Universidade Federal de Santa Catarina
Centro Tecnológico
Programa de Pós-Graduação em Métodos e Gestão em Avaliação

Fernando Curbani

ALGORITMOS DE APRENDIZAGEM DE MÁQUINA NA PREVISÃO DA EVASÃO NOS CURSOS DE ENGENHARIA DA UNIVERSIDADE FEDERAL DE SANTA CATARINA

Florianópolis/SC

2023

ALGORITMOS DE APRENDIZAGEM DE MÁQUINA NA PREVISÃO DA EVASÃO NOS CURSOS DE ENGENHARIA DA UNIVERSIDADE FEDERAL DE SANTA CATARINA

Fernando Curbani

Dissertação submetida ao Programa de Pós-Graduação em Métodos e Gestão em Avaliação (PPGMGA), da Universidade Federal de Santa Catarina para a obtenção do título de mestre em Métodos e Gestão em Avaliação.
Orientador: Prof. Dr. André Wüst Zibetti
Coorientadora: Profa. Dra. Andreia Zanella

Florianópolis/SC

2023

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

Curbani, Fernando
ALGORITMOS DE APRENDIZAGEM DE MÁQUINA NA PREVISÃO DA EVASÃO NOS CURSOS DE ENGENHARIA DA UNIVERSIDADE FEDERAL DE SANTA CATARINA / Fernando Curbani ; orientador, André Zibetti, coorientadora, Andreia Zanella, 2023.
101 p.

Dissertação (mestrado profissional) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Métodos e Gestão em Avaliação, Florianópolis, 2023.

Inclui referências.

1. Métodos e Gestão em Avaliação. 2. Predição de Evasão Universitária. 3. Aprendizagem de Máquina;. 4. Algoritmos de Classificação. 5. Engenharias. I. Zibetti, André . II. Zanella, Andreia. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Métodos e Gestão em Avaliação. IV. Título.

Fernando Curbani

ALGORITMOS DE APRENDIZAGEM DE MÁQUINA NA PREVISÃO DA EVASÃO NOS CURSOS DE ENGENHARIA DA UNIVERSIDADE FEDERAL DE SANTA CATARINA

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

Prof. Dr. André Wüst Zibetti
Universidade Federal de Santa Catarina

Prof. Dr. Alexandre Gonçalves Silva
Universidade Federal de Santa Catarina

Profa. Dra. Andréa Cristina Konrath
Universidade 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.

ICP-Edu
Documento assinado digitalmente
Marcelo Menezes Reis
Data: 22/06/2023 13:59:10-0300
CPF: ***.917.559-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Dr. Marcelo Menezes Reis
Coordenação do Programa de Pós-Graduação

ICP-Edu
Documento assinado digitalmente
Andre Wust Zibetti
Data: 22/06/2023 11:04:16-0300
CPF: ***.936.800-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Dr. André Wüst Zibetti
Orientador

Florianópolis, 2023.

À Isadora e Mathias(3), pelo
amor de sempre!

Agradecimentos

Em primeiro lugar à Deus, por tudo, Ele foi e é e essencial em todas as minhas conquistas e superações.

Em seguida, aos meus pais (Afonso *in memoriam* e Elizete) pela vida (já bastaria) e também pelas memórias e sorrisos.

À Isadora e Mathias que sempre estiveram ao meu lado me apoiando direta ou indiretamente ao longo de toda a minha trajetória. Amo vocês!

Agradeço aos meus orientadores por conduzir o meu trabalho de pesquisa. À Prof[a]. Dr[a] Andreia Zanella pelo paciente trabalho de revisão da redação deste trabalho. Ao Prof. Dr. André Wüst Zibetti, pelo suporte em suas correções e orientações metodológicas.

Agradeço aos professores que me acompanharam ao longo deste mestrado, pela excelência da qualidade técnica de cada um e também por se dedicarem à bela arte de ensinar, seus ensinamentos foram valiosos para o meu progresso acadêmico e para o complemento de ideias deste trabalho.

Aos professores Frank Siqueira e Antonio Carlos Mariani pelos dados fornecidos.

Ao professor Marcelo Menezes Reis pelas aulas e pela Coordenação do Programa de Pós-Graduação em Métodos e Gestão em Avaliação (PPGMGA).

Às professoras Sandra Regina Salvador Ferreira e Jaciane Lutz Ienczak pela compreensão durante a realização desta dissertação permitindo a conciliação com o meu trabalho.

Aos meus colegas e amigos que estiveram ao meu lado e que foi um prazer em conhecer ao longo do curso.

A todos que direta ou indiretamente fizeram parte de minha formação, o meu muito obrigado.

Pof fim, como disse Snoop Dog: “Eu quero me agradecer por acreditar em mim, quero me agradecer por todo esse trabalho duro. Quero me agradecer por não tirar folgas. Quero me agradecer por nunca desistir. Quero me agradecer por ser generoso e sempre dar mais do que recebo. Quero me agradecer por tentar sempre fazer mais o certo do que o errado. Quero me agradecer por ser eu mesmo o tempo inteiro”.

"To err is human, to forgive divine - but to include errors in your design is statistical."

(Leslie Kish, em *Chance, statistics, and statisticians*, 1977)

Resumo

A evasão universitária é um fenômeno preocupante que ocorre quando estudantes abandonam o ensino superior antes de concluírem seus cursos, essa é uma realidade comum em diversas instituições de ensino no Brasil e do mundo. Este trabalho tem como tema a evasão dos cursos de engenharia de uma universidade federal brasileira. O objetivo do estudo é realizar uma comparação do desempenho de diferentes algoritmos de aprendizagem de máquina na previsão da evasão dos alunos desses cursos. O estudo foi realizado na Universidade Federal de Santa Catarina, com uma amostra inicial de 4394 alunos. Para alcançar esse objetivo, o processo de CRISP-DM (Cross-Industry Standard Process for Data Mining) foi aplicado, utilizando variáveis independentes relacionadas a dados acadêmicos, socioeconômicos e demográficos em três momentos distintos: pré-matrícula, final do primeiro semestre e final do terceiro semestre. Os resultados obtidos revelaram uma Acurácia de 69,20%, 80,47% e 84,52% respectivamente nos três modelos analisados, com uma Sensibilidade máxima de 60,79% no Terceiro Semestre. Além disso, ao equilibrar os dados, a Sensibilidade aumentou para 79,93%. O ajuste do limiar de classificação também possibilitou uma maior capacidade de previsão dos estudantes propensos a evadir. A partir da avaliação dos modelos criados, o algoritmo *Random Forest* foi selecionado por ter trazido os melhores resultados de previsão.

Palavras-chave: Predição de Evasão Universitária; Aprendizagem de Máquina; Engenharias; Algoritmos de Classificação

Abstract

University dropout is a concerning phenomenon where students abandon higher education before completing their courses, and it is a common reality in many educational institutions in Brazil and worldwide. This study focuses on dropout in engineering courses at a Brazilian federal university. The objective is to compare the performance of different machine learning algorithms in predicting dropout among students in these courses. The study was conducted at the Federal University of Santa Catarina, with an initial sample of 4394 students. To achieve this goal, the CRISP-DM (Cross-Industry Standard Process for Data Mining) process was applied, using independent variables related to academic, socioeconomic, and demographic data at three different time points: pre-enrollment, end of the first semester, and end of the third semester. The results revealed an accuracy of 69.20%, 80.47%, and 84.52%, respectively, in the three analyzed models, with a maximum sensitivity of 60.79% in the third semester. Furthermore, by balancing the data, sensitivity increased to 79.93%. Adjusting the classification threshold also improved the ability to predict students at risk of dropout. Based on the evaluation of the created models, the Random Forest algorithm was selected as it provided the best prediction results.

Keywords: Student Dropout Prediction; Machine Learning; Engineering; Classification Algorithms

Lista de ilustrações

Lista de quadros

Lista de tabelas

Lista de abreviaturas e siglas

OCDE: Organização para a Cooperação e Desenvolvimento Econômico

INEP: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

UFSC: Universidade Federal de Santa Catarina

CTC: Centro Tecnológico

ANDIFES: Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior

FORPLAD: Fórum Nacional de Pró-Reitores de Planejamento e de Administração das Instituições Federais de Ensino Superior

IES: Instituição de Ensino Superior

TCG: Taxa de Conclusão de Graduação

TSG: Taxa de Sucesso na Graduação

MEC: Ministério da Educação

Instituto SEMESP: Sindicato das Entidades Mantenedoras de Estabelecimentos de Ensino Superior no Estado de São Paulo

RF: *Random Forest*

SVM: *Support Vector Machines*

KNN: *K-Nearest Neighbors*

RL: Regressão Logística

NB: *Naive Bayes*

Sumário

1 Introdução

A evasão universitária é um fenômeno preocupante que ocorre quando estudantes abandonam o ensino superior antes de concluírem seus cursos, essa é uma realidade comum em diversas instituições de ensino no Brasil e do mundo.

No Brasil, em 2020, a taxa de evasão em cursos presenciais foi de 28,5%, esse número representa uma quantidade significativa de estudantes que abandonaram seus estudos antes da conclusão (SEMESP, 2022). Já na Universidade Federal de Santa Catarina, de acordo com o *Relatório de Gestão 2021*, a taxa de evasão em 2019 foi de 11,33%, o que indica um aumento em relação ao ano anterior, quando a taxa foi de 10,60% (UFSC, 2021).

A nível mundial, um estudo realizado pela Organização para a Cooperação e Desenvolvimento Econômico - OCDE em 2019, mostra que o Brasil possui a média mais baixa entre jovens de 25 e 34 anos que se formam no Ensino Superior, com variações significativas entre os países membros. Por exemplo, no Reino Unido esse número é aproximadamente 85%, na Argentina 40% e no Brasil apenas 21% no período regular de curso (OECD, 2022).

O *Censo da Educação Superior de 2020* do INEP, mostra as lacunas da demanda por meio do ensino quando consideradas apenas a formação da grande área de - *Engenharia, Produção e Construção* - que corresponde a 11,30% das matrículas do ensino superior brasileiro. E ainda, os concluintes representam apenas 12% do total (INEP, 2022b). Além disso, pouco desta força de trabalho é convertida em desenvolvimento tecnológico e empresarial, o que deixa o país numa situação preocupante para o futuro da inovação, pesquisa e desenvolvimento tecnológico (FILHO et al., 2022).

As razões para a evasão dos cursos de Engenharia são o alto nível de exigência e a dificuldade dos cursos, que muitas vezes exigem uma grande quantidade de tempo e esforço por parte dos alunos. Outro fator que pode contribuir para a evasão nesses cursos é a falta de preparação adequada dos estudantes para lidar com a complexidade do conteúdo técnico frente a uma formação básica deficiente. Além disso, alguns estudantes podem se sentir desestimulados por um ambiente de aprendizagem pouco inspirador e falta de perspectivas de carreira na área (CHRISTO; RESENDE; KUHN, 2018).

Os impactos da evasão vão além dos estudantes. Eles refletem também nas instituições de ensino e na sociedade como um todo, incluindo perda de recursos financeiros, humanos e baixa produtividade no ensino. No setor público, são recursos públicos investidos sem o devido retorno. No setor privado, é uma importante perda de receitas. Em ambos os casos, a evasão é uma fonte de ociosidade de professores, funcionários, equipamentos e espaço físico (FILHO et al., 2007). Por isso, é importante entender as causas e consequências

desse problema e buscar soluções para reduzir a taxa de evasão nas universidades.

Em termos históricos, a compreensão do fenômeno é resultado de uma série de estudos realizados a partir da década de 1970, quando a evasão se tornou uma preocupação crescente para as instituições de ensino superior. No início, a evasão universitária foi amplamente vista como um problema de motivação dos estudantes, com muitos pesquisadores argumentando que a evasão era causada principalmente pela falta de interesse dos estudantes nos cursos ou pela incompatibilidade entre as expectativas dos alunos e as exigências do ensino superior. No entanto, pesquisas posteriores mostraram que a evasão é um fenômeno complexo e multicausal, influenciado por uma ampla variedade de fatores, incluindo fatores socioeconômicos, acadêmicos e psicológicos (THELIN, 2010). O fenômeno fora estudado por diversos autores que nessa época criaram modelos teóricos que buscaram explicar suas causas, entre eles: Spady (1970) (modelo sociológico); Tinto (1975) (modelo de compromisso pessoal e institucional); Bean (1980) (modelo de performance, psicossocial e ambiental); Ethington (1990) (modelo psicológico); Cabrera et al. (1999) (modelo socioeconômico); esses autores serviram como base para pesquisas subsequentes e complementares, podendo ser combinados de diferentes maneiras para explicar o fenômeno da evasão.

Nacionalmente, os primeiros estudos sobre evasão datam da década de 1990, quando as preocupações do MEC em relação aos altos índices de evasão nas Universidades Públicas se tornam mais evidentes (ANDIFES, 1996). Na década seguinte, iniciaram-se as discussões sobre a necessidade de democratização do ensino superior e da ampliação do acesso à universidade: a tríade expansão, qualidade e democratização (MEC, 2014). Desde então, a evasão tem sido uma preocupação constante das universidades e dos gestores da educação no país.

A partir dos anos 2000, os estudos sobre evasão universitária começaram a incorporar métodos de análise estatística e modelos teóricos mais sofisticados, permitindo que os pesquisadores identificassem padrões mais precisos nos dados e testassem hipóteses mais elaboradas sobre as causas da evasão. Desde então, a pesquisa sobre evasão universitária continua a evoluir, com novos estudos explorando o papel da tecnologia, das intervenções precoces e das políticas públicas na prevenção da evasão (ALVAREZ; CALLEJAS; GRIOL, 2020).

Hoje, o estudo da evasão universitária é considerado um campo de pesquisa importante no Brasil, sendo objeto de investigação de diversos grupos de pesquisa e instituições de ensino, impulsionando a realização de diversos estudos que procuram evidenciar ou mesmo compreender os motivos que levam à evasão seja no ensino superior privado ou público, na forma presencial ou à distância. A partir desses estudos, tem-se buscado compreender suas causas e consequências, bem como identificar estratégias para prevenir e reduzir a taxa de abandono dos cursos de graduação e pós-graduação

(COIMBRA; SILVA; COSTA, 2021).

Na vanguarda do conhecimento, o foco dos últimos anos está na previsão da evasão mediante utilização de **aprendizado de máquina (*machine learning*)**, mostrando-se uma ferramenta valiosa para auxiliar na redução da evasão universitária (LIZ-DOMíNGUEZ et al., 2019). Com a análise de dados históricos e o uso de algoritmos, é possível identificar os fatores que mais influenciam a evasão e prever quais alunos têm maior probabilidade de abandonar seus cursos (OLIVEIRA et al., 2021).

Com essas informações, as instituições de ensino podem implementar intervenções precoces, oferecer orientação acadêmica ou financeira, fornecer ajuda extra para os alunos que estão com dificuldades ou até mesmo fornecer incentivos para aqueles que estão em risco de evasão, a fim de aumentar as chances de que permaneçam na faculdade e concluam seus cursos com sucesso (BARDAGI; HUTZ, 2005).

A UFSC tem adotado algumas dessas medidas para tentar combater a evasão universitária, incluindo a criação de programas de tutoria, de orientação vocacional e de apoio financeiro aos estudantes em situação de vulnerabilidade socioeconômica. Além disso, a universidade tem investido na melhoria da qualidade do ensino e na criação de políticas de acolhimento aos estudantes ingressantes, visando reduzir os índices de evasão (PRAE, 2022).

A previsão com o auxílio de **aprendizagem de máquina** pode também, ajudar as instituições de ensino a entender melhor os padrões de comportamento dos alunos e identificar tendências em seus desempenhos acadêmicos, permitindo que elas ajustem seus programas de ensino e aprendizagem para melhor atender às necessidades e aumentar as chances de sucesso, principalmente nas áreas tecnológicas que estão em constante evolução (ZHANG et al., 2021).

E para que isso possa ser possível, diversos algoritmos são utilizados, a escolha do melhor depende das características do conjunto dos dados e também dos objetivos da pesquisa. Exemplo desses algoritmos são: Árvores de Decisão, Regressão Logística, Redes Neurais Artificiais (RNA), *Random Forests*, *Support Vector Machines, K-Nearnest Neighbours, ADABoosting, Naive Bayes* entre outros. Cada um desses algoritmos apresenta vantagens e desvantagens em relação à precisão, velocidade de processamento e facilidade de interpretação. É importante selecionar o algoritmo adequado para cada caso e considerar as características do conjunto de dados que permitam chegar a uma melhor previsão da evasão.

1.1 Perguntas da Pesquisa

A partir do tema apresentado e tendo em vista a possibilidade no uso de algoritmos de aprendizagem de máquina na previsão da evasão do ensino superior e, especialmente nos cursos de Engenharias, pergunta-se:

1. Com base nos dados dos estudantes e em diferentes momentos do percurso acadêmico, a partir da pré-matrícula, no fim do primeiro e no fim do terceiro semestre, é possível prever quais estudantes estão propensos a evadir dos cursos de Engenharia da UFSC com algoritmos de aprendizagem de máquina?

2. Qual algoritmo de aprendizagem de máquina apresenta a maior eficácia na previsibilidade desta evasão?

Até o momento ainda não existem estudos usando técnicas de aprendizagem de máquina com dados especificamente nos cursos de Engenharias da UFSC – este trabalho se propõe a realizar este estudo no Centro Tecnológico (CTC/UFSC) campus Florianópolis/SC.

1.2 Justificativa

A evasão é um dos maiores problemas de qualquer nível de ensino, também o é, no ensino superior brasileiro, público e privado. O abandono do aluno sem a finalização dos seus estudos representa perda social, de recursos e de tempo de todos os envolvidos no processo de ensino – alunos, professores, instituições de ensino, o sistema de educação, a sociedade e todo o País (LOBO, 2012).

Segundo Lobo (2019), o desperdício de recursos decorrente da evasão, tanto nas IES públicas como privadas no Brasil, são da ordem de 15 bilhões de reais anuais. Um estudo americano do *Educational Policy Institute*, de 2013, envolvendo quase 1700 IES, calculou que o custo anual da evasão no ensino superior americano é de cerca de 40 bilhões de dólares e uma instituição que consiga baixar a evasão de 20% para 15%, por exemplo, é capaz de aumentar em 8% o seu faturamento.

Para a identificação precoce da evasão nos cursos de Engenharia da UFSC, nessa dissertação será utilizada a técnica de Aprendizagem de Máquina com a utilização de diferentes algoritmos de classificação, com diferentes variáveis independentes e dados rotulados de forma binária entre “Evadido” e “Formado”. Os algoritmos de aprendizagem de máquina têm se mostrado uma ferramenta promissora em realizar previsões e análises de dados com base em modelos estatísticos e padrões históricos.

Isso irá permitir futuramente a Universidade Federal de Santa Catarina antecipar tendências de evasão e identificar padrões para a tomada de decisões adequadas com base nessas informações de modo a mitigar a evasão em conjunto com programas de permanência.

1.3 Objetivo

Para responder às perguntas foram desenvolvidos os objetivos geral e específicos.

1.3.1 Objetivo Geral

Comparar o desempenho de diferentes algoritmos de aprendizagem de máquina através da previsão da evasão dos alunos dos cursos de Engenharia da Universidade Federal de Santa Catarina (UFSC).

1.3.2 Objetivos Específicos

Considerando o desenvolvimento do trabalho e o objetivo geral apresentado, destacam-se os seguintes objetivos específicos:

- Conhecer por meio de uma análise exploratória algumas propriedades dos dados utilizados na previsão;
- Selecionar, a partir de técnicas de pré-processamento de dados, as variáveis independentes mais relevantes para predizer a evasão;
- Realizar a modelagem dos dados com diversos algoritmos de aprendizagem de máquina;
- Avaliar a eficiência das previsões realizadas pelos algoritmos utilizando diferentes métricas de avaliação;
- Realizar o ajuste dos parâmetros do modelo para a melhor performance na previsão.

1.4 Delimitação da Pesquisa

Para que problemas como o Paradoxo de Simpson[1] pudessem ser evitados pelos modelos de aprendizagem de máquina (COSTA et al., 2020), o desenvolvimento deste

[1] O paradoxo de Simpson é um fenômeno em probabilidade e estatística, em que uma tendência aparece em diversos grupos de dados, mas desaparece ou reverte quando esses grupos são combinados, ou seja, o mesmo conjunto de dados pode parecer mostrar tendências opostas, dependendo de como está agrupado (H.WAGNER, 1982)

trabalho se deu apenas nos cursos de Engenharia da Universidade Federal de Santa Catarina no Campus Florianópolis/SC (Centro Tecnológico-CTC), pois os cursos de Engenharia apresentam na sua maioria currículos muito semelhantes permitindo evitar conclusões enganosas através de uma grande generalização de cursos e diferentes currículos. Dessa forma, permitem uma abordagem mais específica e precisa dos dados para os modelos.

A seleção dos algoritmos de classificação foi baseada em uma revisão da literatura, levando em consideração os algoritmos de aprendizagem de máquina que são de fácil interpretação e têm apresentado melhor desempenho de acordo com estudos anteriores (AGRUSTI; BONAVOLONTà; MEZZINI, 2019) (OLIVEIRA et al., 2021).

A delimitação temporal deste estudo foi estabelecida a partir de 2008, pois a partir dessa data tornou-se possível o acompanhamento individual dos alunos por meio de um identificador. O período selecionado vai até 2019, a fim de utilizar dados pré-pandemia.

1.5 Organização do Trabalho

O presente trabalho está organizado da seguinte forma:

Capítulo 1 Introdução, Perguntas, Justificativa e Objetivos.

Capítulo 2 Revisão de literatura abordando o referencial teórico da Evasão no Ensino Superior.

Capítulo 3 Revisão de literatura abordando o referencial teórico de Aprendizagem de Máquina.

Capítulo 4 Revisão Sistemática de Literatura com o foco nos principais Algoritmos, Variáveis Independentes e Métricas de Avaliação utilizados na previsão de evasão universitária através de aprendizagem de máquina.

Capítulo 5 Metodologia utilizada, contextualização dos dados e a etapa de Pré-processamento dos dados.

Capítulo 6 Resultados da Análise Exploratória dos Dados – AED, da Seleção de Atributos (*Feature Selection*) e dos Resultados dos Modelos com as Métricas selecionadas.

Capítulo 7 Discussão com trabalhos relacionados.

Capítulo 8 Considerações finais e possíveis trabalhos futuros.

2 Evasão Universitária

Neste Capítulo, serão abordados alguns conceitos relacionados à evasão universitária, como sua definição, problemas e formas de mensuração.

2.1 Conceito de Evasão

Embora pareça um fenômeno de simples definição, conceituar evasão pode ser um desafio, pois nem sempre o aluno evadiu completamente da instituição ou do sistema de ensino. Para diferenciar os diferentes tipos de evasão, é necessário levar em consideração o contexto específico em que ela ocorre.

O Fórum Nacional de Pró-Reitores de Planejamento e Administração (Forplad), da Andifes, com base no documento da Comissão Especial de Estudos sobre Evasão (ANDIFES, 1996) qualifica a evasão em três tipos principais: a **Microevasão ou Evasão do curso** é o desligamento formal do estudante de um determinado curso (saída do curso, mas permanece na instituição); a **Mesoevasão ou Evasão da instituição** é a perda definitiva do vínculo do discente com a instituição (sai duma instituição e vai para outra) e **Macroevasão ou Evasão do sistema** é quando o discente abandona os estudos (sai do sistema educacional) (FORPLAD, 2016).

O INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira), órgão responsável pelo monitoramento e avaliação do ensino no Brasil, define a evasão como:

> "o abandono do curso antes de sua conclusão por desistência (independente do motivo), representando condição terminativa de promover o aluno a uma condição superior ao do ingresso, no que diz a respeito à ampliação do conhecimento, habilidades e competências para o seu nível de ensino" (INEP, 2017).

Já para Ristoff (1995), o aluno que saiu do curso não é considerado como evadido, o autor considera este tipo de evasão como mobilidade acadêmica, segundo ele:

> "A parcela significativa do que chamamos evasão, no entanto, não é exclusão, mas mobilidade, não é fuga, mas busca, não é desperdício, mas investimento, não é fracasso - nem do aluno nem do professor, nem do curso ou da instituição - mas tentativa de buscar o sucesso ou a felicidade, aproveitando as revelações que o processo natural do crescimento dos indivíduos faz sobre suas reais potencialidades" (RISTOFF, 1995)

Para Biazus (2004) o que caracteriza evasão é a saída do aluno da instituição de ensino, ou de um dos seus cursos, de forma temporária ou definitiva, por qualquer motivo,

desde que não seja a diplomação. O autor considera o trancamento também como uma forma de evasão (temporária) bem como a transferência interna como evasão de curso.

Atualmente, a UFSC considera para o cálculo da taxa de evasão o conceito: "alunos desvinculados da universidade no ano corrente, considerando matrículas canceladas, transferidas para outras instituições, abandonos e jubilações, dividido pelo número de alunos matriculados no ano. As matrículas trancadas não são consideradas evasão" (UFSC, 2020, p.7).

Nesta dissertação será estudada a evasão dos cursos, os alunos classificados como "Evadidos" foram aqueles que tiveram a saída do curso por meio de: transferências, abandono, troca de curso, jubilações, desistências e eliminações administrativas. Os alunos com situação regular, trancados ou falecidos foram excluídos da base de dados e os alunos formados e concluintes foram agrupados na categoria "Formados".

2.2 Indicadores de Evasão

O cálculo da evasão varia de acordo com a forma com que o conceito é definido, quando a definição de evasão e a maneira pela qual é mensurada não são consistentes e claras, podem-se apresentar erros de interpretação, impossibilidade de comparações e, portanto, gerar riscos de decisões e encaminhamentos incorretos ou mesmo desnecessários. Outra questão a ser observada na mensuração da evasão é na forma como as taxas ou dados absolutos são calculados. Definir os elementos que compõem o cálculo dos índices de evasão depende dos objetivos de estudo e da disponibilidade dos dados (FREITAS, 2016).

A coleta de dados para os cálculos de evasão pode ser por **dados individuais do aluno,** por exemplo CPF´s ou outra identificação, ou por **dados agregados por curso ou instituição,** como os dados anuais disponíveis no Censo do Ensino Superior do INEP. Além disso, as análises podem ser feitas de duas maneiras distintas: estudos longitudinais e estudos transversais.

Os **estudos longitudinais** envolvem a coleta de dados de uma mesma amostra de indivíduos ou grupos ao longo do tempo. Isto é, acompanham os mesmos participantes ao longo de um período determinado, fazendo medições repetidas em momentos diferentes. Já os **estudos transversais** envolvem a coleta de dados de diferentes amostras de indivíduos ou grupos em um único momento no tempo. Ou seja, medem as variáveis de interesse em um único momento, em uma amostra representativa da população em questão.

Os cálculos feitos pelo INEP até o ano de 2008, eram feitos utilizando dados fornecidos pelas IES apenas de forma agrupada por curso. A partir do ano de 2009, entretanto, iniciou-se a coleta individualizada dessas informações (por CPF´s). A nova metodologia, assim, possibilitou a justaposição das informações anuais dos indivíduos

para a composição de uma trajetória acadêmica do aluno (acompanhamento longitudinal) (INEP, 2017). O INEP coleta os dados de forma individual, porém divulga os índices de evasão de forma longitudinal e agregada.

Já as formas de cálculo da Taxa de Sucesso na Graduação - TSG (TCU, 2004) e da Taxa de Conclusão de Graduação – TCG (BRASIL, 2007) levam em consideração apenas os dados agregados por curso, vide Quadro 1. Existem vantagens e desvantagens nesses dois índices. A vantagem é a possibilidade de ser calculada a partir de dados públicos disponíveis nas sinopses estatísticas como o *Censo da Educação Superior* do INEP. Porém, Junior et al. (2019), apontam para um desvantagem:

> "Usualmente são consideradas vagas, ingressos e matrículas, mas não os indivíduos propriamente ditos. Com a expansão e a diversificação da educação superior, é cada vez mais comum que alunos façam novo ingresso para mudar de curso ou de instituição. Cada vez mais, a evasão do curso não é evasão propriamente dita, mas mobilidade. Uma forma de dar conta das trajetórias cada vez mais complexas realizadas pelos alunos é acompanhá-los pelo número de seu cadastro de pessoa física (CPF), em uma abordagem longitudinal" (JUNIOR et al., 2019, p.164).

Quando são usados dados agregados para instrumentalizar o cálculo, é inevitável que se diluam as histórias, as trajetórias, as especificidades e, principalmente, as causas diferenciais. Para isso, é necessário que se proceda o acompanhamento individual de cada evadido, para entender suas razões (COIMBRA; SILVA; COSTA, 2021).

Nesse sentido, as ponderações de Ristoff (1995) dizem respeito, em sua perspectiva, que as análises dominantes veem a evasão como exclusão, como perda, fuga ou um resíduo, ignorando o que poderiam ser, na verdade, resultados das aspirações maiores dos seres humanos além dos limites destas instituições e que por vezes nem sempre irão se alinhar ou se apropriar do conhecimento acadêmico.

Conforme Lobo (2012), a evasão universitária, independentemente da razão, resulta em perdas que devem ser analisadas, mesmo que sejam "compensadas" pela ocupação de vagas em outros cursos da instituição de ensino superior (IES), ou até mesmo no mesmo curso. Porém, medir a evasão não é como verificar um "saldo de caixa", ou seja, a simples análise de quantos alunos entraram menos quanto saíram. Contudo, é importante analisar quem entrou e quem saiu e por quais razões, para que seja possível evitar outras perdas pelos mesmos motivos com ações que gerem mudanças e essas só acontecem se entendemos, claramente, o que está ocorrendo (LOBO, 2012).

Valendo-se de divisões entre numeradores e denominadores, as fórmulas tendem a apontar taxas ou percentuais de evasão, não raro entendidos como percentuais de sucesso ou fracasso (COIMBRA; SILVA; COSTA, 2021). Nesse contexto, no Quadro 1, exemplifica algumas mensurações de evasão encontradas na literatura, e sua diferentes formas de cálculo.

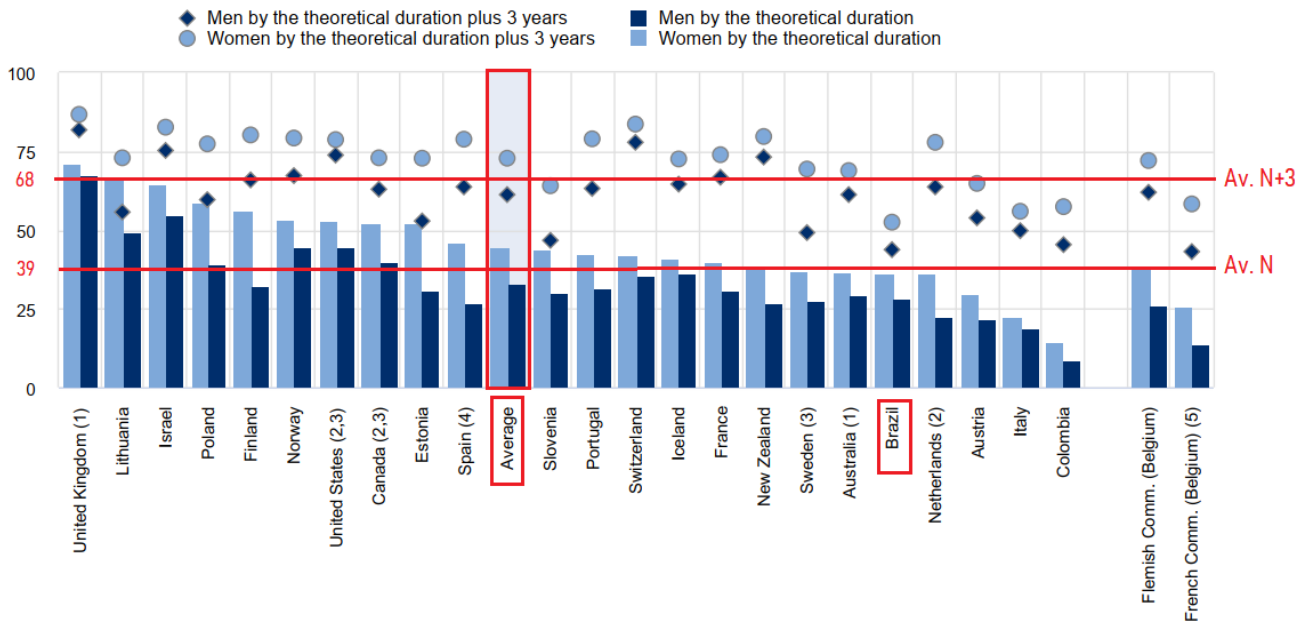
Quadro 1 – **Quadro comparativo das fórmulas de cálculo de evasão**

Referência	Definição	Fórmulas
TCU (2004)	**"TSG – Taxa de Sucesso na Graduação**: razão da quantidade de diplomados no ano n sobre a quantidade de ingressantes no ano (n – k), em que k depende do tempo de previsão de duração do curso", p.3	$TSG = \frac{Nº\ de\ diplomados\ (N_{DI})}{Nº\ total\ de\ alunos\ ingressantes\ (n-k)}$
BRASIL (2007)	**"TCG – Taxa de Conclusão de Graduação:** relação entre o total de diplomados nos cursos de graduação presenciais (DIP) num determinado ano e o total de vagas de ingresso oferecidas pela instituição (ING5) cinco anos antes", p.4	$TCG = \frac{DIP}{ING_5}$
ANDIFES (1996)	"considera-se a série histórica de dados sobre uma geração/turma de alunos ingressantes e o tempo máximo de integralização curricular, por "gerações completas" , p. 28	$\%Evasão = \frac{(N_i - N_d - N_r)}{N_i} \times 100$
SILVA FILHO et al. (2007)	"cálculo básico desse estudo, a comparação entre o número de alunos que estavam matriculados num determinado ano, subtraídos os concluintes, com a quantidade de alunos matriculados no ano seguinte, subtraindo-se deste último total os ingressantes desse ano", p. 645	$E_n = 1 - \frac{(M_n - I_n)}{(M_{n-1} - C_{n-1})}$
LIMA JUNIOR et al.(2019)	**"Taxa Longitudinal de Evasão TLE** (p, q) é definida como o percentual das pessoas físicas que, tendo ingressado no ano q, não obtiveram diploma até o ano de observação p", p.169	$\mathrm{TLE}_{p,q} = \frac{Quantidade\ de\ alunos\ sem\ diploma\ no\ ano\ p\ entre\ os\ que\ ingressaram\ no\ ano\ q}{Quantidade\ de\ alunos\ que\ ingressaram\ no\ ano\ q}$
BRAGA; MIRANDA-PINTO; CARDEAL (1997)	Neste trabalho, os autores observaram que um fator de rematrícula que diminui os valores de evasão em 6%, o que foi incorporado à evasão sendo denominados por estes autores por "geração incompleta", p. 439	$\%\ Evasão = 100\% - \%Formandos - 0{,}06 \times (100\% - \%Formandos)$
SERPA; PINTO (2000)	"a evasão de um ano é a diferença entre o número de ingressos no ano seguinte e a soma da variação da matrícula para o ano seguinte com o número de concluintes do ano em questão", p. 113	$e_x = I_{x+1} - (M_{x+1} - M_x) - C_x$
INEP (2017)	"percentual do número de estudantes que desistiram (desvinculado ou transferido) do curso j até o ano t (acumulado) em relação ao número de ingressantes do curso j no ano T, subtraindo-se o número de estudantes falecidos do curso j do ano T até o ano t", p. 17	$\mathrm{Tda}_{j,T,t} = \frac{\sum_{w=T}^{t}\sum_{i=1}^{n_{3,j,w}} Des_{i,j,t} + \sum_{w=T}^{t}\sum_{i=1}^{n_{4,j,w}} Transf_{i,j,t}}{\sum_{i=1}^{n} IG_{i=j}^{T} - \sum_{w=T}^{t}\sum_{i=1}^{n_{6,j,w}} Fal_{i,j,t}} \times 100$
SEMESP (2022)	A fórmula de cálculo do Instituto Semesp considera alguns motivos de evasão, como as matrículas trancadas, os desvinculados e falecidos	$Taxa\ de\ Evasão = \frac{(Matrículas\ trancadas + Desvinculados + Falecidos)}{(Total\ de\ Matrículas + Matrículas\ trancadas + Desvinculados + Falecidos)}$

Fonte: adaptado de Coimbra, Silva e Costa (2021) e Freitas (2016)

Internacionalmente, a OCDE utiliza o indicador B5 - *Quantos estudantes completam o ensino superior?* - do relatório *Education at a Glance* que se refere à "Taxa de formação bruta de bacharéis e diplomas equivalentes" em relação à população entre 25 e 34 anos (OECD, 2022). A metodologia desse indicador é detalhada na página 210 do documento e envolve a utilização de dois métodos de cálculo distintos: **Coorte verdadeira - *true cohort*** (dados de nível individual) e **Coorte cruzada - *cross cohort*** (dados agregados). O indicador B5 utiliza exclusivamente o método da coorte verdadeira, expresso em porcentagem, e ressalta que os dois métodos **não** são comparáveis entre si. Por meio da coorte verdadeira, o índice B5 apresenta a taxa de conclusão em dois prazos diferentes: a duração teórica (N) dos cursos e a duração teórica acrescida de mais três anos (N+3). Somente países com informações longitudinais são capazes de fornecer tais dados, incluindo o Brasil a partir de 2009 (INEP, 2017). A Figura 1 ilustra a taxa média brasileira em comparação com outros países.

Figura 1 – **Taxa de conclusão de estudantes que ingressaram no ensino superior no período de curso (N) e mais três anos (N+3), por gênero (somente *true cohort*)**



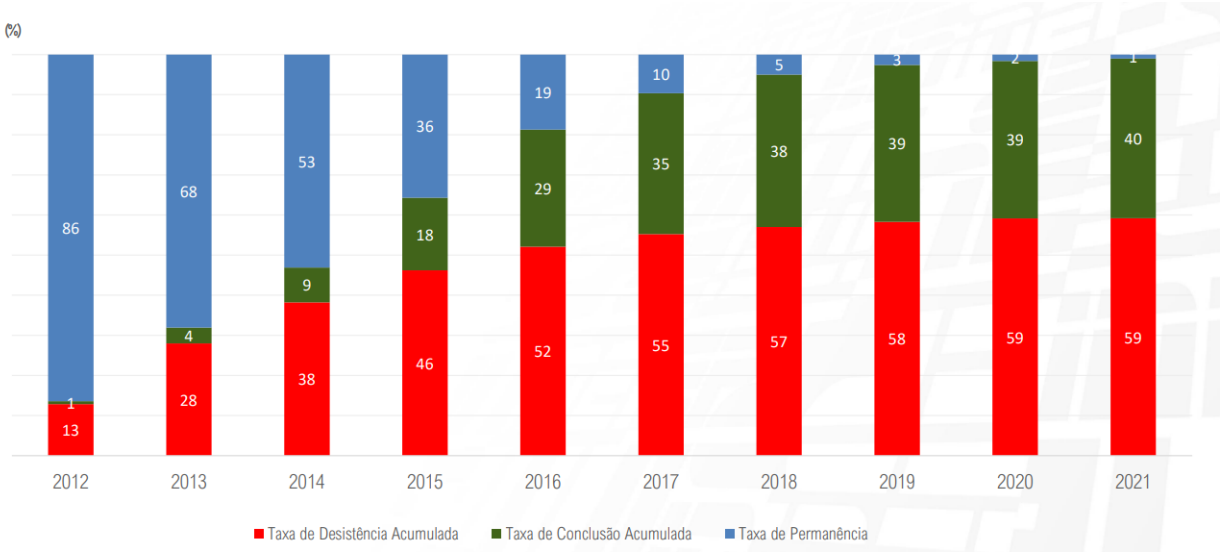
Fonte: adaptado de OECD (2022)

Conforme Figura 1, a média de formação no período regular (N) nos cursos de graduação nos países membros é de 39%, após três anos adicionais (N+3) essa taxa aumenta para 68%. Observa-se também, que em todos os países, as mulheres têm taxas de conclusão mais altas do que os homens em programas de bacharelado. A maior a diferença de gênero para conclusão dentro da duração teórica do programa é de 24 pontos percentuais na Finlândia.

Alguns países, no período regular, possuem uma taxa de formação inferior à do Brasil (33%), como a Colômbia (12%), Itália (21%), Áustria (26%) e Países Baixos (29%). No entanto, quando são considerados três anos adicionais (N+3), esses países apresentam uma taxa de formação acima de 50%, demonstrando uma maior eficácia além do tempo regular. No Brasil, a taxa de formação em N+3 é de 49%. Já o Reino Unido, Suíça, Israel, Nova Zelândia e Estados Unidos possuem uma taxa de formação acima de 75% em N+3.

Nacionalmente, o INEP disponibiliza além da Taxa de desistência acumulada (TDA), também indicadores de permanência (TAP) e conclusão (TCA), calculados a partir do acompanhamento da trajetória (fluxo) dos alunos ingressantes em um determinado ano. Na Figura 2 , são mostrados os três indicadores básicos do INEP para estudantes na coorte 2012-2021. Pode-se observar que em 2021 a taxa de desistência dessa coorte estava em 59%.

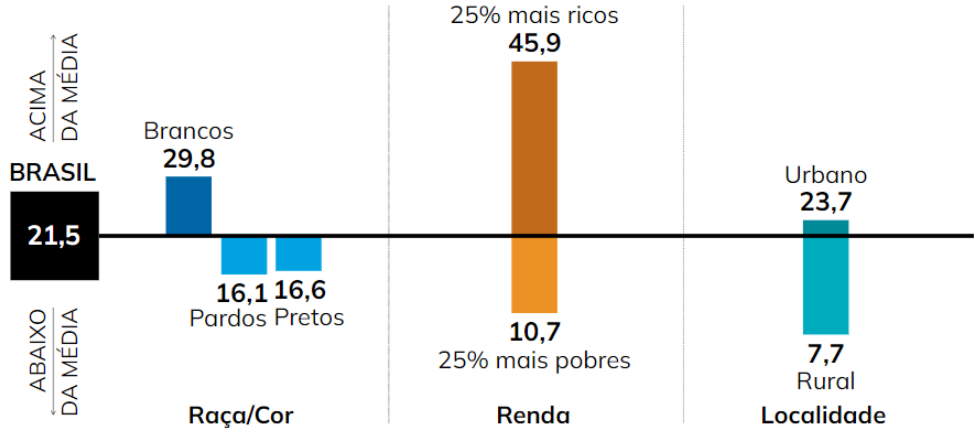
Figura 2 – **Indicadores de trajetória dos estudantes no curso de ingresso na coorte 2012 Brasil 2012-2021**



Fonte: (INEP, 2022a)

De acordo com os dados do Semesp, os anos de 2020 e 2021 apresentaram os maiores índices de evasão da história do ensino privado. Em 2020, aproximadamente 3,42 milhões de estudantes abandonaram os estudos, representando uma taxa de evasão de 36,6%. Já em 2021, esse número aumentou para 3,78 milhões de estudantes, resultando em uma taxa de evasão de 37,2%. Quando analisamos separadamente o ensino à distância, as taxas de evasão foram ainda mais altas, atingindo 40% em 2020 e 43,3% em 2021. Além disso, a inadimplência no ensino privado também cresceu nos anos de 2020 e 2021, atingindo 9,9% e 9,4%, respectivamente. Esses fatores têm impacto especialmente sobre os estudantes em situação de vulnerabilidade social, que muitas vezes precisam conciliar estudos e trabalho (LUDER, 2022).

Figura 3 – **Taxa Líquida de matrícula em relação a raça, renda e localidade Brasil em 2019**



Fonte: MODERNA (2021)

Além disso, um olhar também sobre essas vulnerabilidades, permite compreender nuances do sistema de ensino em que mostra um país desigual também na educação superior. É importante evidenciá-las para que as políticas públicas sejam colocadas em prática de forma mais equitativa. Por isso, é essencial também a análise de alguns recortes: como região, localidade, renda e raça/cor, conforme Figura 3.

Observa-se que apenas 10,7% dos jovens entre 18 e 24 anos pertencentes aos 25% mais pobres possuem matrícula no ensino superior. Em contraste, essa taxa sobe para 45,9% quando se trata dos 25% mais ricos. Além disso, apenas 7,7% da população da zona rural nessa faixa etária está matriculada no ensino superior, em comparação com 23,7% da zona urbana. Também é importante ressaltar a diferença de quase 15% entre a proporção de pretos e pardos em comparação com brancos. Nesse sentido, a meta 8 do Plano Nacional da Educação - PNE, que busca promover a equidade na educação, ainda está longe de ser alcançada:

> "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, 2015).

Na Universidade Federal de Santa Catarina, nos últimos anos, foram realizados avanços significativos em relação à equidade do ensino por meio das Políticas de Ações Afirmativas (PAA). Desde 2008, essas políticas têm garantido uma porcentagem de vagas reservadas para PPI - Pretos, Pardos e Indígenas, além de estudantes provenientes de escolas

públicas. Essas iniciativas têm contribuído para promover maior inclusão e diversidade no ambiente acadêmico da instituição (PRAE, 2022).

2.3 Evasão Centro Tecnológico da UFSC

A Universidade Federal de Santa Catarina (UFSC) é uma das 68 universidades federais do Brasil. Sua sede está localizada em Florianópolis, capital do estado de Santa Catarina. Fundada em 18 de dezembro de 1960, a UFSC tem como missão promover o ensino, a pesquisa e a extensão. A instituição é composta por docentes, técnicos-administrativos em Educação (TAE) e estudantes de diversos níveis de ensino, incluindo graduação, pós-graduação, ensino médio, fundamental e básico.

O indicador de Ensino E2P, apresentado na página da Coordenadoria de Gestão Estratégica (CGE) da Universidade Federal de Santa Catarina (UFSC), fornece dados sobre a taxa de evasão nos anos de 2018 e 2019. No ano de 2018, a taxa de evasão foi de 10,6%, enquanto em 2019 essa taxa foi de 11,33% (UFSC, 2021).

A UFSC estabeleceu metas de evasão até 2024, buscando reduzir os índices e promover a permanência e o êxito dos estudantes. A meta estabelecida para esse período é de 7%, refletindo o compromisso da instituição em criar um ambiente acadêmico favorável ao desenvolvimento dos estudantes (UFSC, 2021).

O Centro Tecnológico (CTC) é uma das unidades de ensino da Universidade Federal de Santa Catarina (UFSC), composta por um total de 15 unidades. Fundado em 1960, o CTC é reconhecido pelos cursos nas áreas tecnológicas, engenharias e arquitetura. O CTC abriga 10 departamentos acadêmicos, oferecendo aos estudantes uma ampla variedade de opções educacionais. São disponibilizados 15 cursos de graduação, abrangendo diversas áreas do conhecimento. Além disso, o centro conta com 14 programas de mestrado, incluindo um programa de mestrado profissional, e 12 programas de doutorado, todos relacionados às áreas tecnológicas, de engenharias e arquitetura.

Taxa de Conclusão de Graduação – TCG (BRASIL, 2007) calculada para os cursos de Engenharia do CTC da UFSC a partir dos dados da Série Histórica da UFSC 1980-2021 (UFSC, 2021), foi de 57% no período de 2016 a 2021, conforme Tabela 1.

Tabela 1 – **TCG calculada para os cursos de Engenharias do CTC dos anos 2016 a 2021**

Código	Curso	2016	2017	2018	2019	2020	2021	Média
201	ENGENHARIA CIVIL	80%	65%	89%	63%	59%	64%	70%
202	ENGENHARIA ELÉTRICA	45%	57%	72%	71%	59%	48%	59%
203	ENGENHARIA MECÂNICA	65%	75%	64%	52%	64%	48%	61%
211	ENGENHARIA SANITÁRIA E AMBIENTAL	51%	68%	73%	63%	70%	48%	62%
212	ENGENHARIA DE PRODUÇÃO CIVIL	73%	74%	75%	58%	2%	121%	67%
213	ENGENHARIA DE PRODUÇÃO ELÉTRICA	44%	50%	50%	54%	2%	61%	44%
214	ENGENHARIA DE PRODUÇÃO MECÂNICA	65%	68%	68%	46%	7%	103%	59%
215	ENGENHARIA DE ALIMENTOS	63%	49%	39%	20%	43%	54%	45%
216	ENGENHARIA QUÍMICA	57%	82%	107%	30%	91%	69%	73%
234	ENGENHARIA DE CONTROLE E AUTOMAÇÃO	66%	58%	79%	43%	17%	96%	60%
235	ENGENHARIA ELETRÔNICA	19%	31%	24%	37%	8%	28%	25%
236	ENGENHARIA DE MATERIAIS	75%	87%	58%	65%	30%	50%	61%
Média total								**57%**

Fonte: o autor, com base em UFSC (2021)

Em instituições como o Instituto Tecnológico de Aeronáutica (ITA) e o Instituto Militar de Engenharia (IME), a TCG média é superior a 95%. Já a taxa de conclusão da graduação média no setor público é de cerca de 60% e, no setor privado, de 40%, com dados analisados pela Confederação Nacional da Indústria (CNI) com base no Censo da Educação Superior (MONACO, 2013).

Resumindo, a evasão possui várias definições a depender do contexto e da forma como a coleta dos dados é feita. Isto impacta diretamente em seu cálculo, o que reflete diferentes indicadores, nem sempre comparáveis entre si. Além disso, os números mostram que internacionalmente o Brasil precisa melhorar no número de formados entre 25 e 34 anos em comparação com os países analisados. Nacionalmente, mesmo numa coorte ampla de dez anos ao final do período temos uma ampla evasão, de 59% nos cursos superiores brasileiros. Também, há uma enorme desigualdade entre as diferentes localidades, raças e tipos de renda, no qual precisam de atenção dos governantes para desenvolver políticas públicas de modo a diminuir a diferença dos que podem estudar no ensino superior para alcançar a meta 8 do Plano Nacional da Educação. Já na UFSC, apesar de termos uma evasão aproximada de 10% no geral, somente nos cursos de Engenharia a Taxa de Conclusão da Graduação (TCG) é apenas de 57% entre 2016 a 2021, ou seja, uma evasão de 43%, tendo muito o que melhorar para estes cursos.

3 Aprendizagem de Máquina

Neste Capítulo, serão discutidos conceitos importantes de aprendizagem de máquina, como viés e variância, tipos de algoritmos e métricas de avaliação.

Este capítulo está organizado da seguinte forma:

- 3.1 **Conceitos**: aborda os principais conceitos de aprendizado de máquina.
- 3.2 **Algoritmos**: aborda os principais algoritmos utilizados nesta previsão.
- 3.3 **Métricas de Avaliação**: conceitos das principais métricas de avaliação utilizadas.

3.1 Conceitos

O matemático especialista em ciências de dados, Clive Humbry, em 2006, cunhou a famosa frase, originalmente em inglês: "*data is the new oil*", comparando os dados ao petróleo e se referindo à importância dos dados para a competitividade econômica dos países no século XX (BAKERTILLY, 2021).

O aprendizado de máquina é o uso de algoritmos para descobrir conhecimento nas bases de dados que ajudem na aplicação de tomada de decisões sobre o futuro (NWANGANGA; CHAPPLE, 2020).

Está inserida no campo da inteligência artificial (IA) que é o segmento da ciência da computação que se concentra na criação de computadores que pensam da maneira como os humanos pensam, na busca por padrões (DSA, 2022).

Diversas são as aplicações de aprendizagem de máquina, como por exemplo: detecção de anomalias; detecção de fraudes bancárias e de seguros; identificação de mensagens de e-mails spam; segmentação de clientes e produtos; carros e drones autônomos; algoritmos de mecanismos de busca; inovações em áreas de segurança da informação; logística; agricultura e saúde, como na detecção de tumores e doenças; na área governamental na detecção e prevenção de crimes, dos desastres naturais e na educação como a previsão de evasão de alunos, entre outras aplicações (GOLLAPUDI, 2016).

O processo de aprendizagem de máquina envolve a análise de dados para identificar padrões e relações, o treinamento de modelos para reconhecer e aplicar esses padrões em novos dados e a validação desses modelos para garantir sua eficácia. Esse processo envolve uma série de conceitos que serão tratados a seguir, como: dados de treino e teste, validação

cruzada, os principais algoritmos utilizados e como ocorre a avalição dos modelos através das métricas de avaliação.

3.1.1 Dados de Treino e Teste

Na fase de modelagem, nos modelos de treinamento supervisionado, os dados rotulados são divididos de forma aleatória em dados de treino e dados de teste. Os dados de treino são apresentados ao algoritmo para que ele aprenda o relacionamento entre as variáveis e crie o modelo. Os dados de teste, por sua vez, são utilizados para avaliar o quanto o algoritmo aprendeu. Ao apresentar os dados de teste ao modelo, as previsões são realizadas com base no que foi aprendido na fase de treinamento. Essas previsões são comparadas com as respostas esperadas para calcular o desempenho do modelo. Uma vez criado e validado, o modelo pode ser utilizado para realizar novas previsões quando for apresentado a novos dados (HACKELING, 2014).

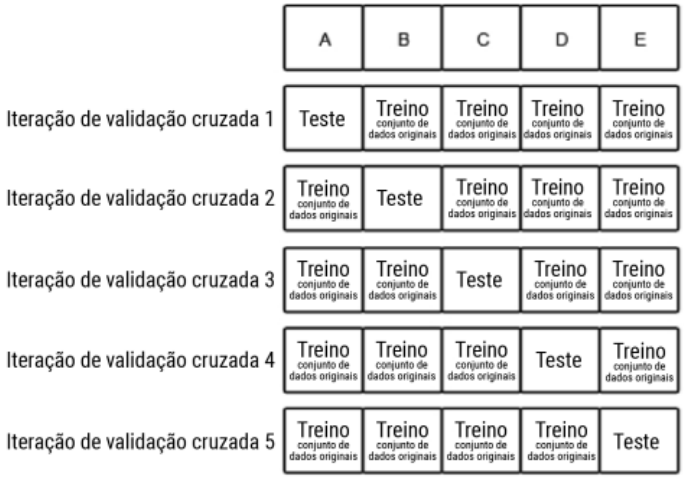
3.1.2 Validação Cruzada

Para que o modelo fique mais robusto e para que a avaliação deste modelo seja mais fidedigna ao resultado de previsão, é utilizada uma técnica chamada de validação cruzada (*cross-validation*) que consiste em dividir o conjunto de dados disponível em x% de "Treino" e y% de "Teste" e escolher k subconjuntos de iterações treino-teste (ou *folds*), o modelo é treinado em n-1 desses subconjuntos e avaliado no subconjunto de teste, esse processo é repetido k vezes (Figura 4), de forma que cada subconjunto seja utilizado uma vez como conjunto como teste e outra como treino, isto faz com que os dados de treino/teste não sejam os mesmos causando apenas um resultado, sendo analisado de forma enviesada. Na validação cruzada, os dados de treino/teste podem ser configurados k (*k-fold*) vezes e cada vez o resultado será diferente, a média dos resultados é por sua vez mais fidedigna do que apenas um resultado na iteração, isso permite configurar os parâmetros do modelo e ajustá-los, melhorando sua previsão e também permite realizar uma comparação de resultados entre diferentes algoritmos. Quando as métricas atingem um patamar estável de resultados, pode-se dizer que o modelo está treinado e pronto para receber os dados de previsão (NWANGANGA; CHAPPLE, 2020). A Figura 4 exemplifica um diagrama de validação cruzada com dados em cinco partições k fold = 5.

3.1.3 Viés e Variância

Viés e Variância são dois conceitos importantes no aprendizado de máquina pois estão diretamente relacionados à capacidade do modelo de generalizar bem para novos dados. Conforme Nwanganga e Chapple (2020, p.20), ao construir um modelo de aprendizado de máquina, ele incluirá algum tipo de erro que pode ser de três formas diferentes:

Figura 4 – **Diagrama exemplo de validação cruzada com cinco partições**



Fonte: Hackeling (2014)

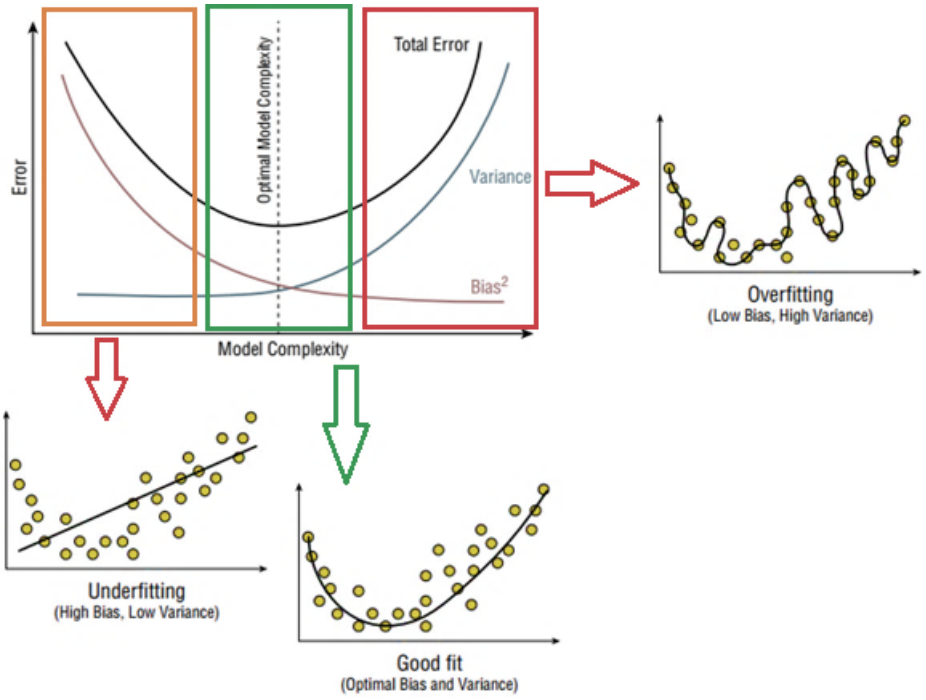
- **Viés (*bias*)** é o tipo de erro que ocorre devido a escolha do modelo de aprendizado de máquina. É o tipo de erro relacionado ao ajuste do modelo escolhido aos dados de treino. Quando o modelo escolhido é incapaz de se ajustar bem ao conjunto de dados, o erro resultante é o viés.
- **Variância (*variance*)** é o tipo de erro usado para avaliar o quão bem o modelo de aprendizado de máquina é representativo dentro do universo destes dados. É a quantidade de erros total do modelo nos dados de teste.
- **Erro irredutível:** também chamado de ruído, ocorre independentemente do algoritmo de aprendizado de máquina e conjunto de dados de treinamento que usamos. É o erro inerente ao problema que estamos tentando resolver.

A Figura 5 abaixo mostra a relação entre viés e variância. Os conceitos de *underfitting, good fit e overfitting* estão relacionados na forma como os modelos se ajustam aos **dados de treino** e posterior resultado nos **dados de teste**.

O conceito de *overfitting* é quando o modelo se ajusta demais aos dados de treino, obtendo baixo viés, e ao confrontar este modelo criado com os dados de teste, o resultado será uma alta variância. Já o conceito contrário é o *underfitting*, quando o modelo se ajusta pouco aos dados de treino (alto viés) porém com os dados de teste apresentam uma baixa variância (NWANGANGA; CHAPPLE, 2020).

Portanto, o melhor resultado é encontrar um equilíbrio entre viés e variância, chamado de *trade-off* viés-variância, para que o modelo seja o mais genérico ao mesmo tempo bem ajustado ao universo dos dados, apresentando baixa variância e baixo viés (*good fit*).

Figura 5 – ***Trade-off* entre viés e variância**



Fonte: adaptado de Nwanganga e Chapple (2020, p.21-22)

Dessa maneira, o desafio do aprendizado de máquina é buscar um modelo que possua baixo viés (modelo bem treinado e ajustado aos dados) e baixa variância (poucos erros nos dados de teste) e para encontrar esse equilíbrio alguns ajustes de hiperparâmetros dos modelos são necessários até obter as melhores métricas na avaliação.

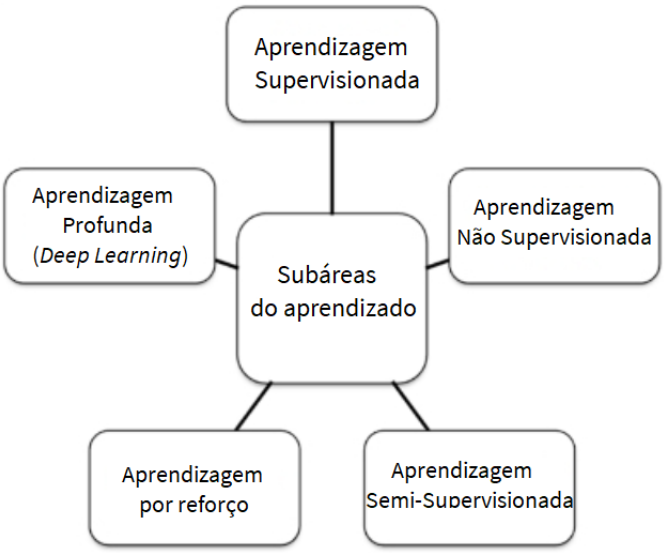
3.1.4 Classificação dos problemas de Aprendizagem de Máquina

Existem algumas formas de classificar as técnicas de aprendizagem de máquina, porém em termos didáticos podemos pensar em como os algoritmos aprendem e o que eles aprendem. Segundo Gollapudi (2016), um dos conceitos-chave em aprendizagem de máquina é como o processo de aprendizado acontece através do método indutivo, segundo a autora essas técnicas subdividem-se em cinco subáreas, conforme Figura 6.

1. **Aprendizagem supervisionada:** A aprendizagem supervisionada cria um modelo através de dados de entrada rotulados que gera a possibilidade de fazer previsões sobre novos dados não rotulados. É neste contexto de aprendizagem que os algoritmos de previsão (preditores) de evasão em educação são utilizados. A aprendizagem supervisionada subdivide-se em problemas de:

    - **Classificação**: predição de respostas categóricas, binárias ou multiclasse, por exemplo: **Evadido/Não Evadido (binária),** Tipo Sanguíneo (multiclasse);

Figura 6 – **Subáreas da aprendizagem de máquina**



Fonte: adaptado de Gollapudi (2016)

- **Regressão**: predição de respostas numéricas, quantitativas contínuas ou discretas, por exemplo: previsão de renda (contínua) ou de idade (discreta).

2. **Aprendizagem não supervisionada**: este tipo de aprendizagem não utiliza rótulos prévios nos dados e busca-se por padrões ocultos, formando agrupamentos por classificações. Alguns usos desse tipo de aprendizado são: Detecção de anomalias; Agrupamento dos dados*(clustering algorithms*); Redução de dimensionalidade das variáveis; entre outros.

3. **Aprendizagem semi-supervisionada:** a aprendizagem semi-supervisionada utiliza tanto dados rotulados quanto não rotulados em conjunto para o melhor aprendizado. É uma forma de aprendizado útil quando é difícil e/ou caro rotular todos os dados disponíveis. A abordagem permite que os modelos usem informações dos dados não rotulados para melhorar a precisão da classificação. Além disso, a aprendizagem semi-supervisionada pode ser útil em casos em que há poucos dados rotulados disponíveis, mas muitos dados não rotulados.

4. **Aprendizagem por reforço:** Aprendizagem por reforço é a aprendizagem que se concentra em maximizar as recompensas do resultado. O algoritmo não recebe a resposta correta mas recebe um sinal de reforço, como recompensa ou punição e realiza uma hipótese através de exemplos e determina se essa hipótese foi boa ou ruim. A aprendizagem por reforço é útil em aplicações onde o agente precisa tomar decisões em ambientes complexos e dinâmicos, como em jogos, robótica, controle de tráfego e muitas outras.

5. **Aprendizagem Profunda (*Deep Learning*)**: O aprendizado profundo é uma subdivisão adicional da aprendizagem de máquina que usa um conjunto complexo de técnicas, conhecidas como redes neurais, para descobrir conhecimento de uma maneira particular (GOLLAPUDI, 2016). Utiliza uma rede de neurônios artificiais (RNA) responsáveis por receber informações de entrada, processá-las e gerar uma saída, é um tipo de aprendizagem de máquina que se inspira no funcionamento do cérebro humano. Nwanganga e Chapple (2020), consideram a aprendizagem profunda como um subcampo altamente especializado do aprendizagem de máquina que é mais comumente usada para análise de imagem, vídeo e som. É uma técnica desenvolvida por vários pesquisadores ao longo do tempo, mais recentemente alguns pesquisadores influentes da área são: Geoffrey Hinton, Yann LeCun, Yoshua Bengio, entre outros (HINTON; OSINDERO; TEH, 2006) (HINTON; SALAKHUTDINOV, 2006) (LECUN; BENGIO; HINTON, 2015). Um exemplo famoso desse tipo de aprendizagem é o sistema *Chat GPT*, e ao questionar ao sistema qual o tipo de algoritmo que ele utiliza em sua construção a resposta foi a seguinte:

   > "Eu sou um modelo de linguagem treinado com base na arquitetura GPT-3.5, que é um modelo de deep learning. Minha capacidade de compreender e produzir linguagem natural é resultado do treinamento com uma grande quantidade de dados de texto, que me permite identificar padrões e relacionamentos complexos em diferentes tipos de texto e gerar respostas coerentes com base no contexto da conversa" (CHATGPT, 2023).

Existem outros autores que podem classificar de forma diferente os processos de aprendizagem de máquina, inclusive separando a Aprendizagem Profunda como um subcampo especialista da Aprendizagem de Máquina e essa por sua vez um subcampo da Inteligência Artificial (NWANGANGA; CHAPPLE, 2020).

Este trabalho utiliza técnicas de **aprendizagem supervisionada** dentro do campo de **classificação** dos dados.

3.2 Algoritmos

A partir da revisão de literatura, foram elencados cinco algoritmos de classificação que são frequentemente utilizados no contexto da previsão da evasão universitária. Os algoritmos são descritos a seguir.

3.2.1 Regressão Logística

A Regressão Logística (também chamada de Regressão *Logit*) (CRAMER, 2003) é comumente utilizada para estimar a probabilidade de uma instância pertencer a uma determinada classe ou não. Para este trabalho como um algoritmo de aprendizagem de máquina, foi utilizada na classificação quando o aluno possa pertencer à classe "Evadido" ou "Formado". Se a probabilidade estimada for maior que 50%, então o modelo prevê que a instância pertence a essa classe (chamada de classe positiva, rotulada como "1" ou "Evadido"), ou então ela prevê que não (isto é, pertence à classe negativa, rotulada "0" ou "Formado"). Isso a transforma em um classificador binário. O resultado da função logística é uma probabilidade entre 0 e 1 representada por uma função derivável sigmoide, em que o eixo y representa a probabilidade e no eixo x a função com suas variáveis preditoras (em conjunto).

O modelo é útil para classificar, pois após ser treinado sua função de custo é ajustada através da máxima verossimilhança e ao receber novas instâncias nos dados de teste, o modelo treinado realiza a previsão dos alunos com o resultado sendo 0 (Formado) ou 1 (Evadido).

3.2.2 KNN – K-Nearest Neighbors

O algoritmo K-Nearest Neighbors (KNN) é uma técnica de aprendizado supervisionado utilizada em problemas de classificação e regressão. Ele é considerado um dos algoritmos mais simples e intuitivos de aprendizado de máquina, é um algoritmo de classificação que atribui uma classe a um novo dado não rotulado com base na classe mais comum dos pontos próximos existentes (SILVERMAN; JONES, 1989).

Primeiramente, os dados já classificados são processados de forma vetorial com distâncias características (interpretadas como coordenadas) na fase de treino, na verdade não se constrói um modelo para este algoritmo, por isso são chamados de *lazy learners* ou *instance-based learners* pois simplesmente identificam as classes na fase de treino (RASCHKA; OLSON, 2015).

Na fase de teste, ao receber um novo dado, este é processado de acordo com suas características mais próximo ou mais distante de uma classe ou outra. Nesta etapa é feito um cálculo de distância entre este ponto e todos os outros pontos, através de uma fórmula

Euclidiana. Então, essas distâncias são colocadas em ordem da menor para a maior e escolhido através de k pontos mais próximos serão avaliados as classes. Por exemplo, k=3 significa que o novo dado será comparado com os 3 vizinhos mais próximos e para escolher a classe desse novo dado é feito uma votação pela quantidade de classes dos vizinhos. Caso os 2 pontos sejam da classe "Evadido" então este novo dado se torna também dessa classe.

Não existe uma fórmula definida para a escolha de qual o valor de k, por isso deve ser feita através de tentativa e erro para a melhor avaliação do modelo. Porém, um k muito pequeno gera o problema de *overffiting* e um k muito grande gera o problema de *underffiting*.

3.2.3 Naive Bayes

Um classificador ainda muito popular para classificação de texto é o Naïve Bayes, ganhou sua popularidade em aplicações de spam de filtragem e-mails. Os classificadores Naïve Bayes são fáceis de implementar, computacionalmente eficientes e tendem a ter um desempenho particularmente bom em conjuntos de dados relativamente pequenos em comparação com outros algoritmos (GéRON, 2019).

O modelo Naive Bayes utiliza o Teorema de Bayes para realizar a classificação criando um modelo probabilístico baseado nos dados de treinamento, ou seja, a partir de exemplos de dados classificados previamente. Com base nesse modelo, ele é capaz de fazer previsões sobre a classe de novos dados.

O classificador é denominado Naive (ingênuo), pois assume que os atributos são condicionalmente independentes, ou seja, a informação de um evento não é informativa sobre nenhum outro. Em alguns casos a frequência de um evento sobre o classificador pode ser zero isso cria o problema de zerar a probabilidade do todo o cálculo, para contornar esse problema utiliza-se o método de Suavização de Laplace (RASCHKA, 2017). Utilizou-se a função *naivebayes* dentro do pacote *Caret* (KUHN, 2019) do *software RStudio* (R CORE TEAM, 2022), esta função dentro deste pacote ajusta automaticamente a função a partir das variáveis fornecidas.

3.2.4 SVM – Support Vector Machine

Uma Máquina de Vetores de Suporte (SVM) é um modelo versátil de Aprendizado de Máquina capaz de realizar classificações lineares ou não lineares, de regressão e até mesmo detecção de *outliers*. Utiliza um limite de decisão linear entre as classes, chamado de **margem**. O objetivo do SVM é encontrar um hiperplano que separe duas classes com a maior margem possível (HEARST et al., 1998).

Para que a margem não se torne tão sensível aos *outliers* o algoritmo permite algumas classificações no "lado" errado da margem e calcula o ajuste desta linha através

de técnicas de validação cruzada permitindo um viés/variância do melhor custo-benefício possível, isto é chamado classificação de **margem suave** (GéRON, 2019).

Nem todos os problemas são linearmente separáveis, podendo alguns serem polinômios complexos (mais de uma dimensão) e para isto a técnica SVM utiliza o truque de Kernel podendo trabalhar com diferentes dimensões no hiperplano, utilizando funções de kernel e encontrando a margem através de Classificadores de Vetores de Suporte sem realmente realizar as transformações do hiperplano, reduzindo o tempo e o custo computacional. A função de transformação de um espaço não linear, truque de kernel, pode ser realizada por Polynomial Kernel no qual possui um parâmetro d (grau) ou *Radial Basis Function* (RBF) Kernel (dimensões infinitas). Para este trabalho o modelo SVM desenvolvido considerou o kernel do tipo RBF (GéRON, 2019).

3.2.5 Árvores de Decisão e *Random Forest*

Floresta Aleatória (*Random Forest*) é um algoritmo de *Ensemble Learning*, ou seja, utiliza a resposta agregada de várias árvores de decisão em conjunto, esta é uma técnica chamada de sabedoria das multidões em que a previsão de um conjunto de previsores será melhor do que com o melhor previsor individual. Para fazer previsões, obtém-se a previsão de todas as árvores individuais e, então, prevê-se a classe com o maior número de votos (BREIMAN, 2001).

O classificador de árvore de decisão usa uma estrutura lógica semelhante a uma árvore para representar a relação entre as variáveis preditoras e a variável dependente. Os resultados potenciais de uma árvore de decisão podem ser discretos (árvore de classificação) ou contínuos (árvore de regressão).

Existem duas implementações mais populares de algoritmos de árvore de decisão que são as Árvores de Classificação e Regressão (CART) – que utiliza a fórmula do índice de Gini e a C5.0 que utiliza a fórmula da Entropia. Ambas as implementações usam uma abordagem semelhante para construção de árvore, conhecido como particionamento recursivo. Essa abordagem divide repetidamente os dados em subconjuntos cada vez menores (IF-THEN) até que alguns critérios de parada sejam atendidos (NWANGANGA; CHAPPLE, 2020).

O critério de parada (profundidade da árvore) geralmente é escolhido para evitar o sobreajuste dos dados (*overfitting*) pois quanto mais profunda for a árvore maior será o sobreajuste aos dados de treino.

3.3 Métricas de Avaliação

Métricas de avaliação são medidas utilizadas para avaliar o desempenho de algoritmos de aprendizagem de máquina. Algumas das principais métricas de avaliação incluem Acurácia, Sensibilidade, Especificidade, Precisão, F1-score, Curva ROC e Área sob a curva ROC (AUC). Cada métrica tem uma interpretação diferente e é adequada para diferentes tipos de problemas e algoritmos.

3.3.1 Matriz de Confusão

Em um problema de previsibilidade de evasão, há duas possíveis respostas (binária): "Evadido" ou "Formado", que são os rótulos que desejamos prever. Além disso, os algoritmos fazem a confrontação da previsão com os dados observáveis, obtendo valores de acerto ou não. Esses dados são classificados em uma tabela 2×2, chamada Matriz de confusão (*Confusion Matrix*), conforme Figura 7, com os seguintes valores:

- **VP (verdadeiro positivo)**: quando o programa previu a evasão e o aluno realmente evadiu;
- **FP (falso positivo)**: quando o programa previu a evasão, mas o aluno formou;
- **FN (falso negativo)**: quando o programa previu a formação, porém o aluno evadiu;
- **VN (verdadeiro negativo)**: quando o programa previu a formação e o aluno realmente se formou.

Figura 7 – **Exemplo de matriz de confusão no *software* R**

```
Confusion Matrix and Statistics

          Real
Previsto  evadido formado
  evadido      VP      FP
  formado      FN      VN
```

Fonte: o autor.

Portanto, a matriz de confusão permite ter um panorama das classificações da previsão em cada algoritmo com os valores absolutos e a partir dela podemos extrair as métricas de avaliação representadas abaixo, com os valores das métricas normalizados de 0 a 1.

3.3.2 Acurácia

A Acurácia mede o percentual de acerto geral do método, é uma técnica básica que indica o percentual de previsão correta. Permite um panorama geral da qualidade do modelo. A Fórmula 3.1 representa o cálculo da Acurácia.

$$Acurácia = \frac{VP + VN}{VP + FN + VN + FP} \tag{3.1}$$

3.3.3 Sensibilidade

Não prever a evasão do aluno e o aluno de fato evadir (Falso Negativo) tem um maior peso do que prever a evasão e o aluno não evadir (Falso Positivo). Portanto, a Sensibilidade é uma métrica muito importante para os estudos da evasão universitária.

A sensibilidade, conhecida também como *Recall* ou *True Positive Rate* (TPR) é uma métrica que avalia a quantidade de verdadeiros positivos (VP) frente ao total de fato positivos observados. Leva em consideração somente o erro de Falso Negativo, a sensibilidade continua alta mesmo com erros de Falsos Positivos.

A Fórmula 3.2 representa o cálculo da Sensibilidade.

$$Sensibilidade = \frac{VP}{VP + FN} \tag{3.2}$$

3.3.4 Especificidade

A Especificidade, *Specificity* ou *True Negative Rate* (TNR) é uma medida que avalia a capacidade de detectar os resultados de fato negativos (VN) frente ao total de negativos. No caso de estudos de evasão, é a habilidade do modelo em prever os casos formação. Na verdade, o foco está mais em identificar a permanência detectando corretamente esses alunos do que a evasão. A Fórmula 3.3 representa o cálculo da Especificidade.

$$Especificidade = \frac{VN}{VN + FP} \tag{3.3}$$

3.3.5 Precisão

A precisão é uma medida de classificação para saber a quantidade de verdadeiros positivos (VP) frente ao total previsto de positivos. É a métrica que calcula dos alunos previstos a evadir, quantos realmente evadiram. A resposta a essa pergunta é dada pela métrica da precisão.

No entanto, esta métrica não olha para o lado da previsão negativa, podemos ter um alto valor de Falsos Negativos (FN) e a precisão continuar sendo alta. A Fórmula 3.4 representa o cálculo da Precisão.

$$Precisão = \frac{VP}{VP + FP} \tag{3.4}$$

3.3.6 F1-Score

A Medida F1, conhecida como: *F1-measure, F1-score*, Métrica F1. A Medida F1 é usada para equilibrar os erros FP e FN das métricas de Sensibilidade e Precisão, pois realiza uma média harmônica entre ambas. Uma característica desta medida é que se uma das métricas for baixa, a medida F1 também será baixa. Portanto, para que esta medida seja alta, tanto a Precisão como a Sensibilidade devem ser altas. Isso implica que o modelo consegue classificar os verdadeiros positivos (VP) de forma eficiente gerando um número de FP e FN baixo. A Fórmula 3.5 representa o cálculo do *F1 Score*.

$$Precisão = 2 \times \frac{Sensibilidade \times Precisão}{Sensibilidade + Precisão} \tag{3.5}$$

A relação das métricas com a Matriz de Confusão pode ser observada na Figura 13.

Figura 8 – **Métricas de Avaliação e a relação com a Matriz de Confusão**

		Predicted		**F1-score** $= 2 \times \frac{Recall \times Precision}{Recall + Precision}$
		Negative (0)	Positive (1)	
Actual	Negative (0)	**True Negative** **TN**	**False Positive** **FP** (Type I error)	**Specificity** $= \frac{TN}{TN + FP}$
	Positive (1)	**False Negative** **FN** (Type II error)	**True Positive** **TP**	**Recall, Sensitivity,** True positive rate (**TPR**) $= \frac{TP}{TP + FN}$
			Precision, Positive predictive value (**PPV**) $= \frac{TP}{TP + FP}$	**Accuracy** $= \frac{TP + TN}{TP + TN + FP + FN}$

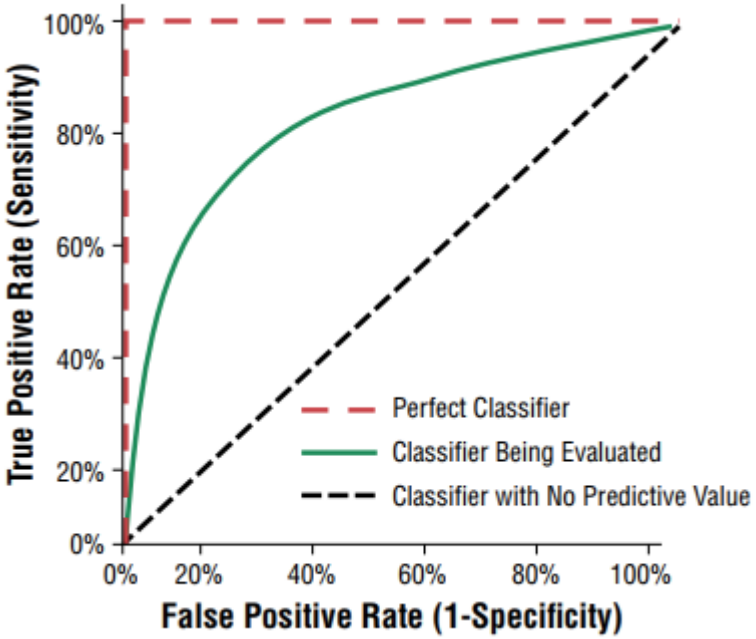
Fonte: o autor.

3.3.7 Curva ROC/AUC

A *Receiver Operating Characteristic* (ROC) e a *Area Under the Curve* (AUC) são métricas comumente utilizadas para avaliar a performance de modelos de classificação. A curva ROC é representada por um gráfico que relaciona a Sensibilidade (TPR - *True Positive Rate*) com a taxa de falso positivo (FPR - *False Positive Rate*), como ilustrado na Figura 9. Essa curva visualiza o compromisso entre falsos positivos e verdadeiros positivos, mostrando a escolha do limiar de classificação (*threshold*). A curva ROC varia de 0 a 1, e quanto mais próxima estiver do canto superior esquerdo, melhor será a capacidade preditiva do modelo. Uma curva perfeita teria TPR igual a 1 e FPR igual a 0.

A *Area Under the Curve* (AUC) é uma medida de desempenho que representa a habilidade do modelo de distinguir entre as classes. Quanto mais próxima de 1 for a AUC, melhor será a performance do modelo. Uma das vantagens da AUC é que ela não é sensível a desbalanceamento de classes, ao contrário da acurácia, por exemplo. A AUC fornece uma medida agregada da qualidade de classificação em diferentes pontos de corte, sendo uma métrica amplamente utilizada na avaliação de modelos de classificação (NWANGANGA; CHAPPLE, 2020).

Figura 9 – **Curva ROC/AUC**



Fonte: Nwanganga e Chapple (2020, p.334)

4 Revisão Sistemática

Uma revisão sistemática de literatura foi feita com o objetivo de avaliar quais os algoritmos mais utilizados dentro da área de **aprendizagem de máquina** no contexto da evasão universitária. Foram selecionados 52 trabalhos utilizando-se para isso duas bases de dados *Scopus e Web of Science* com alguns critérios de elegibilidade, conforme Quadro 2. A busca foi realizada no dia 04 de abril de 2022. Portanto, os artigos foram selecionados até esta data, conforme expressões (*query*) na Tabela 2 abaixo.

Tabela 2 – **Expressões utilizadas na busca das bases de dados WoS e Scopus**

WoS	**Scopus**
TS=(("students dropout"or "student dropout") AND ("university"OR "higher education") AND ("data mining"OR "machine learning") AND (“prediction”)) OR AB=(("students dropout"or "student dropout") AND ("university"OR "higher education") AND ("data mining"OR "machine learning") AND (“prediction”)) OR AK=(("students dropout"or "student dropout") AND ("university"OR "higher education") AND ("data mining"OR "machine learning") AND (“prediction”))	TITLE-ABS-KEY-AUTH(("students dropout"or "student dropout") AND ("university"OR "higher education") AND ("data mining"OR "machine learning") AND (“prediction”))

Fonte: o autor

A revisão sistemática de literatura foi realizada buscando compreender os principais conceitos, abordagens e metodologias empregadas sobre evasão universitária e aprendizagem de máquina. A revisão sistemática teve como objetivo identificar as principais tendências, metodologias e resultados de estudos recentes sobre o tema, a fim de embasar e direcionar as análises e discussões acerca da evasão universitária e as possibilidades de utilização de algoritmos de aprendizagem de máquina.

Quadro 2 – **Quadro dos critérios de elegibilidade**

Critério	Exclusão	Inclusão
Nível de estudo	Ensino fundamental, médio, pós-graduação e técnico	Universitário
Modalidade	À distância (EaD)	Presencial (*on campus)*
Técnica	Estudos teóricos, *Reviews*, técnicas de não previsibilidade	Previsão através de *Machine Learning*
Foco	Permanência ou Retenção	Evasão
Metodologia	Sem métricas ou a metodologia não foi explicada	Com métricas, metodologia e base de dados explicados

Fonte: Do Autor

A revisão focou principalmente em dois aspectos:

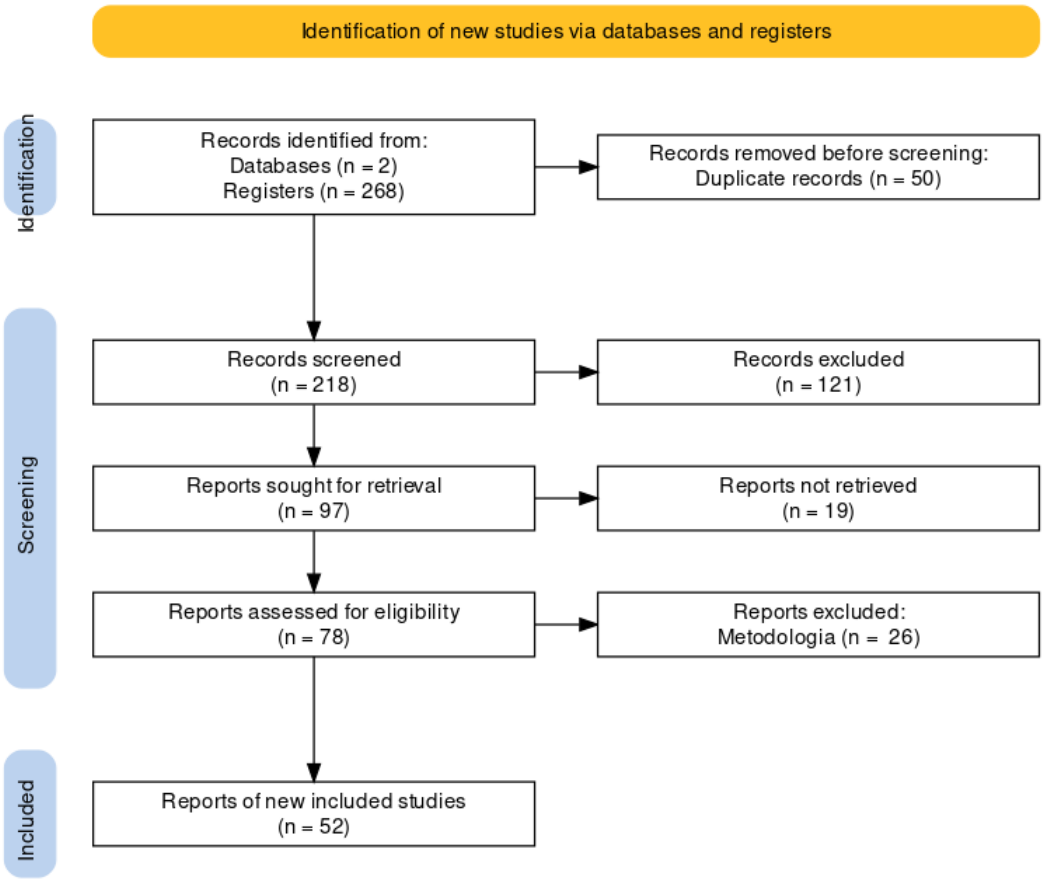
1. Tipos de Algoritmos mais utilizados na comparação;
2. Principais métricas de avaliação.

Na seleção dos artigos, foi utilizado o fluxograma do protocolo de revisão sistemática da *Preferred Reporting Items for Systematic reviews and Meta-Analyses* (PRISMA) (PAGE, 2021). Na Identificação (*Identification*) foram removidas 50 duplicatas através do programa EndNoteWeb® restando 218 artigos para a Triagem de Títulos e Resumos (*Screening*). Em seguida, removeu-se 121 artigos pelos critérios de elegibilidade (vide Quadro 2), sobrando 97 artigos (*Reports sought for retrieval*). Porém, 19 desses não foi possível obter o acesso de texto completo. Ainda assim, ficaram 78 artigos para a leitura completa e extração dos dados. Pelo critério da análise da Metodologia e dos Resultados dos artigos, foram excluídos 26 artigos, finalmente sobrando 52 artigos de referência. A Figura 10 representa esse Processo de seleção.

4.1 Resultados da Revisão Sistemática

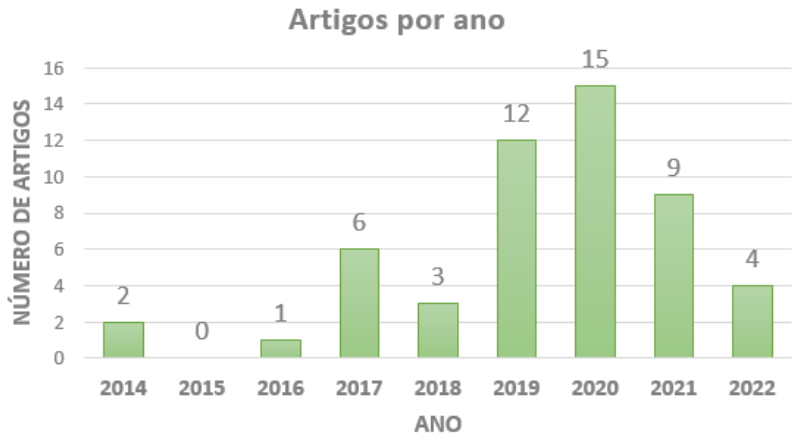
Os resultados mostram um crescimento de artigos na área de aprendizagem de máquina a partir de 2014 com maior número de artigos em 2019, 2020 e 2021, conforme Figura 11 . Os artigos selecionados apresentaram uma evolução maior entre os anos de 2017 e 2021, apenas três artigos foram selecionados em 2022 pois até a data da busca (abril 2022) o ano não estava completo. Comparando com outras revisões do tema, é possível observar que a partir de 2017 também nesses estudos se obteve um crescimento, demonstrando que é uma área recente de estudos e ainda com muito potencial a ser explorada (AGRUSTI; BONAVOLONTà; MEZZINI, 2019) (OLIVEIRA et al., 2021) (LIZ-DOMíNGUEZ et al., 2019) (NAMOUN; ALSHANQITI, 2020).

Figura 10 – **Fluxograma de seleção PRISMA**



Fonte: o autor.

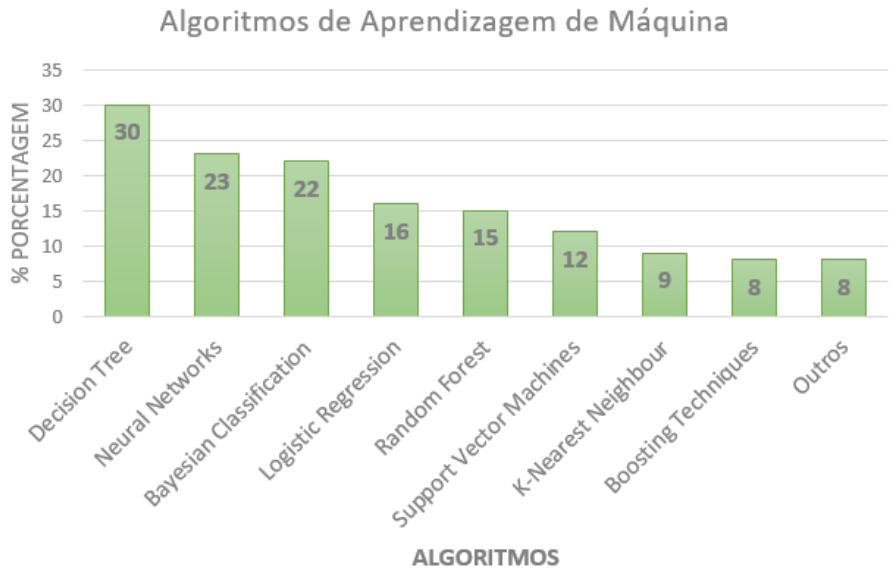
Figura 11 – **Artigos por ano**



Fonte: O autor.

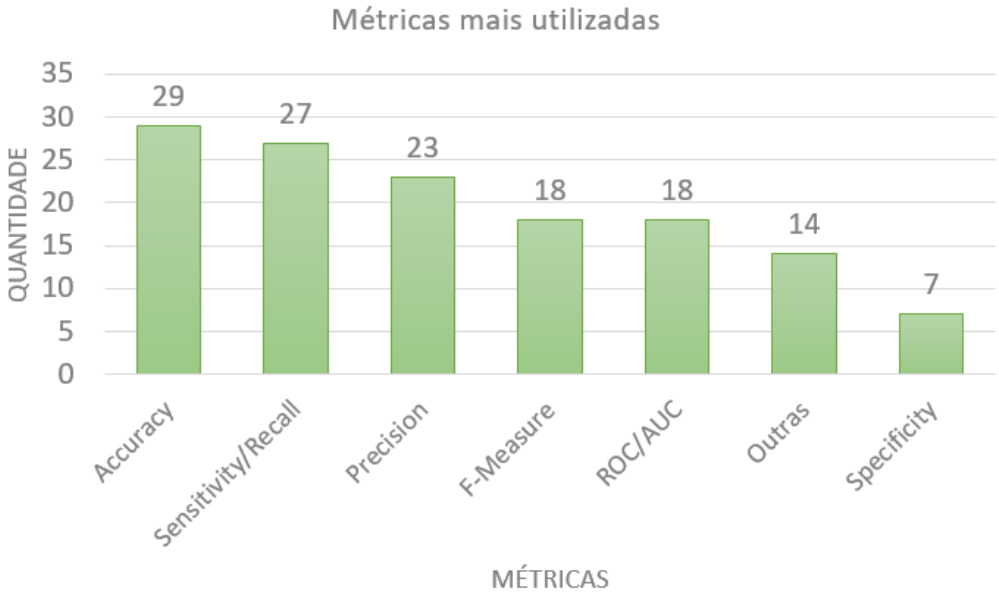
Em relação aos algoritmos de aprendizagem de máquina mais utilizados, estão representados na Figura 12.

Figura 12 – **Algoritmos de Aprendizagem de Máquina mais utilizados**



Fonte: O autor.

Figura 13 – **Métricas mais utilizadas nos artigos selecionados**



Fonte: O autor.

Árvores de decisão, Redes Neurais e Classificações Bayesianas (*Naive Bayes*, por exemplo) são os algoritmos que mais apareceram nos artigos selecionados (52%). Outros algoritmos também aparecem, como Regressão Logística, *Random Forest*, *Support Vector Machines* e *K-Nearest Neighbour*.

Já em relação às métricas de avaliação mais utilizadas, conforme Figura 13, é possível observar que a Acurácia é a métrica mais utilizada. A segunda métrica é a Sensibilidade pois representa os verdadeiros positivos (previstos e evadidos) frente ao total de positivos observados. A Precisão também é utilizada para verificar a previsibilidade do modelo, assim como as medidas *F-measure* e curva ROC. Apesar das medidas *F-measure* e curva ROC aparecerem em quarto e quinto lugar, são medidas mais complexas e mais robustas para avaliar o modelo frente a dados desbalanceados (diferenças significativas entre as classes).

A Especificidade aparece em último lugar, pois o foco dos artigos são nos alunos que evadiram (positivos), essa é uma métrica interessante para analisar os alunos formados, no caso de verdadeiros negativos.

Dentro dos artigos selecionados, 11 artigos apresentaram uma metodologia semelhante a este trabalho, comparando diferentes algoritmos de aprendizagem de máquina em diferentes momentos da trajetória acadêmica dos alunos. Desta forma, os resultados destes artigos foram compilados conforme Tabela 3.

Em resumo, os artigos selecionados mostram que o pré-processamento dos dados com reduções de dimensionalidade e a otimização dos parâmetros dos algoritmos são realizadas para um melhor desempenho na previsão. Alguns autores também fazem uso de diversas ferramentas nesse processo, como no caso de balanceamento dos dados para uma melhor performance. Diversos algoritmos são utilizados na comparação e de diferentes formas, com diferentes momentos no curso dos alunos.

Em geral, nos artigos selecionados, a métrica da Sensibilidade ficou mais alta na medida que possuem mais dados acadêmicos, ou seja, a eficácia da previsão de evasão se torna mais assertiva na medida que os alunos avançam em seus cursos.

Estes resultados da Revisão Sistemática foram importantes para compreender a metodologia utilizada na previsão de evasão universitária por aprendizagem de máquina, selecionar os algoritmos de previsão, conhecer os métodos de seleção de variáveis independentes e também conhecer as diferentes formas de ajustes de parâmetros. Os resultados dos artigos serão comparados com os resultados deste trabalho nos próximos capítulos.

Tabela 3 – **Tabela comparativa dos estudos selecionados com metodologia semelhante**

Autores	Contexto da Previsão de Evasão	Metodologia	Algoritmos de comparação	Resultados
Costa et al. (2017)	Pela disciplina de Introdução à Programação do curso de Ciência da Computação de uma universidade pública brasileira em duas modalidades de ensino EaD (ano 2013) e Presencial (ano 2014)	1. EaD: previsão da evasão nas cinco primeiras semanas e após o primeiro exame e 2. Presencial: previsão da evasão nas três primeiras semanas e após o primeiro exame (Com apenas dados acadêmicos)	1. Neural Networks; 2. Decision Tree, 3. Suport Vector Machine; 4. Naive Bayes (Comparação com pré-processamento e hiperparametrização dos algoritmos)	Os resultados mostraram que o pré-processamento de dados e a otimização de parâmetros influenciam na melhora do resultado. O algoritmo Suport Vector Machine apresentou o melhor resultado com F-measure de 92% EaD e 83% Presencial após o primeiro exame.
Santos et al. (2019)	Entre os cursos de Ciência da Computação (CC), Sistemas da Informação (SI) e Engenharia da Computação (EC) na UFS(Sergipe) com dados de 2010 a 2018	1. Comparação entre os cursos do primeiro ao sexto semestre	1.Decision Tree; 2.KNN; 3.Neural Networks; 4.Suport Vector Machine; 5.Naive Bayes; 6.Random Forest	Os melhores algoritmos foram Decision Tree, Random Forest e Suport Vector Machine, a média da acurácia no curso de CC foi de 66% (Decision Tree e Random Forest), no curso de SI foi de 70% (Random Forest, Suport Vector Machinee Decision Tree) e 72% no curso de EC (Decision Tree).
Hannaford, Cheng e Kunes-Connell (2021)	No curso de enfermagem de uma faculdade privada de Midwest (USA) com dados de 2004 a 2012.	1. Início do primeiro ano (somente demográficos e pré-matrícula), 2. Início do segundo, terceiro, quarto anos, 3. Fim do sexto ano, com dados demográficos, pré-matrícula e de desempenho acadêmico (GPA).	1. Decision Tree (C5.0); 2. Random Forest; 3. XGBoost; 4. Neural Networks; 5. Suport Vector Machine; 6. Naive Bayes; 7. KNN; 8. Logistic Regression	Os resultados mostraram na média uma baixa Sensibilidade no início do curso e no início do primeiro ano (20% e 50%, respectivamente). Porém, uma Acurácia de 73,5% no primeiro ano e 93,5% no terceiro ano com Random Forest.
Fernandez-Garcia et al. (2021)	Nos cursos de engenharia de uma faculdade pública da Espanha com dados de 2012 a 2019	1. Início do primeiro semestre (somente demográficos e pré-matrícula) ; 2. No final do primeiro, segundo, terceiro e quarto semestres (somente com dados de desempenho acadêmico)	1.Gboost; 2. Random Forest; 3.Support Vector Machine; 4. Mix Ensemble	A maior Sensibilidade no modelo de pré-matrícula foi de GBoost com 72,34% e no fim do terceiro semestre com 88,46% com Random Forest
Fernández-Martín et al. (2018)	Cursos do Instituto Tecnológico da Costa Rica (ITCR) dos anos de 2011 a 2013 em quatro sedes diferentes	1. Primeiro semestre com dados sociodemográficos e do curso escolhido; 2. Primeiro semestre com dados de benefícios estudantis e histórico escolar; 3. Fim do primeiro ciclo letivo com todos os dados anteriores mais o de desempenho acadêmico	1. Random Forest; 2. Boosted Trees; 3. Decision Tree; 4. Naive Bayes; 5. Support Vector Machine; 6. Logistic Regression	Os resultados mostram um Sensibilidade baixa nos algoritmos entre 31% a 41% e uma Especificidade alta entre 95% a 98% sendo o algoritmo Random Forest mostrando o melhor desempenho nas métricas ao final do primeiro ciclo.

Continua...

Tabela 3 – ***Tabela comparativa dos estudos selecionados com metodologia semelhante na presente dissertação***

Autores	Objetivo de Prever a Evasão	Metodologia	Algoritmos de comparação	Resultados
Costa et al. (2020)	Curso de Ciência da Computação da UFPel (Brasil) com dados de 2000 a 2020	1. Do curso de CC do primeiro ao terceiro semestre com dados pessoais e acadêmicos	1. Decision Tree; 2. Logistic Regression; 3. Random Forest	O algoritmo com melhores resultados foi Random Forest, porém a diferença entre Logistic Regression e Decision Tree não foi estatisticamente significativa. Atingiu uma Acurácia de 91% e uma Sensibilidade também de 91% e Precisão de 95%.
Huo et al. (2020)	Curso não-tradicionais em faculdades dos Estados Unidos incluindo 24770 registros de 1480 instituições	1. Fim do primeiro ano; 2. Fim do terceiro ano; Com dados demográficos, pré-matrícula, de trabalho, socioeconômicos	1. XGBoost ; 2. Logistic Regression; Comparação com redução de dimensionalidade	A redução de dimensionalidade mostrou-se útil nos resultados. A Sensibilidade atingiu 71,40% e uma Acurácia de 79,82% com modelo XGBoost
Naseem, Chaudhary e Sharma (2022)	Curso de Ciência da Computação de uma universidade do sul do Pacífico com dados de 2013 a 2017	1. Pré-matrícula; 2. Fim do primeiro semestre; 3. Fim do segundo semestre; Com dados demográficos, socioeconômicos, peRandom Forestormance acadêmica e presença online	1. Random Forest; 2. Decision Tree; 3. Naive Bayes; 4. Logistic Regression; 5.KNN	Os resultados mostraram o algoritmo Logistic Regression com melhor perfomance ao fim do terceiro semestre sendo uma Acurácia de 80,3%e uma Sensibilidade de 86%
Dorris, Swann e Ivy (2021)	Cursos de Engenharia numa coorte de 2014 numa faculdade pública	1. Primeiro semestre; 2. Segundo semestre; 3. Terceiro semestre	1. Logistic Regression	O algoritmo apresentou uma Sensibilidade de 63% no primeiro semestre, 67% no segundo e 69% no terceiro semestre.
Kemper, Vorhoff e Wigger (2020)	Curso de Engenharia Industrial da Karlsruhe Insitute of Technology (KIT) Alemanha com dados de 2007 a 2012	1. Primeiro semestre; 2. Segundo semestre; 3. Terceiro semestre; Comparativo entre dados balanceados e desbalanceados	1. Logistic Regression; 2. Decision Tree	Os algoritmos apresentaram melhor perfomance com dados do Terceiro semestre e de forma balanceada. A Sensibilidade da Logistic Regression em 86,2% e Decision Tree em 89,6% com uma Acurácia igual para os dois de 91,8%.
Alvarez, Callejas e Griol (2020)	Curso de Ciência da Computação de todas as províncias de Cuba entre 2013 e 2014	1. Pré-matrícula; 2. Primeiro semestre; 3. Primeiro ano; Com dados pré-matrícula e de perfomance acadêmica em diferentes disciplinas do curso	1. Decision Tree (J48); 2. Neural Networks (MLP)	Os resultados de pré-matrícula foram melhores com Decision Tree tendo uma Acurácia de 59,87% e Sensibilidade de 30,9%. No segundo semestre a NN apresentou uma Sensibilidade de 74,7% e no primeiro ano a NN apresentou uma Acurácia de 96,49% e uma Sensibilidade de 79,8%.

Fonte: o autor

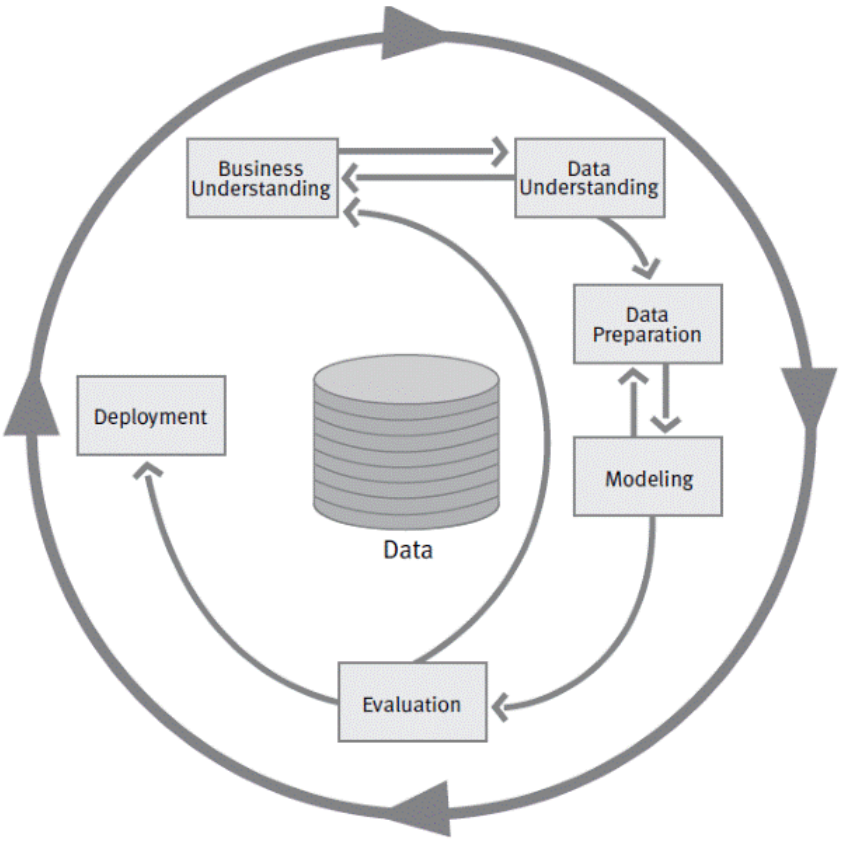
5 Metodologia

O capítulo 5 está subdividido entre a conceptualização da metodologia utilizada *Cross-Industry Standard Process for Data Mining* (CRISP-DM) e aplicação desta metodologia na mineração de dados em educação no presente trabalho.

5.1 Conceito de CRISP-DM

Cross Industry Standard Process for Data Mining (CRISP-DM) é um modelo de processo padrão aberto que descreve abordagens comuns para aprendizagem de máquina. Ele fornece um processo passo a passo e orientações para conduzir projetos de mineração de dados e aprendizagem de máquina, desde a compreensão do problema até a implantação do modelo (CHAPMAN et al., 2000). O CRISP-DM é composto por seis fases: **Entendimento do Negócio, Compreensão dos Dados, Preparação dos Dados, Modelagem, Avaliação e Aplicação** (Figura 14). De acordo com Chapman et al. (2000):

Figura 14 – **Etapas do Processo CRISP-DM**



Fonte: Chapman et al. (2000)

1. **Entendimento do negócio** – Nessa fase, os objetivos do projeto são claramente definidos, juntamente com as questões a serem respondidas e as hipóteses a serem testadas. Também é importante identificar as fontes de dados disponíveis e as limitações desses dados.

2. **Compreensão dos dados** - A fase de compreensão começa com uma coleta inicial de dados e prossegue com as atividades de familiarização. O objetivo é explorar e entender os dados disponíveis, identificando padrões, tendências e anomalias que possam influenciar na construção do modelo de aprendizado de máquina. São utilizadas diversas técnicas de análise, como estatística descritiva, visualização de dados, para identificar e compreender os padrões e relacionamentos existentes entre as variáveis.

3. **Preparação de dados** - A fase de manipulação dos dados consiste em coletar, limpar, integrar, selecionar e transformar os dados brutos em dados úteis para a análise. Essa etapa é crucial, pois a qualidade dos dados influencia diretamente na qualidade dos resultados finais. Além disso, é importante verificar a necessidade de limpeza dos dados, como remover valores ausentes, corrigir erros e padronizar formatos. Também são aplicadas técnicas para reduzir a dimensionalidade dos dados ou seja: seleção de variáveis.

4. **Modelagem** – Esta etapa envolve a implementação de algoritmos de mineração de dados para realizar análises preditivas, na qual as técnicas de aprendizado de máquina são aplicadas aos dados pré-processados. Nessa etapa, são selecionados os algoritmos mais adequados para o problema e o conjunto de dados disponíveis. A partir da escolha dos algoritmos, é feita a implementação do modelo, ajuste dos parâmetros e a realização de treinamento do modelo com os dados de treinamento.

5. **Avaliação** – Fase que envolve a análise do desempenho do(s) modelo(s) através de métricas de avaliação e para garantir que o objetivo do negócio foi atendido e avaliar a sua qualidade. Para isso, podem ser utilizadas diversas métricas de avaliação, como Acurácia, Precisão, Sensibilidade, F1-Score, entre outras.

6. **Aplicação** - Esta etapa envolve a implementação dos modelos no negócio para ser aplicado na prática, é a fase onde o objetivo é implementar o modelo de aprendizado de máquina em um ambiente real e também são realizados testes finais e ajustes necessários para garantir que o modelo esteja pronto para ser utilizado.

5.2 Aplicação da Metodologia CRISP-DM na Predição da Evasão dos cursos do CTC-UFSC

5.2.1 Entendimento do Negócio

Os dados da Universidade são alimentados pelas secretarias dos cursos através do Controle Acadêmico de Graduação (CAGR), compilados pelo Departamento de Administração Escolar (DAE) e mantidos pela Superintendência de Governança Eletrônica e Tecnologia da Informação (SeTIC). As informações são geradas através desses dados pela Secretaria de Planejamento (SEPLAN) na qual estão vinculados o Departamento de Gestão da Informação (DGPI) e a Coordenadoria de Gestão Estratégica (CGE) que além de outras coisas publicam alguns relatórios como: Plano Anual, Plano de Desenvolvimento Institucional (PDI), Relatório de Atividades, Boletim de Dados, Relatório de Gestão, UFSC em Números, entre outros documentos. Na COPERVE são armazenados os dados de pré-matrícula dos alunos, dos vestibulares e do SISU.

Este trabalho é um problema de **Classificação de Aprendizagem** de Máquina através de uma análise preditiva que busca identificar os alunos que estão propensos à evasão acadêmica nos cursos de engenharia do CTC/UFSC.

Alguns termos que serão usados a partir desse momento e devem ser padronizados:

1. **Variável resposta**: classificação final do aluno, é a variável em estudo que é afetada por uma ou mais variáveis independentes. É a variável de interesse na previsão: alunos Evadidos.

2. **Classe**: é o rótulo em que o aluno recebe de acordo com sua situação acadêmica, podendo ser de duas classes **"Evadido"** ou **"Formado"**.

3. **Variável independente**: são as características ou atributos que descrevem os dados de entrada usados para treinar um algoritmo de aprendizagem de máquina e fazer as classificações e as previsões de evasão em novos dados. Ex.: sexo, idade, nota de Matemática no Vestibular, Indice de Aproveitamento Acadêmico (IAA), etc.

4. **Base de Dados**: coleção de dados estruturados (tabelados) com as variáveis independentes e a variavél dependente (em colunas) juntamente com as observações ou registros dos alunos (em linhas). A base de dados deve ser pré-processada antes da aplicação dos algoritmos de aprendizagem de máquina.

5. **Dados de Treino**: são usados para ajustar o modelo e definir os parâmetros dos algoritmos.

6. **Dados de Teste**: avaliam a performance do modelo por meio de métricas em dados que não foram usados durante o treinamento.

7. **Dados de Previsão**: são os dados usados para fazer previsões com o modelo já treinado e validado.

8. **Algoritmo**: conjunto de procedimentos e técnicas que permite que o computador aprenda a partir dos dados de treino e, a partir disso faça previsões.

9. **Modelo**: resultado relativo ao treino/teste do algoritmo de aprendizagem de máquina no comportamento de previsão de evasão para uma determinada fase do aluno:

   a) Pré-Matrícula

   b) Primeiro Semestre

   c) Terceiro Semestre

5.2.2 Compreensão dos Dados

Os dados foram fornecidos em arquivo *.csv* e carregados no *R Studio* (R CORE TEAM, 2022) para as próximas etaps. Dois arquivos principais foram utilizados, para fins de nomenclatura serão chamados de Arquivo 1 e Arquivo 2, que ao final da preparação dos dados resultou em 4394 registros ao total.

1. **Arquivo 1** (dados pré-matrícula): dados socioeconômicos e de desempenho no vertibular. Os dados do **Arquivo 1** continham as seguintes variáveis:

   - **ID**: identificador anonimizado do aluno;
   - **Vestibular**: os vestibulares de 2001 a 2020;
   - **Código do curso**: código dos cursos;
   - **Notas vestibular**: notas das seguintes disciplinas: Biologia, Geografia, Matemática, Língua Estrangeira, Português, Física, História, Química, Ciências Humanas e Sociais;
   - **Chamada**: variável binária categorizada em: 1ª Chamada do Vestibular e Outras Chamadas, em que são as chamadas no respectivo vestibular realizado pelo aluno.
   - **Questionário socioeconômico**: 34 perguntas (Descrição na Tabela 4 em que o candidato deve responder no momento da inscrição do vestibular. O Questionário completo e as opções de resposta podem ser visualizados no Anexo A e dizem respeito desde as informações socioeconômicas do candidato;

Tabela 4 – **Descrição das 34 Perguntas do Questionário Sócio-Econômico**

Questão	Descrição
q1	ESTADO CIVIL
q2	UNIDADE DA FEDERAÇÃO EM QUE VOCÊ RESIDE
q3	UNIDADE DA FEDERAÇÃO EM QUE VOCÊ CONCLUIU O ENSINO FUNDAMENTAL
q4	TIPO DE ESTABELECIMENTO ONDE VOCÊ CURSOU O ENSINO FUNDAMENTAL
q5	UNIDADE DA FEDERAÇÃO EM QUE VOCÊ CONCLUIU OU CONCLUIRÁ O ENSINO MÉDIO
q6	TIPO DE CURSO DE ENSINO MÉDIO QUE VOCÊ CONCLUIU OU CONCLUIRÁ
q7	TIPO DE ESTABELECIMENTO ONDE VOCÊ CURSOU O ENSINO MÉDIO
q8	TURNO EM QUE VOCÊ CURSOU O ENSINO MÉDIO
q9	MARQUE A PRINCIPAL FONTE DE INFORMAÇÃO PELA QUAL VOCÊ TOMOU CONHECIMENTO DO CONCURSO VESTIBULAR DA UFSC
q10	FREQUENTOU OU FREQUENTA CURSO PRÉ-VESTIBULAR
q11	PRINCIPAL MOTIVO QUE O LEVOU A NÃO CURSAR PRÉ-VESTIBULAR
q12	NÚMERO DE VEZES QUE VOCÊ PRESTOU VESTIBULAR PARA A UFSC
q13	PRINCIPAL MOTIVO PARA ESCOLHA DE SUA 1ª OPÇÃO
q14	ASSINALE O QUE VOCÊ ESPERA OBTER NUM CURSO SUPERIOR
q15	CONHECE AS ATIVIDADES QUE DEVERÁ DESENVOLVER NA PROFISSÃO ESCOLHIDA EM 1ª OPÇÃO
q16	INCLUINDO SOMENTE OS QUE MORAM NA SUA CASA, INCLUSIVE VOCÊ, INFORME O NÚMERO DE PESSOAS QUE COMPÕEM A SUA FAMÍLIA
q17	SOME OS SAL. BRUTOS, SEM DEDUÇÕES, DAS PESSOAS DE SEU GRUPO FAM. QUE TRABALHAM, INCLUSIVE O SEU, INDIQUE A RENDA BRUTA
q18	NÍVEL DE INSTRUÇÃO DE SEU PAI
q19	NÍVEL DE INSTRUÇÃO DE SUA MÃE
q20	INDIQUE O PRINCIPAL RESPONSÁVEL PELO SUSTENTO DA SUA FAMÍLIA
q21	PRINCIPAL OCUPAÇÃO DO RESPONSÁVEL PELO SUSTENTO DA SUA FAMÍLIA
q22	IDADE COM QUE COMEÇOU A EXERCER A ATIVIDADE REMUNERADA
q23	SUA OCUPAÇÃO
q24	MARQUE O PRINCIPAL MEIO DE COMUNICAÇÃO QUE VOCÊ UTILIZA PARA SE MANTER INFORMADO SOBRE OS ACONTECIMENTO ATUAIS
q25	POSSUI COMPUTADOR EM SUA RESIDÊNCIA
q26	USA COMPUTADOR
q27	MEIO DE TRANSPORTE QUE VOCÊ MAIS UTILIZA
q28	INICIOU ALGUM CURSO SUPERIOR
q29	INSTITUIÇÃO NA QUAL INICIOU ALGUM CURSO SUPERIOR
q30	INFORME O CURSO SUPERIOR JÁ INICIADO
q31	ACREDITA QUE ORIENTAÇÃO VOCACIONAL AUXILIARIA NA ESCOLHA DE SUA OPÇÃO
q32	DOS ITENS ABAIXO, ASSINALE SUA PREFERÊNCIA
q33	INDIQUE SEU ESPORTE PREDILETO
q34	MOTIVO PRINCIPAL QUE O LEVOU A OPTAR PELO VESTIBULAR DA UFSC

Fonte: COPERVE-UFSC

2. **Arquivo 2** (dados pós-matrícula): dados demográficos e de desempenho acadêmico. Os dados do **Arquivo 2** continham as seguintes variáveis:

   - **ID**: identificador anonimizado do aluno;
   - **Situação**: Variável dependente do estudo. A "situação do aluno" corresponderá à variável resposta no presente trabalho, a qual, na etapa de preparação dos

dados, será reorganizada de modo que os alunos sejam classificados como "Evadido" ou "Formado";

- **Código do curso**: código do curso na opção do Vestibular;
- **Semestre**: do primeiro ao último semestre da Situação do aluno e informa as suas notas em cada período. Usado para filtrar os Modelos do Primeiro e Terceiro Semestres.
- **Sexo**: sexo do aluno em que estava representado por M = masculino e F = feminino;
- **Idade**: idade calculada no ano de ingresso.
- **IAA - Índice de Aproveitamento Acumulado**: O IAA mede o rendimento dos alunos e é calculado é cumulativamente em cada semestre, representado pelo quociente entre o somatório de pontos obtidos e a carga horária matriculada. Entende-se por pontos obtidos o somatório dos produtos das notas pelas cargas horárias matriculadas. Possui uma escala de 0,0 a 10,0.
- **FI**: Frequência Insuficiente, é quando o aluno reprova na disciplina por falta (abaixo de 75% de frequência) conforme Resolução nº 17/CUn/1997 da UFSC.

5.2.3 Preparação dos Dados

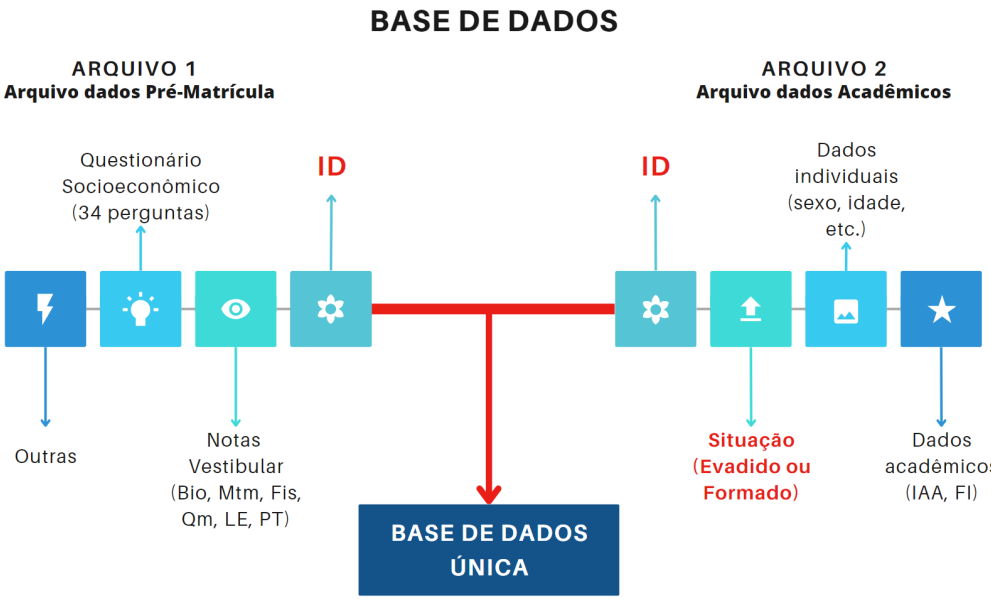
A fase de preparação de dados abrange todas as atividades para construir o conjunto de dados finais a partir dos dados brutos iniciais. Tarefas de preparação de dados são susceptíveis de serem realizadas várias vezes e não necessariamente na ordem prescrita. As tarefas incluem etapas de **Seleção, Limpeza, Construção, Transformação, Integração e Formatação de dados** (CHAPMAN et al., 2000).

5.2.3.1 Integração

Os Arquivos 1 e 2 foram integrados através do Identificador (ID) de cada aluno através da função *"merge"* e *"bind_rows"* do *software R*, representado pela Figura 15. Portanto, ao final foi formada uma base de dados única com todas as variáveis indepententes e também a variável dependente.

Os valores faltantes foram analisados após a etapa de integração, a base continha 275 dados faltantes o que representava 4,93% do total. De acordo com a literatura, 5% de ausência tem sido sugerido como um limite superior máximo para remoção de dados faltantes, como o valor está abaixo de 5% resolveu-se eliminá-las sem prejuízo para a base de dados (MADLEY-DOWD et al., 2019).

Figura 15 – **Integração dos Arquivos 1 e 2 em uma base de dados única**



Fonte: o autor

5.2.3.2 Delimitação

Analisando os currículos dos cursos do Centro Tecnológico, optou-se por delimitar apenas cursos de Engenharia, devido à similaridade do ciclo básico das engenharias nos primeiros semestres, conforme Tabela 5 .

Tabela 5 – **Cursos selecionados CTC-UFSC**

Código curso	Curso
201	Engenharia Civil
202	Engenharia Elétrica
203	Engenharia Mecânica
211	Engenharia Sanitária e Ambiental
212	Engenharia de Produção Civil
213	Engenharia de Produção Elétrica
214	Engenharia de Produção Mecânica
215	Engenharia de Alimentos
216	Engenharia Química
234	Engenharia de Controle e Automação
235	Engenharia Eletrônica
236	Engenharia de Materiais

Fonte: o autor

Os anos selecionados foram de 2008 (ano a partir do qual tornou-se possível o acompanhamento individual dos alunos por meio de um identificador) até 2019, que corresponde ao ano anterior ao início da pandemia de Covid-19.

A partir de 2018, houve uma mudança nas notas do vestibular da UFSC em relação às disciplinas. Uma nova disciplina foi criada pelo agrupamento de Ciências Humanas e Sociais (CHS) e substituiu as disciplinas de Geografia e História. Anteriormente, de 2001 a 2017, as notas de Geografia e História eram avaliadas separadamente. Devido a essa diferença, decidiu-se eliminar as três disciplinas, mantendo apenas as notas de Biologia, Química, Física, Matemática, Português e Língua Estrangeira.

5.2.3.3 Padronização

Nesta fase foram feitas as alterações dos valores inconsistentes:

- **Padronizações de respostas**: as padronizações de dados foram feitas nesta etapa para que todos os registros apresentassem a resposta apropriada. Por exemplo, sexo "f" transformado em "F" por padrão, para feminino. Entre outras transformações para que ficassem no mesmo padrão.

5.2.3.4 Transformação

Esta tarefa inclui operações construtivas de preparação de dados, como a produção de variáveis derivadas de outras, registros inteiramente novos ou novos valores para atributos existentes (CHAPMAN et al., 2000). Nesta fase ocorreu a seguinte transformação:

- **Situação**: a transformação mais importante para a base de dados foi a transformação da situação. Foram selecionados os dados do código situação conforme Tabela 6. Transformando código situação (0 a 13) em classe (Evadido ou Formado). Portanto, Evasão para o presente trabalho é toda forma de saída do curso que não seja a diplomação.
- **Transformação de variáveis**: algumas variáveis foram transformadas para que os algoritmos possam recebê-las no processo de aprendizagem de máquina. Como por exemplo a transformação das respostas do Questionário Socioeconômico do tipo numérico em categórica, outra transformação como a separação de valores contínuos das notas de vírgula para ponto, entre outras.

Tabela 6 – **Transformação da variável dependente nas classes Evadido e Formado**

Código Situação	Situação	Exclusão ou Mantém da base	Classe
0	Regular	Exclusão	-
1	Formado	Mantém	FORMADO
2	Transferido	Mantém	EVADIDO
3	Trancado	Exclusão	-
4	Abandono	Mantém	EVADIDO
5	Troca de curso	Mantém	EVADIDO
6	Jubilado	Mantém	EVADIDO
7	Desistência	Mantém	EVADIDO
8	Eliminado ingresso	Mantém	EVADIDO
9	Eliminado/Cancelado	Mantém	EVADIDO
11	Concluinte	Mantém	FORMADO
13	Falecido	Exclusão	-

Fonte: o autor

5.2.3.5 Seleção de Variáveis Independentes (*Feature Selection)*

Uma parte importante da aprendizagem de máquina é a Seleção de Variáveis (*Feature Selection*) ou redução de dimensionalidade que é o processo de reduzir o número de variáveis em um conjunto de dados, mantendo o máximo de informação possível, isto é, obtendo as Métricas de Avaliação ótimas com o mínimo de variáveis, conhecido como o problema *minimal-optimal* (NILSSON; BJORKEGREN; TEGNER, 2007). É um problema intensamente estudado e atualmente muitos algoritmos foram desenvolvidos para reduzir o conjunto de características para um tamanho gerenciável.

As variáveis independentes servem como entrada para a aprendizagem de máquina e o treinamento do algoritmo para futuras previsões, por isso a quantidade e a relação entre as variáveis podem afetar o desempenho dos algoritmos (FERNANDEZ-GARCIA et al., 2021).

Existem diversos métodos de redução de dimensionalidade: métodos de filtragem (*filter methods*), métodos envelopados (*wrapper methods*) e métodos incorporados (*embedded methods*).

Os métodos de filtragem (*filter methods*) utilizam medidas estatísticas para avaliar a importância de cada característica, como a correlação ou o teste qui-quadrado. Por outro lado, os métodos envelopados (*wrapper methods*) utilizam algoritmos para avaliar um subconjunto de ótimo de variáveis. E por fim, métodos incorporados (*embedded methods*) em que a seleção de variáveis está embutida no próprio algoritmo de aprendizagem de máquina fazendo parte da etapa de treinamento (CHANDRASHEKAR; SAHIN, 2014).

Neste trabalho optou-se pelo método *wrapper* utilizando a técnica do algoritmo *Boruta*.

O algoritmo *Boruta*[1] é uma técnica de seleção de variáveis no pré-processamento

[1]Leva este nome em homenagem à Boruta um deus das florestas na mitologia eslava que auxilia os

para algoritmos de aprendizagem de máquina, que utiliza uma abordagem baseada em *Random Forest* para avaliar a importância de cada variável em uma base de dados. Ele foi projetado por Kursa e Rudnick em 2010 (KURSA; RUDNICKI, 2010) para lidar com problemas de alta dimensionalidade e identificar quais variáveis são realmente relevantes para o modelo.

Para realizar a seleção de variáveis, o algoritmo *Boruta* cria versões duplicadas das variáveis originais, chamadas variáveis sombra. Essas variáveis sombra são geradas aleatoriamente e não têm relação com a variável resposta, elas são usadas para criar uma referência para medir a importância das variáveis originais.

O algoritmo *Boruta* executa as seguintes etapas:

1. O conjunto de variáveis originais, juntamente com as variáveis sombra, é usado para treinar um modelo de aprendizado de máquina usando *Random Forest*.

2. O algoritmo avalia a importância das variáveis comparando suas pontuações com as pontuações das variáveis sombra correspondentes. Se uma variável original tiver um desempenho estatisticamente melhor do que as variáveis sombra correspondentes, ela é considerada importante e é rotulada como um "hit" (acerto).

3. As variáveis consideradas importantes na etapa anterior são marcadas como "confirmadas" e são mantidas no conjunto de variáveis selecionadas.

4. As variáveis originais não consideradas importantes são marcadas como "rejeitadas" e são removidas do conjunto de variáveis.

5. As variáveis sombra relacionadas às variáveis confirmadas são atualizadas: remove-se aquelas que têm menor importância do que a variável original correspondente.

6. O algoritmo *Boruta* repete as etapas acima até que todas as variáveis tenham sido confirmadas ou rejeitadas ou até que um critério de convergência seja atingido. O critério de convergência geralmente é definido como um número fixo de iterações (por padrão 100 iterações).

7. Ao final do processo, as variáveis são classificadas de acordo com o número de hits (acertos) que receberam. Quanto mais hits uma variável teve, mais importante é considerada.

Poderá haver certas variáveis em que o algoritmo não será capaz de tomar uma decisão com a confiança desejada em um número realista de execuções de *Random Forest*. Para isto, o pacote *Boruta* contém uma Função de Correção (*TentativeRoughFix*) que pode

caçadores perdidos na floresta. Em nosso contexto, auxilia com o uso do algoritmo Floresta Aleatória (*Random Forest*) para a escolha das melhores variáveis dos modelos de aprendizagem de máquina.

ser usada para preencher decisões ausentes por meio da simples comparação do *Z-Score* mediano da variável com o *Z-Score* mediano da melhor variável sombra, caso a mediana da variável for maior, o algoritmo considera a variável como importante, caso contrário, classifica como não importante.

Os resultados da Seleção de Variáveis serão apresentadas na **Seção 6.2 - Seleção de Variáveis Independentes**.

5.2.4 Modelagem

Etapa da criação dos modelos, normalmente várias técnicas de modelagem são aplicadas, e seus parâmetros calibrados para otimização. Assim, é comum retornar à Preparação dos Dados durante essa fase.

Foram selecionados de acordo com revisão da literatura cinco algoritmos de classificação buscando uma comparação de melhor performance entre eles:

1. Regressão Logística (RL)
2. *K-Nearest Neighbors* (KNN)
3. *Naive Bayes* (NB)
4. *Support Vector Machine* (SVM)
5. *Random Forest* (RF)

Para este trabalho utilizou-se o pacote *Caret* (KUHN, 2019) do *R Studio* (R CORE TEAM, 2022) conforme Tabela 7. O pacote *caret* (abreviação de *Classification And REgression Training*) é um conjunto de funções que tenta simplificar o processo de criação de modelos preditivos. Os dados foram separados entre treino (70%) e teste (30%). A Técnica de **Validação Cruzada** para este trabalho foi a *k-fold-cross-validation* com $k = 10$.

Tabela 7 – **Parâmetros utilizados nos algoritmos do Pacote *caret* no *software* R**

Algoritmo	**Caret**	**Ajuste parâmetro**
Regressão Logística	*glmnet*	tuneLenght = 4
KNN	*knn*	preprocess = "center"
Naive Bayes	*naive_bayes*	-
Support Vector Machine	*svmRadialSigma*	-
Random Forest	*rf*	ntree = 100

Fonte: o autor

5.2.5 Avaliação

Ao construir um modelo através de um algoritmo de aprendizagem de máquina, utiliza-se métricas apropriadas para avaliar o modelo de predição, esta etapa consiste na etapa de Avaliação (*Evaluation*) na metodologia CRISP-DM.

Buscou-se avaliar os modelos pelas seguintes medidas:

1. Acurácia (ACC)
2. Especificidade (ESP)
3. Sensibilidade (SEN)
4. Precisão (PRE)
5. *F1-Score* (F1S)
6. Curva ROC/AUC (AUC)

5.2.6 Aplicação

A fase de Aplicação da CRISP-DM, não é o objetivo deste trabalho, porém um dos desafios para o futuro próximo, segundo a Pró-Reitoria de Graduação da UFSC, está na:

> "Implantação de um novo sistema de representação curricular e gestão acadêmica, que permita ainda, de maneira complementar, realizar melhor a produção e análise de dados e indicadores referentes aos cursos de graduação da UFSC, **o monitoramento dos índices de reprovação e evasão nos cursos de graduação**, e o aprimoramento dos mecanismos de ocupação de vagas ociosas após o processo de matrícula" UFSC (2020, p.55).

Novos estudos de implantação poderão utilizar o modelo que possui a maior sensibilidade na previsão, identificando os estudantes com maior propensão de evasão e incorporando a ferramenta ao sistema de gestão da graduação na universidade.

6 Resultados

Neste capítulo, são apresentados os resultados da comparação de cinco algoritmos de aprendizagem de máquina na previsão de evasão universitária em cursos de Engenharia. O objetivo deste capítulo é fornecer uma análise detalhada do desempenho dos algoritmos em diferentes estágios da trajetória acadêmica dos estudantes, desde a pré-matrícula até o terceiro semestre.

O capítulo está subdividido como segue:

- Na Seção 6.1 - **Análise Exploratória dos Dados (AED)**. Exploração da estrutura e das características dos dados utilizados na previsão da evasão. Por meio de estatísticas descritivas, visualizações de dados e identificação de padrões, buscou-se compreender melhor o perfil dos estudantes e as possíveis tendências relacionadas à evasão nos cursos de engenharia.
- Na Seção 6.2 - **Seleção de Variáveis Independentes**. Variáveis mais relevantes que influenciam a evasão nos cursos de Engenharia do CTC-UFSC.
- Na Seção 6.3 - **Avaliação dos Algoritmos de Previsão**. Foi analisado o desempenho dos algoritmos nos períodos de pré-matrícula, primeiro semestre e terceiro semestre e utilizadas métricas de desempenho para avaliar a capacidade preditiva de cada algoritmo em cada fase da formação acadêmica.

Ao final deste capítulo, espera-se ter uma compreensão abrangente do desempenho dos algoritmos de aprendizagem de máquina na previsão da evasão universitária em cursos de engenharia do Centro Tecnológico da UFSC, bem como uma visão mais clara dos principais fatores relacionados à evasão nesses cursos. Essas informações serão essenciais para promover ações efetivas de redução da evasão e melhorar a qualidade da formação acadêmica dos estudantes.

Os resultados da **Seleção de Variáveis Independentes e a Análise Exploratória dos Dados (AED)** fazem parte da Etapa da Compreensão e Preparação dos Dados no CRISP-DM. Já os resultados da **Avaliação dos Algoritmos** fazem parte da Modelagem e Avaliação desse processo.

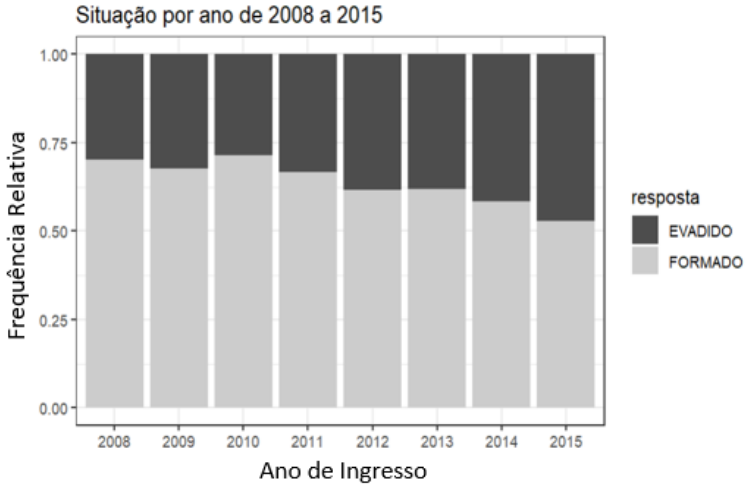
6.1 Análise Exploratória dos Dados (AED)

A análise exploratória dos dados é uma etapa importante para compreender a estrutura e as características dos dados utilizados na previsão. Nesta Seção, são apresentadas as principais análises realizadas, incluindo estatísticas descritivas, visualização de dados e identificação de padrões ou tendências.

6.1.1 Ano de Ingresso

Após analisar os anos selecionados (2008 a 2019), foi observado um aumento progressivo na evasão conforme se aproximava do ano atual (2022). Essa tendência é considerada normal e esperada, uma vez que os alunos não tiveram tempo suficiente para concluir seus cursos. Porém, com o objetivo de garantir evitar distorções entre "Evadidos" e "Formados" que gerem um viés na previsão dos algoritmos, optou-se por limitar a inclusão de dados de ingresso até o ano de 2015 (Figura 16). Essa escolha visa assegurar uma análise mais consistente, considerando um período de tempo significativo e com equilíbrio das tendências Evadidos/Formados para avaliar os fatores que influenciam a evasão universitária desses cursos sem viés.

Figura 16 – **Ano de ingresso dos alunos de Engenharia selecionados de 2008 a 2015**



Fonte: O autor.

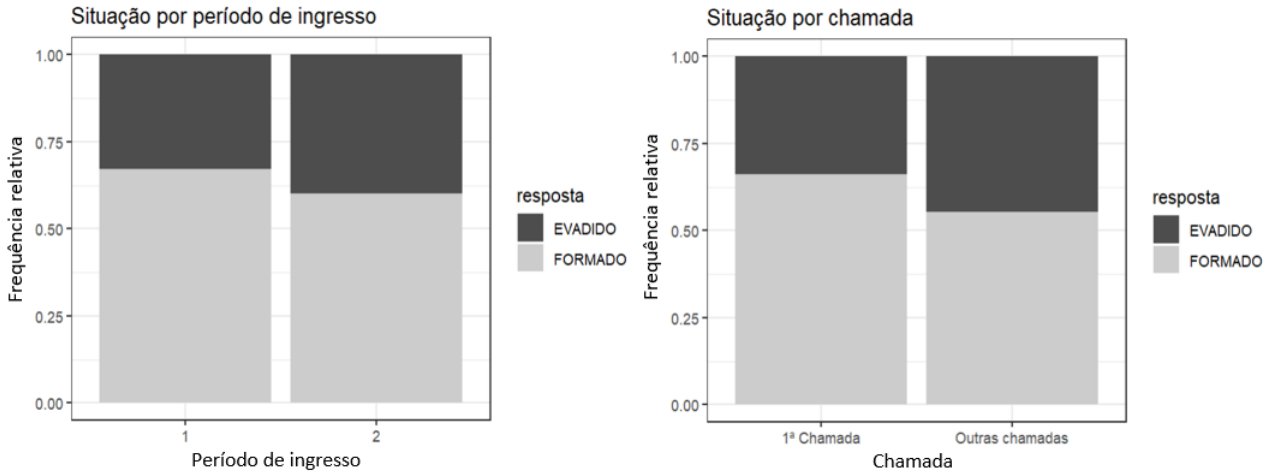
Devido a essa delimitação temporal, todos os dados e análises mostradas a partir deste ponto se referem ao período de 2008 a 2015.

6.1.2 Semestre de Ingresso e Chamada

Ao observar os níveis de evasão em relação ao período de ingresso e a chamada do vestibular, é possível observar que aqueles que ingressaram no primeiro semestre do

ano de matrícula e na primeira chamada apresentam uma taxa de evasão universitária menor. Isso sugere que os alunos que iniciam o curso no primeiro semestre e são admitidos na primeira chamada têm uma maior probabilidade de permanecerem na universidade e concluir os cursos. A Figura 17 mostra os dois perfis de entrada: semestre de ingresso e chamada do vestibular.

Figura 17 – **Perfis de admissão dos candidatos por período de ingresso e chamada do vestibular**

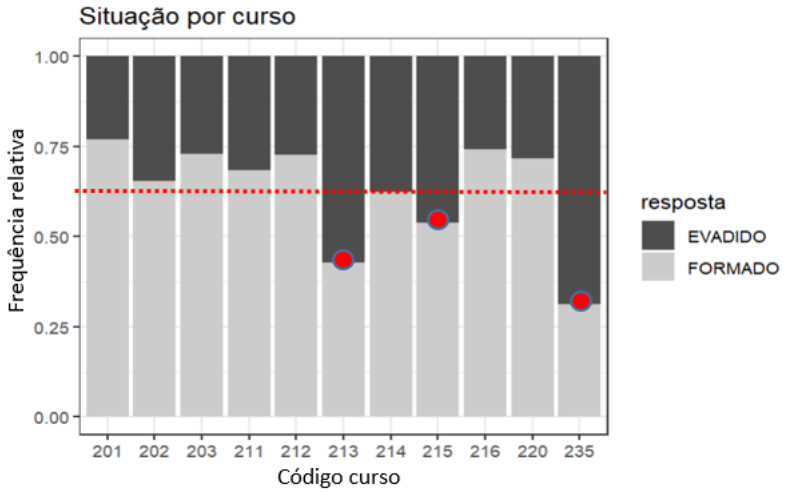


Fonte: O autor.

6.1.3 Evasão por curso

Dentre os 12 cursos selecionados para análise, foi identificado que três deles, nomeadamente **Engenharia de Produção Elétrica (213), Engenharia de Alimentos (215) e Engenharia Eletrônica (235)**, apresentaram uma taxa de formação abaixo da média geral de 62,5% (representada pela linha tracejada vermelha). Essa constatação pode ser visualizada na Figura 18, que exibe o total de alunos formados e evadidos em cada curso no período de 2008 a 2015. Essa discrepância ressalta a importância de se analisar individualmente cada curso, fornecendo *insights* sobre os desafios específicos enfrentados principalmente pelos cursos de Engenharia de Produção Elétrica, Engenharia de Alimentos e Engenharia Eletrônica no que diz respeito à conclusão dos estudos pelos alunos desses cursos.

Figura 18 – **Evasão dos cursos de Engenharia selecionados no CTC-UFSC do período de 2008 a 2015**

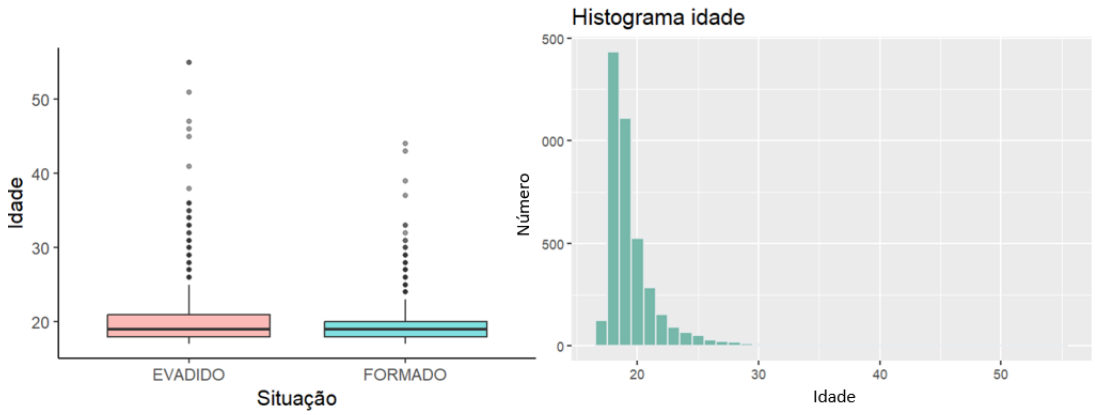


Fonte: o autor.

6.1.4 Idade de ingresso

Ao analisar as idades de ingresso dos alunos nos cursos de engenharia, a Figura 19 ilustra que cerca de 99% dos dados se concentram na faixa etária entre 17 e 30 anos no momento da matrícula. Optou-se por remover as idades acima de 30 anos, pois representavam apenas 0,86% do total.

Figura 19 – **Idade de ingresso dos cursos de Engenharia do CTC-UFSC**



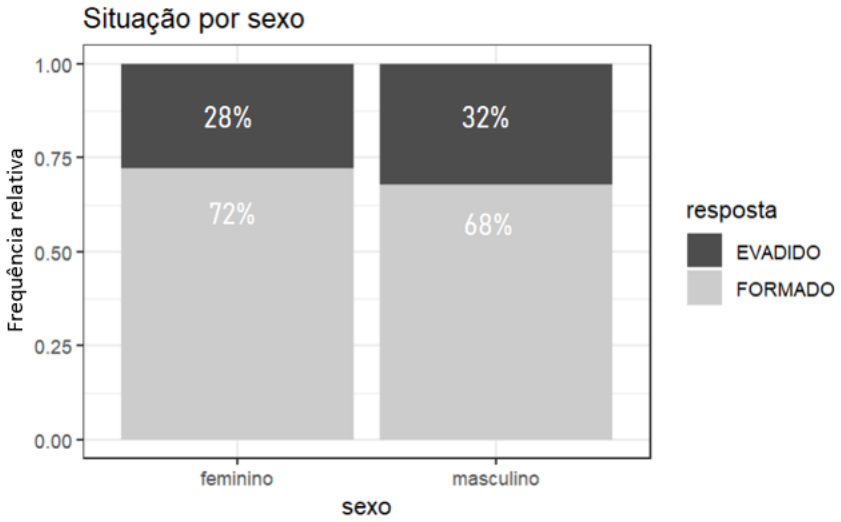
Fonte: o autor.

6.1.5 Sexo

Ao considerar a variável do sexo dos estudantes, é notável que o sexo feminino representa apenas um terço dos dados analisados. Além disso, observa-se uma menor taxa de evasão entre as estudantes do sexo feminino quando comparadas aos estudantes do sexo masculino. Essa diferença na evasão por gênero pode indicar a influência de diversos

fatores, como motivação, interesse e suporte social, que podem desempenhar um papel importante na trajetória acadêmica das estudantes mulheres.

Figura 20 – **Distribuição por sexo nos cursos de Engenharia do CTC-UFSC**

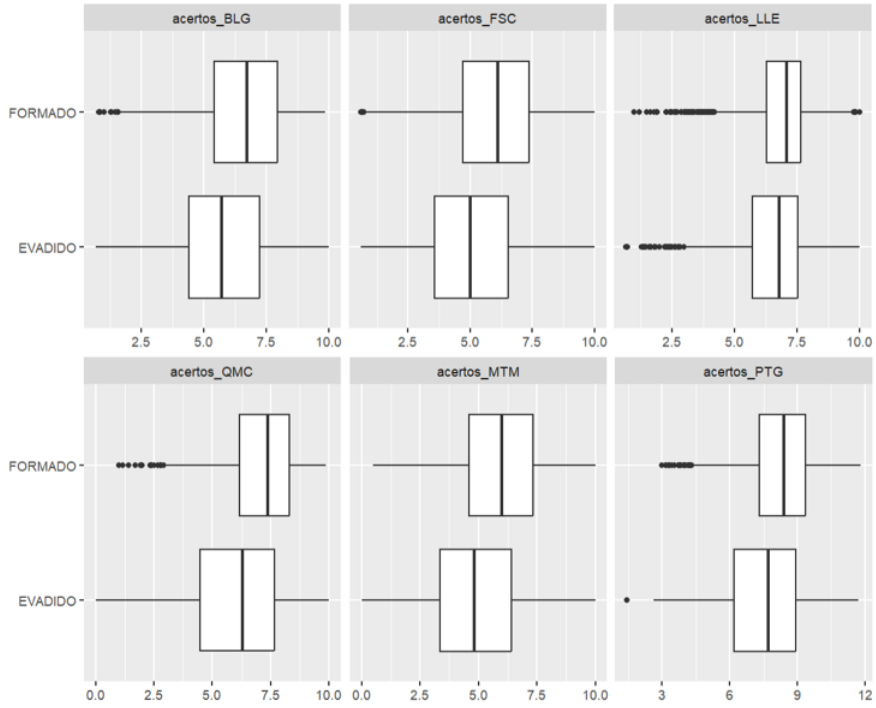


Fonte: o autor.

6.1.6 Notas do Vestibular

Através da análise da Figura 21, é possível observar a relação entre a situação do aluno (“Evadido” ou “Formado”) e suas notas no vestibular em disciplinas específicas, como Biologia, Física, Língua Estrangeira, Química, Matemática e Português. Nessa análise, fica evidente que os alunos formados apresentam notas mais altas em todas as disciplinas do vestibular em comparação aos alunos que evadiram. Essa tendência sugere que um desempenho mais elevado nas disciplinas do vestibular pode estar associado a uma maior probabilidade de conclusão do curso. Tal observação reforça a importância do desempenho acadêmico inicial como um indicador potencial de sucesso acadêmico a longo prazo.

Figura 21 – **Notas do vestibular e a Evasão nos cursos selecionados de Engenharia do CTC-UFSC de 2008 a 2015**



Fonte: o autor.

6.1.7 Delimitações pela Análise Exploratória de Dados

Em resumo, ao concluir a Análise Exploratória de Dados, foram identificadas algumas tendências significativas que auxiliaram na delimitação dos dados durante a fase de Pré-processamento, preparando assim a base de dados para a etapa de Modelagem. Essas tendências foram fundamentais para compreender melhor as características dos dados e direcionar o processo de construção dos modelos preditivos. Ao levar em consideração essas observações, é possível otimizar o desempenho dos algoritmos de aprendizagem de máquina e obter resultados mais precisos na previsão da evasão, são elas:

- Delimitação dos Ano de Ingresso de 2008 a 2015 para ter um equilíbrio na variável dependente;
- Utilização na Previsão do Modelo escolhido o ano de 2016 como Dados de Previsão numa “turma-teste” simulando dados reais;
- Delimitação da idade de ingresso de 17 a 30 com a remoção dos 0,86% com idade acima de 30 anos;

6.2 Seleção de Variáveis Independentes

Por meio do algortimo *Boruta*, buscou-se identificar as principais variáveis independentes que contribuem significativamente para predizer a evasão dos estudantes.

A seleção de variáveis foi feita somente no **Modelo Pré-matrícula** entre elas:

1. Sexo
2. Idade
3. Curso
4. Chamada
5. Notas do Vestibular
6. 34 Variáveis do Questionários Socioeconômico

O algoritmo, após 100 iterações, confirmou todas as variáveis dos itens 1 a 5 (Sexo, Idade, Curso, Chamada, Notas do Vestibular) e, das 34 Questões do Questionário Socioeconômico, o algoritmo confirmou como importantes 21 delas. Na Tabela 8 são apresentadas as variáveis selecionadas juntamente no o valor do *hit* da variável, que, como mencionado no capítulo 3, indica a importância da variável. Para visualizar a análise completa, no Apêndice A são mostrados os resultados do desempenho de todas as variáveis na avaliação conduzida usando o algoritmo Boruta.

Optou-se por utilizar as 21 variáveis confirmadas com a Função de Correção (*TentativeRoughFix*) do algoritmo. Para visualizar por completo, o **Apêndice A** mostra as variáveis selecionadas e o Questionário Socioeconômico pode ser visualizado no **Anexo A**

As variáveis dependentes usadas em cada um dos três modelos explorados no contexto deste trabalho (Pré-matrícula, Primeiro semestre e Terceiro semestre) são mostradas na Tabelas 8, 9 e 10, respectivamente. Observa-se que nos modelos Primeiro Semestre e Terceiro semestre, além das variáveis já selecionadas no modelo Pré-matrícula, são adicionadas as variáveis relacionadas ao desempenho acadêmico (IAA) e reprovação por faltas (FI).

1. **Modelo Pré-matrícula:**

Tabela 8 – **Variáveis Selecionadas do Modelo Pré-matrícula**

Grupo	Variável	normHit Boruta	Tipo
Variável Dependente	resposta	-	Fator
Individuais	sexo	99%	Fator
	idade	100%	Inteiro
	curso	100%	Fator
	chamada	75%	Fator
Notas Vestibular	acertos Biologia	100%	Contínua
	acertos Matemática	100%	Contínua
	acertos Física	100%	Contínua
	acertos Química	100%	Contínua
	acertos Língua Estrangeira	100%	Contínua
	acertos Português	100%	Contínua
Questionário Socioeconômico	Salário Bruto Familiar (q17)	100%	Fator
	Sustento Familiar (q20)	100%	Fator
	Motivo Vestibular (q34)	99%	Fator
	Motivo Escolha Pré-Vestibular (q11)	97%	Fator
	Meio Informação Usual (q24)	96%	Fator
	Tipo de Estabelecimento Ensino Médio (q7)	95%	Fator
	Instrução do Pai (q18)	93%	Fator
	Frequentou Pré-Vestibular (q10)	92%	Fator
	Sua Ocupação (q23)	81%	Fator
	Turno do Ensino Médio (q8)	76%	Fator
	Expectativa Curso Superior (q14)	76%	Fator
	Instrução da Mãe (q19)	76%	Fator
	Motivo Escolha 1ª Opção (q13)	72%	Fator
	UF Residência (q2)	71%	Fator
	Tipo de Curso Ensino Médio	63%	Fator
	Hobbie (q32)	63%	Fator
	Estado Civil (q1)	59%	Fator
	Pessoas da Família na Residência (q16)	57%	Fator
	Exerce Atividade Remunerada (q22)	56%	Fator
	Tipo Estabelecimento Ensino Fundamental (q4)	54%	Fator
	Iniciou Ensino Superior (q28)	52%	Fator

Fonte: o autor

2. **Modelo Primeiro Semestre:**

Tabela 9 – **Variáveis Selecionadas do Modelo Primeiro Semestre**

Grupo	Variável	Tipo
Variável Dependente	resposta	Fator
Pré-matrícula	Todas as Variáveis do Modelo Pré-Matrícula	-
Desempenho Primeiro Semestre	IAA do Primeiro Semestre	Contínua
	Frequência Insuficiente Primeiro Semestre	Inteiro

Fonte: o autor

3. **Modelo Terceiro Semestre:**

Tabela 10 – **Variáveis Selecionadas do Modelo Terceiro Semestre**

Grupo	Variável	Tipo
Variável Dependente	resposta	Fator
Pré-matrícula	Todas as Variáveis do Modelo Pré-Matrícula	-
Desempenho Terceiro Semestre	IAA do Terceiro Semestre	Contínua
	Frequência Insuficiente Terceiro Semestre	Inteiro

Fonte: o autor

O total de registros e variáveis nos três modelos estão apresentados na Tabela 11

Tabela 11 – **Total de variáveis e registros dos três modelos**

Modelo	Variáveis	Registros	Evadidos(%)	Formados(%)	70% Treino	30% Teste
Pré-Matrícula	31	4394	1564 (36%)	2830 (64%)	3076	1318
Primeiro Semestre	33	4389	1559 (36%)	2830 (64%)	3073	1316
Terceiro Semestre	33	3925	1099 (28%)	2826 (72%)	2749	1176

Fonte: o autor

Em cada modelo é possível observar na variável dependente a classificação entre "Evadidos" ou "Formados" na situação do aluno fornecida. Além disso, nos três modelos os dados entre "Evadidos" e "Formados" estão desbalanceados, apresentando uma maior proporção de "Formados" nos três modelos.

No Modelo Pré-Matrícula (Tabela 8), em que houve a seleção de variáveis pelo algoritmo *Boruta*, é possível observar que o *normhit*, que é uma medida que indica a proporção de vezes em que uma variável real teve uma importância estatisticamente maior do que a importância média das variáveis "falsas", ficou acima de 50%. Isto significa que a variável em questão teve uma importância estatisticamente significativa em relação às variáveis "falsas" em mais da metade das iterações realizadas pelo algoritmo. Portanto, a variável pode ser considerada importante para o modelo preditivo.

6.3 Avaliação dos Algoritmos de Previsão

A **Avaliação dos Algoritmos de Previsão** desempenha um papel fundamental neste trabalho, pois é nessa etapa que os três modelos de previsão são minuciosamente avaliados. É considerada a parte mais importante do estudo, pois proporciona uma análise aprofundada do desempenho e da eficácia dos algoritmos utilizados. Por meio de métricas e técnicas adequadas, é possível verificar a capacidade de cada modelo em prever a evasão universitária nos cursos de Engenharia do CTC-UFSC. A partir desses resultados, serão obtidos *insights* valiosos que contribuirão para uma visão abrangente sobre a capacidade desses algoritmos em prever a evasão universitária em diferentes momentos da trajetória acadêmica dos estudantes.

O Quadro 3 fornece os resultados dos algoritmos de classificação Regressão Logística, *K-Nearest Neighbors, Naive Bayes, Support Vector Machines e Random Forest* em três Modelos diferentes: Pré-Matrícula, Primeiro Semestre e Terceiro Semestre.

A linha em cinza representa o algoritmo selecionado com a melhor perfomance em cada modelo. O Quadro mostra os valores das métricas utilizadas na ordem: Acurácia, Especificidade, Sensibilidade, Precisão, F1-Score e Área sob a Curva. Também, os valores absolutos da previsão (30% dos dados de teste de cada modelo) estão mostrados nas colunas Matriz de Confusão (VN,FP,VP,FN) em que a classe positiva utilizada nesta dissertação é a Evasão e negativa a Formação, portanto:

- **Verdadeiros Positivos (VP)** = alunos previstos como Evadido e realmente Evadido (acerto na Evasão)
- **Falsos Negativos (FN)** = alunos previstos como Formado, porém na realidade evadiu (erro tipo II)
- **Verdadeiros Negativos (VN)** = alunos previstos como Formado e realmente Formado (acerto na Formação)
- **Falsos Positivos (FP)** = alunos previstos como Evadido, porém na realidade formou (erro tipo I)

Os valores das métricas estão mostrados de 0 a 1, mas podem ser entendidos também na forma de porcentagem, conforme utilizado na literatura.

Quadro 3 – **Quadro com os resultados dos algoritmos de aprendizagem de máquina com os Dados de Teste em cada Modelo**

MODELOS	ALGORITMOS	Métricas de Avaliação						Matriz de Confusão			
		ACC	ESP	SEN	PRE	F1S	AUC	VN	FP	VP	FN
Pré-Matrícula	Regressão Logística	0.6897	0.9058	0.2985	0.6364	0.4064	0.6625	769	80	140	329
	K Nearest Neighbors	0.6419	0.8327	0.2964	0.4947	0.3707	0.6186	707	142	139	330
	Naive Bayes	0.6487	0.9128	0.1706	0.5195	0.2568	0.6216	775	74	80	389
	Support Vector Machines	0.6942	0.9058	0.3113	0.6460	0.4201	0.6752	769	80	146	323
	Random Forest	0.6920	0.8716	**0.3667**	0.6121	0.4587	**0.6726**	740	109	172	297
Primeiro Semestre	Regressão Logística	0.7941	0.9293	0.5482	0.8101	0.6539	0.8193	789	60	256	211
	K Nearest Neighbors	0.7591	0.9152	0.4754	0.7551	0.5834	0.7497	777	72	222	245
	Naive Bayes	0.7211	0.9270	0.3469	0.7232	0.4689	0.7253	787	62	162	305
	Support Vector Machines	0.8047	0.9293	**0.5782**	0.8182	0.6775	**0.8465**	789	60	270	197
	Random Forest	0.7903	0.9069	0.5782	0.7736	0.6618	0.8069	770	79	270	197
Terceiro Semestre	Regressão Logística	0.8418	0.9433	0.5805	0.7992	0.6725	0.8772	799	48	191	138
	K Nearest Neighbors	0.8350	0.9681	0.4924	0.8571	0.6255	0.8191	820	27	162	167
	Naive Bayes	0.7645	0.8560	0.5289	0.5878	0.5568	0.7669	725	122	174	155
	Support Vector Machines	0.8588	0.9599	0.5988	0.8528	0.7036	0.8553	813	34	197	132
	Random Forest	0.8452	0.9374	**0.6079**	0.7905	0.6873	**0.8649**	794	53	200	129

ACC – Acurácia; ESP – Especificidade; SEN – Sensibilidade; PRE – Precisão; F1S – F1 Score; AUC – *Area Under Curve*;

VN – Verdadeiro Negativo; FP – Falso Positivo; VP – Verdadeiro Positivo; FN – Falso Negativo

Fonte: o autor.

Cada algoritmo foi avaliado dando prioridade para a métrica Sensibilidade - SEN, pois quanto maior esta métrica para a classificação da Evasão, maior é o acerto de Verdadeiros Positivos (acertos de Evadidos) e menor os Falsos Negativos. O critério da Sensibilidade é a métrica mais utilizada em técnicas de aprendizagem de máquina literatura para prever com eficiência os alunos propensos a evadir (HANNAFORD; CHENG; KUNES-CONNELL, 2021) (FERNANDEZ-GARCIA et al., 2021) (COSTA et al., 2020).

No contexto da Evasão, é importante minimizar os Falsos Negativos, ou seja, não classificar erroneamente alunos como formandos, porém ao final evadiram.

A Área sob a Curva - AUC foi outra métrica em que foi levada em consideração, pois quanto maior a AUC maior a capacidade de previsão dos modelos e maior a Sensibilidade e menor os Falsos Positivos no Modelo.

Outras métricas no geral foram consideradas, como Precisão - PRE (acertos da previsão da Evasão), Acurácia - ACC (Acerto Geral), F1-Score (equilíbrio entre Sensibilidade e Precisão) e Especificidade - ESP (acertos na Formação).

Além disso, a Matriz de Confusão com os números absolutos de VP (quanto maior melhor) e FN (quanto menor melhor) foram considerados para a escolha do melhor algoritmo.

Em geral, a Acurácia dos algoritmos variou de 65% a 85%, apresentando um aumento à medida que os alunos avançaram nos cursos. Esse padrão é esperado, uma vez que o desempenho acadêmico, medido por índices como o Índice de Aproveitamento Acadêmico (IAA) e reprovações por Frequência Insuficiente (FI), tendem a estar mais fortemente correlacionados com a variável dependente ao longo do curso.

Os índices de Sensibilidade variaram entre 30% e 60%, enquanto os índices de Especificidade foram consistentemente altos, todos acima de 80%. Isso indica que, de forma geral, os algoritmos foram mais eficazes em prever os alunos que se formam do que aqueles que evadem.

No **Modelo de Pré-matrícula**, no qual não há informações disponíveis sobre o desempenho acadêmico dos alunos, o algoritmo *Random Forest* demonstrou a melhor performance, alcançando a maior Sensibilidade (0,3667), porém ainda muito baixa. Resultados semelhantes foram encontrados no estudo de Alvarez, Callejas e Griol (2020), que obteviveram uma Sensibilidade de 0,309 utilizando o algoritmo de Árvore de Decisão (J48) no período pré-matrícula. Esses resultados indicam a eficácia do algoritmo *Random Forest* na previsão de evasão em um estágio inicial do processo acadêmico, mesmo sem dados específicos sobre o desempenho dos alunos.

Ao incorporar dados de desempenho acadêmico no **Modelo do Primeiro Semestre**, foi observada uma melhora significativa nas métricas de avaliação. A Acurácia média aumentou para 77%, em comparação com os 67% do Modelo de Pré-matrícula. Os

algoritmos *Random Forest* e *Support Vector Machine* (SVM) apresentaram desempenhos semelhantes, porém o SVM teve um número menor de falsos positivos. Ambos os algoritmos apresentaram 197 falsos negativos, indicando que ainda há um desafio em identificar corretamente os casos de evasão. Esses resultados destacam a importância de incorporar dados de desempenho acadêmico para melhorar a precisão do modelo na previsão de evasão durante o primeiro semestre dos cursos de Engenharia.

No **Modelo do Terceiro Semestre**, a melhora em relação ao primeiro semestre não foi tão significativa quanto a observada na transição da pré-matrícula para o primeiro semestre, o que significa que o Modelo de Previsão de Evasão já no primeiro semestre pode ser utilizado de forma eficaz. O algoritmo *Random Forest* continuou apresentando os melhores resultados, porém muito próximos ao *Support Vector Machine*. RF obteve uma acurácia de 84,52% e a maior Sensibilidade entre os testes, alcançando 61% (0,6079). Esses resultados indicam que tanto o algoritmo *Random Forest* quanto o *Support Vector Machine* são os melhores algoritmos para a previsão de evasão durante o primeiro e terceiro semestre dos cursos, fornecendo informações valiosas para intervenções precoces e estratégias de retenção de alunos.

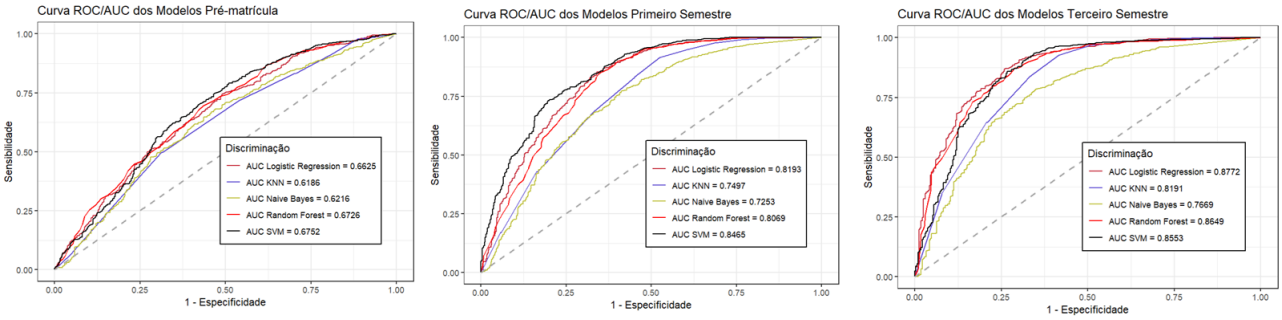
Entre os algoritmos de aprendizado de máquina utilizados, o KNN e o *Naive Bayes* apresentaram os piores desempenhos em termos de previsão de evasão. Embora esses algoritmos não tenham se saído tão bem quanto os outros, eles seguiram a mesma tendência geral dos resultados. É importante mencionar que o *Naive Bayes* teve uma Sensibilidade particularmente baixa no Modelo de Pré-matrícula, porém, nas demais análises, a diferença não foi tão significativa.

Em resumo, os resultados obtidos mostraram que, num primeiro momento, a Sensibilidade dos modelos não atingiu valores elevados. Por outro lado, a Especificidade foi consistentemente alta. A melhora entre o Modelo do Primeiro Semestre e o Modelo do Terceiro Semestre foi discreta, o que sugere que a previsão de evasão já pode ser feita com uma boa precisão desde o primeiro semestre. Os resultados indicam que os modelos têm potencial para serem usados como uma ferramenta de previsão precoce tanto da evasão universitária quanto de formação, fornecendo informações valiosas para intervenções e suporte aos alunos desde o início de seus cursos.

A Figura 22 resume os resultados da Área sob a Curva (AUC) dos algoritmos em cada Modelo. É possível observar uma maior área sob a curva na medida que os alunos avançam no tempo, isso indica uma tendência de maior previsibilidade dos modelos. Essa tendência reforça a importância de considerar o progresso acadêmico dos alunos ao realizar a previsão da evasão universitária.

A linha pontilhada significa um modelo sem nenhum valor de previsão e quanto mais próximo ao canto superior esquerdo melhor a classificação do modelo na previsão de "Evadidos" e "Formados".

Figura 22 – **AUC dos Modelos de Previsão**



Fonte: o autor.

6.3.1 Balanceamento de Dados do Modelo Terceiro Semestre

Ao considerar o Modelo do Terceiro Semestre com as melhores métricas de previsão, optou-se por realizar novas modelagens com dados balanceados deste período.

No Modelo do Terceiro Semestre foi observada uma maior discrepância entre o número de alunos Evadidos (28%) e Formados (72%). Em situações em que há uma discrepância significativa entre as classes, é essencial realizar um balanceamento, de modo a obter um equilíbrio entre as classes (KEMPER; VORHOFF; WIGGER, 2020). Para isso, foi aplicado a Técnica de Sobreamostragem Sintética de Minorias (SMOTE: *Synthetic Minority Oversampling Technique*).

A SMOTE é uma técnica bastante difundida na literatura para tratar do problema de desbalanceamento na classe da variável dependente (COSTA et al., 2017) (MIRANDA; GUZMáN, 2017) (TENPIPAT; AKKARAJITSAKUL, 2020) (FLORES; HERAS; JULIAN, 2022). A técnica funciona gerando amostras sintéticas da classe minoritária através da interpolação de dados entre cada caso da classe de destino e seus vizinhos mais próximos. O algoritmo cria novos exemplos que combinam características do caso de destino com as características de seus vizinhos, utilizando o algoritmo KNN (CHAWLA et al., 2002). Essa abordagem permite aumentar a representatividade da classe minoritária e melhorar o desempenho dos modelos de aprendizado de máquina na previsão.

Além do algoritmo SMOTE, foi utilizado um balanceamento “Natural” obtendo amostras aleatórias fixas de 1000 registros de alunos “Evadidos” e 1000 registros de alunos “Formados”, obtidos pela base de dados do Terceiro Semestre Sem Balanceamento. Os números dos registros em cada Modelo encontram-se na Tabela 12.

Tabela 12 – **Dados Terceiro Semestre Balanceados**

Modelo Terceiro Semestre	Registros	Evadidos(%)	Formados(%)	70% Treino	30% Teste
Sem Balanceamento	3925	1099 (28%)	2826 (72%)	2749	1176
Balanceamento Natural	2000	1000 (50%)	1000 (50%)	1400	600
Balanceamento SMOTE	3955	1978 (50%)	1977 (50%)	2769	1186

Fonte: o autor

Foi utilizado o algoritmo *Random Forest*, para a comparação dos Modelos de Balanceamento do Terceiro Semestre. Os resultados se encontram na Tabela 13.

Tabela 13 – **Resultados *Random Forest* Modelo Terceiro Semestre Balanceado**

Modelo	**ACC**	**ESP**	**SEN**	**PRE**	**F1S**	**AUC**
Sem Balanceamento	0,8452	0,9374	0,6079	0,7905	0,6873	0,8649
Balanceamento Natural	0,7983	0,8529	0,7438	0,8369	0,7867	0,8601
Balanceamento SMOTE	0,8390	0,8786	0,7993	0,8681	0,8323	0,9132

Fonte: o autor

Foi possível observar uma melhora significativa na métrica da Sensibilidade dos Modelos com os dados balanceados, que aumentou de 60% para 80%. Esses resultados corroboram com o estudo de Kemper, Vorhoff e Wigger (2020, p.39), que demonstraram que o balanceamento dos dados se torna uma abordagem válida para prever com maior eficácia os alunos que evadem.

6.4 Previsão do Modelo

Na fase de validação dos modelos desenvolvidos, foram realizados treinos e testes utilizando dados de alunos que ingressaram de 2008 a 2015. Estes testes tem como objetivo avaliar a qualidade dos modelo e ajustá-los para quando novos dados forem inseridos no modelo (dados de previsão).

6.4.1 Dados de Previsão: Ano 2016

Para os dados de previsão, foi utilizado o Ano de 2016 como um ano experimental do modelo para o primeiro e o terceiro semestres, como uma simulação de uma turma real na previsão.

Utilizou-se também, o algoritmo *Randon Forest*, que teve o melhor desempenho na modelagem com dados de 2008 a 2015. Além disso, foram utilizadas as mesmas variáveis independentes nos modelos do primeiro e terceiro semestre, resultando em duas bases

de dados exclusivas para o ano de 2016. A Tabela 14 apresenta o total de registros dos Modelos.

Tabela 14 – **Base de Dados do Ano 2016 Modelo Primeiro e Terceiro Semestre**

Modelos 2016	**Registros**	**Evadidos (%)**	**Formados (%)**
Primeiro Semestre	366	223 (61%)	143 (39%)
Terceiro Semestre	306	164 (54%)	142 (46%)

Fonte: o autor

Ao executar o algoritmo *Random Forest* para os dois Modelos de 2016, observou-se que a Sensibilidade ficou acima de 60% (Tabela 15), indicando que o algoritmo teve sucesso em identificar a maioria dos alunos que evadiram. No entanto, é importante ressaltar que a Especificidade apresentou um resultado próximo a 1 (0,9720 - primeiro semestre e 0,9859 - terceiro semestre), o que significa que o algoritmo acertou quase todos os alunos que se formaram. Isso demonstra a alta capacidade de acerto do modelo na previsão da formação dos alunos.

Tabela 15 – **Tabela com os resultados da previsão para o ano de 2016**

Modelo 2016	**Métricas de Avaliação**						**Matriz de Confusão**			
	ACC	**ESP**	**SEN**	**PRE**	**F1S**	**AUC**	**VN**	**FP**	**VP**	**FN**
Primeiro Semestre	0.7596	0.9720	0.6233	0.9720	0.7596	0.8984	139	4	139	84
Terceiro Semestre	0.8137	0.9859	0.6646	0.9820	0.7927	0.9189	140	2	109	55

Fonte: o autor

6.4.2 Ajuste da *Threshold* de Classificação

O baixo número de Falsos Positivos na previsão dos Modelos de 2016, resultou em duas consequências:

1. Alta taxa de Especificidade (acerto de alunos formados)
2. Alta taxa de Precisão (previsoes corretas de evasão)

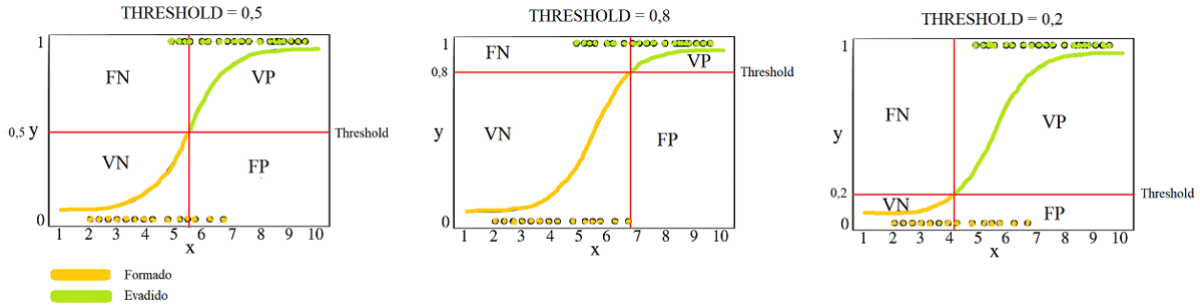
Em outras palavras, o algoritmo acertou cerca de 97% e 98% das previsões de evasão no primeiro e terceiro semestre, respectivamente, para o grupo de alunos previstos como evadidos.

No entanto, em relação ao grupo real de alunos evadidos, o algoritmo previu corretamente em média apenas 64%, deixando passar ou não identificando os outros 36% que foram alunos previstos como formados, porém evadiram (Falsos Negativos) é um erro considerável quando se deseja acertar ao máximo a evasão.

Para aumentar a abrangência na detecção desses casos, é possível ajustar "artificialmente" a classificação do algoritmo, reduzindo o valor de sua *threshold*.

De forma ilustrativa, por padrão, a *threshold* de probabilidade é ajustada em 50%. Acima desse valor, o algoritmo classifica como positiva (“Evadido”), e abaixo como negativa (“Formado”), como ilustrado na Figura 23. Ao ajustar a reta *threshold* para cima, a previsão se torna mais negativa (aumentando os Verdadeiros Negativos), e ao ajustar para baixo, mais positiva (abrangendo mais os casos de Verdadeiros Positivos).

Figura 23 – ***Threshold* de Classificação**



Fonte: o autor.

No entanto, é importante ressaltar que há um *tradeoff* a ser considerado ao fazer esse ajuste fino. Aumentar a taxa de Sensibilidade (TPR) implica em uma diminuição na taxa de Especificidade.

Por exemplo, ao ajustar a *threshold* apenas para o Modelo do Terceiro Semestre de 2016 em 43,5%, a Sensibilidade aumenta para 80%, o que reduz os Falsos Negativos para 33. No entanto, esse ajuste também leva a um aumento nos Falsos Positivos para 9, o que resulta em uma diminuição da Precisão. Essa compensação entre Sensibilidade e Especificidade deve ser cuidadosamente considerada ao ajustar a *threshold*, levando em conta o contexto e as necessidades específicas do problema em questão. Os detalhes desses ajustes podem ser visualizados na Tabela 16.

Tabela 16 – **Tradeoff de ajuste da Threshold de Classificação**

Threshold	**ACC**	**ESP**	**SEN**	**PRE**	**F1S**	**VN**	**FP**	**VP**	**FN**
0,7123	0,7843	1,0000	0,5976	1,0000	1,0000	142	0	98	66
0,5000	0,8137	0,9859	0,6646	0,9820	0,7927	140	2	109	55
0,4350	0,8627	0,9366	0,7988	0,9357	0,8198	133	9	131	33
0,2045	0,8562	0,8873	0,8293	0,8947	0,8608	126	16	136	28
0,1565	0,8497	0,8028	0,8902	0,8391	0,8639	114	28	146	18
0,1019	0,6928	0,3944	0,9512	0,6446	0,7685	56	86	156	8
0,0626	0,5817	0,0985	1,0000	0,5616	0,7193	14	128	164	0

Fonte: o autor

É possível observar que existe um custo de classificação, na medida em que vamos baixando a linha de corte, menor é o erro de Falsos Negativos, porém ao custo de um aumento nos Falsos Positivos (inversamente proporcionais).

Isso é especialmente relevante quando o objetivo é apenas identificar os verdadeiros evasores, priorizando o aumento dos Verdadeiros Positivos sem considerar a questão da formação e focando exclusivamente na evasão. Nesse caso, a utilização de um limiar (*threshold*) mais baixo pode ser uma solução adequada.

No entanto, para alcançar um equilíbrio entre FN = 0 *versus* FP = 0, o limiar deve ser definido em um valor entre 0.7123 e 0.0626, conforme Tabela 16. Observa-se que o custo de obter zero Falsos Negativos é maior do que o custo de obter zero Falsos Positivos, pois para zerar os Falsos Negativos a linha de corte teve que ser quase zerada (0,0626) enquanto que para zerar os Falsos Positivos a linha subiu 0,2123 pontos. Isso indica que os alunos evasores estão mais dispersos no modelo do que os que se formam, apresentando uma maior dificuldade em comparação com a previsão da formação.

6.5 Modelo Final

Para a escolha na prática de um Modelo para a Previsão dos Dados de Evasão e que possa futuramente ser aplicado em um Sistema Acadêmico que auxilie os Gestores de Educação na tomada de decisões e na ação antecipada do problema, sugere-se:

- Utilizar algoritmo de *Random Forest* ou em segunda opção *Support Vector Machine* com os parâmetros ajustados conforme Tabela 7.
- Uma *threshold* ajustada entre 16% e 44% pois apresentou o melhor ajuste entre Sensibilidade e Precisão do Modelo (F1-Score alto) juntamente com a melhor Acurácia (acertos totais) - vide Tabela 17.

Tabela 17 – ***Threshold* recomendada para utilização nos algoritmos de classificação de previsão de evasão**

Threshold	**FP**	**FN**	**F1-Score**	**Sensibilidade**	**Acurácia**
0,44	2,94%	10,78%	82%	80%	86%
0,20	5,23%	9,15%	86%	83%	86%
0,16	9,15%	5,88%	86%	89%	85%

Fonte: o autor

Na prática, quando o Modelo realiza uma previsão, o gestor educacional terá apenas essa previsão como informação disponível. Essa previsão pode se tornar realidade ou não, uma vez que a situação ainda não foi concretizada naquele momento. Portanto, é responsabilidade do gestor transformar essa previsão em um número máximo de Falsos Positivos, ou seja, situações em que o Modelo previu a evasão, porém os alunos acabaram se formando. Isso significa que a previsão desempenhou seu papel de auxiliar na prevenção do problema.

7 Discussão

Diversos trabalhos relacionados à modelagem preditiva de evasão universitária comparam algoritmos de aprendizagem de máquina e seleção das diferentes variáveis independentes (Algoritmos vs. Variáveis).

O **Pré-processamento** é uma parte importante do processo de aprendizagem de máquina, especialmente a seleção de variáveis independentes, pois uma grande quantidade de variáveis pode causar problemas de dimensionalidade, por isso diferentes formas de seleção são utilizadas. Urbina-Nájera, Camino-Hampshire e Barbosa (2020) fizeram uma análise de diferentes métodos de seleção de atributos, posteriormente utilizando dois métodos de seleção (*GainRatioAttributeEval e InfoGainAttributeEval*) reduziram de 56 para 27 atributos, aumentando a Acurácia e a Sensibilidade do modelo de Árvore de Decisão de 74,6% para 92,6%. Fernandez-Garcia et al. (2021) utilizaram de engenharia de atributos (*Feature Engineering*) para reduzir de 40 atributos a partir de 3 bases diferentes para somente 16. Opazo et al. (2021), utilizou *forward selection* para reduzir a dimensionalidade das variáveis. Esses exemplos ilustram como a etapa de pré-processamento e seleção de variáveis é crucial no processo como um todo, o que pode levar a diferenças nos resultados de avaliação dos algoritmos.

Em um estudo conduzido por Naseem, Chaudhary e Sharma (2022), utilizaram o algoritmo *Boruta* para a análise de variáveis, resultando em uma redução de 13 para 10 variáveis confirmadas. Na presente dissertação, foi utilizado o mesmo algoritmo para a seleção de variáveis no Modelo Pré-Matrícula, resultando numa redução de 44 variáveis para 31. Também, a Análise Exploratória dos Dados auxiliou na delimitação do período analisado (de 2008 a 2015) e dos dados de predição para o ano de 2016. Isso ressalta a importância de explorar diferentes abordagens na seleção das variáveis na predição de evasão universitária.

De acordo com os artigos pesquisados na Revisão Sistemática (Capítulo 4), a média de algoritmos utilizados na comparação de previsão de evasão universitária foi de 2,75. Os algoritmos mais comumente empregados foram Árvores de Decisão (*Decision Tree*), Redes Neurais (*Neural Networks*), Classificadores de Bayes (*Bayesian Classification*), Regressão Logística (*Logistic Regression*), Floresta Aleatória (*Random Forest*), *Support Vector Machines*, *K-Nearest Neighbors* (KNN), Algoritmos *Boosting* entre outros. Nesta dissertação, optou-se pelos algoritmos de classificação *Random Forest, Naive Bayes*, Regressão Logística, *Support Vector Machine e K-Nearest Neighbour.*

Em relação aos resultados das **Métricas de Avaliação**, Costa et al. (2017, p.251) compararam a efetividade de 4 algoritmos na previsão de evasão em duas modalidades

de curso (presencial e EAD) na disciplina de introdução à programação, em diferentes momentos (desde a primeira semana até o primeiro exame), apresentando um F1-Measure de 50% a 82% sendo o modelo de Árvore de Decisão com o melhor resultado. Santos et al. (2019) fizeram a comparação entre três cursos de Tecnologia da Informação da Universidade Federal de Sergipe (Ciências da Computação, Sistemas da Informação e Engenharia da Computação) obtendo uma acurácia média de 66%, 70% e 72% respectivamente. Hannaford, Cheng e Kunes-Connell (2021, p.5) obtiveram 73,5% de acurácia com Random Forest e 81,3% de acurácia com diversos algoritmos na pré-matrícula e no primeiro ano respectivamente, para cursos de enfermagem. Fernandez-Garcia et al. (2021, p.13), encontraram 72,34% de Acurácia na pré-matrícula com *Gradient Boosting* e 88,46% no terceiro semestre com *Random Forest* em cursos de engenharia de uma universidade pública espanhola.

Já Naseem, Chaudhary e Sharma (2022), encontraram uma Acurácia menor nos modelos de pré-matrícula, primeiro semestre e segundo semestre, respectivamente de 59,73%, 75,47% e 80,3% em comparação com os dados da presente dissertação, que resultou numa Acurácia de 69,20% no modelo pré-matrícula, 80,47% no modelo primeiro semestre e 84,52% no modelo terceiro semestre, com os algoritmos *Random Forest e Support Vector Machine*. Além disso, a Sensibilidade foi aumentando chegando a 60,79% no terceiro semestre com *Random Forest*, a Especificidade ficou constantemente alta (aproximadamente acima de 85%) desde o modelo de pré-matrícula, o que indicou para este trabalho uma boa previsão da formação dos alunos.

Outros estudos têm abordado o fenômeno complexo da evasão universitária utilizando metodologias diversas, variáveis explicativas distintas e explorando diferentes contextos e cursos. Por exemplo, Huo et al. (2020) investigaram a evasão em cursos superiores não tradicionais, empregando modelos de Regressão Logística e XGBoost em um grande conjunto de dados com 24.770 registros de 1.480 instituições. Eles alcançaram uma Sensibilidade máxima de quase 80%. Santos, Siebra e Oliveira (2014), Costa et al. (2017) e Fernandez-Garcia et al. (2021) utilizaram Sistemas de Gestão da Aprendizagem, como o Moodle e o Ambiente Virtual de Aprendizagem (AVA), para identificar variáveis de interação do aluno EaD com o sistema, como participação em fóruns, discussões, exercícios praticados e horas de acesso. Eles desenvolveram modelos preditivos de evasão em cursos à distância. Hasbun, Araya e Villalon (2016) investigaram como as atividades extracurriculares dos alunos podem ajudar a prever a evasão, utilizando algoritmos de árvore de decisão e 18 variáveis independentes. Eles alcançaram uma Acurácia de 79,29%. Já Sultana, Khan e Abbas (2017) investigaram como a inclusão de atributos não cognitivos pode aumentar a Acurácia dos modelos de Árvore de Decisão em cursos de Engenharia Elétrica. Seus resultados indicaram que a combinação de atributos cognitivos (como desempenho acadêmico) com atributos não cognitivos contribui para melhorar a Acurácia dos modelos preditivos.

No contexto das universidades públicas brasileiras alguns estudos de aprendizagem de máquina foram realizados. Estudos desenvolvidos por Kantorski et al. (2016) preveem a evasão em cursos presencias de universidades públicas com dados pessoais, acadêmicos, sociais e econômicos. Lanes e Alcântara (2018) apresentam um estudo que visa identificar a evasão de aluno no primeiro ano do sistema acadêmico da Universidade Federal do Rio Grande (FURG). Santos et al. (2019) estudaram a previsão em três cursos tecnológicos da Universidade Federal do Sergipe, utilizando seis tipos de algoritmos, do primeiro a o sexto semestre. Costa et al. (2020) apresentaram bons resultados através de *Random Forest* com uma Sensibilidade 91,41% em cursos de Ciência da Computação de 2000 a 2020 na Universidade Federal de Pelotas. Na presente dissertação, a *threshold* foi ajustada no Modelo do Terceiro Semestre com dados de previsão do ano de 2016, entre 16% e 44%, atingindo uma Sensibilidade entre 80% a 89% e uma Acurácia aproximada de 85% o que indica que o ajuste deste parâmetro aumenta a Sensibilidade e a Acurácia do Modelo.

Por fim no Estado de Santa Catarina, Primão (2022) utilizou os algoritmos de Árvore de Decisão, Redes Neurais Artificiais (RNA), *XGBoost* e *Multi Layer Perceptron* (MLP) para comparar a eficiência na previsão de evasão em cursos de graduação presenciais e EaD do Instituto Federal de Santa Catarina (IFSC) em duas bases, uma nos anos pré-pandemia (2017,2018 e 2019) e outra nos anos de pandemia (2020 e 2021), com o algoritmo *XGBoost* performando melhor nas duas bases, apresentando uma Sensibilidade de 97,53% em dados pré-pandemia e 90,32% em dados durante a pandemia. Mioranza (2020) faz um estudo de previsão de evasão nos cursos de graduação do Instituto Federal Catarinense (IFC) nos 14 campi distribuídos no Estado de Santa Catarina, com dados de pré-matrícula e após o primeiro semestre, utilizando os mesmos algoritmos: *Support Vector Machine, Random Forest*, Regressão Logística, KNN e *Naive Bayes*, com resultados de 66% de Acurácia com SVM nos dados pré-matrícula e a maior Especificidade atingida nesses dados antes do início das aulas foi de 66,35% no algoritmo *Naive Bayes* e após o primeiro semestre as métricas se tornaram um pouco mais precisas com 75,52% de Acurácia no algoritmo de Regressão Logística e uma Especificidade máxima de 67,53% com *Naive Bayes*, o que corrobora com a presente dissertação em que na medida que os Modelos avançam com o período dos cursos, as métricas de avaliação se tornam mais precisas na previsão da evasão.

8 Conclusão

Este estudo apresentou os resultados da previsão de evasão de alunos nos cursos de Engenharia do Centro Tecnológico da UFSC em três momentos distintos: pré-matrícula, primeiro semestre e terceiro semestre. Foi realizado um processo de comparação entre diferentes algoritmos de aprendizagem de máquina, seguindo a metodologia CRISP-DM. Portanto, o objetivo geral de realizar a comparação de diferentes algoritmos na previsão de evasão foi alcançado.

Além disso, foi possível conhecer por meio da análise exploratória dos dados as variáveis que mais influenciam na previsão, selecionar estas variáveis independentes e realizar a modelagem com sucesso, permitindo a avaliação pelas métricas selecionadas. Também, ao final, um ajuste dos parâmetros tornou-se necessário para um melhor modelo de aplicação.

Para responder à primeira questão de pesquisa, foram utilizadas duas bases de dados distintas, contendo variáveis socioeconômicas de cadastro, notas do vestibular e índices de desempenho acadêmico nos primeiros e terceiros semestres. Os dados passaram por um pré-processamento e foram aplicados em diferentes algoritmos de modelagem, demonstrando que é possível prever a evasão dos alunos utilizando técnicas de aprendizagem de máquina.

Quanto à segunda questão, observou-se que a acurácia aumenta à medida que o aluno avança em seu percurso universitário. O modelo de *Random Forest* com os dados do terceiro semestre apresentou os melhores resultados, com uma Sensibilidade de 60,79% e Acurácia de 84,52%. Além disso, verificou-se que o balanceamento dos dados aumenta a Sensibilidade do modelo para quase 80%. Testes adicionais foram realizados com dados do ano de 2016, confirmando que os algoritmos apresentaram alta Especificidade desde a fase de pré-matrícula, ou seja, foram mais eficazes na identificação dos alunos que concluem o curso. Também o ajuste do limiar de classificação (*threshold*) entre 16% e 44% apresentou bons resultados aumentando a Sensibilidade e a Acurácia do Modelo.

8.1 Trabalhos Futuros

Devido à complexidade do fenômeno de evasão, que envolve uma variedade de causas sociais, psicológicas, familiares, financeiras, institucionais e de desempenho, sugere-se para estudos futuros o enriquecimento de variáveis independentes, com maior variedade de fatores, incorporando diferentes índices que abordem os diversos aspectos não explorados neste trabalho, como: integração social, fatores psicológicos, contextos familiares, de

moradia e outros aspectos financeiros.

Além disso, outros algoritmos de previsão podem ser usados para testes comparativos como o uso de algoritmos de Redes Neurais Artificiais, a exemplo, *Multi Layer Perceptron* (MLP) e outros algoritmos de classificação como *Boosting*.

Por fim, uma forma de aplicação prática dos modelos para a sua incorporação ao sistema de Controle de Graduação (CAGR) da UFSC como uma ferramenta para se obter um sistema de alerta individual do aluno, identificando quais alunos possuem maiores probabilidades de evasão numa coorte de algum curso específico de forma experimental, podendo verificar se a ferramenta teve êxito na previsão e prevenção desta evasão universitária.

Referências

AGRUSTI, F.; BONAVOLONTà, G.; MEZZINI, M. University Dropout Prediction through Educational Data Mining Techniques: A Systematic Review. **Journal of e-Learning and Knowledge Society**, p. 161–182 Pages, out. 2019. Artwork Size: 161-182 Pages Publisher: Journal of e-Learning and Knowledge Society. Disponível em: <https://www.je-lks.org/ojs/index.php/Je-LKS_EN/article/view/1135017>.

ALVAREZ, N. L.; CALLEJAS, Z.; GRIOL, D. Predicting Computer Engineering students' dropout in Cuban Higher Education with pre-enrollment and early performance data. **Journal of Technology and Science Education**, v. 10, n. 2, p. 241, set. 2020. ISSN 2013-6374. Disponível em: <https://www.jotse.org/index.php/jotse/article/view/922>.

ANDIFES. **Diplomação, Retenção e Evasão nos Cursos de Graduação em Instituições de Ensino Superior Públicas**. [S.l.], 1996. 134 p.

BAKERTILLY. Dados: o novo petróleo do mundo e combustível para o futuro. ago. 2021. Disponível em: <https://bakertillybr.com.br/dados-novo-petroleo/>.

BARDAGI, M.; HUTZ, C. S. Evasão universitária e serviços de apoio ao estudante: uma breve revisão da literatura brasileira. 2005.

BEAN, J. P. Dropouts and turnover: The synthesis and test of a causal model of student attrition. **Research in Higher Education**, v. 12, n. 2, p. 155–187, 1980. ISSN 0361-0365, 1573-188X. Disponível em: <http://link.springer.com/10.1007/BF00976194>.

BIAZUS, C. A. **SISTEMA DE FATORES QUE INFLUENCIAM O ALUNO A EVADIR-SE DOS CURSOS DE GRADUAÇÃO NA UFSM E NA UFSC: UM ESTUDO NO CURSO DE CIÊNCIAS CONTÁBEIS**. Tese (Doutorado) — Universidade Federal de Santa Catarina, 2004. Disponível em: <https://repositorio.ufsc.br/bitstream/handle/123456789/87138/206162.pdf?sequence=1&isAllowed=y>.

BRASIL. **Reestruturação e Expansão das Universidades Federais - Diretrizes Gerais**. [S.l.]: Plano de Desenvolvimento da Educação. Brasília-DF, 2007.

_____. PLANO NACIONAL DE EDUCAÇÃO PNE 2014-2024 LINHA DE BASE. **Brasília-DF**, p. 404, 2015. Disponível em: <https://download.inep.gov.br/publicacoes/institucionais/plano_nacional_de_educacao/plano_nacional_de_educacao_pne_2014_2024_linha_de_base.pdf>.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001. ISSN 08856125. Disponível em: <http://link.springer.com/10.1023/A:1010933404324>.

CABRERA, A. F.; NORA, A.; TERENZINI, P. T.; PASCARELLA, E.; HAGEDORN, L. S. Campus Racial Climate and the Adjustment of Students to College: A Comparison between White Students and African-American Students. **The Journal of Higher Education**, v. 70, n. 2, p. 134–160, mar. 1999. ISSN 0022-1546, 1538-4640. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/00221546.1999.11780759>.

CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. **Computers & Electrical Engineering**, v. 40, n. 1, p. 16–28, jan. 2014. ISSN 00457906. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0045790613003066>.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R. **CRISP-DM: Step-by-step data mining guide**. [S.l.], 2000. Disponível em: <http://www.statoo.com/CRISP-DM.pdf>.

CHATGPT. **Pergunta ao ChatGPT: "Qual o tipo de algoritmo você utiliza?"**. 2023. Disponível em: <https://openai.com/blog/chatgpt>.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, jun. 2002. ISSN 1076-9757. Disponível em: <https://www.jair.org/index.php/jair/article/view/10302>.

CHRISTO, M. M. S.; RESENDE, L. M. M. d.; KUHN, T. D. C. G. POR QUE OS ALUNOS DE ENGENHARIA DESISTEM DE SEUS CURSOS – UM ESTUDO DE CASO. **Nuances: estudos sobre Educação**, v. 29, n. 1, dez. 2018. ISSN 2236-0441, 1413-9855. Disponível em: <http://revista.fct.unesp.br/index.php/Nuances/article/view/4391>.

COIMBRA, C. L.; SILVA, L. B. e.; COSTA, N. C. D. A evasão na educação superior: definições e trajetórias. **Educação e Pesquisa**, v. 47, p. e228764, 2021. ISSN 1678-4634, 1517-9702. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1517-97022021000100713&tlng=pt>.

COSTA, A. G.; QUEIROGA, E.; PRIMO, T. T.; MATTOS, J. C. B.; CECHINEL, C. Prediction analysis of student dropout in a Computer Science course using Educational Data Mining. In: **2020 XV Conferencia Latinoamericana de Tecnologias de Aprendizaje (LACLO)**. Loja, Ecuador: IEEE, 2020. p. 1–6. ISBN 978-1-72819-268-0. Disponível em: <https://ieeexplore.ieee.org/document/9381166/>.

COSTA, E. B.; FONSECA, B.; SANTANA, M. A.; ARAúJO, F. F. de; REGO, J. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. **Computers in Human Behavior**, v. 73, p. 247–256, ago. 2017. ISSN 07475632. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0747563217300596>.

CRAMER, J. The Origins of Logistic Regression. **SSRN Electronic Journal**, 2003. ISSN 1556-5068. Disponível em: <http://www.ssrn.com/abstract=360300>.

DORRIS, D.; SWANN, J.; IVY, J. A Data-driven Approach for Understanding and Predicting Engineering Student Dropout. In: **2021 ASEE Virtual Annual Conference Content Access Proceedings**. Virtual Conference: ASEE Conferences, 2021. p. 36575. Disponível em: <http://peer.asee.org/36575>.

DSA. 17 Casos de Uso de Machine Learning. **Data Science Academy**, mar. 2022. Disponível em: <https://blog.dsacademy.com.br/17-casos-de-uso-de-machine-learning/>.

ETHINGTON, C. A. A psychological model of student persistence. **Research in Higher Education**, v. 31, n. 3, p. 279–293, jun. 1990. ISSN 0361-0365, 1573-188X. Disponível em: <http://link.springer.com/10.1007/BF00992313>.

FERNANDEZ-GARCIA, A. J.; PRECIADO, J. C.; MELCHOR, F.; RODRIGUEZ-ECHEVERRIA, R.; CONEJERO, J. M.; SANCHEZ-FIGUEROA, F. A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data. **IEEE Access**, v. 9, p. 133076–133090, 2021. ISSN 2169-3536. Disponível em: <https://ieeexplore.ieee.org/document/9548895/>.

FERNáNDEZ-MARTíN, T.; SOLíS-SALAZAR, M.; HERNáNDEZ-JIMéNEZ, M. T.; MOREIRA-MORA, T. E. A Multinomial and Predictive Analysis of Factors Associated with University Dropout. **Revista Electrónica Educare**, v. 23, n. 1, jun. 2018. ISSN 1409-4258. Disponível em: <http://www.revistas.una.ac.cr/index.php/EDUCARE/article/view/9038>.

FILHO, A. Bastos do C.; VALLE, E. Cordeiro do; ARAUJO, L. L.; PIETTA, E. V. Z.; GONçALVES, L. P. Retenção e Evasão em Cursos de Engenharia: uso de tecnologia para proporcionar a Aprendizagem Social. **RENOTE**, v. 20, n. 1, p. 273–283, ago. 2022. ISSN 1679-1916. Disponível em: <https://seer.ufrgs.br/index.php/renote/article/view/126673>.

FILHO, R. L. L. e. S.; MOTEJUNAS, P. R.; HIPóLITO, O.; LOBO, M. B. d. C. M. A evasão no ensino superior brasileiro. **Cadernos de Pesquisa**, v. 37, n. 132, p. 641–659, dez. 2007. ISSN 0100-1574. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-15742007000300007&lng=pt&tlng=pt>.

FLORES, V.; HERAS, S.; JULIAN, V. Comparison of Predictive Models with Balanced Classes Using the SMOTE Method for the Forecast of Student Dropout in Higher Education. **Electronics**, v. 11, n. 3, p. 457, fev. 2022. ISSN 2079-9292. Disponível em: <https://www.mdpi.com/2079-9292/11/3/457>.

FORPLAD. **GT - Taxa de Sucesso, Evasão e Retenção nas IFES**. VITÓRIA-ES: [s.n.], 2016. Disponível em: <http://docplayer.com.br/86157311-Forplad-gt-taxa-de-sucesso-evasao-e-retencao-nas-ifes.html>.

FREITAS, R. S. **A ocorrência da evasão do ensino superior: uma análise das diferentes formas de mensurar**. Tese (Mestre em Educação) — Universidade Estadual de Campinas, Campinas, abr. 2016. Disponível em: <http://acervus.unicamp.br/index.asp?codigo_sophia=970585>.

GOLLAPUDI, S. **Practical Machine Learning**. [S.l.: s.n.], 2016.

GéRON, A. **Mãos à Obra Aprendizado de Máquina com Scikit-Learn e TensorFlow**. [S.l.]: Alta Books, 2019. OCLC: 1162720427. ISBN 978-85-508-0381-4.

HACKELING, G. **Mastering machine learning with scikit-learn: apply effective learning algorithms to real-world problems using scikit-learn**. Birmingham: Packt Publ, 2014. (Packt open source). ISBN 978-1-78398-836-5.

HANNAFORD, L.; CHENG, X.; KUNES-CONNELL, M. Predicting nursing baccalaureate program graduates using machine learning models: A quantitative research study. **Nurse Education Today**, v. 99, p. 104784, abr. 2021. ISSN 02606917. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0260691721000411>.

HASBUN, T.; ARAYA, A.; VILLALON, J. Extracurricular Activities as Dropout Prediction Factors in Higher Education Using Decision Trees. In: **2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)**. Austin, TX, USA: IEEE, 2016. p. 242–244. ISBN 978-1-4673-9041-5. Disponível em: <http://ieeexplore.ieee.org/document/7756969/>.

HEARST, M.; DUMAIS, S.; OSUNA, E.; PLATT, J.; SCHOLKOPF, B. Support vector machines. **IEEE Intelligent Systems and their Applications**, v. 13, n. 4, p. 18–28, jul. 1998. ISSN 1094-7167. Disponível em: <http://ieeexplore.ieee.org/document/708428/>.

HINTON, G. E.; OSINDERO, S.; TEH, Y.-W. A Fast Learning Algorithm for Deep Belief Nets. **Neural Computation**, v. 18, n. 7, p. 1527–1554, jul. 2006. ISSN 0899-7667, 1530-888X. Disponível em: <https://direct.mit.edu/neco/article/18/7/1527-1554/7065>.

HINTON, G. E.; SALAKHUTDINOV, R. R. Reducing the Dimensionality of Data with Neural Networks. **Science**, v. 313, n. 5786, p. 504–507, jul. 2006. ISSN 0036-8075, 1095-9203. Disponível em: <https://www.science.org/doi/10.1126/science.1127647>.

HUO, H.; CUI, J.; HEIN, S.; PADGETT, Z.; OSSOLINSKI, M.; RAIM, R.; ZHANG, J. Predicting Dropout for Nontraditional Undergraduate Students: A Machine Learning Approach. **Journal of College Student Retention: Research, Theory & Practice**, v. 24, n. 4, p. 1054–1077, fev. 2020. ISSN 1521-0251, 1541-4167. Disponível em: <http://journals.sagepub.com/doi/10.1177/1521025120963821>.

H.WAGNER, C. Simpson's paradox in real life. **The American Statistician**, Taylor Francis, v. 36, n. 1, p. 46–48, 1982. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/00031305.1982.10482778>.

INEP. **Metodologia de Cálculo dos Indicadores de Fluxo da Educação Superior**. [S.l.]: Brasília, 2017.

_____. **Apresentação do Censo da Educação Superior 2021**. Brasília-DF: [s.n.], 2022.

_____. **Censo da Educação Superior 2021 - Divulgação dos resultados**. Brasília-DF: [s.n.], 2022.

JUNIOR, P. L.; BISINOTO, C.; MELO, N. S. d.; RABELO, M. Taxas longitudinais de retenção e evasão: uma metodologia para estudo da trajetória dos estudantes na educação superior. **Ensaio: Avaliação e Políticas Públicas em Educação**, v. 27, n. 102, p. 157–178, mar. 2019. ISSN 1809-4465, 0104-4036. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-40362019000100157&tlng=pt>.

KANTORSKI, G.; FLORES, E. G.; SCHMITT, J.; HOFFMANN, I.; BARBOSA, F. Predição da Evasão em Cursos de Graduação em Instituições Públicas. In: . Uberlandia, Minas Gerais, Brasil: [s.n.], 2016. p. 906. Disponível em: <http://www.br-ie.org/pub/index.php/sbie/article/view/6776>.

KEMPER, L.; VORHOFF, G.; WIGGER, B. U. Predicting student dropout: A machine learning approach. **European Journal of Higher Education**, v. 10, n. 1, p. 28–47, jan. 2020. ISSN 2156-8235, 2156-8243. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/21568235.2020.1718520>.

KUHN, M. **The caret Package**. 2019. Disponível em: <https://topepo.github.io/caret/index.html>.

KURSA, M. B.; RUDNICKI, W. R. Feature Selection with the **Boruta** Package. **Journal of Statistical Software**, v. 36, n. 11, 2010. ISSN 1548-7660. Disponível em: <http://www.jstatsoft.org/v36/i11/>.

LANES, M.; ALCâNTARA, C. Predição de Alunos com Risco de Evasão: estudo de caso usando mineração de dados. In: . Fortaleza, Ceará, Brasil: [s.n.], 2018. p. 1921. Disponível em: <http://br-ie.org/pub/index.php/sbie/article/view/8191>.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, maio 2015. ISSN 0028-0836, 1476-4687. Disponível em: <http://www.nature.com/articles/nature14539>.

LIZ-DOMíNGUEZ, M.; CAEIRO-RODRíGUEZ, M.; LLAMAS-NISTAL, M.; MIKIC-FONTE, F. A. Systematic Literature Review of Predictive Analysis Tools in Higher Education. **Applied Sciences**, v. 9, n. 24, p. 5569, dez. 2019. ISSN 2076-3417. Disponível em: <https://www.mdpi.com/2076-3417/9/24/5569>.

LOBO, M. B. d. C. M. PANORAMA DA EVASÃO NO ENSINO SUPERIOR BRASILEIRO: ASPECTOS GERAIS DAS CAUSAS E SOLUÇÕES. **São Paulo.**, dez. 2012. Disponível em: <https://www.institutolobo.org.br/core/uploads/artigos/art_087.pdf>.

LOBO, R. **O ACOLHIMENTO DO ESTUDANTE E A EVASÃO**. 2019. Disponível em: <https://www.institutolobo.org.br/core/uploads/artigos/anexo_0bb0fc8e6e7f0b695673cb173caa8f83.pdf>.

LUDER, A. Quase 3,5 milhões de alunos evadiram de universidades privadas no Brasil em 2021. **G1 - Educação**, jan. 2022. Disponível em: <https://g1.globo.com/educacao/noticia/2022/01/02/quase-35-milhoes-de-alunos-evadiram-de-universidades-privadas-no-brasil-em-2021.ghtml>.

MADLEY-DOWD, P.; HUGHES, R.; TILLING, K.; HERON, J. The proportion of missing data should not be used to guide decisions on multiple imputation. v. 110, p. 63–73, 2019. ISSN 08954356. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0895435618308710>.

MEC. **A democratização e expansão da educação superior no país 2003 – 2014**. [S.l.], 2014. Disponível em: <http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=16762-balanco-social-sesu-2003-2014&Itemid=30192>.

MIORANZA, D. **EVASÃO NOS CURSOS DE GRADUAÇÃO DO INSTITUTO FEDERAL CATARINENSE: UM ESTUDO A PARTIR DA MINERAÇÃO DE DADOS**. Tese (Dissertação (mestrado profissional)) — Universidade Federal de Santa Catarina, Centro Tecnológico, 2020. Disponível em: <https://repositorio.ufsc.br/bitstream/handle/123456789/216662/PMGA0051-D.pdf?sequence=-1&isAllowed=y>.

MIRANDA, M. A.; GUZMáN, J. Análisis de la Deserción de Estudiantes Universitarios usando Técnicas de Minería de Datos. **Formación universitaria**, v. 10, n. 3, p. 61–68,

2017. ISSN 0718-5006. Disponível em: <http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-50062017000300007&lng=en&nrm=iso&tlng=en>.

MODERNA. **Anuário Brasileiro da Educação Básica**. 2021. Disponível em: <https://www.moderna.com.br/anuario-educacao-basica/2020/educacao-superior.html>.

MONACO, R. **Mais da metade dos estudantes abandona cursos de engenharia**. 2013. Disponível em: <https://noticias.portaldaindustria.com.br/noticias/educacao/mais-da-metade-dos-estudantes-abandona-cursos-de-engenharia/>.

NAMOUN, A.; ALSHANQITI, A. Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review. **Applied Sciences**, v. 11, n. 1, p. 237, dez. 2020. ISSN 2076-3417. Disponível em: <https://www.mdpi.com/2076-3417/11/1/237>.

NASEEM, M.; CHAUDHARY, K.; SHARMA, B. Predicting Freshmen Attrition in Computing Science using Data Mining. **Education and Information Technologies**, v. 27, n. 7, p. 9587–9617, ago. 2022. ISSN 1360-2357, 1573-7608. Disponível em: <https://link.springer.com/10.1007/s10639-022-11018-3>.

NILSSON, R.; BJORKEGREN, J.; TEGNER, J. Consistent Feature Selection for Pattern Recognition in Polynomial Time. 2007.

NWANGANGA, F.; CHAPPLE, M. **Practical machine learning in r:
**. 1st ed. ed. Indianapolis: John Wiley and Sons, 2020. ISBN 978-1-119-59151-1.

OECD. **Education at a Glance 2022: OECD Indicators**. OECD, 2022. (Education at a Glance). ISBN 978-92-64-58258-3 978-92-64-95055-9 978-92-64-34164-7 978-92-64-59292-6. Disponível em: <https://www.oecd-ilibrary.org/education/education-at-a-glance-2022_3197152b-en>.

OLIVEIRA, C. F. de; SOBRAL, S. R.; FERREIRA, M. J.; MOREIRA, F. How Does Learning Analytics Contribute to Prevent Students' Dropout in Higher Education: A Systematic Literature Review. **Big Data and Cognitive Computing**, v. 5, n. 4, p. 64, nov. 2021. ISSN 2504-2289. Disponível em: <https://www.mdpi.com/2504-2289/5/4/64>.

OPAZO, D.; MORENO, S.; MIRANDA, E. Álvarez; PEREIRA, J. Analysis of First-Year University Student Dropout through Machine Learning Models: A Comparison between Universities. **Mathematics**, v. 9, n. 20, p. 2599, out. 2021. ISSN 2227-7390. Disponível em: <https://www.mdpi.com/2227-7390/9/20/2599>.

PAGE, M. J. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. p. n160, 2021. ISSN 1756-1833. Disponível em: <https://www.bmj.com/lookup/doi/10.1136/bmj.n160>.

PRAE. **Pró-Reitoria de Assuntos Estudantis**. 2022. Disponível em: <https://prae.ufsc.br/>.

PRIMãO, A. P. **USO DE ALGORITMOS DE MACHINE LEARNING PARA PREVER A EVASÃO ESCOLAR NO ENSINO SUPERIOR: UM ESTUDO NO INSTITUTO FEDERAL DE SANTA CATARINA**. Tese (Doutorado), 2022. Disponível em: <https://repositorio.ufsc.br/bitstream/handle/123456789/238320/PPAU0264-D.pdf?sequence=-1&isAllowed=y>.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2022. Disponível em: <https://www.R-project.org/>.

RASCHKA, S. **Naive Bayes and Text Classification I - Introduction and Theory**. arXiv, 2017. ArXiv:1410.5329 [cs]. Disponível em: <http://arxiv.org/abs/1410.5329>.

RASCHKA, S.; OLSON, R. S. **Python machine learning: unlock deeper insights into machine learning with this vital guide to cutting-edge predictive analytics**. Birmingham Mumbai: Packt Publishing open source, 2015. (Open source community experience distilled). ISBN 978-1-78355-513-0.

RISTOFF, D. Evasão: exclusão ou mobilidade. 1995.

SANTOS, K. J. de O.; MENEZES, A. G.; CARVALHO, A. B. de; MONTESCO, C. A. E. Supervised Learning in the Context of Educational Data Mining to Avoid University Students Dropout. In: **2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)**. Maceió, Brazil: IEEE, 2019. p. 207–208. ISBN 978-1-72813-485-7. Disponível em: <https://ieeexplore.ieee.org/document/8820813/>.

SANTOS, R. N. d.; SIEBRA, C. d. A.; OLIVEIRA, E. S. Uma Abordagem Temporal para Identificação Precoce de Estudantes de Graduação a Distância com Risco de Evasão em um AVA utilizando Árvores de Decisão. In: . Dourados, Mato Grosso do Sul, Brasil: [s.n.], 2014. p. 262. Disponível em: <http://br-ie.org/pub/index.php/wcbie/article/view/3224>.

SEMESP. Mapa do Ensino Superior no Brasil. **Instituto Semesp**, n. 12ª, 2022.

SILVERMAN, B. W.; JONES, M. C. E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951). **International Statistical Review / Revue Internationale de Statistique**, v. 57, n. 3, p. 233, dez. 1989. ISSN 03067734. Disponível em: <https://www.jstor.org/stable/1403796?origin=crossref>.

SPADY, W. G. Dropouts from higher education: An interdisciplinary review and synthesis. **Interchange**, v. 1, n. 1, p. 64–85, abr. 1970. ISSN 0826-4805, 1573-1790. Disponível em: <http://link.springer.com/10.1007/BF02214313>.

SULTANA, S.; KHAN, S.; ABBAS, M. A. Predicting performance of electrical engineering students using cognitive and non-cognitive features for identification of potential dropouts. **The International Journal of Electrical Engineering & Education**, v. 54, n. 2, p. 105–118, abr. 2017. ISSN 0020-7209, 2050-4578. Disponível em: <http://journals.sagepub.com/doi/10.1177/0020720916688484>.

TCU. **ORIENTAÇÕES PARA O CÁLCULO DOS INDICADORES DE GESTÃO. Decisão TCU Nº 408/2002**. [S.l.]: Tribunal de Contas da União, 2004.

TENPIPAT, W.; AKKARAJITSAKUL, K. Student Dropout Prediction: A KMUTT Case Study. In: **2020 1st International Conference on Big Data Analytics and Practices (IBDAP)**. Bangkok, Thailand: IEEE, 2020. p. 1–5. ISBN 978-1-72818-106-6. Disponível em: <https://ieeexplore.ieee.org/document/9245457/>.

THELIN, J. R. The Attrition Tradition in American Higher Education: Connecting Past and Present. 2010.

TINTO, V. Dropout from Higher Education: A Theoretical Synthesis of Recent Research. **REVIEW OF EDUCATIONAL RESEARCH**, v. 45, n. 1, 1975.

UFSC. **Indicadores de Ensino PDI 2020-2024**. 2020. Disponível em: <https://pdi.paginas.ufsc.br/files/2020/02/1.-Indicadores-ENSINO-2020-v.3.pdf>.

_____. **Relatório de Gestão 2021**. [S.l.], 2021. 256 p. Disponível em: <http://dpgi.seplan.ufsc.br>.

UFSC. **UFSC Série Histórica 1980-2021**. 2021. Disponível em: <https://dpgi-seplan.ufsc.br/files/2021/08/S%C3%A9rie-Hist%C3%B3rica-2008-2021.pdf>.

URBINA-NáJERA, A. B.; CAMINO-HAMPSHIRE, J. C.; BARBOSA, R. C. Deserción escolar universitaria: Patrones para prevenirla aplicando minería de datos educativa. **RELIEVE - Revista Electrónica de Investigación y Evaluación Educativa**, v. 26, n. 1, out. 2020. ISSN 1134-4032. Disponível em: <https://revistaseug.ugr.es/index.php/RELIEVE/article/view/17345>.

ZHANG, Y.; YUN, Y.; AN, R.; CUI, J.; DAI, H.; SHANG, X. Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis. **Frontiers in Psychology**, v. 12, p. 698490, dez. 2021. ISSN 1664-1078. Disponível em: <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.698490/full>.

APÊNDICE A – Boruta Selection

Tabela 18 – **Variáveis Selecionadas por ordem de normHits Boruta**

Variável	normHits	Decisão	TentativeRoughFix
curso	1,00000	Confirmed	Confirmed
idade	1,00000	Confirmed	Confirmed
q17_Sal.Bruto.fam	1,00000	Confirmed	Confirmed
q20_Sustento.fam	1,00000	Confirmed	Confirmed
acertos_BLG	1,00000	Confirmed	Confirmed
acertos_MTM	1,00000	Confirmed	Confirmed
acertos_FSC	1,00000	Confirmed	Confirmed
acertos_QMC	1,00000	Confirmed	Confirmed
acertos_LLE	1,00000	Confirmed	Confirmed
acertos_PTG	1,00000	Confirmed	Confirmed
sexo	0,98990	Confirmed	Confirmed
q34_Motivo.vest	0,98990	Confirmed	Confirmed
q11_Motivo.pre_vest	0,96970	Confirmed	Confirmed
q24_Meio.info	0,95960	Confirmed	Confirmed
q7_Tipo.EM	0,94949	Confirmed	Confirmed
q18_Instruc.Pai	0,92929	Confirmed	Confirmed
q10_Curso.pre_vest	0,91919	Confirmed	Confirmed
q23_Sua.ocupac	0,80808	Confirmed	Confirmed
q8_Turno.EM	0,75758	Confirmed	Confirmed
q14_Obter.curso	0,75758	Confirmed	Confirmed
q19_Instruc.Mae	0,75758	Confirmed	Confirmed
q13_Motivo.escolha	0,71717	Confirmed	Confirmed
q2_UF.resid	0,70707	Confirmed	Confirmed
q6_TipoCurso.EM	0,62626	Tentative	Confirmed
q32_Hobbie	0,62626	Tentative	Confirmed
q1_estado.civil	0,58586	Tentative	Confirmed
q16_Pessoas.fam	0,56566	Tentative	Confirmed
q22_Idade.Ativ.rem	0,55556	Tentative	Confirmed
q4_Tipo.EF	0,53535	Tentative	Confirmed
q28_Iniciou.ES	0,51515	Tentative	Confirmed
q12_NumeroXvest	0,45455	Tentative	Rejected
q3_UF.EF	0,43434	Tentative	Rejected
q26_Usa.computador	0,43434	Tentative	Rejected
q30_Curso.ES	0,43434	Tentative	Rejected
q29_Inst.ES	0,42424	Tentative	Rejected
q5_UF_EM	0,41414	Tentative	Rejected
q25_Computador	0,37374	Tentative	Rejected
q33_Esporte	0,36364	Tentative	Rejected
q27_Transporte	0,19192	Rejected	Rejected
q9_Info.Vest	0,01010	Rejected	Rejected
q15_Ativ.profissao	0,01010	Rejected	Rejected
q31_Orienta.voca	0,01010	Rejected	Rejected

Continua...

Tabela 18 – ***Variáveis Selecionadas por ordem de normHits***

Variável	normHits	Decisão	TentativeRoughFix
q21_Ocupac.sustento	0,00000	Rejected	Rejected

Fonte: o autor

ANEXO A – Questionário Socioeconômico

Tabela 19 – **Questionário Socioeconômico Inscrição Vestibular UFSC**

Questão	Descrição	Respostas
q1	ESTADO CIVIL	1) Solteiro; 2) Casado; 3) Viúvo; 4) Desquitado; 5) Divorciado; 6) Outros
q2	UNIDADE DA FEDERAÇÃO EM QUE VOCÊ RESIDE	1) Acre; 2)Alagoas; 3) Amapá; 4) Amazonas; 5) Bahia; 6) Ceará; 7) Distrito Federal; 8) Espírito Santo; 9) Goiás; 10) Maranhão; 11) Mato Grosso; 12) Mato Grosso do Sul; 13) Minas Gerais;14) Pará;15) Paraíba;16) Paraná;17) Pernambuco;18) Piauí;19) Rio Grande do Norte;20) Rio Grande do Sul; 21) Rio de Janeiro;22) Rondônia;23) Roraima; 24) Santa Catarina;25) São Paulo;26) Sergipe;27) Tocantins;28) Outros países
q3	UNIDADE DA FEDERAÇÃO EM QUE VOCÊ CONCLUIU O ENSINO FUNDAMENTAL	1) Acre; 2)Alagoas; 3) Amapá; 4) Amazonas; 5) Bahia; 6) Ceará; 7) Distrito Federal; 8) Espírito Santo; 9) Goiás; 10) Maranhão; 11) Mato Grosso; 12) Mato Grosso do Sul; 13) Minas Gerais;14) Pará;15) Paraíba;16) Paraná;17) Pernambuco;18) Piauí;19) Rio Grande do Norte;20) Rio Grande do Sul; 21) Rio de Janeiro;22) Rondônia;23) Roraima; 24) Santa Catarina;25) São Paulo;26) Sergipe;27) Tocantins;28) Outros países
q4	TIPO DE ESTABELECIMENTO ONDE VOCÊ CURSOU O ENSINO FUNDAMENTAL	1)Todo em Escola Pública;2)Todo em Escola Particular;3)Maior parte em Escola Pública;4)Maior parte em Escola Particular;5)Escolas Comunitárias;6)Outros
q5	UNIDADE DA FEDERAÇÃO EM QUE VOCÊ CONCLUIU OU CONCLUIRÁ O ENSINO MÉDIO	1) Acre; 2)Alagoas; 3) Amapá; 4) Amazonas; 5) Bahia; 6) Ceará; 7) Distrito Federal; 8) Espírito Santo; 9) Goiás; 10) Maranhão; 11) Mato Grosso; 12) Mato Grosso do Sul; 13) Minas Gerais;14) Pará;15) Paraíba;16) Paraná;17) Pernambuco;18) Piauí;19) Rio Grande do Norte;20) Rio Grande do Sul; 21) Rio de Janeiro;22) Rondônia;23) Roraima; 24) Santa Catarina;25) São Paulo;26) Sergipe;27) Tocantins;28) Outros países

Continua...

Tabela 19 – ***Questionário Socioeconômico Inscrição Vestibular UFSC***

Questão	Descrição	Respostas
q6	TIPO DE CURSO DE ENSINO MÉDIO QUE VOCÊ CONCLUIU OU CONCLUIR	1) Regular, sem ênfase em qualquer área; 2) Profissionalizante na área agrícola; 3) Profissionalizante na área de comércio ou serviços; 4) Profissionalizante na área de saúde;5) Profissionalizante na área de magistério; 6) Profissionalizante na área industrial;7) Supletivo ou Madureza;8) Outros
q7	TIPO DE ESTABELECIMENTO ONDE VOCÊ CURSOU O ENSINO MÉDIO	1)Todo em Escola Pública;2)Todo em Escola Particular;3)Maior parte em Escola Pública; 4)Maior parte em Escola Particular; 5)Escolas Comunitárias; 6)Outros
q8	TURNO EM QUE VOCÊ CURSOU O ENSINO MÉDIO	1)Todo diurno; 2)Todo noturno; 3)Maior parte no diurno; 4)Maior parte no noturno
q9	MARQUE A PRINCIPAL FONTE DE INFORMAÇÃO PELA QUAL VOCÊ TOMOU CONHECIMENTO DO CONCURSO VESTIBULAR DA UFSC	1)Jornal; 2)Televisão; 3)Rádio; 4)Cartaz ou *folder*; 5)Colégio em que estuda; 6)Internet; 7)Parentes, amigos
q10	FREQUENTOU OU FREQUENTA CURSO PRÉ-VESTIBULAR	1) Não; 2) Sim, por menos de 1 semestre; 3) Sim, por 1 semestre; 4) Sim, por 1 ano; 5) Sim, por mais de 1 ano
q11	PRINCIPAL MOTIVO QUE O LEVOU A NÃO CURSAR PRÉ-VESTIBULAR	1) O Colégio prepara para o Vestibular; 2) O Colégio oferece pré-vestibular "integrado"ao curso;3) Dificuldades econômicas; 4) O horário do pré-vestibular coincidia com o horário de trabalho; 5) Julgou que poderia estudar sozinho; 6) Não havia nenhum pré-vestibular nas proximidades da sua casa; 7) Não é o seu caso (fez pré-vestibular)
q12	NÚMERO DE VEZES QUE VOCÊ PRESTOU VESTIBULAR PARA A UFSC	1)Nenhuma; 2)Uma; 3)Duas; 4)Três; 5)Quatro ou mais
q13	PRINCIPAL MOTIVO PARA ESCOLHA DE SUA 1ª OPÇÃO	1) Menor relação candidato/vaga; 2) Prestígio econômico; 3) Prestígio social; 4) Mais adequada às suas aptidões; 5) Influência da família e/ou amigos; 6) Influência de professores; 7) Outros motivos
q14	ASSINALE O QUE VOCÊ ESPERA OBTER NUM CURSO SUPERIOR	1) Aumento de conhecimento e cultura geral; 2) Melhoria da situação profissional atual; 3)Formação profissional voltada para o futuro emprego; 4)Formação teórica voltada para a pesquisa; 5)Outras
q15	CONHECE AS ATIVIDADES QUE DEVERÁ DESENVOLVER NA PROFISSÃO ESCOLHIDA EM 1ª OPÇÃO	1)Sim; 2)Não

Continua...

Tabela 19 – ***Questionário Socioeconômico Inscrição Vestibular UFSC***

Questão	Descrição	Respostas
q16	INCLUINDO SOMENTE OS QUE MORAM NA SUA CASA, INCLUSIVE VOCÊ, INFORME O NÚMERO DE PESSOAS QUE COMPÕEM A SUA FAMÍLIA	1) 1 pessoa; 2) 2 pessoas; 3) 3 pessoas; 4) 4 pessoas; 5) 5 pessoas; 6) Acima de 5 pessoas
q17	SOME OS SAL. BRUTOS, SEM DEDUÇÕES, DAS PESSOAS DE SEU GRUPO FAM. QUE TRABALHAM, INCLUSIVE O SEU, INDIQUE A RENDA BRUTA	1) Até 1 salário mínimo; 2) Acima de 1 até 3 sal. mín.; 3) Acima de 3 até 5 sal. mín.; 4) Acima de 5 até 7 sal. mín.; 5) Acima de 7 até 10 sal. mín.; 6) Entre 10 e 20 sal. mín.; 7) Entre 20 e 30 sal. mín.; 8) Acima de 30 sal. mín.;
q18	NÍVEL DE INSTRUÇÃO DE SEU PAI	1) Não alfabetizado; 2) Lê e escreve, mas nunca esteve na escola; 3) Fundamental incompleto; 4) Fundamental completo; 5) Médio incompleto; 6) Médio completo; 7) Superior incompleto; 8) Superior completo; 9)Pós-Graduação
q19	NÍVEL DE INSTRUÇÃO DE SUA MÃE	1) Não alfabetizado; 2) Lê e escreve, mas nunca esteve na escola; 3) Fundamental incompleto; 4) Fundamental completo; 5) Médio incompleto; 6) Médio completo; 7) Superior incompleto; 8) Superior completo; 9)Pós-Graduação
q20	INDIQUE O PRINCIPAL RESPONSÁVEL PELO SUSTENTO DA SUA FAMÍLIA	1) Pai; 2) Mãe; 3) Pai e Mãe; 4) Você próprio; 5) Cônjuge; 6)Parente; 7)Outro(s)
q21	PRINCIPAL OCUPAÇÃO DO RESPONSÁVEL PELO SUSTENTO DA SUA FAMÍLIA	1) Profissional liberal; 2)Empresário; 3) Servidor Público; 4)Empregado Empresa Privada; 5)Empregado Rural / Agricultor; 6) Proprietário Rural; 7) Não trabalha; 8) Desempregado; 9) Outro
q22	IDADE COM QUE COMEÇOU A EXERCER A ATIVIDADE REMUNERADA	1) Antes de 14 anos; 2) Entre 14 e 16 anos; 3) Entre 16 e 18 anos; 4) Após 18 anos; 5) Nunca trabalhou
q23	SUA OCUPAÇÃO	1) Profissional liberal; 2)Empresário; 3) Servidor Público; 4)Empregado Empresa Privada; 5)Empregado Rural / Agricultor; 6) Proprietário Rural; 7) Não trabalha; 8) Desempregado; 9) Outro
q24	MARQUE O PRINCIPAL MEIO DE COMUNICAÇÃO QUE VOCÊ UTILIZA PARA SE MANTER INFORMADO SOBRE OS ACONTECIMENTOS ATUAIS	1) Jornal; 2) Televisão; 3) Rádio; 4) Revista; 5) Internet; 6) Conversas com outras pessoas; 7) Não tenho me mantido informado
q25	POSSUI COMPUTADOR EM SUA RESIDÊNCIA	1) Sim, com acesso à Internet; 2) Sim, sem acesso à Internet; 3) Não
q26	USA COMPUTADOR	1) Sim, só para lazer; 2) Sim, para trabalhos escolares e/ou profissionais; 3) Sim, no trabalho; 4) Não
q27	MEIO DE TRANSPORTE QUE VOCÊ MAIS UTILIZA	1) Bicicleta; 2) Carro próprio ou da família; 3) Moto; 4) Ônibus; 5) Outros

Continua...

Tabela 19 – ***Questionário Socioeconômico Inscrição Vestibular UFSC***

Questão	Descrição	Respostas
q28	INICIOU ALGUM CURSO SUPERIOR	1) Sim, mas abandonei; 2) Sim, estou cursando; 3) Sim, mas já concluí; 4) Sim, já concluí um e estou cursando outro; 5) Sim, já concluí um e abandonei outro; 6) Não
q29	INSTITUIÇÃO NA QUAL INICIOU ALGUM CURSO SUPERIOR	1) Não iniciei; 2) UFSC; 3) Outra Instituição de Santa Catarina; 4) Outra Instituição fora de Santa Catarina
q30	INFORME O CURSO SUPERIOR JÁ INICIADO	1) Não iniciei curso superior; 2) Administração; 3) Agronomia; 4) Arquitetura e Urbanismo; 5 Biblioteconomia; 6 Ciências Biológicas; 7 Ciências da Computação; 8 Ciências Contábeis; 9 Ciências Sociais; 10 Direito; 11 Ciências Econômicas; 12 Educação Física; 13 Enfermagem; 14 Engenharia; 15 Farmácia; 16 Filosofia; 17 Física; 18 Geografia; 19 História; 20 Jornalismo; 21 Comunicação e Expressão Visual; 22 Letras; 23 Matemática; 24 Medicina; 25 Nutrição; 26 Odontologia; 27 Pedagogia; 28 Psicologia; 29 Química; 30 Serviço Social; 31 Veterinária; 32 Outros
q31	ACREDITA QUE ORIENTAÇÃO VOCACIONAL AUXILIARIA NA ESCOLHA DE SUA OPÇÃO	1) Sim, para dizer qual profissão devo seguir; 2) Sim, para auxiliar a conhecer as profissões e o mercado de trabalho; 3) Sim, para auxiliar a pensar na melhor opção para mim; 4) Sim, para ajudar a me conhecer melhor; 5) Sim, para auxiliar a conviver com a família e a sociedade e assumir o que quero; 6) Não
q32	DOS ITENS ABAIXO, ASSINALE SUA PREFERÊNCIA	1) Artes Plásticas/Artesanato; 2) Cinema/Vídeo; 3) Dança; 4) Música; 5) Teatro; 6) Literatura; 7) Esporte; 8) Outros
q33	INDIQUE SEU ESPORTE PREDILETO	1) Basquete; 2) Capoeira; 3) Caratê; 4) Futebol de campo; 5) Futebol de salão; 6 Futebol suíço; 7 Handebol; 8 Judô; 9 Natação; 10 Polo aquático; 11 Surf; 12 Tênis de campo; 13 Vela; 14 Voleibol; 15 Voleibol de areia; 16 Remo; 17 Xadrez; 18 Outros; 19 Não tem interesse
q34	MOTIVO PRINCIPAL QUE O LEVOU A OPTAR PELO VESTIBULAR DA UFSC	1) É a única no Estado que oferece o curso pretendido; 2) É a que oferece o melhor curso pretendido; 3) É a que oferece o curso pretendido em horário adequado; 4) O curso pretendido é pouco procurado, o que facilita a classificação; 5) É de fácil acesso (proximidade de casa, prática locomoção etc.); 6) Na realidade, gostaria de estudar em outra universidade; 7) Por ser pública e gratuita, satisfazendo as condições sócio-econom. da família

Fonte: o autor