UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO SOCIOECONÔMICO
PROGRAMA DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO UNIVERSITÁRIA

Aline Pacheco Primão

USO DE ALGORITMOS DE *MACHINE LEARNING* PARA PREVER A EVASÃO ESCOLAR NO ENSINO SUPERIOR: UM ESTUDO NO INSTITUTO FEDERAL DE SANTA CATARINA

Florianópolis – SC
2022

Aline Pacheco Primão

USO DE ALGORITMOS DE *MACHINE LEARNING* PARA PREVER A EVASÃO ESCOLAR NO ENSINO SUPERIOR: UM ESTUDO NO INSTITUTO FEDERAL DE SANTA CATARINA

Dissertação submetida ao Programa de Pós-Graduação em Administração Universitária da Universidade Federal de Santa Catarina para obtenção do título de Mestre em Administração Universitária.
Orientador: Professor Leonardo Flach, Dr.

Florianópolis – SC
2022

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

Primão, Aline
Uso de Algoritmos de Machine Learning para Prever a Evasão Escolar no Ensino Superior: Um Estudo no Instituto Federal de Santa Catarina / Aline Primão ; orientador, Leonardo Flach, 2022.
131 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Sócio-Econômico, Programa de Pós-Graduação em Administração Universitária, Florianópolis, 2022.

Inclui referências.

1. Evasão Escolar. 2. Machine Learning. 3. XGBoost. 4. MultiLayer Perceptron. I. Flach, Leonardo . II. Universidade Federal
de Santa Catarina. Programa de Pós-Graduação em Administração Universitária. III. Título.

Aline Pacheco Primão

Uso de Algoritmos de *Machine Learning* para Prever a Evasão Escolar no Ensino Superior: um Estudo no Instituto Federal de Santa Catarina

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

Professora Andressa Sasaki Vasques Pacheco, Dra.
Universidade Federal de Santa Catarina

Professor Mateus Grellert da Silva, Dr.
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 Administração Universitária.

Documento assinado digitalmente
Leonardo Flach
Data: 15/08/2022 09:37:33-0300
CPF: 007.462.289-70
Verifique as assinaturas em https://v.ufsc.br

Coordenação do Programa de Pós-Graduação

Documento assinado digitalmente
Leonardo Flach
Data: 15/08/2022 09:37:12-0300
CPF: 007.462.289-70
Verifique as assinaturas em https://v.ufsc.br

Prof. Leonardo Flach, Dr.
Orientador

Florianópolis, 27 de julho de 2022.

RESUMO

Conseguir prever a evasão escolar em instituições públicas de ensino superior é importante para formular ações que ajudem o estudante em seu desenvolvimento educacional. Para isso, as técnicas de *Machine Learning* (ML) podem colaborar com as instituições a prever a evasão. Este trabalho tem como objetivo propor um modelo usando algoritmos de *Machine Learning* para prever a evasão escolar no Instituto Federal de Santa Catarina (IFSC). Para isso, foram analisados os fatores que impactam na evasão nas instituições de ensino superior por meio da literatura existente, gerada uma planilha com as características importantes, após isso, foram utilizados os algoritmos *Decision Tree,* usados como *baseline*, *Artificial Neural Network* e *XGBoost* para desenvolver um modelo de previsão da evasão escolar do IFSC e conseguir analisar os anos de 2017, 2018 e 2019 que representam os dados antes da pandemia da Covid-19 e os anos 2020 e 2021 com os dados durante a pandemia. Os dois modelos (*XGBoost* e MLP) se mostraram melhores que o *baseline* das duas bases analisadas, porém o modelo *XGBoost* se mostrou superior. No *DataFrame* antes da pandemia, o algoritmo *XGBoost* obteve o F1-Score de 97,53%, já no algoritmo MLP o *F1-Score* foi de 93,83%. No df_durante_pandemia, o algoritmo *XGBoost* apresentou *F1_score* igual a 90,32%. Já o algoritmo MLP obteve 80% de *F1_Score*. Outra análise importante foi em relação à importância das variáveis, já que, para o *DataFrame* com os dados de antes da pandemia, a variável que apresentou maior relevância foi a de número de disciplinas concluídas, seguida de forma de ingresso, média geral do discente, renda familiar *per capita* e campus. Outrossim, para o *DataFrame* com os dados durante pandemia, a variável mais importante foi a idade do discente, seguida de forma de ingresso, curso do discente, naturalidade do discente e média geral do discente. Sendo assim, é possível verificar que a forma de ingresso é a segunda variável mais importante, tanto antes da pandemia como durante a pandemia, e a média geral do discente encontra-se entre as cinco principais variáveis nos dois *DataFrames*. A partir da avaliação do modelo criado, por ter trazido os melhores resultados, o algoritmo *XGBoost* foi selecionado para criar um modelo ajustado. Dessa forma, foram testados novos hiperparâmetros e retiradas três variáveis que não apresentaram significância estatística. Foi mostrado que o modelo ajustado não alterou o resultado do *Dataframe* antes da pandemia, porém para o *DataFrame* durante a pandemia obteve melhores resultados.

Palavras-chave: Evasão Escolar. *Machine Learning*. XGBoost. *MultiLayer Perceptron*.

ABSTRACT

Being able to predict school dropout in public higher education institutions is important to formulate actions that help students in their educational development. For this, Machine Learning (ML) techniques can collaborate with institutions to predict dropout. This work aims to propose a model using Machine Learning algorithms to predict school dropout at Instituto Federal de Santa Catarina (IFSC). For this, the factors that impact dropout in higher education institutions were analyzed through the existing literature, generated a spreadsheet with the important characteristics, after that, the Decision Tree algorithms were used, used as baseline, Artificial Neural Network and XGBoost to develop an IFSC dropout prediction model and be able to analyze the years 2017, 2018 and 2019 that represent data before the Covid-19 pandemic and the years 2020 and 2021 with data during the pandemic. The two models (XGBoost and MLP) were better than the baseline of the two analyzed bases, but the XGBoost model was superior. In the DataFrame before the pandemic, the XGBoost algorithm obtained an F1-Score of 97.53%, while in the MLP algorithm the F1-Score was 93.83%. In df_during_pandemia, the XGBoost algorithm presented F1_score equal to 90.32%. The MLP algorithm obtained 80% of F1_Score. Another important analysis was in relation to the importance of the variables, since, for the DataFrame with data from before the pandemic, the variable that showed the greatest relevance was the number of courses completed, followed by the form of admission, overall student average, per capita household income and campus. Furthermore, for the DataFrame with the data during the pandemic, the most important variable was the student's age, followed by the form of admission, the student's course, the student's place of birth and the student's general average. Therefore, it is possible to verify that the form of admission is the second most important variable, both before the pandemic and during the pandemic, and the general average of the student is among the five main variables in the two DataFrames. From the evaluation of the created model, for having brought the best results, the XGBoost algorithm was selected to create an adjusted model. Thus, new hyperparameters were tested and three variables that did not show statistical significance were removed. It was shown that the adjusted model did not change the result of the Dataframe before the pandemic, but for the DataFrame during the pandemic it obtained better results.

Keywords: School Dropout. Machine Learning. XGBoost. MultiLayer Perceptron.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANN	Artificial Neural Network
BD	Banco de Dados
CEFET-SC	Centro Federal de Educação Tecnológica de Santa Catarina
CERFEAD	Centro de Referência de Educação à Distância
COLAB	Colaboratory Google
CRISP-DM	Cross-Industry Standard Process for Data Mining
DF	DataFrame
DT	Decision Tree
EAD	Educação à Distância
FIC	Formação Inicial e Continuada
Float	Número de Ponto Flutuante
FN	False Negative
FP	False Positive
HCA	Hierarchical Clustering
IDHM	Índice de Desenvolvimento Humano por Município
IEDMS	International Educational Data Mining Society
IFES	Instituições Federais de Ensino Superior
IFSC	Instituto Federal de Santa Catarina
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
INSS	Instituto Nacional do Seguro Social
Int	Integer
IoT	Internet of Things
KNN	K-Nearest Neighbor
KNNImputer	K-Nearest Neighbor Imputer
LGPD	Lei Geral de Proteção de Dados Pessoais
LDB	Lei de Diretrizes e Bases da Educação Nacional
LLE	Locally-Linear Embedding
MDE	Mineração de Dados Educacionais
ML	Machine Learning
MLP	MultiLayer Perceptron
MLPClassifier	MultiLayer Perceptron Classifier
NaN	Not a Number
PCA	Principal Component Analysis
PNP	Plataforma Nilo Peçanha
PPE	Plano Estratégico de Permanência e Êxito
PPGAU	Programa de Pós-Graduação em Administração Universitária
QDA	Quadratic Discriminant Analysis
RNA	Redes Neurais Artificiais
RFP	Renda Familiar Per Capita
SESu/MEC	Secretaria de Educação Superior do Ministério da Educação e Desporto
SIGAA	Sistema Integrado de Gestão Acadêmica
SiSU	Sistema de Seleção Unificada

Str	String
SVM	Support Vector Machine
TADA	Taxa de Desistência Anual
TAP	Taxa de Permanência
TCA	Taxa de Conclusão Acumulada
TCAN	Taxa de Conclusão Anual
TODA	Taxa de Desistência Acumulada
TN	True Negative
TP	True Positive
XGB	eXtreme Gradient Boosting
XGBoost	eXtreme Gradient Boosting
XGBClassifier	eXtreme Gradient Boosting Classifier

SUMÁRIO

1 INTRODUÇÃO

A evasão escolar tem sido alvo de estudo de pesquisadores de todo o mundo (PÉREZ *et al.*, 2018; MARQUES, 2020; CASANOVA *et al.*, 2021) por representar um grande obstáculo nas instituições de ensino, afetando sua reputação e sua classificação entre as demais. Sendo assim, quando se trata de instituições públicas, isso pode gerar prejuízo para toda a sociedade, pois a evasão corrobora com a redução do retorno dos investimentos em educação.

Conforme apontam os dados do Censo da Educação Superior fornecido pelo Instituto Nacional de Estudos e Pesquisas Anísio Teixeira (INEP, 2019), o Brasil possui 302 instituições públicas, sendo 108 universidades, 11 centros universitários, 143 faculdades e 40 institutos federais e Centros Federais de Educação Tecnológica (CEFETS). Nas IFES, as matrículas de 2005 foram 595 mil e em 2019 foram 1,3 milhão, um aumento de mais de 124%; já os ingressantes de 2005 foram 148 mil e em 2019, 362 mil, um acréscimo de 144%, e, por fim, os concluintes em 2005 foram 92 mil e em 2019 foram 149 mil, um aumento de 61% (INEP, 2019). Ao longo do tempo, é possível perceber que os concluintes não aumentaram na mesma proporção dos ingressantes, permanecendo baixo. Para Marques (2020), o número de alunos concluintes permanece baixo comparado com o número de matriculados, mas, segundo Lima Júnior *et al.* (2019), a evasão não aumentou em todos os cursos da mesma maneira, atingido apenas alguns setores da educação superior.

Silva, Cabral e Pacheco (2020) afirmam que a evasão acarreta uma perda ao estudante, que, além de não obter uma formação, perde tempo e recursos financeiros, porém demonstra uma decisão relacionada a outras formações e propósito individual, já a instituição expressa perda de eficiência (SILVA; CABRAL; PACHECO, 2020), e, para a universidade pública, isso representa um desperdício à sociedade, que perde um futuro profissional, além dos custos atribuídos àquele indivíduo. Sendo assim, é imprescindível para a gestão universitária a compreensão do fenômeno da evasão para, então, realizar o tratamento adequado.

Apesar de ser um dos principais problemas nas instituições de ensino, Sultana, Khan e Abbas (2017) ressaltam que a evasão escolar pode ser controlada e desacelerada, para isso, é necessário que as instituições consigam desenvolver estratégias eficientes para a diminuição de alunos evadidos e, consequentemente, tenha mais estudantes se formando todos os anos, entregando, assim, valor à sociedade.

A Mineração de Dados Educacionais (MDE), apesar de ser um campo emergente, vem ganhando atenção nos últimos anos, já que pode ser utilizada para gerar informações que ajudem na tomada de decisão do processo educacional (SULTANA; KHAN; ABBAS, 2017;

EZZ; ELSHENAWY, 2019; WOTAIFI; AL-SHAMERY, 2019). Para tanto, os dados acadêmicos das instituições de ensino possuem um volume muito grande, e, para se produzir as buscas e as análises desses dados, é necessário que haja uma investigação baseada na extração do conhecimento. Para esse fim, os algoritmos de *Machine Learning* (ML), Aprendizado de Máquina no português, são opções exequíveis (SILVA; ALMEIDA; RAMALHO, 2020).

Os estudos com *Machine Learning* se beneficiam de modelos estatísticos para previsão de risco de algum evento ocorrer (SILVA; ALMEIDA; RAMALHO, 2020). Os algoritmos de ML não conseguem obter uma boa *performance* em todos os tipos de aplicações, e, muitas vezes, exige-se mais de um algoritmo para o ganho de desempenho.

O Instituto Federal de Santa Catarina (IFSC), em 2019, apresentou 18.3% de alunos evadidos nos cursos de graduação, o que significa estar está acima da média nacional de 15.5% (BRASIL, 2021). Com isso, surge uma lacuna de pesquisa, emergindo o seguinte problema de pesquisa: **Qual modelo, utilizando algoritmos de *Machine Learning*, explica a evasão escolar no Instituto Federal de Santa Catarina (IFSC)?**

A pesquisa pretende fazer uma análise dos fatores que mais impactam na evasão escolar dentro das instituições de ensino superior, e, a partir disso, utilizar os algoritmos *Decision Tree, Artificial Neural Network* e *XGBoost* para o treinamento e os testes da base de dados do Instituto Federal de Santa Catarina (IFSC) e, assim, propor um modelo de previsão de evasão escolar para o ensino superior no IFSC.

1.1 OBJETIVOS

Nesta seção, serão apresentados os objetivos, geral específicos, utilizados nesta pesquisa.

1.1.1 Objetivo Geral

Este estudo tem como objetivo propor um modelo usando algoritmos de *Machine Learning* para prever a evasão escolar no Instituto Federal de Santa Catarina antes e durante a pandemia da Covid-19.

1.1.2 Objetivos Específicos

Para atingir o objetivo geral da pesquisa, foram estabelecidos os seguintes objetivos específicos:

a) Analisar fatores que impactam na evasão escolar por meio da literatura existente e selecionar as variáveis a serem utilizadas na pesquisa;
b) Analisar os algoritmos de *Machine Learning: Decision Tree (DecisionTreeClassifier)*, *Artificial Neural Network – MultiLayer Percetron (MLPClassifier)* e *XGBoost (XGBClassifier)*;
c) Realizar pré-processamento dos dados do IFSC a partir das variáveis analisadas e adquiridas;
d) Desenvolver um modelo para previsão de evasão escolar no ensino superior do IFSC usando os algoritmos de ML analisados;
e) Aplicar o modelo de previsão de evasão escolar desenvolvido para analisar dados antes da pandemia da Covid-19 (2017, 2018 e 2019) e durante a pandemia da Covid-19 (2020 e 2021).

1.2 JUSTIFICATIVA E CONTRIBUIÇÕES

Com o intuito de analisar pesquisas anteriores sobre o tema proposto ou temas relacionados, fazer uma discussão da literatura e justificar o trabalho, foi realizada uma pesquisa bibliométrica, em julho de 2021, com as seguintes palavras-chave: *Machine Learning* AND *University* AND *Higher Education* nas bases de dados Scopus e Web of Science a partir do ano de 2010. Não foi utilizada a palavra-chave *University Dropout*, pois a intenção é encontrar todas as pesquisas relacionadas à gestão universitária e, assim, analisar qual a maior preocupação das instituições de ensino superior com base nos seus objetivos.

Foram encontrados 244 artigos, excluídos os repetidos e os que não se enquadraram na pesquisa, após a leitura dos títulos, restaram 63 artigos. Logo após, foram retirados os que não estavam disponíveis na íntegra, então permaneceram 58 artigos. Na sequência com a leitura dos resumos, foram excluídos os que não eram apresentados como foco na gestão universitária, ficando 37 artigos. E, por fim, após a leitura na íntegra, cinco foram excluídos por não apresentarem os algoritmos de *Machine Learning*, restando, assim, um total de 32 artigos.

Esta pesquisa utilizou a recomendação Prisma de Moher *et al.* (2009) apresentada no fluxograma da Figura 1.

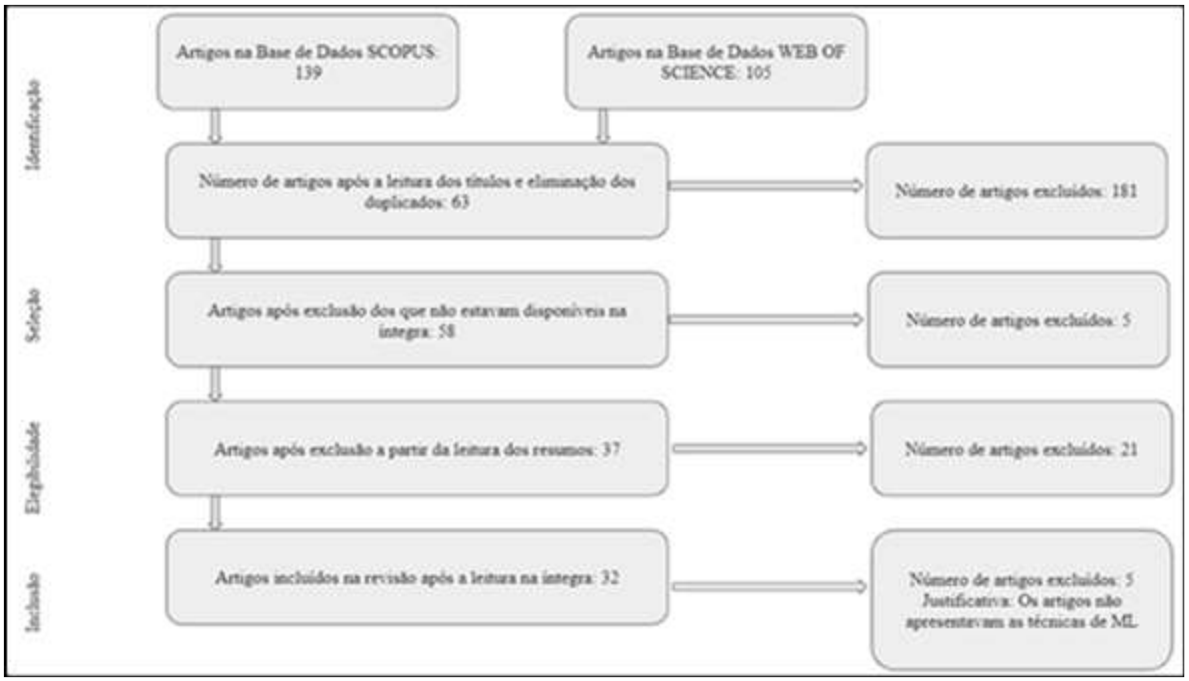


Figura 1 – Fluxograma Pesquisa Bibliométrica Prisma
Fonte: Elaborada pela autora desta dissertação (2022)

A partir da pesquisa, observou-se que os anos de 2019 e 2020 foram os mais evidentes, fornecendo 26 dos 32 artigos. Também se verificou que 29 destes eram na língua inglesa, dois em espanhol e um em português.

O Continente asiático é o mais significativo nas publicações, com 21 representações, sendo que a China é retratada em cinco artigos e Taiwan em quatro, seguidos da Europa com 12 e da América com nove representações. A Figura 2 representa a disposição dos artigos por continente.

Continente	Quantidade
Ásia	21
Europa	12
América	9
África	4
Oceania	1

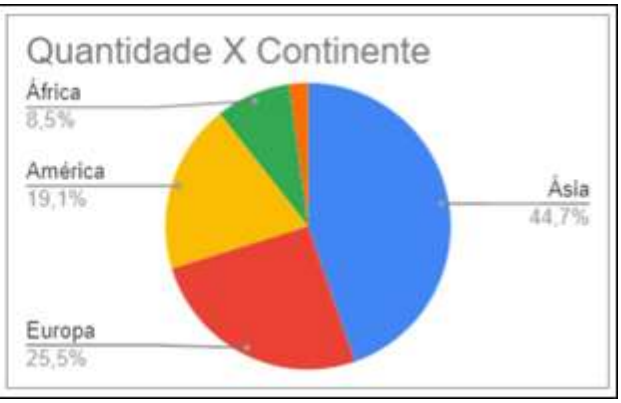


Figura 2 – Divisão dos artigos por continente
Fonte: Elaborada pela autora desta dissertação (2022)

No que se refere aos objetivos dos artigos, a Tabela 1 mostra a quantidade de artigos por objetivos similares.

Tabela 1 – Artigos por objetivos similares

Objetivos	Quantidade
Análise de Procrastinação de alunos	2
Prever notas dos alunos	2
Prever alunos em risco de reprovação	5
Prever resultados acadêmicos dos alunos	6
Revisão de textos em pesquisas de opiniões de alunos	1
Avaliar dados técnicos de alunos	1
Conhecer número efetivo de alunos numa plataforma	1
Análise de evasão para evitar desistência, abandono ou atrito	6
Prever número de alunos com baixo engajamento nos cursos	1
Propor sistema para prever caminho para alunos no ano preparatório da faculdade	1
Explorar os comportamentos de aprendizado gerados por alunos	1
Propor um modelo preditivo com taxas de precisão aprimoradas a partir da análise de dados de alunos	1
Previsão de desempenho acadêmico	2
Estudar que tipos de modelo de previsão tem melhor desempenho	1
Promover o uso intuitivo de técnicas de análise de evasão	1

Fonte: Elaborada pela autora desta dissertação (2022)

As palavras-chave mais utilizadas foram *Machine Learning* com 13 citações, *Higher Education*/Educação Superior/*Educación Superior* com nove citações, *Educational Data Mining* com oito citações e *Learning Analytics* com seis citações. A Figura 3 apresenta as vezes que a palavra-chave foi citada em um artigo.

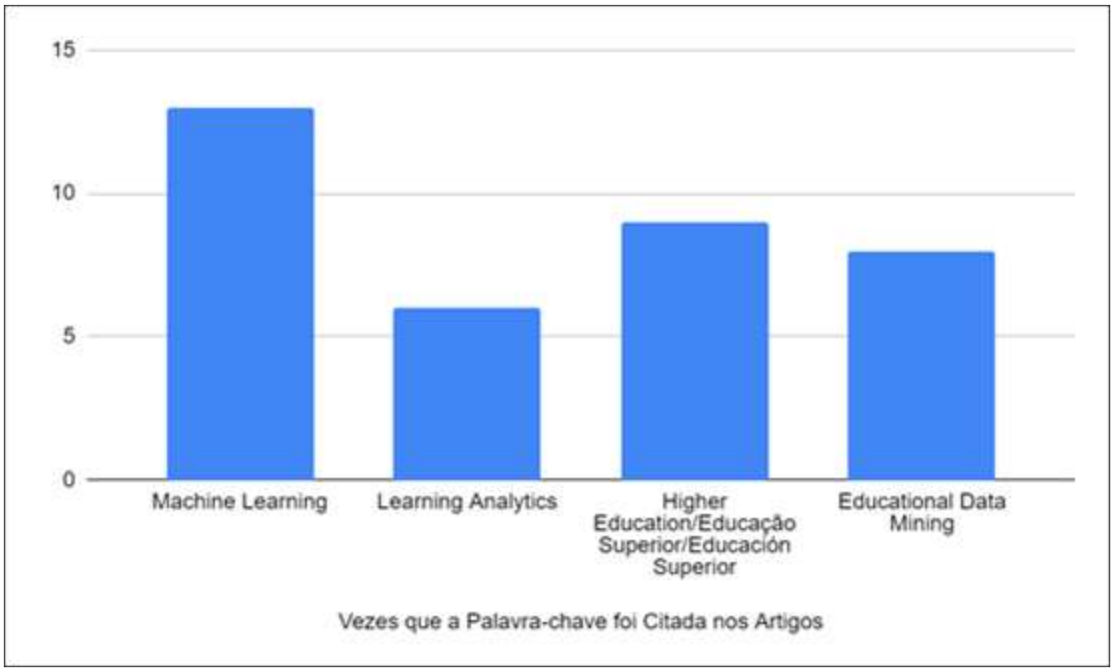


Figura 3 – Palavras-chave apresentadas nos artigos
Fonte: Elaborada pela autora desta dissertação (2022)

Já na questão referente à coleta de dados, a maioria das pesquisas utilizou somente a coleta nos bancos de dados dos sistemas acadêmicos com representação de 26 artigos, as pesquisas que utilizaram coleta de dados dos sistemas acadêmicos, juntamente com um ou mais questionários, somam quatro artigos, e as pesquisas que levaram em conta apenas questionários somam dois artigos.

Na literatura, existem vários estudos sobre evasão escolar que buscam identificar os motivos que levam o aluno evadir, por exemplo, Prim e Fávero (2013), Fortunato e Gontijo (2020) ou Hoffman, Nunes e Muller (2016), que procuram construir um modelo de conhecimento organizacional sobre evasão.

Diante da revisão bibliométrica, quando se fala de uso de métodos de *Machine Learning* para previsão de evasão escolar no contexto brasileiro, as pesquisas encontraram Silva, Almeida e Ramalho (2020), esses autores utilizaram os algoritmos de ML para prever o risco de reprovação em uma disciplina específica, não se aprofundando no contexto geral da graduação de uma instituição; Costa *et al.* (2017), que compararam as eficácias de técnicas de Mineração de Dados Educacionais para identificar alunos em risco; e Freitas *et al.* (2020), que desenvolveram um sistema IoT para usabilidade da gestão universitária. O presente estudo é focado em cursos de graduação, sejam eles na modalidade presencial e a distância, com o intuito de prever a evasão.

Conforme apontam os dados da Plataforma Nilo Peçanha (PNP), no ano de 2019, nos Institutos Federais, a média de evasão escolar nos cursos de graduação no país foi de 15,53%,

sendo que a Região Norte foi a que obteve a menor taxa com 13,23%, e a Região Sul foi a que obteve a maior taxa de evasão escolar com 19,27% (BRASIL, 2021). Desse modo, usar os dados de um instituto federal da Região Sul será de extrema importância, já que essa região apresenta os piores cenários de evasão escolar do país.

Outra especificidade é que nenhum estudo ainda foi realizado no Programa de Pós-Graduação em Administração Universitária (PPGAU) aplicando as técnicas de *Machine Learning*, por isso, este estudo é inovador e poderá trazer grandes benefícios ao programa em pesquisas futuras, visto que é assunto emergente e que vem ganhando muita atenção em diversas áreas e, em especial, na gestão universitária.

Este trabalho deve contribuir, primeiramente, para que as instituições consigam prever os alunos com maior probabilidade de evasão escolar e, com isso, desenvolver ações, a fim de ajudar os alunos com dificuldades a concluírem seus estudos. Segundo, com a diminuição da taxa da evasão escolar, é possível melhorar o retorno dos investimentos em educação e prover valor para a sociedade. E, por fim, com maior número de concluintes nas instituições de ensino superior, melhora a possibilidade de sucesso futuro desses estudantes.

1.3 ESTRUTURA DO TRABALHO

Após a contextualização e a justificativa do tema elaboradas na introdução e a apresentação das hipóteses e dos objetivos da pesquisa, apresenta-se no segundo tópico a revisão bibliográfica sobre a evasão escolar no ensino superior, apontando indicadores do Brasil, indicadores do Instituto Federal de Santa Catarina, *Machine Learning* e algoritmos de *Machine Learning*. Para o desenvolvimento dos objetivos, foram analisados os fatores que impactam na evasão escolar no ensino superior por meio da literatura existente, e, a partir dos fatores analisados, gerar a lista de variáveis a serem investigadas; Analisar algoritmos de *Machine Learning Decision Tree (DecisionTreeClassifier)*, *Artificial Neural Network (MLPClassifier)* e *XGBoost (XGBClassifier)*; Realizar o pré-processamento a partir dos dados do IFSC, e, a partir disso, o treinamento e o teste, usando a linguagem de programação *Python,* utilizando a ferramenta *Colab* do *Google*; Propor um modelo para previsão de evasão escolar no ensino superior federal brasileiro. Após, comparar os dados do modelo criado antes da pandemia (2017, 2018 e 2019) com os dados durante a pandemia da Covid-19 (2020 e 2021). A Figura 4 apresenta o Processo da Dissertação que fornece quatro etapas: Introdução, Revisão Bibliográfica, Desenvolvimento dos Objetivos e Resultados e Trabalhos Futuros.

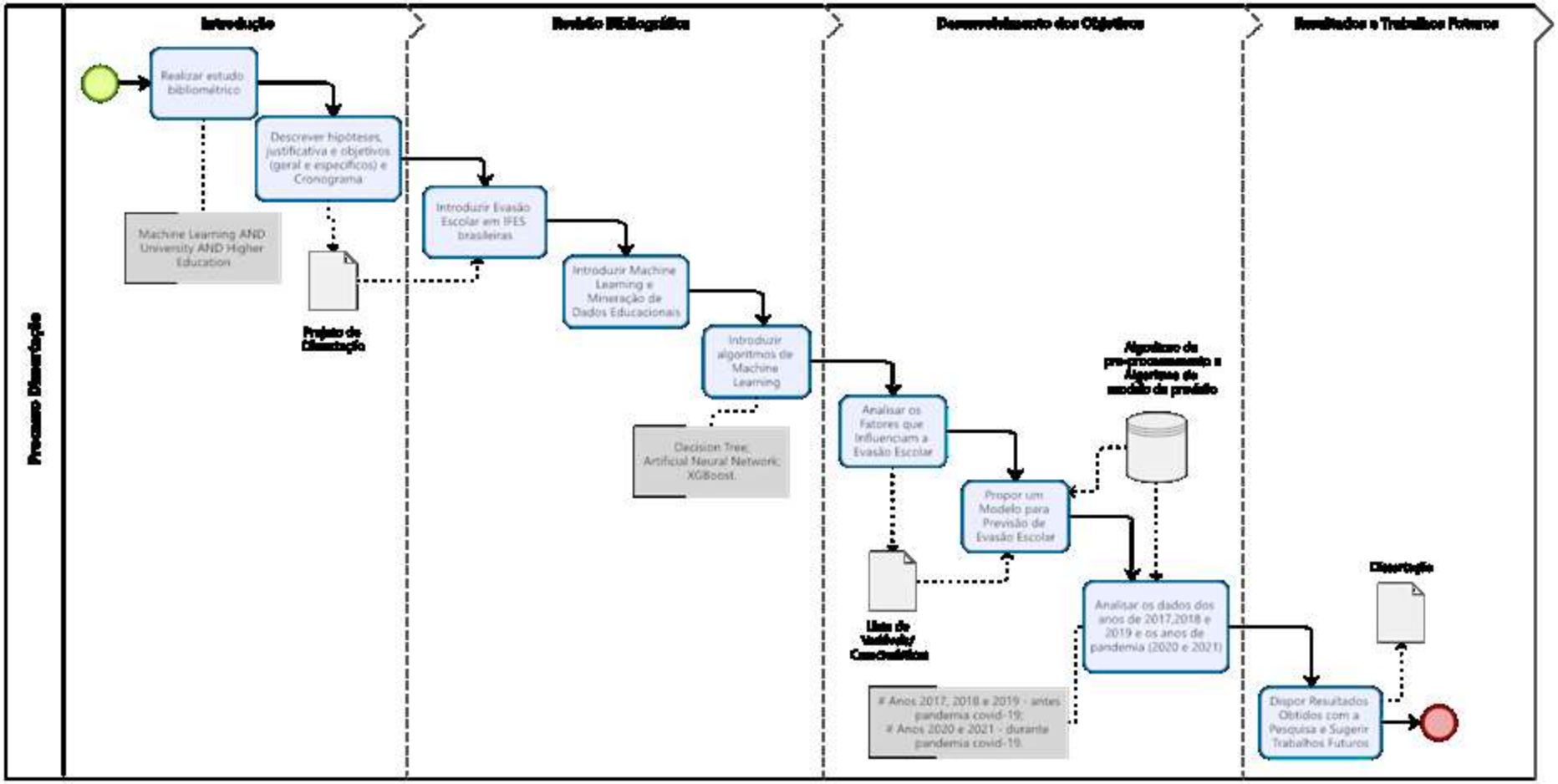


Figura 4 – Processo da dissertação
Fonte: Elaborada pela autora desta dissertação (2022)

2 REFERENCIAL TEÓRICO

Esta seção fornece a fundamentação teórica que levará em conta os objetivos da pesquisa e deve abordar os seguintes temas: Evasão no Ensino Superior; Indicadores da educação superior no Brasil e Indicadores no IFSC; *Machine Learning*; e Algoritmos de *Machine Learning*, conforme processo de dissertação definido na Seção 1.3 Estrutura do Trabalho, no intuito de sustentar a elaboração e a execução da pesquisa.

2.1 EVASÃO NO ENSINO SUPERIOR

A evasão é considerada uma das maiores preocupações do ensino superior, principalmente quando se trata de instituições públicas, pois estas precisam garantir resultados relevantes e afirmar a diplomação de estudantes para o mercado de trabalho (ANDIFES, 1996).

Flores (2017) esclarece que a palavra *evasão* vem do latim *evasio* e significa fuga, saída, abandono, fracasso, insucesso, mas a terminologia adotada varia conforme o autor, porém, todos que inciam um curso têm o mesmo propósito: o de finalizar um curso.

De acordo com Silva, Cabral e Pacheco (2020), Costa e Gouveia (2018) e Prestes e Fialho (2018), não existe concordância da literatura a respeito do conceito de evasão. A partir dessa avaliação, o estudo de Costa e Gouveia (2018) forneceu um quadro com definições para retenção e persistência, pois eles mostraram que, para autores que diferenciam esses conceitos, há diversas definições, já para os autores que não diferenciam ou os autores que apenas definem retenção não há consenso sobre os termos. Outra questão que os autores mostram é que a maior parte das definições é proveniente de pesquisas internacionais, sendo principalmente norte-americanas (COSTA; GOUVEIA, 2018).

A Andifes (1996) definiu a evasão no ensino superior como sendo a saída definitiva do aluno do curso de origem sem concluí-lo. Essa evasão pode ser dividida em três tipos, conforme destacam Andifes (1996) e Prestes e Fialho (2018):

a) Evasão do curso: o estudante desliga-se do curso por abandono, deixando de matricular-se, desistência oficial, transferência ou mudança de curso, ou ainda exclusão por norma institucional.
b) Evasão da instituição: o estudante desliga-se da instituição em que está matriculado.
c) Evasão do sistema: o estudante desliga-se de forma definitiva ou temporária do ensino superior.

De acordo com Brasil (2015), a evasão é definida quando decorre do desligamento do aluno de um curso, sendo que pode ser em diferentes situações: abandono, pedido de cancelamento de matrícula, transferência interna ou externa, entre outros. Para fins de pesquisa, esse será o conceito utilizado para definir a evasão e selecionar as variáveis utilizadas para a criação do modelo, pois esta é a definição utilizada pela Plataforma Nilo Peçanha (PNP) que dispõe dos dados dos Institutos Federais.

Segundo Lima Júnior *et al.* (2019), a evasão não é um fenômeno recente, porém tem ganhado maior importância nas últimas décadas devido à expansão na educação superior. Em sua pesquisa, Pérez *et al.* (2018) descrevem que mais da metade dos alunos que ingressam no ensino superior no Chile acaba abandonando o curso. Ainda, segundo os autores, três a cada dez estudantes que abandonam seus cursos fazem isso no primeiro período da faculdade (PÉREZ *et al.*, 2018).

Para Delen (2010), as altas taxas de evasão escolar afetam o planejamento de matrículas e trazem uma sobrecarga de trabalho para recrutar novas alunos, já para os alunos, desistir antes de obter um diploma representa que o potencial humano não foi explorado e gera um baixo retorno dos investimentos da instituição.

Conforme descreve a Andifes (1996), conhecer mais a fundo o fenômeno da evasão, somente é possível a partir de um programa integrado de pesquisas que estabeleça elo entre os níveis, identifique as causas internas e externas e, assim, consiga dimensionar a totalidade característica de uma avaliação do ensino superior público brasileiro.

Na seção 2.1.1 são apresentados indicadores importantes ao longo dos anos no Brasil, esses indicadores são relevantes para se entender a evolução da educação superior federal e a evasão escolar nas IFES brasileiras.

2.1.1 Indicadores na Educação Superior no Brasil

Segundo Costa e Gouveia (2018), os estudos sobre evasão no Brasil começaram nos anos de 1995 e 1996 por meio de uma comissão especial pela então Secretaria de Educação Superior do Ministério da Educação e Desporto (SESu/MEC). Essa comissão avaliou boa parte das instituições federais de ensino superior quanto aos índices de diplomação, de retenção e de evasão dos cursos de graduação (COSTA; GOUVEIA, 2018).

Desde então, o Brasil tem sofrido intensas transformações na educação superior, com a publicação da Lei Geral de Diretrizes e Bases da Educação Nacional (LDB), que incentivou a expansão de matrículas por meio do crescimento das instituições, de cursos e de vagas, dando

autonomia para as instituições públicas de ensino elaborarem, aprovarem e executarem planos de investimentos, de ações e orçamentários, criando, organizando e extinguindo cursos e programas de educação (BRASIL, 1996).

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) atua em três frentes no país: Avaliação e Exames Educacionais, Gestão do Conhecimento e Estudos Educacionais e Pesquisas Estatísticas e Indicadores Educacionais. Nesta última frente, o INEP realiza o Censo da Educação Superior e dispõe de Indicadores de Fluxo da Educação Superior, os quais foram utilizados nesta pesquisa.

O Censo da Educação Superior é um instrumento de pesquisa realizado anualmente sobre as instituições de ensino superior, de cursos de graduação e sequenciais, alunos e docentes. Os dados utilizados no censo são sobre infraestrutura das instituições, vagas oferecidas, candidatos, matrículas, ingressantes, concluintes e docentes e estão dispostos no Sistema e-MEC, o qual é mantido pelas instituições (INEP, 2019).

De acordo com os dados do INEP, em 1996, o Brasil tinha 57 Instituições Federais de Ensino Superior (IFES) que totalizavam 1.581 cursos. Nos últimos dados disponibilizados de 2019, o país contava com 110 IFES e 6.669 cursos (INEP, 2019). A Tabela 2 apresenta a evolução de IFES e os cursos ao longo dos anos, nela é possível observar que, de 2004 a 2006, obteve-se uma elevação nas instituições e uma baixa em 2008, isso se deve à reforma da educação superior promovida no governo de Luiz Inácio Lula da Silva, que começou a ser desenvolvida em 2003, embora, apesar da diminuição das IFES no período, os cursos continuaram em ascensão (BRASIL, 2005).

Das instituições federais, 40 são Institutos Federais (IF) e Centros Federais de Educação Tecnológica (CEFET), 63 Universidades, seis Faculdades e um Centro Universitário. Os cursos de graduação, 1.718, estão em IFs e Cefets, as universidades representam 4.928 cursos, já as faculdades possuem 22 cursos, e o Centro universitário um curso (INEP, 2019).

Tabela 2 – Evolução das IFES

Ano	Número de IFES	Cursos
1996	57	1.581
1998	57	1.338
2000	61	1.996
2002	73	2.316
2004	87	2.450
2006	105	2.785
2008	93	3.460
2010	99	5.326

Ano	Número de IFES	Cursos
2012	103	6.303
2014	107	6.177
2016	107	6.234
2019	110	6.669

Fonte: Elaborada com base nos dados do Censo da Educação Superior (INEP, 2019)

Na compilação dos dados dos censos desde 1996 a 2019, identifica-se a evolução da educação superior federal em que o número de ingresso de alunos foi de 78.077 em 1996 para 362.558 em 2019, um aumento de mais de 364%, ao mesmo tempo o total de concluintes em 1996 foi de 46.187, e, em 2019, 146.367, uma elevação de 224%.

A Figura 5 mostra a evolução desde 1996 da educação superior de ingressantes e concluintes, a cor azul representa os ingressantes e em vermelho os concluintes, com isso, é possível analisar que elas estão cada vez mais distantes uma da outra.

Os concluintes de 2019 em instituições públicas federais foram 149.673, sendo que 130.185 foram em Universidades, 460 em centros universitários, 365 em faculdades e 18.663 em IFs e Cefets (INEP, 2019).

Ano	Ingresso	Concluintes
1996	78.077	46.187
1998	89.160	51.419
2000	117.507	59.098
2002	122.491	72.054
2004	122.899	90.269
2006	141.989	84.813
2008	186.280	85.634
2010	279.811	99.945
2012	304.885	111.165
2014	346.991	128.084
2016	342.986	146.367
2019	362.558	149.673

Figura 5 – Ingressantes e Concluintes de 1996 a 2019
Fonte: Elaborada com base nos dados do Censo da Educação Superior (INEP, 2019)

Quando analisadas as matrículas dos últimos 10 anos de censo, nota-se que, em 2009, as instituições públicas federais obtiveram 839.397 matrículas em cursos de graduação, e, em 2019, o número foi de 1.335.256, o que representa um aumento de 59,1%. As Universidades representam 83,5% (1.114.468) das matrículas da rede federal, e os Institutos apontam 16,2% (215.843), perfazendo 99,7% do número total das matrículas em cursos de graduação da rede federal. As matrículas presenciais obtiveram aumento de um ano para o outro, já as matrículas

em cursos de Educação a Distância apresentaram altos e baixos números ao longo dos anos (INEP, 2019). A Tabela 3 representa os dados ao longo de 10 anos em percentual do valor de matrículas totais no ano.

Tabela 3 – Matrículas ao longo dos últimos 10 anos do Censo da Educação Superior

Ano	Matrículas	Matrículas no Presencial	Matrículas no EAD
2010	938.656	88,84%	11,16%
2011	1.032.936	89,75%	10,25%
2012	1.087.413	90,60%	9,40%
2013	1.137.851	91,88%	8,12%
2014	1.180.068	91,82%	8,18%
2015	1.214.635	93,29%	6,71%
2016	1.249.324	94,10%	5,90%
2017	1.306.351	92,24%	7,76%
2018	1.324.984	92,98%	7,02%
2019	1.335.254	93,92%	6,08%

Fonte: Elaborada com base nos dados do Censo da Educação Superior (INEP, 2019)

O Censo de 2019 também exibiu os percentuais de participação das áreas gerais de conhecimento dos cursos de graduação (INEP, 2019). Nesse caso, é fornecido apenas a categoria administrativa (privada/pública), não separando as instituições públicas federais de estaduais e municipais. Os cursos na área de "Educação" são os maiores representantes nas instituições públicas, perfazendo 35,8%, seguidos por "Engenharia de Produção e Construção" com 15,7%, após "Negócios, Administração e Direito" com 10,5%, "Saúde e Bem-estar" com 8,1%, com 5,9% estão "Agricultura, Silvicultura, Pesca e Veterinária" e "Ciências Naturais, Matemática e Estatística", "Computação e Tecnologia da Informação e Comunicação" com 5,7%, "Artes e Humanidades" com 5,5%, "Ciências Sociais, Comunicação e Informação" com 5,4% e, por fim, "Serviços" com 1,6%. A Tabela 4 fornece os dados aqui apresentados.

Tabela 4 – Participação das áreas gerais de conhecimentos em cursos

Área Geral do Conhecimento	% Pública
Educação	35,8
Engenharia, produção e construção	15,7
Negócios, administração e direito	10,5
Saúde e bem-estar	8,1
Agricultura, silvicultura, pesca e veterinária	5,9
Ciências naturais, matemática e estatística	5,9
Computação e tecnologias da informação e comunicação (TIC)	5,7
Artes e humanidade	5,5

Área Geral do Conhecimento	% Pública
Ciências sociais, comunicação e informação	5,4
Serviços	1,6

Fonte: Elaborada com base nos dados do Censo da Educação Superior (INEP, 2019)

Para a compreensão dos indicadores fornecidos pelo Censo da Educação Superior do INEP (2019), faz-se necessário saber a definição de três dimensões de situações de vínculos do aluno:

a) Permanência: aluno que possui um vínculo ativo com o curso com situação "cursando" ou matrícula trancada".
b) Desistência: aluno encerra seu vínculo com o curso, com situação "desvinculado do curso" ou "transferido para outro curso da mesma IES".
c) Conclusão: o aluno também encerra seu vínculo com o curso, porém com situação de "formado".

A partir das três dimensões definidas, é possível analisar o cálculo dos indicadores de fluxo dos estudantes e indicadores anuais:

> Taxa de Permanência (TAP): percentual do número de estudantes com vínculos ativos (cursando ou trancado) ao curso j no ano tem relação ao número de estudantes ingressantes do curso j no ano T, subtraindo-se o número de estudantes falecidos do curso j do ano T até o ano t.
> Taxa de Desistência Acumulada (TDA): 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.
> Taxa de Conclusão Acumulada (TCA): percentual do número de estudantes que se formaram no curso j até o ano t do curso j 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.
> Taxa de Conclusão Anual (TCAN): percentual do número de estudantes que se formaram no curso j no ano tem relação ao número de ingressantes do curso j no ano T, subtraindo-se o número de estudantes falecidos do curso j até o ano t.
> Taxa de Desistência Anual (Tada): percentual do número de estudantes que saíram (desvinculado ou transferido) do curso j no ano tem relação ao número de estudantes ingressantes no curso j do ano T, subtraindo-se o número de estudantes falecidos do curso j até o ano t. (INEP, 2019)

O Censo 2019 utilizou seis coortes (ano de ingresso entre 2010 e 2015), até o ano de 2019 para gerar seus indicadores, assim as coortes ficaram: 2010-2019, 2011-2019, 2012-2019, 2013-2019, 2014-2019 e 2015-2019. O gráfico gerado (Figura 6) demonstrou que, no final do acompanhamento (2019), 52% dos ingressantes na rede federal desistiram de seus cursos, sendo

que 36% desistiram até o quarto ano. Outro fato levantado foi que em 10 anos 46% haviam concluído e 2% ainda permaneciam em seus cursos (INEP, 2019).

Referente à Taxa de Conclusão Acumulada (verde), ela se torna ascendente entre o quarto e quinto ano de acompanhamento, estabilizando a Taxa de Desistência Acumulada (vermelho) que diminui seu crescimento nos últimos anos de acompanhamento (INEP, 2019).

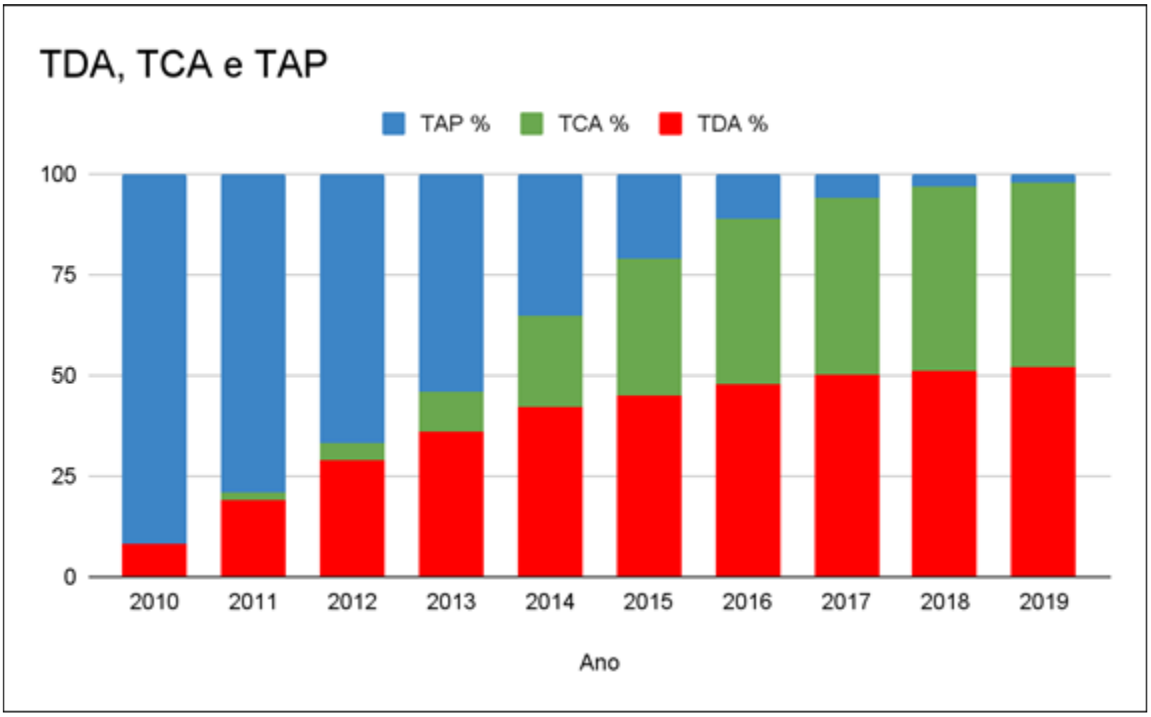


Figura 6 – Acompanhamento de coortes na trajetória de ingressantes em cursos de graduação em IFES
Fonte: Elaborada a partir do Resumo Técnico Censo da Educação Superior 2010 (INEP, 2019)

Outra avaliação realizada com todas as Instituições de Ensino Superior (IES) foi de que cursos presenciais obtiveram TCA de 40% e TDA de 59% ao longo dos 10 anos contra os 36% de TCA e 63% de TDA dos cursos de educação a distância. As duas modalidades em 2019 apresentavam 1% de TAP (INEP, 2019). Diante das avaliações apresentadas, é possível observar que os cursos da modalidade presencial atingiram os melhores resultados com maior conclusão e menor desistência.

Quando avaliado o grau acadêmico dos ingressantes de 2010 das IES, percebe-se grande distinção entre eles no que se refere ao indicador de conclusão anual (TCAN). Os cursos de Bacharelado do quarto ao sexto ano são os mais expressivos, já no grau de Licenciatura, o quarto ano é mais representativo, por fim, nos cursos Tecnológicos, o terceiro ano é o maior em TCAN (INEP, 2019). A taxa de conclusão anual de todos os graus é condizente com o tempo médio de integralização dos cursos.

Na taxa de desistência anual (Tada), o segundo ano é o que possui a maior desistência para todas as coortes, seguido do terceiro ano para a maior parte das coortes (apenas na coorte de 2010-2019, o terceiro ano empata com o primeiro) (INEP, 2019).

Quando analisados apenas os Institutos Federais (IFs) em 2019, estes somam 40 instituições, e, nos cursos de graduação (licenciatura, bacharelado e tecnologia), 483 unidades, 2.153 cursos, 271.071 matrículas, 75.768 ingressantes e 16.279 concluintes (BRASIL, 2021).

No final de 2019, das 271.071 matrículas em graduações nos IFs, 219.618 continuavam em curso, 16.279 tinham sido concluídas e 35.144 tinham evadido (BRASIL, 2021).

A seguir, será fornecida uma avaliação dos fatores que influenciam na evasão no ensino superior e, a partir daí, serão analisadas e selecionadas as características/variáveis importantes para a evasão.

2.2 FATORES QUE INFLUENCIAM NA EVASÃO NO ENSINO SUPERIOR

Para satisfazer o primeiro objetivo da pesquisa, foram analisados os fatores que influenciam na evasão escolar presentes na literatura, e, a partir disso, foi desenvolvida uma planilha com a lista de variáveis para a pesquisa.

Desde os anos de 1970, muitos autores tentam avaliar fatores que possam influenciar na evasão escolar por meio de seus modelos e, assim, seja possível ajudar as instituições a melhorarem o seu desempenho institucional, porém, esse é um assunto muito difícil de analisar, sendo deveras particular. A Andifes (1996) classificou os fatores que influenciam na evasão escolar em três ordens: relacionados ao próprio estudante; relacionados ao curso ou a instituição; fatores socioculturais e econômicos externos.

Diante disso, alguns autores acreditam que as maiores taxas de evasão escolar se dão no primeiro ano de faculdade (SCHMITT *et al.,* 2021; HERBAUT, 2020; BARGMANN; THIELE; KAUFFELD, 2021). Segundo Bargmann, Thiele e Kauffeld (2021), no início dos estudos, a decisão da carreira dos alunos é considerada alta, com o passar do tempo, a evasão, ou o aluno, se livra da possibilidade de abandono ou constrói a intenção. Para o autor, no primeiro ano, os estudantes enfrentam fatores pessoais e organizacionais complexos, pois ainda não estão adaptados com a estrutura institucional (BARGMANN; THIELE; KAUFFELD, 2021). Ainda mais específicos, Casanova *et al.* (2021) acreditam que as primeiras semanas na instituição são fundamentais para a evasão escolar.

Para Prestes e Fialho (2018), é difícil destacar os fatores que predominam na evasão escolar, porém os autores alegam que a natureza está relacionada com aspectos psicológicos e

individuais. Os fatores financeiros podem ser de situações familiares, socioculturais e de trabalho, já os de natureza acadêmica podem estar relacionados com a trajetória escolar, didática de ensino e estado emocional (PRESTES; FIALHO, 2018).

Na pesquisa de Marques (2020), verificou-se que a evasão se deve a diversos fatores, como baixo nível de comprometimento do aluno com o curso, falta de apoio familiar, baixa participação das atividades acadêmicas, instalações precárias, baixo desempenho escolar e falta de perspectivas na carreira. Já Shirasu e Albuquerque (2016) apontam que os fatores que mais influenciam na evasão escolar são as repetidas reprovações e a falta de interesse em estudar.

De acordo com Herbaut (2020), não existe um consenso de que o histórico social desempenha papel relevante na evasão escolar, a maioria dos estudos é voltado para cursos de bacharelado e não tem preocupação com outros programas de ensino superior. Segundo o autor, o fracasso acadêmico ainda no primeiro ano do ensino superior e o comportamento do abandono ainda são desconhecidos (HERBAUT, 2020).

Ashour (2019) dividiu os motivos que acarretam a evasão escolar em oito questões: a) demografia, histórico familiar, *status* social e escolaridade; b) cultura e temas específicos (exclusivo para estudantes dos Emirados Árabes Unidos); c) temas específicos de gênero (exclusivo para homens ou mulheres); d) aptidões acadêmicas pré-universitária; e) engajamento acadêmico e social; f) fatores ambientais (exclusivo para alunos que não são dos Emirados Árabes Unidos); g) fatores financeiros (exclusivo para expatriados); h) fatores institucionais.

Outra pesquisa que avaliou os fatores que afetam na taxa de abandono é a de Jung e Kim (2017), os autores exploraram fatores regionais e institucionais da evasão de estudantes estrangeiros na Coreia do Sul. Os resultados do trabalho apontaram que fatores institucionais, como tipo, tamanho, matrícula e desempenho do aluno, assim como fatores regionais, como produto interno bruto, inflação e número de estrangeiros na região, influenciam nas taxas de evasão escolar (YUNG; KIM, 2017).

Lima Junior *et al.* (2019) verificaram que a renda média dos alunos que abandonaram seus cursos costuma ser parecida com a renda dos que concluem. Segundo os autores, existem estudantes que não dependem de obter um diploma para que sua independência financeira seja obtida e acabam largando o curso por considerarem ser mais vantajoso. Um fator que parece ser relevante para a permanência ou não de um estudante é a origem social dele (LIMA JÚNIOR *et al.*, 2019). Contrapondo Lima Júnior *et al.* (2019), Hoffman, Nunes e Muller (2016) acreditam que, na maioria dos casos, a principal razão para que os alunos não prossigam nos cursos é a falta de recursos financeiros para que o aluno consiga se manter e terminar seus estudos. Na mesma linha, Silva Filho *et al.* (2007) também afirmam que o recurso financeiro é

o principal problema que afeta os estudantes, além disso, o autor sustenta que a taxa de evasão no primeiro ano de curso é duas a três vezes maior que nos demais anos.

Costa e Gouveia (2018) destacam que os fatores apresentados pelos modelos clássicos de evasão (TINTO, 1975; SPADY, 1971; BEAN, 1980) que influenciam direta ou indiretamente a permanência dos estudantes na graduação, e que são os mais citados na literatura, são:

a) Preparação acadêmica: qualidade do estudo anterior à graduação e preparação para o trabalho.
b) Integração social: relação do estudante com atividades da instituição, relacionamento com colegas e outros estudantes fora do ambiente acadêmico de rotina.
c) Integração acadêmica: quanto mais integrado à instituição o estudante estiver, maior será seu compromisso com o curso. Isso inclui relacionamento com professores dentro e fora da sala de aula, com colegas, participação de grupos de estudos, horas dedicadas a estudo, entre outros.
d) Compromisso com a instituição: esses fatores podem ser influenciados pela integração social, percepções do estudante quanto ao ambiente educacional, assim como particularidades individuais com interesses, competências, anseios. Segundo Costa e Gouveia (2018), esse é um dos fatores defendidos por Spady (1971) que afeta diretamente o desempenho de um estudante.
e) Compromisso com o objetivo: compromisso do estudante com o objetivo de concluir a graduação, principalmente no primeiro ano de graduação. Esse compromisso pode ser influenciado pela qualidade do curso, utilidade do diploma com o esforço necessário para conclusão do curso.
f) Ambiente: inclui recursos financeiros, trabalho e relações familiares, que, direto ou indiretamente, durante o curso são fatores que podem afetar a vida acadêmica. Segundo Costa e Gouveia (2018), as finanças possuem grande impacto nos estudantes.
g) Características demográficas: currículo e competências preexistentes que os estudantes trazem no ensino médio. Segundo Costa e Gouveia (2018), esse é um dos maiores preditores na graduação.

Segundo Hoffman, Nunes e Muller (2016), nos últimos anos, as instituições de ensino superior têm realizado diversos estudos e discussões sobre evasão, porém faltam estatísticas confiáveis e uma metodologia de apuração e de medição. Já para Silva Filho *et al.* (2007), são

raras as instituições de ensino superior brasileiras que possuem programas de combate à evasão escolar compostos de um planejamento de ações, acompanhamento de resultados e experiências bem-sucedidas.

2.2.1 Avaliação e Seleção de Características

Esta seção descreve as variáveis utilizadas em pesquisas anteriores voltadas para a evasão escolar no ensino superior. Dessa forma, será possível analisar e selecionar as características que mais se enquadram na pesquisa.

Na pesquisa de Costa *et al.* (2017), os autores usaram variáveis diferentes para o curso presencial e a distância, a fim de avaliar as reprovações dos estudantes em um curso de programação introdutória em uma universidade pública brasileira. Para a educação a distância, o autor extraiu: idade, sexo, estado civil, cidade, renda, matrícula do aluno, período, aula, semestre, campus, frequência de acesso do aluno no sistema, participação no fórum de discussões, quantidade de arquivos recebidos e visualizados, uso de ferramentas educacionais fornecidas pelo sistema (blog, glossário, quiz, wiki, mensagem), ano de matrícula no curso, *status* da disciplina, desempenho do aluno nas atividades semanais e desempenho do aluno nas provas. Para os cursos presenciais, Costa *et al*. (2017) coletaram idade, sexo, estado civil, cidade, renda, matrícula do aluno, período, aula, semestre, campus, ano de matrícula no curso, *status* da disciplina, quantidade de exercícios realizados pelo aluno, número de exercícios corretos, desempenho do aluno nas atividades semanais e desempenho do aluno em provas.

Sistema acadêmico de uma universidade do Reino Unido, *Moodle* e uso de questionários foram as escolhas de Adejo e Connolly (2018) para selecionar as variáveis em sua pesquisa envolvendo previsão escolar. Por meio dos questionários, os autores selecionaram as variáveis: *status* econômico dos pais, horas de trabalho, qualificação de ingresso, horas média de estudo, apoio familiar, satisfação no curso, impacto da tecnologia, tipo de aprendizagem, estado de saúde, primeira universidade, adaptação, apoio universitário e conhecimento prévio do curso (ADEJO; CONNOLLY, 2018). A partir do sistema acadêmico, foram definidas as variáveis idade, sexo, etnia, localização de moradia, campus, forma de entrada na universidade, qualificação e deficiência. Por fim, do *Moodle*, foram extraídas as variáveis: tempo total de *login* do estudante, número de recursos visualizados, número de tentativas de testes enviados, número de fóruns visualizados e número de discussões em fóruns lidas ou visualizadas (ADEJO; CONNOLLY, 2018).

Outros autores também usufruíram de mais de uma fonte de dados, como é o caso de Muñiz *et al.* (2019), que dividiram sua pesquisa em duas etapas, sendo que a primeira extraiu dados do sistema institucional: dados de identificação, sexo, local de nascimento, nacionalidade, deficiência, tamanho da família, qualificação dos pais e ocupações atuais, nota média do ensino médio, pontuação do exame de admissão universitária, idade quando admitido, data da primeira matrícula, prioridades indicadas no aplicativo de admissão do curso, área do conhecimento correspondente ao curso do aluno, número de créditos inscritos, créditos passados e pontuação média, bolsa, situação acadêmica atual e destino de transferência, quando houver. Em uma segunda etapa, foi realizado um questionário para catalogar: estado civil, nível de renda, tipo de moradia durante o curso, motivação para escolha do curso e universidade, participação em atividades de boas-vindas para calouros e opinião deles, tempo gasto com estudo, trabalho e trabalhos domésticos, avaliação dos requisitos do programa de satisfação com pontuações, avaliação das relações pessoais, intenção de abandono e razões, satisfação com a universidade e, se caso o aluno desistiu, a situação atual e satisfação com os resultados de sua decisão (MUÑIZ *et al.*, 2019).

O trabalho de Casanova *et al*. (2021), assim como o de Muñiz *et al.* (2019), obteve os dados a partir do sistema acadêmico e questionários, porém Casanova *et al.* (2021) usaram os dados de estudantes recém-matriculados no curso, e o questionário foi encaminhado entre seis e oito semanas após seu início. Os autores coletaram: sexo, idade, escolaridade dos pais (educação fundamental, básica ou ensino superior), formação acadêmica e opções vocacionais, curso matriculado, curso de primeira opção, expectativa de finalização do curso. Os questionários encaminhados para os estudantes obtinham 17 questões relacionadas à satisfação com a educação, exaustão acadêmica, intenção de abandono. Casanova *et al*. (2021) verificaram em sua pesquisa que estudantes sobrecarregados devido às atividades acadêmicas ou que demonstraram insatisfação com o curso possuem maior risco de abandono. Esse risco aumenta quando os estudantes não estão no curso de primeira escolha ou na instituição de primeira escolha (CASANOVA *et al*., 2021).

Entre as pesquisas avaliadas, apenas a de Silva, Almeida e Ramalho (2020) avaliou variáveis dos docentes. Os autores avaliaram a reprovação de um estudante em uma disciplina chamada *Cálculo Diferencial e Integral I* em todos os cursos de uma universidade federal brasileira, para isso, definiram as variáveis por meio de quatro dimensões (discente, docente, curso e turma). Na dimensão *discente*, foram definidas as variáveis nota do vestibular total, nota do vestibular total em matemática, casado, migrante, raça, sexo, idade ingresso, cotista, período de ingresso e forma de ingresso. A dimensão docente possui as variáveis: tempo de graduação,

doutorado, publicação no ano, estrangeiro, dedicação exclusiva e sexo. A dimensão curso é definida pelas variáveis local do campus, UF curso, UF centro. E, por fim, na dimensão turma, foram especificadas as variáveis turno, carga horária, média da nota do vestibular, média da nota do vestibular em matemática e taxa de cotista (percentual de discentes cotistas na turma). Os autores utilizaram técnicas de *Machine Learning* e geraram um *ranking* das dez variáveis que mais influenciam no *status* final do discente em cada técnica utilizada (SILVA; ALMEIDA; RAMALHO, 2020). Por fim, as variáveis mais constantes foram: nota do vestibular total em matemática, nota vestibular total, se o docente possui doutorado, estrangeiro ou tem dedicação exclusiva, se o discente é cotista, migrante, idade de ingresso, sexo do estudante, publicação no ano do docente, forma de ingresso no ensino superior, média da nota do vestibular, média da nota do vestibular em matemática e taxa de cotistas. Por meio da classificação do *ranking*, Silva, Almeida e Ramalho (2020) concluíram que todas as dimensões de variáveis foram importantes na *performance* de previsão dos modelos.

Alguns autores usaram o desempenho acadêmico e anterior do aluno para fazer suas previsões. É o caso de Adekitan e Salau (2019) que utilizaram as notas do ensino médio, nível de participação das aulas, assiduidade, notas intermediárias, relatórios de laboratórios, notas de tarefas de casa, pontuação de seminários, conclusão de tarefas e notas gerais para prever a evasão escolar no ensino superior. Já Ezz e Elshenawy (2019) usaram um conjunto de variáveis separadas em duas partes. Dados preparatórios relacionados ao curso: notas de todas as atividades do curso (média de pontuações, notas de exames orais, notas provas práticas, se existirem), nota final, *status* final do aluno. A segunda parte das variáveis está relacionada com as notas do último ano do curso: total de alunos, notas específicas e notas totais do último ano finalizado para alunos que ainda estão no programa (EZZ; ELSHENAWY, 2019).

Além do desempenho acadêmico (média cumulativa de notas, avaliação interna, avaliação externa, atividades extracurriculares, histórico do ensino médio e atividades sociais, Suharjito (2019) utilizou dados demográficos para prever o aluno com risco de abandono. Segundo Suharjito (2019), nos primeiros dois anos, o gênero também influencia na qualidade do aprendizado, assim como as características de idade, restrições financeiras, ausência do aluno, influência dos pais, oportunidade de emprego e estado civil. Beaulac e Rosenthal (2019) usaram o nome do curso, departamento do curso, semestre, valor do crédito do curso e a nota obtida pelo aluno para prever se um aluno concluiria sua graduação e, num segundo momento, os cursos que mais atraiam os alunos.

Silva, Cabral e Pacheco (2020) propuseram um modelo preditivo estatístico para gestão da evasão acadêmica para uma instituição pública de ensino superior brasileira a partir da

avaliação de quatro cursos de graduação EaD da universidade. Para isso, utilizaram as variáveis sexo, cor, estado civil, UF residência, UF polo, reside cidade polo, categoria de ingresso, renda familiar, tamanho da família, tempo de deslocamento até o polo, onde estudo maior parte do ensino médio? (público/privado), experiência no ensino superior (nunca ingressou/já concluiu/já concluiu/ingressou, mas não concluiu, está cursando), experiência no EaD, frequência de uso do computador, local de acesso à internet, tipo de conexão à internet, nível de conhecimento para uso do computador e internet.

Freitas *et al.* (2020) utilizaram os dados socioeconômicos sexo (masculino/feminino), raça (branco, pardo, preto, asiático, indígena, não declarado), origem do ensino médio (privado/federal/estadual/municipal/filantrópico/outro), distância da instituição, renda familiar e índice de desenvolvimento humano por município para identificar os estudantes passíveis de evasão escolar e propor uma ferramenta de previsão de abandono para alunos ingressantes no ensino superior utilizando modelos de *Machine Learning*.

Zulfiker *et al.* (2020) utilizaram dados de uma universidade de Bangladesh para prever as notas de um aluno e as variáveis do ambiente de aprendizado relativos ao semestre de um curso: atendimentos de um aluno, aluno retomou um tópico; durante o semestre, o aluno deve ter três questionários em um determinado curso, aluno respondeu a todos os questionários, média das notas obtidas nos questionários, notas obtidas no exame do meio do semestre, aluno encaminhou as tarefas, aluno executou as apresentações e nota final do aluno após o exame final.

O Apêndice B apresenta a listagem completa de variáveis por autor. A partir dessa listagem, foi feita a compilação dos dados e foram definidas as variáveis que podem ser coletadas dos bancos de dados institucionais (Tabela 5). Variáveis que identificam o estudante ou qualquer outra pessoa envolvida ou, ainda, firam a LGPD não foram incluídas na contagem, pois não serão de forma alguma utilizadas.

Tabela 5 – Variáveis utilizadas na pesquisa

Variáveis	Referências	Quantidade
Sexo	Casanova *et al.* (2021); Silva *et al.* (2020.2); Freitas *et al.* (2020); Silva *et al.* (2020.1); Suharjito (2019); Muñiz *et al.* (2019); Adejo e Connolly (2018); Costa *et al.* (2017).	8
Nota média disciplinas concluídas	Muñiz *et al.* (2019); Suharjito (2019); Beaulac e Rosenthal (2019); Costa *et al.* (2017); Adekitan e Salau (2019); Ezz e Elshenawy (2019); Zulfiker *et al.* (2020).	7
Idade	Costa *et al.* (2017); Adejo e Connolly (2018); Suharjito (2019); Casanova *et al.* (2021); Silva *et al.* (2020.2); Muñiz *et al.* (2019).	6

Variáveis	Referências	Quantidade
Renda familiar per capita	Costa *et al.* (2017); Adejo e Connolly (2018); Muñiz *et al.* (2019); Suharjito (2019); Silva *et al.* (2020.1); Freitas *et al.* (2020).	6
Estado civil	Costa *et al.* (2017); Muñiz *et al.* (2019); Suharjito (2019); Silva *et al.* (2020.1); Silva *et al.* (2020.2).	5
Cidade do aluno	Costa *et al.* (2017); Adejo e Connolly (2018); Muñiz *et al.* (2019); Silva *et al.* (2020.1); Silva *et al.* (2020.2).	5
Raça	Adejo e Connolly (2018); Silva *et al.* (2020.1); Freitas *et al.* (2020); Silva *et al.* (2020.2).	4
Campus/cidade campus	Costa *et al.* (2017); Adejo e Connolly (2018); Silva *et al.* (2020.1); Silva *et al.* (2020.2).	4
Ano/semestre ingresso	Costa *et al.* (2017); Muñiz *et al.* (2019); Silva *et al.* (2020.2).	3
Tipo de ingresso	Adejo e Connolly (2018); Silva *et al.* (2020.1); Silva *et al.* (2020.2).	3
Origem do ensino anterior	Freitas *et al.* (2020); Silva, Cabral e Pacheco (2020).	2
Turno curso	Silva, Almeida e Ramalho (2020); Costa *et al.* (2017).	2
Curso	Casanova *et al.* (2021); Beaulac e Rosenthal (2019).	2
Reprovações	Shirasu e Albuquerque (2016); Ezz e Elshenawy (2019)	2
Semestre	Costa *et al.* (2017); Beaulac e Rosenthal (2019).	2
Índice de desenvolvimento humano por município	Freitas *et al.* (2020).	1
Disciplinas concluídas	Muñiz *et al.* (2019).	1
Tipo de aprendizagem	Adejo e Connolly (2018).	1

Fonte: Elaborada pela autora desta dissertação (2022)

Na próxima seção (2.3), serão apresentados a mineração de dados educacionais e as técnicas de *Machine Learning*, os tipos de aprendizado e os algoritmos de ML, fornecendo autores que já utilizaram os modelos em suas pesquisas voltadas para a gestão universitária.

2.3 MINERAÇÃO DE DADOS EDUCACIONAIS (MDE) E *MACHINE LEARNING* (ML)

A International Educational Data Mining Society (IEDMS, 2021) é uma sociedade internacional de Mineração de dados Educacionais, ela apoia o desenvolvimento de pesquisas na área. A IEDMS define a Mineração de Dados Educacionais (MDE) como sendo um campo emergente que estuda, desenvolve e utiliza métodos para entender o cenário da gestão educacional e melhorar o ambiente para os estudantes (IEDMS, 2021).

A MDE busca, principalmente, construir modelos de alunos, em que a maioria das pesquisas analisadas na avaliação bibliométrica produziu modelos para prever o aprendizado

baseado em comportamentos. O processo de MDE inclui a seleção do conjunto de dados a ser analisado, o pré-processamento desses dados, a transformação dos dados no formato aceitável aos algoritmos, a aplicação dos algoritmos de ML sobre os dados e a interpretação e avaliação de resultados (SULTANA; KHAN; ABBAS, 2017).

Na última década, o aprendizado de máquina, mais conhecido do inglês, *Machine Learning*, tem ganhado evidência devido às grandes estruturas de banco de dados que surgiram. Conseguir analisar e trabalhar com grandes volumes de dados se torna praticamente impossível, assim, Domingos (2017, p. 26) exemplifica que "[...] enquanto um cientista talvez passe sua vida inteira criando e testando algumas centenas de hipóteses, um sistema de *Machine Learning* pode fazer o mesmo em uma fração de segundo".

Mitchell (1997) descreve que *Machine Learning* (ML) envolve a pesquisa de um grande espaço de hipóteses possíveis para determinar a que melhor se adapta aos exemplos de treinamentos disponíveis e outras restrições ou conhecimentos anteriores. Para isso, ML aborda a questão de como construir algoritmos que melhoram seu desempenho em alguma tarefa por meio da experiência (MITCHELL, 1997).

Para Mitchell (1997), os algoritmos de ML provaram ser de grande valor prático em uma extensa diversidade de aplicações, como: problemas de mineração de dados, grandes bancos de dados que podem conter regularidades implícitas e que podem ser descobertos automaticamente, domínios mal compreendidos em que os seres humanos podem não possuir o conhecimento necessário para desenvolver algoritmos eficazes e, campos em que o algoritmo deve se adaptar dinamicamente às mudanças nas condições.

Harrison (2020) descreve o fluxo de trabalho comum em ML a partir da metodologia *Cross-Industry Standard Process for Data Mining* (CRISP-DM), do português, Processo Padrão do Mercado para Mineração de Dados. Esse processo apresenta os passos que podem ser seguidos para a melhoria contínua (HARRISON, 2020).

A Figura 7 apresenta o Processo CRISP-DM, em que a primeira atividade é fazer uma pergunta, ela é necessária para a classificação, no caso da pesquisa prever se um estudante tem a probabilidade de evasão escolar da graduação. Depois da pergunta, é importante a coleta dos dados a serem avaliados, para isso, é preciso saber os atributos/características relevantes para o modelo (fatores que levam os estudantes a evadirem) e, após isso, fazer a limpeza dos dados, garantindo que estejam no formato em que os algoritmos consigam interpretar, retirando os que possuem valores ausentes e, assim, possa ser criado o modelo. Em seguida, ocorre a normalização dos dados (pré-processamento), cujos dados serão padronizados para não tratarem as variáveis com escalas maiores como mais importantes do que as escalas menores,

para isso, são padronizados para que tenham o valor de média igual a zero e desvio-padrão igual a 1 (isso não é feito em algoritmos que trabalham com árvores, pois elas tratam cada atributo por si só). Posteriormente, são realizados a divisão dos dados em treinamento, que utiliza o maior percentual dos dados, e o teste. Logo em seguida, é possível criar o modelo utilizando os algoritmos escolhidos e treiná-los com os dados divididos anteriormente, avaliar o modelo com os dados de teste e implantar, caso obtenha um bom resultado, do contrário, é necessário mudar a pergunta e voltar ao passo de coleta de dados (HARRISON, 2020).

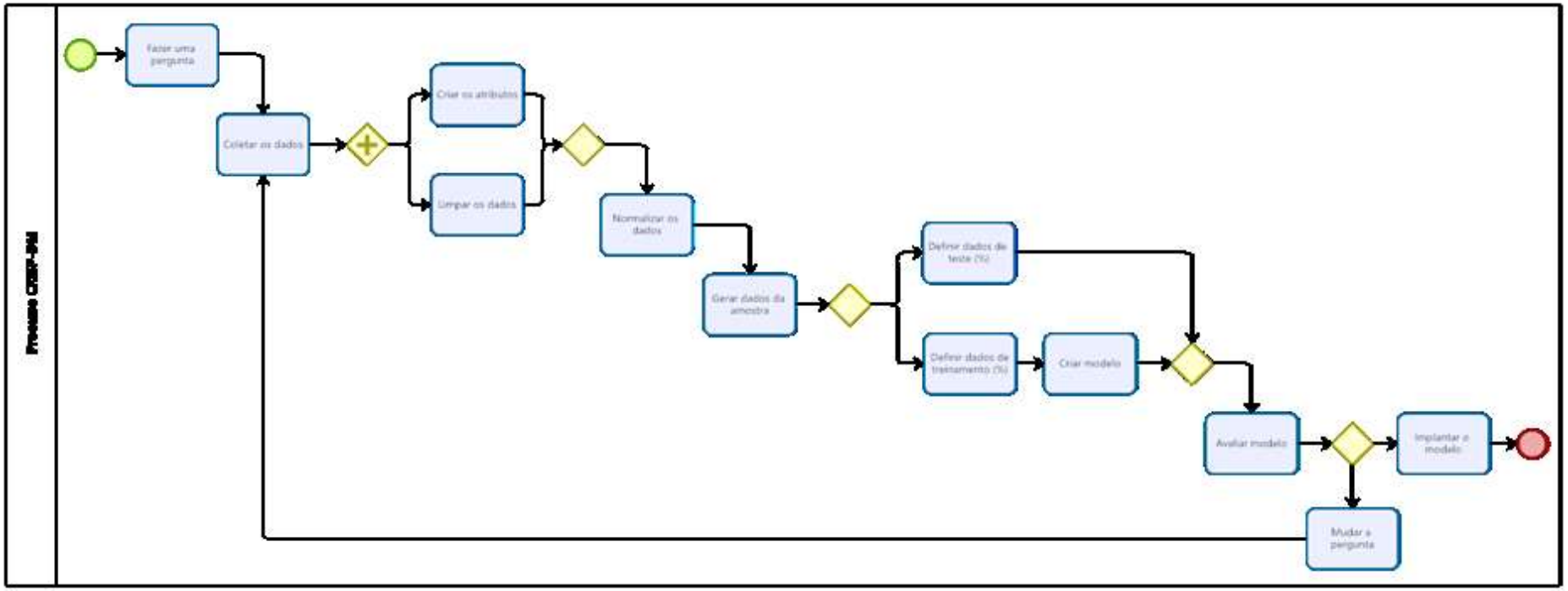


Figura 7 – Processo CRISP-DM
Fonte: Harrison (2020)

2.3.1 Tipos de Aprendizagem

É possível dividir os tipos de aprendizado pelo esforço humano que será necessário para implementá-las (PÁSCOA, 2018). Géron (2019) discorre que o aprendizado pode ser classificado conforme a quantidade e o tipo de supervisão durante o treinamento. Grus (2016) o classifica em quatro tipos (supervisionado, não supervisionado, semissupervisionado e *on-line*), porém o autor não trata do aprendizado *on-line*. Já Géron (2019) classifica como sendo supervisionado, não supervisionado, semissupervisionado, aprendizado por esforço, *on-line* e em lote. Nesse caso, serão especificados apenas os modelos supervisionados, não supervisionados e semissupervisionados, pois são os mais amplamente utilizados nas pesquisas relacionadas.

Modelos Supervisionados: são aqueles modelos que possuem um conjunto de dados etiquetados/rotulados com a resposta correta para o aprendizado (GRUS, 2016; GÉRON, 2019). Esse aprendizado é usado principalmente em problemas de classificação (KUMAR; KRISHNA, 2020), mas a regressão também pode ser utilizada. Dessa forma, o objetivo do aprendizado supervisionado é construir um modelo de classificação pelo qual o computador aprende sobre os dados e, então, consegue prever a saída; para construir o modelo, são necessárias entradas (mapear novos dados) e respostas (saídas desejadas) (KUMAR; KRISHNA, 2020).

Entre os algoritmos de aprendizado supervisionado, estão *K-Nearest Neighbors* (KNN), Regressão Linear, Regressão Logística, S*upport Vector Machine* (SVM), *Decision Trees*, *Random Forests, Naive Bayes* e *Artificial Neural Networks* (algumas redes neurais são não supervisionadas) (GÉRON, 2019; KUMAR; KRISHNA, 2020).

Um exemplo bem popular para esse tipo de aprendizado é de ensinar um computador a identificar e a distinguir animais, como um porco de um cachorro. Para isso, são inseridas imagens de porcos e rotuladas como “porco”, da mesma forma são inseridas imagens de cachorros e rotuladas como “cachorro”. Após isso, no modelo, são feitos treinamentos com os dados que identifiquem corretamente o resultado pretendido. Depois, usa-se os dados nos algoritmos de ML para que eles testem seus aprendizados classificando as imagens corretamente (PÁSCOA, 2018).

Dois problemas básicos podem ser retratados em modelos supervisionados: classificação e regressão. Na classificação, são definidos rótulos entre os existentes para identificar algo. Já na regressão, a ideia é prever um alvo de valor numérico a partir de um conjunto de características chamadas de previsores (GÉRON, 2019). Xiao e Yi (2020)

descrevem que, para aplicações educacionais, os pesquisadores costumam empregar algoritmos de classificação.

Modelos Não Supervisionados: não existem etiquetas/rótulos com resposta correta para aprendizado (GRUS, 2016; GÉRON, 2019). Assim, esse tipo de aprendizado não possui saídas específicas, que, geralmente, mostram as semelhanças entre os dados, analisam padrões de interesse, enfim, tentam aprender mais sobre os dados (KUMAR; KRISHNA, 2020).

Dos algoritmos mais importantes, estão *Clustering* (K-Means, *Clustering* Hierárquico [HCA], Maximização da Expectativa), Visualização e Redução da Dimensionalidade (Análise de Componentes Principais [PCA], *Kernel* PCA, *Locally-Linear Embedding* [LLE]) e Aprendizado da Regra da Associação (*Apriori*, *Eclat*) (GÉRON, 2019).

Um exemplo apresentado por Páscoa (2018, p. 30) com modelos não supervisionados é usar os algoritmos para "[...] identificar segmentos de clientes com atributos similares de marketing e recomendar itens".

Modelos Semissupervisionados: apenas alguns dados são etiquetados/rotulados (GRUS, 2016), a maior parte dos dados não rotulados e poucos rotulados (GÉRON, 2019). Géron (2019) afirma que esses modelos são uma combinação de modelos supervisionados e não supervisionados.

Para extrair informações importantes, a MDE pode colaborar com a análise detalhada de alunos do banco de dados institucional. Para isso, serão utilizados os algoritmos *Decision Tree*, *Artificial Neural Network* (*Multilayer Perceptron*) e *XGBoost* para a criação do modelo.

2.4 ALGORITMOS DE *MACHINE LEARNING*

Silva, Almeida e Ramalho (2020) sustentam que técnicas de *Machine Learning* têm sido eficientes para tratar previsão do risco de um evento ocorrer, principalmente quando se relaciona com o grau de automatização do processo de modelagem, estimativa, testes e definição do melhor modelo de predição.

Para Géron (2019), o principal desafio de ML se divide em algoritmos ruins e dados ruins. Entre os dados ruins, podem estar: quantidade insuficiente de dados de treinamento, pois a maioria dos algoritmos precisa de uma grande quantidade de dados para funcionar corretamente; dados de treinamento não representativos, se a amostra for muito pequena ou não representativa do objeto que se quer estudar.

Em ML os métodos podem ser lineares, o que geralmente não são suficientes para responder a problemas mais complexos, para isso, modelos não lineares podem ser usados, pois eles possuem estruturas não lineares a respeito dos parâmetros a estimar (PÁSCOA, 2018).

Segundo Géron (2019), os principais desafios de *Machine Learning* estão em problemas de dados ruins ou algoritmos ruins, que podem ocorrer devido à: Quantidade insuficiente de dados de treinamento, em que são necessários muitos dados para que boa parte dos algoritmos de ML funcione corretamente. Também são necessários milhares de exemplos, mesmo para resolver problemas simples; Dados de treinamento não representativos, no qual dificilmente o modelo treinado fará previsões precisas, com isso, é muito importante que o conjunto de dados seja representativo em casos que se deseja generalizar. Amostras muito pequenas gerarão um ruído de amostragem, e amostras muito grandes podem não representar se o modelo de amostragem for falho (a isso dá-se o nome de viés de amostragem); Dados de baixa qualidade, com muitos erros, *outliers* e ruídos deixarão o sistema com mais dificuldade de detectar padrões implícitos, podendo não ter um bom desempenho. Para isso, é importante a limpeza dos dados de treinamento; Características irrelevantes devem ser suprimidas, pois o sistema somente conseguirá aprender corretamente se os dados de treinamento tiverem características relevantes suficientes e poucas características não relevantes. O processo para criar um bom conjunto de características é chamado de *feature engineering*, o qual passa por três subprocessos (seleção de características, extração das características e criação de novas características ou coletar novos dados); *Overfitting* os dados de treinamento, o sobreajuste significa que o modelo funciona bem com os dados de treinamento, porém não generaliza bem; *Underfitting* os dados de treinamento é o oposto do *overfitting*. O modelo é muito simples para aprendizado da estrutura tácita dos dados (GÉRON, 2019).

Os algoritmos de *Machine Learning* foram desenvolvidos para solucionar diversos problemas distintos, sendo que nenhum subjuga os demais em todos os tipos de conjunto de dados. Delen (2010) utilizou *Artificial Neural Network (Multilayer Perceptron)*, *Decision Tree*, *Support Vector Machine*, Regressão Logística e *Baggings* para desenvolver um modelo e identificar nos estudantes ingressantes os que possuem maior probabilidade de abandono após o primeiro ano, apontando as variáveis mais importantes e aplicando análise de sensibilidade nos modelos desenvolvidos. O autor usou as técnicas de ML individualmente e em conjunto, sendo que SVM foi o algoritmo que teve o melhor desempenho. O autor afirma que algoritmos usados em conjunto com outros obtiveram melhor *performance* quando comparados com individuais e que dados balanceados trouxeram melhor desempenho do que dados não balanceados independentemente do algoritmo utilizado. Do mesmo modo, o algoritmo *Decision*

Tree fornece uma visão mais transparente de onde e de como fazem quando comparado com *Support Vector Machine* (DELEN, 2010).

De acordo com Sultana, Khan e Abbas (2017), nos países europeus mais desenvolvidos, a taxa de abandono no curso de Engenharia Elétrica é de 40% a 50% durante o primeiro ano, e algumas disciplinas chegam a 80%, sendo um grande problema para a gestão universitária (SULTANA; KHAN; ABBAS, 2017). Para tanto, os autores acreditam que uma solução para controlar a taxa de evasão escolar é adotar um mecanismo de previsão que avise os estudantes sobre seu potencial desempenho ruim para que assim eles possam melhorar suas notas. Para isso, Sultana, Khan e Abbas (2017) usaram as técnicas *Decision Tree*, Regressão logística, *Naive Bayes* e *Artificial Neural Networks* em sua pesquisa.

Com o intuito de prever as reprovações no estágio inicial, o suficiente para que sejam tomadas providências a fim de reduzir estas falhas, Costa *et al.* (2017) selecionaram os algoritmos *Naive Bayes*, *Decision Tree*, *Artificial Neural Network* e *Support Vector Machine*. Para os autores, os resultados obtidos com os algoritmos selecionados foram parecidos, porém o *Support Vector Machine* foi o que atingiu melhor eficácia no trabalho (COSTA *et al.*, 2017).

A fim de desenvolver um modelo analítico para prever alunos com baixo engajamento das diferentes atividades de um curso num ambiente virtual de aprendizado, Hussain *et al.* (2018) se beneficiaram dos algoritmos baseados em árvores de decisão (*Decision Tree*, J48, *Classification and Regression Tree*, JRIP, *Gradient Boosting*) e *Naive Bayes*.

Adejo e Connolly (2018) empregaram *Decision Tree, Artificial Neural Network, Support Vector Machine* e *Ensemble Methods Baggins, Boosting* e *Stacking* para criar e testar sete modelos distintos a partir da junção dos algoritmos e diferenciação das fontes de dados utilizadas para previsão de desempenho acadêmico para estudantes, já que os autores conseguiram verificar que um modelo híbrido de classificadores *ensemble* com as três fontes de dados foi considerado o melhor.

Com o objetivo de extrair regras que ajudem a gestão universitária a identificar caminhos de abandono, ajudando a prever o fenômeno antes que aconteça, evitando a desistência de estudantes por meio de medidas adequadas para aumentar a persistência, Muñiz *et al.* (2019) avaliaram a evasão escolar em uma universidade da Espanha que operou os algoritmos de ML (C45, *Random Forest*, *Classification and Regression Tree*, *Naive Bayes* e *Support Vector Machine*).

Mia *et al.* (2019) utilizaram os algoritmos *Support Vector Machine*, *Naive Bayes*, *Regressão Logística*, JRip, J48, *Multilayer Perceptron* e *Random Forest* para prever o número de alunos registrados em um semestre para ajudar no pré-planejamento de uma universidade

privada de Bangladesh. Eles compararam os classificadores e, em sua pesquisa *Support Vector Machine* obtiveram melhor precisão *Random Forest* da menor precisão (MIA *et al.*, 2019).

Os estudos de Suharjito (2019) tiveram o propósito de analisar e de encontrar uma melhor solução de modelagem na identificação de preditores de abandono de alunos em uma universidade de Jacarta. Para isso, os autores exploraram os algoritmos *K-Nearest Neighbor*, *Naive Bayes* e *Decision Tree*. Ao combinar os algoritmos com método *Ensemble Classifier* e testado várias vezes a precisão, eles tiveram melhor desempenho. Entre os algoritmos utilizados na pesquisa, *K-Nearest Neighbors* foi o que obteve melhor desempenho (SUHARJITO, 2019).

Adekitan e Salau (2019) aplicaram os algoritmos *Probabilistic Neural Network*, *Random Forest*, *Decision Tree*, *Naive Bayes*, *Tree Ensemble* e Regressão Logística de forma independente para prever a Média Cumulativa de Notas Final dos alunos com dados dos três primeiros anos de graduação. O algoritmo que apresentou melhor *performance* foi Regressão Logística. Por fim, os autores combinaram todos os algoritmos em um modelo para obter os benefícios de cada um em conjunto (ADEKITAN; SALAU, 2019).

Tendo o escopo de desenvolver uma metodologia para recomendar um departamento adequado para o estudante, baseado no seu desempenho acadêmico do ano preparatório de uma universidade do Egito, Ezz e Elshenawy (2019) empregaram *Support Vector Machine*, *K-Nearest Neighbor*, *Random Forest*, Regressão Linear e *Quadratic Discriminant Analysis* (QDA). Assim, foi selecionado um algoritmo para cada departamento conforme a *performance* alcançada, QDA foi o melhor para o Departamento Urbano, RF obteve melhor desempenho no Departamento de Mecânica e Departamento de Civil, KNN foi o escolhido como melhor no Departamento de Arquitetura e Departamento Elétrico e Regressão Linear foi selecionado para o Departamento de Mineração e Petróleo e SVM para o Departamento de Informática. Além disso, o algoritmo *Random Forest* foi o que obteve melhor desempenho entre todos com 82,57% no Departamento de Engenharia Civil (EZZ; ELSHENAWY, 2019).

A pesquisa de Silva, Almeida e Ramalho (2020) teve o objetivo de identificar o risco de reprovação de alunos do ensino superior em uma universidade federal brasileira, desse modo, os autores empregaram Regressão Logística, *K-Nearest Neighbors*, *Naive Bayes*, *Support Vector Machines*, *Decision Tree Based Methods* (C *trees*, *Baggins*, *Random Forest*, *Boosting*) e *Penalized Methods* (*Ridge*, *Lasso*, *Elastic Net*) (SILVA; ALMEIDA; RAMALHO, 2020). Da mesma forma, Zulfiker *et al.* (2020) utilizaram oito algoritmos de ML (*Support Vector Machine*, Regressão Logística, *K-Nearest Neighbor*, *Decision tree*, *AdaBoost*, *Multilayer Perceptron*, Classificador de árvore extra e Classificador de votação ponderada) para prever as notas finais dos alunos e analisar se o aluno está em risco de reprovação no exame final, foi

comparado o resultado dos algoritmos para identificar o que oferece melhor desempenho. Os autores utilizaram um conjunto de dados de diferentes cursos e departamentos, somando informações de 400 alunos, sendo que a abordagem proposta obteve uma precisão de 81,73% com classificadores de votação ponderada, utilizando diferentes rótulos de classes, o que, segundo Zulfiker *et al.* (2020), é uma proposta suficientemente confiável. Quando os autores utilizaram apenas dois rótulos de classes, o resultado aumentou significativamente para 93,26%. O algoritmo que obteve menor precisão foi o de Regressão Logística (66,35% com diferentes rótulos de classes), porém, quando apenas com dois rótulos de classes obtiveram um aumento para 81,73% (ZULFIKER *et al.*, 2020).

Com a finalidade de analisar os fatores que impactam o desempenho dos alunos no último ano de curso e de propor um modelo preditivo com taxas de precisão aprimoradas em comparação com outros no mesmo conjunto de dados, Gamie, El-Seoud e Salama (2020) aplicaram os algoritmos de *Machine Learning Support Vector Machine, Decision Tree* e *Neural Networks* em cada partição (combinação de características); Após, aproveitou o algoritmo *Random Forest* como técnica de ensacamento em cada partição (combinação de recursos); Aplicou *XgBoost* e *AdaBoost* com *Decision Tree* como aprendiz base de cada partição (combinação de recursos); Selecionou o melhor classificador junto com a melhor combinação de grupos; Aumentou o *Support Vector Machine* linear e não linear e o *Random Forest*, salvou os resultados e comparou a precisão da classificação. O modelo seguiu as seguintes etapas: inicializar grupos de recursos com base em fontes de dados; gerar combinações possíveis de grupos de repetição a fim de detectar a melhor combinação (GAMIE; EL-SEOUD; SALAMA, 2020).

Do mesmo modo de Gamie, El-Seoud e Salama (2020), seguindo níveis de classificação, Vidhya e Vadivu (2020) desenvolveram um modelo de classificação de estudantes, para isso, na primeira classificação, foi usada *Support Vector Machine, Naive Bayes* e J48 para o primeiro treinamento, e, para o segundo nível de treinamento, foram usados os *ensembles methods Baggins* e *Stacking* (VIDHYA; VADIVU, 2020).

Com o objetivo de prever o aprendizado de alunos no conjunto de dados de amostra de uma plataforma de aprendizagem, Xiao e Yi (2020) dividiram seus dados aleatoriamente em dois conjuntos de treinamento e dois conjuntos de teste. Para os treinamentos, Xiao e Yi (2020) utilizaram os algoritmos k-Nearest Neighbor, Support Vector Machine e Random Forest e realizou a precisão. Após, para o teste, o autor escolheu os algoritmos Random Forest e Decision Tree, em que a precisão dos dois ficaram muito próximas, RF obteve 81,9% e DT 82,9% (XIAO; YI, 2020).

Alguns estudos preferiram utilizar apenas um algoritmo para fazer suas previsões. Dessa maneira, Beaulac e Rosenthal (2019) utilizaram o algoritmo *Random Forest* para prever se o aluno concluirá seu curso de graduação e quais cursos atraem mais os alunos. Os autores justificam que RF é de fácil uso, rápido para treinar e consegue superar modelos de Regressão Linear em precisão de previsões. Por fim, Beaulac e Rosenthal (2019) ajustaram os classificadores de RF e os compararam com dois modelos de Regressão Logística para prever se um aluno conclui seu curso. Da mesma forma, Fernández, Gil e Mora (2019) utilizaram apenas o algoritmo *Decision Tree* para realizar a previsão de desempenho para estudantes do curso de engenharia de sistemas de computação de uma universidade do Equador. Segundo os autores, a técnica utilizada foi escolhida porque seus resultados podem ser interpretados e explicados com facilidade por meio de gráficos que resumem o modelo de regras de decisão implícitas (FERNÁNDEZ; GIL; MORA, 2019). Por fim, Chui *et al.* (2020) propuseram um modelo reduzido de avaliação com o *Support Vector Machine* para prever alunos em riscos e possíveis alunos a abandonarem seus cursos. Após, os autores realizaram uma comparação com pesquisas relacionadas, assim, concluíram que seu modelo poderia ser adotado, pois reduz o tempo de treinamento quando o conjunto de dados fornecido é grande (CHUI *et al.*, 2020).

Seguindo o mesmo caminho de Beaulac e Rosenthal (2019), Fernández, Gil e Mora (2019), Chui *et al.* (2020) e Musso, Hernández e Cascallar (2020) utilizaram redes neurais artificiais (*Multilayer Perceptron* e *Attention Network Test)* a fim de propor um modelo de *Machine Learning* para prever resultados educacionais de alunos durante sua trajetória acadêmica, também identificaram a contribuição de cada variável para cada um dos vários resultados. Com a pesquisa, os autores concluíram que é possível identificar os preditores de resultados educacionais positivos e os relacionados ao fracasso, podendo, nestes últimos, realizar intervenções precoces para que as instituições consigam mudar os resultados negativos (MUSSO; HERNÁNDEZ; CASCALLAR, 2020).

A fim de desenvolver um modelo para prever o abandono acadêmico numa faculdade do Peru, Gismondi e Huiman (2021) utilizaram MLP e *XGBoost*. Como resultado, os autores acreditam que com um modelo de previsão criado aumentaria a taxa de graduação em uma determinada faculdade de 28,89% a 50,13% para 40,37% a 58,47%. Também se observou que variáveis acadêmicas, como número de semestres e média dos alunos e variáveis demográficas do tipo internet e celular, foram as que mais influenciaram a predição do modelo (GISMONDI; HUIMAN, 2021).

Para atender ao segundo e terceiro objetivos específicos da pesquisa, serão analisados os seguintes algoritmos de *Machine Learning: Decision Tree* – amplamente utilizados. Pode

ser simples de entender e ter pouca preparação dos dados, porém, por ser considerado um algoritmo de menor precisão, este será utilizado como *baseline* do modelo. *Artificial Neural Network* (*Multilayer Perceptron*) – pode trabalhar com grandes volumes de dados e de grande complexidade, também é bastante utilizado em pesquisas relacionadas. E, por fim, o algoritmo *XGBoost*, por ser uma evolução das árvores de decisão (*boosting*) mais recente.

2.3.2 Árvores de Decisão (*Decision Tree* – DT)

A árvore de decisão é um método que usa uma estrutura de uma árvore para representar um número de possíveis caminhos de decisão e um resultado para cada um dos caminhos (GRUS, 2016). As árvores de decisão, do inglês *Decision Tree* (DT), são modelos de fácil entendimento e interpretação e usam um processo totalmente transparente, podendo lidar com muitos atributos numéricos e categóricos, também conseguem classificar dados dos atributos faltantes (GRUS, 2016). Nos problemas de regressão, a previsão para uma observação é igual a média ou a moda das observações de treinamento da região à qual ela pertence. As regras de divisão dos segmentos se resumem em árvores.

De acordo com Mitchell (1997), as árvores de decisão classificam as instâncias, da raiz até algum nó folha, que fornece a classificação da instância. Cada nó especifica um teste de algum atributo da instância. Em outras palavras, a árvore vai sofrendo diversas divisões conforme as respostas que são obtidas em cada nó até chegar à folha, em que se encontra a resposta final. Páscoa (2018) chama de estratégia de "dividir para conquistar", em que se divide um problema complexo em subproblemas menores e mais simples de serem resolvidos.

De acordo com Harrison (2020), uma árvore de decisão assemelha-se a ir a um médico, quando este faz uma série de perguntas a fim de determinar o que provoca seus sintomas. Da mesma forma, uma DT cria uma série de questionamentos para prever uma classe-alvo (resultado). Outro exemplo mais visual criado por Mitchell (1997) (Figura 6) traz o jogo de tênis, no qual se supõe que se pretenda jogar tênis numa manhã, para isso, é necessário verificar as condições meteorológicas. São dados os atributos/características (Estado do Tempo – ensolarado/nublado/chuvoso, Umidade, Vento) e os classificadores (sim/não). Se a resposta for sim, por exemplo, o tempo está ensolarado e a umidade normal, então está propício para o jogo de tênis (MITCHELL, 1997).

Desse modo, a árvore é vista de "cabeça-para-baixo", no caso da Figura 8, o "Tempo" é a raiz – o ponto de partida da árvore –, após os nós (Ensolarado, Nublado, Chuvoso), depois os ramos (Umidade, Vento) que conduzem até as folhas as quais são os pontos de decisão da

árvore. Conforme os problemas a serem resolvidos pela árvore se tornam mais complexos, vão crescendo mais ramos na DT, porém a base do raciocínio permanece inalterada (PÁSCOA, 2018). O trajeto da raiz até a folha corresponde a uma regra de classificação, a qual realiza uma combinação de testes dos atributos, e a árvore, a separação dessas conjunções (MITCHELL, 1997).

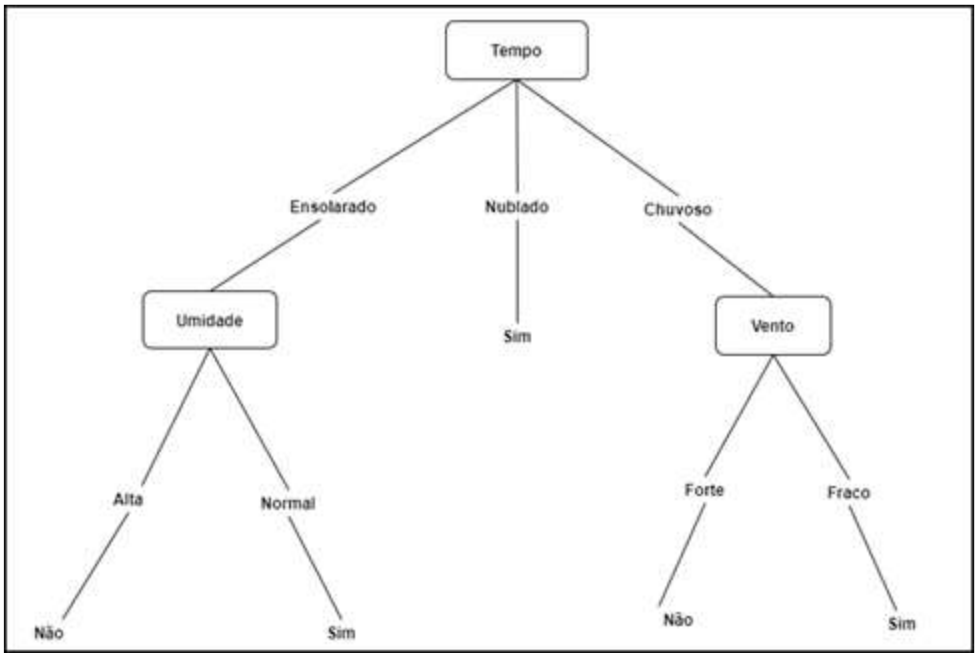


Figura 8 – Estrutura de uma árvore de decisão
Fonte: Mitchell (1997)

Mitchell (1997) levantou a questão de que é necessário controlar o crescimento de uma árvore de decisão, porém é complicado determinar a melhor dimensão, principalmente quando há ruído na amostra ou quando as amostras de treino são muito pequenas, para isso, se dá o nome de *Overfitting* (MITCHELL, 1997; PÁSCOA, 2018). Para solucionar esse problema, é efetuado o procedimento de poda que estabelece um limite de parada, o qual pode ser efetuado uma pré-poda que a segmentação dos nós termina e transforma o nó em uma folha e em pós-poda, que, após a construção da árvore, remove ramos inteiros a partir de um determinado nó (PÁSCOA, 2018).

Nos problemas de classificação, pode ser usado o algoritmo padrão CART (*Classification And Regression Tree* – Árvore de Classificação e Regressão) que utiliza uma medida de índices para tomada de decisão chamada de Gini, com isso, o algoritmo percorre os atributos em um laço até encontrar o valor que forneça a menor probabilidade de erro de classificação (HARRISON, 2020), em outras palavras, o índice de Gini controla a poda verificando se vale a pena quebrar um nó, quando o valor de Gini chega a zero, o nó não pode

mais ser dividido, ela, então, passa a ser uma folha. A Figura 9 apresenta um exemplo de árvore de decisão que utiliza a medida de Gini.

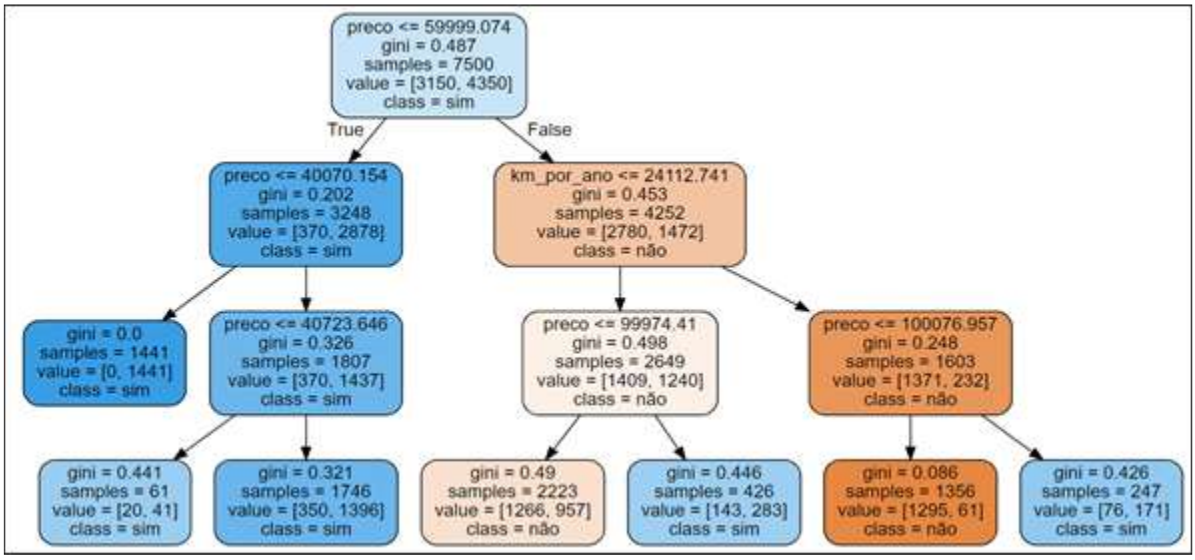


Figura 9 – Classificação Gini
Fonte: Elaborada pela autora desta dissertação (2022)

Além do coeficiente de Gini, que é aplicado por padrão, as árvores de decisão também podem usar a medida de impureza da entropia (*entropy*), a qual é utilizada como uma medida de coeficiente, já que a entropia de um conjunto é considerada zero quando ela contém instâncias de apenas uma classe (GÉRON, 2019). Segundo Géron (2019), utilizar o coeficiente de Gini ou a entropia, traz como resultado árvores semelhantes, porém Gini é um pouco mais rápida para calcular e, nos casos em que elas diferem, o índice de Gini conduz-se a isolar a classe mais frequente no seu próprio ramo da árvore, já a entropia inclina-se a criar árvores mais equilibradas (GÉRON, 2019; DECISIONTREECLASSIFIER, 2022).

A aplicação de árvores de decisão possui diversas vantagens, segundo Harrison (2020) e Grus (2016): intuitivas, simples de entender e usam preditores qualitativos; suporte para dados não numéricos; pouca preparação dos dados; suporte para trabalhar com relacionamentos não lineares; a importância dos atributos é revelada e fácil de explicar.

Em conformidade com Grus (2016), Géron (2019) e Harrison (2020), as Árvores de Decisão possuem as desvantagens de obter menor precisão e serem ruins para lidar com relacionamentos lineares, com isso, uma pequena mudança em um número pode levar a um caminho diferente, as DTs são extremamente dependentes dos dados de treinamento e possuem inconsistências – pequenas mudanças podem gerar grandes alterações nos resultados.

Grus (2016) descreve que é computacionalmente muito difícil conseguir uma DT perfeita para um determinado conjunto de dados, com isso, mesmo sendo complexo, tenta-se

construir árvores boas o bastante. Um problema acontece quando não conseguem generalizar bem os dados desconhecidos e as árvores são construídas com dados de treinamento *overfitting*, isso pode ser corrigido com florestas aleatórias – ver Seção 2.4.4 (GRUS, 2016).

O algoritmo *DecisionTreeClassifier* da biblioteca Sklearn possui diversos parâmetros que podem ser utilizados para melhorar a árvore: o número mínimo de amostras em que o nó deve ter antes que possa ser dividido (min_samples_split), número mínimo de amostras em que o nó da folha deve ter (min_samples_leaf), número máximo de nós da folha (max_leaf_nodes), número máximo de características que são avaliadas para a divisão em cada nó (max_features), profundidade máxima da árvore (max_depth), fração ponderada mínima da soma total de pesos de todas as amostras de entrada necessárias para estar em um nó folha (min_weight_fraction_leaf), divisão de um nó caso esta é a divisão induzir uma diminuição da impureza maior ou igual a este valor (min_impurity_decrease), entre outros (GÉRON, 2019; DECISIONTREECLASSIFIER, 2022).

Os algoritmos que utilizam uma árvore de decisão são amplamente utilizados em trabalhos voltados para a gestão universitária. A seguir, constam pesquisas que utilizaram pelo menos um algoritmo de árvores de decisão: Costa *et al.* (2017), Zulfiker *et al.* (2020), Delen (2010), Sultana, Khan e Abbas (2017), Adejo e Connolly (2018), Suharjito (2019), Adekitan e Salau (2019), Mia *et al.* (2019), Vidhya e Vadivu (2020), Gamie, El-Seoud e Salama (2020), Xiao e Yi (2020), Muñiz *et al*. (2019) e Fernández, Gil e Mora (2019).

As árvores de decisão podem ser mais simples como as *Decisions Trees* exemplificadas nesta seção, ou mais complexas (usam mais de uma árvore de decisão e técnicas mais recentes de *Machine Learning* – regularização) como *Baggins, Random Forests, Boosting, Stacking* (e *Ensemble Methods* derivados).

2.3.3 *Boosting*

Os métodos de *Ensemble Learning* utilizam múltiplas árvores, procurando melhorar o processo de decisão, entre eles, estão: *Boosting, Bagging, Random Forests* e derivados. Eles possuem as vantagens de gerar previsão consensual única, aumento da acurácia de previsão, porém possuem a desvantagem de eventual piora na interpretação. Para Géron (2019), modelos *ensemble* geralmente terão melhores desempenho do que modelos individuais, assim, *boosting* tendem a ter melhores precisões do que *Decision Tree*.

Os métodos de *boosting*, também chamados de *hypothesis boosting*, seguem a ideia de treinar os previsores sequencialmente, cada um tentando corrigir o anterior por meio de uma

combinação de vários classificadores fracos em classificadores fortes (GÉRON, 2019). Segundo Silva, Almeida e Ramalho (2020), classificadores fracos possuem uma predição anterior melhor do que uma classificação aleatória por apresentar uma taxa de erro menor, conseguindo assim, construir um classificador responsável pela predição final mais forte (SILVA; ALMEIDA; RAMALHO, 2020). A Figura 10 apresenta a representação de um *boosting*, no qual as árvores com os dados originais são treinadas e avaliadas, após os dados ponderados, esses dados são treinados e avaliados de forma sequencial (GEEKS, 2022).

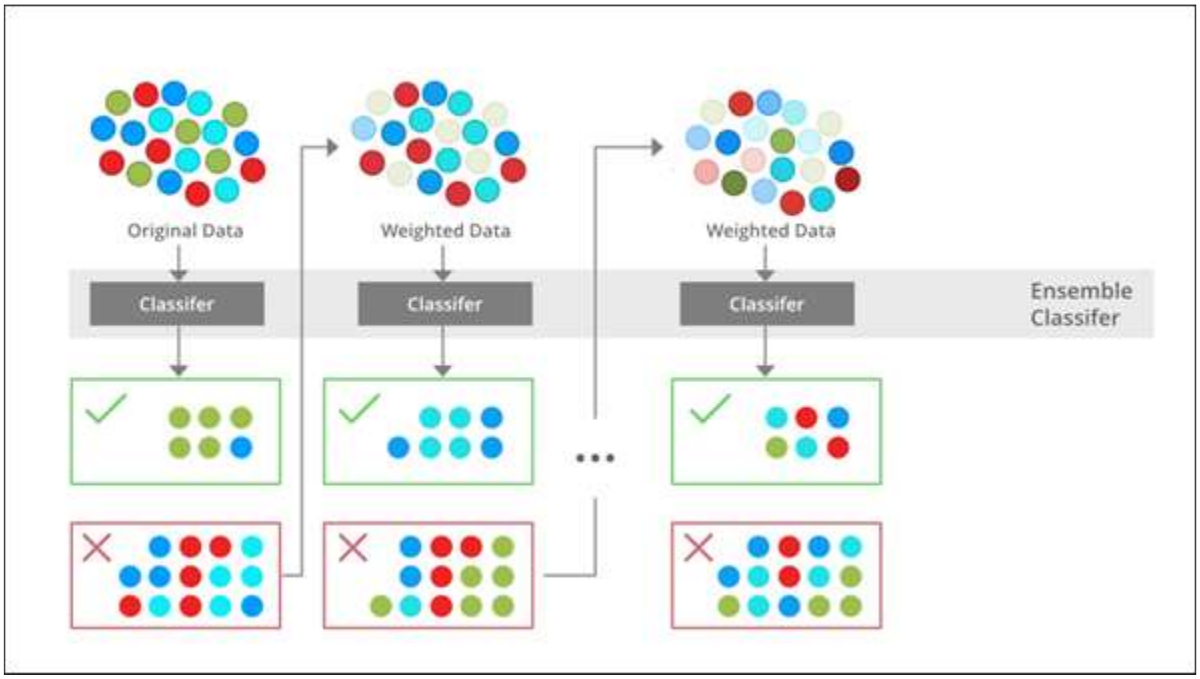


Figura 10 – Treinamento boosting
Fonte: Geeks (2022)

Gradient Boosting é um *boosting* bem popular, ele inclui previsores sequencialmente a um conjunto, tentando ajustar o novo previsor aos erros residuais feitos pelo previsor anterior (GÉRON, 2019). O *XGBoost* (*eXtreme Gradient Boosting*) é um algoritmo de ML escalonável que tenta melhorar o processo de desempenho do *Gradient Boosting* por meio da otimização dos *softwares* e *hardwares* e tem oferecido resultados de última geração para uma gama de problemas produzindo resultados superiores a outras técnicas, usando árvores de decisão com menos recursos computacionais em menor tempo (XGBOOST, 2022).

Segundo Harrison (2020), utilizar o XGBoost ao invés do Gradient Boosting tende a trazer melhores resultados, pois ele inclui paralelização e regularização (evita *overfitting*). Dessa forma, o modelo *XGBoost* cria uma árvore fraca e, após, melhora as árvores subsequentes com o intuito de diminuir os erros residuais, tentando capturar e tratar qualquer padrão, mesmo

que aleatório nos erros encontrados, o algoritmo então irá combinar as previsões das árvores criadas para gerar sua previsão após o final do treinamento, o treinamento é repetido inserindo novas árvores com a capacidade de tratar os erros residuais, assim como os erros de árvores anteriores, que, combinados com as árvores anteriores, geram a previsão final (HARRISON, 2020).

Segundo Chen e Guestrin (2016), o sucesso do *XGBoost* é a escalabilidade em todos os cenários, a qual se deve a vários sistemas importantes e otimizações algorítmicas. Dessa forma, em uma única máquina, o algoritmo é escalonado para bilhões de exemplos em configurações distribuídas ou com memória limitada, fazendo com que o algoritmo funcione mais de dez vezes mais rápido do que outras soluções existentes (CHEN; GUESTRIN, 2016; XGBOOST, 2022). Esse algoritmo compreende a regularização para evitar modelos muito complexos que levam ao *overfitting*. Ele aceita variáveis esparsas para o treinamento, o que faz com que o algoritmo trabalhe com dados faltantes sem necessidade de pré-processamento e um esboço de quantil ponderado para aprendizado aproximado de árvores (CHEN; GUESTRIN, 2016).

O modelo XGB tem as propriedades de Eficiência na execução, tendo a vantagem de poder construir as árvores em paralelo (*boosting* não conseguem executar em paralelo, fazendo com que cada previsor só possa ser treinado após o previsor anterior ter sido treinado e avaliado) e utilizando a opção chamada n_jobs para informar o número de CPUs e a GPU para melhorar seu desempenho; no pré-processamento, nos dados, não é necessário realizar escalonamento, porém é preciso codificar os dados de categoria. Contém parâmetro (*early_stopping_rounds* = N) que consegue evitar superadequação, interrompendo o treinamento caso não tenha melhora após N (número definido por meio do parâmetro) rodadas; Interpretação dos resultados incluindo importância de atributos (HARRISON, 2020). Outro parâmetro (*eval_set*) apresentado pelo *XGBoost* faz com que o modelo pare de criar árvores caso as métricas de avaliação não tenham melhorado após o número de rodadas definidas (HARRISON, 2020).

Mais um parâmetro apresentado pelo modelo (*importance_type*) consegue gerar a importância das variáveis por meio do peso (número de vezes que a variável aparece nas árvores), também sendo possível mudar o valor *default* (peso) pelo ganho médio que a variável é usada ou o número de amostras afetado por uma separação, podendo ser usado também para verificar a importância de variáveis dentro de outros modelos. Após, é possível gerar o gráfico ou uma representação textual das árvores, também sendo possível apresentar a versão gráfica da árvore (Figura 11) (HARRISON, 2020).

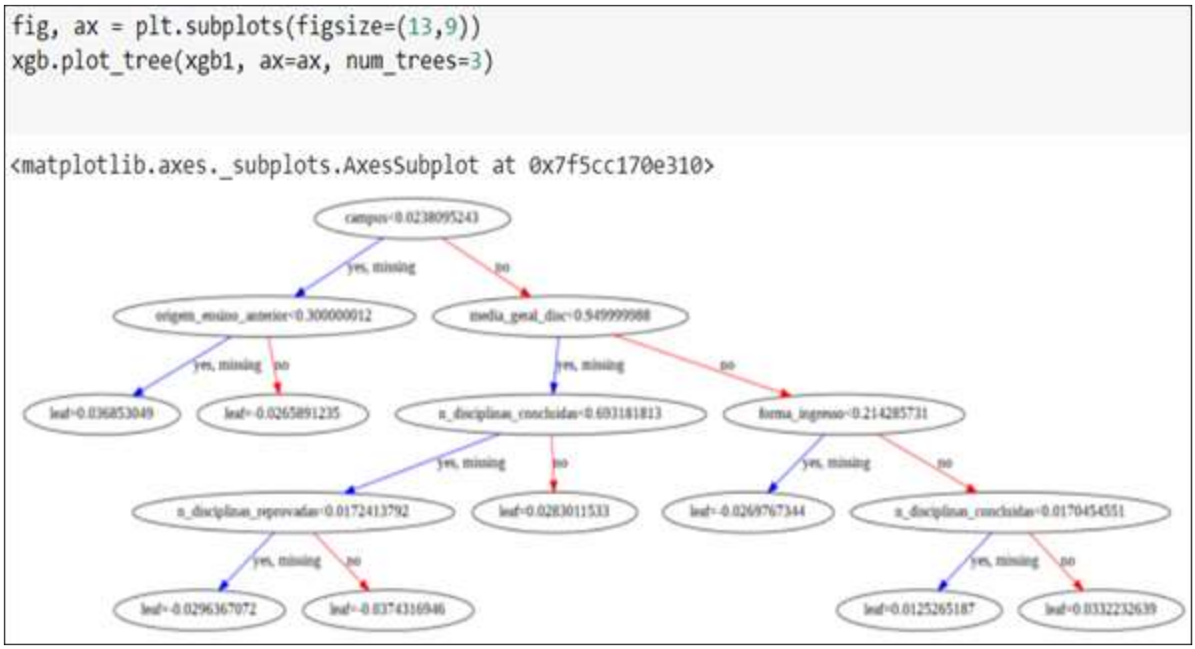


Figura 11 – Treinamento gerado pelo XGBoost
Fonte: Elaborada pela autora desta dissertação (2022).

Por fim, o *XGBoost* é um pacote de código aberto portátil e reutilizável que suporta várias classificações ponderadas e funções de classificação de objetivos e funções de objetivos definidas pelo usuário (CHEN; GUESTRIN, 2016).

Os autores Silva, Almeida e Ramalho (2020), Adejo e Connolly (2018), Gamie, El-Seoud e Salama (2020) e Zulfiker *et al.* (2020) utilizaram *boosting* e derivados para fazer previsões relacionadas com a gestão universitária, sendo que Gamie, El-Seoud e Salama (2020) e Gismondi e Huiman (2021) usaram *XGBoost* para classificação e Pillai (2019) utilizou o *XGBoost* para regressão.

2.3.4 Redes Neurais Artificiais (*Artificial Neural Network* – ANN ou RNA)

As RNAs podem ser utilizadas para problemas de classificação e de regressão. Géron (2019) defende que o uso de RNAs supera muitas outras técnicas de ML quando usado para problemas grandes e complexos, podendo ser treinados em um espaço razoável de tempo.

As Redes Neurais Artificiais (RNAs), do inglês *Artificial Neural Network* (ANN), foram inspiradas pela observação do funcionamento do cérebro humano, construídas a partir de teias complexas de neurônios interconectados. Warren McCulloch e Walter Pitts criaram o modelo do neurônio biológico que depois ficou conhecido como neurônio artificial (GÉRON, 2019).

Mitchell (1997, p. 1) explica que as

> RNAs são construídas a partir de um conjunto densamente interconectado de neurônios simples, onde cada neurônio leva um número de entradas de valor real (possivelmente as saídas de outros neurônios e produz uma única saída de valor real (que pode se tornar a entrada para muitos outros neurônios).

Em outras palavras, é uma rede genuína de múltiplos neurônios do tipo discriminadores lineares, em que os neurônios são organizados em várias camadas, com uma "camada de entrada", que recebe as entradas diretamente, uma segunda camada chamada de "camada oculta", que recebe as saídas da primeira camada como entrada e assim por diante, até chegar na "camada de saída" (KOVÁCS, 2006). Para Géron (2019), as redes neurais artificiais estão gradualmente se tornando diferentes das biológicas e alguns pesquisadores já descartam a comparação e até mesmo o uso do nome "neurônios", trocando para "unidades".

A Figura 12 fornece uma rede neural artificial realizando cálculos lógicos simples em que: a Rede 1 – Função Identidade, o neurônio C, que recebe dois sinais de A, somente é ativado quando o neurônio A estiver ativo; a Rede 2 – Execução de E lógico, o neurônio C somente será ativado quando os dois neurônios A e B estiverem ativos; a Rede 3 – Execução de OU lógico, o neurônio C é ativado quando um ou os dois neurônios (A, B) estiverem ativos; já a Rede 4 – Cálculo de proposição lógica ligeiramente mais complexo, o neurônio C é ativado apenas quando o neurônio A estiver ativo e o neurônio B estiver desligado (GÉRON, 2019).

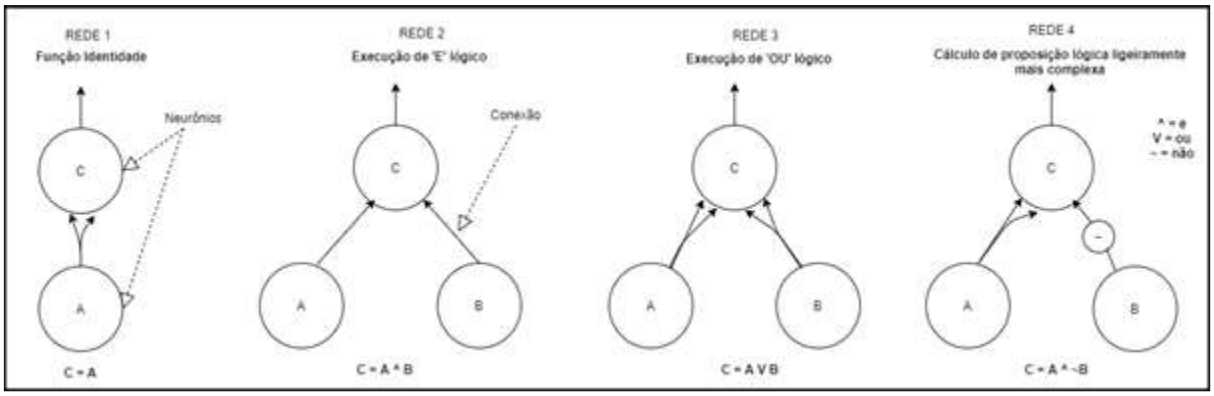


Figura 12 – RNA realizando cálculos lógicos simples
Fonte: Géron (2019)

A rede neural mais simples é chamada de *Perceptron*, desenvolvida no final dos anos 1950 por Frank Rosenblatt, constituída de um único neurônio com *n* entradas binárias (GRUS, 2016; GÉRON, 2019). Essas redes podem resolver problemas simples (como AND, OR e NOT). E, quando é necessário resolver problemas mais complexos, são necessárias redes neurais chamadas *Feed-Forward,* que apresentam as chamadas "camadas ocultas", assim elas possuem camadas de neurônios, em que cada uma é ligada na seguinte (GRUS, 2016; GÉRON, 2019). Com isso, ela apresenta uma camada de entradas que as transmite sem modificações,

uma ou mais camadas ocultas compostas por neurônios com saídas da camada anterior que realizam o cálculo e passam o resultado para a próxima camada e, por fim, uma camada de saída que produz as saídas finais. Para simplificar, é inserido o bias (polarização) como uma característica extra, com isso, cada neurônio, que não seja de entrada, recebe uma entrada polarizada que é sempre igual a 1 (GRUS, 2016). O bias também é chamado de neurônio de viés (GÉRON, 2019). Um exemplo desse tipo de rede de múltiplas camadas é *Multilayer Perceptron* (MLP).

Na Figura 13 tem-se a representação de uma RNA XOR "OR, mas não AND", nela é possível observar a transmissão de uma saída de um neurônio AND e um neurônio OR em um neurônio que possui a segunda entrada, mas não a primeira entrada com resultado da rede XOR (GRUS, 2016).

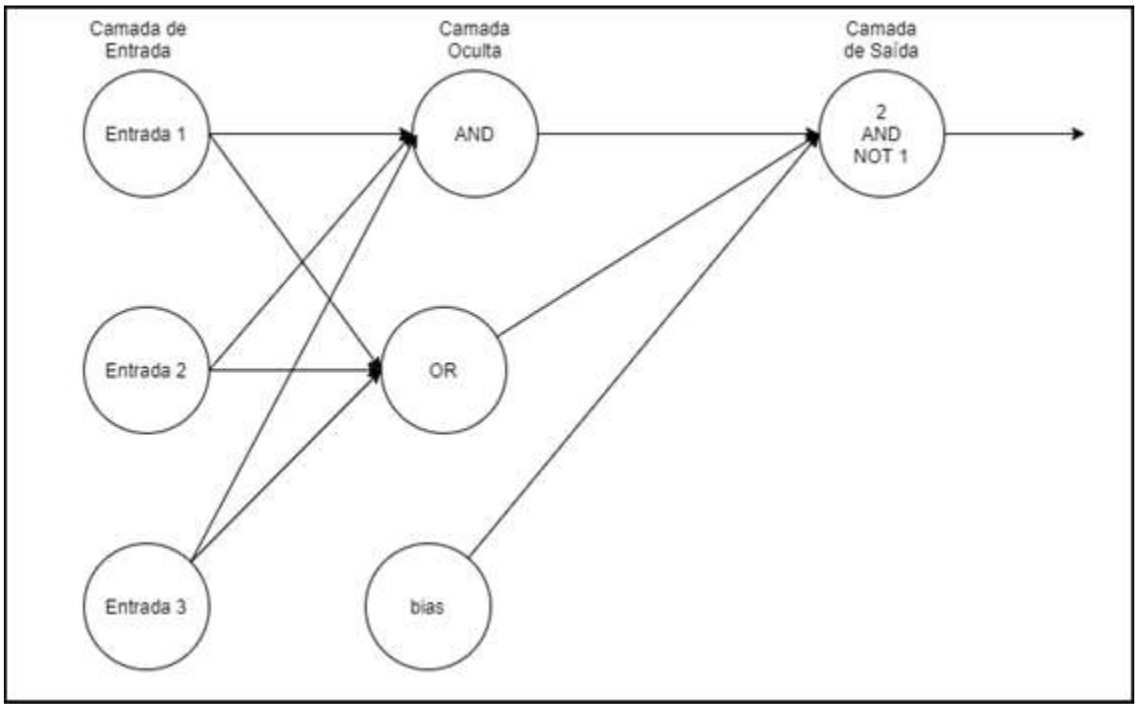


Figura 13 – Rede Neural para XOR
Fonte: Grus (2016)

O algoritmo de *Multilayer Perceptron* da biblioteca Sklearn para classificação é o *MLPClassifier*, do português classificador de *perceptron* de várias camadas, que possui diversos parâmetros que melhoram a classificação: número máximo de iterações (*max_iter*), ativação para a camada oculta (*activation*), i-ésimo elemento que representa o número de neurônios na i-ésima camada oculta (*hidden_layer_sizes*), otimizador de peso, em que a configuração padrão (*adam*) tem melhores resultados de tempo de treinamento e validação para conjuntos de dados grandes e quando conjuntos de dados pequenos têm melhor desempenho

com a configuração ‘ibfgs’ (*solver*), taxa de aprendizado inicial (*learning_rate_init*), cronograma de taxa de aprendizado para atualizações de peso (*learning_rate*), embaralhamento das amostras em cada iteração (*shuffle*), interrupção do treinamento para a pontuação de validação sem melhorias (*early_stopping*), entre outros.

Musso, Hernández e Cascallar (2020) utilizaram redes neurais em sua pesquisa, a fim de prever resultados dos estudantes (*Attention Network Test; Multilayer Perceptron).* Outros autores que incluíram pelo menos um algoritmo de redes neurais artificiais foram Gamie, El-Seoud e Salama (2020), Adekitan e Salau (2019), Suharjito (2019), Adejo e Connolly (2018), Sultana, Khan e Abbas (2017), Delen (2010), Zulfiker *et al.* (2020), Costa *et al.* (2017) e Gismondi e Huiman (2021).

A partir da revisão teórica, surgiram as seguintes hipóteses da pesquisa:

H₁: Usando as técnicas de *Machine Learning* será possível criar regras para prever a evasão de todos os cursos de graduação do IFSC.

H₂: Algumas variáveis podem ser de difícil interpretação ou que não foram possíveis de serem coletadas.

H₃: É possível identificar variáveis que são relevantes para a conclusão dos cursos de graduação do IFSC.

No próximo capítulo, serão apresentados os procedimentos metodológicos aplicados à pesquisa, como a classificação da pesquisa, os dados e a amostra utilizados e a definição dos modelos de pesquisa.

3 MÉTODO DE PESQUISA

Quanto à filosofia, esta pesquisa analisou as quatro filosofias de Saunders, Lewis e Thornhill (2019), positivismo, realismo, interpretativismo e pragmatismo, definindo que esta pesquisa pode ser considerada positivista, já que a lógica e a matemática podem ser válidas por estabelecerem as regras da linguagem, constituindo-se de um conhecimento *a priori*, e, com isso, é independente da experiência (TERENCE; ESCRIVÃO FILHO, 2006). Para os positivistas, o que não pode ser verificado empiricamente são sentenças sem sentido, portanto, para cada enunciado que faça sentido, deve ser possível decidir se ele é falso ou verdadeiro (ALVES-MAZZOTTI; GEWANDSZNAJDER, 1999). Este trabalho busca, por meio de métodos estatísticos (algoritmos de *Machine Learning*) a sua validação. O Positivismo, na visão filosófica da ontologia, é independente e objetivo, quanto à epistemologia, utiliza-se somente de dados e de informações críveis; quanto à axiologia, o pesquisador não impõe seus valores, mantendo sua posição independente; e quanto às técnicas de coleta de dados, é quantitativo, estruturado e utiliza grandes amostras (SAUNDERS; LEWIS; THORNHILL, 2019).

Quanto ao pensamento ou à abordagem, esta pesquisa é considerada dedutiva. Essa abordagem é racional, procura confirmar uma hipótese e, para isso, se utiliza de métodos estatísticos. Nesse método, se todas as premissas forem consideradas verdadeiras, a conclusão também deve ser verdadeira (FREGONEZE *et al.*, 2014).

Quanto ao propósito desta pesquisa, ela é explicativa, segundo os objetivos, para Gil (2008), esse tipo de pesquisa tem a preocupação de identificar fatores que determinam ou contribuem a suceder um fenômeno, aprofundando o conhecimento da realidade, pois explica o porquê das coisas, sendo uma pesquisa mais complexa e delicada. Segundo Gil (2008), a pesquisa explicativa é uma continuação das pesquisas descritivas e exploratórias, pois, para a sua realização, é necessário que o fenômeno a ser estudado esteja descrito e detalhado de maneira suficiente.

Quanto à estratégia ou tipo de pesquisa, é por método quantitativo, obedecendo ao paradigma clássico positivista. Esse método permite chegar a verdades universais, é reprodutível e generalizável. Para isso, envolve coletar e analisar dados numéricos e aplicar testes estatísticos, tanto na fase de coleta de dados quanto no tratamento (ALVES-MAZZOTTI; GEWANDSZNAJDER, 1999). Esta pesquisa busca, por meio de métodos estatísticos representados pelos algoritmos de *Machine Learning*, gerar treinamento e testes de uma base de dados para analisar dados da evasão escolar.

O recorte temporal é longitudinal, pois se conhece o efeito e se busca a causa. Também nesse tipo de pesquisa, a análise se dá ao longo de um determinado tempo, delimitado por períodos.

A coleta e análise de dados ocorre por meio preferencialmente de fontes secundárias, em registros de arquivos, e os dados são coletados diretamente do banco de dados institucional. Mais detalhes na Seção 3.2 Dados e Amostra. A Figura 14 mostra a classificação da pesquisa conforme prevê a "*The Research Onion*" de Saunders, Lewis e Thornhill (2019).

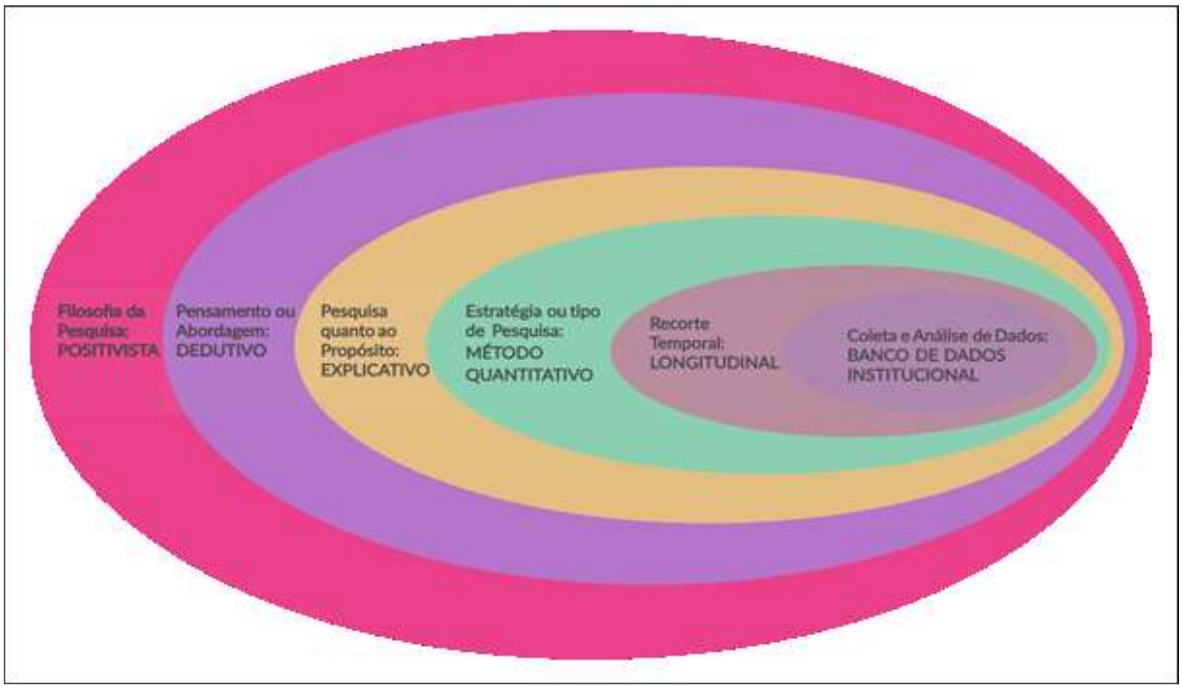


Figura 14 – Classificação da pesquisa
Fonte: Elaborada pela autora desta dissertação (2022)

3.1 POPULAÇÃO E AMOSTRA

A população inicial selecionada para o estudo foi composta de duas bases de dados, a primeira com os dados dos estudantes ingressantes dos anos de 2017, 2018 e 2019, e, para análise, uma segunda base com os dados dos anos de 2020 e 2021, nos cursos de graduação dispostos em 44 cursos únicos, presenciais e a distância (EaD) do Instituto Federal de Santa Catarina (IFSC).

Após o pré-processamento dos dados, a base de dados permaneceu com 7.759 linhas (cada linha representando um estudante) e 19 colunas, em que cada uma representa uma característica (variável). Após a primeira avaliação do modelo, observou-se as características mais importantes, e as três menos importantes foram retiradas do modelo, fazendo uma segunda avaliação com 16 variáveis (colunas).

Para introduzir os dados utilizados, será apresentada a instituição – Instituto Federal de Santa Catarina (IFSC) e seus indicadores nas próximas seções.

3.1.1 Instituto Federal de Santa Catarina (IFSC)

O Instituto Federal de Santa Catarina (IFSC) foi criado em Florianópolis no ano 1909 a partir do Decreto n. 7.566, de 23 de setembro, com o nome de Escola de Aprendizes Artífices de Santa Catarina, pelo então presidente Nilo Peçanha, com o objetivo de proporcionar formação profissional aos filhos de classes menos favorecidas, entre 10 e 13 anos de idade, oferecendo cursos do ensino primário e formações em desenho, tipografia, entre outros. Em 1937, devido à legislação, a instituição teve a primeira troca de nomenclatura para Liceu Industrial de Florianópolis, com a finalidade de propagar a educação profissional, ela focava em formar profissionais que atendessem às demandas das indústrias que estavam em ascensão. Após isso, a instituição passou por diversos nomes, ampliando seus cursos e abrangendo outras cidades, como São José e Jaraguá do Sul, até que, em 2008, a Lei n. 11.892 transformou o então Centro Federal de Educação Tecnológica de Santa Catarina (CEFET-SC) em Instituto Federal de Santa Catarina. O IFSC passou a ofertar cursos de educação profissional de nível médio, pesquisa aplicada, ensino superior e pós-graduação *lato sensu* e *stricto sensu*, expandindo-se em todas as regiões do estado com 22 *Campi* e um centro de educação a distância (CERFEAD) (IFSC, 2021).

Os *Campi* do IFSC possuem autonomia administrativa e oferta própria de cursos, elaboradas com base em necessidades locais. Atualmente, a instituição atua em três modalidades da educação profissional e tecnológica: Formação Inicial e Continuada (FIC), Educação Profissional Técnica de Nível Médio e Educação Superior Tecnológica de Graduação (IFSC, 2018).

Como instituto federal, a instituição deve dispor do maior percentual de vagas para cursos profissionais técnicos de nível médio e educação de jovens e adultos (50%), e 20% para cursos de licenciatura e programas de formação pedagógica (BRASIL, 2008).

Na próxima seção, serão avaliados os indicadores do Instituto Federal de Santa Catarina fornecidos pela Plataforma Nilo Peçanha (PNP), que possui os dados da Rede Federal de Educação Profissional, Científica e Tecnológica SETEC/MEC.

3.1.2 Indicadores do Instituto Federal de Santa Catarina

Os dados utilizados nesta seção foram retirados da Plataforma Nilo Peçanha (PNP) de 2020, com o ano base de 2019, selecionando o campo “Instituição” como IFSC e o campo “Tipo Curso” como Bacharelado, Licenciatura e Tecnologia, com o intuito de obter apenas cursos de graduação. Outras limitações realizadas são explicadas ao longo do texto.

A Plataforma Nilo Peçanha (PNP) é um ambiente virtual de coleta, validação e disseminação das estatísticas oficiais da Rede Federal de Educação Profissional, Científica e Tecnológica (Rede Federal) e tem como objetivo reunir dados relativos ao corpo docente, discente, técnico-administrativo e aos gastos financeiros das unidades da Rede Federal, para fins de cálculo dos indicadores de gestão monitorados pela Secretaria de Educação Profissional e Tecnológica do Ministério da Educação (SETEC/MEC).

O IFSC possui 23 unidades, 573 cursos, 44.724 matrículas, 24.033 ingressantes, 10.537 concluintes, 27.949 vagas e 109.372 inscritos entre curso de graduação (bacharelado, licenciatura e tecnologia), técnicos (subsequentes e integrados), qualificação profissional (FIC), especialização (*lato sensu*) e mestrado acadêmico (*stricto sensu*). Os cursos de graduação estão distribuídos em 22 unidades, nas quais são 70 cursos, 10.213 matrículas, 3.074 ingressantes, 785 concluintes, 3.103 vagas e 18.480 inscritos no ano de 2019 (BRASIL, 2021). A Tabela 6 dispõe sobre a distribuição dos cursos no IFSC.

Tabela 6 – Distribuição de cursos no IFSC em 2019

	Unidades	Cursos	Matrículas	Ingressantes	Concluintes	Vagas	Inscritos
Geral	23	573	44.724	24.033	10.537	27.949	109.372
Graduação	22	70	10.213	3.074	785	3.103	18.480

Fonte: Plataforma Nilo Peçanha (BRASIL, 2021)

Os cursos de graduação no IFSC são de três tipos: bacharelado, licenciatura e tecnologia. Os cursos de bacharelado são 23, e no ano de 2019 foram 4.470 matrículas, 1.371 ingressantes, 138 concluintes, 1.344 vagas e 8.178 inscritos; 14 cursos são de licenciatura com 1.135 matrículas, 254 ingressantes, 108 concluintes, 293 vagas e 944 inscritos. Por fim, são 33 cursos de tecnologia, 4.408 matrículas, 1.449 ingressantes, 539 concluintes, 1.466 vagas e 9.358 inscritos. A Tabela 7 disponibiliza a distribuição dos cursos de graduação do IFSC por tipo de curso.

Tabela 7 – Distribuição dos cursos de graduação do IFSC por tipo de curso

Tipo de Curso	Cursos	Matrículas	Ingressantes	Concluintes	Vagas	Inscritos
Bacharelado	23	4.670	1.371	138	1.344	8.178
Licenciatura	14	1.135	254	108	293	944
Tecnologia	33	4.408	1.449	539	1.466	9.358

Fonte: Plataforma Nilo Peçanha (BRASIL, 2021)

A Tabela 8 traz os cursos de graduação do IFSC divididos por eixo e subeixo tecnológico.

Tabela 8 – Distribuição dos cursos por eixo e subeixo

Eixo Tecnológico	Subeixo Tecnológico	Cursos	Matr.[1]	Ingr.[2]	Concl.[3]	Vagas	Insc.[4]
	Meio ambiente	1	101	41	12	50	41
Ambiente e saúde	Saúde	4	496	157	58	155	3.077
	Automação	5	1.055	234	76	229	1.013
Controle e processos	Elétrica	7	1.997	595	27	613	3.152
industriais	Mecânica	4	660	218	0	180	1.602
Desenvolvimento	Desenvolvimento						
educacional e social	educacional	14	1.135	254	108	293	944
Gestão e negócios	Gestão e negócios	2	590	103	208	81	208
Informação e	Informática	6	1.024	348	61	338	1.992
comunicação	Telecomunicações	3	285	72	19	72	391
Infraestrutura	Civil	5	646	205	28	203	649
	Agroindústria	2	72	25	2	40	45
Produção alimentícia	Alimentos	4	262	85	45	146	266
Produção cultural e							
design	Design	5	845	318	83	278	1.834
	Mecânica	1	348	94	27	91	408
Produção industrial	Química	1	95	46	9	43	189
Recursos naturais	Agrícola	3	251	140	0	140	595
Turismo, hospitalidade	Hospitalidade	2	272	94	95	95	1.520
e lazer	Turismo	1	79	45	56	56	554

[1]Matr. = Matrículas. [2]Ingr. = Ingressantes. [3]Concl. = Concluintes. [4]Insc. = Inscritos.
Fonte: Plataforma Nilo Peçanha (BRASIL, 2021)

Ao analisar o sexo e a faixa etária dos alunos matriculados na graduação do IFSC no ano de 2019, é possível verificar que a instituição possui quase o dobro de alunos do sexo masculino quando comparado com o feminino em todos os conjuntos de idades analisados. Outro ponto relevante é que a idade com mais alunos está entre 20 e 24 anos com 4.189 matrículas, seguida de 25 a 29 anos com 1.885 matrículas e de 15 a 19 anos com 1.589 matrículas. A Tabela 9 fornece o percentual das matrículas por idade e sexo e os seus totais.

Tabela 9 – Distribuição dos cursos por idade e sexo

Sexo	Total	15-19 anos (%)	20-24 anos (%)	25-29 anos (%)	30-34 anos (%)	35-39 anos (%)	40-44 anos (%)	45-49 anos (%)	50-54 anos (%)	55-59 anos (%)	+[1] 60 anos (%)
Fem.[2]	3.520	15,88	41,46	18,57	9,11	6,89	4,25	1,85	1,21	0,56	0,22
Masc.[3]	6.693	15,43	40,77	18,44	10,44	6,90	3,85	2,05	1,20	0,60	0,40
Total Geral	10.213	15,56	41,02	18,44	9,99	6,90	4,00	1,98	1,20	0,58	0,33

[1]Pessoas maiores de 60 anos. [1]Fem. = feminino. [1]Masc. = masculino.
Fonte: Plataforma Nilo Peçanha (BRASIL, 2021)

Quando analisado o percentual geral dos cursos de graduação na PNP que declaram sua classificação racial, nota-se que esse percentual é de 74,17% dos matriculados. Já no IFSC, apenas 29,41% dos alunos matriculados declararam sua classificação racial, o que dificulta saber a totalidade da instituição, já que se sabe apenas sobre uma minoria dos matriculados. Dos que declaram, 51,36% são brancos, 36,62% são pardos e 9,59% são pretos.

Quanto à renda familiar dos matriculados, 69,06% declaram, e 28,58% têm renda *per capita* maior que 0,5 e até 1,0 salário, 22,06% maior que 1,0 até 1,5 salários, 19,62% de 0 até 0,5 salário e 19,52% maior que 1,5 até 2,5 salários. Na Tabela 10, é feito o comparativo de classificação racial e de renda familiar entre o percentual dos declarados.

Tabela 10 – Classificação racial X renda familiar (% do total geral)

Renda Familiar	Amarela	Branca	Indígena	Parda	Preta	Não Declarada	Total Geral
Total Geral	**0,58%**	**15,11%**	**0,14%**	**10,77%**	**2,82%**	**70,58%**	**10.213**
0<RFP[1]<=0,5	0,51%	12,64%	0,22%	19,65%	5,14%	61,84%	1.384
0,5<RFP<=1,0	0,84%	12,95%	0,25%	15,23%	4,22%	66,51%	2.016
1,0<RFP<=1,5	0,64%	17,36%	0,13%	12,98%	3,66%	65,23%	1.556
1,5<RFP<=2,5	1,09%	16,05%	0,29%	10,82%	2,32%	69,43%	1.377
2,5<RFP<=3,5	1,00%	25,00%	0,00%	8,50%	2,75%	62,75%	400
RFP>3,5	0,63%	47,81%	0,00%	9,06%	2,19%	40,31%	320
Não Declarada	0,12%	11,49%	0,00%	3,39%	0,79%	84,21%	3160

[1]RFP = Renda Familiar *per capita*.
Fonte: Plataforma Nilo Peçanha (BRASIL, 2021)

A próxima seção traz os dados de evasão nos cursos de graduação (bacharelado/licenciatura/tecnologia) do IFSC.

3.1.3 Evasão nos Cursos de Graduação no IFSC

Primeiramente, faz-se necessário entender a classificação de matrículas definidas, para isso, a Plataforma Nilo Peçanha (BRASIL, 2021) apresenta cinco tipos de classificação de matrícula para alunos dados como evadido que foram conceituadas no Documento Orientador para Superação da Evasão e Retenção na Rede Federal de Educação Profissional, Científica e Tecnológica (BRASIL, 2014):

a) Abandono (Evadido): o estudante abandona o curso, não realizando a renovação da matrícula ou formalizando o desligamento/desistência.
b) Cancelada: o estudante solicita o trancamento de uma matrícula.
c) Desligada (Desistente): o estudante comunica formalmente, de forma espontânea, o desejo de não permanecer no curso.
d) Transferência Externa: quando o aluno solicita transferência para outra instituição.
e) Transferência Interna: quando o aluno solicita mudança de curso dentro da própria instituição.

Na avaliação da situação das matrículas de 2019 do Instituto Federal de Santa Catarina, com os filtros do Tipo de Curso em Bacharelado, Licenciatura e Tecnologia, foram obtidos os resultados da Tabela 11, em que 7.564 matrículas estavam em curso, 785 concluintes e 1.864 evadidos, o que representa, respectivamente, 74,07%, 7,69% e 18,25% (BRASIL, 2021).

Tabela 11 – Situação das matrículas dos cursos de graduação do IFSC em 2019

	Em Curso		**Concluintes**	**Evadidos**				
	Em Fluxo	Retido	Em Fluxo	Abandono	Cancelada	Desligada	Transf. Externa	Transf. Interna
	6.254	1.310	785	903	7	919	34	1
Total		7.564	785					1.864

Fonte: Plataforma Nilo Peçanha (BRASIL, 2021)

Os cursos da modalidade presencial podem ser separados por turnos: integral, matutino, noturno e vespertino. No turno integral são 2.519 matrículas em curso, 165 concluídas e 421 evadidos, respectivamente, 81,13%, 5,31% e 13,56%. O turno matutino possui 909 alunos em curso, 85 concluídos, 194 evadidos, o que representa 76,52%, 7,15% e 16,33%, respectivamente. Nos cursos de turno noturno são 3.279 matrículas em curso, 246 concluídos, 890 evadidos, respectivamente, 74,27%, 5,57% e 20,16%. Por fim, no turno vespertino são 541

matrículas em curso, 26 concluídos, 170 evadidos, 73,4%, 3,53%, 23,06%, respectivamente (BRASIL, 2021).

Os cursos da modalidade a distância (EaD), na PNP, estão classificados como "Não se aplica a nenhum" turno, 316 em curso, 263 concluintes e 189 evadidos, 41,15% em curso, 34,24% concluintes e 24,61% evadidos. A Tabela 12 apresenta as situações de matrícula modalidade X turno mais detalhadamente.

Tabela 12 – Situação das matrículas dos cursos de graduação do IFSC em 2019 - Modalidade X Turno

Turno	**Em Curso**		**Concluintes**	**Evadidos**				
Modalidade Presencial	Em Fluxo	Retido	Em Fluxo	Abandono	Cancelada	Desligada	Transf. Externa	Transf. Interna
Integral	2.056	463	165	218	3	183	16	1
Matutino	775	134	85	98	0	92	4	0
Noturno	2.683	596	246	421	4	457	8	0
Vespertino	458	83	26	101	0	63	6	0
Modalidade EaD	282	34	263	65	0	124	0	0

Fonte: Plataforma Nilo Peçanha (BRASIL, 2021)

A próxima seção apresenta o Plano Estratégico de Permanência e Êxito dos Estudantes do IFSC, o qual fornece uma análise dos fatores de evasão na instituição.

3.1.4 Plano Estratégico de Permanência e Êxito dos Estudantes do IFSC

A Nota Técnica n. 282/SETEC/MEC, de 9 de julho de 2015, orienta as instituições da rede federal brasileira na construção de Planos Estratégicos Institucionais para a Permanência e Êxito dos Estudantes e dispõe do diagnóstico das causas da evasão e retenção e implementação de políticas e ações para ampliar a permanência e êxito de estudantes nas instituições da rede federal (BRASIL, 2015).

O Plano Estratégico de Permanência e Êxito do Estudante do Instituto Federal de Santa Catarina (PPE-IFSC) foi aprovado em agosto de 2018 e tem os seguintes objetivos:

> Geral: Promover a permanência e êxito dos estudantes em todos os níveis e modalidades de ensino ofertados no IFSC, por meio de um conjunto de estratégias e ações que visam o enfrentamento da evasão e retenção.
> Específicos: Analisar a problemática da evasão e retenção de estudantes no IFSC; Mobilizar os câmpus para a discussão e enfrentamento das causas e consequências da evasão e retenção; Implantar estratégias de intervenção para enfrentamento dos fatores mais recorrentes de evasão e retenção; Monitorar e avaliar as ações em andamento ou a serem desenvolvidas; Levantar subsídios para o aprimoramento dos processos de ingresso e acesso dos estudantes;

> Promover a formação continuada de servidores com foco na permanência e êxito dos estudantes. (IFSC, 2018, p. 40-45)

Na Seção 5 deste plano são apresentadas as causas da evasão e retenção dos estudantes no Instituto Federal de Santa Catarina, fornecidas no entendimento da comunidade acadêmica. O Quadro 1 apresenta os fatores de evasão fornecidos pelo PPE institucional que foram divididos em três dimensões: externas à instituição, individual do estudante e interna à instituição. Eles foram divididos também em fatores gerais e específicos, cujos fatores gerais foram expostos em ordem decrescente, do maior para o menor número de fatores específicos levantados. Os autores também explicam que alguns fatores específicos estão ligados a apenas uma dimensão, porém podem apresentar algumas limitações no enquadramento e significado de algumas causas de evasão/retenção, principalmente as ligadas à dimensão "Individual do Estudante" (IFSC, 2018). Vale lembrar que o quadro apresentado no plano (IFSC, 2018) foi adaptado apenas para os cursos de graduação, portanto, fatores específicos que continham na tabela e não eram considerados para os cursos de graduação foram excluídos.

Quadro 1 – Fatores de evasão e retenção dos estudantes no PPE-IFSC

Dimensão	Fator Geral	Fator Específico
Externo à Instituição	Conjuntura social, econômica e política	Redução do investimento na rede federal e perda de orçamento;
		Aumenta a vulnerabilidade socioeconômica do estudante;
		Distância residência-campus e dificuldades de transporte para o deslocamento;
		Fragilidade das políticas para a educação profissional e tecnológica;
		Falta ou custo elevado da moradia.
	Valorização da profissão	Desvalorização social da profissão;
		Falta de perspectiva profissional em relação à empregabilidade;
		Baixa remuneração do profissional formado;
		Dificuldade e/ou impossibilidade de registro nos conselhos profissionais;
		Falta de reconhecimento dos cursos superiores de tecnologia.
	Antecedência escolar	Dificuldades no uso de novas tecnologias;
		Déficit na formação pregressa do estudante;
		Baixa qualidade do ensino fundamental público.
Individual do Estudante	Adaptação à vida acadêmica	Dificuldade de adaptação do estudante à rotina escolar;
		Dificuldade de adaptação do estudante à metodologia do curso;
		Falta de assiduidade e/ou pontualidade;
		Indisponibilidade de tempo para estudar fora do horário de aula e/ou participar de atividades de monitoria/nivelamento;
		Dificuldades de relacionamento com outros estudantes;
		Problemas pessoais;
		Medo de reprovação ou de repetir o período;
		Desconhecimento do perfil do curso;

Dimensão	Fator Geral	Fator Específico
		Problemas disciplinares.
	Motivação em relação ao curso	Falta de identificação ou desinteresse pelo curso;
		Desmotivação para estudar ou concluir a formação;
		Mudança de interesse pessoal ou profissional;
		Falta de maturidade para escolha da profissão;
		Ingresso em outro curso.
	Habilidade de estudo	Dificuldades de aprendizagem;
		Falta de hábito ou disciplina de estudo;
		Muito tempo afastado do sistema formal de ensino;
		Falta de conhecimento sobre a área do curso escolhido.
	Situação familiar	Problemas familiares;
		Problemas de saúde pessoal ou na família;
		Precisa trabalhar para se sustentar ou sustentar a família;
		Necessidade de cuidar do(s) filho(s) no horário do curso;
		Falta de apoio da família.
	Relação estudo-trabalho	Local de trabalho que não flexibiliza a carga horária do trabalhador estudante;
		Dificuldade de conciliar estudo e trabalho;
		Dificuldades para realizar atividades extraclasse em função do trabalho;
	Personalidade	Baixa autoestima;
		Falta de aptidão para o curso escolhido;
		Falta de maturidade para encarar o curso.
Interno à Instituição	Aspecto didático-pedagógicos	Inadequação do projeto pedagógico do curso;
		Inadequação da metodologia de ensino ao perfil dos estudantes;
		Exigência de pré-requisitos para cursar atividades de recuperação paralela;
		Falta de atualização e de flexibilidade curricular;
		Mudanças curriculares ao longo da oferta do curso;
		Deficiência na formação pedagógica dos docentes;
		Falta de visitas técnicas e aulas práticas;
		Desrespeito a inclusão social e a diversidade;
		Descontextualização ou desatualização dos cursos com a realidade local/regional;
		Dificuldades na relação docente-estudante.
	Gestão acadêmica do curso	Turnos e horários de oferta incompatíveis com a demanda;
		Alterações no horário de aula por motivos diversos;
		Falta de servidores para o suporte ao trabalho docente;
		Dificuldade de realização de aulas práticas no período noturno;
		Rotatividade de docentes em algumas disciplinas;
		Falta de informação e orientação sobre processos acadêmicos (validação, cancelamento, trancamento etc.).
	Programas institucionais para o estudante	Insuficiência de recursos para os programas de assistência estudantil;
		Redução dos valores do auxílio financeiro (PAEVS);

Dimensão	Fator Geral	Fator Específico
		Falta de alimentação escolar;
		Redução dos programas de fomento à pesquisa;
		Demora no recebimento inicial do auxílio financeiro (PAEVS).
	Infraestrutura	Falta de equipamentos/insumos nos laboratórios;
		Falta de docentes em algumas áreas por demora no processo de contratação;
		Falta de reformas na infraestrutura física;
		Dificuldade de acesso devido à localização do câmpus;
		Falta de infraestrutura para atender as necessidades da permanência do estudante de período integral.
	Divulgação e ingresso	Chamadas de matrícula avançando no semestre letivo;
		O processo seletivo não contempla as especificidades em termos de curso e público.

Fonte: Adaptado de IFSC (2018)

3.2 TÉCNICAS DE COLETA DE DADOS

A técnica de coleta de dados utilizada na pesquisa é de fontes secundárias, pois a veracidade e a quantidade de informações que os bancos de dados institucionais possuem são de grande valia. As informações foram coletadas na forma de registro em arquivos, e os dados serão analisados a partir dessa coleta no Sistema Integrado de Gestão Acadêmica (SIGAA), bem como no Sistema de Ingresso do IFSC, diretamente do banco de dados institucional. Assim, a instituição encaminhou dois arquivos no formato ".csv" com os dados necessários para a pesquisa.

Quando necessário algum dado (variável) que não constava no banco de dados institucional, foi retirado da pesquisa, pois por causa da Lei Geral de Proteção de Dados Pessoais (LGPD), a instituição não permitiu a aplicação de questionários com os estudantes. As matrículas dos alunos, recebidas a fim de identificar as informações (linhas de dados) de um mesmo estudante, foram transformadas em números sequenciais após o tratamento, com isso, a base de dados antes do tratamento das informações não pode ser repassada de forma alguma, somente os dados tratados usados para a criação do modelo de previsão, os quais não possuem nenhuma informação que identifique o estudante.

3.3 VARIÁVEIS OU CATEGORIAS DE ANÁLISE

A análise e interpretação dos dados se deu por meio de Estatística Multivariada obtida através dos algoritmos de *Machine Learning*, esses algoritmos se utilizam de técnicas

estatísticas para realizar as previsões de riscos se algo vir a acontecer. Para isso, foram definidas variáveis importantes para as previsões de acordo com os estudos efetuados na Seção 2.2 deste documento.

Dados que precisam ser retirados de ambientes virtuais de aprendizado (*Moodle*) não foram coletados/disponibilizados pela instituição, assim, foram excluídas as seguintes variáveis: "Tempo total de *login* do estudante em plataforma de aprendizado", "Número de recursos visualizados em plataforma", "Número de tentativas de testes enviados em plataforma", "Número de fóruns visualizados em plataforma", "Número de discussões em fóruns lidas ou visualizadas em plataforma", "Desempenho do aluno nas atividades semanais", "Atendimentos de um aluno", "Aluno retomou um tópico", "Durante o semestre, o aluno deve ter três questionários em um determinado curso", "Aluno respondeu a todos os questionários", "Média das notas obtidas nos questionários", "Notas obtidas no exame do meio do semestre", "Aluno encaminhou as tarefas", "Aluno executou as apresentações" e "Uso de ferramentas educacionais fornecidas pelo sistema (blog, glossário, quiz, wiki, mensagem)". Outros dados que não foram possíveis de coleta pelo fato de estarem em outro banco de dados da instituição (Ingresso), o qual não foi possível ligar o então aluno ao candidato: "Histórico do ensino médio – média final" e "Nota/Qualificação ingresso". Variáveis como "Tipo de Moradia Durante o Curso", "Trabalho/Horas de Trabalho" e "Curso de Primeira Opção" não existem em banco de dados.

3.3.1 Seleção de Atributos (*Feature Selection*)

Grus (2016) descreve que características podem ser consideradas quaisquer entradas fornecidas ao modelo. O autor especifica que, dependendo do tipo de característica utilizado para o modelo, restinge-se o tipo de algoritmo a ser utilizado. Por exemplo, quando se pretende manipular características "Sim-ou-Não", é possível utilizar *Naive Bayes*, algoritmos de regressão requerem características numéricas, já outros algoritmos, como árvores de decisão, podem utilizar características numéricas e categóricas (GRUS, 2016).

Outra questão importante a ser analisada é a dimensionalidade. Para Harrison (2020), com o aumento de dimensões dos dados, eles se tornam mais esparsos e, com isso, podem dificultar a obtenção do resultado. Assim, quando adicionadas mais dimensões, algoritmos que utilizam cálculo de vizinhança (por exemplo, KNN) perdem a sua utilidade. Grus (2016) afirma que, dependendo da situação, pode ser melhor diminuir apenas para dimensões importantes, empregando um pequeno número de características. Segundo Grus (2016), é possível que um

ocorra *underfitting* no modelo caso não possua características suficientes, já quando o modelo tiver muitas características, é possível ocorrer o *overfitting*.

Géron (2019) descreve que é necessário que haja uma quantidade de dados grande para que a maioria dos algoritmos de ML funcione bem, mesmo para problemas simples. Também é importante selecionar o máximo de características relevantes e poucas irrelevantes, pois atributos irrelevantes podem trazer um efeito negativo no modelo (GÉRON, 2019; HARRISON, 2020).

A criação do conjunto de características para o treinamento, chamado por Géron (2019) de *feature engineering*, envolve a seleção das características (selecionar as características mais relevantes entre as existentes), extração das características (combinar características para desenvolver uma mais relevante) e criação de novas características (na coleta de novos dados). A seguir, será realizada a descrição das variáveis (características) definidas para esta análise.

Idade (idade_disc): idade do estudante, entre 17 e 82 anos. Com isso, foram transformadas em faixas etárias adaptadas da Plataforma Nilo Peçanha (Tabela 13) (BRASIL, 2021).

Tabela 13 – Descrição da variável idade_disc

Variável	Descrição
idade_disc	Até 19 anos;
	De 20 a 24 anos;
	De 25 a 29 anos;
	De 30 a 34 anos;
	De 35 a 39 anos;
	De 40 a 44 anos;
	Acima de 45 anos.

Fonte: Elaborada pela autora desta dissertação (2022)

Renda *per capita* familiar (renda_pcf): variável que representa a renda familiar obtida por meio da divisão da renda da família pelo número de membros da família. Com isso, foram transformadas em intervalos seguindo a Plataforma Nilo Peçanha (Tabela 14) (BRASIL, 2021).

Tabela 14 – Descrição da variável renda_pcf

Variável	Descrição
renda_pfc	Intervalo de 0 até 0,5 salários mínimos;
	Acima de 0,5 até 1 salários mínimos;
	Acima de 1 até 1,5 salários mínimos;

Variável	Descrição
	Acima de 1,5 até 2,5 salários mínimos;
	Acima de 2,5 até 3,5 salários mínimos;
	Acima de 3,5.

Fonte: Elaborada pela autora desta dissertação (2022)

Raça (raca_disc): representa a classificação racial dos estudantes (Tabela 15).

Tabela 15 – Descrição da variável raca_disc

Variável	Descrição
raca_disc	Branco (a);
	Pardo (a);
	Preto (a);
	Amarela (a);
	Indígena.

Fonte: Elaborada pela autora desta dissertação (2022)

Sexo (sexo_disc): esta variável representa o sexo do discente: Feminino; Masculino.

Campus (campus): esta variável representa os *Campi* da instituição, que possui 23 unidades com cursos de graduação: Araranguá, Canoinhas, Caçador, Chapecó, Criciúma, Florianópolis, Florianópolis Continente, Garopaba, Gaspar, Itajaí, Jaraguá do Sul Centro, Jaraguá do Sul Rau, Joinville, Lages, Palhoça Bilíngue, São Carlos, São José, São Lourenço do Oeste, São Miguel do Oeste, Tubarão, Urupema, Xanxerê e Centro de Referência EaD.

Turno curso/aulas (turno_curso): turno do curso frequentado pelo estudante (Tabela 16).

Tabela 16 – Descrição da variável turno_curso

Variável	Descrição
turno_curso	Matutino;
	Vespertino;
	Noturno;
	Integral.

Fonte: Elaborada pela autora desta dissertação (2022)

Tipo de Aprendizagem (tipo_aprendizagem): tipo de aprendizagem do curso estudado pelo discente: Presencial; EaD.

Curso (curso): nome dos 44 cursos únicos da instituição: Engenharia Civil; Engenharia Elétrica; Engenharia Mecatrônica; Engenharia Eletrônica; Engenharia de Controle e Automação; Engenharia Civil; Engenharia Mecânica; Engenharia de Telecomunicações; Engenharia Química; Engenharia de Alimentos; Engenharia de Produção; Mecânica Industrial; Fabricação Mecânica; Processos Químicos; Design de Produto; Construção de Edifícios; Gestão da Tecnologia da Informação; Ciência da Computação; Sistemas de Informação; Sistemas de Telecomunicações; Análise e Desenvolvimento de Sistemas; Tecnologia em Sistemas para Internet; Matemática; Química; Física; Ciências da Natureza com Habilitação em Física; Educação Profissional e Tecnológica; Gestão Pública; Processos Gerenciais; Gestão de Turismo; Gestão Hospitalar; Gestão de Agronegócio; Hotelaria; Gestão Ambiental; Alimentos; Viticultura e Enologia; Gastronomia; Produção Multimídia; Design; Design de Moda; Agronomia; Radiologia; Sistemas de Energia; Enfermagem; Pedagogia Bilíngue (Libras-Português).

Naturalidade do Discente (naturalidade_disc): cidade de origem do estudante, foram encontradas 949 cidades distintas.

Cidade do Campus (cidade_campus): representa a cidade em que o campus está localizado: Araranguá; Canoinhas; Caçador; Chapecó; Criciúma; Florianópolis; Garopaba; Gaspar; Itajaí; Jaraguá do Sul; Joinville; Lages; Palhoça; São Carlos; São José; São Lourenço do Oeste; São Miguel do Oeste; Tubarão; Urupema; Xanxerê.

Tipo de ingresso (forma_ingresso): forma como o estudante ingressou na instituição:

Tabela 17 – Descrição da variável forma_ingresso

Variável	Descrição
forma_ingresso	Ingresso com Prova; Retorno de Egresso; Transferência Externa; Transferência Interna; Transferência Ex-Officio; SiSU (ENEM); Ingresso sem prova; Vaga remanescente.

Fonte: Elaborada pela autora desta dissertação (2022)

Origem do ensino médio/anterior (origem_ensino_anterior): tipo de instituição anterior do estudante (Tabela 18).

Tabela 18 – Descrição da variável origem_ensino_anterior

Variável	Descrição
origem_ensino_anterior	Federal; Estadual; Municipal; Privada.

Fonte: Elaborada pela autora desta dissertação (2022)

Nota média de disciplinas do estudante (media_geral_disc): esta variável representa a média das notas de disciplinas que o estudante já concluiu (aprovado ou reprovado) na graduação: Tipo numérico de 0 a 10.

Reprovações (n_disciplinas_reprovadas): número de disciplinas reprovadas pelo estudante: 0 a 29.

Disciplinas concluidas (n_disciplinas_concluidas): soma das disciplinas concluidas pelo estudante: 0 a 88.

Índice de desenvolvimento humano do município (IDHM) (idhm_minicipio): último índice de desenvolvimento humano composto disponibilizado pelo município em que o curso está situado: Florianópolis – 0.847; Jaraguá do Sul – 0.803; Joinville – 0.809; Lages – 0.770; Criciúma – 0.788; Gaspar – 0.765; Palhoça – 0.757; Canoinhas – 0.757; São Miguel do Oeste – 0.801; Caçador – 0.735; São José – 0.809; Itajaí – 0.795; Araranguá – 0.760; Tubarão – 0.796; Chapecó – 0.790; Xanxerê – 0.775; São Carlos – 0.769; Garopaba – 0.753; Urupema – 0.699.

Ano/Semestre ingresso (ingresso_disc): ano e semestre em que o estudante ingressou na graduação (Tabela 19).

Tabela 19 – Descrição da variável ingresso_disc

Variável	Descrição
ingresso_disc	01/2017; 02/2017; 01/2018; 02/2018; 01/2019; 02/2019; 01/2020; 02/2020; 01/2021;

Variável	Descrição
	02/2021.

Fonte: Elaborada pela autora desta dissertação (2022)

Semestre atual (semestre_atual): semestre atual em que o estudante se encontra, não necessariamente é o semestre do curso e sim o tempo em que o estudante se encontra na instituição: 2 a 11.

Estado civil (estado_civil): variável que especifica o estado civil do estudante (Tabela 20).

Tabela 20 – Descrição da variável estado_civil

Variável	Descrição
estado_civil	Solteiro (a);
	Casado (a);
	Separado (a) Consensualmente;
	Separado (a) Judicialmente;
	Divorciado (a);
	Desquitado (a);
	Viúvo (a).

Fonte: Elaborada pela autora desta dissertação (2022)

***Status* atual do discente (status_atual_disc):** é a variável de saída e representa o *status* atual do discente: Ativo; Curso concluído.

3.4 TIPOS DE DADOS

Os dados usados nesta pesquisa são de fontes secundárias. Usando os dados do Instituto Federal de Santa Catarina (IFSC) do Sistema Integrado de Gestão Acadêmica (SIGAA) e do Sistema de Ingresso da instituição. Sendo que a base de dados está delimitada por todos os cursos de graduação (bacharelado, licenciatura e tecnologia) da instituição com ingresso nos anos de 2017, 2018, 2019, 2020 e 2021.

No ano de 2017, foram 2.953 ingressantes em 61 cursos de graduação em 22 unidades, nas modalidades: presencial e a distância. Em 2018, foram 3.140 ingressantes em 70 cursos de graduação das 22 unidades, nas modalidades presencial e a distância. No ano de 2019, foram 3.074 ingressantes em 70 cursos de graduação de 22 unidades, nas modalidades: presencial e a

distância. Em 2020, foram 2.999 ingressantes em 71 cursos de graduação das 22 unidades, nas modalidades presencial e a distância (BRASIL, 2021). O ano de 2021 ainda não foi disponibilizado na Plataforma Nilo Peçanha, da qual foram extraídas as informações (BRASIL, 2021).

Para a criação do modelo de previsão, foram utilizadas duas coortes: antes da pandemia da Covid-19 e durante a pandemia da Covid-19. A coorte antes da pandemia com os dados dos anos de 2017, 2018 e 2019 foi utilizada apenas esses anos com o intuito de não ter interferência da pandemia da Covid-19 (df_antes_pandemia). A coorte durante pandemia com os dados dos anos de 2020 e 2021 foi utilizada para fins de análise e para verificar as diferenças ocorridas durante a pandemia da Covid-19 (df_durante_pandemia).

3.5 METODOLOGIA ADOTADA PARA DESENVOLVIMENTO DO MODELO

A pesquisa valeu-se da metodologia *Cross-Industry Standard Process for Data Mining (CRISP-DM)*, e seu processo está definido na Seção 2.4 (Figura 5). A seguir, é fornecido o detalhamento de cada atividade definida no processo e de como elas foram usadas no presente trabalho.

Fazer uma pergunta: Qual modelo utilizando algoritmos de *Machine Learning* explicam a evasão escolar no IFSC?

a) Obter visão geral do problema – fatores da evasão no IFSC - (2.1.4);
b) Definir variáveis a serem utilizadas a partir da literatura existente (2.2);
c) Avaliar os algoritmos de *Machine Learning Decision Tree*, *Artificial Neural Network* (*Multilayer Perceptron*) e *XGBoost* (2.4).

Coleta dos dados: com as variáveis definidas, coletar os dados da base de dados institucional a partir dos sistemas acadêmicos (SIGAA e Sistema de Ingresso).

a) Foi recebido dois arquivos .csv pela instituição com os dados (*dados_dissertacao.csv* e *dissertacao_formandos.csv*);
b) As colunas incluídas neste conjunto de dados continham: identificação discente, gênero, etnia, data de nascimento, renda familiar, quantidade de membros na família, estado civil, município de naturalidade do discente, endereço do estudante (logradouro, descrição rua, bairro, município, estado, CEP), campus, endereço campus (logradouro, descrição rua, número, complemento, bairro, município, estado, CEP), curso, turno do curso, forma de ingresso, ano ingresso, período ingresso, disciplina, ano da disciplina, período da disciplina, média da

disciplina, *status* matrícula, tipo de instituição do ensino anterior, status atual do discente.

Limpeza dos dados: garantir que os dados estejam no formato em que os algoritmos possam interpretar. Para isso, foram realizados os passos definidos a seguir:

a) *Importação dos dados:* importar os arquivos recebidos para o *Python* (na ferramenta *Colab* do *Google*). Com isso, a biblioteca do *Python* chamada *Pandas* é capaz de ler as planilhas (neste caso, *.csv*) e convertê-las em um *DataFrame* (estruturas de dados bidimensionais alinhados de forma tabular em linhas e colunas) (HARRISON, 2020);

b) *Exclusão de valores duplicados:* com a ajuda da biblioteca *Pandas* do *Python*, averiguar e excluir linhas duplicadas no *DataFrame*;

c) *Limpar valores ausentes:* boa parte dos algoritmos de ML falham ao receber valores ausentes (NaN) (HARRISON, 2020). Essa limpeza pode se dar por exclusão dos dados ausentes ou imputação (usar técnicas para substituir os valores NaN). Na pesquisa em questão, optou-se por usar as técnicas de imputação sempre que possível para permanecer uma maior quantidade de informações;

d) *Criação de variáveis:* a partir das variáveis dos arquivos recebidas, criar as variáveis necessárias/definidas para gerar modelo da instituição;

e) *Análise descritiva das variáveis:* esta análise identifica *outliers*, dados muito diferentes dos demais que podem causar anomalias nos resultados dos algoritmos e, assim, tratar esses dados. Esse tratamento foi realizado nas variáveis renda_pcf, n_disciplinas_reprovadas e n_disciplinas_concluidas;

f) *Transformar dados categóricos em dados numéricos:* a maioria dos algoritmos de *Machine Learning* necessita de que os dados estejam do tipo numérico (*int* ou *float*), para isso, usar métodos que façam essa transformação. A Tabela 13 informa os dados que eram inicialmente categóricos e receberam o tratamento.

Normalização dos dados: os algoritmos são sensíveis quando os valores numéricos de entrada têm escalas muito diferentes (GÉRON, 2019). Para resolver isso, foi utilizado o método *MinMaxScaler* (normalização) da biblioteca *sckit-learn* do *Python*, o qual redimensiona os valores para que fiquem no intervalo entre 0 e 1 (SKLEARN, 2022). O método subtrai o valor mínimo e divide pelo valor máximo menos o valor mínimo.

Gerar dados da amostra: após as etapas anteriores, os dados estão prontos para treinamento do modelo, com isso, é possível dividi-los em treinamento e teste. Das pesquisas

analisadas, a divisão dos dados considera de 70 a 80% dos dados para treinamento e de 20 a 30% para teste.

a) Foi empregado 80% dos dados para treino e 20% para teste.

Criar o modelo: utilizar os algoritmos avaliados na Seção 2.4 para criar o modelo com os dados de treinamento.

a) O algoritmo *DecisionTreeClassifier* (algoritmo de classificação de árvores de decisão) foi utilizado como *baseline*: o algoritmo *baseline* serve como comparativo para avaliar a qualidade dos outros algoritmos. O *DecisionTreeClassifier* foi escolhido por trabalhar com métodos de classificação básicos e ser amplamente utilizado;
b) Os algoritmos *XGBClassifier* (algoritmo de classificação de *XGBoost*) e o algoritmo *MLPClassifier* (algoritmo de classificação de *Multilayer Perceptron*) foram usados para a criação do modelo. O algoritmo *XGBoost* vem sendo utilizado por muito e apresentando bons resultados, como é o caso de Gismondi e Huiman (2021), que obtiveram a melhor precisão com o algoritmo. O algoritmo MLP apresentou os melhores resultados na pesquisa de Mduma, Kalegele e Machuve (2019), já que os autores avaliaram quatro classificadores supervisionados, e o *MultiLayer Perceptron* mostrou melhor desempenho quando melhorado seus hiperparâmetros.

Avaliar o modelo com os dados de teste: esta etapa envolve avaliar o modelo criado com os dados de teste e utilizar as métricas de avaliação da classificação, as métricas são utilizadas para verificar se os modelos de ML são bons, traz informações do desempenho do modelo (GÉRON, 2019). As métricas utilizadas para avaliar o modelo foram:

a) *Matriz de Confusão:* é usada em modelos de classificação, ela é uma tabela que indica os erros e os acertos do modelo, comparando com os resultados esperados. Segundo Harrison (2020), o classificador binário (saída zero ou um) pode ter quatro resultados de classificação: *True Positives* (TP), *True Negatives* (TN), *False Positives* (FP) e *False Negative* (FN). Mia *et al.* (2019) definem TP como sendo um classificador que prevê o *status* de registro positivo de um estudante, e a saída real indica o *status* de registro positivo do estudante, TN como o classificador prevê o *status* de registro do estudante como negativo, e a saída real indica o *status* do registro negativo do estudante, FP como um classificador prevê o *status* de registro positivo de um estudante, porém a saída real indica o *status* do registro negativo do estudante e, por fim, FN como o classificador prevê o *status* de registro de um

estudante como negativo, mas a saída real indica o *status* de registro positivo do estudante. Com isso, percebe-se que os resultados corretos são TP e TN. O FP é chamado de Erro Tipo 1, e o FN é chamado de Erro Tipo 2. A partir das definições apresentadas, surge a Figura 15. Do mesmo modo que Mia *et al.* (2019), Freitas *et al.* (2020) também se aproveitaram da Matriz de Confusão para avaliar seus modelos.

Figura 15 – Classificador binário da matriz de confusão
Fonte: Elaborada pela autora desta dissertação (2022)

b) *Accuracy (Acurácia):* é a métrica mais simples e amplamente usada. A acurácia mede a média geral do acerto do modelo ao classificar as classes. Ou seja, ela é o percentual das predições corretas (HARRISON, 2020). Dessa forma, *Accuracy = (TP + TN) / (TP + TN + FP + FN)*. Na implementação do modelo, foi usada a função *accuracy_score* do *sklearn.metrics*. Essa métrica foi utilizada por Silva, Almeida e Ramalho (2020), Zulfiker *et al.* (2020) e Freitas *et al.* (2020).

c) *Precision (Precisão):* traz a informação de quantas observações o modelo classificou corretamente como 1. Ou seja, é o percentual de predições positivas que estavam corretas (HARRISON, 2020). Dessa forma, *Precision = TP / (TP + FP)*. Na implementação do modelo, foi usada a função *precision_score* do *sklearn.metrics*. *Precision* foi a métrica mais utilizada entre os autores avaliados, entre eles, estão: Hussain *et al.* (2018), Beaulac e Rosenthal (2019), Adekitan e

Salau (2019), Zulfiker *et al.* (2020), Iatrellis *et al.* (2020), Adejo e Connolly (2018), Freitas *et al.* (2020), Chui *et al.* (2020) e Xiao e Yi (2020).

d) *Recall (Revocação):* também chamado de *sensitivity* (sensibilidade). Essa métrica fornece a taxa de detecção (HARRISION, 2020). Ou seja, o percentual de valores positivos classificados corretamente. Assim, *Recall = TP / (TP + FN)*. Na implementação do modelo, foi usada a função *recall_score* do *sklearn.metrics*. Essa métrica também foi bastante aplicada entre os autores avaliados: Hussain *et al.* (2018), Zulfiker *et al.* (2020), Iatrellis *et al.* (2020), Adejo e Connolly (2018), Freitas *et al.* (2020), Delen (2010), Deo *et al.* (2020) e Silva, Almeida e Ramalho (2020).

e) *F1 – Score:* média harmônica entre *recall* e *precision*. Dessa maneira, *F1 = (2 * Precision * Recall) / (Precision + Recall)*. Na implementação do modelo, foi usada a função *f1_score* do *sklearn.metrics*. Costa *et al* (2017), Hung *et al* (2020), Zulfiker *et al.* (2020), Adejo e Connolly (2018) e Freitas *et al.* (2020) empregaram essa métrica como avaliação de modelos.

Após a avaliação do modelo a partir das métricas, foi selecionado o algoritmo com os principais resultados para desenvolver um modelo melhorado, ao mesmo tempo, foi realizada a exclusão das três piores variáveis. Após, foram feitas novas avaliação com as métricas e avaliados os novos resultados (Capítulo 4).

Implantar o modelo: após os testes e avaliações do modelo, é possível realizar a implantação na instituição (esta fase entra como trabalhos futuros da pesquisa).

Nas próximas seções será apresentado o detalhamento do pré-processamento dos dados (limpeza, Imputação e Normalização dos Dados).

3.5.1 Limpeza, Imputação e Normalização dos Dados

Os dados crus extraídos do banco de dados da instituição (IFSC) contêm 160.271 registros e 36 variáveis que foram recebidos da instituição em dois arquivos: *dados_dissertacao.csv* e *dissertacao_formados.csv*. A base de dados *dados_dissertacao.csv* foi a primeira recebida, porém, ao analisá-la, percebeu-se que os dados de estudantes que entraram na graduação nos anos especificados e tinham concluído seu curso não estavam presentes. Com isso, a instituição encaminhou uma nova base de dados *dissertacao_formados.csv* com os dados faltantes para a pesquisa (discentes com *status* formado). Dessa forma, foram importadas as

duas bases de dados e usada a função *'concat'* para juntar as duas bases em um único *DataFrame*.

A Tabela 21 representa os dados coletados e, à direita, o percentual de valores vazios. Nesses dados, um discente pode ter um ou mais registros, a depender da quantidade de disciplinas cursadas, em que cada registro vai representar um componente curricular. Foram excluídas oito variáveis que representavam o endereço do estudante, pois elas possuíam mais de 60% dos dados faltantes (*Logradouro* – 61,14%, *Rua* – 61,14%, *Número* – 61,14%, *Complemento* – 71,41%, *Bairro* – 61,14%, *Município* – 61,14%, *Estado* – 61,14%, *CEP* – 61,14%). Com esse percentual alto de dados faltantes, a imputação e a criação da variável *distancia_residencia_campus*, que representam a distância da casa do estudante até o campus (FREITAS *et al.,* 2020; SILVA; CABRAL; PACHECO, 2020), poderiam não representar a realidade. Da mesma forma, sete variáveis que representavam o endereço do campus foram excluídas (*Logradouro*, *Rua*, *Número*, *Complemento*, *Bairro*, *Estado*, *CEP*). Após as exclusões das variáveis que não seriam mais utilizadas, foi modificado o nome das variáveis restantes no *DataFrame*.

Tabela 21 – Dados do BD e percentual faltante (NaN)

Variável	Percentual Faltante
id_discente	0%
Sexo	0%
Raca	0,03%
data_nasc	0,22%
renda_familiar	8,25%
membros_familia	8,25%
estado_civil	24,38%
naturalidade_disc	1,44%
Campus	0%
municipio_campus	0%
Curso	0%
turno_curso	0%
forma_ingresso	0,31%
ano_ingresso	0%
periodo_ingresso	0%
componente_curricular	0%
ano_componente_curricular	0,33%

Variável	Percentual Faltante
periodo_componente_curricular	0,33%
media_componente	5,90%
status_componente_disc	0%
origem_ensino_anterior	4,70%
status_atual_disc	0%

Fonte: Elaborada pela autora desta dissertação (2022)

3.5.1.1 Limpeza e Imputação dos Dados

A seguir são apresentadas as etapas de exclusão de valores duplicados e a criação e tratamento de variáveis que fazem parte da fase de limpeza e imputação dos dados.

Exclusão de valores duplicados: nesta etapa, primeiro verificou-se a existência de linhas iguais por meio da função *df[df.duplicated()]*, e após constatar que não possuía nenhuma linha duplicada, analisou-se se existia alguma linha em que as variáveis *id_discente*, *curso*, *ano_ingresso*, *periodo_ingresso*, *componente_curricular*, *ano_componente_curricular*, *periodo_componente_curricular*, *media_componente* e *status_componente_disc* eram iguais, assim, foram encontradas 27.740 duplicações, o que equivale a 17,31% do *DataFrame*, os quais foram removidos, utilizando o critério de manter o primeiro dado apresentado, permaneceram 132.531 registros em 22 variáveis.

Criação e tratamento de variáveis: a variável *forma_ingresso* apresentou 12 tipos (SiSU-ENEM, Ingresso sem prova, Ingresso com prova, Vaga remanescente, Transferência externa, Transferência interna, Retorno de egresso, Acordo de cooperação técnica IFSC/INSS, Transferência ex-officio, Aluno especial, Certific e Intercambista). Foram excluídos os estudantes do tipo Intercambista, Certific, Aluno especial e Acordo de cooperação técnica IFSC/INSS, pois não se trata de alunos regulares de graduação da instituição. Com isso, permaneceram 131.968 registros no *DataFrame*. A variável apresentou 374 dados faltantes, para o tratamento destes, foi realizada a imputação com a função *SimpleImputer*, usando a estratégia constante de substituir os dados NaN (*Not a Number*) por SiSU-ENEM, pois esse tipo é o que representa a maior parte da instituição. Após isso, a variável categórica foi transformada em números de zero a sete (0 a 7), e, então, foi agrupado pelo *id_discente* e criado o *DataFrame grupo_forma_ingresso['forma_ingresso']*.

Para criar a variável *media_geral_disc*, foi usada a variável *media_componente*, que representa a nota do estudante em uma determinada disciplina. A *media_componente* possuía 8.193 dados faltantes, para isso, na imputação, foi utilizado o algoritmo *K-Nearest Neighbor* –

KNN (*KNNImputer*), em que foi definido o K=3, pois nos testes foi o que trouxe todas as imputações preenchidas (com 1 e 2 as imputações não foram feitas em todos os NaNs). Após, foi agrupado pelo *id_discente*, calculada a média de cada aluno e criado o *DataFrame grupo_discente['media_geral_disc']*.

Da variável *media_geral_disc*, é possível analisar os valores na Tabela 22, extraída do algoritmo criado para o pré-processamento, em que *count* representa a contagem de discentes únicos, *mean* a média geral de todos os estudantes, *std* o desvio-padrão das médias, *min* a média mínima registrada, os quartis 25%, 50% e 75%, e, por fim, *max* que representa a média máxima registrada.

Tabela 22 – Dados da variável media_geral_disc

	media_geral_disc
Contagem	7759
Média	6,9398
Desvio-Padrão	2,0987
Mínimo	0
25%	6
50%	8
75%	8
Máximo	10

Fonte: Elaborada pela autora desta dissertação (2022)

A variável *idade_disc* foi criada a partir da variável que apresenta a data de nascimento do estudante *data_nasc.* Essa *Data_nasc* possuia 330 dados faltantes, por isso, foi utilizado o método de imputação *SimpleImputer*, usando a estratégia "*most_frequent*" para substituir os dados NaN (*Not a Number*) pela data de nascimento mais frequente. Logo após, com a *data_nasc*, foi criada a variável idade do discente *idade_disc.* A idade do discente foi transformada em atributos categóricos de faixas etárias, as quais foram adaptadas da Plataforma Nilo Peçanha (BRASIL, 2021). Com a idade em faixas etárias, foi feita a normalização dos dados com a função *LabelEncoder* que a transformou em dados numéricos entre zero e seis (0 e 6). Depois disso, foi agrupado pelo *id_discente* e criado o *DataFrame grupo_idade_disc['idade_disc']*.

A variável *renda_pcf* foi gerada a partir das variáveis *renda_familiar* (valor da renda total da família) e *membros_familia* (que apresentam o número de membros presentes na família). As duas variáveis utilizadas possuíam 7.951 dados faltantes cada uma, por isso, foi

utilizado o método de imputação *SimpleImputer*, usando a estratégia "*mean*" para substituir os dados NaN (*Not a Number*) pela média de renda familiar e membros da família, que foram arredondadas e depois foi calculada a renda *per capita* familiar, dividindo a renda familiar pelos membros da família.

Com a renda *per capita* familiar gerada, foi analisado o intervalo que trouxe de R$00,00 até R$1.400.00,00. A partir disso, foi feita a análise descritiva da variável, a fim de encontrar e tratar *outliers* (valores atípicos). Analisada a metodologia *box-plot* em que são calculados os quartis, interquartis e limites inferiores e superiores, chegou-se à conclusão de que a metodologia não fez muito sentido, pois excluiria muitas informações, pois a renda seria até R$ 3.250,00. Com isso, foram sendo testados limites superiores (manuais) gerando o *box-plot* até R$ 12.500, um gráfico aceitável e, assim, alterado o valor dos *outliers* pela mediana, pois ela é menos sensível aos *outliers* do que a média. A Figura 16 (gráfico gerado no algoritmo e criado na linguagem python) fornece a distribuição da renda *per capita* antes e depois da alteração de dados discrepantes.

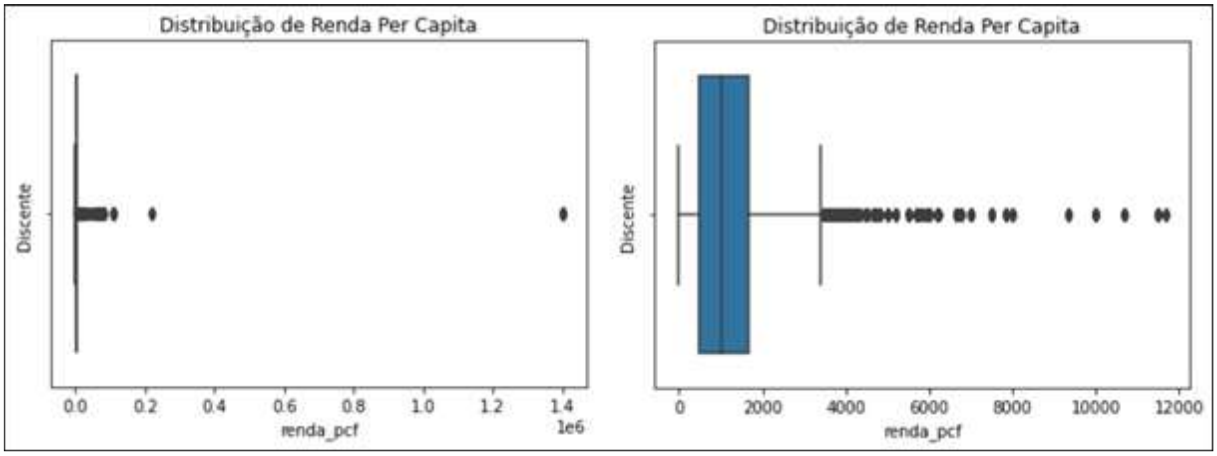


Figura 16 – Distribuição de renda *per capita* antes/após tratar *outliers*
Fonte: Elaborada pela autora desta dissertação (2022)

Após o tratamento de *outliers* foi feito o arredondamento da variável *renda_pcf* e transformada em intervalos seguindo a PNP (BRASIL, 2021) e, a seguir, transformada em dados numéricos. Com isso, foram criados seis intervalos (0 a 5) como segue: Intervalo de 0 até 0,5 salários mínimos (0); Acima de 0,5 até 1 salários mínimos (1); Acima de 1 até 1,5 salários mínimos (2); Acima de 1,5 até 2,5 salários mínimos (3); Acima de 2,5 até 3,5 salários mínimos (4); Acima de 3,5 salários (5). Depois disso, foi agrupado pelo *id_discente* e criado o *DataFrame grupo_renda_pcf ['renda_pcf']*.

A variável *raca_disc* possuía 41 dados faltantes e 7.180 dados com o tipo “Atualizar junto ao R.A.”. Para realizar o tratamento, esses dois dados foram chamados de “Não declarada”. Após, foi feita a categorização personalizada transformando “Branco(a)”, “Pardo(a)”, “Preto(a)”, “Amarelo(a)”, “Indígena” e “Não declarada” para 0, 1, 2, 3, 4 e 5, respectivamente. Após, foi agrupado pelo *id_discente* e criado o *DataFrame grupo_disc_raca [‘raca’]*.

A variável *sexo_disc* não possuía dados faltantes, porém além de M (Masculino) e F (Feminino), ela possuía o N (nulo) que representa o estudante que não informou o sexo. Para o tratamento, os 1.088 dados N foram transformados em NaN e depois utilizado o método *ffill* que preenche os dados vazios com o valor da linha anterior. Para a transformação dos dados categóricos para numéricos, foi utilizada a função *LabelEncoder* que os transformou M em 1 e F em 0. Para finalizar, foi agrupado pelo *id_discente* e criado o *DataFrame grupo_disc_sexo [‘sexo_disc’]*.

A variável *estado_civil* possuía 34.866 dados faltantes, também continha um tipo ‘Não Informado’ com 996 dados, por isso, o primeiro passo foi transformar estes em NaN e, assim, fazer a imputação dos 35.862 dados faltantes a partir do método *bfill* que usa a informação da próxima linha. Para transformar os dados categóricos em numéricos, foi utilizada a função *CategoricalDtype*, em que as codificações não seguem a posição dos elementos das categorias, assim, as numerações são atribuídas inicialmente em zero na medida em que os valores aparecem. Por fim, foi agrupado pelo *id_discente* e criado o *DataFrame grupo_estado_civil [‘estado_civil’]*.

Ao analisar a variável *campus*, foram encontrados 59 dados do tipo *Instituto Federal de Santa Catarina,* porém, nenhum curso é lotado diretamente na raiz IFSC, dessa forma, foram verificados o curso e a cidade aos quais pertenciam os dados, em que a cidade de Araranguá era a presente nos cadastros do curso. Assim, foram alterados os dados do campus *Instituto Federal de Santa Catarina* para o *Campus Araranguá*. Para transformar os dados categóricos em numéricos, foi utilizada a categorização personalizada, à qual foram atribuídos valores de zero a 21. Por fim, foi agrupado pelo *id_discente* e criado o *DataFrame grupo_campus [‘campus’]*.

A variável *idhm_municipio* foi criada a partir da variável *municipio_campus* e atribuído o valor do índice de desenvolvimento humano dos municípios de Santa Catarina disponibilizado de 2010 (último disponível) (SANTA CATARINA, 2010). Os valores incorporados na nova variável foram: Florianópolis – 0.847, Jaraguá do Sul – 0.803, Joinville – 0.809, Lages – 0.770, Criciúma – 0.788, Gaspar – 0.765, Palhoça – 0.757, Canoinhas – 0.757,

São Miguel do Oeste – 0.801, Caçador – 0.735, São José – 0.809, Itajaí – 0.795, Araranguá – 0.760, Tubarão – 0.796, Chapecó – 0.790, Xanxerê – 0.775, São Carlos – 0.769, Garopaba – 0.753, Urupema – 0.699. Por fim, foi agrupado pelo *id_discente* e criado o *DataFrame grupo_idhm_municipio ['idhm_municipio']*.

Na variável *turno_curso*, foi utilizada a função *LabelEncoder* para o pré-processamento, transformando de categórica para numérica, ficando os dados de zero a três. Logo depois, foi agrupado pelo *id_discente* e criado o *DataFrame grupo_turno_curso['turno_curso']*.

A variável *tipo_aprendizagem* foi criada a partir da variável *campus*. Para isso, foi selecionado os *campi* do tipo zero (Centro de Referência de Educação a Distância) em que estão lotados os cursos do tipo EaD e, em seguida, transformada a variável do tipo *boolean* (0,1) em que um (1) ficou todos os demais *campi* diferentes de Centro de Referência de Educação a Distância. Na sequência, foi agrupado pelo *id_discente* e criado o *DataFrame grupo_tipo_aprendizagem['tipo_aprendizagem']*.

Ao listar os dados da variável *curso* a partir da função *value_counts*, foram encontrados vários cursos com grafias diferentes. Para resolver essa questão, foram alterados os nomes para deixar a grafia padronizada, para isso, foram analisados todos os cursos de graduação no *site*[1] da instituição. Depois, foi usada a função *LabelEncoder* para transformar os dados categóricos em numéricos (0 a 43) e agrupado pelo *id_discente*, criando o *DataFrame grupo_curso_disc['curso_disc']*.

A variável *naturalidade_disc* possuía 2.178 dados vazios, para isso, foi utilizado o método de imputação *SimpleImputer*, usando a estratégia "*most_frequent*" para substituir os dados NaN (*Not a Number*) pela cidade mais frequente. Em seguida, foi utilizada a função *LabelEncoder* para transformar os dados categóricos em numéricos (0 a 948) e agrupado pelo *id_discente* criando o *DataFrame grupo_naturalidade['naturalidade_disc']*.

Na variável *cidade_campus*, foi empregada a função *LabelEncoder* para transformar os dados categóricos em numéricos (0 a 18) e agrupado pelo *id_discente* criando o *DataFrame grupo_cidade_campus['cidade_campus']*.

A variável *origem_ensino_anterior* possuía 7.002 dados faltantes (NaN) e apresentava 8.056 como "Não Informada". Os dados ditos como "Não Informados" foram transformados em NaN, o que totalizou 15.058 dados faltantes. Após, para a imputação dos dados, foi utilizado o método de imputação *SimpleImputer*, usando a estratégia "*constant*" com o valor "*Estadual"* para substituir os dados NaN (*Not a Number*) pela cidade mais frequente. Em seguida, foi

[1] https://www.ifsc.edu.br/cursos.

utilizada a função *LabelEncoder* para transformar os dados categóricos em numéricos (0 a 5) e agrupado pelo *id_discente* criando o *DataFrame grupo_origem_ensino_anterior ['origem_ensino_anterior']*.

A variável *n_disciplinas_reprovadas* e a variável *n_disciplinas_concluidas* foram criadas a partir da variável *status_componente_disc*. Após, foram excluídas as linhas em que o *status* do componente era igual à "Excluída" e "Cancelada". Em seguida, foram alterados os *status* "Aprovado", "Validado RE", "Validade RS" e "Cumpriu" para "Aprovado (a)", e os *status* "Reprovado", Rep. Falta", "Não concluído" e "Indeferido" para Reprovado(a). Com o "*status_componente_disc*" apenas com o tipo "Aprovado(a)" e "Reprovado(a)", foram criadas as variáveis *dummy* com a função *get_dummies*, em que é gerada uma coluna para disciplinas Aprovado(a)s, que coloca 1 quando aprovado e 0 quando reprovado, e uma coluna para disciplinas Reprovado(a)s, em que é inserido 1 quando reprovado e 0 quando aprovado. Em seguida, foi feito o agrupamento pelo *id_discente* criando o *DataFrame grupo_disc_reprovadas['n_disciplinas_reprovadas']*, fazendo a soma das disciplinas em que o discente foi reprovado e criado o *DataFrame grupo_disc_concluídas ['n_disciplinas_concluídas']*, fazendo a soma das disciplinas em que o discente foi aprovado.

Com a descrição das duas variáveis criadas, foram analisados o intervalo que trouxe (0 a 65), a variável *n_disciplinas_reprovadas* e (0 a 614) e a variável *n_disciplinas_concluidas*. Com isso, foi feita a análise descritiva das variáveis para encontrar e tratar *outliers* (valores atípicos). Com isso, foram testados limites superiores para *n_disciplinas_reprovadas* até chegar no valor 30, gerando o *box-plot* e, assim, alterado o valor dos *outliers* pela mediana. A Figura 17 (gráfico gerado no algoritmo criado em linguagem python) fornece a distribuição das disciplinas reprovadas antes e depois da alteração de dados discrepantes.

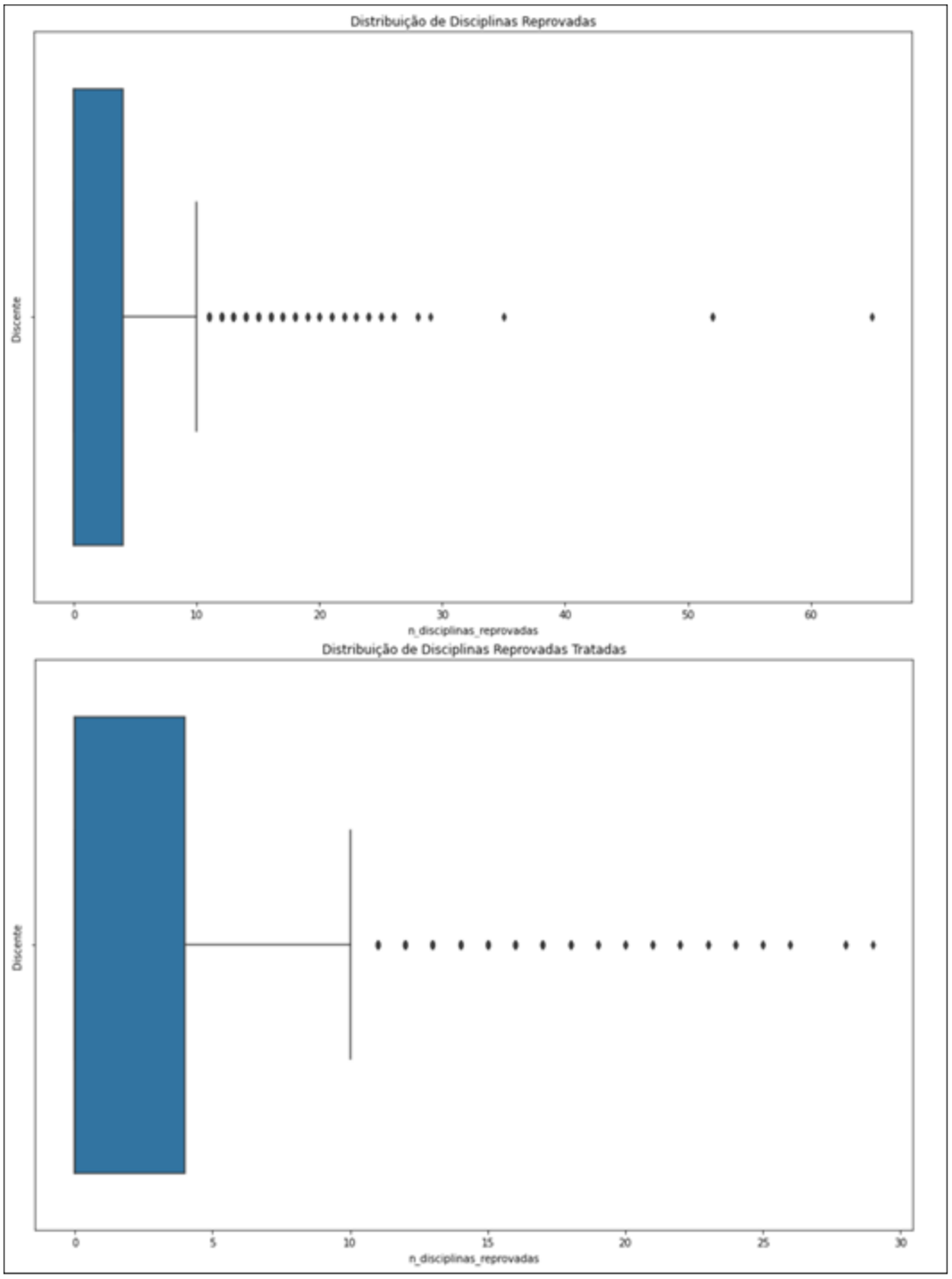


Figura 17 – Distribuição de disciplinas reprovadas antes/após tratar *outliers*
Fonte: Elaborada pela autora desta dissertação (2022)

Para a variável *n_disciplinas_concluídas*, foi inserido o limite superior de 90, pois, conforme análise dos cursos de graduação do IFSC, esse é o número máximo de disciplinas obrigatórias e optativas de um curso, gerando o *box-plot* e, assim, alterado o valor dos *outliers* pela mediana. A Figura 18 (gráfico gerado no algoritmo criado em linguagem python) fornece a distribuição das disciplinas concluídas antes e depois da alteração de dados discrepantes.

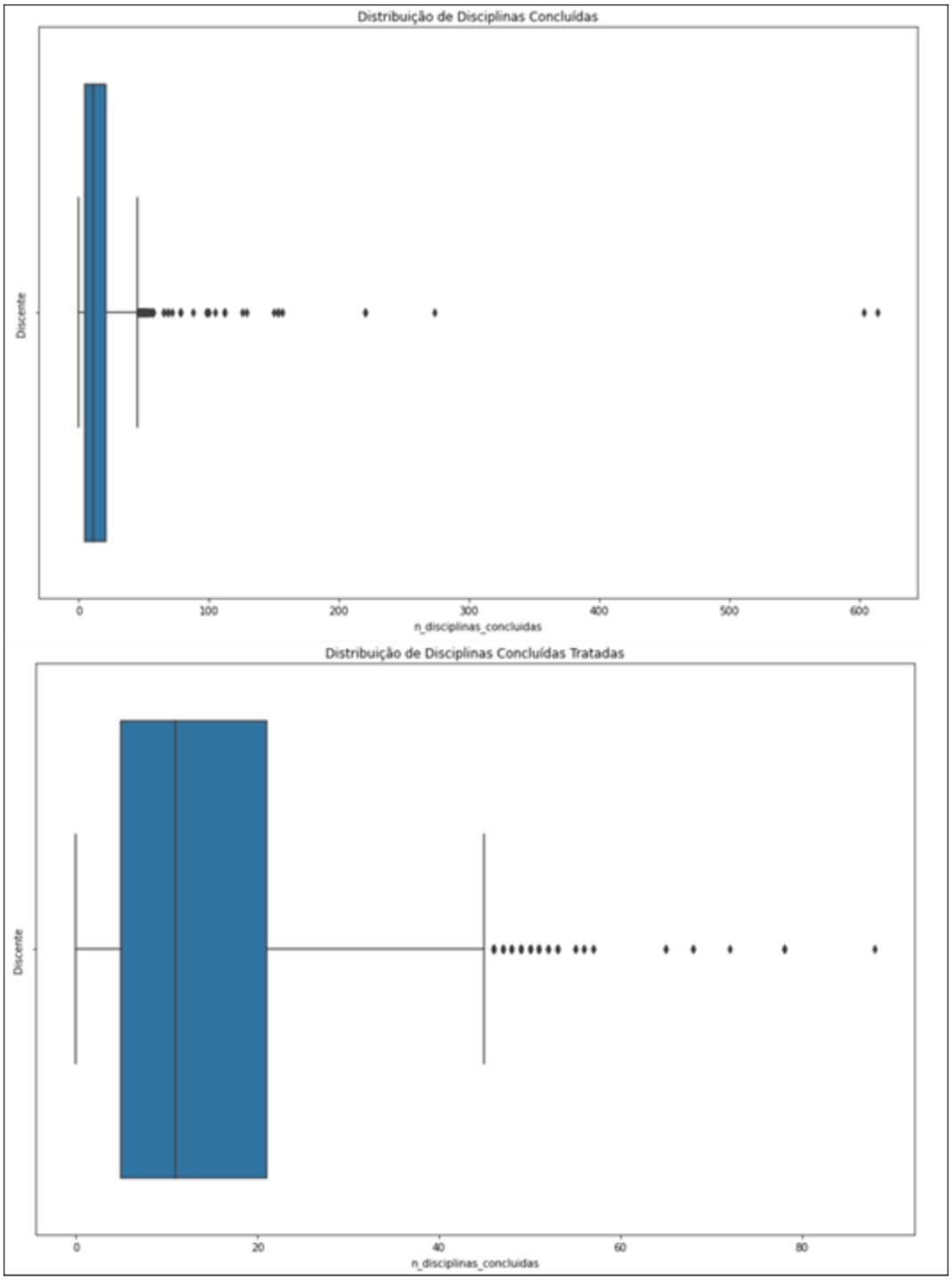


Figura 18 – Distribuição de disciplinas concluídas antes/após tratar *outliers*
Fonte: Elaborada pela autora desta dissertação (2022)

A variável *ingresso_disc* foi criada a partir das variáveis *ano_ingresso* e *periodo_ingresso*. No primeiro momento, foram transformadas as duas variáveis em *string* (*str*) e depois usada a função *map* para juntar as duas variáveis, criando *ingresso_disc*. Logo após, realizada a categorização personalizada (0 a 9) e criado o *DataFrame grupo_ingresso_disc[‘ingresso_disc’]*.

A variável *semestre_atual* representa o tempo do estudante na instituição em semestres e foi gerada a partir da variável *ingresso_disc* a partir da categorização personalizada (2 a 11). A seguir, foi criado o *DataFrame grupo_semestre_atual ['semestre_atual']*.

A variável *status_atual_disc* é a variável de saída (resultado). Foi realizada a categorização personalizada em que o *status* é definido como Ativo, Trancado, Ativo – Formando e Ativo – Graduando foram atribuídos valor zero (0) e, para o *status* Concluído, foi definido como um (1). Logo após, foi agrupado pelo *id_discente* e criado o *DataFrame grupo_status_atual_disc ['status_atual_disc']*.

A Tabela 23 apresenta os dados das 20 variáveis finais do *DataFrame* que será utilizado para criação do modelo de previsão escolar, em que *status_atual_disc* é a saída, e as demais são as entradas para o modelo de previsão.

Tabela 23 – Tratamento das variáveis do DataFrame

Variável	Tipo	Origem outras variáveis?	Dados Faltantes	Imputação	Intervalo
forma_ingresso	Cat[1]	Não	374	*SimpleImputer* - *constant* - "SiSU (ENEM)"	0 a 7
media_geral_disc	Num[2]	média_componente	8.193	*KNNImputer*	0 a 10
idade_disc	cat	data_nasc	330	*SimpleImputer – most_frequent*	0 a 6
	num	renda_familiar	7.951	*SimpleImputer – mean*	0 a 5
renda_pcf	num	membros_familia	7.951	*SimpleImputer – mean*	0 a 5
raca_disc	cat	Não	41 NaN 7.160 "Atualizar junto ao R.A."	*fillna – ffill*	0 a 4
sexo_disc	cat	Não	1.088	*fillna – ffill*	0 e 1
estado_civil	cat	Não	35.862	*fillna – bfill*	0 a 7
Campus	cat	Não	0	não	0 a 21
idhm_municipio	num	cidade_campus	0	não	0 a 18
turno_curso	cat	Não	0	não	0 a 3
tipo_aprendizagem	num	campus	0	não	0 e 1
curso_disc	cat	Não	0	não	0 a 69
naturalidade_disc	cat	Não	2.178	*SimpleImputer – most_frequent*	0 a 948
cidade_campus	cat	Não	0	não	0 a 18

Variável	Tipo	Origem outras variáveis?	Dados Faltantes	Imputação	Intervalo
origem_ensino_anterior	cat	Não	7.002 NaN 8.056 "Não informado"	*SimpleImputer* - *constant* - "Estadual"	0 a 5
n_disciplinas_reprovadas	num	status_componente_disc	0	não	0 a 29
n_disciplinas_concluidas	num	status_componente_disc	0	não	0 a 88
	num	ano_ingresso	0	não	-
ingresso_disc	num	periodo_ingresso	0	não	0 a 10
semestre_atual	num	ingresso_disc	0	não	2 a 11
status_atual_disc (saída)	cat	Não	0	não	0 e 1

[1]Categórico. [2]Numérico.
Fonte: Elaborada pela autora desta dissertação (2022)

3.5.1.2 Criação de DataFrame

Com os 19 *DataFrames* das variáveis criados, foi utilizada a função *Join* para transformá-los em um único *DataFrame* (dados_pre_processados).

A variável *ano_ingresso* foi inserida no *DataFrame* criado para ajudar na divisão e na criação de dois novos *DataFrames* df_antes_pandemia (anos 2017, 2018 e 2019) e *df_durante_pandemia* (anos 2020 e 2021).

Com o *DataFrame dados_pre_processados*, foi redefinido o *index* para retirar o número da matrícula dos estudantes e incluir um número sequencial, sem identificação nenhuma do estudante, usando a função *range*. Assim, ficaram 7.759 linhas e 20 colunas finais. A Tabela 24 representa os dados finais.

Tabela 24 – Variáveis finais e tipo

	Variável	Tipo
0	id_discente	Float
1	renda_pcf	Float
2	Raca_disc	Int
3	sexo_disc	Float
4	Campus	Float
5	status_atual_disc	Int
6	Turno_curso	Int
7	tipo_aprendizagem	Float

	Variável	Tipo
8	curso_disc	Int
9	naturalidade_disc	Int
10	cidade_campus	Int
11	forma_ingresso	Float
12	origem_ensino_anterior	Float
13	media_geral_disc	Float
14	n_disciplinas_reprovadas	Int
15	n_disciplinas_concluídas	Float
16	idhm_municipio	Float
17	ingresso_disc	Float
18	semestre_atual	Float
19	estado_civil	Float

Fonte: Elaborada pela autora desta dissertação (2022)

O *DataFrame* df_antes_pandemia (anos 2017, 2018 e 2019) permaneceu com 4.566 dados (linhas), em que cada uma representa um estudante, e 1.061 eram de ingressantes de 2017, 1.603 ingressantes de 2018 e 1.902 ingressantes do ano de 2019. Do *status* atual desses estudantes, 4.156 estão ativos e 410 concluíram seu curso.

Já o *DataFrame df_durante_pandemia* (anos 2020 e 2021) ficou com 3.193 dados, e obteve 1.751 ingressantes do ano de 2020 e 1.442 ingressantes do ano de 2021. Do *status* atual desses estudantes, 3.060 continuam no curso e 133 concluíram seu curso. É importante lembrar que os cursos tecnólogos são os cursos que possuem menor duração (6 a 7 semestres), porém, a forma do ingresso do discente analisado pode ter sido por meio do SiSU (ENEM), Ingresso sem Prova, Ingresso com Prova, Vaga Remanescente, Transferência Externa, Transferência Interna, Retorno de Egresso ou Transferência ex-Officio, o que justifica o fato de apontar os estudantes que concluíram seu curso após dois anos ingressados na instituição.

3.5.1.3 Normalização dos Dados

Primeiramente, foram analisadas as escalas das variáveis. Para reproduzir a Figura 19 (gráfico gerado no algoritmo criado em linguagem python), foram utilizadas as variáveis vistas como tendo as maiores variações entre elas e que poderiam mostrar melhor a realidade: *campus*, *curso_disc*, *idhm_municipio*, *naturalidade_disc*, *origem_ensino_anterior*, *sexo_disc* e *tipo_aprendizagem* e gerado a partir do *boxplot*.

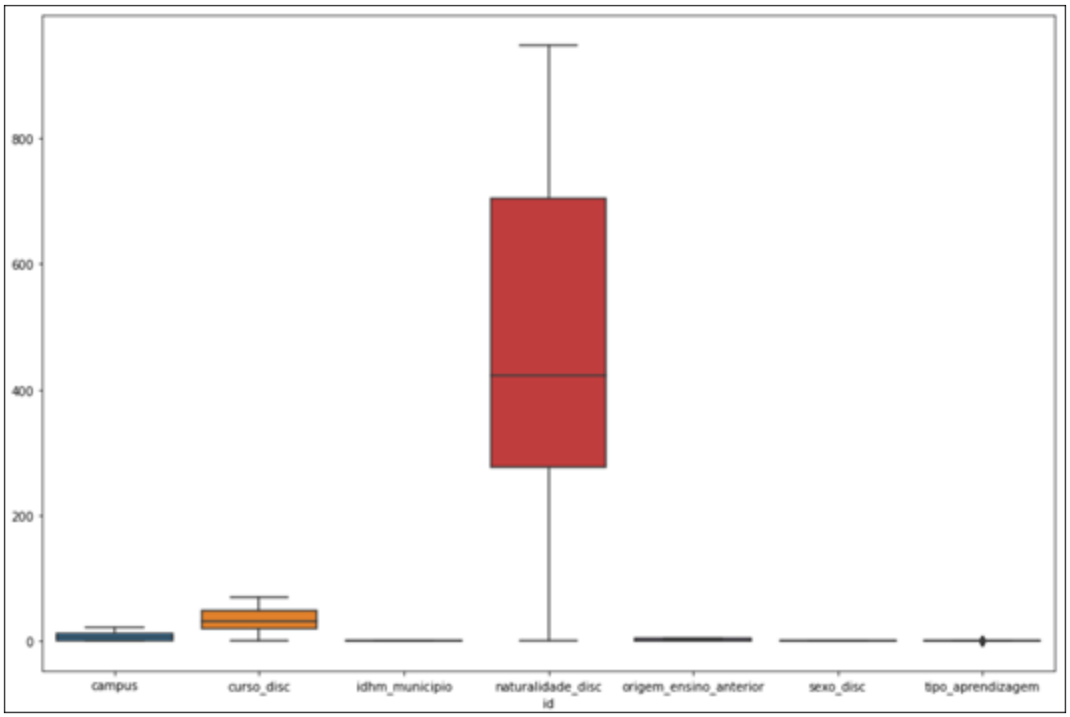


Figura 19 – Boxplot DataFrame final
Fonte: Elaborada pela autora desta dissertação (2022)

Logo em seguida, foi realizada a normalização dos dados da base df_antes_pandemia a partir do método *MinMaxScaler*. Assim, o *bloxplot* do *DataFrame* normalizado ficou da Figura 20 (gerado no algoritmo criado em linguagem *Python*) em que o eixo X representa as variáveis, e o eixo Y, o valor atribuído. O mesmo processo foi repetido no *DataFrame df_durante_pandemia*. Para fins de entendimento da Figura 20, segue a legenda do eixo X (variáveis): 0 = idade_disc; 1 = renda_pcf; 2 = raca_disc; 3 = sexo_disc; 4 = campus; 5 = status_atual_disc; 6 = turno_curso; 7 = tipo_aprendizagem; 8 = curso_disc; 9 = naturalidade_disc; 10 = cidade_campus; 11 = forma_ingresso; 12 = origem_ensino_anterior; 13 = media_geral_disc; 14 = n_disciplinas_reprovadas, 15 = n_disciplinas_concluidas; 16 = idhm_municipio; 17 = ingresso_disc; 18 = semestre_atual; e 19 = estado_civil.

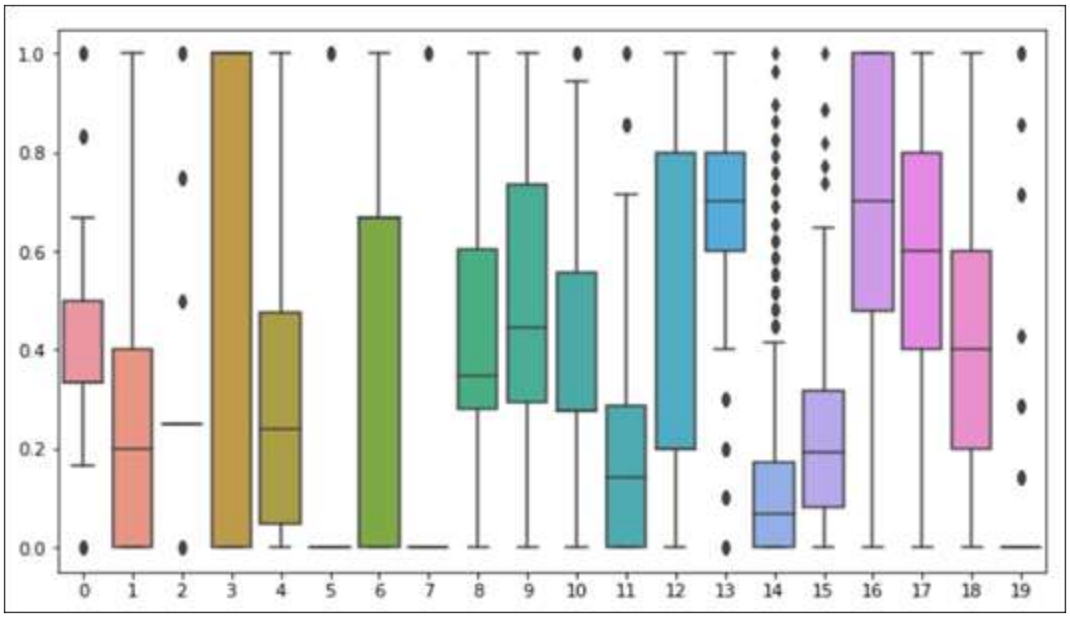


Figura 20 – Boxplot df_antes_pandemia normalizado
Fonte: Elaborada pela autora desta dissertação (2022)

Com o pré-processamento dos dados finalizado em todas as suas etapas (Limpeza e imputação dos dados, Criação do *DataFrame* e Normalização dos dados), é possível, então, desenvolver o modelo. No Capítulo 4 serão apresentadas as etapas de criação e de avaliação do modelo de evasão do IFSC.

4 MODELO DE EVASÃO ESCOLAR DO IFSC

O modelo de evasão escolar do IFSC foi desenvolvido na Linguagem *Python*, no *Notebook* da Google chamado de *Colaboratory* (*Colab*). O modelo está disponível no *link* compartilhado no Apêndice C.

No primeiro momento, foram definidas as variáveis de entrada do modelo dada por x = (idade_disc, renda_pcf, raca_disc, sexo_disc, campus, turno_curso, tipo_aprendizagem, curso_disc, naturalidade_disc, cidade_campus, forma_ingresso, origem_ensino_anterior, media_geral_disc, n_disciplinas_reprovadas, n_disciplinas_concluidas, idhm_municipio, ingresso_disc, semestre_atual e estado_civil) e a variável de saída do modelo dada por y = (status_atual_disc).

Depois, foram definidos 20% dos dados do *DataFrame* para teste e 80% dos dados para treino. Assim, para o df_antes_pandemia, foram 3.652 dados usados para treino e 914 para teste, totalizando 4.566 dados.

Em seguida, foi criado o algoritmo de *baseline*, para isso, foi usado *DecisionTreeClassifier*. As métricas do modelo foram: *Accuracy* = 98,58% de predições corretas; *Precision* = 89,66% das predições positivas corretas; *Recall* = 95,12% dos valores positivos classificados corretamente e; *F1 Score* = 92,31% de média harmônica entre *Precision* e *Recall*.

A Tabela 25 apresenta todas as métricas anteriormente, e a coluna '*Total*' representa a quantidade de dados, na qual é possível observar 832 dados em que a resposta é 0 (não concluído) e 82 como 1 (concluído). Os dados não foram balanceados, ou seja, os dados continuaram com classificações minoritárias (82 classificações como 1).

Tabela 25 – Métricas do modelo *baseline* do df_antes_pandemia

Valor	Precision	Recall	F1-Score	Total
0	-	-	-	832
1	89,66%	95,12%	92,31%	82
Accuracy			98,58%	914

Fonte: Elaborada pela autora desta dissertação (2022)

Outra métrica avaliada nos modelos foi a Matriz de Confusão, no caso do modelo *baseline* (Figura 21), que apresentou 78 classificações *True Positive* (TP) e 823 classificações *True Negative* (TN), as quais são os resultados corretos esperados, também apresentou quatro

classificações *False Positive* (FP) e nove classificações *False Negative* (FN), que são as classificações erradas.

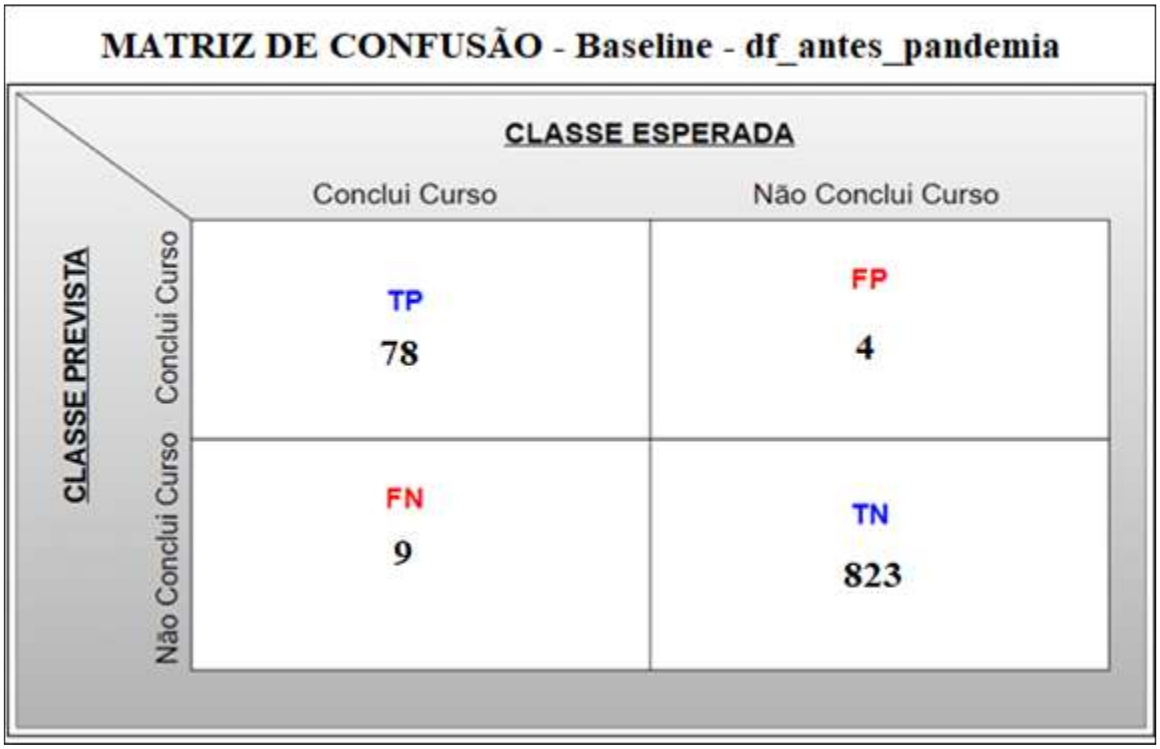


Figura 21 – Matriz de Confusão do modelo *baseline* de df_antes_pandemia
Fonte: Elaborada pela autora desta dissertação (2022)

Com o modelo que servirá de base para os demais (*baseline*), criado e avaliado, são executados então o treinamento e a avaliação dos demais modelos *XGBoost* e *Multilayer Perceptron.*

4.1 TREINAMENTO COM O ALGORITMO *XGBOOST*

Para criar o modelo com o algoritmo *XGBoost*, foi testada a função *GridSearchCV*[2] para efetuar a busca dos melhores hiperparâmetros para o modelo, no entanto, observou-se o processamento extremamente lento, mesmo diminuindo a quantidade de estimadores, por isso, optou-se, nesse modelo, por alterar os hiperparâmetros manualmente. Assim, o hiperparâmetro *learning_rate*, que apresenta a taxa de aprendizagem do modelo, melhorou quando alterada para 0.02, o hiperparâmetro *n_estimators*, que aponta a quantidade de árvores a serem rodadas no modelo, mostrou melhores resultados com 700 árvores. O hiperparâmetro *max_depth*, que apresenta a profundidade máxima das árvores, obteve melhores resultados com a profundidade

[2] GridSearchCV faz uma pesquisa exaustiva sobre os valores de parâmetros especificados para um estimador.

4. Por fim, o último hiperparâmetro alterado dos padrões foi *colsample_bynode*, que representa a fração das colunas a serem usadas pelos nós, e melhorou ao diminuir para 0.75. Dessa forma, obteve-se a configuração dos hiperparâmetros do modelo segundo a Figura 22.

```
from pandas.core.common import random_state
from xgboost import XGBClassifier

xgb = XGBClassifier (learning_rate = 0.02,
                     n_estimators=700,
                     random_state=5,
                     max_depth=4,
                     min_child_weight = 1,
                     subsample=1,
                     colsample_bynode = 0.75,
                     num_parallel_tree = 1
                     )
xgb.fit(treino_x, treino_y)
xgb_predict = xgb.predict(teste_x)
```

Figura 22 – Hiperparâmetros do modelo XGBoost
Fonte: Elaborada pela autora desta dissertação (2022)

Em seguida, foram avaliadas as métricas do modelo XGB: *Accuracy* = 99,56% de predições corretas; *Precision* = 98,75% das predições positivas corretas; *Recall* = 96,34% dos valores positivos classificados corretamente e; *F1 Score* = 97,53% de média harmônica entre *Precision* e *Recall*. A Tabela 26 apresenta todas as métricas anteriores para o modelo *XGBoost*. A coluna '*Total*' representa a quantidade de dados, em que é possível observar 832 dados em que a resposta é 0 (não concluído) e 82 como 1 (concluído).

Tabela 26 – Métricas do modelo XGBoost do df_antes_pandemia

Valor	Precision	Recall	F1-Score	Total
0	-	-	-	832
1	98,75%	96,34%	97,53%	82
Accuracy			99,56%	914

Fonte: Elaborada pela autora desta dissertação (2022)

Ao avaliar a Matriz de Confusão do modelo *XGBoost* (Figura 23) apresentou-se 79 classificações *True Positive* (TP) e 831 classificações *True Negative* (TN) as quais são os resultados corretos esperados, também apresentou três classificações *False Positive* (FP) e uma classificação *False Negative* (FN), que é a classificação errada.

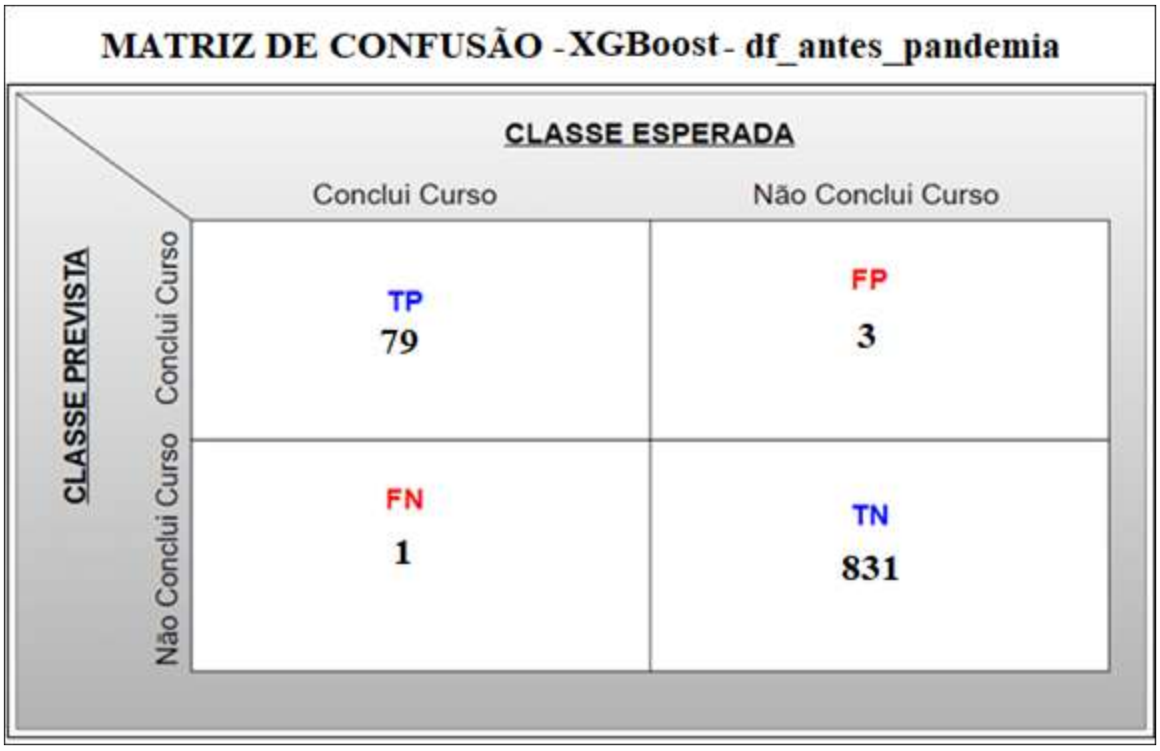


Figura 23 – Matriz de Confusão do modelo XGBoost do df_antes_pandemia
Fonte: Elaborada pela autora desta dissertação (2022)

4.2 TREINAMENTO COM O ALGORITMO *MULTILAYER PERCEPTRON*

Em modelos desenvolvidos a partir de árvores de decisão, como é o caso do *XGBoost,* não é necessário fazer escalonamento dos dados depois de um treinamento, porém, em Redes Neurais Artificiais, como é o caso do *MultiLayer Perceptron,* o escalonamento é importante para melhores resultados (GÉRON, 2019). Por isso, antes do treinamento do modelo MLP, foi usada a função *StanderScaler*, a qual faz a média dos dados e divide pelo desvio-padrão, em seguida, é feito o treinamento do escalador e reescalados os dois conjuntos, transformando o 'treino_x' e o 'teste_x' em novos conjuntos com novas escalas.

Depois de reescalar os conjuntos, foi feito o treinamento com o modelo MLP, alterando os hiperparâmetros *max_iter = 300*, que apresenta o número máximo de iterações, e *n_iter_no_change = 12,* que representa o número máximo de camadas. Optou-se por usar o padrão usado no modelo *XGBoost* de apenas alterar os hiperparâmetros manualmente, sem criar, pois, assim, é possível verificar o melhor modelo para os dados fornecidos seguindo o mesmo padrão dos modelos. A Figura 24 apresenta os hiperparâmetros utilizados no modelo MLP.

```
from sklearn.neural_network import MLPClassifier

# Iniciar modelo
mlp = MLPClassifier(random_state=1,
                    max_iter=300, # padrão é 200
                    n_iter_no_change= 12, # padrão é 10
                    )
modelo_mlp = mlp.fit(raw_treino_x, treino_y)
mlp_predict = mlp.predict(raw_teste_x)
```

Figura 24 – Treinamento do modelo MLP
Fonte: Elaborada pela autora desta dissertação (2022)

Em seguida, foram avaliadas as métricas do modelo MLP: *Accuracy* = 98,91% de predições corretas; *Precision* = 95,00% das predições positivas corretas; *Recall* = 92,68% dos valores positivos classificados corretamente e; *F1 Score* = 93,83% de média harmônica entre *Precision* e *Recall*. A Tabela 27 apresenta todas as métricas anteriormente para o modelo MLP. A coluna 'Total' representa a quantidade de dados, em que é possível observar 832 dados em que a resposta é 0 (não conclui curso) e 82 como 1 (conclui curso).

Tabela 27 – Métricas do modelo MLP do df_antes_pandemia

Valor	Precision	Recall	F1-Score	Total
0	-	-	-	832
1	95,00%	92,68%	93,83%	82
Accuracy			98,91%	914

Fonte: Elaborada pela autora desta dissertação (2022)

Ao avaliar a Matriz de Confusão do modelo MLP (Figura 25), foram apresentadas 76 classificações *True Positive* (TP) e 828 classificações *True Negative* (TN), as quais são os resultados corretos esperados, ela também apresentou seis classificações *False Positive* (FP) e quatro classificações *False Negative* (FN), que são as classificações erradas.

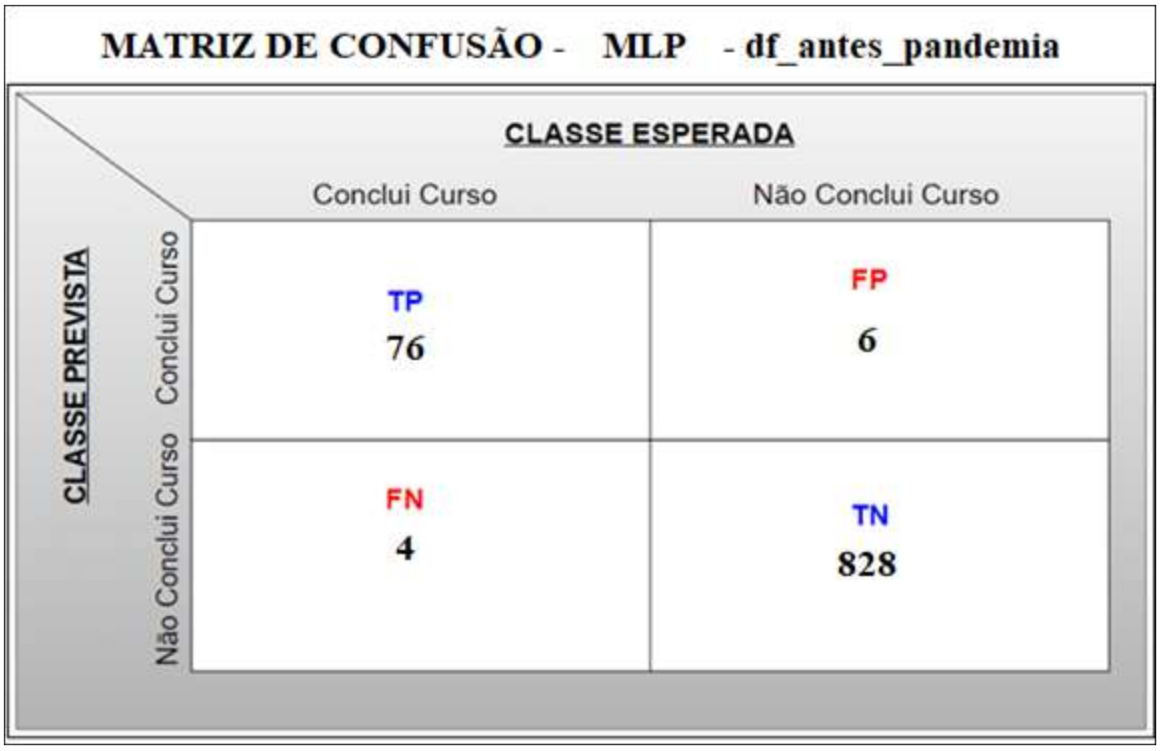


Figura 25 – Matriz de Confusão do modelo MLP de df_antes_pandemia
Fonte: Elaborada pela autora desta dissertação (2022)

Com os modelos que representam a base de dados referentes aos anos 2017, 2018 e 2019 e fornecem os dados antes da pandemia da Covid-19 treinados e avaliados, é possível então repetir todo o processo para a base de dados referente aos dados dos anos de 2020 e 2021 que fornecem dados durante a pandemia da Covid-19.

4.3 TREINAMENTO DO MODELO COM DADOS DE 2020 E 2021

As variáveis e a divisão dos dados de treino e de teste seguiram o mesmo procedimento do *DataFrame* df_antes_pandemia. Assim, para o *df_durante_pandemia*, foram 2.394 dados usados para treino e 799 para teste, totalizando 3.193 dados.

Em seguida, foi criado o algoritmo de *baseline*, para isso, foi usado *DecisionTreeClassifier*. As métricas obtidas do modelo foram: *Accuracy* = 97,87% de predições corretas; *Precision* = 75,00% das predições positivas corretas; *Recall* = 72,73% dos valores positivos classificados corretamente; e *F1 Score* = 73,85% de média harmônica entre *Precision* e *Recall*. A Tabela 28 apresenta todas as métricas anteriormente para o modelo baseline. A coluna ‘Total’ representa a quantidade de dados, em que é possível observar 766 dados em que a resposta é 0 (não concluído) e 33 como 1 (concluído).

Tabela 28 – Métricas do modelo *baseline* do df_durante_pandemia

Valor	Precision	Recall	F1-Score	Total
0	-	-	-	766
1	75,00%	72,73%	73,85%	33
Accuracy			97,87%	799

Fonte: Elaborada pela autora desta dissertação (2022)

Ao avaliar a Matriz de Confusão do modelo *baseline* (Figura 26), foram obtidas 24 classificações *True Positive* (TP) e 758 classificações *True Negative* (TN), as quais são os resultados corretos esperados, ela também apresentou nove classificações *False Positive* (FP) e oito classificações *False Negative* (FN), que são as classificações erradas.

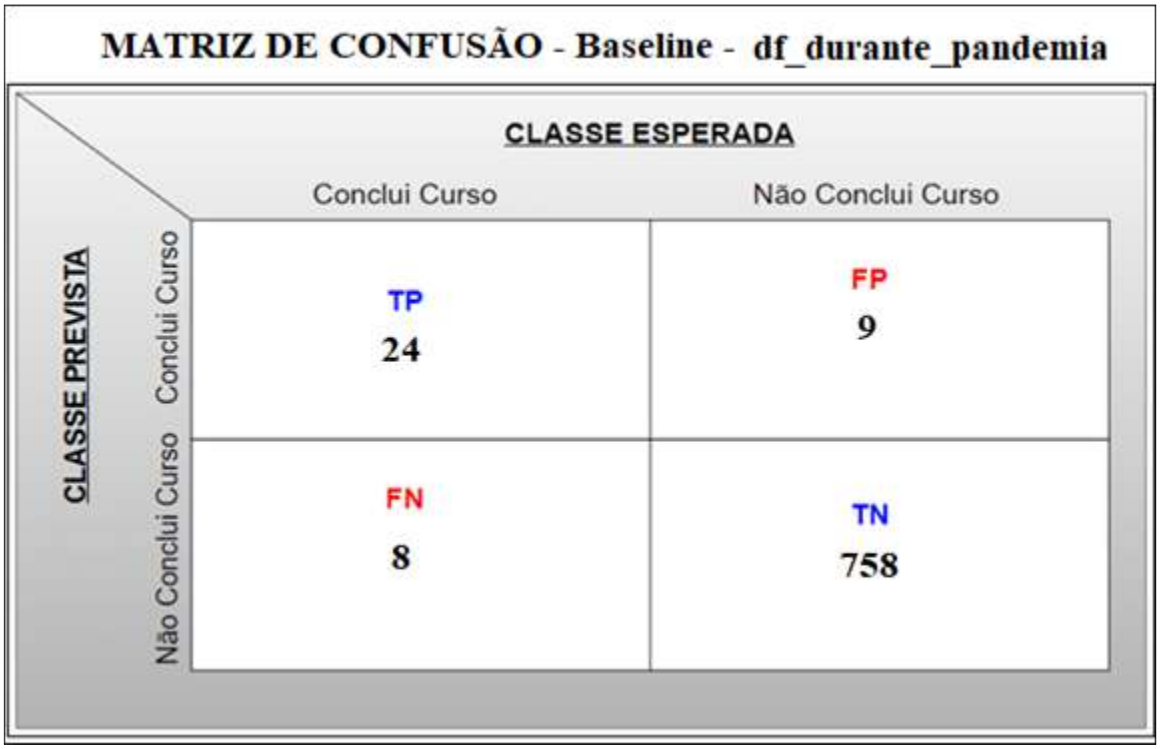


Figura 26 – Matriz de Confusão do modelo *baseline* de df_durante_pandemia
Fonte: Elaborada pela autora desta dissertação (2022)

Após criar o modelo baseline, foi feito o treinamento com o algoritmo *XGBoost*, seguindo os mesmos hiperparâmetros definidos para o *Dataframe* df_antes_pandemia. As métricas obtidas com o modelo foram: *Accuracy* = 99,25% de predições corretas; *Precision* = 96,55% das predições positivas corretas; *Recall* = 84,85% dos valores positivos classificados corretamente e; *F1 Score* = 90,32% de média harmônica entre *Precision* e *Recall*. A Tabela 29 apresenta todas as métricas anteriores para o modelo XGB. A coluna 'Total' representa a

quantidade de dados, em que é possível observar 766 dados em que a resposta é 0 (não concluído) e 33 como 1 (concluído).

Tabela 29 – Métricas do modelo XGBoost do df_durante_pandemia

Valor	Precision	Recall	F1-Score	Total
0	-	-	-	766
1	96,55%	84,85%	90,32%	33
Accuracy			99,25%	799

Fonte: Elaborada pela autora desta dissertação (2022)

Ao avaliar a Matriz de Confusão do modelo XGB (Figura 27), foram obtidas 28 classificações *True Positive* (TP) e 765 classificações *True Negative* (TN), as quais são os resultados corretos esperados, ela também apresentou cinco classificações *False Positive* (FP) e uma classificação *False Negative* (FN), que é a classificação errada.

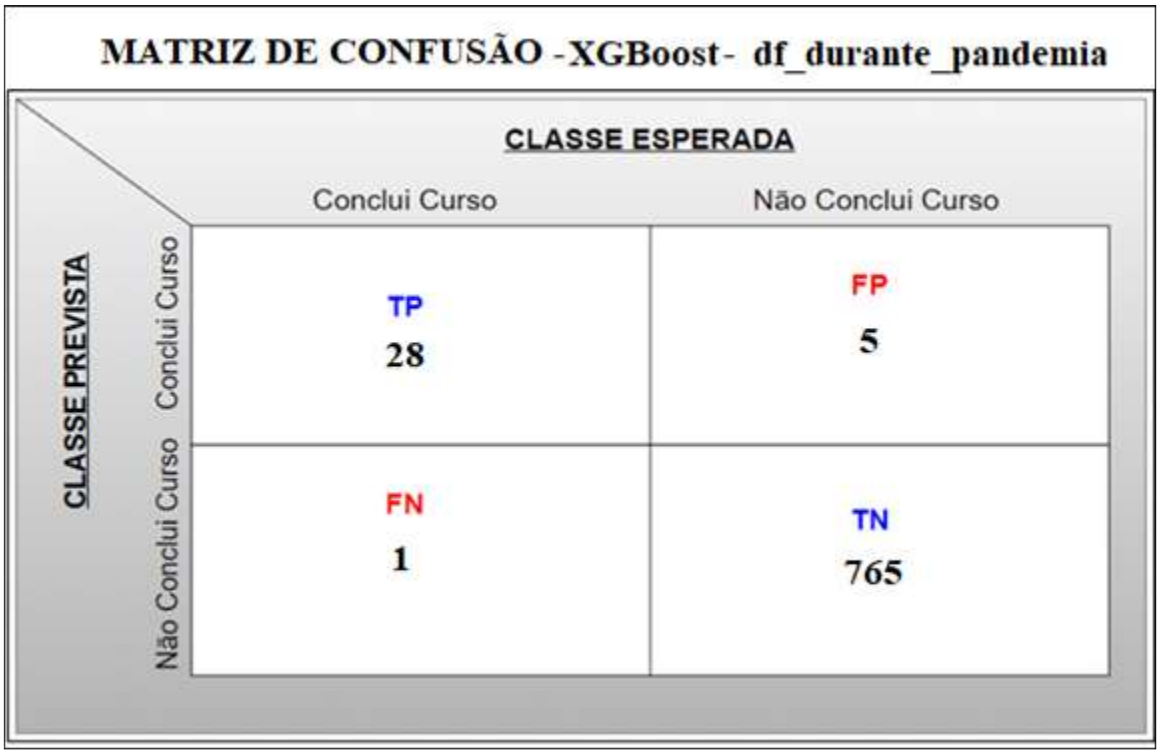


Figura 27 – Matriz de Confusão do modelo XGBoost de df_durante_pandemia
Fonte: Elaborada pela autora desta dissertação (2022)

Depois, foi realizado o treinamento com o algoritmo MLP, seguindo os mesmos hiperparâmetros definidos para o *Dataframe* df_antes_pandemia, assim como o escalonador para repreparação dos dados. As métricas obtidas com o modelo foram: *Accuracy* = 98,50% de predições corretas; *Precision* = 88,89% das predições positivas corretas; *Recall* = 72,73% dos

valores positivos classificados corretamente e; *F1 Score* = 80,00% de média harmônica entre *Precision* e *Recall*. A Tabela 30 apresenta todas as métricas anteriormente para o modelo MLP. A coluna '*Total*' representa a quantidade de dados, em que é possível observar 766 dados em que a resposta é 0 (não concluído) e 33 como 1 (concluído).

Tabela 30 – Métricas do modelo MLP do df_durante_pandemia

Valor	Precision	Recall	F1-Score	Total
0	-	-	-	766
1	88,89%	72,73%	80,00%	33
Accuracy			98,50%	799

Fonte: Elaborada pela autora desta dissertação (2022)

Ao avaliar a Matriz de Confusão do modelo MLP (Figura 28), obteve-se 24 classificações *True Positive* (TP) e 763 classificações *True Negative* (TN), as quais são os resultados corretos esperados, tela também apresentou nove classificações *False Positive* (FP) e três classificações *False Negative* (FN), que são as classificações erradas.

Figura 28 – Matriz de Confusão do modelo MLP de df_durante_pandemia
Fonte: Elaborada pela autora desta dissertação (2022)

4.4 AVALIAÇÃO DO MODELO

A avaliação do modelo foi realizada a partir de cinco métricas: *Accuracy*, *Precision*, *Recall*, *F1-Score* e Matriz de Confusão. Ao avaliar a *Accuracy*, que apresenta a taxa de acerto do modelo, para o df_antes_pandemia, o *baseline* DT obteve 98,58% de acerto, o algoritmo MLP obteve 98,91%, tendo uma melhora de 0,33% em comparação ao *baseline*, e o algoritmo *XGBoost* obteve 99,56%, uma melhora de 0,98% em comparação ao *baseline*. A *Precision,* que fornece o percentual de predições positivas que estavam corretas, do df_antes_pandemia para o *baseline* foi de 89,66%, o algoritmo MLP obteve 95%, uma melhora de 5,34% em comparação ao *baseline*, e o algoritmo *XGBoost* obteve 98,75%, uma melhora de 9,09% em comparação ao *baseline*. A *Recall,* que fornece a taxa de detecção do modelo, do df_antes_pandemia para o *baseline* foi de 95,12%, o algoritmo MLP obteve 92,68%, uma piora de 2,44% em comparação ao *baseline*, e o algoritmo *XGBoost* obteve 96,34%, uma melhora de 1,22% em comparação ao *baseline*. O *F1-Score,* que fornece a média harmônica entre *recall* e *precision*, do df_antes_pandemia para o *baseline* foi de 92,31%, o algoritmo MLP obteve 93,83%, uma melhora de 1,52% em comparação ao *baseline*, e o algoritmo *XGBoost* obteve 97,53%, uma melhora de 5,22% em comparação ao *baseline*.

Ao avaliar a *Accuracy* para o df_durante_pandemia, o *baseline* DT obteve 97,87% de acerto, o algoritmo MLP obteve 98,50%, tendo uma melhora de 0,63% em comparação ao *baseline*, e o algoritmo *XGBoost* obteve 99,25%, uma melhora de 1,38% em comparação ao *baseline*. A *Precision,* do df_durante_pandemia para o *baseline* foi de 75%, o algoritmo MLP obteve 88,89%, uma melhora de 13,89% em comparação ao *baseline*, e o algoritmo *XGBoost* obteve 96,55%, uma melhora de 21,55% em comparação ao *baseline*. A *Recall* do df_durante_pandemia para o *baseline* foi de 72,73%, o algoritmo MLP obteve 72,73% o mesmo valor do *baseline*, e o algoritmo *XGBoost* obteve 84,85%, uma melhora de 12,12% em comparação ao *baseline*. O *F1-Score* do df_durante_pandemia para o *baseline* foi de 73,85%, o algoritmo MLP obteve 80%, uma melhora de 6,15% em comparação ao *baseline*, e o algoritmo *XGBoost* obteve 90,32%, uma melhora de 10,32% em comparação ao *baseline*. A Tabela 31 apresenta todas as métricas dos modelos. Para melhor visualização dos modelos na tabela, os melhores resultados de cada métrica são apresentados na cor verde, os piores resultados na cor vermelha e os que obtiveram resultados iguais na cor azul.

Tabela 31 – Métricas de avaliação nos modelos

	df_antes_pandemia			df_durante_pandemia		
	Baseline - DT	XGBoost	MLP	Baseline - DT	XGBoost	MLP
Accuracy	98,58%	99,56%	98,91%	97,87%	99,25%	98,50%
Precision	89,66%	98,75%	95%	75%	96,55%	88,89%
Recall	95,12%	96,34%	92,68%	72,73%	84,85%	72,73%
F1-Score	92,31%	97,53%	93,83%	73,85%	90,32%	80%

Fonte: Elaborada pela autora desta dissertação (2022)

Como visto, em todas as métricas analisadas, o modelo *XGBoost* apresentou os melhores resultados, se mostrando constante em seus resultados, enquanto o modelo com MLP apresentou resultados sempre piores que o *XGBoost* e até igual (*Recall* – df_durante_pandemia) ou mesmo pior (*Recall* – df_antes_pandemia) que o modelo *baseline* (*Decision Tree*).

Os resultados do df_antes_pandemia demonstram, de forma geral, ser melhores que o df_durante_pandemia, na medição da *Accuracy* as duas bases apresentaram boa *performance* dos modelos, porém, nas demais métricas, o df_antes_pandemia se mostrou melhor em todos os algoritmos.

Quando analisada a Matriz de Confusão dos modelos (Figura 22), sabendo que a classe esperada é a resposta real e a classe prevista a resposta dada pelo modelo, é possível aferir que no *DataFrame* df_antes_pandemia, o *baseline* conseguiu prever 78 dos 82 resultados positivos (conclui curso = 1), o modelo *XGBoost* se mostrou melhor que o *baseline*, acertando 79 dos 82 resultados positivos, porém, o modelo MLP apresentou resultado abaixo do *baseline*, acertando 76 dos 82 resultados positivos. Dessa forma, o modelo MLP teve maior erro de resultados positivos (6), bem acima do *baseline* (4), já o modelo XGB obteve bom resultado (3). Quando analisados os resultados negativos, o *baseline* conseguiu prever 823 das 832 respostas negativas (Não conclui curso = 0), o modelo *XGBoost* conseguiu prever 831 das 832 respostas negativas, sendo muito melhor que o *baseline* e, por fim, o modelo MLP conseguiu prever 828 das 832 respostas negativas, conseguindo também ser melhor que o *baseline*. Dessa forma, o modelo MLP teve maior erro de resultados negativos (4) quando comparado ao modelo XGB que obteve apenas um erro, porém, quando comparado ao *baseline*, os dois modelos mostraram bons resultados.

Ao avaliar a Matriz de Confusão do *DataFrame* df_durante_pandemia, o *baseline* conseguiu prever 24 dos 33 resultados positivos (conclui curso = 1), o modelo *XGBoost* se mostrou melhor que o *baseline*, acertando 28 dos 33 resultados positivos, já o modelo MLP apresentou o mesmo resultado do *baseline*, acertando 24 dos 33 resultados positivos. Dessa

forma, o modelo MLP previu nove resultados positivos de forma errada, tendo a mesma previsão do *baseline*, já o modelo XGB obteve melhor resultado errando apenas cinco previsões positivas. Quando analisados os resultados negativos, o *baseline* conseguiu prever 758 das 766 respostas negativas (Não conclui curso = 0), o modelo *XGBoost* conseguiu prever 765 das 766 respostas negativas, sendo muito melhor que o *baseline* e, por fim, o modelo MLP conseguiu prever 763 das 766 respostas negativas, conseguindo também ser melhor que o *baseline*. Dessa forma, o modelo MLP teve maior erro de resultados negativos (3) quando comparado ao modelo XGB que obteve apenas um erro, porém, quando comparado ao *baseline* (8 erros), os dois modelos mostraram bons resultados.

A Matriz de Confusão (Figura 29) conseguiu corroborar com as outras métricas averiguadas anteriormente (*Accuracy*, *Precision*, *Recall* e *F1_Score*), mostrando a superioridade do modelo *XGBoost* que se manteve constante em suas previsões, sempre melhor que o modelo *baseline* (DT). Do mesmo modo, o modelo MLP apresentou resultados inconstantes, tendo caso em que foi igual ou pior que o modelo *baseline.*

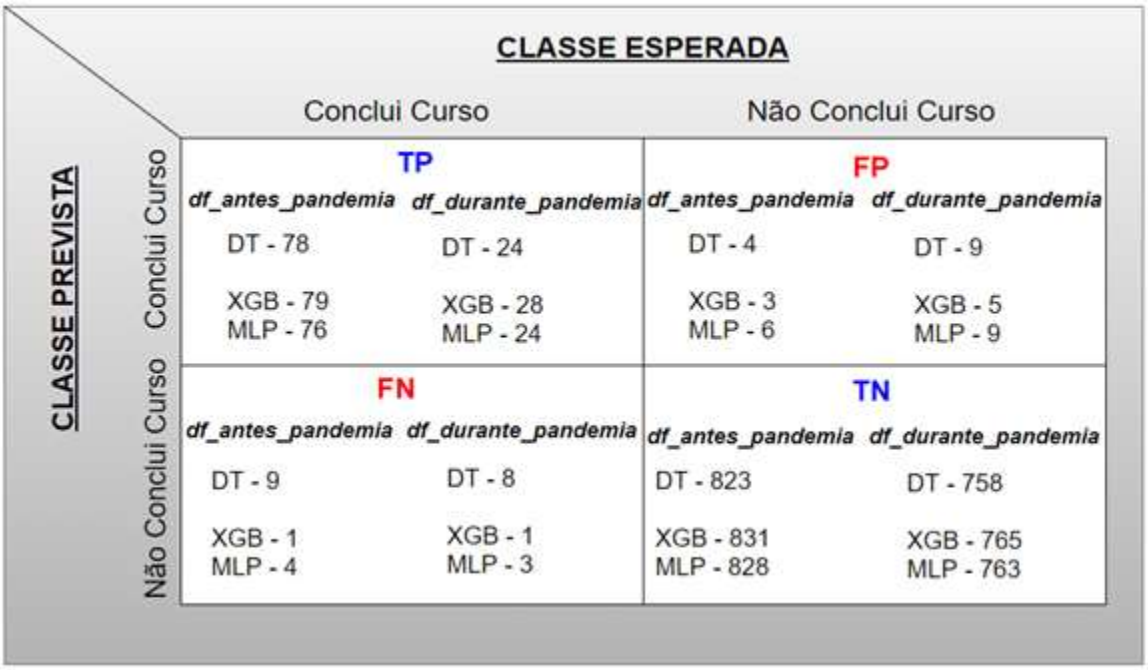


Figura 29 – Matriz de Confusão – Comparação entre os modelos
Fonte: Elaborada pela autora desta dissertação (2022)

Para avaliar a importância das variáveis, o algoritmo *XGBoost* oferece algumas funções que fornecem o gráfico e uma lista com as variáveis a partir do nível de relevância no modelo. Nesse caso, foi utilizada a função *plot_importance*, a qual mostra a importância das variáveis a partir do peso (quantas vezes um atributo aparece nas árvores) por meio do gráfico, ela

representa o ganho médio em todos os nós em que a variável foi usada (HARRISION, 2020). As demais configurações básicas dos gráficos são apresentadas na Figura 30, apenas foi alterada a variável ‘xgb1’ que representa as variáveis de df_antes_pandemia para ‘xgb2’ que representa as variáveis de df_durante_pandemia.

```
fig, ax = plt.subplots(figsize=(6,4))
xgb.plot_importance(xgb1,
                    ax=ax,
                    title='Importância das Variáveis df_antes_pandemia',
                    xlabel='F score', # nome/rótulo no eixo x
                    ylabel='Variáveis', # nome/rótulo no eixo y
                    grid=False, # inclusão de grid (grades)
                    show_values=True)
```

Figura 30 – Importância das variáveis usando XGBoost
Fonte: Elaborada pela autora desta dissertação (2022)

No *DataFrame* df_antes_pandemia (Figura 31), a variável que apresentou maior relevância foi n_disciplinas_concluidas (com 978 aparição nas árvores), seguida de forma_ingresso (470 aparições), media_geral_disc (428 aparições), renda_pcf (406 aparições) e campus (392 aparições). A menos importante para o df foi considerada a variável estado_civil com nove aparições, seguida de raca_disc com 10 aparições, tipo_aprendizagem com 29 aparições, sexo_disc com 37 aparições e semestre_atual com 65 aparições. Nesse caso, é possível perceber a extrema importância da variável n_disciplinas_concluídas que teve mais que o dobro de aparições da segunda colocada (forma_ingresso).

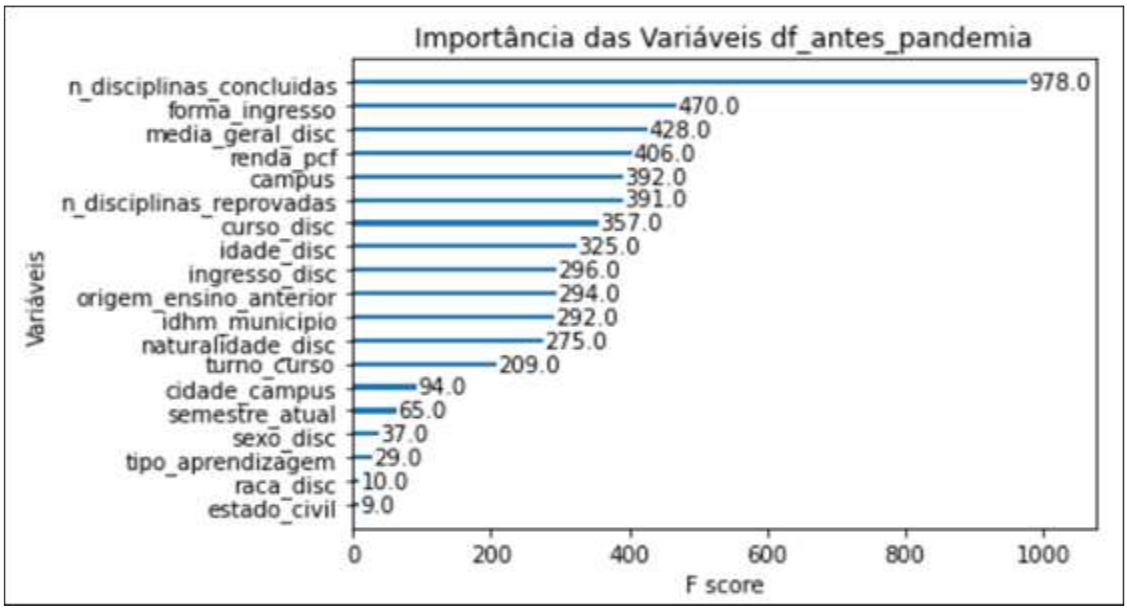


Figura 31 – Importância das variáveis usando XGBoost – df_antes_pandemia
Fonte: Elaborada pela autora desta dissertação (2022)

Para o *DataFrame* df_durante_pandemia (Figura 32), a variável mais importante foi idade_disc (com 513 aparições nas árvores), seguida de forma_ingresso (460 aparições), curso_disc (365 aparições), naturalidade_disc (354 aparições) e media_geral_disc (299 aparições). A variável menos importante foi considerada a variável tipo_aprendizagem com 17 aparições, seguida de estado_civil com 19 aparições, raca_disc com 23 aparições, semestre_atual com 72 aparições e sexo_disc com 125 aparições. Nesse caso, é possível perceber a importância das variáveis idade_disc e forma_ingresso sob as demais.

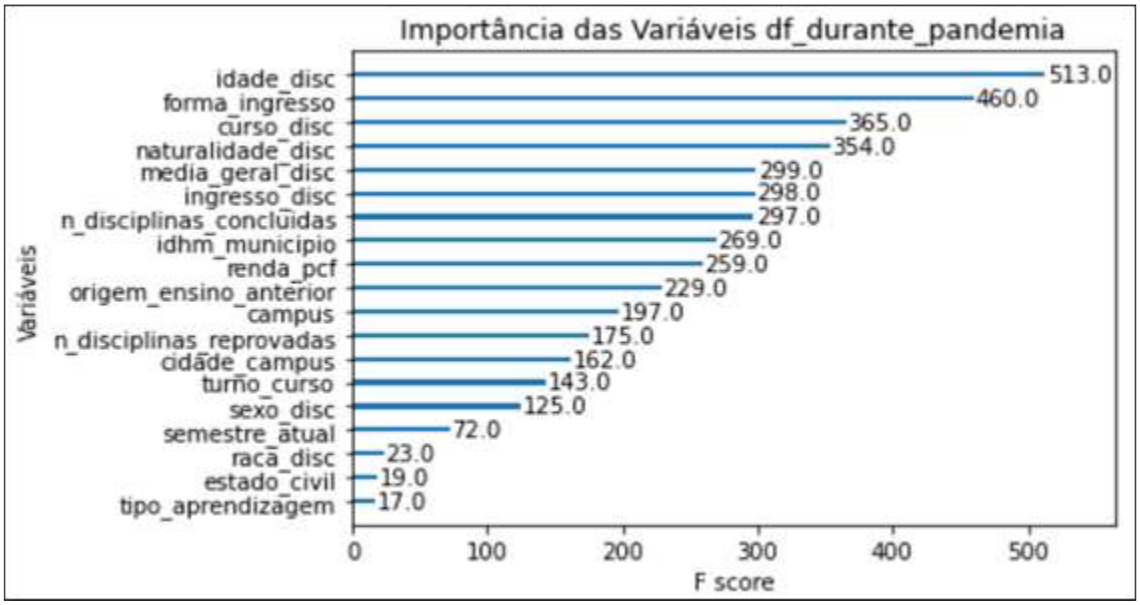


Figura 32 – Importância das variáveis usando XGBoost – df_durante_pandemia
Fonte: Elaborada pela autora desta dissertação (2022)

A forma de ingresso (forma_ingresso) do estudante mostrou-se a segunda variável mais importante nas duas bases de dados (df_antes_pandemia e df_durante_pandemia). Outros autores, como Silva, Almeida e Ramalho (2020) e Delen (2010), obtiveram resultados parecidos. Na pesquisa de Silva, Almeida e Ramalho (2020), a variável forma de ingresso foi importante nos seis modelos avaliados. O modelo de Delen (2010) considerou forma de ingresso e ficou com oito variáveis, a mais importante. Vale ressaltar que o estudo de Delen (2010) avaliou uma universidade privada e, com isso, a maior parte das variáveis que se tornaram mais importantes que forma de ingresso estão voltadas para a ajuda financeira.

A idade do discente (idade_disc) foi considerada a variável mais importante da base de dados durante a pandemia (df_durante_pandemia), porém, antes da pandemia (df_antes_pandemia), a idade do discente era apenas a oitava colocada em importância, concluindo que a idade se tornou mais importante para se manter no curso de graduação. A variável idade evidenciou importância mesmo antes da pandemia nos estudos de Silva, Almeida

e Ramalho (2020) e esteve presente entre as dez mais importantes variáveis em quatro dos seis modelos analisados. Já a pesquisa de Muñiz *et al.* (2019) concluiu que a idade do discente está entre as duas características de maior importância para a sua permanência até a conclusão do curso.

O número de disciplinas concluídas (n_disciplinas_concluidas) foi a variável mais importante antes da pandemia, já durante a pandemia, ela foi a sétima colocada. Estudos que precedem, como os de Muñiz *et al.* (2019), usaram as disciplinas concluídas como um fator importante para a formação do discente.

A média geral das notas obtida pelo estudante (media_geral_disc) também se mostrou entre as variáveis mais importantes das duas bases de dados. Os resultados corroboram com estudos anteriores, como o de Beaulac e Rosenthall (2019), que definiram as notas como a variável mais importante, e o de Gismondi e Huiman (2021), que também trabalharam com o algoritmo *XGBoost*, em que as médias, juntamente com o semestre do estudante, foram as variáveis que mais influenciam na previsão dos estudantes evadirem.

Nos dois gráficos (Figura 31 e Figura 32), pode-se ver que as variáveis menos importantes foram tipo_aprendizagem, estado_civil e raca_disc entre as três últimas e semestre_atual e sexo_disc entre as cinco piores variáveis.

Com a análise do modelo a partir das métricas de avaliação e das variáveis por meio da função de importância das variáveis, é possível, então, desenvolver um modelo ajustado. Para isso, serão retiradas as três variáveis menos importantes para o modelo (tipo_aprendizagem, estado_civil e raca_disc) e concentrar-se no melhor algoritmo (*XGBoost*) para verificar se é possível melhorar o modelo, obtendo um modelo ótimo para a previsão de evasão escolar do IFSC. Na próxima seção (4.5), serão fornecidos os detalhes do modelo ajustado.

4.5 MODELO AJUSTADO GERADO APÓS AVALIAÇÃO

Após o desenvolvimento do modelo de evasão escolar do IFSC, utilizando os algoritmos *XGBoost* e *MultiLayer Percetron*, e a avaliação dele, constatou-se que o algoritmo *XGBoost* obteve os melhores resultados nas duas bases de dados (df_antes_pandemia e df_durante_pandemia) diante de todas as métricas utilizadas. Também se percebeu, por meio da avaliação da importância das variáveis com a função *plot_importance* do algoritmo *XGBoost*, que, tanto no *DataFrame* df_antes_pandemia quanto em df_durante_pandemia, as três piores variáveis foram raca_disc, estado_civil e tipo_aprendizagem.

Nesse sentido, a fim de refinar o modelo conseguindo o seu ápice, foi aplicado o algoritmo *XGBoost e* tentou-se aprimorar os hiperparâmetros com as mesmas bases de dados, porém retirando as três piores variáveis, treinando e testando o modelo ajustado. Após, foram utilizadas as mesmas métricas (*Accuracy*, *Precision*, *Recall*, *F1-Score* e Matriz de Confusão) e comparadas com os resultados do primeiro modelo, a fim de verificar a evolução do modelo ajustado.

Para esse propósito, as variáveis x (entradas) do modelo ajustado foram: idade_disc, renda_pcf, sexo_disc, campus, turno_curso, curso_disc, naturalidade_disc, cidade _campus, forma_ingresso, origem_ensino_anterior, media_geral_disc, n_disciplinas_reprova das, n_disciplinas_concluidas, idhm_municipio, ingresso_disc e semestre_atual. A variável y (saída) permaneceu status_atual_disc.

Após isso, foi mantido 20% dos dados do *DataFrame* para teste e 80% dos dados para treino. Assim, para o df_antes_pandemia, foram 3.652 dados usados para treino e 914 para teste, totalizando 4.566 dados.

Em seguida, foi utilizada a função *RandomizedSearchCV*[3] da biblioteca do Sklearn para tentar melhorar os hiperparâmetros do treinamento do modelo *XGBoost*. Dessa forma, foi importada a função *RandomizedSearchCV* e foram criados os valores dos hiperparâmetros para teste 'param_rand', conforme apresentado na Figura 33. Depois, foi chamado o algoritmo *XGBClassifier* e foram definidos os parâmetros *n_iter*=32 e *scoring* = '*accuracy*' e realizado o treino. Assim, foram processados 160 treinamentos, o que demorou cerca de sete minutos para a finalização destes.

[3] A função *RandomizedSearchCV* faz uma pesquisa aleatória de hiperparâmetros, implementando um método '*fit*' e um método '*score*'. Os parâmetros do estimador usados para aplicar os métodos são otimizados por pesquisa de validação cruzada sobre as configurações de parâmetros. Diferente do *GridSearchCV*, ele não testa todos os valores, mas um número fixo de configurações de parâmetro das distribuições especificadas é amostrado. Este número é dado por '*n_iter*' (RANDOM, 2022).

```
# TUNAGEM RANDOMIZADA
# Efetuar os treinos de forma randomizada - se torna melhor que o GridSearchCV
# pq ela testa muito mais de forma aleatória e mais rápida
from pandas.core.common import random_state
from xgboost import XGBClassifier
import xgboost as xgb
from sklearn.model_selection import RandomizedSearchCV

param_rand = {'min_child_weight': [1, 3, 5],
              'gamma': [0.5, 1, 1.5, 2],
              'subsample': [0.8, 1, 2],
              'colsample_bytree': [0.6, 0.8, 1],
              'max_depth': [3, 4, 5, 6, 8],
              'learning_rate': [0.01, 0.02, 0.03, 0.04],
              'n_estimators': [500, 700, 750, 800, 850, 900],
              'colsample_bytree': [0.3, 0.5, 0.75, 0.8],
              'colsample_bynode': [0.3, 0.5, 0.75, 0.8]}

xgb_rd=XGBClassifier()
xgb_rand = RandomizedSearchCV(xgb_rd, param_rand, n_iter=32,
                              scoring = "accuracy", cv = 5, verbose=True,
                              n_jobs=-1, random_state=5)
xgb_rand.fit(treino_x, treino_y)
```

Figura 33 – Treinamento com XGBoost – modelo ajustado com RandomizedSearchCV
Fonte: Elaborada pela autora desta dissertação (2022)

Como resultados, obteve-se a configuração dos melhores parâmetros e estimadores sendo colsample_bynode = 0.75, colsample_bytree = 0.75, gamma = 2, learning_rate = 0.04, max_depth = 4, min_child_weight = 1, n_estimators = 800 e subsample = 1 representados na Figura 34. Também foram obtidos os melhores estimadores e foi definida a acurácia para avaliar o modelo, assim, o resultado do treinamento com as configurações dada pela função *RandomizedSearchCV* foi igual a 99,37%.

```
# Melhores parâmetros dos randomizados
xgb_rand.best_params_

{'colsample_bynode': 0.75,
 'colsample_bytree': 0.75,
 'gamma': 2,
 'learning_rate': 0.04,
 'max_depth': 4,
 'min_child_weight': 1,
 'n_estimators': 800,
 'subsample': 1}

# Melhor Estimador
xgb_rand.best_estimator_

XGBClassifier(base_score=0.5, booster='gbtree', callbacks=None,
              colsample_bylevel=1, colsample_bynode=0.75, colsample_bytree=0.75,
              early_stopping_rounds=None, enable_categorical=False,
              eval_metric=None, gamma=2, gpu_id=-1, grow_policy='depthwise',
              importance_type=None, interaction_constraints='',
              learning_rate=0.04, max_bin=256, max_cat_to_onehot=4,
              max_delta_step=0, max_depth=4, max_leaves=0, min_child_weight=1,
              missing=nan, monotone_constraints='()', n_estimators=800,
              n_jobs=0, num_parallel_tree=1, predictor='auto', random_state=0,
              reg_alpha=0, reg_lambda=1, ...)

print("A acurácia do RandomizedSearchCV: %.5f" % (xgb_rand.best_score_))

A acurácia do RandomizedSearchCV: 0.99370
```

Figura 34 – Resultado treinamento XGBoost com RandomizedSearchCV
Fonte: Elaborada pela autora desta dissertação (2022)

O valor da acurácia do modelo otimizado com *RandomizedSearchCV* foi considerado pior que o modelo inicial (*Accuracy* = 99,56%), além do tempo de processamento elevado. Por isso, foi descartado treinamento randomizado e optou-se por tentar melhorar o modelo manualmente novamente.

Assim, o parâmetro *learning_rate*, que apresenta a taxa de aprendizagem do modelo, melhorou quando alterada para 0.02, o parâmetro *n_estimators*, que aponta a quantidade de árvores a serem rodadas no modelo, mostrou melhores resultados com 700 árvores. O parâmetro *max_depth*, que apresenta a profundidade máxima das árvores, obteve melhores resultados com a profundidade 4. O parâmetro alterado dos padrões foi *colsample_bynode*, que representa a fração das colunas a serem usadas pelos nós, melhorou ao diminuir para 0.75. Os demais hiperparâmetros testados estão descritos na forma de comentário na Figura 35, a qual traz o algoritmo com os testes realizados, já que na maior parte deles se manteve a configuração padrão do algoritmo por trazer os melhores resultados ou não o alterar. Esse treinamento levou em torno de 10 segundos para ser finalizado.

```
from pandas.core.common import random_state
from xgboost import XGBClassifier

# melhorando hiperparâmetros do modelo
# https://xgboost.readthedocs.io/en/latest/python/python_api.html#xgboost.XGBClassifier
xgb1 = XGBClassifier (learning_rate = 0.02,
                      n_estimators = 700, # n_estimators = quantidade de árvores que o modelo vai rodar - padrão é 100.
                      random_state = 42, #padrão é zero
                      max_depth = 4, # nós - padrão é 3, melhorou com 4 e piorou com 5 ou mais
                      min_child_weight = 1, # padrão é 1, não mudou valor ao aumentar ou diminuir
                      subsample = 1, # piorou usando outros valores, entao deixamos o padrão 1
                      colsample_bynode = 0.75, # melhorou em relação ao padrão 1, fração de colunas por nó
                      num_parallel_tree = 1, # não fez diferença aumentando o treinamento paralelo - deixamos padrão 1
                      colsample_bylevel = 1, # padrão é 1 - não alterou resultados - fração de colunas a serem usadaspo nível
                      gamma = 0, # padrão é 0 não mudou resultados - controla a poda - de 0 a infinito.
                      base_score = 0.5, # padrão 0.5 não mudou resultados - previsao inicial
                      max_delta_step = 1, # padrão 0, não mudou resultados - 1 a 10, deixa as atualizações + conservadoras
                      importance_type = 'weight', # padrão é gain - tipo de importância - não alterou resultado
                      reg_lambda = 1, # padrão 1 - não alterou resultado - raiz dos quadrados dos pesos
                      reg_alpha = 0, #padrão 0 - média dos pesos
                      booster = 'gbtree', #padrão gbtree- dart=mesmo resultado, mas treinamento +demorado, gblinear=resultado bem pior
                      n_jobs = -1,
                      scale_pos_weight = 1 #padrão 1
                      )

xgb1.fit(treino_x,
         treino_y,
         early_stopping_rounds=10,
         eval_set=[(teste_x, teste_y)])

xgb1_predict = xgb1.predict(teste_x)
acuracia_xgb1 = accuracy_score(teste_y, xgb1_predict)*100
print("A acurácia do modelo base usando XGBoost foi %.3f%%" % acuracia_xgb1)
```

Figura 35 – Treinamento com XGBoost – modelo ajustado manualmente
Fonte: Elaborada pela autora desta dissertação (2022)

Para o *DataFrame* df_durante_pandemia, repetiu-se a análise com a função *RandomizedSearchCV* (Figura 23) e observou-se resultado melhor que o modelo inicial, a acurácia (*Accuracy*) foi de 99,42% e o processamento dos treinamentos demorou três minutos.

A configuração dos melhores parâmetros e estimadores foi colsample_bynode =0.75, colsample_bytree = 0.3, gamma = 0.5, learning_rate = 0.03, max_depth = 5, min_child_weight = 3, n_estimators = 500 e subsample = 1, que estão representados na Figura 36 que também apresentou os melhores estimadores.

```
# Melhores parâmetros dos randomizados
xgb_rand2.best_params_

{'colsample_bynode': 0.75,
 'colsample_bytree': 0.3,
 'gamma': 0.5,
 'learning_rate': 0.03,
 'max_depth': 5,
 'min_child_weight': 3,
 'n_estimators': 500,
 'subsample': 1}

# Melhor Estimador
xgb_rand2.best_estimator_

XGBClassifier(base_score=0.5, booster='gbtree', callbacks=None,
              colsample_bylevel=1, colsample_bynode=0.75, colsample_bytree=0.3,
              early_stopping_rounds=None, enable_categorical=False,
              eval_metric=None, gamma=0.5, gpu_id=-1, grow_policy='depthwise',
              importance_type=None, interaction_constraints='',
              learning_rate=0.03, max_bin=256, max_cat_to_onehot=4,
              max_delta_step=0, max_depth=5, max_leaves=0, min_child_weight=3,
              missing=nan, monotone_constraints='()', n_estimators=500,
              n_jobs=0, num_parallel_tree=1, predictor='auto', random_state=0,
              reg_alpha=0, reg_lambda=1, ...)

print("A acurácia do RandomizedSearchCV: %.5f" % (xgb_rand2.best_score_))

A acurácia do RandomizedSearchCV: 0.99415
```

Figura 36 – Resultados com XGBoost – modelo ajustado com RandomizedSearchCV
Fonte: Elaborada pela autora desta dissertação (2022)

No entanto, ao repetir os mesmos ajustes (Figura 33) dos hiperparâmetros manualmente ao df_antes_pandemia, foram obtidos os resultados melhores que o modelo inicial, a melhoria se deve ao parâmetro reg_alpha = 1, em que o padrão é zero e representa a regularização em L1 em pesos. Porém, a acurácia foi pior que o modelo ajustado com *RandomizedSearchCV*. Os resultados das métricas foram: *Accuracy* = 99,37%, *Precision* = 96,67%, *Recall* = 97,88% e *F1-Score* = 92,06%.

O resultado da matriz de confusão também alterou para o df_durante_pandemia, em que TP = 29, FP = 4, FN = 1 e TN = 765. Com isso, repara-se que os TPs, verdadeiros positivos, melhorou em relação ao modelo inicial, acertando um resultado de concluir o curso a mais, diminuindo os FPs, falsos positivos de cinco para quatro.

A Tabela 32 mostra os resultados obtidos com *XGBoost* nos dois *DataFrames* durante o modelo inicial, ajustado com *RandomizedSearchCV* e manualmente. Os resultados em vermelho foram os piores, os em azul os que permaneceram os mesmos e os em verde foram os melhores resultados.

Tabela 32 – Métricas de avaliação do modelo ajustado

	df_antes_pandemia			df_durante_pandemia		
	Modelo Evasão	Modelo Ajustado com RandomizedSearchCV	Modelo Ajustado	Modelo Evasão	Modelo Ajustado com RandomizedSearchCV	Modelo Ajustado
Accuracy	99,56%	99,37%	99,56%	99,25%	99,42%	99,37%
Precision	98,75%		98,75%	96,55%		96,67%
Recall	96,34%		96,34%	84,85%		97,88%
F1-Score	97,53%		97,53%	90,32%		92,06%

Fonte: Elaborada pela autora desta dissertação (2022)

Depois de diversos testes, o melhor treinamento permaneceu com os mesmos valores do modelo de previsão inicial para o df_antes_pandemia e melhorou com o modelo ajustado para df_durante_pandemia, alterando o valor do parâmetro reg_alpha para 1, com isso, para o modelo ajustado final, manteve-se o algoritmo *XGBoost* com hiperparâmetros inseridos manualmente, em que o treinamento, além de trazer os melhores resultados, foi muito mais rápido (em torno de 10 segundos para o df_antes_pandemia e 11 segundos para o df_durante_pandemia). Com isso, a configuração final para o modelo de previsão de evasão escolar do IFSC resultou nos hiperparâmetros apresentados na Figura 37 para os dois *DataFrames*.

```
# melhorando hiperparâmetros do modelo
xgb2 = XGBClassifier (learning_rate = 0.02,
                      n_estimators = 700,
                      random_state = 42,
                      max_depth = 4,
                      colsample_bynode = 0.75,
                      reg_alpha = 1,
                      )

xgb2.fit(treino2_x,
         treino2_y,
         eval_set=[(teste2_x, teste2_y)])

xgb2_predict = xgb2.predict(teste2_x)
```

Figura 37 – Resultados dos hiperparâmetros do modelo final de previsão
Fonte: Elaborada pela autora desta dissertação (2022)

No próximo capítulo serão apresentadas as Considerações finais desta pesquisa, além de sugerir trabalhos futuros para este tema.

5 CONSIDERAÇÕES FINAIS E SUGESTÕES PARA TRABALHOS FUTUROS

O presente trabalho teve como propósito principal propor um modelo usando algoritmos de *Machine Learning* para prever a evasão escolar no Instituto Federal de Santa Catarina (IFSC). Para o desenvolvimento do modelo, foi utilizada a metodologia CRISP-DM, como modelo *baseline*, foi utilizado o algoritmo *DecisionTreeClassifier* e, para o desenvolvimento do modelo, os algoritmos *XGBClassifier* e *MLPClassifier*. Para a implementação do modelo, foi utilizada a ferramenta *Colaboratory* do *Google* e a linguagem de programação *Python*. O estudo levou em consideração dois conjuntos de dados separados, anteriormente à pandemia da Covid-19 (df_antes_pandemia) e durante a pandemia da Covid-19 (df_durante_pandemia), totalizando 7.759 estudantes únicos dos cursos de graduação (bacharel, licenciatura e tecnológico) do IFSC.

Após avaliar os fatores que impactam na evasão escolar na literatura existente, análise dos algoritmos *Decision Tree*, *MultiLayer Perceptron* e *XGBoost* e coletar os dados da instituição, foi realizada a fase de pré-processamento desses dados. O pré-processamento dos dados foi de extrema relevância para a desenvolvimento do modelo, pois a base de dados recebida da instituição possuía muitas informações faltantes, 1.046.213 no total em 160.271 linhas e 36 colunas (variáveis). Com isso, foram excluídos os dados de endereço dos estudantes, pois mais de 61% das informações eram faltantes, e as informações do endereço do campus, pois os dois endereços serviriam para criar a distância da residência do aluno até o campus, também foram excluídas linhas duplicadas do *DataFrame*. Após, deu-se início ao tratamento e à criação das variáveis, em que a maior parte das variáveis obtidas possuía dados faltantes, com isso, foram realizadas técnicas de imputações, as variáveis que eram do tipo categórico foram transformadas em numéricas, em algumas variáveis, foi realizada estatística descritiva para verificar *outliers* (dados muito discrepantes) e feito o tratamento dessas informações. Por fim, os dados foram agrupados em discentes únicos e o *DataFrame* final resultou em 7.759 linhas (estudantes) e 19 colunas (variáveis). Em seguida, foi construída a análise descritiva das informações, foi analisada a variação dos dados e foi dividido o *DataFrame* final em dois momentos: df_antes_pandemia que abarca os dados de 2017, 2018 e 2019 e o df_durante_pandemia que engloba os anos de 2020 e 2021. Para finalizar o pré-processamento, foi feita a normalização dos dados (deixar as variáveis em intervalos entre 0 e 1) e exportado os dados.

Para a criação do modelo de previsão da evasão escolar do IFSC, foram definidos os algoritmos *XGBClassifier* e *MLPClassifier* e aperfeiçoados por meio da utilização melhoria de

hiperparâmetros, depois comparados com o modelo *baseline*, usando *DecisionTreeClassifier,* como este foi o *baseline* não foi melhorado seus parâmetros.

Como métrica de avaliação dos modelos, foram utilizadas *Accuracy*, *Precision*, *Recall*, *F1-Score* e Matriz de Confusão. Os dois modelos (*XGBoost* e MLP) se mostraram melhores que o *baseline* das duas bases analisadas, porém o modelo *XGBoost* se mostrou muito superior. No df_antes_pandemia a métrica *F1_score* do algoritmo *XGBoost* foi de 97,53%; já o algoritmo MLP obteve *F1-Score* de 93,83%. No df_durante_pandemia, o algoritmo *XGBoost* apresentou *F1_Score* de 90,32%, já o algoritmo MLP obteve 80% de *F1_Score*. De forma geral, o algoritmo XGBoost apresentou-se melhor que o algoritmo *MultiLayer Perceptron*.

Por fim, foi analisada a importância das variáveis, em que, para o df_antes_pandemia, a variável que apresentou maior relevância foi n_disciplinas_concluidas, seguida de forma_ingresso, media_geral_disc, renda_pcf e campus. Do mesmo modo, para o *DataFrame* df_durante_pandemia, a variável mais importante foi idade_disc seguida de forma_ingresso, curso_disc, naturalidade_disc e media_geral_disc. Já as três piores variáveis foram as mesmas para o df_antes_pandemia e para o df_durante_pandemia, apenas alterando a ordem: raca_disc, tipo_aprendizagem e estado_civil.

Com os resultados da avaliação do modelo, foi possível construir um modelo ajustado sem utilizar as piores variáveis (raca_disc, estado_civil e tipo_aprendizagem) para o aprendizado, usando apenas o algoritmo com os melhores resultados na primeira rodada (*XGBoost*) para tentar afinar o modelo de evasão escolar. Seus resultados mostraram que a extração das piores variáveis não alterou a *performance* do modelo, ratificando a trivialidade destas. Do mesmo modo, o df_antes_pandemia mostrou que algoritimo *XGBoost* da primeira rodada utilizou bons hiperparâmetros e não foi possível melhorá-lo utilizando a função *RandomizedSearchCV*, ao invés disso, ele obteve resultados piores do que o da primeira rodada ajustado manualmente. Porém, para o df_durante_pandemia, conseguiu-se obter melhor resultado alterando o padrão do parâmetro reg_alpha para 1. Dessa forma, para o modelo ajustado final, manteve-se o algoritmo *XGBoost* com hiperparâmetros inseridos manualmente, alterando apenas o parâmetro reg_alpha, além dos demais do modelo inicial para o XGB (treinamento muito mais rápido) e as 16 variáveis (idade_disc, renda_pcf, sexo_disc, campus, turno_curso, curso_disc, naturalidade_disc, cidade_campus, forma_ingresso, origem_ensino_anterior, media_geral_disc, n_disciplinas_reprovadas, n_disciplinas_concluidas, idhm_municipio, ingresso_disc, semestre_atual), pois eles trouxeram os melhores resultados entre todos os testados.

Uma das principais limitações do modelo foi não ter sido possível coletar dados dos estudantes na forma de questionários, os quais poderiam colaborar com as variáveis mais importantes na visão dos estudantes que já passaram pelo processo de evasão ou estão em andamento na graduação, entendendo quais os motivos que os levaram a tal fato ou os que os levariam. Tsai *et al.* (2020), Musso, Hernández e Cascallar (2020) e Xiao e Yi (2020) sugerem a inclusão de variáveis relacionadas a aspectos comportamentais para aumentar a precisão dos modelos de previsão. Do mesmo modo, Tsai *et al.* (2020) e Deo *et al.* (2020) acreditam que aspectos relativos ao engajamento e à família são importantes para um estudante evadir ou não de um curso.

Outro limitante foi conseguir utilizar dados apenas do sistema acadêmico institucional (SIGAA) e do Sistema de Ingresso do IFSC, não podendo utilizar dados do Ambiente Virtual de Aprendizagem Institucional (*Moodle*) que traria diversas variáveis utilizadas por diferentes autores, como Zulfiker *et al.* (2020), Gamie, El-Seoud e Salama (2020), Gamie *et al.* (2019), Manzanares *et al.* (2018), Adejo e Connolly (2018) e Costa *et al.* (2017).

Da mesma forma, a quantidade de valores ausentes encontrados nas bases de dados limita resultados os mais próximos da realidade possível, mesmo quando se trabalha com boas técnicas de imputação de dados, além do tempo despendido no tratamento dos dados. Assim, se a instituição conseguir melhorar a coleta/manipulação de seus dados, é possível melhorar a qualidade e a precisão das previsões de evasão escolar. Dessa forma, a pesquisa contribuiu para que o Instituto Federal de Santa Catarina forme uma lista de variáveis/características importante para a avaliação, fazendo com que os estudantes permaneçam no curso de graduação até sua conclusão e possibilitando que a instituição foque em melhorar essas variáveis em seus sistemas acadêmicos.

5.1 TRABALHOS FUTUROS

Como trabalhos futuros, sugere-se aprofundar os estudos no algoritmo *XGBoost*, agregando melhor os hiperparâmetros para o algoritmo sem sofrer com o processamento extremamente lento que se mostrou em alguns testes (como foi a tentativa do uso da função *GridSearchCV* e da função *RandomizedSearchCV*). Assim como sugere-se explorar e implementar outras funcionalidades disponibilizadas pelo algoritmo, como é o caso da biblioteca *xgbfir* que fornece várias medidas para avaliar a importância das variáveis (HARRISON, 2020).

Segundo Yang *et al.* (2020), o número ideal de características para ser utilizado em um método de classificação é vital, dessa forma, seria interessante avaliar mais a fundo as variáveis utilizadas e fazer diferentes combinações, assim como feito em estudo de Gamie, El-Seoud e Salama (2020), incluindo e testando novas combinações com o propósito de encontrar formas de coletar dados da plataforma *Moodle*.

Com as informações obtidas para a base de estudantes para os cursos de graduação do IFSC, é possível desenvolver programas direcionados que abordem especificamente os fatores que são identificados em estudantes com a probabilidade de evasão, por isso, avaliar a base de dados após a pandemia, comparando os três momentos (antes, durante e depois), se faz necessário para identificar as possíveis diferenças e as mudanças relevantes aos efeitos da pandemia nos estudantes do IFSC. Por fim, após as melhorias no modelo focado no algoritmo *XGBoost*, testes e avaliações, propõe-se criar uma ferramenta mais acessível ao público da instituição para futura implantação.

REFERÊNCIAS

ADEJO, O. W.; CONNOLLY, T. Predicting Student Academic Performance Using Multi-Model Heterogeneous Ensemble Approach. **Journal of Applied Research in Higher Education**, [*s.l.*], 2018. DOI: 10.1108/JARHE-09-2017-0113.

ADEKITAN, Aderibigbe Israel; SALAU, Odunayo. The impact of engineering students' performance in the first three years on their graduation result using educational data mining. **Heliyum,** [*s.l.*], v. 5, n. 2, 2019. DOI: 10.1016/j.heliyon.2019.e01250.

ALVES-MAZZOTTI, Alda Judith; GEWANDSZNAJDER, Fernando. **O método nas ciências naturais e sociais:** Pesquisa Quantitativa e Qualitativa. 2. ed. São Paulo: Pioneira Thomson Learning. 1999. ISBN: 85-221-0133-7.

ANDIFES – ASSOCIAÇÃO NACIONAL DOS DIRIGENTES DAS INSTITUIÇÕES FEDERAIS DE ENSINO SUPERIOR. Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas. **Resumo do Relatório apresentado a ANDIFES, ABRUEM e SESu/MEC pela Comissão Especial**. [*S.l.*]: Andifes, 1996. p. 55-65.

ASHOUR, Sanaa. Analysis of the attrition phenomenon through the lens of university dropouts in the United Arab Emirates. **Journal of Applied Research in Higher Education,** [*s.l.*], v. 12, n. 2, p. 357-374, 2019. DOI: 10.1108/JARHE-05-2019-0110.

BARGMANN, Carina; THIELE, Lisa; KAUFFELD, Simone. Motivation matters: predicting students' career decidedness and intention to drop out after the first year in higher education. **Higher Education,** [*s.l.*], março de 2021. DOI: 10.1007/s10734-021-00707-6.

BEAN, John P. Dropouts and turnover: the synthesis and test of a causal model of student attrition. **Research in Higher Education,** [*s.l.*], v. 12, n. 2, p. 155-187, junho de 1980.

BEAULAC, Cédric; ROSENTHAL, Jeffrey S. Predicting university students' academic success and major using random forests. **Research in Higher Education,** [*s.l.*], v. 60, n. 7, p. 1.048-1.064, 2019. DOI: 10.1007/s11162-019-09546-y.

BRASIL. Ministério da Educação. **Anteprojeto de lei da reforma da educação superior.** 2005. Disponível em: http://portal.mec.gov.br/arquivos/pdf/acs_finalreforma 280705.pdf. Acesso em: 28 out. 2021.

BRASIL, Ministério da Educação. Secretaria de Educação Tecnológica. **Documento Orientador da Evasão e Retenção da Educação Profissional, Científica e Tecnológica**. 2014. Disponível em: http://portal.mec.gov.br/index.php?option=com_docman&view=download &alias=110401-documento-orientador-evasao-retencao-vfinal&category_slug=abril-2019-pdf&Itemid=30192. Acesso em: 10 nov. 2021.

BRASIL. **Lei n. 11.892 de 29 de dezembro de 2008**. Institui a Rede Federal de Educação Profissional, Científica e Tecnológica, cria os Institutos Federais de Educação, Ciência e Tecnologia, e dá outras providências. Brasília, DF: Presidência da República, 2008. Disponível em: http://www.plan alto.gov.br/ccivil_03/_ato2007-2010/2008/lei/l11892.htm. Acesso em: 10 nov. 2021.

BRASIL. **Lei n. 9.394 de 20 de dezembro de 1996**. Lei Geral de Diretrizes e Bases da Educação Nacional - LDB. Brasília, DF: Presidência da República, 1996. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l9394.htm. Acesso em: 10 nov. 2021.

BRASIL. **Plataforma Nilo Peçanha**. Brasília, DF: Ministério da Educação, [2021]. Disponível em: http://platafo rmanilo pecanha.mec.gov.br/. Acesso em: 4 out. 2021.

BRASIL. **Nota Técnica n. 282/SETEC/MEC, de 9 de julho de 2015.** Informa e orienta as instituições da rede federal sobre a construção dos planos estratégicos institucionais para a permanência e êxito dos estudantes. Brasília, DF: Ministério da Educação, 2015.

CASANOVA, Joana R. *et al*. Dimensionality and reliability of a screening instrument for students at-risk of dropout out from higher education. **Studies in Educational Evaluation,** [*s.l.*], v. 68, março de 2021. DOI: 10.1016/j.stueduc.2020.100957.

CHEN, Tianqi; GUESTRIN, Carlos. XGBoost: a scalable tree boosting system. **KDD16,** San Francisco, CA, USA, p. 13-17, august, 2016.

CHUI, K. *et al*. Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. **Computer in Human Behavior**, [*s.l.*], v. 107, p. 2.733-2.746, dezembro de 2020. DOI: 10.1016/j.chb.2018.06.032.

COSTA, Oberdan; GOUVEIA, Luis Borges. Modelos de retenção de estudantes: abordagens e perspectivas. **Revista Eletrônica de Administração – REAd**, [*s.l.*], v. 24, n. 3, p. 155-182. set.-dez. 2018. DOI: 10.1590/1413-2311.226.85489.

COSTA, Evandro B. *et al*. Joilson Evaluating the effectiveness of educational data mining techniques for early prediction of student's academic failure in introductory programming courses. **Computers in Human Behavior,** [*s.l.*], v. 73, p. 247-256, 2017. DOI: 10.1016/j.chb.2017.01.047.

DECISIONTREECLASSIFIER. **Documentação oficial sklearn para o algoritmo DecisionTreeClassifier**. [2022]. Disponível em: https://scikit-learn.org/stable/modules/generated/ sklearn.tree.DecisionTreeClassifier.html. Acesso em: 4 jun. 2022.

DELEN, Dursun. A comparative analysis of machine learning techniques for student retention management. **Decision Support Systems**, [*s.l.*], v. 49, n. 4, p. 498-506, 2010. DOI: 10.1016/j.dss.2010.06.003.

DEO, R. C. *et al*. Modern Artificial Intelligence Model Development for Undergraduate Student Performance Prediction: An Investigation on Engineering Mathematics Courses. **IEEE Access**, [*s.l.*], v. 8, p. 136697-136724, 2020. DOI: 10.1109/ACCESS.2020.3010938.

DOMINGOS, Pedro. O Algoritmo Mestre. Como a busca pelo algoritmo de machine learning definitivo recriará nosso mundo. **The Master Algorithm.** 1. ed. [*S.l.*]: Novac Editora Ltda. 2017. ISBN: 978-85-7522-542-4.

DURSO, Samuel de Oliveira. **Características do processo de evasão dos estudantes do curso de Ciências Contábeis de uma universidade pública brasileira.** 2015. 198p. Dissertação (Mestrado em Ciências Contábeis) – Universidade Federal de Minas Gerais, Belo Horizonte, MG, 2015.

EZZ, M.; ELSHENAWY, A. Adaptive Recommendation System Using Machine Learning Algorithms for Predicting Student's best Academic Program. **Education and Information Technologies,** [*s.l.*], v. 25, n. 4, p. 2.733-2.746, 2019. DOI: 10.1007/s10639-019-10049-7.

FERNÁNDEZ, Diego Buenaño; GIL, David; MORA, Sergio Luján. Application of machine learning in predicting performance for computer engineering students: a case study. **Sustainability**, [*s.l.*], v. 11, 2019. DOI: 10.3390/su11102833.

FLORES, Evandro Gomes. **Modelo de gestão do conhecimento para acompanhamento de tendência à evasão em cursos de graduação presencial.** 2017. 73p. Dissertação (Mestrado em Engenharia de Produção) – Universidade Federal de Santa Maria, Santa Maria, RS, 2017.

FORTUNATO, Maria Rocha; GONTIJO, Simone Braz Ferreira. Causas da Evasão Escolar na Percepção dos Estudantes Evadidos: o Caso do Curso de Licenciatura. **Rev. Bras. de Iniciação Científica (RBIC),** Itapetininga, v. 7, n. 1, p. 55-76, jan.-mar. 2020.

FREGONEZE, Gisleine Bartolomei *et al*. **Metodologia Científica**. 184 páginas. Londrina: Editora e Distribuidora Educacional S.A., 2014. 184p.

FREITAS, Francisco A. da S. *et al*. IoT Systems for School Dropout Prediction Using Machine Learning Techniques Based on Socioeconomic Data. **Electronics,** [*s.l.*], v. 9, n. 10, p. 1-14, 2020. DOI: 10.3390/electronics9101613.

GAMIE, E.; EL-SEOUD, M. S. A.; SALAMA, M.A. Comparative Analysis for Boosting Classifier in the Context of Higher Education. **International Journal of Emerging Technologies in Learning,** [*s.l.*], v. 15, n. 10, p. 16-26, 2020. DOI: 10.3991/ijet.v15i10.13663.

GAMIE, A. E. *et al*. Multi-dimensional analysis to predict students' grades in higher education. **International Journal of Emerging Technologies in Learning,** Cairo, Egypt. V. 14, n. 2, p. 4-15, 2019. DOI: DOI: 10.3991/ijet.v14i02.9905.

GEEKS FOR GEEKS. **XGBoost.** [2022]. Disponível em: https://www.geeksforgeeks.org/xgboost/. Acesso em: 30 maio 2022.

GÉRON, Aurélien. **Mãos à obra:** Aprendizado de máquina com Scikit-learn & TensorFlow – conceitos, ferramentas e técnicas para a construção de sistemas inteligentes. 1. ed. Tradução de Rafael Contatori de: Hands-on machine learning with Scikit-learn & TensorFlow. Rio de Janeiro: Atlas Books, 2019. ISBN: 978-85-508-0381-4.

GIL, Antonio Carlos. **Métodos e Técnicas de Pesquisa Social.** 6. ed. São Paulo: Atlas, 2008. ISBN: 978-85-224-5142-5.

GISMONDI, Hugo E. C.; HUIMAN, Luis V. U. **Multilayer neural networks for predicting academic dropout at the National University of Santa – Peru**. [*S.l.*]: IEEE Xplore, 2021. ISBN: 978-1-6654-0719-9.

GRUS, Joel. **Data science do zero:** primeiras regras com o python. 1. ed. Tradução de Welington Nascimento, de Data Science from scratch: first principles with python. Rio de Janeiro: Atlas Books, 2016. ISBN: 978-85-7608-998-8.

HARRISON, Matt. **Machine learning:** guia de referência rápida – trabalhando com dados estruturados em Python. 1. ed. São Paulo: Editora Novatec, 2020. ISBN: 978-85-7522-817-3.

HERBAUT, Estelle. Overcoming failure in higher education: social inequalities and compensatory advantage in dropout patterns. **Acta Sociologica,** [*s.l.*], v. 64, n. 4, 2020. DOI: 10.1177/0001699320920916.

HOFFMAN, Ivan Londero; NUNES, Raul Ceretta; MULLER, Felipe Martins. As informações do Censo da Educação Superior na Implementação da Gestão do Conhecimento Organizacional sobre Evasão. **Gestão & Produção**, São Carlos, v. 26, n. 2, 2016. DOI: 10.1590/0104-530X-2852-19.

HUNG, H. C. *et al*. Applying Educational Data Mining to Explore Student's Learning Patterns in the Flipped Learning Approach for Coding Education. **Symmetry**, [*s.l.*], v. 12, n. 2, 2020. DOI: 10.3390/sym12020213.

HUSSAIN, M. *et al*. Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores. **Hindawi – Computational Intelligence and Neuroscience**, [*s.l.*], v. 2.018, 2018. DOI: 10.1155/2018/6347186.

IATRELLIS, O. *et al*. A two-phase machine learning approach for predicting student outcomes. **Education and Information Technologies**, [*s.l.*], 2020. DOI: 10.1007/s10639-020-10260-x.

IEDMS – INTERNATIONAL EDUCATIONAL DATA MINING SOCIETY. [2021]. Disponível em: https://educationaldatamining.org/. Acesso em: 20 out. 2021.

IFSC – INSTITUTO FEDERAL DE SANTA CATARINA. **Histórico do Instituto Federal de Santa Catarina**: Linha do Tempo. [2021]. Disponível em: https://www.ifsc.edu.br/linha-do-tempo. Acesso em: 17 fev. 2021.

IFSC – INSTITUTO FEDERAL DE SANTA CATARINA. **Plano estratégico de permanência e êxito dos estudantes do IFSC:** Resolução CONSUP n. 23, de 21 de agosto de 2018. Disponível em: http://cs.ifsc.edu.br/portal/files/consup_resolucao23_2018_plano_de_permanencia_e_exito.pdf. Acesso em: 19 nov. 2021.

INEP – INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **Metodologia de indicadores e trajetória de curso**. Brasília, DF: Ministério da Educação, 2017. Disponível em: https://download.inep.gov.br/informacoes_estatisticas/indicadores_educacionais/2017/metodologia_indicadores_trajetoria_curso.pdf. Acesso em: 20 jan. 2021.

INEP – INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **Censo da educação superior 2019**. [2019]. Disponível em: https://download.inep.gov.br/educacao_superior/censo_superior/documentos/2020/Press_Kit_Censo_Superior_2019.pdf. Acesso em: 20 jan. 2021.

IZBICKI, Rafael; SANTOS, Tiago Mendonça dos. **Aprendizado de máquina**: uma abordagem estatística. 1. ed. São Carlos, SP: [s.n.], 2020. ISBN: 978-65-0002-410-4.

JUNG, Jisun; KIM, Yangson. Exploring regional and institutional factors of international students' dropout: the South Korea case. **Higher Education Quarterly,** [*s.l.*], v. 72, p. 141-159, 2017. DOI: 10.1111/hequ.12148.

KILLEN, Roy. **Programming and assessment for quality teaching and learning.** [*S.l.*]: Cengage Learning. 1. ed. 2005. ISBN: 0170122476.

KOVÁCS, Zsolt László. **Redes Neurais Artificiais**: fundamentos e aplicações. 4. ed. São Paulo: Editora Livraria da Física, 2006. ISBN: 978-85-8832-514-2.

KUMAR, V. Uday; KRISHNA, Azmira. Advanced prediction of performance of a student in an university using machine learning techniques. *In*: INTERNATIONAL CONFERENCE ON ELECTRONICS AND SUSTAINABLE COMMUNICATION SYSTEMS (ICESC). Agosto de 2020. **Anais** [...]. [*S.l*], agosto de 2020. DOI. 10.1109/ICESC48915.2020.9155557.

LIMA JÚNIOR, Paulo *et al*. Taxas longitudinais de retenção e evasão: uma metodologia para estudo da trajetória dos estudantes na educação superior. **Ensaio: Avaliação de Políticas Públicas na Educação,** Rio de Janeiro, v. 27, n. 102, p. 157-178, 2019. DOI: 10.1590/S0104-40362018002701431.

MANZANARES, M. C. S. *et al*. Detección del Alumno en Riesgo en Titulaciones de Ciências de la Salut: Aplicación de Técnicas de Learning Analytics. **European Journal of Investigation in Health, Psychology and Education**, [*s.l.*], 2018.

MARQUES, Felipe Tumenas. A volta aos estudos dos alunos evadidos do ensino superior brasileiro. **Caderno de Pesquisa**, São Paulo, v. 50, n. 178, p. 1.061-1.077, 2020. DOI: 10.1590/198053147158.

MCKINNEY, Wes. **Python para análise de dados**: tratamento de dados com Pandas, NumPy e IPython. 1. ed. São Paulo: Editora Novatec, 2018. ISBN: 978-85-7522-751-0.

MDUMA, Neema; KALEGELE, Khamisi; MACHUVE, Dina. Machine learning approach for reducing students dropout rates. **International Journal of Advanced Computer Research**, [*s.l.*], v. 9, n. 42, 2019. ISSN: 2249-7277. DOI: http://dx.doi.org/10.19101/IJACR.2018.839045. 2019.

MIA, M. *et al*. Registration status prediction of students using machine learning in the context of private university of Bangladesh. **International Journal of Innovative Technology and Exploring Engineering,** [*s.l.*], v. 9, n. 1, p. 2.594-2.600, 2019. DOI: 10.35940/ijitee.A5292.119119.

MITCHELL, Tom M. **Machine Learning.** 1. ed. [*S.l.*]: McGraw-Hill Education, 1997. ISBN: 978-00-7042-807-2.

MOHER, D. *et al*. **The PRISMA Group**. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. 2009. Disponível em: www.prisma-statement.org. Acesso em: 10 out. 2021.

MUÑIZ, Luis Rodriguez *et al.* Dropout and transfer paths: what are the risk profiles when analyzing university persistence with machine learning techniques? **Plos One,** [*s.l.*], v. 14, n. 6, 2019. DOI: 10.1371/journal.pone.0218796.

MUSSO, M.; HERNÁNDEZ, C.; CASCALLAR, E. Predicting Key Educational Outcomes in Academic Trajectories: a Machine-learning Approach. **Higher Education,** [*s.l.*], v. 80, n. 5, p. 875-894, 2020. DOI: 10.1007/s10734-020-00520-7.

NIKOLIC, Nikola; GRLJEVIC, Olivera; KOVACEVIC, Aleksandar. Aspect-based sentiment analysis of reviews in the domain of Higher. **Electronic Library**, [*s.l.*], v. 38, n. 1, p. 44-64, 2020. DOI: 10.1108/EL-06-2019-0140.

NIEMBA, Armando. O abandono dos estudantes no ensino superior: o modelo de Vicent Tinto. **Revista Amazônica**, LAPESAM/GMPEPPE/UFAM/CNPq, v. XIII, n. 1, p. 195-211, jan-jun, 2021.

PÁSCOA, Mariana I. F. **Os desafios de machine learning**: aplicações no mercado financeiro. 2018. 50p. Dissertação (Mestrado de Economia Industrial) – Universidade de Coimbra, Portugal, 2018.

PÉREZ, Alexis Matheu *et al*. Prediction model of first-year student desertion at Universidad Bernardo O'Higgins (UBO). **Educação e Pesquisa,** São Paulo, v. 44, 2018. DOI: 10.1590/S1678-4634201844172094.

PRESTES, Emília Maria da Trindade; FIALHO, Marília Gabriela Duarte. Evasão na educação superior e gestão institucional: o caso da Universidade Federal da Paraíba. Ensaio: **Avaliação de Políticas Públicas na Educação,** [*s.l.*], v. 26, n. 100, p. 869-889, 2018.

PRIM, Alexandre Luis; FÁVERO, Jéferson Deleon. Motivos da Evasão Escolar nos Cursos de Ensino Superior de uma Faculdade na Cidade de Blumenau. **Tecnologias para Competitividade Industrial,** Florianópolis, n. Especial Educação, p. 53-72, 2013.

RANDOM. Search Cros Validation. **Documentação Oficial Sklearn RandomizedSearchCV**. [2022]. Disponível em https://scikit-

learn.org/stable/modules/generated/sklearn.model_selection. RandomizedSearchCV.html. Acesso em: 1º jul. 2022.

SALES JÚNIOR, Jaime Souza. **Uma análise estatística dos fatores de evasão e permanência de estudantes de graduação presencial da UFES.** 2013. 111p. Dissertação (Mestrado Profissional de Gestão Pública) – Universidade Federal do Espírito Santo, Vitória, 2013.

SANTA CATARINA. **Índice de desenvolvimento humano dos municípios de Santa Catarina do ano de 2010.** [2010]. Disponível em: https://drive.google.com/file/d/1EZnkk3m1dDRYwtGe0S0WFQB 9pvjZ7gDN/view. Acesso em: 18 mar. 2022.

SANTOS, Hellen Geremias dos. **Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina.** 2018. 206p. Tese (Doutorado de Ciências) – Universidade de São Paulo, São Paulo, 2018.

SAUNDERS, Mark; LEWIS, Philip; THORNHILL, Adrian. **Research methods for business students.** 8. ed. [*S.l.*]: Pearson, 2019. ISBN: 978-1-292-20878-7.

SCHMITT, Jeovani *et al*. WWH-dropout scale: when, why and how to measure propensity to drop out of undergraduate courses. **Journal of Applied Research in Higher Education,** [*s.l.*], v. 13, n. 2, p. 540-560, 2021.

SHIRASU, Maitê Rimekká; ALBUQUERQUE, Ronaldo de. Determinantes da Evasão e Repetência Escolar. *In*: XLIII ENCONTRO NACIONAL DE ECONOMIA. ANPEC-Associação Nacional dos Centros de Pós-Graduação em Economia, 2016. **Anais** [...]. [*S.l.*], 2016.

SILVA, Fernanda Cristina da; CABRAL, Thiago Luiz de Oliveira; PACHECO, Andressa Sasaki Vasques. Evasão ou permanência? Modelos preditivos para a gestão do ensino superior. **AAPE – Arquivos Analíticos de Políticas Educativas**, [*s.l.*], v. 28, n. 149, 2020.

SILVA, Andréa Ferreira da; ALMEIDA, Aléssio Tony Cavalcanti de; RAMALHO, Hilton Martins de Brito. Predição do risco de reprovação no ensino superior usando algoritmos de machine learning. **Teoria e Prática em Administração,** [*s.l.*], v. 10, n. 2, p. 58-80, jul-dez., 2020. DOI. 10.21714/2238-104X2020v10i2-51124.

SILVA FILHO, Roberto Leal Lobo *et al*. A evasão no ensino superior brasileiro. **Cadernos de Pesquisa,** [*s.l.*], v. 37, n. 132, p. 641-659, set.-dez., 2007.

SKLEARN. **Documentação oficial do método de pré-processamento MinMaxScaler**. [2022]. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing. MinMaxScaler.html. Acesso em: 13 abr. 2022.

SPADY, William G. Dropouts from higher education: an interdisciplinary review and synthesis. **Interchange,** [*s.l.*], p. 64-85, abril de 1970.

SPADY, William G. Dropouts from higher education: toward an empirical model. **Interchange,** [*s.l.*], p. 38-62, setembro de 1971.

SUHARJITO, Nindhia Hutagaol. Predictive modelling of student dropout using ensemble classifier method in higher education. **Advances in Science, Technology and Systems Journal**, [*s.l.*], v. 4, n. 4, p. 206-211, 2019. DOI: 10.25046/aj040425.

SULTANA, Sara; KHAN, Sharifullah; ABBAS, Muhammad A. Predicting performance of electrical engineering students using cognitive and non-cognitive features for identification of

potential dropouts. **International Journal of Electrical Engineering Education,** [*s.l.*], v. 54, p. 105-118, 2017. DOI: 10.1177/0020720916688484.

TERENCE, Ana Cláudia Fernandes; ESCRIVÃO FILHO, Edmundo. Abordagem quantitativa, qualitativa e a utilização da pesquisa-ação nos estudos organizacionais. *In*: XXVI ENCONTRO NACIONAL DE ENGENHARIA DE PRODUÇÃO – ENEGEP, v. 9, Fortaleza, CE, 2006. **Anais** [...]. Fortaleza, CE, 2006.

TINTO, Vicent. Classrooms as communities: exploring the educational character of student persistence. **The Journal of Higher Education**, London, v. 68, n. 6, p. 599-624, Nov.-Dec. 1997.

TINTO, Vicent. Dropout from higher education: a theoretical synthesis of recent research. **Review of Educational Research**, Washington, DC, v. 45, n. 1, p. 89-125, Mar. 1975.

TINTO, Vicent. **Leaving college**: rethinking the causes and cures of student attrition. 2. ed. Chicago: University of Chicago Press, 1993.

TSAI, S. *et al*. Precision education with statistical learning and deep learning: a case study in Taiwan. **International Journal of Educational Technology in Higher Education**, [*s.l.*], v. 17, 2020. DOI: 10.1186/s41239-020-00186-2.

VIDHYA, R.; VADIVU, G. Towards developing an ensemble based two-level student classification model (ESCM) using advanced learning patterns and analytics. **Journal of Ambient Intelligence and Humanized Computing**, [*s.l.*], v. 12, p. 7.095-7.105, 2020. DOI: 10.1007/s12652-020-02375-3.

WOTAIFI, T.; AL-SHAMERY, E. Mining of Completion Rate of Higher Education Based on Fuzzy Feature Selection Model and Machine Learning Techniques. **International Journal of Recent Technology and Engineering,** [*s.l.*], November, 2019. DOI: 10.31219/osf.io/wjbfk.

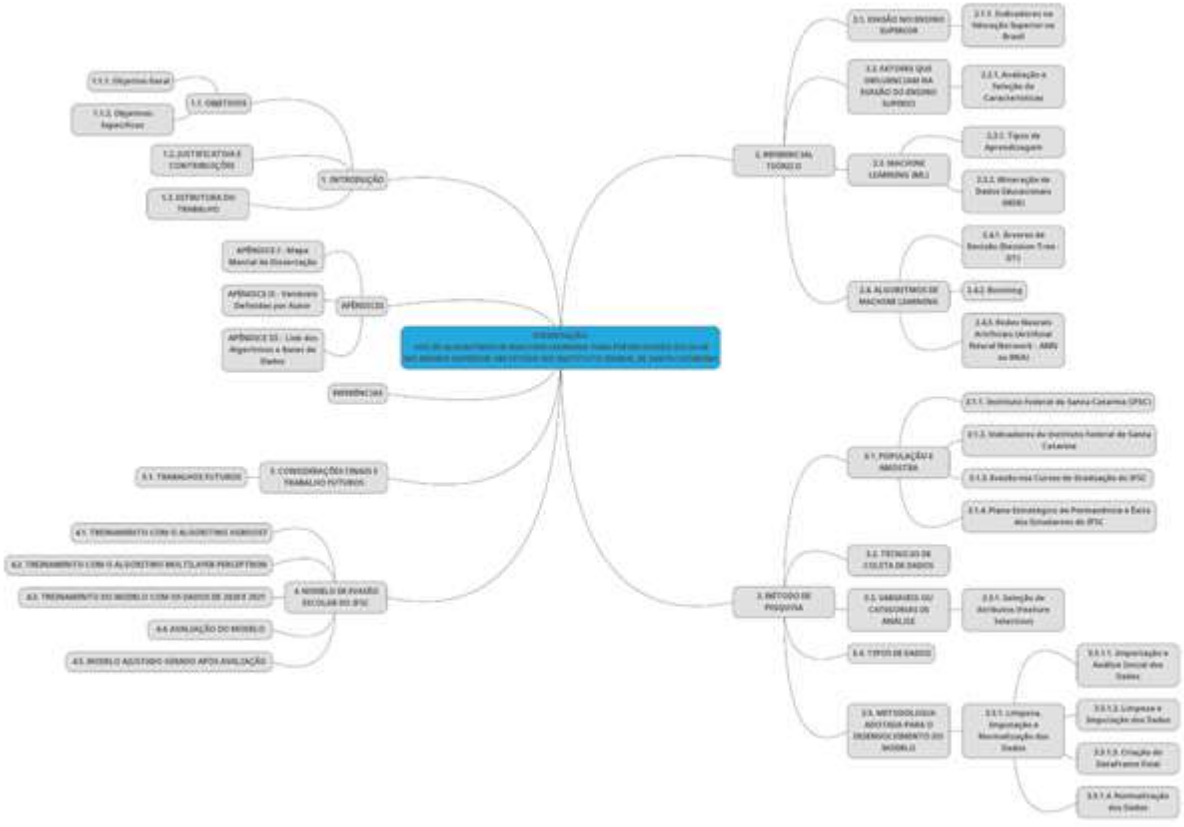
YANG, Y. *et al.* Predicting course achievement of university students based on their procrastination behaviour on Moodle. **Soft Computing**, Germany, v. 24, p. 18.777-18.793, Jul., 2020. DOI: doi.org/10.1007/s00500-020-05110-4.

XGBOOST. **Documentação XGBoost**. [2022]. Disponível em: https://xgboost.readthedocs.io/en/ latest/. Acesso em: 5 jan. 2022.

XIAO, M.; YI, H. Research on Adaptive Learning Prediction Based on XAPI. **International Journal of Information and Education Technology**, [*s.l.*], v. 10, n. 9, p. 679-684, 2020. DOI: 10.1002/cae.22235.

ZULFIKER, M. S. *et al*. Predicting student's performance of the private universities of Bangladesh using machine learning approaches. **International Journal of Advanced Computer Science and Applications,** [*s.l.*], v. 11, n. 3, p. 672-679, 2020. DOI: 10.14569/IJACSA.2020.0110383.

APÊNDICE A – MAPA MENTAL DA DISSERTAÇÃO



APÊNDICE B – VARIÁVEIS DEFINIDAS POR AUTOR

Autor	Variáveis
Costa *et al.* (2017)	Idade
	Sexo
	Estado civil
	Cidade
	Renda
	Matrícula do aluno
	Período
	Aula
	Semestre
	Campus
	frequência de acesso do aluno no sistema (para EaD)
	participação no fórum de discussões (para EaD)
	quantidade de arquivos recebidos e visualizados (para EaD)
	uso de ferramentas educacionais fornecidas pelo sistema (blog, glossário, quiz, wiki, mensagem) (para EaD)
	quantidade de exercícios realizados pelo aluno (para presencial)
	número de exercícios corretos (para presencial)
	ano de matrícula no curso
	status da disciplina
	desempenho do aluno nas atividades semanais
	desempenho do aluno nas provas
Adejo e Connolly (2018)	*Status* econômico dos pais
	horas de trabalho
	qualificação de ingresso
	horas média de estudo
	apoio familiar
	satisfação no curso
	impacto da tecnologia
	tipo de aprendizagem
	estado de saúde
	primeira universidade
	Adaptação
	apoio universitário
	conhecimento prévio do curso
	Idade
	Sexo
	Etnia
	localização de moradia
	Campus
	forma de entrada na universidade

Autor	Variáveis
	Qualificação
	Deficiência
	tempo total de *login* do estudante em plataforma de aprendizado
	número de recursos visualizados em plataforma
	número de tentativas de testes enviados em plataforma
	número de fóruns visualizados em plataforma
	número de discussões em fóruns lidas ou visualizadas em plataforma
Muñiz *et al*. (2019)	dados de identificação
	Sexo
	local de nascimento
	Nacionalidade
	Deficiência
	tamanho da família
	qualificação dos pais e ocupações atuais
	nota média do ensino médio
	pontuação do exame de admissão universitária
	idade quando admitido
	data da primeira matrícula
	prioridades indicadas no aplicativo de admissão do curso
	área do conhecimento correspondente ao curso do aluno
	número de créditos inscritos
	créditos passados
	pontuação média
	Bolsa
	situação acadêmica atual destino de transferência (quando houver)
	estado civil
	nível de renda
	tipo de moradia durante o curso
	motivação para escolha do curso e universidade
	participação em atividades de boas-vindas para calouros e opinião deles
	tempo gasto com estudo
	trabalho e trabalhos domésticos
	avaliação dos requisitos do programa de satisfação com pontuações
	avaliação das relações pessoais
	intenção de abandono e razões
	satisfação com a universidade
	(se caso o aluno desistiu) a situação atual e satisfação com os resultados de sua decisão
Adekitan e Salau (2019)	notas do ensino médio
	nível de participação das aulas
	Assiduidade
	notas intermediárias
	relatórios de laboratórios

Autor	Variáveis
	notas de tarefas de casa
	pontuação de seminários
	conclusão de tarefas
	notas gerais
Suharjito (2019)	média cumulativa de notas
	avaliação interna
	avaliação externa
	atividades extracurriculares
	histórico do ensino médio
	atividades sociais
	Idade
	restrições financeiras
	ausência do aluno
	influência dos pais
	oportunidade de emprego
	estado civil
	Sexo
Beaulac e Rosenthal (2019)	ID do aluno
	nome do curso
	departamento do curso
	Semestre
	valor do crédito do curso
	nota obtida pelo aluno
Ezz e Elshenawy (2019)	notas de todas as atividades do curso (média de pontuações, notas de exames orais, notas provas práticas, se existirem)
	nota final
	status final do aluno
	total de alunos
	notas específicas
	notas totais
Silva *et al.* (2020a)	Sexo
	Cor
	estado civil
	UF residência
	UF polo
	reside cidade polo?
	categoria de ingresso
	renda familiar
	tamanho da família
	tempo de deslocamento até o polo
	onde estudo maior parte do ensino médio?
	experiência no ensino superior (nunca ingressou/já concluiu/já concluiu/ingressou, mas não concluiu, está cursando)

Autor	Variáveis
	experiência no EaD
	frequência de uso do computador
	local de acesso à internet
	tipo de conexão à internet
	nível de conhecimento para uso do computador e internet
Freitas *et al.* (2020)	Sexo
	Raça
	origem do ensino médio
	distância da instituição
	renda familiar
	índice de desenvolvimento humano por município
Silva *et al.* (2020b)	nota do vestibular total
	nota do vestibular total em matemática
	Casado
	Migrante
	Raça
	Sexo
	idade ingresso
	Cotista
	período de ingresso
	forma de ingresso
	tempo de graduação do docente
	docente com doutorado
	publicação no ano do docente
	docente estrangeiro
	docente com dedicação exclusiva
	sexo docente
	local do campus
	UF curso
	UF centro
	Turno
	carga horária turma
	média nota vestibular turma
	média nota vestibular matemática turma
	taxa de cotista
Zulfiker *et al.* (2020)	atendimentos de um aluno
	aluno retomou um tópico
	durante o semestre, o aluno deve ter 3 questionários em um determinado curso
	aluno respondeu todos os questionários
	média das notas obtidas nos questionários
	notas obtidas no exame do meio do semestre
	aluno encaminhou as tarefas

Autor	Variáveis
	aluno executou as apresentações
	nota final do aluno após o exame final
Casanova *et al.* (2021)	Sexo
	Idade
	escolaridade dos pais (educação fundamental, básica ou ensino superior)
	formação acadêmica
	opções vocacionais
	curso matriculado
	curso de primeira opção
	expectativa de finalização do curso

APÊNDICE C – *LINK* DOS ALGORITMOS E DAS BASES DE DADOS

Pré-processamento dos dados:
https://colab.research.google.com/drive/15Nx3pyPWI3uXBZCLQD5rR6nHRBmZNhoR?usp=sharing

Modelo de Evasão Escolar IFSC:
https://colab.research.google.com/drive/1S7p1nL4uAVlw6LBTnM8BozHmDAlc1c4p?usp=sharing

Modelo de Evasão Ajustado com XGboost:
https://colab.research.google.com/drive/12yWlPjzmH1a62kVHoTIjtrKd7gnyx2d6?usp=sharing

Base de Dados Pré-Processada sem Normalização e sem Identificação dos Estudantes – df_antes_pandemia:
https://drive.google.com/file/d/12RuqT8u6FHpUjtfpJT-s-W8eAvqtrh01/view?usp=sharing

Base de Dados Pré-Processada sem Normalização e sem Identificação dos Estudantes – df_durante_pandemia:
https://drive.google.com/file/d/1uR_IG0AiRdFDoAOmcmHjRRk6S5SHlc-d/view?usp=sharing

Base de Dados Normalizada - df_antes_pandemia:
https://drive.google.com/file/d/1prTUJE4rKFp1NyXhuTFF08M15zQMcK01/view?usp=sharing

Base de Dados Normalizada - df_durante_pandemia:
https://drive.google.com/file/d/1ZPRjqkwJRtN4L2pk3cvX7RBYibPo4PTc/view?usp=sharing

Bases de Dados originais recebidas da instituição e usada no algoritmo de pré-processamento: Estas bases de dados não podem ser disponibilizadas devido a Lei Geral de Proteção de Dados Pessoais – LGPD por conter dados pessoais e que identificam o estudante, ficando sob cuidados do autor da pesquisa por 1 ano após a data de defesa desta dissertação, devendo ser excluída após este prazo conforme acordado com a instituição.