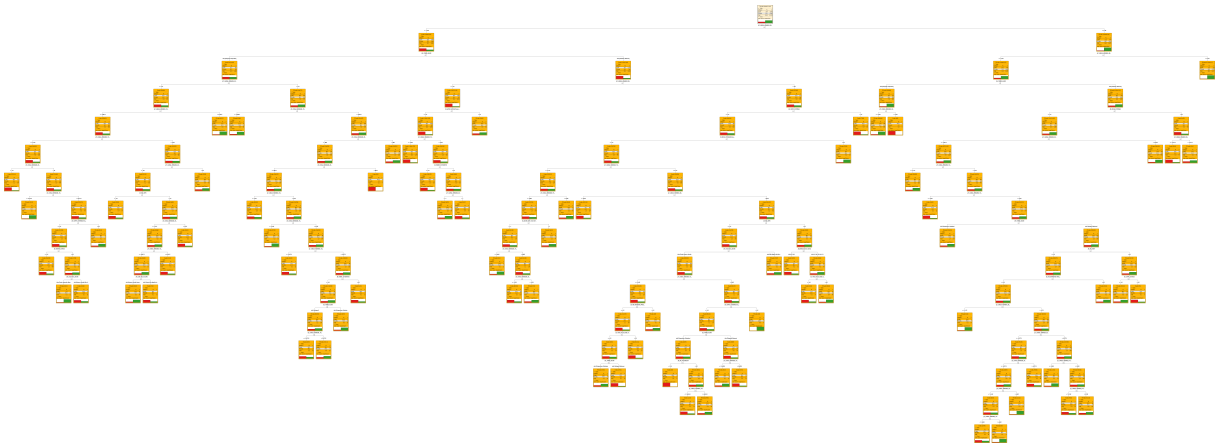
Fonte: Próprio Autor.

Figura 41 – Decision Tree utilizada para descobrir perfis estudantis nos dois primeiros anos cursados.



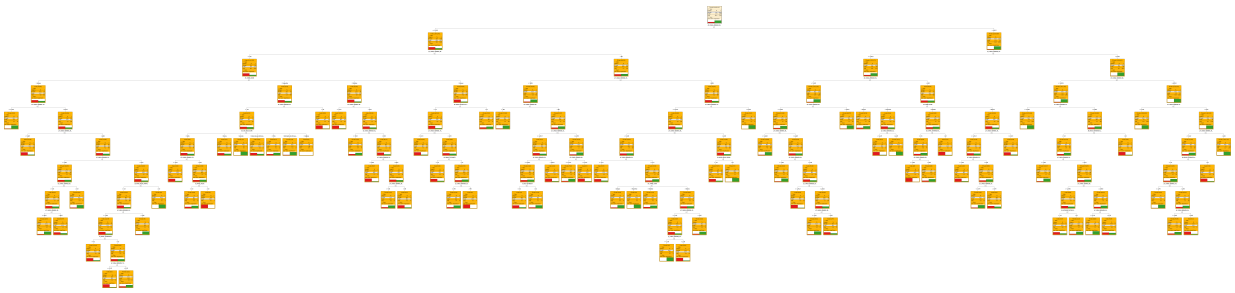
Fonte: Próprio Autor.

Figura 42 – Decision Tree utilizada para descobrir perfis estudantis nos três anos cursados.



Fonte: Próprio Autor.

Figura 43 – Decision Tree utilizada para descobrir perfis estudantis nos quatro anos cursados.



Fonte: Próprio Autor.