UNIVERSIDADE FEDERAL DE SANTA CATARINA

CENTRO TECNOLÓGICO

PROGRAMA DE PÓS-GRADUAÇÃO EM MÉTODOS E GESTÃO EM AVALIAÇÃO

Diego Trentin Mioranza

Evasão nos Cursos de Graduação do Instituto Federal Catarinense: um estudo a partir da Mineração de Dados

Florianópolis

2020

Diego Trentin Mioranza

Evasão nos Cursos de Graduação do Instituto Federal Catarinense: um estudo a partir da Mineração de Dados

Dissertação submetida ao Programa de Pós-Graduação em Métodos e Gestão em Avaliação da Universidade Federal de Santa Catarina para a obtenção do título de Mestre em Métodos e Gestão em Avaliação.
Orientador: Prof. Marcelo Menezes Reis, Dr.

Florianópolis
2020

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Mioranza, Diego
Evasão nos Cursos de Graduação do Instituto Federal Catarinense : um estudo a partir da Mineração de Dados / Diego Mioranza ; orientador, Marcelo Menezes Reis, 2020.
120 p.

Dissertação (mestrado profissional) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Métodos e Gestão em Avaliação, Florianópolis, 2020.

Inclui referências.

1. Métodos e Gestão em Avaliação. 2. Evasão Discente. 3. Mineração de Dados Educacionais. I. Menezes Reis, Marcelo. II. Universidade Federal de Santa Catarina. Programa de Pós Graduação em Métodos e Gestão em Avaliação. III. Título.

Diego Trentin Mioranza

Evasão nos Cursos de Graduação do Instituto Federal Catarinense: um estudo a partir da Mineração de Dados

O presente trabalho em nível de mestrado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Dr. Marcelo Menezes Reis
Universidade Federal de Santa Catarina

Prof. Dr. Pedro Alberto Barbetta
Universidade Federal de Santa Catarina

Prof. Dr. Carlos Andres Ferrero
Instituto Federal de Santa Catarina

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de mestre em Métodos e Gestão em Avaliação.

Documento assinado digitalmente
Marcelo Menezes Reis
Data: 09/06/2020 10:05:09-0300
CPF: 888.917.559-15

Coordenação do Programa de Pós-Graduação

Documento assinado digitalmente
Marcelo Menezes Reis
Data: 09/06/2020 10:05:36-0300
CPF: 888.917.559-15

Prof. Marcelo Menezes Reis, Dr.
Orientador

Florianópolis, 2020.

Este trabalho é dedicado à minha esposa.

RESUMO

Este trabalho tem como tema a evasão discente no ensino superior de um instituto federal de educação, ciência e tecnologia. O objetivo deste estudo é propor uma metodologia para identificar precocemente os alunos com maior propensão a evadir de modo a tornar possível que ações de intervenção sejam tomadas em tempo hábil para mitigar o problema. A pesquisa utiliza técnicas de Mineração de Dados Educacionais (EDM) para classificar os discentes ingressantes de acordo com sua propensão a evadir a partir de dados cadastrais e acadêmicos dos estudantes. Os resultados demonstram que é possível prever, com base nos dados cadastrais obtidos antes do início das aulas, a situação da matrícula dos alunos com 66,04% de acurácia e predizer com acerto de 66,3% os estudantes que têm sua matrícula cancelada. Após o primeiro semestre letivo a previsão se torna mais precisa, com os algoritmos de mineração de dados atingindo 75,52% de acerto geral e 67,53% de acerto das evasões. Além da previsão, foi possível identificar quais as variáveis mais relevantes para a permanência ou não no curso.

Palavras-chave: Evasão Discente. Mineração de Dados Educacionais. Algoritmos de Classificação.

ABSTRACT

This work has as its theme the student dropout in higher education of a Brazilian federal institute of education, science and technology. The goal of this study is to propose a methodology to identify early the students most likely to dropout the courses, in order to do intervention and mitigate the problem in a short period of time. The research applies Educational Data Mining (EDM) techniques to classify incoming students according to their propensity to evade using students' registration and academic data. The results show that it is possible to predict, based on the registration data obtained before the beginning of classes, the situation of students' enrollment with 66.04% accuracy and predict with 66.3% correctness of students who have their enrollment canceled. After the first semester, the forecast becomes more accurate, with the data mining algorithms reaching 75.52% of general accuracy and 67.53% of dropout accuracy. In addition to the forecast, it was possible to identify which variables are most relevant for student retention.

Keywords: Student Dropout. Educational Data Mining. Classification Algorithms.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

INTRODUÇÃO

É incontestável a importância do papel que a educação exerce na sociedade. Ela é responsável por preparar os indivíduos para o exercício da cidadania e qualificá-los para o trabalho (BRASIL, 1996). Os benefícios de uma população educada são reconhecidos universalmente, transcendem as questões materiais e financeiras e influenciam positivamente todos os aspectos da vida em sociedade.

No Brasil, a educação é um direito social expresso no artigo 205 da Constituição Federal, cabendo ao Estado e à família o dever de garanti-lo (BRASIL, 1988). A atuação governamental no âmbito da educação segue as diretrizes apresentadas no Plano Nacional de Educação – PNE, aprovado em 2014 e com duração de 10 anos. O PNE tem como objetivo erradicar o analfabetismo, universalizar o atendimento escolar e elevar o nível de escolaridade da população (BRASIL, 2014).

No entanto, muitos obstáculos dificultam que estes objetivos sejam alcançados e que se tenha garantida a plena formação dos indivíduos. A própria criação de condições adequadas e universais de acesso à educação já se constitui em uma tarefa de difícil consecução, especialmente no Brasil pela vasta dimensão geográfica e por uma economia ainda em desenvolvimento. Além disso, a oferta de vagas por si só não garante o efetivo aproveitamento das matrículas e a permanência dos estudantes, ao que se soma o desafio de promover a melhoria da qualidade do ensino e o êxito acadêmico.

No contexto destas problemáticas educacionais brasileiras a taxa de evasão se configura como um dos indicadores importantes para se pensar as condições do estudante dar continuidade aos seus estudos, bem como para avaliar a efetividade das políticas de promoção da educação. O fato de muitos ingressantes sequer finalizarem seus cursos é um problema que há tempos preocupa os gestores e estudiosos da área, tendo se tornado objeto de inúmeras investigações que envolvem não só suas causas e consequências (NERY, 2009; AMBIEL, 2018; SILVA, 2018) , mas também sua correta quantificação. É sobre este último aspecto que trata este trabalho.

A evasão afeta em maior ou menor grau todos os níveis de ensino no Brasil. Ainda que o ensino básico obrigatório compreenda dos 4 aos 17 anos (BRASIL, 2009), a taxa de evasão pode superar 12% no ensino médio, segundo informações publicadas no site do Ministério da Educação (MEC, 2017). No ensino superior a realidade é ainda mais preocupante, de acordo

com levantamento realizado pelo Sindicado das Entidades Mantenedoras de Estabelecimentos de Ensino Superior no Estado de São Paulo (SEMESP, 2019) a taxa de evasão registrada para os cursos de graduação no Brasil supera os 25%. Inclui-se aí, também os cursos de graduação dos Institutos Federais, instituições criadas com foco inicial no ensino médio técnico, mas que ampliaram sua atuação contemplando o ensino superior e a pós-graduação. Este trabalho se insere neste cenário uma vez que teve como objeto os cursos de graduação do Instituto Federal Catarinense (IFC), distribuídos em 14 campi no estado de Santa Catarina.

Antes de tratar sobre o objeto especifico deste trabalho, a quantificação da evasão no IFC, cabe salientar alguns modelos que historicamente foram usados para descrever as condições em que o fenômeno da evasão acontece. Destaca-se inicialmente o modelo de Spady (1970), precursor dos mais importantes estudos sobre o tema na literatura especializada e para o qual a escola deve ser compreendida como um sistema social, com sua própria estrutura, valores e dinâmicas. Para o autor a evasão escolar representa uma forma de fuga daqueles indivíduos que não se sentem afiliados à comunidade escolar ou não se integram ao conjunto de valores ali praticados. Esse entendimento foi proveniente da aplicação da Teoria do Suicídio de Durkheim (1961) - que tratava dos aspectos psicológicos da fuga dos indivíduos da própria sociedade - ao contexto educacional, considerando a escola como o microcosmo das relações sociais.

Mais tarde, Tinto (1975) aprimorou o estudo de Spady acrescentando o conceito econômico de custo benefício ao modelo. Segundo Tinto, além dos aspectos psicológicos da aplicação do modelo de Durkheim ao âmbito escolar, a decisão de evadir estava também condicionada à percepção individual de atividades alternativas mais atraentes. Essa noção explica que o estudante busca investir seu tempo, energia e foco naquilo que considera lhe trazer mais benefício do que estar na escola.

Atualmente, no Brasil, muitos estudos têm sido desenvolvidos para auxiliar a compreensão acerca das causas da evasão (AMBIEL, 2015; NERI, 2009). Os autores costumam dividir as causas em dois aspectos, os institucionais e os individuais. Os aspectos institucionais são aqueles que dizem respeito à própria escola: infraestrutura, qualidade do corpo docente, serviços oferecidos, grade curricular, ou seja, todas as características que podem ser influenciadas pelos gestores, tanto no que se refere ao ambiente e as formas de interação social, como ao desenvolvimento educacional (AMBIEL, 2015). Quanto aos aspectos individuais, ganham destaque aqueles relacionados aos desejos, interesses e expectativas do estudante bem

como sua bagagem familiar, experiência pregressa e características vocacionais (TINTO, 1975; LIMA, PIMENTEL, 2017).

Após estes estudos iniciais que contribuíram para o desenvolvimento de um necessário arcabouço teórico para entender o complexo fenômeno da evasão e com o desenvolvimento de modernas ferramentas computacionais abriram-se novos campos de estudo e, por conseguinte, novas maneiras de abordar o problema. Uma delas é a aplicação de modelos estatísticos para medir a propensão a evadir dos estudantes, com destaque para a Mineração de Dados aplicada ao contexto educacional (MANHÃES, 2015; AMARAL, 2016; KANTORSKY ET AL, 2015; FUNCHAL, 2016; AIRES et al, 2017; BITENCOURT, FERRERO, 2019).

Esta forma de abordagem busca quantificar a probabilidade de um aluno ter sua matrícula cancelada concentrando o foco naqueles estudantes com maior chance de evadir e, a partir disso, prever se o estudante vai ou não continuar o curso. Nesta perspectiva, institucionalmente é mais importante que os modelos de mineração de dados apresentem melhor desempenho na previsão das evasões. Possíveis erros de previsão dos estudantes que permanecem nos cursos acabam sendo menos relevantes que falhas na previsão daqueles que evadem. Por isso, boa parte dos estudos com este objetivo adotam métricas específicas para avaliar os modelos que apresentam o melhor desempenho na identificação das evasões.

Para além dos estudos acadêmicos, a (não) continuidade dos estudos é um problema governamental que inclui ações de fiscalização e gestão pública. À medida que as taxas de evasão são publicadas e o tema ganha relevância, os órgãos de controle brasileiros, como Ministério Público (MP) e Tribunal de Contas da União (TCU), passam a monitorar com mais intensidade as instituições de ensino públicas e exigir dos gestores ações concretas para redução do problema. Como exemplo, em 2013 o TCU realizou auditoria na Rede Federal de Educação Profissional objetivando avaliar a atuação dos Institutos Federais de Educação com relação à caracterização da evasão e as medidas para reduzi-la. A partir dessa averiguação recomendou-se que a Secretaria de Educação Profissional e Tecnológica (Setec/MEC) instituísse, em conjunto com os Institutos Federais, um plano voltado ao tratamento da evasão na rede (TCU, 2013).

No IFC, o quadro de evasão é impreciso e de complexa quantificação devido, principalmente, à implantação de um sistema de gestão integrado (SIGA) ter acontecido apenas em 2014. Até então coexistiam diferentes sistemas de gestão acadêmica, o que dificultou a integração das informações principalmente aquelas anteriores a 2014. Então, desde 2014, todos

os dados acadêmicos dos cursos superiores encontram-se em um sistema único para todos os *campi* conferindo maior confiabilidade aos dados.

Para os demais níveis de ensino, principalmente aqueles técnicos de nível médio que representam pelo menos 50% dos estudantes do IFC, o setor responsável pela tecnologia da informação no IFC continua trabalhando na customização do sistema de modo a atender as especificidades dos diferentes níveis de ensino. A completa implantação do SIGA para todos os níveis de ensino está prevista para o ano de 2020. Por isso, foram contemplados no presente estudo apenas os cursos superiores do IFC.

De acordo com um relatório interno (IFC, 2018) a maior parte da evasão nos cursos superiores do IFC acontece no primeiro ano de curso, com alguns cursos chegando a superar a taxa de 50% de evasão no primeiro ano letivo. Esta parece ser a realidade de outras instituições de ensino superior. Segundo Silva Filho et al (2007) a taxa de evasão no primeiro ano do curso chega a ser até três vezes maior que nos anos subsequentes. Mais recentemente, em pesquisa com alunos do curso de Estatística da Universidade Federal da Paraíba – UFPB, Silva et al (2018) expõe que a chance dos ingressantes desistirem do curso após o primeiro ano é de 45,85%. A particularidade da evasão incidir com maior intensidade nos primeiros semestres do curso suscita a necessidade das instituições tomarem medidas sem demora para mitigar o problema.

Nesse sentido, em atendimento à recomendação do TCU, a Setec/MEC publicou a Nota Informativa nº 138 que orienta a elaboração dos Planos Estratégicos Institucionais para a Permanência e Êxito dos Estudantes. De acordo com a nota, todos os Institutos Federais deveriam elaborar o referido plano contemplando diagnósticos das causas da evasão e implementação de ações de modo a ampliar as oportunidades de permanência e êxito dos estudantes.

Em abril de 2019 o IFC aprovou seu Plano de Permanência e Êxito (IFC, 2019). A partir de um diagnóstico quantitativo e qualitativo foram propostas estratégias institucionais gerais e ações de intervenção específicas para cada curso. Na etapa de diagnóstico verificou-se que a maior parte da evasão era identificada sem tempo hábil para que os envolvidos tomassem providências, ficando clara a necessidade da identificação precoce do fenômeno. Um dado interessante que ilustra a dimensão da evasão logo nos primeiros semestres letivos é que da totalidade dos ingressantes em cursos superiores do IFC no ano de 2017, 18,4% não fizeram

rematrícula para o segundo semestre. E, ao final do segundo semestre, 41,5% não se rematricularam para o semestre seguinte.

A partir das questões apresentadas pergunta-se: Seria possível identificar precocemente os alunos propensos a evadir dos cursos de graduação presenciais do IFC? Seria a *Educational Data Mining* – EDM alternativa efetiva para auxiliar neste processo?

O desenvolvimento deste trabalho seguirá com a exposição dos seus objetivos, passando-se para a fundamentação teórica de alguns conceitos de evasão e de como mineração de dados se insere nesse contexto, por fim, a aplicação da abordagem proposta no âmbito do Instituto Federal Catarinense.

Estruturalmente o trabalho está dividido da seguinte forma: no primeiro capítulo estão descritos os objetivos e a justificativa para este trabalho. No segundo capítulo é apresentada a fundamentação teórica do trabalho, são descritos os conceitos de evasão, qual a realidade da evasão no âmbito do ensino superior dos institutos federais de educação, o contexto em que o estudo se desenvolve com a caracterização da instituição de ensino trabalhada e como a EDM pode auxiliar no combate à evasão. No capítulo 3 é apresentada a metodologia adotada bem como as etapas da metodologia aplicadas aos IFC. O capítulo 4 concentra os resultados encontrados ao longo do processo de mineração de dados realizado. E, por fim, o capítulo 5 traz as conclusões do estudo, as limitações e considerações a respeito de trabalhos futuros.

1.1 OBJETIVOS

Nas seções abaixo estão descritos o objetivo geral e os objetivos específicos da dissertação.

1.1.1 OBJETIVO GERAL

Propor metodologia para identificar precocemente os alunos propensos a abandonar os estudos nos cursos de graduação presenciais do IFC – Instituto Federal Catarinense, utilizando técnicas de mineração de dados.

1.1.2 OBJETIVOS ESPECÍFICOS

• Identificar as variáveis mais relevantes para o entendimento do problema;

• Comparar as técnicas de mineração de dados aplicadas;

• Identificar o(s) algoritmo(s) que apresentar(em) maior acurácia e especificidade de classificação.

1.2 JUSTIFICATIVA

O tema da evasão discente, a despeito da significativa quantidade de estudos realizados, ainda se constitui num empecilho para o desenvolvimento educacional no país. Dessa forma, o fato de fazer parte do quadro funcional do IFC e, portanto, ter acesso às informações institucionais, além do contato frequente com a realidade enfrentada na oferta de ensino superior despertaram especial interesse em contribuir para uma gestão mais eficiente das ações de permanência e êxito acadêmico.

A relevância da questão também pode ser vista pela ótica do gasto público necessário para manter uma estrutura de oferta de vagas, que em muitos casos tornam-se ociosas. Segundo informações da própria Rede Federal de Educação Profissional, Científica e Tecnológica (PNP, 2019) o gasto corrente por matrícula do IFC em 2018 foi de 18.540,83 reais. Embora esse valor não diga respeito especificamente às matrículas de curso superior, pois engloba também outros níveis de ensino, demonstra a magnitude do problema e, ao mesmo tempo, representa um chamado para contribuir no sentido de alterar essa realidade.

1.3 DELIMITAÇÃO DO TRABALHO

O presente trabalho foi desenvolvido com informações cadastrais e acadêmicas dos estudantes de ensino superior presencial do Instituto Federal Catarinense – IFC. Embora a evasão seja um problema que afeta praticamente todos os níveis de ensino, foi necessário restringir a análise deste estudo aos dados do ensino superior devido à falta de disponibilidade de dados para os demais níveis de ensino. Como evidenciado anteriormente, na parte introdutória do estudo, atualmente apenas os registros dos estudantes de cursos superiores estão totalmente informatizados no IFC.

Outra definição que delimita o objeto deste trabalho é o período de tempo analisado. A base de dados utilizada compreende registros a partir de 2010, ano de criação do instituto, até o segundo semestre de 2018. No entanto, é a partir da implantação de um sistema de gestão

integrado em 2014 que os registros acadêmicos passaram a ser tratados de forma institucional, sem a coexistência de controles paralelos e específicos para cada *campus*. Neste mesmo ano o IFC aderiu ao SISU, o que trouxe informações adicionais em um formato uniforme para processo de seleção.

Estes dois aspectos trouxeram maior confiabilidade para os registros acadêmicos da instituição, pois forçaram a adoção de procedimentos unificados de cadastros e armazenamento centralizado das bases de dados. Por estas razões dados anteriores a 2014 não foram utilizados neste trabalho.

Além disso, sabe-se que a evasão tem maior incidência no primeiro ano de estudos (SILVA FILHO ET AL, 2007; IFC, 2017 e SILVA ET AL, 2018) com significativa redução das taxas de evasão para os anos subsequentes. Com base nessas evidências os dados de ingressantes mais recentes (2018) também foram descartados. Portanto, foram objeto deste estudo dados cadastrais e acadêmicos registrados durante 4 anos, 2014 a 2017.

Outro aspecto que precisa ser abordado é o tipo de evasão que será considerado no estudo. Dentre os conceitos de evasão que serão detalhados no capítulo 3, este trabalho adota o conceito de evasão do curso em virtude das informações e dados disponíveis e selecionados do IFC, bem como da necessidade de se estabelecer um recorte de dados passíveis de serem analisados no tempo disponível para a pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

A fundamentação teórica deste estudo está dividida em quatro seções. Inicialmente tratar-se-á dos conceitos de evasão abordando trabalhos seminais para o entendimento do conceito e a definição oficial de evasão em nível nacional. Em seguida será retratada a situação atual da evasão no âmbito dos Institutos Federais de Educação com foco nos cursos presenciais de ensino superior. Na terceira seção, será apresentado o IFC, instituição em que se desenvolveu o estudo, com ênfase no Plano Estratégico de Permanência e Êxito. Por fim, será tratada a conceituação da mineração de dados e sua aplicação no contexto educacional.

2.1 CONCEITOS DE EVASÃO

Já em 1975, Tinto (1975) apontava a importância de uma adequada delimitação do conceito de evasão para entender o problema. Segundo ele os resultados contraditórios encontrados nos estudos até então eram provenientes da não distinção entre a evasão permanente e aquela derivada de uma saída temporária ou transferência para outra instituição. Em sua avaliação, esse tipo de falha conceitual poderia impactar negativamente nas políticas para o ensino superior.

Outra questão importante foi levantada por Bueno (1993), que definiu a evasão como sendo a decisão espontânea do estudante de abandonar os estudos, não podendo ser confundida com “exclusão”, que seria a incapacidade da escola e tudo que a cerca de fornecer ao indivíduo condições de estudo. Tampouco, segundo Ristoff (1999, p. 125), mistura-se com a ideia de “mobilidade”, que se refere à migração entre cursos ou instituições, resultado do próprio aproveitamento das “revelações que o processo natural do crescimento dos indivíduos faz sobre suas reais potencialidades.”

Ainda hoje não há um consenso a respeito do conceito de evasão. No entanto, uma iniciativa importante, proveniente do âmbito governamental, foi a formação da Comissão Especial de Estudos sobre a Evasão nas Universidades Públicas Brasileiras, encarregada de aprofundar o tema e contribuir concretamente para que as Instituições Federais de Ensino Superior - IFES reduzissem suas taxas de evasão. Esta comissão, composta por representantes do MEC e indicados pelas IFES, deu origem a definição de evasão oficialmente aceita até hoje (MEC, 1996), e estabeleceu três tipos, que são:

> 1. Evasão do 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;
> 2. Evasão da instituição: quando o estudante desliga-se da instituição na qual está matriculado;
> 3. Evasão do sistema: quando o estudante abandona de forma definitiva ou temporária o ensino superior (MEC, 1996, p. 15).

Desta tipificação, este estudo abordará o primeiro tipo - evasão do curso - em virtude das informações e dados disponíveis e selecionados do IFC, bem como da necessidade de se estabelecer um recorte de dados passíveis de serem analisados no tempo disponível para a pesquisa. Assim, a evasão da instituição não será abordada pois optou-se por excluir os dados sobre transferências internas e externas, e as informações sobre evasão do sistema, além de possuírem acesso restrito, não possuem a dinâmica de atualização anual que seria necessária para esta análise[1].

Embora não seja objetivo deste estudo discorrer sobre as causas da evasão, muito se tem estudado no sentido de entender as motivações que levam um aluno a abandonar os estudos. O estudo de Tinto (1975) cria um modelo teórico que associa a integração social e a integração acadêmica do estudante como fatores fundamentais para a permanência ou não na instituição. Em grande parte os trabalhos desenvolvidos no contexto nacional evidenciaram a importância e desenvolveram desdobramentos dos atributos elencados por Tinto (BARDAGI; HUTZ, 2008; RIBEIRO, 2005; AMBIEL, 2015, SCHMITT, 2018).

Segundo Tinto (1975), dentro do quadro geral das características individuais que influenciam a decisão de abandonar os estudos encontra-se a bagagem familiar como relevante fator determinante da persistência na escola. Compreende-se bagagem familiar como a estrutura familiar, a capacidade intelectual e financeira dos pais de oferecer ao indivíduo condições mais ou menos adequadas para o desenvolvimento de habilidades individuais. Bardagi e Hutz (2008) também apontam que o apoio parental é decisivo não só no desenvolvimento de habilidades individuais cognitivas, mas também no equilíbrio emocional: capacidade de enfrentar as constantes mudanças no ambiente social, capacidade de lidar com questões psicológicas como ansiedade, frustrações, euforias e capacidade de manter o nível de resiliência necessário para a conclusão do curso.

[1] Os dados que permitem observar a evasão do sistema requerem autorização especial junto ao INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.

Aliado às questões individuais estão os aspectos vocacionais. Em condições ideais espera-se que os indivíduos, ao confrontarem-se com o dilema da escolha de um curso, optem por aquele mais alinhado as suas competências, desejos, expectativas, sonhos, ou seja, aquele curso que representa sua vocação. No entanto, para fazer a escolha mais precisa é necessário um nível de autoconhecimento que, muitas vezes, contrasta com o grau de maturidade de um jovem recém saído da adolescência. A condição ideal assume igualmente que o indivíduo tem pleno conhecimento, não só de todas as alternativas à disposição, mas também das carreiras profissionais a elas relacionadas (NERY, 2009).

Infelizmente não é esta a realidade. É muito comum o estudante, ao se deparar com as disciplinas do curso escolhido, perceber que tomou uma decisão equivocada, que o curso é incompatível com as expectativas formadas em seu imaginário e, portanto, desista do curso (PIMENTEL; LIMA, 2017). Isso demonstra as consequências do (des)alinhamento entre a vocação individual e o curso escolhido.

Outra perspectiva apresentada no trabalho de Tinto (1975), corroborada por Nery (2009) e Pimentel e Lima (2017), está relacionada ao emprego da ideia de custo benefício à decisão de evadir. O conceito econômico que representa mais precisamente a intenção de Tinto é o conceito de custo de oportunidade, apresentado no início do século XX pelo economista austríaco Friederich von Wieser na obra chamada “Valor Natural” de 1914. O custo de oportunidade assume que os agentes econômicos são racionais e que, associada a uma decisão, há um custo decorrente d a renúncia de uma opção alternativa. Na educação, esse conceito pode ser exemplificado como o custo do aluno não trabalhar, não se divertir ou não se dedicar a outra atividade para frequentar a escola e investir seu tempo na aquisição de conhecimento. Portanto, quando o estudante percebe que o benefício de estudar é menor que o custo da renúncia ao trabalho ou prazeres acaba abandonado os estudos.

Essa noção está intimamente relacionada a oportunidades alternativas que se apresentam ao estudante, sejam oportunidades fora do âmbito escolar ou, até mesmo, alternativas de estudos em outras instituições. Neste caso, não se trata de abandono dos estudos, mas sim do curso. Embora não represente uma perda para o sistema de ensino, sob a ótica da instituição refere-se a uma vaga desocupada que dificilmente será preenchida.

Uma forma das instituições de ensino reduzirem o custo de oportunidade associado ao abandono do curso é a adoção de uma política de assistência estudantil que compreenda a

concessão de incentivos financeiros para a permanência nos cursos. Michelotto (2019) verifica que esse tipo de estímulo tem um impacto positivo na probabilidade de permanência estudantil.

Por fim, outro importante aspecto a ser considerado remete às interações sociais dos indivíduos (TINTO, 1975; BARDAGI; HUTZ, 2008, SCHMITT, 2018). O estudante, assim como qualquer pessoa, influencia e tem seu comportamento influenciado pela rede de contatos que compõe seu círculo social, desde aqueles mais próximos como a família e amigos íntimos como também ícones da sociedade ou pessoas bem sucedidas que podem representar alguma referência ao estudante.

Em certo grau, cada uma dessas interações sociais contribui para a forma como o estudante encara a escola. Numa perspectiva mais ampla, toda a sociedade, ao valorizar determinadas profissões em detrimento de outras, emite sinais aos estudantes que, em alguns casos, podem ser determinantes para permanência na escola. Sobretudo quando o estudante percebe que a profissão escolhida não é valorizada pelo senso comum e a insistência na carreira não vai se traduzir no reconhecimento e sucesso esperado.

Num âmbito mais restrito a família atua de modo especial. Como já relatado anteriormente, é ela - a família - que condiciona o conjunto de experiências formativas do indivíduo, no entanto em certas situações a família pode intervir diretamente na escolha do estudante. Seja submetendo o estudante a certas condições materiais, financeiras e sócio ambientais que irão determinar as escolhas do jovem ou até mesmo pressionando-o a seguir carreira específica por crença pessoal.

Diante disso, percebe-se que o fenômeno da evasão se trata de um processo complexo, multifacetado e de difícil compreensão. No entanto, a dificuldade do problema não impede que esforços sejam dispensados para buscar alternativas que reduzam a quantidade de sua incidência e minimizem os impactos do abandono escolar.

2.2 EVASÃO NO ENSINO SUPERIOR DOS INSTITUTOS FEDERAIS DE EDUCAÇÃO E TECNOLOGIA

Desde a promulgação da LDB - Lei de Diretrizes e Bases da Educação Nacional – em 1996, o Brasil vem experimentando forte crescimento na oferta de ensino superior (BRASIL, 1996). Ao longo do tempo o governo federal lançou uma série de programas com objetivo de ampliar o acesso à educação superior e aumentar a taxa de escolarização da população, dentre

eles pode-se destacar o ProUni[2], FIES[3], REUNI[4], SISU[5], UAB[6] e PNAES[7]. Além disso, em 2014 foram determinadas as diretrizes, metas e estratégias para a política educacional durante o período de 2014 a 2024 num documento chamado Plano Nacional de Educação – PNE (BRASIL, 2014). Nele consta um conjunto de metas a serem alcançadas ao longo do tempo, sendo algumas específicas para o ensino superior, como por exemplo:

> Elevar a escolaridade média da população de 18 (dezoito) a 29 (vinte e nove) anos, de modo a alcançar, no mínimo, 12 (doze) anos de estudo no último ano de vigência deste Plano, para as populações do campo, da região de menor escolaridade no País e dos 25% (vinte e cinco por cento) mais pobres, e igualar a escolaridade média entre negros e não negros declarados à Fundação Instituto Brasileiro de Geografia e Estatística - IBGE. (BRASIL, 2014 – Meta 8).

A soma desses esforços resultou em efetivo avanço no número de matrículas no ensino superior. No gráfico abaixo, pode-se observar a magnitude desse crescimento, contexto em que o Brasil passou de 1,87 milhões de matriculados em 1996 para 8,05 milhões em 2016, aumento de 330% em 20 anos. Embora haja registro de expansão nas matrículas tanto na rede pública quanto na rede privada de ensino, foi na rede privada que o crescimento foi mais acentuado.

---

[2] PROUNI – Programa Universidade para Todos: programa que concede bolsas em instituições privadas de ensino superior para estudantes de baixa renda que cursaram ensino médio exclusivamente em escolas pública.

[3] FIES - Fundo de Financiamento ao Estudante de Ensino Superior: fundo de concessão de créditos para estudantes matriculados em instituições não gratuitas.

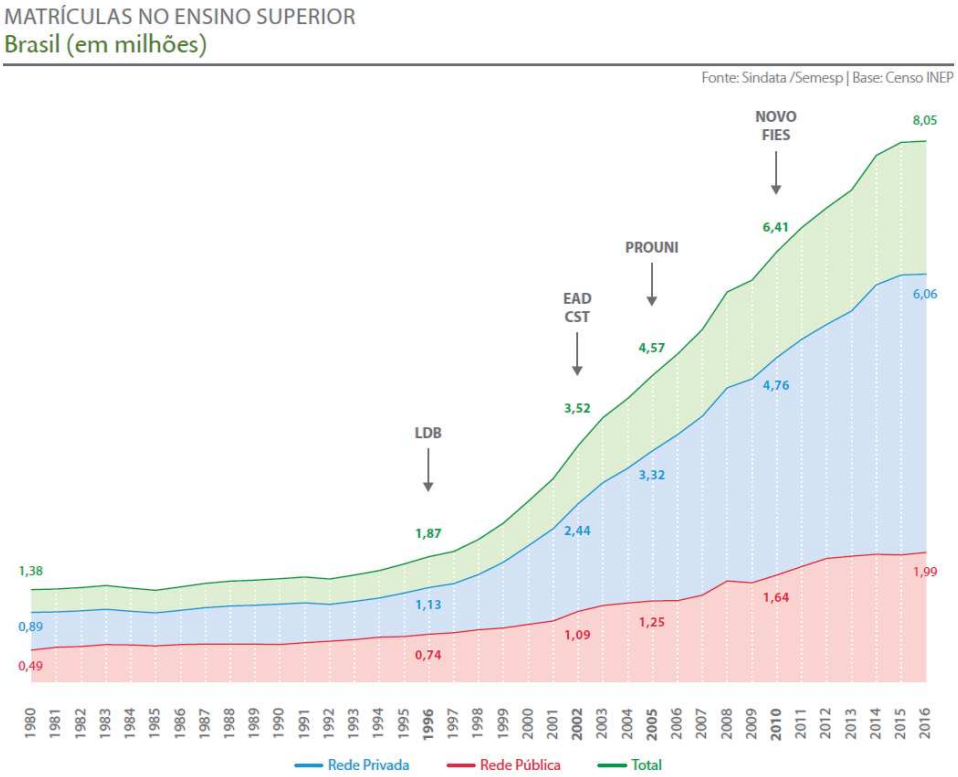
[4] REUNI – Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais: programa tem, dentre outros objetivos, a intenção de expandir o acesso ao ensino superior.

[5] SISU – Sistema de Seleção Unificada: plataforma digital de acesso às vagas de ensino superior ofertadas pela IES que aderirem ao sistema.

[6] UAB – Universidade Aberta do Brasil: sistema de apoio às IFES para ofertar cursos de nível superior e pós-graduação na modalidade de educação à distância.

[7] PNAES – Plano Nacional de Assistência Estudantil: plano de apoio à permanência de estudantes de baixa renda em cursos de graduação das IFES

Figura 1: Evolução das matrículas de ensino superior no Brasil.



Fonte: Sindata/Semesp. Base Censo INEP

O bem vindo fato do acesso à educação superior estar mais facilitado hoje que noutras épocas não é um fim em si mesmo, o ponto de chegada, muito menos significa que os ingressantes saiam das IES capacitados e preparados para uma carreira profissional. Muito se tem questionado sobre a qualidade do ensino superior ofertado no Brasil (CUNHA, 2004; DURHAM, 2009). Uma limitação à qualidade do ensino superior apontada por Durham (2009) é a deficiência na formação anterior à chegada do estudante nos cursos de graduação. A educação tem como característica ser um processo cumulativo, desenvolvido em etapas, de modo que o aproveitamento de determinado conhecimento depende do cumprimento das etapas anteriores. Logo, deficiências na formação primária têm impacto negativo na aprendizagem dos estudantes ao longo da carreira acadêmica escolar e acabam refletindo na qualidade do ensino superior.

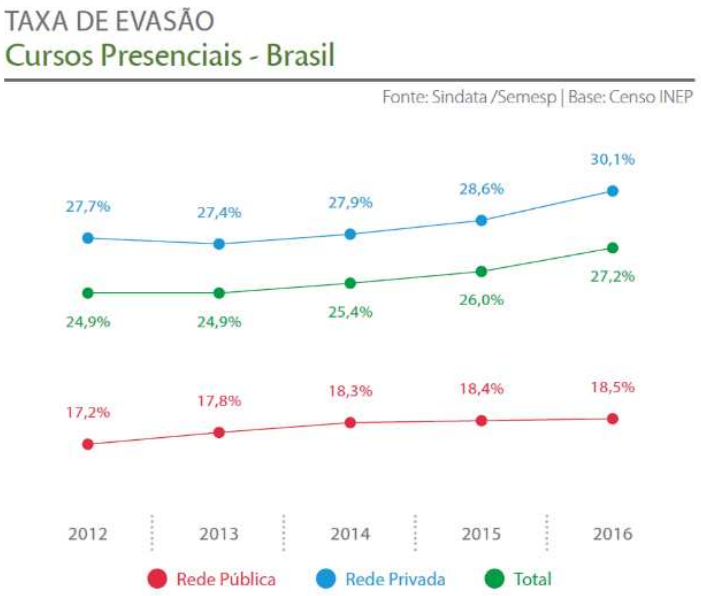
Para traçar um panorama da qualidade dos cursos e instituições de ensino superior no país, o governo federal criou o Sistema Nacional de Avaliação da Educação Superior – SINAES (BRASIL, 2004). Assumindo que a etapa de avaliação é essencial para o processo de melhoria

da qualidade, o SINAES constitui-se em uma série de instrumentos que objetivam melhorar a educação superior e orientar a expansão da oferta. A partir do resultado da aplicação desses instrumentos são criados subsídios para o desenvolvimento de políticas de melhoria da qualidade por parte do poder público, bem como direcionam atuação das próprias instituições de ensino no sentido de melhorar seu mérito e valor enquanto instituição de ensino.

Não é propósito deste trabalho discorrer acerca da qualidade, ou falta dela, do ensino superior, muito menos tratar sobre o perfil do formando, neste momento as atenções estão voltadas para uma ocorrência anterior à conclusão do curso: o fato de muitos ingressantes sequer finalizarem seus cursos. O abandono precoce dos estudos é uma realidade vivenciada em todos os níveis de ensino, conforme já citado na introdução e, a despeito de todos esforços oficiais, continua sendo um obstáculo distante de ser superado, sobretudo no ensino de nível superior.

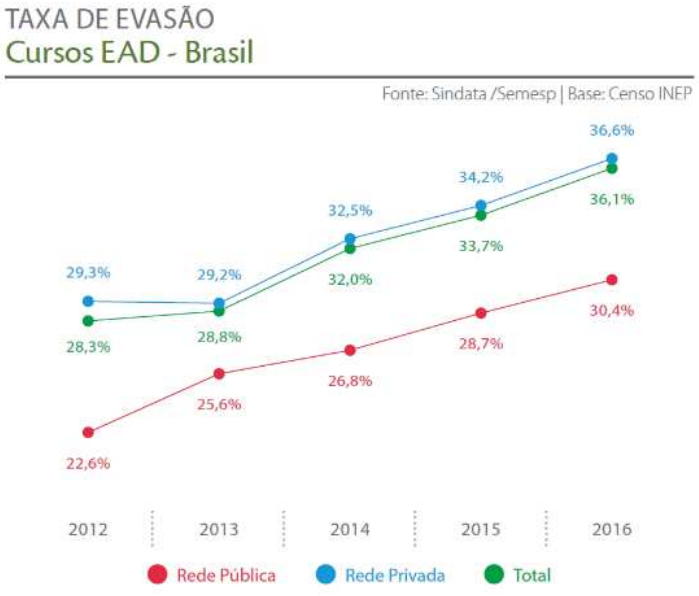
Segundo estudo do Instituto Lobo de 2017, as taxas de evasão nos cursos de nível superior brasileiros se mantiveram em torno de 22% durante os últimos 15 anos, pouco maior para cursos à distância (EAD) e menor para cursos presenciais. Essa situação preocupante merece atenção de toda sociedade, ainda mais quando se sabe que a maior parte da oferta de ensino é financiada direta ou indiretamente com orçamento público, caso do Brasil. Recentemente dados mais atualizados corroboram essa realidade, conforme pode-se observar no gráfico abaixo:

Figura 2: Evolução da taxa de evasão no ensino superior presencial no Brasil[8].



Fonte: Sindata/Semesp. Base Censo INEP

Figura 3: Evolução da taxa de evasão no ensino superior EAD no Brasil.



Fonte: Sindata/Semesp. Base Censo INEP

[8] A taxa de evasão total, indicada pela linha verde, não se trata de uma média, mas sim a taxa de evasão de todos os cursos superiores presenciais sem distinção da rede de ensino. Assim, as taxas totais anuais ficam mais próximas das taxas da rede privada pelo fato da rede privada atender maior quantidade de estudantes.

Observa-se que no ano de 2016 a taxa de evasão nos cursos superiores presenciais foi de 27,2%, significando que mais de um quarto dos ingressantes não completam seus estudos no curso escolhido. Quando se examina a situação dos cursos EAD, o quadro é ainda mais grave, taxas de evasão atingindo 36,1%, ou seja, menos de dois terços dos alunos concluem os estudos. Estas altas taxas de evasão observadas, independentemente do tipo de instituição ou modalidade de ensino por si só suscitam preocupações, além disso percebe-se uma aceleração do aumento das taxas de evasão nos últimos anos. A trajetória ascendente das taxas sugere que os esforços institucionais para garantir a permanência dos estudantes foram insuficientes ou ainda não surtiram o efeito desejado.

Na Rede Federal de Educação Profissional e Tecnológica as estatísticas oficiais são coletadas, validadas e publicadas num ambiente virtual chamado Plataforma Nilo Peçanha - PNP. A PNP, lançada em 2018, representa um avanço importante na consolidação dos indicadores de gestão dos IFs pois até então a apresentação de tais indicadores ficava a cargo de cada instituição, frequentemente sem padrão e obtidas a partir de diferentes metodologias. De acordo com os dados publicados em 2019, ano base 2018, a taxa de evasão para os cursos de graduação dos IFs é de 14,5% (PNP, 2019).

A despeito de todos os esforços para levantar informações fidedignas sobre os IFs, uma série de obstáculos estão postos e dificultam a obtenção de dados precisos sobre a evasão nesse tipo de instituição. A própria ideia de informar uma taxa de evasão única, que compreenda todos os cursos, neste caso os superiores, de uma IES ou grupo de IES encerra uma dificuldade metodológica pois precisa considerar no cálculo cursos ainda não integralizados, ou seja, cursos que ainda não completaram o ciclo de estudos da primeira turma. A agregação desse tipo de situação no cálculo faz com que a taxa observada seja menor do que a calculada caso sejam considerados apenas os cursos já integralizados, ou caso sejam calculadas taxas de evasão para cada turma.

Outra dificuldade encontrada é a heterogeneidade das mais de 600 unidades de ensino contempladas na PNP. A diversidade entre as instituições se apresenta nos diferentes procedimentos de registro acadêmico. Algumas instituições demonstram verdadeiro zelo com a adequada inserção das informações em seus sistemas de gestão elaborando programas de melhoria nos processos de registro, enquanto outras ainda não se atentaram para a importância de dados precisos para as tomadas de decisão.

Além disso, as regras de negócio definidas pelas IES fazem com que cada instituição tenha requisitos próprios para determinar em que momento a matrícula de um estudante é cancelada. No IFC, por exemplo, a organização didática para cursos superiores de graduação define que a matrícula de um estudante só é cancelada após 4 semestres sem que haja rematrícula (IFC, 2012). Até o quarto semestre após abandonar o curso o estudante tem sua matrícula classificada como TRANCADA e conserva o direito de realizar a rematrícula e retomar os estudos. Esse tipo de situação faz com que, oficialmente a taxa de evasão seja menor nos primeiros semestres do curso, o que não reflete a realidade com precisão. É possível que em outras IFES tecnicidades parecidas encubram o cálculo exato da evasão.

2.3 O IFC

Historicamente ligadas ao ensino técnico, os institutos federais de educação vêm ganhando relevância na educação superior, principalmente após as sucessivas expansões patrocinadas pelo governo federal. Criados por lei em 2008 (BRASIL, 2008), os Institutos Federais de Educação, Ciência e Tecnologia são a junção das antigas Escolas Agrotécnicas e Colégios Agrícolas com os Cefets, formando uma rede de ensino técnico e profissionalizante atualmente composta por 40 Institutos presentes em todos os estados da federação.

A lei de criação dos Institutos Federais define como uma das finalidades dos IFs "ofertar educação profissional e tecnológica, em todos os seus níveis e modalidades" (BRASIL, 2008). Esta finalidade delimita um escopo ampliado de atuação dos institutos, permitindo o desenvolvimento tanto de cursos de nível médio quanto de nível superior. A lei traz também algumas limitações na atuação dos IFs, como a obrigatoriedade de garantir pelo menos 50% de suas vagas para cursos de nível médio, prioritariamente na forma de cursos integrados[9], e 20% das vagas ofertadas para ensino superior serem destinadas a cursos de licenciatura.

Esta ambivalência na oferta de diferentes níveis de ensino diferencia os IFs das tradicionais instituições de ensino superior. Para além desta diferença, os institutos também se distinguem pelo seu caráter tecnológico profissionalizante que visa atender demandas sociais e desenvolver os arranjos produtivos locais. Por exemplo, no IFC, onde a pesquisa foi realizada, este caráter fica evidente no *campus* Concórdia que verticalizou o eixo tecnológico de produção

[9] Curso integrado oferece uma formação geral integrada a uma formação profissional. Neste tipo de curso o aluno cursa as disciplinas do currículo normal do ensino médio e disciplinas do curso técnico escolhido.

alimentícia, em sintonia com o potencial da região, reconhecidamente grande produtora de alimentos.

Com a Reitoria localizada na cidade de Blumenau, o IFC está presente em 15 cidades do Estado de Santa Catarina, atuando com a missão de "proporcionar educação profissional, atuando em Ensino, Pesquisa e Extensão, comprometida com a formação cidadã, a inclusão social, a inovação e o desenvolvimento regional" (IFC, 2018). Ao todo são cerca de 147 cursos ofertados, atendendo 15.663 alunos regularmente matriculados em 147 cursos ativos. Do total de cursos ativos 51 são de nível técnico, 44 cursos de graduação, 16 cursos de pós-graduação e o restante é composto por cursos de curta duração ou que atendam alguma demanda específica como ensino a jovens e adultos. Com relação à distribuição dos alunos nos diferentes níveis de ensino, a maior concentração de estudantes se encontra em cursos de nível médio, 7.663 alunos, e nos cursos de nível superior, com 6.625 discentes (IFC, 2019). Observa-se, a partir dos números apresentados, que 42,30% dos alunos do IFC estão vinculados a cursos de nível superior.

Uma característica importante do IFC, fruto do seu processo constitutivo, é a descentralização da gestão. A junção de instituições antigas, algumas com mais de 50 anos de atividade, impôs a dificuldade de unificar processos de gestão particulares, desenvolvidos ao longo de décadas, desconectados uns dos outros e transformá-los numa gestão unificada de caráter institucional.

A dificuldade de acomodação verifica-se também na integração dos sistemas de gestão acadêmica. Até 2018 apenas os cursos de ensino superior estavam completamente informatizados e contavam com registros históricos desde o ano de 2013. Por isso que o presente estudo se limitará a analisar os cursos de nível superior. Além disso, diversos estudos indicam que a evasão no ensino superior é maior que noutros níveis de ensino (SILVA FILHO, 2007; SEMESP, 2018).

Tabela 1 – Quantidade de cursos de ensino superior no IFC ofertados para ingresso em 2018.

Curso	**Quantidade**
Agronomia	4
Análise e Desenvolvimento de Sistemas	2
Ciências Agrícolas	1
Ciência da Computação	2
Design de Moda	1
Engenharia da Computação	1
Engenharia de Alimentos	1
Engenharia de Controle e Automação	2
Engenharia Elétrica	3
Engenharia Mecânica	1
Engenharia Mecatrônica	1
Gestão de Turismo	1
Física	2
Logística	4
Química	2
Matemática	4
Medicina Veterinária	2
Negócios Imobiliários	1
Pedagogia	5
Redes de Computadores	3
Sistemas de Informação	2
Sistemas para Internet	1
Total	**43**

Fonte: autoria própria

Ao todo foram ofertados pelo IFC 43 cursos de nível superior como pode ser observado na tabela 1. Destacam-se os cursos de Pedagogia com oferta em 5 *campi* e Matemática e Agronomia, ambos ofertados em 4 *campi* diferentes. Além disso pode-se destacar os cursos ligados à computação e informática que totalizam 11 cursos quando somados os cursos de Análise e Desenvolvimento de Sistemas, Ciência da Computação, Engenharia da Computação, Redes de Computadores, Sistemas de Informação e Sistemas para Internet.

Então, a pesquisa foi realizada num contexto de instituição pública de ensino, multicampi, com presença em 15 cidades do estado de Santa Catarina e com características particulares de oferta de ensino em vários níveis. Além disso, o caráter tecnológico se evidencia pela busca do alinhamento entre as atividades de ensino, pesquisa e extensão com as necessidades dos arranjos produtivos em que os *campi* estão inseridos.

2.3.1 PLANO ESTRATÉGICO DE PERMANÊNCIA E ÊXITO DOS ESTUDANTES DO IFC

Em 2019 o IFC aprovou por meio de uma resolução de seu conselho superior o Plano Estratégico Institucional para a Permanência e Êxito (IFC, 2019). O plano deriva de um compromisso assumido pelo Ministério da Educação na figura da Secretaria de Educação Profissional e Tecnológica (SETEC/MEC) junto ao Tribunal de Contas da União. O objetivo é orientar as medidas de promoção da permanência e êxito dos estudantes da rede federal de ensino profissional e tecnológico.

A partir dessa exigência, uma comissão foi instituída no âmbito do IFC para elaborar o plano válido por dois anos em alinhamento com Plano de Desenvolvimento Institucional. Para subsidiar o trabalho uma proposta de metodologia foi apresentada pela SETEC/MEC (2015), sugerindo a realização de diagnósticos quantitativos e qualitativos sobre a situação da evasão.

O diagnóstico quantitativo realizado no IFC se baseou em dados da Plataforma Nilo Peçanha e do próprio sistema de gestão acadêmica do IFC. No caso dos cursos superiores, além das taxas de evasão registradas durante os anos de 2015 a 2017, foi apresentada a situação em 2018 dos ingressantes em cursos superiores no ano de 2017, ou seja, um ano depois. Nesse levantamento foram verificadas taxas de evasão superiores a 50% em alguns cursos, isto é, após um ano mais da metade dos estudantes desses cursos já não frequentavam as aulas. Esta é a situação dos cursos de Ciências Agrícolas (*campus* Araquari) e de Física (*campus* Rio do Sul).

Com relação ao diagnóstico qualitativo, foram elencados alguns motivos recorrentes pelos quais os estudantes desistem de frequentar os cursos. Dentre eles pode-se citar a desinformação acerca do curso escolhido, dificuldade em acompanhar o ritmo das aulas, além de problemas pessoais tais como financeiros, de saúde e falta de tempo. Constatações que não diferem muito do que foi apresentado na seção 2.1 deste trabalho.

O plano ainda prevê estratégias de intervenção específicas para cada curso ofertado. Pode-se perceber dentre as ações propostas duas linhas principais de atuação: a primeira delas é no sentido de prevenir a evasão por meio de um atendimento ao educando mais ativo oferecendo orientação para organização de uma rotina de estudos, fortalecimento das monitorias e divulgação dos programas oferecidos pela instituição. A outra linha de atuação é, uma vez que a evasão foi identificada, adotar o procedimento de contatar o estudante para entender os motivos e assim aprimorar as ações de prevenção.

Ainda estão previstas avaliações e monitoramento semestrais do plano com objetivo de verificar se as ações propostas estão sendo colocadas em prática, identificar quais são as dificuldades enfrentadas e, por fim, averiguar se as ações estão refletindo em menores taxas de evasão. As respostas desse monitoramento permitirão que os envolvidos no processo possam buscar soluções para aprimorar a estratégia.

2.4 MINERAÇÃO DE DADOS E EDM – *EDUCATIONAL DATA MINING*

A capacidade de criação de dados nunca foi tão grande desde que as ferramentas digitais passaram a fazer parte do dia a dia da sociedade. A quantidade de dados produzidos diariamente no mundo é incrivelmente grande, espera-se que em 2020 todo o universo digital atinja 44 *zettabytes*[10]. Isto significa uma quantidade de dados quase 10 vezes superior à existente em 2013 (WEF, 2019). Além do que, à medida que novas tecnologias são inventadas e cada vez mais pessoas ingressam no mundo digital, o ritmo de geração de dados tende a acelerar.

A partir disso surge um problema fundamental: como extrair dessa quantidade volumosa de dados informações relevantes que auxiliem o processo de tomada de decisão?

Para lidar com esse problema e obter um entendimento geral dos dados a ponto de tirar conclusões pertinentes, diversos campos de estudo são integrados como: computação, estatística e matemática, *machine learning*, bases de dados, inteligência artificial e apresentação de dados. O que liga estes campos de estudo é o objetivo final de identificar padrões que representem o conhecimento armazenado em grandes bases de dados (HAN; KAMBER; PEI, 2011).

Historicamente a noção de encontrar informações relevantes em bases de dados tem recebido uma variedade de nomes: descoberta de informações, colheita de informações, arqueologia de dados, entre outros. Atualmente, Mineração de Dados é o termo mais popularmente utilizado para descrever a atividade, embora alguns ainda entendam que a mineração de dados é apenas uma etapa de um processo maior chamado *Knowledge Discovery in Databases* (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Segundo Han, Kamber e Pei (2011) “Mineração de Dados é o processo de descoberta de padrões interessantes e conhecimento a partir de grandes quantidades de dados”. A partir

[10] 1 *zettabyte* = 1.000^7 = 1.000.000.000.000.000.000.000 bytes.

desta definição subentende-se um conjunto de etapas de identificação da estrutura dos dados ou descrição de alguma característica específica do conjunto de dados. Tipicamente o processo de mineração envolve atividades de preparação dos dados como seleção, limpeza e transformação para, em seguida, aplicar as técnicas de reconhecimento de padrões e, por fim, apresentar os resultados.

Lidar com grandes bases de dados é uma necessidade presente em praticamente todas as áreas, nas mais diversas atividades. Por isso, a mineração de dados já foi aplicada com sucesso em muitos ambientes diferentes – finanças, saúde, seguros, segurança, dentre outros. Atualmente, ao simplesmente navegar pela internet um indivíduo já produz uma quantidade de dados suficiente para que milhares de algoritmos caracterizem seu padrão de comportamento e transformem essa informação em ativo de importante valor comercial.

Uma aplicação da mineração de dados que se desenvolveu muito na última década e ainda vem ganhando relevância é no contexto educacional. O progresso da aplicação de métodos de identificação de padrões relevantes em dados educacionais foi tão relevante que se transformou em uma área de pesquisa específica de estudos chamada *Educational Data Mining* – EDM (BAKER, 2011).

Um reflexo do crescimento da mineração de dados no contexto educacional foi a fundação da *International Educational Data Mining Society* - IEDMS no ano de 2011 com objetivo de promover a pesquisa científica na área de EDM. Além disso a IEDMS busca estimular a participação da comunidade envolvida com a EDM por meio da realização de congressos anuais e da edição de um jornal focado no tema[11].

Segundo Baker (2011, p.4) EDM "é definida como a área de pesquisa que tem como principal foco o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais". Ainda de acordo com Baker (2011) a condição essencial para a utilização da EDM é a disponibilização de dados estruturados e em quantidade suficiente. Nesse sentido, com a disseminação da utilização de softwares educacionais desenvolvem-se cada vez mais ambientes ideais para a aplicação da EDM.

As principais aplicações da EDM estão relacionadas ao desenvolvimento de modelos de conhecimento do estudante, tanto na identificação da proficiência dos alunos em

[11] Estas informações encontram-se no próprio site da *Journal of Educational Data Mining*: http://educationaldatamining.org/.

determinada área de estudo, quanto no reconhecimento de características comportamentais que auxiliem a gestão escolar no oferecimento de serviços cada vez mais adequados ao perfil do estudante.

Embora a EDM possa ser utilizada em inúmeras situações, pode-se de maneira geral destacar duas grandes áreas de aplicação. Uma diz respeito à utilização dos dados estudantis para aprimorar o ambiente de aprendizado de modo a permitir aos estudantes diferentes formas de aquisição de conhecimento. Isso ganha ainda maior relevância quando se insere no processo de ensino a colaboração de ambientes interativos de aprendizagem.

Outra importante aplicação está relacionada ao uso dos dados discentes para otimizar os processos de gestão administrativa das instituições de ensino. Isso permite que se desenvolvam ações mais assertivas e focadas no que se deseja aperfeiçoar. É justamente neste contexto que este estudo se insere, uma vez que busca fornecer à gestão institucional do IFC mais uma ferramenta de combate à evasão.

Uma fundamental vantagem deste tipo de abordagem é a redução substancial dos custos tradicionais da pesquisa. Além disso, reduz a necessidade de mão de obra para coletar, formatar e digitalizar os dados obtidos, bem como dispensa a necessidade de conduzir experimentos in loco que podem durar dias ou até semanas. A partir de dados confiáveis, a obtenção de resultados em tempo reduzido, com precisão, permite que as ações de melhoria sejam antecipadas e produzam efeito mais rapidamente.

Os principais métodos de EDM são apresentados por Baker (2011, p.5) da seguinte forma[12]:

> Predição (*Prediction*)
> - Classificação (*Classification*)
> - Regressão (*Regression*)
> - Estimação de Densidade (*Density Estimation*)

> Agrupamento (*Clustering*)
> Mineração de relações (*Relationship Mining*)
> - Mineração de Regras de associação (*Association Rule Mining*)
> - Mineração de Correlações (*Correlation Mining*)
> - Mineração de Padrões Sequenciais (*Sequential Pattern Mining*)
> - Mineração de Causas (*Causal Mining*)

> Destilação de dados para facilitar decisões humanas (*Distillation of Data for Human Judgment*)
> Descobertas com modelos (*Discovery with Models*)
>
> (BAKER, 2011, p.5)

[12] Apresentação muito parecida com a realizada por Fayyal et al (1996) se referindo a etapa de Mineração de Dados no processo KDD.

Como se pode observar, existem diversos métodos de EDM que podem ser utilizados a depender do objetivo a ser atingido. A sintonia entre o objetivo proposto e o método de EDM mais adequado é uma etapa fundamental do processo de Mineração de dados.

Uma recorrente aplicação de EDM, muito em função da relevância do problema, está relacionada com a identificação da propensão à evasão dos estudantes. Segundo Baker (2011, p.5), "na área da predição, a meta é desenvolver modelos que deduzam aspectos específicos dos dados", ou seja, deduzir o comportamento de uma determinada variável em função de outras variáveis. Quando se trata de evasão, a variável a ser prevista é binária ou categórica, portanto, o método de EDM mais adequado a ser utilizado é a classificação.

Neste sentido, muitos estudos têm sido realizados nos mais diferentes contextos na tentativa de prever a evasão e municiar os gestores com relevantes informações para auxiliar o desenvolvimento de programas de permanência estudantil (MANHÃES, 2015; AMARAL, 2016; KANTORSKY ET AL, 2015; FUNCHAL, 2016; AIRES et al, 2017; BITENCOURT, FERRERO, 2019).

Manhães (2015) propôs uma arquitetura baseada em EDM com objetivo de prever o desempenho acadêmico dos graduandos dos cursos de engenharia da Universidade Federal do Rio de Janeiro (UFRJ). O estudo parte do pressuposto que o desempenho acadêmico é a chave que permite identificar a propensão a evadir dos estudantes. Os resultados encontrados mostram que é possível prever com aproximadamente 80% de acerto o desempenho acadêmico a cada semestre.

Na mesma linha, Funchal (2016) e Kantorski et. al. (2015) aplicam metodologia parecida em outros contextos. A partir de dados acadêmicos e cadastrais dos estudantes buscam prever a evasão nos cursos analisados. Ambos estudos alcançam resultados interessantes com acurácias elevadas: 95,5% (FUNCHAL, 2016) e 98% (KANTORSKY ET. AL., 2015).

Amaral (2016) desenvolve um estudo com uma particularidade significativa, propõe aplicar a EDM apenas aos dados cadastrais dos estudantes. Segundo ele, fazer uso de dados acadêmicos requer um período de espera até que essas informações estejam disponíveis, reduzindo o tempo hábil de adoção de medidas por parte da instituição de ensino. O estudo foi realizado com discentes da Universidade Federal de Pernambuco – UFPE e obteve acurácia de acerto em torno de 70% na previsão de evasão.

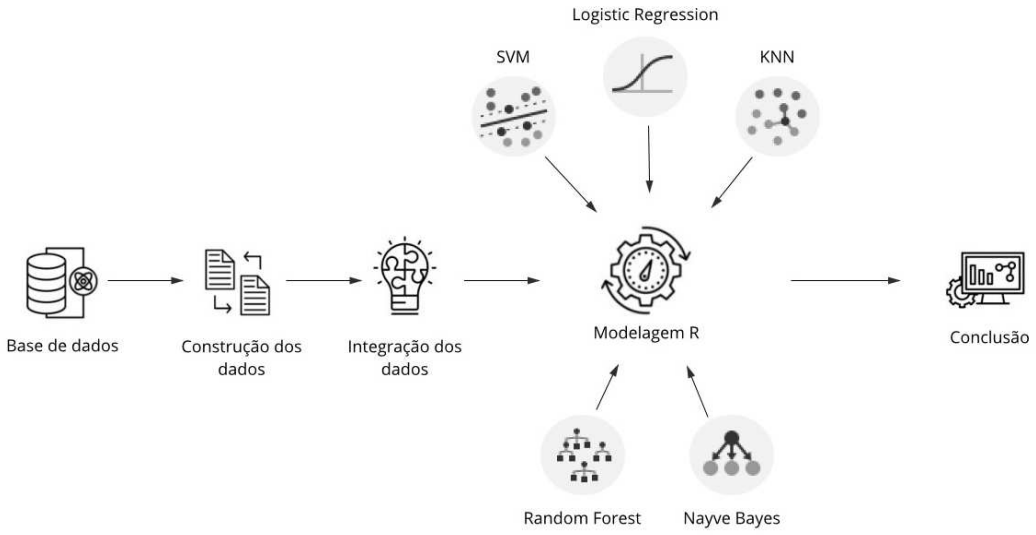
Bitencourt e Ferrero (2019), em seu estudo de classificação baseado em árvores de decisão, acrescentam aos dados cadastrais dos estudantes o número de faltas durante os 7 primeiros dias de aula. Com essas informações foi possível prever com 86% de precisão as evasões. Esta forma de utilização de dados acadêmicos logo nos primeiros dias de aula constitui-se numa maneira de reduzir o tempo necessário para obtenção de resultados que auxiliem a gestão institucional na elaboração de estratégias para a permanência estudantil.

Percebe-se, a partir destes estudos, que há um *trade-off* na utilização ou não de dados acadêmicos na EDM, se por um lado as informações acadêmicas são relevantes e aumentam a precisão dos modelos, a demora em obter os resultados dificulta a tomada de ação por parte dos envolvidos no processo. Outra constatação importante é que não houve muita diferença na acurácia de classificação entre os algoritmos utilizados nos estudos, nem verificada alguma técnica específica de mineração com resultados muito superiores às outras.

Dentre os muitos modelos de classificação disponíveis, alguns se destacam por aparecer com frequência em estudos acadêmicos: Nayve Bayes, Support Vector Machines, Regressão Logística e Decision Tree (MANHÃES, 2015; MACHADO, 2015; AMARAL, 2016; AIRES, 2017; BITENCOURT; FERRERO, 2019). Eventualmente, a depender da ferramenta utilizada os nomes podem alterar um pouco, mas os fundamentos são semelhantes, por exemplo a combinação de várias Árvores de Decisão resulta em um modelo chamado *Random Forest.*

Com base nesses estudos e na ferramenta computacional utilizada (R), foram selecionados 5 modelos de classificação para identificar a propensão à evasão dos ingressantes nos cursos de ensino superior do IFC. Os modelos bem como o fluxo do processo de EDM podem ser visualizados abaixo:

Figura 4: Fluxo do processo de mineração de dados realizado neste estudo.



Fonte: autoria própria

Podem-se observar no fluxo acima as etapas do processo de mineração de dados a serem desenvolvidas neste trabalho. O ponto de partida é a base de dados da instituição, a partir disso, conforme a necessidade e de acordo com os objetivos propostos, as variáveis são selecionadas, modificadas ou, até mesmo, são criadas novas variáveis até atingir um conjunto de dados pronto para a próxima etapa. Na fase de modelagem os algoritmos de classificação são aplicados a esses dados e, fundamentados nos resultados obtidos, são tiradas as conclusões pertinentes.

Finalizada esta parte da dissertação, necessária para fundamentação teórica do trabalho, apresenta-se na sequência a metodologia utilizada, com especial atenção para mineração de dados e, finalmente expor os resultados e conclusões.

3 METODOLOGIA

Neste capítulo será apresentada a metodologia utilizada, *Cross-Industry Standard Process for Data Mining* (CRISP-DM), suas etapas, bem como a aplicação da metodologia no IFC. Este capítulo está dividido basicamente em duas partes: a primeira diz respeito à apresentação da metodologia utilizada neste trabalho e o ciclo do processo de mineração de dados incluindo as etapas de entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem e desenvolvimento. A segunda e maior parte do capítulo refere-se à aplicação dessa metodologia no IFC com base nos dados disponibilizados.

3.1 *CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING* (CRISP-DM)

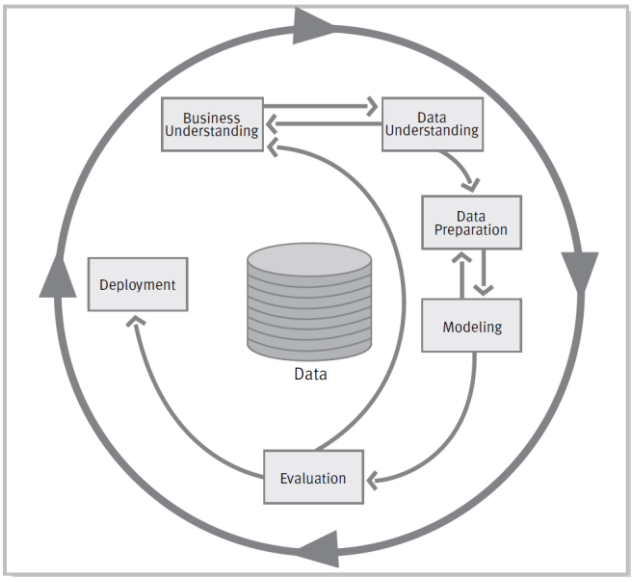
Conforme visto na fundamentação teórica, o processo de mineração de dados pode ser empregado aos mais diferentes campos de atividade como medicina, finanças, comportamento, entre outros. No entanto, uma dificuldade encontrada inerente a esta utilização intersetorial era desenvolver uma metodologia que pudesse ser empregada em atividades tão distintas. Foi o que fizeram Chapman et al (2000) quando desenvolveram a metodologia *Cross-Industry Standard Process for Data Mining* (CRISP-DM), que pode ser entendida como uma aplicação do processo de mineração de dados num contexto amplo. (AZEVEDO; SANTOS, 2008).

Essa metodologia permite que as técnicas de mineração de dados sejam operadas por diferentes tipos de usuários nas mais distintas áreas de conhecimento. O modelo CRISP-DM consiste na proposta de um ciclo padrão para os projetos de mineração de dados e pode ser perfeitamente aplicado ao contexto educacional (EDM). É por este motivo que a metodologia CRISP-DM foi escolhida para a realização deste estudo.

De acordo com pesquisa realizada pelo KDnuggets (2014), relevante website sobre análise de dados, verificou-se que a metodologia CRISP-DM era a mais utilizada pelos respondentes. No levantamento realizado em 2014 a CRISP-DM obteve aproximadamente 42% dos registros como a metodologia mais utilizada para mineração de dados, em seguida aparecem as metodologias próprias com 27% (KDNUGGETS, 2014).

As etapas deste processo podem ser observadas na figura abaixo e, em seguida, é apresentada a descrição de cada etapa:

Figura 5: – Fases do Modelo de Referência CRISP-DM.



Fonte: Chapman et al, 2000 CRISP-DM

O processo inicia-se com o entendimento do contexto de negócio em que a metodologia será empregada. A partir deste entendimento, é definido o problema a ser tratado e são desenhadas as linhas gerais para o atingimento dos objetivos.

A segunda etapa consiste na coleta inicial e entendimento dos dados. Os procedimentos iniciais de coleta e organização dos dados permitem que o usuário adquira familiaridade com a base, identificando problemas de qualidade dos dados e descobrindo os primeiros *insights* a respeito do problema.

A fase de preparação dos dados abrange todas as atividades necessárias para construção da base de dados final, que será objeto da aplicação dos modelos de mineração. Normalmente, esta é uma etapa realizada várias vezes, sem uma ordem pré-determinada. Podem-se destacar nesta etapa as atividades de limpeza, seleção e transformação dos dados.

Com a base de dados construída são selecionados os modelos de mineração mais apropriados a serem aplicados aos dados, tendo em vista os objetivos definidos. É bastante comum existirem várias técnicas para o mesmo problema de mineração, assim como não é raro reconhecer a necessidade de retornar à etapa de preparação dos dados para a correta aplicação dos modelos.

A etapa seguinte consiste na avaliação dos resultados obtidos com a aplicação do(s) modelo(s) proposto(s). Além disso, este é o estágio em que são conferidos todos os passos do

processo realizados até o momento no sentido de buscar identificar algum aspecto relevante para o negócio que não tenha sido considerado.

O estágio final do processo CRISP-DM trata-se da efetiva aplicabilidade do conhecimento descoberto a partir da aplicação do modelo desenvolvido. A aplicabilidade dos resultados está intimamente ligada ao objetivo proposto. Podem estar relacionados ao desenvolvimento de relatórios simples até a implementação de estruturas mais complexas. Mesmo quando se tratam de modelos com objetivo limitado ao avanço do conhecimento de determinado campo de estudo, as informações descobertas precisam ser organizadas e apresentadas de uma forma que o consumidor final consiga interpretar e absorver o conhecimento produzido.

3.2 APLICAÇÃO DA METODOLOGIA AO IFC

Nesta seção serão aplicadas as etapas descritas na metodologia CRISP-DM ao contexto do IFC.

3.2.1 ENTENDIMENTO DO NEGÓCIO

Do ponto de vista da instituição educacional trabalhada, o IFC, a evasão se configura um problema de difícil solução e com consequências tanto para o Instituto quanto para a sociedade como um todo. Esta situação suscita a adoção de medidas que favoreçam a permanência dos estudantes em seus cursos por parte dos agentes ligados à educação – governo, gestores educacionais, professores, pesquisadores -, sobretudo pela própria instituição de ensino que possui contato direto com os estudantes envolvidos.

Para que a atuação institucional seja eficiente, no sentido de reduzir a quantidade de alunos que desistem dos cursos, e tenha maiores possibilidades de sucesso, é muito importante identificar e concentrar os esforços naquele grupo de estudantes com maior propensão a evadir. Além disso, é imprescindível que a atuação institucional seja precoce, de preferência preventiva. Deste modo, a tarefa de identificar aqueles estudantes mais propensos a evadir pode ser realizada utilizando métodos de mineração de dados, mais especificamente EDM. Essa abordagem permite obter resultados em tempo reduzido sem que haja elevado dispêndio financeiro.

Busca-se, a partir de dados cadastrais e de desempenho coletados pelo IFC desde o processo seletivo para entrada nos cursos, bem como dados acadêmicos do período inicial de estudos, criar modelos que prevejam quais dos novos entrantes têm maiores possibilidade de evadir. Dessa forma os professores, gestores, coordenadores podem se preparar com antecedência, ficando alerta aos sinais emitidos pelos estudantes e atuar de acordo com cada caso em tempo de evitar/reduzir a desistência dos cursos.

Embora a EDM seja uma abordagem eficiente em termos de gastos financeiros e tempo necessário para obtenção de resultados, seu sucesso está condicionado à qualidade dos dados coletados. A evasão já é, por natureza, um problema multifacetado, sua ocorrência deve-se a muitas diferentes motivações e isso torna difícil para uma instituição de ensino captar informações suficientes para fazer um diagnóstico precoce do problema.

A despeito de todas as dificuldades na obtenção de dados relevantes para entender o fenômeno e realizar uma aplicação eficaz da EDM, tem-se ainda obstáculos oriundos de erros processuais, falhas humanas, equívocos nas definições conceituais, dentre outros, para o correto registro das informações estudantis e isso tem impacto na qualidade das bases de dados mantidas pela instituição de ensino. Estra dificuldade é bastante relevante numa instituição como o IFC, historicamente descentralizada e como significativas diferenças de infraestrutura entre os *campi*.

3.2.2 ENTENDIMENTO DOS DADOS

Os dados primários utilizados neste estudo são provenientes de duas fontes: do Sistema Integrado de Gestão Acadêmica - SIGA utilizado pelo IFC e do processo seletivo realizado pelo SISU. Os dados do SIGA, são registrados no momento da matrícula do estudante e contém informações cadastrais e acadêmicas, ou seja, o histórico acadêmico do aluno. Por outro lado, as informações disponibilizadas ao IFC pelo SISU são compostas basicamente por registros cadastrais informados pelos próprios estudantes no momento da inscrição para a prova do ENEM e por dados referentes ao desempenho dos candidatos na prova.

A extração dos dados no SIGA foi realizada em novembro de 2018 e compreendeu registros desde 2010 até o segundo semestre de 2018. Embora existam *campi* cujo surgimento precede muito a criação do IFC, a consolidação dos registros só tomou forma com a implantação do SIGA. O processo gradual de consolidação de todos os registros em um sistema único fez

com que o resgate de informações antigas fosse parcial, por este motivo não foi possível o acesso a dados anteriores ao ano de 2010. Ao todo foram extraídos registros de 8.985 matrículas ao longo desse período.

Quanto aos dados do SISU, relativos à participação dos candidatos na prova do ENEM, são compostos por arquivos anuais relativos à aplicação do ENEM naquele ano. Em 2014 a IFC aderiu ao SISU e passou a ofertar, inicialmente, 50% das vagas de ensino superior por meio desse sistema. Até então era realizado vestibular: elaborado, coordenado e aplicado pela própria instituição. Em 2016, o Instituto passou a destinar 100% das vagas ao SISU e, desde então, apenas os casos de vagas remanescentes são preenchidos por procedimentos próprios.

Para que os candidatos tenham acesso às vagas disponibilizadas pelo SISU, é preciso ter realizado a prova do ENEM. Esta prova é composta por uma redação e mais 4 áreas de conhecimento: linguagens, ciências humanas, ciências da natureza e matemática. A partir do desempenho dos candidatos, são calculadas ao todo 6 notas, 5 referentes às áreas de conhecimento e 1 nota global que representa a nota final do candidato.

Então, para compor este conjunto inicial de dados foram selecionadas nas duas bases de dados todas as informações disponíveis ao nível do aluno, ou que pudessem ser vinculadas a uma matrícula específica. A partir deste conjunto inicial de dados, foram pré-selecionadas as variáveis relacionadas à escolha do estudante, seu desempenho no ENEM, sua situação sócio econômica ou que sejam chave para o agrupamento das duas fontes de dados. Aquelas variáveis não relacionadas a esses aspectos ou que, embora pessoais, fossem irrelevantes para o entendimento do problema como nome, número de telefone, entre outros, foram desconsideradas. As variáveis pré-selecionadas podem ser observadas nos quadros abaixo e serão detalhadas nas seções posteriores:

Quadro 1: Variáveis extraídas do SIGA.

Código	Variável
V01	Status
V02	CPF
V03	Data de Nascimento
V04	Endereço
V05	Município
V06	Unidade Federativa
V07	Campus
V08	Curso
V09	Índice de Rendimento Acadêmico no IFC
V10	Forma de Ingresso
V11	Grau
V12	Período de início do curso
V13	Tipo de Escola

Fonte: autoria própria

Quadro 2: Variáveis extraídas do SISU.

Código	Variável
V02	CPF
V03	Data de Nascimento
V04	Endereço
V05	Município
V06	Unidade Federativa
V13	Tipo de Escola
V14	Turno
V15	Nota Linguagens
V16	Nota Ciências Humanas
V17	Nota Ciências Naturais
V18	Nota Matemática
V19	Nota Redação
V20	Nota Inscrito

Fonte: autoria própria

Percebe-se nos quadros acima que algumas informações extraídas são semelhantes, principalmente quando se tratam de informações cadastrais e da escolha do estudante. No entanto, existem outras informações que podem ser complementares supondo que possa haver alguma relação entre a situação da matrícula dos alunos e o resultado obtido no exame de ingresso.

A situação do vínculo do aluno com a instituição, representada pelo atributo V01 é a variável fundamental para desenvolver os modelos de classificação. É o dado que nos diz se o aluno concluiu com sucesso seu curso ou se, em algum momento, interrompeu sem retorno os estudos. Nos registros do IFC essa variável está classificada da seguinte forma:

ATIVO: alunos com matrícula ativa são aqueles frequentando regularmente as aulas. Não se trata de um status final pois o estudante ainda pode concluir ou desistir do curso.

CONCLUIDO: são estudantes que concluíram com sucesso seu curso. Por se tratar de uma situação final estes registros serão utilizados na análise.

FORMADO: semelhante aos concluídos, o status FORMADO se refere àqueles estudantes que concluíram com sucesso seu curso. A existência de diferentes termos para se referir à mesma situação se deve à coexistência de diferentes sistemas de registro acadêmico. Com a adoção do SIGA os dados foram migrados para um sistema único. Para fins de análise, os alunos nesta situação serão considerados concluintes.

FORMANDO: são os estudantes que já integralizaram o curso, ou seja, já concluíram todos os componentes curriculares, mas ainda não encerraram seu vínculo com a instituição devido à pendência de entrega de alguma documentação exigida. Como se tratam de alunos que foram aprovados nas disciplinas e, portanto, obtiveram sucesso acadêmico, este status de matrícula será considerado concluído.

TRANCADO: este status acadêmico compreende duas situações distintas. A primeira diz respeito aos estudantes que por livre e espontânea vontade solicitaram trancamento com a intenção de retomar os estudos num futuro próximo. A outra situação se refere aos estudantes que evadiram a menos de quatro semestres. O IFC, em sua organização didática para os cursos superiores, só cancela a matrícula do aluno que não frequentar as aulas durante quatro semestres consecutivos. Isso faz com que um aluno trancado seja um potencial evadido. Tendo em vista que a maior parte dos alunos qualificados como TRANCADO se refere a alunos sem frequência, nem registro de rematrícula, ou seja, sem perspectiva de retorno considerou-se, para este estudo, o status TRANCADO como evadido.

CANCELADO: refere-se às matrículas dos estudantes que não concluíram os cursos, seja porque solicitaram o cancelamento de sua matrícula, porque abandonaram o curso ou, até mesmo, porque não conseguiram concluir o curso dentro do prazo previsto pelas normas didáticas da instituição. Este status representa o insucesso acadêmico, ou seja, a situação de matrícula fundamental para entender o fenômeno da evasão.

EXCLUIDO: este registro se refere às matrículas que, diferente das canceladas, foram excluídas por motivos técnicos ou judiciais. Eventualmente, por falha humana a matrícula foi realizada equivocadamente e, portanto, precisa ser excluída. Outra situação de matrículas canceladas acontece quando, por decisão judicial, o estudante deve se desligar do curso. Normalmente são eventos raros e por não representarem o insucesso acadêmico este tipo de registro não será utilizado na análise.

Com relação às demais variáveis apresentadas em seguida, pode-se compreender a relevância e as particularidades de cada uma, bem como sua relação com a situação da matrícula, conforme descrição abaixo:

V01 – Status: Refere-se à situação da matrícula do aluno, fundamental para identificar se o aluno evadiu ou permanece frequentando o curso escolhido. Para este trabalho este atributo é a variável resposta, ou seja, a variável a ser prevista.

V02 – CPF do discente, utilizado como chave para o agrupamento de diferentes bases de dados. Como este atributo não apresenta relação com o a permanência ou não do estudante na instituição não será considerado pelos algoritmos de classificação.

V03 – Data de nascimento: A partir desse dado é possível calcular a idade do estudante quando da entrada no curso. A relevância dessa variável se deve à hipótese de que a idade do aluno influencia a propensão a evadir do mesmo.

V04 – Endereço do discente: A relevância deste atributo advém da hipótese de que a distância entre a residência do aluno e a escola influenciam na decisão de evadir. Para tanto é preciso garantir que o cálculo da distância esteja corretamente relacionado ao campus em que o aluno está vinculado.

V05 – Município: Esta variável está relacionada com a mesma hipótese da distância (V04), sendo utilizada como complementação ao endereço do estudante.

V06 – Unidade Federativa: Idem V04.

V07 – Campus: A importância desta variável deriva da influência que o ambiente da escola exerce sobre a decisão de permanência do estudante. Como tratam-se de diversos *campi*, há possibilidade de que o fenômeno da evasão se comporte de forma diferenciada a depender do *campus* onde o curso está localizado.

V08 – Curso: A variável curso compreende a ideia de que diferentes cursos apresentam taxas distintas de evasão.

V09 – Índice de rendimento acadêmico (IRA): Trata-se da média das notas obtidas pelo aluno ao longo do curso. Depreende-se desta variável a hipótese de que quanto melhor o rendimento acadêmico do estudante menos incentivo há para abandonar o curso.

V10 – Forma de ingresso: trata-se da etapa do processo seletivo na qual o estudante ingressou na instituição. O processo seletivo do IFC compreende três etapas: primeira chamada do SISU, cadastro de reserva e vagas não ocupadas. Esta variável encerra a hipótese de que estudantes aprovados nas etapas finais do processo seletivo demonstrem menor interesse em concluir o curso.

V11 – Grau: refere-se ao tipo de curso superior, se bacharelado, licenciatura ou tecnólogo. A relevância do atributo é oriunda da hipótese de que as características gerais do curso escolhido têm influencia a decisão de evadir.

V12 – Período de início do curso. Este atributo está relacionado com a hipótese da variável V03, a partir dessas variáveis é possível calcular a idade do discente no momento de ingresso no curso.

V13 – Tipo de escola: trata-se da escola onde o estudante completou o ensino médio, se pública ou privada. A relevância desta variável advém da possibilidade de haver relação entre o tipo de escola em que o estudante cursou e a evasão.

V14 – Turno: Trata-se do turno escolhido pelo discente para frequentar o curso. A importância deste atributo decorre da hipótese de que o horário das aulas possa influenciar a propensão a evasão dos estudantes, principalmente quando considerado em conjunto com os demais atributos.

V15 – Nota Linguagens: Trata-se da nota obtida pelo estudante na prova de linguagens do ENEM. A pertinência desta variável decorre da possibilidade dos estudantes com melhor desempenho na prova do ENEM, terem maior facilidade nos estudos de ensino superior e, portanto, tenham menor propensão a evadir.

V16 – Nota Ciências Humanas: Trata-se da nota obtida pelo estudante na prova de ciências humanas do ENEM.

V17 – Nota Ciências Naturais: Trata-se da nota obtida pelo estudante na prova de ciências naturais do ENEM.

V18 – Nota Matemática: Trata-se da nota obtida pelo estudante na prova de matemática do ENEM.

V19 – Nota Redação: Trata-se da nota obtida pelo estudante na prova redação do ENEM. Hipótese semelhante a V15.

V20 – Nota Inscrito: Trata-se da nota final obtida pelo estudante no ENEM obtida a partir da média simples das 5 provas realizadas. Hipótese semelhante a V15.

3.2.3 PREPARAÇÃO DOS DADOS

A etapa seguinte ao entendimento inicial dos dados, suporte para o desenvolvimento do estudo, é a preparação dos dados com o objetivo de produzir uma base de dados única, limpa e no formato adequado para a etapa posterior de modelagem. A fase de preparação pode ser subdividida em cinco partes: seleção, limpeza, construção, integração e formatação (CHAPMAN ET AL, 2000).

Parte dos critérios para seleção de variáveis já foi definido na fase de extração inicial dos dados, quais são: informações do ensino superior, dados ao nível do aluno representantes da sua escolha, nível socioeconômico ou desempenho acadêmico e informações chave para o agrupamento de diferentes fontes de dados. A partir da posse desses dados outras delimitações se fazem necessárias tendo em vista o objetivo de criar um modelo de previsão da propensão a evadir dos alunos a ser utilizado pela instituição nos futuros ingressos de ensino superior.

Como o processo de seleção dos candidatos do IFC é realizado pelo SISU desde 2014 e a perspectiva atual é a continuidade dessa sistemática de seleção, definiu-se a utilização das informações a partir de 2014, sendo os dados anteriores descartados. Além de agregar mais informações, a decisão de adoção do SISU representou um marco para o processo seletivo do IFC. Um efeito colateral benéfico dessa decisão foi intensificar o movimento de institucionalização das ações no IFC, autarquia bastante descentralizada por características históricas, uma vez que o processo seletivo de todos os cursos superiores precisaria estar

enquadrado no modelo SISU. Ou seja, estimulou o IFC a padronizar os procedimentos de seleção e por consequência trouxe mais confiabilidade para os dados.

A fase de limpeza se constitui numa atividade uniformização e melhoria da qualidade dos dados por meio da detecção de dados incorretos, sua correção quando possível ou até mesmo a exclusão do registro. No caso da uniformização é comum, na realidade do IFC, os diferentes *campi* adotarem nomenclaturas distintas para expressar a mesma situação. Uma das diversas situações pode ser exemplificada pelo registro de "transferência interna" ser tratado por alguns como "mudança de curso", ou ainda, a utilização da expressão "formado" por alguns e "concluído" por outros.

A fase de construção, por sua vez, consiste na transformação de atributos ou na criação de novas variáveis. A transformação de variáveis pode ser necessária por diversas razões: simplificação dos dados para melhor representá-lo, exigência da ferramenta computacional para realizar a modelagem, atendimento a pressupostos estatísticos, dentre outros. A criação de variáveis, por outro lado, é a criação de um novo atributo a partir dos dados existentes. Neste trabalho, por exemplo, a variável tempo percorrido foi formada a partir dos endereços dos estudantes e do *campus* onde a matrícula estava vinculada.

Normalmente estas fases são as mais demoradas do processo de mineração de dados, pois é constante a necessidade de retorno a fases anteriores até a obtenção de uma base de dados adequada à fase de modelagem. Neste estudo a limpeza e construção dos dados serão tratadas na análise de cada variável selecionada na etapa de entendimento dos dados. Em cada análise de variável serão abordados os aspectos de limpeza e transformações que se fizeram necessárias.

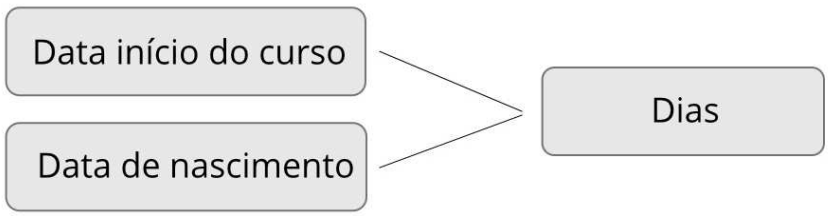
Obter uma base de dados limpa, com as construções e adaptações de variáveis realizadas, não significa que as variáveis da base sejam efetivamente relevantes para explicar o fenômeno que se pretende estudar. Uma forma de verificar se as variáveis inicialmente selecionadas têm relação com a variável resposta é por meio de regressões. Assim é possível verificar se cada variável individualmente tem influência na determinação do status de matrícula.

3.2.3.1 TRANSFORMAÇÃO DAS VARIÁVEIS

V03 - Data de nascimento: inicialmente foram corrigidos valores discrepantes provavelmente resultantes de erros de digitação no momento da matrícula do aluno. Valores como, por exemplo, nascimentos anteriores a 1900 ou posteriores à data atual foram reportados ao setor de registro acadêmico e o corrigidos na base de dados.

Além disso, a data de nascimento em si não é uma informação relevante para o objetivo deste trabalho. A hipótese subentendida nessa variável é que a idade do aluno durante a realização do curso tenha alguma influência na propensão a evadir. A informação mais precisa é a idade do estudante no momento de início do curso. Para obter esse dado foi preciso transformar essa variável, subtraindo a data de nascimento da data de início do curso. Assim, criou-se a variável chamada "dias", calculada a partir da diferença em dias entre estas duas datas, conforme pode ser observado na figura abaixo:

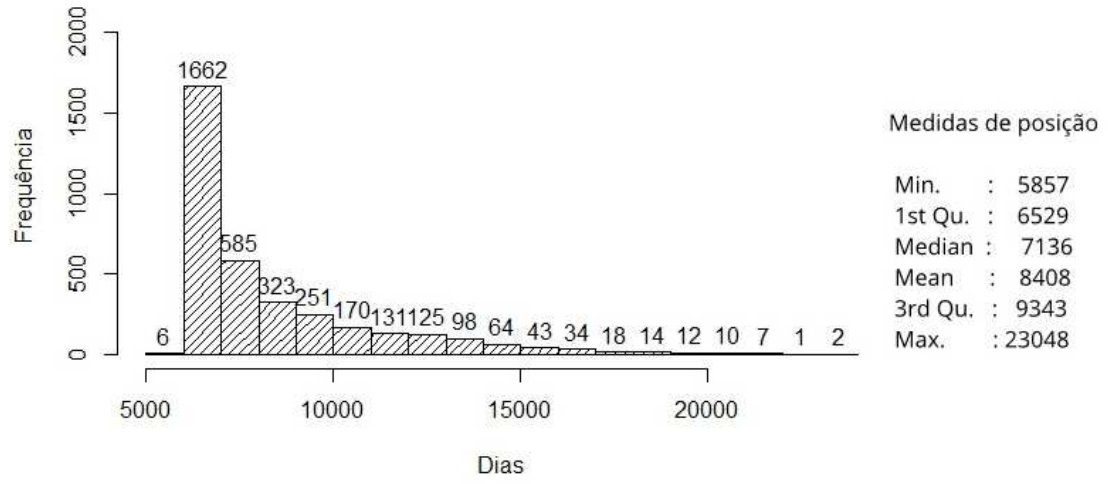
Figura 6: Ilustração da criação da variável Dias.



Fonte: autoria própria

Quanto à distribuição da variável dias, pode-se observar no gráfico 1 que a maioria dos estudantes possuía entre 6 e 8 mil dias de vida no momento do ingresso no curso, ou seja, tinham aproximadamente entre 17 e 22 anos. Isto indica uma distribuição assimétrica com valores concentrados na faixa inicial da distribuição e cauda na direção dos valores maiores.

Figura 7: Distribuição de frequência da variável dias e medidas de posição.



Fonte: autoria própria

As medidas de posição da distribuição podem ser vistas na tabela ao lado direito do gráfico. A média (*mean*) de idade dos alunos ao ingressarem nos cursos superiores do IFC é de 8.408 dias, equivalente à aproximadamente 23 anos. O aluno mais novo registrado tinha pouco mais de 16 anos (5.857 dias) de idade enquanto o mais velho ostentava pouco mais de 63 anos (23.048 dias) no início das aulas.

Pode-se também observar a assimetria da distribuição pela diferença de amplitude entre os quartis. Enquanto a diferença entre a mediana (*median*) e o primeiro quartil (*1st qu.*) é de 607 dias a diferença entre o terceiro quartil (*3rd qu.)* e a mediana é de 2.207 dias. Isto significa que os registros estão mais concentrados nos menores valores da distribuição.

A assimetria observada não impede a realização de regressão logística, pois uma das vantagens de utilizar deste tipo de regressão é a não exigência de normalidade na distribuição da variável independente (HAIR ET AL, 2009). Logo, é possível fazer a regressão para verificar a influência da variável independente, neste caso dias, sobre a variável resposta (status). O resultado da regressão pode ser observado abaixo.

```
glm(formula = status ~ dias, family = "binomial", data = p3033)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6123  -0.9625  -0.9335   1.3427   1.4662

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.216e+00  1.099e-01 -11.064  < 2e-16 ***
```

```
dias         9.546e-05  1.233e-05   7.745 9.59e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4783.7  on 3555  degrees of freedom
Residual deviance: 4722.3  on 3554  degrees of freedom
AIC: 4726.3

Number of Fisher Scoring iterations: 4

> exp(m$coefficients)
(Intercept)        dias
  0.2963011   1.0000955
```

A primeira linha observada na regressão significa o comando utilizado no software R para realizar a regressão. Nele são indicadas quais variáveis fazem parte do modelo, qual a função empregada, neste caso a logit[13] simbolizada pela expressão “binomial”, e a base de dados utilizada.

Em seguida é apresentada a distribuição dos desvios residuais (*deviance residuals*). Esta é uma medida de ajuste do modelo que será retomada mais adiante quando comparados o desvio residual (*residual deviance*) e o desvio residual do modelo nulo (*null deviance*).

Na próxima parte do resultado da regressão são mostrados os coeficientes logísticos (*estimate*), seus erros padrões[14] (*std. error*), os z estatísticos[15] (*z value*) e os p-valores (*pr (>|z|)*) associados à estimação dos coeficientes. A interpretação se concentra no sinal do coeficiente, na magnitude do impacto na probabilidade prevista e na significância da sua estimação (p-valor).

No caso examinado, percebe-se que o coeficiente da variável dias é positivo, isto é, quanto maior a idade do aluno maior é a probabilidade de ter sua matrícula cancelada. Para cada dia adicional a razão de chance de evasão aumenta em 0,00009546. Além disso, a estimativa é estatisticamente significante, pois o p-valor calculado ficou abaixo de 0,05, ou seja, é uma variável relevante para o entendimento do problema.

---

[13] Logit é o cálculo que transforma o valor da probabilidade em uma variável métrica usada como variável dependente no modelo de regressão. Seu cálculo é dado pelo logaritmo das razões de chance (HAIR ET AL, 2009).

[14] Semelhante ao desvio padrão de um conjunto de dados, o erro padrão informa a amplitude esperada do coeficiente em múltiplas amostras dos dados (HAIR ET AL, 2009)

[15] Z Estatístico é calculado dividindo coeficiente da regressão pelo seu erro padrão. Trata-se de um valor de significância da estimação do coeficiente da regressão.

Abaixo dos coeficientes são exibidas as medidas de ajuste do modelo, dadas pela *deviance* e pelo AIC (*Akaike Information Criterion*). Ambas as medidas têm interpretação semelhante, quanto menor o valor melhor é o ajuste preditivo do modelo. A medida do *deviance* é mais facilmente interpretável no caso apresentado, pois permite a comparação direta entre o modelo nulo, ou seja, sem a variável dias, e o modelo com a inclusão da variável. Na regressão realizada percebe-se que houve melhora no ajuste com o *deviance* reduzindo de 4783,7 do modelo nulo para 4722,3 com a inclusão da variável dias.

Para evitar repetições desnecessárias, nas próximas regressões apresentadas serão analisados diretamente os pontos mais relevantes para o entendimento da relação entre a variável independente observada e o status da matrícula do aluno. Quais sejam: a interpretação do coeficiente e razão de chance, significância da estimação do coeficiente e medida de ajuste preditivo do modelo.

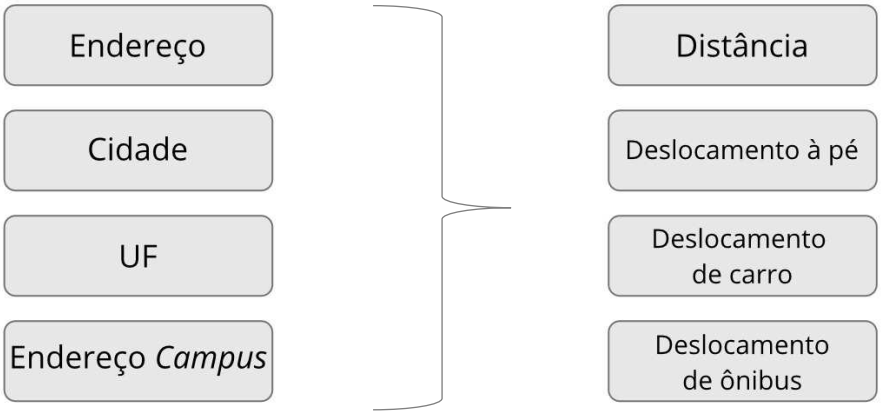
V04 - Endereço do discente: assim como a data de nascimento, o endereço do estudante não carrega consigo alguma informação relevante, mas a partir dele é possível gerar um atributo com potencial de explicar parte do fenômeno da evasão.

Com o endereço do estudante e o endereço do *campus* escolhido é possível calcular a distância e o tempo percorrido entre a residência do estudante até o campus. Para isso foi utilizado o pacote *Distance Matrix Advance*[16] do Google que permite fazer o cálculo automatizado da distância e tempo percorrido entre duas coordenadas geográficas.

Com esse procedimento foi possível criar quatro variáveis numéricas baseadas no endereço do estudante: distância, tempo de deslocamento a pé, tempo de deslocamento de carro e tempo de deslocamento de ônibus, sendo a distância medida em quilômetros e as variáveis de tempo em minutos.

[16] *Distance Matrix Advance* é um serviço oferecido pela empresa Google LLC que permite o cálculo automatizado da distância e tempo de deslocamento entre múltiplas origens e destinações.

Figura 8: Ilustração da transformação das variáveis de endereço.



Fonte: autoria própria

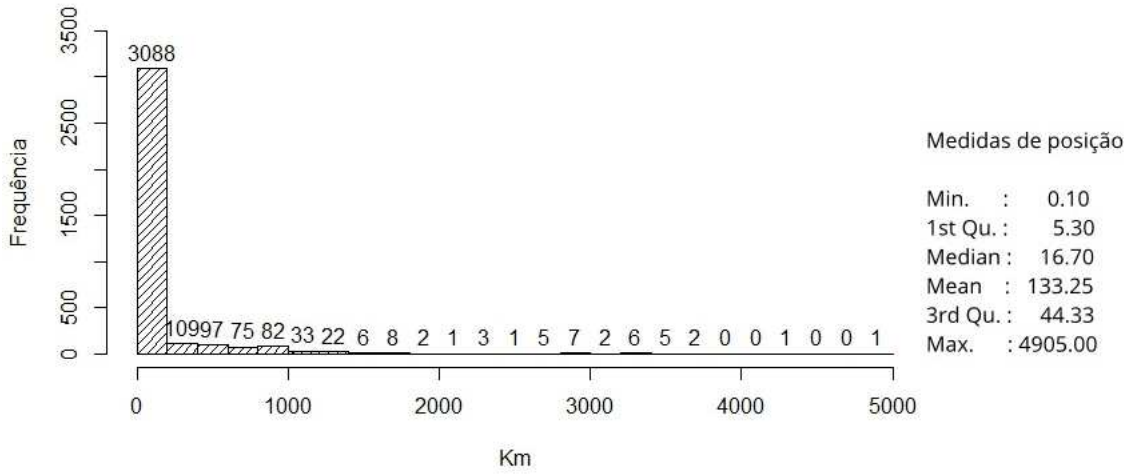
No caso dos dados de tempo percorrido há um detalhe que precisa ser observado. Diferente da distância que é um dado numérico estático, o tempo de deslocamento pode ser maior ou menor dependendo do horário em que se toma a medida. Por exemplo, dois quilômetros podem ser percorridos em menos tempo numa cidade do interior que no horário de pico de uma grande cidade. Então, para colher os dados mais ajustados à realidade vivenciada pelos estudantes o tempo de deslocamento foi calculado 30 minutos antes no início das aulas.

Embora sejam dados correlacionados, espera-se que o tempo percorrido seja mais explicativo que a distância para o problema da evasão pois é o tempo de deslocamento que efetivamente impacta na rotina do estudante. Ou seja, é possível que o tempo de deslocamento tenha mais influência na propensão a evadir dos estudantes.

Além disso, ao extrair os dados da ferramenta do Google e vinculá-los às matrículas dos estudantes percebeu-se que o tempo de deslocamento de ônibus apresentava muitos dados faltantes, muito acima de 5% do total de registros. Isso se deve ao fato de muitas localidades não possuírem linhas de transporte coletivo, logo faltaram informações e por isso esse dado foi descartado. O mesmo problema não aconteceu com a distância, deslocamento de carro e a pé pois não dependem da disponibilidade de nenhum serviço, basta haver uma rota mapeada que o cálculo pode ser realizado.

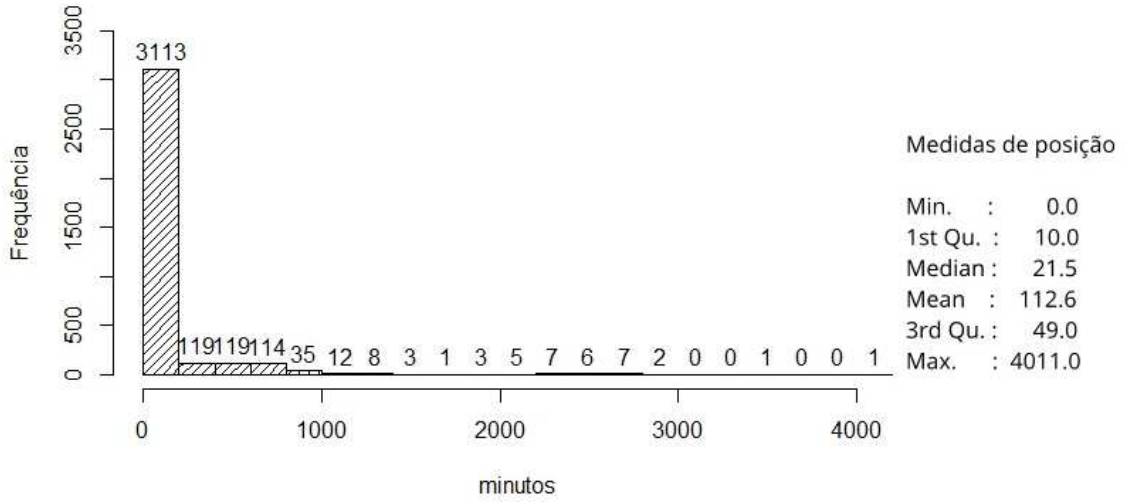
Assim, com base nas informações de distância e deslocamento foram elaborados histogramas da distribuição das informações, conforme pode ser observado nas figuras abaixo:

Figura 9: Distribuição de frequência da variável distância e medidas de posição.



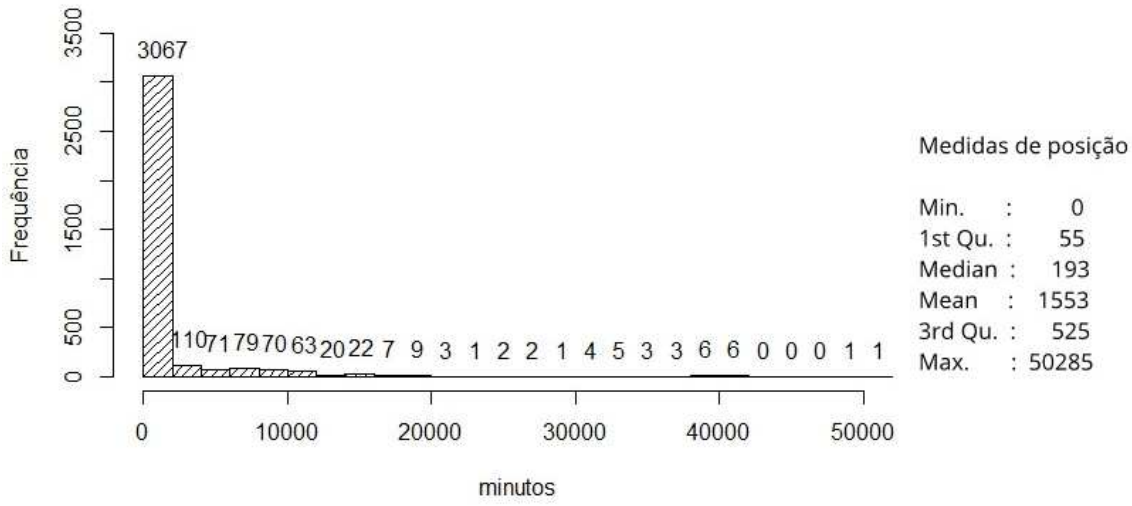
Fonte: autoria própria

Figura 10: Distribuição de frequência da variável deslocamento de carro e medidas de posição.



Fonte: autoria própria

Figura 11: Distribuição de frequência da variável deslocamento a pé e medidas de posição.



Fonte: autoria própria

As três variáveis calculadas (distância, tempo de deslocamento de carro e tempo de deslocamento a pé) apresentam distribuições muito parecidas, com assimetria positiva, isto é, com a maior parte dos registros concentrados à esquerda da do eixo X. Em que pese a diferença de medida e escala, depreende-se, a partir destas distribuições, que a grande maioria dos estudantes reside próximo ao campus escolhido. Além disso, há dados discrepantes, provavelmente referentes a alunos provenientes de outras regiões em que há necessidade de fixar residência nos arredores do *campus* escolhido sem, no entanto, realizar a atualização cadastral junto ao registro acadêmico da unidade.

A falta de atualização dos registros de endereço faz com que a média de distância até a escola seja 133,25 km e, consequentemente, a média de tempo de deslocamento também seja alta, 112,6 minutos de carro e 1.553 minutos a pé. Por outro lado, a mediana da distância mostra que mais da metade dos estudantes residem a menos de 16,7 km de distância do *campus*, portanto levam menos de 21,5 minutos de carro ou 193 minutos a pé para se deslocar de cada para a escola.

A diferença entre os quartis também demonstra a assimetria na distribuição dos dados, com a diferença entre a mediana e o primeiro quartil sendo menor que a diferença entre o terceiro quartil e a mediana. O mesmo comportamento pode ser observado nas três distribuições.

A semelhança nas distribuições reforça a ideia de que estas três variáveis não são independentes entre si. A existência de correlação entre as variáveis independentes pode causar problema de multicolinearidade na classificação, o que dificulta a interpretação de todas as variáveis na análise. “A multicolinearidade ocorre quando qualquer variável independente é altamente correlacionada com um conjunto de outras variáveis independentes” (HAIR ET AL, 2009, p.151), ainda segundo Hair et al (2009) à medida que a multicolinearidade aumenta é mais difícil verificar o efeito de cada variável sobre a variável resposta. O teste de correlação entre as variáveis independentes é uma das últimas etapas realizadas na fase de preparação dos dados, antes de efetivamente aplicar os modelos de classificação, conforme será visto na seção 3.2.3.2.

A seguir são apresentadas as regressões logísticas para cada uma das três variáveis derivadas do endereço do aluno: distância, deslocamento de carro e deslocamento a pé.

Distância

```
glm(formula = status ~ dist, family = "binomial", data = p3033)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.025  -1.009  -1.009   1.356   1.356

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.107e-01  3.621e-02 -11.343   <2e-16 ***
dist         9.482e-06  8.807e-05   0.108    0.914
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4783.7  on 3555  degrees of freedom
Residual deviance: 4783.7  on 3554  degrees of freedom
AIC: 4787.7

Number of Fisher Scoring iterations: 4

> exp(m$coefficients)
(Intercept)        dist
  0.6631756   1.0000095
```

Deslocamento de Carro

```
glm(formula = status ~ carro, family = "binomial", data = p3033)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.023  -1.009  -1.009   1.356   1.356

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.106e-01  3.656e-02 -11.232   <2e-16 ***
carro        1.026e-05  1.134e-04   0.091    0.928
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4783.7  on 3555  degrees of freedom
Residual deviance: 4783.7  on 3554  degrees of freedom
AIC: 4787.7

Number of Fisher Scoring iterations: 4

> exp(m$coefficients)
(Intercept)       carro
  0.6632472   1.0000103
```

Deslocamento a pé

```
glm(formula = status ~ ape, family = "binomial", data = p3033)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.024  -1.009  -1.009   1.356   1.356

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.107e-01  3.622e-02 -11.339   <2e-16 ***
ape          7.748e-07  7.577e-06   0.102    0.919
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4783.7  on 3555  degrees of freedom
Residual deviance: 4783.7  on 3554  degrees of freedom
AIC: 4787.7

Number of Fisher Scoring iterations: 4

> exp(m$coefficients)
(Intercept)         ape
  0.6632158   1.0000008
```

Observa-se, a partir das regressões acima, que individualmente as três variáveis apresentam coeficientes logísticos estimados positivos, indicando que um aumento nas variáveis independentes aumenta a probabilidade do evento (evasão) ocorrer. O resultado apresentado está alinhado ao que seria esperado quando se imagina que quanto mais longe é a residência do estudante. No entanto, essa observação não é ratificada pela significância dos coeficientes.

Quando se examina a significância da estimação dada pelo p-valor (Pr(>|z|) dos coeficientes, tem-se os valores de 0,914 para a distância, 0,928 para o deslocamento de carro e 0,919 para o deslocamento a pé. Nota-se que nenhum deles atinge o nível de significância mínimo de 0,05, nível aceitável para indicar, com confiança de 95%, que a variável independente tem influência sobre a variável dependente. Diante disso, pode-se dizer que pelas regressões logísticas realizadas nenhuma das três variáveis analisadas comprova sua influência sobre o status da matrícula do aluno.

Além disso, pode-se analisar também o ajuste preditivo dos modelos comparando os modelos com a inclusão das variáveis com um modelo nulo, isto é, sem inclusão de variáveis independentes. A comparação se dá pelo valor desvio (*deviance*), obtido a partir do valor de verossimilhança, quanto menor o valor melhor é o ajuste do modelo proposto. Nas regressões apresentadas acima, verifica-se, para os três casos, que não houve melhora no ajuste dos modelos permanecendo o valor 4783.7 com a inclusão da variável.

Constatou-se, então, que as variáveis independentes distância, deslocamento de carro e deslocamento a pé não apresentam coeficientes logísticos com nível de significância aceitável, informação corroborada com a pequena melhora do ajuste preditivo dos modelos com a inclusão destas variáveis. Com base nestas constatações, as referidas variáveis não serão utilizadas na etapa seguinte de modelagem dos algoritmos de classificação.

V05 – Município: este é um atributo que combinado com o endereço do estudante e a unidade federativa fornece o endereço completo de residência do discente. Vide V04.

V06 - Unidade Federativa: idem ao município, este atributo complementa as informações necessárias para endereço do estudante. Vide V04.

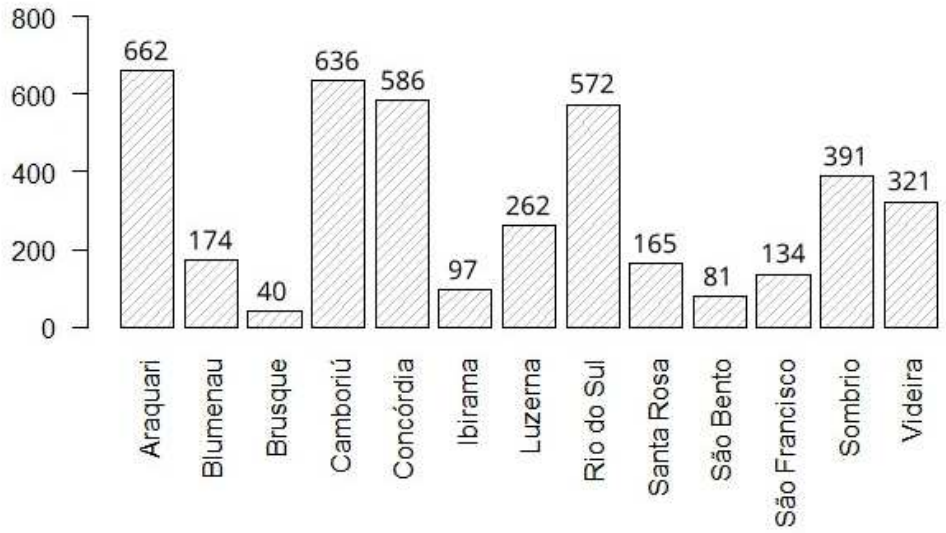
V07 - *Campus*: conforme abordado anteriormente na etapa de entendimento dos dados, o *campus* escolhido é uma variável categórica que traz a informação do ambiente de convivência encontrado pelo estudante. É verdade que características pessoais e da turma frequentada fazem parte da construção dessa atmosfera estudantil, e que isso influencia de alguma forma a decisão de permanência do aluno. No entanto, parte desse ambiente depende da infraestrutura física do *campus*, das ações desportivas e culturais desenvolvidas na unidade de ensino, dos projetos de ensino, pesquisa e extensão, além de uma cultura própria de acolhimento que pode diferir entre *campi* da mesma instituição.

É nesse sentido que a variável *campus* pode auxiliar na mensuração da propensão a evadir dos novos ingressantes na instituição, carregando consigo parte das informações relativas ao ambiente encontrado pelo estudante no momento do ingresso na instituição.

Não houve necessidade de alteração nesta variável, pois se trata de uma variável categórica em que qualquer modificação acarretaria perda de informações sem justificativa

lógica. Além disto, como todas as matrículas necessariamente precisam estar vinculadas a um *campus* específico não houve registro de dados faltantes.

Figura 12: Distribuição dos estudantes por *campus*.



Fonte: autoria própria

Observa-se no histograma acima que há diferenças na distribuição dos alunos por campus, o campus Araquari, com 662 alunos (16,06%), apresenta o maior número de alunos de ensino superior do IFC, seguido do campus Camboriú com 636 (15,43%) alunos. De outro lado os menores campi em número de alunos são Brusque (0,97%), Ibirama (2,35%) e São Bento do Sul (1,97%), cada um com menos de 100 alunos com registro no ensino superior.

Quando se testa a influência da variável campus sobre o status do aluno por meio da regressão logística, obtém-se um coeficiente logístico para cada campus em relação ao campus Araquari. Pode-se observar na regressão abaixo que a maioria dos campi apresentam coeficientes negativos, sugerindo que nesses campi a probabilidade do aluno ter sua matrícula cancelada é menor que em Araquari. O campus Concórdia, por exemplo, apresenta uma razão se chance de aproximadamente 0,44, ou seja, de cada 100 matrículas canceladas em Araquari apenas 44 são canceladas em Concórdia. A exceção fica por conta de Brusque e Sombrio que possuem coeficientes positivos e com razões de chance levemente superiores a Araquari. São justamente esses dois coeficientes (0.932674 e 0.449732) além do campus São Francisco do

Sul (0,050969) que não atingem o nível de significância mínimo de 0,05, enquanto todos os demais têm sua estimação dentro do nível de significância aceitável, ou seja, menor de 0,05.

```
glm(formula = status ~ campus, family = "binomial", data = p3035)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2501  -1.0315  -0.8825   1.3006   1.6120

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           0.07254    0.07778   0.933 0.351036
BLUMENAU             -0.63893    0.17588  -3.633 0.000280 ***
BRUSQUE               0.02754    0.32604   0.084 0.932674
CAMBORIU             -0.35747    0.11166  -3.201 0.001368 **
CONCORDIA            -0.81473    0.11773  -6.920 4.50e-12 ***
IBIRAMA              -0.64433    0.22528  -2.860 0.004235 **
LUZERNA              -0.65321    0.15047  -4.341 1.42e-05 ***
RIO_DO_SUL           -0.42582    0.11517  -3.697 0.000218 ***
SANTA_ROSA_DO_SUL    -1.05337    0.19133  -5.506 3.68e-08 ***
SAO_BENTO_DO_SUL     -0.99749    0.25840  -3.860 0.000113 ***
SAO_FRANCISCO_DO_SUL -0.37329    0.19126  -1.952 0.050969 .
SOMBRIO               0.09666    0.12788   0.756 0.449732
VIDEIRA              -0.61506    0.13947  -4.410 1.03e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5592.4  on 4120  degrees of freedom
Residual deviance: 5484.3  on 4108  degrees of freedom
AIC: 5510.3

Number of Fisher Scoring iterations: 4

> exp(m$coefficients)
                (Intercept)                    BLUMENAU
                  1.0752351                   0.5278544
                    BRUSQUE                    CAMBORIU
                  1.0279270                   0.6994434
                   CONCORDIA                     IBIRAMA
                  0.4427595                   0.5250165
                     LUZERNA                  RIO_DO_SUL
                  0.5203735                   0.6532348
          SANTA_ROSA_DO_SUL             SAO_BENTO_DO_SUL
                  0.3487609                   0.3688047
       SAO_FRANCISCO_DO_SUL                      SOMBRIO
                  0.6884631                   1.1014870
                    VIDEIRA
                  0.5406081
```

Quando se analisa o ajuste do modelo com a inclusão da variável *campus* por meio da comparação entre o *deviance* do modelo nulo e aquele obtido com a inclusão da variável, nota-se que há melhora no ajuste preditivo. O *deviance* inicial de 5592,4 passa a ser de 5484,3 com a inclusão do *campus* na análise.

Esta melhora de ajuste pode ser entendida como a parte do fenômeno da evasão que é explicada pelo campus escolhido pelo estudante. No entanto, como houve três casos em que a relação entre o campus e o status da matrícula do estudante não atingiram o nível de

significância estatística aceitável, a variável campus será utilizada na etapa de mineração de dados excluindo-se os alunos dos campi Brusque, São Francisco do Sul e Sombrio.

Abaixo pode-se observar o resultado da regressão com a exclusão dos *campi* não significativos. Obviamente há uma redução do número de estudantes considerados, antes eram 4.121 alunos, agora são 3.556, e os coeficientes dos *campi* remanescentes não se alteraram.

```
glm(formula = status ~ campus, family = "binomial", data = p3033)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2084  -1.0315  -0.8825   1.3006   1.6120

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        0.07254    0.07778   0.933 0.351036
BLUMENAU          -0.63893    0.17588  -3.633 0.000280 ***
CAMBORIU          -0.35747    0.11166  -3.201 0.001368 **
CONCORDIA         -0.81473    0.11773  -6.920 4.50e-12 ***
IBIRAMA           -0.64433    0.22528  -2.860 0.004235 **
LUZERNA           -0.65321    0.15047  -4.341 1.42e-05 ***
RIO_DO_SUL        -0.42582    0.11517  -3.697 0.000218 ***
SANTA_ROSA_DO_SUL -1.05337    0.19133  -5.506 3.68e-08 ***
SAO_BENTO_DO_SUL  -0.99749    0.25840  -3.860 0.000113 ***
VIDEIRA           -0.61506    0.13947  -4.410 1.03e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4783.7  on 3555  degrees of freedom
Residual deviance: 4707.0  on 3546  degrees of freedom
AIC: 4727

Number of Fisher Scoring iterations: 4

> exp(m$coefficients)
        (Intercept)          BLUMENAU          CAMBORIU         CONCORDIA
          1.0752351         0.5278544         0.6994434         0.4427595
            IBIRAMA           LUZERNA        RIO_DO_SUL   SANTA_ROSA_DO_SUL
          0.5250165         0.5203735         0.6532348         0.3487609
    SAO_BENTO_DO_SUL          VIDEIRA
          0.3688047         0.5406081
```

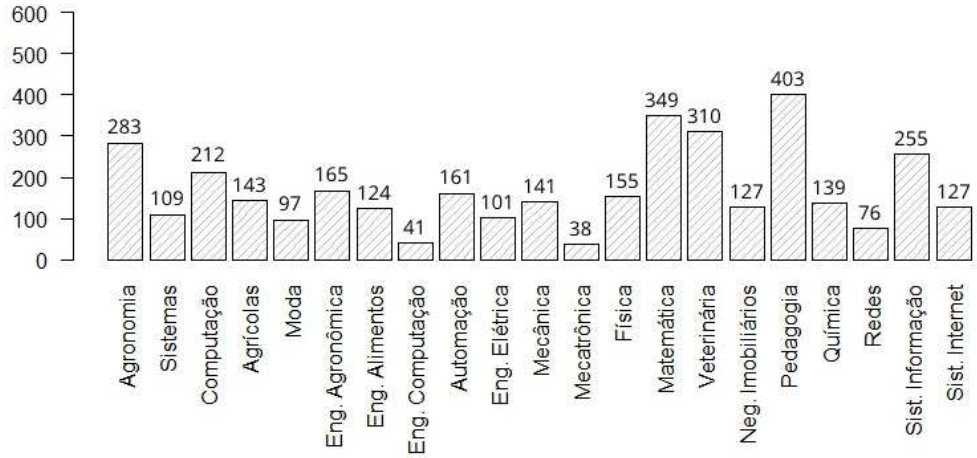
Diferença significativa pode ser verificada no ajuste do modelo, enquanto o *deviance* da regressão com todos os campi era de 5.484,3 com a exclusão dos *campi* não significativos o *deviance* foi reduzido para 4.707,0. Isto demonstra que a exclusão daqueles *campi* da análise melhora consideravelmente o potencial de predição do modelo justificando portanto que se mantenham apenas os *campi* com nível de significância aceitável na sequência do trabalho.

V08 – Curso: o curso escolhido pelo estudante compreende uma série de fatores que podem influenciar sua decisão de evadir ou permanecer estudando. Parte deles está relacionado ao ambiente proporcionado pela infraestrutura voltada ao curso, corpo docente, currículo do

curso. O curso também tangencia um aspecto pessoal importante, a vocação. É sabido que a vocação tem influência decisiva na decisão dos alunos trocar de curso ou abandonar os estudos para seguir outro caminho (LIMA, PIMENTEL, 2017).

Não houve necessidade de alteração desta variável, nem foram registrados dados faltantes.

Figura 13: Distribuição dos estudantes por curso.



Fonte: autoria própria

No histograma acima, pode-se observar como se dá a distribuição dos estudantes nos 23 cursos ofertados e, destes, quais os cursos com maior número de alunos registrados no período analisado. O curso de Pedagogia com 11,33 % dos alunos (403), o curso de Matemática com 9,81% dos alunos (349) e o curso de Veterinária com 8,72% dos alunos (310) são aqueles com maior número de estudantes. De outro lado, os menores cursos quanto à quantidade de alunos são Engenharia da Computação com 1,15% dos alunos (38) e Mecatrônica com 1,07% dos alunos (38).

O resultado da regressão logística realizada com a variável curso, à semelhança daquela feita com a variável *campus*, apresenta parte dos cursos com coeficiente logístico positivo enquanto outros têm coeficiente negativo. A significância de tais coeficientes também é diversa com alguns deles ficando com p-valor acima de 0,05. A presença de coeficientes não significativos indica que não é possível afirmar que estes cursos tenham real associação com a determinação do status da matrícula dos alunos.

No entanto, embora a significância não tenha ficado dentro do padrão aceitável para todos os cursos, o ajuste do modelo melhorou com a inclusão dos cursos. O *deviance* reduziu de 4466.3 no modelo nulo para 4167.4 com a inclusão da variável.

Poder-se-ia excluir da análise os cursos com baixa significância e manter para a fase de mineração de dados apenas aqueles cursos com p-valor abaixo de 0,05. Neste caso, entretanto, haveria considerável perda de informações para uma variável que demonstrou melhora do ajuste do modelo. Por isso, optou-se pode manter a variável cursos na etapa de aplicação dos algoritmos de classificação.

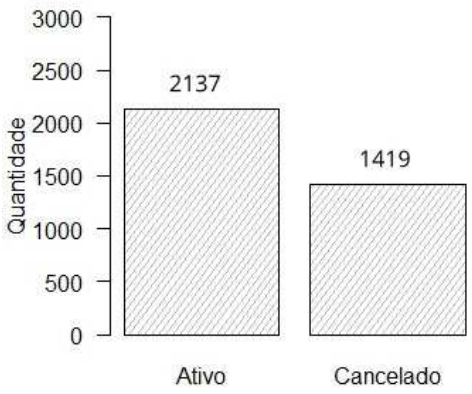
```
glm(formula = status ~ curso, family = "binomial", data = p3033)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6105  -0.9862  -0.7552   1.1841   1.8597

Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                         -0.91382    0.13151  -6.948 3.69e-12 ***
ANALISE_E_DESENVOLVIMENTO_DE_SISTEMAS  0.67415    0.23350   2.887  0.00389 **
CIENCIA_DA_COMPUTACAO                0.78155    0.19039   4.105 4.04e-05 ***
CIENCIAS_AGRICOLAS                   1.47343    0.21798   6.759 1.39e-11 ***
DESIGN_DE_MODA                       0.34203    0.24899   1.374  0.16954
ENGENHARIA_AGRONOMICA               -0.06701    0.21875  -0.306  0.75935
ENGENHARIA_DE_ALIMENTOS             -0.18479    0.24557  -0.752  0.45175
ENGENHARIA_DE_COMPUTACAO             0.25704    0.35463   0.725  0.46857
ENGENHARIA_DE_CONTROLE_E_AUTOMACAO   0.44583    0.20863   2.137  0.03260 *
ENGENHARIA_ELETRICA                  0.45025    0.24304   1.853  0.06394 .
ENGENHARIA_MECANICA                  0.02222    0.22735   0.098  0.92214
ENGENHARIA_MECATRONICA               0.01588    0.38111   0.042  0.96677
FISICA                               1.42809    0.21177   6.744 1.55e-11 ***
MATEMATICA                           1.09194    0.16985   6.429 1.29e-10 ***
MEDICINA_VETERINARIA                -0.62011    0.19849  -3.124  0.00178 **
NEGOCIOS_IMOBILIARIOS                0.44893    0.22478   1.997  0.04580 *
PEDAGOGIA                           -0.19474    0.17492  -1.113  0.26556
QUIMICA                              1.89135    0.23133   8.176 2.93e-16 ***
REDES_DE_COMPUTADORES                1.45282    0.27174   5.346 8.98e-08 ***
SISTEMAS_DE_INFORMACAO               0.78029    0.18180   4.292 1.77e-05 ***
SISTEMAS_PARA_INTERNET               0.89807    0.22089   4.066 4.79e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4783.7  on 3555  degrees of freedom
Residual deviance: 4434.6  on 3535  degrees of freedom
AIC: 4476.6

Number of Fisher Scoring iterations: 4


> exp(m$coefficients)
                        (Intercept)     ANALISE_E_DESENVOLVIMENTO_DE_SISTEMAS
                          0.4009901                                 1.9623558
                CIENCIA_DA_COMPUTACAO                       CIENCIAS_AGRICOLAS
                          2.1848574                                 4.3641975
                     DESIGN_DE_MODA                    ENGENHARIA_AGRONOMICA
                          1.4078057                                 0.9351852
               ENGENHARIA_DE_ALIMENTOS                ENGENHARIA_DE_COMPUTACAO
                          0.8312757                                 1.2930956
   ENGENHARIA_DE_CONTROLE_E_AUTOMACAO                     ENGENHARIA_ELETRICA
                          1.5617907                                 1.5686977
```

```
ENGENHARIA_MECANICA                ENGENHARIA_MECATRONICA
          1.0224691                          1.0160037
              FISICA                         MATEMATICA
          4.1707109                          2.9800450
MEDICINA_VETERINARIA          NEGOCIOS_IMOBILIARIOS
          0.5378843                          1.5666350
           PEDAGOGIA                           QUIMICA
          0.8230453                          6.6283301
REDES_DE_COMPUTADORES          SISTEMAS_DE_INFORMACAO
          4.2751323                          2.1820988
SISTEMAS_PARA_INTERNET
          2.4548611
```

V08 - Status: esta é a variável resposta, a variável que se deseja prever. Como no sistema de gestão do IFC existem diferentes status para a situação da matrícula, conforme foi mencionado na etapa de entendimento dos dados, e algumas delas acabam por reproduzir informações semelhantes para o contexto desse trabalho, houve a necessidade modificações para melhor atingir os objetivos do estudo. Os status ATIVO, CONCLUÍDO, FORMADO e FORMANDO foram agrupados e representam o sucesso acadêmico. Enquanto os status TRANCADO e CANCELADO foram agrupados e considerados como insucesso acadêmico.

Com a exclusão das matrículas excluídas e o agrupamento das matrículas afins em dois status básicos dicotômicos, Ativo (60,10%) ou Cancelado (39,90%), tem-se as condições de sucesso e insucesso necessárias para a criação dos modelos de classificação. A estrutura final para a aplicação dos algoritmos pode ser visualizada na figura abaixo:

Figura 14: Distribuição dos estudantes por status de matrícula.



Fonte: autoria própria

V09 - Índice de rendimento acadêmico (IRA): refere-se à média das notas dos estudantes nas disciplinas cursadas. Este atributo não é um registro armazenado na base de dados da instituição. O que fica registrado são as notas das disciplinas e o IRA é calculado automaticamente no momento da extração, somando todas as notas e dividindo pelo número de disciplinas cursadas até o momento.

A forma de cálculo dessa variável dificulta a comparação pois para os concluintes o dado é definitivo uma vez que o curso está finalizado e para aqueles estudantes que ainda não concluíram o curso o IRA ainda pode ser alterado. Portanto, não é possível comparar o IRA calculado no mesmo estágio do curso, como por exemplo o IRA ao final do primeiro semestre letivo. Além do que, a identificação das possíveis evasões é útil se for realizada no início do curso, a tempo da instituição de ensino tomar alguma medida, e para isso o IRA ao final do curso não tem muita utilidade.

Por estes motivos o IRA não será utilizado neste estudo. No entanto a informação do rendimento acadêmico dos estudantes, sobretudo no início do curso, é relevante e pode auxiliar na identificação daqueles estudantes mais propensos a evadir. Para isso, buscou-se num segundo momento junto ao setor de tecnologia da informação do IFC a lista de disciplinas cursadas pelos estudantes no primeiro semestre e o resultado, se aprovados ou reprovados. Este dado traz a informação do rendimento acadêmico do estudante no mesmo estágio do curso, final do primeiro semestre, e atende à necessidade da identificação precoce, já que é um dado que pode ser utilizado logo ao final do primeiro semestre do curso.

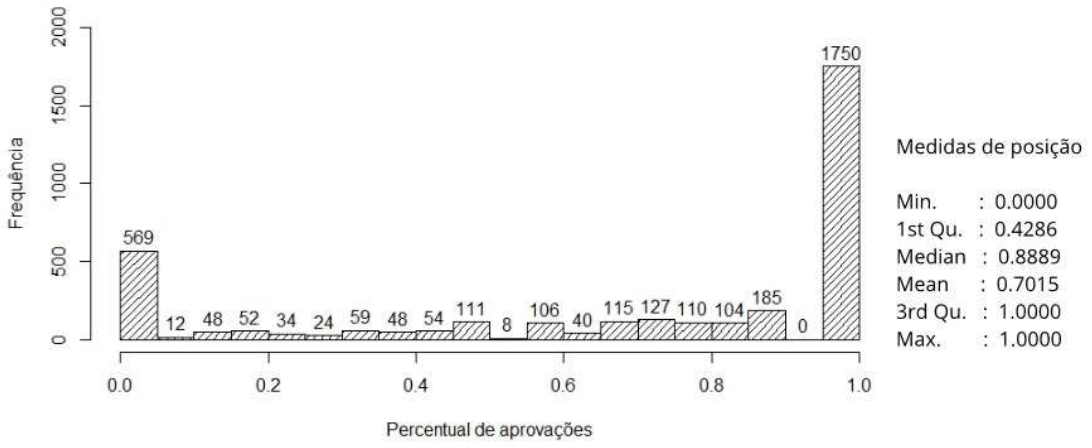
A partir das aprovações e reprovações no primeiro semestre pode-se calcular diretamente a quantidade de disciplinas aprovadas, trazendo a ideia de quanto maior o número de aprovações melhor o desempenho do estudante. Contudo, o número de disciplinas varia entre os cursos. Por exemplo, o curso regular de Química prevê 5 disciplinas no primeiro semestre enquanto nos cursos de período integral como Agronomia estão previstas até 8 disciplinas. A comparação do número de disciplinas aprovadas, neste caso, não pode ser realizada, pois um estudante de Agronomia com aprovação em 6 disciplinas das 8 cursadas seria considerado mais proficiente que outro aluno que tivesse sido aprovado em todas as 5 disciplinas cursadas do curso de Química.

Dessa forma, para lidar com a diferença do número de disciplinas entre os cursos, foi calculado o percentual de aprovação de cada aluno. Este novo atributo varia de 0 a 1, com 0 representando aqueles estudantes que reprovaram em todas as disciplinas e 1 representando os

estudantes que obtiveram aprovação máxima. Assim é possível utilizar o percentual de aprovação nos modelos de classificação pois é um dado representativo da mesma etapa de ensino para todos os estudantes e considera a diferença entre o número de disciplinas cursadas.

A distribuição dos alunos de acordo com o percentual de aprovação pode ser observada no gráfico 7, nele percebe-se que dois conjuntos de estudantes se destacam. Inicialmente aqueles que não registraram aprovações compreendendo 569 alunos e por fim aqueles que foram aprovados em todas as disciplinas, caso de 1750 alunos. Entre esses dois conjuntos de estudantes a frequência é menor, não passando de 200 alunos em cada grupo.

Figura 15: Distribuição de frequência da variável percentual de aprovações nas disciplinas do primeiro semestre.



Fonte: autoria própria

O percentual médio de aprovações (*mean*) é de 0,7015. A grande concentração de alunos com percentual máximo ou próximo do máximo de aprovações faz com que a mediana (*Median = 0,8889*) esteja localizada bem à direita da distribuição. O posicionamento dos quartis também reflete a concentração dos alunos próximo ao 1, estando o terceiro quartil (*3rd Qu.*) localizado no 1 e o primeiro quartil (*1st Qu.*) no 0,4286, próximo à metade da amplitude da distribuição.

A verificação da relação entre o percentual de aprovação e o status da matrícula pode ser verificado na regressão logística abaixo. Verifica-se que o coeficiente logístico estimado é negativo, ou seja, quanto maior o percentual de aprovação menor a probabilidade do estudante ter sua matrícula cancelada. A razão de chance calculada (0.04576893) traz a informação de que a cada ponto percentual adicional reduz a chance de evasão em aproximadamente 4%.

O grau de confiança para o coeficiente estimado atende o nível de significância mínimo desejado (<0,05), evidenciando que o percentual de aprovação é um dado explicativo de parte do fenômeno da evasão. Isso também pode ser observado pela melhora no ajuste do modelo medido pela diferença de *deviance* entre o modelo nulo e o modelo com a inclusão da referida variável, houve redução do *deviance* de 4783,7 para 3828,8. Cabe salientar que esta foi a maior redução do *deviance* observada.

```
glm(formula = status ~ perc_aprov, family = "binomial", data = p3033)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9439  -0.6764  -0.6764   0.6691   1.7818

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.72556    0.09074   19.02   <2e-16 ***
perc_aprov  -3.08415    0.11433  -26.98   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4783.7  on 3555  degrees of freedom
Residual deviance: 3828.8  on 3554  degrees of freedom
AIC: 3832.8

Number of Fisher Scoring iterations: 4

> exp(m$coefficients) # razão de chance
(Intercept)  perc_aprov
 5.61568403  0.04576893
```

O fato da variável percentual de aprovação ficar disponível em tempo diferente das demais variáveis consideradas é uma particularidade que precisa ser trabalhada de modo específico. Tem-se dois momentos, o primeiro refere-se às informações prévias ao estudante iniciar as aulas, o segundo momento, além das informações prévias, agrega a variável chamada aqui de percentual de aprovação. Então, neste trabalho esses momentos distintos serão analisados separadamente conforme será explicado no decorrer do estudo.

V10 - Forma de ingresso: ao analisar os dados extraídos verificou-se grande variedade de categorias em razão de nomenclaturas diferentes utilizadas para descrever a mesma informação. Isto se deve à migração de antigas bases para o SIG. Após o procedimento de limpeza dos dados, excluindo dados faltantes, alterando quando possível os erros de registro e agrupando as diferentes nomenclaturas obteve-se as seguintes categorias que a seguir serão detalhadas:

Aluno Especial
Transferência Interna
Transferência Externa
SISU - AC
SISU - EPBRNPPI
SISU - EPBRPPI
SISU - EPQRNPPI
SISU – EPQRPPI
Vagas não ocupadas - AC
Vagas não ocupadas – EPBRNPPI
Vagas não ocupadas - EPBRPPI
Vagas não ocupadas - EPQRNPPI
Vagas não ocupadas – EPQRPPI

A forma de ingresso Aluno Especial foi descartada, pois representa o compromisso do estudante com alguma disciplina isolada e não tem relação com o sucesso ou insucesso no curso. Transferência Interna e Externa também não foram consideradas neste estudo pois ambos os casos acontecem ao longo do curso e não se referem a uma condição inicial.

Observa-se também que as categorias restantes trazem duas informações distintas, a etapa do processo seletivo (SISU e Vagas não ocupadas) e a cota para a qual o estudante se candidatou (AC, EPBRNPPI, EPBRPPI, EPQRNPPI e EPQRPPI). É possível então separar as duas variáveis já que são essencialmente aspectos distintos do processo seletivo.

A etapa do processo seletivo diz respeito ao tipo de processo, ou seja, à fase do processo seletivo em que o estudante foi chamado e realizou a matrícula. O processo seletivo do IFC compreende duas fases. A etapa inicial, chamada aqui de SISU, corresponde à primeira chamada do processo seletivo. Os candidatos melhor classificados de acordo com a nota do ENEM são chamados a realizar a matrícula dentro de um prazo estabelecido em edital. Decorrido este prazo, faz-se uma nova chamada para preencher as vagas e assim sucessivamente até o preenchimento integral das vagas ofertadas. Se, ainda assim, restarem vagas é publicado um edital de vagas não ocupadas possibilitando que qualquer pessoa, independentemente de ter realizado o ENEM, possa pleitear o ingresso na instituição.

Ao separar a informação da etapa do processo da forma de ingresso cria-se uma variável categórica que identifica em qual estágio do processo seletivo ocorreu o ingresso do estudante. A hipótese que fundamenta esta variável é que as primeiras vagas são preenchidas por estudantes de maior proficiência e teriam, portanto, menor dificuldade em cumprir com as exigências do curso, uma vez que é sabido que uma das causas da evasão é a dificuldade em acompanhar o conteúdo ministrado (MEC, 1996)

A variável Forma de Ingresso, por sua vez, está relacionada às ações afirmativas as quais os ingressantes são vinculados. Desde 2012, com a promulgação da Lei 12.711 art. 1º, conhecida popularmente como Lei das Cotas, 50% das vagas de ensino superior em instituições federais vinculadas ao MEC são reservadas para pessoas que atendem determinados critérios de renda e raça.

> Art. 3º Em cada instituição federal de ensino superior, as vagas de que trata o art. 1º desta Lei serão preenchidas, por curso e turno, por autodeclarados pretos, pardos e indígenas e por pessoas com deficiência, nos termos da legislação, em proporção ao total de vagas no mínimo igual à proporção respectiva de pretos, pardos, indígenas e pessoas com deficiência na população da unidade da Federação onde está instalada a instituição, segundo o último censo da Fundação Instituto Brasileiro de Geografia e Estatística – IBGE (BRASIL, 2012)

Basicamente os candidatos são classificados em categorias distintas de acordo com a cota escolhida:

AC – Ampla Concorrência;

EPBRPPI – Escola pública, baixa renda, preto, pardo ou indígena;

EPBRNPPI – Escola pública, baixa renda, não preto, pardo ou indígena;

EPQRPPI – Escola pública, qualquer renda, preto, pardo ou indígena;

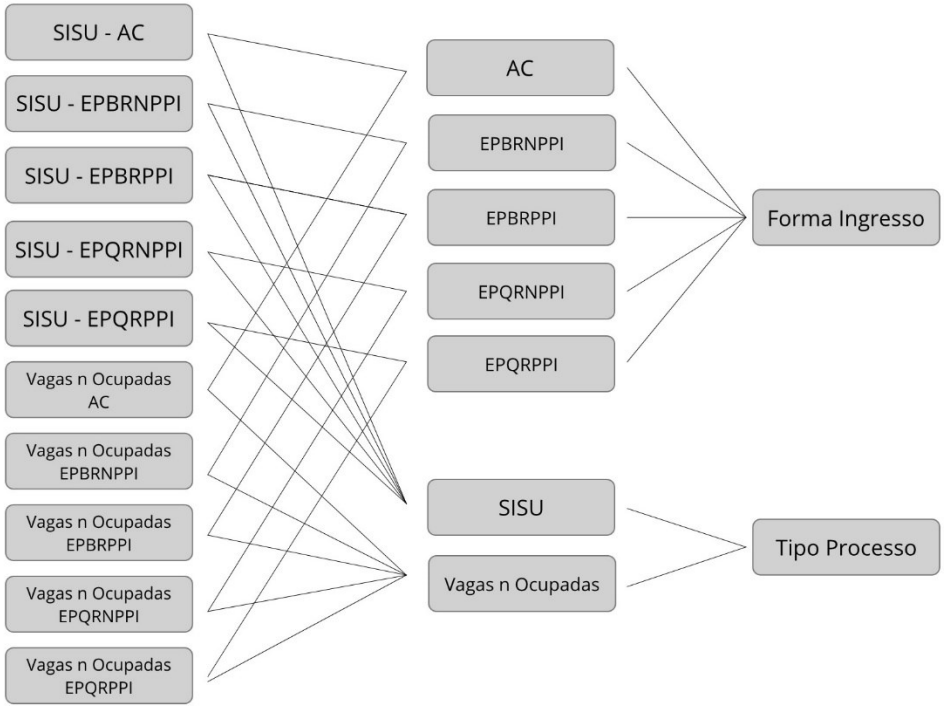
EPQRNPPI – Escola pública, qualquer renda, não preto, pardo ou indígena;

Discricionariamente o IFC permite que todos os candidatos inscritos no processo seletivo concorram primeiramente na Ampla Concorrência – AC – e, num segundo momento, as vagas são destinadas aos cotistas. Isso resulta na possibilidade de haver alunos inscritos por cota dentro do grupo de estudantes AC. Atualmente não há formas de evitar isso, nem maneiras de identificar pelo sistema de gestão do IFC se o aluno aprovado por Ampla Concorrência fez sua inscrição em alguma cota.

Os agrupamentos e modificações realizadas resultaram em duas variáveis conforme pode ser observado na figura abaixo. A partir disso as variáveis são tratadas separadamente e,

caso atendam aos requisitos de significância e ajuste preditivo dos modelos, podem ser utilizadas na etapa seguinte de modelagem dos algoritmos de classificação.

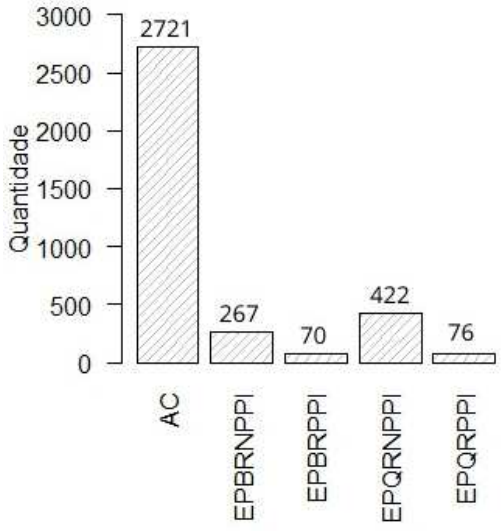
Figura 16: Fluxo de transformação da variável forma de ingresso.



Fonte: autoria própria

A variável Forma de ingresso, como mostrado anteriormente, apresenta 5 categorias relacionadas as ações afirmativas em que os estudantes estão distribuídos conforme gráfico abaixo:

Figura 17: Distribuição dos estudantes de acordo com a forma de ingresso.



Fonte: autoria própria

A categoria Ampla Concorrência (AC) concentra mais de 76% dos estudantes, 2.721 do total de 3.556 alunos. Dos 24% restantes, distribuídos entre as quatro categorias restantes, as cotas EPBRPPI e EPQRPPI, representativas dos estudantes pretos, pardos e indígenas, possuem apenas 70 e 76 alunos respectivamente. Essa grande diferença entre a AC e as demais cotas é explicada pela própria sistemática do processo seletivo do IFC estabelecido em edital que acaba misturando candidatos cotistas entre a ampla concorrência.

Quando se analisa a regressão logística dessa variável, os coeficientes logísticos obtidos permitem interpretar que os estudantes das cotas não pretos, pardos e indígenas (EPBRNPPI e EPQRNPPI) têm menor probabilidade de evadir que os estudantes de ampla concorrência, pois os coeficientes de ambos são negativos. Por outro lado, os coeficientes das cotas EPBRPPI e EPQRPPI, positivos, sugerem que os alunos são mais propensos a evadir que aqueles da ampla concorrência.

Essas interpretações seriam válidas se o nível de significância da estimação dos coeficientes atingisse um nível de confiança aceitável (<0,05). Não é o caso para a variável forma de ingresso. Nenhum coeficiente estimado alcançou p-valor aceitável, por isso a variável não será utilizada na aplicação dos modelos de mineração de dados.

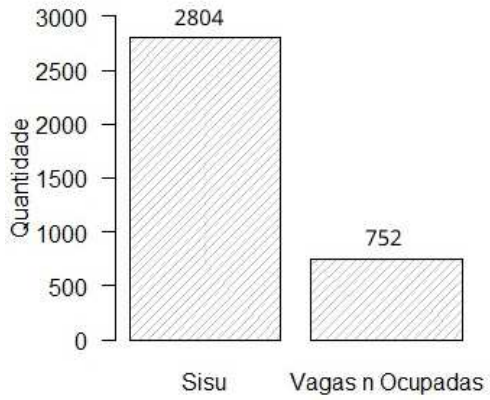
O fato da variável não conseguir explicar parte do fenômeno é observado também quando se olha para o ajuste preditivo do modelo com a inclusão da variável em comparação

com um modelo nulo. A redução do *deviance* foi praticamente inexistente, passando de 4783,7 do modelo nulo para 4780,3 quando se insere a forma de ingresso no modelo como pode ser observado abaixo. Por estes motivos a forma de ingresso não será considerada na etapa seguinte do estudo.

```
glm(formula = status ~ forma_ingresso, family = "binomial", data = p3033)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.153  -1.007  -1.007   1.358   1.394

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           -0.415275   0.039171 -10.602   <2e-16 ***
   EPBRNPPI           -0.081592   0.132134  -0.617    0.537
   EPBRPPI             0.358117   0.242330   1.478    0.139
   EPQRNPPI            0.001905   0.106882   0.018    0.986
   EPQRPPI             0.203966   0.233999   0.872    0.383
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4783.7  on 3555  degrees of freedom
Residual deviance: 4780.3  on 3551  degrees of freedom
AIC: 4790.3

Number of Fisher Scoring iterations: 4

> exp(m$coefficients)
          (Intercept)              EPBRNPPI               EPBRPPI
            0.6601586             0.9216478             1.4306326
             EPQRNPPI               EPQRPPI
            1.0019066             1.2262565
```

Com relação à variável tipo de processo, criada a partir da forma de ingresso, verifica-se pelo gráfico 9 que a maior parte dos alunos, 78,85%, teve seu ingresso durante as chamadas classificatórias do SISU, ou seja, primeira fase do processo seletivo. Apenas 21,15% das vagas precisaram ser preenchidas por editais de vagas não ocupadas.

Figura 18: Distribuição dos estudantes de acordo com o tipo de processo.



Fonte: autoria própria

A regressão logística realizada a partir da variável independente tipo de processo apresenta coeficiente positivo para a categoria vagas não ocupadas. Isto indica que em relação aos inscritos pela classificação do SISU, aqueles ingressantes por edital de vagas não ocupadas tem maior chance de evadir. A razão de chance de um aluno ingressante por vagas não ocupadas ter sua matrícula cancelada é 2,685 maior que um aluno classificado no SISU, ou seja, para cada aluno do SISU com matrícula cancelada 2,685 provenientes de vagas não ocupadas cancelaram.

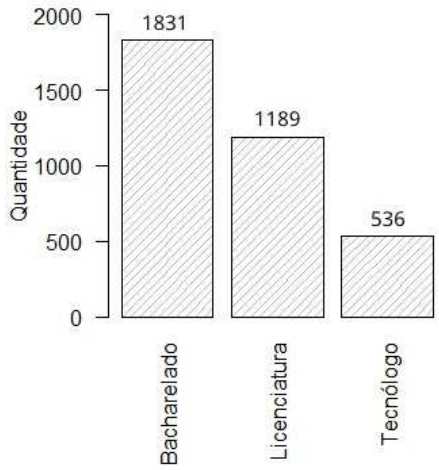
```
glm(formula = status ~ tipo_processo, family = "binomial", data = p3033)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.334  -0.925  -0.925   1.453   1.453

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -0.62752    0.03964  -15.83   <2e-16 ***
vagas_n_ocupadas           0.98774    0.08405   11.75   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4783.7  on 3555  degrees of freedom
Residual deviance: 4642.6  on 3554  degrees of freedom
AIC: 4646.6

Number of Fisher Scoring iterations: 4

> exp(m$coefficients) # razão de chance
                  (Intercept)              vagas_n_ocupadas
                    0.5339168                     2.6851690
```

A significância da estimativa do coeficiente logístico para a variável tipo de processo logrou estar abaixo de 0,05, conforme pode ser observado pelo p-valor na regressão acima, o que significa que alcançou um nível de confiança aceitável para utilizar a variável na etapa posterior de modelagem dos algoritmos de classificação.

V11 – Grau do curso: esta variável está relacionada com o curso escolhido pelo estudante. Por isso, assim como na variável curso, não se verificaram dados faltantes ou erros de registro.

A distribuição da variável nas três categorias pode ser observada no histograma abaixo. Nele percebe-se que os cursos de bacharelado possuem 51,49% dos alunos (1.831), seguido dos cursos de licenciatura com 33,44% (1.189) e, por fim, os cursos de tecnólogo com 15,07% dos estudantes (536).

Figura 19: Distribuição dos estudantes de acordo com o grau do curso.



Fonte: autoria própria

Na regressão logística realizada para verificar se há influência da variável grau no status da matrícula do aluno (abaixo), observou-se que tanto o coeficiente logístico dos cursos de licenciatura quanto o coeficiente dos cursos de tecnólogo são positivos, o que indica maior probabilidade dos alunos nessas modalidades de cursos evadirem em relação aqueles frequentando os cursos de bacharelado. Quando se olha para a razão de chance, observa-se que um estudante de licenciatura tem 1,9619 maior probabilidade de evadir que um estudante de

bacharelado e um aluno de curso tecnólogo tem chance 71,4% maior de ter sua matrícula cancelada que um aluno de bacharelado.

```
glm(formula = status ~ grau, family = "binomial", data = p3033)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1553  -1.0991  -0.8884   1.1996   1.4971

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.72610    0.04985 -14.565  < 2e-16 ***
    licenciatura  0.67394    0.07650   8.810  < 2e-16 ***
    tecnologo     0.53899    0.10007   5.386  7.2e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4783.7  on 3555  degrees of freedom
Residual deviance: 4697.9  on 3553  degrees of freedom
AIC: 4703.9

Number of Fisher Scoring iterations: 4

> exp(m$coefficients) # razão de chance
     (Intercept)     licenciatura        tecnologo
       0.4837925        1.9619573        1.7142710
```

A significância dos coeficientes, observada pelo p-valor, calculadas para a variável grau encontram-se abaixo de 0,05, isto é, conferem à estimativa pelo menos 95% de confiança. Ao se examinar o ajuste preditivo do modelo, nota-se que há uma melhora com a inclusão da variável. O valor do desvio (*deviance*) é reduzido de 4783,7 para 4697,9 com a inserção da variável. Por estas razões a variável grau será incluída na etapa seguinte de mineração de dados.

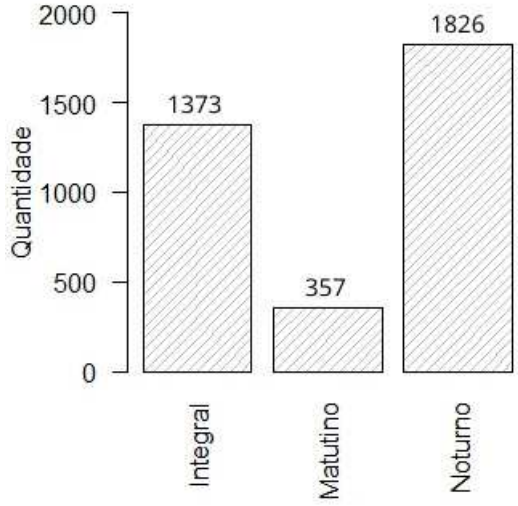
V12 - Período de início do curso: vide V03.

V13 - Tipo de escola: esta variável apresenta duas categorias que representam se o estudante cursou o ensino médio em escola pública ou particular. Embora não se tenha verificado erros de registro, foram observados 754 dados faltantes do total de 3556 registros. Essa quantidade de *missing values* equivale a mais de 20% dos dados. Por esta razão a variável tipo de escola será descartada da análise.

V14 – Turno: esta variável está dividida em três categorias: integral, para os cursos com aulas na parte da manhã e tarde, matutino para os cursos com aulas apenas no período da

manhã e noturno para os cursos com aulas no período da noite. Não se verificaram dados faltantes ou erros de registro para esta variável.

Figura 20: Distribuição dos estudantes conforme o turno do curso.



Fonte: autoria própria

Conforme pode ser observado no gráfico acima, o turno com maior quantidade de alunos é o Noturno com 1.826 (51,35%) estudantes, seguido dos cursos de turno integral com 1.373 (38,61%) alunos enquanto o turno com menor número de alunos é o Matutino com apenas 357 (10,04%).

A verificação da influência da variável turno no status da matrícula dos alunos, realizada pela regressão logística abaixo, demonstra que a probabilidade de um estudante ter sua matrícula cancelada é maior nos turnos matutino e noturno quando comparados com aqueles estudantes de período integral. A razão de chance encontrada é de aproximadamente 2,10 para o turno matutino e 2,22 para o turno noturno, ou seja, para cada 100 alunos de período integral com matrícula cancelada espera-se que 210 do matutino e 222 do noturno tenham sua matrícula cancelada.

Ao se olhar para os cursos integrais, observa-se algumas particularidades que podem explicar essa diferença de probabilidades, embora uma pesquisa mais aprofundada fosse necessária. Cogita-se que essa diferença de probabilidade entre os turnos se dê devido a parte dos cursos integrais contarem com alojamento para os estudantes, como é o caso do curso de

Agronomia do *campus* Santa Rosa do Sul. Além disso, tem-se que os cursos de Medicina Veterinária, cursos de maior concorrência e menor chance de evadir do IFC, conforme pode ser observado nas razões de chance da variável curso, serem de período integral.

Com relação à significância dos coeficientes logísticos estimados, verifica-se que estão dentro do esperado, isto é, p-valor menor de 0,05, para confirmar a influência do turno no status da matrícula. O que é ratificado pela melhora do ajuste do modelo com a inclusão da variável, passando de 4783,7 para 4664,0. Por estes motivos a variável turno será incluída nos modelos de classificação a serem desenvolvidos.

```
glm(formula = status ~ turno, family = "binomial", data = p3033)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1294  -1.1294  -0.8225   1.2262   1.5801

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.91018    0.05966 -15.256  < 2e-16 ***
matutino       0.74736    0.12181   6.135  8.5e-10 ***
noturno        0.79615    0.07588  10.493  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4783.7  on 3555  degrees of freedom
Residual deviance: 4664.0  on 3553  degrees of freedom
AIC: 4670

Number of Fisher Scoring iterations: 4

> exp(m$coefficients) # razão de chance
  (Intercept)      matutino      noturno
    0.4024515     2.1114121     2.2169827
```

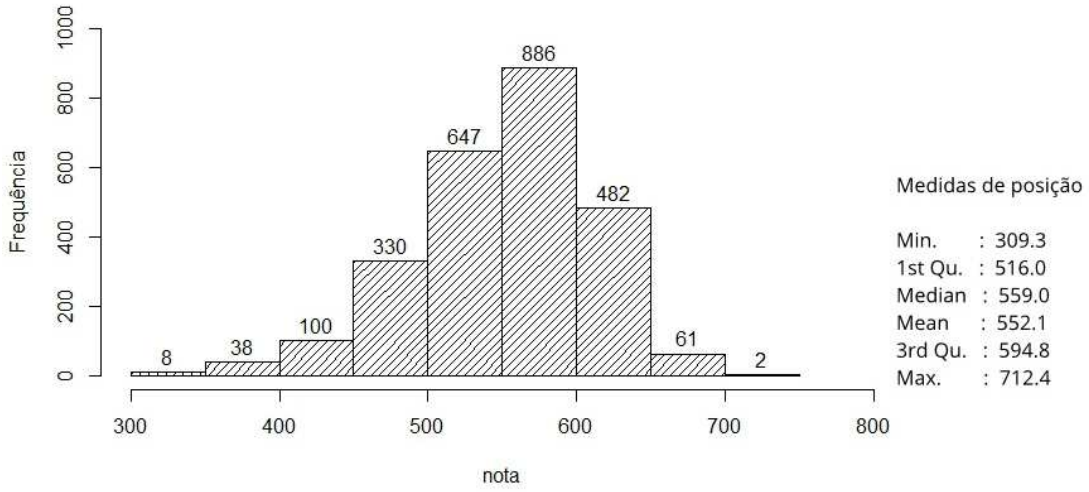
V15 - Nota Linguagens: não são todos os estudantes que apresentam as variáveis referentes ao desempenho nas provas do ENEM. Conforme explicado anteriormente, aqueles alunos que ingressaram na instituição pelos editais de vagas não ocupadas não realizaram as provas e, portanto, não apresentam esses registros de desempenho. O mesmo acontece com aqueles alunos que ingressaram por meio de vestibular próprio realizado nos anos de 2014 e 2015, antes do IFC ofertar todas as vagas pelo SISU. Assim do total de 3.556 alunos da base de dados trabalhada 2.554 possuem notas do ENEM. O mesmo acontece para todas as notas do ENEM que serão apresentadas a seguir.

Como a informação das notas está disponível apenas para uma parte dos estudantes, a base de dados será separada para permitir que se analise a relevância dessas variáveis sem a

necessidade de exclusão dos demais registros. O procedimento de separação de bases será apresentado em seção específica deste trabalho.

Com relação especificamente à prova de linguagens, pode-se observar no histograma abaixo que existe uma leve assimetria negativa na distribuição das notas, com a maior parte dos registros se concentrando no intervalo entre 500 e 600. A média das notas foi de 552,1 pontos e a mediana pouco superior, 559, o que corrobora com a constatação de leve assimetria negativa, devido principalmente ao registro de notas baixas, algumas com pontuação mais de 100 pontos menores que a média. A nota mínima registrada foi de 309,3 enquanto a máxima foi de 712,4. A diferença entre os quartis é relativamente parecida, sendo 43 pontos a diferença entre a mediana e o primeiro quartil enquanto a diferença entre o terceiro quartil e a mediana foi de 35,8.

Figura 21: Distribuição de frequência da variável nota de linguagens e medidas de posição.



Fonte: autoria própria

Na regressão realizada com as notas da prova de linguagens obteve-se um coeficiente logístico negativo, indicando que quanto melhor o desempenho do estudante menor a probabilidade do aluno ter sua matrícula cancelada. A razão de chance calculada para essa variável expressa que cada ponto adicional na prova reduz aproximadamente 0,19% na chance do estudante evadir.

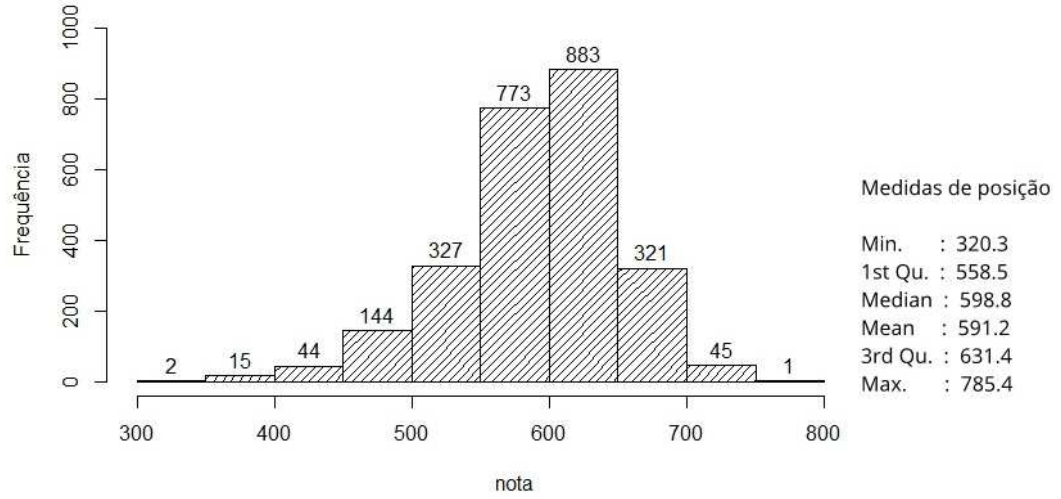
A confiança de tal estimativa está acima de 95% pois o p-valor calculado ficou abaixo de 0,05. Portanto, pode-se dizer que a nota de linguagens tem influência no status da matrícula

dos alunos para este universo de estudantes pesquisado. No entanto, ao analisar o ajuste do modelo observa-se que a redução no *residual deviance*, de 3.337,2 para 3.333,0, é pequena demonstrando que embora haja influência da variável no status da matrícula do aluno essa influência é módica a ponto de colocar em dúvida a manutenção desta variável na aplicação dos algoritmos de classificação.

```
glm(formula = status ~ NU_NOTA_L, family = "binomial", data = p3033)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0769  -0.9486  -0.9188   1.4113   1.5218

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.186683   0.377873   0.494   0.6213
NU_NOTA_L   -0.001383   0.000682  -2.028   0.0425 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3337.2  on 2553  degrees of freedom
Residual deviance: 3333.0  on 2552  degrees of freedom
  (1002 observations deleted due to missingness)
AIC: 3337

Number of Fisher Scoring iterations: 4

> exp(m$coefficients) # razão de chance
(Intercept)   NU_NOTA_L
  1.2052447   0.9986177
```

V16 - Nota Ciências Humanas: visualmente percebe-se que a distribuição das notas da prova de ciências humanas é bastante parecida com a da prova de linguagens, levemente assimétrica para a esquerda. A maior parte das notas está concentrada no intervalo entre 550 e 650 pontos conforme pode ser observado no gráfico abaixo:

Figura 22: Distribuição de frequência da variável nota de ciências humanas e medidas de posição.



Fonte: autoria própria

A média observada é de 591,2, acima daquela vista na prova de linguagens (552,1), sugerindo que em média os estudantes obtiveram melhor resultado no teste de ciências humanas que no teste de linguagens. Esse melhor desempenho também pode ser reparado quando se examina a nota mínima (320,3) e máxima (785,4), ambas acima das registradas na prova de linguagens.

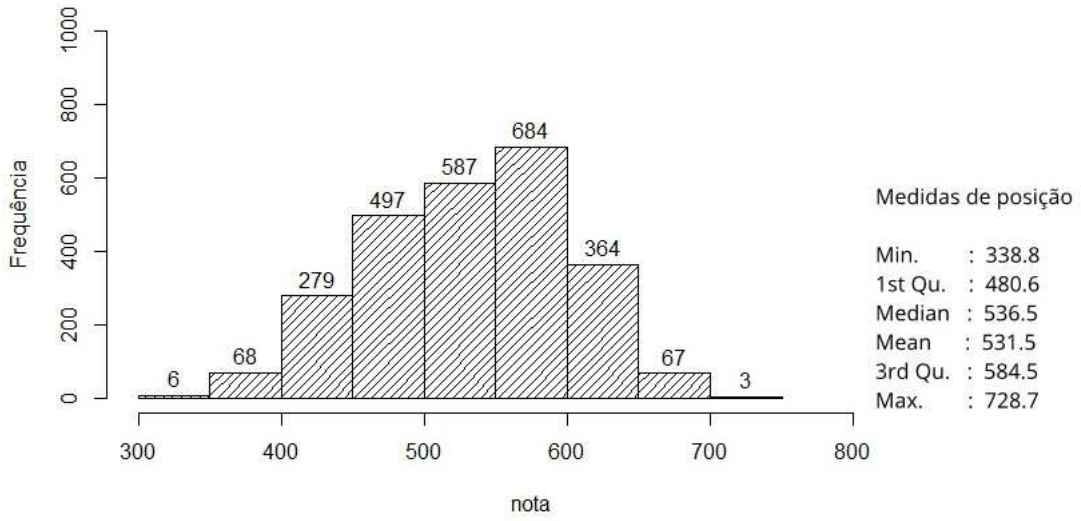
A leve assimetria também pode ser notada pela pequena diferença entre a média (591,2) e a mediana (598,8) e, ao mesmo tempo, se traduz em uma diferença interquartílica de 40,3 para *Mediana – 1st Qu.* e 32,60 para *3rd Qu. – Mediana*. Isto ocorre porque aqueles estudantes com menores notas são mais numerosos e estão mais distantes da massa de estudantes com notas próximas à média que os mais proficientes.

A nota em ciências humanas utilizada como variável independente para explicar parte da evasão pode ser verificada na regressão logística abaixo. Nela observa-se que o coeficiente estimado é negativo, indicando que quanto maior a nota obtida menor a probabilidade do estudante ter sua matrícula cancelada. No entanto, essa estimativa não atinge o nível de significância mínimo medido pelo p-valor (0,19), ou seja, embora o coeficiente sugira algumas interpretações, parecidas inclusive com a nota em linguagens, a estimativa não é confiável ao nível de 95%. Dessa forma, a nota da prova de ciências humanas não será utilizada na aplicação dos algoritmos de classificação.

```
glm(formula = status ~ NU_NOTA_CH, family = "binomial", data = p3033)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0392  -0.9463  -0.9288   1.4194   1.4915

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.0469786  0.4060673  -0.116    0.908
NU_NOTA_CH  -0.0008967  0.0006843  -1.310    0.190

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3338.0  on 2554  degrees of freedom
Residual deviance: 3336.3  on 2553  degrees of freedom
AIC: 3340.3

Number of Fisher Scoring iterations: 4

> exp(m$coefficients) # razão de chance
(Intercept)  NU_NOTA_CH
  0.9541078   0.9991037
```

V17 - Nota Ciências Naturais: a análise visual da distribuição das notas da prova de ciências naturais é semelhante à distribuição normal, com ligeira assimetria à esquerda, como as outras duas provas analisadas até o momento, e apresenta, diferente das anteriores, um formato mais achatado devido a uma quantidade considerável de estudantes que obtiveram score entre 400 e 600 pontos.

Figura 23: Distribuição de frequência da variável nota de ciências naturais e medidas de posição.



Fonte: autoria própria

A média das notas é de 531,5 pontos, muito próxima da mediana de 536,5. A diferença entre os quartis também é parecida, 55,9 (*Mediana – 1st Qu.*) e 48,0 (*3rd Qu. – Mediana*), corroborando com a análise visual de uma distribuição parecida com a normal padrão mas com leve assimetria negativa. A menor nota registrada foi de 338,8 enquanto a maior foi de 728,7.

Na regressão logística feita, o coeficiente estimado, à semelhança das duas provas anteriores, também ficou negativo sugerindo que a probabilidade de evadir decresce à medida que o score na prova fica mais alto. A estimativa alcançou nível de significância esperado ficando com p-valor abaixo de 0,05, ou seja, com pelo menos 95% de confiança é rejeitada a hipótese de que a nota na prova de ciências naturais não tem influência no status da matrícula dos alunos. Este resultado é corroborado com a melhora no ajuste do modelo com a inclusão da variável medido pelo *deviance*. Para o modelo nulo o *deviance* é 3.338,0 diminuindo para 3326,5 quando é incluída a variável.

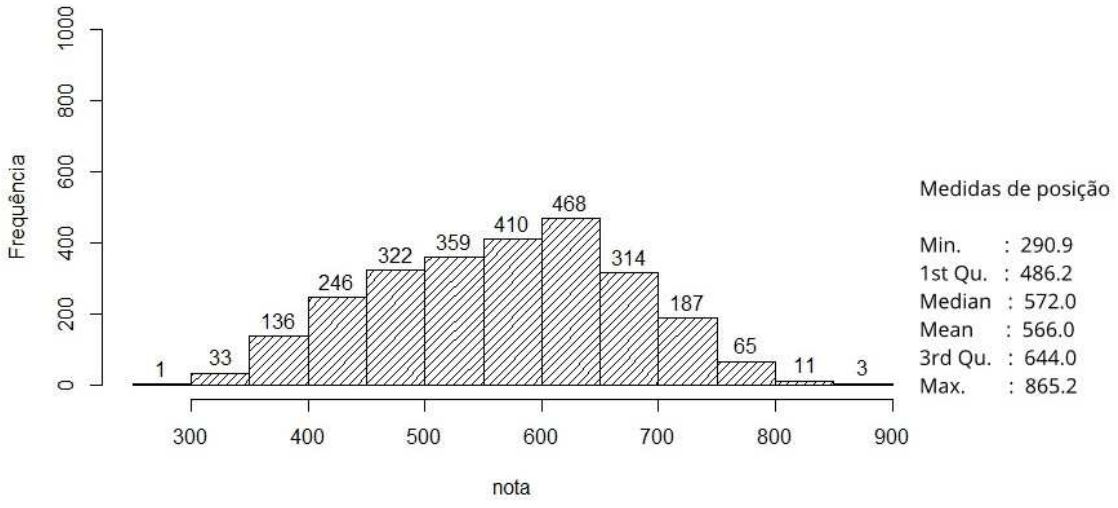
Assim como no teste de linguagens, embora haja significância estatística para a manutenção da variável, a melhora do ajuste do modelo é bastante reduzida. Isto pode ser constatado também pela razão de chance ser muito próxima a 1 (0,9979698) evidenciando mais uma vez que a influência dessa variável é diminuta.

```
glm(formula = status ~ NU_NOTA_CN, family = "binomial", data = p3033)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0932  -0.9532  -0.8958   1.3909   1.6117

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.5006954  0.3188876   1.570 0.116385
NU_NOTA_CN  -0.0020322  0.0005979  -3.399 0.000676 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3338.0  on 2554  degrees of freedom
Residual deviance: 3326.5  on 2553  degrees of freedom
AIC: 3330.5

Number of Fisher Scoring iterations: 4

> exp(m$coefficients) # razão de chance
(Intercept)  NU_NOTA_CN
  1.6498682   0.9979698
```

V18 - Nota Matemática: na prova de matemática os scores dos estudantes foram um pouco diferentes. Visualmente, percebe-se que a distribuição se assemelha à normal padrão e com um achatamento da distribuição. É possível visualizar também que, diferente das provas anteriores, há incidência de algumas notas acima de 750 pontos.

Figura 24: Distribuição de frequência da variável nota de matemática e medidas de posição.



:

Fonte: autoria própria

A média registrada é de 566,0 pontos, muito aproximada da mediana (572,0). A amplitude da distribuição é maior que a observada nas distribuições dos testes analisados anteriormente, sendo 290,9 o menor score registrado e 865,2 o maior.

Assim como nos testes anteriores, na prova de matemática também existe uma leve assimetria negativa, o que pode ser verificado, além da pequena desigualdade entre média e mediana, pela diferença interquartílica de 85,8 para *Mediana – 1st Qu.* e 72,0 para *3rd Qu. – Mediana*. Nota-se também que o achatamento da distribuição se reflete em maiores diferenças interquartílicas. Enquanto nas distribuições vistas anteriormente essa diferença ficava na casa dos 30, 40 pontos na prova de matemática é de aproximadamente 80 pontos.

A análise da regressão logística realizada, abaixo, apontou um coeficiente negativo, indicando que quanto maior a nota na prova de matemática menor a probabilidade do aluno ter sua matrícula cancelada. A partir da razão de chance calculada é possível afirmar que a cada ponto adicional na nota da prova a probabilidade de evadir é reduzida em 0,01%.

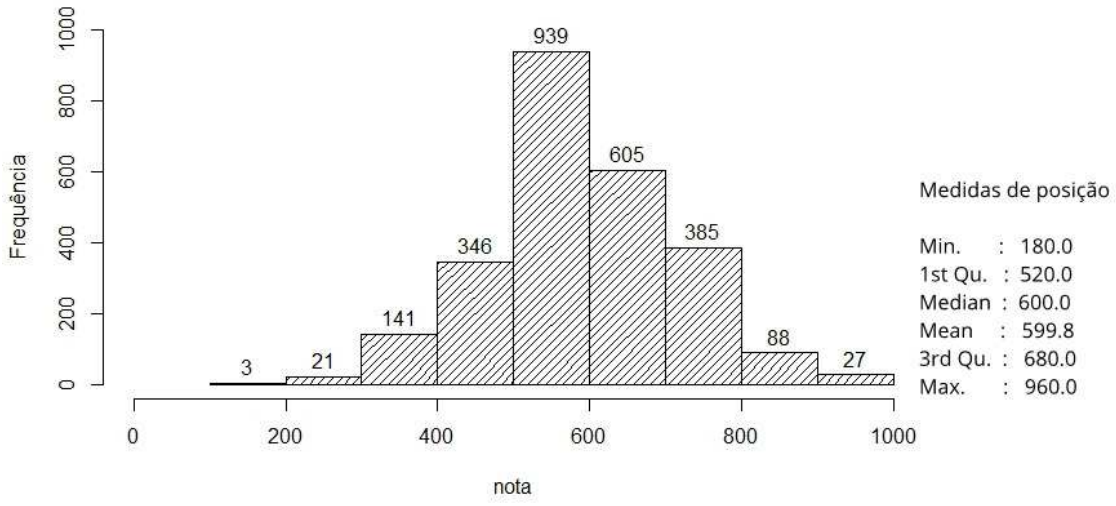
Estas estimativas atendem ao nível mínimo de confiança de 95% pois seu p-valor é menor de 0,05, portanto atestam com alto grau de confiança que a variável independente influencia a variável resposta. Uma ligeira melhora no ajuste do modelo também pode ser observada a partir da redução do *deviance* após a inclusão das notas de matemática.

```
glm(formula = status ~ NU_NOTA_M, family = "binomial", data = p3033)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0402  -0.9524  -0.9101   1.4006   1.5639

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.0183305  0.2243101  -0.082   0.9349
NU_NOTA_M   -0.0009893  0.0003917  -2.526   0.0115 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3338.0  on 2554  degrees of freedom
Residual deviance: 3331.7  on 2553  degrees of freedom
AIC: 3335.7

Number of Fisher Scoring iterations: 4

> exp(m$coefficients) # razão de chance
(Intercept)   NU_NOTA_M
  0.9818365   0.9990112
```

A relação entre o desempenho dos estudantes na prova de matemática e o status da matrícula é muito parecido com as provas anteriores, estatisticamente significante mas pouco relevantes. Razão de chance muito próxima a 1 (0,9990112) e diminuição do *residual deviance* bastante reduzida (6,3 unidades) fazem com que seja necessário avaliar até que ponto é válido incluir esse tipo de variável nos modelos de classificação.

V19 - Nota Redação: a análise visual da distribuição das notas de redação apresenta o formato da normal padrão, no entanto sem aquele achatamento verificado na distribuição das notas de matemática. Uma particularidade encontrada nas notas de redação é a existência de notas no intervalo entre 950 e 1000 pontos, 27 estudantes atingiram pontuação próxima da máxima do teste.

Figura 25: Distribuição de frequência da variável nota de redação e medidas de posição.



Fonte: autoria própria

A média das notas registradas é de 599,8 enquanto a mediana é de 600,0. A amplitude da distribuição é 780 pontos, a maior entre todas as provas, sendo a menor nota de 180 pontos e a maior 960. A diferença entre os quartis evidencia o formato da normal padrão observada, sendo 80 para *Mediana – 1st Qu.* e 80 para *3rd Qu. – Mediana*.

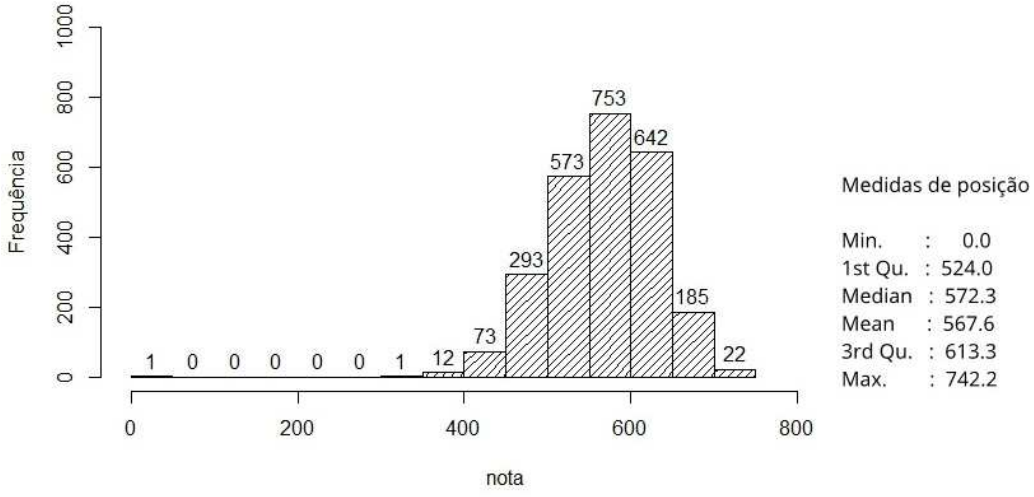
Quando se realiza a regressão logística para verificar se a nota de redação tem influência no status da matrícula observa-se que o coeficiente, à semelhança das demais provas, também é negativo, indicando que quanto melhor o desempenho do estudante menor a probabilidade de evadir. O fato do coeficiente logístico calculado ser significante evidencia que há efetivamente alguma influência desta variável independente na variável que se deseja explicar. Isto pode também ser verificado pela melhora no ajuste do modelo quando é incluída a variável, o *deviance* diminui de 3.338,0 para 3.326,4.

Tem-se aqui situação análoga àquelas observadas nos testes anteriores, com exceção da prova de ciências humanas que não demonstrou significância estatística. Variável estatisticamente significante e, ao mesmo tempo, pouca redução do *deviance*, isto é, pouco relevantes.

```
glm(formula = status ~ NU_NOTA_R, family = "binomial", data = p3033)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1370  -0.9607  -0.8998   1.3902   1.6182

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.1114860  0.2056834   0.542 0.587800
NU_NOTA_R   -0.0011522  0.0003387  -3.402 0.000669 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3338.0  on 2554  degrees of freedom
Residual deviance: 3326.4  on 2553  degrees of freedom
AIC: 3330.4

Number of Fisher Scoring iterations: 4

> exp(m$coefficients) # razão de chance
(Intercept)   NU_NOTA_R
  1.1179381   0.9988485
```

V20 - Nota Inscrito: a distribuição das notas finais do ENEM pode ser observada no histograma abaixo. Nele verifica-se que a distribuição é muito parecida com a normal padrão e a grande maioria das notas estão concentradas no intervalo entre 450 e 650 pontos.

Figura 26: Distribuição de frequência da variável nota do inscrito e medidas de posição.



Fonte: autoria própria

A média registrada de 567,6 é parecida com a mediana de 572,3. Assim como nas outras distribuições das provas do ENEM, há uma pequena assimetria negativa evidenciada pela pequena diferença entre média e mediana e pelas diferenças interquartílicas, 48,3 para *Mediana – 1st Qu.* e 41,0 para *3rd Qu. – Mediana.* Curiosamente há um dado discrepante, um estudante obteve média zero, provavelmente alguma forma de desclassificação haja vista que não houve outras notas zero nas provas.

Quanto à regressão logística realizada, observa-se um coeficiente negativo, não poderia ser diferente pois a nota final do candidato é uma média das provas anteriormente analisadas. Isto indica que quanto maior a nota, menor a probabilidade do estudante ter sua matrícula cancelada. A significância da estimativa do coeficiente atente ao nível de confiança de 95% (p-valor < 0,05), portanto pode-se de dizer que há alguma influência do desempenho final do estudante no status da matrícula. Esta influência também pode ser observada pela melhora no ajuste do modelo com a inclusão da variável independente, passando de um *deviance* de 3.338 para 3.326,4.

```
glm(formula = status ~ NU_NOTA_R, family = "binomial", data = p3033)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1370  -0.9607  -0.8998   1.3902   1.6182

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.1114860  0.2056834   0.542 0.587800
NU_NOTA_R   -0.0011522  0.0003387  -3.402 0.000669 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3338.0  on 2554  degrees of freedom
Residual deviance: 3326.4  on 2553  degrees of freedom
AIC: 3330.4

Number of Fisher Scoring iterations: 4

> exp(m$coefficients) # razão de chance
(Intercept)   NU_NOTA_R
  1.1179381   0.9988485
```

Novamente faz-se o questionamento sobre a validade da inclusão de uma variável estatisticamente significante, mas pouco relevante. Para encaminhar essa dúvida serão rodados os modelos de classificação com as notas do ENEM e sem as notas do ENEM. Assim será possível verificar se a inclusão dessas variáveis melhora ou não o desempenho dos algoritmos

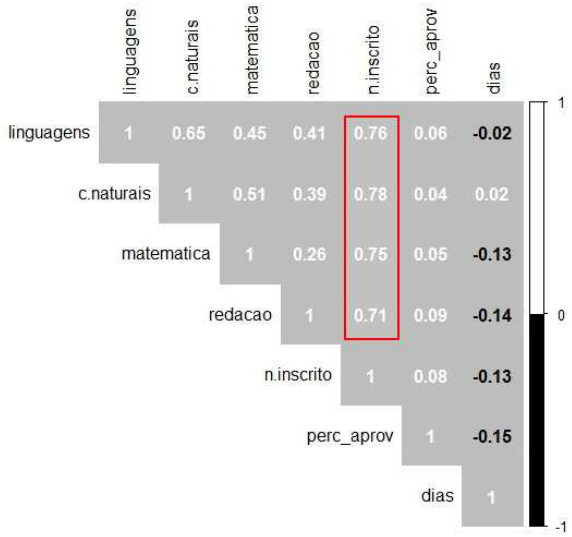
3.2.3.2 Teste de Multicolinearidade

Antes de partir para a divisão das bases de dados e aplicação dos modelos de classificação, é preciso ter cautela com uma situação. Segundo Hair et. al. (2009), um cuidado a ser tomado é a presença de multicolinearidade entre as variáveis independentes. Num cenário ideal, as variáveis independentes teriam alta correlação com a variável dependente, mas seriam independentes entre si. A multicolinearidade, portanto, é entendida como a alta correlação entre duas ou mais variáveis independentes.

Podem-se verificar na matriz abaixo os coeficientes de correlação de Pearson[17] calculados para as variáveis numéricas da base de dados. Na área indicada em vermelho, observa-se que a nota final do candidato no Enem tem forte correlação com as demais notas da prova. Evidentemente essa correlação faz sentido visto que a nota final é uma média das notas nas demais provas. Entre as outras provas há correlação moderada para fraca, com o coeficiente máximo 0,51 (matemática x ciências naturais) e o mínimo 0,26 (matemática x redação). Para as variáveis aprovação e dias, ambas com correlação muito próximas a zero em relação as outras variáveis, pode-se afirmar que são totalmente independentes.

[17] O coeficiente de correlação de Pearson mede o grau de correlação entre duas variáveis, além de indicar se a correlação é positiva ou negativa. Sua interpretação é dada da seguinte forma:
- Valores entre 0 e 0,3 em módulo significa correlação desprezível;
- Valores entre 0,3 e 0,5 em módulo significa correlação fraca;
- Valores entre 0,5 e 0,7 em módulo significa correlação moderada;
- Valores entre 0,7 e 0,9 em módulo significa correlação forte;
- Valores acima de 0,9 em módulo significa correlação muito forte.

Figura 27: Matriz de correção das variáveis quantitativas.



Fonte: autoria própria

A presença de alta correlação entre as variáveis independentes tem efeitos prejudiciais na estimação dos coeficientes de regressão, em casos extremos, pode até impedir a estimação de tais coeficientes. Seu efeito direto é inflar desnecessariamente os erros padrões dos coeficientes estimados reduzindo, assim, a confiança da estimativa (AKINWANDE; DIKKO; SAMSON, 2015).

Uma forma de diagnosticar a existência de multicolinearidade entre as variáveis independentes é o *Variance Inflation Factor* – VIF que indica quanto a variância do coeficiente estimado aumenta na presença de variáveis independentes correlacionadas. O valor mínimo do VIF é 1, indicando que não há multicolinearidade, VIF entre 5 e 10 indica que pode haver problema e VIF maior de 10 pode-se assumir que há multicolinearidade e, portanto, os coeficientes apresentam problemas de estimação (AKINWANDE; DIKKO; SAMSON, 2015).

Quadro 3: VIF para cada variável testada.

Variável	VIF
Linguagens	4.832796
Ciências Naturais	5.118716
Matemática	7.112069
Redação	8.421111
Nota Inscrito	42.733986
Percentual Aprovação	1.031057
Dias	1.081465

Fonte: autoria própria

Seguem, no quadro acima, os VIF das variáveis numéricas calculadas por meio do pacote Faraway no software estatístico R. Nela observa-se que as notas das provas do ENEM ficaram com VIF acima de 5, sendo que a nota final do inscrito obteve um valor de 42.73 indicando que há existência de multicolinearidade e, portanto, essa variável precisa ser retirada da análise.

Ao refazer o cálculo (quadro 4), agora sem a variável nota final do inscrito, percebe-se que os VIF ficaram consistentemente baixos, todos com valores abaixo de 2. Como os valores VIF ficaram em nível aceitável após a retirada da nota do inscrito nenhum teste adicional foi realizado e as variáveis restantes utilizadas nas próximas etapas do estudo.

Quadro 4: VIF para cada variável com a retirada da variável Nota Inscrito.

Variável	VIF
Linguagens	1.628596
Ciências Naturais	1.812354
Matemática	1.567886
Redação	1.251609
Percentual Aprovação	1.030607
Dias	1.066551

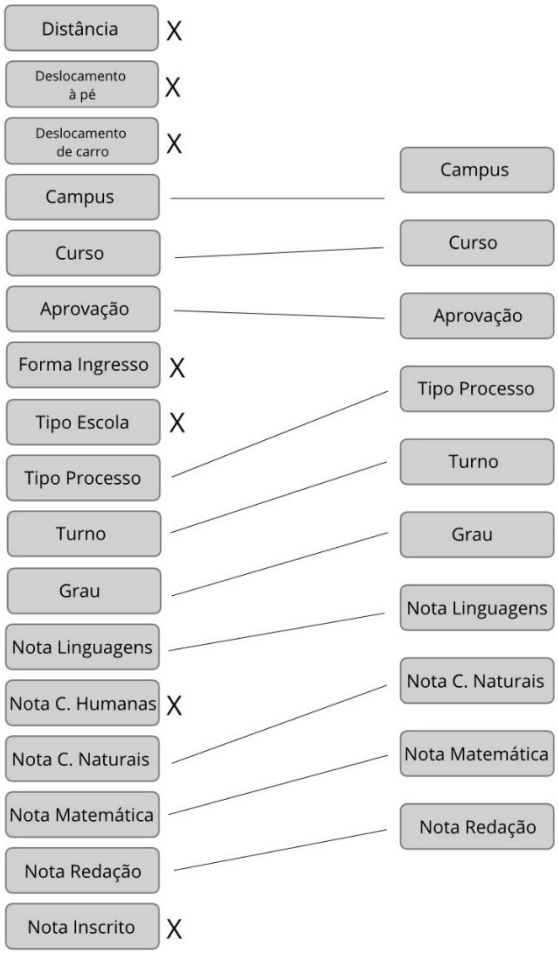
Fonte: autoria própria

3.2.3.3 Conjunto de Variáveis Final

Diante da análise individual de cada atributo e do teste de multicolinearidade, algumas variáveis não evidenciaram ter real influência sobre a variável resposta. As variáveis de distância e deslocamento (V3) não apresentaram nível de significância suficiente para serem utilizadas como variáveis independentes na modelagem dos algoritmos de classificação. O mesmo ocorreu com as variáveis Forma de ingresso (V10), e Nota Ciências Humanas (V16), portanto foram descartadas. Por outro lado, a variável Tipo de Escola (V13) foi excluída devido à presença de uma quantidade elevada de valores ausentes. Por motivo diferente, multicolinearidade, a variável nota final do inscrito no ENEM também foi eliminada.

Os procedimentos de seleção de variáveis adotados até o momento atendem ao objetivo específico deste trabalho de identificar as variáveis mais relevantes para o entendimento do problema. O conjunto final de variáveis a serem utilizadas na etapa de modelagem dos algoritmos de classificação pode ser observado na figura abaixo:

Figura 28: Apresentação visual da seleção das variáveis finais.



Fonte: autoria própria

Com relação à quantidade de registros da base de dados, após as delimitações de período temporal, exclusão dos registros faltantes e demais exclusões provenientes de erros ou de estudantes cujas características não eram objeto do estudo, passou dos 8.985 registros iniciais para 3.556. A tipologia das variáveis finais a serem utilizadas na etapa seguinte de modelagem pode ser observada no quadro abaixo:

Quadro 5: Tipologia das variáveis finais.

Variável	Tipo
Campus	Categórica
Curso	Categórica
Tipo processo	Categórica
Turno	Categórica
Grau	Categórica
Idade em dias	Numérica
Nota Linguagens	Numérica
Nota Ciências Naturais	Numérica
Nota Matemática	Numérica
Nota Redação	Numérica
Aprovações	Numérica

Fonte: autoria própria

3.2.3.4 Padronização

O penúltimo procedimento adotado antes da efetiva aplicação dos modelos de classificação é a padronização das variáveis numéricas. Conforme pode ser observado no quadro 5, algumas variáveis são numéricas e possuem amplitudes de valores diferentes. A variável dias, por exemplo, tem valor máximo de 23.048 e mínimo de 5.857, enquanto a variável Aprovação apresenta valor máximo de 1 e mínimo de 0. Ao se aplicar os modelos de classificação, os atributos de maior escala tendem a ter maior influência na análise devido a sua escala de valores ser maior. No entanto, isso não necessariamente significa que o atributo de maior escala seja mais importante (HAIR ET. AL., 2009).

Para que os modelos de classificação tenham melhor desempenho é preciso transformar as variáveis numéricas de modo que todas estejam na mesma ordem de grandeza. Assim garante-se que a aplicação dos algoritmos não fique enviesada em favor das variáveis de maior grandeza.

Não há um método definitivo de transformação dessas variáveis. Os métodos de normalização e padronização mais comuns são o Min-Max, o *Z-score* e a normalização por escala decimal. Segundo Mohamad e Usman (2013), o método de padronização *Z-score* obteve melhor desempenho em trabalhos de mineração de dados quando comparado com os outros métodos. Por este motivo, este método será utilizado neste trabalho para padronizar as variáveis numéricas.

Segundo Hair et al. (2009), “a forma mais comum de padronização é a conversão de cada variável em escores padrão (também conhecidos como escores Z) pela subtração da média e divisão pelo desvio-padrão”. (HAIR ET. AL., 2009, p.445). Isto permite comparar diretamente o efeito de cada variável independente apresentada sobre o status da matrícula, variável dependente.

A fórmula de padronização é definida por:

$$x_{ij} = Z(x_{ij}) = \frac{x_{ij} - \overline{x_j}}{\sigma_j} \quad (1)$$

3.2.3.5 Separação das Bases de Dados

Para responder o questionamento sobre se a inclusão das notas dos alunos no ENEM melhoraria o desempenho dos algoritmos de mineração de dados, optou-se por formar uma base separada e avaliar os modelos com e sem nota para o mesmo conjunto de estudantes. A partir desse teste separado é possível concluir se as notas do ENEM auxiliam na estimação dos modelos.

Além disso, como pôde ser observado na análise das aprovações dos alunos nas disciplinas do primeiro semestre, há uma diferença temporal entre as variáveis coletadas antes do início das aulas e a variável acadêmica (aprovações) utilizada neste trabalho. Este dado de desempenho acadêmico, embora relevante, é processado ao final do semestre, portanto pelo menos seis meses após o conjunto de dados iniciais.

Então, para que a instituição tenha informações relevantes com rapidez para basear suas ações logo no momento de ingresso, a aplicação dos modelos de classificação se deu de forma separada. Num primeiro momento, abre-se mão da precisão em nome da rapidez realizando a classificação com as variáveis disponíveis antes do ingresso na instituição, ou seja, sem a variável aprovações. Posteriormente, incluindo esta variável, tem-se uma previsão mais precisa, porém com atraso de um semestre na análise.

3.2.3.6 Variáveis Dummy

O conjunto de variáveis a ser considerado neste estudo é composto por variáveis numéricas e categóricas. No entanto, parte dos modelos de classificação que serão utilizados aceitam apenas variáveis numéricas. Uma forma de resolver esse problema é transformar a variável categórica em *n-1* variáveis numéricas chamadas "*dummies*". Assim as variáveis categóricas assumem valores 0 e 1, sendo 0 a ausência e 1 a presença do atributo.

Por exemplo, a variável Turno apresenta três categorias: integral, matutino e noturno. Ao transformar essa variável em *dummy* são criadas duas variáveis *dummy_integral* e *dummy_matutino*. O resultado ficaria:

- *dummy_integral* = 1, caso o estudante pertença ao turno integral;
- *dummy_integral* = 0, caso o estudante não pertença ao turno integral;
- *dummy_matutino* = 1, caso o estudante pertença ao turno matutino;
- *dummy_matutino* = 0, caso o estudante não pertença ao turno matutino.

Nota-se que a quantidade de variáveis *dummies* criadas é *n-1* categorias. A última variável, no caso exemplificado o turno noturno, será a exclusão das demais, ou seja, quando *dummy_integral* = 0 e *dummy_matutino* = 0. Este é um cuidado para evitar haja multicolinearidade devido a variáveis *dummy* redundantes, "*dummy trap*", o que inviabilizaria a interpretação dos modelos.

Dessa forma, todas as variáveis categóricas utilizadas foram transformadas em *dummies* para que possam ser utilizadas nos modelos de classificação.

3.2.3.7 Validação Cruzada

Um procedimento padrão quando se busca desenvolver um modelo de predição é dividir os dados em dois conjuntos distintos: um de treinamento para estimar os parâmetros do modelo e outro de teste para validar o modelo. Algumas técnicas são empregadas para evitar que se valide um modelo embasado em apenas uma partição dos dados.

Uma das técnicas mais utilizadas para avaliar o desempenho de um modelo de predição é a "k-fold cross validation", nela os dados são divididos em k partes mutuamente exclusivas e

de mesmo tamanho chamadas “folds”. O procedimento de treinamento e teste é realizado k vezes, em que a cada repetição uma parte (fold) é utilizada como conjunto de validação com os parâmetros estimados pelas demais partes (HAN; KAMBER; PEI, 2011). Ainda segundo Han, Kamber e Pei (2011), é recomendado que a base de dados seja dividida em 10 folds para estimar a acurácia de um modelo como forma de reduzir o viés e a variância dos resultados.

Neste trabalho, com auxílio do pacote caret do R, o conjunto de dados é inicialmente dividido em duas partes, 70% para treinamento e 30% para teste. E, para cada modelo de mineração de dados proposto, a técnica de validação cruzada k-fold é aplicada à base de treinamento dividindo-a aleatoriamente em 10 partes (10-fold). Este procedimento é repetido 3 vezes. A partir da média das 3 repetições do 10-fold cross validation são definidos os parâmetros dos modelos de mineração de dados que posteriormente serão aplicados ao conjunto de teste para verificar suas acurácias.

3.3 MODELAGEM

A etapa de modelagem da metodologia CRISP-DM diz respeito à construção dos modelos, neste caso, modelos de classificação, na tentativa de prever o comportamento da variável resposta em função das variáveis independentes selecionadas. Neste estudo, a variável a ser estimada é o status da matrícula dos alunos de ensino superior do IFC em função das variáveis selecionadas.

Os modelos a serem aplicados foram baseados em trabalhos semelhantes realizados com estudantes de cursos superiores de outras instituições de ensino (MANHÃES, 2015; MACHADO, 2015; AMARAL, 2016; AIRES, 2017). Nestes estudos, os algoritmos de classificação apresentaram desempenho preditivo satisfatório, sugerindo que possam exibir desempenho similar para alunos do IFC embora a aplicação para este estudo seja em um contexto diferente.

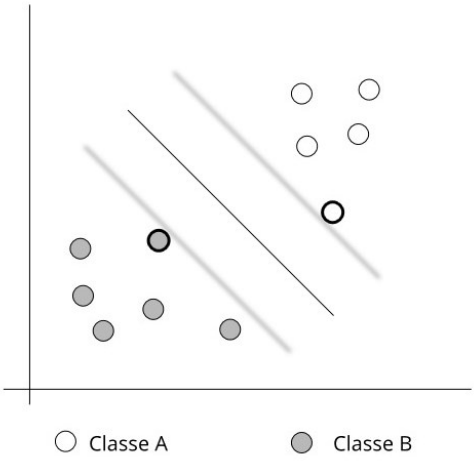
A seguir serão apresentados os modelos utilizados neste estudo. Não se pretende explicar profundamente os elementos estatísticos de tais modelos, apenas trazer as noções gerais que fundamentam o desenvolvimento destes algoritmos de classificação.

3.3.1 SVM – SUPPORT VECTOR MACHINES

O SVM busca dividir as classes por meio da aplicação de modelos lineares de classificação implementando limites de classes não lineares. Na prática o SVM faz uso dos modelos lineares, cuja limitação é representar limites lineares de separação das classes quando aplicado a problemas de classificação, para dividir as classes de forma não linear. O artifício, truque de kernel, é transformar os dados em um mapeamento não linear, ou seja, alterar o espaço utilizado em um espaço não linear. Logo, uma reta traçada nesse novo espaço representa uma linha não linear no espaço original (HAN; KAMBER; PEI, 2011).

O SVM busca dividir uma base de dados em classes por meio da aplicação de modelos lineares de classificação. Dentre todos os possíveis limites de separação, o modelo opta pelo limite mais distante dos pontos aproximados ao limite entre as classes. Uma vez definido este limite (figura 28), qualquer novo dado pode ser classificado, se ficar à esquerda do limite será cinza, se ficar à direita branco.

Figura 29: Representação visual do classificador SVM.



Fonte: autoria própria

A aplicação direta do SVM funciona para dados linearmente separáveis, no entanto problemas do mundo real não são linearmente separáveis. Com dados não lineares não se consegue traçar um limite para transformar os dados para outros espaços. Para lidar com estes casos pode-se adicionar uma dimensão por meio da transformação chamada de truque do kernel

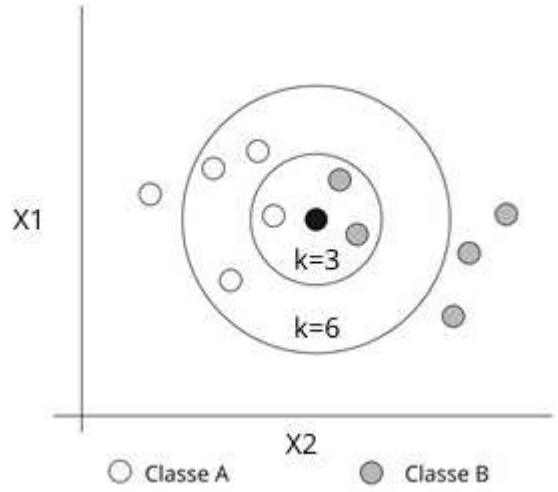
e com isso permitir a separação as classes. Na prática o SVM faz uso dos modelos lineares, cuja limitação é representar limites lineares de separação das classes quando aplicado a problemas de classificação, para dividir as classes de forma não linear (HAN; KAMBER; PEI, 2011).

A função de transformação para um espaço não linear, truque de kernel, pode ser realizada de algumas formas diferentes. Para este trabalho o modelo SVM desenvolvido considerou o kernel do tipo Radial.

3.3.2 KNN – K-NEAREST NEIGHBORS

O KNN busca classificar os dados por analogia a partir das características de um ou mais objetos vizinhos àquele considerado. Para isso são calculadas as distâncias entre o novo dado e aqueles já classificados e então definida a qual classe o novo dado pertence. Assim, pode-se estabelecer qual o número de k-vizinhos utilizado para classificar. Por exemplo, se k=3, então a classificação depende das características dos três vizinhos mais próximos, se k=6, então a classificação será de acordo com os seis vizinhos mais próximos, e assim por diante (HAN; KAMBER; PEI, 2011).

Figura 30: Representação visual do classificador KNN.



Fonte: autoria própria

A partir da figura acima percebe-se a importância da definição do número de k-vizinhos. Normalmente com o aumento de k o modelo fica mais preciso, pois baseia-se em um

número maior de vizinhos para fazer a classificação, até um certo ponto ótimo em que o aumento de k não melhora a precisão do modelo.

Neste trabalho, o número de vizinhos próximos (k) encontrado pelo modelo KNN foi 9. Então, a classificação de cada estudante como ativo ou cancelado considerou a as características dos 9 estudantes mais próximos.

3.3.3 NAIVE BAYES

Naive Bayes é um classificador baseado no teorema de Bayes. De acordo com a fórmula do teorema, abaixo, busca-se encontrar a probabilidade da ocorrência do evento A dado que B ocorreu. Assume-se então a independência entre as variáveis preditoras, ou seja, considera que cada característica contribui de forma independente para a probabilidade de acerto.

$$Teorema\ de\ Bayes\colon\ \ P(A|B)\ = \frac{(P(B|A)\ P(A))}{(P|B)} \tag{2}$$

Este modelo permite que as independências condicionais de classe sejam definidas entre subconjuntos de variáveis, fornecendo um modelo de relações causais, no qual a aprendizagem pode ser realizada (HAN; KAMBER; PEI, 2011).

Existem algumas maneiras de afinar o modelo de classificação Naive Bayes, como o método de suavização de probabilidades Laplace ou a utilização de kernel para os dados numéricos. No entanto este estudo utiliza o modelo disponibilizado no pacote caret do R que por padrão desconsidera qualquer tipo de suavização das probabilidades ou a utilização de função kernel.

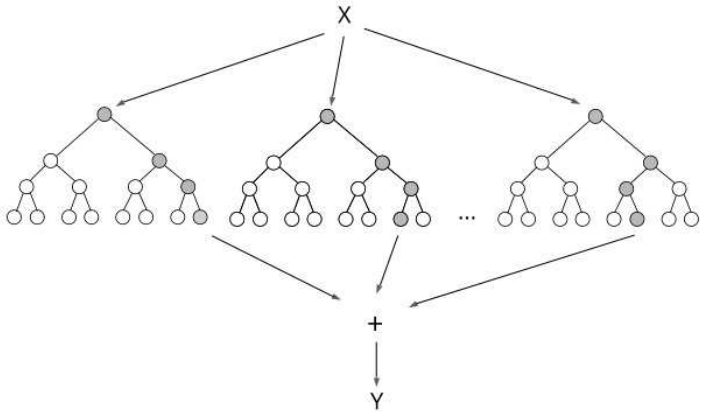
3.3.4 RANDOM FOREST

Random Forest é um modelo preditivo derivado das árvores de decisão. Em uma árvore de decisão cada nó interno representa uma variável com um ramo até um nó filho e uma folha sendo o valor previsto para a variável resposta considerando as outras variáveis, que na árvore é representado pelo caminho do nó raiz ao nó folha. Uma árvore de decisão descreve uma estrutura de uma árvore onde o nó folha representa a classificação e as ramificações que levam àquela classificação.

A profundidade da árvore de decisão pode ser definida uma vez que a precisão do modelo não é diretamente influenciada pelo tamanho da árvore. Modelos muito profundos podem inclusive aumentar muito o tempo de processamento para classificar sem necessariamente melhorar a acurácia do modelo. No pacote randomForest para o R utilizado, dada a limitação de profundidade, o próprio pacote indica qual o número de nós ótimo para aquela classificação proposta.

O Random Forest é uma evolução dessa técnica, em que são feitas múltiplas árvores de decisão em um ambiente de treinamento e a moda de classe encontrada é a saída do modelo. Neste estudo foram definidos como parâmetros do classificador o número de 500 árvores de decisão com duas variáveis utilizadas para escolher cada nó.

Figura 31: Representação visual do classificador Random Forest.



Fonte: autoria própria

3.3.5 REGRESSÃO LOGÍSTICA

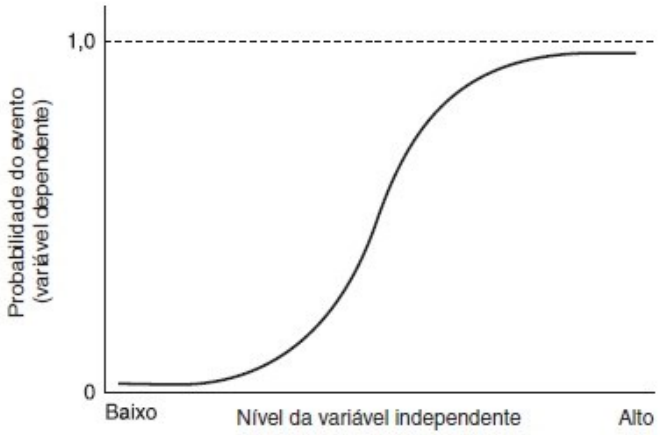
A análise de regressão é um modelo probabilístico que busca entender como uma variável (dependente) é afetada por outra variável (independente). Como se trada de variável depende binária (0 e 1), a regressão logística apresenta a probabilidade de classificação do evento observado em função das variáveis independentes.

O método de regressão logística utilizado nos estudos realizados foi o GLM - *Generalized Linear Model* disponível no pacote caret do R. Por padrão, esse método não utiliza

nenhum parâmetro de afinamento que possibilite induzir a melhora do desempenho do classificador.

Visualmente se pode observar a representação da curva logística na figura 31.

Figura 32: Representação visual do classificador Regressão Logística.



Fonte: autoria própria

Os modelos de regressão buscam verificar a relação entre uma ou mais variáveis explicativas e uma variável dependente (FAVERO ET AL, 2009). Basicamente o intuito da análise de regressão é entender como uma variável (dependente) é afetada por outra variável (independente).

Inicialmente as regressões eram formuladas para explicar variáveis numéricas, podendo ser regressões simples, quando uma variável independente busca explicar a variável resposta, ou regressões múltiplas, quando há mais de uma variável independente. No entanto, em muitas situações busca-se entender a probabilidade de um evento específico acontecer como, por exemplo, a ocorrência de uma doença na área da saúde, de inadimplência no mercado de crédito ou de sinistro no mercado de seguros. Para estes casos, em que a variável a ser explicada é binária, uma forma especial de regressão foi desenvolvida, chamada de regressão logística.

Uma dessas situações é a tratada neste estudo. Como o fenômeno de interesse trata-se de um evento binário, aluno com matrícula cancelada ou ativa, a regressão logística é o tipo de regressão indicado para analisar a relevância das variáveis em relação a variável resposta.

A regressão logística é semelhante à regressão múltipla em muitos aspectos, a diferença fica por conta da estimação dos coeficientes. Enquanto a regressão múltipla utiliza o método

dos mínimos quadrados, a regressão logística estima os coeficientes por meio do método da máxima verossimilhança (HAIR ET AL, 2009). Por isso os coeficientes da regressão logística têm um significado um pouco diferente daqueles estimados na regressão múltipla.

Como pôde ser observado na análise individual de cada variável, a interpretação dos coeficientes logísticos inicia pela direção da relação, se positiva ou negativa, e indica como variações na variável independente afetam a variável dependente. Por exemplo, um coeficiente positivo retrata que um aumento na variável independente é associado a uma maior probabilidade de o evento ocorrer (HAIR ET AL, 2009). A magnitude do impacto na probabilidade prevista é mais facilmente entendida em termos de razão de chance, ou seja, qual a chance do evento ocorrer em um grupo e a chance de ocorrer em outro grupo. Um exemplo deste trabalho, conforme será visto mais adiante, é a chance de um aluno de licenciatura ter sua matrícula cancelada em comparação com um aluno de bacharelado.

A estimativa dos coeficientes pode também ser avaliada quanto à sua significância estatística. Se o coeficiente observado apresenta valor de significância (p-valor) menor de 0,05 rejeita-se a hipótese de o coeficiente ser zero e, portanto, assume-se com 95% de confiança que a variável em questão tem influência na variável resposta (FÁVERO ET AL, 2009). A significância estatística é, no sentido de manter nos modelos de classificação apenas as variáveis que apresentam relação com a variável resposta, um critério de manutenção ou exclusão das variáveis.

Outra maneira de verificar se a variável independente analisada exerce influência sobre a variável dependente é comparar o ajuste preditivo dos modelos com e sem a variável independente. A medida de ajuste fornecida no software R é o desvio residual (*deviance*), em que quanto menor seu valor melhor é o ajuste do modelo. Dessa forma, o teste de significância e o *deviance* auxiliam na verificação da influência das variáveis independentes.

Então, além da Regressão Logística ser um relevante classificador, o resultado da regressão de cada variável independente em relação à variável resposta se configura em importante ferramenta para selecionar, dentre as muitas variáveis disponíveis, quais são as que mais auxiliam no desempenho dos modelos de predição.

3.4 AVALIAÇÃO

Para fins de avaliação do desempenho de cada modelo de classificação apresentado, serão expostos os resultados dos estudos realizados em cada conjunto de dados descrito na seção 3.2.3.5. O espaço amostral compreende os registros discentes dos ingressantes em cursos superiores do IFC entre os anos de 2014 e 2017 abarcando um total de 3.556 alunos após todo o processo de limpeza e preparação dos dados.

Como o objetivo é realizar uma avaliação criteriosa de cada modelo e, assim, verificar qual deles obteve o melhor desempenho é preciso descrever quais serão as métricas utilizadas para comparar os resultados. Comumente a métrica mais utilizada em trabalhos semelhantes é a acurácia de classificação (MANHÃES, 2015; AMARAL, 2016). Outras medidas complementares são a sensibilidade e especificidade de classificação pois ajudam a entender o desempenho do modelo à luz do problema trabalhado. A depender do autor pesquisado, essas métricas complementares são apresentadas com diferentes nomenclaturas como precisão, *recall*, revocação, entre outros. No entanto, a nomenclatura adotada nesse trabalho segue abaixo (HAN; KAMBER; PEI, 2011):

A escolha do critério mais adequado a ser utilizado depende da definição de qual classe é considerada positiva e negativa. Neste estudo a classe considerada positiva é a permanência no curso, enquanto a classe negativa é o registro da evasão, ou seja, a matrícula cancelada.

Acurácia de classificação: apresenta o percentual de classificações corretas que o modelo atingiu, ou seja, do total de casos apresentados quantos foram corretamente previstos pelo modelo. Como se trata de um percentual, a escala da medida varia de 0 a 1. No caso deste estudo, a acurácia considera tanto os acertos de matrículas ativas (verdadeiros positivos) quanto os acertos de matrículas canceladas (verdadeiros negativos) observadas no conjunto de teste.

$$Acurácia = \frac{Verdadeiros\ Positivos + Vernadeiros\ Negativos}{Total} \quad (3)$$

Sensibilidade de classificação: reflete a proporção dos casos positivos corretamente previstos pelo modelo. No caso deste trabalho a sensibilidade representa a quantidade de matrículas ativas

acertadamente identificadas pelo modelo sobre o total de matrículas ativas presentes no conjunto de dados. Assim como a acurácia, a escala da sensibilidade varia entre 0 e 1.

$$Sensibilidade = \frac{Verdadeiros\ Positivos}{Verdadeiros\ Positivos + Falsos\ Negativos} \quad (4)$$

Especificidade de classificação: representa a proporção dos casos negativos corretamente identificados pelo modelo. Para este trabalho a especificidade significa o percentual de alunos evadidos acertadamente previstos. Os valores que a especificidade pode adotar varia entre 0 e 1.

$$Especificidade = \frac{Verdadeiros\ Negativos}{Verdadeiros\ Negativos + Falsos\ Positivos} \quad (5)$$

Embora a acurácia de classificação seja o indicador mais utilizado para medir o desempenho dos algoritmos pois, de fato, mede o desempenho integral do classificador, a especificidade desempenha um papel fundamental para este estudo. Como o objetivo principal é identificar precocemente os alunos propensos a abandonar o curso, é a especificidade que apresenta a melhor previsão para os estudantes com matrícula cancelada.

O ideal seria lograr um modelo com alta acurácia, alta sensibilidade e alta especificidade. Entretanto, na prática os modelos, a depender de suas características, podem ser mais eficazes em um aspecto que outro. Um modelo com alta sensibilidade e baixa especificidade, por exemplo, tem como característica acertar muito as previsões daqueles estudantes que permanecem nos cursos. Isso significa que os estudantes com potencial de evadir passam despercebidos.

Em termos de institucionais é mais importante a identificação das possíveis evasões, isto é, que o modelo tenha, portanto, maior especificidade. Assim, o corpo gestor da instituição pode atuar de maneira mais efetiva no amparo àqueles estudantes. Possíveis erros de previsão dos estudantes que permanecem nos cursos acabam sendo menos relevantes que falhas na previsão daqueles que evadem. Por este motivo o principal critério a ser utilizado na avaliação dos modelos de classificação é a especificidade.

Os estudos realizados serão descritos na seção a seguir e apresentados os resultados de cada modelo de classificação nos diferentes cenários.

3.5 DESENVOLVIMENTO

A fase de desenvolvimento, conforme visto no início do capítulo, é a última etapa da metodologia CRISP-DM e consiste na efetiva aplicação da solução encontrada. A automatização da identificação daqueles estudantes com maior propensão a evadir é algo além do alcance deste estudo, pois as transmissões das informações produzidas precisam estar interligadas ao ambiente institucional acadêmico e, portanto, contar com apoio e suporte da diretoria responsável pela tecnologia da informação no IFC.

Além disso, qualquer solução que compreenda o dispêndio de esforços institucionais envolvendo diferentes áreas do IFC, mesmo no âmbito do Plano Estratégico de Permanência e Êxito, necessita da aprovação e efetivo envolvimento das pró-reitorias e diretorias envolvidas. Por isso, esta etapa da metodologia acaba constituindo-se numa proposta de desenvolvimento de um sistema de alertas encaminhados aos professores e demais profissionais envolvidos diretamente com o atendimento estudantil.

4 RESULTADOS

Neste capítulo estão reunidos os principais resultados encontrados ao longo do trabalho, desde o início do processo de EDM até a etapa de avaliação dos algoritmos utilizados. Conforme o fluxo de mineração de dados ilustrado na figura 4 da seção 2.4, o processo inicia-se a partir das bases de dados, neste caso a partir de duas fontes principais de dados: o SIGA e o SISU. São apresentados ainda os resultados dos dois estudos 4.1 e 4.2, realizados com a aplicação dos algoritmos de mineração de dados.

Já na etapa inicial do processo verificou-se que nem todos os registros da base SIGA tinham correspondência com as informações do SISU, ou seja, nem todos os ingressantes haviam realizado a prova do SISU para seu referente ano de ingresso. A significativa quantidade de estudantes sem nota do ENEM, 1.002 estudantes do total de 3.556 considerados após os procedimentos de limpeza e exclusão de dados faltantes, configurou-se em uma das razões porque o conjunto de estudantes com notas do ENEM foi objeto de um estudo específico (estudo 1).

Na fase de preparação dos dados, as variáveis foram analisadas e foram realizadas regressões logísticas com objetivo de selecionar as variáveis com real influência na situação da matrícula do aluno. A partir deste procedimento, devido à falta de significância estatística, foram descartadas as seguintes variáveis: distância (V04), deslocamento de carro (V04), deslocamento a pé (V04), forma de ingresso (V10) e nota ciências humanas (V16). As variáveis deslocamento de ônibus (V04) e tipo de escola (V13) foram excluídas pois continham dados faltantes acima do mínimo aceitável de 5%.

Além destas, foram excluídas três categorias da variável *campus* (V07) pois também não alcançaram o nível mínimo de significância para comporem o conjunto de variáveis finais. Os estudantes dos *campi* Brusque, São Francisco do Sul e Sombrio, um total de 565 alunos, foram excluídos da análise por este motivo.

Algo parecido ocorreu com a variável curso (V08), na qual parte deles revelou baixa significância estatística. Diante da possibilidade de exclusão de parte dos cursos e, por conseguinte, perda de significativa quantidade de dados optou-se pela manutenção da variável cursos na íntegra. Além do aspecto técnico da quantidade de registros, a decisão de manter a variável tem amparo na literatura sobre o tema, a qual revela que a escolha do curso do exerce influência na decisão evadir principalmente por aspectos vocacionais e de desempenho

acadêmico. (BARDAGI; HUTZ, 2008; RIBEIRO, 2005; NERY, 2009; AMBIEL, 2015, PIMENTEL; LIMA, 2017).

Outra situação que merece ser destacada é o resultado do teste de multicolinearidade realizado na seção 3.2.3.2. Neste teste a variável nota do inscrito (V20) apresentou alta correlação com as outras notas da prova no ENEM (figura 7) e precisou ser excluída de acordo com o critério VIF que indica a presença ou não de multicolinearidade entre as variáveis independentes.

Ainda na etapa de preparação dos dados, momento em que se verificou a influência de cada variável independente sobre a variável resposta, observou-se que a variável de maior correlação com o status da matrícula é a variável percentual de aprovações, dado derivado do índice de rendimento acadêmico (V09). A melhora no ajuste do modelo com a inclusão deste dado corrobora a hipótese de que a dificuldade em acompanhar o ritmo das aulas influi na decisão de continuar os estudos.

Com relação aos estudos, foram realizados dois testes com objetivos distintos. O estudo 1, abaixo, foi feito com o propósito de verificar se a previsão dos modelos de mineração melhora com a inclusão das variáveis de desempenho nas provas do ENEM. Individualmente as notas apresentaram significância estatística aceitável nas regressões logísticas, no entanto a melhora no ajuste do modelo foi muito pequena quando comparado com o ajuste do modelo nulo. Esse comportamento sugeriu que talvez as notas, embora significantes, não contribuem para a construção de modelos mais precisos, tornando-os mais complexos sem necessidade.

4.1 ESTUDO 1

O primeiro estudo consiste na verificação da melhoria ou não do desempenho dos classificadores com a presença das notas do ENEM. Como visto anteriormente, individualmente cada nota do ENEM é estatisticamente significante, mas tem pouca influência no status da matrícula do aluno.

Então para verificar se as notas selecionadas, em conjunto, têm potencial de melhorar o desempenho dos algoritmos de mineração de dados, foram separados os dados dos estudantes que haviam feito a prova do ENEM durante o período analisado, totalizando 2.554 estudantes. Com base nesse conjunto de dados foram aplicados os algoritmos de mineração em dois momentos: primeiramente sem as notas e em seguida realizada a mesma análise com as notas.

Na tabela abaixo se pode verificar o resultado de cada classificador sem a utilização das notas do ENEM para a o conjunto de dados dos estudantes que realizaram a prova. A coluna "Acerto sem Informação" significa a proporção de ativos na base de teste dos algoritmos, ou seja, equivale ao acerto de um modelo nulo. A acurácia é o acerto geral do modelo enquanto a especificidade é o acerto dos casos de evasão.

Tabela 2: Resultados do estudo 1.1, sem as notas do ENEM.

Modelo	**Acerto sem informação**	**Acurácia**	**Especificidade**
Regressão Logística	0.6405	0.6837	0.3345
Support Vector Machines	0.6405	0.6588	0.2945
Random Forest	0.6405	0.6444	0.0254
k-Nearest Neighbors	0.6405	0.6418	0.3382
Naive Bayes	0.6405	0.6405	0.0000

Fonte: autoria própria

Observa-se no estudo 1.1 que de modo geral a acurácia dos modelos ficou ligeiramente acima da acurácia da classe majoritária, sugerindo que as variáveis independentes deste conjunto de dados não continham informações suficientes para atingir índices de acerto consistentes na previsão da evasão.

Pode-se destacar o modelo de Regressão Logística, que obteve a maior acurácia de classificação, com 68,37% de acerto, ao mesmo tempo em que atingiu a segunda melhor taxa de acerto entre os casos de evasão, 33,45%. No quesito especificidade, o modelo com melhor precisão foi o k-Nearest Neighbors com 33,82% de acerto na identificação da evasão. Por outro lado, o modelo Naive Bayes não obteve um desempenho capaz de superar o modelo nulo.

Já no estudo 1.2, cujo objetivo é verificar se as notas do ENEM auxiliam na previsão dos modelos de classificação, os resultados podem ser observados na tabela abaixo:

Tabela 3: Resultados do estudo 1.2, com as notas do ENEM.

Modelo	Acerto sem informação	Acurácia	Especificidade
Regressão Logística	0.6405	0.6863	0.3273
Support Vector Machines	0.6405	0.6549	0.2836
Random Forest	0.6405	0.6484	0.0400
k-Nearest Neighbors	0.6405	0.6405	0.0000
Naive Bayes	0.6405	0.6196	0.2873

Fonte: autoria própria

Repara-se que com a inclusão das notas não houve melhoria significativa nos resultados. O modelo de Regressão Logística novamente apresentou o melhor desempenho, com a acurácia passando de 68,37% para 68,63%. Uma melhora marginal que revela a pouca relevância da proficiência dos estudantes nas provas do ENEM para a determinação do status da matrícula no ensino superior.

Também não houve alteração relevante na especificidade dos modelos. Pode-se dizer inclusive que houve piora na especificidade dos modelos, exceto no Random Forest. É interessante notar também que o modelo k-Nearest Neighbors, que no estudo 1.1, atingiu desempenho regular, piorou o desempenho com a inclusão das notas. Cabe ressaltar que nenhum algoritmo de mineração superou o acerto sem informação (35,95%), os que mais se aproximaram deste valor foram o k-Nearest Neighbors com 33,82% de acerto nas evasões sem a inclusão das notas do ENEM e a Regressão Logística com 32,73% de especificidade quando inclusas as notas.

Diante disso, pode-se afirmar que as notas dos estudantes nas provas do ENEM não auxiliam a previsão da evasão no contexto analisado. Como resultado dessa constatação, essas variáveis não serão utilizadas nos estudos seguintes.

4.2 ESTUDO 2

O segundo estudo realizado se refere à aplicação dos modelos de classificação em todo o conjunto de registros selecionado para este trabalho, total de 3.556 estudantes. O objetivo é avaliar qual dos modelos de classificação apresenta melhor desempenho à luz dos critérios de acurácia e especificidade.

A aplicação se dá em dois momentos distintos: no estudo 2.1 os modelos são testados para os dados obtidos antes do estudante iniciar efetivamente as aulas, espera-se resultados menos precisos, mas obtidos com maior rapidez. No segundo teste, estudo 2.2, é incluída no conjunto de variáveis a informação da taxa de aprovação dos estudantes ao final do primeiro semestre letivo, neste caso espera-se melhoria na acurácia dos classificadores não obstante o acesso a essa informação venha um pouco mais tarde.

Tabela 4: Resultados do estudo 2.1, antes do início das aulas.

Modelo	**Acerto sem informação**	**Acurácia**	**Especificidade**
Support Vector Machines	0.6013	0.6604	0.3600
Random Forest	0.6013	0.6585	0.2188
Regressão Logística	0.6013	0.6520	0.4141
k-Nearest Neighbors	0.6013	0.6473	0.4400
Naive Bayes	0.6013	0.6088	0.6635

Fonte: autoria própria

Como é possível observar nos resultados do estudo 2.1, de modo geral os classificadores atingem acurácia aproximadamente 5% superior à acurácia da classe majoritária. Com exceção do Naive Bayes, todos os outros apresentam em torno de 65% de acerto do status da matrícula dos alunos. O modelo de melhor desempenho neste teste é o Support Vector Machines com 66,04% de acurácia.

Quando se analisa a especificidade dos modelos, verifica-se bastante diferença nos resultados, com o melhor modelo (Naive Bayes) atingindo 66,35% de acerto das matrículas canceladas e o pior (Random Forest) prevendo corretamente apenas 21,88% das evasões. Curiosamente o classificador Naive Bayes, que obteve a menor acurácia, é aquele que apresenta o melhor desempenho na previsão do principal evento deste trabalho, a evasão.

Os resultados apresentados no estudo 2.1, conforme abordado anteriormente, consideram apenas as informações de registros cadastrais dos alunos preenchidos no momento da matrícula. Para o próximo teste, estudo 2.2, agregar-se-á a este conjunto de variáveis um dado de proficiência acadêmica do estudante, taxa de aprovações no primeiro semestre. O resultado pode ser observado na tabela abaixo:

Tabela 5: Resultados do estudo 2.2 após o início das aulas.

Modelo	Acerto sem informação	Acurácia	Especificidade
Regressão Logística	0.6013	0.7552	0.5624
Random Forest	0.6013	0.7373	0.5741
Support Vector Machines	0.6013	0.7280	0.5224
k-Nearest Neighbors	0.6013	0.7036	0.4894
Naive Bayes	0.6013	0.6370	0.6753

Fonte: autoria própria

Percebe-se, na comparação com o estudo 2.1, que há uma melhora considerável na acurácia dos modelos. O classificador de melhor desempenho foi a Regressão Logística, com 75,52% de acerto no status das matrículas, seguido do Random Forest com 73,73% de acerto. A exceção do Naive Bayes, os demais modelos atingiram desempenho superior a 70% de acurácia na classificação.

Ao avaliar a especificidade das classificações, percebem-se resultados mais homogêneos que os encontrados no estudo 2.1. Novamente, o modelo Naive Bayes apresentou a maior especificidade, 67,53% de acertos para os casos de matrícula cancelada, resultado consideravelmente superior ao obtido pelo segundo melhor classificador, Random Forest com 57,41% de acerto.

Os resultados apresentados estão alinhados à expectativa de melhora dos critérios de desempenho dos modelos, acurácia e especificidade, com a inclusão da variável de desempenho acadêmico. Ao mesmo tempo, percebe-se que a especificidade aumentou muito pouco com a inclusão da variável aprovações.

De modo geral os classificadores apresentaram valores de acurácia e especificidade similares. A exceção ficou por conta do Naive Bayes que apresentou baixa acurácia em todos os testes realizados, mas se destacou positivamente na identificação das evasões, resultado que corrobora a constatação de Manhães (2015). Mesmo assim, 67% de taxa de acerto para as matrículas canceladas, embora seja importante, mostra que o fenômeno da evasão passa por fatores que não são captados, ou apenas em parte, pelos registros cadastrais e acadêmicos disponíveis na instituição trabalhada.

O comportamento singular do algoritmo Naive Bayes nos testes realizados é interessante. No estudo 1 sequer obteve resultado analisável, em todo estudo 2 logrou as piores

acurácias, no entanto quando se considera o critério de avaliação mais importante para o problema deste estudo, a especificidade, o algoritmo ostenta o melhor desempenho no estudo mais relevante.

5 CONCLUSÃO

Neste trabalho foram abordados conceitos relacionados à evasão discente, um problema complexo que tem consequências indesejadas para a sociedade e principalmente para os estudantes que abandonam os estudos. Foi apresentado o panorama nacional da evasão no ensino superior, com foco no ensino presencial da Rede Federal de Educação Profissional e Tecnológica.

Em seguida tratou-se dos conceitos relacionados à Mineração de Dados, e como ela se insere no âmbito educacional a ponto de criar uma nomenclatura própria *Educational Data Mining - EDM*. Quais os principais métodos de EDM ligados à previsão da evasão e como eles podem ser aplicados no contexto de uma instituição de ensino superior multicampi e com as particularidades do IFC.

A metodologia utilizada buscou atender a proposta de gerar conhecimentos relevantes a partir da base de dados da instituição de ensino foco deste. Com base nos dados disponíveis foi possível entender cada variável trabalhada e avaliar se havia efetiva relação com o fenômeno da evasão.

À luz do problema proposto relativamente à proposta de metodologia para identificar precocemente os alunos mais propensos a evadir, os resultados mostram que a EDM pode se configurar como mais uma ferramenta de auxílio às instituições de ensino superior. Um bom ponto de partida para o aprimoramento de políticas de permanência e êxito no âmbito das instituições federais de educação, ciência e tecnologia.

5.1 LIMITAÇÕES DO ESTUDO

Uma das maiores limitações enfrentadas foi a falta de informações sobre os estudantes armazenadas nos bancos de dados. A própria característica descentralizada da instituição trabalhada afetou a confiabilidade dos dados mais antigos limitando a abrangência temporal da análise. Além disso, informações com potencial de auxiliar na explicação do problema como as cotas não puderam ser utilizadas devido à forma como o processo seletivo se desenvolve no IFC.

Mesmo assim, a acurácia e especificidade de classificação obtidos nesse contexto de poucos dados relevantes são suficientes para proporcionar um panorama preliminar da evasão.

Esta identificação preliminar permite que outras abordagens sejam postas em prática com maior eficiência e assertividade.

5.2 TRABALHOS FUTUROS

Diante dos obstáculos encontrados devido à carência de informações relevantes nos bancos de dados da instituição trabalhada para entender o contexto da evasão, resta como sugestão para trabalhos futuros investigar mais profundamente quais variáveis adicionais poderiam ser coletadas para identificar melhor o problema. E, a partir disso, empregar os procedimentos necessários para registrar adequadamente tais informações nas bases de dados da instituição.

Uma interessante análise a ser realizada é considerar o número de semestres cursados pelos estudantes e o campus como variáveis controle, seja por meio de regressão multinível ou integrando-os como efeito aleatório. Dessa forma pode ser esclarecido que a probabilidade de evasão pode variar ao longo do curso e também pode não ser a mesma para diferentes *campi*.

Pode-se ainda destacar a necessidade, uma vez adotada a metodologia de identificação precoce dos alunos mais propensos a evadir, do desenvolvimento de um mecanismo de transmissão dessas informações para os professores e demais profissionais envolvidos diretamente com os estudantes. Uma forma interessante de lidar com a disseminação dessas informações é integrar ao sistema de gestão acadêmica da instituição alguma forma de alerta que permita ao professor tomar previamente ciência da possibilidade de determinados alunos evadirem.

Além disso, a aplicação desse tipo de abordagem em outros contextos educacionais pode trazer novas informações e consolidar a utilização da mineração de dados como ferramenta de auxílio à gestão acadêmica. Uma dessas aplicações e que pode ser desenvolvida na mesma instituição de ensino trabalhada é expandir a análise para outros níveis de ensino.

REFERÊNCIAS

AKIWANDE, Michael Olusegun; DIKKO, Hussaini Garba; SAMSON, Agboola. Variance Inflation Factor: As a condition for the Inclusion of Suppressor Variable(s) in Regression Analysis. **Open Journal of Statistics**, v.5, p. 754-767. 2015. Disponível em: <https://www.scirp.org/html/11-1240578_62189.htm>. Acesso em: 20 jan. 2020.

ALMEIDA, Leandro S.; SOARES, Ana Paula C.; FERREIRA, Joaquim Armando. Questionário de Vivências Acadêmicas (QVA-r**)**: avaliação do ajustamento dos estudantes universitários. **Aval. psicol**., Porto Alegre, v. 1, n. 2, p. 81-93, nov. 2002. Disponível em: <http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1677-04712002000200002&lng=pt&nrm=iso>. Acesso em: 13 set. 2019.

AMARAL, Marcelo Gomes do. **Mineração de Dados Aplicada à Classificação do Risco de Evasão de Discentes Ingressantes em Instituições Federais de Ensino Superior**. Universidade Federal de Pernambuco. Dissertação de mestrado. 132 p. Disponível em: <https://repositorio.ufpe.br/handle/123456789/19502>. Acesso em: 07 jan. 2018.

AMBIEL, Rodolfo Augusto Matteo; SANTOS, Acácia Aparecida Angeli; DALBOSCO, Simone Nenê Portela. Motivos para evasão, vivências acadêmicas e adaptabilidade de carreira em universitários. **Psico,** Porto Alegre, v. 47, n. 4, p. 288-297, 2016. Disponível em: <http://revistaseletronicas.pucrs.br/ojs/index.php/revistapsico/article/view/23872/0>. Acesso em: 07 jan. 2018.

AZEVEDO, Ana Isabel Rojão Lourenço; SANTOS, Manuel Filipe. **KDD, SEMMA and CRISP-DM**: a parallel overview. Florianópolis: Instituto Politécnico do Porto. Instituto Superior de Contabilidade e Administração do Porto, 2008. Disponível em: <http://hdl.handle.net/10400.22/136>. Acesso em: 10 mai. 2019.

BARDAGI, Marucia Patta; FERRERO, Carlos Andres. Predição de Risco de Evasão de Alunos Usando Métodos de Aprendizado de Máquina em Cursos Técnicos. **Rev. bras. orientac. prof**, São Paulo , v. 9, n. 2, p. 31-44, dez. 2008. Disponível em: <http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1679-33902008000200005&lng=pt&nrm=iso>. Acesso em: 22 mar. 2018.

BITENCOURT, Priscilla Busin; HUTZ, Cláudio Simon. Apoio parental percebido no contexto da escolha inicial e da evasão de curso universitário. VIII Congresso Brasileiro de Informática na Educação (CBIE 2019). **Anais dos Workshops do VIII Congresso Brasileiro de Informática na Educação (WCBIE 2019)**, Brasília, p. 149-158, nov. 2019. Disponível em: < https://www.br-ie.org/pub/index.php/wcbie/article/view/8956>. Acesso em: 05 mai. 2020.

BRASIL, **Constituição**. Constituição da República Federativa do Brasil. 1988. Disponível em: <http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm>. Acesso em: 21 ago. 2019.

_____. **Lei n° 9394**, 20 de dezembro de 1996. *Lei de Diretrizes e Bases da Educação Nacional*. Brasília, 20 de dezembro de 1996. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/l9394.htm>. Acesso em 15 jan. 2018.

_____. **Lei n° 10.861**, 14 de abril de 2004. Institui o Sistema Nacional de Avaliação Superior – SINAES. 14 de abril de 2004. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2004/lei/l10.861.htm>. Acesso em 16 jan. 2019.

_____. **Decreto n° 6.096**, 24 de abril de 2007. Institui o Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais Reuni. Brasília, 24 de abril de 2007. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2007-2010/2007/Decreto/D6096.htm>. Acesso em 15 jan. 2018.

_____. **Lei n° 11.892**, 29 de dezembro de 2008. Institui a Rede Federal de Educação Profissional, Científica e Tecnológica e Cria os Institutos Federais de Educação. Brasília, 29 de dezembro de 2008. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2007-2010/2008/Lei/L11892.htm>. Acesso em 15 jan. 2018.

_____. **Emenda Constitucional n° 59**, 11 de novembro de 2009. Brasília, 11 de novembro de 2009. Disponível em:<http://www.planalto.gov.br/ccivil_03/constituicao/Emendas/Emc/emc59.htm>. Acesso em 13 out. 2019.

_____. **Lei n° 12.711**, 29 de agosto de 2012. Brasília. Brasília, 29 de nagosto de 2012. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/l12711.htm. Acesso em 13 out. 2019.

_____. **Lei n° 13.005**, 25 de junho de 2014. Aprova o Plano Nacional de Educação – PNE. Brasília, 25 de junho de 2014. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2014/Lei/L13005.htm>. Acesso em 15 jan. 2018.

_____. **SETEC/MEC**. Nota Informativa n. 138/2015/DPE/DDR/SETEC/ MEC, de 09 de julho de 2015. Informa e orienta as Instituições da Rede Federal sobre a construção dos Planos Estratégicos Institucionais para a Permanência e Êxito dos Estudantes, 2015. 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 25 ago. 2019

_____. **SETEC/MEC**. Plataforma Nilo Peçanha – PNP 2019 (Ano Base 2018), 2019. Disponível em: < http://plataformanilopecanha.mec.gov.br/2019.html>. Acesso em 28 jan. 2020.

_____. **Tribunal de Contas da União**. Acórdão nº 506/2013 – TCU – Plenário, de 13 de março de 2013. Disponível em:

<https://contas.tcu.gov.br/etcu/ObterDocumentoSisdoc?seAbrirDocNoBrowser=true&codArqCatalogado=8995767>. Acesso em 25 ago. 2019.

_____. **Ministério da Educação**. Site. 2017 Disponível em: <http://portal.mec.gov.br/ultimas-noticias/211-218175739/50411-evasao-no-ensino-medio-supera-12-revela-pesquisa-inedita>. Acesso em 13 out. 2019.

CHAPMAN, Pete et al. **CRISP-DM**: Step-by-step data mining guide. The Modeling Agency, [S.l.], 2000. Disponível em: <https://the-modeling-agency.com/crisp-dm.pdf>. Acesso em: 08 jun. 2018.

CUNHA, Luiz Antônio. Desenvolvimento desigual e combinado no ensino superior: Estado e mercado. **Educ. Soc**. [online]. v .25, n.88, p. 795-817, 2004.Disponível em: <https://doi.org/10.1590/S0101-73302004000300008. Acesso em: 20 out. 2018.

DURHAM, Eunice Ribeiro. A qualidade do ensino superior. **Revista @mbienteeducação.** v. 2, n. 1, p. 9-14, mês, 2009. Disponível em: <http://publicacoes.unicid.edu.br/index.php/ambienteeducacao/article/view/475>. Acesso em: 10 ago. 2019.

FUNCHAL, João P. da Silva; RODRIGUES, Alex S. P.; BORGES, Eduardo Nunes. Um modelo preditivo para estudo da evasão na graduação utilizando mineração de dados. **RETEC - Revista de Tecnologias**, v. 9, n. 3, p. 75-79, fev. 2017. Disponível em: <https://www.fatecourinhos.edu.br/retec/index.php/retec/article/view/262>. Acesso em: 10 ago. 2019.

GUIMARAES, Sueli Édi Rufini; BORUCHOVITCH, Evely. O estilo motivacional do professor e a motivação intrínseca dos estudantes: uma perspectiva da Teoria da Autodeterminação. **Psicol. Reflex. Crit**., Porto Alegre , v. 17, n. 2, p. 143-150, 2004 . Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-79722004000200002&lng=en&nrm=iso>. Acesso em: 10 ago. 2019.

HAIR, Joseph F. et al. **Análise Multivariada de Dados**. Porto Alegre: Bookman, 2009.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining**: Concepts and Techniques. Waltham: Morgan Kaufmann, 3th edition, 2011.

IFC. Instituto Federal Catarinense. **Planejamento Estratégico IFC 2018-2021**. Blumenau, 2018. Disponível em: < http://ifc.edu.br/2018/03/23/planejamento-estrategico-ifc-2/ >. Acesso em 10 de ago. 2019.

_____. **Censo Interno** – Maio 2019. Blumenau, 2019. Disponível em: < http://ifc.edu.br/2014/08/11/censo-interno/ >. Acesso em 10 de ago. 2019.

_____. **Relatório Técnico**: Evasão Discente nos Cursos Superiores do IFC. Blumenau, 2018.

_____. **Plano Estratégico Institucional para a Permanência e Êxito dos Estudantes do Instituto Federal Catarinense 2019-2021**. Blumenau, 2019. Disponível em: < http://estudante.ifc.edu.br/2019/05/02/plano-estrategico-institucional-para-a-permanencia-e-o-exito-dos-estudantes-do-ifc/ >. Acesso em 11 de jun. 2019.

_____. **Organização Acadêmica dos Cursos Superiores de Graduação**. Blumenau, 2012. Disponível em: < http://ifc.edu.br/wp-content/uploads/2014/05/RESOLUCAO-057-2012-org-didatica-SUP.pdf>. Acesso em 11 de jun. 2019.

KANTORSKI, Gustavo Zanini et al. Uma visão do futuro: previsão de evasão em cursos de graduação presenciais de universidades públicas: o caso do curso de zootecnia. **XV Colóquio internacional de gestão universitária – CIGU.** Desafios da Gestão Universitária no Século XXI. Mar del Plata – Argentina, 2, 3 e 4 de dezembro de 2015.

KDNUGGETS. What main methodology are you using for your analytics, data mining, or data science projects? **KDNUGGETS**, [S.l.], 2014. Disponível em: <http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html>. Acesso em: 02 nov. 2019.

LIMA, Keycinara Batista; PIMENTEL, Elisabeth Tavares. T. Vocação Profissional e Impactos na Evasão Universitária. **Revista de Estudos e Investigación Psicologia y Educación,** Coruña, 3, 2017. Disponível em: <https://www.researchgate.net/publication/321843496_Vocacao_Profissional_e_Impactos_na_Evasao_Universitaria >. Acesso em: 10 nov. 2018.

MICHELOTTO, Marcele Arruda. Impactos de Incentivos Financeiros sobre o Sucesso Acadêmico Empregando Modelos de Regressão Multinível. 2019. 144 p. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Métodos e Gestão em Avaliação, Florianópolis, 2019.

MOHAMAD, Ismail; USMAN Dauda. Standardization and Its Effects on K-Means Clustering Algorithm. **Research Journal of Applied Sciences, Engineering and Technology**, v. 6, n. 17, p. 3299-3303, 2013. Disponível em: < https://www.semanticscholar.org/paper/Standardization-and-Its-Effects-on-K-Means-Mohamad-Usman/1d352dd5f030589ecfe8910ab1cc0dd320bf600d >. Acesso em: 12 jan 2020.

NERI, Marcelo. **Motivos da evasão escolar**. Rio de Janeiro: Fundação Getúlio Vargas, 2009. Disponível em: <https://cps.fgv.br/pesquisas/motivos-da-evasao-escolar>. Acesso em: 15 jan 2018.

PIMENTEL, Fernando S. C.; LIMA, Mônica. R. F. Evasão na EAD: O caso do curso de pós-graduação em EDHDI/UFAL. **Debates em Educação**, Maceió, v. 10, n. 21, p.185 - 199, mai./ago. 2018. Disponível em: <http://www.seer.ufal.br/index.php/debateseducacao/article/view/3397/pdf>. Acesso em 13 mai. 2019.

RIBEIRO, Marcelo A. O projeto profissional familiar como determinante da evasão universitária – um estudo preliminar. **Revista Brasileira de Orientação Profissional ABOP**, Porto Alegre, v. 6, n. 2, p. 55-70, 2005. Disponível em: <http://pepsic.bvsalud.org/pdf/rbop/v6n2/v6n2a06.pdf > Acesso em 12 jul. 2019.

RISTOFF, Dilvo. **Evasão:** Exclusão ou Mobilidade. UFSC, Florianópolis, 1995.

_____. **Universidade em foco**: reflexões sobre a Educação Superior. Florianópolis: Insular, 1999.

SCHMITT, Jeovani. Construção de uma escala de propensão à evasão estudantil em cursos de graduação. 2018. 174 p. Tese (Doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia de Produção, Florianópolis, 2018. Disponível em: <http://www.bu.ufsc.br/teses/PEPS5709-T.pdf&gt. Acesso em: 15 jan 2018.

SILVA, Alisson de Oliveira; et al. Modelos de sobrevivência aplicados à evasão dos alunos de estatística da UFPB. **Interscientia,** v. 6, n. 2, p. 134 – 145, 2018. Disponível em: < https://periodicos.unipe.br/index.php/interscientia/article/view/860>. Acesso em: 16 fev 2019.

SILVA, Glauco P. Análise de evasão no ensino superior: uma proposta de diagnóstico de seus determinantes. **Avaliação,** Campinas, v. 18, n. 2, p. 311-333, 2013. Disponível em: < http://www.scielo.br/scielo.php?pid=S1414-40772013000200005&script=sci_abstract&tlng=pt>. Acesso em 2 nov 2019.

SPADY, Willian G. Dropout from Higher Education: An interdisciplinary review and synthesis. **American Journal of Sociology**, Chicago, v. 1, p. 64–85, Apr. 1970.

TINTO, Vicent. Dropout from Higher Education: A theoretical synthesis of recent research. **Review of Educational Research**, v. 45, n. 1, p. 89-125, Winter 1975.

_____. **Leaving college**: rethinking the causes and cures of student attrition. 2. ed. Chicago: The University of Chicago, 1993.

_____. Classrooms as Communities: Exploring the Education Character of Student Persistence. **Journal of Higher Education**, v. 68, n. 6, p. 599-624, Nov-Dez, 1997

SILVA FILHO, Roberto Leal Lobo et al. A Evasão no Ensino Superior Brasileiro. **Caderno de Pesquisa**, São Paulo, v.37, n. 132, p. 641 - 659, set./dez. 2007.

SILVA FILHO. Roberto Leal Lobo. **A Evasão no Ensino Superior Brasileiro** – Novos Dados**.** Instituto Lobo / Lobo & Associados Consultoria. 2017. Disponível em: <http://www.institutolobo.org.br/imagens/pdf/artigos/art_088.pdf> . Acesso em: 22 jan. 2018.

WORLD ECONOMIC FORUM. How much data is generated each day? **WEF**, 2019. Disponível em: < https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/>. Acesso em: 22 fev. 2020.