



**UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
UNIVERSIDADE DO ESTADO DO RIO GRANDE DO NORTE
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA
COMPUTAÇÃO**



FERNANDA FERREIRA DO NASCIMENTO

**IMPLEMENTAÇÃO E AVALIAÇÃO DE DIFERENTES
MODELOS DE MACHINE LEARNING APLICADOS À
PREDIÇÃO DE ESTUDANTES EM RISCO DE EVASÃO
ESTUDANTIL EM DIFERENTES CURSOS DO ENSINO
SUPERIOR.**

MOSSORÓ

2022

FERNANDA FERREIRA DO NASCIMENTO

**IMPLEMENTAÇÃO E AVALIAÇÃO DE DIFERENTES
MODELOS DE MACHINE LEARNING APLICADOS À
PREDIÇÃO DE ESTUDANTES EM RISCO DE EVASÃO
ESTUDANTIL EM DIFERENTES CURSOS DO ENSINO
SUPERIOR.**

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação - associação ampla entre a Universidade do Estado do Rio Grande do Norte e a Universidade Federal Rural do Semi-Árido, para a obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Paulo Gabriel Gadelha Queiroz-

Coorientadora: Profa. Dra. Angélica Félix de Castro- UFERSA

MOSSORÓ

2022

© Todos os direitos estão reservados a Universidade Federal Rural do Semi-Árido. O conteúdo desta obra é de inteira responsabilidade do (a) autor (a), sendo o mesmo, passível de sanções administrativas ou penais, caso sejam infringidas as leis que regulamentam a Propriedade Intelectual, respectivamente, Patentes: Lei nº 9.279/1996 e Direitos Autorais: Lei nº 9.610/1998. O conteúdo desta obra tomar-se-á de domínio público após a data de defesa e homologação da sua respectiva ata. A mesma poderá servir de base literária para novas pesquisas, desde que a obra e seu (a) respectivo (a) autor (a) sejam devidamente citados e mencionados os seus créditos bibliográficos.

d244i do Nascimento, Fernanda Ferreira .
IMPLEMENTAÇÃO E AVALIAÇÃO DE DIFERENTES
MODELOS DE MACHINE LEARNING APLICADOS À PREDIÇÃO
DE ESTUDANTES EM RISCO DE EVASÃO ESTUDANTIL EM
DIFERENTES CURSOS DO ENSINO SUPERIOR / Fernanda
Ferreira do Nascimento. - 2022.
145 f. : il.

Orientador: Paulo Gabriel Gadelha Queiroz.
Coorientador: Angélica Félix de Castro.
Dissertação (Mestrado) - Universidade Federal
Rural do Semi-árido, Programa de Pós-graduação em
Ciência da Computação, 2022.

1. Evasão Estudantil. 2. Modelos de
classificação. 3. Métricas de avaliação de
algoritmos. 4. Machine Learning. I. Queiroz,
Paulo Gabriel Gadelha , orient. II. de Castro,
Angélica Félix, co-orient. III. Título.

Ficha catalográfica elaborada por sistema gerador automático em conformidade
com AACR2 e os dados fornecidos pelo(a) autor(a).
Biblioteca Campus Mossoró / Setor de Informação e Referência
Bibliotecária: Keina Cristina Santos Sousa e Silva
CRB: 15/120

O serviço de Geração Automática de Ficha Catalográfica para Trabalhos de Conclusão de Curso (TCC's) foi desenvolvido pelo Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (USP) e gentilmente cedido para o Sistema de Bibliotecas da Universidade Federal Rural do Semi-Árido (SISBI-UFERSA), sendo customizado pela Superintendência de Tecnologia da Informação e Comunicação (SUTIC) sob orientação dos bibliotecários da instituição para ser adaptado às necessidades dos alunos dos Cursos de Graduação e Programas de Pós-Graduação da Universidade.

FERNANDA FERREIRA DO NASCIMENTO

**IMPLEMENTAÇÃO E AVALIAÇÃO DE DIFERENTES MODELOS DE
MACHINE LEARNING APLICADOS À PREDIÇÃO DE ESTUDANTES
EM RISCO DE EVASÃO ESTUDANTIL EM DIFERENTES CURSOS
DO ENSINO SUPERIOR**

Dissertação apresentada ao Programa de
Pós-Graduação em Ciência da
Computação para a obtenção do título
de Mestra em Ciência da Computação.

Aprovada em: 21/12/2022

BANCA EXAMINADORA

Documento assinado digitalmente
 PAULO GABRIEL GADELHA QUEIROZ
Data: 20/03/2023 10:03:52-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Paulo Gabriel Gadelha Queiroz-
Orientador e Presidente - Universidade Federal Rural do Semi-Árido - UFERSA

Documento assinado digitalmente
 ANGELICA FELIX DE CASTRO
Data: 20/03/2023 14:42:48-0300
Verifique em <https://validar.iti.gov.br>

Profa. Dra. Angélica Félix de Castro- UFERSA
Primeiro Membro

Documento assinado digitalmente
 ARAKEN DE MEDEIROS SANTOS
Data: 20/03/2023 14:58:43-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Araken de Medeiros Santos-
Membro Interno - Universidade Federal Rural do
Semi-Árido – UFERSA

Documento assinado digitalmente
 KAMILA RIOS DA HORA RODRIGUES
Data: 27/03/2023 08:56:48-0300
Verifique em <https://validar.iti.gov.br>

Profa. Dra. Kamila Rios da Hora Rodrigues-
Membro Externo - Universidade de São Paulo - USP

Aos meus familiares, que agora terão a primeira
mestra na família. Em especial à minha filha,
Raíssa Gabrielly, que sempre divide a minha
atenção entre ela e os estudos.

AGRADECIMENTOS

Há alguns anos este momento parecia um sonho impossível. José Maria Amaranto disse que: "*O sonho sonhado sozinho é um sonho solitário. O sonho sonhado junto pode ser, um sonho realizável ou discutível, para que se torne realidade.*"

Em primeiro lugar, agradeço a Deus por ter me concedido a oportunidade, coragem, fé, sabedoria e resiliência necessária para enfrentar as dificuldades encontradas ao longo dessa jornada.

À minha família, que sempre acreditou em mim, me apoiou e nunca me deixou desistir dos meus sonhos, em especial ao meu inesquecível avô Faé, *in memória*, sinto a presença e alegria com minhas conquistas a todo instante.

Ao meu orientador, Paulo Gabriel, que me acolheu desde o primeiro contato, sempre teve paciência comigo e sempre me ensinou pelo exemplo, principalmente a ser um professor humano. Isso fez e faz diferença na minha atuação docente. Palavras são poucas para expressar toda a minha admiração e gratidão.

À minha coorientadora Angélica Félix, por todos os ensinamentos, compreensão, companheirismo e alegria. Uma pessoa incrível e um exemplo de mulher forte e perseverante.

Ao Grupo de Pesquisa na área de Banco de Dados Não-Convencionais e Mineração de Dados (carinhosamente, grupo evasão escolar) que me receberam e que compartilharam comigo momentos de aprendizagens e descontração.

Ao querido Aislan Freitas, que não somente me ajudou a compreender o universo da evasão estudantil e dos dados, literalmente com a mão na massa, mas também não permitiu que eu desistisse quando estive em meus piores momentos.

Ao amigo e colega de grupo, Lucas Cesar, que contribuiu do início ao fim desta pesquisa e também foi um grande ombro amigo.

Aos companheiros colegas de mestrado que sempre colaboraram uns com os outros: "ninguém solta a mão de ninguém."

Por fim, agradeço a todos que, mesmo sem saber, colaboraram para o meu crescimento acadêmico, como estudante, pesquisadora e professora. Todos fizeram parte dessa conquista.

EPÍGRAFE

“A ciência nunca resolve um problema sem criar pelo menos outros dez”.

(George Bernard Shaw.)

RESUMO

Um dos maiores desafios da educação no Brasil é a evasão estudantil. A evasão estudantil é o ato de deixar de frequentar as aulas, abandonar o ensino em decorrência de um ou mais fatores. Esse problema afeta a qualidade de ensino, impactando os âmbitos econômicos e ambientais, refletindo negativamente no desenvolvimento social. Identificar as causas da evasão estudantil, bem como detectar previamente os perfis de estudantes considerados possíveis evasores em todos os níveis de ensino, tornou-se um tema de grande interesse e tem desencadeado pesquisas nacionais e internacionais. Para tanto, objetiva-se com este trabalho, desenvolver e avaliar modelos de *Machine Learning* capazes de realizar a predição, de forma precoce, de estudantes com alto risco de evasão estudantil, nos cursos de Ciência e Tecnologia, Engenharia Agrícola e Ambiental e Biotecnologia da Universidade Federal Rural do Semi-Árido. A abordagem foi validada, utilizando dados de ex-alunos dos cursos superiores de Ciência e Tecnologia, Engenharia Agrícola e Ambiental e Biotecnologia da Universidade Federal Rural do Semi-Árido. Neste trabalho foram utilizadas técnicas de classificação em modelos de *Machine Learning*. Na validação da abordagem foram empregados sete algoritmos: *K-Nearest Neighbor*, *Naive Bayes*, *Decision Tree*, *Random Forest*, *Perceptron*, *Multilayer perceptron* e *Support vector machine*, sendo que o modelo *Support vector machine* obteve os melhores resultados dentre as métricas: *Accuracy*, *Precision*, *Recall* e *F1 Score*. Em relação aos resultados, observou-se que: estudantes que não ingressaram na sua primeira opção de curso, têm uma maior probabilidade de evadir-se; estudantes que não possuem um conhecimento prévio na área do seu curso, têm uma maior probabilidade de evadir-se; o maior índice de evasão ocorre no primeiro ano de curso; estudantes que reprovam em uma ou mais disciplinas no primeiro ano de curso, têm uma maior probabilidade de evadir-se; estudantes do gênero feminino têm uma maior probabilidade de evadir-se; o mesmo modelo de *Machine Learning* pode ser aplicado para grupos de estudantes de diferentes cursos.

Palavras-chave: Evasão Estudantil; Machine Learning; Modelos de Classificação; Métricas de Avaliação de algoritmos.

ABSTRACT

One of the biggest challenges of education in Brazil is student dropout. Student evasion is the act of failing to attend classes, abandoning teaching as a result of one or more factors. This problem affects the quality of education, impacting the economic and environmental spheres, reflecting negatively on social development. Identifying the causes of student dropout, as well as previously detecting the profiles of students considered possible dropouts at all levels of education, has become a topic of great interest and has triggered national and international research. Therefore, the objective of this work is to develop and evaluate Machine Learning models capable of predicting, at an early stage, students at high risk of dropping out, in Science and Technology, Agricultural Engineering and Environment and Biotechnology at UFERSA. The approach has been validated, using data from former students of the Science and Technology, Agricultural and Environmental Engineering and Biotechnology courses at the Universidade Federal Rural do Semi-Árido. In this work, classification techniques were used in Machine Learning models. In the validation of the approach, seven Machine Learning algorithms were used: K-Nearest Neighbor, Naive Bayes, Decision Tree, Random Forest, Perceptron, Multilayer perceptron, and the Support vector machine model obtained the best results among the metrics Accuracy, Precision, Recall and F1 Score. Regarding the results, it was observed that: students who did not enter their first course option are more likely to drop out; students who do not have prior knowledge in the area of their course are more likely to drop out; the highest dropout rate occurs in the first year of the course; students who fail one or more subjects in the first year of the course are more likely to drop out; female students are more likely to drop out; the same Machine Learning model can be applied to groups of students from different courses.

Keywords: Student Dropout; Machine Learning; Classification Models; Algorithm Evaluation Metrics.

LISTA DE FIGURAS

Figura 1 – Evolução média dos indicadores de trajetória dos ingressantes de 2011 em cursos de graduação - Brasil - 2011-2020	19
Figura 2 – Comparativo de Concluintes e Desistentes da UFERSA por semestre .	20
Figura 3 – Quantitativo de abandono de cursos superiores da UFERSA por semestre	21
Figura 4 – Cursos da UFERSA com maior taxa de evasão	23
Figura 5 – Representação da metodologia do Design Science Research	27
Figura 6 – Etapas do CRISP-DM	29
Figura 7 – Gráfico de importância de cada feature	35
Figura 8 – Fluxo do treinamento com k-Fold	36
Figura 9 – Representação do funcionamento de um algoritmo de classificação . .	43
Figura 10 – Exemplo de modelo de regressão	43
Figura 11 – Sequencia do algoritmo KNN	45
Figura 12 – Decision Tree	46
Figura 13 – Perceptron	47
Figura 14 – MLP	48
Figura 15 – SVM	49
Figura 16 – Métricas para avaliações de modelos de ML baseados em técnicas de classificação	51
Figura 17 – Etapas do MSL	56
Figura 18 – Quantidade de trabalhos avaliados em cada etapa do mapeamento . . .	62
Figura 19 – Ferramentas de MD	65
Figura 20 – Técnicas de MD	68
Figura 21 – Principais algoritmos utilizados para MD	69
Figura 22 – Fatores indutores investigados nas causas da evasão estudantil	70
Figura 23 – Principais métricas de avaliação de modelos de ML	72
Figura 24 – <i>Ranking</i> das obras selecionadas com base em critérios de qualidade. .	74
Figura 25 – Situação em relação ao curso	80
Figura 26 – Aproveitamento acadêmico do curso de CeT	82
Figura 27 – Aproveitamento acadêmico do curso de Engenharia Agrícola e Ambiental e Biotecnologia	82
Figura 28 – Renda dos estudantes de Ciência e Tecnologia	83

Figura 29 – Renda dos estudantes de Engenharia Agrícola e Ambiental e Biotecnologia	84
Figura 30 – Escolaridade dos pais dos estudantes de CeT	86
Figura 31 – Escolaridade dos pais dos estudantes de Engenharia Agrícola e Ambiental e Biotecnologia	86
Figura 32 – Escolaridade das mães dos estudantes de Ciência e Tecnologia	87
Figura 33 – Escolaridade das mães dos estudantes de Engenharia Agrícola e Ambiental e Biotecnologia	87
Figura 34 – Relação de moradia e mudança para a cidade da Universidade	88
Figura 35 – Tipo de escola onde cursou o ensino médio	89
Figura 36 – Relação entre a evasão e o tipo de escola onde cursou o ensino médio .	90
Figura 37 – Relação entre a evasão e bolsa de estudo	91
Figura 38 – Relação entre a evasão e opção de curso dos estudantes de Ciência e tecnologia	92
Figura 39 – Relação entre a evasão e opção de curso dos estudantes de Engenharia Agrícola e Ambiental e Biotecnologia	92
Figura 40 – Frequência de evasão entre os cursos	95
Figura 41 – Métrica de avaliação Recall	97
Figura 42 – Métrica de avaliação Precision	98
Figura 43 – Métrica de avaliação F1 Score	98
Figura 44 – Métrica de avaliação ROC_AUC	99
Figura 45 – Métrica de avaliação Accuracy	100
Figura 46 – Confusion Matrix	101
Figura 47 – Comportamento dos dados em cada conjunto	103
Figura 48 – Teste de Friedman	104

LISTA DE TABELAS

Tabela 1 – Descrição da base de dados	32
Tabela 2 – Descrição da base de dados	33
Tabela 3 – Características do aprendizado supervisionado e do aprendizado não supervisionado	42
Tabela 4 – Características dos algoritmos de classificação e regressão	44
Tabela 5 – Termos de busca	59
Tabela 6 – Fontes de Pesquisa, strings de busca e retorno	61
Tabela 7 – Trabalhos Seleccionados	63
Tabela 8 – Trabalhos Seleccionados	64
Tabela 9 – Características dos algoritmos mais utilizados nos trabalhos seleccionados	68
Tabela 10 – Características dos algoritmos mais utilizados nos trabalhos seleccionados	69
Tabela 11 – Características dos algoritmos mais utilizados nos trabalhos seleccionados	70
Tabela 12 – Bases de Dados abertas	71
Tabela 13 – Comparação das melhores métricas entre os dois conjuntos de dados .	102

LISTA DE ABREVIATURAS E SIGLAS

AI	Artificial Intelligence
AUC	Area under the ROC Curve
CC	Ciência da Computação
CE	Critério de Exclusão
CI	Critério de Inclusão
CQ	Critérios de Qualidade
CRISP-DM	Cross-Industry Standard Process for Data Mining
DRA	Divisão de Registro Acadêmico
DSR	Design Science Research
EDM	Educacional Data Mining
FI	Fator de Impacto
IA	Inteligência Artificial
IDEB	Índice de Desenvolvimento da Educação Básica
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
KDD	Knowledge Discovery in Databases
KNN	K-Nearest Neighbors
MD	Mineração de Dados
MEC	Ministério da Educação
ML	Machine Learning
MLP	Multilayer Perceptron
MSL	Mapeamento Sistemático da Literatura
NB	Naive Bayes
PIB	Produto Interno Bruto
PNE	Plano Nacional da Educação
PPGCC	Programa de Pós-Graduação em Ciência da Computação
PROPLAN	Pró-Reitoria de Planejamento
Q1	Questão 1
Q2	Questão 2
Q3	Questão 3
Q4	Questão 4
Q5	Questão 5

QC	Questões Conceituais
QP	Questões Práticas
QT	Questões Tecnológicas
ROC	Receiver Operating Characteristic Curve
SEMMA	Sample, Explore, Modify, Model e Assess
SSPS	Statistical Package for the Social Sciences
SVM	Support Vector Machine
UERN	Universidade do Estado do Rio Grande do Norte
UFERSA	Universidade Federal Rural do Semi-Árido

SUMÁRIO

1	INTRODUÇÃO	18
1.1	Motivação	20
1.2	Questões de Pesquisa	23
1.3	Objetivos	25
<i>1.3.1</i>	<i>Objetivo Geral</i>	<i>25</i>
<i>1.3.2</i>	<i>Objetivos Específicos</i>	<i>25</i>
1.4	Hipóteses	25
1.5	Organização do Trabalho	26
2	PROCEDIMENTOS E MÉTODOS	27
2.1	Entendimento do negócio (<i>Business Understanding</i>)	30
2.2	Entendimento dos Dados (<i>Data Understanding</i>)	30
2.3	Preparação dos dados (<i>Data Preparation</i>)	33
2.4	Modelagem (<i>Modeling</i>)	35
2.5	Avaliação (<i>Evaluation</i>)	36
2.6	Ferramentas tecnológicas	37
2.7	Considerações Finais	37
3	FUNDAMENTAÇÃO TEÓRICA	39
3.1	Evasão estudantil	39
3.2	Machine Learning	41
<i>3.2.1</i>	<i>Algoritmos de Machine Learning</i>	<i>44</i>
3.3	Métricas para avaliação de modelos de <i>Machine Learning</i>	49
3.4	Análise Estatística	52
3.5	Considerações Finais	52
4	MAPEAMENTO SISTEMÁTICO DA LITERATURA	54
4.1	Procedimentos e métodos	56
<i>4.1.1</i>	<i>Questões de pesquisa</i>	<i>57</i>
<i>4.1.2</i>	<i>Critérios de seleção</i>	<i>57</i>
<i>4.1.3</i>	<i>Critérios de qualidade</i>	<i>58</i>
<i>4.1.4</i>	<i>Definição das Palavras-chave, String de Busca e Fontes</i>	<i>59</i>
<i>4.1.5</i>	<i>Processo de seleção</i>	<i>59</i>

4.2	Condução do Estudo	60
4.3	Resultados e Discussões	65
4.4	Considerações Finais	77
5	RESULTADOS	79
5.1	Análise comparativa entre as bases	79
5.1.1	<i>Situação em relação ao curso</i>	80
5.1.2	<i>Distribuição da universidade</i>	81
5.1.3	<i>Aproveitamento acadêmico no primeiro ano de curso</i>	81
5.1.4	<i>Renda mensal da família</i>	83
5.1.5	<i>Raça</i>	84
5.1.6	<i>Gênero</i>	85
5.1.7	<i>Escolaridade do pai</i>	86
5.1.8	<i>Escolaridade da mãe</i>	86
5.1.9	<i>Relação de moradia e mudança para a cidade da Universidade</i>	87
5.1.10	<i>Instituição onde cursou o Ensino médio</i>	88
5.1.11	<i>Conhecimento prévio na área do curso</i>	90
5.1.12	<i>Obtenção de bolsa durante o curso</i>	90
5.1.13	<i>Percepção do estudante com relação aos pré-requisitos</i>	91
5.1.14	<i>Primeira opção de curso</i>	92
5.1.15	<i>Filhos</i>	93
5.2	Comparação entre análise da presente pesquisa e o trabalho de Marques (2020)	93
5.3	Resultados da implementação dos modelos de ML	94
5.3.1	<i>Recall</i>	97
5.3.2	<i>Precision</i>	97
5.3.3	<i>F1 Score</i>	98
5.3.4	<i>Roc e AUC</i>	99
5.3.5	<i>Accuracy</i>	100
5.3.6	<i>Confusion Matrix</i>	100
5.3.7	<i>Análise estatística</i>	103
5.4	Considerações Finais	105
6	CONSIDERAÇÕES FINAIS	106

6.1	Contribuições	110
6.2	Limitações	111
6.3	Trabalhos futuros	111
	REFERÊNCIAS	113
A	APENDICE - FICHA DE EXTRAÇÃO DE DADOS	126
B	APENDICE - TUTORIAL DETALHADO DA CONSTRUÇÃO DA ANÁLISE EXPLORATÓRIA DE DADOS	128
C	APENDICE - TUTORIAL DETALHADO DA IMPLEMENTAÇÃO DOS MODELOS DE ML	137

1 INTRODUÇÃO

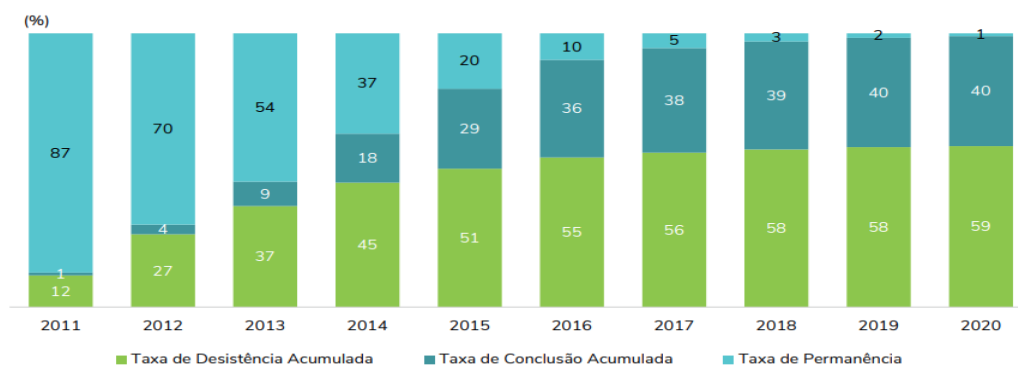
O Brasil é um país continental, caracterizado pela diversidade étnica, social e cultural, composto por cinco regiões e 27 unidades federativas. Neste país, apenas 17,4% das pessoas de 25 anos ou mais possuem o ensino superior completo (IBGE, 2019). A taxa de escolarização líquida (que mede o percentual de jovens de 18 a 24 anos matriculados no ensino superior em relação ao total da população da mesma faixa etária) é de apenas 18,1%, configurando-se abaixo da meta (33% até o ano de 2024) estimada pelo Plano Nacional da Educação (PNE), regido pela LEI N° 13.005/2014 (PNE, 2014). Apesar de o país possuir uma média de quase 2 milhões de concluintes (1,9 milhão em 2019) no ensino médio anualmente, estes estudantes não chegam até o ensino superior ou evadem antes de concluir o curso.

Barros (2021) afirma que, as perdas pessoais ao longo da vida, por jovem que não conclui sequer a educação básica, é de R\$ 290 mil a R\$ 395 mil, o que pode gerar para a sociedade, perdas de aproximadamente 220 bilhões de reais. Essa perda representa 2,5% do Produto Interno Bruto (PIB) nacional.

Um dos maiores desafios da educação no Brasil é a evasão estudantil. A evasão estudantil é o ato de deixar de frequentar as aulas, abandonar o ensino em decorrência de um ou mais fatores. Riffel e Malacarne (2010) definem evasão como o ato de evadir, fugir, abandonar, sair, desistir, não permanecer. Esse problema afeta a qualidade de ensino, impactando os âmbitos econômicos e ambientais, refletindo negativamente no desenvolvimento social.

No ensino superior, conforme o último censo da Educação Superior, divulgado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) e pelo Ministério da Educação (MEC) (MEC, 2022), referente ao ano de 2020, foram contabilizados 8.680.354 (oito milhões, seiscentos e oitenta mil, trezentos e cinquenta e quatro) estudantes matriculados na graduação. Entretanto, apenas 40% (quarenta por cento) desta quantidade de estudantes concluem o curso, pois a maioria entra nos dados estatísticos de abandono estudantil. A taxa de desistência acumulada, para estudantes que ingressaram no ensino superior em 2011 e foram acompanhados até 2020, indica que 40% dos ingressantes em 2011 concluíram seu curso de ingresso, 59% desistiram e 1% nele permanecem, conforme dados divulgados pelo INEP (MEC, 2022) e ilustrados na Figura 1.

Figura 1 – Evolução média dos indicadores de trajetória dos ingressantes de 2011 em cursos de graduação - Brasil - 2011-2020



Fonte: Elaborado pelo Inep com base nos dados do Censo da Educação Superior (2022)

Devido a esses prejuízos, identificar as causas da evasão estudantil em todos os níveis de ensino tornou-se um tema de grande interesse e tem desencadeado pesquisas nacionais e internacionais, tais como os trabalhos de Meca *et al.* (2020), Agrusti, Bonavolontá e Mezzini (2019), Marques *et al.* (2019), Mioranza *et al.* (2020) e Chung *et al.* (2019). Um dos pioneiros no desenvolvimento de pesquisas acerca da evasão e responsável pela indicação dos fatores que a desencadeiam foi Tinto (1975), se consolidando como principal estudo para compreender as causas da evasão estudantil.

Detectar previamente os perfis de estudantes considerados como possíveis evasores é importante porque essa informação pode auxiliar na tomada de decisões, para que os órgãos competentes possam desenvolver estratégias, tais como alterações curriculares e metodológicas, para mitigar esse risco (LOBO, 2017).

Uma alternativa para auxiliar na investigação das causas da evasão estudantil é a Mineração de Dados Educacionais (do inglês *Educational Data Mining (EDM)*): que é uma área emergente de Mineração de Dados (MD) na qual são desenvolvidos métodos e algoritmos de MD, de modo a compreender os dados em contextos educacionais, produzidos por estudantes e professores, considerando os fatores relacionados, tais como o ambiente no qual estão inseridos, a maneira como os atores interagem, entre outros (COSTA *et al.*, 2013). No contexto da evasão estudantil, a EDM vem sendo utilizada para desenvolver modelos preditivos de estudantes com possibilidade de evadir-se da instituição de ensino.

A EDM envolve tarefas de análise de dados gerados nas interações dos estudantes, desempenho acadêmico, fatores relacionados à instituição, dados demográficos, dentre

outros. Por meio dessa análise surgem subsídios para responder questões relacionadas ao desenvolvimento de soluções que auxiliem na permanência e sucesso dos estudantes. Essa tarefa, contudo, tem um alto nível de complexidade, considerando que, os dados analisados, por vezes, se tratam de diferentes características individuais dos estudantes a serem analisadas e tratadas de maneiras distintas.

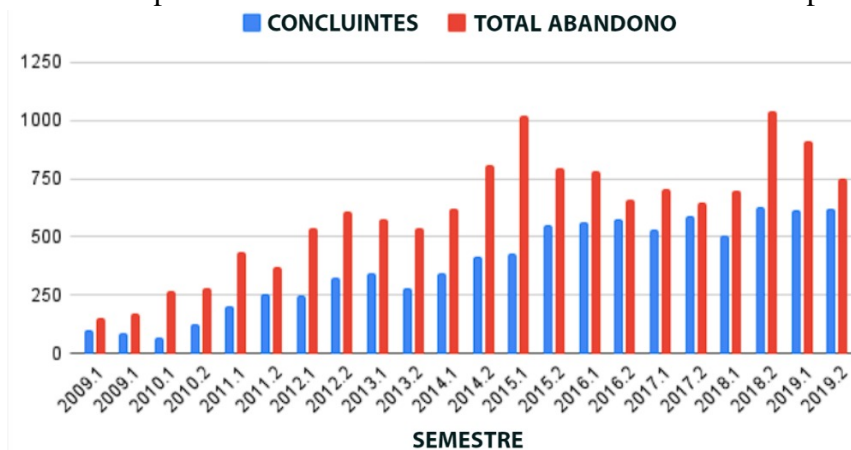
Dentre as áreas da computação que se aliam à EDM se destaca a *Machine Learning* (ML) (traduzido para o português como o Aprendizado de Máquina), definido por Faria (2014) como o campo da Ciência da Computação que concentra o uso dos algoritmos para imitar a maneira como os humanos aprendem, melhorando gradualmente sua precisão. Os algoritmos de ML são usados para realizar predição, baseado em características conhecidas e aprendidas por meio dos dados. Para a problemática da evasão estudantil, a técnica mais empregada é a de classificação (SCHEUER; MCLAREN, 2012).

1.1 Motivação

O último estudo do INEP, divulgado pelo MEC (2019) apresenta a Universidade Federal Rural do Semi-Árido (UFERSA), na 11ª posição do *ranking* de evasão, entre as universidades Federais. No ano de 2019 a UFERSA teve um percentual de abandono de 21,8%.

Dados fornecidos pela UFERSA, por meio da Divisão de Registro Acadêmico (DRA) (2020), apontam que desde o semestre 2009.1 até o semestre 2019.1, as taxas de abandono superam as taxas de conclusão (Figura 2).

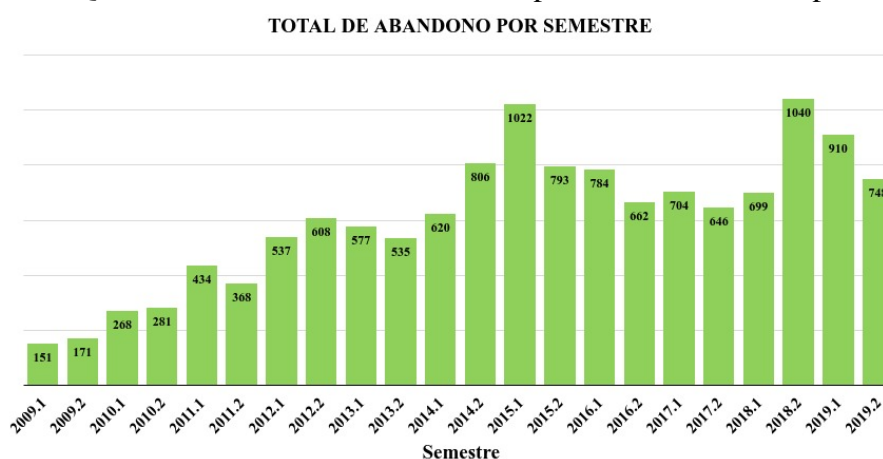
Figura 2 – Comparativo de Concluintes e Desistentes da UFERSA por semestre



Fonte: Divisão de Registros Acadêmicos da UFERSA (2020)

Ainda por meio dos dados da DRA (Figura 3), se observa uma quantidade considerável de estudantes que evadem da Universidade semestralmente, seja por desligamento voluntário, pela não renovação de matrícula, ou não comparecimento as aulas. A Universidade busca, constantemente, por soluções para mitigar a evasão.

Figura 3 – Quantitativo de abandono de cursos superiores da UFERSA por semestre



Fonte: Divisão de Registros Acadêmicos da UFERSA (2020)

Nesse cenário, uma pesquisa que contribuiu com essa problemática foi iniciada em 2018 e concluída em 2020 por Leonardo Torres Marques (MARQUES, 2020), discente do Programa de Pós-Graduação em Ciência da Computação (PPGCC), programa esse que é uma associação ampla entre a Universidade do Estado do Rio Grande do Norte (UERN) e a UFERSA. O trabalho de Marques (2020) aplicou a *EDM* a fim de detectar as causas da evasão estudantil dos estudantes do curso de Ciência da Computação (CC) da UFERSA, com base em um conjunto de fatores indutores. A pesquisa obteve os seguintes resultados:

- Uma abordagem de descoberta de conhecimento foi proposta para identificar os determinantes da evasão de estudantes no curso superior de CC da UFERSA. No que se refere aos determinantes responsáveis pela decisão de não permanência no curso, evidenciou-se que há um conjunto amplo de fatores associados à evasão, dentre os quais destacam-se: ano de ingresso e ano de saída da universidade, instituição que cursou o ensino médio, cidade onde o estudante residia, renda mensal familiar, raça, bolsa, aproveitamento acadêmico e grau de instrução dos pais (MARQUES, 2020);
- Com base na literatura e nas análises realizadas neste estudo, constatou-se que não existe um único e melhor algoritmo para previsão da evasão estudantil que possa ser aplicado em qualquer situação para fornecer os melhores resultados, devido ao fato

dos padrões de evasão serem influenciados por muitos fatores (MARQUES, 2020);

- A abordagem foi testada com cinco modelos diferentes, destacando-se que em dois deles (*Support Vector Machine e Adaboost*) se alcançou a excelente acurácia de noventa e oito por cento (98%) de acertos, por meio da aplicação em um conjunto de dados e seleção de variáveis representados pelas características de concludentes e evadidos do curso de CC da UFERSA (MARQUES, 2020).

A pesquisa de Marques (2020) cumpriu os objetivos de detectar as causas da evasão dos estudantes do curso de CC e conseguir identificar estudantes em risco de evasão. No entanto, a Universidade é composta por outros cursos superiores que também passam pelo mesmo problema de evasão de estudantes. Ainda assim, os modelos desenvolvidos na pesquisa não foram aplicados em outros cursos.

A pesquisa de Marques (2020) apontou como trabalho futuro a criação de uma ferramenta para os gestores que consiga prever quais estudantes estão em risco potencial de evadir, seja por meio de alerta ou relatórios e quais medidas podem ser tomadas para mitigar esses riscos. Adicionalmente, acredita-se que a ampliação do estudo, com a adição de novas variáveis, pode apresentar resultados ainda melhores. Duas das variáveis candidatas são: a escolha da universidade e do curso em que este estudante está matriculado. Essas variáveis são importantes para verificar a hipótese levantada por diversos professores: estudantes que ingressam na segunda opção de curso, ou que não ingressaram na sua primeira opção de universidade, tem maior chance de evadir-se.

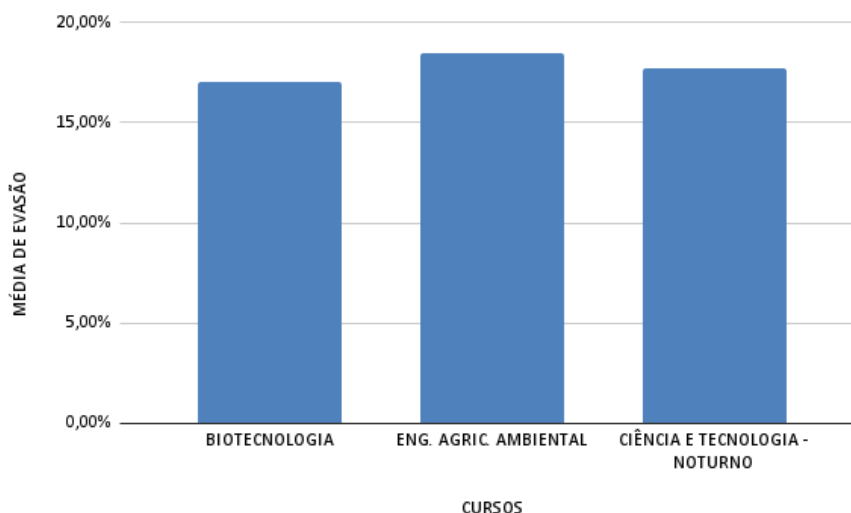
Outra condição a ser considerada é a investigação de fatores indutores distintos para cada área acadêmica, pois se observa que, motivos que levam um estudante de Ciências Exatas ou Engenharias a abandonar o curso, não necessariamente são determinantes para que estudantes de Ciências Sociais, Saúde ou Ciências Humanas desistam do seu curso.

Por fim, dada a crescente aliança entre MD e ML, se torna relevante a aplicação de modelos de ML para a predição de estudantes de diferentes cursos superiores da UFERSA com potencial risco de abandonar seus estudos.

Por meio dos dados fornecidos pela DRA, é possível identificar que os cursos superiores da UFERSA que possuem maior taxa de evasão são (Figura 4): Engenharia Agrícola e Ambiental (18,46% de evasão), Bacharelado em Ciência e Tecnologia (17,70% de evasão) e Biotecnologia (17,05% de evasão). Para fins de cálculo da média de evasão se considerou: o número de matrículas e o número de abandono (que compreende os dados

de abandono, trancamentos e cancelamentos) no período compreendido entre 2015.1 e 2020.1. Salienta-se que, para fins desta pesquisa, foram examinados apenas os cursos da modalidade presencial.

Figura 4 – Cursos da UFERSA com maior taxa de evasão



Fonte: Divisão de Registros Acadêmicos da UFERSA (2020)

Os processos de identificação das causas da evasão estudantil e retenção dos estudantes da UFERSA continuam manuais, subjetivos, empíricos, sujeitos à falha e ainda dependendo do esforço da gestão. Essas condições dificultam o acompanhamento das necessidades individuais de cada estudante e a identificação dos que estão vulneráveis e propensos a abandonar a Universidade. Dessa forma, a adoção de MD e ML viabiliza a detecção precoce de grupos de estudantes com risco de evasão estudantil.

1.2 Questões de Pesquisa

Diante das motivações apresentadas, a questão geral de pesquisa é:

QGP: Como realizar a predição, de forma precoce, de estudantes dos cursos de Bacharelado em Ciência e Tecnologia, Engenharia Agrícola e Ambiental, e Biotecnologia, com alto risco de evasão estudantil?

Conforme o problema da pesquisa, a Questão Geral de Pesquisa pode ser fragmentada em outros três grupos de perguntas: Questões Conceituais (QC), Questões Tecnológicas (QT) e Questões Práticas (QP). As QC descrevem sobre as noções necessárias à percepção dos problemas e sobre fatos que evidenciam porque as coisas ocorrem da

maneira que estão. As QT aportam o estado da arte das tecnologias para desenvolver o modelo de predição de evasão. Por fim, as QP procuram reconhecer a adequação prática da proposta para o mundo real. O conjunto de questões de pesquisa é descrito a seguir:

- **QC1 -Quais fatores podem influenciar no processo de evasão estudantil?**
- **QT1 – Quais as abordagens encontradas na literatura para o desenvolvimento de modelos de predição de evasão estudantil?**
 - QT 1.1 – Quais as etapas necessárias para a criação de um modelo preditivo de estudantes com alto risco de evasão estudantil?
- **QT2 – Quais as principais ferramentas utilizadas para a aplicação de ML na predição de evasão estudantil?**
- **QP - Como avaliar a qualidade do modelo preditivo?**
 - QP 1.1 - Quais métricas podem ser consideradas adequadas para a constatação dos resultados?
 - QP 1.2 – Como validar o processo de avaliação dos modelos de ML desenvolvidos?

1.3 Objetivos

1.3.1 *Objetivo Geral*

O objetivo geral desta pesquisa consiste em propor uma metodologia complementar a de Marques (2020), contemplando, de forma ampla, todas as etapas para predição e detecção de estudantes com alto risco de evasão em diferentes cursos superiores.

1.3.2 *Objetivos Específicos*

Destaca-se como objetivos específicos deste trabalho:

1. Compreender, por meio da literatura, o estado da arte relacionado ao desenvolvimento de modelos de predição, no âmbito da evasão estudantil e suas metodologias, técnicas, ferramentas utilizadas e resultados alcançados.
2. Propor um modelo preditivo que proporcione a identificação da possibilidade de evasão estudantil em diferentes cursos superiores.
3. Avaliar o modelo preditivo, por meio de métricas de avaliação e análise estatística.
4. Avançar no estado da arte apresentando nuances para o contexto do trabalho, diferentes do que foi encontrado na literatura, por meio de uma metodologia que garanta a reprodutibilidade.

1.4 Hipóteses

Em complemento aos objetivos citados, este trabalho também busca identificar as seguintes hipóteses levantadas a partir de trabalhos anteriores e de sessões de *brainstorming* realizadas com os pesquisadores da presente pesquisa. São elas:

1. Estudantes que não ingressaram na sua primeira opção de curso, têm uma maior probabilidade de evadir-se.
2. Estudantes que não possuem um conhecimento prévio na área do seu curso, têm uma maior probabilidade de evadir-se.
3. O maior índice de evasão ocorre no primeiro ano de curso.
4. Estudantes que reprovam em uma ou mais disciplinas no primeiro ano de curso, têm uma maior probabilidade de evadir-se.

5. Estudantes do turno noturno, têm uma maior probabilidade de evadir-se.
6. Estudantes do gênero feminino têm uma maior probabilidade de evadir-se.
7. Estudantes que cursaram o ensino médio em escolas públicas, têm uma maior probabilidade de evadir-se.
8. Estudantes oriundos de outros estados, têm uma maior probabilidade de evadir-se.
9. Estudantes com pais separados, têm uma maior probabilidade de evadir-se.
10. O mesmo modelo de ML (com as mesmas variáveis (*features* e configurações) pode ser aplicado a grupo de estudantes de quaisquer cursos superiores e obter resultados semelhantes.

Ainda em relação ao escopo deste trabalho, observa-se que o mesmo aborda processos de ML, utilização de algoritmos existentes e consolidados na área de MD, definição das melhores variáveis preditivas, mas não aborda o projeto de novos algoritmos de ML.

1.5 Organização do Trabalho

No Capítulo 2 é descrita a metodologia de trabalho, com destaque para as etapas de entendimento dos dados, a qual realiza o comparativo entre as bases de dados, aplicação dos modelos de ML e avaliação por meio de métricas. No Capítulo 3 é apresentado o Referencial Teórico acerca da temática da evasão estudantil e *machine learning*, algoritmos e métricas de avaliação. No Capítulo 4 é apresentado um Mapeamento Sistemático da Literatura, que caracteriza os principais trabalhos da literatura que tratam do problema evasão estudantil, bem como as ferramentas, algoritmos de ML, métricas de avaliação, processos e fatores utilizados para desvendar as causas da evasão estudantil, no qual se descrevem os principais trabalhos relacionados com os objetivos de pesquisa, avaliados por meio do referido Mapeamento Sistemático da Literatura. No Capítulo 5 são apresentados os resultados do trabalho, com as respectivas discussões sobre os experimentos realizados nos três cursos superiores. Por fim, as considerações finais são relatadas no Capítulo 6.

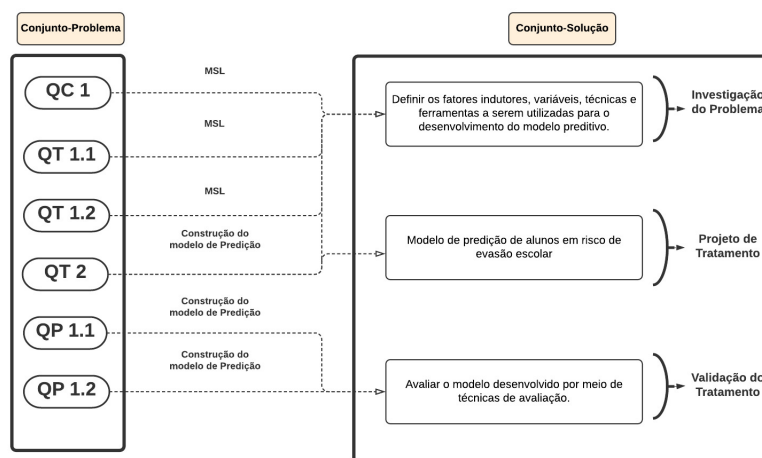
2 PROCEDIMENTOS E MÉTODOS

Neste capítulo são apresentados procedimentos e métodos da presente pesquisa, ressaltando-se os procedimentos de pré-processamento dos dados e as estratégias para a implementação dos modelos de ML, guiados pela metodologia adotada.

Esta pesquisa segue o *Design Science Research (DSR)*, que é uma metodologia centrada na investigação e geração de conhecimento no processo de concepção de artefatos. HEVNER *et al.* (2004) definem o DSR como “um processo rigoroso para projetar artefatos, resolver os problemas observados, fazer contribuições à pesquisa, avaliar os projetos e comunicar os resultados para o público adequado.”

A metodologia de pesquisa adotada é caracterizada por um conjunto de métodos de pesquisa que associam o conjunto-problema ao conjunto-solução, conforme ilustrado na Figura 5).

Figura 5 – Representação da metodologia do Design Science Research



Fonte: adaptada de HEVNER *et al.* (2004).

Para alcançar consolidar a proposta, este trabalho seguiu as etapas detalhadas a seguir:

- Para identificar os fatores indutores e explorar as principais abordagens e técnicas utilizadas no contexto da evasão estudantil, de modo a fortalecer o amadurecimento na área do estudo e de identificar lacunas que possam ser sanadas mediante a pesquisa a ser realizada, o ponto de partida foi o estudo dos trabalhos de Marques (2020) e Marques *et.al* (2019), que nortearam os objetivos e métodos a serem seguidos.

Por meio desta pesquisa foi possível compreender os principais conceitos acerca da evasão estudantil, EDM e o cenário de evasão estudantil na UFERSA.

- Após adquirir o amadurecimento da área do estudo, o próximo passo foi realizar uma pesquisa exploratória acerca do quadro teórico que norteia a presente pesquisa, objetivando formular o conteúdo científico nas etapas desenvolvidas e o estudo dos conceitos da área do trabalho proposto. Para tanto, realizou-se um levantamento de informações mediante a pesquisas consolidadas, para que, a partir do levantamento, fosse possível produzir conhecimento sobre evasão estudantil, EDM e ML, com o intuito de compreender como as pesquisas são desenvolvidas, quais as bases de dados utilizadas, quais os algoritmos e métricas de avaliação que mais se adéquam e, especialmente, quais os principais fatores que podem influenciar na decisão do estudante de abandonar a universidade. Esse conhecimento foi a principal referência para a compreensão e observação dos dados considerados no modelo de predição desenvolvido nesta pesquisa. As fontes digitais utilizadas para a pesquisa foram: *ACM Digital Library*, *IEEE Explorer*, *Scopus*, Portal de Periódicos da CAPES e *textitScience Direct*. Como resultado desta etapa tem-se o MSL (capítulo 4).

Em paralelo a esta fase também foi realizado o estudo das disciplinas de Ciência de Dados e Aprendizagem de Máquinas. Estas foram fundamentais para a compreensão de como ocorre todo o fluxo em Ciência de Dados, as práticas de implementação dos algoritmos de ML.

- Detido o conhecimento acerca dos fatores indutores, técnicas e ferramentas de ML adequadas para o contexto da evasão estudantil tratados na literatura, o próximo passo foi identificar quais os cursos superiores da UFERSA possuem maior taxa de evasão. Essa descoberta se deu por meio da análise dos dados de registro acadêmico, disponibilizados pelo DRA. Para fins de cálculo da média de evasão utilizou-se o número de matrículas e o número de abandono (que compreende os dados de abandono, trancamentos e cancelamentos) no período compreendido entre 2015.1 e 2020.1. Salienta-se que para fins desta pesquisa foram considerados apenas os cursos da modalidade presencial. Os cursos que apresentaram maior taxa de evasão foram: Engenharia Agrícola e Ambiental (18,46% de evasão), Bacharelado em Ciência e Tecnologia (17,70% de evasão) e Biotecnologia (17,05% de evasão).
- Uma vez definidos os cursos de graduação que farão parte do estudo de caso e os

estudantil, entre os cursos superiores da UFERSA. Adicionalmente, por esses cursos se classificarem em 3 áreas distintas, pensou-se em avaliar a efetividade de um mesmo modelo de ML em diferentes áreas acadêmicas, bem como observar o comportamento do mesmo conjunto de informações coletadas para os três cursos. Inicialmente, pretendia-se utilizar a base de dados da própria Universidade para realizar o presente trabalho. No entanto, após muitas tentativas, não foi possível o acesso à base e então os pesquisadores construíram duas bases de dados que decorrer deste capítulo.

2.1 Entendimento do negócio (*Business Understanding*)

Nesta etapa do CRISP-DM a finalidade é contextualizar os objetivos e os dados para que seja possível compreender a relevância dos dados naquele modelo de negócio específico.

Para o presente trabalho, esta etapa foi composta por reuniões entre os pesquisadores e especialistas na temática da evasão estudantil. Por fim, os pesquisadores realizaram um mapeamento sistemático da literatura sobre a evasão estudantil, a partir do qual buscou-se identificar tecnologias de mineração de dados, ferramentas de ML, fatores indutores da evasão estudantil, bases de dados abertas e métricas de avaliação que têm sido exploradas para desvendar as possíveis causas da evasão estudantil.

Os objetivos desse trabalho seguem o direcionamento dos trabalhos iniciais realizados por Marques (2020), que apresentou uma abordagem de descoberta de conhecimento, em banco de dados, com o objetivo de identificar grupos de estudantes em risco de evasão estudantil no curso de Bacharelado em Ciência da Computação da UFERSA.

Na pesquisa de Marques (2020) foi realizado uma coleta por meio do *google forms* para a construção da base de dados. Após a coleta, essa base passou pelas fases de análise, pré-processamento e em seguida, foram empregados cinco modelos de ML, avaliados pela métrica da acurácia. O trabalho cumpriu os objetivos propostos, mostrando-se eficiente na identificação de estudantes em risco de evasão para o curso de Ciência da Computação.

2.2 Entendimento dos Dados (*Data Understanding*)

Nesta etapa verifica-se a qualidade dos dados, integridade, distribuição de valores e conformidade de governança de dados.

A coleta de dados se deu em dois momentos distintos. Inicialmente, todo o processo de coleta, análise exploratória, implementação e aplicação dos modelos de ML foi realizado para o curso de Bacharelado em Ciência e Tecnologia. Após todo o processo concluído com esse curso, o estudo foi replicado com os cursos de Biotecnologia e Engenharia Agrícola e Ambiental, conforme se detalha a seguir.

1. A primeira base de dados foi construída a partir de uma pesquisa, disponibilizada aos ex-alunos (concluintes ou desistentes), do curso de Ciência e Tecnologia dos *campi* Caraúbas, Mossoró, Angicos e Pau dos Ferros, por meio de formulário eletrônico *google forms*, enviado pela PROGRAD no período compreendido entre 17/12/2021 e 01/02/2022. O formulário foi composto por perguntas pessoais, acadêmicas, demográficas e socioeconômicas dos estudantes.
2. A segunda base de dados foi construída semelhante à primeira, a partir de uma pesquisa, disponibilizada aos ex-alunos (concluintes ou desistentes), dos cursos de Biotecnologia e Engenharia Agrícola e Ambiental, ambos os cursos situados no *campus* Mossoró e para o curso de Engenharia Agrícola e Ambiental o questionário foi expandido para outras universidades do Brasil. O meio de coleta foi formulário eletrônico *google forms*, enviado pela PROGRAD para os estudantes da UFERSA e para as demais universidades do País foi realizada a tentativa de envio via e-mail para as coordenações dos cursos, a fim de que essas coordenações propagassem o questionário para os antigos estudantes da instituição. O período foi compreendido entre 01/03/2022 e 09/06/2022. O formulário foi composto por perguntas pessoais, acadêmicas, demográficas e socioeconômicas dos estudantes.

Observa-se que, no decorrer deste trabalho, a base de dados denominada como primeira base dados (ou primeiro conjunto de dados) se refere ao estudo com o curso de Ciência e Tecnologia. A segunda base de dados (ou segundo conjunto de dados) se refere aos cursos de Biotecnologia e Engenharia Agrícola e Ambiental.

As Tabelas 11 e 2 apresentam os atributos (informações) coletados e que foram comuns para às duas bases. Essas variáveis em que foram submetidas ao processo do CRISP-DM.

Tabela 1 – Descrição da base de dados

Atributo	Descrição
var_1	Variável classificadora que indica se o estudante é concluinte ou desistente
var_2	Qual campus e turno o estudante frequentou
var_3	Ano e semestre de ingresso
var_4	Ano e semestre de conclusão ou desistência
var_5	Indica se o curso foi a primeira opção de curso do estudante
var_6	Indica se o estudante possuía algum conhecimento prévio na área antes de ingressar no curso.
var_7	Indica o tipo de escola que o estudante cursou o ensino médio e caso tenha estudado em escola pública, qual foi o tipo de instituição (federal, estadual, militar)
var_8	Indica o tipo de escola que o estudante cursou o ensino fundamental e caso tenha estudado em escola pública, qual foi o tipo de instituição (federal, estadual, militar)
var_9	Indica a cidade/estado onde o estudante morava antes de ingressar na Universidade
var_10	Indica se o estudante mudou para morar na mesma cidade do campus que se matriculou ao ingressar na Universidade
var_11	Indica se o estudante morou na vila acadêmica
var_12	Indica o gênero do estudante (masculino, feminino ou não informado)
var_13	Indica a característica étnico-racial com a qual o estudante se identifica.
var_14	Indica o estado civil do estudante
var_15	Indica se o estudante possui filhos
var_16	Indica se o estudante trabalha
var_17	Indica a renda mensal da família ao ingressar na Universidade
var_18	Indica se o estudante foi contemplado com alguma bolsa durante o período que esteve no curso
var_19	Indica o aproveitamento acadêmico, até o segundo semestre do curso (se houve aprovações ou reprovações)
var_20	Indica se o estudante acredita que durante o período em que esteve no curso teve o devido acompanhamento da Universidade
var_21	Indica se o estudante acredita que a grade curricular do curso é adequado ao que se exige no mercado de trabalho
var_22	Indica se o estudante acredita que a cadeia de pré-requisitos das disciplinas do curso é adequada
var_23	Indica se o estudante considera que os critérios de avaliação do desempenho do estudante são adequados

Tabela 2 – Descrição da base de dados

Atributo	Descrição
var_24	Indica se o estudante considera que a infraestrutura da instituição (como laboratórios, bibliotecas ou equipamentos de informática) é adequada
var_25	Indica se o estudante considera que os professores apresentam deficiências pedagógicas, desinteresse ou outros problemas
var_26	Indica se o estudante considera que as carreiras relacionadas ao curso são socialmente reconhecidas
var_27	Indica se o estudante considera que o profissional formado é pago adequadamente
var_28	Indica se os pais do estudante eram casados durante o período em que o mesmo frequentava a Universidade
var_29	Indica o grau de instrução do pai do estudante
var_30	Indica o grau de instrução da mãe do estudante
var_31	Indica se a universidade a qual o estudante ingressou foi a primeira opção do estudante

Além dessas informações, o questionário foi adaptado antes da segunda aplicação (para os cursos de Biotecnologia e Engenharia Agrícola e Ambiental), no qual foram adicionadas as seguintes informações: universidade para obter a informação de qual universidade o estudante é egresso; curso, para identificar a qual curso o respondente estudou; estado, para obter a informação de qual estado o estudante respondente é oriundo; na informação de gênero foi adicionada a opção na qual o usuário prefere não informar em que gênero se identifica; por fim, foi adicionada a seguinte pergunta: "*a universidade em que estudou sempre foi sua primeira opção de curso?*".

Salienta-se que o questionário que deu origem a essa base de dados teve como viés o trabalho de Marques 2020, com algumas adaptações para a adição de perguntas advindas dos questionamentos e hipóteses desta pesquisa.

Os dados foram analisados para estudar os atributos disponíveis, volumetria e qualidade dos dados para serem tratados posteriormente. Nesta etapa também foram gerados gráficos de análise univariada e multivariada com o objetivo visualizar dados, responder hipóteses formuladas, gerando *insights* relevantes para a pesquisa.

2.3 Preparação dos dados (*Data Preparation*)

A preparação dos dados é necessária para efetuar a preparação dos conjuntos de dados, que serão utilizados na etapa de modelagem, sendo considerada como uma das etapas mais importantes porque afeta diretamente a possibilidade de bons resultados das

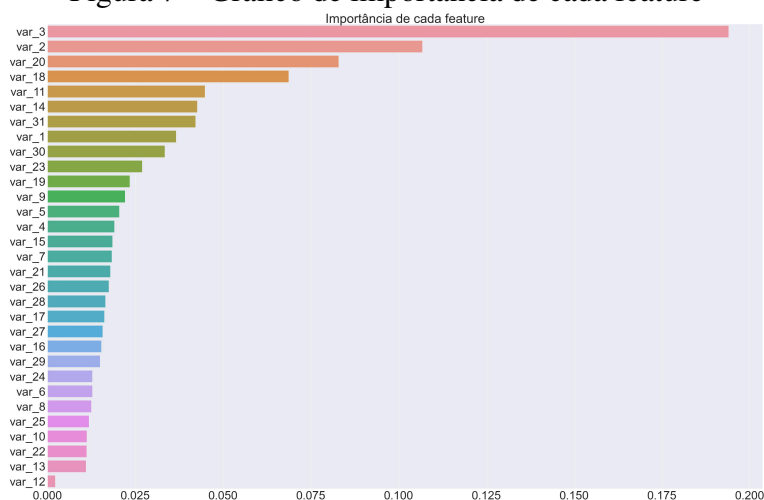
fases seguintes. Geralmente esta etapa é composta por cinco tarefas:

1. Selecionar os dados: no primeiro conjunto de dados foram utilizadas todas as características e já no segundo conjunto foram utilizadas todas as colunas com a exceção da data e hora da resposta do estudante.
2. Limpar dados: frequentemente, essa é a tarefa mais demorada. Uma prática comum durante essa tarefa é corrigir, imputar ou remover valores erroneamente inseridos nos conjuntos de dados. Para os dados da presente pesquisa foram realizadas padronizações das categorias tendo em vista que todas as características são do tipo categórica⁰.
3. Construir dados: derivar novos atributos que serão úteis. Nos conjuntos trabalhados neste estudo foram calculados o tempo da jornada acadêmica do estudante no curso, porém, ao realizar a correlação, não apresentou resultados significativos.
4. Integrar dados: essa fase possibilita a integração de diferentes fontes de dados. Nesse estudo, os conjuntos de dados foram oriundos do questionário desenvolvido. Existia a possibilidade de integrar com outras bases de dados da Universidade, como foi planejado inicialmente, porém, conforme explanado anteriormente, os pesquisadores deste trabalho não conseguiram acesso aos dados da UFERSA.
5. Formatar dados: foram realizadas conversões de variáveis categóricas para numéricas já que os modelos de ML utilizam apenas números.

Ao final desta etapa, foi gerado um gráfico para definir a importância de cada *feature* (denominação para dados que são preparados para se trabalhar com análise e mineração de dados). O gráfico de *feature importance* apresenta, em proporções, quão importante aquela *feature* é para o modelo. O desenvolvimento do gráfico é detalhado no apêndice B. Conforme ilustradas na Figura 7, as quinze primeiras características do gráfico são consideradas as melhores *features* ao que tange o problema conclusão ou evasão (após a décima quinta variável, a proporção estabiliza e todas as variáveis possuem a mesma importância para o problema em estudo). As melhores *features* que foram aplicadas na primeira análise para o curso de Ciência e Tecnologia e, posteriormente, foram replicadas para o segundo conjunto de dados na fase de treinamento e teste dos modelos de ML.

⁰ Em estatística, uma variável categórica é uma variável que pode assumir apenas um número limitado de opções. Um exemplo que pode representar variáveis categóricas é o tipo sanguíneo de uma pessoa: A, B, AB ou O (POTDAR *et al.*, 2017).

Figura 7 – Gráfico de importância de cada feature



Fonte: autoria própria (2022)

2.4 Modelagem (*Modeling*)

Esta etapa consiste em construir modelos ou aplicar técnicas de ML, definir os dados aleatoriamente para treino e teste, bem como definir a quantidade de vezes que esses dados vão iterar e alternar entre treino e teste. No presente trabalho, os modelos de ML foram implementados com os algoritmos: *KNN*, *Decision Tree*, *Naive Bayes (NB)*, *Random Forest*, *Perceptron*, *MLP* e *SVM* e para cada implementação, 80% da base foi reservada para treino e 20% da base para teste. Cada algoritmo de ML rodou 50 vezes fazendo a validação cruzada de *10k-folds*.

A validação cruzada (do inglês *Cross Validation*) (REFAEILZADEH *et al.*, 2009) é uma técnica que auxilia na divisão dos subconjuntos de dados de treino e teste. Observa-se que o *K* significa o número de subdivisões (iguais) e o *Fold* significa cada um dos blocos de cada *K*. Neste estudo, definiu-se $K = 10$; *Fold*: o modelo foi treinado com 80% dos dados, dividido entre 9 subconjuntos de treino e 1 subconjunto de teste em cada interação. Os 20% de dados que foram separados, foram usados na fase de validação, em que foram comparadas as métricas durante a fase de treinamento e a fase de validação. Na Figura 8 é ilustrado o fluxo do treinamento com *k-Fold*.

Figura 8 – Fluxo do treinamento com k-Fold



Fonte: autoria própria (2022)

Por meio da validação cruzada, se garante que, a cada rodada, os dados que antes eram utilizados para teste, na rodada seguinte passam a ser utilizados para treino e vice-versa.

Os parâmetros adotados para este estudo, tais como a divisão da base para treino e teste (50 vezes com validação cruzada de *10k-folds*) seguem o que é indicado pela literatura como mais adequado para aplicações em modelos de ML, conforme ressaltava Freitas (2020).

Caso os resultados sejam melhoráveis, a metodologia do CRISP-DM é definida para voltar até a etapa anterior (preparação de dados) e realizar ajuste nos dados disponíveis.

2.5 Avaliação (*Evaluation*)

Nesta etapa ocorre a avaliação dos modelos de ML implementados, visando checar se o modelo elaborado apresenta resultados válidos e corretos. Caso os resultados das avaliações de modelos não sejam satisfatórios, a metodologia do CRISP-DM permite que os pesquisadores retornem para as etapas anteriores, afim de revisar, entender e melhorar os modelos. Esta etapa também permite comparar os algoritmos e constatar quais deles obtém os melhores resultados quando aplicados à problemática estudada.

No presente trabalho, as métricas de avaliação utilizadas foram: *Accuracy*, *Precision*, *Recall* e *F1 Score*, que conforme apresentadas na seção 2, possuem particularidades que se complementam.

Os resultados das métricas de avaliação passaram ainda por uma análise estatística,

por meio dos testes de Friedman e Nemenyi. Os resultados de cada métrica de avaliação são considerados para a aplicação desses testes.

2.6 Ferramentas tecnológicas

Para o desenvolvimento do presente trabalho foram utilizadas algumas tecnologias que serão descritas nesta seção.

A linguagem de programação utilizada foi a *Python*, por meio do ambiente de desenvolvimento integrado (*IDE*) Anaconda (ROLON-MÉRETTE *et al.*, 2016). O Anaconda IDE é uma forma de distribuição *open source* da linguagem de programação *Python* e é uma plataforma poderosa para a ciência de dados porque oferece todos os pacotes necessários para a ciência de dados.

Foram utilizadas algumas bibliotecas da linguagem de programação *Python* para ML: *Pandas* (MCKINNEY, 2015), para análise e manipulação dos dados; *Seaborn* e *Matplot* (BISONG, 2019) para geração de gráficos estatísticos; *Numpy* (OLIPHANT, 2006) para armazenar os dados de treinamento, bem como os parâmetros dos modelos de ML em seus *arrays* e o *Scikit-learn* (KRAMER, 2016) que é uma biblioteca da linguagem Python desenvolvida especificamente para aplicações práticas de ML. A biblioteca *Scikit-learn* dispõe de ferramentas simples e eficientes para o desenvolvimento dos modelos preditivos de dados.

Para a realização dos testes de análise e estatística foi utilizada a linguagem de programação R (TEAM, 2000), por meio do ambiente *R studio*. Para os cálculos dos testes de Friedman e Nemenyi foi utilizada a biblioteca *TSUtils* (JANUSCHOWSKI *et al.*, 2019).

2.7 Considerações Finais

Neste capítulo foi apresentada a metodologia do trabalho, ressaltando-se os procedimentos de pré-processamento dos dados e as estratégias para desenvolvimento dos modelos de ML, guiados pela metodologia CRISP-DM.

No próximo capítulo são apresentados os resultados do trabalho, juntamente com as discussões a respeito dos experimentos correspondentes à análise univariada das variáveis consideradas como importantes para o problema estudado, bem como a aplicação dos modelos de ML nos conjuntos de dados, correspondente aos três cursos abordados.

Os apêndices B e C contém tutoriais que detalham as etapas do CRISP-DM implementadas na presente pesquisa para o primeiro conjunto de dados.

3 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados os conteúdos que fundamentam a proposta deste trabalho, destacando-se os aspectos relacionados a evasão estudantil, *Machine Learning* e métricas de avaliação de algoritmos.

3.1 Evasão estudantil

A evasão estudantil é vista como um grande problema que afeta qualquer nível de ensino (LOBO, 2012). Seu conceito é amplamente definido como o fenômeno que expressa o número de educandos de um grau de ensino ou de uma série escolar que abandonam definitiva ou temporariamente a instituição de ensino (JOSÉ *et al.*, 2010). No entanto, existem muitas definições diferentes na literatura, geralmente relacionadas a uma concepção temporal e que torna o tema controverso (XAVIER; MENESES, 2020).

Algumas terminologias encontradas na literatura relacionados à temática da evasão estudantil: *dropout* (DESCHACHT; GOEMAN, 2015), (CAIRNS *et al.*, 1989), (XAVIER; MENESES, 2020); *truancy* (ATTWOOD; CROLL, 2006), (REID, 2005); *attrition* (FIGUEIRA, 2015), (VELEZ, 1989); *withdrawal* (COOK, 1956), (NADASEN, 2016); *failure* (BRADSHAW *et al.*, 2008), (VOGT *et al.*, 1987) *retention* (MACY, 2015), dentre outros, que por vezes aparentam serem sinônimos em suas terminologias e em outros momentos abordam a temática de maneira distinta. Essa imprecisão em uma única definição chega a ser um problema, devido à falta de padronização e determinação de como é medida, abordada e pesquisada (ASHBY, 2004).

A terminologia mais comum e mais citada pelos pesquisadores é ***dropout* (abandono)**, que é definida de maneiras variadas: abandono de uma instituição ou programa em um período de tempo (BROCK, 2014); desistência dos cursos em decorrência dos exames realizados (TAN; SHAO, 2015). Em outros estudos o termo é empregado sem uma definição clara, mas abordando a temática do abandono ou desistência de um curso, ou instituição (FRANKO, 2015).

No Brasil, o Índice de Desenvolvimento da Educação Básica (IDEB) (IDEB, 2012) utiliza o termo "**abandono escolar**" e, o define como o afastamento do estudante do sistema de ensino e desistência das atividades escolares, sem solicitar transferência. Entende-se que, o estudante em situação de abandono, pode retornar no próximo período

letivo.

Outro termo comumente utilizado no Brasil é a "evasão estudantil", que é definida por Nacif *et.al* (2020) como a situação em que o estudante, seja reprovado ou aprovado, não efetua a matrícula para dar continuidade aos estudos no ano seguinte. A presente pesquisa adotou o termo evasão estudantil em todo o texto, dado que a presente pesquisa foi realizada com estudantes que saíram do seu curso de origem de forma permanente.

Segundo Rumberger (1995), a evasão estudantil é um processo complexo, dinâmico e cumulativo de saída do estudante. Isso se deve as diversidades que compõem o conjunto de circunstâncias individuais, institucionais e sociais e, em virtude desses fatos, é crescente o número de autores que buscam identificar as causas da evasão estudantil, tais como: Meca *et al.* (2020), Agrusti *et al.* (2019), Marques *et al.* (2019), Mioranza *et al.* (2020), Chung, Lee (2019) e Eric (2019).

Um dos pioneiros no desenvolvimento de pesquisas acerca da evasão estudantil e dos fatores que a desencadeiam foi Tinto (1975), se consolidando como principal estudo para compreender as causas da evasão estudantil. No Brasil, a principal referência desta temática é a Comissão Especial de Estudos da evasão, que destaca três fatores principais relacionados a desistência de estudantes, a saber: as **características individuais do estudante**, os **fatores internos às instituições** e os **fatores externos às instituições** (MEC, 2019).

Os fatores denominados como características individuais, representam dados inerentes às características individuais do estudante. Exemplos dessas características são: fatores acadêmicos (desempenho do estudante nas avaliações, assiduidade, participação em atividades extracurriculares), falta de habilidade de estudo, repetência, formação anterior precária, dentre outros (MARQUES, 2020), (TINTO, 1975), (SILVA *et al.*, 2015). Estes, além de serem os mais pesquisados no campo da evasão, também são os mais aplicados na construção de modelos de ML.

Fatores internos à instituição estão relacionados à estrutura da instituição, grade curricular, corpo docente desinteressado e/ou com qualificação insuficiente, ausência de professores e falta de refeitório no campus (MARQUES, 2020), (TINTO, 1975), (MATOS *et al.*, 2015). Geralmente estes fatores são investigados e combinados aos dados de características individuais.

Os fatores externos à instituição são fatores relacionados à vulnerabilidade soci-

oeconômica como, por exemplo, o estudante não conseguir conciliar estudo e trabalho, falta de políticas públicas adequadas para a educação, entre outras (MARQUES, 2020), (TINTO, 1975).

3.2 Machine Learning

Machine Learning - ML é um subcampo da ciência da computação definido por Samuel (1959) como "o campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados". Mitchell (1997) apresenta a seguinte definição formal: "*Diz-se que um programa de computador aprende pela experiência (E), com respeito a algum tipo de tarefa (T) e performance medida por (P), se sua performance na tarefa (T), medida por (P), melhora com a experiência (E).*"

O ML concentra seus estudos no desenvolvimento de algoritmos capazes de aprender com seus erros e realizar previsões sobre um conjunto de dados ou representações de experiências. Os algoritmos de ML tem sua construção baseada em entrada de dados amostrais, de modo a realizar previsões ou decisões guiadas por esses dados.

O crescimento do ML se deu a partir da Inteligência Artificial (IA) do inglês, *Artificial Intelligence (AI)*. Desde o início dos estudos da IA, alguns pesquisadores já se interessavam em fazer com que máquinas aprendessem a partir de dados. Durante a década de 90, o ML conquistou sua independência, se desmembrando da IA e concentrando seus esforços em métodos e modelos derivados da estatística e da teoria da probabilidade (LANGLEY *et al.*, 2011).

No que tange à categorização, o subcampo do ML se divide pelo tipo de aprendizado e possui quatro categorias: Aprendizado supervisionado (SINGH *et al.*, 2016), Aprendizado não supervisionado (MAHESH, 2020), Aprendizado semi-supervisionado (MATSUBARA, 2004) e Aprendizado por reforço (MARTINS *et al.*, 2007). Dessas categorias, duas se destacam como principais para o problema abordado na presente pesquisa. São elas:

1. **Aprendizado supervisionado:** neste tipo de aprendizado, os algoritmos aprendem a executar determinada tarefa a partir de resultados conhecidos e conseguem produzir padrões e hipóteses gerais usando instâncias fornecidas externamente para prever o destino de instâncias futuras (SINGH *et al.*, 2016).
2. **Aprendizado não supervisionado:** o objetivo dessa categoria de algoritmos é descobrir novos padrões nos dados. Para tanto, diferentemente da categoria super-

visionada, as saídas a serem estimadas não são fornecidas. Por este motivo, os dados são caracterizados como não-rotulados. Desse modo, os algoritmos devem descobrir, por meio da exploração dos dados, os possíveis relacionamentos entre eles (MAHESH, 2020).

Na Tabela 3 são descritas algumas características que distinguem os tipos de aprendizado supervisionado e não supervisionado.

Tabela 3 – Características do aprendizado supervisionado e do aprendizado não supervisionado

Aprendizado Supervisionado	Aprendizado não supervisionado
São treinados usando dados rotulados.	São treinados usando dados não rotulados.
Recebe feedback direto para verificar se está prevendo a saída correta ou não.	Não recebe nenhum feedback.
Prevê a saída.	Encontra os padrões ocultos nos dados.
Os dados de entrada são fornecidos ao modelo junto com a saída.	Apenas os dados de entrada são fornecidos ao modelo.
O objetivo é treinar o modelo para que ele possa prever a saída quando receber novos dados.	O objetivo é encontrar os padrões ocultos e insights úteis do conjunto de dados desconhecido.
Necessita de supervisão para treinar o modelo.	Não necessita de supervisão para treinar o modelo.
Pode ser categorizada em problemas de Classificação e Regressão.	Pode ser classificado em problemas de Agrupamento e Associações.
Produce um resultado preciso.	Pode fornecer resultados menos precisos em comparação com o aprendizado supervisionado.

É importante salientar que existem outras categorias de aprendizado:

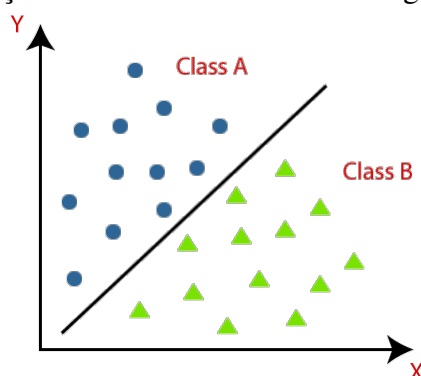
A depender da categoria de aprendizado que será utilizado para uma determinada problemática, os modelos de ML podem ser divididos em classes de algoritmo para cada tipo de aprendizado. As principais classes de algoritmos são: **Classificação e Regressão**.

- **Classificação:** pode ser enquadrada no aprendizado supervisionado e o seu objetivo é identificar a categoria de um determinado conjunto de dados ou observações e, em seguida, classificar a nova observação em várias classes ou grupos. Na construção desse tipo de algoritmo, as entradas são divididas em duas ou mais classes categóricas e o aprendiz deve produzir um modelo que vincula entradas não vistas a uma ou mais dessas classes (IGNACIO, 2021). No algoritmo de classificação, uma função de saída discreta (y) é mapeada para a variável de entrada (x): $y = f(x)$, onde y = saída categórica. Para melhor compreensão da notação da classificação, é ilustrada por meio de um gráfico presente na Figura 9.

Conforme pode ser observado na Figura 9, a linha que divide o primeiro quadrante separa o gráfico em duas classes distintas, representadas como classe A e classe B. Essas classes possuem características que são semelhantes entre si e diferentes de outras classes.

Comumente, os algoritmos de classificação são aplicados aos problemas relacionados à evasão estudantil. Outro tipo de aplicação comum é o reconhecimento de imagens. Os principais algoritmos de classificação são: *K-Nearest Neighbors (KNN)*, *Decision*

Figura 9 – Representação do funcionamento de um algoritmo de classificação



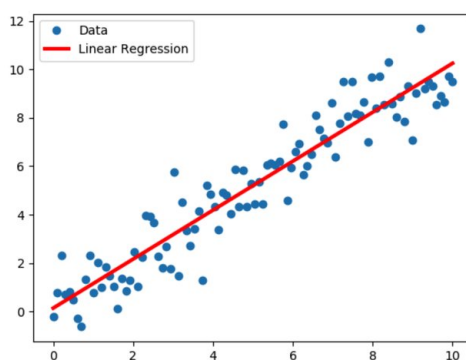
Fonte: Ignacio, 2021

Tree, NB, Random Forest, Perceptron, Multilayer Perceptron (MLP) e Support Vector Machine (SVM), que serão abordados na subseção 3.2.1.

- **Regressão:** a regressão também é parte do aprendizado supervisionado, no entanto, diferentemente da classificação, o objetivo é identificar uma categoria em escala contínua, podendo até ser uma probabilidade, por meio da previsão de um valor numérico. A regressão ajuda a encontrar a correlação entre as variáveis, permitindo prever a variável de saída contínua com base em uma ou mais variáveis preditoras (BONACCORSO, 2017).

Por meio da regressão, é possível visualizar uma linha ou curva que passa por todos os pontos de dados no gráfico do preditor de destino de forma que a distância vertical entre os pontos de dados e a linha de regressão seja mínima (BONACCORSO, 2017). Tran (2019) pontua o comportamento de um modelo de regressão como: "A distância entre os pontos de dados e a linha indica se um modelo capturou um relacionamento forte ou não". Essa informação pode ser ilustrada na Figura 10.

Figura 10 – Exemplo de modelo de regressão



Fonte: Bonaccorso, 2017

Por meio da Figura 10 é possível observar que o modelo de regressão linear fornece uma linha reta inclinada que representa a relação entre as variáveis. O eixo x representa a variável independente (variável preditora), o eixo y representa a variável dependente (variável alvo), e os pontos circulares representam os dados.

Algumas das aplicações para a regressão são: previsão de preço de um imóvel, probabilidade de superlotação de um estoque, previsões financeiras como a aumento do limite de crédito de um cliente ou qualquer outro valor quantitativo.

Os principais algoritmos de regressão são: *Linear Regression*, *Ridge Regression*, *Neural Network Regression*, *Lasso Regression*, *Decision Tree Regression*, *Random Forest*, *KNN* e *SVM*. Observa-se que, alguns algoritmos de ML podem ser utilizados tanto para problemas de Classificação quanto para problemas de Regressão. Na Tabela 4 são descritas algumas características que distinguem os algoritmos para problemas de classificação e regressão.

Tabela 4 – Características dos algoritmos de classificação e regressão

Classificação	Regressão
A variável de saída deve ser um valor discreto.	A variável de saída deve ser de natureza contínua ou valor real.
A tarefa do algoritmo de classificação é mapear o valor de entrada(x) com a variável de saída discreta(y).	A tarefa do algoritmo de regressão é mapear o valor de entrada (x) com a variável de saída contínua (y).
São usados com dados discretos.	São usados com dados contínuos.
Tenta encontrar o limite de decisão, que pode dividir o conjunto de dados em diferentes classes.	Tenta encontrar a linha de melhor ajuste, que pode prever a saída com mais precisão.

Fonte: Elaborada pela autora (2022)

3.2.1 Algoritmos de Machine Learning

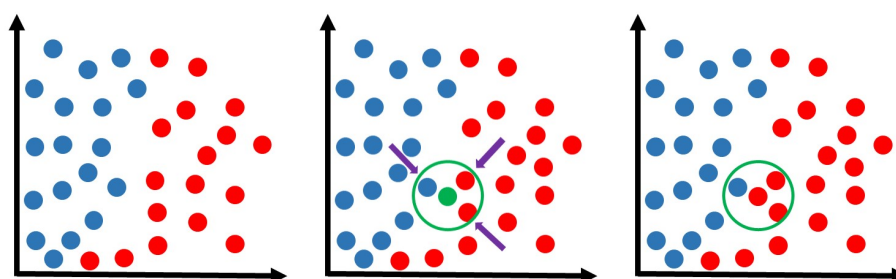
Na seção anterior se verificou que, os algoritmos de classificação e regressão têm como base o aprendizado supervisionado. No presente estudo, por se tratar de identificar uma classe de estudantes que podem evadir da universidade, se utilizou a classificação. Portanto, alguns algoritmos dentro dessa categoria são abordados nesta seção.

- **KNN**: é um algoritmo no qual o aprendizado é baseado "no quão similar" são os dados de um determinado elemento e seus *k-vizinhos*. O algoritmo KNN armazena todos os dados disponíveis e classifica um novo ponto de dados, assumindo a semelhança entre o novo dado e os dados disponíveis.

O KNN pertence à categoria de algoritmos supervisionados de ML e pode ser utilizando tanto para classificação quanto para regressão, dependendo da abordagem do problema em que o algoritmo está sendo aplicado. Na maioria dos casos este algoritmo é aplicado para problemas de classificação (MANI; ZHANG, 2003). Para

o problema de classificação, a saída do algoritmo KNN é uma associação de classe. Um objeto é classificado pelos votos de seus vizinhos, com o objeto sendo atribuído à classe mais comum entre seus k vizinhos mais próximos. Para o problema de regressão, a saída do algoritmo KNN é o valor da propriedade para o objeto. Este valor é a média dos valores de k vizinhos mais próximos. A sequência ilustrada na Figura 11 apresenta o funcionamento do KNN.

Figura 11 – Sequencia do algoritmo KNN



Fonte: Didática Tech (2022)

Conforme ilustrado na Figura 11, a aplicação do KNN funciona da seguinte maneira: se considera um conjunto de dados dividido em duas classes (azul e vermelha), o modelo KNN é treinado de modo a aprender as características de cada uma das classes, de modo que, ao inserir um novo dado (verde), o modelo possa definir de qual classe esse novo dado está mais próximo e assim classificá-lo.

- **NB:** é um algoritmo classificador que vem de uma família de classificadores probabilísticos e sua base é o Teorema de Bayes (RISH *et al.*, 2001).

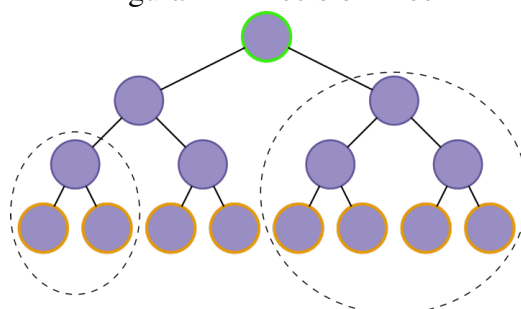
O funcionamento do NB é baseado na atribuição de classe mais provável a um dado exemplo descrito por seu vetor de características. Para cada valor de um vetor de dados, o NB atribui uma probabilidade de pertencer a determinada classe ou problema. O modelo Naive Bayes possui 3 variações: Gaussiano, Multinomial e Bernoulli.

- **Decision Tree (Árvore de Decisão):** é um algoritmo da categoria supervisionado, que pode ser utilizado tanto para problemas de classificação como para problemas de regressão. Este algoritmo busca imitar a capacidade de pensamento humano ao decidir por meio de uma estrutura de árvore, na qual os nós internos representam as características de um conjunto de dados, os ramos representam as regras de decisão e cada nó folha representa o resultado (QUINLAN, 1996).

As decisões tomadas pela *decision tree* com base nas características do conjunto de dados, começando a partir do nó raiz da árvore e comparando os valores do atributo raiz com o atributo registro (conjunto de dados) e seguindo para o próximo nó. No nó seguinte, o algoritmo compara novamente o valor do atributo com os outros sub-nós e avança. Ele continua o processo até atingir o nó folha da árvore.

De modo detalhado, o algoritmo *decision tree* conterà duas classes e cada uma levará a uma decisão distinta, conforme ilustrado na Figura 12.

Figura 12 – Decision Tree



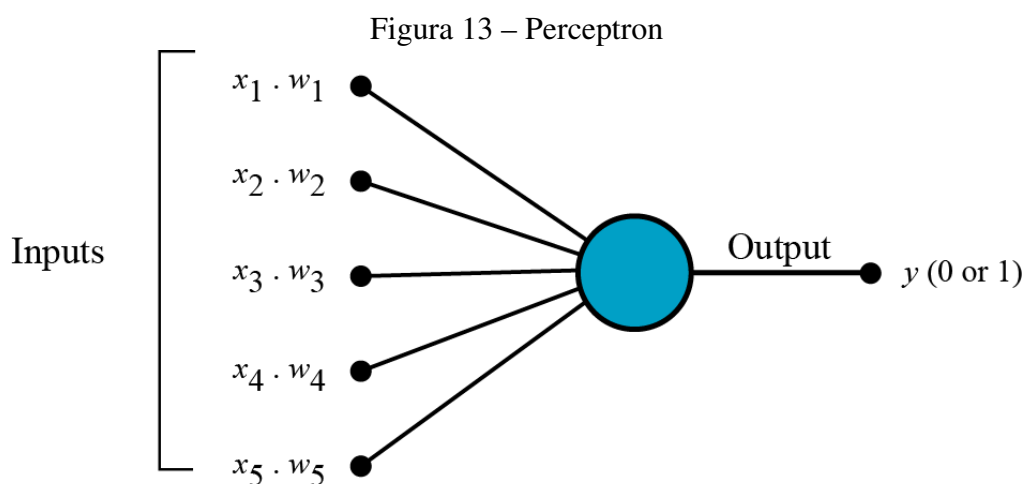
Fonte: Fernandes (2019)

A raiz da árvore, destacada na imagem com o círculo verde, representa o conjunto de dados, as linhas pretas que ligam os nós são denominadas de ramos. As arestas saindo da raiz e indo para a esquerda e para a direita representam as classes de decisão, de modo que cada nó representa o conjunto de características de cada classe. Os outros nós e ramos corresponderiam então a outras perguntas e as folhas (círculos laranja) dariam a classificação final. O número total de perguntas feitas nesse processo é chamado de profundidade (ou altura) dessa árvore. Esse algoritmo é baseado na ideia de entropia (ou ganho de informação). A entropia de um conjunto de dados é uma métrica que representa a aleatoriedade em seus valores. O seu valor é zero quando não há aleatoriedade (todos os elementos do dataset tem a mesma classificação) e aumenta conforme o conjunto de dados fica mais impuro (aleatório) (FRNANDES, 2019).

- **Random Forest (Floresta aleatória):** é um algoritmo de ML pertencente à categoria de aprendizado supervisionado, que pode ser utilizado tanto para problemas de classificação como para problemas de regressão. Sua estrutura é composta por uma coleção de árvores de decisão de um determinado conjunto de dados, com o objetivo de combinar a saída das árvores de decisão, calculando a média dessas saídas, para

alcançar um único resultado (CUTLER *et al.*, 2012). Esse algoritmo de ML possui três hiperparâmetros principais: tamanho do nó, o número de árvores e o número de recursos amostrados. O maior número de árvores na floresta leva a uma maior precisão e evita o problema de *overfitting* que é um fenômeno que ocorre quando o modelo de ML aprende muito bem sobre determinado conjunto dados trabalhados, porém, se mostra ineficaz para prever novos resultados.

- **Perceptron**: é um algoritmo classificador linear (binário), caracterizado como supervisionado, que pode ser aplicado em problemas de classificação de dados. Este algoritmo de ML é estruturado no formato de uma rede neural artificial de camada única (com apenas um neurônio) que possui quatro parâmetros principais: valores de entrada (nós de entrada), pesos e viés, soma líquida e uma função de ativação (GALLANT *et al.*, 1990). Na Figura 13 o funcionamento da *perceptron* é lustrado.



Fonte: Data Science Academy (2022)

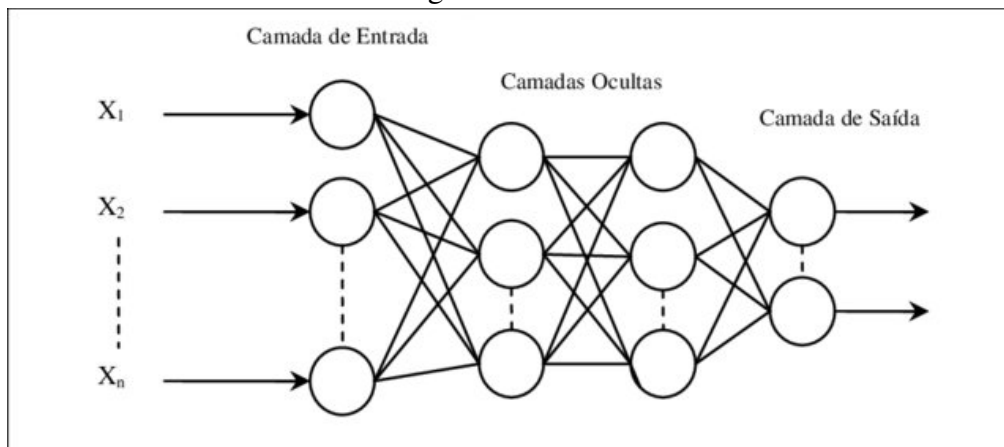
O funcionamento do *perceptron* é iniciado com a multiplicação de todos os valores de entrada e seus respectivos pesos, seguido pela soma desses valores, a fim de criar uma soma ponderada. A soma ponderada obtida é aplicada a uma função de ativação, também conhecida como função degrau f para obter a saída desejada, mapeada entre os valores necessários (0,1) ou (-1,1).

Observa-se que, o *perceptron* só pode ser utilizado para classificar os conjuntos linearmente separáveis e é comumente aplicado a problemas complexos, devido ao seu estado de constante avanço.

- **MLP**: é um algoritmo de ML semelhante à perceptron, mas que em sua composição

conta com mais de uma camada de neurônios em alimentação direta (GARDNER; DORLING, 1998), conforme ilustração na Figura 14.

Figura 14 – MLP



Fonte: Adaptado de Gardner (1998).

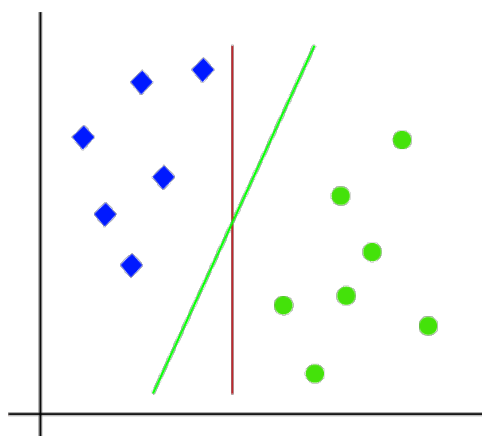
O modelo MLP também é conhecido como algoritmo *Backpropagation* e é executado em duas etapas:

1. **Forward:** estágio no qual as funções de ativação começam na camada de entrada no estágio avançado e terminam na camada de saída;
2. **Estágio Backward:** estágio no qual os valores de peso são modificados conforme a exigência do modelo. Nesse estágio, o erro entre a saída real e a demandada se originou na camada de saída e terminou na camada de entrada.

Um algoritmo MLP possui maior poder de processamento do que um *perceptron* de uma camada e pode processar padrões lineares (binários) e não lineares.

- **SVM:** é um algoritmo da categoria supervisionado, que pode ser utilizado tanto para problemas de classificação como para problemas de regressão. No entanto, sua principal aplicação em ML é como um classificador linear binário não probabilístico, que atua definindo o hiperplano de separação ideal entre duas classes (WANG; HU, 2005). O hiperplano traçado pelo SVM é como uma linha que separa pontos que representam uma classe distinta, conforme ilustrado na Figura 15.

Figura 15 – SVM



Fonte: Adaptado de Wang (2005).

Um exemplo de aplicação do *SVM* é o problema da evasão estudantil. Aplicado a um conjunto de dados de estudantes concluintes e desistentes. O algoritmo *SVM* é capaz de traçar um hiperplano (apenas uma linha reta) para separar essas duas classes, considerado como o limite de decisão.

3.3 Métricas para avaliação de modelos de *Machine Learning*

Algoritmos de ML são tecnologias que vêm se mostrando eficientes para a previsão de dados. Entretanto, a depender do problema, um determinado algoritmo pode se mostrar ineficiente, enquanto outro algoritmo pode apresentar o melhor desempenho (MONARD; BARANAUSKAS, 2003). Dessa forma, se faz necessária a utilização de métricas de avaliação que possibilitem comparar o desempenho, bem como as limitações dos algoritmos. Diante disso, essa seção descreve as principais métricas de avaliação, frequentemente utilizadas pela comunidade de ML, para comparar dois ou mais algoritmos e que foram aplicadas no presente estudo.

- **Matriz de Confusão:** é uma métrica que é extraída por meio de uma matriz que possibilita determinar o desempenho dos modelos de classificação para um conjunto de dados de teste nos quais a variável resposta é categórica. A matriz é dividida em duas dimensões: os valores previstos pelo modelo e os valores reais, que são os valores verdadeiros para as observações dadas juntamente com o número total de previsões.

Por meio da matriz de confusão, é possível calcular os diferentes parâmetros do

modelo, como a exatidão, precisão, dentre outros. Desse modo, a matriz de confusão pode servir como parâmetro para outras métricas como a *recall* e *precision*.

- **Acurácia***Accuracy*: esta métrica é considerada uma das mais simples de implementar, conhecida como a métrica global, dado que seu resultado pode ser calculado como a razão entre o número de previsões corretas para o número total de previsões. Esta métrica funciona bem apenas se houver igual número de amostras pertencentes a cada classe, podendo comprometer sua avaliação, caso o conjunto de dados seja desbalanceado.
- **Precision**: esta métrica determina a exatidão do modelo em classificar uma amostra como positiva. Seu cálculo é dado pela razão entre o número de amostras positivas classificadas corretamente e o número total de amostras classificadas como positivas (correta ou incorretamente). When *et.al.* (2012) pontuam que "a *precision* reflete a confiabilidade do modelo em classificar as amostras como positivas". Para tanto, além da métrica classificar todas as amostras positivas como positivas, é importante ela não classificar erroneamente uma amostra negativa como positiva.
- **Recall**: é uma métrica semelhante à *precision*, no entanto, ela objetiva calcular a proporção de positivos reais identificados incorretamente (GRANDINI *et al.*, 2020). O cálculo é feito pelo número de resultados positivos corretos, dividido pelo número de todas as amostras que deveriam ter sido identificadas como positivas. Quanto maior o *recall*, mais amostras positivas foram detectadas.
- **F1 Score**: é uma métrica de ML utilizada em modelos de classificação, que tem como base a combinação das métricas *precision* e *recall* em uma única métrica. O *F1 Score* é definido como a média harmônica de *precision* e *recall* (GRANDINI *et al.*, 2020). Desse modo, a presente métrica atribui peso igual às duas métricas, de tal forma que um modelo obterá uma pontuação alta na *F1 Score* se a *precision* e o *recall* forem altos. Por outro lado, um modelo obterá uma pontuação *F1 Score* baixa se *precision* e o *recall* forem baixos. Por fim, um modelo poderá obter uma pontuação *F1 Score* média se uma das duas métricas que formam a combinação obtiver pontuação baixa e a outra obtiver pontuação alta, ou se as duas métricas obtiverem um valor médio.
- **Receiver Operating Characteristic Curve (ROC) e Area under the ROC Curve (AUC)**: é uma métrica para avaliar o desempenho do modelo de classificação por meio

de um gráfico em curva. A *ROC* representa um gráfico para mostrar o desempenho do modelo em diferentes níveis de limiar. A curva *ROC* é desenhada sob dois parâmetros: a taxa de verdadeiros positivos e a taxa de falsos positivos. De forma complementar, a *AUC* calcula o desempenho em todos os limites e fornece uma medida agregada. O valor da *AUC* possui variação entre 0 (um modelo com previsão 100% errada terá uma *AUC* de 0,0) e 1 (modelos com previsões 100% corretas terão uma *AUC* de 1,0) (MUSCHELLI, 2020).

Na Figura 16 é ilustrada uma síntese (disposta em forma de tabela) das métricas que são base para a *matriz de confusão* em uma comparação feita por Hossin e Sulaiman (2015).

Figura 16 – Métricas para avaliações de modelos de ML baseados em técnicas de classificação

Métrica	Fórmula	Foco da avaliação
<i>Accuracy</i>	$\frac{tp + tn}{tp + fp + tn + fn}$	Em geral, esta métrica mede a proporção de previsões corretas sobre o número total de instâncias avaliadas.
<i>Precision (p)</i>	$\frac{tp}{tp + fp}$	A precisão é usada para medir os padrões positivos previstos corretamente a partir do total de padrões previstos em uma classe positiva.
<i>Recall (r)</i>	$\frac{tp}{tp + tn}$	O Recall é usado para medir a fração de padrões positivos classificados corretamente.
<i>F1-Score</i>	$\frac{2 * p * r}{p + r}$	Representa a média harmônica entre os valores das métricas <i>Recall</i> e <i>Precision</i> .

- *tp* significa *true positive*: percentual predito positivo corretamente sobre o total que era de fato positivo.
- *tn* significa *true negative*: percentual predito negativo corretamente sobre o total que era de fato negativo.
- *fp* significa *false positive*: percentual predito positivo sobre o total que era de fato negativo.
- *fn* significa *false negative*: percentual predito negativo sobre o total que era de fato positivo.

Fonte: autoria própria com base em Hossin e Sulaiman (2015).

3.4 Análise Estatística

A análise estatística aplicada aos modelos de *machine learning* é eficiente para validar estatisticamente o desempenho dos modelos (VERMA; RANGA, 2020). Os testes de Friedman (FRIEDMAN, 1937) e Nemenyi (DUNN, 1961) são os principais testes estatísticos aplicados com a finalidade de comparar o desempenho de algoritmos classificadores de maneira correta.

Os testes de Friedman e Nemenyi trabalham com duas hipóteses: A hipótese nula (H_0) considera que não há diferença de desempenho entre os classificadores (todos possuem desempenho semelhante); a hipótese H_1 considera que existe pelo menos um classificador com desempenho significativamente diferente de pelo menos um outro classificador.

O teste de Friedman é um teste não-paramétrico, aplicado para verificar se existe diferença significativa entre três ou mais classificadores que estão sendo testados (H_0) (FRIEDMAN, 1940). Esse teste não tem parâmetros para determinar onde está a diferença significativa. Para fazer o teste de Friedman, se utiliza os valores médios de, pelo menos, 4 métricas de avaliação como referência.

O teste de Nemenyi é um pós-teste realizado para identificar qual o classificador possui desempenho significativo. O teste Nemenyi calcula a diferença entre as classificações médias entre todos os classificadores. Se essa diferença for maior ou igual a uma distância crítica, se constata que os classificadores são significativamente diferentes entre si (HERBOLD, 2020).

3.5 Considerações Finais

Neste capítulo foram apresentados os principais conceitos acerca da evasão estudantil. Explanaram-se os principais conceitos referentes ao *Machine Learning*, além das definições conceituais sobre os principais algoritmos e métricas de avaliação e análise estatística aplicada a avaliação de desempenho dos modelos de ML.

No próximo capítulo são discutidos os trabalhos relacionados, obtidos por meio de um Mapeamento Sistemático da Literatura, que serviram como norteador para a presente pesquisa, para obter conhecimento sobre o quadro teórico, objetivando formular o conteúdo científico nas etapas desenvolvidas e o estudo dos conceitos da área do trabalho proposto. O mapeamento sistemático também foi utilizado como referencial técnico para o

desenvolvimento dos experimentos realizados neste trabalho.

4 MAPEAMENTO SISTEMÁTICO DA LITERATURA

Neste capítulo são apresentados os conteúdos que fundamentam a proposta deste trabalho, se destacando os aspectos relacionados a evasão estudantil, *Machine Learning* e métricas de avaliação de algoritmos.

Os direcionamentos conceituais e técnicos da presente pesquisa tiveram ponto de partida em um Mapeamento Sistemático da Literatura (MSL). Um MSL apresenta-se como o processo para coleta, avaliação e sistematização de trabalhos primários, apresentando uma abordagem quantitativa, capaz de fornecer dados para identificação de tendências de estudos, bem como lacunas a serem exploradas (CAMPOS *et al.*, 2020). Para conhecer as pesquisas desenvolvidas na EDM, bem como suas características, técnicas e ferramentas aplicadas à evasão estudantil, se faz necessário recorrer a uma metodologia de pesquisa confiável e repetível, como um MSL.

Petersen (2008) acredita que o estudo feito por meio do MSL é estruturado por fornecer publicações mediante relatórios e resultados de pesquisa, abrangendo quantitativos de estudos, os quais podem ser categorizados e ter um mapa de seus resultados vistos num resumo visual. Ademais, no que diz respeito a análise de resultados, possibilita verificar a frequência de publicações por categorias e, dessa forma, visualizar a ênfase dada em pesquisas anteriores e, assim, identificar lacunas numa pesquisa atual, com o intuito de sugerir ou instigar novas hipóteses, novos estudos, um estudo adicional e contextualizar novas atividades de pesquisa.

O presente MSL, tem como ponto de partida uma Revisão Sistemática da Literatura (RSL) desenvolvida por Marques (2020) que considerou o período compreendido entre 2008 e 2018 (10 anos), nas bases: *IEEEExplore*, *ACM Digital Library*, *Science Direct* e *Scopus* e objetivou identificar os fatores indutores da evasão estudantil, as abordagens de descoberta de conhecimento, as ferramentas e técnicas de MD que têm sido utilizadas para descoberta das causas da evasão estudantil. Diferente da RSL previamente executada, a presente pesquisa adotou um MSL, visto que o número de trabalhos retornados foi consideravelmente maior, permitindo obter um contexto geral do andamento das pesquisas relacionadas a evasão estudantil nos três últimos anos.

Na revisão sistemática executada por Marques (MARQUES, 2020), foram selecionados 16 trabalhos. Os resultados obtidos por meio da análise dos trabalhos selecionados foram:

- É possível verificar que as ferramentas *Waikato environment for Knowledge Analysis (Weka)* e *Mplus*, se destacam como as ferramentas de MD mais utilizadas para auxiliar a desvendar causas da evasão estudantil, utilizadas em três (3) trabalhos (MARQUES, 2020);
- A técnica de Classificação, que foi utilizada em 4 (quatro) trabalhos e a técnica de Equações Estruturais empregada em três (3) trabalhos, foram as técnicas de MD mais utilizadas para descoberta das causas da evasão estudantil, mediante os trabalhos selecionados (MARQUES, 2020);
- Em relação aos fatores indutores para a evasão estudantil, se constatou que, em todos os trabalhos retornados, foram investigadas possíveis causas relacionadas às características individuais do estudante. Essa investigação ocorre por meio da análise dos dados armazenados nas instituições ou levantamento de dados por meio de questionários, aplicados para coletar informações da perspectiva do estudante. Marques (2020) ressalta que, fatores relacionados as características internas à instituição e características externas à instituição, foram pouco explorados em conjunto.

A presente pesquisa apresenta o planejamento e execução de um MSL, propondo uma atualização expandida à pesquisa de Marques (2020), determinando como horizonte de pesquisa os anos de 2018 a 2021, de modo a oferecer subsídios para novos estudos, detectar lacunas e potencialidades de aprofundamento nessas áreas.

Diante do contexto apresentado, este MSL tem como objetivos: (I) identificar as ferramentas mais utilizadas na MD para a descoberta das causas da evasão estudantil no ensino presencial; (II) identificar quais as técnicas ML têm sido utilizadas na predição de estudantes com alto risco de evasão estudantil no ensino presencial, identificando suas vantagens e desvantagens; (III) Desvendar quais fatores indutores têm sido investigados para a descoberta das causas da evasão estudantil no ensino presencial; (IV) Investigar a existência de bases abertas que forneçam dados escolares de estudantes; (V) Identificar quais métricas de avaliação de modelos de ML foram aplicadas nos experimentos realizados.

Por se tratar de uma atualização à pesquisa de Marques (2020), alguns parâmetros adotados pelo autor em sua pesquisa também são adotados no presente estudo. No entanto, visto que alguns objetivos foram alterados, o protocolo também passou por algumas alterações, conforme se apresenta na próxima sub-secção.

4.1 Procedimentos e métodos

Este MSL é fundamentado pelas diretrizes de Petersen (PETERSEN *et al.*, 2008) e Kitchenham (KITCHENHAM, 2004), segundo as quais, o protocolo do mapeamento, que corresponde ao seu planejamento, consiste na definição de questões de pesquisa, horizonte de pesquisa, idioma, critérios de inclusão e exclusão, definição de palavras-chave, *strings* de busca e fontes de busca. A seguir os detalhes do planejamento desta pesquisa são apresentados.

Para este MSL foram definidas as seguintes etapas, conforme ilustrado na figura 17:

Figura 17 – Etapas do MSL



Fonte: autoria própria (2022)

1. definição do protocolo de pesquisa: nesta etapa foram definidas as questões de pesquisa para condução das coletas, eleição das bases de dados, definição dos termos e criação da *string* de busca. Ainda nesta etapa, foram definidos os critérios de exclusão e qualidade e foi elaborado o formulário de extração de dados para análise;
2. seleção de estudos primários: tendo em vista o protocolo estabelecido, procedeu-se à coleta e seleção de estudos de forma criteriosa e sistematizada, passando por 3 fases de seleção;
3. extração de dados: este processo se deu com a leitura dos artigos selecionados e a extração de informações, conforme os atributos definidos no protocolo deste mapeamento;
4. análise dos resultados: os dados inseridos no documento de extração foram consolidados em forma de gráficos, com o intuito de possibilitar a análise segundo as questões de pesquisa definidas.

4.1.1 Questões de pesquisa

Por meio desta pesquisa, se idealizou compreender as causas de evasão estudantil que já foram investigadas por outros pesquisadores e os artefatos desenvolvidos como algoritmos e técnicas de MD e ML. Dessa forma, as questões de pesquisa que nortearam este trabalho são:

- **Questão 1 (Q1):** Quais ferramentas de MD têm sido utilizadas para a descoberta das causas da evasão estudantil?
- **Questão 2 (Q2):** Quais técnicas de ML têm sido utilizadas para a descoberta das causas da evasão estudantil?
- **Questão 3 (Q3):** Quais fatores indutores têm sido investigados para a descoberta das causas da evasão estudantil?
- **Questão 4 (Q4):** Quais bases abertas fornecem dados escolares de estudantes?
- **Questão 5 (Q5):** Quais métricas de avaliação de modelos de ML foram aplicadas nos experimentos realizados?

Além das questões formuladas por Marques (2020), este trabalho considerou outros dois pontos: investigar as técnicas de ML que tem sido utilizadas pelos pesquisadores e identificar bases de dados abertas que fornecem dados escolares.

O horizonte desta pesquisa considera estudos realizados após o período já analisado em outros trabalhos que se encerrou em 2018. Dessa forma, no presente estudo, foram mapeadas as pesquisas desenvolvidas entre os anos de 2018 e 2021. Observa-se que as buscas foram conduzidas no mês de janeiro de 2022. Além do horizonte de pesquisa, também foi definido o idioma dos trabalhos buscados. Para este MSL foi definido que serão aceitos e analisados apenas trabalhos em Inglês, por se tratar do idioma oficial para a divulgação de pesquisas científicas.

4.1.2 Critérios de seleção

Os Critério de Inclusão (CI) e Critério de Exclusão (CE) são estabelecidos para auxiliar na condução de um MSL, com a finalidade de contribuir na seleção dos estudos (FUZETO; BRAGA, 2016). Para um trabalho ser selecionado é suficiente atender a, pelo menos, um critério de inclusão. Na seleção dos estudos para este MSL, foram aplicados os

seguintes (CI):

- Critério de Inclusão (CI1): o trabalho discute ou aplica ferramentas de MD relacionadas ao problema da evasão estudantil no ensino presencial;
- Critério de Inclusão (CI2): o trabalho aplica técnicas de ML relacionadas ao problema da evasão estudantil no ensino presencial;
- Critério de Inclusão (CI3): o trabalho apresenta fatores indutores ou dados relacionados ao problema da evasão estudantil no ensino presencial;
- Critério de Inclusão (CI4): o trabalho apresenta bases de dados abertas que forneçam dados escolares de estudantes;
- Critério de Inclusão (CI5): o trabalho apresenta as métricas de avaliação de modelos de ML que foram aplicadas nos experimentos realizados durante a pesquisa.

Por outro lado, para se excluir um trabalho é suficiente que ela atenda a um dos critérios de exclusão definidos a seguir.

- Critério de Exclusão (CE1): trabalhos que utilizam MD ou ML no contexto escolar, mas não abordam a evasão estudantil;
- Critério de Exclusão (CE2): estudos que não tratem das causas da evasão estudantil no ensino presencial;
- Critério de Exclusão (CE3): artigos em outros idiomas que não o Inglês;
- Critério de Exclusão (CE4): em casos de estudos duplicados apenas o mais recente será aceito.

4.1.3 Critérios de qualidade

Os Critérios de Qualidade (CQ) têm como finalidade avaliar aspectos metodológicos dos estudos, tais como, a relevância do tema de pesquisa e a adoção de métodos que conduzam aos objetivos propostos (KITCHENHAM, 2004). Observa-se que os critérios de qualidade não excluem trabalhos. Os artigos selecionados foram avaliados de acordo com 02 critérios de qualidade, apresentados a seguir.

- Critério de Qualidade (CQ1): trabalhos publicados em veículos de divulgação científica com Fator de Impacto (FI) entre 0,1 e 0,8 (meio ponto), entre 0,9 e 1,9 (1 ponto), entre 2,0 e 2,9 (2 pontos), entre 3,0 e 5,0 (3 pontos) e maior que 5,0 (5 pontos).
- Critérios de Qualidade (CQ2): o trabalho responde às questões de pesquisa (QP).

Será adicionado 1 ponto para cada QP que o trabalho responder.

4.1.4 Definição das Palavras-chave, String de Busca e Fontes

A partir das questões de pesquisa, foram definidos as palavras-chave e respectivos sinônimos dispostos na Tabela 5. Com base nas palavras-chave e seus respectivos sinônimos, se definiu a seguinte *string* de busca: ((*“Data Mining” OR “Machine Learning” OR “artificial intelligence” OR “Deep Learning” OR “Knowledge discovery” OR “data extraction” OR prediction OR detection OR “database”*) AND (*“school dropout” OR “school evasion” OR “school retention” OR “school truancy” OR “school failure”*)).

Tabela 5 – Termos de busca

Palavra	Sinônimo em Inglês
Data Mining	<i>Knowledge Discovery, Data Extraction, Predicting, Prediction, Detection, Detecting, Database, Machine Learning, Deep Learning, Artificial Intelligence</i>
School Dropout	<i>School Drop-Out, School Evasion, School Retention, School Truancy, School Failure</i>

Para se chegar a *string* utilizada, foram realizados testes, nos quais foram aplicadas diferentes combinações dos termos supracitados e verificados quais destas combinações retornavam o maior número de trabalhos para cada uma das fontes de busca. Os testes resultaram na adição de *“Machine Learning, Artificial Intelligence e Deep Learning”* por se tratarem de técnicas que vêm sendo utilizadas em trabalhos relacionados a evasão estudantil e que retornaram uma maior quantidade de trabalhos.

As bases de dados eletrônicas selecionadas para este MSL são: *IEEE Xplore* (<https://ieeexplore.ieee.org/Xplore/home.jsp>), *ACM Digital Library* (<https://dl.acm.org/>), *Science Direct* (<https://www.sciencedirect.com/>), *Scopus* (<https://www.scopus.com/home.uri>), e Portal de Periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) ([www-periodicos-capes-gov-br](http://www.periodicos-capes.gov.br)). Essas bases foram escolhidas, pois têm se consolidado e se tornaram referências no campo da Ciência da Computação.

4.1.5 Processo de seleção

O processo de seleção dos estudos foi dividido em três fases: coleta de trabalhos, seleção inicial e seleção final. As fases de seleção dos estudos são descritas a seguir.

- **Fase 1 - Coleta de trabalhos:** coleta dos trabalhos relacionados aos termos de busca,

retornados em cada base de dados eletrônicas.

- **Fase 2 - Seleção Inicial:** leitura de títulos e resumos das publicações coletadas na fase anterior e eleição daquelas que satisfazem aos critérios de inclusão. Trabalhos que não se enquadram em nenhum critério de inclusão são excluídos com base em alguns critérios de exclusão.
- **Fase 3 - Seleção Final:** leitura integral dos artigos selecionados na fase anterior e reavaliação dos trabalhos com base na aplicação dos critérios de inclusão e exclusão.

Salienta-se que todas as fases passaram pela validação de outros pesquisadores, bem como foram utilizados 3 artigos como artigos de controle.

Observa-se ainda que, a fim de extrair informações relevantes dos trabalhos selecionados, foi desenvolvido um formulário de extração de dados (Apêndice A).

4.2 Condução do Estudo

Inicialmente, foram realizadas buscas nas bases de dados selecionadas. Observa-se que as *strings* de busca foram adaptadas para cada uma das bases, de modo a se aquedar as suas particularidades, mas sem apresentar diferença semântica em relação à *string* apresentada na seção anterior.

A principal alteração aconteceu na base *Science Direct*, por aceitar apenas 8 palavras por pesquisa. Dessa forma, se fez necessária a realização de três buscas com as seguintes *Strings*:

- ((*"Data Mining"* OR *"Knowledge discovery"* OR *"data extraction"*) AND (*"school dropout"* OR *"school evasion"* OR *"school retention"* OR *"school truancy"* OR *"school failure"*));
- ((*prediction* OR *detection* OR *"database"*) AND (*"school dropout"* OR *"school evasion"* OR *"school retention"* OR *"school truancy"* OR *"school failure"*));
- ((*"Machine Learning"* OR *"Deep Learning"* OR *"artificial intelligence"*) AND (*"school dropout"* OR *"school evasion"* OR *"school retention"* OR *"school truancy"* OR *"school failure"*)).

Na Tabela 6, são apresentadas como as *strings* foram dispostas em cada base de dados, por meio do campo de pesquisa avançada.

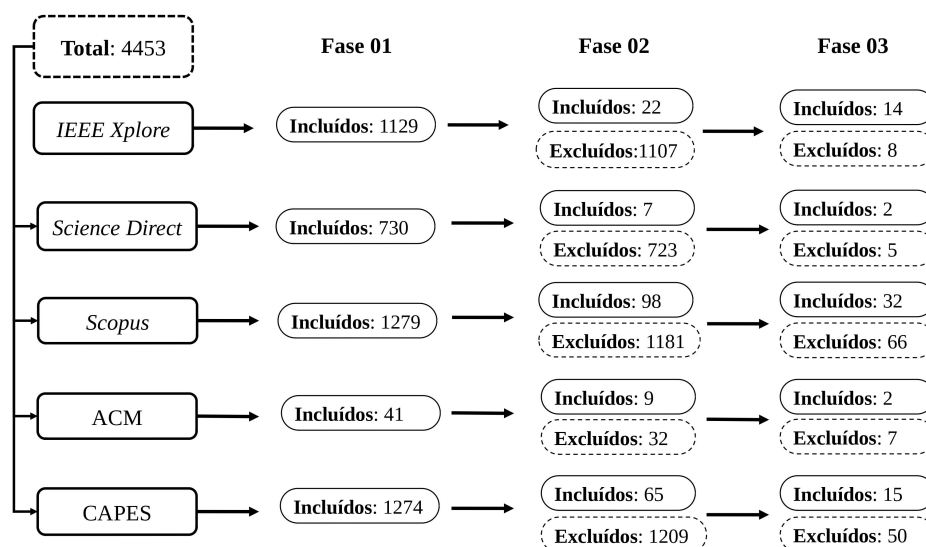
Tabela 6 – Fontes de Pesquisa, strings de busca e retorno

Base	String	Retorno
IEEE XPLORE	((“Data Mining” OR “Machine Learning” OR “artificial intelligence” OR “Deep Learning” OR “Knowledge discovery” OR “data extraction” OR prediction OR detection OR “database”) AND (“school dropout” OR “school evasion” OR “school retention” OR “school truancy” OR “school failure”))	1.129
ACM Digital Library	((“Data Mining” OR “Machine Learning” OR “artificial intelligence” OR “Deep Learning” OR “Knowledge discovery” OR “data extraction” OR prediction OR detection OR “database”) AND (“school dropout” OR “school evasion” OR “school retention” OR “school truancy” OR “school failure”))	41
Science Direct	Busca 1: ((“Data Mining” OR “Knowledge discovery” OR “data extraction”) AND (“school dropout” OR “school evasion” OR “school retention” OR “school truancy” OR “school failure”)) Busca 2: ((prediction OR detection OR “database”) AND (“school dropout” OR “school evasion” OR “school retention” OR “school truancy” OR “school failure”)) Busca 3: ((“Machine Learning” OR “Deep Learning” OR “artificial intelligence”) AND (“school dropout” OR “school evasion” OR “school retention” OR “school truancy” OR “school failure”))	730
Scopus	((“Data Mining” OR “Machine Learning” OR “artificial intelligence” OR “Deep Learning” OR “Knowledge discovery” OR “data extraction” OR “prediction” OR “detection” OR “database”) AND (“school dropout” OR “school evasion” OR “school retention” OR “school truancy” OR “school failure”))	1.279
CAPES	((“Data Mining” OR “Machine Learning” OR “artificial intelligence” OR “Deep Learning” OR “Knowledge discovery” OR “data extraction” OR prediction OR detection OR “database”) AND (“school dropout” OR “school evasion” OR “school retention” OR “school truancy” OR “school failure”))	1.274

Para cada retorno (buscas 1, 2 e 3), coletaram-se os trabalhos retornados e, ao final, houve a junção destes trabalhos, excluindo as repetições.

As buscas nas bases de dados selecionadas com as strings adaptadas resultaram em um total de 4453 trabalhos retornados. Na fase 2, os artigos foram selecionados aplicando os critérios de inclusão e exclusão, a partir da leitura dos títulos e resumos dos artigos coletados na fase anterior. Ao final desta etapa, foram incluídos 201 trabalhos. Na última fase da seleção, os artigos selecionados na fase anterior foram lidos integralmente e os critérios de inclusão e exclusão foram reaplicados. Na Figura 18 são exibidos os números de artigos incluídos e excluídos para cada plataforma, em cada uma das fases.

Figura 18 – Quantidade de trabalhos avaliados em cada etapa do mapeamento



Fonte: autoria própria (2022)

Após a seleção final, os dados apresentados pelos trabalhos foram coletados e usados para responder às questões de pesquisa, conforme se apresenta na sub-seção anterior.

O processo de avaliação dos trabalhos ocorreu da seguinte forma: 4 pesquisadores participaram do processo de avaliação de trabalhos. Na etapa de avaliação dos resumos, o filtro inicial foi feito por dois pesquisadores que apresentaram os dados para debate e discussão com os outros pesquisadores. Na etapa de leitura completa, um pesquisador fez a leitura de todos os trabalhos e apresentou a tabela de seleção com todas as justificativas para discussão com os outros pesquisadores. Além disso, 72 trabalhos geraram dúvidas na sua avaliação e foram lidos pelos 4 pesquisadores. Todos os trabalhos incluídos foram debatidos e adicionados apenas após a confirmação da unanimidade por parte dos 4 pesquisadores.

Os trabalhos que geraram dúvidas foram debatidos pela equipe de pesquisa e enquadrados como incluído ou excluído, após decisão por consenso. Nesta fase, foram selecionados 65 artigos, conforme apresentados nas Tabelas 7 e 8.

Tabela 7 – Trabalhos Selecionados

ID	Título	Autor(es)
#1	Analyze and Predicting the Student Academic Performance Using Data Mining Tools	(DEEPIKA; SATHVANARAVANA, 2018)
#2	A Comparison of Logistic Regression and Linear Discriminant Analysis in Predicting of Female Students Attrition from School in Bangladesh	(HASAN, 2019)
#3	Determination of Dropout Student Profile Based on Correspondence Analysis Technique	(BARROS <i>et al.</i> , 2019b)
#4	Machine learning model for detecting high school students as candidates for drop-out from a study program	(PAŠIĆ; KUČAK, 2020)
#5	Optimization of Weight Backpropagation with Particle Swarm Optimization for Student Dropout Prediction	(SARI <i>et al.</i> , 2019)
#6	Predicting Dropout in Higher Education Based on Secondary School Performance	(NAGY; MOLONTAY, 2018)
#7	Predicting Dropout Using High School and First-semester Academic Achievement Measures	(KISS <i>et al.</i> , 2019)
#8	Predicting Louisiana Public High School Dropout through Imbalanced Learning Techniques	(OROOJI; CHEN, 2019)
#9	Prediction of student's retention in first year of engineering program at a technological chilean	(DOMBROVSKAIA <i>et al.</i> , 2020)
#10	Student Dropout Prediction: A KMUTT Case Study	(TENPIPAT; AKKARAJITSAKUL, 2020)
#11	Dropout early warning systems for high school students using machine learning	(CHUNG; LEE, 2019)
#12	Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country	(CRUZ-JESUS <i>et al.</i> , 2020)
#13	Early Warning Methodology for dropping out of university degrees	(MECA <i>et al.</i> , 2020)
#14	University Dropout Prevention through the Application of Big Data	(SHIAU, 2020)
#15	Early Dropout Prediction Model: A Case Study of University Leveling Course Students	(SANDOVAL-PALIS <i>et al.</i> , 2020)
#16	Early Multi-criteria Detection of Students at Risk of Failure	(ILIEVA; YANKOVA, 2020)
#17	IoT System for School Dropout Prediction Using Machine Learning Techniques Based on Socioeconomic Data.	(FREITAS <i>et al.</i> , 2020)
#18	Machine learning approach for reducing students dropout rates	(MDUMA <i>et al.</i> , 2019)
#19	Predictive Models for Imbalanced Data: A School Dropout Perspective	(BARROS <i>et al.</i> , 2019a)
#20	Using academic analytics to predict dropout risk in engineering courses	(LIMA <i>et al.</i> , 2018)
#21	University dropout: Prevention patterns through the application of educational data mining	(NÁJERA <i>et al.</i> , 2020)
#22	University dropout prediction through educational data mining techniques: A systematic review	(AGRUSTI <i>et al.</i> , 2019)
#23	The machine learning-based dropout early warning system for improving the performance of dropout prediction	(LEE, 2005)
#24	Supervised learning in the context of educational data mining to avoid university students dropout	(KELLY <i>et al.</i> , 2019)
#25	University Dropout: A Prediction Model for an Engineering Program in Bogotá, Colombia	(ACERO <i>et al.</i> , 2019)
#26	Student dropout prediction	(BITTENCOURT <i>et al.</i> , 2020)
#27	The data revolution comes to higher education: identifying students at risk of dropout in Chile	(HIPPEL; HOFFLINGER, 2021)
#28	Student dropout from foundation program at modern college of business & science, Sultanate of Oman	(ERIC, 2019)
#29	Risk Factors for Dropping out of High School: A Review of Contemporary, International Empirical Research	(RIPAMONTI, 2018)
#30	Predicting the Risk of Academic Dropout with Temporal Multi-Objective Optimization	(JIMENEZ <i>et al.</i> , 2019)
#31	Predicting the Enrollment and Dropout of Students in the Post-Graduation Degree using Machine Learning Classifier	(BISWAS <i>et al.</i> , 2019)
#32	Predicting students' academic drop out and failures using data mining techniques	(VENKATESAN <i>et al.</i> , 2019)

Tabela 8 – Trabalhos Selecionados

ID	Título	Autor(es)
#33	Predicting student dropouts in higher education using supervised classification algorithms	(SERRA <i>et al.</i> , 2018)
#34	Predicting school dropout with administrative data: new evidence from Guatemala and Honduras	(ADELMAN <i>et al.</i> , 2018)
#35	Knowledge Discovery in Educational Databases in Indian Educational System: A Case Study of GHRCE, Nagpur	(BAHEL <i>et al.</i> , 2019)
#36	From Lab to Production: Lessons Learnt and RealLife Challenges of an Early Student-Dropout Prevention System	(ORTIGOSA <i>et al.</i> , 2019)
#37	Ensemble regression models applied to dropout in higher education	(SILVA <i>et al.</i> , 2019)
#38	Educational data mining: An application of regressors in predicting school dropout	(NASCIMENTO <i>et al.</i> , 2018)
#39	Dropout prediction system to reduce discontinue study rate of information technology students	(LIMSATHITWONG <i>et al.</i> , 2018)
#40	Beyond Early Warning Indicators: High School Dropout and Machine Learning	(SANSONE, 2018)
#41	Analytical approach for predicting dropouts in higher education	(JAISWAL <i>et al.</i> , 2019)
#42	Analysis of Student-at-Risk of Dropping out (SARDO) Using Decision Tree: An Intelligent Predictive Model for Reduction	(TIMBAL, 2019)
#43	A survey of machine learning approaches and techniques for student dropout prediction	(MDUMA <i>et al.</i> , 2019)
#44	“Big Data” in Educational Administration: An Application for Predicting School Dropout Risk	(SORENSEN, 2019)
#45	Classification and predictive analysis of educational data to improve the quality of distance learning courses	(BRANDÃO <i>et al.</i> , 2019)
#46	Information Systems Students’ Study Performance Prediction Using Data Mining Approach	(GUNAWAN, 2019)
#47	Students Disruption in Business Computer at the University of Phayao, Thailand	(NUANKAEW <i>et al.</i> , 2020)
#48	Parenting practices during middle adolescence and high school dropout	(AFIA <i>et al.</i> , 2019)
#49	Preliminary validation of the dropout risk inventory for middle and high school students	(VAUGHN <i>et al.</i> , 2020)
#50	An analytical investigation of the characteristics of the dropout students in higher education	(MORTAGY <i>et al.</i> , 2018)
#51	Dropout at Danish vocational schools: does the school’s health promotion capacity play a role? A survey- and register-based prospective study	(THORNBURG, 2001)
#52	School connectedness and high school graduation among maltreated youth	(LEMKIN <i>et al.</i> , 2018)
#53	WWH-dropout scale: when, why and how to measure propensity to drop out of undergraduate courses	(LEMKIN <i>et al.</i> , 2020)
#54	Who intends to learn and who intends to leave? The intention to leave education early among students from inclusive and regular classes in primary and secondary schools	(SCHWAB, 2018)
#55	User profiling to overcome the social effects of student dropout	(GILBERT; HAMID, 2019)
#56	The association of identity and motivation with students’ academic achievement in higher education	(MEENS <i>et al.</i> , 2018)
#57	Exploring young people’s perceptions about secondary school: Critical issues and improvements to prevent	(BUSS <i>et al.</i> , 1987)
#58	Exploring the factors associated with the school dropout	(BOYACI, 2019)
#59	Dropout and transfer paths: What are the risky profiles when analyzing university persistence with machine learning techniques?	(RODRÍGUEZ-MUÑOZ <i>et al.</i> , 2019)
#60	Classification of province based on dropout rate using C4.5 algorithm	(PERTIWI <i>et al.</i> , 2017)
#61	A longitudinal study of factors predicting students intentions to leave upper secondary school in Norway	(HAUGAN <i>et al.</i> , 2019)
#62	‘There is no free education nowadays’: youth explanations for school dropout in Indonesia	(ZUILKOWSKI <i>et al.</i> , 2019)
#63	A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data	(FERNÁNDEZ-GARCÍA <i>et al.</i> , 2021)
#64	Analysis of First-Year University Student Dropout through Machine Learning Models: A Comparison between Universities	(OPAZO <i>et al.</i> , 2021)
#65	Patterns of persistence among engineering students at a South African university: A decision tree analysis	(BENGESAI; POCKOCK, 2021)

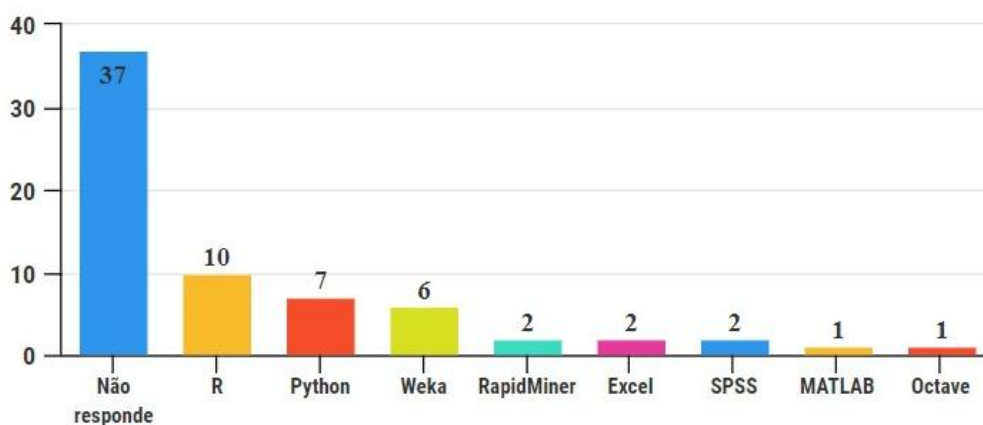
4.3 Resultados e Discussões

Nesta seção são apresentadas as respostas para as questões de pesquisa, baseadas no estudo dos trabalhos selecionados. A análise dessas questões é apresentada a seguir.

1 - Quais ferramentas de MD têm sido utilizadas para a descoberta das causas da evasão estudantil?

Na Figura 19 é representado o quantitativo de uso das ferramentas de MD nos trabalhos selecionados. É possível verificar que o software R, Python e Weka se destacam pela utilização em 10 (15%), 7 (11%) e 6 (10%) trabalhos, respectivamente. As ferramentas, *RapidMiner*, *Excel*, *SPSS*, foram utilizadas 2 (6%) trabalhos cada. Enquanto as ferramentas *MATLAB* e *Octave* foram utilizadas em apenas 1 (3%) trabalho, respectivamente.

Figura 19 – Ferramentas de MD



Fonte: autoria própria (2022)

Observa-se que, 37 (55%) dos trabalhos selecionados não respondem a esta pergunta. Acredita-se que o motivo pelos quais os artigos selecionados não mencionam as ferramentas de MD utilizadas é pelo fato de que alguns destes trabalhos não aplicam a MD de forma prática, se tratando de pesquisas bibliográficas acerca da temática da EDM aplicada à evasão estudantil. Infelizmente, isso dificulta um princípio importante do método científico que é a reprodutibilidade.

O *software* R foi utilizado nos trabalhos: #2, #8, #11, #13, #20, #22, #23, #25, #36, e #42. O R é um ambiente de *software* de licença livre, projetado para executar computação estatística e gráficos. O R é amplamente utilizado na academia, pesquisa, engenharia e aplicações industriais. O *software* possui um grande número de pacotes com modelos, fórmulas e testes estatísticos, auxiliando na análise de dados (GERMAN *et al.*, 2013).

A linguagem de programação *Python* foi utilizada nos trabalhos #12, #17, #19, #24, #26, #63 e #64. Esta linguagem é de licença livre e além de ser utilizada para MD, pode ser utilizada para o uso geral, como: *data science*, *machine learning*, desenvolvimento de aplicativos, automação de *scripts* e desenvolvimento de aplicações *web* (STANČIN; JOVIĆ, 2019).

A ferramenta *Weka* foi utilizada nos trabalhos #1, #22, #32, #45, #46, #59 e consiste em uma coleção de algoritmos de aprendizado de máquina para tarefas de MD, de licença livre. A ferramenta *Weka* é aplicada para MD, para processar *big data* e *deep learning* (RATRA; GULIA, 2020).

O *Statistical Package for the Social Sciences (SPSS)* foi utilizado pelos trabalhos #22 e #65 e consiste em um pacote de software aplicado à análise estatística interativa (SOFYAN; KURNIAWAN, 2009).

A plataforma de ciência de dados baseada na nuvem para profissionais de dados *RapidMiner* (NAIK; SAMANT, 2016), foi utilizada nos trabalhos de #6 e #39. O editor de planilhas *Microsoft Excel* foi utilizado por #20 e #50.

O MATLAB é um software interativo de alto desempenho voltado para o cálculo numérico, que também é utilizado para explorar, modelar e analisar dados (ELHORST, 2014). Este, foi utilizado no trabalho #5. O *Octave*, utilizado pelo trabalho #4, é uma linguagem computacional, desenvolvida para computação matemática, que na MD é utilizado para Regressão Linear (SILVA; MOODY, 2014).

Constata-se que as ferramentas R, *Weka* e *Python* são as mais populares na área de MD, pois além de disponibilizarem licença livre, com ampla disponibilização de documentação e resolução de problemas por meio de comunidades.

2 - Quais técnicas de ML têm sido utilizadas para a descoberta das causas da evasão estudantil?

As técnicas de predição frequentemente utilizadas pela EDM são a classificação e a regressão (COSTA *et al.*, 2013). Estas foram as técnicas que predominaram nos trabalhos selecionados na presente pesquisa.

A técnica de classificação foi utilizada em 67% dos trabalhos selecionados. Os trabalhos foram: #1, #2, #4, #5, #6, #7, #8, #9, #10, #11, #12, #13, #14, #15, #17, #19, #20, #21, #22, #23, #24, #25, #26, #27, #30, #31, #32, #33, #35, #36, #39, #40, #41, #42, #43, #44, #45, #46, #50, #58, #59, #60, #63 e #65. Essa técnica consiste em associar

objetos a determinadas classes, objetivando prever automaticamente a qual classe um novo dado pode ser associado. Um exemplo aplicado à evasão estudantil, consiste em analisar múltiplas variáveis que estão relacionadas as características dos estudantes e associá-las a classes de estudantes que evadem ou permanecem no curso.

A Regressão foi explorada em 16% dos experimentos. São eles: #1, #7, #13, #16, #18, #34, #35, #37, #38, #41, #48. A Regressão é uma técnica de MD que busca modelar o relacionamento de variáveis independentes (chamadas preditoras) com uma variável dependente (chamada resposta). As variáveis preditoras são os atributos dos registros, e a resposta é o que se quer predizer.

A análise de correspondência, abordada por #3, #14, o que corresponde a 3% dos trabalhos selecionados, é uma técnica de análise multivariada, adequada para dados categóricos, que permite analisar graficamente as relações existentes por meio da redução de dimensionalidade do conjunto de dados.

A análise fatorial, utilizada pelo trabalho #50 (1.5%), é uma técnica da estatística multivariada que objetiva a redução do número de variáveis iniciais em uma determinada análise, minimizando a perda de informação. Essa técnica cria um grupo reduzido de fatores comuns não observados a partir da interpretação das inter-relações existentes no conjunto de variáveis observadas.

Costa (2013) determina que as técnicas de EDM são categorizadas mediante as sub-áreas da MD e que a sub-área de predição tem se destacado porque seus métodos podem ser utilizados para estudar quais aspectos de um modelo são importantes para predição.

Observa-se que parte dos trabalhos analisados apresentam a combinação de técnicas de classificação e regressão (Figura 20). Isso acontece porque cada técnica apresenta particularidades e essa combinação depende do problema a ser trabalhado. Na Tabela 9 se apresentam algumas características que distinguem as técnicas de classificação e regressão.

Figura 20 – Técnicas de MD



Fonte: autoria própria (2022)

Tabela 9 – Características dos algoritmos mais utilizados nos trabalhos selecionados

Base para comparação	Classificação	Regressão
Comportamento	Modela uma função por meio da qual os dados são previstos em rótulos de classe discretos.	Modelo criado no qual o mapeamento de objetos é feito em valores contínuos.
Valores de previsão	Valores discretos	Valores contínuos.
Algoritmos	Árvore de decisão, regressão logística, etc.	Árvore de regressão (floresta aleatória), regressão linear, etc.
Natureza dos dados previstos	Não ordenado.	Pedido.
Método de cálculo	Precisão de medição.	Medição do erro quadrático médio da raiz.

Fonte: autoria própria (2022)

Observa-se também que alguns algoritmos de ML podem ser utilizados tanto para problemas de Classificação quanto para problemas de Regressão. Na Figura 21 é ilustrada uma nuvem de palavras que relaciona os principais algoritmos utilizados pelos trabalhos selecionados e na Tabela 10 são apresentados os algoritmos que mais se destacaram na nuvem de palavras, suas características e as técnicas que os mesmos pertencem.

Figura 21 – Principais algoritmos utilizados para MD



Fonte: autoria própria (2022)

Tabela 10 – Características dos algoritmos mais utilizados nos trabalhos selecionados

Algoritmo	Definição	Utilizado na Técnica
Decision Tree	Usado para prever a classe ou o valor da variável de destino, aprendendo regras de decisão simples, inferidas de dados anteriores (dados de treinamento).	Regressão e Classificação
Logistic Regression	Usado para classificar os dados em duas ou mais classes	Classificação
Randon Forest	Combina a saída de várias árvores de decisão para chegar a um único resultado	Regressão e Classificação
Artificial Neural Network	Imita as operações do cérebro humano para reconhecer as relações entre grandes quantidades de dados	Classificação e Regressão
Naive Bayes	Funciona com base no teorema de probabilidade de Bayes para prever a classe de conjuntos de dados desconhecidos.	Classificação
Support Vector Machine (SVM)	Funciona criando uma linha ou um hiperplano que separa os dados em classes	Regressão e Classificação

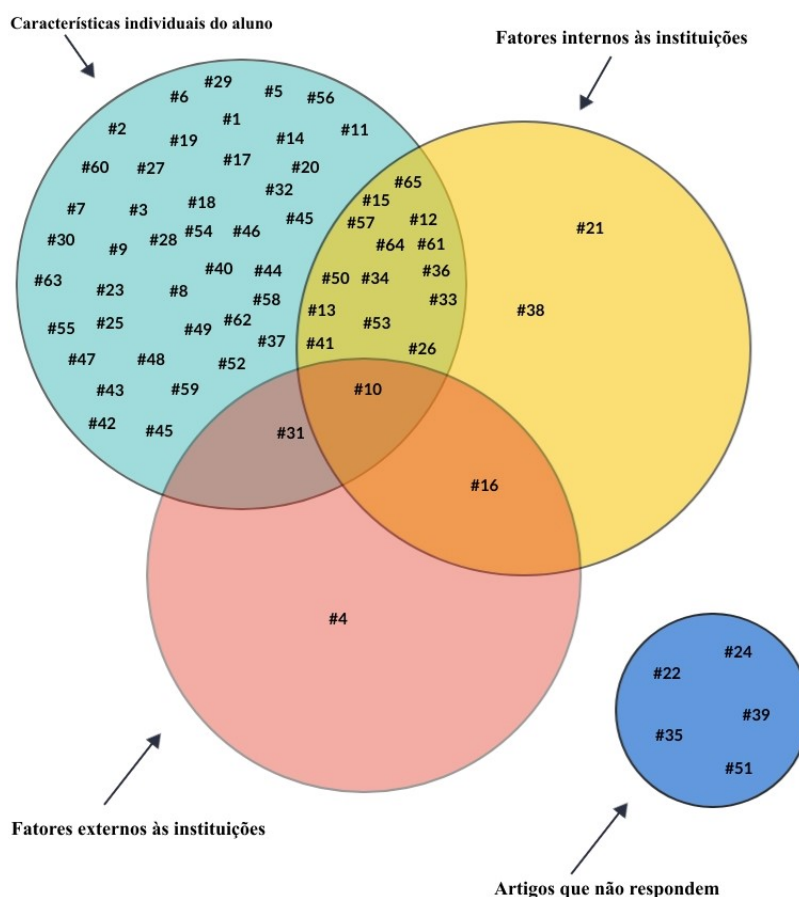
3 - Quais fatores indutores têm sido investigados para a descoberta das causas da evasão estudantil?

Na Tabela 11 são exibidos os fatores indutores e principais variáveis relacionadas ao risco de evasão estudantil, representados pelos identificadores (IDs) dos trabalhos selecionados para este mapeamento e, na Figura 11, são ilustradas as categorias dos fatores indutores nas causas da evasão estudantil investigados em cada trabalho. Por meio da figura é possível visualizar que, parte dos trabalhos aborda a investigação de mais de uma categoria de fatores indutores, mas apenas 1 trabalho investiga a evasão considerando as três categorias de fatores indutores.

Tabela 11 – Características dos algoritmos mais utilizados nos trabalhos selecionados

Fator	ID	Principais variáveis
Características Individuais do estudante	#1, #2, #3, #7, #8, #9, #10, #12, #13, #14, #15, #17, #18, #19, #20, #23, #25, #26, #28, #29, #31, #33, #34, #36, #37, #38, #42, #43, #45, #46, #47, #48, #49, #50, #51, #53, #54, #55, #56, #57, #58, #59, #60, #62, #63	gênero, idade, cor, renda, quantidade de pessoas na mesma casa, se vive com os pais, escolaridade dos pais, estado civil, possuir filhos, uso de substâncias, doenças, oriundo de escola pública, frequência nas aulas, faltas, reprovações, notas de português e matemática no ensino médio, médias no ensino médio, engajamento na escola, realiza atividades extracurriculares,
Interno a instituição	#10, #12, #13, #15, #16, #21, #26, #33, #34, #36, #37, #39, #42, #51, #54, #58, #62	recursos do corpo docente, presença de mentoria ou aconselhamento, ambiente escolar adequado, acompanhamento acadêmico, programas de apoio, oferece atividades extracurriculares,
Externo a Instituição	#4, #10, #16, #31	situação profissional, ensino anterior de baixa qualidade, pouca oferta no mercado de trabalho

Figura 22 – Fatores indutores investigados nas causas da evasão estudantil



Fonte: autoria própria (2022)

Os resultados desta questão constataam que, os principais fatores indutores investigados para a descoberta das causas da evasão estudantil ainda são relacionados às características individuais do estudante, analisando dados armazenados nas instituições ou levantando dados por meio de questionários, sob a perspectiva do estudante. Fatores

relacionados às características internas à instituição vêm sendo combinadas com as características individuais para traçar um perfil do estudante. Fatores externos à instituição são pouco considerados quando comparado aos demais fatores.

4 - Quais bases abertas fornecem dados escolares de estudantes?

Dos 65 trabalhos selecionados, apenas 15 (23%) atuaram com bases abertas. A maioria dos trabalhos são realizados com dados de estudantes de uma determinada instituição e por se tratar de dados pessoais, estes não são disponibilizados.

Na Tabela 12 são apresentados os IDs dos trabalhos e as respectivas bases abertas utilizadas pelos mesmos.

Tabela 12 – Bases de Dados abertas

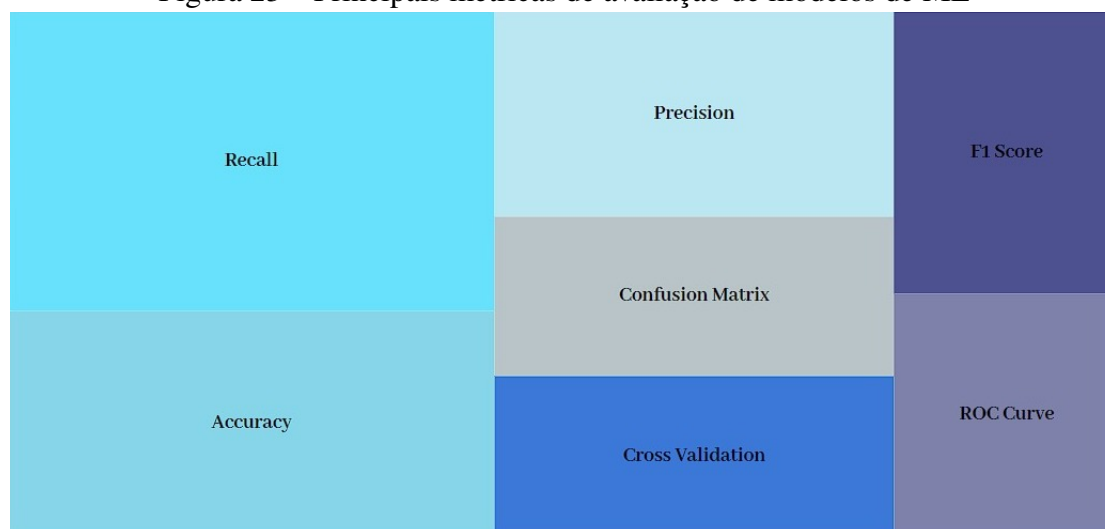
ID	Base de Dados
#1	<i>Dataset from the University of California at Irvine (UCI) Machine Learning Repository.</i>
#2	Fundo das Nações Unidas para a Infância (UNICEF)
#8	<i>Social Research and Evaluation Center da Louisiana State University</i>
#10	<i>KMUTT Registrar's Office</i>
#11	<i>National Education Information System (NEIS) in Korea</i>
#12	Conjunto de dados anonimizado fornecido pela Direcção-Geral de Estatística da Educação e Ciência (DGEEC) do Ministério da Educação de Portugal
#18	Dados de transcrição da <i>George Mason University (GMU)</i> , dados de transcrição da Universidade de Minnesota (UMN), dados de UMNLMs e dados de MOOC da <i>Stanford University</i>
#23	<i>National Education Information System (NEIS) da Coréia do Sul</i>
#24	<i>Department of Computer Science in the UFS</i>
#27	Departamento de Avaliação, Medição e Registros Educacionais (DEMRE) - Chile
#34	Seis conjuntos de dados retirados da coleção UCI
#37	Censo da Educação Superior e os Indicadores de Fluxo da Educação Superior. Ambos estão disponíveis no Instituto Nacional de Estudos e Pesquisas Educacionais (INEP) 2013
#38	INEP, referentes aos microdados do Censo Escolar e Taxa de Desempenho Escolar, ambos do ano de 2016
#44	Dados administrativos do Departamento de Instrução Pública da Carolina do Norte (NC DPI)
#58	Dados do <i>Family Structure Survey (FSS)</i> coletados em 2016 de 35.475 membros da família com mais de 15 anos de idade, pelo Instituto de Estatística da Turquia (<i>TurkStat</i>).

5 - Quais métricas de avaliação de modelos de ML foram aplicadas nos experimentos realizados?

Métricas de avaliação são utilizadas com a finalidade de avaliar a qualidade do modelo por meio da validação do desempenho do modelo aplicado. A avaliação do modelo pode determinar se ele terá bom desempenho na previsão do destino em dados novos e futuros.

Dos 65 trabalhos selecionados, 61% (40 trabalhos) aplicam uma ou mais métricas para avaliar os modelos de ML. Na Figura 23 são ilustradas as métricas que mais foram aplicadas nos trabalhos selecionados.

Figura 23 – Principais métricas de avaliação de modelos de ML



Fonte: autoria própria (2022)

A métrica *Recall* ou *Sensibility* (DOMBROVSKAIA *et al.*, 2020), foi aplicada em 16 (26%) trabalhos: #4, #7, #9, #10, #12, #17, #21, #25, #31, #33, #34, #40, #41, #46 e #60.

A métrica *Accuracy* (DOMBROVSKAIA *et al.*, 2020) foi aplicada em 12 (18%) trabalhos selecionados: #7, #10, #12, #15, #17, #21, #24, #25, #31, #33, #34 e #60.

A métrica *Precision* foi aplicada nos trabalhos #4, #7, #8, #10, #17, #21, #31, #46 e #60, correspondem a 13% dos trabalhos selecionados.

A *F1 Score* foi utilizada em 10% dos trabalhos selecionados: #4, #9, #10, #17, #31, #40 e #46.

A *Cross Validation* (STANČIN; JOVIĆ, 2019) foi aplicada nos trabalhos #5, #6, #8, #15, #24, #37 e #40, correspondendo a 10% dos trabalhos selecionados.

A métrica de *Confusion Matrix* (STANČIN; JOVIĆ, 2019) foi aplicada pelos trabalhos #10, #11, #13, #21, #41, #47 e #61, correspondendo a 10% dos trabalhos selecionados aplicando esta métrica.

É importante destacar que, por meio da matriz de confusão, é possível calcular os diferentes parâmetros do modelo, como a exatidão, precisão, dentre outros. Desse modo, a matriz de confusão pode servir como parâmetro para outras métricas como a *recall* e *precision*.

Receiver Operating Characteristic Curve - ROC foi utilizada em 6 (9%) trabalhos: (#6, #8, #11, #12, #15 e #23).

Salienta-se que no capítulo 3 todas as principais métricas de avaliação de algoritmos foram fundamentadas.

Outras métricas foram aplicadas em um menor número de trabalhos: *Probability of Correct Classification - PCC* foi aplicada no trabalho #1 (correspondendo a 1.5% dos trabalhos selecionado); *Root Mean Square Error (RMSE)* também foi aplicada apenas no trabalho #1; *Area under the receiver operating characteristic (AUROC)*, foi aplicada no trabalho #2; *Missclassification Error (ME)*, aplicada pelo trabalho #4; *Sensitivity* aplicada nos trabalhos #10, #31, #33 e #34 (4% dos trabalhos selecionados); *Lift* utilizada nos trabalhos #12 e #27 (3% dos trabalhos selecionados); *Specificity*, *False Positive Rating - FPR*, *False Negative Rating - FNR*, ambas foram aplicadas pelo trabalho #31; *Overall* foi utilizada apenas pelo trabalho #34; *Absolute Mean error (MAE)*, foi aplicada nos trabalhos #38 e #39; *Mean Square Error (MSE)* foi aplicada nos trabalhos #38 e #41.

Observa-se que: (I) a maioria dos trabalhos aplicou duas ou mais métricas de avaliação nos experimentos realizados; (II) os trabalhos #16, #22, #28, #29, #30, #35, #36, #42, #44, #48, #49, #50, #51, #52, #53, #54, #55, #56, #57, #58, #62 e #63 não respondem à pergunta de pesquisa por não aplicarem de forma prática os algoritmos de ML, bem como métricas de avaliação. Estes trabalhos tratam de revisões sistemáticas, estudos exploratórios ou estudos de caso. O #3 não responde à pergunta por se tratar de um estudo de visualização de dados que não aplica algoritmos de ML; (III) por fim, os trabalhos #14, #20, #32, #45 e #59 não informam a utilização de métricas de avaliação de modelos de ML nos experimentos realizados.

Ao final da extração de dados, coleta das respostas para as perguntas de pesquisa e aplicação dos critérios de qualidade, se obteve um *ranking* de trabalhos selecionados, com base na pontuação aplicada aos critérios de qualidade, ilustrados na Figura 24.

Figura 24 – *Ranking* das obras selecionadas com base em critérios de qualidade.

POSICÃO	#ID	F.I	CQ1	CQ2	TOTAL
1	#10	3.79	3	5	8
2	#13	3.7	3	5	8
3	#17	2397	3	5	8
4	#8	2.97	2	5	7
5	#11	2393	2	5	7
6	#15	3251	3	4	7
7	#23	2.47	2	5	7
8	#36	3.72	3	4	7
9	#63	3367	3	4	7
10	#5	1.2	1	5	6
11	#6	1.2	1	5	6
12	#7	1.2	1	5	6
13	#12	1857	1	5	6
14	#19	1.3	1	5	6
15	#26	2.60	2	4	6
16	#27	2125	2	4	6
17	#30	3720	3	3	6
18	#44	2140	2	4	6
19	#59	2740	2	4	6
20	#64	2258	2	4	6
21	#65	2197	2	4	6

POSICÃO	#ID	F.I	CQ1	CQ2	TOTAL
22	#1	0.82	0.5	5	5.5
23	#2	0.67	0.5	5	5.5
24	#9	0.44	0.5	5	5.5
25	#32	0.108	0.5	5	5.5
26	#14	0	0	5	5
27	#20	1.20	1	4	5
28	#24	2.60	2	3	5
29	#25	0	0	5	5
30	#37	2.0	2	3	5
31	#45	0	0	5	5
32	#33	0.159	0.5	4	4.5
33	#42	0.243	0.5	4	4.5
34	#46	0.34	0.5	4	4.5
35	#58	0.33	0.5	4	4.5
36	#18	1.20	1	3	4
37	#21	1.20	1	3	4
38	#39	0	0	4	4
39	#48	2.284	2	2	4
40	#60	2.40	2	2	4

POSICÃO	#ID	F.I	CQ1	CQ2	TOTAL
41	#22	0.69	0.5	3	3.5
42	#31	0.6	0.5	3	3.5
43	#38	0.402	0.5	3	3.5
44	#41	0.48	0.5	3	3.5
45	#40	1.79	1	2	3
46	#47	2.4	2	1	3
47	#49	2.393	2	1	3
48	#50	0	0	3	3
49	#51	2.521	2	1	3
50	#52	2.845	2	1	3
51	#3	0.78	0.5	2	2.5
52	#34	0.82	0.5	2	2.5
53	#43	0.358	0.5	2	2.5
54	#54	1.815	1	1	2
55	#56	1.809	1	1	2
56	#61	1.619	1	1	2
57	#62	1.607	1	1	2
58	#28	0.120	0.5	1	1.5
59	#29	0.51	0.5	1	1.5
60	#53	0.061	0.5	1	1.5
61	#57	0.675	0.5	1	1.5
62	#35	0	0	1	1
63	#55	0	0	1	1

Fonte: autoria própria (2022)

Dentre os trabalhos selecionados, os que mais se relacionam com a presente pesquisa são:

- O trabalho de Barros *et.al* (2019b) no qual a pesquisa é desenvolvida por meio de técnica de análise de correspondência para relacionar os motivos de desistências de estudantes e investigar como o uso de técnicas de visualização podem identificar as principais características socioeconômicas e demográficas de estudantes evadidos. Para validar o estudo, foram analisados dados educacionais de estudantes dos cursos de ensino médio integrado de um Instituto Federal no Brasil.
- Por meio de técnicas de análise de correspondência, o trabalho de Lima *et.al* (2018) analisou uma amostra de 1.844 estudantes, entre graduados e evadidos, no período de 2007 a 2015 para analisar perfis e graus de estudantes com base em notas, número

de tentativas e outros indicadores de desempenho. O trabalho também apresentou uma proposta de um modelo baseado em árvores de decisão, objetivando a geração de instruções padronizadas, de fácil interpretação e permitindo a adição de diversos resultados possíveis, contribuindo para o processo de tomada de decisão.

- O trabalho de Deepika *et.al* (2018) se concentra em aplicar diferentes métodos de categorização com base nas características dos estudantes para classificar os estudantes com base em informações pessoais, colhidas pela instituição de ensino. Para tanto, os pesquisadores usaram a ferramenta *Weka* com os algoritmos de classificação e regressão: *Linear Regression*, *Radom Forest* e *Support Vector Machine*. Como pontos positivos desse trabalho destaca-se a utilização de um repositório de dados abertos (*University of California at Irvine (UCI) Machine Learning Repository* [<http://UCI/machinelearning/repository>]), bem como o enfoque em classificar os estudantes com base no desempenho dos estudantes nas disciplinas de Matemática e Português. Como pontos de melhoria, observam-se: a pesquisa não deixa clara as etapas de análise e preparação dos dados, dificultado a reprodutibilidade; a pesquisa não faz uso de métricas conhecidas para avaliar o desempenho dos algoritmos.
- Pašić e Kučak (2020) desenvolveram um modelo de ML a partir de dados coletados de escolas de ensino médio em que os estudantes cursaram antes de ingressar na Universidade, a fim de calcular a probabilidade do estudante não concluir a Universidade. O modelo de ML ainda classifica o estudante que não concluirá e recomenda para este estudante outro programa de estudos. O trabalho foi realizado por meio da ferramenta *Octave* com algoritmos de regressão. Para avaliação dos modelos, os autores utilizaram: *Recall*, *Precision* e *F1-Score*. Apesar dos excelentes resultados alcançados na aplicação dos modelos de ML por meio das métricas de avaliação, dos cálculos a serem realizados pelos algoritmos de regressão e das hipóteses bem definidas, o trabalho deixa dúvidas com relação à preparação dos dados, a quais os algoritmos de regressão foram utilizados e quantas vezes cada algoritmo rodou com a validação cruzada. O número do conjunto de dados (161) e os atributos escolhidos para execução dos modelos (4 atributos) são considerados pequenos.
- No estudo de Sari *et.al* (2019), a previsão de evasão de estudantes foi realizada para os estudantes que tiveram a possibilidade de ultrapassar o período máximo

de estudo. As previsões são feitas a partir de dados no banco de dados acadêmico do estudante, utilizando os dados do índice de desempenho acadêmico de cada semestre e frequência às aulas. Os algoritmos utilizados foram backpropagation (BP) otimizada com *Particle Swam Optimization (PSO)*. A avaliação do modelo de classificação é feita com validação cruzada de 10 vezes: os dados são divididos em dados de treino e teste, treinados e o teste é feito 10 vezes até obter a maior precisão. O trabalho apresenta de forma resumida uma metodologia clara, os dados, algoritmos e métricas utilizadas. Como pontos de melhorias, destaca-se que a pesquisa poderia detalhar melhor a seleção dos atributos que foram submetidos ao modelo na etapa de preparação dos dados; aplicação de uma métrica complementar à acurácia, pois essa significa apenas o acerto global do algoritmo. Acredita-se que, adicionar o resultado da matriz de confusão ao presente estudo traria uma melhor compreensão do comportamento do modelo de ML, porcentagem de previsões corretas e incorretas.

- No trabalho de Nagy e Molontay (2018), foram aplicados e avaliados vários algoritmos de ML para identificar estudantes em risco e prever a evasão de estudantes em programas universitários com base nos dados disponíveis no momento da inscrição (desempenho no ensino médio, dados pessoais). Também apresentaram uma plataforma de apoio à decisão baseada em dados para a diretoria de educação e partes interessadas. Os algoritmos aplicados foram: *Naive Bayes*, *KNN*, *Linear Models* e *Deep Learning*, por meio da ferramenta *RapidMiner* e para avaliar a performance de cada modelo, utilizaram a curva ROC. Dentre os trabalhos relacionados apresentados, o trabalho de Nagy e Molontay (NAGY; MOLONTAY, 2018) apresenta a melhor proposta: maior conjunto de dados, metodologia clara e definida, aplicação de alguns algoritmos de ML, bem como a avaliação por meio das principais métricas de avaliação de desempenho. O trabalho ainda apresenta como diferencial a disponibilização da plataforma de auxílio. Alguns pontos que poderiam ser melhorados tem relação aos resultados das métricas que alcançaram entre 50% e 80% de acerto. Acredita-se que, mais testes com os demais atributos disponíveis, bem como alterações nos parâmetros de configuração dos classificadores elevariam os resultados da pesquisa.

Diante dos trabalhos citados, este trabalho tem como principal contribuição propor a utilização de análise gráfica com métodos estatísticos para medir o nível de associação

entre as variáveis coletadas, de modo a confirmar as hipóteses que guiam o estudo, tem foco exclusivo em algoritmos de categorias supervisionados e que são aplicados para problemas de classificação que é o mais adequado para o problema da evasão estudantil. Destaca-se ainda que, o presente estudo se preocupa em aplicar diferentes métricas de avaliação em cada um dos modelos desenvolvidos e aplicados em diferentes cursos de ensino superior. Além de todos os pontos mencionados, pretende-se aplicar e apresentar uma metodologia clara e definida, para poder ser replicada, garantindo assim a reprodutibilidade da pesquisa.

4.4 Considerações Finais

Neste capítulo foram apresentadas as principais referências de trabalhos relacionados com o objetivo desta pesquisa, coletadas a partir da condução de um MSL acerca da evasão estudantil, em que se buscou elencar ferramentas, técnicas, fatores indutores e bases de dados abertas que foram utilizadas para predição das causas do problema de evasão, nos dois últimos anos. Buscaram-se trabalhos a partir de 2018, até o ano de 2021.

A busca pelos trabalhos resultou na pré-seleção de 4453 trabalhos, dentre os quais sessenta e cinco (65) foram incluídos para a extração de dados. A partir dos resultados, evidenciou-se que a ferramenta R se consolida como uma das mais utilizadas para a descoberta das causas da evasão estudantil, por meio da aplicação de técnicas e algoritmos nos conjuntos de dados. Com esse pacote, se encontram modelos, fórmulas e testes estatísticos que auxiliam na análise de dados.

Verificou-se que, a técnica de classificação vem sendo bastante utilizada nas previsões de tendência de evasão estudantil. Os algoritmos de classificação mais utilizados nos trabalhos foram: *decision tree*, *logistic regression*, *random forest*, *naive bayes*, *support-vector machine*, e *multilayer perceptron*. Para selecionar o algoritmo que melhor resolve determinado problema se faz necessário analisar alguns fatores, como: complexidade, tempo e quantidade de dados. Os algoritmos mencionados se mostram adequados para o problema da evasão estudantil.

Os principais fatores indutores para a evasão estudantil, se relacionam com as características individuais do estudante, analisando dados armazenados nas instituições ou levantando dados por meio de questionários, da perspectiva do estudante. Fatores relacionados às características internas à instituição vêm sendo combinadas com as características individuais para traçar um perfil do estudante. Fatores externos à instituição são pouco

considerados quando comparados aos demais fatores. Faz-se necessária uma investigação minuciosa para descobrir se isso se dá por falta de dados relacionados a estes fatores ou a pouca influência dos mesmos na problemática da evasão estudantil.

Com relação à pergunta de utilização de bases abertas, apenas quinze (15) atuaram com bases abertas. Isso significa que existem poucas fontes de dados abertas para se trabalhar a evasão de estudantes. Boa parte dos pesquisadores atua com dados privados das instituições a qual tem ligação.

Para avaliar os experimentos realizados, a maioria dos trabalhos selecionados (40 trabalhos) aplicaram duas ou mais métricas de avaliação. As métricas mais aplicadas foram: *Recall*, *Accuracy*, *Precision*, *F1 Score*, *Cross validation*, *Confusion Matrix* e *Receiver Operating Characteristic Curve - ROC*.

Por fim, se apresentam como limitações deste trabalho:

- Algumas pesquisas não foram consideradas por não deixar claro em sua metodologia se tratava das causas da evasão estudantil ou desempenho do estudante.
- Em algumas pesquisas, a evasão não era tratada como estudo principal e sim como uma consequência do fraco desempenho do estudante.
- Grande parte das pesquisas deixaram lacunas em seus métodos, no sentido de não justificar a escolha dos algoritmos de ML utilizados, técnicas e métricas adotadas, bem como em qual nível de ensino se tratavam os dados (educação básica, ensino médio, superior ou pós-graduação) ou qual cursos superiores eram considerados na pesquisa. Isso dificulta a replicação do trabalho, impedindo a reprodutibilidade.

Consideram-se duas principais ameaças à validade do presente estudo:

1. Não se tem como garantir que a *string* de busca utilizada é a melhor possível. Alguma palavra-chave importante pode ter sido omissa. Para mitigar esta ameaça, foram realizados diversos testes com algumas opções de *strings* antes da realização da pesquisa;
2. Como a quantidade de trabalhos encontrada no mapeamento foi relevante, não se aplicou a técnica de *snowball* que teria potencial de encontrar possíveis trabalhos relevantes. Este problema foi mitigado pela quantidade considerável de trabalhos avaliados.

No próximo capítulo a metodologia de trabalho é apresentada, com ênfase para a análise e compreensão de dados, aplicação e avaliação dos modelos de ML desenvolvidos.

5 RESULTADOS

Neste capítulo apresentam-se os resultados e discussões do trabalho, constituídos da descrição e avaliação do desempenho dos modelos de ML implementados. Os experimentos foram definidos pela implementação e avaliação dos modelos de ML conforme os objetivos estabelecidos no início da pesquisa.

Durante a fase de treinamento, 80% de cada base foi reservada para treino e 20% de cada base para teste. Cada algoritmo de ML rodou 50 vezes fazendo a validação cruzada de *10k-folds*, conforme detalhado na seção 2.4.

O desempenho dos modelos foi avaliado por meio das métricas *Accuracy*, *Precision*, *Recall* e *F1 Score* e pelo testes estatísticos de Friedman e Nemenyi.

Este capítulo apresenta os resultados das análises dos dois conjuntos de dados compostos pelos dados de estudantes egressos e evadidos dos cursos de Ciência e Tecnologia, Biotecnologia ou Engenharia Agrícola e Ambiental.

5.1 Análise comparativa entre as bases

Esta seção tem por finalidade realizar uma análise comparativa entre os conjuntos de dados que compõe a presente pesquisa, de modo a observar como as variáveis selecionadas para compor o modelo de ML se comportam para cada curso estudado. O primeiro conjunto formado por dados dos estudantes de Ciência e Tecnologia e o segundo conjunto formado pelos dados dos estudantes de Biotecnologia ou Engenharia Agrícola e Ambiental.

No CRISPDM essa análise corresponde à etapa de *business understanding*, que é uma etapa essencial para ter uma compreensão completa do problema estudado, de modo a auxiliar na especificação dos desafios. Ainda nesta etapa, deve-se avaliar esses desafios e chegar a uma solução ou soluções, além de identificar estratégias de como mitigar o risco de evasão estudantil nos cursos selecionados.

A partir deste ponto o primeiro conjunto de dados será representado pelo código **C1** e o segundo conjunto de dados será representado pelo código **C2**.

A presente análise é centrada nas variáveis que foram classificadas como as mais importantes, diante da aplicação da técnica de *feature importance*, que é uma pontuação atribuída as *features* de um modelo de ML que define o quão “importante” é uma variável para a previsão do modelo. A técnica de *feature importance* foi aplicada ao C1 de dados e

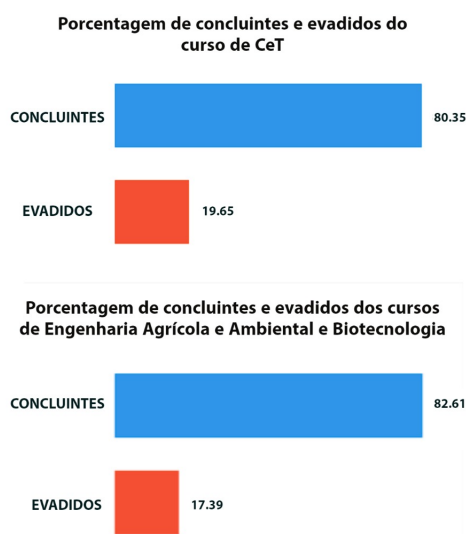
as variáveis selecionadas foram utilizadas tanto no C1, quanto no C2.

A seguir, apresenta-se o comportamento das principais variáveis para os dois conjuntos de dados.

5.1.1 Situação em relação ao curso

Ao se fazer uma análise comparativa entre as duas bases, observou-se que elas apresentam um volume de dados diferente. Os estudantes que cursaram Ciência e Tecnologia (C1) representam 229 amostras, divididas em 184 concluintes e 45 desistentes. O conjunto dos estudantes que cursaram Biotecnologia ou Engenharia Agrícola e Ambiental (C2), é composta por 92 amostras (76 concluintes e 16 desistentes), entre as quais 77 estudantes são egressos ou evadidos do curso de Biotecnologia e 15 do curso de Engenharia Agrícola e Ambiental. Conforme ilustrado na Figura 25, o curso de Ciência e Tecnologia apresenta maior evasão na UFERSA com 19,65% e os demais cursos que compõe o C2 (17,39%).

Figura 25 – Situação em relação ao curso



Fonte: autoria própria (2022)

Salienta-se que, o C2 é composto pela junção dos cursos de Biotecnologia e Engenharia Agrícola e Ambiental e mesmo com essa junção, o conjunto ainda apresenta volume menor. Observa-se que, ao longo da pesquisa, o projeto consistia em avaliar os dados dos estudantes de forma separada para cada curso, fazendo uso da base de dados da UFERSA, no entanto, houve após algumas tentativas não foi possível trabalhar com os dados da UFERSA. Entretanto, após diversas tentativas frustradas de obtenção dos dados

com o setor responsável, optou-se por tentar coletar os dados diretamente dos egressos e desistentes, por meio de formulários eletrônicos. Observa-se que os formulários foram amplamente divulgados, via *e-mail* e nas redes sociais de ex-alunos da UFERSA. Após 2 meses de divulgação, o número de respostas obtidas era inferior a 30. Dessa forma, optou-se por aumentar o prazo para as respostas e ampliar o escopo da pesquisa ao divulgar o formulário entre ex-alunos de outras universidades no país.

Observa-se que ambos os conjuntos de dados tem mais informações de estudantes que concluíram o curso. Vale destacar que isso torna o problema desbalanceado, já que o número de estudantes evadidos/desistentes é bem menor comparado com os concluintes.

5.1.2 Distribuição da universidade

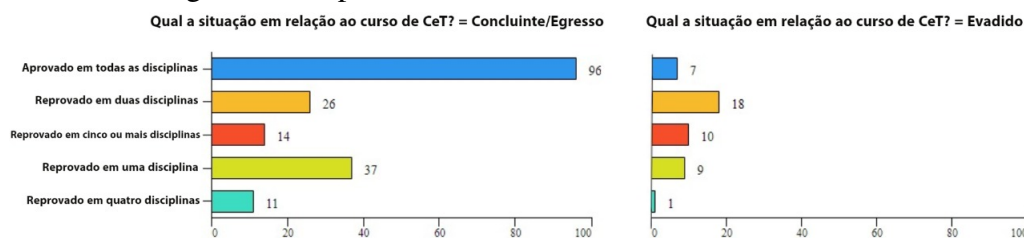
O C1, foi construído apenas por estudantes do curso de Ciência e Tecnologia, todas as amostras com egressos ou evadidos da UFERSA. Por outro lado, o C2 é menos homogêneo em relação aos cursos e universidades dos egressos e evadidos, pois apresenta dados de ex-alunos do curso de Biotecnologia da UFERSA e dados de ex-alunos do curso de Engenharia Agrícola e Ambiental de 3 universidades diferentes.

5.1.3 Aproveitamento acadêmico no primeiro ano de curso

Alguns estudos apontam que a evasão estudantil apresenta maior taxa durante o primeiro ano de curso (LANES 2018, SILVA 2007, BARROS 2004).

No que tange ao aproveitamento acadêmico no primeiro ano de curso, 44% dos estudantes do curso de Ciência e Tecnologia (C1) foram aprovados em todas as disciplinas. Relacionando-se o aproveitamento no primeiro ano de curso com o fator de evasão, o C1 obteve o maior percentual de evasão entre os estudantes que reprovaram em duas disciplinas (40%), seguido de estudantes reprovados em cinco ou mais disciplinas (22%), reprovados em uma disciplina (20%) e aprovados em todas as disciplinas (15%). Por fim, apenas um estudante que foi reprovado em quatro disciplinas evadiu (2%). O percentual de aproveitamento *versus* conclusão e evasão é ilustrado na Figura 26.

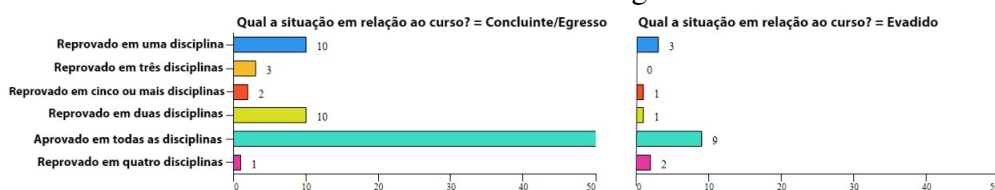
Figura 26 – Aproveitamento acadêmico do curso de CeT



Fonte: autoria própria (2022)

No C2, 64% dos estudantes obtiveram aprovação em todas as disciplinas durante o primeiro ano de curso. Com relação à evasão, o maior percentual está no conjunto de estudantes aprovados em todas as disciplinas, 56% (9 estudantes), seguido dos reprovados em 4 disciplinas, 18% (3 estudantes), reprovado em quatro disciplinas 12% (2 estudantes) e por fim, reprovado em duas disciplinas e reprovado em uma disciplina 12% (2 estudantes), conforme ilustrado na Figura 27.

Figura 27 – Aproveitamento acadêmico do curso de Engenharia Agrícola e Ambiental e Biotecnologia



Fonte: autoria própria (2022)

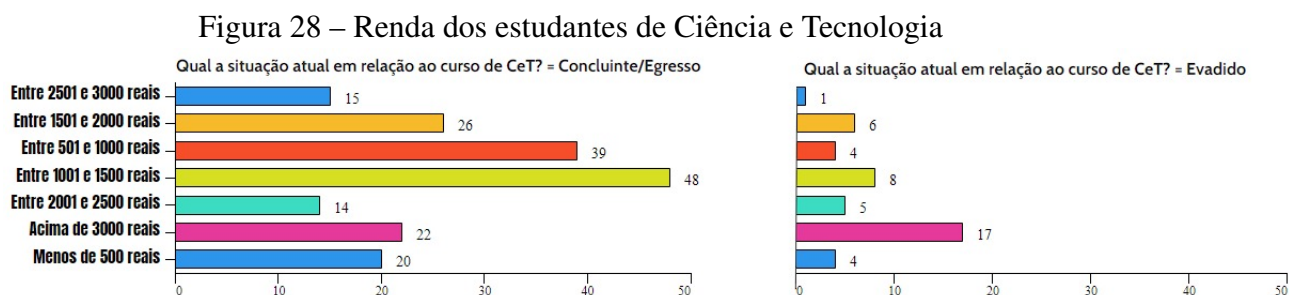
Observa-se que os dois conjuntos de dados apresentam comportamentos distintos no que tange ao fator evasão / aproveitamento acadêmico. O C2 se comporta diferente do primeiro, bem como de outros estudos que foram apresentados nos capítulos 3 e 4, nos quais estudantes que aprovam em todas as disciplinas têm maior probabilidade de permanecer no curso. Para o caso do C2, ser aprovado em todas as disciplinas e desistir do curso pode elencar outros fatores que necessitam ser investigados. Segundo Fagundes *et.al.*, o aproveitamento acadêmico de um estudante pode ser influenciado por alguns fatores como:

- Desempenho acadêmico durante o ensino médio, que se relaciona com o nível de preparação dos estudantes ao ingressar no ensino superior. Quanto melhor o desempenho no ensino médio, maior o nível de preparação do estudante para o ensino superior;

- A bagagem de conhecimentos prévios que os mesmos carregam e que pode influenciar no seu aproveitamento entre as disciplinas;
- A possibilidade das grades curriculares apresentarem certo nível de complexidade para determinados estudantes, sendo mais voltada para área de ciências exatas no curso de Ciência e Tecnologia e mais voltado para Ciências da Natureza nos cursos de Biotecnologia e Engenharia Agrícola e Ambiental;
- Falta de identificação com o curso. Em alguns casos o estudante entra no curso sem conhecer a grade curricular e acabam se frustrando no decorrer do curso.

5.1.4 Renda mensal da família

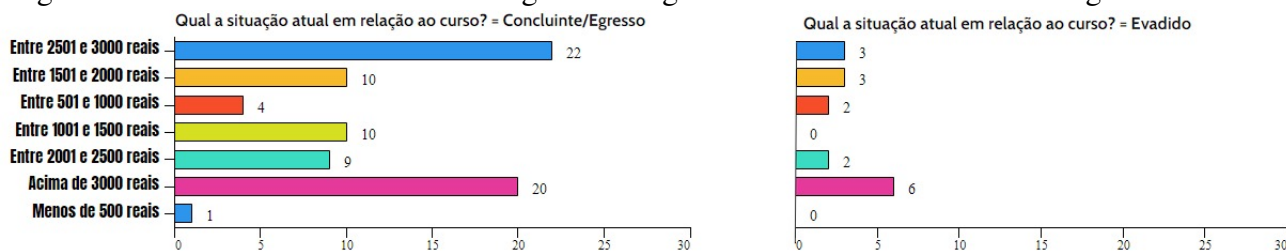
Em relação à renda mensal dos ex-alunos, cujos dados foram agrupados no conjunto C1 observa-se: 47% possuem renda acima de um salário mínimo (entre R\$1.501,00 e acima de R\$ 3.000,00); 44% possuem renda entre R\$501,00 e R\$ 1.500,00 (equivalente a aproximadamente um salário mínimo) e cerca de 9% possuem renda de menos de um salário mínimo (abaixo de R\$ 501,00), conforme ilustrado na figura 28. Ainda com relação à renda, observa-se que, o maior percentual de evasão está entre os estudantes que possuem renda acima de R\$ 3.000,00.



Fonte: autoria própria (2022)

No C2, observa que a condição de renda dos estudantes é superior ao C1: 79% possuem renda acima de um salário mínimo (entre R\$1.501,00 e acima de R\$ 3.000,00); 17% possuem renda entre R\$501,00 e R\$ 1.500,00 (equivalente a aproximadamente um salário mínimo) e cerca de 1% possuem renda de menos de um salário mínimo (abaixo de R\$ 501,00), conforme ilustrado na figura 29. Ainda com relação à renda, observa-se que, o maior percentual de evasão está com os estudantes que possuem renda acima de R\$ 3.000,00.

Figura 29 – Renda dos estudantes de Engenharia Agrícola e Ambiental e Biotecnologia



Fonte: autoria própria (2022)

Realizando-se um comparativo entre o perfil dos cursos, percebe-se que os estudantes do C1 possuem condições financeiras inferiores aos estudantes do C2. Entretanto, no que tange à evasão, os dois conjuntos se comportam de maneiras semelhantes no perfil de estudantes que evadiram. Esse resultado não corrobora com a análise de Marques (2020), bem como outros autores mencionados nos capítulos 3 e 4 que defendem que um dos fatores para que o estudante evada é a baixa condição financeira, que o faz desistir do curso para, em alguns casos, trabalhar para contribuir com a renda familiar (VELOSO; ALMEIDA, 2002).

Tal fato pode ser observado da perspectiva que a UFERSA dispõe de programas de assistência estudantil para estudantes de baixa renda. Dados da Pró-Reitoria de Planejamento (PROPLAN), (UFERSA, 2022) demonstram que somente no ano de 2022 a UFERSA destinou, aproximadamente, 7.081.000,00 (sete milhões e oitenta e um mil reais) de recursos para assistência estudantil, tais como: auxílio moradia, auxílio-creche, auxílio emergencial e outros. Dentre esses recursos, 2.003.000 (dois milhões e três mil reais) foram destinando exclusivamente ao auxílio permanência acadêmica.

5.1.5 Raça

Ao descrever como os estudantes consideram-se racialmente, é possível perceber que os dois conjuntos são compostos majoritariamente de estudantes brancos (45% no C1 e 48% no C2) e estudantes pardos (43% no conjunto C1 e 38% no conjunto C2). Para os estudantes que se consideram negros, o C1 corresponde a 9% e o C2 10%.

No que tange ao problema da evasão, no C1 os estudantes que se declaram amarelos e brancos destacam-se pelo fato de apresentarem o maior índice de evasão: 50% e 23% respectivamente.

No C2, 37% da evasão são dos estudantes que se consideram brancos, seguidos dos

estudantes que se consideram negros com 31%. Analisando-se o universo dos estudantes que se consideram negros, 100% deles evadiram. Ao analisar os dados do IBGE (2019) é possível mensurar que a evasão para as pessoas negras é maior em relação as demais raças.

5.1.6 Gênero

No que se refere à relação entre gênero e evasão, 100% dos estudantes do curso de Ciência e Tecnologia que optaram por não informar seu gênero, evadiram da instituição. Tal resultado pode ser analisado da perspectiva de que os cursos de tecnologia, como Ciência e Tecnologia, são formados majoritariamente pelo gênero masculino. Dados fornecidos pela DRA (2020) da UFERSA, apontam que, somente no campus central, o quantitativo de pessoas consideradas pertencentes ao gênero masculino entre o período de 2009.1 e o período de 2020.1, foi de, aproximadamente, 68.7%.

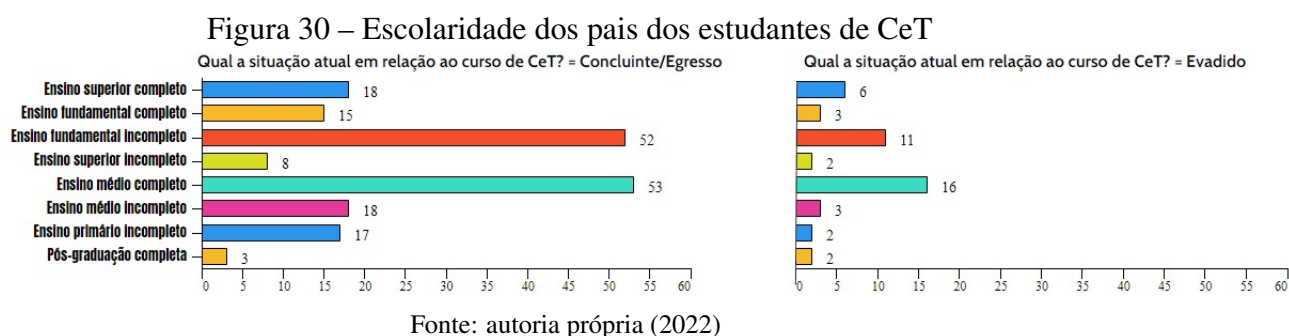
Carvalho *et.al* (2006) defende que essa questão é uma representação que se formou ao longo da história e que marcou as relações de gênero para excluir as mulheres e outros gêneros formados pelas denominações LGBTQIA+ da apropriação do conhecimento tecnológico. Se a tecnologia tem sido interpretada como um universo masculino, por outro lado, este padrão está passando por um processo de transformação em que, cada vez mais, a discussão sobre os desafios da permanência de estudantes LGBTQIA+ e a procura de mulheres por cursos de tecnologia vem crescendo (INSTITUTE, 2021).

No semestre de 2021.2, a taxa de matrícula de mulheres no C1 foi de 57% (DRA, 2020), superando, pela primeira vez, o número de ingressantes do gênero masculino. Cabe à Universidade, traçar medidas como o fomento da participação de mulheres em cursos de tecnologia, bem como estratégias para a permanência destas até o final do curso, tais como a promoção de: palestras, mesas redondas, oficinas e minicursos com a participação de estudantes egressas que atuem no mercado de trabalho; apresentações realizadas por mulheres reconhecidas no ambiente empresarial e acadêmico e com conhecimento destacado para abordar a questão do gênero e da participação de mulheres na tecnologia.

O comportamento do C2 diferiu: as pessoas do sexo masculino estão com 56% de evasão e as pessoas do sexo feminino com 44% evasão.

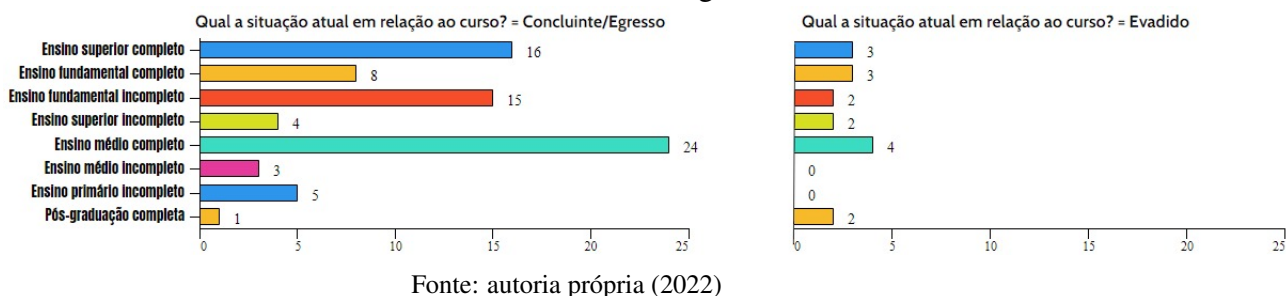
5.1.7 Escolaridade do pai

Em relação ao nível de escolaridade dos pais, no C1 (Ciência e Tecnologia) os pais da maioria dos estudantes (30%) possuem ensino médio completo, seguido de 28% que possuem ensino fundamental incompleto. No que tange à evasão, 16% dos estudantes que evadiram são dos pais que possuem ensino médio completo e apenas 10% dos pais apresentam ensino superior completo, conforme ilustrado na Figura 30.



No C2, os pais dos estudantes de Biotecnologia ou Engenharia Agrícola e Ambiental apresentam o dobro do volume para o ensino médio completo (30%), superior completo (20%), ensino fundamental incompleto (16%). O maior índice de evasão foi o de estudantes que possuem pai com o ensino médio completo (25%) conforme se ilustra na Figura 31.

Figura 31 – Escolaridade dos pais dos estudantes de Engenharia Agrícola e Ambiental e Biotecnologia

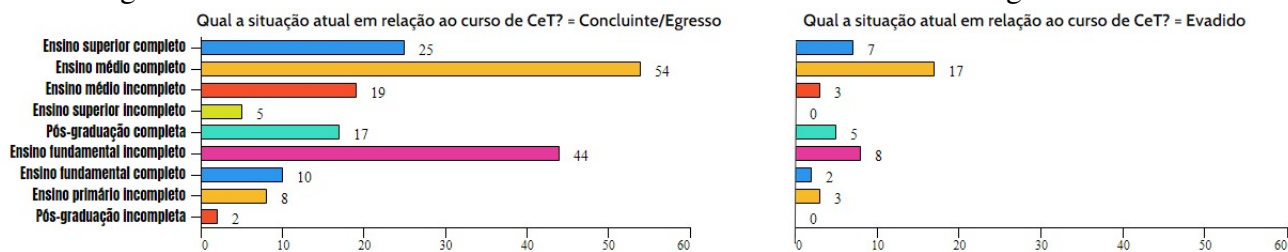


5.1.8 Escolaridade da mãe

As mães dos estudantes de ambos os conjuntos de dados possuem níveis escolares em sua maioria superior aos pais. As mães dos participantes do C1 apresentam: 9% de pós-graduação completa, 14% de nível superior, 31% nível médio completo e 23% de

ensino fundamental incompleto. Com relação à evasão, as mães da maioria dos estudantes que evadiram possuem ensino médio completo (37%), conforme se ilustra na Figura 32.

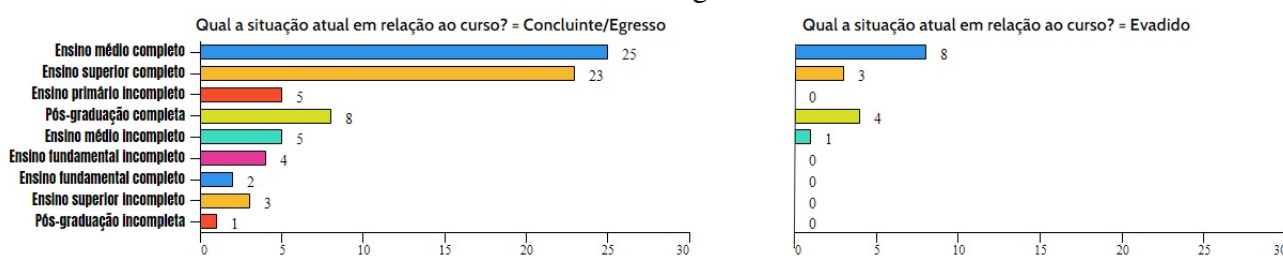
Figura 32 – Escolaridade das mães dos estudantes de Ciência e Tecnologia



Fonte: autoria própria (2022)

O C2 apresenta percentual maior no nível superior, 13% de pós-graduação completa, 28% de nível superior e 35% nível médio completo. No que tange à evasão, 50% dos estudantes que desistiram da Universidade possuem mães com ensino médio completo, conforme é detalhado na Figura 33.

Figura 33 – Escolaridade das mães dos estudantes de Engenharia Agrícola e Ambiental e Biotecnologia



Fonte: autoria própria (2022)

Destaca-se uma pesquisa do IBGE (2018), a trajetória escolar desigual, relacionado a papéis de gênero e entrada precoce dos homens no mercado de trabalho, as mulheres atingem em média um nível de instrução superior ao dos homens. Essa informação pode ser associada com a informação de renda, na qual a renda dos participantes do C2 foi maior, corroborando com uma pesquisa feita pelo IBGE (2018) que pessoas de níveis acadêmicos mais elevados apresentam maiores salários.

5.1.9 Relação de moradia e mudança para a cidade da Universidade

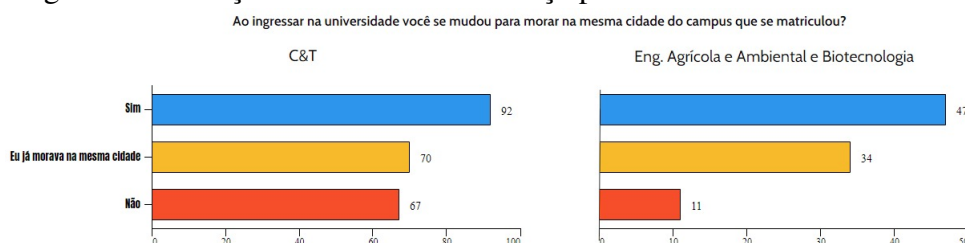
A mudança de cidade é algo que se reflete em ambos os conjuntos. O percentual de 40% do C1 representa os estudantes que mudaram sua residência. Boa parte dos discentes já residia na cidade do campus, não havendo a necessidade de mudança. Quando se trata

de trafegar entre a cidade natal e a cidade do campus, o curso de Ciência e Tecnologia possui um maior percentual de estudantes que realizam esse traslado em dias de aula, com 29% desse conjunto.

O C2 contou com mais estudantes que mudaram sua residência (51%). Vale destacar que apenas 12% dos evadidos não mudaram sua cidade, cenário interessante já que normalmente esses estudantes precisam de algumas horas para sair de sua cidade e chegar até o campus. Nos cursos de Biotecnologia e Engenharia Agrícola e Ambiental nenhum estudante que residiu na vila acadêmica evadiu/desistiu do curso.

A Figura 34 ilustra o detalhamento dos dados acerca da moradia dos estudantes para os dois conjuntos de dados.

Figura 34 – Relação de moradia e mudança para a cidade da Universidade



Fonte: autoria própria (2022)

Estudantes que optam por residir na vila acadêmica por vezes possui um perfil de baixa renda, avaliado por suas característica socioeconômicas, como renda familiar, cidade onde reside, nível acadêmico dos pais, entre outros fatores. Esses cenários são motivadores para o discente buscar transformar sua situação e consequentemente de sua família por meio do desenvolvimento educacional.

5.1.10 Instituição onde cursou o Ensino médio

O C1 apresenta 68,12% dos estudantes oriundos de escola pública e 31,88% dos estudantes oriundos de instituições privadas.

No C2, antes de ingressar na Universidade, 60% dos estudantes são oriundos de instituições privadas, já 40% dos estudantes são de instituições públicas.

Essas informações podem ser observadas na Figura 35.

Figura 35 – Tipo de escola onde cursou o ensino médio



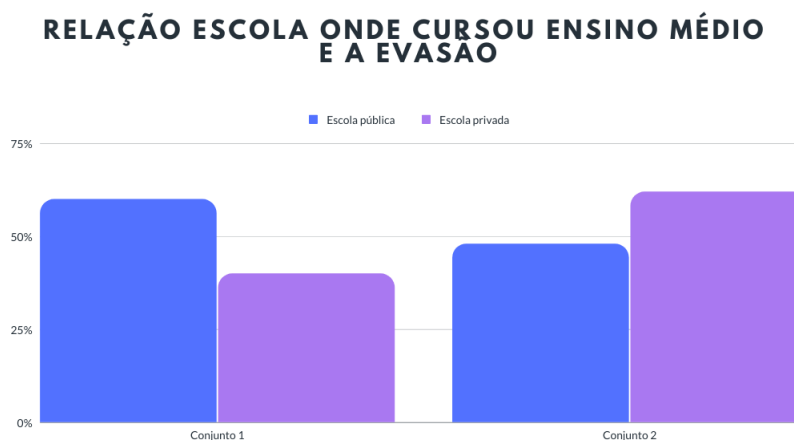
Fonte: autoria própria (2022)

O perfil de estudantes e as origens de ensino se divergem: estudantes do C1, em sua maioria, são oriundos de escola pública e os do C2 vieram de escolas privadas.

Para os estudantes que estudaram o ensino médio em escola pública, em ambos os casos foram de instituições Federais ou Estaduais, porém, em ambos os conjuntos de dados, como também, para os estudantes do curso de Biotecnologia e Engenharia Agrícola e Ambiental, a instituição Estadual é predominante, seguida da Instituição Federal. Vale destacar que a concorrência e o volume de oferta de vagas para as instituições Federais são maiores que as estaduais.

Seguindo a mesma linha, o maior índice de evasão para o C1 esteve entre estudantes oriundos de escola pública: no C1, 60% dos estudantes que evadiram pertenceram a escolas públicas no ensino médio; no C2, 62% dos estudantes que evadiram eram oriundos de escolas privadas durante o ensino médio, conforme se ilustra na Figura 36.

Figura 36 – Relação entre a evasão e o tipo de escola onde cursou o ensino médio



Fonte: autoria própria (2022)

5.1.11 *Conhecimento prévio na área do curso*

Identificou-se que antes de ingressar nos cursos, os estudantes do conjunto de Ciência e Tecnologia apresentam maior número de estudantes que já possuía algum conhecimento prévio do curso (45%), diferente do curso de Biotecnologia ou Engenharia Agrícola e Ambiental, nos quais apenas 6% obteve algum conhecimento antes de ingressar no curso. Atualmente, as áreas correlatas a tecnologia e afins contam com vastos treinamentos, tutoriais e materiais de estudo, principalmente *on-line*. Essa gama de materiais disponíveis para estudo pode influenciar no resultado do curso de Ciência e Tecnologia.

No C1, se tratando do grupo dos evadidos, 55% dos estudantes não tinha nenhum conhecimento técnico sobre o curso. No C2, 93% dos estudantes evadidos não detinha nenhum conhecimento técnico sobre o curso, reforçando o contexto supracitado.

5.1.12 *Obtenção de bolsa durante o curso*

Aproximadamente 60% dos estudantes do curso de Ciência e Tecnologia não foram contemplados com bolsa durante sua jornada acadêmica. Entretanto, 72% dos estudantes do curso de Biotecnologia ou Engenharia Agrícola e Ambiental foram contemplados com bolsa.

No que diz respeito à relação de estudantes evadidos e bolsa de estudos, no C1,

82% dos estudantes que evadiram não obtiveram bolsa de estudo durante sua passagem na graduação. De modo semelhante, no C2, 56% dos estudantes que evadiram não obtiveram bolsa de estudo, conforme se detalha na Figura 37.

Figura 37 – Relação entre a evasão e bolsa de estudo



Fonte: autoria própria (2022)

A oferta de bolsas pode ser um fator contribuinte para a não evasão do estudante, pelo fato de manter o estudante engajado com as ações que as bolsas exigem, bem como a remuneração da bolsa que pode auxiliar o estudante a custear suas despesas pessoais.

5.1.13 Percepção do estudante com relação aos pré-requisitos

A presente pergunta teve o objetivo de captar a percepção do estudante acerca da cadeia de pré-requisitos do seu curso, de modo a reconhecer que a cadeia de pré-requisitos é adequada ou não.

No C1 73% dos estudantes consideram a cadeia de pré-requisitos do seu curso adequada e 27% dos estudantes não consideram adequada. Se tratando da evasão, 62% dos estudantes que evadiram consideram a cadeia de pré-requisitos do seu curso adequada e 38% não consideram adequada.

No C2, 62% dos estudantes considera a cadeia de pré-requisitos do seu curso adequada e 38% dos estudantes não considera adequada. Com relação à evasão, 81% dos estudantes que evadiram considera a cadeia de pré-requisitos do seu curso adequada e 19% não considera adequada.

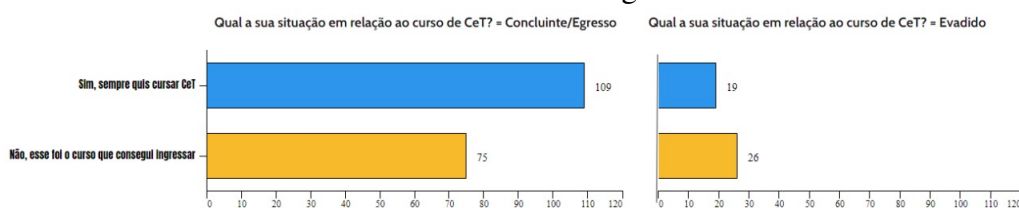
Observa-se que, o nível de insatisfação com a cadeia de pré-requisitos não foi um fator decisivo para a maioria dos estudantes que evadiram. Vale destacar que os níveis

acadêmicos no geral entre os pais e a renda familiar são fatores mais decisivos quanto a essa exigência.

5.1.14 Primeira opção de curso

Com relação ao ingresso no curso, no C1, 55% dos estudantes informaram que a escolha foi a primeira opção e 45% não entraram em sua primeira opção de curso. Relacionando com o fator evasão, conforme ilustrado na Figura 38, 42% dos estudantes que evadiram estavam na sua primeira opção de curso, enquanto a maioria (58%) dos estudantes que evadiram afirmam que o curso de Ciência e Tecnologia foi apenas o curso que conseguiram ingressar.

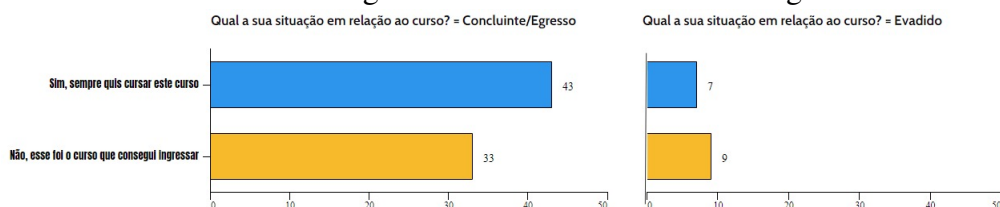
Figura 38 – Relação entre a evasão e opção de curso dos estudantes de Ciência e tecnologia



Fonte: autoria própria (2022)

No C2, 54% dos estudantes informaram que a escolha foi a primeira opção e 46% não entraram em sua primeira opção de curso. Analisando-se os evadidos, 44% dos estudantes que evadiram estavam na sua primeira opção de curso, enquanto a maioria (56%) dos estudantes que evadiram afirmam que o curso de Ciência e Tecnologia foi apenas o curso que conseguiram ingressar, conforme ilustrado na Figura 39.

Figura 39 – Relação entre a evasão e opção de curso dos estudantes de Engenharia Agrícola e Ambiental e Biotecnologia



Fonte: autoria própria (2022)

Para ambos os conjuntos, observa-se que, não estar em um curso que era sua primeira opção pode ser um fator decisivo para que o estudante evada. Vale destacar que a

complexidade do curso, falta de entendimento e um conjunto de outros fatores favorecem a não permanência do estudante e o encerramento do ciclo acadêmico.

5.1.15 Filhos

Durante o período em que permaneceram no ensino superior, 87% dos estudantes que formam o C1 não possuíam filhos e 13% possuíam.

Todos os estudantes do C2 afirmaram não possuir filhos durante a jornada acadêmica na Universidade.

Na falta do atributo idade não é possível calcular se os estudantes do curso de Ciência e Tecnologia possuem filhos quando são jovens ou não, porém, o fator filho pode, em alguns casos, ser considerado como algo que contribui para evasão de estudantes, principalmente nos casos das pessoas do gênero feminino, sabendo de todas as atribuições que o papel exige. Essa afirmação pode ser confirmada por estudos como o de Gaioso (2005), que, em uma ampla reflexão sobre a evasão e suas diversas causas, a partir da visão do estudante, identificou que o nascimento de filhos é apontado como um dos problemas que contribuem para a evasão de estudantes.

5.2 Comparação entre análise da presente pesquisa e o trabalho de Marques (2020)

O trabalho de Marques 2020 não foi centrado no entendimento dos dados, portanto, não foi possível comparar o comportamento de todas as características trabalhadas e o fenômeno da evasão. No entanto, alguns pontos podem ser comparados com a presente pesquisa:

- No trabalho de Marques (2020) o conjunto de dados de estudantes do curso de Ciência da computação foi de tamanho menor: 100 ex-alunos, entre os quais 52 (51%) foram de desistentes e 48 (49%) de concluintes, o que tornou a base de dados um pouco mais balanceada do que as bases trabalhadas no presente estudo.
- No trabalho de Marques (2020) o conjunto de dados foi composto apenas por estudantes da UFERSA.
- No estudo realizado por Marques (2020) constatou-se que a renda mensal de estudantes que concluíram o curso de Ciência da Computação da UFERSA, é melhor em detrimento aos que evadiram.

- Com relação ao aproveitamento acadêmico dos estudantes, foi observado que, no trabalho de Marques (2020) os estudantes reprovados em até duas disciplinas tem maior probabilidade de evadir-se e que os aprovados em todas as disciplinas, tem maior probabilidade de concluir. O estudo não apresenta porcentagens.
- Com relação à raça dos estudantes que participaram do estudo, no trabalho de Marques (2020) foi observado que os estudantes pardos e os estudantes negros tem maior probabilidade de evadir-se e que os estudantes brancos, tem maior probabilidade de concluir. O estudo não apresenta porcentagens.

5.3 Resultados da implementação dos modelos de ML

Nesta seção são apresentados os resultados obtidos usando diferentes classificadores. Cada conjunto de dados, em seu momento, foi dividido para que a mesma quantidade de dados de treinamento de cada classe fosse distribuída igualmente. Em seguida, 80% (de cada conjunto de dados) dos registros informados pelos estudantes concluintes e desistentes foram utilizados e o restante das amostras foi utilizado para testes. Cada conjunto de dados foi normalizado entre -1 e 1. Todos os modelos de ML implementados neste trabalho passaram por 50 treinamentos e validação cruzada com *10-folds*, conforme já foi detalhado na seção 2.4.

As características analisadas na seção anterior foram aplicadas aos modelos de ML para obter o resultado da previsão e posteriormente avaliados por meio das métricas de avaliação: ano e semestre de conclusão ou desistência, ano e semestre de ingresso do estudante na Universidade, renda mensal da família, aproveitamento acadêmico até o segundo semestre do curso, raça, mudou-se para morar na mesma cidade da Universidade, nível de escolaridade da mãe, campus e turno os quais o estudante estava matriculado, oriundo de escola pública no ensino médio, conhecimento prévio, contemplado com alguma bolsa durante o período que esteve no curso, cadeia de pré-requisitos, primeira opção de curso, possuir filhos.

Os estudantes do curso de Ciência e Tecnologia apresentam um percentual de evasão um pouco maior do que o curso de Biotecnologia ou Engenharia Agrícola e Ambiental, respectivamente 20% e 17%, conforme ilustrado na Figura 40. Nesse contexto, observa-se que a volumetria do conjunto de dados de Ciência e Tecnologia foi maior que a do C2.

Figura 40 – Frequência de evasão entre os cursos

Alunos - CeT				Alunos - Biotec e Ambiental			
	Situação	Contagem	Frequência (%)		Situação	Contagem	Frequência (%)
0	Concluente	184	80.000000	0	Concluente	76	83.000000
1	Evadido/Desistente	45	20.000000	1	Evadido/Desistente	16	17.000000

Fonte: autoria própria (2022)

Analisar a evasão dentro de um contexto estudantil não é algo trivial e apresenta suas complexidades, entre elas destacam-se: análise do desempenho acadêmico; análise do cognitivo; análise do comportamento do estudante; análise do âmbito socioeconômico.

As melhores configurações, decorridos os ajustes necessários para cada classificador, foram:

- Para o classificador perceptron foi utilizada taxa de aprendizagem entre 0,01 a 0,5. A taxa de aprendizagem n (também um valor entre 0 e 1) diz o quão rápido a rede chega ao seu processo de classificação: um valor muito pequeno causa demora a convergir, enquanto um valor muito alto pode levar para valores fora do ajuste e nunca convergir (BLOCK, 1962).
- O classificador *Decision Tree* foi treinado com o parâmetro *criterion entropy e gini*, que são critérios para calcular o ganho de informação. Tanto o *entropy* quanto o *gini* são medidas de impureza de um nó. Um nó com várias classes é impuro, enquanto um nó com apenas uma classe é puro (OLANOW; KOLLER, 1998).
- Para o classificador *Naive Bayes* os hiperparâmetros utilizados foram *var_smoothing*, que é um cálculo de estabilidade para alargar a curva e conseguir alcançar amostras que estão mais distantes da média de distribuição. Esse hiperparâmetro recebeu o *np.logspace* que retorna números espaçados uniformemente em uma escala de *log* com valores entre 0 e -9.
- O classificador *Random Forest* recebeu os parâmetros: “*gini*”, “*entropy*” que é um parâmetro específico da árvore para medir a qualidade de uma divisão. Os critérios “*gini*” para a impureza *Gini* e “*entropy*” para o ganho de informação; o parâmetro *max_depth* que é a profundidade máxima da árvore, definida com um *array* com os números 2, 3, 5 10 e 20; *min_samples_split* que atribui o número mínimo de amostras necessárias para dividir um nó interno, configurados com um *array* contendo os valores 5, 10, 20, 50, 100 e 200. Por fim, o parâmetro *n_estimators* é o número de

árvores na floresta.

- Para o classificador SVM foram definidos como kernel: *linear*, *poly*, *Radial Basis* e *sigmoid*. A Função Kernel é um método usado para receber dados como entrada e transformá-los na forma necessária de processamento de dados (PATLE; CHOUHAN, 2013). Um parâmetro de regularização que controla a compensação entre alcançar um baixo erro de treinamento e um baixo erro de teste que é a capacidade de generalizar seu classificador (CHERKASSKY; MA, 2004). Para o parâmetro de regularização deste estudo, foi atribuído uma variação de 0,1 a 10. Essa variação na regularização foi adotada, pois com um valor menor, obtemos hiperplano de pequena margem e com um valor maior, obtemos hiperplano de maior valor. O valor de regularização que alcançou a melhor precisão foi escolhido para classificar todas as instâncias de teste.
- O hiperparâmetro número de vizinhos k do KNN variou de 3 a 11, selecionando apenas os números ímpares. Foi possível chegar a esse hiperparâmetro, adotando diferentes valores para k . O valor de k que alcançou a melhor precisão foi escolhido para classificar todas as instâncias de teste.
- O classificador MLP foi treinado com função de ativação *sigmoid*, tangente hiperbólica e *relu*, a função de otimização para ambos é o *Adam* e a taxa de aprendizagem variou de 0,1 a 0,5. Em relação à quantidade de neurônios, a rede neural MLP inicialmente começou com 5 e terminou com 65 neurônios.

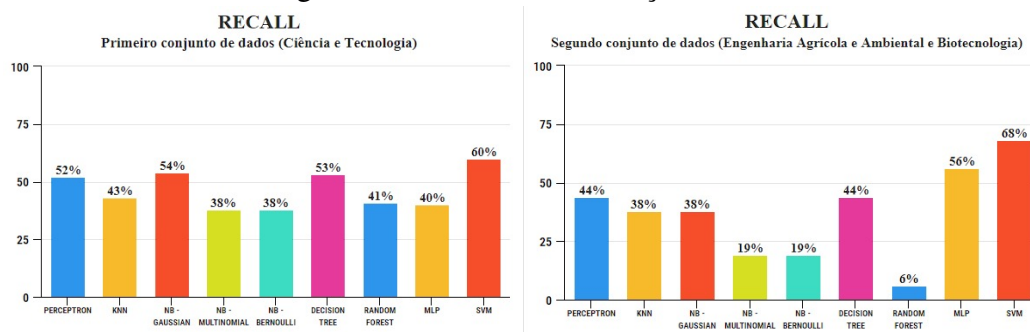
Ressalta-se que, ao se trabalhar com algoritmos de ML, não existe um parâmetro inicial para o processo de treinamento e teste dos modelos. Porém, utiliza-se da boa prática de definir um intervalo de valores para que os algoritmos encontrem o melhor parâmetro para o conjunto de dados. No presente estudo, o intervalo de parâmetros adotados para todos os algoritmos foram pautados nas referências de utilização de pesquisadores renomados no campo da ML, tais como: Mduma *et.al* (2019), Soresen (2019), Guanawan (2019) e AFIA *et.al* (2019). Os valores que alcançaram melhores resultados foram escolhidos para classificar todas as instâncias de teste.

Muitas métricas de avaliação de algoritmo podem ser consideradas ao determinar a adequação de um modelo de ML aplicado. Neste trabalho foram utilizadas as métricas: *Recall*, *Precision*, *F1 Score*, *Curvas ROC e AUC* e por fim *Accuracy*. O resultado dessas métricas de avaliação, são apresentados a seguir.

5.3.1 Recall

A métrica de avaliação *Recall* avalia a sensibilidade em relação ao dado. Na Figura 41 ilustram-se os resultados dos algoritmos.

Figura 41 – Métrica de avaliação Recall



Fonte: autoria própria (2022)

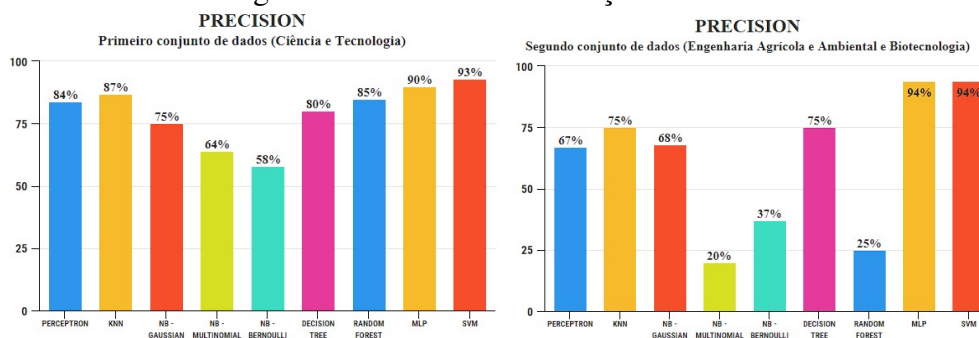
Observa-se que, o SVM obteve melhor avaliação, com diferença de 7% para o *Naive Bayes*, o segundo melhor avaliado para o C1 de dados, enquanto no C2, o SVM e o MLP obtiveram melhores avaliações próximas a 60%.

A avaliação corrobora com a explicação sobre a métrica *Precision* que por apresentar mais amostras no conjunto para estudantes concluintes, eleva o nível de dificuldade de aprendizagem do modelo para a métrica de Recall, pois, o modelo erra consideravelmente entre as amostras que são dos estudantes evadidos.

5.3.2 Precision

A segunda métrica avaliada foi a *Precision* que é uma métrica que quantifica o número de previsões positivas corretas. Os resultados dos algoritmos são observados na Figura 42.

Figura 42 – Métrica de avaliação Precision



Fonte: autoria própria (2022)

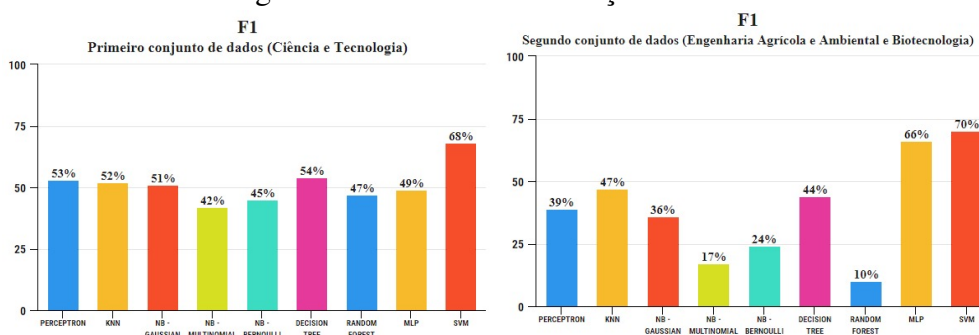
Essa métrica representa a taxa de acerto de uma determinada classe, neste caso, em ambos os conjuntos é natural que apresente resultados acima de 92%, devido se tratar de um problema desbalanceado em que existem mais informações de estudantes concluintes do que estudantes evadidos, fato esse que eleva o nível de assertividade do modelo. Neste caso o modelo construído consegue predizer com mais certeza quem pode concluir o curso.

Observa-se que, o SVM obteve melhor avaliação, acima de 90%, com diferença de 3% para o MLP que foi o segundo melhor avaliado para o C1, enquanto no C2, SVM e MLP apresentaram a mesma porcentagem.

5.3.3 *F1 Score*

A terceira métrica avaliada foi a *F1 Score* que objetiva combinar as métricas de *Precision* e *Recall* em uma única métrica. Na Figura 43 são ilustrados os resultados dos algoritmos.

Figura 43 – Métrica de avaliação F1 Score



Fonte: autoria própria (2022)

Como a *F1 Score* é destinada a mensurar a harmonia entre as métricas *recall* e

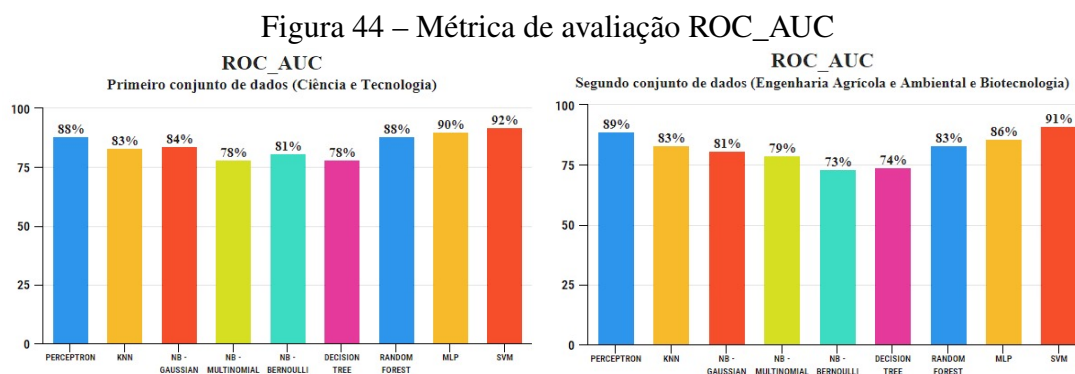
a *precision* é natural que se em uma dessas duas métricas apresente valor percentual baixo, influencie o resultado da *F1 Score*. Conforme mencionado anteriormente, o modelo treinado possui uma excelente precisão para predizer os estudantes concluintes e um baixo percentual quando se trata dos estudantes evadidos.

Com percentual em ambos os conjuntos acima de 68%, é possível perceber que o *kernel linear* possui uma leve vantagem na harmonização entre *precision* e *recall* com 70%.

No C1, observa-se que, o SVM obteve melhor avaliação, com diferença de 14% para o algoritmo de árvore de decisão (*Decision Tree*) que foi o segundo melhor avaliado. No C2, o MLP obteve melhor avaliação, com diferença de 3% para o algoritmo SVM que foi o segundo melhor avaliado.

5.3.4 Roc e AUC

A quarta métrica avaliada foi a *ROC e AUC* que mostra o quão bom o modelo criado pode distinguir entre duas características. Aplicado ao presente estudo, avalia o quão bons os modelos criados podem distinguir entre a classificação de estudantes concluintes ou desistentes. Na Figura 44 os resultados da avaliação dos algoritmos são ilustrados:

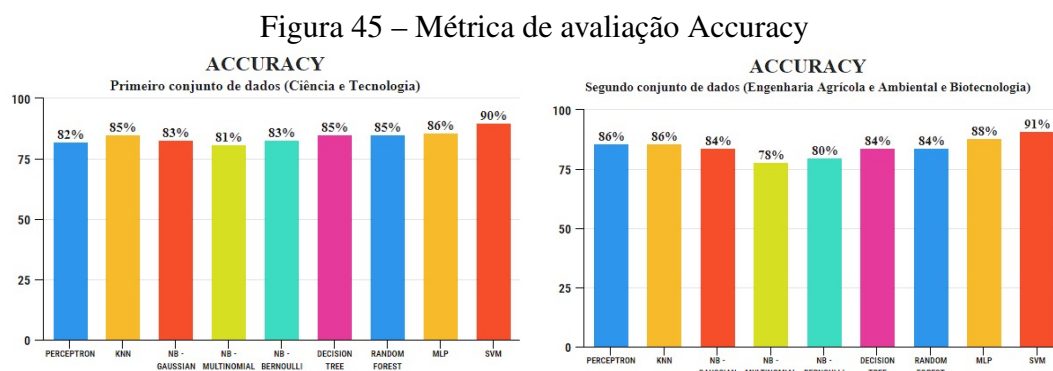


Fonte: autoria própria (2022)

Esta métrica é ideal para classes dicotômicas e, no caso do presente estudo em que se têm classes de concluintes e evadidos, ela pode distinguir com precisão. Ambos os conjuntos apresentaram métricas acima de 91% para o C1 e acima de 89% para o C2. Diante do fato de existirem mais amostras dos estudantes concluintes do que os evadidos, o modelo consegue distinguir bem quem são os concluintes.

5.3.5 Accuracy

Por fim, a métrica *Accuracy* avalia o acerto global de cada modelo. Na Figura 45 os resultados da avaliação dos algoritmos são ilustrados:



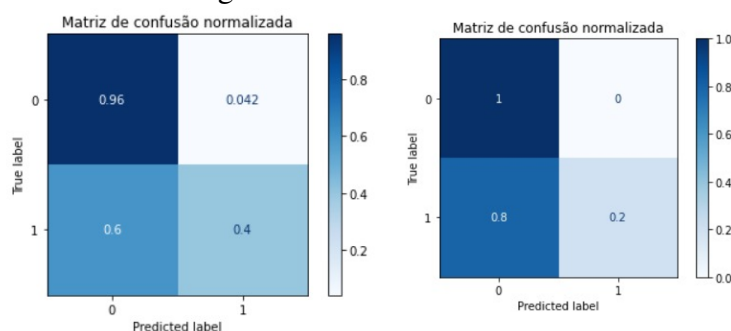
Fonte: autoria própria (2022)

Observa-se que, globalmente, ambos os conjuntos obtiveram resultados acima de 80%. Esse resultado se deve ao fato de que a acurácia considera os acertos globais. O modelo desenvolvido nesse estudo consegue acertar bem quando se trata de estudantes concluintes.

5.3.6 Confusion Matrix

No campo de ML, uma *confusion matrix* (matriz de confusão) é uma tabela que permite a visualização do desempenho de um algoritmo de classificação. A seguir, na Figura 46 ilustram-se os resultados da matriz de confusão do algoritmo SVM, para avaliar o desempenho deste algoritmo que se destacou como o que obteve melhor avaliação dentre todas as métricas de avaliação utilizadas no presente estudo.

Figura 46 – Confusion Matrix



Fonte: autoria própria (2022)

Interpretando o resultado da matriz de confusão, observa-se que:

Verdadeiro Positivo: o conjunto da classe real, o resultado foi correto. O estudante concluiu e o modelo previu que ele concluiria com taxa de 0,96 (96% de acerto) para o C1 e 1 (100% de acerto) para o C2.

Falso Positivo: no conjunto da classe real, o resultado foi incorreto para o primeiro modelo: o estudante concluiu e o modelo previu que não iria concluir com taxa de 0,042 (4,2%) e para o C2 (0%).

Falso Negativo: no conjunto da classe real, o resultado foi incorreto. O estudante não evadiu e o modelo previu que ele iria evadir-se com taxa de 0,6 (60%) para o C1 e 0,8 (80%) para o C2.

Verdadeiro Negativo: no conjunto da classe real, o resultado foi correto. O estudante evadiu e o modelo previu que ele evadiria com taxa de 0,4 (40%) no C1 e 0,2 (20%) no C2.

O modelo mostrou-se eficiente para prever quem conclui e isso evidencia-se pelo recall, que apresentou percentual baixo, quando comparado as demais métricas. O modelo de ML desenvolvido consegue prever quem conclui, mas se for algo diferente disso, apresenta dificuldades. Essa dificuldade do modelo prever quem evade pode ser interpretada sob duas óticas:

1. Em ambos os conjuntos de dados, a quantidade de amostras de estudantes que não concluíram o curso foi menor do que a quantidade de estudantes que concluíram o curso, resultando em conjuntos de dados desbalanceados, tornando-se insuficiente para que o modelo aprendesse com exatidão quais os estudantes com potencial para evadir-se.

2. O C2 pode ter apresentado resultados inferiores ao primeiro pelo fato de que o modelo foi desenvolvido para *features* significativas do C1 e foi apenas replicado ao C2. Dessa forma, podemos levantar a hipótese de que a busca por diferentes variáveis significativas para o C2 poderia melhorar os resultados.

Ao implementar os modelos de ML, espera-se que o modelo apresente o maior percentual da métrica em estudo. Os classificadores dos dois conjuntos de dados foram apresentados aos mesmos modelos de ML e os parâmetros de cada modelo foram ajustados para ambos os conjuntos de forma igual. O classificador SVM foi o que obteve os melhores resultados em todas as métricas, conforme apresentado na Tabela 13.

Tabela 13 – Comparação das melhores métricas entre os dois conjuntos de dados

Métrica	Conjunto 1: Ciência e Tecnologia <i>kernel RBF</i>	Conjunto 2: Biotecnologia Engenharia Agrícola e Ambiental <i>kernel linear</i>
<i>Accuracy</i>	90,41%	91,30%
<i>Precision</i>	92,70%	93,75%
<i>Recall</i>	60,60%	68,75%
<i>F1-Score</i>	68,28%	70,23%
<i>ROC_AUC</i>	92,30%	91,44%

Acredita-se que o resultado destaque do SVM deve-se ao fato de que este algoritmo apresenta bom desempenho em classes linearmente separáveis, como as classes trabalhadas no presente estudo.

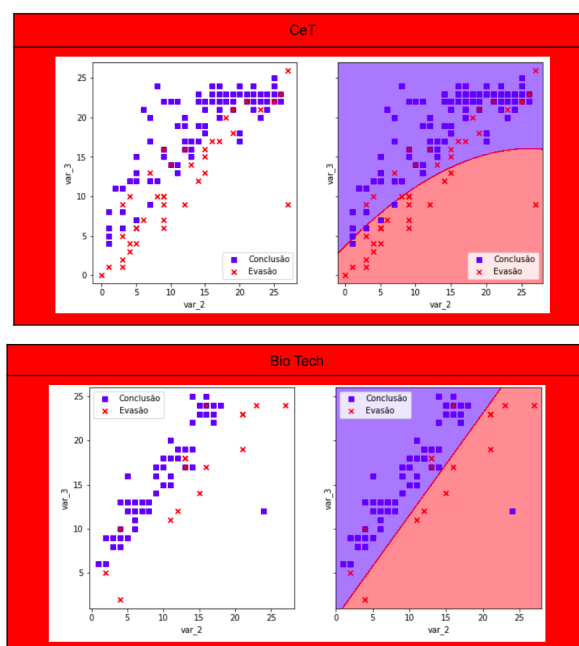
O diferencial entre os modelos é o *Kernel* em que cada um apresentou resultados distintos, mesmo sendo as métricas e características selecionadas, os dados de cada conjunto divergem.

Durante o treinamento do modelo, o conjunto do curso de Ciência e Tecnologia apresentou melhores resultados utilizando o *kernel RBF* e o curso de Biotecnologia ou Engenharia Agrícola e Ambiental utilizando o mesmo classificador, porém com *kernel linear*.

Quando um conjunto de dados apresenta em suas métricas os melhores resultados com um *kernel RBF* significa que o conjunto apresenta uma distribuição aproximada a de uma gaussiana, em que a sua distribuição pode ser considerada normal. Já para o conjunto que apresenta resultados com *kernel linear* é possível mensurar e traçar uma reta que melhor distingue os dados. Para os dois conjuntos de dados do presente estudo, esse

comportamento é ilustrado na Figura 47.

Figura 47 – Comportamento dos dados em cada conjunto



Fonte: autoria própria (2022)

Por meio da ilustração, observa-se que, os dados do C1, a distribuição das características estão próximas, enquanto no C2, é possível distinguir e separar linearmente as características dos concluintes e evadidos, tornando-se mais fácil a identificação e separação das classes.

5.3.7 Análise estatística

A análise estatística de modelos de ML é um processo importante para avaliar a eficácia dos modelos implementados. Ela permite aos cientistas de dados comparar diferentes algoritmos em termos de precisão, acurácia e robustez.

Conforme explanado no Capítulo 3, o teste estatístico de Friedman é uma técnica usada para comparar vários modelos de aprendizagem de máquina. Este teste é amplamente utilizado quando há mais de dois algoritmos a serem comparados e não há suposições de normalidade sobre os dados.

O teste foi aplicado para os dois casos da presente pesquisa, conforme ilustrado na Figura 48.

Figura 48 – Teste de Friedman



Fonte: autoria própria (2022)

Ao interpretar a figura, observa-se que:

- Cada algoritmo implementado e seu respectivo valor p (o valor p é uma medida da força das evidências estatísticas contra a hipótese nula (H_0), que é calculado por meio da média de desempenho do algoritmo em cada execução ao qual foi submetido). Se o valor p for menor que um nível de significância pré-definido, é porque há uma diferença significativa no desempenho deste algoritmo. Caso contrário, não há diferença significativa;
- a barra observada no lado direito da figura apresenta os grupos de algoritmos que não apresentam diferença estatística, o que significa que ao se selecionar qualquer um destes para o problema aplicado, o desempenho será o mesmo.

Ao analisar os resultados é possível identificar que, de acordo com o teste de Friedman, não existe distância crítica entre os modelos para ambos os casos. Isso reforça que mesmo que um dos modelos tenha apresentado resultados melhores em uma métrica, não existe diferença significativa ao selecionar qualquer um dos citados para implementação.

O teste de Nemenyi é uma técnica usada para identificar as diferenças significativas entre os modelos após o teste de Friedman. Esse teste é baseado na análise de resíduos e permite aos cientistas de dados visualizar e interpretar as diferenças entre os modelos de ML. Ao aplicar o teste de Nemenyi, observa-se que existe uma diferença significativa no primeiro conjunto de dados ao comparar o NB com os demais modelos, já que a diferença

é maior do que a diferença crítica de 1.644. No segundo conjunto de dados, o teste de Nemenyi não apresentou diferenças significativas.

5.4 Considerações Finais

Neste capítulo foram apresentados os resultados do trabalho, constituídos da descrição e avaliação do desempenho dos classificadores baseados em cinco métricas de avaliação: *Accuracy*, *Precision*, *Recall*, *F1-Score*, *ROC_AUC* e também a *Confusion Matrix*. Nos experimentos relatados, o algoritmo SVM foi avaliado como o de melhor desempenho para predição de estudantes em risco de evasão. Ao aplicar o teste estatístico de Friedman, se constata que não existe distância crítica entre os modelos para ambos os casos estudados. Entretanto, ao se aplicar o pós-teste (teste de Nemenyi) o primeiro conjunto de dados apresenta o algoritmo NB com diferenças significativas em relação aos demais algoritmos implementados.

Salienta-se que o modelo se mostra mais eficaz em prever estudantes com chances de conclusão e que esse comportamento pode ter ocorrido por duas possibilidades: em ambos os conjuntos a quantidade de amostras de estudantes que não concluíram o curso foi menor que a quantidade de estudantes que concluíram o curso, resultando em conjuntos de dados desbalanceados; as *features* utilizadas no C2 podem não ser as melhores para definir o perfil que representa os cursos estudados no C2.

6 CONSIDERAÇÕES FINAIS

A evasão estudantil é um problema complexo, com limitações relacionadas a traçar perfis de diferentes estudantes com características acadêmicas, sociais, demográficas e comportamentais distintas. A presente pesquisa teve como objetivo propor uma metodologia complementar a de Marques (2020), contemplando, de forma ampla, todas as etapas para predição e detecção de estudantes com alto risco de evasão em diferentes cursos superiores. Para tanto, a metodologia desenvolvida foi aplicada nos cursos de Engenharia Agrícola e Ambiental, Bacharelado em Ciência e Tecnologia e Biotecnologia da UFERSA, que serviram como estudos de caso.

A Revisão Sistemática da Literatura permitiu ampliar o conhecimento sobre o desenvolvimento de modelos de ML, baseado em diferentes técnicas e métricas de avaliação, bem como os diferentes fatores indutores que levam um estudante a evadir-se da Universidade, além de retornar trabalhos relevantes para o contexto desta pesquisa. As principais pesquisas que foram avaliadas no MSL serviram como referencial técnico para os procedimentos práticos dos experimentos, constituídos por etapas de pré-processamento dos dados, treinamento dos modelos e avaliação dos modelos de classificação a partir de métricas de desempenho.

O trabalho foi desenvolvido segundo a abordagem do CRISP-DM e foi realizado em duas etapas. Na primeira etapa foi realizada para o curso de Ciência e Tecnologia, englobando todos os *campi* da UFERSA, nos quais o curso é ofertado. Nesta etapa, os dados foram coletados, o conjunto de dados foi criado e passou pelas fases de análise, pré-processamento, aplicação dos modelos de ML e avaliação por meio das métricas. Nesta etapa obtiveram-se excelentes resultados com métricas acima de 90% de acerto.

A partir dos resultados da primeira etapa, foi realizada uma segunda etapa com os cursos de Biotecnologia, ofertado na UFERSA *campus* Mossoró e com o curso de Engenharia Agrícola e Ambiental, também ofertada no *campus* Mossoró.

Os pesquisadores da presente pesquisa realizaram diversos esforços para popular a base de dados do segundo conjunto com informações de uma quantidade considerável de estudantes, no entanto, não foi possível. Para o curso de Engenharia Agrícola e Ambiental expandiu-se o universo da pesquisa para outras Universidades do país e mesmo assim, ainda se obteve baixo número de respostas ao questionário. Dado esse cenário, o segundo conjunto de dados foi mais difícil de se trabalhar. As fases de análise, pré-

processamento, aplicação dos modelos de ML e avaliação por meio das métricas foram realizadas seguindo os mesmos parâmetros aplicados no primeiro conjunto de dados e o resultado se assemelhou, obtendo-se métricas acima de 90% de acerto.

O algoritmo SVM gerou os melhores resultados de predição em ambos os conjuntos de dados, na qual foram alcançadas avaliações acima de 90%. Ao aplicar o teste estatístico de Friedman, se constata que não existe distância crítica entre os modelos para ambos os casos estudados. Entretanto, ao se aplicar o pós-teste (teste de Nemenyi) o primeiro conjunto de dados apresenta o algoritmo NB com diferenças significativas em relação aos demais algoritmos implementados.

Ao final do trabalho, foi possível responder às perguntas de pesquisa formuladas no início do projeto:

QC1 -Quais os fatores que podem influenciar no processo de evasão estudantil: fatores indutores relacionados às características individuais, internas à instituição e externas à instituição, que foram minuciosamente estudados durante o desenvolvimento do MSL apresentado no capítulo 3. Para o presente estudo, as principais características que se mostraram relevantes para o problema da evasão estudantil foram: ano e semestre de conclusão ou desistência; ano e semestre de ingresso; renda mensal da família; aproveitamento acadêmico no primeiro ano de curso; raça; nível de escolaridade do pai; mudança de residência para morar na mesma cidade da Universidade; nível de escolaridade da mãe; tipo de instituição que estudou no ensino médio; se o estudante foi contemplado com bolsa durante o período que esteve matriculado; cadeia de pré-requisitos das disciplinas do curso ser adequada na concepção do estudante; o curso o qual se matriculou ter sido a primeira opção; possuir filhos.

Observa-se que essas variáveis foram analisadas para cada curso e que em alguns casos o perfil se mostrou distinto. No caso dos cursos de Engenharia Agrícola e Ambiental e Biotecnologia, os estudantes têm um perfil econômico melhor, o nível de escolaridade dos pais é mais elevado, tem prevalência de pessoas que se declaram do gênero feminino, que não possuem filhos, que não possuíam conhecimento técnico acerca do curso. Entretanto, no curso de Ciência e Tecnologia os estudantes são de rendas inferiores, pais com menor escolaridade, possuem filhos, boa parte já possuía conhecimento técnico prévio na área de estudo, tem prevalência de pessoas que se declaram do gênero masculino ou que preferem não se identificar, que não foram contemplados com bolsa e que não escolheram o curso

como sua primeira opção

Todas essas características foram detalhadas no capítulo de resultados.

QT 1 – Quais as abordagens encontradas na literatura para o desenvolvimento de modelos de predição de evasão estudantil? As principais abordagens aplicadas são: *Sample, Explore, Modify, Model e Assess (SEMMA)*, *Knowledge Discovery in Databases (KDD)* e *CRISP-DM*. O presente estudo adotou o *CRISP-DM*, conforme se detalha no capítulo 2.

QT 1.1 - Quais as etapas necessárias para a criação de um modelo preditivo de estudantes com alto risco de evasão estudantil? As etapas do *CRISP-DM* compreendem: Entendimento do negócio (*Business Understanding*), Entendimento dos Dados (*Data Understanding*), Preparação dos dados (*Data Preparation*), Modelagem (*Modeling*) e Avaliação (*Evaluation*), detalhadas no capítulo que trata dos materiais e métodos da presente pesquisa.

QT2 – Quais as principais ferramentas utilizadas para a aplicação de ML na predição de evasão estudantil? Conforme detalhado no capítulo de trabalhos relacionados, a linguagem de programação *Python* foi utilizada em grande parte dos trabalhos da MSL que especificam as ferramentas utilizadas. No presente estudo ela foi utilizada em conjunto com algumas bibliotecas para ML: *Pandas*; *Seaborn* e *Matplot*; *Numpy* e o *Scikit-learn*;

QP - Como avaliar a qualidade do modelo preditivo? Conforme se detalha no capítulo de fundamentação teórica, as métricas de avaliação possibilitam comparar o desempenho, bem como as limitações dos algoritmos de ML.

QP 1.1 - Quais métricas podem ser consideradas adequadas para a constatação dos resultados? No capítulo de MSL constatou-se que os trabalhos retornados aplicam uma ou mais métricas para avaliar os modelos de ML. Dentre elas as principais são: *Accuracy*, *Precision*, *Recall*, *F1-Score*, *ROC_AUC*, *confusion matrix* e *testes estatísticos como os de Friedman e o pós-teste (Nemenyi)*.

QP 1.2 – Como validar o processo de avaliação dos modelos de ML desenvolvidos? No capítulo de MSL também foi possível constatar que não existe uma melhor métrica para validar o processo, mas que cada métrica apresenta pontos importantes que podem definir a eficiência ou deficiência do modelo para o problema estudado. Conforme apresentado no capítulo dos resultados, o modelo de ML desenvolvido mostrou-se eficiente para prever quem conclui e isso evidencia-se pelo baixo *recall*. O modelo de

ML desenvolvido consegue prever quem conclui, mas se for algo diferente disso, ele apresenta dificuldades. Acredita-se que a dificuldade do modelo aprender quem evade é devido à baixa quantidade de amostras, o que tornou os dados de estudantes que evadiram insuficientes para o modelo aprender.

Os objetivos específicos foram alcançados em diferentes etapas da pesquisa:

1. Após a pesquisa e realização do MSL (Capítulo 4), foi possível compreender o estado da arte relacionado ao desenvolvimento de modelos de predição, no âmbito da evasão estudantil e suas metodologias, técnicas, ferramentas utilizadas e o panorama das pesquisas na área.
2. A presente pesquisa propôs, em toda a sua extensão, um modelo preditivo que proporciona a identificação da possibilidade de evasão estudantil em diferentes cursos superiores. A pesquisa foi além: propôs o modelo, bem como uma metodologia clara e possível de ser replicada.
3. O modelo preditivo foi avaliado por meio de métricas de avaliação (alcançando excelentes percentuais de acerto) e por métodos de análise estatística.
4. A pesquisa conseguiu avançar no estado da arte e apresentou nuances para o contexto do trabalho, diferentes do que foi encontrado na literatura, conforme foi discorrido neste documento.

As hipóteses, que nortearam a análise realizada, foram:

1. Estudantes que não ingressaram na sua primeira opção de curso, têm uma maior probabilidade de evadir-se. Essa hipótese foi confirmada e apresentada na sessão de resultados da análise comparativa entre as bases.
2. Estudantes que não possuem um conhecimento prévio na área do seu curso, têm uma maior probabilidade de evadir-se. Essa hipótese foi confirmada e apresentada na sessão de resultados da análise comparativa entre as bases.
3. O maior índice de evasão ocorre no primeiro ano de curso. Essa hipótese foi confirmada e apresentada na sessão de resultados da análise comparativa entre as bases.
4. Estudantes que reprovam em uma ou mais disciplinas no primeiro ano de curso, têm uma maior probabilidade de evadir-se. Essa hipótese foi confirmada e apresentada na sessão de resultados da análise comparativa entre as bases.
5. Estudantes do turno noturno, têm uma maior probabilidade de evadir-se. Essa

hipótese não foi confirmada nessa pesquisa. Recomenda-se que seja considerada em trabalhos futuros.

6. Estudantes do gênero feminino têm uma maior probabilidade de evadir-se. No caso dessa hipótese, ela se confirma para o curso de Ciência e Tecnologia, mas não é um fator determinante para os cursos de Biotecnologia e Engenharia Agrícola e Ambiental.
7. Estudantes que cursaram o ensino médio em escolas públicas, têm uma maior probabilidade de evadir-se. No caso dessa hipótese, ela se confirma para o curso de Ciência e Tecnologia, mas não se confirma para os cursos de Biotecnologia e Engenharia Agrícola e Ambiental.
8. Estudantes oriundos de outros estados, têm uma maior probabilidade de evadir-se. Essa hipótese não foi confirmada nessa pesquisa. Recomenda-se que seja considerada em trabalhos futuros.
9. Estudantes com pais separados, têm uma maior probabilidade de evadir-se. Essa hipótese não foi confirmada nessa pesquisa. Recomenda-se que seja considerada em trabalhos futuros.
10. O mesmo modelo de ML pode ser aplicado para grupos de estudantes de diferentes cursos. Essa hipótese foi confirmada no desenvolvimento deste trabalho, em que os mesmos modelos foram aplicados para 3 cursos superiores da UFERSA. Observa-se que, utilizando-se dos mesmos modelos de ML, configurações (*features*) e algoritmos, o modelo conseguiu realizar a predição dos dados. Entretanto, os resultados das métricas de avaliação foram distintos. Essa distinção pode ser interpretada pelo volume de dados desigual entre os dois conjuntos de dados, as classes de estudantes concluintes e desistentes que são desbalanceadas e por fim, pelas *features* utilizadas (talvez o segundo conjunto de dados obtivesse melhor resposta com outra combinação de variáveis).

6.1 Contribuições

Como principal contribuição deste trabalho, destaca-se a apresentação de uma metodologia ampla, contemplando todas as etapas para detectar previamente a probabilidade de alunos de diferentes cursos superiores evadirem. A metodologia aplicada foi pautada na literatura, no entanto, apresenta nuances na descrição de cada etapa desenvolvida,

facilitando a reprodutibilidade em outros cenários.

Ainda como contribuições, destaca-se a ampliação da pesquisa realizada por Marques (2020) nos seguintes aspectos:

- ampliação da pesquisa bibliográfica na qual foram adicionadas novas questões de pesquisa e retorno de uma maior quantidade de trabalhos relacionados;
- utilização da metodologia CRISP-DM;
- adição de novas variáveis correlacionadas ao fenômeno da evasão;
- implementação em 3 cursos superiores distintos;
- maior número de algoritmos implementados;
- utilização de diferentes métricas de avaliação e aplicação de testes estatísticos.

6.2 Limitações

Dentre as principais limitações do trabalho, é relevante enfatizar a dificuldade para construir a base de dados. Apesar dos incansáveis esforços dos pesquisadores e o auxílio da PROGRAD, na UFERSA, não foi possível alcançar uma quantidade expressiva de estudantes, em especial dos desistentes. Tornando-se relevante desenvolver, em futuros experimentos, novos procedimentos com bases da própria Universidade.

Destaca-se também como limitação da pesquisa a falta de trabalhos que descrevem detalhadamente a metodologia. Apesar da quantidade de trabalhos retornados no MSL, a maioria dificulta a reprodutibilidade pelo fato de não descrever em detalhes a metodologia adotada.

6.3 Trabalhos futuros

Em trabalhos futuros, recomenda-se:

Realizar o aprimoramento da aprendizagem dos modelos preditivos, de modo a realizar uma nova análise e estudo de variáveis significativas para o segundo conjunto de dados e também a aquisição de novos dados para ambos os conjuntos. São exemplos de variáveis:

- se os pais do estudante se divorciaram durante o período em que cursou a faculdade;
- se o estudante já foi diagnosticado ou acredita que sofre com problemas de déficit de atenção, tais como TDAH ou autismo;

- se o estudante já foi diagnosticado ou acredita que sofre com problemas de depressão;
- perguntar abertamente o motivo da desistência do estudante.

Após uma nova coleta, que esses novos dados possam ser estudados, correlacionando-os com a decisão do estudante de abandonar o curso.

Trabalhar, de forma unitária, os dois conjuntos de dados que foram utilizados como estudo de caso na presente pesquisa.

Aplicar o modelo implementado na presente pesquisa com os dados da Universidade, bem como os dados adquiridos por Marques (2020) em sua pesquisa.

Implementar os algoritmos em forma de comitês em ambos os conjuntos de dados da presente pesquisa.

Aplicar estratégias para cortar dados desbalanceados ou gerar dados aleatórios, garantindo um conjunto mais robusto.

Por fim, para trabalhos futuros, recomenda-se ainda a implementação dos modelos desenvolvidos aplicados a outros cursos superiores da UFERSA.

REFERÊNCIAS

- ACADEMY, D. S. O perceptron. In: ATA SCIENCE ACADEMY. **Deep Learning Book**. 2022. p. 200. Disponível em: <<https://www.deeplearningbook.com.br/o-perceptron-parte-1/>>.
- ACERO, A.; ACHURY, J.; MORALES, J. University dropout: A prediction model for an engineering program in bogotá, colombia. In: **8th Research in Engineering Education Symposium: Making Connections, REES**. [S.l.: s.n.], 2019. p. 483–490.
- ADELMAN, M.; HAIMOVICH, F.; HAM, A.; VAZQUEZ, E. Predicting school dropout with administrative data: New evidence from guatemala and honduras. **Education Economics**, Taylor & Francis, v. 26, n. 4, p. 356–372, 2018.
- AFIA, K.; DION, E.; DUPÉRÉ, V.; ARCHAMBAULT, I.; TOSTE, J. Parenting practices during middle adolescence and high school dropout. **Journal of Adolescence**, v. 76, p. 55–64, 2019. ISSN 0140-1971. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0140197119301496>>.
- AGRUSTI, F.; BONA VOLONTÀ, G.; MEZZINI, M. University dropout prediction through educational data mining techniques: A systematic review. **Journal of E-Learning and Knowledge Society**, v. 15, n. 3, p. 161–182, 2019.
- ASHBY, A. Monitoring student retention in the open university: Definition, measurement, interpretation and action. **Open Learning: The Journal of Open, Distance and e-Learning**, Taylor & Francis, v. 19, n. 1, p. 65–77, 2004.
- ATTWOOD, G.; CROLL, P. Truancy in secondary school pupils: Prevalence, trajectories and pupil perspectives. **Research papers in education**, Taylor & Francis, v. 21, n. 4, p. 467–484, 2006.
- AYELE, W. Y. Adapting crisp-dm for idea mining: a data mining process for generating ideas using a textual dataset. **International Journal of Advanced Computer Sciences and Applications**, v. 11, n. 6, p. 20–32, 2020.
- BAHEL, V.; BAJAJ, P.; THOMAS, A. Knowledge discovery in educational databases in indian educational system: A case study of ghrce, nagpur. In: IEEE. **2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)**. [S.l.], 2019. p. 235–239.
- BARROS, R. P. de. **Consequências da violação do direito à educação**. [S.l.]: Editora Autografia, 2021.
- BARROS, T. M.; NETO, P. A. S.; SILVA, I.; GUEDES, L. A. Predictive models for imbalanced data: A school dropout perspective. **Education Sciences**, Multidisciplinary Digital Publishing Institute, v. 9, n. 4, p. 275, 2019.
- BARROS, T. M.; SILVA, I.; GUEDES, L. A. Determination of dropout student profile based on correspondence analysis technique. **IEEE Latin America Transactions**, IEEE, v. 17, n. 09, p. 1517–1523, 2019.
- BARROSO, M. F.; FALCÃO, E. B. Evasão universitária: o caso do instituto de física da ufrj. **IX Encontro Nacional de Pesquisa em Ensino de Física**, v. 9, p. 1–14, 2004.

BENGESAI, A. V.; POCKOCK, J. Patterns of persistence among engineering students at a south african university: A decision tree analysis. **South African Journal of Science**, Academy of Science of South Africa, v. 117, n. 3-4, p. 1-9, 2021.

BISONG, E. Matplotlib and seaborn. In: **Building machine learning and deep learning models on google cloud platform**. [S.l.]: Springer, 2019. p. 151-165.

BISWAS, A. M. A. A.; MIA, M. J.; NOWRIN, I.; RITU, N. A. Predicting the enrollment and dropout of students in the post-graduation degree using machine learning classifier. 2019.

BITTENCOURT, I. I.; CUKUROVA, M.; MULDER, K.; LUCKIN, R.; MILLÁN, E. **Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6-10, 2020, Proceedings, Part I**. 1st ed.. ed. Springer International Publishing;Springer, 2020. (Lecture Notes in Computer Science 12163). ISBN 978-3-030-52236-0,978-3-030-52237-7. Disponível em: <<http://gen.lib.rus.ec/book/index.php?md5=006e4cf9fd88f344c740ab4b4ecf9484>>.

BLOCK, H.-D. The perceptron: A model for brain functioning. i. **Reviews of Modern Physics**, APS, v. 34, n. 1, p. 123, 1962.

BONACCORSO, G. **Machine learning algorithms**. [S.l.]: Packt Publishing Ltd, 2017.

BOYACI, A. Exploring the factors associated with the school dropout. **International Electronic Journal of Elementary Education**, 2019.

BRADSHAW, C. P.; O'BRENNAN, L. M.; MCNEELY, C. A. Core competencies and the prevention of school failure and early school leaving. **New directions for child and adolescent Development**, Wiley Online Library, v. 2008, n. 122, p. 19-32, 2008.

BRANDÃO, I. V.; COSTA, J. P. C. L. da; SANTOS, G. A.; PRACIANO, B. J. G.; JÚNIOR, F. C.; JÚNIOR, R. T. de S. Classification and predictive analysis of educational data to improve the quality of distance learning courses. In: **2019 Workshop on Communication Networks and Power Systems (WCNPS)**. [S.l.: s.n.], 2019. p. 1-6.

BROCK, K. R. **Identifying the factors that predict degree completion for entirely online community college students**. Tese (Doutorado) — Capella University, 2014.

BUSS, J. F.; ROSENBERG, A. L.; KNOTT, J. D. **Vertex Types in Book-Embeddings**. Amherst, MA, USA, 1987.

CAIRNS, R. B.; CAIRNS, B. D.; NECKERMAN, H. J. Early school dropout: Configurations and determinants. **Child development**, JSTOR, p. 1437-1452, 1989.

CAMPOS, A. de; GALAFASSI, C.; BASTIANI, E.; PAZ, F. J.; CAMPOS, R. L.; WIVES, L. K.; CAZELLA, S. C.; REATEGUI, E. B.; BARONE, D. A. C. Mineração de dados educacionais e learning analytics no contexto educacional brasileiro: um mapeamento sistemático. **Informática na educação: teoria & prática**, v. 23, n. 3 Set/Dez, 2020.

CARVALHO, M. G. de; FEITOSA, S.; SILVA, V. C. da. Relações de gênero entre alunos e alunas em uma instituição de educação tecnológica brasileira. **Revista Tecnologia e Sociedade**, Universidade Tecnológica Federal do Paraná, v. 2, n. 3, p. 87-135, 2006.

CHERKASSKY, V.; MA, Y. Practical selection of svm parameters and noise estimation for svm regression. **Neural networks**, Elsevier, v. 17, n. 1, p. 113–126, 2004.

CHUNG, J. Y.; LEE, S. Dropout early warning systems for high school students using machine learning. **Children and Youth Services Review**, Elsevier, v. 96, p. 346–353, 2019.

COOK, E. S. An analysis of factors related to withdrawal from high school prior to graduation. **The Journal of Educational Research**, Taylor & Francis, v. 50, n. 3, p. 191–196, 1956.

COSTA, E.; BAKER, R. S.; AMORIM, L.; MAGALHÃES, J.; MARINHO, T. Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. **Jornada de Atualização em Informática na Educação**, v. 1, n. 1, p. 1–29, 2013.

CRUZ-JESUS, F.; CASTELLI, M.; OLIVEIRA, T.; MENDES, R.; NUNES, C.; SA-VELHO, M.; ROSA-LOURO, A. Using artificial intelligence methods to assess academic achievement in public high schools of a european union country. **Heliyon**, Elsevier, v. 6, n. 6, p. e04081, 2020.

CUTLER, A.; CUTLER, D. R.; STEVENS, J. R. Random forests. In: **Ensemble machine learning**. [S.l.]: Springer, 2012. p. 157–175.

DEEPIKA, K.; SATHVANARAVANA, N. Analyze and predicting the student academic performance using data mining tools. In: IEEE. **2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)**. [S.l.], 2018. p. 76–81.

DESCHACHT, N.; GOEMAN, K. The effect of blended learning on course persistence and performance of adult learners: A difference-in-differences analysis. **Computers & Education**, Elsevier, v. 87, p. 83–89, 2015.

DOMBROVSKAIA, L.; JOSÉ, P.; RODRÍGUEZ, P. Prediction of student's retention in first year of engineering program at a technological chilean university. In: IEEE. **2020 39th International Conference of the Chilean Computer Science Society (SCCC)**. [S.l.], 2020. p. 1–4.

DRA. **Divisão de Registro Acadêmico**. 2020. Disponível em: <<https://dra.ufersa.edu.br/>>.

DUNN, O. J. Multiple comparisons among means. **Journal of the American statistical association**, Taylor & Francis, v. 56, n. 293, p. 52–64, 1961.

ELHORST, J. P. Matlab software for spatial panels. **International Regional Science Review**, SAGE Publications Sage CA: Los Angeles, CA, v. 37, n. 3, p. 389–405, 2014.

Student dropout from foundation program at modern college of business science, sultanate of oman. v. 8, 2019.

FARIA, S. M. S. M. L. d. **Educational data mining e learning analytics na melhoria do ensino online**. Tese (Doutorado), 2014.

FERNÁNDEZ-GARCÍA, A. J.; PRECIADO, J. C.; MELCHOR, F.; RODRIGUEZ-ECHEVERRIA, R.; CONEJERO, J. M.; SÁNCHEZ-FIGUEROA, F. A real-life machine learning experience for predicting university dropout at different stages using academic data. **IEEE Access**, IEEE, v. 9, p. 133076–133090, 2021.

FIGUEIRA, R. J. **The applicability of Tinto's model of student retention in online learning A faculty perspective**. [S.l.]: Wilmington University (Delaware), 2015.

FILHO, R. L. L. S.; MOTEJUNAS, P. R.; HIPÓLITO, O.; LOBO, M. B. d. C. M. A evasão no ensino superior brasileiro. **Cadernos de pesquisa**, SciELO Brasil, v. 37, p. 641–659, 2007.

FRANKO, D. L. **Increasing online academic success and persistence in higher education using coaching**. Tese (Doutorado) — Northcentral University, 2015.

FREITAS, F. A. d. S.; VASCONCELOS, F. F.; PEIXOTO, S. A.; HASSAN, M. M.; DEWAN, M.; ALBUQUERQUE, V. H. C. d. *et al.* Iot system for school dropout prediction using machine learning techniques based on socioeconomic data. **Electronics**, Multidisciplinary Digital Publishing Institute, v. 9, n. 10, p. 1613, 2020.

FRIEDMAN, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. **Journal of the american statistical association**, Taylor & Francis, v. 32, n. 200, p. 675–701, 1937.

FRIEDMAN, M. A comparison of alternative tests of significance for the problem of m rankings. **The Annals of Mathematical Statistics**, JSTOR, v. 11, n. 1, p. 86–92, 1940.

FRNANDES, G. **Modelos de Predição | Decision Tree**. 2019. Disponível em: <<https://medium.com/turing-talks/turing-talks-17-modelos-de-predi%C3%A7%C3%A3o\discretionary{-}{ }{ }{ }decision\discretionary{-}{ }{ }{ }tree\discretionary{-}{ }{ }{ }610aa484cb05>>.

FUZETO, R.; BRAGA, R. Um mapeamento sistemático em progresso sobre internet das coisas e educação à distância. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. [S.l.: s.n.], 2016. v. 5, n. 1, p. 1334.

GAIOSO, N. P. d. L. O fenômeno da evasão escolar na educação superior no brasil. **Brasília, DF: Universidade Católica de Brasília**, p. 20, 2005.

GALLANT, S. I. *et al.* Perceptron-based learning algorithms. **IEEE Transactions on neural networks**, v. 1, n. 2, p. 179–191, 1990.

GARDNER, M. W.; DORLING, S. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. **Atmospheric environment**, Elsevier, v. 32, n. 14-15, p. 2627–2636, 1998.

GERMAN, D. M.; ADAMS, B.; HASSAN, A. E. The evolution of the r software ecosystem. In: IEEE. **2013 17th European Conference on Software Maintenance and Reengineering**. [S.l.], 2013. p. 243–252.

GILBERT, J.; HAMID, S. User profiling to overcome the social effects of student dropout. In: ZAMAN, H. B.; SMEATON, A. F.; SHIH, T. K.; VELASTIN, S.; TERUTOSHI, T.; ALI, N. M.; AHMAD, M. N. (Ed.). **Advances in Visual Informatics**. Cham: Springer International Publishing, 2019. p. 663–671. ISBN 978-3-030-34032-2.

GRANDINI, M.; BAGLI, E.; VISANI, G. Metrics for multi-class classification: an overview. **arXiv preprint arXiv:2008.05756**, 2020.

GUNAWAN, H. Catherine: Information systems students' study performance prediction using data mining approach. In: **Fourth International Conference on Informatics and Computing**. [S.l.: s.n.], 2019. p. 1–8.

HASAN, M. N. A comparison of logistic regression and linear discriminant analysis in predicting of female students attrition from school in bangladesh. In: IEEE. **2019 4th International Conference on Electrical Information and Communication Technology (EICT)**. [S.l.], 2019. p. 1–3.

HAUGAN, J.; FROSTAD, P.; MJAAVATN, P. A longitudinal study of factors predicting students' intentions to leave upper secondary school in norway. **Soc Psychol Education**, 2019.

HERBOLD, S. Autorank: A python package for automated ranking of classifiers. **Journal of Open Source Software**, v. 5, n. 48, p. 2173, 2020.

HEVNER, A. R.; MARCH, S. T.; PARK, J.; RAM, S. Design science in information systems research. **MIS quarterly**, JSTOR, p. 75–105, 2004.

HIPPEL, P. T. V.; HOFFLINGER, A. The data revolution comes to higher education: identifying students at risk of dropout in chile. **Journal of Higher Education Policy and Management**, Taylor & Francis, v. 43, n. 1, p. 2–23, 2021.

HOSSIN, M.; SULAIMAN, M. N. A review on evaluation metrics for data classification evaluations. **International journal of data mining & knowledge management process**, Academy & Industry Research Collaboration Center (AIRCC), v. 5, n. 2, p. 1, 2015.

HUBER, S.; WIEMER, H.; SCHNEIDER, D.; IHLENFELDT, S. Dmme: Data mining methodology for engineering applications—a holistic extension to the crisp-dm model. **Procedia Cirp**, Elsevier, v. 79, p. 403–408, 2019.

IBGE. **Estatísticas de gênero**. 2018. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv101551_informativo.pdf>.

IBGE. **Pesquisa Nacional por Amostra de Domicílios Contínua - Os Sistema Educacional Brasileiro**. 2019. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv101736_informativo.pdf>.

IDEB. **Índice de Desenvolvimento da Educação Básica. Formação em Ação**. 2012. Disponível em: <http://www.nre.seed.pr.gov.br/cascavel/File/CIENCIAS_IndicedeDesenvolvimentodaEducaÃoBásica.pdf>.

IGNACIO, L. F. F. **Aprendizado de máquina: da teoria à aplicação**. Volta Redonda, 2021.

ILIEVA, G.; YANKOVA, T. Early multi-criteria detection of students at risk of failure. **TEM Journal**, UIKTEN-Association for Information Communication Technology Education and . . . , v. 9, n. 1, p. 344–350, 2020.

INSTITUTE, A. B. **2020 Top Companies for Women Technologists Results**. 2021. Disponível em: <<https://anitab.org/research-and-impact/top-companies/2020-results/>>.

JAISWAL, G.; SHARMA, A.; YADAV, S. K. . Analytical approach for predicting dropouts in higher education. **Technology Educations**, 2019.

JANUSCHOWSKI, T.; GASTHAUS, J.; WANG, Y. B. Open-source forecasting tools in python. 2019.

JIMENEZ, F.; PAOLETTI, A.; SANCHEZ, G.; SCIAVICCO, G. Predicting the risk of academic dropout with temporal multi-objective optimization. **IEEE Transactions on Learning Technologies**, IEEE, v. 12, n. 2, p. 225–236, 2019.

JOSÉ, A. R.; BROILO, C. L.; ANDREOLI, G. S. A evasão na unipampa–diagnosticando processos, acompanhando trajetórias e itinerários de formação. **Universidade Federal do Pampa**, 2010.

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

KISS, B.; NAGY, M.; MOLONTAY, R.; CSABAY, B. Predicting dropout using high school and first-semester academic achievement measures. In: IEEE. **2019 17th International Conference on Emerging eLearning Technologies and Applications (ICETA)**. [S.l.], 2019. p. 383–389.

KITCHENHAM, B. Procedures for performing systematic reviews. **Keele, UK, Keele University**, v. 33, n. 2004, p. 1–26, 2004.

KRAMER, O. Scikit-learn. In: **Machine learning for evolution strategies**. [S.l.]: Springer, 2016. p. 45–53.

LANES, M.; ALCÂNTARA, C. Predição de alunos com risco de evasão: estudo de caso usando mineração de dados. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. [S.l.: s.n.], 2018. v. 29, n. 1, p. 1921.

LANGLEY, P. *et al.* The changing science of machine learning. **Machine learning**, Citeseer, v. 82, n. 3, p. 275–279, 2011.

LEE, N. Interview with bill kinder: January 13, 2005. **Comput. Entertain.**, v. 3, n. 1, Jan.-March 2005. Disponível em: <<http://doi.acm.org/10.1145/1057270.1057278>>.

LEMKIN, A.; KISTIN, C. J.; CABRAL, H. J.; ASCHENGRAU, A.; BAIR-MERRITT, M. School connectedness and high school graduation among maltreated youth. **Child Abuse & Neglect**, v. 75, p. 130–138, 2018. ISSN 0145-2134. Highlighting Education and Learning in the Context of Childhood Abuse, Neglect, and Related Stressors. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0145213417301722>>.

LEMKIN, A.; KISTIN, C. J.; CABRAL, H. J.; ASCHENGRAU, A.; BAIR-MERRITT, M. Wwh-dropout scale: when, why and how to measure propensity to drop out of undergraduate courses. **Journal of Applied Research in Higher Education, ahead-of-print(ahead-of-print)**, 2020.

LIMA, J.; ALVES, P.; PEREIRA, M.; ALMEIDA, S. Using academic analytics to predict dropout risk in engineering courses. In: ACADEMIC CONFERENCES AND PUBLISHING INTERNATIONAL LIMITED. **17th European Conference on e-Learning ECEL 2018**. [S.l.], 2018. p. 316–321.

LIMSATHITWONG, K.; TIWATTHANONT, K.; YATSUNGNOEN, T. Dropout prediction system to reduce discontinue study rate of information technology students. In: **2018 5th International Conference on Business and Industrial Research (ICBIR)**. [S.l.: s.n.], 2018. p. 110–114.

LOBO, M. Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. **Associação Brasileira de Mantenedoras de Ensino Superior. Cadernos**, v. 25, p. 14, 2012.

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

MACY, T. G. **The effect of web-based instruction on retention of non-traditional students in a rural comprehensive university.** Tese (Doutorado) — Eastern Kentucky University, 2015.

MAHESH, B. Machine learning algorithms-a review. **International Journal of Science and Research (IJSR)**. [Internet], v. 9, p. 381–386, 2020.

MANI, I.; ZHANG, I. knn approach to unbalanced data distributions: a case study involving information extraction. In: ICML. **Proceedings of workshop on learning from imbalanced datasets**. [S.l.], 2003. v. 126, p. 1–7.

MARQUES, L. T. Mateo: uma abordagem de descoberta de conhecimento para desvendar as causas da evasão escolar. Universidade Federal Rural do Semi-Árido, 2020.

MARQUES, L. T.; CASTRO, A. F. D.; MARQUES, B. T.; SILVA, J. C. P.; QUEIROZ, P. G. G. Mineração de dados auxiliando na descoberta das causas da evasão escolar: Um mapeamento sistemático da literatura. **RENOTE**, v. 17, n. 3, p. 194–203, 2019.

MARTINS, W.; AFONSECA, U. R.; NALINI, L. E.; GOMES, V. M. Tutoriais inteligentes baseados em aprendizado por reforço: concepção, implementação e avaliação empírica. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. [S.l.: s.n.], 2007. v. 1, n. 1, p. 550–559.

MATOS, I. A.; VASCONCELOS, D. M. B. P.; SANTOS, M. C. O. d. . a evasão escolar no ifto campi araguaína e porto nacional: da estatística ao enfrentamento. **IV Colóquio Internacional sobre Educação Profissional e Evasão Escolar: Caderno de Resumos Expandidos**, 2015.

MATSUBARA, E. T. **O algoritmo de aprendizado semi-supervisionado co-training e sua aplicação na rotulação de documentos.** Tese (Doutorado) — Universidade de São Paulo, 2004.

MCKINNEY, W. Pandas, python data analysis library. URL <http://pandas.pydata.org>, 2015.

MDUMA, N.; KALEGELE, K.; MACHUVE, D. Machine learning approach for reducing students dropout rates. **International Journal of Advanced Computer Research**, 2019.

MEC, M. d. E. **Censo da Educação Superior.** 2019. Disponível em: <<https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior/resultados>>.

MEC, M. d. E. **Censo da Educação Superior**. 2022. Disponível em: <https://download.inep.gov.br/publicacoes/institucionais/estatisticas_e_indicadores/resumo_tecnico_censo_da_educacao_superior_2020.pdf>.

MECA, I.; RABASA, A.; SOBRINO, E.; LÓPEZ-ESPÍN, J. J. Early warning methodology for dropping out of university degrees. In: **Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality**. [S.l.: s.n.], 2020. p. 245–249.

MEENS, E. E.; BAKX, A. W.; KLIMSTRA, T. A.; DENISSEN, J. J. The association of identity and motivation with students' academic achievement in higher education. **Learning and Individual Differences**, v. 64, p. 54–70, 2018. ISSN 1041-6080. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1041608018300657>>.

MIORANZA, D. T. *et al.* Evasão nos cursos de graduação do instituto federal catarinense: um estudo a partir da mineração de dados. 2020.

MITCHELL, T. M.; MITCHELL, T. M. **Machine learning**. [S.l.]: McGraw-hill New York, 1997. v. 1.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. **Sistemas inteligentes-Fundamentos e aplicações**, Manole, v. 1, n. 1, p. 32, 2003.

MORO, S.; LAUREANO, R.; CORTEZ, P. Using data mining for bank direct marketing: An application of the crisp-dm methodology. EUROSIS-ETI, 2011.

MORTAGY, Y.; BOGHIKIAN-WHITBY, S.; HELOU, I. An analytical investigation of the characteristics of the dropout students in higher education. **Issues in Informing Science and Information Technology**, v. 15, p. 30, 2018. ISSN 0190-7409. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0190740919314641>>.

MUSCHELLI, J. Roc and auc with a binary predictor: a potentially misleading metric. **Journal of classification**, Springer, v. 37, n. 3, p. 696–708, 2020.

NACIF, P. G.; MENDES, R.; LÁZARO, A.; EVARISTO, M. Inclusão educacional na pandemia. 2020.

NADASEN, D. **Innovations and student success in online learning: A systematic review of how innovations affect student retention**. Tese (Doutorado) — University of Maryland University College, 2016.

NAGY, M.; MOLONTAY, R. Predicting dropout in higher education based on secondary school performance. In: IEEE. **2018 IEEE 22nd international conference on intelligent engineering systems (INES)**. [S.l.], 2018. p. 000389–000394.

NAIK, A.; SAMANT, L. Correlation review of classification algorithm using data mining tool: Weka, rapidminer, tanagra, orange and knime. **Procedia Computer Science**, Elsevier, v. 85, p. 662–668, 2016.

NÁJERA, A. B. U.; HAMPSHIRE, J. C. C.; BARBOSA, R. C. University dropout: Prevention patterns through the application of educational data mining. 2020.

NASCIMENTO, R. L. S. do; JUNIOR, R. B. das N.; NETO, M. A. de A.; FAGUNDES, R. A. de A. Educational data mining: An application of regressors in predicting school dropout. In: PERNER, P. (Ed.). **Machine Learning and Data Mining in Pattern Recognition**. Cham: Springer International Publishing, 2018. p. 246–257. ISBN 978-3-319-96133-0.

NUANKAEW, P.; NUANKAEW, W.; PHANNIPHONG, K.; BUSSAMAN, S. Students disruption in business computer at the university of phayao, thailand. In: **2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT NCON)**. [S.l.: s.n.], 2020. p. 95–100.

OLANOW, C. W.; KOLLER, W. C. An algorithm (decision tree) for the management of parkinson's disease: treatment guidelines. **Neurology**, AAN Enterprises, v. 50, n. 3 Suppl 3, p. S1–S1, 1998.

OLIPHANT, T. E. **A guide to NumPy**. [S.l.]: Trelgol Publishing USA, 2006. v. 1.

OPAZO, D.; MORENO, S.; ÁLVAREZ-MIRANDA, E.; PEREIRA, J. Analysis of first-year university student dropout through machine learning models: A comparison between universities. **Mathematics**, Multidisciplinary Digital Publishing Institute, v. 9, n. 20, p. 2599, 2021.

OROOJI, M.; CHEN, J. Predicting louisiana public high school dropout through imbalanced learning techniques. In: IEEE. **2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)**. [S.l.], 2019. p. 456–461.

ORTIGOSA, A.; CARRO, R. M.; BRAVO-AGAPITO, J.; LIZCANO, D.; ALCOLEA, J. J.; BLANCO, From lab to production: Lessons learnt and real-life challenges of an early student-dropout prevention system. **IEEE Transactions on Learning Technologies**, v. 12, n. 2, p. 264–277, 2019.

PAŠIĆ, Đ.; KUČAK, D. Machine learning model for detecting high school students as candidates for drop-out from a study program. In: IEEE. **2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)**. [S.l.], 2020. p. 1140–1144.

PATLE, A.; CHOUHAN, D. S. Svm kernel functions for classification. In: IEEE. **2013 International Conference on Advances in Technology and Engineering (ICATE)**. [S.l.], 2013. p. 1–9.

PERTIWI, A. G.; WIDYANINGTYAS, T.; PUJANTO, U. Classification of province based on dropout rate using c4. 5 algorithm. In: IEEE. **2017 International Conference on Sustainable Information Engineering and Technology (SIET)**. [S.l.], 2017. p. 410–413.

PETERSEN, K.; FELDT, R.; MUJTABA, S.; MATTSSON, M. Systematic mapping studies in software engineering. In: **12th International Conference on Evaluation and Assessment in Software Engineering (EASE) 12**. [S.l.: s.n.], 2008. p. 1–10.

PNE. **Plano Nacional da Educação**. 2014. Disponível em: <<https://pne.mec.gov.br/>>.

- POTDAR, K.; PARDAWALA, T. S.; PAI, C. D. A comparative study of categorical variable encoding techniques for neural network classifiers. **International journal of computer applications**, v. 175, n. 4, p. 7–9, 2017.
- QUINLAN, J. R. Learning decision tree classifiers. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 28, n. 1, p. 71–72, 1996.
- RATRA, R.; GULIA, P. Experimental evaluation of open source data mining tools (weka and orange). **Int. J. Eng. Trends Technol**, v. 68, n. 8, p. 30–35, 2020.
- REFAELZADEH, P.; TANG, L.; LIU, H. Cross-validation. **Encyclopedia of database systems**, Springer, v. 5, p. 532–538, 2009.
- REID, K. The causes, views and traits of school absenteeism and truancy: An analytical review. **Research in education**, SAGE Publications Sage UK: London, England, v. 74, n. 1, p. 59–82, 2005.
- RIFFEL, S. M.; MALACARNE, V. Evasão escolar no ensino médio: o caso do colégio estadual santo agostinho no município de palotina. **O professor PDE e os desafios da escola pública paranaense**, v. 1, p. 01–24, 2010.
- RIPAMONTI, E. Risk factors for dropping out of high school: A review of contemporary, international empirical research. **Adolescent Research Review**, Springer, v. 3, n. 3, p. 321–338, 2018.
- RISH, I. *et al.* An empirical study of the naive bayes classifier. In: **IJCAI 2001 workshop on empirical methods in artificial intelligence**. [S.l.: s.n.], 2001. v. 3, n. 22, p. 41–46.
- RODRÍGUEZ-MUÑIZ, L. J.; BERNARDO, A. B.; ESTEBAN, M.; DÍAZ, I. Dropout and transfer paths: What are the risky profiles when analyzing university persistence with machine learning techniques? **Plos one**, Public Library of Science San Francisco, CA USA, v. 14, n. 6, p. e0218796, 2019.
- ROLON-MÉRETTE, D.; ROSS, M.; ROLON-MÉRETTE, T.; CHURCH, K. Introduction to anaconda and python: Installation and setup. **Python for research in psychology**, v. 16, n. 5, p. S5–S11, 2016.
- RUMBERGER, R. W. Dropping out of middle school: A multilevel analysis of students and schools. **American educational Research journal**, Sage Publications, v. 32, n. 3, p. 583–625, 1995.
- SAMUEL, A. L. Machine learning. **The Technology Review**, v. 62, n. 1, p. 42–45, 1959.
- SANDOVAL-PALIS, I.; NARANJO, D.; VIDAL, J.; GILAR-CORBI, R. Early dropout prediction model: A case study of university leveling course students. **Sustainability**, Multidisciplinary Digital Publishing Institute, v. 12, n. 22, p. 9314, 2020.
- SANSONE, D. Beyond early warning indicators: High school dropout and machine learning. **Oxford Bulletin of Economics and Statistics**, 2018.
- SARI, E. Y.; SUNYOTO, A. *et al.* Optimization of weight backpropagation with particle swarm optimization for student dropout prediction. In: IEEE. **2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)**. [S.l.], 2019. p. 423–428.

SCHEUER, O.; MCLAREN, B. M. Educational data mining. **Encyclopedia of the Sciences of Learning**, Citeseer, p. 1075–1079, 2012.

SCHWAB, S. Who intends to learn and who intends to leave? the intention to leave education early among students from inclusive and regular classes in primary and secondary schools. **School Effectiveness and School Improvement**, Routledge, v. 29, n. 4, p. 573–589, 2018. Disponível em: <<https://doi.org/10.1080/09243453.2018.1481871>>.

SERRA, A.; PERCHINUNNO, P.; BILANCIA, M. Predicting student dropouts in higher education using supervised classification algorithms. In: SPRINGER. **International Conference on Computational Science and Its Applications**. [S.l.], 2018. p. 18–33.

SHIAU, Y. University dropout prevention through the application of big data. In: **Proceedings of the 2020 3rd International Conference on Information Management and Management Science**. [S.l.: s.n.], 2020. p. 1–7.

SILVA, A. L. d.; QUEIROZ, A. P. T. d.; SANTOS, V. C. A. d. Evasão na educação profissional: identificando caminhos de permanência sob o olhar da assistência estudantil. **V Colóquio Internacional sobre Educação Profissional e Evasão Escolar: Caderno de Resumos Expandidos**, 2015.

SILVA, I.; MOODY, G. B. An open-source toolbox for analysing and processing physionet databases in matlab and octave. **Journal of open research software**, NIH Public Access, v. 2, n. 1, 2014.

SILVA, P. M. da; LIMA, M. N. C. A.; SOARES, W. L.; SILVA, I. R. R.; FAGUNDES, R. A. de A.; SOUZA, F. F. de. Ensemble regression models applied to dropout in higher education. In: **2019 8th Brazilian Conference on Intelligent Systems (BRACIS)**. [S.l.: s.n.], 2019. p. 120–125.

SINGH, A.; THAKUR, N.; SHARMA, A. A review of supervised machine learning algorithms. In: IEEE. **2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)**. [S.l.], 2016. p. 1310–1315.

SOFYAN, Y.; KURNIAWAN, H. Teknik analisis statistik terlengkap dengan software spss. **Salemba Infotek, Jakarta**, 2009.

SORENSEN, L. C. “big data” in educational administration: An application for predicting school dropout risk. **Educational Administration Quarterly**, SAGE Publications Sage CA: Los Angeles, CA, v. 55, n. 3, p. 404–446, 2019.

STANČIN, I.; JOVIĆ, A. An overview and comparison of free python libraries for data mining and big data analysis. In: IEEE. **2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)**. [S.l.], 2019. p. 977–982.

TAN, M.; SHAO, P. Prediction of student dropout in e-learning program through the use of machine learning method. **International journal of emerging technologies in learning**, v. 10, n. 1, 2015.

TEAM, R. C. R language definition. **Vienna, Austria: R foundation for statistical computing**, v. 3, n. 1, 2000.

TECH, D. **O que é e como funciona o algoritmo KNN**. 2022. Disponível em: <<https://didatica.tech/o-que-e-e-como-funciona-o-algoritmo-knn/#:~:text=Em%20resumo%2C%20o%20KNN%20tenta,quest%C3%A3o%20ser%C3%A1%20classificada%20nesta%20categoria.>>

TENPIPAT, W.; AKKARAJITSAKUL, K. Student dropout prediction: A knutt case study. In: IEEE. **2020 1st International Conference on Big Data Analytics and Practices (IBDAP)**. [S.l.], 2020. p. 1–5.

THORNBURG, H. **Introduction to Bayesian Statistics**. 2001. Disponível em: <<http://ccrma.stanford.edu/~jos/bayes/bayes.html>>.

TIMBAL, M. A. Analysis of student-at-risk of dropping out (sardo) using decision tree: An intelligent predictive model for reduction. **International Journal of Machine Learning and Computing**, 2019.

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

TRAN, H. A survey of machine learning and data mining techniques used in multimedia system. **no**, v. 113, p. 13–21, 2019.

UFERSA, N. **Assistência Estudantil**. 2022. Disponível em: <<https://numeros.ufersa.edu.br/assistencia-estudantil/>>.

VAUGHN, M. G.; ROBERTS, G.; FALL, A.-M.; KREMER, K.; MARTINEZ, L. Preliminary validation of the dropout risk inventory for middle and high school students. **Children and Youth Services Review**, v. 111, p. 104855, 2020. ISSN 0190-7409. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0190740919314641>>.

VELEZ, W. High school attrition among hispanic and non-hispanic white youths. **Sociology of Education**, JSTOR, p. 119–133, 1989.

VELOSO, T. C. M.; ALMEIDA, E. P. de. Evasão nos cursos de graduação da universidade federal de mato grosso, campus universitário de cuiabá—um processo de exclusão. **Série-Estudos-Periódico do Programa de Pós-Graduação em Educação da UCDB**, 2002.

VENKATESAN, R.; MANIKANDAN, V.; YUVARAJ, D.; AHAMED, A. M. U. Predicting students' academic drop out and failures using data mining techniques. 2019.

VERMA, A.; RANGA, V. Machine learning based intrusion detection systems for iot applications. **Wireless Personal Communications**, Springer, v. 111, p. 2287–2310, 2020.

VOGT, L. A.; JORDAN, C.; THARP, R. G. Explaining school failure, producing school success: Two cases. **Anthropology & Education Quarterly**, Wiley Online Library, v. 18, n. 4, p. 276–286, 1987.

WANG, H.; HU, D. Comparison of svm and ls-svm for regression. In: IEEE. **2005 International conference on neural networks and brain**. [S.l.], 2005. v. 1, p. 279–283.

WEN, J.; LI, S.; LIN, Z.; HU, Y.; HUANG, C. Systematic literature review of machine learning based software development effort estimation models. **Information and Software Technology**, Elsevier, v. 54, n. 1, p. 41–59, 2012.

XAVIER, M.; MENESES, J. A literature review on the definitions of dropout in online higher education. In: **EDEN Conference Proceedings**. [S.l.: s.n.], 2020. p. 73–80.

ZUILKOWSKI, S. S.; SAMANHUDI, U.; INDRIANA, I. ‘there is no free education nowadays’: youth explanations for school dropout in indonesia. **Compare: A Journal of Comparative and International Education**, Routledge, v. 49, n. 1, p. 16–29, 2019. Disponível em: <<https://doi.org/10.1080/03057925.2017.1369002>>.

A APENDICE - FICHA DE EXTRAÇÃO DE DADOS

A seguir apresenta-se o documento denominado de ficha de extração que foi elaborado pelos autores da presente pesquisa para extrair as informações dos artigos selecionados no MSL.

Informações Gerais			
ID:			
Título:			
Idioma: Inglês			
Fonte: DOI:	Número de páginas:	Data da publicação:	Qualis / F.I:
Resumo do Trabalho:			
Critérios de Inclusão:			
Respostas às questões de Pesquisa: Q1: Q2: Q3: Q4			
Observações:			

B APENDICE - TUTORIAL DETALHADO DA CONSTRUÇÃO DA ANÁLISE EXPLORATÓRIA DE DADOS

A seguir apresenta-se o tutorial detalhado da análise exploratória de dados realizada no primeiro conjunto de dados dessa pesquisa (dados dos egressos curso de Ciência e Tecnologia da UFRESA). O tutorial foi desenvolvido pela autora da pesquisa na ferramenta Colaboratory Notebook, contendo os códigos implementados. Esses códigos podem ser reproduzidos em qualquer outra ferramenta de Ciência de Dados, desde que seguido corretamente e aplicado a um conjunto de dados contendo os mesmos atributos da presente pesquisa. Caso utilizado em outros atributos, o resultado pode sofrer alterações. Ressalta-se que a etapa de análise exploratória de dados é de fundamental importância e deve ocorrer antes da etapa de implementação dos algoritmos de ML.

Tutorial da análise exploratória dos dados

Nessa fase o conjunto de dados precisa ter realizado todos os ajustes de entendimento, análises, pré-processamento e transformação de dados

As bibliotecas utilizadas serão Perceptron, KNN, Florestas aleatórias, Gaussiana, árvores de decisão, Perceptron multicamadas, Máquinas de vetores de suporte, busca em grade, pandas, numpy, matplotlib e seaborn para os gráficos

```
# Importando bibliotecas
from sklearn.model_selection import train_test_split
from sklearn.linear_model import Perceptron
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.naive_bayes import MultinomialNB
from sklearn.naive_bayes import BernoulliNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.neural_network import MLPClassifier
from sklearn.svm import SVC
from sklearn.model_selection import GridSearchCV
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

# Lendo o csv
df = pd.read_csv('selecionados.csv', sep=';')
df.head()

# Criando um dataframe somente com as features
X = df.drop('var_0',axis=1)

# Criando um dataframe somente com as classes
y = df['var_0']

# Definindo as métricas de avaliação
scoring = ['accuracy', 'precision','recall', 'f1', 'roc_auc']

# Variaveis para armazenar o melhor classificador de cada modelo
per_best_estimator = ""
per_best_score = 0
knn_best_estimator = ""
```

```

_ _
knn_best_score = 0
nb_best_estimator = ""
nb_best_score = 0
nb_distribution = ""
dt_best_estimator = ""
dt_best_score = 0
rf_best_estimator = ""
rf_best_score = 0
mlp_best_estimator = ""
mlp_best_score = 0
svc_best_estimator = ""
svc_best_score = 0
svc_kernel = ""

```

▼ PERCEPTRON

Nessa fase será instânciado o perceptron e definido os parâmetros para que a busca em grade encontre os melhores resultados

```

# Instanciando o classificador
per = Perceptron()

# Definindo alguns parametros para teste
params = {
    'penalty': ['l2','l1','elasticnet']
}

# Instanciando o gridsearch e executando o modelo

for score_item in scoring:
    print('Métrica de avaliacao: ', score_item)
    grid_search = GridSearchCV(estimator=per,
                               param_grid=params,
                               cv = 4,
                               n_jobs=-1, verbose=1, scoring=score_item)

    #%%time
    grid_search.fit(X, y)
    best = grid_search.best_estimator_
    print('Melhor classificador: ',best)
    print('Score: ',grid_search.best_score_)
    print()
    if (per_best_score < grid_search.best_score_):
        per_best_estimator = best
        per_best_score = grid_search.best_score_

```

▼ KNN

```

# Instanciando o classificador
knn = KNeighborsClassifier(n_jobs=-1)

# Definindo alguns parametros para teste
params = {
    'n_neighbors': [2,3,5,10,20],
    'metric': ['euclidean', 'manhattan', 'chebyshev']
}

# Instanciando o gridsearch e executando o modelo

for score_item in scoring:
    print('Métrica de avaliacao: ', score_item)
    grid_search = GridSearchCV(estimator=knn,
                               param_grid=params,
                               cv = 4,
                               n_jobs=-1, verbose=1, scoring=score_item)

    #%%time
    grid_search.fit(X, y)
    best = grid_search.best_estimator_
    print('Melhor classificador: ',best)
    print('Score: ',grid_search.best_score_)
    print()
    if (knn_best_score < grid_search.best_score_):
        knn_best_estimator = best
        knn_best_score = grid_search.best_score_

```

▼ Naive Bayes

```

# Chamada do classificador
nb = GaussianNB()

# Definindo parâmetros
params = {'var_smoothing': np.logspace(0,-9, num=10)}
for score_item in scoring:
    print('Métrica de avaliacao: ', score_item)
    grid_search = GridSearchCV(estimator=nb,
                               param_grid=params,
                               cv = 4,
                               n_jobs=-1, verbose=1, scoring=score_item)

    #%%time
    grid_search.fit(X, y)
    best = grid_search.best_estimator_
    print('Melhor classificador: ',best)
    print('Score: ',grid_search.best_score_)
    print()
    if (nb_best_score < grid_search.best_score_):
        nb_best_estimator = best

```

```

        nb_best_score = grid_search.best_score_
        nb_distribution = 'Gaussian'

# Chamada do classificador

nb = MultinomialNB()

# Definindo parâmetros
params = {
    'alpha': [0.1,0.2,0.3,0.5,0.7,1.0]
}
for score_item in scoring:
    print('Métrica de avaliacao: ', score_item)
    grid_search = GridSearchCV(estimator=nb,
                               param_grid=params,
                               cv = 4,
                               n_jobs=-1, verbose=1, scoring=score_item)

    #%%time
    grid_search.fit(X, y)
    best = grid_search.best_estimator_
    print('Melhor classificador: ',best)
    print('Score: ',grid_search.best_score_)
    print()
    if (nb_best_score < grid_search.best_score_):
        nb_best_estimator = best
        nb_best_score = grid_search.best_score_
        nb_distribution = 'Multinomial'

# Chamada do classificador

nb = BernoulliNB()

# Definindo parâmetros
params = {
    'alpha': [0.1,0.2,0.3,0.5,0.7,1.0]
}
for score_item in scoring:
    print('Métrica de avaliacao: ', score_item)
    grid_search = GridSearchCV(estimator=nb,
                               param_grid=params,
                               cv = 4,
                               n_jobs=-1, verbose=1, scoring=score_item)

    #%%time
    grid_search.fit(X, y)
    best = grid_search.best_estimator_
    print('Melhor classificador: ',best)
    print('Score: ',grid_search.best_score_)
    print()
    if (nb_best_score < grid_search.best_score_):
        nb_best_estimator = best
        nb_best_score = grid_search.best_score_
        nb_distribution = 'Bernoulli'

```

▼ Decision Tree

```
# Chamada do classificador
dt = DecisionTreeClassifier(random_state=42)

# Definindo parâmetros
params = {
    'criterion': ['gini','entropy'],
    'max_depth': [2,3,5,10,20],
    'min_samples_leaf': [5,10,20,50,100,200]
}
for score_item in scoring:
    print('Métrica de avaliacao: ', score_item)
    grid_search = GridSearchCV(estimator=dt,
                               param_grid=params,
                               cv = 4,
                               n_jobs=-1, verbose=1, scoring=score_item)

    #%%time
    grid_search.fit(X, y)
    best = grid_search.best_estimator_
    print('Melhor classificador: ',best)
    print('Score: ',grid_search.best_score_)
    print()
    if (dt_best_score < grid_search.best_score_):
        dt_best_estimator = best
        dt_best_score = grid_search.best_score_
```

▼ Random Forest

```
# Chamada do classificador
rf = RandomForestClassifier(random_state=42)

# Definindo parâmetros
params = {
    'criterion': ['gini','entropy'],
    'max_depth': [2,3,5,10,20],
    'min_samples_leaf': [5,10,20,50,100,200],
    'n_estimators': [10,25,30,50,100,200]
}
for score_item in scoring:
    print('Métrica de avaliacao: ', score_item)
    grid_search = GridSearchCV(estimator=rf,
                               param_grid=params,
                               cv = 4,
                               n_jobs=-1, verbose=1, scoring=score_item)

    #%%time
    grid_search.fit(X, y)
    best = grid_search.best_estimator_
    print('Melhor classificador: ',best)
    print('Score: ',grid_search.best_score_)
    print()
    if (rf_best_score < grid_search.best_score_):
```

```
rf_best_estimator = best
rf_best_score = grid_search.best_score_
```

▼ MLP

```
# Chamada do classificador
mlp = MLPClassifier(max_iter=100)

# Definindo parâmetros
params = {
    'hidden_layer_sizes': [(10,30,10),(20,)],
    'activation': ['tanh', 'relu'],
    'solver': ['sgd', 'adam'],
    'alpha': [0.0001, 0.05],
    'learning_rate': ['constant','adaptive'],
}

for score_item in scoring:
    print('Métrica de avaliacao: ', score_item)
    grid_search = GridSearchCV(estimator=mlp,
                               param_grid=params,
                               cv = 4,
                               n_jobs=-1, verbose=1, scoring=score_item)

    #%%time
    grid_search.fit(X, y)
    best = grid_search.best_estimator_
    print('Melhor classificador: ',best)
    print('Score: ',grid_search.best_score_)
    print()
    if (mlp_best_score < grid_search.best_score_):
        mlp_best_estimator = best
        mlp_best_score = grid_search.best_score_
```

▼ SVM

```
#kernel rbf
svc_kernel = 'rbf'
svc = SVC(kernel='rbf')

# Definindo parâmetros
params = {'C': [0.1, 1, 10, 100],
          'gamma': [1,0.1, 0.01, 0.001]}

for score_item in scoring:
    print('Métrica de avaliacao: ', score_item)
    grid_search = GridSearchCV(estimator=svc,
                               param_grid=params,
                               cv = 4,
```

```

n_jobs=-1, verbose=1, scoring=score_item)

#%%time
grid_search.fit(X, y)
best = grid_search.best_estimator_
print('Melhor classificador: ',best)
print('Score: ',grid_search.best_score_)
print()
if (svc_best_score < grid_search.best_score_):
    svc_best_estimator = best
    svc_best_score = grid_search.best_score_

#kernel linear
svc_kernel = 'linear'
svc = SVC(kernel='linear')

# Definindo parâmetros
params = {'C': [0.1,1,10]}

for score_item in scoring:
    print('Métrica de avaliacao: ', score_item)
    grid_search = GridSearchCV(estimator=svc,
                               param_grid=params,
                               cv = 4,
                               n_jobs=-1, verbose=1, scoring=score_item)

    #%%time
    grid_search.fit(X, y)
    best = grid_search.best_estimator_
    print('Melhor classificador: ',best)
    print('Score: ',grid_search.best_score_)
    print()
    if (svc_best_score < grid_search.best_score_):
        svc_best_estimator = best
        svc_best_score = grid_search.best_score_

#kernel polinomial
svc_kernel = 'polinomial'
svc = SVC(kernel='poly')

# Definindo parâmetros
params = {'C': [0.1, 1, 10],
          'gamma': [1,0.1,0.01],
          'degree': [3,4]}

for score_item in scoring:
    print('Métrica de avaliacao: ', score_item)
    grid_search = GridSearchCV(estimator=svc,
                               param_grid=params,
                               cv = 4,
                               n_jobs=-1, verbose=1, scoring=score_item)

    #%%time

```

```
grid_search.fit(X, y)
best = grid_search.best_estimator_
print('Melhor classificador: ',best)
print('Score: ',grid_search.best_score_)
print()
if (svc_best_score < grid_search.best_score_):
    svc_best_estimator = best
    svc_best_score = grid_search.best_score_
```

C APENDICE - TUTORIAL DETALHADO DA IMPLEMENTAÇÃO DOS MODELOS DE ML

A seguir apresenta-se o tutorial detalhado da implementação dos algoritmos de ML e métricas de avaliação para o primeiro conjunto de dados (dados dos egressos curso de Ciência e Tecnologia da UFRESA). O tutorial foi desenvolvido na ferramenta Colaboratory Notebook, contendo os códigos implementados. Esses códigos podem ser reproduzidos em qualquer outra ferramenta de Ciência de Dados, desde que seguido corretamente e aplicado a um conjunto de dados contendo os mesmos atributos da presente pesquisa. Caso utilizado em outros atributos, o resultado pode sofrer alterações. Ressalta-se que a etapa de análise exploratória de dados é de fundamental importância e deve ocorrer antes da etapa de implementação dos algoritmos de ML.

Tutorial de desenvolvimento dos modelos de ML

Nessa fase o conjunto de dados precisa ter realizado todos os ajustes de entendimento, análises, pré-processamento e transformação de dados

As bibliotecas utilizadas serão Perceptron, KNN, Florestas aleatórias, Gaussiana, árvores de decisão, Perceptron multicamadas, Máquinas de vetores de suporte, busca em grade, pandas, numpy, matplotlib e seaborn para os gráficos

```
# Importando bibliotecas
from sklearn.model_selection import train_test_split
from sklearn.linear_model import Perceptron
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.naive_bayes import MultinomialNB
from sklearn.naive_bayes import BernoulliNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.neural_network import MLPClassifier
from sklearn.svm import SVC
from sklearn.model_selection import GridSearchCV
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

# Lendo o csv
df = pd.read_csv('selecionados.csv', sep=';')
df.head()

# Criando um dataframe somente com as features
X = df.drop('var_0',axis=1)

# Criando um dataframe somente com as classes
y = df['var_0']

# Definindo as métricas de avaliação
scoring = ['accuracy', 'precision', 'recall', 'f1', 'roc_auc']

# Variaveis para armazenar o melhor classificador de cada modelo
per_best_estimator = ""
per_best_score = 0
knn_best_estimator = ""
```

```

_ _
knn_best_score = 0
nb_best_estimator = ""
nb_best_score = 0
nb_distribution = ""
dt_best_estimator = ""
dt_best_score = 0
rf_best_estimator = ""
rf_best_score = 0
mlp_best_estimator = ""
mlp_best_score = 0
svc_best_estimator = ""
svc_best_score = 0
svc_kernel = ""

```

▼ PERCEPTRON

Nessa fase será instânciado o perceptron e definido os parâmetros para que a busca em grade encontre os melhores resultados

```

# Instanciando o classificador
per = Perceptron()

# Definindo alguns parametros para teste
params = {
    'penalty': ['l2','l1','elasticnet']
}

# Instanciando o gridsearch e executando o modelo

for score_item in scoring:
    print('Métrica de avaliacao: ', score_item)
    grid_search = GridSearchCV(estimator=per,
                               param_grid=params,
                               cv = 4,
                               n_jobs=-1, verbose=1, scoring=score_item)

    #%%time
    grid_search.fit(X, y)
    best = grid_search.best_estimator_
    print('Melhor classificador: ',best)
    print('Score: ',grid_search.best_score_)
    print()
    if (per_best_score < grid_search.best_score_):
        per_best_estimator = best
        per_best_score = grid_search.best_score_

```

▼ KNN

```

# Instanciando o classificador
knn = KNeighborsClassifier(n_jobs=-1)

# Definindo alguns parametros para teste
params = {
    'n_neighbors': [2,3,5,10,20],
    'metric': ['euclidean', 'manhattan', 'chebyshev']
}

# Instanciando o gridsearch e executando o modelo

for score_item in scoring:
    print('Métrica de avaliacao: ', score_item)
    grid_search = GridSearchCV(estimator=knn,
                               param_grid=params,
                               cv = 4,
                               n_jobs=-1, verbose=1, scoring=score_item)

    #%%time
    grid_search.fit(X, y)
    best = grid_search.best_estimator_
    print('Melhor classificador: ',best)
    print('Score: ',grid_search.best_score_)
    print()
    if (knn_best_score < grid_search.best_score_):
        knn_best_estimator = best
        knn_best_score = grid_search.best_score_

```

▼ Naive Bayes

```

# Chamada do classificador
nb = GaussianNB()

# Definindo parâmetros
params = {'var_smoothing': np.logspace(0,-9, num=10)}
for score_item in scoring:
    print('Métrica de avaliacao: ', score_item)
    grid_search = GridSearchCV(estimator=nb,
                               param_grid=params,
                               cv = 4,
                               n_jobs=-1, verbose=1, scoring=score_item)

    #%%time
    grid_search.fit(X, y)
    best = grid_search.best_estimator_
    print('Melhor classificador: ',best)
    print('Score: ',grid_search.best_score_)
    print()
    if (nb_best_score < grid_search.best_score_):
        nb_best_estimator = best

```

```

        nb_best_score = grid_search.best_score_
        nb_distribution = 'Gaussian'

# Chamada do classificador

nb = MultinomialNB()

# Definindo parâmetros
params = {
    'alpha': [0.1,0.2,0.3,0.5,0.7,1.0]
}
for score_item in scoring:
    print('Métrica de avaliacao: ', score_item)
    grid_search = GridSearchCV(estimator=nb,
                               param_grid=params,
                               cv = 4,
                               n_jobs=-1, verbose=1, scoring=score_item)

    #%%time
    grid_search.fit(X, y)
    best = grid_search.best_estimator_
    print('Melhor classificador: ',best)
    print('Score: ',grid_search.best_score_)
    print()
    if (nb_best_score < grid_search.best_score_):
        nb_best_estimator = best
        nb_best_score = grid_search.best_score_
        nb_distribution = 'Multinomial'

# Chamada do classificador

nb = BernoulliNB()

# Definindo parâmetros
params = {
    'alpha': [0.1,0.2,0.3,0.5,0.7,1.0]
}
for score_item in scoring:
    print('Métrica de avaliacao: ', score_item)
    grid_search = GridSearchCV(estimator=nb,
                               param_grid=params,
                               cv = 4,
                               n_jobs=-1, verbose=1, scoring=score_item)

    #%%time
    grid_search.fit(X, y)
    best = grid_search.best_estimator_
    print('Melhor classificador: ',best)
    print('Score: ',grid_search.best_score_)
    print()
    if (nb_best_score < grid_search.best_score_):
        nb_best_estimator = best
        nb_best_score = grid_search.best_score_
        nb_distribution = 'Bernoulli'

```

▼ Decision Tree

```
# Chamada do classificador
dt = DecisionTreeClassifier(random_state=42)

# Definindo parâmetros
params = {
    'criterion': ['gini','entropy'],
    'max_depth': [2,3,5,10,20],
    'min_samples_leaf': [5,10,20,50,100,200]
}

for score_item in scoring:
    print('Métrica de avaliacao: ', score_item)
    grid_search = GridSearchCV(estimator=dt,
                               param_grid=params,
                               cv = 4,
                               n_jobs=-1, verbose=1, scoring=score_item)

    #%%time
    grid_search.fit(X, y)
    best = grid_search.best_estimator_
    print('Melhor classificador: ',best)
    print('Score: ',grid_search.best_score_)
    print()
    if (dt_best_score < grid_search.best_score_):
        dt_best_estimator = best
        dt_best_score = grid_search.best_score_
```

▼ Random Forest

```
# Chamada do classificador
rf = RandomForestClassifier(random_state=42)

# Definindo parâmetros
params = {
    'criterion': ['gini','entropy'],
    'max_depth': [2,3,5,10,20],
    'min_samples_leaf': [5,10,20,50,100,200],
    'n_estimators': [10,25,30,50,100,200]
}

for score_item in scoring:
    print('Métrica de avaliacao: ', score_item)
    grid_search = GridSearchCV(estimator=rf,
                               param_grid=params,
                               cv = 4,
                               n_jobs=-1, verbose=1, scoring=score_item)

    #%%time
    grid_search.fit(X, y)
    best = grid_search.best_estimator_
    print('Melhor classificador: ',best)
    print('Score: ',grid_search.best_score_)
    print()
    if (rf_best_score < grid_search.best_score_):
```

```
rf_best_estimator = best
rf_best_score = grid_search.best_score_
```

▼ MLP

```
# Chamada do classificador
mlp = MLPClassifier(max_iter=100)

# Definindo parâmetros
params = {
    'hidden_layer_sizes': [(10,30,10),(20,)],
    'activation': ['tanh', 'relu'],
    'solver': ['sgd', 'adam'],
    'alpha': [0.0001, 0.05],
    'learning_rate': ['constant','adaptive'],
}

for score_item in scoring:
    print('Métrica de avaliacao: ', score_item)
    grid_search = GridSearchCV(estimator=mlp,
                               param_grid=params,
                               cv = 4,
                               n_jobs=-1, verbose=1, scoring=score_item)

    #%%time
    grid_search.fit(X, y)
    best = grid_search.best_estimator_
    print('Melhor classificador: ',best)
    print('Score: ',grid_search.best_score_)
    print()
    if (mlp_best_score < grid_search.best_score_):
        mlp_best_estimator = best
        mlp_best_score = grid_search.best_score_
```

▼ SVM

```
#kernel rbf
svc_kernel = 'rbf'
svc = SVC(kernel='rbf')

# Definindo parâmetros
params = {'C': [0.1, 1, 10, 100],
          'gamma': [1,0.1, 0.01, 0.001]}

for score_item in scoring:
    print('Métrica de avaliacao: ', score_item)
    grid_search = GridSearchCV(estimator=svc,
                               param_grid=params,
                               cv = 4,
```

```

n_jobs=-1, verbose=1, scoring=score_item)

#%%time
grid_search.fit(X, y)
best = grid_search.best_estimator_
print('Melhor classificador: ',best)
print('Score: ',grid_search.best_score_)
print()
if (svc_best_score < grid_search.best_score_):
    svc_best_estimator = best
    svc_best_score = grid_search.best_score_

#kernel linear
svc_kernel = 'linear'
svc = SVC(kernel='linear')

# Definindo parâmetros
params = {'C': [0.1,1,10]}

for score_item in scoring:
    print('Métrica de avaliacao: ', score_item)
    grid_search = GridSearchCV(estimator=svc,
                               param_grid=params,
                               cv = 4,
                               n_jobs=-1, verbose=1, scoring=score_item)

    #%%time
    grid_search.fit(X, y)
    best = grid_search.best_estimator_
    print('Melhor classificador: ',best)
    print('Score: ',grid_search.best_score_)
    print()
    if (svc_best_score < grid_search.best_score_):
        svc_best_estimator = best
        svc_best_score = grid_search.best_score_

#kernel polinomial
svc_kernel = 'polinomial'
svc = SVC(kernel='poly')

# Definindo parâmetros
params = {'C': [0.1, 1, 10],
          'gamma': [1,0.1,0.01],
          'degree': [3,4]}

for score_item in scoring:
    print('Métrica de avaliacao: ', score_item)
    grid_search = GridSearchCV(estimator=svc,
                               param_grid=params,
                               cv = 4,
                               n_jobs=-1, verbose=1, scoring=score_item)

    #%%time

```

```
grid_search.fit(X, y)
best = grid_search.best_estimator_
print('Melhor classificador: ',best)
print('Score: ',grid_search.best_score_)
print()
if (svc_best_score < grid_search.best_score_):
    svc_best_estimator = best
    svc_best_score = grid_search.best_score_
```