INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO RIO GRANDE DO SUL – *CAMPUS* PORTO ALEGRE
MESTRADO PROFISSIONAL EM INFORMÁTICA NA EDUCAÇÃO

NARA MILBRATH DE OLIVEIRA

A EVASÃO EM CURSOS SUPERIORES DE TECNOLOGIA: UMA ABORDAGEM BASEADA EM MODELAGEM PREDITIVA

PORTO ALEGRE - RS
2019

NARA MILBRATH DE OLIVEIRA

A EVASÃO EM CURSOS SUPERIORES DE TECNOLOGIA: UMA ABORDAGEM BASEADA EM MODELAGEM PREDITIVA

Dissertação apresentada junto ao Programa de Pós-graduação *Stricto Sensu* – Mestrado Profissional em Informática na Educação do Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul – *Campus* Porto Alegre, como requisito para obtenção do título de Mestre em Informática na Educação.

Orientador: Mariano Nicolao
Coorientadora: Silvia de Castro Bertagnolli

PORTO ALEGRE - RS
2019

Dados Internacionais de Catalogação na Publicação (CIP)

O48e Oliveira, Nara Milbrath de
A evasão em cursos superiores de tecnologia: uma abordagem baseada em modelagem preditiva / Nara Milbrath de Oliveira. – 2020.
153 f.: il ; 30 cm.

Dissertação (Mestrado) – Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul – Campus Porto Alegre. Mestrado Profissional em Informática na Educação. Porto Alegre, 2020.

Orientadora: Prof. Dr. Mariano Nicolao.
Coorientadora: Profª. Dra. Silvia de Castro Bertagnolli

1. Educação. 2. Ensino superior tecnológico. 3. Evasão. I. Nicolao, Mariano. II. Bertagnolli, Silvia de Castro. III. Título.

CDU 004:37

Elaborada por Débora Cristina Daenecke Albuquerque Moura - CRB10/2229.

AGRADECIMENTOS

Agradeço aos meus filhos, Filipe e Mariana, por estarem comigo nesta caminhada, e serem a razão e o suporte para eu ser quem sou e chegar aonde cheguei. À minha família e amigos pela torcida e por entenderem minha ausência.

Agradeço ao professor Mariano Nicolao e à professora Sílvia de Castro Bertagnolli pela orientação, incentivo e confiança durante este tempo de aprendizagem. Mais que o título de mestre, a amizade e o compromisso com vocês não me permitiram desistir.

Agradeço ao amigo Vitor Bertoncello, por me incentivar a cursar o mestrado, bem como pela ajuda com a extração dos dados dos sistemas. À colega Sabrina Eufrásio, pela revisão da bibliografia. Aos demais colegas de trabalho que choraram, riram, abraçaram, incentivaram e apoiaram nos momentos complicados do trabalho e da vida, em especial a Aline Mesquita, a Angélica Costa, a Andréia Pruinelli, o Marcelo Gonçalves da Silva e a Olívia Tavares.

Agradeço aos professores Márcio Bigolin, Rafael Pinto e à professora Denise Regina Pechmann que, com muita paciência, ouviram longas defesas da construção da pesquisa e auxiliaram na elucidação das dúvidas. À professora Sheila Staudt, pelas revisões do abstract e do resumo.

Por fim, meu muito especial agradecimento à minha “dupla dinâmica”, “fiéis escudeiras”, minhas colegas do Setor de Registro do *Campus* Canoas, Aline da Silveira Muniz e Cintia Lauriane Steindorff Jhanke. Sem o apoio de vocês, tenho certeza, teria desistido. Obrigada pelas escutas, pelos conselhos, pelo apoio incondicional e pela amizade.

Obrigada a todos e a todas que não me deixaram *evadir* do meu propósito e nem de mim.

RESUMO

Na última década, o número de ingressantes em instituições de ensino superior aumentou devido, em grande parte, a políticas educacionais que vêm promovendo uma ampliação do acesso. No contexto da educação pública, além da ampliação do acesso, as Instituições Federais de Ensino Superior têm como objetivo promover a permanência e o êxito dos seus alunos. Como parte dessa realidade, o Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul visa, entre outros fins, à melhoria do desempenho acadêmico e atua, preventivamente, no que diz respeito às situações de retenção e evasão. Na perspectiva de contribuir com a prevenção deste problema, identificando, de maneira precoce, acadêmicos com propensão a abandonar o curso, essa dissertação tem como objetivo criar um modelo para predizer a evasão nos cursos superiores de tecnologia. Sendo assim, optou-se por usar técnicas de Mineração de Dados, dentro de um processo de Descoberta de Conhecimento em Banco de Dados (KDD – *Knowledge Discovery in Databases*). Este trabalho caracteriza-se como um estudo de caso e utilizou dados acadêmicos e sociodemográficos, armazenados nos Sistemas de acompanhamento acadêmico dos estudantes, do IFRS Campus Canoas. A fase de pré-processamento dos dados gerou um Modelo de Dados que poderá ser usado como ponto de partida em outras pesquisas realizadas pelo IFRS. Para construção do modelo de predição, foram avaliados cinco algoritmos de classificação, sendo que o *Decision Tree*, demonstrou melhor desempenho, atingindo a acurácia de 82% na fase de treinamento do Modelo e 60,42% na validação em novos dados. Este trabalho pretende trazer contribuições significativas no que tange ao processo de tomada de decisão dos gestores da instituição, em relação às ações de permanência e êxito, a partir das predições indicativas de evasão, bem como estimular a realização de outros trabalhos sobre a evasão utilizando técnicas de mineração de dados.

Palavras-chave: KDD. Predição. Evasão. Cursos Superiores de Tecnologia.

ABSTRACT

In the last decade, the number of new students entering into higher education institutions has increased, largely due to educational policies that have been promoting increased access. In the context of public education, in addition to expanding access, the Federal Institutions of Higher Education aim to promote the permanence and success of their students. As part of this reality, the Federal Institute of Education, Science and Technology of Rio Grande do Sul aims, among other purposes, at improving academic performance and acts preventively regarding retention and dropout. In order to contribute to the prevention of this problem, early identifying academics with a propensity to drop out, this dissertation aims to create a model to predict dropout in higher technology courses. Therefore, we chose to use Data Mining techniques, within a Knowledge Discovery in Databases (KDD) process. This work is characterized as a case study and used academic and sociodemographic data stored in the IFRS *Campus* Canoas Student Tracking Systems. The data preprocessing phase generated a Data Model that could be used as a start point for other researches concerning the IFRS. To construct the prediction model, five classification algorithms were evaluated. Among them, the Decision Tree showed better performance, reaching an accuracy of 82% in the training phase of the Model and 60.42% in the application of new data. This paper intends to bring significant contributions regarding the decision-making process of the institution's managers, regarding the permanence and success actions, based on the indicative evasion predictions, as well as to stimulate the accomplishment of other works on the evasion using techniques of data mining.

Keywords: KDD. Prediction. Evasion. Higher Technology Courses.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE GRÁFICOS

LISTA DE SIGLAS

AAE	Assessoria de Assistência Estudantil
AE	Assistência Estudantil
CAE	Coordenações de Assistência Estudantil
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CIAAPE	Comissão Interna de Acompanhamento de Ações de Permanência e Êxito dos Estudantes
CONIF	Conselho Nacional das Instituições da Rede Federal de Educação Profissional, Científica e Tecnológica
CONSUP	Conselho Superior
DCBD	Descoberta de Conhecimento em Banco de Dados
DDR/SETEC	Diretoria de Desenvolvimento da Rede Federal de Educação Profissional, Científica e Tecnológica
ENEM	Exame Nacional do Ensino Médio
ETFC	Escola Técnica Federal de Canoas
EPCT	Educação Profissional, Científica e Tecnológica
EPT	Educação Profissional e Tecnológica
GTPAE	Grupo de Trabalho Permanente em Assistência Estudantil
IES	Instituições de Ensino Superior
IFES	Instituições Federais de Ensino Superior
IFRS	Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul
IFs	Institutos Federais
Inep	Instituto Pesquisa e Estatística Anísio Teixeira
MEC	Ministério da Educação
KDD	Knowledge Discovery in Databases
LDBEN	Lei de Diretrizes e Bases da Educação Nacional
MD	Mineração de Dados
PA	Percentual de Aproveitamento
PAE	Política de Assistência Estudantil
PDI	Plano de Desenvolvimento Institucional

PEPEEIFRS	Plano Estratégico de Permanência e Êxito dos Estudantes do IFRS
PEPEE	Plano Estratégico de Permanência e Êxito dos Estudantes
PNE	Plano Nacional de Educação
PNP	Plataforma Nilo Peçanha
PROEJA	Programa Nacional de Integração da Educação Profissional com a Educação Básica na Modalidade de Educação de Jovens e Adultos
PROUNI	Programa Universidade Para Todos
QS	Questionário Sociodemográfico
REVALIDE	Rede de Coleta, Validação e Disseminação das Estatísticas da Rede Federal de Educação Profissional, Científica e Tecnológica
RF	Rede Federal
RFEPCT	Rede Federal de Educação Profissional, Científica e Tecnológica
SETEC/MEC	Secretaria de Educação Profissional e Tecnológica do Ministério da Educação
SETEC	Secretaria de Educação Profissional e Tecnológica
SRE	Setor de Registro Escolar
SIA	Sistema de Informações Acadêmicas
SIAF	Sistema Integrado de Administração Financeira do Governo Federal
SIAPE	Sistema Integrado de Administração de Recursos Humanos
SIFRS	Sistema IFRS
SISTEC	Sistema Nacional de Informações da Educação Profissional e Tecnológica
Sisu	Sistema de Seleção Unificada
TAM	Termo de Acordo de Metas e Compromissos
TICs	Tecnologias da Informação e Comunicação
TCU	Tribunal de Contas da União

SUMÁRIO

1 INTRODUÇÃO

Observa-se que o número de vagas no ensino superior aumentou na última década, seja pelas políticas públicas como o Programa Universidade Para Todos (PROUNI), seja pelo aumento do número e pela ampliação das instituições públicas e privadas que ofertam cursos desse nível de ensino. A reestruturação da Rede Federal de Educação Profissional, Científica e Tecnológica (RFEPCT) em 2008, através da Lei nº 11.892, de 29 de dezembro de 2008 (BRASIL, 2008), contribuiu consideravelmente para esse aumento. A nova organização ampliou o número de instituições vinculadas à Rede Federal. O processo de interiorização, o aumento do número de cursos e da oferta de vagas abriu caminho para a democratização do ensino superior, possibilitando, a milhares de jovens, maior qualificação e formação. A organização multicampi dos Institutos Federais (IFs), acrescida da proposta de adequação e fortalecimento dos arranjos produtivos das regiões nas quais estão inseridos, faz com que cada *campi* tenha características peculiares, culminando com identidades distintas. De acordo com a realidade da comunidade na qual está inserido cada *campi* acolhe uma diversidade de sujeitos, com realidades e interesses diferentes.

Para garantir meios e condições que assegurem a permanência e o êxito, ou seja, a conclusão do curso pelo estudante e, por consequência, o resgate de um direito constitucional e da cidadania desses sujeitos, o Ministério da Educação (MEC), por meio da Secretaria de Educação Profissional e Tecnológica (SETEC), instituiu um grupo de trabalho, através da Portaria SETEC nº 39, de 22 de novembro de 2013 (BRASIL, 2013a), composto por representantes da Rede Federal e da própria secretaria. Este grupo tinha como proposta de trabalho definir, conceituar e dimensionar os fenômenos da evasão e da retenção[1], bem como categorizar e definir as causas e ações para superá-los, baseados na literatura sobre o assunto e no diagnóstico feito pela própria Rede Federal.

[1] Evasão é a interrupção do aluno no ciclo do curso: o estudante pode ter abandonado o curso, não ter realizado a renovação da matrícula ou formalizado o desligamento/desistência do curso. Por outro lado, a retenção consiste da não conclusão do curso no período previsto, fator concorrente para o aumento da propensão em relação à evasão (BRASIL, 2014, p. 20).

E, a partir desses estudos, foi organizado um documento orientador, o qual fornece os “subsídios para o planejamento de ações, para o enfrentamento do fenômeno da evasão e da retenção” na Rede Federal (BRASIL, 2014).

Com base nesse documento orientador, o Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul (IFRS) construiu sua proposta para o Plano Estratégico de Permanência e Êxito (PEPEEIFRS), aprovado pelo Conselho Superior da Instituição (CONSUP), através da Resolução nº 064, de 23 de outubro de 2018, cujo principal objetivo é “propor medidas para superar a evasão e a retenção/reprovação dos estudantes” (IFRS, 2018a). Estudar a evasão a partir dos fatores que motivaram a saída dos ex-alunos é uma alternativa apontada pela comissão de Acompanhamento de Ações de Permanência e Êxito dos Estudantes (CIAAPE), a qual constatou que os motivos são os mais variados e de diferentes ordens. Mesmo conhecendo as possíveis razões, ainda não se sabe precisar se e quando um aluno abandonará seu curso. Para tanto, seria necessário acompanhar a vida de cada estudante dentro e fora da instituição, o que é impossível. Mas, por outro lado, é imprescindível que ações sejam executadas, além de se propor metodologias que permitam analisar informações dos alunos individualmente e que apontem indicativos de propensão à evasão.

Neste contexto, identificou-se como problema de pesquisa desta dissertação: como antecipar ou predizer a propensão de um determinado aluno evadir?

A evasão é o estágio final de um processo multiforme, resultante de uma série de fatores e problemas individuais, sociais, familiares e da instituição de ensino, que culminam com o abandono (RUMBERGER, 2001; FINI, DORE e LÜSCHER, 2013). Dessa forma, com base nas causas da evasão levantadas pela instituição para a construção do seu PEPEEIFRS e descritas na literatura e estudos sobre este problema, é possível buscar indicadores que possam aparecer precocemente, nos dados que constituem o perfil acadêmico e sociodemográfico dos alunos. Não é interesse dessa dissertação analisar, explicar ou justificar os motivos e os dados dos alunos que já evadiram, mas usá-los para identificar outros estudantes/discentes com propensão a sair da instituição sem êxito.

Com o objetivo de colaborar com a melhoria dos indicadores de eficiência e eficácia do IFRS, especialmente no que tange a uma formação humana e profissional exitosa e de qualidade, a partir da ampliação das condições de permanência e êxito dos acadêmicos, essa dissertação tem em sua proposta uma alternativa para identificar antecipadamente aqueles alunos com propensão à evasão. Neste sentido, propõe-se o uso do processo de Descoberta de Conhecimento em Banco de Dados, (KDD – *Knowledge Discovery in Databases)*, com o emprego de técnicas de Mineração de Dados (MD), para criar um modelo preditivo que identifique os alunos com propensão a evadir. A proposta consiste em usar dados pré-existentes dos acadêmicos, com o intuito de extrair conhecimentos a respeito das características dos que concluíram os cursos e dos que saíram sem êxito. Em seguida, construir um “modelo” de predição, que permita identificar com antecedência a propensão à evasão em outros estudantes que ainda estejam com a matrícula em situação regular.

Neste trabalho de pesquisa utilizou-se como estratégia de investigação um estudo de caso, valendo-se dos seguintes procedimentos metodológicos: pesquisa bibliográfica, pesquisa documental e pesquisa exploratória. Quanto à natureza e ao uso dos dados, a pesquisa pode ser caracterizada como qualitativa e quantitativa. O estudo de caso foi realizado no *Campus* Canoas, do Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul (IFRS) e desenvolveu o processo de *KDD* nos dados de aproximadamente 938 alunos dos três cursos superiores de tecnologia desta unidade. As informações para a organização do modelo de dados foram extraídas dos sistemas de acompanhamento acadêmico (SIA - Sistema de Informações Acadêmicas e SIFRS - Sistemas IFRS), correspondentes ao período entre 2011 e 2017.

1.1 **OBJETIVOS**

A partir da identificação do problema de pesquisa dessa dissertação e da maneira pela qual será abordado, tem-se como objetivo geral:

Criar um modelo preditivo para identificar, de maneira precoce, acadêmicos dos cursos superiores de tecnologia do IFRS *Campus* Canoas com propensão à evasão, através de técnicas de mineração de dados, dentro de um processo de Descoberta de Conhecimento em Banco de Dados.

Para alcançar esse objetivo, foram delineados os seguintes objetivos específicos:

- Desenvolver o processo de Descoberta de Conhecimento em Banco de Dados (*KDD – Knowledge Discovery in Databases*), em dados dos sistemas de acompanhamento acadêmico utilizados pelo IFRS *Campus* Canoas;
- Selecionar dados ligados a características e desempenho acadêmico dos alunos dos cursos superiores de tecnologia do IFRS *Campus* Canoas;
- Realizar a etapa de pré-processamento do *KDD*, nos dados selecionados;
- Usar técnicas de mineração de dados para criar um modelo preditivo, com base nos dados pré-processados dos alunos, integrantes dos cursos superiores de tecnologia;
- Criar um modelo que possa predizer a tendência ou propensão dos alunos à evasão, utilizando a ferramenta *RapidMiner*.

1.2 **MOTIVAÇÃO E JUSTIFICATIVA**

A autora do presente trabalho atua como coordenadora do setor de Registro Escolar (SRE) do Campus Canoas, do Instituto Federal de Educação Ciência e Tecnologia do Rio Grande do Sul (IFRS), desde 2010, e tem acompanhado o processo de ingresso e de conclusão dos alunos nos diversos cursos do *Campus* Canoas. Nesse período, foi possível perceber o grande número de alunos que desistem durante o caminho. Os motivos dessa desistência, na trajetória de formação, ou abandono do curso, passam por causas pessoais e familiares; além de questões internas da própria instituição, no caso de não adaptação com o sistema de ensino ou curso; ou, ainda, atinge fatores externos à instituição, como deslocamentos, transporte, troca de horário de trabalho, novo emprego, entre outras possíveis motivações. Estes dados confirmam-se em conversas informais e na justificativa posta, pelo aluno, no formulário preenchido no momento que solicita o cancelamento da matrícula ou a transferência para outra instituição.

O campus Canoas reflete a realidade do IFRS e de muitas outras instituições (IFRS, 2018a). De acordo com dados coletados pelo SRE desta unidade do IFRS, do total de 1641 ingressantes no *campus*, entre 2011 e 2017, nos quatro cursos técnicos integrados ao ensino médio, no curso de Manutenção e Suporte em Informática na modalidade do Programa de Integração da Educação Profissional ao Ensino Médio na Modalidade Educação de Jovens e Adultos (PROEJA), na licenciatura e nos três cursos de nível superior de tecnologia, tem-se um percentual de 40,46% de alunos que saíram dos seus cursos sem concluí-los. Desse percentual uma pequena parte (8,13%) transferiu-se para outras instituições, principalmente nos cursos integrados, pela obrigatoriedade da educação básica; outra solicitou o cancelamento da matrícula (28,61%) com alguma justificativa, mas a maioria não renovou a matrícula (63,25%), ou seja, abandonou o curso e a instituição. Outro dado bastante importante, e que ajudou a definir a delimitação do estudo, é que as desistências, neste período, ocorreram em maior número nos cursos de tecnologia, cerca de 47% dos alunos ingressantes.

Nessa pesquisa, a preocupação se volta para estes percentuais, os quais representam alunos que ocupam uma vaga pública e, mesmo com as ações de permanência e de êxito desenvolvidas pela instituição, abandonam o curso sem alcançar êxito, acarretando prejuízos pessoais, institucionais e sociais. A motivação é identificar com antecedência possíveis candidatos à evasão, aqueles alunos que abandonam o curso sem aviso prévio e, portanto, sem procurar auxílio dos serviços de apoio ao educando; e com esta informação auxiliar nas decisões e ações institucionais, mitigando, com isso, o desfecho de evasão.

Portanto, neste primeiro momento, fez-se um recorte no total de alunos que saíram dos cursos sem concluí-los e tem-se como tema desta dissertação, a evasão nos cursos superiores de tecnologia no *Campus* Canoas do IFRS.

Nem sempre é possível acompanhar o que ocorre com cada aluno, individualmente, e perceber com antecedência situações indicadoras de problemas e dificuldades pessoais, ou com os processos educacionais e de conhecimento, com relacionamentos interpessoais dentro da instituição e tantos outros, para que se possa agir no sentido de evitar a evasão. Pensando no sucesso desses sujeitos, ou seja, na permanência e no êxito e, portanto, na efetiva conclusão do curso, buscou-se um outro caminho que, para além de compreender as causas da evasão, pudesse dar um passo à frente, prever um desfecho futuro, para que os gestores possam balizar suas ações no sentido de minimizá-las.

Com o crescimento da geração de dados, dispositivos, sistemas e tecnologias no campo educacional, técnicas que estavam sendo usadas para análise, tratamento e exploração de dados armazenados em outras áreas (medicina, ciências, negócios, engenharia e outras), passaram a ser empregadas na educação. A Mineração de Dados é uma delas, que utilizando informações oriundas de ambientes educacionais e também administrativos das instituições de ensino, possibilita que análises detalhadas e voltadas para um melhor planejamento e execução de ações sejam realizadas. (BAKER, et al., 2011; SILVA e SILVA, 2015 e SILVA et al., 2017).

Desta forma, optou-se por adotar técnicas de mineração de dados, dentro de um processo de *KDD*, nos dados extraídos dos sistemas de acompanhamento acadêmico, para descobrir conhecimentos sobre os alunos que evadiram e os que concluíram com êxito e, dessa forma, criar um modelo preditivo que possa mostrar com antecedência a tendência ou propensão de outros alunos à evasão. Este estudo de caso foi realizado no Campus Canoas, com a perspectiva de que, a partir dos resultados obtidos, se incentive a organização dos dados produzidos pelos sistemas de acompanhamento dos estudantes utilizados pelo IFRS, em uma base de dados padronizada e estruturada – de preferência única – além da realização de novos estudos sobre a predição da evasão.

1.3 **ESTRUTURA DO DOCUMENTO**

O texto da presente proposta encontra-se organizado em 9 capítulos, conforme descrevem os próximos parágrafos.

O Capítulo 2 discorre sobre o tema da evasão, através de estudo bibliográfico e documental, buscando uma revisão do conceito, dos fatores envolvidos e as principais causas, principalmente no campo da educação superior.

O Capítulo 3, através de estudo documental, apresenta como o problema da evasão vem sendo abordado na Rede Federal, descrevendo, também, como está sendo feito o monitoramento, os conceitos assumidos e as formas de cálculo da evasão.

O Capítulo 4 coloca em evidência o IFRS e o *Campus* Canoas, contando um pouco de suas histórias e constituição, e mostrando as estratégias para permanência e êxito dos estudantes, contrapondo a evasão.

O Capítulo 5 apresenta alguns aspectos teóricos relacionados ao processo de *KDD*, o que é e quais são suas etapas, destacando os métodos para etapa de Mineração de Dados.

O Capítulo 6 descreve, de forma resumida, os trabalhos relacionados, selecionados através da busca realizada no catálogo de teses e dissertações da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), e descreve as técnicas de mineração, os algoritmos e as ferramentas utilizadas por cada autor.

O Capítulo 7 enuncia os aspectos metodológicos utilizados para desenvolver essa proposta: escolha do tipo de pesquisa, delimitação do local e dos sujeitos, a seleção e tratamento dos dados, o processo de *KDD* e a ferramenta utilizada;

O Capítulo 8 detalha as etapas do processo de *KDD*: a seleção e extração dos dados; o trabalho limpeza, transformação e integração dos dados na etapa de pré-processamento; os experimentos realizados na fase de mineração com algoritmos classificadores para criação do modelo preditivo e a validação e análise do modelo.

O Capítulo 9 elenca as conclusões em relação ao problema da evasão no IFRS, a aplicação do processo de *KDD* para predição da evasão e ao modelo de predição criado. Aborda algumas considerações finais sobre a necessidade de padronização e disponibilidade das informações, sobre a importância do que já vem sendo feito para a prevenção da evasão e sobre o trabalho futuro.

2 EVASÃO NO ENSINO SUPERIOR

A evasão ocorre em todos os níveis de ensino, na rede pública e privada, mas para melhor entendê-la e buscar alternativas para minimizá-la, é preciso fazer recortes de acordo com as características que constituem a comunidade educacional em que o problema se evidencia, sempre mantendo uma visão dos aspectos macros que a influência. A evasão é um desafio para os pesquisadores de diferentes áreas do conhecimento que se dedicam a investigar os problemas relacionados à educação, por ser ela um fenômeno complexo, que pode ser analisado de muitos ângulos e levando em consideração diferentes aspectos. Ela ocorre em todos os níveis de ensino, na rede federal, estadual e municipal, no ensino privado, no ensino presencial e/ou a distância, e causa prejuízos irreparáveis nos âmbitos pessoal, institucional e social.

Devido à importância do tema para a educação de maneira geral e, mais especificamente, para as instituições e os sistemas de ensino, pesquisadores têm tentado conceituar, compreender as razões, os fatores causadores, seus desdobramentos e as consequências, analisando de forma qualitativa e quantitativa os dados disponibilizados por órgãos oficiais e/ou coletados individualmente em seus estudos.

Um dos primeiros estudos sobre a evasão no Brasil de forma sistemática, envolvendo a participação de Instituições de Ensino Superior (IES) públicas em âmbito nacional, com o intuito de compreender as causas e buscar alternativas para minimizar o problema, foi realizado pela Comissão Especial para Estudos da Evasão nas Universidades Brasileiras (1996). “Antes deste trabalho, os estudos realizados, sobretudo na segunda metade dos anos 80, enfatizavam apenas levantamentos estatísticos e estudos de casos de forma fragmentada, realizados por iniciativa do MEC (Ministério da Educação) e de universidades públicas.” (MEC, 2014, p. 16)

A Comissão Especial para Estudos da Evasão foi instituída pelo MEC em março de 1995, com representantes indicados pelos dirigentes das IFES (Instituições Federais de Ensino Superior) e do MEC e contou com a participação efetiva de 89,7% das universidades federais do país. Em sua primeira reunião, a comissão estabeleceu como objetivos específicos do estudo:

> 1. Aclarar o conceito de evasão, considerando suas dimensões concretas: evasão de curso, evasão da instituição e evasão do sistema de ensino superior; 2. Definir e aplicar metodologia homogeneizadora de coleta e tratamento de dados; 3. Identificar as taxas de diplomação, retenção e evasão dos cursos de graduação das IESP do país; 4. Apontar causas internas e externas da evasão, considerando as peculiaridades dos cursos e das regiões do país; 5. Definir estratégias de ação voltadas à redução dos índices de evasão nas universidades públicas brasileiras (COMISSÃO ESPECIAL DE ESTUDOS SOBRE A EVASÃO NAS UNIVERSIDADES PÚBLICAS BRASILEIRAS – ANDIFES/ABRUEM/SESu/MEC, 1996, p. 15).

No que diz respeito ao primeiro objetivo, a Comissão Especial de Estudos da Evasão (ANDIFES/ABRUEM/SESu/MEC, 1996, p. 26), caracterizou a evasão considerando suas dimensões concretas:

- **evasão de curso**: quando o estudante desliga-se do curso superior em situações diversas, tais como: abandono (deixa de matricular-se), desistência (oficial), transferência ou reopção (mudança de curso), exclusão por norma institucional;
- **evasão da instituição**: quando o estudante desliga-se da instituição na qual está matriculado;
- **evasão do sistema**: quanto o estudante abandona, de forma definitiva ou temporária, o ensino superior.

O trabalho da comissão vem sendo usado como referência por pesquisadores do tema como Baggi e Lopes (2011), Johann (2012), Amaral (2013), Figueiredo e Salles (2017), que o reafirmam como um marco para os estudos da evasão no Brasil.

Para Johann (2012, p. 65) a evasão “é um fenômeno caracterizado pelo abandono do curso, rompendo com o vínculo jurídico estabelecido, não renovando o compromisso ou sua manifestação de continuar no estabelecimento de ensino.”

De acordo com o trabalho de Dore e Lüscher sobre a permanência e a evasão dos estudantes na educação profissional técnica de nível médio do Estado de Minas Gerais:

> A evasão escolar tem sido associada a situações tão diversas quanto a retenção e repetência do aluno na escola, a saída do aluno da instituição, a saída do aluno do sistema de ensino, a não

> conclusão de um determinado nível de ensino, o abandono da escola e posterior retorno. Refere-se ainda àqueles indivíduos que nunca ingressaram em um determinado nível de ensino, especialmente na educação compulsória, e ao estudante que concluiu um determinado nível de ensino, mas se comporta como um *dropout.* (DORE e LÜSCHER, 2011, p. 775)

Em função da diversidade de situações que podem ser caracterizadas como evasão, os pesquisadores que trabalham com esta problemática têm adotado diferentes formas de analisá-la e compreendê-la, de acordo com o nível de ensino, com seus objetivos e com a realidade que estão pesquisando, o que caracteriza grande parte dos trabalhos como estudo de caso. É preciso ter bem definido qual evasão está em estudo. Referindo-se ao ensino superior, a Comissão de Estudos da Evasão aponta:

> Este cuidado, além de evitar o risco de generalizações ou simplificações desfiguradoras da realidade, permite qualificar adequadamente os dados quantitativos indicadores do desempenho das instituições universitárias. (ANDIFES/ABRUEM/SESu/MEC, 1996, p. 25)

Russel Rumberger (2001, p. 4 a 6), importante pesquisador sobre a evasão nos Estados Unidos, afirma que entender por que os alunos abandonam a escola é a chave para lidar com esse grande problema educacional. Porém, identificar suas causas é extremamente difícil porque ela é influenciada por fatores relacionados tanto ao aluno individualmente, quanto à família, à escola e à comunidade na qual ele vive. Ele considera um fenômeno complexo e apresenta duas perspectivas diferentes para entendê-lo: uma é a individual e a outra é a institucional, que se concentra nos fatores contextuais encontrados nas famílias, escolas, comunidades e colegas dos alunos. A perspectiva individual enfoca os atributos dos estudantes, tais como os seus valores, atitudes e comportamentos, bem como seu envolvimento ou engajamento na vida escolar e como estes contribuem para suas decisões de abandonar ou não os estudos. Segundo ele, mesmo com algumas diferenças, todas as teorias desenvolvidas anteriormente sobre o tema indicam que há duas dimensões para o engajamento: o acadêmico, relacionado à aprendizagem, e o social, que diz respeito às relações estabelecidas com colegas, professores e funcionários. E todas sugerem que abandonar a escola é apenas o estágio final de um processo dinâmico e cumulativo de desengajamento.

Coadunando com o entendimento de que a evasão é o estágio final de um processo, Fini, Dore e Lüscher (2013, p. 256) argumentam que por ser determinada por um complexo de causas individuais e institucionais “[...] uma de suas características mais marcante é a de ser um fenômeno evolutivo e processual, cujo entendimento requer levar em conta a dimensão temporal em que ele se processa [...].” Entender a evasão dessa forma é importante para evidenciar alguns indicadores que aparecem precocemente nas articulações e na sucessão das experiências na trajetória acadêmica dos estudantes, e realizar o planejamento e intervenções de prevenção.

Em seu estudo longitudinal sobre o espaço ocupado pela modalidade de trancamento dentro do fenômeno da evasão, Polydoro (2000, p. 126 e 127) considera trancamento uma forma de evasão, pois a similaridade dos processos torna-se evidente quanto aos motivos alegados pelos alunos para solicitá-lo e “[...] é uma minoria que retorna à IES, entre esses acadêmicos, apenas cerca de metade efetivamente se reintegra à graduação no ano letivo posterior ao trancamento.” De acordo com a sua pesquisa, o trancamento tem uma característica de provisoriedade que pode estar associada a uma circunstância do seu momento de vida ou a uma dúvida do estudante, uma incerteza em relação a abandonar ou não os estudos, mas é uma decisão que se relaciona mais à permanência do que à provisoriedade.

Mesmo que este trabalho não considere o trancamento como uma forma de evasão, como o faz Polydoro, pois o aluno ainda está vinculado ao seu curso e à instituição, essa característica de provisoriedade e de dúvida apontada, quanto a permanecer ou não no curso, demonstra que pode ser considerado como um indicador precoce dentro do processo evolutivo da evasão.

Considerando a evasão o resultado de um processo complexo, para melhor compreendê-la é preciso identificar e entender as particularidades de suas variáveis individuais, institucionais, sócio-culturais e econômicas, mas também as inter-relações que ocorrem entre elas. Essas três dimensões de análise foram apontadas no relatório da Comissão Especial para Estudos da Evasão nas Universidades Brasileiras (ANDIFES/ABRUEM/SESu/MEC, 1996, p. 117) e vêm sendo adotadas por outros pesquisadores como Dore e Lüscher (2011) e Severino et al (2013).

De acordo com Dore e Lüscher (2011, p. 782 e 783), as pesquisas nacionais e internacionais sobre as causas da evasão no nível superior revelaram como principais indicadores o conjunto de condições e antecedentes familiares (escolaridade dos pais, nível socioeconômico, entre outros); a dificuldade de conciliar estudo com trabalho; desconhecimento do curso e/ou imaturidade na escolha profissional; desilusão com o curso; desestímulo do mercado de trabalho; desprestígio das carreiras e não absorção do profissional pelo mercado de trabalho; fraco desempenho acadêmico no primeiro ano do curso devido à formação precária na educação básica; repetência; dificuldade com o corpo docente; dificuldade de adaptação à estrutura do curso. Para as duas autoras, as pesquisas devem incluir:

> [...] necessariamente, além das motivações individuais, os fatores associados à esfera de competência e de atuação da instituição escolar; por exemplo, as áreas tecnológicas em que os cursos são ofertados, as práticas pedagógicas, a programação das disciplinas, os programas de estágio e de outras práticas profissionais, os processos de avaliação, a formação docente, dentre outros aspectos. (DORE e LÜSCHER, 2011, p. 785)

Severino (2013) reforça que compreender as causas do abandono é importante para evitar as experiências de fracasso dos estudantes:

> O entendimento das causas que levam o estudante ao abandono da sala de aula é relevante e necessário a fim de que os esforços dos diversos segmentos – governo, comunidade escolar, família e aluno não sejam infrutíferos, mas que contribuam para evitar o aprofundamento das experiências de fracasso e a decepção pessoal pelas quais passam os alunos evadidos, que são os que mais se prejudicam no processo. (SEVERINO, et al., 2013, p. 4)

Segundo Johann (2012, p. 12) o aluno é o principal prejudicado com a evasão, pois “vive o sentimento de fracasso”, “concebendo uma autoimagem de incapacidade e inferioridade”, principalmente se esta estiver relacionada às sucessivas reprovações, e sente seus efeitos em seu futuro profissional devido à falta de capacitação e habilitação.

João Batista Amaral (2013) buscou a visão dos alunos evadidos para identificar e analisar as causas da evasão nos cursos superiores de graduação do IFCE - *Campus* Sobral, nos anos de 2010 e 2011. Para tanto, analisou as respostas de 35 alunos a um questionário enviado por e-mail a 140 alunos evadidos nesse

período. Dos vinte e um fatores estudados pela pesquisa, os quais os alunos teriam que graduar, numa escala de significância de 1 a 5, a influência de cada fator na sua decisão de deixar o curso, apenas seis foram considerados relevantes para a evasão. Os fatores foram sistematizados em dois grandes grupos quanto à fonte:

> [...] os relativos à **dimensão interna**, ou relacionados à situação dos estudantes evadidos; e os relativos à **dimensão externa** ou ligados à instituição. Como fatores internos, a pesquisa apontou a compatibilização do curso com necessidade de trabalhar, as condições socioeconômicas enfrentadas pelos alunos, descoberta de novos interesses e ingresso/opção por novo curso e a insatisfação com o curso, comprometendo o desempenho nas disciplinas. Na dimensão externa, a pesquisa revelou a falta de ações institucionais para evitar a evasão e a dificuldade de acesso aos benefícios do programa de assistência ao educando. (AMARAL, 2013, p. 83. *grifo nosso*).

De acordo com Amaral, é uma forma reducionista de ver a questão pensar que a instituição deve se preocupar apenas com fatores externos, pois sua ação pode fazer-se presente nos fatores motivadores da evasão que dizem respeito à situação do aluno. Nesse sentido, a IES em primeiro lugar precisa "[...] sentir-se responsável, entre outros fatores, pela permanência e pela satisfação do aluno", o que a levará a desenvolver estratégias para identificar os problemas acadêmicos com maior antecedência e agir de forma preventiva, e, em segundo lugar, "[...] reconhecer a existência do aluno trabalhador e da aluna mãe, oferecendo condições para que os mesmos possam acompanhar e concluir o curso" (AMARAL, 2013, p. 83).

De acordo com as análises de Schmitt (2014, p. 5), nestes últimos quarenta anos a evasão na educação superior foi examinada:

> [...] a partir de múltiplas perspectivas e paradigmas epistemológicos, entre os quais se verificam estudos de enfoques econômico ou socioeconômico, sociológicos, socioeducativos, pedagógicos, psicológicos, interacionistas, culturais, organizacionais, entre outros. A partir de todas essas visões, os estudiosos vêm caracterizando a evasão estudantil na educação superior com um fenômeno complexo, multifatorial, contextual, dinâmico e transitório.

Pode-se perceber que os conceitos, visões e concepções entre os autores que tratam do fenômeno da evasão variam de acordo com a amplitude ou foco da pesquisa e de quais fatores estão sendo priorizados, em muitos aspectos não havendo consenso.

Segundo Silva (2016, p. 621) há um ponto de consenso:

> Existe unanimidade nas pesquisas sobre como lidar com o fenômeno da evasão escolar: **prevenir**, adotar procedimentos para identificar os alunos que estão em situação de risco de abandonar a escola e envidar todos os esforços possíveis para impedir que isso ocorra. Depois que o aluno sai da escola, o seu retorno é muito mais difícil. A prevenção do abandono escolar depende muito dos recursos mobilizados pelas instituições e professores (*grifo nosso*).

Concordando com o que é consenso nas pesquisas, pode-se dizer que para reduzir os percentuais de evasão a números aceitáveis não basta o levantamento estatístico depois que a decisão de abandonar o curso e/ou a instituição já aconteceu. É fundamental detectar a possibilidade quando ela ainda está embrionária, identificar o estudante com propensão a evadir, ou seja, atacar o problema de forma preventiva.

A evasão escolar no ensino superior é um fenômeno que requer múltiplos olhares, por sua complexidade, e precisa ser analisada dentro de um contexto macro e micro. No contexto macro, é preciso compreender os sistemas de ensino, a história das instituições, as políticas públicas para esse setor e todos os fatores conjunturais; no contexto micro, as individualidades, a realidade da comunidade, os processos institucionais, o perfil dos alunos, dos cursos, as práticas escolares dentro e fora da sala de aula. E ter consciência de que todos esses fatores se inter-relacionam e influenciam na escolha dos alunos de continuarem estudando ou não.

Alguns pesquisadores apontam a necessidade de criar políticas de estudo e diagnóstico da evasão de forma sistemática e longitudinal, levando em consideração aspectos qualitativos e quantitativos para a construção de uma visão mais ampla do problema. Ao mesmo tempo, sinalizam que os dados produzidos pelas instituições devem ser divulgados pelos órgãos oficiais, de maneira mais confiável e com informações precisas, para facilitar o cálculo dos indicadores e as pesquisas que podem ser geradas a partir deles. (BAGGI E LOPES, 2011; DORE e LÜSCHER, 2011). Neste sentido, o MEC vem empenhando esforços para melhorar o monitoramento dos indicadores de evasão e retenção, principalmente no âmbito da educação pública federal, com o objetivo de superação desses problemas nos cursos técnicos e de graduação.

3 MONITORAMENTO DA EVASÃO NA REDE FEDERAL DE EDUCAÇÃO PROFISSIONAL, CIENTÍFICA E TECNOLÓGICA

A Rede Federal completou seu centenário em 2009. Ao longo desse século de existência, que iniciou com a criação das primeiras escolas de Aprendizes e Artífices, em 23 de setembro de 1909, a rede passou por várias mudanças, sendo a mais recente no ano de 2008. A Lei nº 11,892, de 29 de dezembro de 2008, transformou as escolas técnicas e agrotécnicas e os Centros Federais de Educação Tecnológica (CEFETS) em institutos federais e instituiu a Rede Federal de Educação Profissional e Tecnológica (BRASIL, 2008).

De acordo com o portal da Rede Federal[2] e o portal do Conselho Nacional das Instituições da Rede Federal de Educação Profissional, Científica e Tecnológica (CONIF[3]), hoje a RFEPCT está constituída por 38 institutos federais, presentes em todos os Estados brasileiros; dois CEFETS; 25 escolas vinculadas à universidades; o Colégio Pedro II e uma universidade tecnológica - englobando 644 *campi*, em que atuam cerca de 80 mil servidores (professores e técnicos administrativos) para atender mais de um milhão de alunos matriculados (BRASIL, 2016b; CONSELHO NACIONAL DAS INSTITUIÇÕES DA REDE FEDERAL DE EDUCAÇÃO PROFISSIONAL, CIENTÍFICA E TECNOLÓGICA (CONIF), 2018).

Atuando na educação profissional e tecnológica nas diferentes modalidades de ensino (superior, básica e profissional), a Rede Federal vivencia a maior expansão da sua história. Cumprindo um papel importante na inclusão e na formação de trabalhadores qualificados e cidadãos atuantes, a Rede leva à grandes centros urbanos e, da mesma forma, à pequenas cidades do interior do Brasil, a educação gratuita e de qualidade para oportunizar desenvolvimento tecnológico, econômico e social. Considerando a importância dessa atuação e do compromisso assumido com a sociedade, é fundamental a construção e ampla utilização de mecanismos que acompanhem e expressem o nível de alcance das metas e objetivos firmados.

[2] Portal da Rede Federal: disponível em http://redefederal.mec.gov.br/
[3] CONIF: disponível em http://portal.conif.org.br/br/

A Secretaria de Educação Profissional e Tecnológica do Ministério da Educação (SETEC/MEC) é responsável pela elaboração e coordenação das políticas para a Educação Profissional e Tecnológica (EPT) de todo Brasil. Sua atribuição é formular, implementar, monitorar, avaliar e induzir políticas, programas e ações que assegurem a ampliação da oferta e a melhoria da eficiência, da eficácia e da efetividade das instituições federais de educação profissional (BRASIL, 2017). Nesse sentido, vem aprimorando conceitos, cálculos, indicadores de gestão e sistemas de informação que permitam a coleta de dados de forma atualizada, sistêmica, padronizada as quais possam, com isso, auxiliar no monitoramento e na avaliação das instituições.

No ano seguinte à institucionalização da Rede, a SETEC criou um mecanismo de registro e divulgação dos dados da educação profissional e tecnológica e de validação de diplomas de cursos de educação profissional técnica de nível médio, denominado Sistema Nacional de Informações da Educação Profissional e Tecnológica (SISTEC[4]). O SISTEC é um programa eletrônico do governo federal e foi instituído e implantado pelo Ministério da Educação (MEC) em setembro de 2009 (BRASIL, 2009). Nele, as instituições de ensino ofertantes de educação profissional e tecnológica inserem as informações sobre os cursos técnicos de nível médio e os cursos de qualificação profissional, incluindo matrícula, frequência, concluintes, entre outros dados.

O cadastro e o acesso ao sistema são permitidos às instituições participantes e aos seus usuários e as duas modalidades são feitas de forma distinta para alunos, órgãos validadores, unidades de ensino e seus representantes. Cabe ressaltar que o acesso dos alunos para confirmação da validade do seu diploma está na página inicial. Para obter mais informações, pesquisadores e o público em geral precisam solicitar acesso para terem uma senha disponibilizada.

Os dados da educação superior da Rede Federal também são coletados através do Censo da Educação Superior (CENSUP), realizado anualmente pelo Instituto de Pesquisa e Estatística Anísio Teixeira (INEP). Segundo informações constantes no portal do Inep[5], o Censup é o instrumento de pesquisa mais completo do Brasil sobre as instituições de educação superior. Os dados são coletados a

[4] SISTEC: disponível em http://portal.mec.gov.br/sistec-inicial/sistec-apresentacao
[5] Inep: http://portal.inep.gov.br/web/guest/censo-da-educacao-superior

partir do preenchimento dos questionários, por parte das IES e por importação de dados do Sistema e-MEC[6], e reúne informações sobre as instituições de ensino superior, seus cursos de graduação presencial ou a distância, cursos sequenciais, vagas oferecidas, inscrições, matrículas, ingressantes e concluintes e informações sobre docentes nas diferentes formas de organização acadêmica e categoria administrativa. A partir de 2009, o Censo traz as informações de alunos e profissionais individualmente, permitindo análises mais aprimoradas; o acompanhamento minucioso das políticas do setor e seus participantes, bem como o planejamento e a avaliação de políticas públicas, além de contribuir no cálculo de indicadores de qualidade como o Cálculo Preliminar de Curso (CPC) e Índice Geral de Cursos (IGC). Todas essas informações são divulgadas pelo InepData em forma de sinopse estatística e microdados, assim como o resumo técnico.

A coleta e a divulgação dos dados da educação superior pelo INEP, através do sistema CenSup[7], possibilita maior transparência sobre o percentual da população que busca o ensino superior e sobre a oferta e ampliação de vagas, bem como sobre o número de alunos que obtêm sucesso na sua formação, além de permitir estudos e pesquisas sobre esse nível de ensino de uma forma geral. Porém, o recorte específico das informações da Rede Federal não é possível, pois são apresentados dados dos IFs e CEFETs sem distinção dos dados das 25 escolas vinculadas a universidades, do Colégio Pedro II e da Universidade Tecnológica do Paraná.

No final do ano de 2013, através da Portaria nº 39/2013 SETEC/MEC, foi criado um grupo de trabalho com a atribuição de elaborar relatório dos índices de evasão, retenção e conclusão, desagregados para diferentes modalidades de cursos; bem como elaborar manual de orientação para o combate à evasão, incluindo o diagnóstico de aluno ingressante com propensão à evasão, identificação das causas e utilização de monitorias, tutorias e reforço escolar (BRASIL, 2013a). O grupo composto por membros da própria SETEC e de representantes da Rede Federal, foi uma das várias frentes de trabalho organizadas para elaborar um plano de ação em resposta ao Acórdão nº 506, de 2013 (BRASIL, 2013b). O documento criado com o nome de Documento Orientador para a Superação da Evasão e

[6] e-MEC: disponível em http://emec.mec.gov.br/

[7] Censup: disponível em http://sistemascensosuperior.inep.gov.br/censosuperior_2017

Retenção na Rede Federal de Educação Profissional, Científica e Tecnológica tem o propósito de:

> [...] orientar o desenvolvimento de ações capazes de ampliar as possibilidades de permanência e êxito dos estudantes no processo formativo oferecido pelas instituições da Rede Federal, respeitadas as especificidades de cada região e território de atuação. Assim, oferecem-se subsídios para a criação de **planos estratégicos institucionais** que contemplem o diagnóstico das causas de evasão e retenção e a implementação de políticas e ações administrativas e pedagógicas de modo a ampliar as possibilidades de permanência e êxito dos estudantes no processo educativo. (BRASIL, 2014, p. 4. *grifo nosso*)

A construção do documento envolveu as instituições na elaboração de diagnósticos locais sobre evasão e retenção em cursos técnicos e de graduação; indicação de causas e medidas de combate; participação em oficina para elaborar a proposta para o plano estratégico de intervenção e monitoramento para superação desses problemas.

Pode-se pensar esse trabalho como uma retomada da preocupação com a evasão nas instituições públicas federais, e uma ampliação do trabalho da Comissão Especial para Estudos da Evasão, pois a proposta, além de diagnóstico da evasão, objetivou indicar as causas e as medidas de combate desse problema na Rede Federal. O GT sistematizou as causas da evasão apontadas pelas instituições, separadas em três categorias, descritas no quadro abaixo, tendo por referência a proposta da Comissão Especial de 1996.

Quadro 1 - Fatores e causas da evasão

FATORES	CAUSAS
FATORES INDIVIDUAIS: são aqueles ligados às características do estudante.	● adaptação à vida acadêmica; ● capacidade de aprendizagem e habilidade de estudo; ● compatibilidade entre a vida acadêmica e as exigências do mundo do trabalho; ● descoberta de novos interesses ou novo processo de seleção; ● encanto ou motivação com o curso escolhido; ● escolha precoce da profissão; ● qualidade da formação escolar anterior; ● informação a respeito do curso; ● outras questões de ordem pessoal ou familiar; ● participação e envolvimento em atividades acadêmicas; ● personalidade; ● questões de saúde do estudante ou de familiar; ● questões financeiras do estudante ou da família.
FATORES INTERNOS: são problemas relacionados à infraestrutura, ao currículo, à gestão administrativa e didático-pedagógica da instituição	● atualização, estrutura e flexibilidade curricular; ● cultura institucional de valorização da docência; ● existência e abrangência dos programas institucionais para o estudante (assistência estudantil, iniciação científica, monitoria); ● formação do professor; ● gestão acadêmica do curso (horários, oferta de disciplinas etc.); ● gestão administrativa e financeira da unidade de ensino; ● inclusão social e respeito à diversidade; ● infraestrutura física, material, tecnológica e de pessoal para o ensino; ● motivação do professor; ● processo de seleção e política de ocupação das vagas; ● questões didático-pedagógicas; ● relação escola-família.
FATORES EXTERNOS: relacionam-se às políticas públicas, ou ausência delas, às dificuldades financeiras do estudante de permanecer no curso e às questões inerentes à futura profissão	● avanços tecnológicos, econômicos e sociais; ● conjuntura econômica e social; ● oportunidade de trabalho para egressos do curso; ● políticas governamentais para a educação profissional e tecnológica e para a educação superior; ● questões financeiras da instituição; ● reconhecimento social do curso; ● valorização da profissão.

Fonte: construído pela autora, a partir do Documento Orientador para a Superação da Evasão e Retenção na RFEPCT. (BRASIL, 2014).

O grupo de trabalho adotou como conceito de evasão:

> [...] a interrupção do aluno no ciclo do curso. Em tal situação, o estudante pode ter abandonado o curso, não ter realizado a renovação da matrícula ou formalizado o desligamento/desistência do curso. Por outro lado, a retenção consiste da não conclusão do curso no período previsto, fator concorrente para o aumento da propensão em relação à evasão. (BRASIL, 2014, p. 20)

Assim, cumprindo o seu propósito de orientar Rede Federal na elaboração do Plano Estratégico de Intervenção e Monitoramento para Superação da Evasão, o Documento Orientador (BRASIL, 2014, p. 29) indica que as metas e as ações do plano deverão estar previstas no PDI das instituições e os resultados no relatório anual de gestão institucional, apresentando uma proposta metodológica para sua construção, composta por quatro fases:

- Fase 1: instituição de comissão interna;
- Fase 2: elaboração de diagnóstico quantitativo;
- Fase 3: elaboração de diagnóstico qualitativo;
- Fase 4: consolidação do plano estratégico.

Todo o trabalho sistematizado no Documento Orientador foi repassado às instituições da Rede Federal através da Nota Informativa nº 138/2015/DPE/DDR/SETEC/MEC, de 9 de julho de 2015, cujo assunto informa e orienta as instituições da Rede Federal sobre a construção dos Planos Estratégicos Institucionais para Permanência e Êxito dos Estudantes. A nota também determina que os diagnósticos e os planos estratégicos, após a aprovação dos conselhos superiores das instituições, sejam encaminhados para a recém-criada Comissão Permanente de Acompanhamento das Ações de Permanência e o Êxito dos Estudantes da Rede Federal. (BRASIL, 2015a).

A Comissão Permanente de Acompanhamento das Ações de Permanência e Êxito foi instituída pela Portaria nº 23, de 10 de julho de 2015 (BRASIL, 2015b), com caráter consultivo, propositivo e de assessoramento das instituições da Rede Federal, com a finalidade de:

I. orientar as instituições da RF na elaboração e aperfeiçoamento dos planos estratégicos para a permanência e êxito dos estudantes da Rede Federal;

II. receber, analisar e propor melhorias aos planos estratégicos para a permanência e o êxito dos estudantes;

III. monitorar e acompanhar a execução dos planos estratégicos nas instituições da Rede Federal;

IV. propor mecanismos de divulgação das ações institucionais e dos seus resultados.

A SETEC publicou, em agosto do mesmo ano, a Portaria nº 25 (BRASIL, 2015c) a qual define conceitos e estabelece fatores para realização dos cálculos dos indicadores de gestão das instituições da RFEPCT. Os conceitos definidos são:

a) **Aluno Ingressante** (em um dado período): é o aluno que realiza matrícula inicial no período e tem seu registro associado a um ciclo de matrícula de curso no SISTEC;
b) **Aluno Matriculado** (em um dado período): é o aluno com a situação “Em curso” no SISTEC em pelo menos um dia no período considerado e que não esteja retido por tempo maior do que a duração do seu ciclo;
c) **Aluno Retido**: é o aluno que permanece matriculado por período superior ao tempo previsto para integralização do curso;
d) **Aluno-Equivalente**: definido na Portaria MEC nº 818/2015, é calculado a partir do produto do Aluno Matriculado pelo Fator de Equiparação de Carga Horária de curso e pelo Fator de Esforço de Curso, ou seja: Aluno-Equivalente = Aluno Matriculado X Fator de Equiparação de Carga Horária X Fator de Esforço de Curso;
e) **Ciclo de Matrícula**: envolve a oferta de um curso com uma carga horária definida, com a mesma data de início e de previsão de término, com o objetivo de englobar um conjunto de matrículas de alunos no SISTEC, para a obtenção de uma mesma certificação ou diploma.

No artigo 10 desta mesma Portaria, o MEC anuncia que a SETEC publicará o manual com os indicadores, suas fórmulas de cálculo, critérios de agregação, período de abrangência e demais informações necessárias, utilizando os conceitos definidos nesta portaria. Em 2016, a SETEC evoluiu o Manual para Produção e Análise dos Indicadores da Rede Federal de EPCT para o Manual para cálculo dos indicadores de gestão das instituições da Rede Federal de EPCT, levando em consideração a evolução nos métodos de cálculo, a extração dos dados que compõem os indicadores, as novas regulamentações e legislações que ampliam sua finalidade, acrescentando novos indicadores.

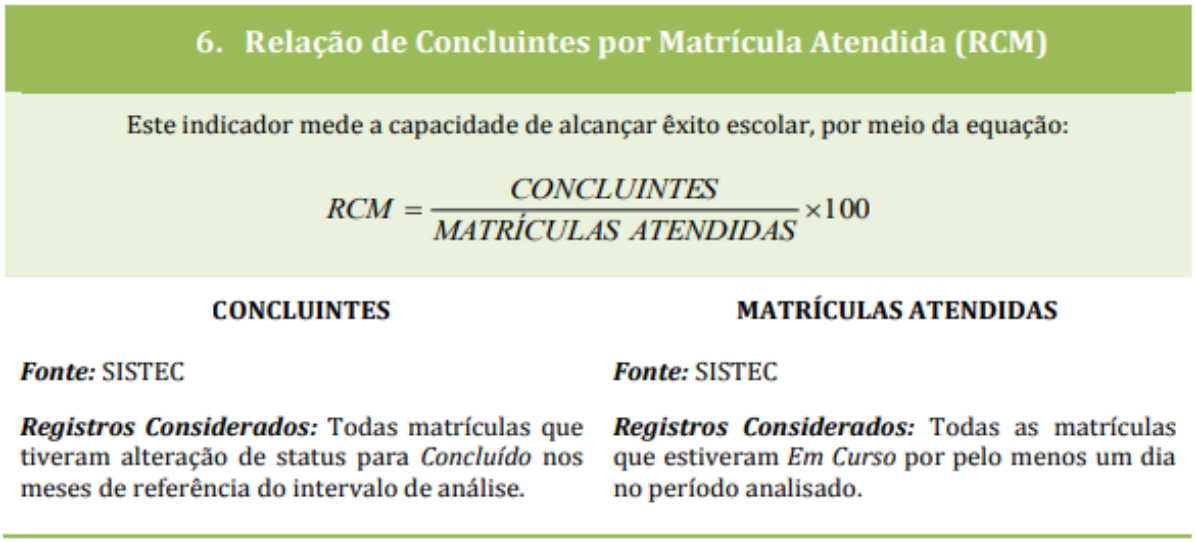
Destacam-se, a seguir, os principais conceitos que estão definidos no manual para cálculo dos indicadores de gestão das instituições da Rede Federal de EPCT (BRASIL, 2016a), além dos já citados acima, e a forma de cálculo de dois indicadores, os quais são importantes para este trabalho.

- **Desligado**: é o aluno que solicita o cancelamento de sua matrícula junto à secretaria da unidade escolar;

- **Evadido**: é o aluno que não possui nenhuma possibilidade regulamentar de retorno ao curso no mesmo ciclo de matrícula, geralmente por faltas além de 25% e não trancamento de matrícula;
- **Integralizado Fase Escolar (Integralizado)**: é o aluno que concluiu disciplinas, módulos ou créditos, mas que por não ter sido aprovado no estágio obrigatório ou ter concluído o TCC, ainda não está apto a colar grau e não é considerado “concluinte”;
- **Matrículas Atendidas:** corresponde ao número total de matrículas na instituição dentro de um determinado período de tempo, independentemente da situação atual da matrícula. Em síntese, corresponde ao total de matrículas que estiveram “em curso” por pelo menos um dia, dentro do período de análise;
- **Matrículas Finalizadas (Finalizados)**: refere-se às matrículas que foram finalizadas, independentemente do êxito ou não do aluno. Nesta modalidade, o aluno pode ter concluído, evadido, desligado ou transferido;
- **Número de alunos retidos (Retidos)**: é o número de alunos que permanece matriculado por período superior ao tempo previsto para integralização do curso;
- **Número de concluintes (Concluintes)**: concluinte é o aluno que integralizou todas as fases do curso, incluindo disciplinas, módulos ou créditos, estágio obrigatório, Trabalho de Conclusão de Curso (TCC), etc e está apto a colar grau;
- **Transferido Externo**: o aluno é transferido de uma unidade para outra unidade de ensino;
- **Vagas Ofertadas**: número de vagas ofertadas, por curso e *campus* dentro do período em análise, em editais de oferta de vagas por meio do SISU, ENEM, vestibular, processos seletivos, sorteios e/ou outras formas de ingresso.

Considerando a Relação de Concluintes por Matrícula Atendida (RCM) – esse indicador possibilita quantificar, em termos percentuais, o número de alunos que concluíram seu curso com êxito. Pode ser calculado de forma individual para cada curso ou para a instituição como um todo.

Figura 1 - Cálculo da Relação de Concluintes por Matrícula Atendida



6. Relação de Concluintes por Matrícula Atendida (RCM)

Este indicador mede a capacidade de alcançar êxito escolar, por meio da equação:

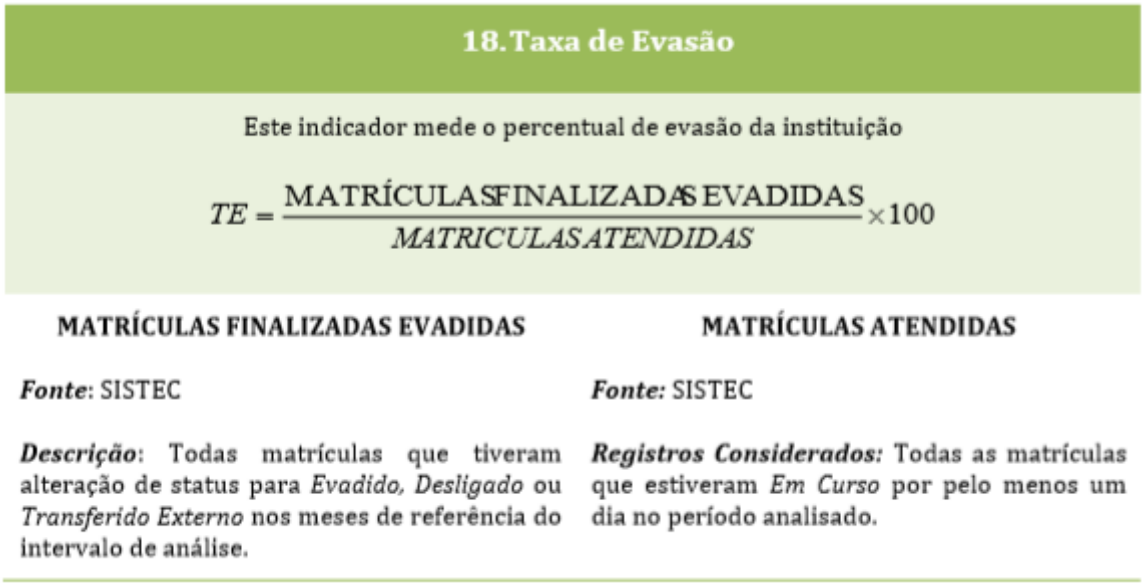
$$RCM = \frac{CONCLUINTES}{MATRÍCULAS\ ATENDIDAS} \times 100$$

CONCLUINTES	MATRÍCULAS ATENDIDAS
Fonte: SISTEC	***Fonte:*** SISTEC
Registros Considerados: Todas matrículas que tiveram alteração de status para *Concluído* nos meses de referência do intervalo de análise.	***Registros Considerados:*** Todas as matrículas que estiveram *Em Curso* por pelo menos um dia no período analisado.

Fonte: Manual para cálculo dos indicadores de gestão das Instituições da RFEPCT. (BRASIL, 2016a).

Outro indicador é a Taxa de Evasão, a qual permite quantificar, em termos percentuais, o número de alunos que interromperam o ciclo do seu curso, incluindo aqui as situações de abandono, não renovação da matrícula, solicitação de desligamento/desistência e transferência externa.

Figura 2 - Cálculo da Taxa de Evasão



18. Taxa de Evasão

Este indicador mede o percentual de evasão da instituição

$$TE = \frac{MATRÍCULAS\ FINALIZADAS\ EVADIDAS}{MATRICULAS\ ATENDIDAS} \times 100$$

MATRÍCULAS FINALIZADAS EVADIDAS	MATRÍCULAS ATENDIDAS
Fonte: SISTEC	***Fonte:*** SISTEC
Descrição: Todas matrículas que tiveram alteração de status para *Evadido, Desligado* ou *Transferido Externo* nos meses de referência do intervalo de análise.	***Registros Considerados:*** Todas as matrículas que estiveram *Em Curso* por pelo menos um dia no período analisado.

Fonte: Manual para cálculo dos indicadores de gestão das Instituições da RFEPCT. (BRASIL, 2016a).

Os conceitos, formas de cálculos e os indicadores aqui apresentados, e constantes no manual para cálculo dos indicadores de gestão das instituições da RFEPCT, vem sendo padronizados desde 2012 pela Rede Federal e pela SETEC, com o objetivo de realizar ”[...] uma análise mais abrangente da situação do ensino tecnológico nas instituições e, a partir delas, uma análise da situação da educação tecnológica do país e sua evolução.” (BRASIL, 2016a, p. 31) que aponte e justifique tanto resultados positivos como negativos desses indicadores em resposta aos acórdãos estabelecidos com o Tribunal de Contas da União (TCU), quanto ao cumprimento do Termo de Acordo de Metas e Compromissos (TAM), constante da Lei nº 11,892, a qual institui a RF e os IFs e ao Plano Nacional de Educação (PNE).

No caminho crescente da organização, formulação e implementação de ações para monitorar, avaliar e induzir políticas, programas para o aperfeiçoamento e qualificação EPT, a SETEC instituiu a Plataforma Nilo Peçanha[8] – PNP e a Rede de Coleta, Validação e Disseminação das Estatísticas da Rede Federal de Educação Profissional, Científica e Tecnológica – REVALIDE, através da Portaria nº 1, de 3 de janeiro de 2018. De acordo com a portaria, a PNP é um ambiente virtual de coleta, validação e disseminação das estatísticas oficiais da RFEPCT e reunirá dados relativos ao corpo docente, discente, técnico-administrativo e de gastos financeiros das unidades da Rede Federal, para fins de cálculo dos indicadores de gestão monitorados pela SETEC/MEC. Quanto à REVALIDE, estrutura colaborativa responsável pelas informações contidas na PNP, tem como participantes os responsáveis pelo registro acadêmico local (RA) de cada unidade de ensino da Rede Federal; os diretores de cada unidade de ensino; os Pesquisadores Institucionais (PIs), ou cargo equivalente, responsável pela produção da estatística educacional de cada instituição; os dirigentes máximos de cada instituição da Rede Federal; a Diretoria de Desenvolvimento da Rede Federal de Educação Profissional, Científica e Tecnológica (DDR/SETEC). Consta, também, na portaria, as competências de todos os envolvidos na coleta e validação das informações (BRASIL, 2018).

As bases das quais são extraídos os dados PNP, de acordo com o calendário organizado pela SETEC, para compor as estatísticas e indicadores da Rede Federal

[8] Plataforma Nilo Peçanha: disponível em https://www.plataformanilopecanha.org/

são: o Sistema Integrado de Administração Financeira (SIAFI), o Sistema Integrado de Administração de Recursos Humanos (SIAPE), o Sistema Nacional de Informações da Educação Profissional e Tecnológica (SISTEC) e o Formulário para Informações sobre a Política de Propriedade Intelectual das Instituições Científicas, Tecnológicas e de Inovação do Brasil (FORMICT). Depois de carregados para PNP, os dados serão qualificados através da observação das regras de consistência pela REVALIDE, garantindo a confiabilidade das estatísticas educacionais. (MORAES et al, 2018, p. 33).

A plataforma apresenta os dados dos indicadores de cada instituição, disponibilizados ao público em dezessete painéis, com colunas expansíveis e filtros que possibilitam a escolha de múltiplas opções simultâneas, além de apresentação, referência metodológica e glossário, constituindo-se, dessa forma, num importante instrumento para gestores e pesquisadores em educação, com o intuito de tratar acerca da produção de conhecimento sobre a EPT no Brasil.

Figura 3 - Estrutura geral dos painéis da PNP

Fonte: Moraes et al, (2018, p. 28).

A padronização dos conceitos e formas de cálculos são importantes para orientar os gestores e setores da instituição responsáveis pela coleta e organização dos dados, os quais darão origem aos indicadores e análise posterior de cada um deles. Ter uma plataforma com dados padronizados e organizados de forma sistemática favorece as pesquisas, o acompanhamento e avaliações pelas próprias instituições, mas é preciso lembrar que o aumento da qualidade e a diminuição de problemas como a evasão envolvem outras dimensões além da constatação dos números. A organização do Plano Estratégico de Intervenção e Monitoramento para Superação da Evasão, é um conjunto de procedimentos que poderá se transformar num instrumento que desencadeará uma ação mais eficaz para redução dos problemas causados pela evasão. Esta expectativa só se concretizará, na medida em que as ações alcançarem o pleno envolvimento de cada instituição, no levantamento não só dos números estatísticos, mas das causas do abandono relacionadas aos fatores individuais, internos e externos. Para tanto, será necessário o engajamento de todos os seus segmentos.

4 A EVASÃO NO CONTEXTO DO IFRS

Instituído pela Lei nº 11.892, no dia 29 de dezembro de 2008, o Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul (IFRS) é uma instituição de educação superior, básica e profissional, pluricurricular e multicampi, especializada na oferta de educação profissional e tecnológica nas diferentes modalidades de ensino. Criado pela união do Centro Federal de Educação Tecnológica de Bento Gonçalves, da Escola Técnica Federal de Canoas e da Escola Agrotécnica Federal de Sertão, é uma autarquia, detentora de autonomia administrativa, patrimonial, financeira, didático-pedagógica e disciplinar.

Completando dez anos de atuação na educação profissional e tecnológica, com sua reitoria situada na cidade de Bento Gonçalves, hoje possui dezessete *campi*, conta com cerca de 20 mil alunos em mais de 200 opções de cursos técnicos e superiores de diferentes modalidades e PROEJA. Oferece, também, cursos de pós-graduação (especialização e mestrado) e cursos de formação inicial e continuada (cursos rápidos). Tem aproximadamente 1.020 professores e 950 técnicos administrativos (IFRS, 2018b).

De acordo com seu Estatuto (IFRS, 2016, p. 2 e 3), o IFRS observa em sua atuação os seguintes princípios norteadores:

I. compromisso com a justiça social, equidade, cidadania, ética, preservação do meio ambiente, transparência e gestão democrática;
II. verticalização do ensino e sua integração com a pesquisa e a extensão;
III. eficácia nas respostas de formação profissional, difusão do conhecimento científico e tecnológico e suporte aos arranjos produtivos locais, sociais e culturais;
IV. inclusão de pessoas com necessidades educacionais especiais e deficiências específicas;
V. natureza pública e gratuita do ensino, sob a responsabilidade da União;
VI. inclusão social de pessoas afrodescendentes, indígenas e em situação de vulnerabilidade social.

Seus princípios formam o lastro para sua atuação em diferentes regiões do Estado do Rio Grande do Sul e para o cumprimento de sua missão:

> Promover a educação profissional, científica e tecnológica, gratuita e de excelência, em todos os níveis e modalidades, através da articulação entre ensino, pesquisa e extensão, em consonância com as demandas dos arranjos produtivos locais, formando cidadãos capazes de impulsionar o desenvolvimento sustentável. (IFRS, 2014, p. 18)

Preocupado com o desenvolvimento de sua ação acadêmica e a manutenção da qualidade da oferta do ensino, com vistas a cumprir sua missão e seu compromisso social, o IFRS tem como um dos seus objetivos:

> Estimular, por meio da criação de políticas, a ampliação continuada das condições de permanência dos estudantes no IFRS, considerando a necessidade de viabilizar a igualdade de oportunidades, contribuir para a melhoria do desempenho acadêmico e agir, preventivamente, nas situações de retenção e evasão. (IFRS, 2016, p. 4).

1.1.1 AS ESTRATÉGIAS DO IFRS PARA PERMANÊNCIA E ÊXITO DOS ESTUDANTES E O COMBATE À EVASÃO.

Com base na análise dos ambientes internos e externos da instituição, empreendida para construção do Plano de Desenvolvimento Institucional – compreendido entre 2014 e 2018, foram estabelecidos cinco objetivos estratégicos e metas da área de ensino do IFRS. Os objetivos número um e cinco estão diretamente ligados à redução da evasão e à permanência do estudante no IFRS. (IFRS, 2014).

Quadro 2 - Objetivos e metas da área de ensino do IFRS relacionados à evasão

Objetivos	Metas
1. Criar observatório da evasão e retenção discente no IFRS	1. construir e consolidar o instrumento para levantamento dos dados de evasão e retenção no IFRS;
	2. aplicar, junto aos *campi,* o instrumento de estudo dos números de evasão e retenção no IFRS e iniciar a análise dos dados;
	3. realizar seminário bianual de análise dos dados e planejamento de ações para o combate à evasão e retenção;
	4. acompanhar e avaliar, junto aos *campi,* as ações de superação dos índices de evasão e retenção identificados a partir da análise de dados e do seminário bianual.
5. Consolidar a Política de Assistência Estudantil do IFRS.	1. implementar, junto aos *campi* do IFRS, ações de permanência e êxito dos estudantes em consonância com a Política de Assistência Estudantil do IFRS e com a legislação vigente;
	2. fomentar as ações de inclusão de forma articulada com os Napnes e Neabis dos *campi* do IFRS.

Fonte: adaptado da Pró-Reitoria de Ensino e Comitê de Ensino (IFRS, 2014).

Em relação ao objetivo nº 5, da área de Ensino do PDI (2014-2018), o IFRS já havia aprovado a sua Política de Assistência Estudantil (PAE) em dezembro de 2013, contendo as ações para promover o acesso, a permanência e o êxito dos estudantes, em consonância com o Programa Nacional de Assistência Estudantil (PNAE), com o Projeto Pedagógico Institucional (PPI) e com o Plano de Desenvolvimento Institucional, os quais começavam a ser postos em prática, juntamente com a organização da estrutura e do modo de funcionamento dos diferentes órgãos da Assistência Estudantil (IFRS, 2013, p. 1).

De acordo com a proposta da Política de Assistência Estudantil do IFRS (IFRS, 2013, p. 6), a estrutura da Assistência Estudantil (AE) deve ser a seguinte:

I. **Assessoria de Assistência Estudantil (AAE)**: com estrutura mínima de dois servidores, dentre os quais Assistentes Sociais, Pedagogos, Psicólogos e Técnicos em Assuntos Educacionais, possui a função de planejar, implementar e acompanhar a Política de Assistência Estudantil do IFRS, em conjunto com os demais órgãos da AE;
II. **Grupo de Trabalho Permanente em Assistência Estudantil do IFRS (GTPAE)**: é um órgão colegiado propositivo, consultivo, atuando na área da

Assistência Estudantil e que auxilia a Assessoria de Assistência Estudantil na implementação, regulação, planejamento, acompanhamento e avaliação da PAE do IFRS, seus programas, projetos e ações. É composto pela AAE e pelos coordenadores das Coordenações de Assistência Estudantil de cada *campi;*

III. **Coordenações de Assistência Estudantil (CAE)**: são compostas por um coordenador, indicado pela direção do *campus*, e uma equipe técnica mínima formada por um(a) pedagogo(a), um(a) psicólogo(a) e um(a) assistente social. É o órgão que, subordinado às suas Direções-Gerais e de Ensino, possui em seu âmbito a função de planejar, executar e acompanhar a PAE, trabalhando de forma integrada às demais coordenações e setores do seu *campus*, e à Comissão de Assistência Estudantil local;

IV. **Comissões de Assistência Estudantil**: são órgãos dos *campi* que possuem em seu âmbito a função de apoiar as CAE no planejamento, execução e acompanhamento da Política de Assistência Estudantil. Regulamentadas por Regimento Interno Próprio, serão compostas pelo Coordenador da AE, por 2 (dois) servidores docentes e 2 (dois) servidores do segmento técnico-administrativos e por 2 (dois) discentes, eleitos entre seus pares.

Quanto ao público atendido, o texto da Política de Assistência Estudantil IFRS (IFRS, 2013, p. 3), em seu Art. 4º afirma:

> A Assistência Estudantil do Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul possui um amplo escopo de atenção, oferecendo condições para a melhoria do desempenho acadêmico dos estudantes e agindo, preventivamente, nas situações de retenção e evasão, incluindo, desde Ações de Caráter Universal, até Programas de Benefícios, atingindo, desse modo, diferentes públicos dentro da comunidade escolar.

Isso significa que as CAE propiciam a todos os estudantes, sem distinção, o trabalho de apoio e acompanhamento, oferecidos pelas equipes multiprofissionais diariamente, e, também, buscam propiciar a igualdade de oportunidades e a melhoria das condições socioeconômicas daqueles que necessitam, através dos programas de benefícios. Quanto a estes, o Art. 23 deixa claro que “têm por finalidade subsidiar as despesas dos estudantes beneficiados [...] com vistas a ampliar suas condições de permanência e êxito acadêmico, bem como reduzir os índices de retenção e evasão escolar no IFRS”. A Assistência Estudantil de cada *campus* é responsável pela organização do processo de concessão, através de edital (IFRS, 2013, p. 12).

A organização da Assistência através da PAE juntamente com o empenho da instituição no cumprimento das metas do seu PDI, desencadearam uma

visibilidade maior das ações das CAEs, e da AE como um todo. Este fato contribuiu para maior apoio e engajamento pelos outros setores da instituição, bem como para facilitar o acesso dos alunos que buscam apoio pedagógico, psicológico e/ou benefícios. Uma série de reuniões do GT (Grupo de Trabalho) e encontros das equipes das CAE tiveram como pauta a preocupação com a vulnerabilidade social e econômica dos estudantes e com os prejuízos causados por estas no desenvolvimento dos alunos. Com destaque para: a permanência e êxito dos estudantes; a organização conjunta de editais para a concessão dos benefícios, de forma a facilitar a solicitação dos alunos e a distribuição justa dos recursos; o envolvimento da AE na política de ingresso e no próprio processo; a necessidade de criar um instrumento para conhecer o perfil sociodemográfico dos alunos, para melhor acompanhá-los na trajetória acadêmica; o problema da evasão e da retenção dos alunos e as formas de combatê-los.

O objetivo nº 1 da área de Ensino, bem como suas metas, estabelecidas no PDI (2014-2018), vem sendo desenvolvidas através do Plano Estratégico de Permanência e Êxito dos Estudantes do IFRS (PEPEEIFRS). Em 2014, em resposta à solicitação da SETEC/MEC, através do Documento Orientador para Superação da Evasão e Retenção da Rede Federal, foi criado um grupo de trabalho para elaboração do Plano Estratégico. A partir de novas orientações em 2015, o foco das ações passou a ser a permanência e o êxito dos estudantes. Com esse intuito, foi criada a Comissão Interna de Acompanhamento de Ações de Permanência e Êxito dos Estudantes (CIAAPE) e as subcomissões nos *campi*, com a missão de elaborar em conjunto o PEPEEIFRS.

O Plano Estratégico de Permanência e Êxito dos Estudantes (PEPEE) foi aprovado pelo Conselho Superior do IFRS (CONSUP) em outubro de 2018, e é o resultado de um processo coletivo, cujo objetivo é propor medidas para superar a evasão e a retenção/reprovação dos estudantes. Para tanto, o documento descreve em seis capítulos a contextualização do IFRS, apresentando sua caracterização, organização e oferta educacional; expõe as bases conceituais relativas à permanência e ao êxito; detalha os aspectos metodológicos para elaboração e implementação do Plano Estratégico; apresenta o diagnóstico de indicadores quantitativos e qualitativos e, por fim, sistematiza as metas e as ações previstas,

com estratégias de monitoramento, de modo a garantir a efetividade das mesmas, fechando com algumas considerações. (IFRS, 2018a).

A metodologia para elaboração do Plano Estratégico (IFRS, 2018a, p. 17 a 19) obedeceu às seguintes fases, ficando de acordo com o Documento Orientador para a Superação da Evasão e Retenção na Rede Federal de Educação Profissional:

Fase 1: formação de uma comissão e subcomissões de acompanhamento de ações de permanência e êxito dos estudantes – CIAAPE;

Fase 2: construção dos indicadores quantitativos;

Fase 3: diagnóstico qualitativo dos fatores de evasão/reprovação;

Fase 4: elaboração do plano estratégico institucional;

Fase 5: elaboração dos Planos Estratégicos dos *campi*.

Para realizar a análise quantitativa sobre evasão, retenção, permanência e êxito dos estudantes os dados foram extraídos do SISTEC e da PNP, realizados os cálculos dos indicadores Taxa de Conclusão; Taxa de Evasão; Taxa de Retenção; Taxa de Matrícula Continuada Regular; Taxa de Matrícula Continuada Retida; Índice de Permanência e Êxito; Taxa de Efetividade Acadêmica; Índice de Eficácia; e Índice de Eficiência Acadêmica, de acordo os conceitos estabelecidos pelas Portarias nº 818/2015-MEC e 25/2015-SETEC, pelo termo de acordo e metas MEC/SETEC-IFRS e pelo Acórdão nº 2.267/2005-TCU.

Quanto à análise qualitativa, a CIAAPE realizou um questionário no Google Drive, sobre as causas da evasão, o qual foi respondido pelos Estudantes em Curso, Estudantes Evadidos e Servidores (técnicos e docentes). As causas elencadas pelos segmentos foram organizadas de acordo com os três fatores categorizados pelo Documento Orientador para a Superação da Evasão e Retenção na Rede Federal de Educação Profissional Científica e Tecnológica (BRASIL, 2014) da seguinte forma:

Quadro 3 - Causas de evasão e retenção apontadas no questionário online

Fatores individuais dos estudantes	**Fatores internos à instituição**	**Fatores externos à instituição**
- Problemas financeiros; - Dificuldades no acompanhamento dos conteúdos; - Reprovação em semestres/anos anteriores; - Busca pela conclusão dos estudos em outra Instituição após reprovação; - Incompatibilidade de horário entre trabalho e estudos (cansaço, muitos acabam optando pelo trabalho que lhes garante sobrevivência); - Curso não atende às expectativas; - Perfil diferente do curso escolhido; - Problemas familiares; - Recolocação profissional; - Falta de tempo para estudar fora da instituição; - Distância entre casa/instituição; - Relacionamento com os colegas; - Perda dos prazos referentes aos fluxos durante período de matrículas; - Migração de curso.	- Nível de exigência muito rigoroso; - Inserção do curso no mercado de trabalho local; - Relacionamento com alguns professores.	- Falta de transporte para se deslocar até a instituição.

Fonte: Plano Estratégico de Permanência e Êxito dos Estudantes do IFRS. (IFRS, 2018a, p. 48).

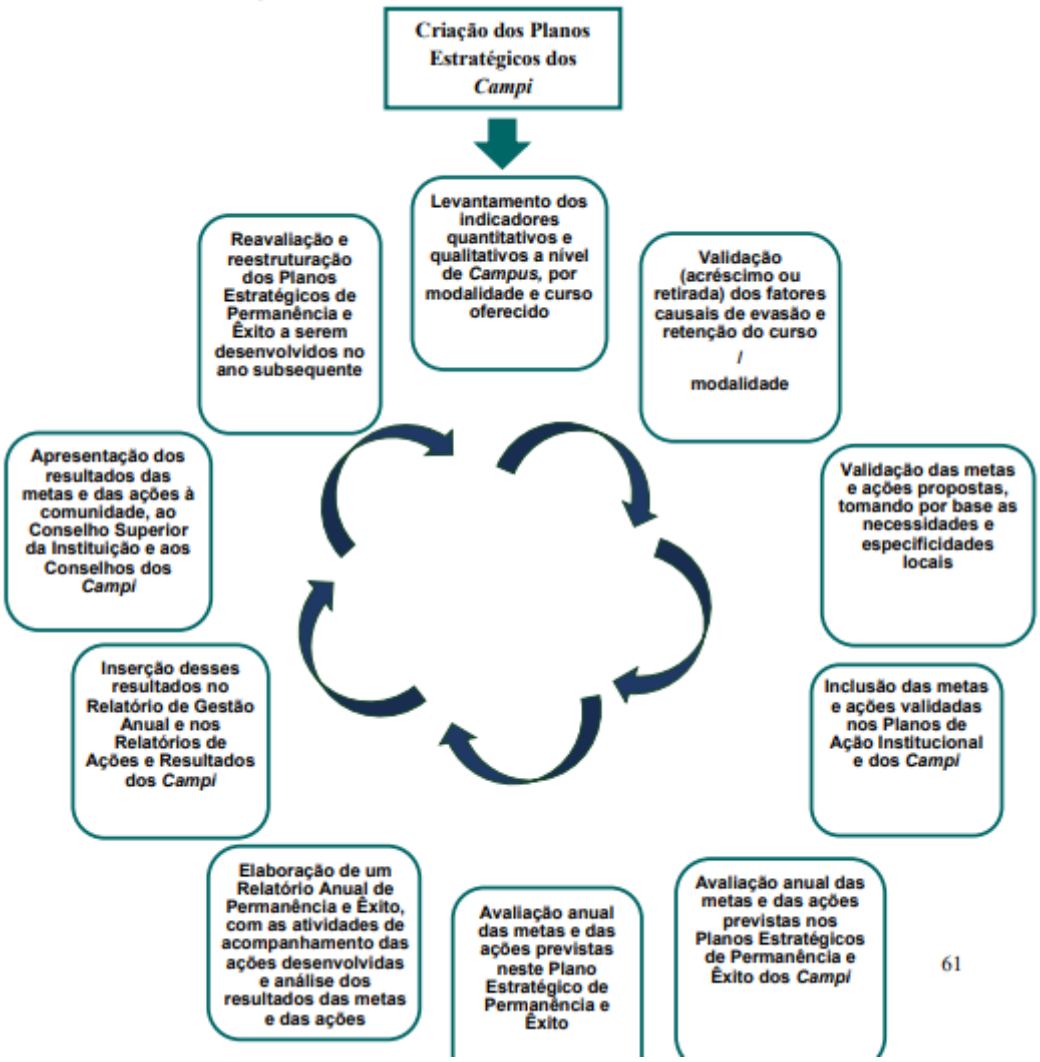
Após a análise dos indicadores, a CIAAPE (IFRS, 2018a, p. 64 e 65) estabeleceu metas quantitativas e qualitativas para permanência e o êxito dos estudantes do IFRS em seu processo de formação integral, as quais seguem descritas abaixo:

1. QUANTITATIVAS: as mesmas explicitadas no Termo de Acordo de Metas e Compromissos firmado com o MEC, em 2010, pois não foram atingidas, ou seja, índices de Eficiência da Instituição de 90% e de Eficácia da Instituição de 80%;
2. QUALITATIVAS: foram definidas de acordo com as principais causas elencadas através do questionário. Para cada meta foram propostas ações, com prazos determinados e o agente responsável por sua execução. São elas:
   2.1 construção de Programa Institucional de Formação Continuada para Servidores (Docentes e Técnicos Administrativos);

2.2 acompanhamento dos estudantes com equipe multidisciplinar e multiprofissional;
2.3 instituição de diretrizes de comunicação e eventos;
2.4 monitoramento, avaliação e acompanhamento dos cursos do IFRS;
2.5 desenvolvimento de programas institucionais com ações coordenadas pelas Pró-reitorias, com vistas à consolidação da identidade institucional e do sentimento de pertencimento dos atores sociais envolvidos;
2.6 articulação institucional interna e externa, objetivando minimizar fatores estruturais que contribuem com a evasão;
2.7 criação do Observatório de Permanência e Êxito de estudantes no IFRS.

A Figura 4 esquematiza o ciclo anual de monitoramento dos indicadores, das metas e das ações contidas Planos Estratégicos de Permanência e Êxito dos *campi*.

Figura 4 - Ciclo anual de atividades de monitoramento e avaliação dos Planos Estratégicos de Permanência e Êxito dos campi



Fonte: Plano Estratégico de Permanência e Êxito do IFRS. (IFRS, 2018a, p. 61).

A soma desses dois movimentos, a organização da Política de Assistência Estudantil e o Plano Estratégico dos sujeitos do IFRS, tem como intuito apoiar os estudantes nas suas trajetórias escolares e acadêmicas, melhorando suas condições de permanência e êxito. Desta forma, a instituição trabalha preventivamente para minimizar a evasão e a retenção e, com isso, os prejuízos pessoais, institucionais e sociais causados por esses problemas:

> Espera-se, portanto, num curto prazo, apresentando também perspectivas de médio e longo prazo, que a instituição – a partir dos seus diversos sujeitos ativos – empreenda esforços para promover ainda mais a inclusão social por meio não apenas da democratização do acesso, mas da oferta de condições de permanência e êxito dos estudantes. (IFRS, 2018a, p. 8).

No próximo capítulo, encontram-se descritas algumas características do *campus* Canoas do IFRS no qual está sendo desenvolvida a pesquisa. É abordado o problema da evasão neste *campus*, de maneira especial no que diz respeito ao ensino superior, e o que está sendo feito até o momento para minimizá-lo.

4.1 O *CAMPUS* CANOAS DO IFRS

O *Campus* Canoas do IFRS foi criado pela Lei nº 11.534, de 25 de outubro de 2007, como Escola Técnica Federal de Canoas (ETFC). A Portaria nº 1.068, de 13 de novembro de 2007, atribuiu ao Centro Federal de Educação Tecnológica de Pelotas (CEFET-RS) o encargo de adotar as medidas necessárias à implantação da Escola Técnica Federal de Canoas. Posteriormente, em 18 de abril de 2008, a Portaria nº 488 transferiu essa tarefa ao Centro Federal de Educação Tecnológica de Bento Gonçalves (CEFET-BG).

Com a reorganização da Rede Federal de Educação Profissional e Tecnológica, através da Lei nº 11.892, a ETFC passou a compor o IFRS, juntamente ao Centro Federal de Educação Tecnológica de Bento Gonçalves, a Escola Técnica da UFRGS, o Colégio Técnico Industrial Prof. Mário Alquati, de Rio Grande e a Escola Agrotécnica Federal de Sertão. Dessa forma, o *Campus* Canoas,o *Campus* Bento Gonçalves, o *Campus* Porto Alegre, o *Campus* Rio Grande e o *Campus* Sertão foram as primeiras unidades do IFRS.

Nesse processo de implantação foram realizadas audiências públicas com a participação de diversos setores da comunidade canoense, com o objetivo de identificar as necessidades de qualificação, requalificação ou reconversão profissional dos trabalhadores empregados ou desempregados, para atuar em sintonia com os arranjos produtivos locais.

Em 27 de agosto de 2010 iniciaram as aulas no *Campus* Canoas com uma turma do Curso Técnico em Manutenção e Suporte em Informática Integrado ao Ensino Médio PROEJA, duas turmas do Curso Técnico de Nível Médio Subsequente em Eletrônica e duas turmas do Curso Técnico de Nível Médio Subsequente em Informática. No primeiro semestre de 2011 começaram as primeiras turmas dos cursos técnicos integrados ao ensino médio e dos superiores de tecnologia.

4.1.1 Os cursos do *Campus* Canoas

A partir de 2014 e estendendo até hoje, o *Campus* Canoas possui um curso na modalidade de jovens e adultos (PROEJA), três cursos técnicos integrados ao nível médio e quatro cursos superiores, sendo três deles de tecnologia e um de licenciatura.

O **Curso Técnico em Manutenção e Suporte em Informática Integrado ao Ensino Médio - modalidade PROEJA -** Eixo Tecnológico de Informação e Comunicação - autorizado pela Resolução nº 075/2010, do Conselho Superior do IFRS e alterado pela Resolução nº 067/2011 - CS-IFRS, passou a ser ministrado no turno da noite a partir de 2011 e possui 73 alunos matriculados. O curso de PROEJA desde seu início possui um processo de ingresso diferenciado dos demais cursos, sem prova, em que os critérios levam em consideração a idade do candidato, o tempo de afastamento da escola e a renda, com o objetivo de promover a inclusão e oferecer educação profissional e elevação da escolaridade para jovens e adultos que não concluíram os estudos no tempo dito “regular”. Em 2019, o eixo tecnológico será alterado para a área de Comércio, com o intuito de atender às demandas locais.

O **Curso Técnico em Informática Integrado ao Ensino Médio** - Eixo Tecnológico de Informação e Comunicação – autorizado pela Resolução nº 156/10

- CS-IFRS, e alterado pela Resolução nº 009/13 – Conselho de *Campus* (CONCAMP) do *Campus* Canoas do IFRS. Este foi extinto no final de 2015, em função de adequação no nome do curso ao Catálogo Nacional de Cursos Técnicos do Ministério da Educação e Cultura (MEC). Eram ofertadas 30 vagas para ingresso anual, com seleção através de prova. Em 2018, os últimos 50 alunos concluíram o quarto ano e o curso.

O **Curso Técnico em Administração Integrado ao Ensino Médio** – autorizado pela Resolução nº 155/10 - CONSUP-IFRS, e alterado pela Resolução nº 008/13- CONCAMP-IFRS Canoas. Esse curso começou em 2011, oferta 30 vagas para o ingresso anual com seleção através de prova. Em 2019, possui 119 alunos cursando regularmente.

O **Curso Técnico em Eletrônica Integrado ao Ensino Médio** – autorizado pela Resolução nº 10/2013 - CONCAMP-IFRS Canoas. O curso de eletrônica começou em 2014 e oferta 24 vagas para ingresso anual, com seleção através de prova. Em 2019, possui 82 alunos matriculados.

O **Curso Técnico em Desenvolvimento de Sistemas Integrado ao Ensino Médio** - Eixo Tecnológico de Informação e Comunicação - aprovado pela Resolução nº 07/16 - CONCAMP-IFRS Canoas. Este curso começou em 2016, substituindo o Técnico em Informática Integrado e oferta 30 vagas para ingresso anual, com seleção através de prova. Em 2019, 104 alunos estão matriculados.

O **Curso Superior de Tecnologia em Automação Industrial -** Eixo Tecnológico Controle e Processos Industriais – reconhecido pela Portaria N° 651, de 10 de dezembro de 2013 - DOU nº 240 de 11/12/2013, renovado pela Portaria nº 1343 de 15 de dezembro de 2017, DOU nº 241 de 18/12/2017. Este curso tem duração de três anos e meio, divididos em 7 semestres. Começou em 2011 e oferta quinze vagas através de processo seletivo próprio e quinze vagas através da nota da prova do ENEM, no turno da noite.

O **Curso Superior de Tecnologia em Logística** - Eixo Tecnológico Gestão e Negócios – reconhecido pela Portaria n°408, de 30 de agosto de 2013 - DOU nº 169 de 02/09/2013, renovado pela Portaria nº 432, de 15 de maio de 2017, DOU nº 93 de 17/05/2017. Este curso tem duração de três anos, divididos em 6 semestres.

Começou em 2011 e oferta dezoito vagas através de processo seletivo próprio e dezoito vagas através da nota da prova do ENEM, no turno da noite.

O **Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas** - Eixo Tecnológico de Informação e Comunicação – reconhecido Portaria nº 876, de 12 de novembro de 2015, DOU nº 217, de 13/11/2015. Este curso tem duração de três anos, divididos em seis semestres. Começou no segundo semestre de 2012 e oferta quinze vagas através de processo seletivo próprio e quinze vagas através da nota da prova do ENEM, no turno da manhã.

O **Curso Superior de Licenciatura em Matemática** - reconhecido pela portaria nº 972, de 6 de setembro de 2017, DOU nº 173, de 8 de setembro de 2017. Este curso tem duração de quatro anos, divididos em oito semestres. Começou em 2013 e oferta vinte vagas através de processo seletivo próprio e vinte vagas através da nota da prova do ENEM, no turno da manhã.

Os cursos superiores possuem oferta anual, com exceção do Curso de Tecnologia em Análise e Desenvolvimento de Sistemas que, a partir de 2018, começou a ofertar ingresso semestral. Em 2019, o Curso de Logística também passará a ofertar vagas semestralmente. A cada oferta, um número fixo de vagas é disponibilizado de acordo com o Projeto Pedagógico (PPC) de cada curso. Até 2017 metade das vagas para ingresso, nos cursos superiores, era ofertada pelo Sisu, sendo substituído pela nota do Enem, não sendo mais necessária a inscrição no cadastro para participar da seleção.

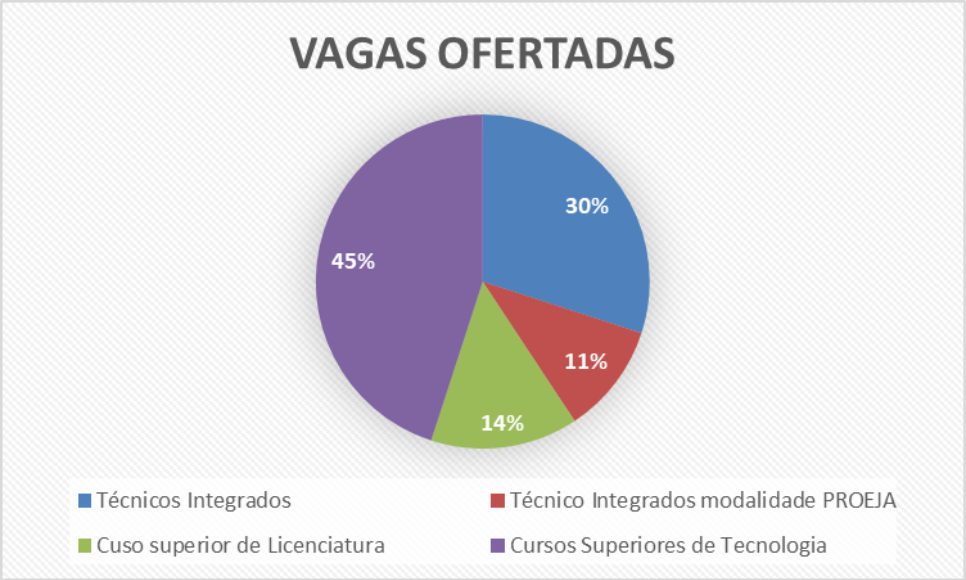
Quadro 4 - Número de vagas ofertadas por curso, do *Campus* Canoas - IFRS

CURSOS	VAGAS OFERTADAS 2017	VAGAS OFERTADAS 2018	VAGAS OFERTADAS 2019
Técnico em Administração Integrado ao Ensino Médio	30	30	30
Técnico em Eletrônica Integrado ao Ensino Médio	24	24	24
Técnico em Desenvolvimento de Sistemas Integrado ao Ensino Médio	30	30	30
Técnico em Manutenção e Suporte em Informática- PROEJA	30	30	0
Curso Técnico em Comércio Integrado à Modalidade de Educação de Jovens e Adultos (PROEJA)	0	0	30
Licenciatura em Matemática	40	40	40
Superior de Tecnologia em Análise e Desenvolvimento de Sistemas	30	60	60
Superior de Tecnologia em Automação Industrial	30	30	30
Superior de Tecnologia em Logística	36	36	72

Fonte: construção da autora, a partir dos dados do SRE do IFRS - *Campus* Canoas.

O Gráfico 1 ilustra a proporção de vagas em cada uma das modalidades de ensino, de acordo com o número ofertado em 2018, sendo que é possível observar que o maior percentual de oferta de vagas ocorre nos cursos superiores de tecnologia, foco do presente estudo.

Gráfico 1 - Percentual de vagas ofertadas, por modalidade de ensino, no ano de 2018, no IFRS-Canoas.



Fonte: criado pela a autora, a partir dos dados do Setor de Registros Escolares.

4.1.2 A Evasão no *Campus* Canoas

O *Campus* Canoas está localizado na região metropolitana da capital do Estado, atendendo aos arranjos produtivos da cidade e da região, através de seus cursos distribuídos nos eixos tecnológicos de Controle e Processos Industriais, Informação e Comunicação e Gestão e Negócios Seus cursos são bastante procurados por jovens em busca de formação técnica e adultos trabalhadores, empregados e desempregados, em busca de formação ou qualificação para melhor se inserir no mundo do trabalho.

Esses dois fatores - atender aos arranjos produtivos locais e ter grande procura por seus cursos - possibilitam ao *Campus* Canoas ser uma instituição fomentadora do progresso tecnológico, cultural, econômico e social da região, através do ensino, da pesquisa e da extensão, fazendo com que se cumpra seu papel de instituição pública e gratuita de excelência. Porém, o fantasma da evasão, como problema a ser minimizado, também preocupa e, assim como o restante do IFRS, o *campus* está empenhado em buscar alternativas para combatê-lo e fazer com que seus estudantes sejam exitosos em sua formação.

De 2011 a 2017, período que selecionamos para análise nesta pesquisa, ingressaram no *Campus* Canoas 1641 (um mil seiscentos e quarenta e um) alunos, distribuídos nas modalidades de curso, conforme tabela abaixo.

Quadro 5 Nº de ingressantes, por modalidade de curso, de 2011 a 2017.

Modalidade de curso	**Total de alunos que ingressaram no período entre 2011 e 2017**
Técnicos integrados ao ensino médio	518
Técnico integrado PROEJA	196
Superiores de tecnologia	765
Superior de Licenciatura em Matemática	162

Fonte: criado pela autora, a partir dos dados do SRE do IFRS - *Campus* Canoas.

Ao final de 2017, levando-se em consideração o total de ingressantes, 261 haviam concluído seus cursos; 54 pediram transferência para outra instituição; 190 cancelaram a matrícula e 420 abandonaram a instituição, revelando uma situação preocupante de 40,46 % (quarenta vírgula quarenta e seis por cento) saídas sem êxito, ou seja, sem a conclusão do curso. O Quadro 6 apresenta, em percentuais, essas situações por modalidade de ensino.

Quadro 6 - Percentual de concluintes, transferidos, desligados e evadidos por modalidade de ensino, entre 2011 e 2017, no Campus Canoas.

Modalidades	Total de Ingressantes	Saídas com Êxito	Saídas sem Êxito		
		% de Concluintes	% de Transferidos	% de Desligados	% de Evadidos
Técnicos integrados	518	32,05	8,11	6,76	6,37
Técnico Integrado PROEJA	196	18,37	1,02	6,63	52,55
Superior de tecnologia	765	7,71	1,18	13,59	32,55
Licenciatura em Matemática	162	0,62	0,62	23,46	21,60

Fonte: criado pela autora, a partir dos dados da planilha “Dados de Matrícula por ano letivo” do SRE do IFRS – Canoas.

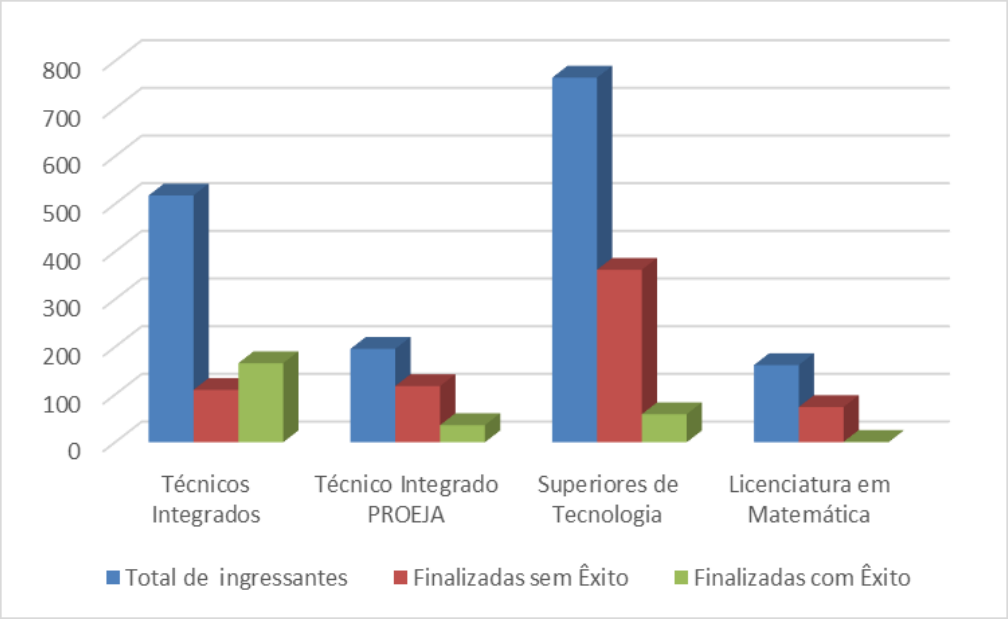
Ao analisar o Quadro 6, é possível perceber que a modalidade com o percentual maior de evasão é o PROEJA, mas também é a segunda que mais formou alunos em relação ao número de ingressantes. A segunda modalidade com o maior percentual de evadidos é representada pelos cursos superiores de

tecnologia, com percentual pouco expressivo de formandos em relação ao número de ingressantes. A modalidade com percentual maior de concluintes congrega os cursos técnicos integrados, que também apresentam um menor número de evadidos. O Curso Superior de Licenciatura, que começou em 2014, teve apenas uma concluinte em 2017. Cabe lembrar que o conceito de “evadido” usado na tabela é a situação na qual o(a) aluno(a) abandona o curso, sem formalizar sua saída à instituição e “desligados” são os alunos que solicitaram formalmente o cancelamento de sua matrícula. Os conceitos e os cálculos aqui empregados foram feitos com base no manual para cálculo dos indicadores de gestão das Instituições da RFEPCT (BRASIL, 2016a), comentado anteriormente.

Cabe esclarecer como e quando a ausência do aluno no curso e na instituição é considerada como abandono e, consequentemente, evasão. De acordo com o Art. 120 da Organização Didática (OD) do IFRS, aprovada pela Resolução nº 046, de 08 de maio de 2015, e alterada pela Resolução nº 086, de 17 de outubro de 2017 - ambas do Conselho Superior (CONSUP) da instituição - considera-se evasão “quando o estudante não tiver renovado a matrícula por dois períodos letivos consecutivos, caracterizando o abandono de curso.” E esclarece no parágrafo único: “o status do estudante até o término do período descrito no caput do artigo será registrado como trancamento automático” (IFRS, 2017).

Dessa forma, o aluno com status “trancado” e em “trancamento automático” ou ainda, de acordo com a nomenclatura do S.I.A. -sistema acadêmico utilizado pelo IFRS e pelo *Campus* Canoas - em “trancamento total”, não compõem o cálculo da evasão, pois, de acordo com o Art. 117, o aluno em trancamento possui a matrícula ativa: “entende-se por trancamento da matrícula o ato formal pelo qual se dá a interrupção temporária dos estudos, sem a perda do vínculo do estudante com a instituição.” O aluno precisa renovar seu trancamento a cada semestre por, no máximo, 50% do tempo do curso, considerando períodos letivos consecutivos ou não. Caso não o faça, terá a matrícula cancelada (IFRS, 2017). No Gráfico 2 é possível visualizar a diferença do percentual de saídas com e sem êxito para cada modalidade.

Gráfico 2 - Saídas com e sem êxito para cada modalidade de ensino



Fonte: criado pela autora, a partir das informações da planilha “Dados de Matrícula por ano letivo” do SR do IFRS - Canoas.

Os conceitos de saída com êxito e sem êxito são utilizados pelo SISTEC, já apresentado anteriormente, e “Saída sem êxito” é o mesmo que “Matrículas finalizadas evadidas” empregado pelo Manual da Rede Federal (BRASIL, 2014).

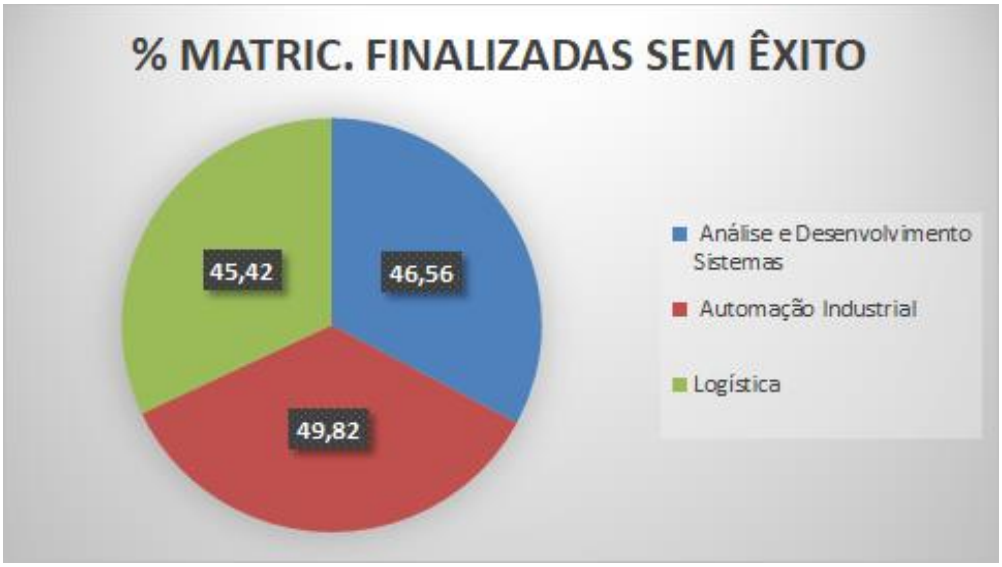
Fica evidente, ao se observar o gráfico, que os cursos de tecnologia têm o maior número de ingressantes no período entre 2011 e 2017, porém são também os em que grande parte dos estudantes não atingem seu propósito de concluir um curso superior. É mais preocupante, ainda, quando se observa que estes são cursos de tecnologia, procurados por trabalhadores que desejam a inserção no mercado de trabalho ou uma melhor posição por meio de melhor qualificação. Além disso, são cursos que visam à formação de um profissional com postura crítica, ativa e consciente do seu papel social e profissional e que contribua para o avanço científico e tecnológico do país. A saída do curso sem concluí-lo representa prejuízo tanto para a instituição que mobiliza recursos financeiros, espaço físico, professores e técnicos qualificados na formação dos alunos, quanto para o aluno(a), que provavelmente terá mais dificuldade de colocação no mercado de trabalho, cada vez mais competitivo, além de carregar a sensação de fracasso. A sociedade como um todo também sente os reflexos, pois carece de cidadãos atuantes e profissionais qualificados em diversas áreas.

Os dados quantitativos sobre a evasão no *Campus* Canoas justificam a necessidade de buscar alternativas para minimizar esse problema, bem como explicam a opção por esta temática para a realização este trabalho, que busca identificar os(as) alunos(as) com propensão à evasão, por intermédio do processo de KDD e, em especial, nos cursos de tecnologia.

4.1.3 A evasão nos cursos de tecnologia

Em uma análise mais apurada dos dados quantitativos dos cursos de tecnologia, pode-se verificar que o percentual de matrículas finalizadas sem êxito, em relação ao número de ingressantes em cada curso, situa-se numa faixa entre 45% e 50%, ou seja, não há uma discrepância muito grande entre os cursos.

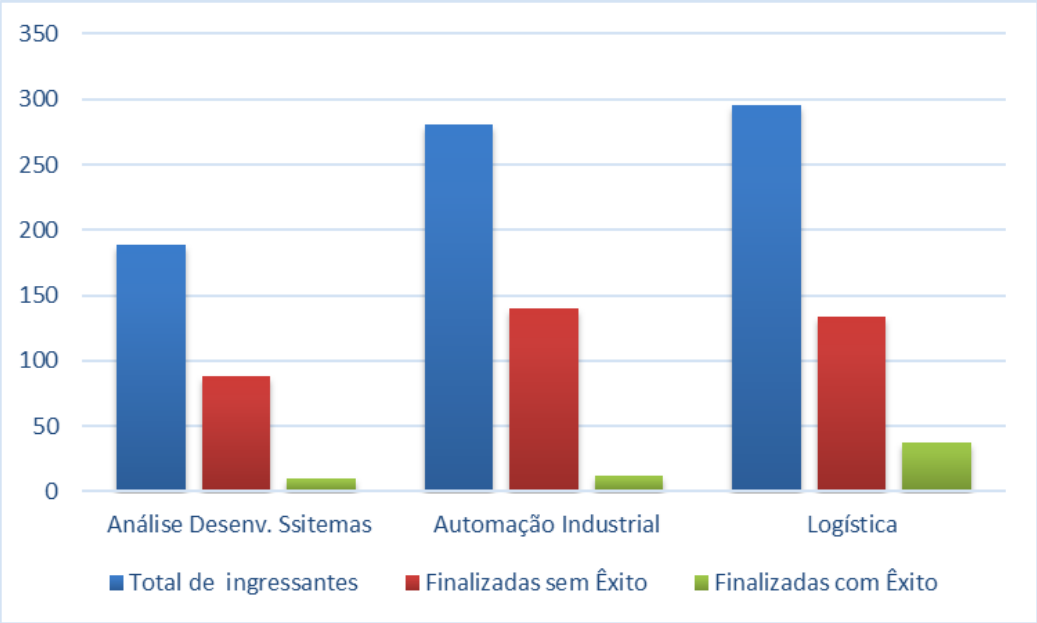
Gráfico 3 - Percentual de evasão em cada curso superior de tecnologia



Fonte: criado pela autora, a partir das informações da planilha “Dados de Matrícula por ano letivo”, do SRE do IFRS – Canoas.

Os dados mostram que a metade dos alunos que ingressam nos cursos de tecnologia saíram sem concluí-los e que o total daqueles que conquistaram o diploma é baixo.

Gráfico 4 - Relação entre ingressantes, matrículas finalizadas sem êxito e com êxito



Fonte: elaborado pela autora, a partir das informações da planilha “Dados de Matrícula por ano letivo” do SRE do IFRS – Canoas.

Esse percentual de abandono em cada curso evidencia a necessidade de um trabalho coletivo da direção, das coordenações dos cursos e das equipes pedagógicas e de Assistência Estudantil, para identificar dificuldades e problemas dos alunos, sejam eles individuais, relacionados à instituição ou de fora dela. É preciso ações integradas que ofereçam apoio e alternativas aos estudantes, para que estes alcem o objetivo da formação profissional de qualidade.

No Quadro 7, tem-se o detalhamento das formas de saídas sem êxito e os respectivos percentuais em cada curso. Novamente, percebe-se que os números são similares entre os cursos, mas é possível destacar que o Curso de Tecnologia em Logística tem formado mais alunos.

Quadro 7 - Percentuais de saídas dos cursos superiores de tecnologia

Modalidades	Total de ingressantes	Saídas com êxito	Saídas sem êxito		
		% de Concluintes	% de Transferidos	% de Desligados	% de Evadidos
Análise Desenvol. Sistemas	189	5,29	1,06	12,70	32,80
Automação Industrial	281	4,27	0,71	16,37	32,74
Logística	295	12,54	1,69	11,53	32,20

Fonte: construção da autora, a partir das informações da planilha “Dados de Matrícula por ano letivo” do SRE do IFRS – Canoas.

No Quadro 7, pode-se verificar que as transferências são inexpressivas. No *Campus* Canoas não há transferência interna, porque os cursos são de eixos tecnológicos diferentes, portanto a saída de um curso também significa a saída da instituição. Muitos dos que deixam o curso tem como objetivo o ingresso em outra instituição via processo seletivo e não através de processo de transferência, mas esses casos não podem ser confirmados, a não ser que o aluno manifeste sua intenção, no momento de sua saída. Mesmo sendo confirmada a ida para outra instituição, estas situações são computadas como desligamentos. Pode-se observar, no Quadro 7, que o curso de Tecnologia em Automação Industrial tem um percentual maior de alunos desligados.

Os alunos “desligados” são aqueles que solicitaram o cancelamento da matrícula e, antes que se efetive o desligamento, são encaminhados para uma conversa com a equipe do Setor de Assistência Estudantil ou, pelo menos, preencham um formulário com os motivos de saída. Essa conversa tem como objetivo evitar a saída e/ou entender os motivos. Por outro lado, os alunos “evadidos” são aqueles que saem sem aviso, abandonam, e são os que apresentam um percentual maior nos três cursos. Nesses casos, o olhar atento das equipes em relação a sinais, pistas no comportamento dos alunos, tais como excesso de faltas não justificadas, trancamentos de matrícula, solicitação de auxílio estudantil (recebendo ou não), e reprovações podem evitar o desfecho negativo para o aluno e para instituição, no caso, a evasão. O *Campus* Canoas desenvolve essas e outras ações para permanência e êxito que serão descritas no próximo subcapítulo. Além disso, a criação de mecanismos como o que está propondo esse trabalho poderá ajudar bastante no direcionamento desse olhar, mudando as ações perante as

formas de abandono dos cursos, que deixarão de ser reativas, podendo fazer a diferença entre o ficar e o sair.

4.1.4 Estratégias de permanência e êxito do *Campus* Canoas

As ações de permanência e êxito executadas pelo *Campus* Canoas estão em consonância com o Plano Estratégico de Permanência e Êxito do Instituto Federal do RS, descrito anteriormente (IFRS, 2018a). O PEPEEIFRS foi aprovado recentemente, porém várias atividades já estavam sendo realizadas, com intuito de apoiar o desenvolvimento e o aprendizado dos alunos; possibilitar experiências de pesquisa e extensão, aproximando-os da realidade do mundo da ciência, do trabalho e dos arranjos sociais; acompanhá-los em suas trajetórias acadêmicas e incentivá-los a participar, fazendo-os sentirem-se pertencentes e parte do *campus*.

A partir de uma análise documental realizada no IFRS *Campus* Canoas, é possível descrever algumas das ações de permanência e êxito:

A. **Horário de atendimento docente aos alunos**: todos os docentes disponibilizam e divulgam aos alunos duas horas de atendimento além do horário de aula, normalmente no turno inverso, no qual são retomados conteúdos de acordo com as dificuldades individuais e organizadas estratégias de estudos.
B. **Monitoria**s: os alunos têm à sua disposição monitores de várias disciplinas propedêuticas e técnicas, em uma sala própria, para tirar dúvidas dos conteúdos e receber ajuda na realização de exercícios.
C. **Projetos de ensino, pesquisa e extensão**: além de fortalecer a tríade ensino pesquisa e extensão, como princípio do IFRS, os projetos têm como objetivo estimular a participação dos alunos em trabalhos extracurriculares, que ampliem seus conhecimentos dentro e fora da área dos cursos; o vínculo com seu curso e com o *campus*; bem como o espírito de participação, colaboração, solidariedade, proatividade, independência, responsabilidade, necessários para formação da cidadania.
D. **Atendimento psicológico**: o profissional de psicologia no ambiente educacional está aberto a escutar e acolher as demandas dos sujeitos, tentando encontrar alternativas para a melhoria da experiência neste espaço, além de buscar o bem-estar pessoal. Através da análise, escuta e intervenção, procura auxiliar e prevenir eventuais problemas que podem surgir durante a vida acadêmica, a saber: momentos de dificuldades pessoais, familiares, de interação com colegas e professores, entre outros. Algumas das ações desenvolvidas são: acolhimento e orientação; atendimento individual e familiar; atividades

com grupos, e em sala de aula; busca ativa de estudantes com problemas de frequência; orientação e manutenção do benefício para estudantes em sofrimento psíquico e atendimento a partir de encaminhamentos e discussões entre professores, colegas de equipe e conselhos de classe, entre outras ações.

E. **A Coordenadoria de Assistência Estudantil (CAE)**: é o setor que trata das situações que envolvam questões de ensino/aprendizagem, visando a ajudar nas dificuldades que estejam prejudicando estas áreas. Para tanto, trabalham com assuntos como hábitos de estudo, organização de horários, aproveitamento, frequência, atendimento relacionado a questões emocionais, relacionamento com colegas e com a instituição, atendimento aos pais, atendimento e acompanhamento social. Quando necessário, também propicia assessoramento ao corpo docente. A CAE também é responsável pelo programa de Benefício da Assistência Estudantil (BAE) e, ainda, participa das atividades de acolhimento aos estudantes, propõe oficinas de formação estudantil e pesquisa e difunde os dados sobre o diagnóstico sociodemográfico do *campus*. Em resumo, trata-se de um setor que busca viabilizar a igualdade de oportunidades e contribuir para a melhoria do desempenho acadêmico, implementando ações pedagógicas, psicológicas e sociais que contribuam para a permanência discente e para a melhoria de sua qualidade de vida.

F. **Apoio pedagógico**: o *campus* possui o setor de Apoio Pedagógico, vinculado às coordenações de ensino, de pesquisa e de extensão, composto por servidores de diferentes cargos: pedagogo, técnico em assuntos educacionais, assistentes de aluno e assistentes em administração. Algumas atividades realizadas: reuniões periódicas com os estudantes dos cursos integrados, para tratar de questões da turma referentes às relações aluno-aluno, professor-aluno e curso-aluno; mediação da relação aluno-professor; acompanhamento dos discentes quanto ao cumprimento da Organização Didática; auxílio em questões de saúde; organização da entrega de boletins e entrega direta aos pais, após a data da entrega oficial; agendamento de reuniões com pais e professores; organização das atividades relativas aos editais da pesquisa e extensão e atendimento aos bolsistas de pesquisa e extensão e atendimento aos alunos no encaminhamento de estágio curricular.

G. **Benefício da Assistência Estudantil**: é um programa institucional, baseado em uma política pública federal de assistência estudantil e de acordo com a Política de Assistência Estudantil do IFRS, que tem por finalidade subsidiar os estudantes em despesas relacionadas às questões escolares de modo a fortalecer suas condições de acesso, aproveitamento e permanência nas atividades acadêmicas.

H. **Programa de monitoria**: coordenado pela Diretoria de Ensino, tem como finalidade apoiar as ações de ensino por meio da oferta de monitoria aos estudantes regularmente matriculados no *campus*. Os estudantes participantes do programa poderão exercer monitoria remunerada, com concessão de bolsas de monitoria ou monitoria voluntária. O programa tem como objetivos: promover a melhoria dos processos de ensino e de

aprendizagem dos cursos oferecidos; iniciar os estudantes na prática de monitoria; estimular o desenvolvimento da criatividade na busca da socialização de saberes, aprimorando o processo formativo de profissionais enquanto cidadãos; contribuir para a permanência e o êxito dos estudantes matriculados nos cursos do IFRS – *Campus* Canoas.

I. **Bolsas de ensino, pesquisa e extensão**: através do Programa Institucional de Bolsas de Ensino (PIBEN), do Programa de Bolsas de Iniciação Científica e/ou Tecnológica (PROBICT) e do Auxílio Institucional à Produção Científica e/ou Tecnológica (AIPCT) e do Programa Institucional de Bolsas de Extensão do IFRS (PIBEX), o *Campus* Canoas oferece bolsas de ensino, pesquisa e extensão, divulgadas por meio de edital aos estudantes regularmente matriculados. As bolsas têm o objetivo de incentivar os discentes a participarem em atividades de ensino, pesquisa e extensão, com o intuito de proporcionar-lhes o conhecimento metodológico das ações, a formação integral, o desenvolvimento da sensibilidade social e da solidariedade, o espírito crítico e participativo e pró-ativo. Os valores estão vinculados ao tempo de atuação semanal que poderá ser de 4,8,12 ou 16 horas semanais.
J. **Alimentação**: os alunos dos cursos técnicos integrados, incluindo o PROEJA, recebem merenda escolar diariamente, no intervalo do turno das aulas. São distribuídos aproximadamente 375 (trezentos e setenta e cinco) lanches por dia.
K. **Calendário de matrícula e rematrículas**: o setor de Registros Escolar, com o apoio do setor de comunicação do *campus*, disponibiliza nos murais e na página do *campus* informações sobre os fluxos e os prazos das matrículas e das rematrículas, bem como envia informações, através do e-mail, para os alunos que estão afastados temporariamente, em processo de trancamento do semestre, sobre as datas do período de reingresso. Esse cuidado é para que os alunos não percam os prazos e o vínculo com a instituição.

Os projetos desenvolvidos pelo *campus* também possibilitam a integração e participação dos alunos em eventos, oficinas, cursos, palestras, atividades culturais e que desenvolvam habilidades artísticas através de oficinas em áreas diversificadas, a saber: teatro, quadrinhos, roteiro audiovisual, fotografia e jogos.

A forma como o aluno se percebe dentro da instituição busca e constrói seu conhecimento e a forma como convive e se relaciona com colegas, com os professores e com os demais membros da comunidade escolar. Em outras palavras, como se dá seu engajamento acadêmico ou de aprendizagem e o engajamento social ou de convivência, sendo este crucial sobre sua decisão de evadir ou permanecer estudando. (RUMBERGER, 2001).

Todas essas atividades desenvolvidas pelo *Campus* Canoas buscam integrar os alunos, fazer com que se sintam parte da instituição, que atuem como sujeitos construtores de sua trajetória e de seu conhecimento, que tenham constância e perseverança no propósito da sua formação profissional e que, no final do percurso acadêmico, estejam aptos a receber o diploma e atuarem como cidadãos conscientes e profissionais qualificados, em prol da sua comunidade.

4.1.5 Sistemas de Informações Acadêmicas no *Campus* Canoas do IFRS

Esta seção tem como foco descrever os sistemas de armazenamento e gerenciamento das informações do itinerário formativo dos alunos dos cursos do *Campus* Canoas e as planilhas de acompanhamento dos números relativos à situação desses estudantes. A intenção é compreender quais são as formas de acesso aos dados, tanto para o cálculo dos números da evasão quanto para a aplicação do processo de KDD.

4.1.5.1 Sistema de Informações Acadêmicas (SIA)

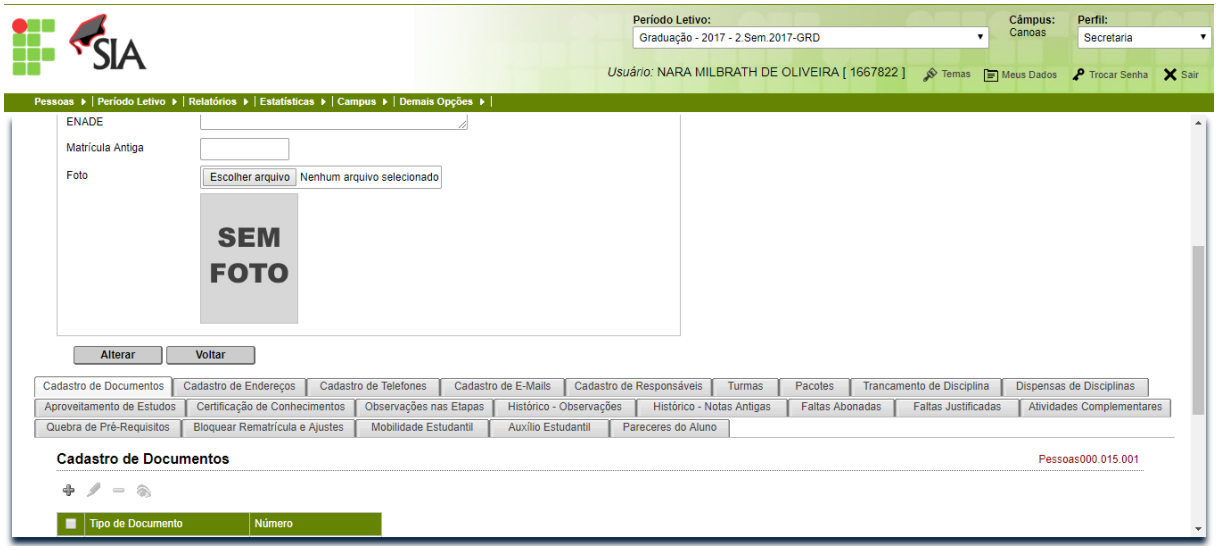
Os Sistemas de Informações Acadêmicas desempenham um papel crucial na organização das instituições de ensino e, com o avanço das Tecnologias da Informação e Comunicação (TICs), não apenas armazenam informações da vida acadêmica do aluno, como organizam a maior parte dos processos do ensino, fazendo a gestão acadêmica da instituição.

O Instituto Federal do Rio Grande do Sul (IFRS), em setembro de 2010, passou a utilizar para o registro de suas atividades acadêmicas o sistema de informações acadêmico desenvolvido e utilizado pela Universidade Federal de Rio Grande (FURG), em um processo de cessão ao IFRS.

Denominado Sistema de Informações Acadêmicas (SIA), possibilita o armazenamento e gerenciamento dos dados do itinerário formativo dos alunos de cursos superiores de uma universidade. O SIA foi adaptado pelo IFRS para, na época, atender seus oito *campi*, com níveis de ensino e modalidades de curso distintas. Inicialmente, era um “grande banco de dados” com informações cadastrais de professores, alunos, cursos e disciplinas, que foi sendo modelado pelo IFRS, para atender às suas necessidades, possibilitando, atualmente, vários processos

para o acompanhamento acadêmico, dentre eles: matrícula online, nas disciplinas para os cursos superiores; diário de classe online para todos os cursos e a extração de relatórios e de documentos acadêmicos para os alunos.

Figura 5 - Tela demonstrativa do SIA - informações no cadastro de aluno



Fonte: construção da autora.

Desde 2014, o IFRS está em processo em implantação de outro sistema e pretende manter de forma integrada as informações administrativas e acadêmicas do IFRS. Este processo de implantação ainda está em andamento e o *Campus* Canoas ainda não o utiliza.

Mesmo com os esforços de personalização do SIA (Figura 5) às rotinas do IFRS, este não atende a todas as necessidades cadastrais do *Campus* Canoas, sendo deficiente em situações como: cadastro de trancamento, período de trancamento, formas de ingresso, notas parciais, atividades complementares entre outras. Há situações em que suas rotinas de registro de informações são fragmentadas em diversas telas ou subprocessos, ocasionando demora e erros. Outra limitação importante é a falta de relatórios com dados quantitativos, por exemplo, relatório de número de alunos, de número de ingressantes, de alunos em situação de trancamento, de evadidos, de transferidos, de formados, de reprovados e outros.

4.1.5.2 Sistema de Informação do IFRS (SIFRS)

Para tentar suprir estas deficiências, a Coordenadoria de Tecnologia da Informação do *Campus* Canoas desenvolveu um sistema web para atuar em paralelo ao SIA, fornecendo ferramentas de gestão mais alinhadas às demandas existentes no *campus*, no sentido de complementar o funcionamento do sistema disponibilizado pela Reitoria do IFRS. A este sistema desenvolvido pelo *Campus* Canoas, apresentado na Figura 6, foi dado o nome de Sistemas IFRS (SIFRS). O SIFRS possui uma base de dados própria e não interage diretamente com a base de dados do SIA.

Este sistema passou a ser utilizado no *campus* para atender determinados processos, entre eles cita-se: gestão do processo seletivo, gestão do processo de matrículas, aplicação do Questionário Sociodemográfico (QS) ao público discente, gestão de planos de ensino, conselhos de classe, assistência estudantil, entre outros.

Para a integração de informações necessárias entre o SIFRS e o SIA são realizadas rotinas de importação e exportação de dados entre os sistemas. Por exemplo, os alunos matriculados pelo SIFRS, são importados pelo SIA. Posteriormente, as informações de vínculos de alunos, professores e turmas são transferidas do SIA para o SIFRS, para oferecer suporte a funcionalidades como, a elaboração de relatórios para o conselho de classe.

O SIFRS, portanto, atua de forma complementar ao sistema acadêmico oficial da instituição. Com a implantação do novo sistema no IFRS, será verificada a possibilidade de integração de tais funcionalidades existentes a este novo sistema, de modo a não causar uma regressão no fornecimento dos serviços.

Figura 6 - Tela demonstrativa do SIFRS, apresentando a gestão das matrículas realizadas

Fonte: construção da autora.

4.1.5.3 Planilhas do Setor de Registro Escolar

Além dos dois sistemas, o setor de Registro Escolar mantém informações em planilhas sobre o número de matrículas dos cursos e do total do *campus*, números de alunos matriculados, evadidos, trancados, transferidos, cancelados, formados, organizados por curso e ano letivo e a situação de matrícula do aluno no curso, individualmente, a cada semestre.

Essas planilhas foram criadas para serem usadas como apoio ao SIA, para armazenamento, organização e compartilhamento de informações com outros setores. Tais planilhas foram criadas com o Microsoft Office Excel, que é um software editor de planilhas (folhas de cálculo) e armazenadas no Google Drive. As planilhas criadas encontram-se descritas a seguir:

a) CONTROLE Nº DE ALUNOS MATRICULADOS: essa planilha (Figura 7) contém: o número de matrículas ativas, número de matrículas trancadas, número de alunos em mobilidade acadêmica e número total de matrículas, por cursos e por modalidade de ensino, organizada por semestre.

Figura 7 - Tabela “Controle nº de alunos matriculados

Período 2017/2

Curso	nº matrículas ativas	nº matrículas trancadas	nº alunos CSF*	nº total de matrículas
Integrado em Administração	113	-		
Integrado em Eletrônica	74	-		
Integrado em Informática	50	-		
Integrado em Desenvolvimento de Sistemas	65	-		
PROEJA	73	-		
Licenciatura em Matemática	74	14		88
Superior em Análise e Desenvolvimento de Sistemas	80	11		91
Superior em Automação	116	14		130
Superior em Logística	109	13		122
Total	**754**	**52**	**0**	**806**

Modalidade	nº matrículas ativas	nº matrículas trancadas	nº alunos CSF*	nº total de matrículas
Integrados	302	0	-	302
PROEJA	73	0	-	73
Superiores	379	52	0	431

* em Mobilidade Acadêmica pelo Ciência Sem Fronteiras

Fonte: construção da autora.

b) DADOS MATRÍCULAS POR ANO LETIVO: essa planilha (Figura 8) é organizada por ano letivo e contém: o número de vagas ofertadas, número de alunos que ingressaram em cada semestre, número de matrículas ativas de ingressos anteriores, número de evadidos, transferidos, cancelados, formados, total de matriculados e reprovados de cada curso.

Figura 8 - Tabela “Dados de Matrícula por Ano Letivo”

DADOS MATRÍCULA 2017

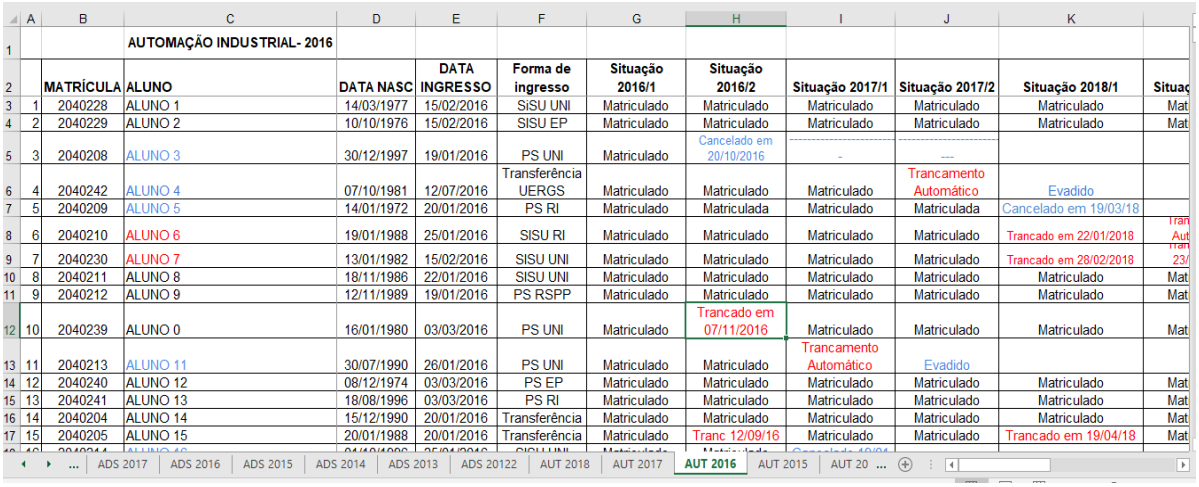
Curso	V.O.	Ingressos 2017/1	Ingressos 2017/2	Alunos ativos	Evadidos	Transferidos	I.C.A.I.A *	Cancelados	Formados	Total Matriculados	Reprovados
Técnico em Administração Integrado ao Ensino Médio	30	30		90	2	5	1	2	21	90	3
Técnico em Eletrônica Integrado ao Ensino Médio	24	24		61	2	6		4	9	64	9
Técnico em Informática Integrado ao Ensino Médio - em Extinção	0	0		52	1	1		1	23	26	3
Técnico em Desenvolvimento de Sistemas Integrado ao Ensino Médio	30	30		39	1	1		1		66	12
Licenciatura em Matemática	40	43	2	72	14		2	15		88	
Superior de Tecnologia em Análise e Desenvolvimento de Sistemas	30	30		82	12		2	7	2	91	
Superior de Tecnologia em Automação Industrial	30	41	7	108	15	1		8	3	129	
Superior de Tecnologia em Logística	36	40	3	125	16	1	1	12	17	122	
Técnico em Manutenção e Suporte em Informática - PROEJA	30	22		60	24			3	13	42	9
TOTAL		260	12	689						718	

Fonte: construção da autora.

c) ALUNOS CURSOS SUPERIORES: esta planilha (Figura 9) contém dados dos alunos dos cursos superiores desde 2011, separados nas abas por ano de ingresso de todos os cursos. Em cada aba, temos a lista de alunos que

ingressaram, pelo Processo Seletivo próprio, Sisu/Enem, transferência e ingresso de diplomados, nos dois semestres do ano letivo. Nas colunas da planilha, temos o número de matrícula, nome, data de nascimento, data de ingresso, forma de ingresso, situação a cada semestre (matriculado, trancado, transferido, cancelado e evadido), telefone, e-mail, CPF e RG de cada aluno. Essa planilha começou a ser usada desde a primeira turma dos cursos, contendo apenas o nome do aluno, data de nascimento, CPF e RG e atualmente, ela é compartilhada com outros setores e contém todas as informações descritas anteriormente. Para o escopo deste trabalho, é importante considerar duas informações nessa planilha, que são a forma de ingresso e a situação do aluno. As colunas da "situação do aluno" fornecem a informação da condição da matrícula do aluno a cada semestre. Com isso, pode-se saber quando e a forma que o aluno saiu da instituição, quantos e em que semestre houve trancamento, quantos semestre já cursou e quantos ainda faltam do prazo para a sua conclusão.

Figura 9 - Figura 9 - Planilha "Alunos Cursos Superiores".



	A	B	C	D	E	F	G	H	I	J	K	
1			AUTOMAÇÃO INDUSTRIAL- 2016									
2		MATRÍCULA	ALUNO	DATA NASC	DATA INGRESSO	Forma de ingresso	Situação 2016/1	Situação 2016/2	Situação 2017/1	Situação 2017/2	Situação 2018/1	Situaç
3	1	2040228	ALUNO 1	14/03/1977	15/02/2016	SiSU UNI	Matriculado	Matriculado	Matriculado	Matriculado	Matriculado	Mat
4	2	2040229	ALUNO 2	10/10/1976	15/02/2016	SISU EP	Matriculado	Matriculado	Matriculado	Matriculado	Matriculado	Mat
5	3	2040208	ALUNO 3	30/12/1997	19/01/2016	PS UNI	Matriculado	Cancelado em 20/10/2016	-	---		
6	4	2040242	ALUNO 4	07/10/1981	12/07/2016	Transferência UERGS	Matriculado	Matriculado	Matriculado	Trancamento Automático	Evadido	
7	5	2040209	ALUNO 5	14/01/1972	20/01/2016	PS RI	Matriculado	Matriculada	Matriculado	Matriculada	Cancelado em 19/03/18	
8	6	2040210	ALUNO 6	19/01/1988	25/01/2016	SISU RI	Matriculado	Matriculado	Matriculado	Matriculado	Trancado em 22/01/2018	Tran Aut
9	7	2040230	ALUNO 7	13/01/1982	15/02/2016	SISU UNI	Matriculado	Matriculado	Matriculado	Matriculado	Trancado em 28/02/2018	Tran 23/
10	8	2040211	ALUNO 8	18/11/1986	22/01/2016	SISU UNI	Matriculado	Matriculado	Matriculado	Matriculado	Matriculado	Mat
11	9	2040212	ALUNO 9	12/11/1989	19/01/2016	PS RSPP	Matriculado	Matriculado	Matriculado	Matriculado	Matriculado	Mat
12	10	2040239	ALUNO 0	16/01/1980	03/03/2016	PS UNI	Matriculado	Trancado em 07/11/2016	Matriculado	Matriculado	Matriculado	Mat
13	11	2040213	ALUNO 11	30/07/1990	26/01/2016	PS UNI	Matriculado	Matriculado	Trancamento Automático	Evadido		
14	12	2040240	ALUNO 12	08/12/1974	03/03/2016	PS EP	Matriculado	Matriculado	Matriculado	Matriculado	Matriculado	Mat
15	13	2040241	ALUNO 13	18/08/1996	03/03/2016	PS RI	Matriculado	Matriculado	Matriculado	Matriculado	Matriculado	Mat
16	14	2040204	ALUNO 14	15/12/1990	20/01/2016	Transferência	Matriculado	Matriculado	Matriculado	Matriculado	Matriculado	Mat
17	15	2040205	ALUNO 15	20/01/1988	20/01/2016	Transferência	Matriculado	Tranc 12/09/16	Matriculado	Matriculado	Trancado em 19/04/18	Mat

ADS 2017 | ADS 2016 | ADS 2015 | ADS 2014 | ADS 2013 | ADS 20122 | AUT 2018 | AUT 2017 | AUT 2016 | AUT 2015 | AUT 20 ...

Fonte: construção da autora.

Os dois sistemas e planilhas descritos acima acumulam dados com informações importantes para se conhecer os alunos do *Campus* Canoas. A forma de armazenamento das informações em planilhas e a permissão para extração dos dados dos sistemas no mesmo formato possibilitam o uso de técnicas e métodos para identificar informações válidas, novas e potencialmente úteis.

No próximo capítulo, é apresentado o processo para descoberta de conhecimento em base de dados, as técnicas e os métodos mais utilizados e mais

adequados para prever uma situação, no caso específico desta pesquisa, buscando predizer os alunos com propensão a não concluírem seu curso.

5 O KDD

Com o uso crescente da tecnologia em todas as áreas diversas empresas, organizações governamentais e não governamentais geram e armazenam quantidades cada vez maiores de dados. Saber analisar esses dados, de todos os tipos e de origens diversas, de forma mais rápida e precisa, pode ajudar em inúmeras pesquisas, comprovando hipóteses ou fornecendo novas pistas para solução de problemas simples ou complexos. Com isso, tem-se considerável aumento da eficácia e da eficiência, bem como significativo crescimento e produtividade em várias áreas. O grande desafio é extrair de conjuntos de dados informações novas, úteis e válidas para aumentar o conhecimento, pois este representa fator determinante para a tomada de decisões pelas empresas, organizações, governos, cientistas e sociedades, de hoje e de amanhã.

5.1 O QUE É O KDD?

Novas técnicas e ferramentas vêm sendo desenvolvidas para dar conta da natureza diferenciada e da quantidade de dados que surgem por conta das novas tecnologias. O *KDD* - Descoberta de Conhecimento em Bancos de Dados (do Inglês *Knowledge Discovery in Databases*) - é uma delas. Fayyad, Piatetsky-Shapiro e Smyth definem o KDD como um processo não trivial para identificar informações válidas, novas, potencialmente úteis e definitivas; e padrões compreensíveis em dados. Segundo eles, esse termo foi cunhado no primeiro workshop do *KDD*, realizado em 1989, e considerá-lo “não trivial” significa dizer que não é um processo corriqueiro, exige busca dos “Dados” que são as informações armazenadas e “padrão”[9] é um subconjunto desses dados (FAYYAD et al., 1996). Talvez, o mais importante dessa definição é compreender que o KDD é um “processo”, ou seja, pressupõe etapas com as quais novos padrões, válidos para serem replicados em outros dados de forma compreensível, serão descobertos, trazendo benefícios para o usuário. O objetivo é a descoberta desses novos padrões, úteis e replicáveis.

[9] Podemos imaginar, como exemplo, um fábrica de carros, cujo padrão ou modelo de determinada montadora seja carros populares. Carros populares é um padrão de carro, entre outros modelos fabricados.

Fayyad, Piatetsky-Shapiro, & Smyth (1996), representam as etapas do processo KDD da seguinte forma:

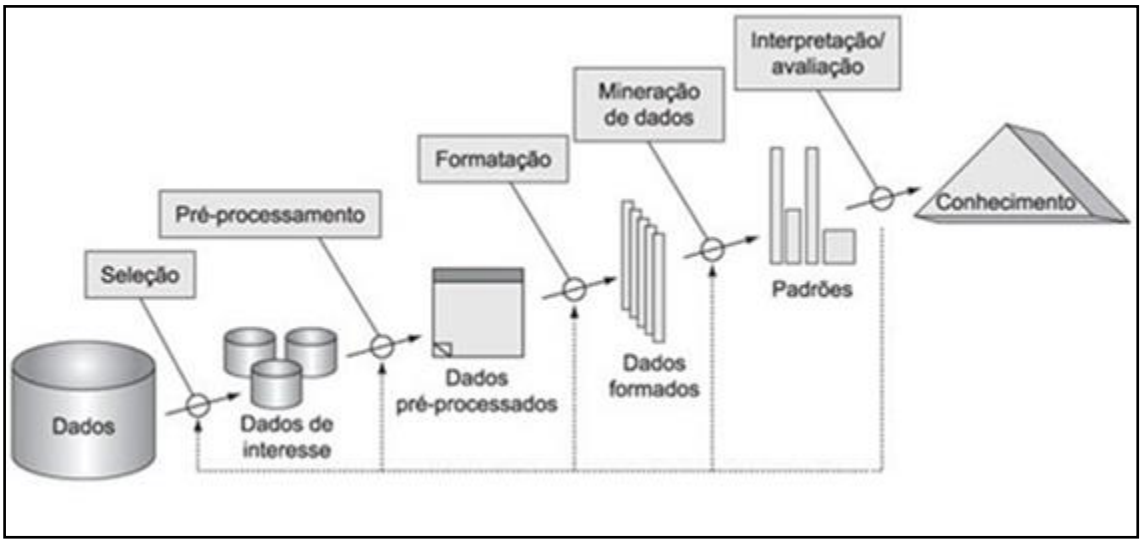
- Seleção de dados: seleção de um conjunto de dados, da qual se deseja extrair conhecimento, baseada no objetivo de se realizar o processo de KDD e na compreensão do domínio. A seleção pode usar toda a base de dados ou apenas um subgrupo de interesse (FAYYAD et al., 1996). No caso deste trabalho, a seleção é do subgrupo dos alunos dos cursos de tecnologia;
- Pré-processamento: dependendo da qualidade e da origem dos dados selecionados, vários procedimentos precisam ser realizados nesta etapa, conhecidos como limpeza de dados. Por esta característica, ela é considerada, por muitos autores, como a mais demorada e trabalhosa. O objetivo é a melhora da qualidade dos dados brutos. A melhora da qualidade e da compreensão dos dados geralmente melhora a qualidade da análise final (TAN et al, 2009); torna mais fáceis e rápidos o ajuste de parâmetros do modelo e o seu uso posterior e leva à construção de modelos mais fiéis (FACELI, et al, 2011). As atividades envolvem eliminação de ruídos, retirada de dados sem pertinência ou representatividade, inconsistentes, errôneos, duplicados, preenchimento de dados faltantes, “todos problemas comuns que ocorrem em bases de dados reais e que podem ser questionados aos especialistas do domínio[10]”(NEVES, 2003). De acordo com Neves (2003), o papel do especialista no domínio é fundamental em todo processo. Corroboramos com esse pensamento, pois a realização destes procedimentos, em grande parte, só foi possível nesta pesquisa pelo conhecimento da área abordada e pelo acesso às informações geradas e organizadas pelo setor de Registro do *Campus* Canoas. Além da qualificação dos dados, nessa fase é possível que ocorra a transformação de atributos, fusão de dados de fontes diferentes e seleção de registros e características que são importantes à tarefa de mineração;
- Formatação dos dados: é a transformação dos dados em um formato aceito pela técnica de mineração que se pretende usar e/ou ferramenta, para a extração de padrões;

[10] Especialistas do domínio - pessoa, usuário “[...] que conhece muito bem o problema a ser resolvido sabendo explicá-lo ao analista de mineração de dados [...]” (NEVES, 2003).

- Mineração de dados: é a aplicação de um algoritmo específico para extrair padrões (modelos) dos dados. Nesta etapa, é escolhido o método de mineração de dados trabalhando com sumarização, classificação, regressão, agrupamento/CLUSTERING, entre outros, que melhor se adaptem aos objetivos da aplicação do processo de KDD. Em outras palavras, são selecionados modelos e parâmetros, de acordo com a natureza dos dados e o algoritmo da mineração mais adequado para gerar padrões destes dados. (FAYYAD et al.,1996, p. 42);
- Interpretação ou avaliação: nesta última etapa os padrões encontrados são interpretados, analisados e avaliados para aquisição do conhecimento útil e novo, que poderá ser usado diretamente, ser incorporado em outro sistema para ação adicional ou, simplesmente, documentado e relatado às partes interessadas. (FAYYAD et al.,1996, p. 42). De acordo com SILVA e SILVA (2014, p. 575), as análises podem ser quantitativas ou qualitativas, dependendo da tarefa de mineração que foi escolhida.

O processo de *KDD* é interativo, pois seu sucesso depende de inúmeras decisões feitas pelo usuário, de preferência conhecedor do domínio e interativo, já que o fluxo das etapas não é linear, podendo ocorrer um retorno a quaisquer uma das etapas quando for necessário, conforme ilustrado na Figura 10 (FAYYAD et al.,1996, p. 42).

Figura 10 - Representação do processo de *KDD*



Fonte: FAYYAD et al.,1996.

5.2 A ETAPA DE MINERAÇÃO DE DADOS

A mineração de dados (do Inglês *Data Mining*) deve ser usada para descobrir conhecimentos em grandes quantidades de dados de diversas áreas como medicina, bioquímica, física, astrofísica, sistemas de *marketing*, análise de crédito, vendas e, também, na área da educação. Grande parte dos trabalhos desenvolvidos se concentra nesta etapa do processo de *KDD*, tendo-a como a mais importante.

De acordo com Pang-Ning Tan e seus colaboradores:

> A mineração de dados é uma tecnologia que combina métodos tradicionais de análise de dados com algoritmos sofisticados para processar grandes volumes de dados. Ela também abriu oportunidades interessantes para explorar e analisar novos tipos de dados e para se analisar tipos antigos de novas maneiras. (TAN et al., 2009, p. 1).

Os métodos ou técnicas usadas nesta etapa estão diretamente relacionadas aos objetivos de uso da mineração de dados. Segundo FAYYAD et al. (1996, p. 43) os objetivos podem ser de dois tipos: verificação e descoberta. Se o objetivo é a verificação, o sistema se limitará a verificar hipóteses pré-estabelecidas, já na descoberta, o sistema autonomamente evidencia novos padrões. O objetivo de descobrir padrões úteis ou a meta de descoberta é subdividida em previsão, que envolve o uso de algumas variáveis (atributos) da base de dados, com o objetivo de prever valores desconhecidos ou futuros de outras variáveis de interesse, e a descrição se concentra em encontrar padrões compreensíveis para o ser humano, para descrever os dados. Nesta pesquisa, estamos principalmente interessados na descoberta de padrões que possam prever valores futuros, ou seja, que possam prever o futuro comportamento dos alunos em relação a sua situação final no curso.

TAN et al. (2009, p. 8) se referem à previsão e à descrição como tarefas de mineração de dados. Na tarefa de previsão, o objetivo é “prever o valor de um determinado atributo baseado nos valores de outros atributos.” Os atributos, cujos valores são conhecidos, são chamados de variáveis independentes ou explicativas, enquanto que o atributo a ser previsto é chamado de variável dependente ou alvo. E o objetivo da tarefa descritiva é “derivar padrões (correlações, tendências, grupos, trajetória e anomalias) que resumem os relacionamentos subjacentes nos dados.”

TAN et al. (2009) descrevem quatro das tarefas centrais da mineração de dados:

1) modelagem de previsão;
2) análise de associação;
3) análise de grupo;
4) detecção de anomalias.

Como mencionado anteriormente, esta pesquisa tem interesse em prever comportamentos futuros. Desta forma, o presente trabalho se concentra na tarefa de modelagem de previsão ou, também, chamada de análise preditiva.

De acordo com Silva et al. (2016), a análise preditiva é um processo “que permite descobrir o relacionamento existente entre os exemplares de um conjunto de dados, descritos por uma série de características (atributos descritivos), e os rótulos a eles associados (atributo de classe)”.

Para TAN et al. (2009, p. 9) a modelagem de previsão “se refere à tarefa de construir modelo para a variável alvo como uma função das variáveis explicativas.” Dependendo do tipo dos valores do atributo-alvo (também chamado de rótulo da classe, atributo de categorização, atributo de classe, variável dependente), a modelagem de previsão pode ser chamada de classificação, quando a variável dependente é discreta e chamada de regressão quando a variável dependente é contínua. Por exemplo, prever se um aluno terá aprovação ou não em uma disciplina, de acordo com suas notas do semestre é uma tarefa de classificação, porque a variável-alvo é binária (aprovado ou reprovado). Porém, se for necessário prever notas futuras, baseadas em anteriores, é uma tarefa de regressão, porque nota é um atributo de valor contínuo (3,2; 3,5; 4,2; 6,0). As duas tarefas têm como objetivo “aprender um modelo que minimize o erro entre os valores previsto e real da variável-alvo.”

No processo de construção de um modelo preditivo, um conjunto de dados, cujas classes são conhecidas, é submetido a um algoritmo de aprendizagem para que este construa um modelo que será aplicado ao conjunto de teste, cujos dados não possuem os rótulos de classe conhecidos. “O modelo gerado pelo algoritmo de

aprendizagem deve se adaptar bem aos dados de entrada e prever corretamente os rótulos de classes de registros que ele nunca viu antes”. (TAN et al., 2009, p. 174).

6 TRABALHOS RELACIONADOS

A busca por trabalhos relacionados foi realizada no catálogo de teses e dissertações da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) e compreende o período entre 2014 e 2017, usando várias chaves de buscas, relacionadas ao tema desta proposta de dissertação. O objetivo do rastreio foi localizar o maior número de trabalhos com informações consolidadas, principalmente os que correlacionam mineração de dados e evasão. A pesquisa na Plataforma Sucupira é considerada de grande importância pelo fato das teses e dissertações serem fornecidas diretamente à Capes pelos programas de pós-graduação de todo o país, que se responsabilizam pela veracidade dos dados.

O processo de busca por trabalhos no portal da Capes utilizou como termos de pesquisa: mineração de dados educacionais, resultando em 52 trabalhos, sendo que destes, 24 têm alguma relação com o tema dessa pesquisa, ou seja, usam mineração para buscar conhecimento, em base de dados educacionais, sobre a evasão, retenção e desempenho dos alunos, no ensino presencial ou a distância. Com a chave “KDD”, foram encontrados 50 resultados, sendo que dois trabalhos já haviam sido selecionados na busca anterior e os demais relacionados a outras áreas. Com a associação da palavra evasão, KDD AND EVASÃO, a plataforma relacionou apenas dois trabalhos já selecionados. Usando-se a busca com associação de termos “mineração AND evasão” resultou na localização de 10 trabalhos diferentes. Buscou-se, também, a associação de duas expressões, palavras chaves do trabalho, quais sejam: "evasão no ensino superior" AND "mineração de dados" e quatro trabalhos foram localizados, apenas dois dentro do período pesquisado, sendo um já selecionado anteriormente.

Outras chaves de busca foram utilizadas, mas não foram encontrados trabalhos na Plataforma Sucupira, são elas: mineração de dados na educação; *KDD* na educação; mineração e *KDD*; "mineração na evasão"; "mineração evasão" (com aspas); mineração evasão (sem aspas); “mineração e *KDD* na evasão”; "*KDD* e mineração na evasão".

A pesquisa na Plataforma Sucupira resultou num total de 40 trabalhos que, de alguma forma, tem relação com essa pesquisa, distribuídos no período analisado, como esquematiza o Gráfico 5.

Gráfico 5 - Trabalhos relacionados Portal da CAPES, por ano de publicação

Fonte: construção da autora, com base nos dados do Portal da Capes.

O Gráfico 5 ilustra o aumento do número de trabalhos envolvendo a Descoberta de Conhecimento em Banco de Dados e dados educacionais. Algo interessante, que foi observado, é que levando-se em consideração as áreas de concentração em que foram desenvolvidas essas dissertações, apenas quatro foram na área da educação. As demais foram em áreas ligadas à informática e à computação, a saber: Ciência da Computação, Engenharia de Sistemas Computacionais, Banco de Dados, Controle e Otimização de Processos Industriais, Computação Aplicada e outras. A expectativa é que estes trabalhos consigam aproximar as comunidades da educação e da mineração de dados, com o intuito de trazer boas análises e descobertas de conhecimentos nos dados educacionais, promovendo melhorias nessa área.

Os quarenta trabalhos foram selecionados tendo como foco níveis e modalidade de ensino diferentes, de acordo com o Gráfico 6.

Gráfico 6 - Trabalhos relacionados Portal da CAPES, por níveis e modalidades de ensino

Fonte: construção da autora, com base nos dados do Portal da Capes.

Com base no gráfico, é possível observar que a maioria dos trabalhos se concentra em cursos de nível superior, tendo os cursos presenciais um número de trabalhos 20% menor em comparação aos cursos à distância. Os trabalhos de pesquisa realizados nos cursos superiores à distância utilizam dados de interação dos alunos nos Ambientes Virtuais de Aprendizagem (AVAS) ou de questionários respondidos pelos sujeitos envolvidos no processo de ensino a distância, diferentemente da proposta deste trabalho.

Sendo assim, foram selecionados os 12 trabalhos realizados nos cursos de nível superior presenciais, mesmo nível e modalidade de ensino do público-alvo desta dissertação, além dos dois trabalhos que utilizam dados de cursos superiores nas modalidades a distância e presencial.

O trabalho de Santana (2015), avaliou a eficácia de algoritmos de predição em dados da disciplina de Programação Introdutória, nas modalidades a distância e presencial. Por meio de seu estudo, verificou que o algoritmo *Vetor de Suport* (*SVM*, *Support Vector Machine*), apresenta melhores resultados dentre as técnicas de predição analisadas por ele, alcançando uma taxa de *f-measure* de 83% no ensino presencial e 92% no ensino a distância. Esta observação se deu principalmente após a realização das etapas de pré-processamento e ajustes de parâmetros dos algoritmos, confirmando que é possível identificar os estudantes propensos ao insucesso no início da disciplina. O trabalho do autor, assim como

este trabalho, usa algoritmos de predição para identificar de forma precoce o insucesso dos estudantes, porém analisa o desempenho em uma única disciplina em duas modalidades de ensino.

O estudo de Machado (2015) identificou questões relevantes, de acordo com os 23 coordenadores de curso, em relação ao problema da evasão nos cursos de graduação presenciais, através de um questionário contendo 15 questões do tipo fechada escalar, uma questão de livre resposta e duas questões do tipo fechada única. Os dados de 20 anos do curso de Ciência da Computação de uma instituição de ensino superior, foram minerados visando explorar as questões apontadas pelos coordenadores de curso. Utilizando o software *WEKA*, aplicou o algoritmo *Apriori* com o apoio do algoritmo *J48* e extraiu regras para cada uma das questões mais pertinentes. Os resultados encontrados permitiram identificar disciplinas chaves, associadas tanto ao sucesso quanto ao insucesso do aluno no curso. Com base nas regras encontradas, foi possível afirmar que o desempenho dos estudantes nas disciplinas do primeiro ano do curso está fortemente associado à decisão de desistir ou permanecer no curso. O trabalho de Machado confirma a questão levantada por outros autores do tema evasão, constantes no referencial teórico e também incluída neste trabalho, representada pelo desempenho nas disciplinas iniciais do curso. Por outro lado, não leva em consideração as características sociais, culturais e econômicas que podem influenciar a decisão de deixar o curso e o próprio rendimento do aluno.

Manhães (2015), em sua tese, propôs a arquitetura *EDM Wave*, que “engloba todo o processo de descoberta de conhecimento em dados (pré-processamento, mineração de dados e pós-processamento)”. A arquitetura e os modelos propostos utilizaram apenas dados acadêmicos de graduandos da Universidade Federal do Rio de Janeiro (UFRJ), que variaram no tempo, no período de 16 anos. Um aspecto importante do trabalho foi a análise dos doze algoritmos de classificação. Os resultados mostraram que a “arquitetura proposta é capaz de predizer o desempenho acadêmico dos graduandos a cada semestre letivo com precisão em torno de 80%”, e, ainda, identifica as principais variáveis que distinguem aqueles que concluem dos que não concluem o curso. A arquitetura *EDM Ware* acrescenta funcionalidades aos Sistemas de Gestão Acadêmicos (SGA) legados, como a previsão do desempenho acadêmico a cada semestre, bem como a presença de

alertas, indicando os alunos que estão em risco de evasão. Os resultados obtidos por Manhães, com o estudo do desempenho dos algoritmos de classificação utilizando dados acadêmicos, servem de referência para este trabalho, que usa dados acadêmicos e sociodemográficos. Além de incluir dados sociodemográficos como atributos, buscou-se identificar a propensão da não conclusão do curso da forma mais precoce possível, não sendo realizada a previsão de desempenho a cada semestre.

O trabalho de Costa (2015) apresenta uma proposta diferenciada, tanto em relação à abordagem da mineração, quanto às dificuldades para concluir um curso. Diferente dos demais, ele propõe avaliar métodos e técnicas para identificar as disciplinas nas quais os alunos têm grande dificuldade para cursar nos currículos. Para atingir esse objetivo, as grades curriculares e os históricos dos alunos foram representados por grafos. Usando métodos baseados na mineração em grafos, uma especialidade da mineração de dados, buscou-se identificar os caminhos mais longos (também chamados de caminhos críticos ou caminhos mais custosos) em currículos de graduação, ou seja, identificar caminhos que possam estar provocando maior retenção e, com isso, dificultando a conclusão dos cursos de graduação. Para avaliar o modelo proposto, foram utilizados dados reais de alunos formados de quatro cursos de graduação da Universidade Federal Fluminense, dos últimos dez anos, focando, apenas, no problema da retenção no ensino superior. Os resultados permitem gerar hipóteses que ofereçam aos coordenadores e professores de instituições de ensino superior novas possibilidades de identificação do aumento da evasão e da retenção nos cursos de graduação. Costa apresenta um trabalho diferente desta dissertação, no qual se refere à técnica de mineração de dados utilizada e ao foco na retenção, considerada uma das causas da evasão por muitos autores. Desta forma, amplia-se a perspectiva de trabalhos futuros, abordando os fatores internos da instituição e os aspectos do currículo e do curso.

Oliveira Júnior (2016), propõe uma abordagem computacional para a identificação de padrões a serem utilizados na análise da evasão de estudantes em cursos presenciais de graduação da Universidade Tecnológica Federal do Paraná. O autor apresenta um método para seleção dos melhores atributos para a tarefa de classificação, que considera as classes “haverá evasão” e “não haverá evasão”. Os experimentos foram realizados com dados consolidados em um *Data Warehouse*,

que permitiu investigar a evasão entre 1980 e 2014. A pesquisa abordou os problemas mais comuns que ocorrem na mineração de dados educacionais, como a seleção do subconjunto de atributos, dados desbalanceados, valores discrepantes e sobreajuste. Os resultados experimentais apresentaram como atributos mais relevantes a previsão da evasão, indicando a contribuição da criação de atributos na tarefa de mineração de dados. Nos experimentos com a base completa, o algoritmo *JRip* obteve acurácia de 87,69% e o algoritmo *J48* obteve acurácia de 85,78%, podendo estes índices serem considerados expressivos no contexto educacional. O trabalho aponta que estas inferências podem apoiar a tomada de decisão pelos gestores educacionais situados nos níveis estratégico, tático e operacional. Os experimentos para criação de atributos, apresentados por Oliveira Junior, mostraram a possibilidade de se criar um único atributo de desempenho acadêmico, a partir das disciplinas do primeiro semestre. Os experimentos de Oliveira Junior indicaram o caminho para resolver o problema do aumento da dimensionalidade, a partir da criação de um único atributo, em substituição ao uso do rendimento em cada disciplina como atributo.

A dissertação de Hoed (2016) apresenta um estudo sobre evasão nos cursos de graduação da área de computação, com base em dados de instituições públicas e privadas, fornecidos pelo INEP, e em dados fornecidos pela Universidade de Brasília (UnB). Nos estudos quantitativos realizados, são obtidas as taxas anuais de evasão e aplicada a técnica estatística de análise de sobrevivência e mineração de regras de associação via algoritmo *Apriori*. É apresentada, também, uma análise qualitativa a partir de questionários aplicados a alunos evadidos em cursos superiores de computação da UnB, com o intuito de analisar as causas de evasão. Inicialmente, são comparadas as evasões nas grandes áreas do conhecimento. Logo após, é feito um detalhamento para a grande área de ciências, matemática e computação. Posteriormente, um estudo de caso em quatro cursos da área de computação foi conduzido na UnB, para levantamento de causas de evasão nesses cursos. Foram obtidas evidências de que a relação candidatos/vagas é inversamente proporcional à evasão e de que os cursos da grande área de ciências, matemática e computação, que requerem maior uso de conhecimentos matemáticos e de abstração algorítmica, possuem maiores taxas de evasão e, ainda, de que o sexo dos estudantes, a forma de ingresso na instituição e ser ou

não cotista afetam as taxas de evasão nos cursos na área de computação. As evidências encontradas por Hoed contribuíram com esta dissertação, no sentido de reforçar a seleção dos atributos mais pertinentes para a construção de um modelo preditivo, o qual possa identificar com antecedência a possibilidade de um aluno evadir.

Caetano (2016), em sua dissertação, apresenta um trabalho de pesquisa na área de mineração de dados educacionais e aprendizado de máquina, cujo principal objetivo é prever o desempenho de alunos em uma disciplina presencial, de um curso de Bacharelado em Ciência da Computação (BCC), de um Centro Universitário situado na grande São Paulo. Seus experimentos envolvem seleção de atributos referentes aos dados do cadastro dos alunos e histórico de notas na disciplina Construção de Algoritmos e Programação I (CAP1), oferecida no 1º semestre do curso de BCC, e usam aprendizado automático, com três dos mais conhecidos algoritmos de classificação, *Naïve Bayes, Nearest Neighbor (1Bk) e J48 (versão Weka* do *C4.5),* com 4-validação cruzada, no ambiente do *WEKA*. Os resultados mostraram que é possível predizer o desempenho de estudantes com precisão em torno de 70% com informações iniciais e em torno de 90%, se usados o conjunto completo de atributos, ou o conjunto reduzido, fazendo uso dos atributos mais expressivos na classe. Assim como Caetano, foram realizados experimentos com algoritmos classificadores e com um número mais reduzido de atributos que o proposto inicialmente, para que se possa simplificar o modelo mantendo a precisão, facilitando seu uso posteriormente.

Motta (2016) faz um estudo exploratório de métodos de classificação, com o objetivo de predizer o desempenho e o abandono de alunos a partir de dados demográficos, sócio-econômicos e resultados acadêmicos. Com base nas tendências encontradas na revisão da literatura, utilizou como principal algoritmo de classificação o *J48*. Segundo ele, a preferência por esse algoritmo se deve ao fato das árvores de decisão permitirem uma análise dos atributos usados nos modelos gerados, mantendo níveis de acurácia aceitáveis, enquanto técnicas de regressão logística e redes Bayesianas, apesar de estudos comparativos terem mostrado melhores resultados, funcionam como uma caixa preta. Suas análises concluíram que a técnica de Resample, que escolhe um subconjunto balanceado dos dados, apresentou melhores resultados que a técnica de *SMOTE*, que gera dados

sintéticos para balancear os dados. Além disso, não foram apresentadas vantagens significativas no uso de técnicas de seleção de atributos, porém notas e aspectos econômicos são atributos que aparecem com frequência nos modelos gerados. Motta concluiu que o uso de classificadores é um caminho promissor para a predição de desempenho e abandono, o que corrobora a proposta deste trabalho.

Amaral (2016) propõe em sua dissertação analisar o perfil dos estudantes ingressantes, com o objetivo de classificá-los através de uma abordagem para aplicação de técnicas diretas de mineração de dados, de acordo com o risco de evasão que apresentam. As experimentações foram conduzidas no ambiente da Universidade Federal de Pernambuco (UFPE), sendo selecionados 16 dados socioeconômicos dos discentes ingressantes, no período de 2009/1 a 2011/2. Utilizando a ferramenta Orange, analisou-se o desempenho de algoritmos classificadores, através de cinco experimentos, no qual foi observada a viabilidade da abordagem proposta, com destaque para o desempenho dos algoritmos *Naive Bayes, Logistic Regression e Classification Tree*, os quais apresentaram acurácia de classificação superior a 70%. Amaral conclui que é válido utilizar apenas dados disponíveis quando do ingresso do discente na instituição, para identificação dos casos de evasão, mesmo que, segundo ele, a acurácia tenha sido inferior a outros trabalhos que utilizam histórico acadêmico, pois a abordagem é capaz de fornecer um indicativo do risco de evasão logo no início da vida acadêmica do discente. O trabalho de Amaral oferece subsídios para esta dissertação, a qual mantém semelhança, pelo uso de dados socioeconômicos, de algoritmos classificadores e pela preocupação em identificar o risco da evasão o mais precocemente.

Também usando classificadores, Sousa (2017), apresenta a Mineração de Dados Educacionais e um experimento envolvendo previsão de provas parciais, realizada através dos dados gerados a partir da interação dos alunos em dois ambientes virtuais, da disciplina presencial de Introdução à Programação de Computadores, da Universidade Federal do Amazonas, e busca classificar os alunos de acordo com as notas obtidas em, no máximo, três classes: satisfatório, insatisfatório e sem conceito (alunos evadidos), usando a ferramenta *WEKA*. O melhor modelo geral desenvolvido obteve acurácia de 78,64%, utilizando o algoritmo *Random*. O trabalho de Sousa fez experimentos interessantes com algoritmos classificadores, que indicam o potencial dos mesmos, porém o fato de

prever o desempenho de alunos em uma disciplina presencial, usando dados de interação em um ambiente virtual, torna sua contribuição pouco relevante para esta dissertação.

O trabalho de Couto (2017) investiga os perfis de alunos ingressantes, de cursos presenciais de graduação da Universidade Federal do Pará, propensos à evasão e à retenção, utilizando Redes Bayesianas. Foram selecionados registros acadêmicos de 98.698 ingressantes até o ano de 2016, contendo informações cadastrais, forma de ingresso, de rendimento acumulado e eficiência acadêmica durante o percurso, num total de 31 atributos. Os dados foram submetidos ao processo de Descoberta de Conhecimento em Base de Dados, especificamente na etapa de Mineração de Dados. Os padrões desejados foram extraídos valendo-se da tarefa de classificação. Em adição, foram realizadas várias análises de desempenhos da Rede Bayesiana junto a outros algoritmos clássicos do aprendizado supervisionado. Em três estudos de casos avaliados, o classificador baseado em Redes Bayesianas apresentou acurácia superior a 82%, condição que legitima a sua utilidade no domínio pesquisado. De acordo com o autor, os resultados atingidos foram satisfatórios e apontaram fortes influências das variáveis índice de eficiência acadêmica, percentual de reprovação e número de trancamentos em componentes curriculares, na propensão à evasão ou à retenção. Em consonância com o trabalho de Couto, foram criados e incluídos dois atributos acadêmicos nesta dissertação, são eles: percentual de aprovação nas disciplinas do primeiro semestre e trancamento no segundo semestre do curso.

No trabalho de Assis (2017) foram utilizados dados do Censo da Educação Superior (CES) e Exame Nacional do Ensino Médio (Enem). Assis conduziu os experimentos nos dados dos alunos da UnB, no qual foram criados modelos de cinco algoritmos de classificação para analisar a evasão de alunos ingressantes em três diferentes níveis de evasão: de curso, de área de estudo e de IES. Entre os algoritmos testados, o CART obteve um desempenho superior, na métrica de sensibilidade e desempenho de cerca de 84% para evasão em nível de curso. Nos demais testes, não houve diferença estatisticamente significativa entre os algoritmos. As principais características identificadas nos alunos que possuem propensão a evadir são: ingressar no primeiro semestre; possuir vínculos com mais de uma IES; obter notas acima da média nos exames do Enem e já ter concluído o

ensino médio no momento que realiza as provas do Exame Nacional do Ensino Médio. Também foi desenvolvido um pacote para o software *R,* em que é possível treinar novos classificadores de evasão, os quais podem ser utilizados para determinar, em qualquer IES ou grupo de IES, quais alunos possuem maior tendência a evadir. O estudo com cinco algoritmos classificadores usando dados do CES e do Enem é bastante relevante, porém identifica características bastante genéricas dos estudantes. Por este fator, utilizaram-se dados dos sistemas de acompanhamento acadêmico da própria instituição e do questionário sociodemográfico preenchido pelos alunos, na busca de uma maior adequação do modelo à realidade dos estudantes.

A seguir é apresentado, no Quadro 8, o resumo dos doze trabalhos descritos, no qual é possível perceber que, com exceção do trabalho de Sousa (2017), que analisa dados de interação em ambientes virtuais, e de Costa (2015), que correlacionou disciplinas e percursos curriculares através da mineração de grafos para estudar a retenção, o enfoque da maioria dos trabalhos selecionados encontra-se no aluno, nas suas características individuais, sociais e econômicas e em seus resultados acadêmicos.

Quadro 8 - Resumo dos trabalhos relacionados

Quadro resumo dos trabalhos relacionados			
AUTOR	**DADOS UTILIZADOS**	**ALGORITMOS EMPREGADOS**	**FERRAMENTA MINERAÇÃO**
SANTANA, Marcelo Almeida. (2015)	Dados da disciplina de Programação Introdutória, nas modalidades a distância e presencial.	O algoritmo *Vetor de Suport (SVM, Support Vector Machine*) teve melhor desempenho se comparado com a *árvore de decisão, naive bayes e rede neural*	*WEKA*
MANHAES, Laci Mary Barbosa. (2015)	Dados acadêmicos dos estudantes, nos dois primeiros semestres letivos entre os anos de 1994 e 2010 (16 anos)	Analisou doze algoritmos de classificação e o *Naive Bayes* apresentou melhor resultado geral, com 80% de acurácia.	*WEKA*
MACHADO, Roger Douglas (2015)	Dados de 20 anos do curso de ciência da computação, de acordo com questões referentes à evasão, levantadas por coordenadores de curso	Algoritmo *Apriori* com o apoio do algoritmo *J48*	*WEKA*
COSTA, Jefferson de Jesus. (2015)	Dados dos históricos escolares de alunos formados e as grades curriculares de cursos de graduação, para aplicação de técnicas e métodos de mineração em grafos.	Algoritmos de mineração em grafos criados pelo autor, devido à especificidade do contexto.	Ferramenta desenvolvida pelo autor

MOTTA, Porthos Ribeiro de Albuquerque. (2016)	Dados demográficos, sócio-econômicos e resultados acadêmicos, oriundos do vestibular e do banco de dados acadêmico	Principal (*J48, Árvore de decisão); regressão logística e redes Bayesianas*	*WEKA*
OLIVEIRA JUNIOR, Jose Gonçalves de. (2016)	Dados acadêmicos consolidados em um Data Warehouse, que permitiu investigar a evasão entre os anos de 1980 e 2014	JRip obteve acurácia de 87,69% e o algoritmo J48 obteve acurácia de 85,78%,	*WEKA*
AMARAL, Marcelo Gomes do. (2016)	Dados socioeconômicos (16 atributos)	CLASSIFICADORES: *Naive Bayes, Logistic Regression e Classification Tree*	*ORANGE*
CAETANO, Maitê Marques. (2016)	Dados do cadastro dos alunos e histórico de notas de uma disciplina, do curso ciência da computação	Algoritmos de classificação, *Naïve Bayes, Nearest Neighbor (1Bk) e J48 (*versão *Weka* do *C4.5)*	*WEKA*
HOED, Raphael Magalhaes. (2016)	Base dados do INEP, UnB e questionário sobre causas da evasão	Regras de associação via algoritmo *Apriori.*	*Software R (versão 3.1.2)*
SOUSA, Marilia Maria Bastos de Araujo Cavalcanti Feitoza. (2017)	Dados gerados a partir da interação dos alunos em dois ambientes virtuais, utilizados em uma disciplina presencial	78,64% com o algoritmo *Random Forest*	*WEKA*
COUTO, Diego da Costa do. (2017)	Dados cadastrais, forma de ingresso, dados de rendimento acumulado e eficiência acadêmica, num total de 31 atributos	Avaliou o desempenho de 9 classificadores e selecionou as Redes Bayesianas	*WEKA*
ASSIS, Lucas Rocha Soares de. (2017)	Dados do Censo da Educação Superior (CES) e Exame Nacional do Ensino Médio (Enem), dos alunos ingressantes da UnB	Cinco algoritmos de classificação; *CART* apresentou melhor desempenho	*Software R*

Fonte: construção da autora.

Ainda, pode-se destacar o fato da maioria dos autores usarem algoritmos classificadores, sendo mais utilizados os algoritmos *Naive Bayes e Árvore de Decisão (J48),* e o Software *WEKA.*

Os trabalhos relacionados motivaram essa dissertação em busca de predizer alunos com propensão à evasão, e, dessa forma, auxiliar na redução desse problema. Acredita-se que técnicas e algoritmos foram replicados, conforme os estudos da área, porém cada trabalho (e este também), é realizado em um contexto específico. Em face disso, o caminho percorrido, as estratégias de coleta e tratamento dos dados, os objetos e atributos selecionados de acordo com a realidade da instituição e com os objetivos estabelecidos, tornaram esse trabalho diferente dos demais. Além disso, foi selecionada como ferramenta para a mineração de dados e criação do modelo, o software *RapidMiner,* ainda pouco utilizado em trabalhos na área da educação. Pode-se apontar, também, como

diferencial, o uso de dados extraídos do Questionário Sociodemográfico, aplicado pela instituição para identificar o perfil dos alunos, preenchido pelos próprios estudantes, o que sustenta informações mais reais, apesar de menos padronizadas.

7 METODOLOGIA

Neste capítulo são especificadas as escolhas metodológicas para se atingir os objetivos propostos, buscando alcançar êxito na construção de um modelo de predição que aponte os alunos com propensão à evasão.

Esta investigação, conforme Lakatos (2003) e Gil (2008), constitui um estudo de caso que utilizou pesquisa exploratória e uma intensa pesquisa documental. Quanto à natureza dos dados, pode ser definida tanto como qualitativa quanto quantitativa.

Inicialmente, foi realizado um estudo bibliográfico e documental, de modo a se obter uma efetiva apropriação da evasão em cursos regulares de nível superior, além de abordar como esse tema vem sendo trabalhado na Rede Federal e no IFRS, além de buscar outras referências relativas à compreensão do processo de *KDD*.

A segunda etapa, pertinente à metodologia, foi a realização de uma análise estatística dos números de evasão no *Campus* Canoas, de modo a identificar variáveis relevantes e estabelecer dados que poderiam ser pertinentes à pesquisa.

Por fim, tem-se a aplicação do processo de *KDD* nos dados selecionados, especificamente dos alunos dos cursos superiores de tecnologia do *Campus* Canoas, para descoberta de conhecimentos em relação ao aluno que não conclui o curso. O objetivo é criar um modelo preditivo que possa indicar, com o máximo de precisão possível, aqueles com propensão à evasão.

7.1 DELIMITAÇÃO DO LOCAL E DOS SUJEITOS

O estudo foi realizado no *Campus* Canoas do Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul (IFRS) e analisou os dados de 938 alunos, dos três cursos superiores de tecnologia desta unidade.

Esse trabalho de pesquisa não interagiu com sujeitos, mas utilizou dados armazenados nos sistemas de informações acadêmicas, de alunos que estiveram matriculados nesta instituição por, pelo menos um dia, no período compreendido entre os anos de 2011 e 2017, nos quais a matrícula não foi cancelada antes do início das aulas. Portanto, são dados de alunos de cursos de tecnologia de três áreas distintas, quais sejam: Análise e Desenvolvimento de Sistemas, Automação

Industrial e Logística os quais, no momento da extração dos dados, estavam com a matrícula ativa (regular ou trancada); formados ou desligados (sem vínculo com a instituição).

Apesar de o *Campus* Canoas também oferecer o Curso Superior de Licenciatura em Matemática, nesse primeiro momento os dados deste curso não serão incluídos no modelo proposto por este trabalho, por se entender que o perfil do aluno que busca a licenciatura é diferente dos que procuram as áreas de tecnologia, o que não impedirá de serem feitas análises com o modelo de mineração posteriormente.

7.2 **A ORIGEM DOS DADOS**

As informações utilizadas para o cálculo e análise dos números relativos às saídas com êxito e sem êxito dos alunos foram extraídas das planilhas do SRE do *Campus* Canoas.

Os dados relativos aos alunos foram extraídos do Sistema de Informação Acadêmica, conhecido como S.I.A, um dos sistemas de informações acadêmicas utilizados pelo IFRS e do SIFRS, sistema utilizado apenas no *Campus* Canoas, e exportados para o formato de planilha de dados. Eles correspondem as informações dos alunos que ingressaram nos três cursos superiores de tecnologia, no período entre 2011 e 2017.

Para identificar as principais variáveis que permitirão criar um modelo para predizer a propensão à evasão, foram inicialmente selecionados 42 itens; compreendendo informações de ingresso, de desempenho acadêmico, cadastrais, econômicas, familiares, sociais, de saúde, entre outras, coletadas no cadastramento para matrícula e, posteriormente, no preenchimento do Questionário Sociodemográfico, respondido pelos alunos, também no SIFRS. No decorrer da etapa de pré-processamento, foram incluídos na base de dados três atributos relativos ao desempenho e ao histórico acadêmico. As informações selecionadas guardam relação com os motivos citados para evasão na bibliografia estudada, mas, principalmente, no Plano Estratégico de Permanência e Êxito do IFRS. Dados como números de documentos, nomes dos pais, telefone, e-mail não foram considerados relevantes.

Após a etapa de pré-processamento dos dados, do processo de *KDD*, foram mantidos 37 itens, dispostos nas colunas da planilha de dados, denominados de atributos, os quais descrevem as características dos objetos dispostos nas linhas. Cada um dos objetos, nesta pesquisa, corresponde a um conjunto de informações de um aluno. Esses atributos farão parte da etapa de mineração de dados para criação do modelo de predição e estão identificados no Apêndice A, juntamente com o tipo, a descrição e os valores possíveis para cada um.

Dos 37 atributos selecionados e pré-processados, têm-se 35 atributos regulares e dois atributos especiais. Os atributos especiais são o número de matrícula, que corresponde ao atributo identificador, o qual fará apenas a distinção dos objetos; e a situação de aluno, que corresponde ao atributo classificador, ou de classe.

7.3 **O MODELO DE PREDIÇÃO**

Para criar o modelo de predição, foi utilizada uma técnica de classificação, pois a variável-alvo “a situação dos alunos” é discreta, ou seja, tem valores finitos e nominais (TAN et al., 2009, p. 9 e 176). De acordo com os valores do atributo classificador, que correspondem à situação da matrícula dos alunos no momento da extração dos dados, os objetos foram divididos em três classes:

a) FORMADO: concluíram o curso com êxito;

b) DESLIGADO: não concluíram o curso, não possuem vínculo com a instituição;

c) REGULAR: estão com matrícula ativa, mantém vínculo com a instituição.

Nesse momento, para atingir o objetivo desta pesquisa, o foco está na situação de saída dos alunos. Sendo assim, foi extraído do conjunto de dados um subconjunto, no qual os objetos têm como valor no atributo “Situação” apenas as informações Formado ou Desligado. Em outras palavras, o subconjunto foi formado pelas duas classes que representam as formas de saída dos cursos, com as quais o modelo foi treinado para predizer, com maior percentual de acurácia possível, a forma de saída daqueles alunos que estão regulares. As duas classes estão assim identificadas e caracterizadas:

a) FORMADO: classe negativa, indica as saídas com êxito; alunos que concluíram o curso;

b) DESLIGADO: classe positiva, indica as saídas sem êxito; alunos que não concluíram o curso.

A classe DESLIGADO foi considerada como positiva em função do objetivo do trabalho, que é identificar os alunos com propensão à evasão, a qual desejamos predizer com maior número de acertos.

Os objetos, cujo valor do atributo “Situação” é a informação Regular, objetos da classe Regular, formaram um segundo subconjunto de dados para validação do modelo de predição após seu treinamento, no qual foi feita a predição dos alunos com propensão à evasão. Antes da aplicação do modelo neste conjunto de dados, as informações da “Situação” dos alunos foram atualizadas com base nas informações da planilha “Alunos Cursos Superiores” (Figura 9), referentes ao início do segundo semestre de 2019.

Para a criação do modelo de predição, utilizando uma técnica de classificação, avaliou-se o desempenho de cinco algoritmos de classificação, implementados na ferramenta de mineração, que suportam a utilização de dados nominais e numéricos, equivalente ao conjunto de dados utilizado nesta dissertação, são eles: *Decision Tree, Random Forest, Naive Bayes, K-NN e Gradient Boosted Tree.* Estes algoritmos estão referendados como tendo bom desempenho em dados educacionais, pelos trabalhos relacionados e descritos na seção 3 e resumidos no Quadro 8. O desempenho foi avaliado em três experimentos e levou-se em consideração os resultados nas seguintes métricas: acurácia, precisão, sensibilidade, especificidade e valor preditivo negativo (VPN).

7.4 **A FERRAMENTA DE MINERAÇÃO DE DADOS**

A ferramenta escolhida para realizar as técnicas de mineração de dados é o *RapidMiner Studio Free 9.4.001*, um dos quatro produtos da plataforma *RapidMiner* (RAPIDMINER, c2010). O *RapidMiner Studio* possui um fluxo de trabalho com designer visual, em que os processos são construídos arrastando os operadores do seu repositório para área de *Design,* o que torna o trabalho mais produtivo. Os resultados são mostrados na área de resultados (*Results*), de acordo com as portas de saída dos operadores que estiverem conectadas às portas de resultados do painel de processos.

Ela foi escolhida por apresentar a possibilidade de uso gratuito e suportar todas as etapas de mineração, incluindo os resultados de visualização, validação e otimização. Para tanto, não é necessário saber programar, pois a ferramenta possui uma interface gráfica para o usuário com mais de 1.500 operadores e comporta o formato no qual os dados extraídos dos sistemas estão organizados.

8 O ESTUDO DE CASO: A CONSTRUÇÃO DO MODELO PREDITIVO

Neste capítulo são descritas todas as etapas do processo de Descoberta de Conhecimento em Banco de Dados (*KDD*), desenvolvidas, para atingir os objetivos propostos para este trabalho.

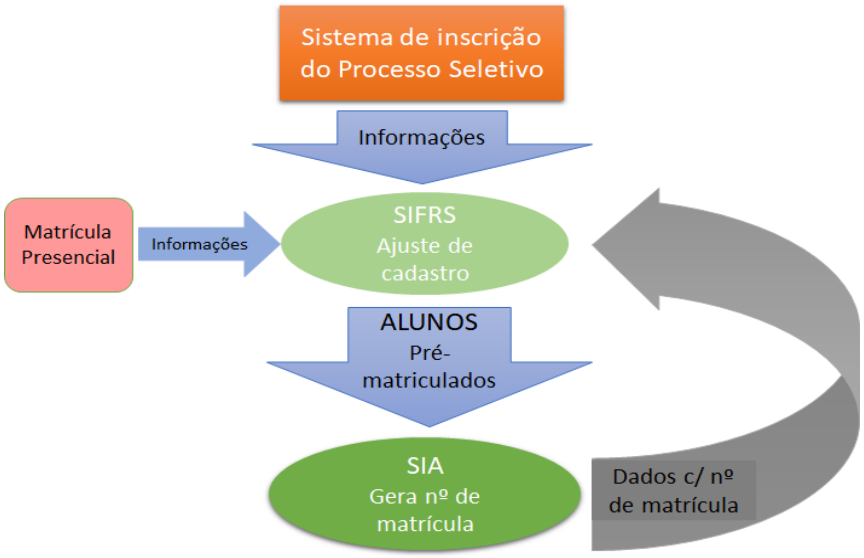
O primeiro passo é ter a definição e o conhecimento prévio da área que se deseja estudar e o que se quer descobrir empregando o processo de KDD, os quais foram estabelecidos com o desenvolvimento da primeira e da segunda parte metodológica deste trabalho. O conhecimento do domínio do estudo e das informações necessárias para delimitá-lo, a definição das subáreas das quais as informações mais relevantes são buscadas para compor a base de dados - qual seja, a evasão nos cursos de tecnologia do *Campus* Canoas do IFRS - foram elaborados no Capítulo 2. As próximas seções irão descrever as demais etapas do processo e as escolhas e decisões tomadas em cada uma.

8.1 SELEÇÃO E CONSTRUÇÃO DA BASE DE DADOS

Conforme já mencionado previamente, a origem dos dados utilizados por este estudo é do SIA e do sistema de apoio desenvolvido no *Campus* Canoas do IFRS, denominado SIFRS. Os dados informados pelos candidatos aprovados, no sistema de inscrição do processo seletivo da instituição, são importados no SIFRS e reaproveitados para o seu processo de pré-matrícula neste mesmo sistema.

Após todos os processos de pré-matrícula, os dados são exportados pelo SIFRS e importados no SIA para que os alunos sejam efetivamente matriculados e ganhem um número de matrícula. Estes números de matrícula são exportados pelo SIA e importados no SIFRS, para que as funcionalidades utilizadas no SIFRS tenham o número de matrícula do aluno relacionado. Este processo é ilustrado na Figura 11.

Figura 11 - Geração dos dados



Fonte: construção da autora, com base nas informações da TI *Campus* Canoas.

Uma das funcionalidades do SIFRS está em possibilitar o preenchimento e armazenamento das informações do Questionário Sociodemográfico (QS), aplicado a todos os alunos, sob a responsabilidade do setor de Assistência Estudantil do *Campus* Canoas. Este questionário é composto por 51 questões e tem como objetivo entender qual é o perfil do estudante matriculado. No início de cada ano letivo, alunos novos e antigos são estimulados a respondê-lo. Porém, como o preenchimento não é obrigatório, nem todos o fazem. As 51 questões estão distribuídas dentro de seis temas:

I. identificação;
II. sobre o grupo familiar dos educandos;
III. vida estudantil;
IV. participação social, cultural e esportiva;
V. sobre o IFRS e o curso;
VI. saúde.

Desta forma, foi possível a realização de uma consulta na base de dados do SIFRS, extraindo informações de alunos e relacionando-as com as respostas dadas ao QS preenchido. Sendo assim, tem-se, na Figura 12, as informações extraídas para compor a base de dados.

Figura 12 - Informações extraídas para compor a base de dados

1. Número de matrícula;
2. Situação da matrícula;
3. Data de nascimento;
4. Sexo;
5. Cor/raça/etnia;
6. Curso;
7. Motivo da escolha do curso;
8. Motivo da escolha do *campus;*
9. Matriz curricular;
10. Turno do curso;
11. Modo de ingresso (vestibular, Sisu ou Enem);
12. Tipo de cota que ingressou;
13. Tipo de instituição que fez o ensino médio (escola pública ou privada);
14. Ano de conclusão do ensino médio;
15. Tempo diário de estudo;
16. Conhecimentos de informática;
17. Nível de escolaridade do pai;
18. Nível de escolaridade da mãe;
19. Áreas de dificuldade na vida escolar;
20. Possui computador em casa;
21. Possui internet em casa;
22. Frequência de acesso à Internet;
23. Estado civil;
24. Número de filhos;
25. Renda per capita (ou renda familiar);
26. Número de pessoas que dependem da renda;
27. Realiza atividade remunerada (Quantas horas semanais?);
28. Recebe Assistência Estudantil;
29. Recebe bolsa de ensino, pesquisa ou extensão;
30. Recebe auxílio do governo;
31. Situação da moradia (própria, alugada, cedida);
32. Cidade onde mora;
33. Bairro onde mora;
34. Meio de transporte utilizado;
35. Atividade que ocupa o tempo além dos estudos;
36. Realiza atividades esportivas;
37. Quantos livros lê por ano;
38. Fonte de informação;
39. Possui necessidade especial;
40. Possui doença crônica;
41. Teve acompanhamento psicológico ou psiquiátrico;
42. Toma medicamento de uso controlado.

Fonte: construção da autora, com base nas informações da TI *Campus* Canoas.

Observa-se que, em acordo com os estudos sobre a evasão citados neste trabalho anteriormente, incluindo os trabalhos correlacionados utilizando mineração de dados, o maior percentual de saída da instituição ocorre no início do curso e em função do baixo desempenho acadêmico. Sendo assim, decidiu-se acrescentar dados acadêmicos dos alunos ao conjunto dos dados que serão minerados. Para tanto, extraiu-se do SIA a informação de desempenho em cada uma das disciplinas cursadas pelos alunos, em seu semestre de ingresso. Com exceção dos alunos transferidos e portadores de diploma, ao ingressar os alunos são compulsoriamente matriculados em todas as disciplinas constantes no primeiro semestre da matriz curricular do seu curso. A extração dessas informações do sistema acadêmico gerou uma nova planilha de dados, sendo necessário realizar a fusão com as informações extraídas do Questionário Sociodemográfico. Outro dado acrescentado foi a informação relativa ao trancamento do segundo semestre letivo do curso. Essa informação foi extraída da planilha Alunos Cursos Superiores (Figura 9) e, também, inserida na base de dados de forma manual.

8.2 **PRÉ-PROCESSAMENTO E CONSTRUÇÃO DO MODELO DE DADOS**

O pré-processamento dos dados, como apontam Tan et al. (TAN et al., 2009), caracterizou-se como uma fase bastante demorada, na qual o conhecimento da área estudada, da origem dos dados e dos valores possíveis e corretos foi bastante exigido. Vários procedimentos de processamento, descritos nesta seção, foram aplicados ao conjunto de dados, para que o formato dos mesmos se adequasse às técnicas da etapa de mineração e, principalmente, para que estes tenham relevância no que diz respeito à identificação da condição de saída do aluno da instituição.

Na primeira análise, foram encontrados objetos fora do período proposto para o estudo, ou seja, alunos que ingressaram em 2018, os quais foram retirados do conjunto de dados. Porém, o principal problema foi o número de informações faltantes em um número grande de objetos, devido ao não preenchimento do QS pelos alunos, gerando a ausência de muitos valores nos atributos.

A solução de preenchimento de valores ausentes com um valor padrão, o qual indicasse que os valores reais dos dados eram desconhecidos, não foi empregada, pois para cada aluno (objeto) o número de atributos com valores

faltantes equivalia a dois terços do total de atributos selecionados e, ainda, do total de objetos, cinquenta por cento apresentava valores ausentes. Sendo assim, optou-se por não correr o risco do algoritmo indutor assumir que o valor desconhecido representasse um conceito importante (FACELI, et al, 2011 e SILVA et al., 2016). Mesmo com uma perda grande de objetos, a escolha foi por preservar o maior número de atributos com informações completas, para que o modelo criado identificasse os alunos com propensão à evasão com maior precisão.

Ainda que tenha sido feita uma pré-seleção das informações mais relevantes do QS em relação aos fatores da evasão elencados no referencial teórico e no PEPEEIFRS, foi necessário eliminar outros atributos redundantes, irrelevantes e inconsistentes para evitar que a tarefa do algoritmo fosse dificultada (FACELI, et al, 2011). A descrição acerca dos atributos excluídos se encontra no Apêndice B.

Os demais problemas encontrados estavam na falta de padronização da escrita dos dados; informações redundantes entre si, que geram valores diferentes; informações conflitantes dentro de um mesmo atributo; atributos com muita variação de valores, e muitos valores com pouca importância para caracterizar o atributo e, ainda, alguns atributos com pouca relevância para caracterizar o objeto. Alguns exemplos:

- no atributo cuja informação é a cor, raça ou etnia, foram encontrados valores com flexão de gênero, preto e preta, no caso é uma redundância; o mesmo na informação do estado civil, constando valores como divorciado e separado, nos dois casos os valores foram normalizados;
- no atributo cuja informação é se teve ou tem acompanhamento psicológico, havia valores conflitantes como sim e não para o mesmo objeto;
- no atributo com a informação da prática de atividade física e desportiva, além do conflito entre sim e não, os valores mostram combinações de 23 possibilidades das formas de atividades física e desportivas apresentadas no QS, gerando um número grande de valores.

A limpeza, e a padronização desses dados, foi realizada para melhorar a qualidade dos mesmos, tornando-os mais apropriados para as técnicas de mineração de dados. Isso só foi possível pelo conhecimento dos dados e de quais valores estão dentro do possível e aceitável para cada atributo e de quais informações são significativas para identificar as características dos alunos.

Um bom exemplo disto foi a retirada dos tipos de atividades físicas e esportivas do atributo, exemplificado anteriormente, pois, neste caso, a informação mais importante é se o aluno realiza ou não atividade. Mantendo apenas a informação de “sim” e “não” foi retirado o excesso de dados e, com isso, o risco de ruídos, evitando, também, dados com baixa qualidade. O atributo transformado em binário, contendo apenas dois valores, possibilita se obter uma maior acurácia e, consequentemente, um modelo melhor.

A opção por incluir dados acadêmicos exigiu um trabalho cuidadoso e minucioso de retirada das informações de um outro conjunto de dados em forma de planilha, extraídos diretamente do SIA. Essa planilha apresenta três abas, uma para cada curso de tecnologia do *campus*, contendo cada uma delas as seguintes informações: curso, matrícula aluno, ano (ano de realização da disciplina), período (semestre do ano letivo em que a disciplina foi ofertada), turma, código da disciplina, disciplina, nota, resultado e situação atual (situação em que a matrícula do aluno se encontrava no momento da extração da informação do sistema acadêmico).

Figura 13 - Imagem da planilha com dados acadêmicos

	Curso	Matrícula	Aluno	Ano	Período	Turma	Código disciplina	Disciplina	Nota	Resultado	Situação atual
2	Sup. Tec. Logística	2050174	ASD	2014	2	LOG2014-2	20143	Gestão de Operações e Logística II	0	RF	Desligado - Evasão
3	Sup. Tec. Logística	2050174	ASD	2014	2	LOG2014-2	20160	Higiene e Segurança do Trabalho	8.7	APR	Desligado - Evasão
4	Sup. Tec. Logística	2050174	ASD	2014	2	LOG2014-2	20114	Inglês	8	APR	Desligado - Evasão
5	Sup. Tec. Logística	2050228	ATS	2016	1	LOG2016-1 IN	20136	Informática Aplicada	7.5	APR	Regular
6	Sup. Tec. Logística	2050228	ATS	2016	1	LOG2016-1 ÁL	20140	Álgebra Linear	9.4	APR	Regular
7	Sup. Tec. Logística	2050228	ATS	2016	1	LOG2016-1 PO	20097	Português Instrumental	7.9	APR	Regular
8	Sup. Tec. Logística	2050228	ATS	2016	1	LOG2016-1 ES	20135	Estatística	7.8	APR	Regular
9	Sup. Tec. Logística	2050228	ATS	2016	1	LOG2016-1 GO	20138	Gestão de Operações e Logística I	8.3	APR	Regular
10	Sup. Tec. Logística	2050228	ATS	2016	2	LOG2016-2 GO	20143	Gestão de Operações e Logística II	8.2	APR	Regular
11	Sup. Tec. Logística	2050228	ATS	2016	2	LOG2016-2 MA	20271	Matemática Financeira	8.1	APR	Regular
12	Sup. Tec. Logística	2050228	ATS	2016	2	LOG2016-2 MO	20141	Modelagem e Simulação	7.5	APR	Regular
13	Sup. Tec. Logística	2050228	ATS	2016	2	LOG2016-2 AD	20270	Administração	7.7	APR	Regular
14	Sup. Tec. Logística	2050228	ATS	2016	2	LOG2016-2 CA	20134	Cálculo	7.8	APR	Regular
15	Sup. Tec. Logística	2050228	ATS	2017	1	LOG2017-1 PE	20469	Pesquisa Operacional	8	APR	Regular
16	Sup. Tec. Logística	2050228	ATS	2017	1	LOG2017-1 DI	20162	Direito Aplicado à Logística	10	APR	Regular
17	Sup. Tec. Logística	2050228	ATS	2017	1	LOG2017-1 CO	20472	Contabilidade	8.3	APR	Regular
18	Sup. Tec. Logística	2050228	ATS	2017	1	LOG2017-1 GI	20471	Gestão da Qualidade	8.4	APR	Regular
19	Sup. Tec. Logística	2050228	ATS	2017	1	LOG2017-1 EC	20273	Economia	9	APR	Regular
20	Sup. Tec. Logística	2050228	ATS	2017	2	LOG2017-2 GO	20145	Gestão de Operações e Logística III	8.1	APR	Regular
21	Sup. Tec. Logística	2050228	ATS	2017	2	LOG2017-2 GI	20153	Gestão Socioambiental	9.1	APR	Regular

Análise de Sistemas | Logística | Automação Industrial

Fonte: construção da autora.

Como é possível analisar na Figura 13, as informações que interessam estavam nas colunas 8, 9 e 10 da planilha, mas para poder transpor os dados era preciso manter a identificação do aluno. Primeiramente, foram retiradas as colunas dispensáveis: curso, aluno, turma, código da disciplina e situação atual. O segundo passo foi filtrar apenas as disciplinas de primeiro semestre, usando filtro nas colunas Disciplina e Período. Com isso, foi possível identificar o aluno pelo número de

matrícula, as disciplinas e as correspondentes notas e resultados obtidos em cada uma delas. Pensando em uma maior segurança, esse conjunto de informações de cada aluno foi copiado e colado em uma nova aba. Para essa nova aba, foram transferidas apenas as informações dos alunos que já estavam selecionados como objetos na base de dados, identificados através do atributo “número de matrícula”.

Figura 14 - Imagem das informações selecionados da planilha dados acadêmicos

1	Matrícula	Ano	Period	Cód. disciplir	Disciplina	Nota	Resultado
10	2050133	2014	1	20138	Gestão de Operações e Logística I	8.1	APR
11	2050133	2014	1	20140	Álgebra Linear	3.4	REP
12	2050133	2014	1	20136	Informática Aplicada	7.8	APR
13	2050133	2014	1	20097	Português Instrumental	8	APR
14	2050133	2014	1	20135	Estatística	6.3	APR
20	2050133	2015	1	20140	Álgebra Linear	0	RF
59	2050303	2017	1	20097	Português Instrumental	8.1	APR
60	2050303	2017	1	20465	Gestão de Operações e Logística I	7.2	APR
77	2050176	2015	1	20140	Álgebra Linear	5.5	REP
78	2050176	2015	1	20097	Português Instrumental	8	APR
79	2050176	2015	1	20136	Informática Aplicada	6.8	APR
106	2050213	2015	1	20140	Álgebra Linear	1.8	RF
107	2050213	2015	1	20135	Estatística	6.6	APR
108	2050213	2015	1	20136	Informática Aplicada	7.4	APR
109	2050213	2015	1	20097	Português Instrumental	7.6	APR

Análise de Sistemas | Logística | Automação Industrial

Fonte: construção da autora.

Como pode ser visto na Figura 14, nas linhas grifadas com cores diferentes, nem todos os alunos cursaram o número máximo de disciplinas do primeiro semestre, que varia de cinco a seis, dependendo da matriz curricular de cada curso. Este fato ocorre em todos os cursos, pois os alunos transferidos e diplomados matriculam-se nas disciplinas com vaga disponível. Foi necessário um olhar atento, pois alunos que reprovaram no semestre de ingresso, refizeram a disciplina no primeiro semestre do ano seguinte (Figura 13). Por este motivo, a coluna do Ano (ano de realização da disciplina) exigiu especial atenção no momento da seleção.

O próximo passo foi decidir quais informações acadêmicas seriam mais consistentes e relevantes para serem usadas como atributos. Era possível usar as notas de cada disciplina do primeiro semestre dos cursos como atributo, mas isso acrescentaria, no mínimo, mais quinze atributos na base de dados, pois esta é

composta por alunos de três cursos. Esta ação aumentaria a dimensionalidade[11] do conjunto de dados e, ao mesmo tempo, haveria dados faltando nas disciplinas que não pertencem ao curso do aluno. Usar o coeficiente de rendimento (média das notas alcançadas nas disciplinas) não teria relevância, em função de não representar o número de aprovações, sucessos alcançados. Alunos com o mesmo coeficiente podem ter diferente número de aprovação nas disciplinas cursadas. Entende-se que usar o percentual de aproveitamento (percentual de aprovação em relação ao total das disciplinas cursadas) é a melhor opção, por representar o número de aprovações, sucesso obtido nas disciplinas, independente de nota.

Apenas o Resultado (aprovado, reprovado e reprovado por falta), obtido em cada disciplina cursada foi selecionado e integrado manualmente à base de dados para o cálculo do percentual de aproveitamento. Os resultados obtidos por cada aluno foram acrescentados ao conjunto de dados, utilizando os recursos de copiar e colar/transpor, fazendo a correspondência com o número de matrícula. Para o resultado de cada disciplina cursada, acrescentou-se uma coluna na base de dados, com uma das informações possíveis: reprovado (REP), aprovado (APR) ou reprovado por falta (RF). Com base nestas informações, dois novos atributos foram criados, um contendo o número de disciplinas cursadas e o outro contendo o percentual de aproveitamento em relação às disciplinas cursadas. Esse processo encontra-se esquematizado pela Figura 15, nos dados marcados em azul.

Figura 15 – Transformação e criação de novos atributos

	Turno	Data de nascimento	Idade	Faixa etária	Sexo	Matriz	Estado civil	Cor/Raça /Etnia	Forma ingresso	Modalidade de Ingresso	Período	Dic 1	Disc 2	Disc 3	Disci 4	Disci 5	Disc. 6	Nº disc. Cursadas	% aproveitamento	Ano Ingresso	Ano conclusão nível anterio	Tempo de conclusão ens. Médio	TCNAE	
2	Noite	19/04/1988	31	4- de 30 a 39 anos	M	20503	Solteiro	Preta	SiSU	C9	2016/1	APR	APR	APR	APR	APR		5	100	2016	2005	11	>8 e ≤12	1-
3	Noite	04/09/1978	41	5- de 40 a 49 anos	M	20503	Casado	Branca	Transferência	NA	2015/1	REP	APR	APR	RF			4	50	2015	1998	17	>12	2-
4	Manhã	03/03/1999	20	1- de 18 a 19 anos	M	20802	Solteiro	Branca	SiSU	C1	2017/1	APR	APR	APR	APR	APR	APR	6	100	2017	2016	1	≤1	1-
5	Manhã	18/03/1999	20	1- de 18 a 19 anos	M	20802	Solteiro	Branca	SiSU	C1	2017/1	APR	APR	APR	APR	APR	APR	6	100	2017	2016	1	≤1	1-
6	Noite	15/10/1990	29	3- de 25 a 29 anos	M	209	Solteiro	Branca	PS IFRS	C1	2016/1	APR	APR	APR	APR	APR		5	100	2016	2007	9	>8 e ≤12	1-
7	Noite	08/09/1995	24	2- de 20 a 24 anos	M	209	solteiro	ND	PS IFRS	C9	2013-1	APR	APR	APR	APR	APR		5	100	2013	2012	1	≤1	1-
8	Manhã	14/01/1999	21	1- de 18 a 19 anos	F	20802	Solteiro	Branca	PS IFRS	C5	2017/1	APR	APR	APR	APR	APR	REP	6	83	2017	2016	1	≤1	1-
9	Manhã	31/12/1996	23	2- de 20 a 24 anos	F	20802	Solteiro	ND	SiSU	C1	2016/1	APR	APR	APR	REP	RF		5	60	2016	2015	1	≤1	1-
10	Manhã	02/05/1993	26	3- de 25 a 29 anos	M	20801	Solteiro	Branca	SiSU	C1	2014/1	APR	APR	APR	REP	APR		4	80	2014	2010	4	>2 e ≤4	2-

Fonte: construção da autora.

[11] Para alguns algoritmos a complexidade computacional cresce à medida em que a dimensionalidade aumenta, ou seja, são acrescentadas novas características aos objetos. (TAN et al., 2009, pag. 6)

Para o cálculo do Percentual de Aproveitamento (PA) foi utilizado o processo de regra de três simples, considerando a aprovação em todas as disciplinas que o estudante estava matriculado com 100% de aproveitamento. O cálculo pode ser representado da seguinte forma:

$$PA = \frac{100tda}{tdm}$$

Onde “tdm” é o número de disciplinas em que o estudante esteve matriculado e “tda” é o número de disciplinas em que alcançou aprovação.

Outras transformações foram realizadas para a criação de novos atributos com dados mais relevantes para essa pesquisa. Dois destes atributos estão assinalados na Figura 15. Na área marcada em verde (coluna 2, 3 e 4 da Figura 15), tem-se a informação de data de nascimento, que foi transformada em idade e depois alocada em faixas de idade pré-estabelecidas. Nas colunas em cinza (colunas 20, 21 e 22 da Figura 15), tem-se a transformação do “Ano de conclusão do nível anterior” para “Tempo de conclusão do nível anterior”, alocado em faixas de tempo, na coluna 23. No item ano de ingresso, existia a data de ingresso, da qual foi retirado o dia e o mês para que fosse possível realizar a subtração do ano de conclusão do nível anterior, resultando no tempo de conclusão do nível anterior.

Acrescentou-se, ainda, o dado sobre trancamento de matrícula, e a opção foi por observar essa situação no segundo semestre, por ser o momento mais precoce que o aluno pode solicitar esta condição para sua situação acadêmica, uma vez que se pretende identificar, de forma mais prematura, o início do processo de abandono. O valor “sim” para trancou o segundo semestre e “não” para não trancou o segundo semestre foi acrescentado individualmente a cada aluno (objeto, linha no conjunto de dados), correspondente à coluna do novo atributo “trancamento”. A informação sobre ter trancado ou não o segundo semestre, foi extraída da planilha “Alunos Cursos Superiores” (Figura 9). Foram considerados trancamentos por solicitação do aluno e trancamentos automáticos.

Observa-se que todas as alterações realizadas na base de dados, quais sejam: exclusão de objetos, inclusão, exclusão e transformação de atributos estão descritas no Apêndice B.

Na última parte do pré-processamento foram definidos os atributos especiais, aquele com valores que identificam os objetos, atributo identificador, ID, e o atributo

que identifica as classes, os quais servirão para o treinamento do algoritmo, atributo de classe ou de saída.

De acordo com o objetivo deste trabalho, que é criar um modelo de predição, ou seja, treinar um algoritmo classificador com as características distribuídas nos atributos para identificar os alunos com propensão à evasão, tem-se como alvo as situações de saída dos alunos da instituição. Na base de dados, a informação sobre a forma de saída dos alunos está no atributo “Situação”. Desta forma, este é o atributo alvo ou classificador. No entanto, os dados registrados originalmente neste atributo possuem 7 valores diferentes, nominais, com nomenclatura utilizada no S.I.A., são eles: Formados; Regulares, Desligado-Desistência, Desligado-Evasão, Desligado-Jubilado e Desligado-Transferência e Trancado Total. Para obter o melhor desempenho na predição, os valores foram agrupados em três classes:

a) REGULAR: inclui os valores Regular e Trancado Total. Em relação à característica do objeto, representa a situação do aluno que está com matrícula ativa, ainda cursando;
b) FORMADO: valor Formado. Em relação à característica do objeto, representa a situação do aluno que concluiu o curso;
c) DESLIGADO: inclui os valores Desligado-Desistência, Desligado-Evasão, Desligado-Jubilado e Desligado-Transferência. Em relação à característica do objeto, representa a situação do aluno que saiu do curso sem concluí-lo.

Como apontado anteriormente, o foco desta pesquisa está nas situações de saída do curso. Em virtude disso, para a criação do modelo de predição foi extraído da base de dados um subgrupo dos dados, contendo objetos das classes Formado e Desligado, por possuírem, respectivamente, características relacionadas aos alunos que concluíram o curso com êxito e aos alunos que não concluíram, sendo estes últimos o objeto desta investigação. Além disto, estas duas situações encontram-se definidas, ao passo que os alunos em situação Regular estão ainda cursando e pertencem ao grupo do qual se deseja predizer a situação futura, ou seja, a forma como irão sair da instituição.

O trabalho de seleção, limpeza, organização dos dados, exclusão, inclusão e criação de atributos, realizado nesta fase de pré-processamento, resultou num conjunto de dados composto por 37 (trinta e sete) atributos e 420 (quatrocentos e

vinte) objetos. O conjunto de atributos, o tipo, os valores possíveis e a explicação para cada um deles encontram-se organizados em um quadro denominado Modelo de Dados, que se encontra disponível no Apêndice A.

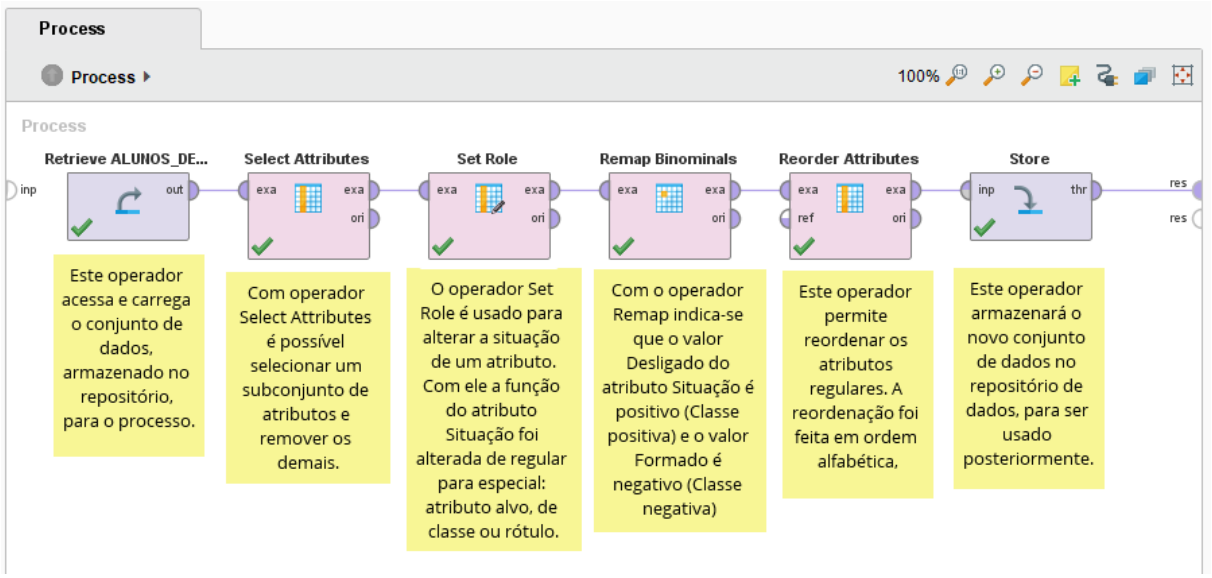
Com os dados organizados, passou-se à etapa de mineração dos dados do processo de *KDD*.

8.3 **MINERAÇÃO DE DADOS**

É nesta etapa que os dados são processados por algoritmos sofisticados para que deles sejam extraídos padrões úteis ou prever um resultado futuro. No caso desta dissertação queremos prever, ou predizer, os alunos com propensão à evasão, usando dados de alunos que concluíram o curso com sucesso (Formado) ou saíram da instituição sem concluí-lo (Desligado). Desta forma, usaremos um método de classificação para construir um modelo preditivo que melhor se adapte ao conjunto de dados.

A construção do modelo foi realizada com a ferramenta de mineração de dados *RapidMiner Studio*. Após a importação do arquivo de dados para o *RapidMiner*, um processo (Figura 16) foi organizado para a seleção dos atributos, indicação do atributo-alvo, da classe positiva e da classe negativa e ordenação dos atributos. Três atributos foram retirados do conjunto de dados neste momento: o ID (número de matrícula), o atributo especial de identificação dos objetos, que não é utilizado para treinamento e teste do modelo; os atributos Nº Disciplina (número de disciplinas que estão matriculados no 1º semestre) e Turno_Curso, pois ambos estão redundantes, fazendo relação direta com os atributos %Aprov (percentual de aprovação no 1º semestre) e Curso, respectivamente.

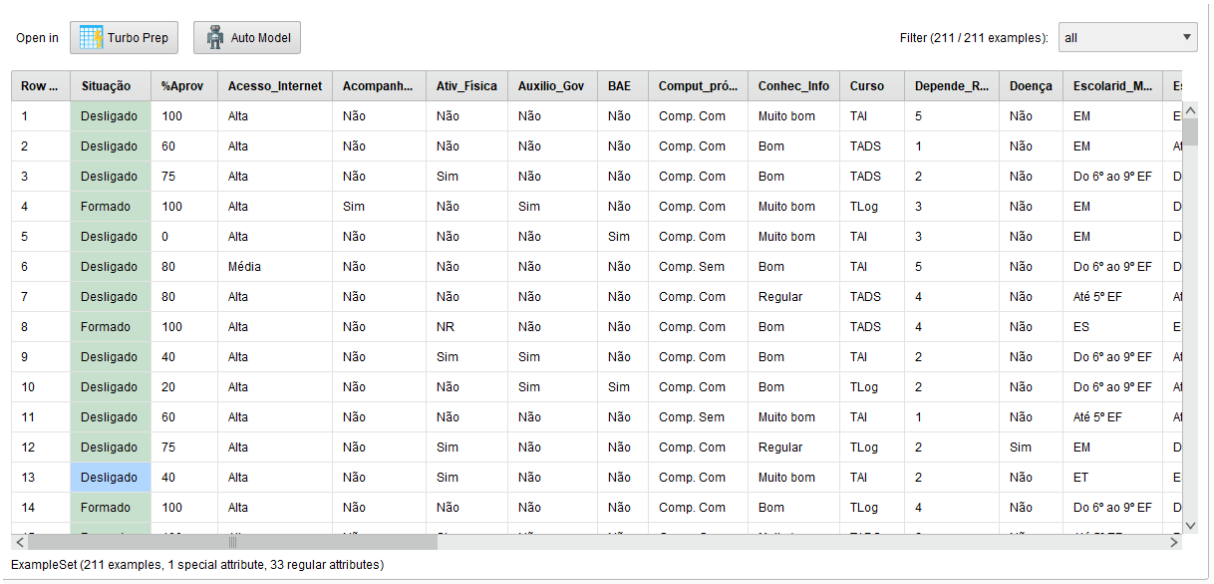
Figura 16 - Processo de ajuste do conjunto de dados no *RapidMiner*



Fonte: construção da autora.

O resultado é um novo conjunto de dados, com 33 atributos regulares e 1 atributo especial, como ilustra a Figura 17. Este novo conjunto foi usado para treinar o modelo de predição.

Figura 17 - Conjunto de dados “DESLIGADO_FORMADO_PROCESSADO”



Row ...	Situação	%Aprov	Acesso_Internet	Acompanh...	Ativ_Física	Auxilio_Gov	BAE	Comput_pró...	Conhec_Info	Curso	Depende_R...	Doença	Escolarid_M...
1	Desligado	100	Alta	Não	Não	Não	Não	Comp. Com	Muito bom	TAI	5	Não	EM
2	Desligado	60	Alta	Não	Não	Não	Não	Comp. Com	Bom	TADS	1	Não	EM
3	Desligado	75	Alta	Não	Sim	Não	Não	Comp. Com	Bom	TADS	2	Não	Do 6º ao 9º EF
4	Formado	100	Alta	Sim	Não	Sim	Não	Comp. Com	Muito bom	TLog	3	Não	EM
5	Desligado	0	Alta	Não	Não	Não	Sim	Comp. Com	Muito bom	TAI	3	Não	EM
6	Desligado	80	Média	Não	Não	Não	Não	Comp. Sem	Bom	TAI	5	Não	Do 6º ao 9º EF
7	Desligado	80	Alta	Não	Não	Não	Não	Comp. Com	Regular	TADS	4	Não	Até 5º EF
8	Formado	100	Alta	Não	NR	Não	Não	Comp. Com	Bom	TADS	4	Não	ES
9	Desligado	40	Alta	Não	Sim	Sim	Não	Comp. Com	Bom	TAI	2	Não	Do 6º ao 9º EF
10	Desligado	20	Alta	Não	Não	Sim	Sim	Comp. Com	Bom	TLog	2	Não	Do 6º ao 9º EF
11	Desligado	60	Alta	Não	Não	Não	Não	Comp. Sem	Muito bom	TAI	1	Não	Até 5º EF
12	Desligado	75	Alta	Não	Sim	Não	Não	Comp. Com	Regular	TLog	2	Sim	EM
13	Desligado	40	Alta	Não	Sim	Não	Não	Comp. Com	Muito bom	TAI	2	Não	ET
14	Formado	100	Alta	Não	Não	Não	Não	Comp. Com	Bom	TLog	4	Não	Do 6º ao 9º EF

Fonte: construção da autora.

8.3.1 Treinamento e teste do modelo

O objetivo desta etapa é construir um modelo que obtenha um bom resultado de predição e, ao mesmo tempo, seja de fácil entendimento para os profissionais da área da educação.

A seleção dos algoritmos de aprendizagem, entre os demais disponíveis no *RapidMiner*, levou em consideração o objetivo do trabalho, que é a predição, e as características do conjunto de dados, o qual possui o valor do atributo-alvo do tipo nominal e binário e os demais atributos dos tipos nominal e numérico. Os cinco algoritmos classificadores selecionados e descritos a seguir estão referenciados entre os que obtiveram melhor performance nas pesquisas relacionadas, elencadas na seção 3, os quais aparecem discriminados de forma resumida no Quadro 8.

Os modelos selecionados estão brevemente descritos a seguir e informações adicionais, para maior conhecimento, podem ser encontradas em TAN et al. (2009), Camilo e Silva (2009), FACELI, et al. (2011) e Silva et al. (2016):

- ***Decision Tree (DT)****:* Árvore de Decisão é uma das técnicas mais populares de mineração. O algoritmo impõe uma regra, um teste sobre cada atributo, os quais formam os nós internos da árvore, que levam a uma tomada de decisão, separando os objetos que possuem características diferentes, que são representados nas folhas. Ele é um modelo de fácil entendimento.
- ***Random Forest (RF)***: Florestas Aleatórias ou Florestas de Decisão Aleatória. O modelo gera várias árvores de decisão, cujas previsões são combinadas por votação uniforme. Cada árvore é induzida a partir de uma amostra com reposição do conjunto de treinamento.
- ***k-nearest neighbors algorithm (k-NN)***: é um método não paramétrico usado para classificação e regressão. Baseado em distância, é conhecido como o algoritmo dos vizinhos mais próximos. Seu processo de aprendizado consiste em memorizar os objetos de treino e encontrar aqueles mais próximos (semelhantes) ao objeto de teste, o rótulo de classe desses vizinhos é atribuída a este objeto.
- ***Gradient Boosted Tree (GBT)***: Árvores Impulsionadas por *Gradient.* Um modelo impulsionado por gradiente (*gradient boosted*) é um conjunto de modelos de árvore de regressão ou classificação. O

impulso é um procedimento de regressão não linear flexível que ajuda a melhorar a precisão das árvores.

- ***Naive Bayes (NB)***: esse modelo pertence a uma família de classificadores probabilísticos simples, baseados na aplicação do teorema de Bayes. Utiliza um princípio estatístico para combinar conhecimento prévio das classes com novas evidências colhidas dos dados.

De acordo com TAN et al. (2009) “A avaliação do desempenho de um modelo de classificação é baseada nas contagens de registros de testes previstos correta e incorretamente pelo modelo”. Estas contagens de acertos e erros para cada classe são organizadas em uma tabela conhecida como Matriz de Confusão. A Figura 18 mostra o modelo de Matriz de Confusão com a posição das classes verdadeiras e preditas, de acordo com a ferramenta *RapidMiner Studio.*

Figura 18 - Matriz de confusão para duas classes

MATRIZ DE CONFUSÃO		**Classes Verdadeiras**	
		Classe (-) FORMADO	Classe (+) DESLIGADO
Classes Preditas	Classe (-) FORMADO	VN	FN
	Classe (+) DESLIGADO	FP	VP

Fonte: construção da autora.

Cada célula da Matriz de Confusão possui um significado. Silva et al. (2016), os descreve da seguinte forma:

- **VN - Verdadeiro Negativo:** classificação correta na classe negativa. O exemplar pertence à classe negativa, e o classificador o classificou como pertencente à classe negativa.
- **FN - Falso Negativo:** classificação incorreta na classe negativa. O exemplar pertence à classe positiva, mas o classificador o classificou como pertencente à classe negativa.

- **VP - Verdadeiro Positivo**: classificação correta na classe positiva. O exemplar pertence à classe positiva, e o classificador o classificou como pertencente à classe positiva.
- **FP - Falso Positivo:** classificação incorreta na classe positiva. O exemplar pertence à classe negativa, mas o classificador o classificou como pertencente à classe positiva.

Para avaliar o desempenho dos algoritmos de classificação e selecionar aquele que melhor se adapte ao conjunto de dados de treinamento foram realizados três experimentos. O Quadro 9 mostra como cada medida de desempenho, utilizada nos experimentos para avaliar os classificadores, é calculada sobre os resultados da Matriz de Confusão, gerada no treinamento e teste de cada modelo, de acordo com os tutoriais do *RapidMiner Studio*.

Quadro 9 - Cálculo das medidas de desempenho, utilizadas para avaliação dos modelos classificadores.

MEDIDA	DESCRIÇÃO DO CÁLCULO
ACURÁCIA	(previsões corretas) / (número de exemplos) = (VP + VN) / (VP + FP + FN + VN)
PRECISÃO	(previsões Verdadeiras Positivas) / (todas previsões positivas) = VP / (VP + FP)
SENSIBILIDADE = RECALL	(previsões Verdadeiras Positivas) / (nº de exemplos positivos) = VP / (VP + FN)
ESPECIFICIDADE	(previsões Verdadeiras Negativas) / (nº de exemplos negativos) = VN / (VN + FP)
VALOR PREDITIVO NEGATIVO (VPN)	(previsões Negativas Verdadeiras) / (todas as previsões negativas) = VN / (VN + FN)

Fonte: construção da autora.

8.3.1.1 Experimento 1

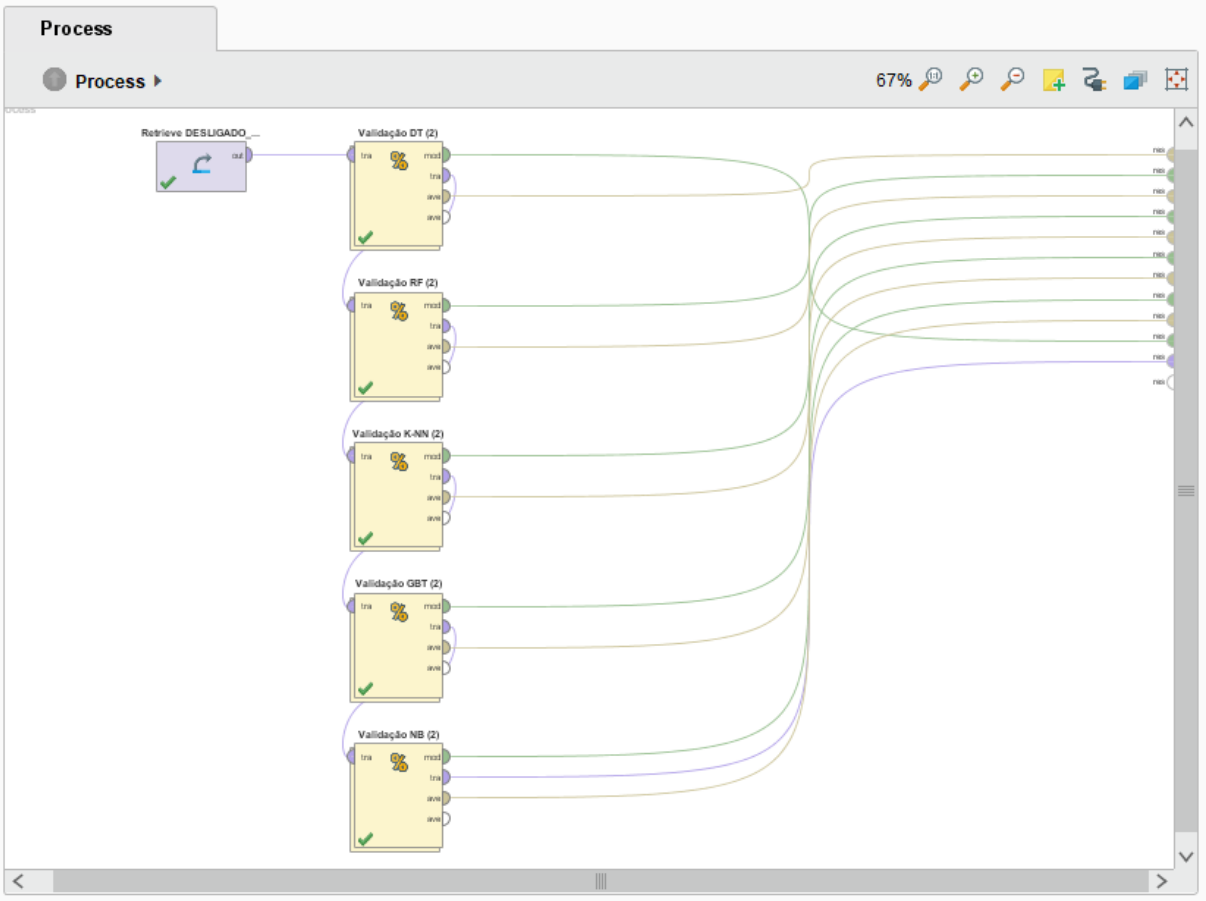
O primeiro experimento foi realizado na área de designer do *RapidMiner Studio* utilizando o operador *Split Validation.* Esse operador executa uma validação simples, ou seja, ele divide aleatoriamente o conjunto de dados em um conjunto de treinamento e conjunto de teste para avaliar o modelo. A configuração dos parâmetros deste operador foi ajustada para que o conjunto de treino contivesse

70% dos objetos e o conjunto de teste 30%. Além disso, a seleção destes objetos foi configurada como estratificada, garantindo que, neste caso em que a classificação é binominal, cada subconjunto contenha aproximadamente as mesmas proporções dos dois valores dos rótulos de classe.

O processo organizado na área de designer do *RapidMiner*, para aplicação do operador *Split Validation,* com cada um dos classificadores (*Decision Tree, Random Forest, K-NN, Gradient Boosted Trees e Naive Bayes)*, utilizou o arquivo de dados descrito na seção 5.3, contendo o conjunto de dados dos 211 alunos desligados e formados.

As validações de cada classificador foram organizadas de forma idêntica e por opção, dispostas no mesmo processo na tela de designer, porém a ferramenta executa uma de cada vez e mostra os resultados de forma individualizada. As Figuras 19 e 20 mostram o designer do processo.

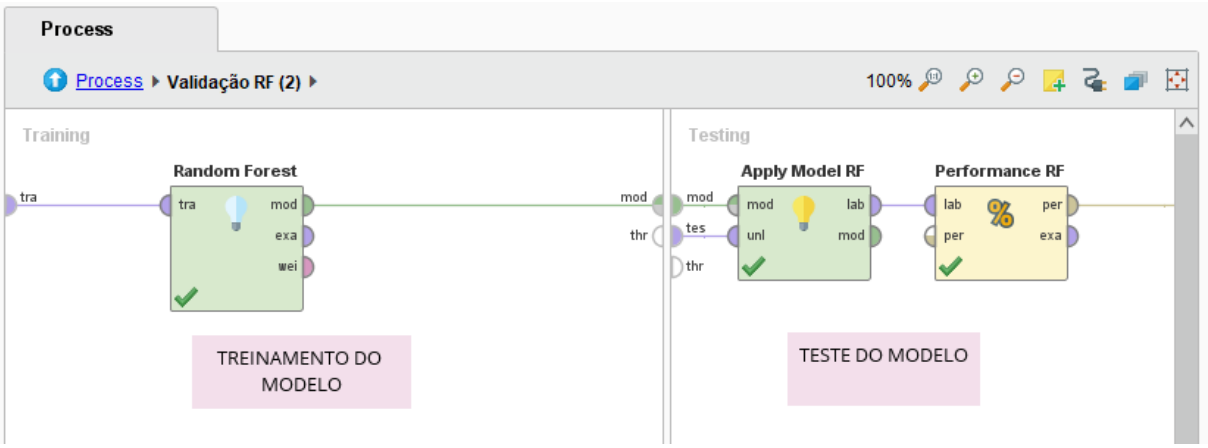
Figura 19 - Processo *Split Validation* com os cinco classificadores



Fonte: construção da autora.

O operador *Split Validation* é um operador aninhado, ou seja, comporta dois subprocessos, o treino e o teste do modelo. O primeiro denominado subprocesso representa o lugar em que ocorre o aprendizado ou a construção do modelo no conjunto de dados de treinamento. O modelo treinado é então aplicado no conjunto de teste, no segundo subprocesso. O desempenho do modelo é medido durante a fase de teste. Os dois subprocessos com o classificador *Random Forest* e os demais operadores são mostrados na Figura 20.

Figura 20 - Subprocessos do operador *Split Validation* com o classificador *Random Forest*



Fonte: construção da autora.

O operador *Apply Model* aplica o modelo treinado nos dados de teste e o *Performance Binominal Classification* mostra o resultado do desempenho do teste do modelo.

8.3.1.1.1 Avaliação dos resultados

Os resultados das métricas de desempenho dos classificadores, selecionadas através do operador *Performance Binomial Classification,* foram organizados no Quadro 10 e demonstram que, neste experimento, os classificadores *Random Forest* e *K-NN* tiveram melhor desempenho em relação aos demais. O *RF* atingiu 92,68% na medida de sensibilidade. Isso significa dizer que ele teve um percentual de erros menor que 10% na predição da classe positiva. A Matriz de Confusão do *RF* mostra que dos 41 exemplos da classe Desligados ele classificou apenas 3 como Formados, porém teve um desempenho baixo na predição da classe negativa, apresentando um percentual de especificidade de 36,36%. Já o *K-NN* mostrou resultados mais equilibrados entre a acurácia, a

sensibilidade e a especificidade, com a média 76,3 % entre elas, ou seja, uma predição melhor e equilibrada nas duas classes.

Quadro 10 - Métricas atingidas pelos classificadores com a *Split Validation*

SPLIT VALIDATION						
Modelos classificadores	Acurácia %	Precisão %	Sensibilidade %	Especificidade %	VPN %	MATRIZ DE CONFUSÃO
DT	66,67	73,81	75,61	50,00	52,38	11 10 11 31
RF	73,02	73,08	92,68	36,36	72,73	8 3 14 38
K-NN	76,19	86,11	75,61	77,27	62,96	17 10 5 31
GBT	69,84	77,5	75,61	59,09	56,52	13 10 9 31
NB	73,02	77,27	81,93	54,55	63,16	12 7 10 34

Fonte: construção da autora.

O quadro também mostra a Matriz de Confusão de cada um dos classificadores e apresenta, em destaque, as melhores performances nas métricas.

8.3.1.2 Experimento 2

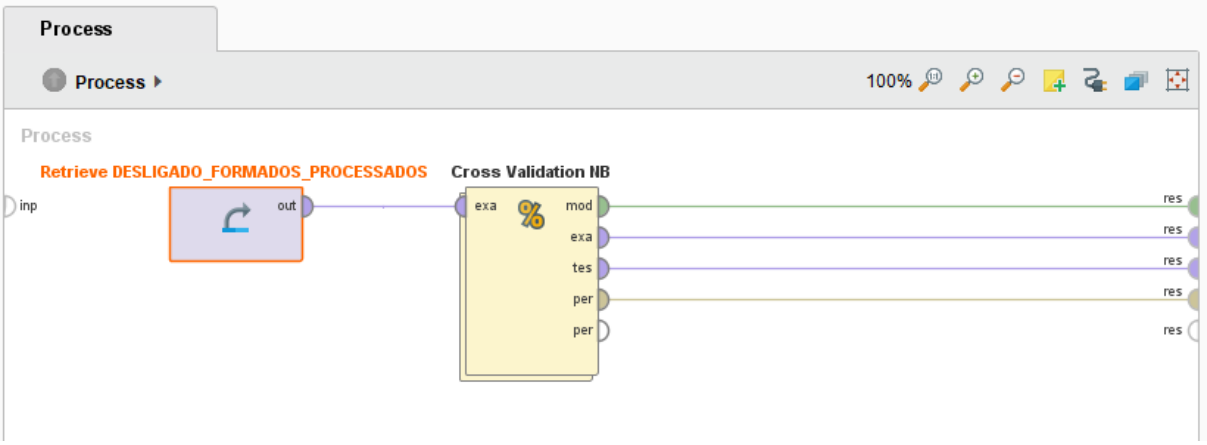
O segundo experimento realizado na área de designer do *RapidMiner Studio* utilizou o operador *Cross Validation.* Este operador executa uma validação cruzada, ou seja, divide o conjunto de dados para treinamento em subconjuntos e, em seguida, para teste de forma independente. O modelo treinado em cada um dos subconjuntos de treinamento é aplicado aos subconjuntos de teste. Os resultados para todos os conjuntos de treino e teste são coletados pelo operador de validação cruzada e a média é construída e entregue como resultado, o que garante uma melhor estimativa de desempenho.

O operador *Cross Validation* foi configurado para dividir os dados em 10 subconjuntos (10 *folds*). Cada subconjunto tem número igual de objetos. Além disso, o número de interações que ocorreram é o mesmo de *folds*. A seleção dos objetos

foi configurada de forma estratificada e os cinco classificadores foram validados um de cada vez.

Neste experimento, os processos foram organizados na área de designer do *RapidMiner* de forma individual para cada um dos cinco classificadores utilizados no experimento anterior, assim como o arquivo de dados descrito na seção 5.3, contendo o conjunto de dados dos 211 alunos desligados e formados. As Figuras 21 e 22 mostram o designer do processo, realizado com o classificador *Naive Bayes*.

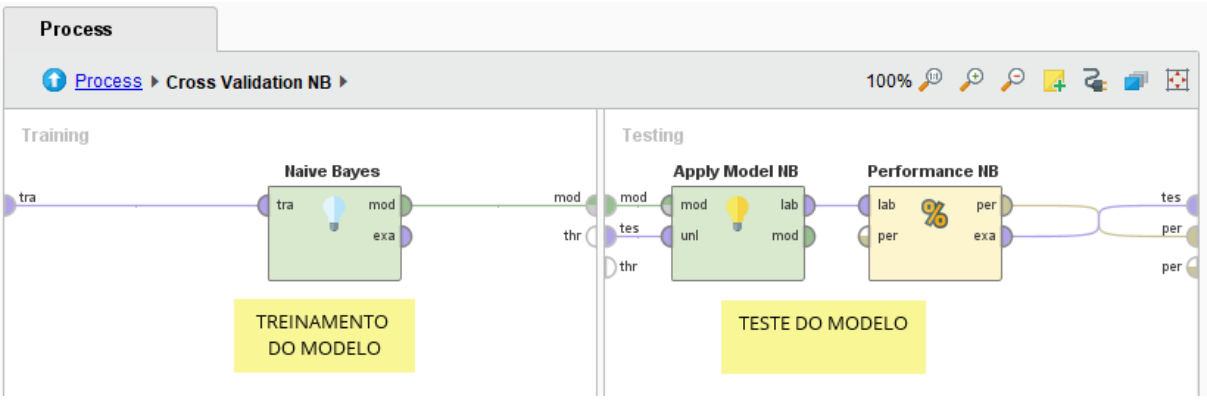
Figura 21 - Processo *Cross Validation*



Fonte: construção da autora.

O operador *Cross Validation*, assim como *Split Validation,* é um operador aninhado e também abriga os subprocessos de treino e teste do modelo, conforme ilustra a Figura 22.

Figura 22 - Subprocessos do operador *Cross Validation* com o classificador *Naive Bayes*



Fonte: construção da autora.

As métricas acurácias, precisão, sensibilidade especificidade e valor preditivo negativo foram selecionadas no operador *Performance Binominal Classification* e os resultados encontram-se no Quadro 11.

8.3.1.2.1 Avaliação dos resultados

Novamente o classificador *Random Forest* teve melhor desempenho, melhorando a acurácia em relação ao experimento anterior e mantendo os percentuais de sensibilidade e de VPN mais altos que os demais classificadores, ou seja, teve o menor número de erros na predição da classe positiva. O Classificador *K-NN* fez a melhor predição para a classe negativa, com 70,67% de especificidade. Contudo, ao se observar as Matrizes de Confusão, constatou-se que ele teve o segundo maior número de erros na classe positiva. O *Decision Tree* e o *Gradient Boosted Tree* apresentaram boa precisão e sensibilidade, conduzindo a um menor número de erros na classe positiva em relação ao *K-NN* e o *NB* e aos erros na classe negativa (Quadro 11).

Quadro 11 - Métricas atingidas pelos classificadores com a *Cross Validation*

CROSS VALIDATION						
Modelos classificadores	Acurácia %	Precisão %	Sensibilidade %	Especificidade %	VPN %	MATRIZ DE CONFUSÃO
DT	73,83	80,05	80,88	61,33	63,89	46 26 29 110
RF	76,23	79,95	84,56	61,33	68,66	46 21 29 115
K-NN	74,31	82,67	76,47	70,67	62,35	53 32 22 104
GBT	75,28	80,79	81,62	64,00	65,75	48 25 27 111
NB	71,95	80,71	75,00	66,67	59,52	50 34 25 102

Fonte: construção da autora.

A Matriz de Confusão também é mostrada no quadro, com os valores de acertos e erros em cada uma das classes e, em destaque, os melhores desempenhos.

8.3.1.3 Experimento 3

Nos experimentos anteriores foram usados os valores padrão (Quadro 12) do *RapidMiner* para configuração dos parâmetros de todos os classificadores selecionados. Neste terceiro experimento, usaremos o operador *Optimize Parameters (Grid)* para ajustar os parâmetros de cada classificador durante a validação cruzada e melhorar seus desempenhos. Este operador executa a validação cruzada para cada um dos classificadores, tantas vezes quantas forem as combinações possíveis entre os parâmetros selecionados, até que a melhor seja encontrada.

A seleção do valor ideal dos parâmetros é baseada no valor de desempenho dos operadores selecionados. Para este experimento, criaram-se processos individuais, aninhando o processo de validação cruzada no operador *Optimize Parameters (Grid)* e selecionaram-se os parâmetros dos algoritmos classificadores para otimização. Os valores padrão dos parâmetros dos algoritmos de aprendizagem, as faixas dos parâmetros para otimização, os valores de parâmetros otimizados e o número de combinações são mostrados no Quadro 12.

Quadro 12 - Parâmetros selecionados para otimização

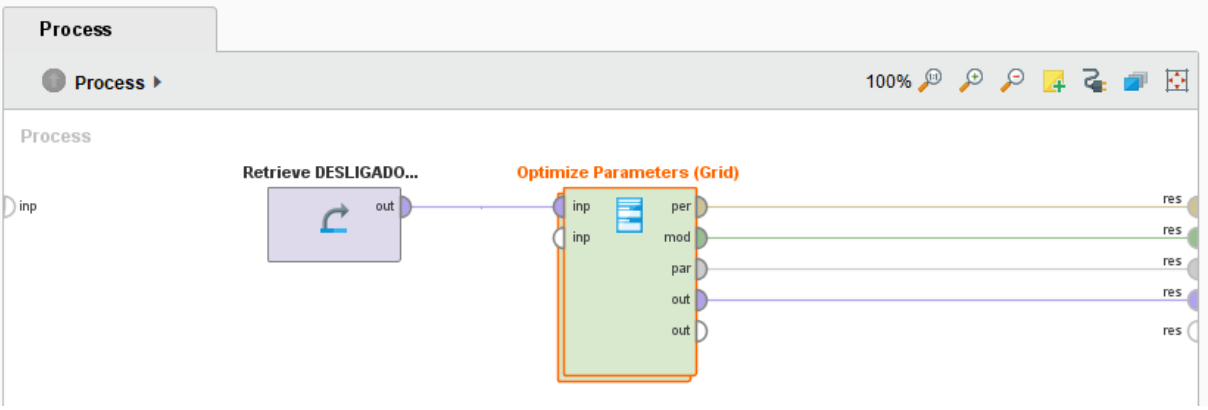
PARÂMETROS OTIMIZADOS					
Algoritmos	**Parâmetros selecionados**	**Parametros padrão do *RapidMiner***	**Faixa selecionada p/ otimização**	**Parâmetros selecionados**	**Nº de combinações**
DT	Critério	Taxa de ganho	Taxa de ganho, ganho de informação, índice gini e acurácia	Taxa de ganho	43560
	Tamanho mínimo de folha	2	Min. 1 Max. 10	4	
	Profundidade máxima	10	Min. 2 Max. 10	9	
	Ganho mínimo	0.01	Min. 0.0 Max. 0.01	0.006	
	Confiança	0.1	Min. 1.0E-7 Max. 0.5	0.300	
RF	Nº de árvores	100	Min. 10 Max. 100	10	4840
	Critério	Taxa de ganho	Taxa de ganho, ganho de informação, índice gini e acurácia	Taxa de ganho	
	Tamanho mínimo de folha	2	Min. 1 Max. 10	3	
	Ganho mínimo	0.01	Min. 0,01 Max. 10	3.007	

K-NN	K-NN.K	5	Min. 1.0 Max. 10	9	10
GBT	Nº de árvores	100	Min. 1.0 Max. 100	80	1331
	Linhas mínimas	10.0	Min. 4.9E-324 Max. 10	4.0	
NB	Correção de Laplace	Não se aplica	Verdadeiro, falso	false	2

Fonte: construção da autora, com base nos tutoriais do *RapidMiner Studio 9.4.001*.

O processo para otimização de cada um dos classificadores foi organizado na área de designer do *RapidMiner,* utilizando o operador *Optimize Parameters (Grid)*. Como este também é um operador aninhado, dentro dele foi colocado o operador *Cross Validation*, com seus subprocessos organizados da mesma forma que no experimento anterior e utilizando o mesmo conjunto de dados. Cada processo de validação otimizada foi organizado alterando apenas os operadores correspondentes aos algoritmos de classificação. As Figuras 23, 24 e 25 mostram o designer do processo e dos subprocessos aninhados.

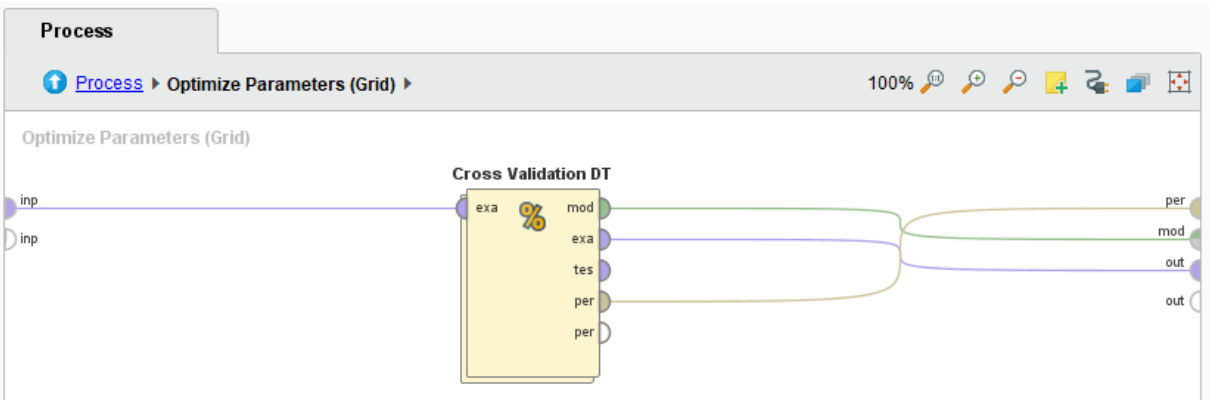
Figura 23 - Processo de otimização com operador *Optimize Parameters (Grid)*



Fonte: construção da autora.

As portas com os resultados de performance (per), modelo (mod), parâmetros (par) e de saída (out) do conjunto de dados foram ligadas às portas de resultados (res) para que estes sejam mostrados na tela de resultados.

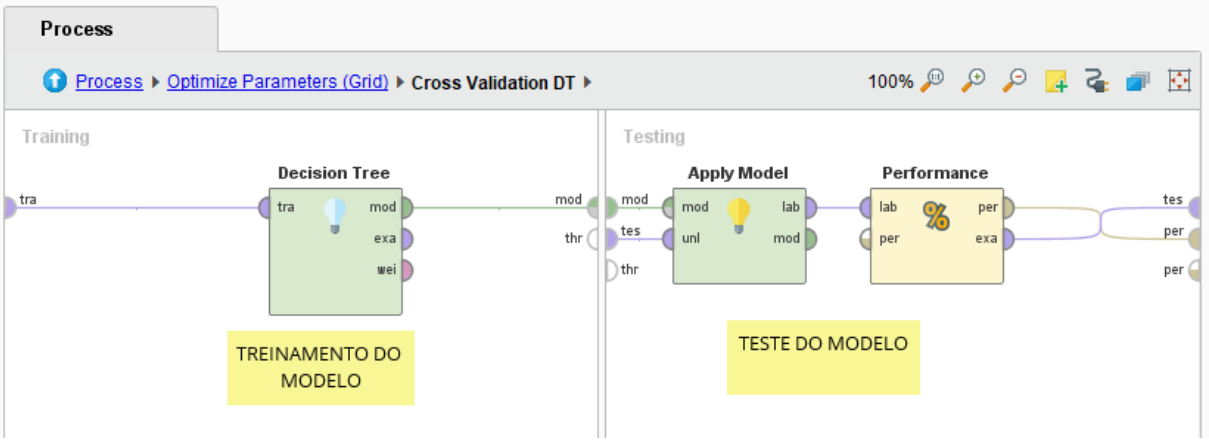
Figura 24 - Subprocesso do operador *Optimize Parameters (Grid)*, Validação Cruzada



Fonte: construção da autora.

A Figura 25 mostra os subprocessos de treinamento e teste do classificador *Decision Tree (DT),* realizados pelo operador *Cross Validation*.

Figura 25 - Subprocessos do operador *Cross Validation*, com o operador *Decision Tree*



Fonte: construção da autora.

Para análise do desempenho dos classificadores, com otimização dos seus parâmetros, foram selecionadas e organizadas, no Quadro 13, as mesmas métricas utilizadas anteriormente.

8.3.1.3.1 Avaliação dos resultados

Com a otimização dos parâmetros, o classificador *Decision Tree (DT)* obteve o melhor desempenho. O *DT* demonstrou melhora em todas as métricas, inclusive no percentual de acertos na predição das duas classes, que passou de 73,83%,

atingido no experimento anterior, para, neste atingir 82,01%. A precisão e a especificidade tiveram uma considerável melhora, o que significa que o modelo diminui os erros de predição para a classe negativa. O *RF* também melhorou seu desempenho e manteve o equilíbrio entre as métricas, apresentando o maior número de acertos na classe positiva em relação aos demais classificadores. O *K-NN* não teve melhora significativa, ficando com desempenho abaixo do *DT* e do *RF* (Quadro 13).

Quadro 13 - Desempenho dos classificadores com otimização dos parâmetros

OTIMIZAÇÃO *CROSS VALIDATION*						
Modelos classificadores	Acurácia %	Precisão %	Sensibilidade %	Especificidade %	VPN %	MATRIZ DE CONFUSÃO
DT	82,01	91,42	79,41	86,67	69,89	65 28 10 108
RF	81,04	84,82	87,50	69,33	75,36	52 17 23 119
K-NN	75,37	83,87	77,21	72,00	63,53	54 31 21 105
GBT	76,32	83,42	79,41	70,67	65,43	53 28 22 108
NB	73,82	83,17	76,52	68,92	62,20	51 31 23 101

Fonte: construção da autora.

8.3.1.4 Avaliação dos experimentos

Os experimentos demonstraram que, entre as formas de divisão dos dados para treinamento e teste, o método de Validação Cruzada, usado no experimento 2, apresentou melhor resultado na validação dos classificadores, por usar de forma fracionada todos os objetos em algum momento no conjunto de treino e no conjunto de teste do modelo, aumentando a precisão da avaliação através da média dos resultados.

A otimização dos parâmetros de cada classificador, com o operador *Optimize Parameters*, resultou em aumento das acurácias, com destaque para o *Decision Tree* e *Random Forest*. Verificou-se que as acurácias variaram entre 73% e 82%. Além desta métrica, a precisão, que é a taxa de acertos da predição na classe

positiva, variou entre 83% e 93% demonstrando um bom desempenho dos algoritmos na construção dos modelos.

De acordo com os resultados do terceiro experimento, três dos cinco classificadores geraram modelos com estimativa de acertos de predição acima de 75%, podendo identificar os alunos com propensão à evasão com probabilidade de erro menor do que 25%, são eles: *K-NN, Random Forest* e *Decision Tree.*

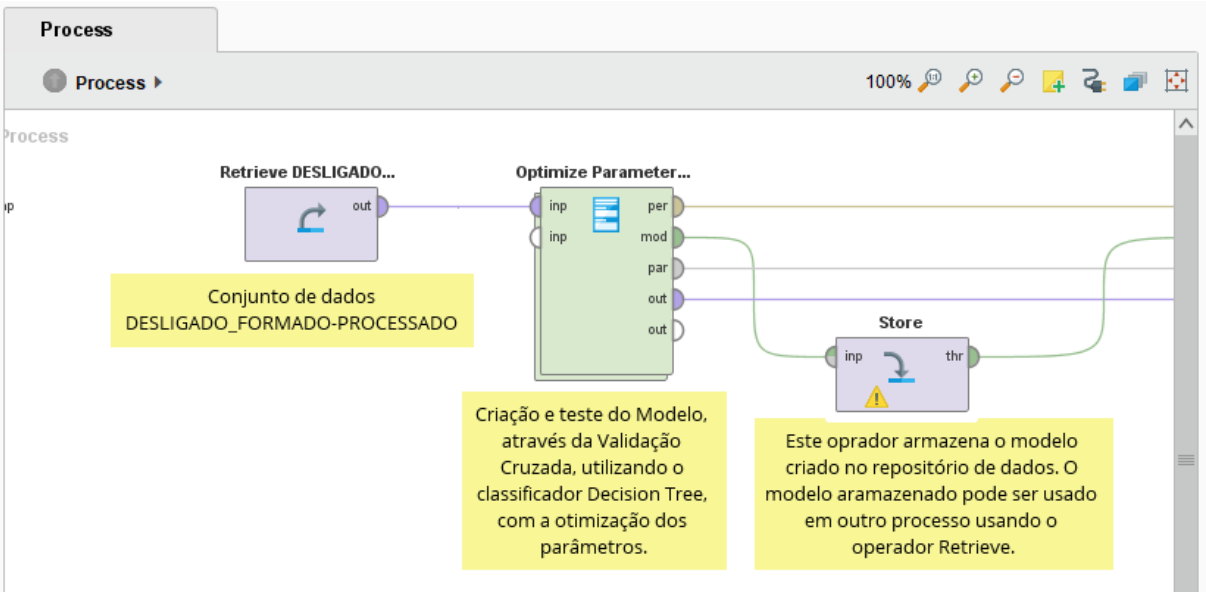
8.3.2 Escolha do modelo

Com base nos resultados, apresentados na seção anterior, selecionou-se o modelo construído pelo classificador *Decision Tree,* que obteve o melhor desempenho com a otimização dos parâmetros. Além de apresentar bom desempenho na classificação dos dados utilizados para treinamento, este classificador apresenta uma representação gráfica de fácil compreensão.

A representação do modelo criado pelo *Decision Tree* é em formato de árvore, o que possibilita identificação dos alunos propensos à evasão de forma direta, sem precisar sua aplicação em uma ferramenta de mineração de dados, o que também pode ser feito. Por ser de fácil interpretação, pode ser utilizado pelas equipes pedagógicas e gestoras sem dificuldades e sem a necessidade de um especialista da área de ciências de dados ou da ciência da computação para fazê-lo. A árvore que representa o Modelo IFRS-CAN pode ser vista na Figura 27.

O processo no *RapidMiner* para criação do Modelo IFRS-CAN é o mesmo utilizado para a otimização dos parâmetros, descrito no Experimento 3 e demonstrado nas Figuras 23, 24 e 25. Com o operador *Store* o Modelo IFRS-CAN foi salvo e armazenado no repositório do *RapidMiner*, para ser aplicado em nova base de dados (Figura 26).

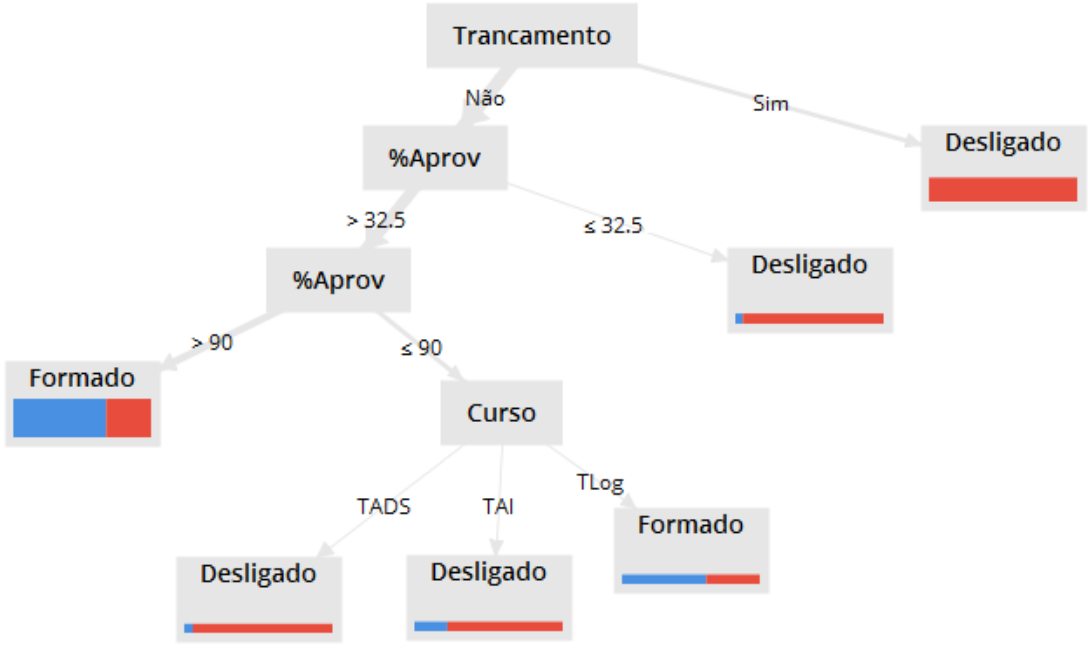
Figura 26 - Criação do Modelo IFRS-CAN



Fonte: construção da autora.

Conforme descrito anteriormente, predizer se um aluno é propenso a evadir pode ser feito de modo direto, respondendo às condições de teste do nó raiz e dos nós internos da árvore construída, que representa o modelo. Iniciando pelo nó raiz, seguindo a ramificação que responde à condição de teste, de acordo com as características do aluno, que levará a um nó interno, no qual uma nova condição de teste será aplicada a outro atributo, ou a um nó folha. Os nós folhas contêm um dos rótulos de classe que será atribuído ao aluno (TAN et al., 2009, pag. 178). Com base nesta explicação, a análise do modelo foi feita na seção 5.4.

Figura 27 - Modelo de predição IFRS-CAN



Fonte: construção da autora.

8.3.3 Validação do modelo em novos dados

A última fase da avaliação do modelo é sua aplicação ou validação em um conjunto de dados diferente daquele no qual foi treinado e testado. O conjunto de dados para validação do Modelo IFRS-CAN, foi constituído pelos dados dos alunos que estavam com a situação da matrícula regular quando da extração e pré-processamento dos dados, bem como dentro do período compreendido entre o segundo semestre de 2018 e o segundo semestre de 2019, momento este em que saíram da instituição. Dos 209 alunos que estavam com situação regular em 2018, 96 saíram da instituição, 48 foram desligados e 48 estão formados. O atributo “Situação” foi atualizado para todos os alunos (objetos) com base nas informações da planilha “Alunos Cursos Superiores” (Figura 9) e, de acordo com a forma de saída da instituição, o valor foi alterado de REGULAR para DESLIGADO ou FORMADO.

O mesmo processo aplicado ao conjunto dos dados utilizados para treinamento, descrito na seção 5.3 e demonstrado na Figura 16, também foi realizado neste novo conjunto. E o resultante pode ser visto na Figura 28.

Figura 28 - Conjunto de dados “DESLIGADO_FORMADO_2019-PROCESSADO”

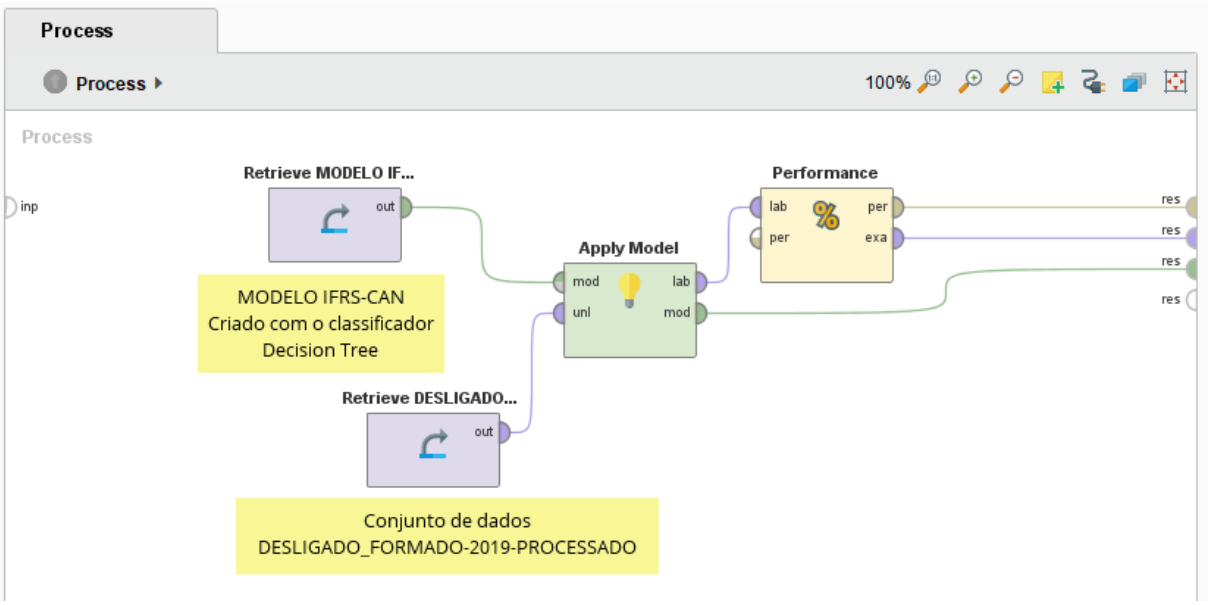
Row No.	Situação	%Aprov	Acesso_Inte...	Acompanha...	Ativ_Fisica	Auxilio_Gov	BAE	Comput_pró...	Conhec_Info	Curso	Depende_R
1	Formado	100	Alta	Não	Sim	Não	Não	Comp. Com	Muito bom	TLog	3
2	Formado	50	Alta	Sim	Sim	Não	Não	Comp. Com	Bom	TLog	4
3	Desligado	100	Alta	Não	Sim	Não	Não	Comp. Com	Bom	TADS	2
4	Desligado	100	Alta	Não	Sim	Não	Não	Comp. Com	Bom	TAI	2
5	Desligado	80	Alta	Sim	Sim	Não	Não	Comp. Com	Bom	TADS	5
6	Formado	40	Média	Não	Sim	Não	Não	Comp. Sem	Ruim	TAI	1
7	Desligado	60	Alta	Sim	Sim	Não	Não	Comp. Com	Bom	TLog	5
8	Formado	0	Baixa	Não	Não	Não	Não	Comp. Com	Bom	TLog	3
9	Desligado	100	Alta	Não	Não	Não	Sim	Comp. Com	Bom	TLog	3
10	Desligado	100	Alta	Não	Sim	Não	Não	Comp. Com	Bom	TADS	3
11	Desligado	100	Alta	Não	Sim	Não	Sim	Comp. Com	Muito bom	TAI	1
12	Formado	80	Alta	Não	Não	Não	Não	Comp. Com	Bom	TAI	4

ExampleSet (96 examples, 1 special attribute, 33 regular attributes)

Fonte: construção da autora.

O Modelo IFRS-CAN foi aplicado ao novo conjunto de dados, formado por 96 objetos. Destes, 48 pertencem à classe DESLIGADO e 48 pertencem à classe FORMADO e 34 atributos, conforme mostra a Figura 29.

Figura 29 - Aplicação do Modelo IFRS-CAN



Fonte: construção da autora.

As métricas utilizadas para avaliar o desempenho dos classificadores na fase de treinamento e teste também foram usadas para avaliar o resultado da validação do Modelo IFRS-CAN. O número de acertos e erros do modelo em cada uma das classes está representado na matriz de confusão gerada pelo *RapidMiner* (Figura 29) e mostra que o modelo teve dificuldade de classificar os objetos da classe positiva (Desligado) neste conjunto de dados. Dos 48 objetos pertencentes a esta classe, 25 foram classificados corretamente, resultando numa sensibilidade de 52,08%.

Figura 30 - Matriz de confusão da aplicação do Modelo IFRS-CAN

accuracy: 60.42%

	true Formado	true Desligado	class precision
pred. Formado	33	23	58.93%
pred. Desligado	15	25	62.50%
class recall	68.75%	52.08%	

Fonte: construção da autora.

Na classe negativa (Formado) o modelo teve um desempenho melhor, predizendo corretamente o dobro de objetos (33 Verdadeiros Positivos) em relação ao número predito para Falsos Positivos, resultando em 68,75% na medida de especificidade. A acurácia do modelo neste conjunto de dados foi de 60,42%, o que representa 58 predições corretas (33 Verdadeiros Positivos e 25 Verdadeiros Negativos), num total de 96 objetos. O *RapidMiner* gerou um arquivo de dados (Figura 31) com as predições, que foi armazenado no seu repositório interno, para estudos posteriores.

Figura 31 - Conjunto de dados DESLIGADO_FORMADO_2019-PROCESSADO com as predições do Modelo IFRS-CAN

Row No.	Situação	prediction(Situação)	confidence(Formado)	confidence(Desligado)	%Aprov	Acesso_Inte...	Acompanha...	Ativ_Fisica	Auxilio_Gov	BAE	Comput_pró...
1	Formado	Formado	0.674	0.326	100	Alta	Não	Sim	Não	Não	Comp. Com
2	Formado	Formado	0.611	0.389	50	Alta	Sim	Sim	Não	Não	Comp. Com
3	Desligado	Formado	0.674	0.326	100	Alta	Não	Sim	Não	Não	Comp. Com
4	Desligado	Formado	0.674	0.326	100	Alta	Não	Sim	Não	Não	Comp. Com
5	Desligado	Desligado	0.059	0.941	80	Alta	Sim	Sim	Não	Não	Comp. Com
6	Formado	Desligado	0.222	0.778	40	Média	Não	Sim	Não	Não	Comp. Sem
7	Desligado	Formado	0.611	0.389	60	Alta	Sim	Sim	Não	Não	Comp. Com
8	Formado	Desligado	0	1	0	Baixa	Não	Não	Não	Não	Comp. Com
9	Desligado	Formado	0.674	0.326	100	Alta	Não	Não	Não	Sim	Comp. Com
10	Desligado	Formado	0.674	0.326	100	Alta	Não	Sim	Não	Não	Comp. Com
11	Desligado	Formado	0.674	0.326	100	Alta	Não	Sim	Não	Sim	Comp. Com
12	Formado	Desligado	0.222	0.778	80	Alta	Não	Não	Não	Não	Comp. Com
13	Desligado	Desligado	0.222	0.778	40	Baixa	Não	Não	Não	Não	Comp. Com

ExampleSet (96 examples, 4 special attributes, 33 regular attributes)

Fonte: construção da autora.

Alguns testes foram realizados para criar um modelo com menor número de erros na predição para novos dados. Os experimentos envolveram equilíbrio das classes e ajustes de parâmetros (de forma manual) no processo de treinamento e teste, redução de atributos, uso dos demais classificadores para criar outros modelos. Porém, o desempenho dos modelos na validação em novos dados foi muito próximo, com a acurácia variando entre 60,42% e 63,54%.

Buscando melhorar o desempenho do Modelo IFRS-CAN e identificar as causas dos erros, analisaram-se os dois conjuntos de dados, tanto o conjunto usado para treinamento, quanto o usado para validação do modelo. Explorando os valores contidos no atributo “Trancamento” (Quadro 14), destacado no nó raiz da árvore criada pelo Modelo, observou-se, no conjunto de treinamento que: 52 objetos têm o valor sim (25% do total de 211 alunos trancaram o segundo semestre). Destes, 100% pertencem à classe Desligado (100% dos alunos que trancaram o segundo semestre não concluíram o curso). No conjunto de aplicação, observou-se que apenas 6 objetos tem o valor sim (apenas 6% do total de 96 alunos trancaram o segundo semestre) e destes três pertencem à classe Desligado e três à classe Formado. Ainda, dos 90 objetos com valor não (94% do total de objetos no conjunto de dados), 50% pertencem à classe Desligado e 50% à classe Formado.

Quadro 14 - Nº de alunos por classe, no atributo Trancamento, nos conjuntos de dados

	Conjunto de treinamento e teste Desligado_Formado - 2018 Total de alunos 211 (objetos)	**Conjunto de validação Desligado_Formado - 2019 Total de 96 alunos (objetos)**
TRANCADO **SIM**	Total de objetos - 52 (25 % de 211) Desligado - 52 (100 %) Formado - **0 (0 %)**	Total de objetos - **6 (6 %** de 96) Desligado - 3 (50 %) Formado - **3 (50 %)**
TRANCADO **NÃO**	Total de objetos - 159 (75 % de 211) Desligado - 84 (53 %) Formado - 75 (47 %)	Total de objetos - **90 (94 %** de 96) Desligado - 45 (50 %) Formado - 45 (45 %)

Fonte: construção da autora.

Outro atributo analisado foi o %Aprov (percentual de aprovação no primeiro semestre) (quadro 15), localizado na árvore como primeiro nó interno. No conjunto

de treinamento, 75 objetos (36% do total de 211) possuem valor menor que 50 para este atributo; destes, apenas 4 pertencem à classe Formado (4% dos 75 objetos), o restante pertence à classe Desligado (95% dos alunos com percentual de rendimento abaixo de cinquenta por cento estão desligados). No conjunto de validação do modelo, 27 objetos (28% do total de 96) possuem valor menor que 50 neste atributo, 7 objetos (26%) pertencem à classe Formado, o restante à classe Desligado (74% dos alunos com percentual de rendimento abaixo de cinquenta por cento estão desligados no conjunto de dados de aplicação).

Quadro 15 - Nº de alunos por classe, no atributo % Aprov, nos conjuntos de dados

	Conjunto de treinamento e teste Desligado_Formado - 2018 Total de alunos 211 (objetos)	**Conjunto de validação Desligado_Formado - 2019 Total de 96 alunos (objetos)**
% APROV ≤ 50%	Total de objetos -75 (36 % de 211) Desligado - 71 (95 %) Formado - **4 (5 %)**	Total de objetos - 27 (28 % de 96) Desligado - 20 (74 %) Formado - **7 (26 %)**
% APROV ≥ 50%	Total de objetos - 136 (64 % de 211) Desligado - 65 (48 %) Formado - 71 (52 %)	Total de objetos - 69 (72 % de 96) Desligado - 28 (41 %) Formado - 41 (59 %)

Fonte: construção da autora.

Os objetos que apresentaram valor “sim” para trancado, no conjunto de validação do modelo, mas que pertencem à classe Formado, assim como aqueles com valor menor que 50% no atributo %Aprov e também pertencem à classe formado, foram classificados como Desligado. Este fato foi comprovado pela análise da planilha gerada pelo *RapidMiner* com as predições (Figura 31). Essas exceções foram consideradas como ruídos pelo classificador, ou seja, objetos rotulados de forma errada, fazendo com que fossem classificados como pertencentes à classe Desligado.

Além das exceções serem consideradas ruídos, o fato, do conjunto de treinamento possuir número reduzido de objetos, impossibilitou que o modelo encontrasse todos os valores possíveis de cada atributo distribuídos nas duas classes de predição. Sendo assim, não foi possível que o modelo identificasse corretamente novos objetos no conjunto de dados de validação. Por consequência, e com o conjunto de validação também pequeno, cada objeto mal classificado

refletiu de forma exponencial na acurácia e nas demais métricas avaliadas. Segundo Manhães (2015, p. 46) “A precisão da resposta do modelo, entre outras coisas, depende da qualidade e da quantidade de dados disponíveis. Neste caso, isto se aplica tanto na criação do modelo, quanto na sua validação ou utilização”.

O Modelo IFRS-CAN também foi aplicado sobre o conjunto de dados dos alunos que ainda se mantém regulares (Figura 32), para que a instituição faça o acompanhamento destes, no sentido de não permitir que as predições na classe positiva se efetivem, buscando impedir a evasão.

Figura 32 - Planilha com a predição do modelo IFRS-CAN sobre os alunos com situação regular

R...	ID	Situação	prediction(S...	confidence(...	confidence(...	%Aprov	Acesso_Inte...	Acompanha...	Ativ_Fisica	Auxilio_Gov	BAE	Comput_pró...	Conhec_Info
1	2080196	?	Formado	0.674	0.326	100	Alta	Não	Não	Não	Não	Comp. Com	Bom
2	2080188	?	Desligado	0.059	0.941	83	Alta	Não	Não	Não	Sim	Comp. Com	Regular
3	2050140	?	Desligado	0	1	0	Alta	Não	Sim	Não	Não	Comp. Com	Bom
4	2040176	?	Desligado	0.222	0.778	60	Alta	Não	NR	Não	Não	Comp. Com	Muito bom
5	2040182	?	Desligado	0.222	0.778	60	Alta	Não	NR	Não	Não	Comp. Com	Regular
6	2040260	?	Formado	0.674	0.326	100	Alta	Sim	Não	Não	Não	Comp. Com	Muito bom
7	2080067	?	Formado	0.674	0.326	100	Alta	Sim	Sim	Não	Não	Comp. Com	Muito bom
8	2050259	?	Formado	0.674	0.326	100	Alta	Não	Sim	Não	Sim	Comp. Com	Regular
9	2080167	?	Desligado	0.059	0.941	83	Alta	Não	Sim	Não	Não	Comp. Com	Muito bom
10	2040196	?	Desligado	0.222	0.778	40	Alta	Não	Não	Não	Não	Comp. Com	Muito bom
11	2080181	?	Desligado	0.059	0.941	83	Alta	Não	Sim	Sim	Não	Comp. Com	Bom
12	2040240	?	Desligado	0.222	0.778	60	Baixa	Não	Não	Sim	Não	Comp. Sem	Regular
13	2040279	?	Desligado	0.050	0.950	0	Alta	Não	Não	Não	Não	Comp. Com	Regular

ExampleSet (113 examples, 5 special attributes, 33 regular attributes)

Fonte: construção da autora.

Para este conjunto de dados os valores no atributo situação ficaram vazios e o número de matrícula foi mantido como atributo especial Id (identidade), para que os alunos possam ser acompanhados. Quando da importação, o *RapidMiner* atribuiu um ponto de interrogação para os valores faltantes no atributo situação. Após a aplicação do modelo, a coluna com a predição foi gerada. Este arquivo com as predições para os alunos que ainda estão regulares será repassado à direção do Campus Canoas, para que tome as decisões quanto à melhor forma de acompanhamento, e estratégias cabíveis; a fim de que os alunos classificados como propensos à evasão permaneçam na instituição e tenham êxito na conclusão do curso.

8.4 **ANÁLISE DO MODELO**

O Modelo IFRS-CAN atingiu um desempenho satisfatório na fase de treinamento e teste, mostrando-se promissor na identificação de alunos com propensão à evasão. Analisando a Matriz de Confusão (Figura 32), gerada a partir das classificações realizadas pelo algoritmo DT, no Experimento 3, sobre o conjunto de dados utilizado para treinamento e teste, observamos que sua taxa de acertos na predição das duas classes (acurácia) é de 82,01%. Outras análises importantes referentes a acertos e erros da classificação dos exemplos em cada uma das classes foram realizadas. Na classe negativa, dos 75 exemplos FORMADOS, 65 foram classificados de forma correta e apenas 10 foram classificados na classe positiva (10 falsos positivos), o que demonstra um valor de especificidade de 86,67%. Já na classe positiva, a medida de sensibilidade é de 79,41%, demonstrando que o total de predições é bom, porém 28 exemplos, do total de 136, foram classificados como formados (28 falsos negativos), o que representa um percentual de 20,59% do total de exemplos nesta classe. Este percentual exige especial atenção, pois mostra que o modelo está classificando erroneamente exemplos da classe DESLIGADO como pertencentes à classe FORMADO, em outras palavras está predizendo como formado alunos com potencial risco à evasão. A precisão, que é calculada pelo número de exemplos classificados corretamente como pertencentes à classe positiva, do total alocado nesta classe é de 91,42%. Pode-se sintetizar a análise afirmando que o modelo classificou corretamente 173 (82%) objetos e incorretamente 38 (18%).

Figura 33 - Matriz de Confusão do Modelo IFRS-CAN, fase de treinamento e teste

accuracy: 82.01% +/- 11.99% (micro average: 81.99%)

	true Formado	true Desligado	class precision
pred. Formado	65	28	69.89%
pred. Desligado	10	108	91.53%
class recall	86.67%	79.41%	

Fonte: construção da autora.

A análise da representação gráfica do Modelo (Figura 26), mostra os principais atributos utilizados pelo modelo para fazer a classificação. No nó raiz tem-se o atributo Trancamento, a condição de teste realizada sobre ele é se o aluno

trancou ou não o segundo semestre. Como resposta, tem-se todos os alunos que trancaram sendo direcionados ao nó folha, classificados como DESLIGADOS, e os que não trancaram seguindo a ramificação para o próximo nó. No segundo nó, uma nova condição é estabelecida por outro atributo, percentual de aproveitamento atingido nas disciplinas do primeiro semestre, na qual aqueles que tiveram aproveitamento menor ou igual a 32%, foram, na grande maioria, classificados como desligados. Os que tiveram aproveitamento acima deste percentual passaram pelo terceiro nó, com a condição de ter aproveitamento maior ou menor que 90% nas disciplinas do primeiro semestre. Um terço dos alunos que obtiverem aproveitamento acima de 90% foi classificado como desligado e dois terços como formados. Por outro lado, os que não atingiram este percentual passaram pelo quarto nó, cuja condição é o curso no qual o aluno esteve matriculado. A grande maioria dos alunos do curso de TADS forma classificados como desligados, menos de um terço do curso de TAI foi classificado como formado e o restante como desligado e no curso de TLog em torno de 60% classificados como formados e o restante como desligados.

O atributo Trancamento no nó raiz confirma os estudos de Polydoro (2000), em que ela considera esta parada temporária no curso como um indício do abandono e mostra a importância de analisar os motivos individuais dos alunos, para que estes sejam incentivados a retornar e recebam o auxílio da instituição, quando possível, na solução de suas dificuldades. Da mesma forma, o percentual de aproveitamento teve um peso importante na classificação dos alunos, apontando para a necessidade de uma análise mais detalhada dos motivos da não aprovação.

O modelo identificou os atributos relacionados ao desempenho acadêmico dos estudantes como tendo maior relação com o atributo classificador “Situação”, sendo estes evidenciados na representação em forma de árvore. Os atributos sociodemográficos tiveram menor relevância na classificação dos alunos, seja na classe Formado ou Desligado. Pode-se considerar a estabilidade desses atributos, presença majoritária de um único valor, como causa da pouca relação com o atributo-alvo e, por consequência, a menor importância para a classificação. A predominância de um valor sobre os demais, em alguns atributos, pode ser constatada durante a fase de pré-processamento. Acredita-se que é prematuro desconsiderar os atributos sociais, econômicos e demográficos, devido ao número reduzido de objetos no conjunto de dados e a importância dada a estes como

causas da evasão na literatura sobre o assunto. Por estes motivos, mantiveram-se todos os atributos constantes no modelo de dados (Apêndice A), para que as características individuais dos alunos, bem como a realidade do campus, fossem contempladas. A redução do número de atributos, para melhorar a performance do algoritmo classificador, assim como a inclusão de outras informações de desempenho acadêmico, devem ser analisados em trabalhos futuros, para qualificar o modelo.

9 CONSIDERAÇÕES FINAIS

A evasão é um problema sério, que faz sombra às políticas de acesso, de ingresso e de aumento do número de vagas nas instituições de ensino. Este fenômeno complexo preocupa o IFRS e o *Campus* Canoas e vem sendo combatido, através de diversas ações, na tentativa de aumentar a permanência e o êxito dos alunos. No *Campus* Canoas se faz necessária uma maior atenção aos cursos superiores de tecnologia, pois são nesses que ingressam um maior número de estudantes, nos quais identificamos o maior percentual de saídas sem êxito, o abandono do curso antes da sua conclusão, ou seja, a evasão. Ademais, são cursos procurados por jovens que desejam entrar no mercado de trabalho e por trabalhadores que buscam formação acadêmica para uma melhor colocação, e a não conclusão resulta no prejuízo pessoal para o aluno, para a instituição e para a sociedade.

A opção, feita nesta dissertação, por abordar o problema da evasão através do processo de *KDD*, mostrou-se muito apropriada, pois os resultados permitem reconhecer que é possível criar um modelo preditivo para identificar alunos com propensão à evasão, Somado a isto, a aquisição de conhecimentos sobre o problema se deu em todas as fases do processo e não apenas na mineração dos dados.

No que se refere à base de dados, esta foi construída a partir de informações extraída dos sistemas de acompanhamento acadêmico (SIA - Sistema de Informações Acadêmicas, SIFRS - Sistemas IFRS e das planilhas do Setor de Registro Escolar) utilizados pelo campus, o que demonstra que um único sistema de registro das informações dos estudantes deve contemplar as mais diversas informações, de modo que documentos complementares não sejam necessários para a predição da evasão.

A preparação dos dados, realizada na fase de pré-processamento do processo de *KDD*, resultou na criação de um Modelo de Dados de referência, que poderá servir de base para a aplicação de estudos similares neste ou em outros campi, bem como no Instituto como um todo. O Modelo de Dados contempla, nos atributos, os dados que representam as possíveis causas da evasão presentes nas pesquisas da área, abarcadas por esta dissertação e no PEPEE do IFRS. Este Modelo de Dados, se aplicado na base de dados do sistema acadêmico do Campus

Canoas (ou em outras bases que o quiserem utilizar), por exemplo, nos remete à necessidade de uma readequação dos dados armazenados, de forma que ela contemple atributos como: trancamentos de curso, período de desligamento do curso, informações de aprovações, reprovações (por nota e frequência) e percentual de rendimento e aproveitamento em relação às disciplinas cursadas a cada semestre, organizados por ano/semestres do curso. Cabe salientar que, para atingir os objetivos desta dissertação, a falta destas informações (atributos) ou a dificuldade de acesso a elas, geraram um custo de tempo e trabalho maior para a elaboração do Modelo de Dados.

O presente trabalho selecionou um modelo preditivo, através de três experimentos de validação, incluindo a otimização de parâmetros, em que foram aplicados cinco algoritmos de classificação sobre uma base de dados dos alunos dos cursos superiores de tecnologia do *Campus* Canoas do IFRS, ingressantes no período entre 2011 e 2017. A partir do resultado dos experimentos, foi selecionado o modelo construído com o classificador *Decision Tree*. O Modelo IFRS-CAN, assim denominado, demonstrou desempenho satisfatório no treinamento, com acurácia de 82%, o que indica uma boa taxa de acertos de predições nas duas classes; 91,42% de precisão e 79,41% de sensibilidade, demonstrando precisão de acertos na classe positiva, ou seja, fazendo uma predição bastante segura dos alunos com propensão à evasão no grupo de teste. Embora a taxa de acertos nas predições do modelo para o conjunto de teste, durante o treinamento, seja de 82%, sua taxa de acertos nas predições para o conjunto de novos dados, fase de validação, foi de 60,42%. Esta discrepância nos valores de acurácia ocorreu porque o modelo não conseguiu mapear todos os possíveis valores, principalmente, para os atributos “Trancamento’ e “%Aprov” e relacioná-los com as duas classes, durante a fase de treinamento. Este fato resultou em erro de generalização durante a Validação do Modelo, porque alguns registros na base de dados (exceções), não condizem com o que foi aprendido por meio dele. Alguns valores no conjunto de treinamento, nestes atributos, estavam relacionados apenas a umas das classes. Também é importante mencionar que o número reduzido de objetos, na fase de treinamento e teste, se refletiu em um limitante para o aprendizado do modelo.

Apesar do Modelo IFRS-CAN não ter demonstrado desempenho equivalente à etapa de treinamento, na fase de validação, ainda assim, este foi superior a 50%, o que é melhor do que uma situação de aposta sobre a conclusão ou não do curso

pelo aluno. Este fato motiva a realização de trabalhos futuros, utilizando o Modelo de Dados criado para a organização das informações dos estudantes, com inclusão de mais objetos (alunos).

Os resultados, principalmente da fase de treinamento, permitem concluir que é possível realizar a predição das formas de saída dos alunos da instituição, objetivando identificar os alunos com propensão à evasão antes do final do segundo semestre do curso. Isto possibilita que a equipe diretiva e de acompanhamento dos estudantes possam tomar decisões e realizar ações preventivas, ainda durante o primeiro ano do curso.

O IFRS, apesar de fazer parte de uma rede centenária, é uma instituição jovem, que está alinhando suas políticas, seus processos e práticas, levando em consideração a diversidade da sua comunidade interna e externa, especialmente no que tange à mitigação de problemas como a evasão e suas consequências. A consolidação da Política de Assistência Estudantil, o Plano Estratégico de Permanência e Êxito dos Estudantes e o Observatório de Permanência e Êxito são bons exemplos do esforço para entender as causas e buscar soluções para reduzir a evasão.

Espera-se que os experimentos realizados nesta dissertação, que teve como estudo de caso o Campus Canoas, despertem a criação e a execução de projetos futuros, envolvendo novos dados dos alunos deste campus e de outros, para busca de conhecimentos sobre a evasão e de alternativas para mitigá-la, para além da mera constatação estatística.

Neste sentido, aponta-se que, mesmo com a experiência e o conhecimento da autora sobre as informações extraídas dos sistemas, das planilhas do SRE e dos valores possíveis em cada atributo, o pré-processamento dos dados exigiu muita atenção, concentração, tempo e uma grande quantidade de trabalho manual. Assim, para que o modelo de dados possa ser utilizado em outros experimentos, com um custo de tempo e trabalho menor para preparação do conjunto de dados, é importante que as informações estejam concentradas em único sistema, ou, pelo menos, sejam geradas em bases similares, com o mesmo formato, que facilite a integração. Além disto, a estruturação e o armazenamento dos dados dos alunos com qualidade, ou seja, sistêmicos, anônimos, com maior número possível de informações padronizadas e de fácil acesso, facilitará novas pesquisas individuais (estudos de caso) ou coletivas, envolvendo o IFRS como um todo. Para tanto,

recomenda-se que o Sistema de Acompanhamento Acadêmico, que está em uso e ou implantação na maioria dos *campi* do IFRS, possa concentrar todas as informações dos alunos descritas no Modelo de Dados.

Destaca-se a importância da assistência aos alunos apontados pelo modelo como propensos a evadir, para que as predições não sejam determinísticas, mas indicativas para intervenção da equipe pedagógica, da Assistência Estudantil e das coordenações de curso. Além disto, o acompanhamento possibilitará outras relações e análises qualitativas, incorporando mais informações sobre a evasão.

Como trabalho futuro, vislumbra-se a possibilidade de se utilizar os dados obtidos através do questionário elaborado pela Comissão Interna de Acompanhamento de Ações de Permanência e Êxito dos Estudantes, o qual será aplicado anualmente de forma padronizada para todos os campi, o que permitirá a criação de uma base de dados ampla e atualizada anualmente. Esta base única possibilitará a experiência de criação de um modelo de predição da evasão para o IFRS, pois o aprendizado deste estará apoiado em informações que retratem a realidade dos alunos de cada campi e do IFRS como um todo.

Por fim, deseja-se que o acompanhamento da evasão dentro do IFRS seja uma constante e que envolva não apenas as equipes pedagógicas, as Assistências Estudantis e as Comissões de Permanência e Êxito, mas técnicos administrativos, professores, coordenadores de cursos e equipe diretiva, na busca de ações mais efetivas no combate à evasão.

REFERÊNCIAS

AMARAL, João Batista. **Evasão discente no ensino superior:** estudo de caso no Instituto Federal de Educação, Ciência e Tecnologia do Ceará. 2013. Dissertação (Mestrado em Políticas Públicas e Gestão da Educação Superior) – Superintendência de Recursos Humanos, Universidade Federal do Ceará, Fortaleza, 2013. Disponível em: http://www.repositorio.ufc.br/handle/riufc/8013. Acesso em: 02 nov. 2018.

AMARAL, Marcelo Gomes do. **Mineração de dados aplicada à classificação do risco de evasão de discentes ingressantes em instituições federais de ensino superior.** 2016. Dissertação (Mestrado em Ciência da computação) - Centro de Informática, Universidade Federal de Pernambuco, Recife, 2016. Disponível em: https://repositorio.ufpe.br/handle/123456789/19502. Acesso em: 15 ago. 2018.

ASSIS, Lucas Rocha Soares de. **Perfil de evasão no ensino superior brasileiro: uma abordagem de mineração de dados.** 2017. Dissertação (Mestrado Profissional em Computação aplicada) – Departamento de Ciência da Computação, Instituto de Ciências Exatas, Universidade de Brasília, Brasília, 2017. Disponível em: https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/trabalhoConclusao/viewTrabalhoConclusao.jsf?popup=true&id_trabalho=6240221. Acesso em: 16 ago. 2018.

BAGGI, Cristiane Aparecida dos Santos; LOPES, Doraci Alves. **Evasão e avaliação institucional no ensino superior**: uma discussão bibliográfica. Avaliação (Campinas), Sorocaba, v. 16, n. 2, p. 355-374, jul 2011. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1414-40772011000200007&lng=en&nrm=iso. Acesso em: 16 de maio de 2018.

BAKER, Ryan; ISOTANI, Seiji; CARVALHO, Adriana. Mineração de dados educacionais: oportunidades para o Brasil. **Revista Brasileira de Computadores em Educação**, Porto Alegre, v. 19, n. 02, p. 03, ago. 2011. Disponível em: http://br-ie.org/pub/index.php/rbie/article/view/1301/1172. Acesso em: 07 maio 2018.

BRASIL. **Lei nº 11.892, de 29 de dezembro de 2008.** Institui a Rede Federal de Educação Profissional, Ciência e Tecnológica, cria os Institutos Federais de Educação, Ciência e Tecnologia. Brasília, DF, 2008. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2008/lei/l11892.htm Acesso em: 10 de mar. de 2018.

BRASIL. Ministério da Educação. Conselho Nacional de Educação. Câmara de Educação Básica. **Resolução nº 3, de 30 de setembro de 2009**. Dispõe sobre a instituição Sistema Nacional de Informações da Educação Profissional e Tecnológica (SISTEC), em substituição ao Cadastro Nacional de Cursos Técnicos de Nível Médio (CNCT), definido pela Resolução CNE/CEB nº 4/99. Brasília, DF: Ministério da Educação, 2009. Disponível em: http://portal.mec.gov.br/dmdocuments/rceb003_09.pdf. Acesso em: 01 nov. 2018.

BRASIL. Ministério da Educação. Secretaria de Educação Profissional e Tecnológica. **Portaria nº 39, de 22 de novembro de 2013.** Institui Grupo de Trabalho para estudos da evasão, retenção e conclusão na Rede Federal. Brasília. DF, 2013a. Disponível em: http://www.andifes.org.br/wp-content/files_flutter/138563852990_MEC-SETEC_-_Port39_-_22-11-13_-_GT_estudo_evasao_retencao_conclusao.pdf Acesso em: 25 jul. 2018.

BRASIL. Tribunal de Contas da União. **Acórdão nº 506/2013**. Relatório de Auditoria. Interessado: Tribunal de Contas da União. Órgão: Secretaria de Educação Profissional e Tecnológica - MEC. Relator: Ministro José Jorge. Brasília, DF, 13 de mar. 2013b. Disponível em: https://contas.tcu.gov.br/etcu/ObterDocumentoSisdoc?seAbrirDocNoBrowser=true&codArqCatalogado=8995767. Acesso em: 10 nov. 2018.

BRASIL. Ministério da Educação. Secretaria de Educação Profissional e Tecnológica. **Documento orientador para a superação da evasão e retenção na Rede Federal de Educação Profissional, Científica e Tecnológica**. Brasília, DF: Ministério da Educação, 2014. Disponível em: http://r1.ufrrj.br/ctur/wp-content/uploads/2017/03/Documento-Orientador-SETEC.pdf. Acesso em: 22 jun. 2018.

BRASIL. Ministério da Educação. Secretaria de Educação Profissional e Tecnológica. Diretoria de Desenvolvimento da Rede Federal. **Nota Informativa nº 138/2015**. Informa e orienta as Instituições da Rede Federal sobre a construção dos Planos Estratégicos Institucionais para Permanência e Êxito dos Estudantes, Interessado: Rede Federal de Educação Científica e Tecnológica. Brasília, DF, 2015a. Disponível em: http://www.iftm.edu.br/proreitorias/ensino/permanenciaeexito/documentos/documentos/2015%20Nota%20Informativa%20n%C2%B0%20138%20_2015_DPE_DDR_SETEC_MEC%282%29.pdf. Acesso em: 21 nov. 2018.

BRASIL. Ministério da Educação. Secretaria de Educação Profissional e Tecnológica. **Portaria nº 23, de 10 de julho de 2015**. Institui e regulamenta a Comissão Permanente de Acompanhamento das Ações de Permanência e o Êxito dos Estudantes da Rede Federal e dá outras providências. Brasília. DF, 2015b. Disponível em: http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=21971-portaria-n23-2015-setec-pdf&Itemid=30192. Acesso em: 25 nov. 2018.

BRASIL. Ministério da Educação. Secretaria de Educação Profissional e Tecnológica. **Portaria nº 25, 13 de agosto de 2015**. Define conceitos e estabelece fatores para fins de cálculo dos indicadores de gestão das Instituições da Rede Federal de Educação Profissional, Científica e Tecnológica. Brasília. DF, ago. 2015c. Disponível em: http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=21991-portaria-n25-2015-setec-pdf&Itemid=30192. Acesso em: 25 nov. 2018.

BRASIL. Ministério da Educação. **Manual para cálculo dos indicadores de gestão das Instituições da Rede Federal de Educação Profissional, Científica e Tecnológica**: 2.0. Brasília. DF: Ministério da Educação, 2016a. Disponível em:

http://ifbemnumeros.ifb.edu.br/manual_de_indicadores_da_rfepct.pdf. Acesso em : 20 jun. 2018.

BRASIL. Ministério da Educação. Rede Federal. **Histórico.** DF, 11 abr. 2016b. Disponível em: http://redefederal.mec.gov.br/historico. Acesso em: 10 jul. 2018.

BRASIL. **Decreto nº 9.005**. Aprova a estrutura regimental e o quadro demonstrativo dos cargos em comissão e das funções de confiança do Ministério da Educação, remaneja cargos em comissão e substitui cargos em comissão do Grupo-Direção e Assessoramento Superiores – DAS por Funções Comissionadas do Poder Executivo - FCPE. Brasília. DF, 14 mar. 2017. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2017/decreto/D9005.htm. Acesso em: 25 nov. 2018.

BRASIL. Secretaria de Educação Tecnológica. **Portaria nº 1, de 3 de janeiro de 2018.** Institui a Plataforma Nilo Peçanha - PNP, a Rede de Coleta, Validação e Disseminação das Estatísticas da Rede Federal de Educação Profissional, Científica e Tecnológica – REVALIDE. Brasília. DF, 4 Jan. 2018. Disponível em: http://pesquisa.in.gov.br/imprensa/jsp/visualiza/index.jsp?data=04/01/2018&jornal=515&pagina=10. Acesso em: 01 dez. 2018.

CAETANO, Maitê Marques**. O uso de técnicas de aprendizado de máquina na predição de desempenho acadêmico de alunos em cursos superiores.** 2016. 175 f. Dissertação (Mestrado Ciência da Computação) - Centro Universitário Campo Limpo Paulista, Campo Limpo Paulista, 2016. Disponível em: https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/trabalhoConclusao/viewTrabalhoConclusao.jsf?popup=true&id_trabalho=3991818. Acesso em: 16 ago. 2018.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. **Mineração de dados: Conceitos, tarefas, métodos e ferramentas.** Universidade Federal de Goiás (UFC), 2009, 1-29. Disponível em: http://www.portal.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf. Acessado em: 03 nov. 2019.

COMISSÃO ESPECIAL DE ESTUDOS SOBRE A EVASÃO NAS UNIVERSIDADES PÚBLICAS BRASILEIRAS – ANDIFES/ABRUEM/SESu/MEC. **Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas**. [S.l.]: ANDIFES, ABRUEM, SESu, MEC, 1996. Disponível em: http://www.andifes.org.br/wp-content/files_flutter/Diplomacao_Retencao_Evasao_Graduacao_em_IES_Publicas-1996.pdf. Acesso em: 01 jun. de 2018.

CONSELHO NACIONAL DAS INSTITUIÇÕES DA REDE FEDERAL DE EDUCAÇÃO PROFISSIONAL, CIENTÍFICA E TECNOLÓGICA (CONIF). **Histórico**.2018. Disponível em: http://portal.conif.org.br/br/rede-federal/historico-do-conif. Acesso em: 10 jul. 2018.

COSTA, Jefferson de Jesus. **Uma abordagem baseada em mineração de grafos para identificação de caminhos críticos em grades e históricos curriculares de cursos de graduação.** 2015. 154 f. Dissertação (Mestrado em

Engenharia de Produção e Sistemas Computacionais) - Instituto de Ciência e Tecnologia, Universidade Federal Fluminense, Rio das Ostras, RJ, 2015. Disponível em: https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/trabalhoConclusao/viewTrabalhoConclusao.jsf?popup=true&id_trabalho=3500922. Acesso em: 15 ago. 2018.

COUTO, Diego da Costa do. **Mineração de dados educacionais aplicada à busca de perfis de alunos em casos de evasão ou retenção: uma abordagem através de redes bayesianas.** 2017. 89 f. Dissertação (Mestrado em Engenharia Elétrica) - Instituto de Tecnologia, Universidade Federal do Pará, Belém, 2017. Disponível em: https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/trabalhoConclusao/viewTrabalhoConclusao.jsf?popup=true&id_trabalho=5065467. Acesso em: 16 ago. 2018.

DORE, Rosemary.; LÜSCHER, Ana Zuleima. Permanência e evasão na educação técnica de nível médio em Minas Gerais. **Cadernos de Pesquisa**, v. 41, n. 144, p. 770-89, dez. 2011. Disponível em: http://doi.org/10.1590/S0100-15742011000300007. Acesso em: 07 de jun. 2018.

FIGUEIREDO, Natália Gomes da Silva; SALLES, Denise Medeiros Ribeiro. Educação Profissional e evasão escolar em contexto: motivos e reflexões. **Ensaio**: aval.pol.públ.Educ., Rio de Janeiro, v. 25, n. 95, p. 356-392, abr. 2017. Disponível em http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-40362017000200356&lng=pt&nrm=iso. Acesso em: 17 nov. 2017.

FACELI, Katti; LORENA, Ana Carolina; GAMA, João; CARVALHO, André Carlos Ponce de Leon Ferreira de. **Inteligência artificial**: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2011

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37-54, 1996. Disponível em: https://doi.org/10.1609/aimag.v17i3.1230 Acesso em: 20 de out. de 2018.

FINI, R.; DORE, R.; LÜSCHER, A. Z. Insucesso, fracasso, abandono, evasão... um debate multifacetado. In: **Formação/profissionalização de professores e formação profissional e tecnológica**: fundamentos e reflexões contemporâneas. Belo Horizonte: Editora PUC-Minas, 2013. p. 235-271.

GIL, Antônio Carlos. **Métodos e técnicas de pesquisa social.** 6. ed. São Paulo: Editora Atlas, 2008.

HOED, Raphael Magalhães. **Análise da evasão em cursos superiores**: o caso da evasão em cursos superiores da área de computação. 2016. 188 f. Dissertação (Mestrado Computação Aplicada) - Instituto de Ciências Exatas, Departamento de Ciência da Computação, Universidade de Brasília, Brasília, 2016. Disponível em: https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/trabalhoConclusao/viewTrabalhoConclusao.jsf?popup=true&id_trabalho=4885451. Acesso em: 15 ago. 2018.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO RIO GRANDE DO SUL. Conselho Superior. **Resolução nº 117, de 16 de dezembro de 2014.** Plano de desenvolvimento institucional do Instituto Federal do RS 2014-2018. Bento Gonçalves, RS: Conselho Superior, 2014. Disponível em: https://ifrs.edu.br/wp-content/uploads/2017/08/PDI-2014-2018.pdf. Acesso em: 15 fev. 2018.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO RIO GRANDE DO SUL. Conselho Superior. **Resolução nº 037, de 19 de abril de 2016.** Estatuto do Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul**.** Bento Gonçalves, RS, Conselho Superior, 2016. Disponível em: https://ifrs.edu.br/wp-content/uploads/2017/08/Estatuto-IFRS-Atual.pdf. Acesso em: 01 mar. 2018.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO RIO GRANDE DO SUL. Conselho Superior. **Resolução nº 086, de 17 de outubro de 2017.** Aprovar as alterações na Organização Didática do Instituto Federal do Rio Grande do Sul, aprovada pela Resolução nº 046, de 08 de maio de 2015**.** Bento Gonçalves, RS, Conselho Superior, 2017. Disponível em: https://ifrs.edu.br/ensino/documentos/organizacao-didatica/ Acesso em: 01 mar. 2018.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO RIO GRANDE DO SUL. Conselho Superior. **Resolução nº 064, de 23 de outubro de 2018.** Plano Estratégico de Permanência e Êxito os Estudantes do Instituto Federal do Rio Grande do Sul. Bento Gonçalves, RS, Conselho Superior, 2018a. Disponível em: https://ifrs.edu.br/wp-content/uploads/2018/10/Resolucao_064_18_Aprovar_Plano_Estrategico_Completo.pdf. Acesso em: 10 nov. 2018.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO RIO GRANDE DO SUL. **Sobre o IFRS.** Bento Gonçalves. 2018b. Disponível em: https://ifrs.edu.br/institucional/sobre Acesso em: 15 de out. 2018.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO RIO GRANDE DO SUL. EDITAL Nº 64/2018, Processo Seletivo Unificado para Ingresso Discente no semestre 2019/1, nos cursos superiores de graduação dos *Campi* do IFRS. Bento Gonçalves. 2018c. Disponível em: https://ingresso.ifrs.edu.br/2019/editais/ Acesso em: 14 fev. 2018.

JOHANN, Cristiane Cabral. **Evasão escolar no Instituto Federal Sul-Rio-Grandense:** um estudo de caso no *campus* Passo Fundo. 2012.119 f. Dissertação (Mestrado em Educação) - Universidade de Passo Fundo, Passo Fundo, 2012. Disponível em: http://tede.upf.br/jspui/handle/tede/739. Acesso em: 16 de maio de 2018.

LAKATOS, Eva Maria; MARCONI, Marina de Andrade**. Fundamentos de metodologia científica.** 5. ed. São Paulo: Atlas, 2003.

MACHADO, Roger Douglas. **Mineração de dados educacionais**: análise da evasão no curso de graduação em ciência da computação. 2015. 129 f.

Dissertação (Mestrado em Sistemas e Processos Industriais) - Universidade de Santa Cruz do Sul, Santa Cruz do Sul, 2015. Disponível em: https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/trabalhoConclusao/viewTrabalhoConclusao.jsf?popup=true&id_trabalho=3044292. Acesso em: 15 ago. 2018.

MANHÃES, Laci Mary Barbosa. **Predição do desempenho acadêmico de graduandos utilizando mineração de dados educacionais.** 2015. 140 f. Tese (Doutorado em Engenharia de Sistemas e Computação) - Coppe, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2015. Disponível em: https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/trabalhoConclusao/viewTrabalhoConclusao.jsf?popup=true&id_trabalho=2362410. Acesso em: 15 nov. 2018.

MORAES, Gustavo Henrique et al**. Plataforma Nilo Peçanha**: guia de referência metodológica. Brasília/DF: Editora Evobiz,2018, 101 p. Disponível em: https://drive.google.com/file/d/1WLWTxdjNej448_VMVGsbC-wLMiT7r-9d/view. Acesso em: 01 dez. 2018.

MOTTA, Porthos Ribeiro de Albuquerque. **Estudo exploratório do uso de classificadores para a predição de desempenho e abandono em universidades.** 2016. 154 f. Dissertação (Mestrado em Ciência da Computação) - Instituto de Informática, Universidade Federal de Goiás, Goiânia, 2016. Disponível em: https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/trabalhoConclusao/viewTrabalhoConclusao.jsf?popup=true&id_trabalho=3872977. Acesso em: 15 nov. 2018.

NEVES, Rita de Cássia David das. **Pré-Processamento no Processo de Descoberta de Conhecimento em Banco de Dados.** 2003. 137 f. Dissertação (Mestrado em Ciências da Computação)- Universidade Federal do Rio Grande do Sul, Porto Alegre, 2003. Disponível em: https://www.lume.ufrgs.br/bitstream/handle/10183/2701/000375412.pdf?...1. Acesso em: 20 nov. 2018.

OLIVEIRA JÚNIOR, José Gonçalves de. **Identificação de padrões para a análise da evasão em cursos de graduação usando mineração de dados educacionais.** 2015. 86 f. Dissertação (Mestrado em Computação Aplicada) - Universidade Tecnológica Federal do Paraná, Curitiba, 2015. Disponível em: http://repositorio.utfpr.edu.br/jspui/handle/1/1995. Acesso em: 15 nov. 2018

POLYDORO, Soely Aparecida Jorge. **O trancamento de matrícula na trajetória acadêmica do universitário**: condições de saída e de retorno à instituição. 2000. 167 p. Tese (Doutorado em Educação) – Universidade Estadual de Campinas, Campinas, 2000. Disponível em: http://www.repositorio.unicamp.br/handle/REPOSIP/253539. Acesso em: 27 jul. 2018.

RAPIDMINER. **RapidMiner**. Boston, c2010. Disponível em: https://rapidminer.com/. Acesso em: 13 mar. 2019.

RUMBERGER, R. W. Why students Drop Out of School and What Can Be Done. UCLA: The Civil Rights Project / Proyecto Derechos Civiles. 2001. Disponível em: https://escholarship.org/uc/item/58p2c3wp Acesso em: 29 jul. 2018.

SANTANA, Marcelo Almeida. **Um estudo comparativo das técnicas de predição na identificação de insucesso acadêmico dos estudantes durante cursos de Programação Introdutória.** 2015. 73 f. Dissertação (Mestrado em em Informática) - Instituto de Computação, Universidade Federal de Alagoas, Maceió, 2015. Disponível em: https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/trabalhoConclusao/viewTrabalhoConclusao.jsf?popup=true&id_trabalho=3106194. Acesso em: 15 nov. 2018.

SCHIMITT, Rafael Eduardo. **A evasão na educação superior:** uma compreensão ecológica do fenômeno como estratégia para a gestão da permanência estudantil. Anais da X Anped Sul - Reunião Científica da ANPED. Florianópolis: outubro de 2014. Disponível em: <http://xanpedsul.faed.udesc.br/arq_pdf/690-0.pdf>. Acessado em: março de 2018.

SEVERINO, Caroline Silva de et al. Evasão no ensino superior do Instituto Federal de Educação, Ciência e Tecnologia do Triângulo Mineiro – IFTM- Câmpus Uberlândia: um olhar sobre o curso de Tecnologia em alimentos. In: COLÓQUIO INTERNACIONAL SOBRE EDUCAÇÃO PROFISSIONAL E EVASÃO ESCOLAR**,** 3.,2013, Belo Horizonte. **Resumos** [...]. Belo Horizonte: FAE/UFMG, 2013. Disponível em: https://drive.google.com/?tab=mo&authuser=0#folders/0B1yMsJLydsHnNmo2WlBqVml2WmM Acesso em: 19 de set. de 2018.

SILVA, Denise Bianca Maduro. Evasão escolar e educação profissional. **Linhas Críticas**, Brasília, DF, v.22, n.49, p. 619-622, 2016. Disponível em: http://periodicos.unb.br/index.php/linhascriticas/article/view/24649/18502. Acesso em: 05 nov. 2018.

SILVA FILHO, R.L.L.; Motejunas, P.R.; Hipólito, O.; Lobo, M.B.C.M. A evasão no ensino superior brasileiro. **Cadernos de Pesquisa**, São Paulo, v. 37, n. 132, p. 641-659, set./dez. 2007. Disponível em: http://www.scielo.br/pdf/cp/v37n132/a0737132.pdf. Acesso em: 05 nov. 2018.

SILVA, Leandro Augusto da; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados:** com aplicações em R. 1. ed. Rio de Janeiro: Elsevier, 2016.

SILVA, Leandro A.; SILVA, Luciano. Fundamentos de mineração de dados educacionais. In: WORKSHOPS DO CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 3., 2014, Dourados, MS. **Anais** [...]. Dourados, MS: UFGD, 2014. p. 568-581. Disponível em: http://dx.doi.org/10.5753/cbie.wcbie.2014.568 Acesso em: 07 de maio de 2018.

SILVA, Leandro Augusto et al. Ciência de Dados Educacionais: definições e convergências entre as áreas de pesquisa. **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**, [S.l.], p. 764, out. 2017.

ISSN 2316-8889. Disponível em: <https://www.br-ie.org/pub/index.php/wcbie/article/view/7462/5258>. Acesso em: 7 de maio de 2018. doi:http://dx.doi.org/10.5753/cbie.wcbie.2017.764.

SOUSA, Marília Maria Bastos de Araújo Cavalcanti Feitoza Fava de. **Mineração de dados educacionais**: previsão de notas parciais utilizando classificação. 2017. 85 f. Dissertação (Mestrado em Informática) - Instituto de Computação, Icomp, Universidade Federal do Amazonas, Manaus, 2017. Disponível em: https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/trabalhoConclusao/viewTrabalhoConclusao.jsf?popup=true&id_trabalho=5943032. Acesso em: 05 dez. 2018.

TAN, Pang-Ning; STEINBACH, Michel; KUMAR, Vipin. **Introdução ao Data mining:** mineração de dados. Rio de Janeiro: Ciência Moderna, 2009.

ZAGO, N. Do acesso a permanência no ensino superior: percursos de estudantes universitários de camadas populares. **Revista Brasileira de Educação**, Rio de Janeiro, v. 11, n. 32, p. 226-237, maio/ago.2006. Disponível em: http://www.scielo.br/pdf/rbedu/v11n32/a03v11n32.pdf. Acesso em: 05 dez.2018.

APÊNDICE A – MODELO DE DADOS: DESCRIÇÃO E VALORES POSSÍVEIS PARA OS ATRIBUTOS

DESCRIÇÃO E VALORES POSSÍVEIS PARA OS ATRIBUTOS				
Nº	**ATRIBUTOS**	**TIPO DE ATRIBUTO**	**VALORES POSSÍVEIS**	**DESCRIÇÃO DOS ATRIBUTOS**
1	ID	Identificador: Numeral Inteiro	204000x; 20400xx; 2040xxx; 205000x; 20500xx; 2050xxx; 208000x	Número de matrícula de cada aluno(a): o primeiro dígito identifica o *campus* do IFRS, o segundo e o terceiro identificam o curso do *campus* e os quatro últimos dígitos são seriais e identificam o aluno no curso.
2	Situação	Rótulo de Classe: Nominal	1. Desligado 2. Formado 3. Regular	Situação em que a matrícula do aluno se encontra. A informação é de como a situação da matrícula dos alunos está no momento da coleta dos dados. Desligados inclui: desistência, evasão, jubilado e transferido Regular inclui: aluno em curso, matrícula trancada e mobilidade acadêmica.
3	Curso	Polinominal	TADS TAI Tlog	Curso no qual o aluno está matriculado 1. Tecnólogo em Análise e Desenvolvimento de Sistemas 2. Tecnólogo em Automação Industrial 3. Tecnólogo em Logística
4	Turno_Curso	Binominal	Manhã Noite	Turno no qual o curso é desenvolvido, ministrado.
5	Faixa_Etária	Polinominal	de 18 a 19 anos de 20 a 24 anos de 25 a 29 anos de 30 a 39 anos de 40 a 49 anos mais de 50 anos	Tempo de nascimento em anos. As faixas etárias são as mesmas estabelecidas na Plataforma Nilo Peçanha e no SISTEC
6	Sexo	Binominal	Masculino Feminino	Sexo do aluno(a)

7	Est.Civil	Polinominal	Casado/União Estável Divorciado/Separado Solteiro	Estado civil do aluno(a). 1. Casado/União Estável 2. Divorciado/Separado 3. Solteiro
8	Raça	Polinominal	Branco Preto Pardo Indígena ND	Cor, raça ou etnia que o aluno(a) se auto declara. As denominações estão de acordo com a Lei de Cotas. ND - não declarou ou soube dizer.
9	Forma_Ingresso	Polinominal	PS IFRS Sisu ENEM Transferência Portador de Diploma	Forma de ingresso do aluno no curso: 1. PS IFRS - Processo Seletivo/Vestibular; 2. Sisu - Sistema de Seleção Unificada; 3. Enem - nota no Exame Nacional do Ensino Médio; 4. Transferência - Edital de Transferência e 5. Portador de Diploma - Edital de Ingresso de Diplomado
10	Modalidade_Ingr	Polinominal	C1 C2 C3 C4 C5 C6 C7 C8 C9 C10 NA	Identifica se o ingresso foi por acesso universal, livre concorrência, ou por alguma reserva de vaga. A legenda das modalidades de ingresso consta no ANEXO A NA - Não se aplica modalidade de ingresso para alunos que ingressaram por transferência ou como portador de diploma.
11	Nº Disciplina	Numérico Inteiro	1 2 3 4 5 6	Mostra o número de disciplinas cursadas no semestre de ingresso do aluno(a). Os alunos ingressantes são matriculados de forma compulsória em todas as disciplinas do primeiro semestre, de acordo com a Matriz curricular de cada curso. Alunos transferido ou portadores de diploma podem escolher as disciplinas que irão cursar
12	%Aprov	Numérico Inteiro	0 17 20 25 33	Mostra o percentual de aproveitamento, aprovações, em relação ao número de disciplinas que o aluno esteve matriculado no primeiro semestre do curso. Foram

			40 50 60 66 67 68 75 80 83 100	computadas reprovações por nota e por falta. Aproveitamento de Estudos não foi considerado como aprovação. 0 nenhuma aprovação; 17% - 1 de 6 disciplinas; 20% - 1 de 5 disciplinas; 25% - 1 de 4 disciplinas; 33% - 1 de 3 disciplinas; 40% - 2 de 5 disciplinas; 60% - 3 de 5 disciplinas; 67% - 2 de 3 ou 4 de 6 disciplinas; 75% - 3 de 4 disciplinas; 80% - 4 de 5 disciplinas; 83% - 5 de 6 disciplinas; 100% - 5 de 5 ou 6 de 6 disciplinas
13	Trancamento	Binominal	Sim Não	Informação sobre o trancamento ou não do segundo semestre. Inclui trancamento a pedido do aluno ou “Automático” pelo SRE, em caso de não rematrícula.
14	Tempo_Escola_ Anterior	Polinominal	≤1 >1 e ≤2 >2 e ≤4 >4 e ≤8 >8 e ≤12 >12	Tempo de afastamento dos estudos regulares ao ingressar no curso. 1. Igual ou menos de 1 ano (≤1) 2. Entre 1 e 2 anos (>1 e ≤2) 3. Entre 2 e 4 anos (>2 e ≤4) 4. Entre 4 e 8 anos (>4 e ≤8) 5. Entre 8 e 12 anos (>8 e ≤12) 6. Mais de 12 anos (>12)
15	Instituição_ensino_médio	Polinominal	Escola Pública Escola Privada ENCCEJA/ENEM	Tipo de Instituição que o aluno cursou o ensino regular anterior. Inclui a possibilidade dos alunos terem feito certificação através da prova do ENEM ou ENCCEJA.
16	Município	Polinominal	Canoas Esteio Porto Alegre São Leopoldo Sapucaia do Sul Outro	Cidade onde mora o(a) aluno(a). Selecionar os municípios com maior número de alunos moradores e os demais municípios identificar como "outro".

17	BAE	Binominal	Sim Não	Informação se o aluno(a) recebe ou não auxílio estudantil.
18	Escolha_Curso	Polinominal	Formação Mercado de trabalho Identificação com a profissão Por exclusão, curso de preferência não oferecido Outros	Motivo da escolha do curso: 1. Complemento de **formação** 2. Disponibilidade de vagas no **mercado de trabalho** 3. **Identificação com a profissão** / Área de seu interesse 4. **Por exclusão**, uma que o **curso de preferência não era oferecido** 5. **Outros**
19	Escolha_Instituição	Polinominal	Apoio oferecido Ensino de qualidade Ensino gratuito Proximidade da residência ou do local de trabalho Outro	Motivo da escolha da instituição: 1. Apoio oferecido 2. Ensino de qualidade 3. Ensino gratuito 4. Proximidade da residência ou do local de trabalho 5. Outro
20	Tempo_Estudo	Polinominal	Nenhum Menos de uma hora Uma hora Duas horas Mais de duas horas	Tempo médio que o(a) aluno(a) dedica diariamente aos estudos, fora do período de aulas.
21	Conhec_Info	Polinominal	Muito bom Bom Regular Ruim	Como o aluno considera seus conhecimentos de informática.
22	Escolarid_Mãe	Polinominal	Até 5º EF Do 6º ao 9º EF EM ES ET Pós-gra NS	Nível de escolaridade da mãe ou responsável. 1. Até 5º EF (Do 1º ao 5º ano do Ensino Fundamental) 2. Do 6º ao 9º EF (Do 6º ao 9º ano do Ensino Fundamental) 3. EM (Ensino Médio completo) 4. ES (Ensino Superior completo) 5. ET (Ensino Técnico completo) 6. Pós-gra (Pós-graduação) 7. NS (Não sei)

23	Escolarid_Pai	Polinominal	Até 5º EF Do 6º ao 9º EF EM ES ET Pós-gra NS	Nível de escolaridade do pai ou responsável. 1. Até 5º EF (Do 1º ao 5º ano do Ensino Fundamental) 2. Do 6º ao 9º EF (Do 6º ao 9º ano do Ensino Fundamental) 3. EM (Ensino Médio completo) 4. ES (Ensino Superior completo) 5. ET (Ensino Técnico completo) 6. Pós-gra (Pós-graduação) 7. NS (Não sei)
24	Comput_ próprio	Polinominal	Comp. Sem Comp. Com Não Comp.	O aluno possui computador em casa ou não, se sim, com ou sem internet. 1. Comp. Sem (Computador sem internet) 2. Comp. Com (Computador com internet) 3. Não possuo computador
25	Acesso_Internet	Polinominal	Alta Média Baixa	Frequência que o aluno acessa a internet. 1. Alta (Diariamente) 2. Média (Semanalmente ou só nos finais de semana) 3. Baixa (Ocasionalmente ou não acessa)
26	Nº Filhos	Numérico Inteiro	0 1 2 3	Se o aluno(a) não tem filhos(as) ou se tem, quantos. 1. 0 - Não tem filhos 2. 1 - Sim, um filho(a) 3. 2 - Sim, dois filhos(as) 4. 3 - Sim, três filhos ou mais
27	Renda	Polinominal	$RFPC \leq 0{,}5\ SM$ $0{,}5 \leq RFPC < 1{,}0\ SM$ $1{,}0 \leq RFPC < 1{,}5\ SM$ $1{,}5 \leq RFPC < 2{,}5\ SM$ $2{,}5 \leq RFPC < 3{,}0\ SM$ $3{,}0 \leq RFPC$	Renda familiar per capita. Faixas de renda usadas pelo SISTEC e pelos relatórios anuais de dados indicadores de gestão, feitos pela SETEC. 1. $RFPC \leq 0{,}5\ SM$ 2. $0{,}5 \leq RFPC < 1{,}0\ SM$ 3. $1{,}0 \leq RFPC < 1{,}5\ SM$ 4. $1{,}5 \leq RFPC < 2{,}5\ SM$ 5. $2{,}5 \leq RFPC < 3{,}0\ SM$ 6. $3{,}0 \leq RFPC$

28	Depende_Renda	Numérico Inteiro	1 2 3 4 5	Número de pessoas que usufruem da renda familiar mensal. 1- Uma 2- Duas 3- Três 4- Quatro 5- Cinco ou mais
29	Trabalho	Polinominal	NTrab Trab. Eventual Trab até 20h/s Trab 21 a 30h/s Trab 31 a 40h/s Trab+40h/s	Aluno(a) que realiza atividade remunerada e a quantidade de horas semanais 1. NTrab (Não trabalha) 2. Trab. Eventual (Sim, mas é trabalho eventual) 3. Trab até 20h/s (Sim, até 20 horas por semana) 4. Trab 21 a 30h/s (Sim, de 21 a 30 horas por semana) 5. Trab 31 a 40h/s (Sim, de 31 a 40 horas por semana) 6. Trab+40h/s (Sim, mais de 40 horas por semana)
30	Auxílio_Gov	Binominal	Sim Não	Informação se a família participa ou não de algum Programa de renda do governo.
31	Moradia	Polinominal	Alugada Própria em pagamento Própria quitada Cedida Área verde	Situação da moradia do(a) estudante: 1. Alugada 2. Própria em pagamento 3. Própria quitada 4. Cedida 5. Área verde
32	Transporte	Polinominal	A pé / carona / bicicleta Transp. coletivo Transp. próprio Locado	Meio de transporte utilizado pelo(a) aluno(a) para chegar ao *Campus*. 1. A pé / carona / bicicleta 2. Transp. coletivo 3. Transp. próprio 4. Locado
33	Ativ_Física	Polinominal	Sim Não NR	Realiza ou não atividades físicas ou esportivas. 1. Sim (Realiza atividades físicas/esportivas) 2. Não (Não realiza atividades físicas/esportivas) 3. NR (Não responderam)

34	Leitura_livros	Polinominal	Nenhum De 1 a 2 livros De 2 a 4 livros De 4 a 6 livros De 6 a 8 livros Acima de 8 livros	Quantidade de livros lidos por ano. 1. Nenhum 2. De 1 a 2 livros 3. De 2 a 4 livros 4. De 4 a 6 livros 5. De 6 a 8 livros 6. Acima de 8 livros
35	Neces_Especial	Binominal	Sim Não	O(a) aluno(a) possui alguma necessidade especial.
36	Doença	Binominal	Sim Não	O(a) aluno(a) possui alguma doença crônica.
37	Acompanha_Psico	Binominal	1. Sim 2. Não	O(a) aluno(a) teve acompanhamento psicológico ou psiquiátrico.

Fonte: construção da autora.

APÊNDICE B - ALTERAÇÕES REALIZADAS NA BASE DE DADOS

ALTERAÇÕES REALIZADAS NA BASE DADOS			
Nº	**ATRIBUTO**	**MODIFICAÇÕES**	**Nº OBJETOS**
1	GERAL	Planilha base extraída do SIFRS, sem alterações	938
2	GERAL	Retirada dos alunos ingressantes de 2018-1 e 2018-2. Total de 134 objetos	804
3	GERAL	Retirada dos alunos que não responderam Questionário Sociodemográfico (QS). Total de 376	428
4	PERÍODO	Inclusão da informação "período de ingresso", em 22 objetos que estavam com a informação em branco. Consulta da informação no SIA.	428
5	RENDA FAMILIAR	Padronização dos dados sobre renda per capita para: RFPC $\leq$ 0,5 SM; 0,5 $\leq$ RFPC $<$ 1,0 SM; 1,0 $\leq$ RFPC $<$ 1,5 SM; 1,5 $\leq$ RFPC $<$ 2,5 SM; 2,5 $\leq$ RFPC $<$ 3,0 SM; 3,0 $\leq$ RFPC	428
6	FAIXA ETÁRIA	O atributo foi criado a partir da data de nascimento, com a qual foi feito o cálculo da Idade, utilizando a função: INT((AGORA)- F2)/365,25). A idade, um atributo numérico contínuo, foi categorizado dentro de uma das faixas etárias estabelecidas: 1- de 18 a 19 anos; 2- de 20 a 24 anos; 3- de 25 a 29 anos; 4- de 30 a 39 anos; 5- de 40 a 49 anos; 6- mais de 50 anos.	428
7	RECEBE BAE	A informação se recebe ou não Bolsa de Auxilio Estudantil (BAE) foi padronizado para sim e não	428
8	SITUAÇÃO	Correção da situação de 12 alunos para "Formados", após conferência com nova planilha base, enviada pelo S.T.I.	428
9	FORMA DE INGRESSO e MODALIDADE DE INGRESSO	Separação das informações forma de ingresso e modalidade de ingresso em dois atributos diferentes, para diminuir o número de valores nos atributos e fazer a diferenciação em forma e modalidade. Preenchimento dos dados que estavam faltando, em objetos relacionado ao curso de Automação Industrial. Informações extraídas da planilha Alunos Cursos Superiores e do SIA.	428
10	ESTADO	Retirada do atributo com a informação do estado onde mora, por serem todos os alunos moradores do RS.	428

11	DISCIPLINAS CURSADAS	Inclusão da situação final nas disciplinas cursadas no primeiro semestre, de cada aluno; copiadas individualmente (aluno por aluno) da planilha repassada pelo programador do S.I.A, na Reitoria do IFRS. O objetivo é calcular o aproveitamento no primeiro semestre.	
12	GERAL	Exclusão de um objeto, por estar duplicado, com as mesmas informações.	427
13	GERAL	Excluído de um objeto, pois aluno é falecido.	426
14	GERAL	Excluídos 6 objetos NÃO computados como alunos, pois desistiram da vaga antes do início das aulas, tendo sido matriculado outro aluno.	420
15	Nº DISCIPLINAS e % de APROVEITA-MENTO	Criação de dois novos atributos a partir das disciplinas cursadas no primeiro semestre: o número de disciplinas cursadas e o percentual de aproveitamento de cada aluno no primeiro semestre.	420
16	TIPO DE INSTITUIÇÃO DO NÍVEL ANTERIOR	Transformação do atributo "Escola de Origem" que continha o nome das escolas onde os alunos cursaram o nível anterior ou a forma de certificação, para o atributo " Tipo de instituição do nível anterior" com valores: 1- Escola pública, 2 - Escola privada e 3- "ENCCEJA/ENEM". As várias formas, as quais estavam registradas a informação sobre a certificação pelo ENEM ou ENCCEJA, foram padronizadas para "ENCCEJA/ENEM".	420
17	SITUAÇÃO	Alteração da situação de quatro alunos de "Desligado-Mudança de Curso" para "Desligado-Desistência". O motivo da desistência foi a mudança de curso, mas os alunos desistiram do 1º curso.	420
18	ESTADO CIVIL	Busca da informação no S.I.A e inclusão do estado civil de 20 alunos, para os quais constava ND (Não disponível). Redução de cinco para três valores diferentes com a junção de Casado/União estável e Divorciado/Separado. Sendo possível três valores: 1- Casado/União Estável; 2- Divorciado/Separado; 3- Solteiro	420
19	RAÇA/COR/ ETNIA	Padronização das respostas do atributo "Raça/Cor/Etnia": retirada a flexão de gênero "Preto e Preta" e redução da expressão "Não sabe ou não declarado" para "ND"	420
20	MODALIDADE DE INGRESSO	Substituição da descrição das reservas de vagas (cotas) pelos códigos usados no Edital do Processo Seletivo Unificado para Ingresso Discente no semestre 2019/1, do IFRS.	420
21	% DE APROVEITA MENTO	Filtragem nos percentuais de aprovação e realizado a segunda conferência em relação ao número de disciplinas aprovadas. Feito algumas correções necessárias.	420

22	MUNICÍPIO ONDE MORA	Quinze alunos tinham como município de moradia Bento Gonçalves. Feito alteração em todos, após busca da informação correta no SIA. Redução do total de municípios para cinco. Os municípios contendo menos de 10 alunos moradores foram agrupados com a denominação de “Outro”.	420
23	TEMPO DE CONCLUSÃO DO NÍVEL ANTERIOR	Criação de novo atributo, para isso foi inserida a coluna com o ano de ingresso no *Campus* Canoas sem a identificação do semestre, coluna "x". A coluna "y" com o ano de conclusão do ensino anterior já existia. Para o cálculo do tempo de conclusão do nível de escolaridade anterior (TCNEA) uma nova coluna foi criada contendo a fórmula " =(X-Y). Obtivemos um atributo numérico. Cada valor foi categorizado dentro de uma das seis faixas de tempo estabelecidas, reduzindo a variação de valores, transformando o numérico para nominal.	420
24	ESCOLHA DO CURSO	Redução da variação da quantidade de valores possíveis de 8 para 5. Os valores contendo motivos para escolha do curso, selecionados por menos de 10 alunos, foram agrupados no valor "outro".	420
25	ESCOLHA DA INSTITUIÇÃO	Redução da variação da quantidade de valores possíveis de 7 para 5. Os valores contendo motivos para escolha da instituição, selecionados por menos de 10 alunos, foram agrupados no valor "outro".	420
26	DIFICULDADE EM UMA ÁREA DO CONHECIMENTO	Exclusão do atributo com a informação da área de conhecimento que o aluno teve dificuldade durante a vida escolar. Motivo: O aluno pode escolher, mais de uma, até treze áreas e ainda a opção de "Não ter dificuldades". As combinações das alternativas escolhidas pelos alunos geraram 202 valores diferentes (respostas diferentes), dados com baixa qualidade pela imprecisão da informação.	420
27	COMPUTADOR PRÓPRIO, COM E SEM INTERNET	Redução da redundância entre os valores, através da padronização das respostas sobre ter computador e internet, para apenas três valores possíveis: 1- Comp. Sem (Computador sem internet); 2- Comp. Com (Computador com internet) e 3- Não possuo computador.	420
28	FILHOS	Padronização e transformação do atributo nominal em numérico. O atributo contém a informação se o aluno(a) não tem filhos(as) ou se tem, quantos. Valor 0 - Não tem filhos; 1 - Sim, um filho(a); 2 - Sim, dois filhos(as); 3 - Sim, três filhos ou mais	420
29	REALIZA ATIVIDADES FÍSICAS/ ESPORTIVAS	Além de responder sim ou não o aluno pode apontar quais atividades realiza. A combinação das alternativas escolhidas gerou muitos valores diferentes (muita variação nas respostas). Foi realizada a redução do número de valores, retirando o tipo de atividade física, mantendo apenas as informações: Sim, realiza atividade física; Não, não realiza atividade física e NR, não respondeu. Com isso foi reduzido, também, as redundâncias e as inconsistências entre os valores.	420
30	LIVROS LIDOS POR ANO	O atributo com a informação do número de livros lidos por ano, apresenta como dados faixas de quantidade de livros lidos por ano. Essas faixas foram padronizadas pois se sobrepunham. Ex: "De 6 a 8 livros" e "De 7 a 8 livros"	420

31	NECESSIDADE ESPECIAL	Padronização das respostas para "Sim" e "Não". Foi retirada a informação sobre o tipo de deficiência, pois era possível escolher entre deficiências elencadas e escrever outras, gerando um número grande de valores e inconsistência quando era assinalado "não" e escrito o nome alguma deficiência.	420
32	DOENÇA CRÔNICA	Padronização das respostas para "Sim" e "Não". Foi retirada a informação sobre o tipo de doença. Além de gerar muitos valores sem relevância, havia possibilidade de marcar "não" e escrever o nome de qualquer doença, gerando inconsistência.	420
33	ACOMPANHAMENTO PSICOLÓGICO	Padronização das respostas para "Sim" e "Não". Foi retirada a informação sobre o tempo de acompanhamento para evitar inconsistência e informações irrelevantes. Para o tempo de acompanhamento o questionário permitia o preenchimento, mesmo quando assinalado o "não".	420
34	GERAL	Exclusão das colunas com "data de nascimento" e "idade", usadas para padronização do atributo " Faixa etária"	420
35	GERAL	Exclusão das colunas com notas das disciplinas do primeiro semestre, usadas para padronização do atributo "% de aproveitamento"	420
36	GERAL	Complementação das informações "Ano de conclusão do nível anterior", "cidade e bairro" onde mora, de 20 objetos. Busca das informações feita no S.I.A. e nas pastas físicas de 5 alunos.	420
37	BAIRRO	Exclusão do atributo Bairro, por ter como valores, 140 bairros, das seis cidades informadas nos endereços dos alunos. Não foi encontrada uma forma de reduzir esse número, para ter uma maior precisão e relevância dos dados, sem que isso não influenciasse a análise dos algoritmos.	420
38	TRANCAMENTO	Inclusão da informação sobre o Trancamento ou não do segundo semestre do curso (Sim ou Não). Incluída de forma manual a partir das informações extraídas da Planilha Alunos Cursos Superiores.	420
39	GERAL	Exclusão das colunas "Ano de ingresso no IFRS", "Ano de conclusão do nível médio", "Tempo de conclusão do ensino médio", usadas para padronização do atributo TCNAE (Tempo de Conclusão do Nível Anterior da Escolaridade) em faixas temporais.	420
40	GERAL	Exclusão do atributo Matriz curricular. O aluno não pode escolher a matriz de ingresso, escolhe o curso e é matriculado na matriz que está ativa, as mais antigas foram extintas sendo os alunos migrados para as mais atuais. A Matriz curricular, portanto, é uma característica do curso e não do aluno.	420

41	GERAL	Reorganização dos intervalos de tempo, das faixas de Tempo de Conclusão do Nível de Escolaridade Anterior, pois estavam sobrepostas.	420
42	CURSO	Redução do nome dos cursos para siglas: TADS, TAI, Tlog.	420
43	NÍVEL DE ESCOLARIDADE DOS PAIS	Os atributos "escolaridade da mãe" e "escolaridade do pai", foram reduzidos o número de valores, mantendo-se apenas as alternativas de escolaridade completa, com exceção do ensino fundamental, sendo remetidas as respostas do nível incompleto para o anterior, por consequência completo.	420
44	ATIVIDADES ALÉM DO ESTUDO	Este tributo foi excluído, pois neste item do QS, há a possibilidade de escolha entre várias atividades e ainda escrever outras que não estejam contempladas. Isso gera uma gama muito grande de informações (dados) de baixa qualidade e imprecisos.	420
45	FONTE DE INFORMAÇÃO	Este atributo foi excluído, pois neste item do QS, também há a possibilidade de múltipla escolha, de um total de sete opções incluindo "Nenhuma das opções". Isso gera uma gama muito grande de informações (dados) de baixa qualidade e imprecisos.	420
46	MEDICAÇÃO DE USO CONTROLADO	Este atributo foi excluído. Neste item havia a possibilidade de assinalar "sim" e "não" e no caso afirmativo, escreve qual medicamento faz uso. Muitos alunos marcaram as duas alternativas ao mesmo tempo e escreveram mais de um medicamento. Devido à imprecisão nos dados e existir um atributo similar, QS 22 sobre doença crônica, o atributo foi excluí-lo.	420
47	ACESSO A INTERNET	Junção de valores redundantes e transformação do atributo. Frequência que o aluno acessa a internet: 1 - Alta (Diariamente); 2 - Média (Semanalmente ou só nos finais de semana) e 3- Baixa (Ocasionalmente ou não acessa).	420
48	SITUAÇÃO	Redução dos valores do atributo em três categorias: Formados; Regulares e Desligados. Todos valores com a designação de desligado (Desligado-Desistência, Desligado-Evasão, Desligado-Jubilado e Desligado-Transferência) estão inseridos no rótulo “Desistência”. O valor Trancado Total, foi inserido no rótulo “Regular”.	420
49	GERAL	Retirada de espaços antes e depois dos valores e feito alinhamento à esquerda.	420

Fonte: construção da autora.

ANEXO A - LEGENDA DAS MODALIDADES DE INGRESSO

LEGENDA DO QUADRO DE VAGAS	
PcD	Pessoa com Deficiência
PPI	Preto, Pardo e Indígena.
C1	Acesso Universal
C2	Pessoa com Deficiência (PcD) que tenha cursado integralmente o Ensino Médio em escola pública, autodeclarada/autodeclarado negra/negro (preta/preto, parda/pardo) ou indígena (PPI), com renda familiar bruta per capita igual ou inferior a 1,5 salário mínimo.
C3	Candidata/Candidato que tenha cursado integralmente o Ensino Médio em escola pública, autodeclarada/autodeclarado negra/negro (preta/preto, parda/pardo) ou indígena (PPI), com renda familiar bruta per capita igual ou inferior a 1,5 salário mínimo.
C4	Pessoa com Deficiência (PcD) que tenha cursado integralmente o Ensino Médio em escola pública, com renda familiar bruta per capita igual ou inferior a 1,5 salário mínimo.
C5	Candidata/Candidato que tenha cursado integralmente o Ensino Médio em escola pública, com renda familiar bruta per capita igual ou inferior a 1,5 salário mínimo.
C6	Pessoa com Deficiência (PcD) que tenha cursado integralmente o Ensino Médio em escola pública, autodeclarada/autodeclarado negra/negro (preta/preto, parda/pardo) ou indígena (PPI), independente de renda.
C7	Candidata/Candidato que tenha cursado integralmente o Ensino Médio em escola pública, autodeclarada/autodeclarado negra/negro (preta/preto, parda/pardo) ou indígena (PPI), independente de renda.
C8	Pessoa com Deficiência (PcD), que tenha cursado integralmente o Ensino Médio em escola pública, independente de renda.
C9	Candidata/Candidato que tenha cursado integralmente o Ensino Médio em escola pública, independente de renda.
C10	Acesso Universal e Reserva de Vagas para Pessoa com Deficiência (PcD), independente de ter cursado integralmente o Ensino Médio em escola pública.

Fonte: Edital IFRS nº 64/2018, Processo Seletivo Unificado para Ingresso Discente no semestre 2019/1 (IFRS, 2018c).