UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO MULTIDISCIPLINAR
PROGRAMA DE PÓS-GRADUAÇÃO INTERDISCIPLINAR EM HUMANIDADES DIGITAIS

DISSERTAÇÃO

ESTUDO DA METODOLOGIA DE *DATA ANALYTICS* APLICADA EM PESQUISAS SOBRE O FENÔMENO DA EVASÃO NO ENSINO SUPERIOR UTILIZANDO A ESTRUTURA DA *DESIGN SCIENCE RESEARCH*

LÁYLA ADVINCULA CANDIDO DE AZEVEDO

2022

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO MULTIDISCIPLINAR
PROGRAMA DE PÓS-GRADUAÇÃO INTERDISCIPLINAR EM HUMANIDADES DIGITAIS

LÁYLA ADVINCULA CANDIDO DE AZEVEDO

Sob a orientação da Professora
Adria Ramos de Lyra

E coorientação do Professor
Carlos Eduardo Ribeiro de Mello

Dissertação submetida como requisito parcial para obtenção do grau de **Mestre em Humanidades Digitais** no Programa de Pós-Graduação Interdisciplinar em Humanidades Digitais, Área de Concentração em Mineração de Dados Digitais.

Nova Iguaçu/RJ
Abril de 2022

Universidade Federal Rural do Rio de Janeiro Biblioteca Central / Seção de Processamento Técnico

Ficha catalográfica elaborada com os
dados fornecidos pelo(a) autor(a)

A994e

Azevedo, Láyla Advincula Candido de , 1979-
Estudo da metodologia de Data Analytics aplicada em pesquisas sobre o fenômeno da evasão no ensino superior utilizando a estrutura da Design Science Research / Láyla Advincula Candido de Azevedo. - Paracambi, 2022.
97 f.: il.

Orientadora: Adria Ramos de Lyra.
Coorientador: Carlos Eduardo Ribeiro de Mello.
Dissertação(Mestrado). -- Universidade Federal Rural do Rio de Janeiro, Programa de Pós-Graduação Interdisciplinar em Humanidades Digitais, 2022.

1. Data Analytics. 2. Design Science Research. 3. Evasão. 4. Ensino Superior. 5. Humanidades Digitais. I. Lyra, Adria Ramos de, 1979-, orient. II. Mello, Carlos Eduardo Ribeiro de, 1983-, coorient. III Universidade Federal Rural do Rio de Janeiro. Programa de Pós-Graduação Interdisciplinar em Humanidades Digitais. IV. Título.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO/IM

TERMO Nº 886 / 2022 - DeptCC/IM (12.28.01.00.00.83)

Nº do Protocolo: 23083.048335/2022-55

Seropédica-RJ, 08 de agosto de 2022.

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO

INSTITUTO MULTIDISCIPLINAR

PROGRAMA DE PÓS-GRADUAÇÃO INTERDISCIPLINAR EM HUMANIDADES DIGITAIS

LÁYLA ADVINCULA CANDIDO DE AZEVEDO

Dissertação / Tese submetida como requisito parcial para a obtenção do grau de **Mestre em Humanidades Digitais**, no Programa de Pós Graduação Interdisciplinar em Humanidades Digitais, Área de Concentração em Mineração de Dados Digitais.

DISSERTAÇÃO / TESE APROVADA EM 28/04/2022.

Conforme deliberação número 001/2020 da PROPPG, de 30/06/2020, tendo em vista a implementação de trabalho remoto e durante a vigência do período de suspensão das atividades acadêmicas presenciais, em virtude das medidas adotadas para reduzir a propagação da pandemia de Covid-19, nas versões finais das teses e dissertações as assinaturas originais dos membros da banca examinadora poderão ser substituídas por documento(s) com assinaturas eletrônicas. Estas devem ser feitas na própria folha de assinaturas, através do SIPAC.

Identificar membros da banca:

ADRIA RAMOS DE LYRA

Dr, UFRRJ

(Orientador / Presidente da Banca)

LEANDRO GUIMARAES MARQUES ALVIM

Dr., UFRRJ

LAURA DE OLIVEIRA FERNANDES MORAES

Dr., CPII

(Assinado digitalmente em 08/08/2022 16:02)
ADRIA RAMOS DE LYRA
PROFESSOR DO MAGISTERIO SUPERIOR
DeptCC/IM (12.28.01.00.00.83)
Matrícula: 1604994

(Assinado digitalmente em 08/08/2022 16:15)
LEANDRO GUIMARAES MARQUES ALVIM
PROFESSOR DO MAGISTERIO SUPERIOR
DeptCC/IM (12.28.01.00.00.83)
Matrícula: 1800852

(Assinado digitalmente em 11/08/2022 14:33)
LAURA DE OLIVEIRA FERNANDES MORAES
ASSINANTE EXTERNO
CPF: 124.359.357-14

Para verificar a autenticidade deste documento entre em **https://sipac.ufrrj.br/public/documentos/index.jsp** informando seu número: **886**, ano: **2022**, tipo: **TERMO**, data de emissão: **08/08/2022** e o código de verificação: **36b54846c7**

DEDICATÓRIA

A Deus, sem O qual nada seria possível, à memória de Welayne, minha mãe, aquela que sempre acreditou que eu poderia conquistar qualquer coisa que eu quisesse. Para Carlos, meu pai por todo o apoio e carinho. Ao meu marido Wendel, companheiro das minhas empreitadas. Ao meu amado filho Bernardo para que ele tenha inspiração para ir mais longe. E a todos aqueles que se dedicam a conhecer um pouco mais e a dividir esse conhecimento com o mundo.

AGRADECIMENTOS

Primeiramente a Deus, por iluminar e abençoar minhas escolhas e por ter me dado asas quando me faltou o chão.

Ao meu marido, Wendel por compreender meus momentos de dedicação a esta pesquisa, pelo apoio e por todo amor que me dedica. Agradeço ao meu filho Bernardo, por ser minha inspiração e motivação.

Ao meus pais. À minha mãe Welayne (em memória), ela estava aqui no começo, me auxiliando e apoiando certa de que eu iria conseguir. E ao meu pai por todo o apoio, especialmente logístico durante as aulas presenciais. A verdade é que eles foram o começo de tudo. Agradeço por todos os livros e enciclopédias e por todo incentivo para que eu estudasse e buscasse o conhecimento.

A minha Orientadora Adria Lyra e a Professora Márcia Denise Pletsch pelo projeto de pesquisa escrito inicialmente, por terem me aceito para desenvolvê-lo. Agradeço também a professora Márcia por me apresentar um universo desconhecido (amei cursar a disciplina de Inclusão na Perspectiva dos Direitos Humanos) e por compreender que as pesquisas acabam nos levando por rumos inesperados.

Ao professor Carlos Eduardo que juntamente com a Adria redirecionaram a pesquisa para que em meio a pandemia de Covid-19 ela pudesse ser executada.

Agradeço ao Instituto Multidisciplinar de Nova Iguaçu/UFRRJ pela oportunidade e por todo apoio para pesquisa especialmente durante o Ensino Remoto. Obrigada pelo apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES).

Aos amigos que fiz durante o curso, não teria conseguido sem vocês, obrigada por me permitir compartilhar minha ansiedade, frustrações, incertezas e alegrias.

Aos meus familiares e amigos, por ouvirem meus desabafos.

RESUMO

AZEVEDO, Láyla Advincula Candido. **Estudo da metodologia de *Data Analytics* aplicada em pesquisas sobre o fenômeno da evasão no ensino superior utilizando a estrutura da *Design Science Research*,** 2022. 97p Dissertação (Mestrado em Humanidades Digitais). Instituto Multidisciplinar, Universidade Federal Rural do Rio de Janeiro, Nova Iguaçu, RJ, 2022.

A evasão escolar é um fenômeno complexo que afeta o desempenho socioeconômico de um país e durante muitas décadas tem sido objeto de estudo de pesquisadores de diversas áreas por todo o mundo. Tendo um caráter interdisciplinar observa-se que estudos sobre o fenômeno da evasão têm se valido de modelos analíticos quantitativos, recorrendo em especial, ao uso de metodologias de análise de dados. Sendo assim, essa dissertação, inserida no campo das Humanidades Digitais, tem por objetivo pesquisar abordagens de *Data Analytics* para o apoio às Políticas de Ensino Superior em cursos de graduação (bacharelados e licenciaturas) na modalidade de ensino presencial, voltadas, especificamente, para o controle e combate à evasão escolar. A pesquisa conduzida abordou duas frentes: (a) revisão sistemática da literatura – onde, como o próprio nome diz, de forma sistemática, utiliza-se critérios de busca para coletar, identificar e selecionar trabalhos científicos relevantes da literatura pertinentes ao tema; e (b) criação de uma metodologia baseada na *Design Science Research* para desenvolver a análise dos trabalhos da literatura. A metodologia proposta é composta por quatro componentes: Enquadramento, Teoria, Modelagem e Protocolo Experimental. O protocolo elaborado para orientar a Revisão Sistemática foi satisfatório retornando 42 artigos para análise. Na análise do Enquadramento, verificou-se que a tarefa de *Data Analytics* mais utilizada é a preditiva, e dentre estas observou-se uma predominância na utilização de técnicas individuais em detrimento dos métodos *ensembles*, sendo a Árvore de Decisão uma das mais utilizadas. Menos de 50% dos estudos definem o termo evasão e 70% deles tratam esse fenômeno como uma tarefa de classificação. Com relação à Teorização, as informações acadêmicas são as mais consideradas para construção dos modelos. Boa parte dos trabalhos parte da teoria de que o desempenho acadêmico é um preditor importante para a evasão. Quanto à Modelagem, avaliou-se que grande parte dos estudos utilizam apenas um conjunto de dados, cuja origem pode ser das informações do sistema acadêmico (fontes internas); pesquisas institucionais, que compõem bases de dados nacionais, regionais e acadêmicas ou de questionários utilizados pelos próprios pesquisadores para adquirir informações mais específicas (fontes externas). Além disso, utilizou-se uma combinação de informações (background demográfico, desempenho/informação escolar anterior e Informações/desempenho acadêmico) para a construção dos modelos. Na análise do Protocolo Experimental, observou-se que o método de ajuste de modelo mais utilizado foi a validação cruzada (*cross-validation)* e a métrica de interesse mais utilizada fora a Acurácia, presente em 26 estudos. Esses resultados e análises levaram a construção de um mapa mental, organizando as principais propostas da literatura. A metodologia proposta com base na DSR foi fundamental para a análise dos trabalhos, possibilitando a identificação das abordagens de *Data Analytics* presentes nos trabalhos investigados de forma ortogonal aos componentes, contribuindo para que futuras pesquisas se beneficiem desta metodologia, especialmente no que diz respeito à criação de artefatos computacionais. O estudo também evidenciou que é possível utilizar a abordagem de *Data Analytics* para lidar com a evasão de alunos, eventualmente contribuindo para mitigar os efeitos deste fenômeno no Ensino Superior.

Palavras-chaves: Humanidades Digitais, *Data Analytics, Design Science Research*, Evasão, Ensino Superior.

ABSTRACT

AZEVEDO, Láyla Advincula Candido. **Study of the Data Analytics methodology applied in research on the phenomenon of dropout in higher education using the structure Design Science Research**, 2022. 97p Dissertation (Masters in Digital Humanities). Multidisciplinary Institute, Federal Rural University of Rio de Janeiro, Nova Iguaçu, RJ, 2022.

School dropout is a complex phenomenon that affects the socioeconomic performance of a country and for many decades has been the object of study by researchers from different areas around the world. Having an interdisciplinary character, it is observed that studies on the phenomenon of dropout have made use of quantitative analytical models, resorting in particular to the use of data analysis methodologies. Therefore, this dissertation, inserted in the field of Digital Humanities, aims to research Data Analytics approaches to support Higher Education Policies in undergraduate courses (bachelors and licentiates) in the face-to-face teaching modality, specifically aimed at the control and fight against school dropout. The research conducted addressed two fronts: (a) systematic literature review – where, as the name implies, search criteria are used systematically to collect, identify and select relevant scientific works from the literature relevant to the topic; and (b) creation of a methodology based on Design Science Research to develop the analysis of works in the literature. The proposed methodology is composed of four components: Framework, Theory, Modeling and Experimental Protocol. The protocol developed to guide the Systematic Review was satisfactory, returning 42 articles for analysis. In the Framing analysis, it was found that the most used Data Analytics task is the predictive one, and among these, there was a predominance in the use of individual techniques to the detriment of the ensembles methods, with the Decision Tree being one of the most used. Less than 50% of studies define the term dropout and 70% of them treat this phenomenon as a classification task. Regarding theorization, academic information is the most considered for the construction of models. Much of the work starts from the theory that academic performance is an important predictor of dropout. As for Modeling, it was evaluated that most studies use only one set of data, whose origin can be from information from the academic system (internal sources); institutional surveys, which comprise national, regional and academic databases or questionnaires used by the researchers themselves to acquire more specific information (external sources). In addition, a combination of information (demographic background, previous school performance/information and academic information/performance) was used to build the models. In the analysis of the Experimental Protocol, it was observed that the most used model adjustment method was cross-validation and the most used metric of interest was Accuracy, present in 26 studies. These results and analyzes led to the construction of a mental map, organizing the main proposals in the literature. The methodology proposed based on the DSR was fundamental for the analysis of the works, allowing the identification of Data Analytics approaches present in the investigated works in an orthogonal way to the components, contributing to future research to benefit from this methodology, especially regarding the creation of computational artifacts. The study also showed that it is possible to use the Data Analytics approach to deal with student dropout, eventually helping to mitigate the effects of this phenomenon in Higher Education.

Keywords: Digital Humanities, Data Analytics, Design Science Research, Droupout, Higher Education.

LISTA DE ILUSTRAÇÕES

FIGURAS

GRÁFICOS

LISTA DE QUADROS E TABELAS

QUADROS

TABELAS

LISTA DE SIGLAS E ABREVIATURAS

AVA	Ambiente Virtual de Aprendizagem
AUC	*Area under the ROC curve*
BDA	*Big Data Analytics*
CEP/cep	Código de Endereçamento Postal
CLV	*Costumer Lifetime Value*
CR	Coeficiente de Rendimento
CRISP	*Cross Industry Standard Proccess*
ROC	*Receiver Operating Characteristic*
DA	*Data Analytics*
DSR	*Design Science Research*
e.g.	*exempli gratia*
EDM	*Educational Data Mining*
ENEM	Exame Nacional do Ensino Médio
FECAP	A Fundação Escola de Comércio Álvares Penteado
GED	Ensino Médio externo/Educado em casa (indicativo)
GPA/GPAX	*Grade Point Average*/Média de pontos das notas acima de 2,0 na admissão
IBGE	Instituto Brasileiro de Geografia e Estatística
IES	Instituição Ensino Superior
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
IELTS	*International English Language Testing System*
KDD	*Knowledg Discovery Databases*
KNN	*K-Nearest-Nearest*
KPCA	*Kernel Principal Component Analysis*
LA	*Learning Analytics*
LOOCV	*Leave-on-out cross-validation*
MAE	Erro Médio Absoluto
MDA	Média da diminuição da Acurácia
MDG	Média da Diminuição da Gini
MEC	Ministério da Educação
MOOC	*Massive Open Online Courses*
MP	*Maximum Profit*
N.º/n.º	Número
NEM	Nota do Ensino Médio
OLAP	*Online Analytical Processing*
PCA	*Principal Component Analysis*
PSU	*Prueba de Selección Universitaria*
RG	Ganho Relativo
RMD	Regime de Matrícula por Disciplina
RMS	Regime de Matrícula Seriado
RMSE	Erro Médio Quadrático da Raiz
SAT	*Scholastic Assessement Test*
SGA	Sistema de Gestão Acadêmica
SIGAA	Sistema Integrado Gestão Acadêmica
STEM	*Science, Technology, Engineering, and Mathematic*
UTA	*Weighted Average University (Admission Test*)

SUMÁRIO

1 INTRODUÇÃO

A evasão de estudantes no Ensino Superior é um fenômeno extremamente complexo, de modo que o desenvolvimento de estratégias de prevenção e mitigação pode exigir o estudo de diferentes áreas do conhecimento, tais como: Ciências Sociais Aplicadas, Ciências Humanas e Ciências Exatas (ABALCO et al., 2019). Nesta dissertação, introduzimos mais uma perspectiva metodológica na investigação desse fenômeno – *a Data Analytics*. Em virtude deste caráter interdisciplinar observamos que os estudos sobre o fenômeno da evasão têm se valido de modelos analíticos quantitativos, recorrendo em especial, ao uso de metodologias de análise de dados (em inglês, *Data Analytics*) (PROVOST e FAWCETT, 2013), onde o objetivo consiste tanto em construir conhecimento a partir de dados e evidências quantitativas quanto propor soluções analíticas que efetivamente contribuam para a diminuição da evasão.

1.1 Contextualização e motivação

Desde que se estabeleceu a educação formal, as taxas de retenção e evasão têm sido fonte de preocupação por parte das Instituições de Ensino Superior (IES) (ALJOHANI, 2016). De acordo com (ROVIRA, PUERTAS, IGUAL, 2017), por exemplo, a Espanha registrou estimativas de taxas de evasão entre 25% e 30%. No Brasil, a taxa de desistência acumulada alcançou valores de cerca de 59%, segundo dados do Ministério da Educação em 2020 (MEC/INEP, 2020). Diante de dados como estes, o fenômeno da evasão tornou-se objeto de inúmeras pesquisas conduzidas ao longo dos anos. Nestas, além da compreensão do fenômeno, boa parte tem como objetivo criar indicadores úteis para identificar relações de causa e efeito, e marcadores de risco relevantes.

Independentemente dos objetivos que motivaram tais pesquisas, acredita-se que os resultados obtidos podem auxiliar instituições, governos e legisladores a elaborar estratégias e conduzir políticas de Ensino Superior voltadas ao combate à evasão. Além disso, esses estudos permitem o desenvolvimento e análise de teorias que apoiam a elaboração de novos modelos de *Data Analytics* (DA) que têm aplicação direta na identificação e mitigação dos riscos de evasão.

1.2 Escopo

A proposta desta pesquisa é investigar o fenômeno da evasão especificamente no Ensino Superior universitário, em cursos de graduação (bacharelados e licenciaturas) na modalidade de ensino presencial. Diferentemente do ensino à distância, onde os ambientes virtuais de aprendizagem coletam inúmeras informações dos alunos ao longo do curso, a modalidade presencial é mais desafiadora, pois não dispõe de ferramentas de coleta de dados com tal nível de capilaridade. Para a maioria das universidades, a realidade consiste apenas na coleta dos registros acadêmicos dos alunos através de sistemas de administração/gerenciamento de matrículas. Neste sentido, monitorar o comportamento de forma granular na modalidade presencial é algo muito mais limitado do que em uma plataforma virtual de ensino à distância, representando um desafio para os pesquisadores.

Na modalidade de ensino presencial, a principal fonte de dados associada aos estudantes está nos sistemas de registro e controle de matrículas, inscrição em disciplinas, frequência e notas. Diversos dados encontrados nos Sistemas de Gestão Acadêmica (SGA) podem ser considerados importantes indicadores educacionais com potencial preditivo para a evasão. Ainda, a partir desses indicadores acadêmicos, como, por exemplo, desempenho acadêmico (e.g. Coeficiente de Rendimento - CR, em universidades públicas brasileiras), é possível analisar eventuais associações com fatores demográficos (e.g. gênero e raça), socioeconômicos (e.g. renda familiar), histórico de desempenho escolar (e.g. notas do Ensino Médio), notas de ingresso/admissão (e.g. desempenho no ENEM – Exame Nacional do Ensino Médio, no caso brasileiro).

1.3 Justificativa

Embora haja diversas propostas na literatura que investigam o problema da evasão sob a perspectiva de *Data Analytics*, não há, até onde se sabe, um estudo que organize sistematicamente essas pesquisas segundo suas principais teorias, estratégias de enquadramento, modelagem analítica e protocolos experimentais. Esta pesquisa de organização do conhecimento na área é fundamental para o avanço tanto do estado da arte quanto da prática.

1.4 Objetivos

1.4.1 Objetivos Gerais

Esta pesquisa, inserida no campo das Humanidades Digitais, tem por objetivo pesquisar metodologias de *Data Analytics* aplicadas ao apoio de Políticas de Ensino Superior, voltadas, especificamente, para o controle e combate à evasão escolar. Para tal, a pesquisa conduzida considerou duas frentes: (a) revisão sistemática da literatura – onde, como o próprio nome diz, de forma sistemática, utilizou-se critérios de busca para coletar, identificar e selecionar trabalhos científicos relevantes da literatura pertinentes ao tema; e (b) uma proposta metodológica de análise baseada no Método de Pesquisa em Ciência do Artificial (em inglês, *Design Science Research*) (DRESCH, LACERDA, ANTUNES, 2014).

1.4.2 Objetivos Específicos

- Propor uma metodologia para análise dos estudos baseada em *Design Science Research* adaptada para *Data Analytics*.
- Criar um protocolo de revisão sistemática para a área.
- Aplicar a metodologia proposta para organizar e categorizar as diferentes abordagens de *Data Analytics* para evasão no Ensino Superior.
- Estruturar os principais eixos de pesquisa a partir de um esquema semântico visual (mapa mental).
- Investigar estudos de caso disponíveis na literatura e seu enquadramento de acordo com a metodologia proposta.
- Sintetizar os resultados obtidos em diversas pesquisas de forma a criar um banco de dados de artigos científicos categorizados de acordo com os elementos metodológicos propostos.

1.5 Resultados Esperados

- Identificar e organizar os estudos acerca do fenômeno da evasão no ensino superior através da abordagem da *Data Analytics*.
- Contribuir com a ampliação e difusão da *Design Science Research* em pesquisas acadêmicas com foco em solucionar problemas aplicados ao mundo real.

- Agregar conhecimento ao campo das Humanidades Digitais a partir da organização das pesquisas e da criação de um mapa mental que sirva como metadado para armazenamento e recuperação deste conhecimento.
- Fornecer subsídios para a elaboração de Políticas Públicas que possam mitigar os problemas associados ao fenômeno da evasão de estudantes no ensino superior.
- Compilar informações que contribuam para a elaboração de ações institucionais em prol da prevenção da evasão e aumento dos índices de retenção e diplomação.

1.6 Organização Geral da Pesquisa

A Figura 1 demonstra os passos da construção desse estudo. O levantamento inicial da literatura permitiu identificar os termos para construção dos critérios de busca utilizados na Revisão Sistemática. Após a seleção dos trabalhos, foi feita a análise dos textos conforme o método de investigação baseado na *Design Science Research* que engloba os componentes da *Data Analytics*. O último passo foi a criação de um Mapa Mental para visualização dos resultados de forma dinâmica.

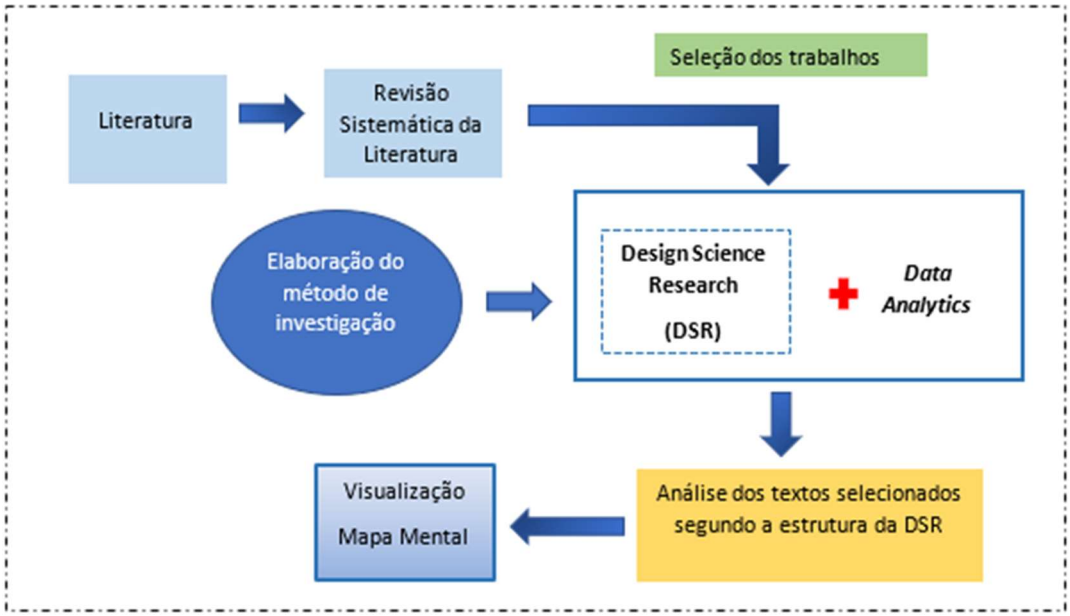


Figura 1. Organização Geral da Pesquisa

2 REFERENCIAL TEÓRICO

Este capítulo tem como objetivo apresentar os principais conceitos que fundamentam esta pesquisa. Primeiramente, apresenta-se uma visão geral sobre o fenômeno da evasão no Ensino Superior, suas principais teorias, características, conceitos e limitações, bem como, seus principais desafios, implicações socioeconômicas e demais questões relacionadas. Em seguida, apresenta-se os conceitos básicos sobre *Data Analytics* (DA), necessários para a compreensão de como se dá sua aplicação na modelagem do fenômeno da evasão, bem como no desenvolvimento de modelos aplicados. A terceira seção aborda a Revisão Sistemática, método utilizado para o desenvolvimento da revisão bibliográfica que introduz rigor metodológico à coleta, identificação e seleção de trabalhos a serem estudados. Por fim, apresenta-se uma visão geral da *Design Science Research* (DSR), paradigma epistemológico que será utilizado para análise dos trabalhos pesquisados, incluindo uma subseção que apresenta estudos que tratam da relação entre DSR e DA.

2.1 A Evasão no Ensino Superior

O estudo do fenômeno da evasão no ensino superior é a chave para entender como diminuir as taxas de desistências, bem como melhorar os índices de diplomação, revertendo-se em benefícios econômicos e sociais diretos. A evasão, do ponto de vista institucional ocasiona, segundo Delen (2010), "perdas financeiras, baixas taxas de diplomação, diminuição do prestígio diante dos stakeholders[1]". Para Engle e Tinto (2007), o aluno que evade representa uma grande perda de capital humano. Por outro lado, a melhora nos níveis educacionais aumenta a competitividade e contribui para o desenvolvimento econômico sustentável. Por isso, há grande interesse das instituições de ensino, legisladores/governos, pais e alunos de que haja mecanismos que possibilitem a permanência dos alunos nos cursos para os quais eles ingressaram até a sua conclusão.

Nos Estados Unidos, dados do Departamento de Educação, citado por Delen (2010), indicam que metade dos alunos que ingressam no ensino superior concluem seus cursos e melhorar as taxas de evasão têm sido um desafio para os gestores destas instituições. Aquelas

[1] São todas as partes interessadas, no caso da evasão escolar são os alunos, gestores das instituições, financiadores, governos, pais, etc.

que não conseguem manter seus alunos perdem visibilidade do público, qualidade e recursos financeiros, devido à diminuição das receitas com matrículas.

No Brasil, de acordo com os dados divulgados pelo MEC/INEP (2020) através do Censo da Educação Superior 2019, a taxa acumulada de desistência em 2019 foi de 59%, indicando também a alta evasão como um grave problema. Este índice compreende a evasão global, por isso, é necessário considerar que há cursos com menores índices de evasão do que outros.

As disciplinas das áreas de Ciência, Tecnologia, Engenharia, Matemática (STEM) possuem os maiores índices de evasão e apresentam um número ainda maior de evadidos quando se trata de alunos de minorias[2] (ALKHASAWNEH e HARGRAVES, 2014).

O estudo do fenômeno da evasão apresenta inúmeros elementos que devem ser considerados para entender este tema tão complexo (e.g. sistemas de ensino, modalidades, regimes acadêmicos), por isso não existe uma pesquisa que consiga agrupar todos os elementos. Segundo Castro e Teixeira (2014) a grande maioria diz respeito a instituições públicas, tem como escopo apenas uma instituição ou mesmo um único curso.

No Brasil, os sistemas de ensino estão divididos entre instituições públicas (federais, estaduais e municipais) financiadas quase que exclusivamente pelas respectivas esferas governamentais e, instituições privadas (comunitárias, confessionais, filantrópicas e particulares), financiadas por instituições mantenedoras, incentivos fiscais e mensalidades pagas pelos alunos (STALLIVIERE, 2006). Quanto aos cursos de graduação ofertados no País, eles são formados por cursos de Bacharelado, Licenciatura e Tecnólogo, ofertados na modalidade presencial ou à distância.

De acordo com o Censo da Educação Superior 2019 a maioria dos alunos de graduação estão matriculados em instituições privadas e apesar do crescimento do Ensino à Distância, o ensino presencial possui 56,1% das matrículas. (MEC/INEP, 2020)

No Brasil, existem dois regimes acadêmicos, o Regime de Matrícula Seriado (RMS) e o Regime de Matrícula por Disciplina (RMD), este também conhecido por sistema de matrícula por créditos. No primeiro, a matrícula funciona como na Educação Básica, sendo o aluno obrigado a se matricular no ano ou semestre. No segundo regime a instituição, oferece um percurso para o aluno trilhar e escolher suas disciplinas, algumas com pré-requisitos. No entanto, de acordo com Poulsen e Bandeira (2014) não há dados para identificar como é a distribuição destes regimes pelas instituições de ensino superior no país.

[2] Minorias (gênero, etnia e raça).

Outra complexidade diz respeito à conceituação do termo evasão, pois ela pode ser definida de muitas maneiras. Lima e Zago (2018) também faz esse apontamento, uma vez que ela "pode ser compreendida como abandono, desistência, fracasso, saída definitiva do curso, da instituição e/ou do sistema escolar". Além disso, encontramos na literatura o termo associado não somente ao abandono, mas a persistência, a retenção ou permanência. Conforme Costa e Gouveia (2018) os diversos estudos sobre o tema apresentam a retenção e a permanência quase como sinônimos, sendo que retenção é vista como a capacidade da instituição em manter o aluno no sistema, enquanto a permanência, é a vontade do aluno de permanecer até conquistar o diploma. De acordo com Casanova et al. (2018) o termo abandono é complexo, pois existem diferentes definições; já o termo permanência é tido como a manutenção do aluno no sistema até que ele complete a graduação.

Segundo a Andifes (1996), existem 3 tipos de evasão: evasão do curso (abandono, desistência, transferência ou reopção, ou exclusão), da instituição e do sistema (abandono definitivo ou temporário do nível superior). Para Delen (2010), a evasão é definida como o número de alunos que não completam a graduação naquela instituição. Rovira, Puertas e Igual (2017) considera o aluno que deixou o sistema universitário, não realizando matrícula por dois semestres consecutivos. Segundo os autores, esta definição permite identificar alunos que abandonaram pela dificuldade enfrentada em seus estudos. E ainda segundo Viloria et al. (2019), a evasão é o abandono de carreira antes da obtenção do grau pretendido, levando em consideração o prazo que impeça a possibilidade de retorno.

É preciso também considerar a origem dos dados. No ensino à distância, os Ambientes Virtuais de Aprendizagem (AVA) proporcionam aos gestores e docentes inúmeras informações geradas pela interação entre o aluno e a plataforma em tempo real, um campo propício para a Análise de Aprendizagem (em inglês, *Learning Analytics*). No entanto, no ensino presencial a maioria das instituições contam apenas com Sistemas de Gerenciamento de Matrículas. Por exemplo, o Sistema Integrado de Gestão de Atividades Acadêmicas (SIGAA) - desenvolvido pela Universidade Federal do Rio Grande do Norte em 2007 (SOUZA e MONTEIRO, 2015) - que gerencia as matrículas de diversas Universidades Públicas Federais no Brasil. Apesar de permitir algumas interações através de fóruns de discussão, o SIGAA ainda é limitado em relação aos AVA, o que requer um entendimento ainda maior sobre quais informações contidas nestes bancos de dados são úteis para o estudo do fenômeno da evasão.

Pesquisadores brasileiros vêm estudando estes sistemas de gestão como fonte de dados relativos à evasão, como no estudo realizado por Manhães e Cruz (2019), que utilizou apenas

dados acadêmicos da Universidade Federal do Rio de Janeiro para prever o perfil do aluno desistente. Um outro exemplo pode ser encontrado no trabalho de Silva (2013), que utilizou informações da base de dados do Centro Universitário FECAP - São Paulo para identificar fatores que retardam ou prolongam a permanência de um aluno na graduação.

Historicamente, a preocupação com a conclusão dos alunos que ingressam no ensino superior surgiu junto com este nível de ensino e desde então inúmeras pesquisas foram conduzidas. Segundo Berger et al. (2002) apud Aljohani (2016) os estudos sobre retenção podem ser agrupados em eras; a partir da década de 60 tem início os estudos sobre prevenção do abandono, seguido pela construção das teorias. As principais teorias e modelos sobre a evasão de alunos surgiram no final desta década e se intensificaram durante os anos 70, algumas delas são tão importantes que fundamentam e orientam estudos desenvolvidos nos dias atuais.

Spady (1970) desenvolveu um modelo teórico sociológico para explicar o processo de abandono, que pode ser explicado através de uma abordagem interdisciplinar envolvendo a interação entre o aluno e a instituição onde seus atributos sofrem influências, expectativas e demandas vindas de várias fontes. O modelo revisado (Spady, 1971) considera as variáveis: background dos alunos, congruência normativa, desempenho acadêmico, satisfação, integração social e apoio de amizades.

Tinto (1975) formulou um modelo teórico para explicar o processo de interação entre o indivíduo e a instituição no nível social e acadêmico que leva diferentes tipos de pessoas a abandonar a educação superior. O modelo conceitual desenvolvido por ele serve de base para inúmeros estudos e considera o background do aluno, fatores institucionais, contato informal, outras experiências na faculdade e os resultados educacionais para prever as decisões de abandono/permanência. Para Vincent Tinto, o processo de abandono é um processo longitudinal da interação entre o indivíduo, o sistema acadêmico e social da faculdade, em que as experiências pessoais vão modificando os objetivos e os compromissos de forma a levar o aluno a persistir ou a várias formas de abandono.

O Modelo Causal de Evasão de Alunos, elaborado por John Bean (1980), foi adaptado da rotatividade de empregados nas organizações de trabalho. A pesquisa pretendeu testar o modelo e elaborar um ranque das variáveis pela capacidade de explicar a evasão. O Modelo da Síndrome do Abandono (Bean, 1985) investigou comportamentos e atitudes que tendem a influenciar a evasão. O desenvolvimento do modelo enfatizou fatores acadêmicos, fatores sociopsicológicos, fatores ambientais e fatores de seleção/socialização, chegando à conclusão que a nota na faculdade contribui mais para a seleção do que a socialização dentro da instituição.

Astin (1984) desenvolveu a Teoria do Envolvimento do Aluno, que se refere a quantidade e qualidade de energia física e psicológica que o aluno investe na experiência na faculdade. As variáveis consideradas são: trabalho acadêmico, participação em atividades extracurriculares e interação dos alunos com a equipe e outras pessoas da instituição.

Os autores Cabrera, Nora e Castañeda (1992), decidiram explorar o papel das finanças em um modelo teórico de permanência. Segundo eles, a ajuda financeira contribui com um papel importante na equalização de oportunidades, na socialização do aluno e no seu compromisso de continuar seus estudos. A base conceitual da estrutura do modelo é oriunda de outros autores e emprega as seguintes variáveis: permanência institucional, a intenção de permanecer, desempenho acadêmico anterior, influência de outros significantes (pais e amigos), finanças e integração social.

Os modelos produzidos buscam entender o fenômeno da evasão levando em consideração diversas variáveis. Dessa forma, os estudos podem ser classificados segundo a sua abordagem em: econômico, psicológico, relacional, institucional ou sociológico como demonstrado na figura 2 (ABALCO et al., 2019). A perspectiva econômica analisa o quanto a ajuda financeira e a percepção do custo-benefício contribuem para aumentar ou diminuir a probabilidade do aluno evadir. Os estudos que tentam avaliar a influência dos traços de personalidade do indivíduo no processo de permanência são classificados como psicológicos. Os estudos conduzidos em uma abordagem sociológica são aqueles que investigam a influência dos fatores externos, como origem e apoio familiar nas decisões de abandono/permanência. A abordagem institucional analisa o quanto as características da instituição de ensino como o tamanho e oferta de serviços poderia explicar o processo de abandono. A última abordagem intitulada por Abalco e colaboradores de relacional[3] é aquela que investiga a integração dos alunos ao sistema social e acadêmico da instituição de ensino superior.

[3] Costa e Gouveia (2018) denominam esta abordagem como interacionista.

Figura 2: Modelos de Abandono Universitário (adaptado de Abalco et al., 2019)

De um modo geral os estudos mencionados acima em sua maioria, são estudos quantitativos e longitudinais, utilizam dados institucionais combinados com *surveys* desenvolvidos em escala Likert e empregam em suas análises técnicas estatísticas. Segundo Rovira, Puertas e Igual (2017), as análises estatísticas são baseadas em hipóteses desenhadas para um problema subjacente, e são melhores para quando o objetivo é entender as variáveis envolvidas no problema.

Além da elaboração de teorias e modelos que tentam explicar e representar o fenômeno da evasão, os estudos contemporâneos têm tido como escopo a *Learning Analytics* (LACAVE, MOLINA-DÍAZ, CRUZ-LEMUS, 2018), *Educational Data Mining* (SARRA, FONTANELLA, DI ZIO, 2018), Sistema baseado em dados "*Data Science*" (ROVIRA, PUERTAS, IGUAL, 2017) e *Data Mining* (DELEN, 2010). As pesquisas são empreendidas com vistas a identificar fatores que levam os alunos a abandonar o ensino superior, aqueles com probabilidade de se transferir para outras instituições ou de área; mitigar a evasão, prever o desempenho acadêmico ou avaliar o impacto dos programas de intervenção e apoio acadêmico na permanência dos alunos com risco de evadir.

> a pesquisa acadêmica relacionada ao fenômeno dos abandonos e/ou baixo desempenho acadêmico no nível universitário foi inicialmente realizada dentro de um contexto de causa e efeito, com parte descritiva das causas e prescritiva para as soluções, onde foi apontado que o estudo dos abandonos é um problema complexo e multidimensional, as últimas pesquisas estão mudando o foco e os esforços das ações

> do aluno para investigar aspectos associados à instituição e suas intervenções para evitar abandonos. (LOZANDO, VIEITES, CALABUIG, 2017) - "tradução da autora"

2.2 As Tecnologias de *Data Analytics* para Apoio e Desenvolvimento de Políticas de Ensino Superior

O fenômeno da evasão é complexo e multidisciplinar. Áreas como Educação, Sociologia, Psicologia, Computação e Economia buscam investigá-lo de forma a entender e mitigar seus efeitos. Muitos dos estudos foram conduzidos em uma perspectiva qualitativa, enquanto outros, têm utilizado os dados para gerar resultados e apoiar a formulação de políticas públicas, como metodologias de análise oriundas da Econometria, que foram utilizadas por Leon e Martines-Filho (2002) para estudar a evasão e reprovação de alunos no ensino superior, a partir de dados do Instituto Brasileiro de Geografia e Estatística - IBGE.

A produção de uma enorme quantidade de dados, na chamada Era Digital, tem propiciado cada vez mais a utilização de métodos quantitativos para análise. Estes dados são criados, coletados e armazenados por empresas e governos, chamados *Big Data* (Grandes Dados - em uma tradução livre). Uma das definições mais utilizadas para o termo Big Data é a dos 5Vs, Volume, Variedade, Velocidade, Veracidade e Valor (SALGANIK, 2018). No entanto, uma definição mais ampla pode ser encontrada em Freitas Junior et al. (2016) onde os autores após um estudo das definições encontradas na literatura concluíram que *Big Data* pode ser sintetizada "como sendo um grande volume de dados estruturados ou não, de fontes diversas, que devem ser gerenciados e analisados de forma peculiar".

Nessa conjuntura apresenta-se a Ciência de Dados (em inglês, *Data Science*) caracterizada como um conjunto de princípios, processos e técnicas que tem por objetivo entender fenômenos via análise de dados automatizada e que no alto nível, guia a extração do conhecimento de dados (PROVOST e FAWCETT, 2013). A Ciência de Dados é uma área interdisciplinar aplicada à análise de dados de diversas naturezas, em especial a econômica e social.

A *Data Analytics* está a serviço da Ciência de Dados, ela entra em ação quando o volume de dados armazenados automaticamente pelas organizações passa por um processo de análise de dados com um propósito específico. Sua utilização, de acordo com Hirave e colaboradores (2018), é um fator crucial para auxiliar qualquer organização que queira tomar decisões baseada em informação significativa.

A *Data Analytics*, também conhecida como Big Data Analytics (BDA), é definida por Elragal e Haddara (2019) como "o uso de técnicas de mineração de dados e estatísticas com o objetivo de encontrar ou prever padrões desconhecidos na big data". Espera-se com o uso da *Data Analytics* não somente solucionar o problema, mas entender todo o processo, analisando o "negócio" como um todo. O Quadro 1 apresenta uma estrutura de abordagem de *Data Analytics* elaborada por Hirave e colaboradores para guiar a criação de um sistema de análise de dados no ensino à distância.

Quadro 1 - Estrutura para análise da abordagem de *Data Analytics*

	ANÁLISE DESCRITIVA (Fundamental)		**ANÁLISE DIAGNÓSTICA (Operacional)**	**ANÁLISE PREDITIVA (Perspicácia)**	**ANÁLISE PRESCRITIVA (Estratégica)**
Pergunta	O que aconteceu no passado?	O que está acontecendo agora?	Por que aconteceu e qual é o relacionamento?	O que acontecerá no futuro?	Como nós deveríamos agir no futuro?
Foco Processo	Comunicado	Medida/ monitoramento dos indicadores chaves de desempenho	Análise de tendência, Análise situacional, Caso raiz, Causa e efeito; e Análise de grupo.	Previsão, Avaliação de probabilidade, Administração de risco e Predição.	Planejamento baseado em evidências, Simulação e formulação de estratégias, Otimização de opções.
Ferramentas e técnicas	Relatórios estáticos e interativos	Painel de controle, Cartão de pontuação de desempenho	Mineração de dados, modelos estatísticos, ferramentas de busca, Rodas de programação, Ferramentas OLAP, Árvores de Decisão	Análise de hipóteses, Aprendizagem de Máquina, Modelagem preditiva, Redes Neurais, Visualização de dados	Modelagem escolha discreta, Programação linear e não-linear, Análise de valor
	Retrospectiva			**Em perspectiva**	

Fonte: Adaptado de Hirave et al. 2018.

Estudos recentes sobre evasão têm utilizado a mineração de dados (DELEN, 2010) para extração de conhecimento. Muitos deles que investigam o fenômeno da evasão no ensino superior na modalidade presencial utilizam o registro acadêmico como fonte de dados e muitos consideram o desempenho acadêmico (nota) como uma importante variável preditiva. Os sistemas de gerenciamento de matrículas das instituições permitem armazenamento de inúmeras informações acerca dos alunos, coletados no ato da matrícula e alimentados ao longo dos cursos.

A análise dos dados gerados pelos sistemas educacionais tornou-se essencial para combater os problemas associados à evasão. Um exemplo pode ser encontrado no trabalho realizado por Delen (2010), onde o autor utiliza técnicas de mineração de dados para auxiliar na tarefa de prever e explicar a natureza do problema do abandono a partir da utilização de dados históricos coletados nos sistemas educacionais da instituição. Alkhasawheh e Hargraves (2014) afirmam, com base na literatura estudada por eles, que as técnicas de mineração de dados são consideradas técnicas robustas e de alta acurácia.

Da mesma forma que Rovira, Puertas e Igual (2017) propõem um sistema baseado em dados para análise da evasão na Universidade de Barcelona, utilizando técnicas de aprendizagem de máquina para prever a evasão de alunos e as notas, outros estudos têm apostado nessas técnicas, aplicando-as à tarefa de previsão da evasão de alunos no ensino superior. A Aprendizagem de Máquina (do inglês, *Machine Learning*) "é um conjunto de técnicas que dá aos computadores a capacidade de aprender sem a intervenção da programação humana" (RASTROLLO-GUERRERO; GÓMEZ-PULIDO; DÚRAN-DOMINGUEZ, 2020). Ela busca investigar como os artefatos computacionais aprendem (ou melhoram o seu desempenho) baseado nos dados (HAN; KAMBER; PEI, 2011).

A figura 3 ilustra os principais tipos de aprendizagem: aprendizagem clássica (supervisionada e não supervisionada); métodos *ensembles*; aprendizagem por reforço e, rede neural profunda (em inglês, *deep learning).*

A aprendizagem supervisionada é aquela, cuja entrada de dados é rotulada, quando a máquina possui um parâmetro para ela trabalhar. Ela pode ser uma tarefa de classificação, quando as variáveis são nominais/categóricas (BEULAC e ROSENTHAL, 2019), quando se trata de evasão, é possível que a partir dos dados informados (*e.g.* idade, gênero, renda) a máquina classifique os alunos segundo a possibilidade de evadir ou não. A aprendizagem supervisionada também pode ser uma tarefa de regressão quando os rótulos ou a variável de resposta é um número (RUSSEL e NORVIG, 2009), ou seja, quando estamos tentando prever um valor, que pode ser a nota de um aluno no próximo semestre.

A aprendizagem não supervisionada consiste em dados sem rotulação, segundo Russel e Norvig (2009) o modelo aprende sem que haja um *feedback* definido. A máquina não é treinada a partir de um rótulo no qual ela colocará o aluno, ela o faz de acordo com seus próprios parâmetros, com base no conjunto de dados. Dentre as tarefas não supervisionadas estão: *clustering* (faz o agrupamento dos dados segundo seu grau de semelhança) (SARRA, FONTANELLA, DI ZIO, 2018); associação (busca elementos que ocorrem em comum dentro

de um conjunto de dados) e, redução de dimensão ou generalização (busca reduzir a dimensão dos dados em um subespaço que capte a essência dos dados).

A Aprendizagem por Reforço consiste no treinamento do modelo de aprendizado para tomar uma sequência de decisões, neste caso, diz-se que a máquina aprende com os próprios erros, numa tentativa de imitar o comportamento humano. Os Métodos *Ensembles* fazem uma combinação de vários modelos em busca de encontrar o melhor modelo preditivo (IAM-ON e BOONGOEN, 2015). Enquanto a *Deep Learning* emprega algoritmos para processar dados e imitar o processamento feito pelo cérebro humano (RUSSEL e NORVIG, 2009).

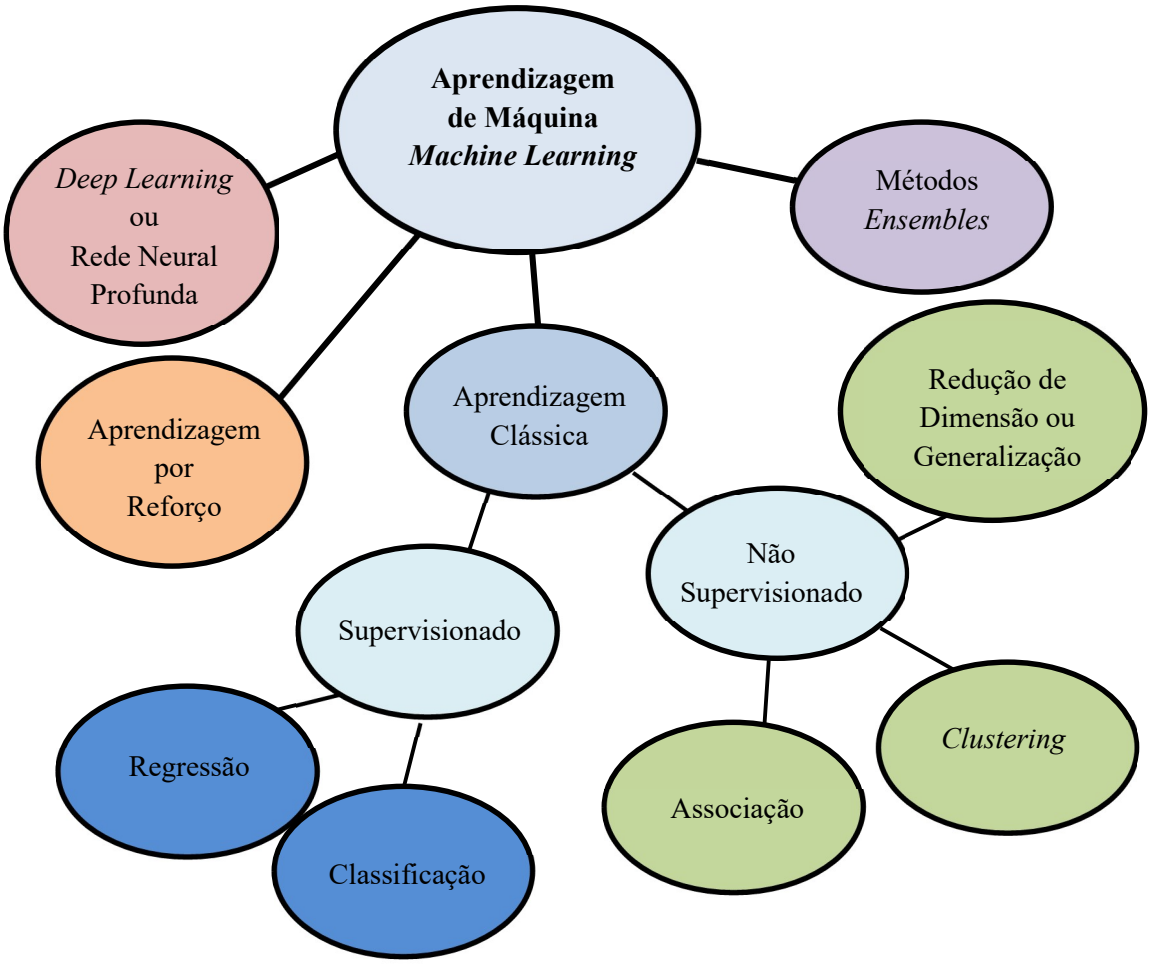


Figura 3. Tipos de Aprendizagem de Máquina (Elaborado pela autora)

2.3 A importância da Revisão Sistemática

Toda pesquisa tem início com uma revisão da literatura. As revisões de literatura podem ser classificadas em narrativas e em revisões bibliográficas sistemáticas de acordo com suas

características e objetivos. As revisões narrativas, ou revisões tradicionais, não possuem preocupação com metodologias, visa conhecer e apresentar o estado da arte sobre determinado assunto. (BOTELHO; CUNHA; MACEDO, 2011). Enquanto as revisões bibliográficas sistemáticas são específicas quanto a questão que pretende ser respondida. Ainda segundo os autores as revisões bibliográficas sistemáticas podem ser classificadas de acordo com sua metodologia em: meta-análise, revisão sistemática, revisão qualitativa e revisão integrativa. À revisão sistemática acrescenta-se dois complementos: os estudos de mapeamento sistemático e as revisões terciárias. (KITCHENHAM et al., 2007)

Quando uma Revisão Sistemática (RS) é empreendida ela busca compreender todo o cenário a respeito da questão em particular sob investigação. Dentro deste aspecto é possível testar hipóteses que ainda não foram testadas em estudos primários, relacionar todo conhecimento já produzido a respeito de determinado assunto, identificando tendência, conceitos, métodos e resultados. (BIOLCHINI et al., 2005).

A história da Revisão Sistemática tem origem na medicina, para apoiar a prática baseada em evidência na área da saúde, ela é um meio de avaliar e interpretar toda pesquisa relevante a respeito de um assunto, área ou fenômeno de interesse. Sendo uma forma de estudo secundário que usa uma metodologia bem definida para análise com o objetivo de evitar o viés durante o processo de pesquisa. (KITCHENHAM et al., 2007).

A prática baseada em evidência segundo Botelho; Cunha; Macedo (2011) tem sido incorporada nas Ciências Sociais através da utilização de métodos que permitem coletar, categorizar, avaliar e sintetizar resultados de pesquisa relacionados ao tema investigado.

A relevância da revisão sistemática da literatura que será conduzida reside na escassez dos trabalhos secundários para o tema específico e em fornecer aos stakeholders da dinâmica educacional terciária um painel dos estudos realizados, dos conhecimentos produzidos, dos resultados obtidos e, se os mesmos, tendo em comum a mesma classe de problemas, possa ser replicado em um contexto particular ou até mesmo generalizado e se, é capaz de solucionar o problema da evasão de alunos no Ensino Superior.

2.3.1 Revisões Sistemáticas sobre Evasão Escolar no Ensino Superior

Encontramos na literatura revisões sistemáticas cujo tema é a evasão escolar no Ensino Superior. No entanto, cada trabalho busca investigar a partir de uma perspectiva específica. Aljohani (2016) realizou uma revisão compreensiva com o objetivo de fornecer aos

pesquisadores, educadores e legisladores uma visão a respeito dos principais estudos e modelos teóricos sobre a retenção de alunos no Ensino Superior, a fim de elencar as técnicas e estratégias utilizadas e os padrões que estão relacionados com a evasão de estudantes. No entanto, seu estudo não apresenta a metodologia que norteou a pesquisa.

Alban e Maurício (2019) apresentaram uma pesquisa cujo foco é a aplicação de Mineração de Dados (do inglês, *Data Mining*) para prever o abandono. Eles estabeleceram um protocolo de três etapas (planejamento, implementação e resultados); a pesquisa buscou por artigos de conferências e journals indexados nas plataformas: Science Direct, ACM Digital Library, IEEE Xplore, Springer, DOAJ, Taylor e Francies, Emerald, Proquest e Ebsco, entre Janeiro de 2006 e Dezembro de 2017. Foi estabelecido um critério de inclusão e exclusão. O critério de busca foi “dropout student” OR “drop out student” OR “dropping student” AND “data mining” aplicada ao título, abstract e palavras-chaves. A análise pretendeu responder a 5 perguntas: 1. Qual técnica foi utilizada para o pré-processamento dos dados? 2. Que fatores afetam a evasão? 3. Quais técnicas são utilizadas para selecionar os fatores? 4. Quais técnicas são utilizadas para revisão e quais são os níveis de confiabilidade? e 5. Quais ferramentas foram usadas?

Em 2019 Liz-Dominguez e seus colaboradores realizaram uma revisão sistemática com o objetivo de explorar o estado da arte sobre o uso da Análise Preditiva (em inglês, *Predictive Analytics*) na Educação Superior. O processo de recuperação dos documentos consistiu de fontes e repositórios online (IEEE Xplore Digital Library, ACM Digital Library, Elsevier, Wiley Online Library, Springer e Google Scholar). Através das seguintes strings (“early warning system”; predictive analysis”, “predictive analytics”, predictive algorithm”, “education”, "university"); excluindo as palavras desastre, médico e saúde. O foco do trabalho é a Análise Preditiva que inclui as várias técnicas que realizam predições de futuros resultados confiáveis em dados históricos e atuais. Os autores abordam a proximidade da *Data Analytics* com outros campos de estudo como a mineração de dados e aprendizagem de máquina e que compartilham técnicas semelhantes. Eles consideraram relevantes os trabalhos que explicam claramente os algoritmos, que tenham sido testados e que apresentaram resultados. Três passos foram utilizados durante o processo de seleção: primeiro, análise de títulos e abstracts descartando aqueles que não estavam relacionados a área educacional; segundo, análise da introdução e conclusão e por último a análise completa dos textos e avaliação da relevância dos mesmos.

O trabalho de Rastrollo-Guerrero, Gómez-Pulido, Dúran-Domínguez (2020) trata de uma pesquisa qualitativa, o estudo analisou artigos recentes a respeito de diferentes técnicas aplicadas para prever o comportamento do estudante. Eles consideraram estudos de diferentes fontes, capítulos de livros, journals e conferências. Os autores excluíram trabalhos considerados sem qualidade suficientes, aqueles sem fator de impacto segundo a ISI Journal Citation Report e sem revisão pelos pares. Também utilizaram como critério de exclusão as conferências que não foram organizadas/apoiadas/publicadas pela IEEE, ACM, Springer ou por organizações ou editoras reconhecidas. A busca nos bancos de dados utilizou os descritores "Predicting students' performance", "Predicting algorithms students", "Machine Learning prediction students" e outros relacionados. A classificação dos artigos foi feita quanto a técnica e quanto aos objetivos. Diferente do que propõe esta revisão, cujo escopo é a evasão no Ensino Superior, o estudo diz respeito à investigação do uso de técnicas computacionais para prever o desempenho, nota e o risco de evasão escolar, infantil, no ensino médio e na universidade.

O objetivo do trabalho realizado por Charitopoulos, Rangouse e Koubourotes (2020) consiste de uma pesquisa realizada em publicações recentes (2010-2018) que usam métodos de habilidades computacionais para responder problemas relacionados à educação baseado em análise de dados educacionais, minerados principalmente em sistemas interativos e a distância. O foco do estudo é a Mineração de Dados Educacionais (*do inglês Educational Data Mining - EDM*) e a Aprendizagem Analítica (*do inglês Learning Analytics - LA*), apresentadas como áreas próximas e distintas que usam a mineração de dados para solucionar problemas educacionais. A metodologia utilizada foi adaptada de Kitchinham et al. (2004). Eles utilizaram a base de dados da Scopus para extrair as publicações a serem analisadas utilizando as seguintes palavras-chaves: "Educational Data Mining", "Learning Analytics" e "Education", "Educational Data Mining" e "Learning Analytics", "Educational Data Mining" ou "(Learning Analytics" e "Education". O estudo se limitou aos estudos primários e a análise inicial foi feita em três passos: (1) análise de título e abstract; (2) aplicação de critérios de inclusão e exclusão aos títulos e abstracts e (3) análise completa dos textos. O Coeficiente Kappa de Cohen foi utilizado para calcular a confiabilidade da seleção, visto ter mais de um revisor.

2.4 *Design Science Research* (DSR)

Como veremos na proposta deste trabalho, uma revisão sistemática será desenvolvida a partir de uma organização dos trabalhos de *Data Analytics* voltados para evasão no ensino

superior sob a égide do paradigma epistemológico da *Design Science Research* (DSR). Portanto, a seguir descrevemos tal paradigma.

Inicialmente, é preciso compreender que a DSR é um método de pesquisa que tem como objetivo instrumentalizar as pesquisas da *Design Science*, também conhecida como Ciência do Projeto ou Ciência do Artificial. Este método desponta na literatura como um complemento aos métodos de investigação realizados sob a ótica das ciências tradicionais (naturais e sociais) onde as pesquisas em geral tem o objetivo de explicar, descrever, explorar ou predizer o que acontece e qual a relação existente entre os fenômenos. Diferente do que ocorre dentro das organizações, onde as pesquisas possuem um caráter prescritivo e são orientadas para a solução de problemas, através do estudo ou criação de um artefato. (DRESCH, LACERDA, ANTUNES, 2014)

A *Design Science* é a ciência responsável por todo o conhecimento que se origina a partir da criação de um artefato e de todo o processo empregado na sua construção. Ela se fundamenta em métodos computacionais e matemáticos que buscam avaliar a qualidade e a efetividade dos artefatos (constructos, modelos, métodos e instanciações), sendo de grande utilidade para fundamentar e/ou analisar as pesquisas, uma vez que o objeto de muitos estudos é a análise de artefatos ou prescrições, objetivando a melhoria de um processo ou a solução de um problema, como no caso deste estudo, cujo escopo é o fenômeno da evasão de alunos. (HEVNER et al., 2004)

Na atual conjuntura, estudos são de natureza prescritiva e muitos artefatos são projetados para solucionar problemas existentes no mundo real e não somente especificidades do mundo acadêmico. Nesse contexto, a *Design Science Research* tem sido um importante método para condução de pesquisas, diminuindo a distância existente entre a teoria e a prática (utilidade do projeto). Com aplicações, segundo Dresch, Lacerda e Anuntes (2014), em diversas áreas como Arquitetura, Ciências Sociais, Educação, Engenharia, Gestão e Sistemas de Informação.

> Como método de pesquisa orientado à solução de problemas, a Design Science Research busca, a partir do entendimento do problema, construir e avaliar artefatos que permitam transformar situações, alterando suas condições para estados melhores ou desejáveis. Ela é utilizada nas pesquisas como forma de diminuir o distanciamento entre a teoria e a prática. (DRESCH, LACERDA, ANTUNES, 2014)

A necessidade de tornar a dicotomia teoria e prática mais próxima leva a duas importantes características da *Design Science Research*: o rigor e a relevância. O Quadro 2 mostra a classificação das pesquisas segundo estas características.

Quadro 2 – Classificação das pesquisas em termo de rigor e relevância

		Relevância	
		Baixo	**Alto**
Rigor teórico e metodológico	**Baixo**	Pesquisa Indesejada	Pesquisa leviana
	Alto	Pesquisa autocentrada	Pesquisa necessária

Fonte: Adaptado de Dresch, Lacerda e Antunes. 2014.

As **pesquisas indesejadas** são aquelas com baixo rigor e baixa relevância, pois não são úteis à solução do problema, não possuem sustentação teórica ou metodológica. As **pesquisas levianas** são aquelas cujos resultados são úteis apenas para uma determinada empresa, possuem baixa fundamentação teórica e alta relevância para um grupo específico. As **pesquisas autocentradas** têm foco no mundo acadêmico, com alto rigor teórico-metodológico, contribuindo muito para ampliação do conhecimento acerca do fenômeno estudado, no entanto possuem baixa relevância para a solução de problemas. As **pesquisas necessárias** são aquelas que combinam alto rigor teórico-metodológico e alta relevância. Esse tipo de pesquisa é o foco da DSR, pois diminui a distância entre a teoria (academia) e a prática (aplicação da solução nas organizações).

A abordagem da *Design Science Research* está centrada em dois pontos: a **construção** e a **avaliação** de um artefato. Ulrich et al. (2018) aponta que o método DSR envolve a realização de três atividades: a construção de um artefato, a coleta dos dados sobre o desempenho do artefato com base em uma avaliação empírica e a reflexão sobre o processo de construção. No entanto, para Hevner et al. (2004) a *Design Science Research* é composta por dois processos: a construção e a avaliação de artefatos sendo o processo de construção do artefato direcionado

para um problema sem solução, enquanto o processo de avaliação direcionado para verificar a utilidade do artefato em solucionar o problema em questão.

Hevner e colaboradores (2004) propõem um roteiro para condução de pesquisas segundo a DSR para ser aplicada na análise de estudos na área de Sistema de Informação. A proposta é que sejam analisados os seguintes tópicos:

1. **Projeto como um artefato:** as pesquisas devem produzir artefatos viáveis, na forma de um constructo, modelo, método ou de uma instanciação.
2. **Relevância do Problema**: a DSR desenvolve soluções para resolver problemas importantes e relevantes para o negócio.
3. **Avaliação do Projeto:** demonstração da utilidade, da qualidade e da eficácia do artefato devem ser rigorosas por meio de métodos de avaliação bem executados.
4. **Contribuição da Pesquisa:** ela deve prover contribuições claras e verificáveis nas áreas específicas dos artefatos desenvolvidos e apresentar fundamentação clara em fundamentos da *design* e/ou metodologias de design.
5. **Rigor da Pesquisa:** a aplicação da pesquisa deve ser baseada em métodos rigorosos, tanto na construção como na avaliação dos artefatos.
6. **Projeto como um processo de Pesquisa:** a busca por artefato efetivo requer a utilização de meios confiáveis para alcançar os fins desejados enquanto satisfazem as leis do ambiente onde o problema ocorre.
7. **Comunicação da Pesquisa:** uma pesquisa conduzida pelo método da *Design Science Research* deve ser apresentada tanto para o público voltado para a tecnologia quanto para aquele orientado à gestão.

Em linhas gerais o pesquisador que empreende uma pesquisa segundo a *Design Science Research* precisa ter em mãos um problema que seja relevante e demonstrar que ainda não existem soluções para este problema ou que seja possível encontrar uma mais adequada. Em seguida o pesquisador apresenta o desenvolvimento do artefato que seja capaz de solucionar o problema em questão. A próxima etapa é a avaliação deste artefato quanto a sua utilidade e viabilidade, bem como expor a contribuição dessa pesquisa para aumentar o conhecimento a respeito do assunto, podendo depois dessa etapa passar para a comunicação da pesquisa de forma que todos possam entender tanto a comunidade acadêmica quanto os profissionais envolvidos.

A Figura 4 demonstra como a condução das pesquisas segundo a DSR se relacionam com o rigor e a relevância, suas principais características. É possível visualizar como o ambiente organizacional e os problemas inerentes a ele e que são passíveis de investigação contribuem para enriquecer o conhecimento já existente. Com base na DSR é possível criar métodos e artefatos, verificando a teoria e os artefatos desenvolvidos por outros pesquisadores que possam ser aplicados ou melhorados, gerando um ciclo contínuo que contribua para melhorar e ampliar a base de conhecimento a partir dos resultados obtidos com as pesquisas.

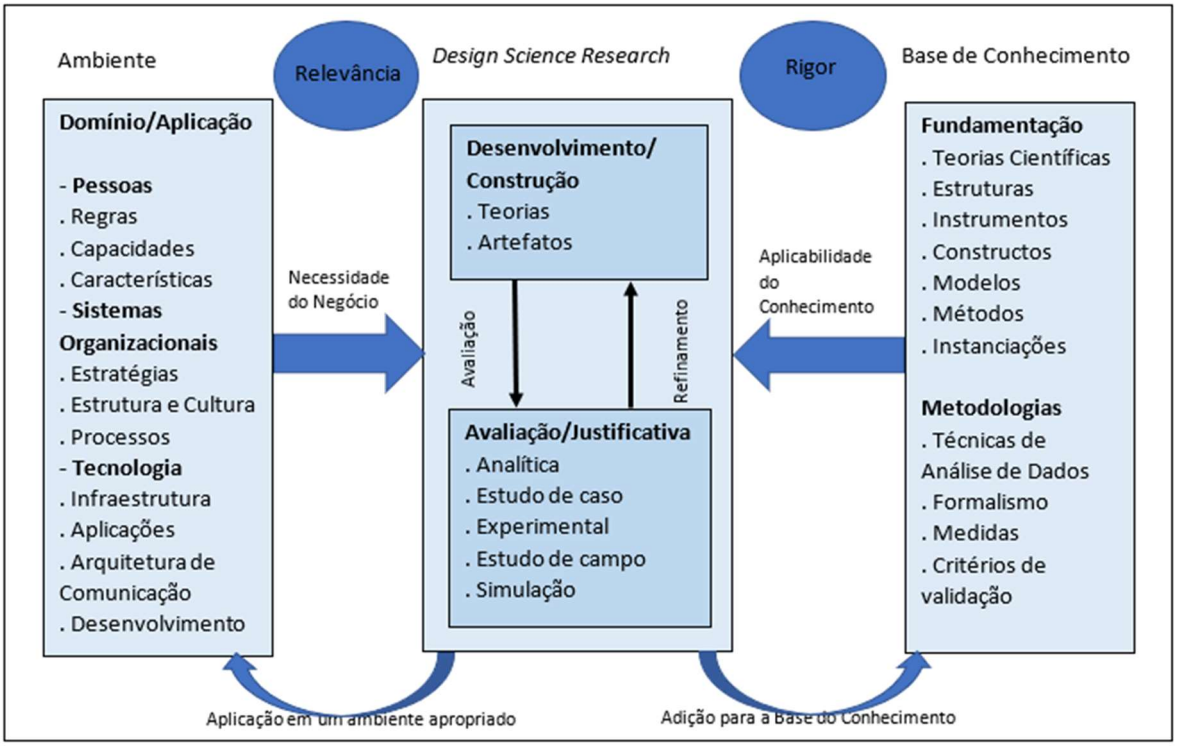


Figura 4 - Ciclo da *Design Science Research* (Adaptado de Hevner et al. 2004)

2.4.1 DSR e Data Analytics

Quando investigamos a literatura em busca de estudos que utilizam a *Design Science Research* e a *Data Analytics* encontramos o trabalho de Ulrich et al. (2018), onde a abordagem da *Design Science* é utilizada para criar uma estrutura que sirva como ponto de partida para pesquisas relacionadas às características das empresas de pequeno e médio porte e implementação de um sistema de *data analytics*. Esta estrutura tem como objetivos estabelecer uma base acadêmica sólida para futuras pesquisas e formar a base para implementação da *data analytics* nas companhias. Após uma revisão da literatura eles constataram que já existem pesquisas sobre a implantação da *data analytics* em diversos tipos de empresas e setores pelo

mundo, no entanto, os estudos demonstram que as empresas já realizam a coleta de dados internamente, mas falham em analisá-los. Para os autores os sistemas de *Data Analytics* precisam preencher os seguintes requerimentos: nível de transparência, nível de complexidade, acesso por todas as funções, orientação da seleção (alvo), análise em tempo real, know-how do usuário, comparabilidade e segurança dos dados.

Outro trabalho que emprega a DSR e a *Data Analytic*s é o realizado por Elragal e Haddara (2019) focado na avaliação, uma vez que ela é a atividade central quando se trata da construção de artefatos e na condução da *Design Science Research*. Sendo assim eles questionam se o surgimento da *big data analytics* impacta na condução das avaliações da *Design Science Research*, exigindo um mecanismo eficiente e confiável para a avaliação dos artefatos. Segundo os autores, existem quatro classes ou gêneros da DSR (Computacional, Otimização, Representação e Sistema de Informação Econômico) e apenas a Computacional e a Otimização estão relacionadas com a *big data analytics*, pois elas possuem grande importância na construção do artefato. Muitos dados são produzidos, dados internos e externos e em um volume cada vez maior. A questão envolve a veracidade dos dados e mecanismo de avaliação confiáveis tanto para aquisição quanto para o pré-processamento dos dados adquiridos, o que influencia na forma de condução da DSR, gerando a preocupação com questões como que tipo, representação, conclusão e interpretação de tais conclusões a respeito dos dados.

Como observado nos estudos citados acima, a DSR é um método aplicado a Sistemas de Informação, mas que pode ser utilizado em outras áreas, e a proposta deste estudo é elaborar uma metodologia que possa ser adaptada para abordagem de *Data Analytics* em estudos relacionados ao fenômeno da evasão no Ensino Superior (figura 7).

É a partir da fundamentação teórica descrita nos parágrafos acima que este projeto de revisão tem início e cuja metodologia é descrita a seguir.

3. METODOLOGIA

Nas seções seguintes serão descritas tanto o processo de revisão quanto o desenvolvimento da proposta de metodologia de análise de trabalhos de *data Analytics*, fundamentada pelo método da *Design Science Research*.

3.1 Desenvolvimento da Metodologia

A análise dos estudos primários selecionados usará o método adaptado da *Design Science Research* (DSR) para a abordagem de desenvolvimento de artefatos de *Data Analytics*, conforme apresentado na figura 5.

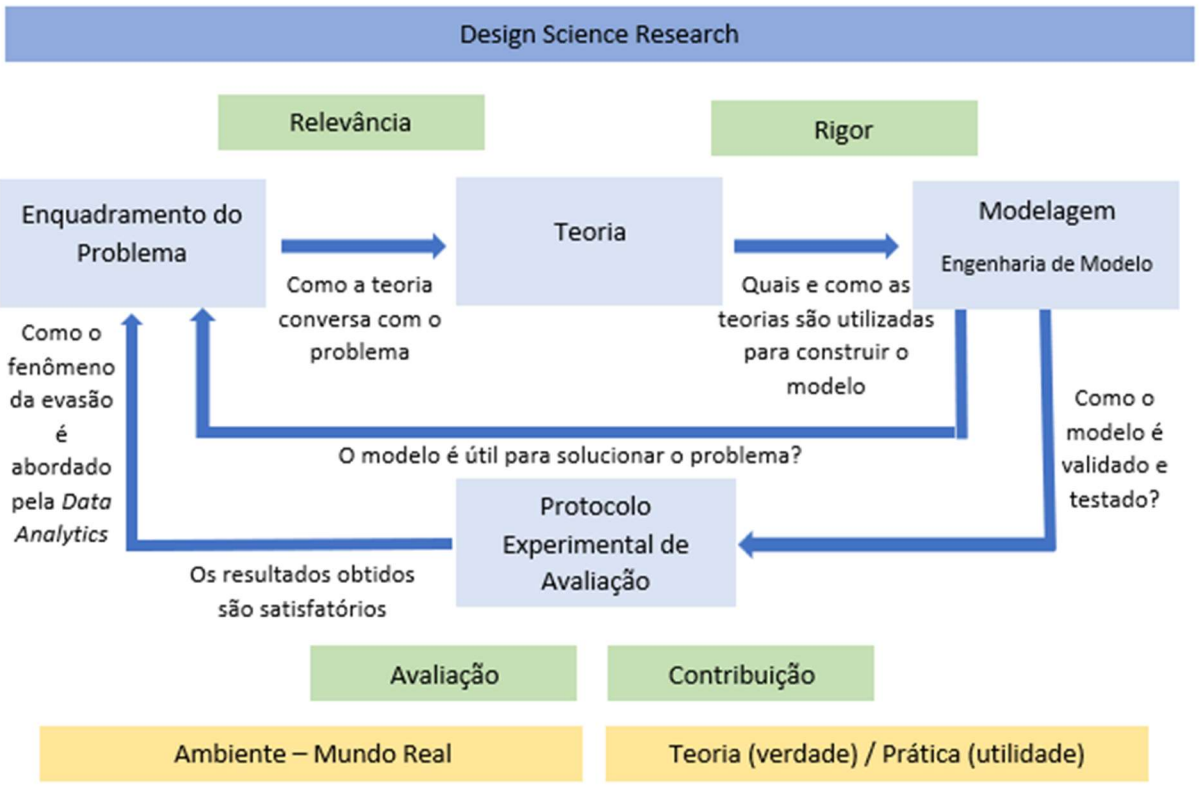


Figura 5. Esquema Metodológico (elaborado pela autora)

Esse esquema metodológico proposto apresenta os quatro componentes que orientarão a análise dos estudos selecionados: **Enquadramento do Problema**, **Teoria, Modelagem ou "Engenharia de Modelo"** e **Protocolo Experimental de Avaliação**.

O fenômeno da evasão investigado nesta pesquisa é inerente ao Ensino Superior ministrado na modalidade presencial. É neste ambiente, composto por alunos, professores,

gestores, instituições de ensino superior e tecnologias (sistemas de gestão/administração de matrículas) onde os problemas relativos ao fenômeno em questão podem ser observados.

É a partir deste cenário que identificamos a "necessidade do negócio", que será definida através do componente Enquadramento do Problema. A partir deste componente, pretende-se identificar como o fenômeno da evasão tem sido tratado por pesquisadores e praticantes. Três estratégias foram estabelecidas para identificar aspectos relativos ao enquadramento do problema, a partir da abordagem de *Data Analytics*. A seguir, apresentamos tais estratégias:

- **Quanto à tarefa do modelo:** deve-se analisar a proposta de enquadramento de acordo com a tarefa que o modelo de *Data Analytics* tem como objetivo. Por exemplo, se o modelo desempenha uma tarefa preditiva, diagnóstica, descritiva, prescritiva ou prognóstica.
- **Quanto à escolha do modelo:** deve-se identificar, a partir da descrição dos modelos adotados nos trabalhos, qual foi o enquadramento da tarefa do modelo definido (ver figura 4). Dependendo do enquadramento adotado, um mesmo problema de negócio pode ser abordado com tarefas diferentes. Por exemplo, o risco de evasão de um aluno pode ser enquadrado como um problema de classificação num caso, enquanto em outro pode ser enquadrado como um problema de detecção de anomalias. Caso não haja declaração explícita no estudo, será identificado a partir da técnica de aprendizagem de máquina que está sendo utilizada;
- **Quanto à definição do problema de negócio:** deve-se identificar a definição de abandono (evasão) utilizada nos trabalhos, tendo em vista as muitas possibilidades já apontadas na literatura.

Esse componente está diretamente relacionado ao critério de *Relevância* da Metodologia da DSR, pois especifica qual o problema real a ser investigado e qual artefato de *Data Analytics* deve ser desenvolvido.

A teoria é o componente que tem como objetivo identificar qual fundamento foi utilizado na construção dos artefatos, investigando do ponto de vista científico quais teorias sobre o fenômeno norteiam o desenvolvimento do projeto. De maneira específica, este componente investigará quais características do fenômeno foram consideradas importantes, pelos pesquisadores, para a construção do modelo (*e.g.* demográficos, financeiros, desempenho escolar anterior e desempenho acadêmico).

O componente Protocolo Experimental de Avaliação diz respeito aos critérios de avaliação que são utilizados, não só para validar e testar o artefato, mas também para refutar ou reforçar hipóteses teóricas sobre o fenômeno. No caso, refere-se a construção do modelo, a origem do conjunto de dados, como eles são divididos para a seleção de parâmetros e para a avaliação do desempenho do modelo. O objetivo é identificar esquemas metodológicos úteis para avaliar o fenômeno da evasão à luz do artefato, de modo que se tenha uma visão verossímil de quanto a solução agrega ao problema real.

Tanto a análise da Teoria quanto dos resultados experimentais, acabam por integrar e ampliar a base de conhecimento. Ambos estão relacionados com a *Design Science Research*, representando a dicotomia teoria e prática. É importante lembrar que para Dresch, Lacerda e Antunes (2015) não existe uma única solução para o problema ou classe de problemas, mas sim uma solução melhor. O Protocolo Experimental está relacionado ao critério de *Avaliação* e ao critério de *Rigor,* que visa garantir a confiabilidade dos resultados. Juntos Teoria e Protocolo Experimental corroboram também para o critério de *Contribuição,* associada a divulgação dos resultados.

O componente Modelagem ou "Engenharia de Modelo" tem como objetivo analisar os artigos em duas questões: como a Teoria foi traduzida em termos de variáveis preditivas e qual é a variável dependente (o alvo). Este componente atende ao critério de *Rigor*.

O método proposto nesse estudo (figura 5) apresenta explicitamente os três elementos presentes na estrutura proposta por Hevner et al. (2004): Ambiente, DSR e Base de Conhecimento. Além de atender aos dois fatores fundamentais para a condução de pesquisas segundo a *Design Science Research*, Rigor e Relevância, juntamente com os critérios de Contribuição e Avaliação.

3.2 Processo de Revisão Sistemática

A condução da revisão sistemática da literatura deste estudo é norteada pelo Modelo de Protocolo de Revisão proposto pelos autores Jorge Biolchini, Paula Gomes Mian, Ana Candida Cruz Natali e Guilherme Horta Travassos. Segundo os autores, o modelo foi elaborado para servir de diretriz para pesquisadores de Engenharia de Software e foi desenvolvido baseado nos protocolos de revisão sistemática da área médica (Biolchini et. al., 2005).

Em uma perspectiva mais específica o processo de revisão sistemática é composto por cinco estágios e, em uma perspectiva macro, a revisão sistemática é composta por três passos (figura 6): planejamento, execução e avaliação da revisão:

- *Planejamento:* formulação do problema, coleta de dados e avaliação dos dados;
- *Execução:* etapa da execução da seleção e extração da informação através do processo de interpretação e análise da informação;
- *Análise dos resultados:* processo de conclusão e apresentação dos dados obtidos.

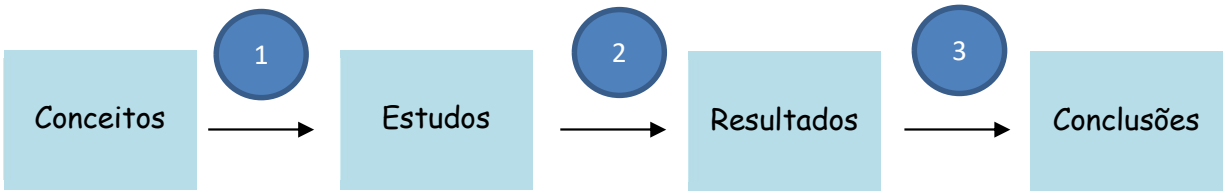


Figura 6. A Revisão Sistemática - Abordagem de 3 passos (Adaptado de Biolchini et al. 2005)

Biolchini e colaboradores (2005) destacam que toda pesquisa começa com a revisão da literatura e que os estudos primários auxiliam no estabelecimento de critérios de inclusão e exclusão que serão utilizados neste processo. Os autores também salientam que o processo de revisão sistemática não é sequencial (figura 7), nele há interação entre as etapas e que ele deve ser planejado antes da execução com registro de todo o processo incluindo os resultados intermediários que surgirem.

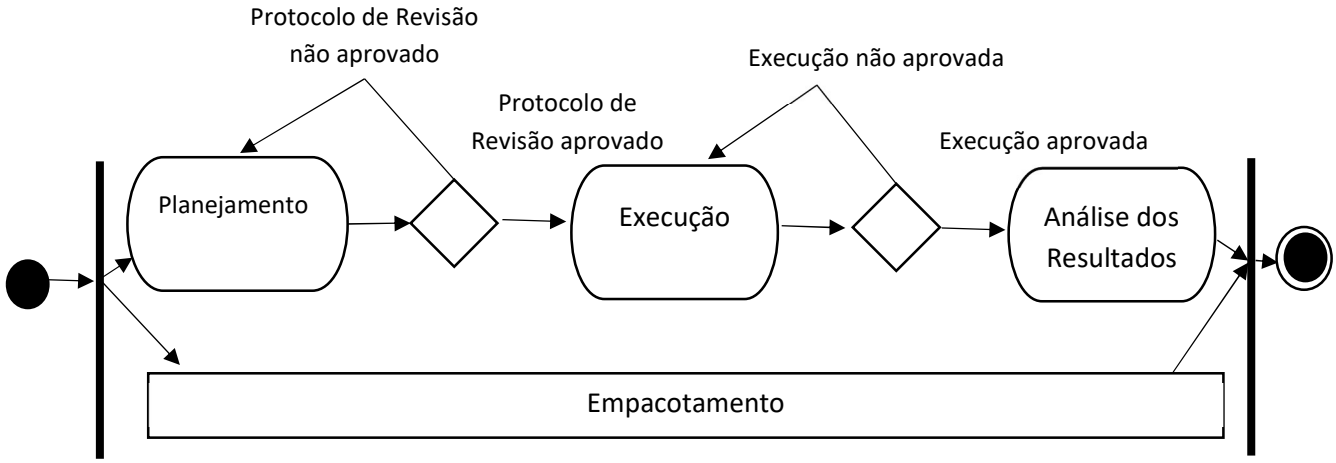


Figura 7. Processo de Revisão Sistemática (Adaptado de Biolchini et al. 2005)

O Planejamento da Revisão, compreende a formulação da pergunta geral que norteia a pesquisa, identificando o foco, a qualidade e a amplitude da questão, ou seja, a definição do problema, pergunta ou perguntas que envolvem o alvo da pesquisa, as palavras-chaves e sinônimos, que neste caso, foram retirados dos estudos realizados previamente. Nesta etapa, também é feita a descrição do que está sendo observado (intervenção), quais mecanismos de controle, os efeitos e medidas que serão usados para verificar os resultados, qual a população que será observada na intervenção, em que áreas a pesquisa será aplicada e onde é definido o projeto experimental.

Ademais, faz parte desta fase, a seleção das fontes, o estabelecimento do critério de inclusão e exclusão, qual o idioma dos estudos primários que serão obtidos, a identificação das fontes, quais os métodos de pesquisa e qual a *string* de busca que será utilizada. Recomenda-se a avaliação das fontes depois da seleção de acordo com o critério de inclusão e exclusão e, checagem das referências para confirmar a confiabilidade das fontes selecionadas (que pode ser feita por especialistas).

O último estágio desta fase é a seleção dos estudos; aqui se responde à pergunta sobre quais estudos serão selecionados através da definição dos critérios de inclusão e exclusão dos estudos, quais os tipos de estudos serão usados e, finalmente, qual será o procedimento de análise desses trabalhos. Ao término desta fase é feita uma avaliação de todo o planejamento antes da execução para verificar possíveis erros.

A fase seguinte é a de Execução da Revisão, onde é feito o registro dos estudos primários obtidos, relatando os trabalhos e a avaliação que foi feita sobre eles. Ao término da seleção é feita a extração das informações, estabelecendo novos critérios de inclusão e exclusão, além da forma como a informação será apresentada e quais resultados serão extraídos.

A extração pode ser objetiva, através da identificação do estudo, da metodologia, dos resultados e dos problemas apresentados pelo estudo; e/ou subjetiva onde pode ser anotada informações adicionais sobre os autores e as impressões pessoais sobre o trabalho (este tipo de extração foi realizada durante a classificação dos estudos de acordo com os critérios do Quadro 2)

O modelo original prevê uma seção para resolução de divergências entre revisores, o que não se aplica nesta pesquisa. Também é recomendado realizar uma avaliação após esta etapa.

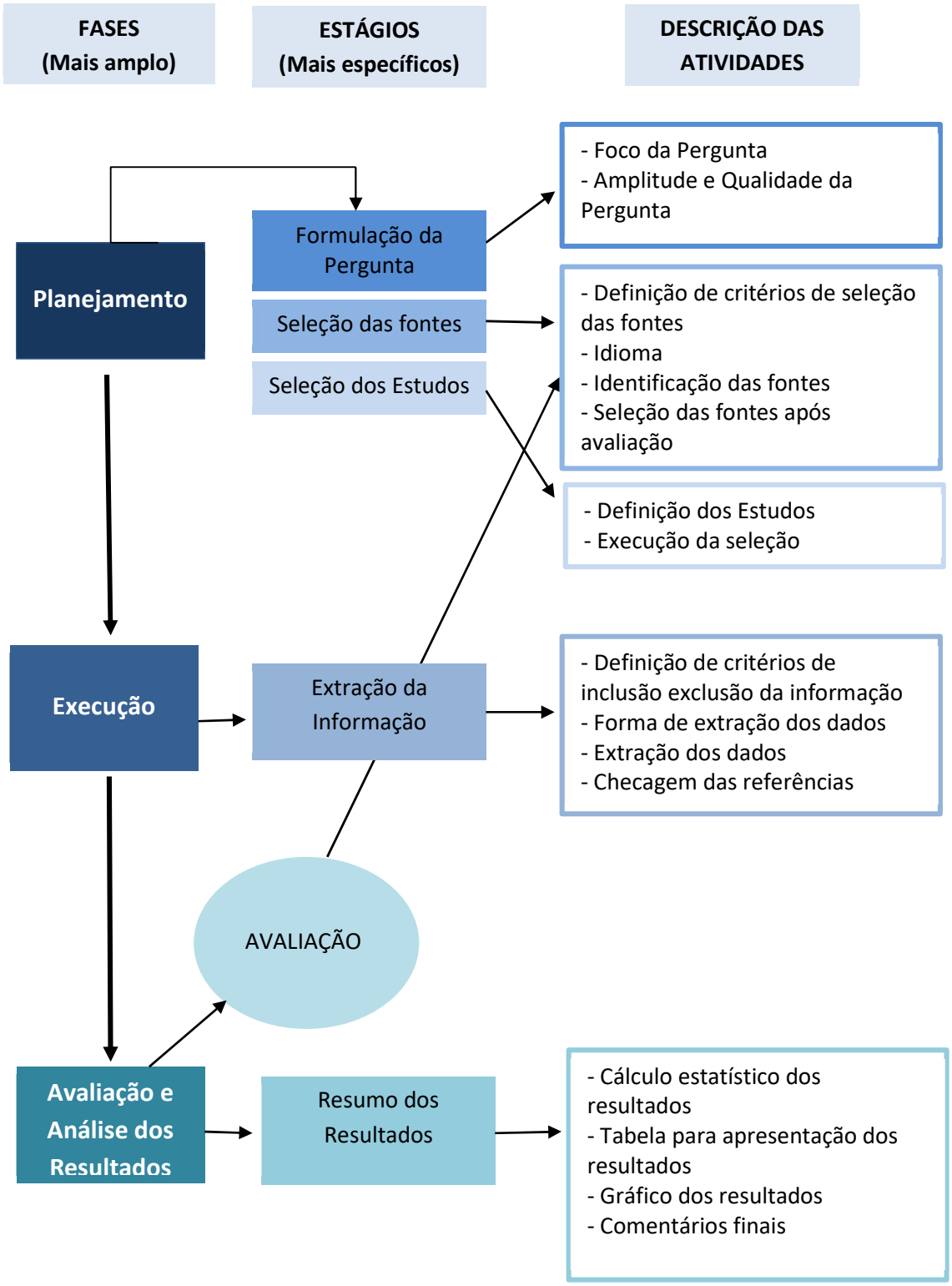


Figura 8 – Protocolo de Revisão (elaborado pela autora)

A última etapa do Protocolo é a Análise dos Resultados, onde os resultados são resumidos e analisados em conformidade com o que foi estabelecido na fase de planejamento.

4 RESULTADOS E DISCUSSÃO

Aqui serão apresentados os resultados da Revisão Sistemática e da análise dos textos, conforme a metodologia proposta no item 3.1.

4.1 Resultado da Revisão

Nesta seção apresentamos os resultados encontrados com base no Protocolo de Revisão (Apêndice A) anteriormente estabelecido. A Revisão Sistemática foi conduzida no sentido de identificar estudos sobre o fenômeno da evasão que utilizam a abordagem de *Data Analytics* para mitigar os efeitos deste fenômeno no Ensino Superior, nos cursos de graduação (Bacharel e Licenciatura) realizados na modalidade presencial. É através desta revisão que serão coletados os dados para serem analisados através da Metodologia da *Design Science Research*.

Optou-se por fontes disponíveis na Web e seguindo a recomendação de Neiva e Silva (2016), as bases escolhidas para execução da busca foram: Web of Science, Scopus, Science Direct, IEEE Explore, Compendex, ACM Digital Library e Springer Link. Além destas, foi feita uma busca na Plos One, em virtude da pesquisa narrativa realizada no início deste estudo ter apontado para publicações sobre o tema disponíveis nesta base.

A condução da pesquisa teve como referência palavras-chaves baseadas na literatura sobre predição da evasão apresentada no Capítulo 2. Sendo assim a *string* adotada foi (Dropout) AND (“Data Analytics” OR “Data Mining” OR “Machine Learning”) AND (“Higher Education” OR College OR University) AND (Retention) AND (Preventing OR Predicting) AND NOT (MOOC OR e-Learning). No entanto, nem todas as bases aceitaram a expressão “AND NOT (MOOC OR e-Learning)” que foi adotada para evitar artigos sobre Ensino à Distância, uma vez que o foco deste estudo é a modalidade presencial.

O objetivo foi estabelecer uma única *string* de busca que englobasse vários termos de forma a viabilizar a pesquisa para uma pessoa, em virtude do tempo, e que pudesse ser utilizada em todas as bases selecionadas de forma padronizada. Algumas bases apresentaram limitações quanto ao uso de operadores booleanos, não sendo possível utilizar todas as palavras encontradas nos estudos sobre evasão e retenção de alunos. Por isso, optou-se pelo uso dos termos “Dropout” e “Retention”, utilizados com maior frequência na literatura. É importante ressaltar que segundo o que foi observado nos estudos sobre o fenômeno da evasão, essas duas palavras são muito representativas desse universo e convém que sejam utilizadas nos termos de

busca, pois o combate ao abandono escolar tanto pode ser estudado do ponto de vista de se evitar o abandono quanto dos esforços para reter o aluno na instituição ou sistema.

Na Scopus foi possível utilizar a string de busca completa e realizar a pesquisa somente nos títulos e no abstract. Compendex, Web of Science e Science Direct (aplicado apenas a artigos de pesquisa) foi necessário excluir os termos "AND NOT (MOOC OR e-Learning)". Na base Springer Link, a pesquisa foi feita em dois grupos – Artigos e Capítulos, utilizando todos os termos. A busca foi realizada apenas na área de Ciência da Computação porque a análise inicial demonstrou que em outras áreas o retorno não era satisfatório. No primeiro grupo restringiu-se a busca em "artigos" e no segundo grupo à "Conferência". Na Plos One o retorno foi de 8 artigos, utilizando todos os termos da busca, no entanto, apenas '1' foi aproveitado; o mesmo aproveitamento da ACM Digital Library de um total de 2 artigos.

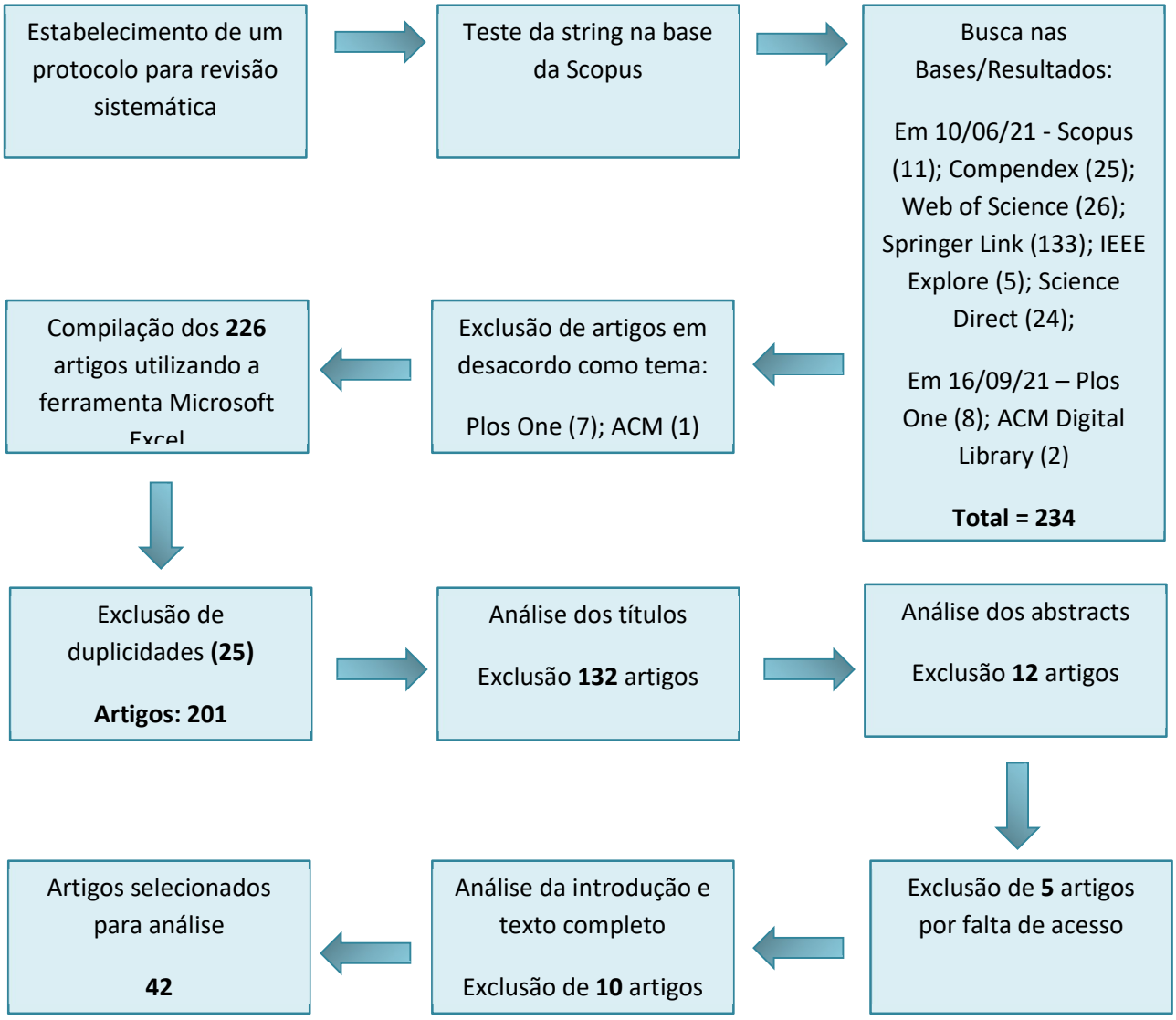


Figura 9. Fluxograma do processo de revisão (elaborado pela autora)

A busca encontrou 226 artigos (tabela 1) e após a exclusão de duplicidades, análise dos títulos, abstract e introdução, permaneceram 57 artigos para análise. Grande parte das exclusões realizadas nesta etapa foi porque os estudos não tratavam de evasão de alunos ou abordavam o ensino à distância. O fluxograma deste processo é demonstrado na figura 9. O aproveitamento nesta etapa foi de aproximadamente 25%.

Tabela 1: Número de artigos por Base

Base	**Nº artigos**
ACM Digital Library	01
Compendex	25
IEEE Explore	05
Plos One	01
Science Direct	24
Scopus	11
Springer Link Artigos	54
Springer Link Capítulos	79
Web of Science	26
Total	**226**

Fonte: Elaborado pela autora

Após a leitura completa dos textos, 42 foram selecionados para análise da Metodologia de *Data Analytics* de acordo a DSR. Nesta etapa foram excluídos cinco artigos pela falta de acesso e devido a conteúdos cujo foco fugia do escopo desta pesquisa (*e.g.* abordavam o desempenho sem fazer associação com a evasão ou com a retenção dos alunos; abordavam outros níveis de ensino).

Estudos que abordavam *Learning Analytics* foram excluídos do processo de análise porque estavam aplicados no contexto do Ensino a Distância, no entanto, optou-se por permanecer com a pesquisa desenvolvida por Gray e Perkins (2018) porque os autores não deixaram claro a modalidade de ensino e porque a coleta de dados/alimentação do sistema foi feita manualmente e não coletada automaticamente através da interação do aluno com o sistema.

A ferramenta Microsoft Excel foi utilizada para organização, seleção e análise dos textos.

Estudos sobre Revisão Sistemática recomendam adotar critérios de inclusão e exclusão de artigos com base na sua relevância para o mundo acadêmico (KITCHINHAM et al., 2004). Uma destas análises considera o Fator de Impacto da fonte onde o artigo é publicado. Apesar desta verificação ter sido realizada neste estudo *a priori* ela foi desconsiderada, pois muitas

publicações apesar de serem consideradas relevantes para a pesquisa não eram indexadas na JCR[4]. O quadro abaixo mostra a relação das fontes dos estudos (em ordem alfabética):

Quadro 3: Lista de Fontes

Continua

Fonte	Autor(ers)
2021 IEEE International Conference on Automation/XXIV Congress of the Chilean Association of Automatic Control (ICA-ACCA)	Peralta et al. (2021)
ACM International Conference Proceeding Series	Chen; Johri; Rangwala (2018)
Advances in Intelligent Systems and Computing	Bilquise; Abdallah; Kobbaey (2019) Ullah et al (2019)
Advances in Soft Computing	Aguilar-Gonzalez e Palafox (2019)
Applications of Computational Intelligence	Pérez; Castellanos; Correal (2018)
Artificial Intelligence in Education	Zhang; Rangwala (2018)
Computers & Education,	Gray e Perkins. (2019)
Computers & Electrical Engineering,	Zeineddine; Braendle; Farah (2021)
Decision Support Systems	Delen (2010); Maldonado et al. (2021); Olaya et al. (2019)
EDM 2019 - Proceedings of the 12th International Conference on Educational Data Mining	Jayaraman; Gerber; Garcia (2019)
Education and Information Technologies	Al-Sudani, Palaniappan (2019); Gil et al. (2021)
Entropy	Palacios et al. (2021)
European Journal of Higher Education	Kemper; Vorhoff; Wigger (2020)
Expert Systems with Applications,	Nandeshwar; Menzies; Nelson (2011)
Frontiers in Education	Kilian; Loose; Kelava (2020)
Future Generation Computer Systems	Kuzilek; Zdrahal; Fuglik (2021)
International Journal of Educational Technology in Higher Education	Tsai,et al. (2020)
International Journal of Machine Learning and Cybernetics	Iam-On e Boongoen (2017)
Journal of College Student Retention-research Theory & Practice	Huo et al. (2020)
Journal of Physics: Conference Series	Wan Yaacob et al (2020)
Journal of Technology and Science Education	Alvarez; Callejas; Griol (2020)
Plos One	Rovira; Puertas; Igual (2017)
Procedia Computer Science	Viloria et al. (2019)
Procedia Manufacturing	Cardona e Cudney (2019)
Proceedings - 2019 Brazilian Conference on Intelligent Systems, BRACIS 2019	Silva et al. (2019)
Proceedings - IEEE 19th International Conference on Advanced Learning Technologies, ICALT 2019	Santos et al (2019)

[4] JCR – Journal Citation Reports

Quadro 3: Lista de Fontes

Conclusão

Proceedings - 2019 Brazilian Conference on Intelligent Systems, BRACIS 2019	Silva et al. (2019)
Proceedings - IEEE 19th International Conference on Advanced Learning Technologies, ICALT 2019	Santos et al (2019)
Proceedings - International Conference of the Chilean Computer Science Society, SCCC	Alfredo (2018)
	Bello et al. (2020)
Proceedings of the 2020 IISE Annual Conference	Patterson et al. 2020)
Proceedings of the 2nd International Conference on Inventive Systems and Control, ICISC 2018	Hegde e Prageeth (2018)
Proceedings of the 8th Research in Engineering Education Symposium, REES 2019 - Making Connections	Acero; Achury; Morales Piñero (2019)
Research in Higher Education	Beaulac e Rosenthal (2019)
Socio-Economic Planning Sciences,	Contini e Salza (2020)
Soft Computing in Data Science	Ahmad Tarmizi et al. (2019)
Studies in Educational Evaluation,	Berka e Marek (2021)
Technology and Innovation in Learning, Teaching and Education	Tampakas et al. (2019)
Technology Trends	Vila et al. (2019)

Fonte: Elaborado pela autora.

Não foi aplicado critério de busca por período. O Gráfico 1 demonstra a quantidade de publicações por ano dos textos selecionados após os critérios de inclusão e exclusão. O gráfico sugere que os estudos envolvendo abordagens de *Data Analytics* ou técnicas de mineração de dados e de aprendizagem de máquinas aplicadas ao fenômeno da evasão no ensino superior é uma prática recente.

Fonte: Elaborado pela autora.

O mapa abaixo demonstra o quanto o fenômeno da evasão é uma preocupação mundial. Nota-se que a maioria dos estudos foram conduzidos nos últimos anos em países do continente americano com destaque para Estados Unidos com 16 publicações e Chile com 12. O Brasil contribui com dois artigos: Silva et al. (2019) e Santos et al. (2019).

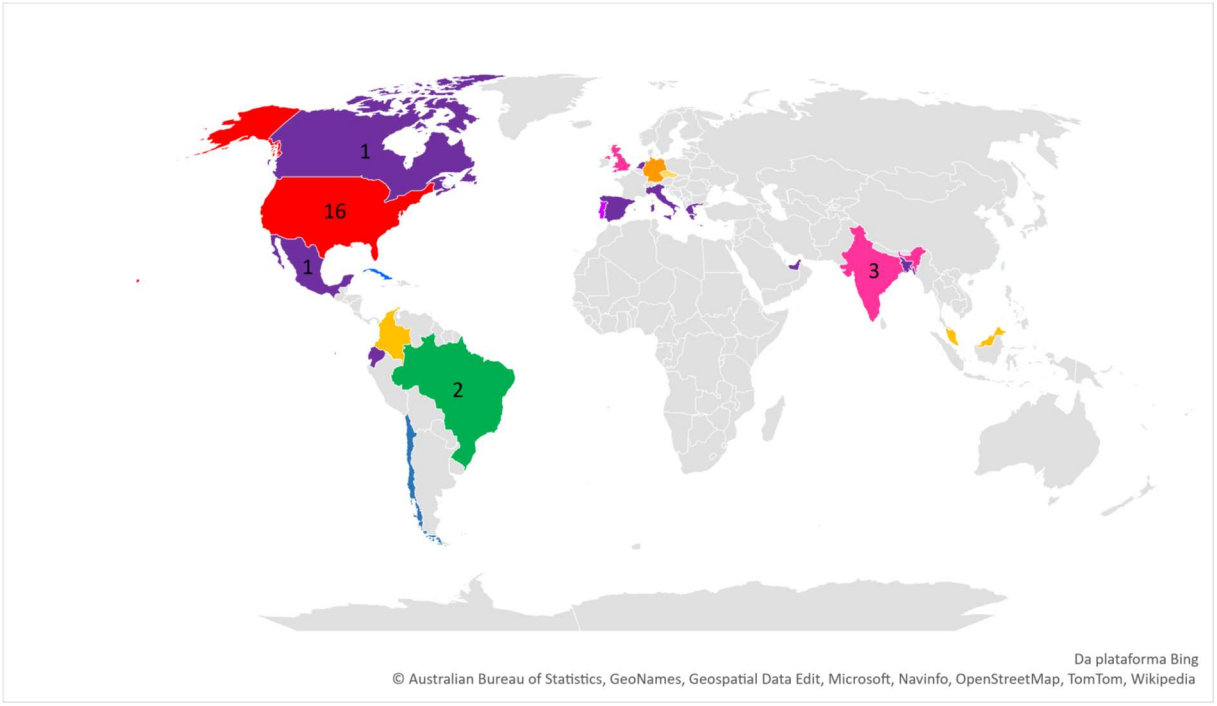
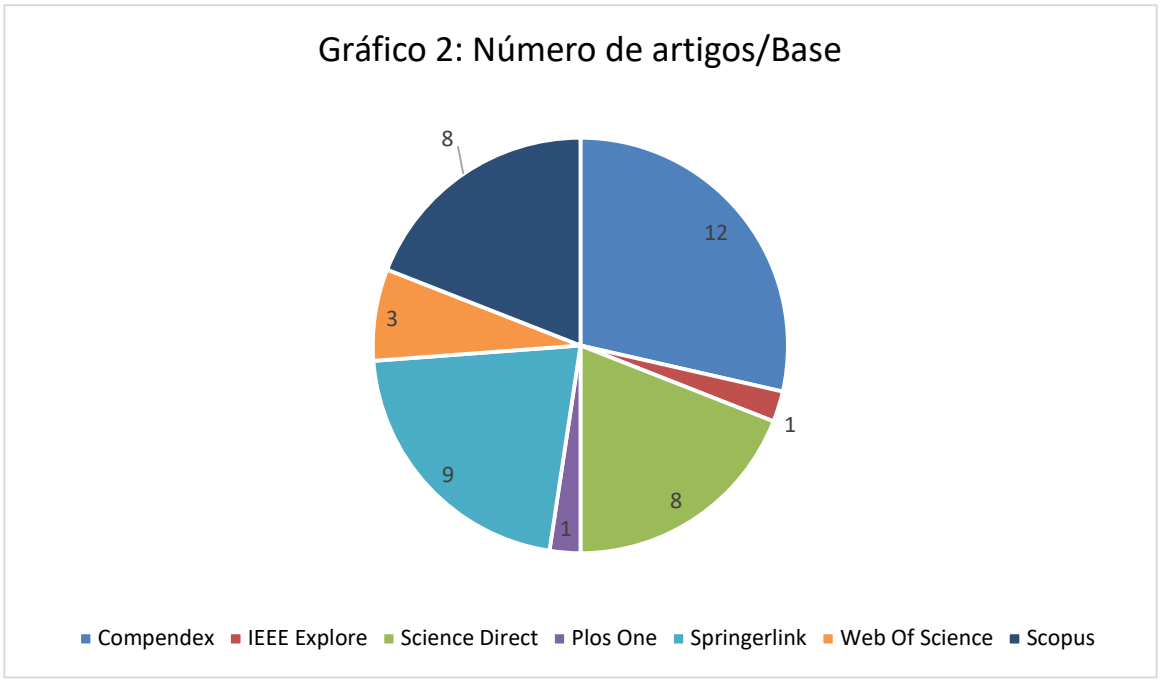


Figura 10: Contribuição dos estudos por país – elaborado pela autora.

O gráfico de pizza (Gráfico 2) demonstra a contribuição das bases para a etapa de análise dos textos.



Fonte: Elaborado pela autora.

Como já citado anteriormente o fenômeno da evasão de alunos pode ser abordado sob muitos aspectos. Para demonstrar, a partir das palavras-chaves dos estudos selecionados foi construída uma Nuvem de Palavras (*Word Cloud* em inglês) que possibilita a visualização do foco dos estudos. O resultado corrobora com a pesquisa bibliográfica inicial que apontou as palavras dropout, retention, machine learning e data mining como termos de busca.

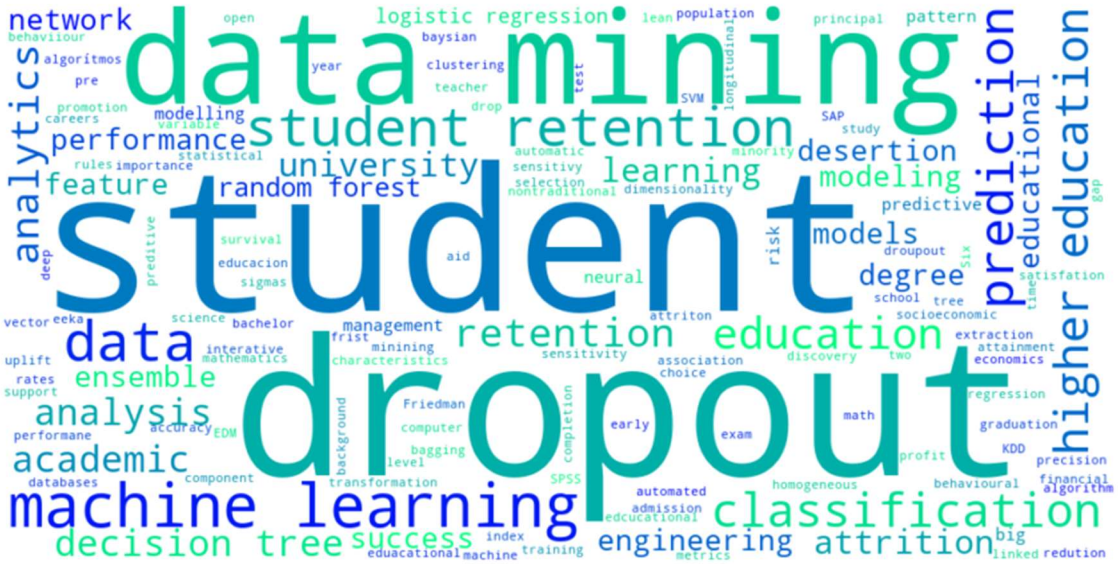


Figura 11: Nuvem de Palavras - elaborado pela autora.

4.2 Resultados da aplicação da estrutura da *Design Science Research*

Como resultado da Revisão Sistemática, 42 artigos foram selecionados para a análise metodológica (Quadro 4). Para realizar a análise da abordagem de *Data Analytics* utilizada pelos autores, a coleta de dados foi baseada em uma estrutura da *Design Science Research* (Figura 5) sobre os seguintes componentes: Enquadramento do Problema, Teoria, Modelagem ou "Engenharia de Modelo" e Protocolo Experimental de Avaliação.

A execução desta etapa contou com o auxílio da ferramenta Microsoft Excel. Um arquivo foi criado, com uma planilha contendo os resultados (linhas) e com colunas representando cada componente a ser investigado. Os resultados são descritos nas seções seguintes:

Quadro 4: Artigos Selecionados

Continua

Título	Autor(res)
A Case Study on Student Attrition Prediction in Higher Education Using Data Mining Techniques	Delen (2010)
A causal modelling for desertion and graduation prediction using Bayesian networks: a Chilean case	Nandeshwar; Menzies; Nelson (2011)
A comparative analysis of machine learning techniques for student retention management	Sarker; Tiropanis; Davis (2014)
A data-driven approach to predict first-year students' academic success in higher education institutions	Rovira; Puertas; Igual (2017)
Bachelor's degree student dropouts: Who tend to stay and who tend to leave?	Iam-On e Boongoen (2017)
Comparative Analysis of Prediction Techniques to Determine Student Dropout: Logistic Regression vs Decision Trees	Alfredo (2018)
Data-driven system to predict academic grades and dropout	Zhang e Rangwala (2018)
Detection of Desertion Patterns in University Students Using Data Mining Techniques: A Case Study	Hegde e Prageeth (2018)
Early Identification of At-Risk Students Using Iterative Logistic Regression	Pérez; Castellanos; Correal (2018)
Enhancing prediction of student success: Automated machine learning approach	Chen; Johri; Rangwala (2018)
Ensemble regression models applied to dropout in higher education	Ahmad Tarmizi et al. (2019)
Higher education student dropout prediction and analysis through educational data mining	Vila et al. (2019)
Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings	Silva et al. (2019)
Integrating lean six sigma and data analytics to improve student retention	Viloria et al. (2019)
Integration of data technology for analyzing university dropout	Ullah et al. (2019)
Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in chile	Cardona e Cudney (2019)
Learning patterns of university student retention	Al-Sudani e Palaniappan (2019)
Linked Data, Data Mining and External Open Data for Better Prediction of at-risk Students	Beaulac e Rosenthal (2019)
Precision education with statistical learning and deep learning: a case study in Taiwan	Aguilar-Gonzalez e Palafox (2019)
Predicting Computer Engineering Students' Dropout In Cuban Higher Education With Pre-Enrollment and Early Performance Data	Tampakas et al (2019)
Predicting Dropout for Nontraditional Undergraduate Students: A Machine Learning Approach	Santos et al. (2019)
Predicting factors of students dissatisfaction for retention	Jayaraman; Gerber; Garcia (2019)
Predicting Math Student Success in the Initial Phase of College With Sparse Information Using Approaches From Statistical Learning	Acero; Achury; Morales Piñero (2019)
Predicting Student Drop-Out in Higher Institution Using Data Mining Techniques	Gray e Perkins (2019)
Predicting Student Drop-Out Rates Using Data Mining Techniques: A Case Study	Patterson et al. (2020)
Predicting student dropout: A machine learning approach	Tsai et al. (2020)
Predicting Student Retention Among a Homogeneous Population Using Data Mining	Alvarez; Callejas; Griol (2020)
Predicting Student Retention Using Support Vector Machines	Huo, H. D. et al (2020)

Quadro 4: Artigos Selecionados

Conclusão

Predicting students' final degree classification using an extended profile	Kilian; Loose; Kelava (2020)
Predicting University Students' Academic Success and Major Using Random Forests	Wan Yaacob et al (2020)
Prediction of Student Attrition Using Machine Learning	Kemper.; Vorhoff; Wigger (2020)
Prediction of Student´s Graduation Time Using a Two-Level Classification Algorithm	Bilquise; Abdallah; Kobbaey (2019)
Redefining profit metrics for boosting student retention in higher education	Contini e Salza (2020)
Running out of STEM: A comparative study across STEM majors of college students At-Risk of dropping out early	Olaya et al. (2019)
Student success prediction using student exam behaviour	Bello et al. (2020)
Supervised learning in the context of educational data mining to avoid university students dropout	Peralta et al. (2021)
Supporting minority student success by using machine learning to identify at-risk students	Gil et al. (2021)
Too few university graduates. Inclusiveness and effectiveness of the Italian higher education system	Berka e Marek (2021)
University dropout: A prediction model for an engineering program in bogota, Colombia	Zeineddine; Braendle; Farah (2021)
Uplift Modeling for preventing student dropout in higher education	Palacios et al. (2021)
Using machine learning methods to identify significant variables for the prediction of first-year Informatics Engineering students dropout	Maldonado et al. (2021)
Utilizing early engagement and machine learning to predict student outcomes	Kuzilek; Zdrahal; Fuglik (2021)

Fonte: Elaborado pela autora.

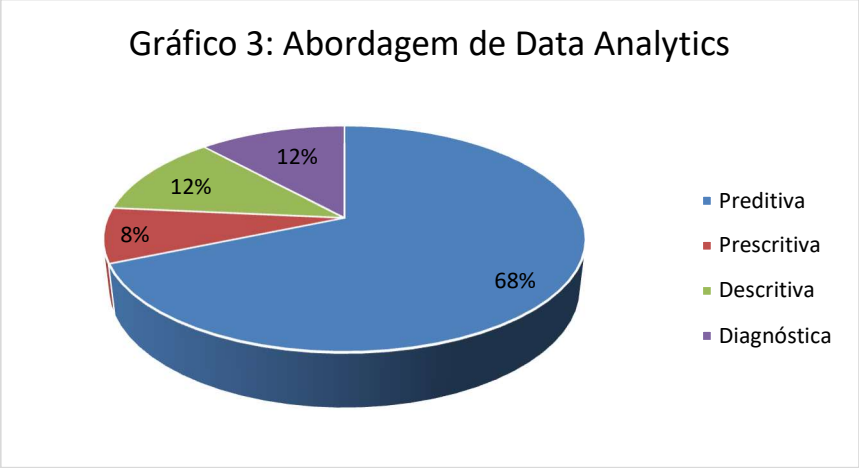
4.2.1. Enquadramento do Problema

O primeiro componente da análise diz respeito a como os artigos realizam o enquadramento do problema.

Quanto a tarefa do modelo:

A pesquisa encontrou trabalhos que abordam as quatro tarefas de *Data Analytics*: Descritiva, Diagnóstica, Preditiva e Prescritiva ou Prognóstica. A análise Descritiva é aquela que trabalha com fatos, busca descrever a situação da evasão de alunos naquele momento, fazendo uso de estatística como média e desvio padrão para consolidação das informações. A análise Diagnóstica são estudos que se preocupam em entender as causas do fenômeno que está sendo observado. Os estudos com análise Preditiva formam a maioria das observações como é possível observar no gráfico 3, onde o objetivo é tentar prever a probabilidade de alunos evadirem de suas instituições, cursos ou das disciplinas nas quais foram matriculados. Por

último, a análise Prescritiva ou Prognóstica, que é aquela que em virtude das informações obtidas através das demais análises é possível propor uma ação que evite ou diminua os riscos de os alunos evadirem. Nesse sentido é importante lembrar que algumas análises são feitas em conjunto e que alguns estudos apresentam mais de um tipo de análise.



Fonte: Elaborado pela autora.

O quadro a seguir apresenta a distribuição dos autores segundo cada tarefa de *Data Analytics*.

Quadro 5: Abordagens de *Data Analytics*

Descritiva	Nandeshwar; Menzies; Nelson (2011); Jayaraman; Gerber; Garcia (2019); Bello et al. (2020); Bilquise; Contini; Salza (2020)
Diagnóstica	Delen (2010); Hegde e Prageeth (2018); Santos et al. (2019); Jayaraman; Gerber; Garcia (2019); Ullah et al. (2019); Viloria et al. (2019); Patterson et al. (2020); Tsai et al. (2020); Berka e Marek (2021)
Preditiva	Delen (2010); Sarker; Tiropanis; Davis (2014); Iam-On e Boongoen (2017); Rovira; Puertas; Igual (2017); Alfredo (2018); Chen; Johri; Rangwala (2018); Zhang e Rangwala (2018); Pérez; Castellanos; Correal (2018); Acero; Achury; Morales Piñero (2019); Aguilar-Gonzalez, S.; Palafox, L. (2019); Ahmad Tarmizi, S.S. et al. (2019); Al-Sudani e Palaniappan (2019); Beaulac e Rosenthal (2019); Cardona e Cudney (2019); Gray e Perkins (2019); Jayaraman; Gerber; Garcia (2019); Silva et al. (2019); Tampakas et al (2019); Ullah et al (2019); Vila et al. (2019); Alvarez; Callejas; Griol (2020); Huo et al (2020); Kilian; Loose; Kelava. (2020); Patterson et al. (2020); Tsai et al. (2020); Wan Yaacob et al. (2020); Berka e Marek (2021); Gil et al. (2021); Kuzilek; Zdrahal; Fuglik (2021); Maldonado et al. (2021); Palacios et al. (2021); Peralta et al. (2021); Zeineddine; Braendle; Farah (2021)
Prescritiva ou Prognóstica	Rovira; Puertas; Igual (2017); Kilian; Loose; Kelava (2020); Olaya et al. (2019); Maldonado et al. (2021)

Fonte: Elaborado pela autora.

Entre os objetivos da análise descritiva observados nos artigos está identificar variáveis que influenciam na evasão de alunos ou encontrar padrões na retenção de alunos (BELLO et al., 2020). Jayaraman, Gerber e Garcia (2019) fazem uma análise em grupos específicos (grupos minoritários: alunos nativos e transferidos de uma instituição tentando identificar as características de cada grupo, enquanto Contini e Salza (2020) pretendem a partir da análise descritiva propor um modelo de estimação de risco.

A análise diagnóstica foi realizada com vistas a identificar quais variáveis são mais importantes nos modelos desenvolvidos para saber se o aluno irá abandonar ou não; ou quais fatores influenciaram na falha acadêmica, quais atributos e as razões que levam os alunos a abandonar antecipadamente, qual o perfil do aluno universitário em risco de abandonar. Ao identificar as características de cada grupo, Jayaraman, Gerber e Garcia (2019) analisam qual característica mais contribui para a evasão em cada grupo (nativos e transferidos). Patterson e seus colaboradores (2020) trabalham para encontrar tendência nos dados, com foco na retenção, para inferir quais causas são mais importantes para distinguir entre os alunos que serão retidos e os que não serão.

Quando se trata da análise preditiva os estudos são muito diversos:

- desenvolver modelos para identificar corretamente os calouros com mais probabilidade de abandonar depois do primeiro ano;
- prever alunos em risco de falhar ou abandonar em semestres futuros;
- prever o desempenho ou o sucesso do aluno;
- prever se o aluno se matriculará no segundo e terceiro ano;
- prever a nota dos alunos;
- comparar modelos preditivos;
- identificar em que momento o abandono acontece;
- detectar a evasão de alunos em cursos específicos ou disciplinas (e.g. Engenharia, Ciência da Computação, disciplinas STEM – Ciência, Tecnologia, Engenharia e Matemática);
- prever se o aluno irá concluir o seu programa/em qual área o aluno concluirá.
- prever antecipadamente alunos que se beneficiariam de uma possível intervenção
- prever a evasão de alunos de grupos minoritários
- prever o momento da graduação do aluno colocando ênfase na identificação de quem está propenso a não concluir os seus estudos em seis anos ou abandonar;
- prever a satisfação do aluno e sua relação com a retenção;

- identificar fatores que predizem a evasão;
- prever resultados de retenções futuras;
- prever o sucesso acadêmico dos alunos de primeiro ano;
- prever através do comportamento no exame do aluno o sucesso no primeiro ano acadêmico) e identificar se o comportamento no exame é um bom fator preditivo;
- identificar quem irá abandonar ou não - a partir da avaliação dos classificadores de abandono através de uma perspectiva baseada no lucro;

Cabe ressaltar que estes objetivos são muito mais complexos porque cada estudo se utiliza de teoria, variáveis e modelagens específicas para seus fins, conforme veremos nas seções seguintes.

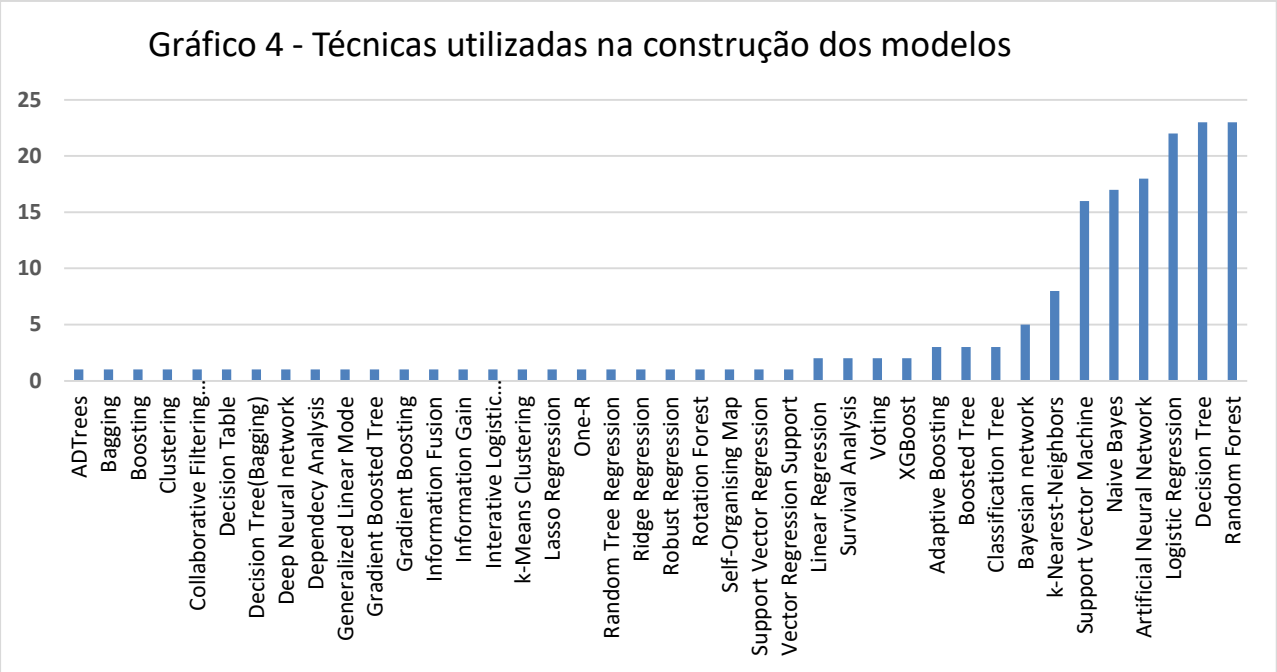
Diversos artigos fazem recomendações em suas conclusões ou considerações finais sobre formas de evitar a evasão de alunos, no entanto apenas quatro artigos possuem foco na abordagem prescritiva, pois a partir de análises preditivas identificam grupos que estejam suscetíveis a um tratamento/intervenção (*e.g*. tutoria, programa de retenção).

Quanto a escolha do modelo:

Segundo Elragal e Haddara (2019) para a construção de um artefato podem ser aplicados centenas de algoritmos diferentes a um conjunto de dados na tentativa de escolher um modelo. Durante a análise foram encontrados 38 artigos que utilizam a classificação para tratar o problema da evasão. Dentre as técnicas utilizadas para classificar a Árvore de Decisão e Regressão Logística são as de uso mais frequente na construção de modelos individuais, seguidas por de Redes Neurais Artificiais e Máquina de Suporte de Vetores.

O algoritmo K-Vizinho mais próximo (em inglês, K-Nearest-Neighbors - K-NN) foi utilizado nos estudos de classificação feitos por Wan Yaacob et al., (2020) e Santos et al. (2019); empregado na fase de pré-processamento dos dados para fazer a redução de dimensionalidade por Iam-On, e Boongoen (2017) e no estudo sobre AutoML (Autoaprendizagem de Máquina) onde foi utilizado como técnica de Agrupamento (em inglês – *Clustering*) pelos autores Zeineddine, Braendle, Farah (2021).

O gráfico 4 apresenta as técnicas utilizadas nas pesquisas.



Fonte: Elaborado pela autora.

A tabela 2 apresenta as técnicas utilizadas nos estudos de acordo com o esquema sobre Aprendizagem de Máquina apresentado anteriormente (figura 3).

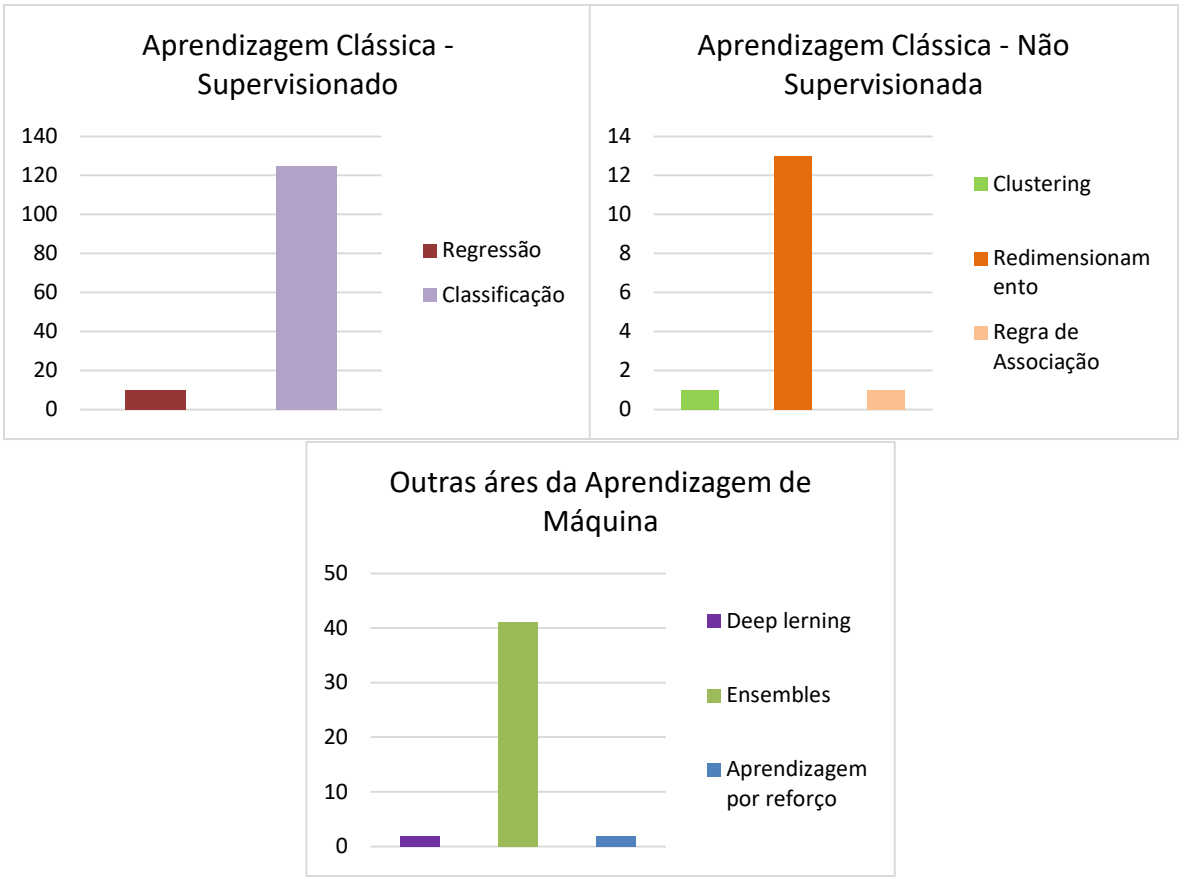
Tabela 2 – Quantidade de Técnicas de Aprendizagem de Máquina

Aprendizagem clássica	
Supervisionado	
Classificação	125
Regressão	10
Não supervisionado	
Agrupamento ou Redimensionamento	14
Regras de Associação	01
Rede Neural Profunda	02
Métodos *Ensembles* (algoritmos)	41
Aprendizagem por Reforço	02

Fonte: Elaborado pela autora

É importante esclarecer que a Aprendizagem por Reforço foi utilizada no estudo realizado por Olaya et. Al (2019). Ela foi utilizada como técnica de pré-processamento da *uplif*.

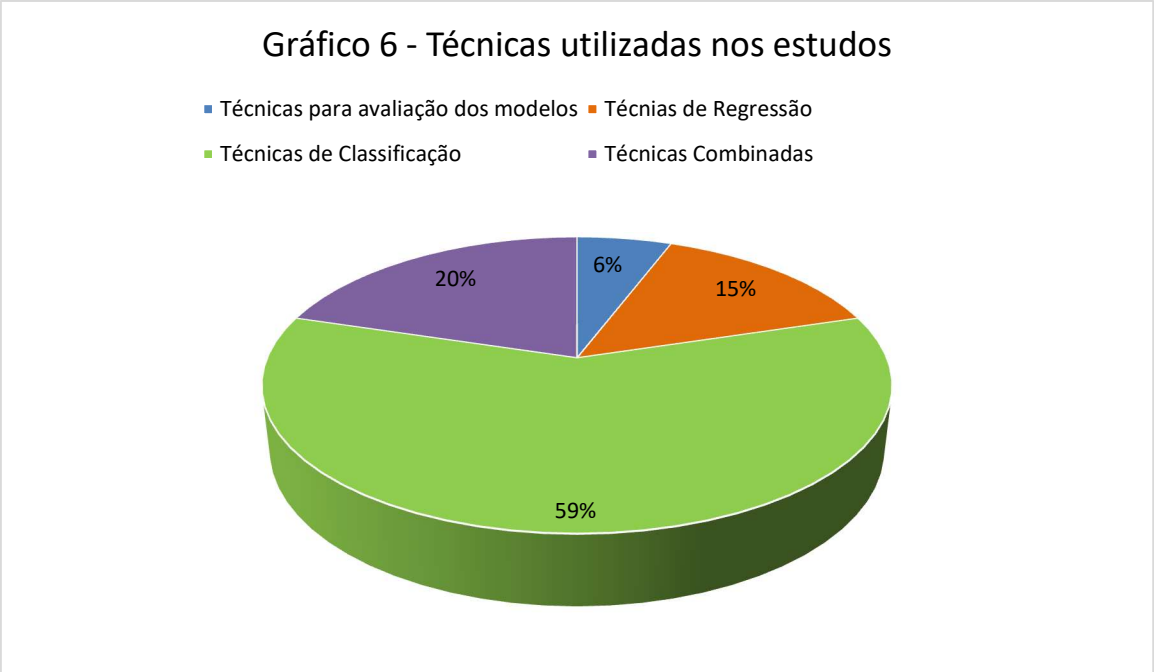
Gráfico 5 – Utilização das Técnicas de Aprendizagem de Máquina



Fonte: Elaborado pela autora.

Através do gráfico 5 podemos visualizar que as técnicas de aprendizagem clássica são as mais utilizadas. Nos estudos sobre o fenômeno da evasão predominam o uso de técnicas de classificação.

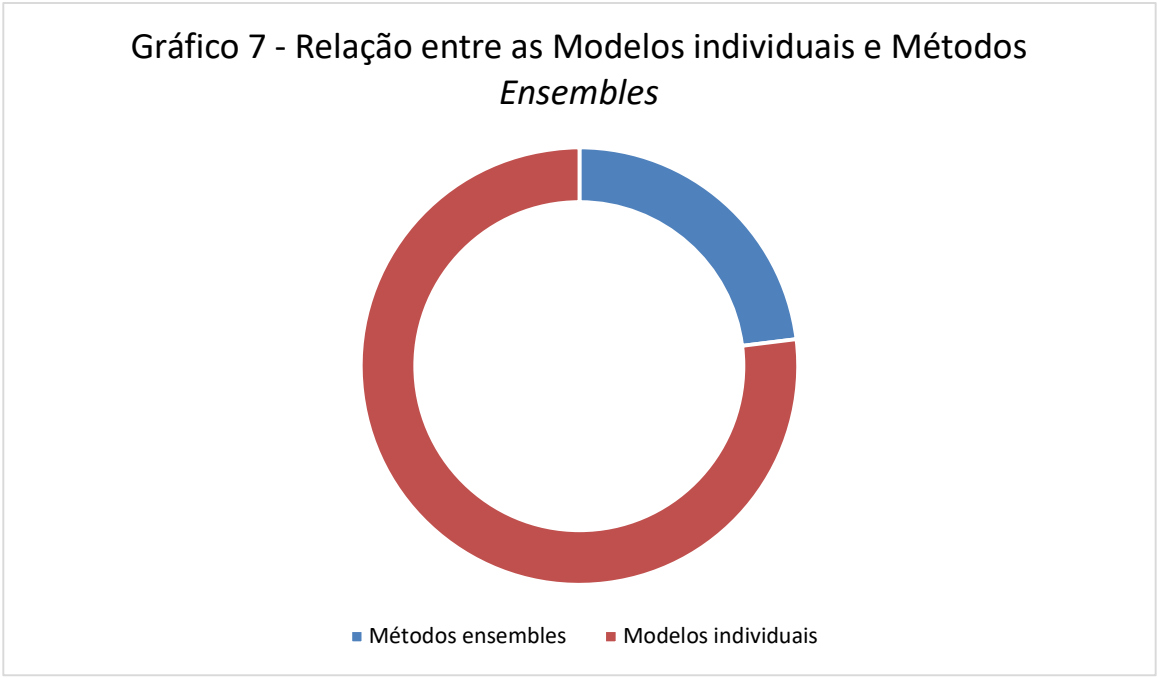
O gráfico 6 apresenta o enquadramento das tarefas conforme apresentado nos estudos. Isso porque o uso de Rede Neural Profunda, Regras de Associação e das técnicas de *Clustering* aparecerem de forma distinta na literatura sobre Aprendizagem de Máquina, no entanto, elas foram utilizadas nos estudos como classificadores. Além disso alguns estudos utilizaram uma combinação de técnicas: métodos ensembles combinados com técnicas de regressão e métodos de ensemble Clustering combinados com técnicas de classificação, apresentadas no gráfico com Técnicas Combinadas.



Fonte: Elaborado pela autora.

Considerando os termos utilizados pelos autores foram enumeradas o uso de 54 técnicas. No entanto, foi feito um agrupamento para unificar as técnicas cujos nomes apareciam de forma diferente, por exemplo *Neural Networks* e *Artificial Neural Networks*, *AdaBoosting* e *Adaptative Boosting*, entre outras, permanecendo o registro de 41 técnicas.

Observou-se nos estudos, a utilização, tanto de modelos individuais quanto de Métodos *Ensembles*, e ao visualizarmos o gráfico 7 percebemos que as primeiras se sobrepõem em termos de quantidade.



Fonte: Elaborado pela autora.

Foram observados três estudos que tratam o fenômeno da evasão no Ensino Superior como um problema de Regressão. Silva et al. (2019) combinam técnicas de regressão com modelos ensembles, além de fazer uso de bootstrap para obter uma amostra do conjunto de dados. O estudo realizado pelos autores Kilian, Losse e Kelava (2020) é um estudo de classificação, no entanto eles utilizam a técnica de regressão Elastic Net para prever algumas variáveis numéricas. As técnicas de Regressão Linear (do inglês, *Linear Regression*) e Suporte a Regressão Vetorial (do inglês, *Suport Vector Regression*) são utilizadas em dois estudos, como pode ser observado no quadro 6.

Quadro 6 – Estudos de Regressão e suas técnicas

Silva et al. (2019)	Gray e Perkins (2019)	Rovira, Puertas e Igual (2017)	Kilian, Loose e Kelava (2020)
Linear Regression Robust Regression Ridge Regression Lasso Regression Support Vector Regression	Random Tree Regression	Collaborative Filtering Recommendation System Linear Regression Support Vector Regression	Elastic Net

Fonte: Elabora do pela autora

Dois estudos utilizam a *Survival Analysis* (Análise de Sobrevivência, em português) para estimar o tempo entre a entrada do aluno no Ensino Superior e o momento do abandono (CHEN, JOHRI e RANGWALA, 2018); (CONTINI e SALZA, 2020), e, por tratarem de variáveis numéricas, estes estudos foram alocados nas tarefas de regressão.

Quanto a definição do problema do negócio:

Esta questão é fundamental para a solução de um determinado problema. É a partir deste ponto que tem início um projeto de análise quantitativa. Entender qual é a definição de evasão permite ao pesquisador saber quais dados serão coletados e como eles devem ser tratados, entre outras questões.

Um dos problemas relacionados ao estudo do fenômeno da evasão no Ensino Superior, apontado no Referencial Teórico, está relacionado com as muitas definições existentes. O que pode ser constatado neste estudo.

Após leitura dos artigos, observou-se que nem todos os estudos fazem uma definição clara do que é evasão referindo-se apenas aos alunos que não concluem o Ensino Superior.

Alguns estudos focam na falha acadêmica enquanto outros no sucesso acadêmico do aluno. Há estudos que trabalham na perspectiva da retenção, no desempenho acadêmico e na conclusão dos estudos.

Os trabalhos que investigam o fenômeno da evasão em um sentido estrito, definem o termo como a não conclusão em um programa para o qual o aluno se inscreveu; a ausência de matrícula em um semestre posterior; a troca para outra área ou curso, e a transferência para outra instituição. O número de créditos em um determinado período também é considerado como medida para evasão, permanência ou conclusão. (KEMPER, VORHOFF e WIGGER, 2020).

Encontramos ainda o estudo feito por Kuzilek, Zdrahal e Fuglik (2021), que trata do tema de forma mais ampla, abordando a efetividade e a inclusividade; a primeira acontece quando a maioria dos alunos que se matriculam conseguem concluir em um tempo razoável e a segunda refere-se à capacidade de fornecer oportunidade para todos.

Quadro 7: Definição de Evasão

Continua

Autores	Definição
Delen (2010)	Se refere aquela pessoa que está deixando o curso por cancelamento, falha ao progredir para o próximo semestre e falha em múltiplos exames.
Nandeshwar; Menzies; Nelson (2011)	A evasão consiste no abandono de um programa de estudos antes de obter o título ou grau correspondente, considerando um tempo suficientemente longo para descartar a possibilidade de reincorporação.
Sarker; Tiropanis; Davis (2014)	Alunos que não concluem o bacharel em sua instituição / número de aluno que não retornam para o segundo ano.
Rovira; Puertas; Igual (2017)	Aluno que não ganha quantidade suficiente de créditos para permanecer na Universidade.
Alfredo (2018)	Alunos que ingressaram em uma carreira e no ano seguinte se inscreveram em uma carreira diferente, aqueles que saíram definitivamente, tiveram perda de matrícula porque não se inscreveram por 3 semestres consecutivos, alunos trancados.
Chen; Johri; Rangwala (2018)	Número de alunos que registrados em um curso não formalizaram matrícula novamente para os próximos dois anos acadêmicos consecutivos.
Zhang e Rangwala (2018)	Abandono ou parada. Alunos que falharam em se registrar no próximo semestre ou que tiveram um GPA igual 0.0 no semestre seguinte.
Aguilar-Gonzalez e Palafox (2019)	Qualquer aluno que deixa a escola ou qualquer outra instituição educacional por qualquer razão antes de concluir o programa de estudos no qual se matriculou sem se transferir para outra instituição.
Beaulac e Rosenthal (2019)	Define evasão como o abandono de carreira antes de obter o grau correspondente, considerando um período de tempo que possibilite o retorno segundo as regras. Tempo considerado para concluir os estudos (3 anos)
Kemper; Vorhoff; Wigger (2020)	Alunos que se inscreveram para 5 créditos em cursos, tiveram sucesso em menos de 18 créditos em cursos e pararam de fazer os cursos por 3 semestres consecutivos são considerados alunos que iniciaram um programa, mas não o concluíram.
Patterson et al. (2020)	Evasão - é a falha para se graduar depois de se matricular em um programa.
Tsai et al. (2020)	Alunos que não apareceram em dois semestres consecutivos.

Quadro 7: Definição de Evasão

Continuação

Gil et al. (2021)	Estudante (ele/ela) que não se matriculou no semestre seguinte ao último semestre de matrícula.
Kuzilek; Zdrahal; Fuglik (2021)	O texto fala de efetividade (onde a maioria dos alunos que se matriculam conseguem concluir em um tempo razoável) e inclusividade (capacidade de fornecer oportunidade para todos)
Palacios et al. (2021)	O abandono ocorre quando um indivíduo matriculado em uma instituição decide abandonar os estudos voluntariamente.

Fonte: Elabora pela autora.

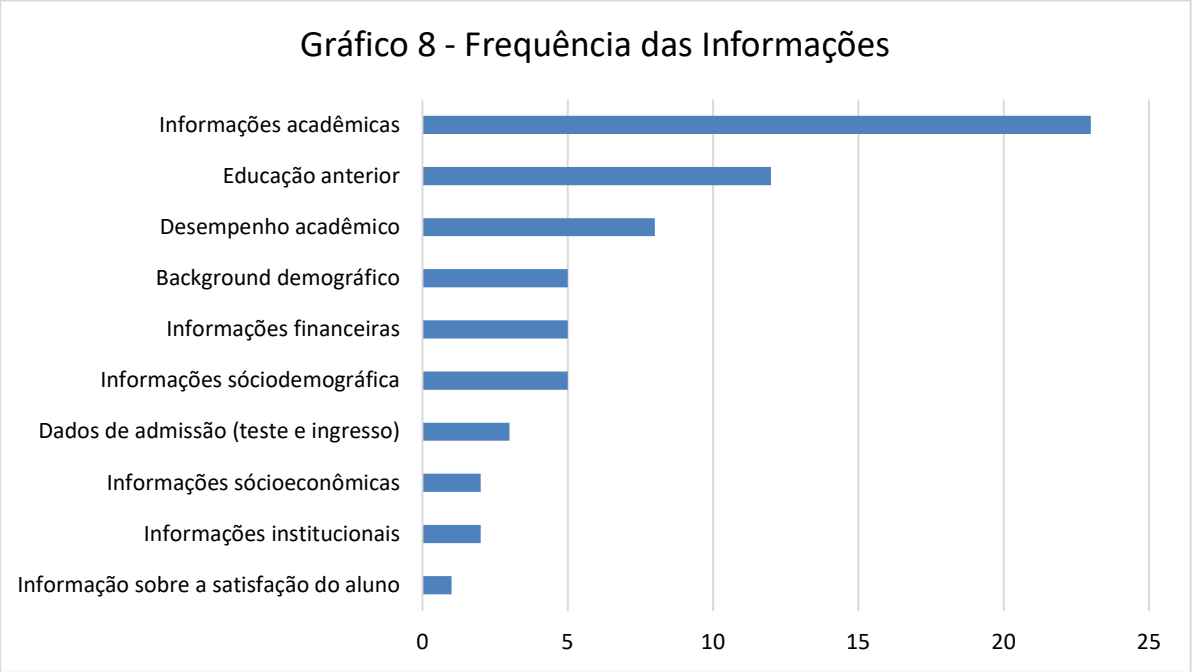
4.2.2 Teoria

As teorias segundo Hevner et al. (2004), são aproximações com a realidade e a base de muitas hipóteses. O objetivo desta seção é descrever em quais teorias os autores se apoiaram para tentar explicar e predizer o fenômeno da evasão. De acordo com as observações feitas os artigos se baseiam nas seguintes informações[5]:

- Informações sociodemográfica;
- Desempenho acadêmico;
- Informações acadêmicas;
- Educação anterior;
- Informações sobre o tipo de admissão;
- Informações institucionais;
- Informações financeiras;
- Background demográfico;
- Informação sobre a satisfação do aluno.
- Informações sobre o comportamento do aluno.

[5] Procurou-se manter as nomenclaturas utilizadas pelos autores.

O gráfico abaixo apresenta a frequência com que estas informações são utilizadas nos artigos. Podemos visualizar que as informações acadêmicas são as mais utilizadas seguida pela educação anterior e o desempenho acadêmico.



Fonte: Elaborado pela autora.

A figura 12 ilustra a construção das hipóteses com as quais os autores trabalharam a questão da evasão de alunos. É possível constatar a referência a diversos autores que se ocuparam de modelar a evasão de alunos dentre eles Vicent Tinto, John Bean, Cabrera e colaboradores e Astin apenas para citar alguns.

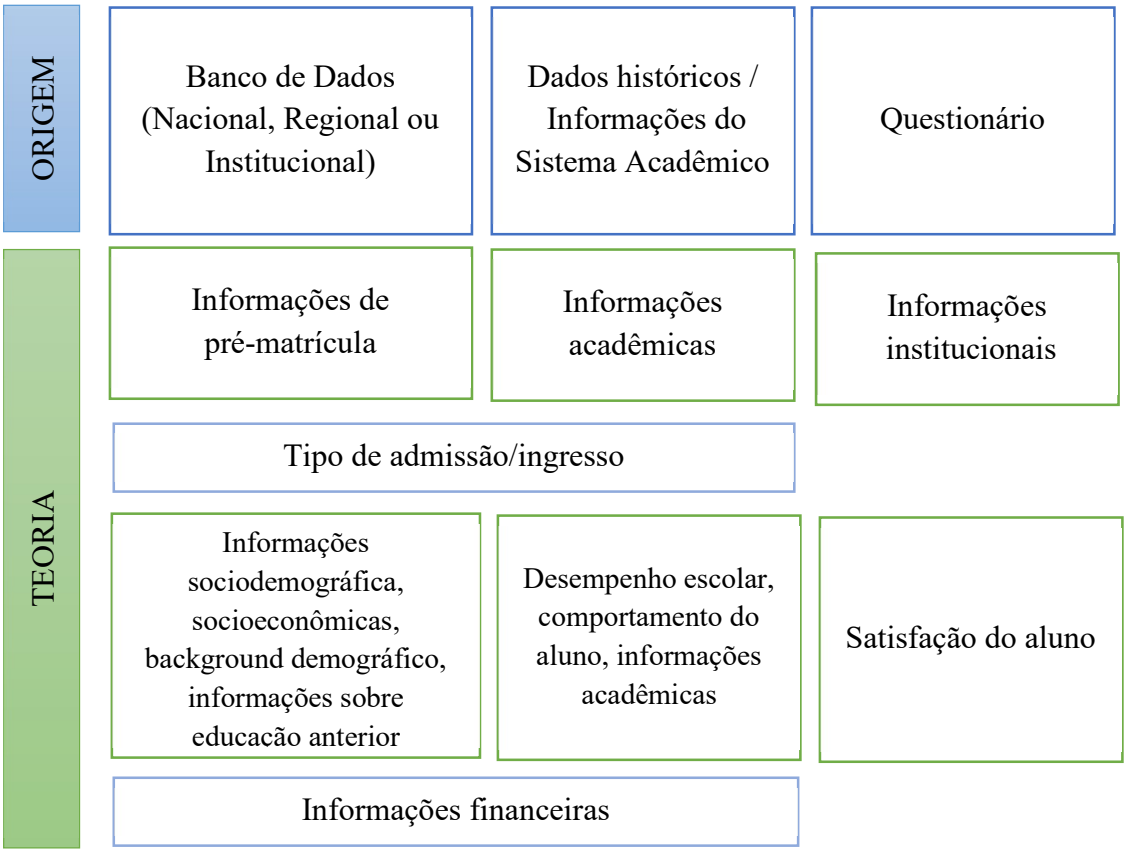


Figura 12 – Teorias utilizadas para explicar e predizer o fenômeno da evasão

Outra questão observada, diz respeito à forma como cada autor traduz essas informações. Sarker, Tiropanis e Davis (2014) utilizam um questionário tradicional elaborado em escala para verificar a integração institucional. Ullah et al. (2019) também se utiliza de questionários com o objetivo de avaliar a satisfação do aluno com relação às questões institucionais.

Quando se trata de informações financeiras, há aquelas relacionadas às questões socioeconômicas (TSAI et al. (2020); PALACIOS et al. (2021)) e há aquelas que se referem a financiamento estudantil, bolsas de estudo ou empréstimos (ZEINEDDINE, BRAENDLE e FARAH, 2021). A educação anterior tanto se refere ao tipo de escola em que o aluno cursou o Ensino Médio, ao tipo de financiamento da escola e ao desempenho acadêmico anterior. O estudo sobre o comportamento do aluno é traduzido pelos autores Kuzilek, Zdrahal, Fuglik, (2021) pela forma como o aluno realiza os exames/provas durante o primeiro semestre no Ensino Superior.

Na seção seguinte será detalhado a modelagem e as variáveis utilizadas para transformar a teoria em realidade.

4.2.3 Modelagem ou “Engenharia de Modelo”

Este componente é analisado em duas etapas. A primeira diz respeito às variáveis utilizadas para a construção dos modelos. Elas foram divididas neste estudo em cinco categorias: background demográfico, financeira, desempenho/informação escolar anterior, admissão, informações/desempenho acadêmico. Algumas variáveis foram agrupadas para facilitar o entendimento do leitor e generalizar os termos para evitar a redundância no texto.

A tabela 3 demonstra a diversidade com que as diversas categoria foram utilizadas nos estudos.

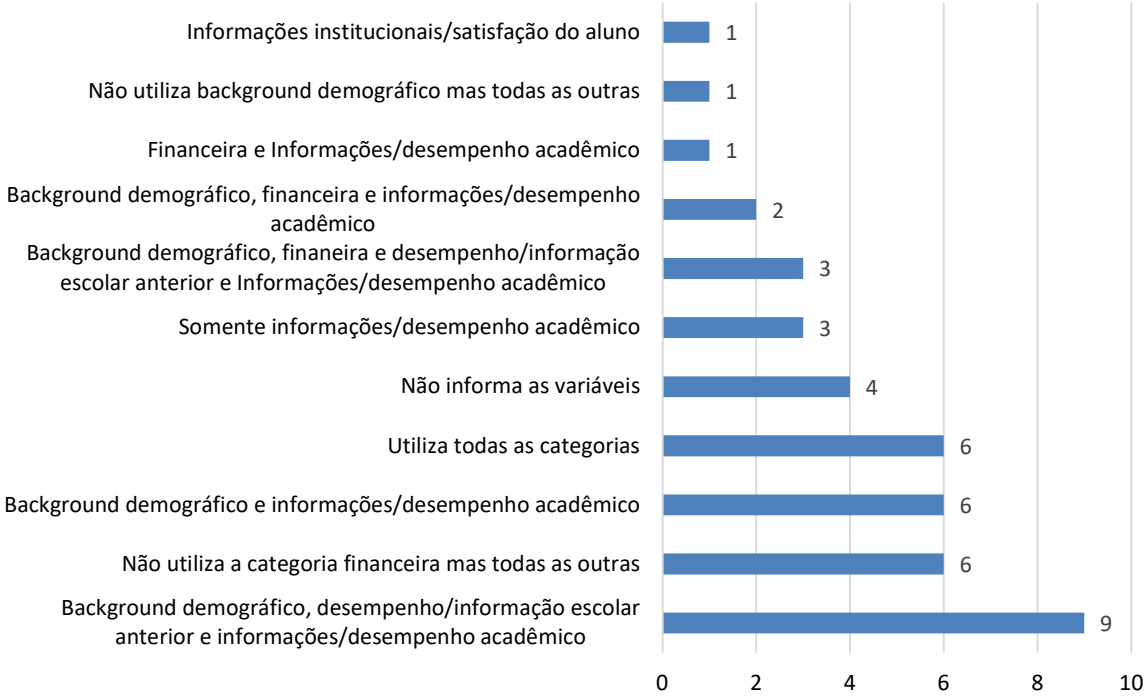
Tabela 3: Distribuição das categorias

Número	Categoria
1	Background demográfico, desempenho/informação escolar anterior e informações/desempenho acadêmico
2	Não utiliza a categoria financeira, mas todas as outras
3	Background demográfico e informações/desempenho acadêmico
4	Utiliza todas as categorias
5	Não informa as variáveis
6	Somente informações/desempenho acadêmico
7	Background demográfico, financeira e desempenho/informação escolar anterior e informações/desempenho acadêmico
8	Background demográfico, financeira e Informações/desempenho acadêmico
9	Financeira e Informações/desempenho acadêmico
10	Não utiliza background demográfico, mas todas as outras
11	Informações institucionais/satisfação do aluno

Fonte: Elaborado pela autora.

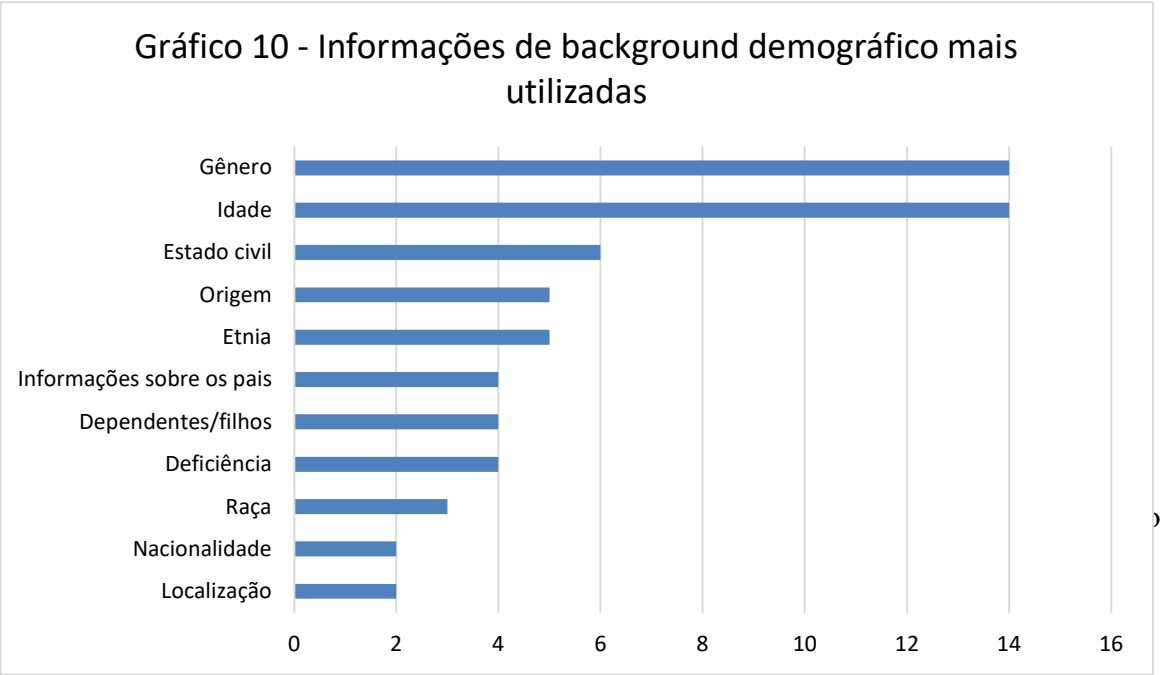
O gráfico abaixo apresenta a distribuição das categorias utilizadas nos estudos.

Gráfico 9 - Distribuição das categorias utilizadas nos estudos



Fonte: Elaborado pela autora.

- **Background demográfico**: incluem dados sobre gênero/sexo, se possui alguma deficiência física ou mental (grau; temporária ou permanente), estado civil, idade, raça, estado/município/província/distrito onde vive, número de dependentes, qual a origem do aluno (localidade), número de pessoas na família, tipo de emprego do pai e da mãe, etnia, idade, cep, estado de moradia permanente, data de nascimento, tipo de ocupação dos pais, ocupação do aluno, nacionalidade, segunda nacionalidade, escolaridade dos pais, ocupação dos pais, grávida ou com filhos, cidadania, responsabilidade familiar/se chefe de família, nome, índice de pobreza da comunidade, empregabilidade, se possui ou não desvantagem (se família de baixa renda, deficiente ou se recebe ajuda financeira), se tem dependentes/ nº de filhos, se solteiro com dependentes, se é primeira geração, número de pais que estão vivos, raça, urbano ou rural, estrato social, status ocupacional, horas trabalhadas, número de pais que estão vivos, ocupação dos pais, número de membros que trabalham, número de membros matriculados em instituição de ensino.



Fonte: Elaborado pela autora.

- **Financeira:** são aquelas variáveis relacionadas a renda familiar, se o aluno recebe algum tipo de ajuda financeira (empréstimo, subsídio/renúncia de matrícula/bolsa de estudos, financiamento, crédito), quintil econômico, se o aluno trabalha em meio período com ajuda financeira, se possui plano de saúde, qual a fonte da ajuda financeira, renda dos pais e se é independente para propostas de ajuda financeira.
- **Desempenho/informação escolar anterior:** GPA (do inglês *Grade Point Average*, que significa média de pontos das notas) do Ensino Médio, data da graduação do ensino médio, tipo de educação secundária (tipo de colégio, tipo de ensino, tipo de financiamento recebido pela instituição, se escola pública ou privada), Sistema Escolar, média de notas considerando apenas algumas disciplinas, pontuação educacional secundário (NEM), fonte de entrada. Se possui diploma do Ensino Médio, se Ensino Médio externo ou educado em casa/GED ou outro equivalente, se possui certificado de conclusão ou nenhum diploma do ensino médio, certificado ou outro equivalente, se fez preparatório, *Stream* Ensino Médio, média, nota em matemática, inglês, árabe, teste de nivelamento de matemática e inglês, Nota do Sistema Internacional de Exame da Língua Inglesa (ouvir, escrever, ler, falar), Nota final do Ensino Médio (Média dos 3 anos acadêmicos), tipo de ensino médio (técnico, geral, noturno), tipo de gênero da escola

(no caso do sistema educacional chileno), tipo de governança escolar, resultado do teste de atitude do Ensino Médio.

- **Admissão:** incluem informações sobre como o aluno ingressou no Ensino Superior, quais os critérios que foram utilizados: exame de qualificação, resultados do processo de admissão, forma/tipo de ingresso, uso do NEM, forma de ingresso: quota regional, admissão direta, admissão condicional, com GPAX acima de 2,0 (IAM-ON e BOONGOEN, 2017), classificação na lista de admissão, peso médio no teste de admissão universitário, PSU score (*La Prueba de Selección Universitaria*), nota de matemática no exame de admissão, informações sobre o exame de admissão realizado em duas etapas (EXANI-II Admission e EXANI-II Diagnostic), SAT scores (sigla em inglês *Standard Admision Test* – pontuação do Teste de Admissão Padronizado), preferências no Procedimento de Aplicação.
- **Informações/desempenho acadêmico:** esta é a categoria onde se encontram a maior variedade de variáveis: ID do aluno, idade na matrícula, curso, área, status do aluno (aprovado, reprovado, evadido ou não, ativo, trancado, desligado, transferido para outro curso na mesma instituição, formado, falecido); faculdade, grau, concentração, créditos ganhos/créditos perdidos, horas ganhas/horas perdidas, horas transferidas, duração, anos em que o aluno permaneceu matriculado, se recebeu remuneração por atividade exercida durante o Ensino Superior, se participou de atividade extracurricular não obrigatória, se o aluno participou de atividade de extensão ou de atividade extracurricular (e.g. monitoria, pesquisa), se possui empréstimo, número de ausências na escola, se trabalha em meio-período, número de assuntos/disciplinas alertadas, se atuou como núcleo de apoio, desempenho acadêmico (rank na classe), notas finais dos alunos dos diferentes cursos durante seus anos acadêmicos, média das notas em todas as disciplinas ou em disciplinas específicas, GPA acumulado anual ou semestral, notas finais em todas as disciplinas, tempo que o aluno levou para realizar as provas/exames, pontuação SAT (do inglês *Scholastic Assessement Test* – Teste de Avaliação Escolástico), se o aluno solicitou algum tipo de auxílio e se este foi aceito ou não (e.g. transporte, moradia, alimentação), se trabalha enquanto está matriculado (tempo integral, meio período ou não trabalha);

Além destes grupos de variáveis existe uma variável **institucional** identificada no estudo dos autores Sarker, Tiropanis e Davis (2014) no qual utilizaram um questionário para

coletar informações relativas à satisfação do aluno com a instituição. campo de estudo, tipo de ocupação, tipo de acomodação, nível de pontuação, interação com os pares, interação equipe-aluno, preocupação da equipe com o desenvolvimento e ensino do aluno, desenvolvimento intelectual e acadêmico, compromisso com os objetivos, compromisso institucional I, II, II, opinião sobre o ensino no curso, intenção, avaliação e feedback, apoio acadêmico, desenvolvimento pessoal e sobre a satisfação geral com a qualidade do curso.

O estudo realizado por Silva et al. (2019) utiliza um conjunto de dados onde as informações dos alunos não estão individualizadas e as variáveis utilizadas não cabem na estrutura de análise da DSR. Eles utilizam dados do Censo da Educação Superior e dos Indicadores de Fluxo da Educação Superior do INEP e consideram os seguintes fatores: número de alunos novos, indicador de permanência do aluno, indicador de conclusão do aluno, número de graduados no curso, número de graduados, taxa de evasão, alunos no turno da noite, número de alunos que permaneceram no curso. Já os autores Santos et al. (2019) não informam as variáveis utilizadas para a modelagem do artefato.

Como cada modelo é construído para alcançar um objetivo, relacionamos abaixo os alvos (aquilo que deseja prever) a partir da análise dos textos. Somente 3 artigos não deixam claro o alvo do estudo. A maioria dos artigos trabalha com classificação binária, mas também se observou a tentativa de classificação em classes não binárias. A relação dos alvos e a quantidade em que foram usados segue abaixo:

- *Binária (2 classes)*: conclui/não conclui (3), evadiu/não evadiu (11), evasão/graduação (3), registrado ou não para o 2º outono (1), sucesso/falha (1), passou/falhou (2), graduou/falhou (1), retido/não retido (1), retenção/evasão (1), satisfação/retenção (1), defendeu/não defendeu (1), boa graduação/graduação básica (1), promoção/não promoção (1);
- *Não binária (3 classes)*: se o aluno retornou depois de 1 ano / se o aluno retornou depois de 2 anos / se o aluno retornou depois de 3 anos (1), Promoção/ repetição/evasão (1);
- *Não binária (múltiplas classes)*: passou/falhou/falha condicional/repetiu/repetiu um semestre;
- Nota dos alunos;
- O curso em que o aluno concluirá;
- Quando o aluno abandonou o curso;

O último componente da estrutura da DSR é o Protocolo Experimental de Avaliação cujos resultados serão apresentados na próxima seção.

4.2.4 Protocolo Experimental de Avaliação

A primeira etapa deste componente diz respeito a como os autores fizeram a construção do problema. Esta análise diz respeito ao desenvolvimento do modelo que passa pela aquisição dos dados e descrição deste processo utilizado nos artigos. De acordo com as observações as principais fontes de dados são:

- informações do sistema acadêmico (fontes internas);
- pesquisas institucionais, que compõem bases de dados nacionais, regionais e acadêmicas;
- questionários utilizados pelos pesquisadores para adquirir informações mais específicas (fontes externas).

Esse passo é muito importante para a *Data Analytics* pois segundo Elragal e Haddara (2019) na *Design Science Research* os artefatos devem ser explicados e avaliados durante todo o processo, garantindo uma boa qualidade para a construção do modelo.

Quanto às etapas e metodologias alguns estudos utilizaram uma metodologia popular em mineração de dados, chamada CRISP (*Cross Industry Standard Proccess*), outros empregam a KDD (*Knowledg Discovery Databases*), Patterson et al. (2020) fizeram uso do *Lean Six Sigma*; Pérez et al (2018) fazem uso de uma metodologia da Ciência de Dados e os demais não fazem referência às metodologias que estão sendo usadas conforme podemos observar na tabela a seguir:

Tabela 4: Metodologias	
Ciência de Dados	1
CRISP	8
KDD	3
Lean Six	1
Outras	29
Total	42

Fonte: Elaborado pela autora

Para melhor entendimento do leitor, alguns esclarecimentos são necessários: **conjunto de dados** são dados que serão analisados (oriundos de fontes internas, externas ou combinados); **experimento ou cenários** são tarefas que implicam em modificação do conjunto de dados inicial em busca de melhores resultados/desempenho. **Modelo** é quando é feita a aplicação de uma técnica a um conjunto de dados. Os termos *dropout, attrition* e outros relacionados utilizados nos estudos foram padronizados nesta pesquisa através do termo **evasão**.

Na tentativa de encontrar um padrão entre os estudos analisados, eles foram organizados em 8 grupos, no que diz respeito a forma de utilização dos dados:

Grupo 1 – utilizam somente um conjunto de dados.
Grupo 2 – utilizam dois ou mais conjuntos de dados, realizando experimentos que tem como foco a mudança na quantidade de variáveis utilizadas em cada um.
Grupo 3 – utilizam dois ou mais conjuntos de dados, realizando experimentos com inserção de novas informações em cada etapa.
Grupo 4 – utilizam dois ou mais conjuntos de dados, realizando experimentos onde uma das entradas do modelo seguinte é o resultado do modelo anterior.
Grupo 5 – utilizam conjunto de dados composto por fontes internas e externas.
Grupo 6 – utiliza um conjunto de dados dividido em grupos.
Grupo 7 – utiliza dois ou mais conjuntos de dados, aplicando as técnicas escolhidas a cada um deles.
Grupo 8 – utiliza dois conjuntos de dados e os transforma em apenas um.

A tabela abaixo mostra a quantidade de artigos relacionados em cada grupo.

Tabela 5: Classificação dos estudos quanto ao uso dos dados

Grupos	*Nº de artigos*
Grupo 1	17
Grupo 2	7
Grupo 3	5
Grupo 4	1
Grupo 5	4
Grupo 6	4
Grupo 7	3
Grupo 8	1

Fonte: Elaborado pela autora.

A figura 13 é um exemplo do Grupo 3. Palácios e colaboradores (2021) utilizaram um modelo criado em um estudo na Universidade Católica de Maue, na Espanha, adaptado ao contexto Chileno. Esse modelo tem como alvo evadiu/não evadiu (ativo) e é elaborado para prever a evasão em quatro níveis.

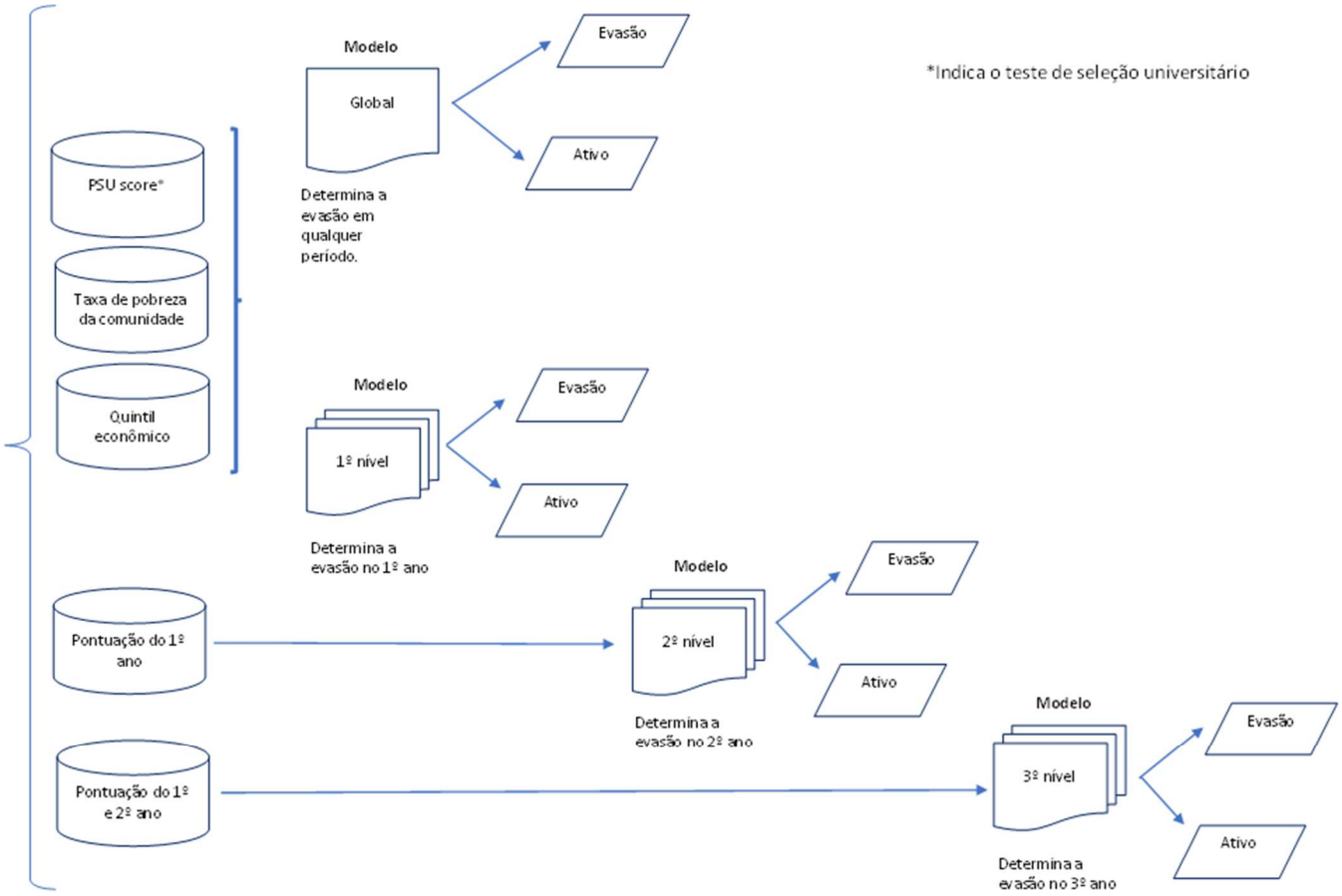


Figura 13 – Esquema de 4 diferentes modelos proposto para prever a retenção/evasão de alunos. (Adaptado de Palacios et al. (2021))

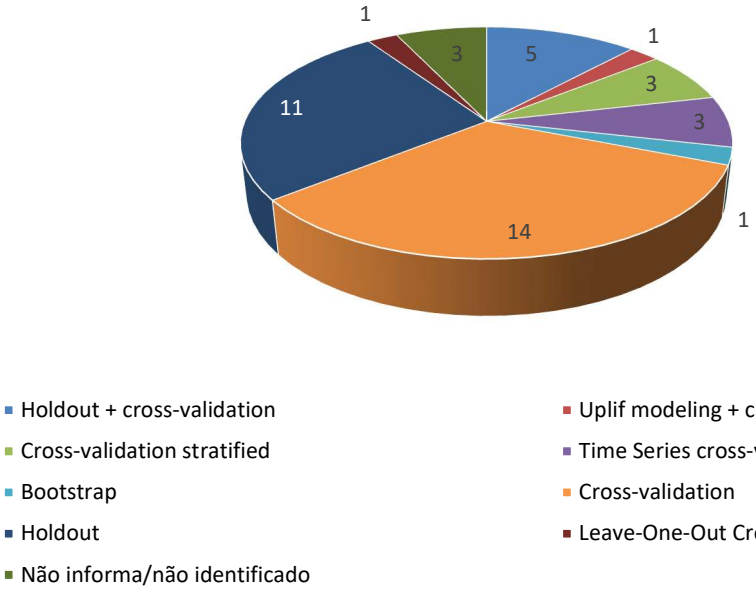
A seguir é descrito como os estudos avaliam os modelos que foram desenvolvidos, como os conjuntos de dados são treinados e testados, quais as métricas para avaliação de desempenho e se os resultados alcançados foram satisfatórios para a solução do problema. Esse componente da análise é importante pois segundo Hevner et al. (2004) "A seleção dos métodos de avaliação deve estar apropriadamente de acordo com o artefato desenvolvido e as métricas de avaliação escolhidas" (HEVNER et al., pp 86, 2004)

Durante a leitura dos textos foram encontrados oito métodos de validação de modelos e em aproximadamente 10% dos estudos não havia descrição do método utilizado para dividir o conjunto de dados em treino e teste. São eles:

- *Holdout*
- *Holdout com cross-validation*
- *Cross-validation*
- *Bootstrap*
- *Uplift modeling com cross-validation*
- *Times series cross-validation*
- *Cross-validation stratified*
- *Leave-on-out cross-validation ou LOOCV*

É importante ressaltar que a *Uplift modeling* foi mantida como método de validação, pois se caracteriza como uma nova proposta utilizada por Olaya et al. (2019) para divisão dos conjuntos de dados que se pretende selecionar. O gráfico a seguir demostra a proporção em que os métodos foram utilizados nos estudos.

Gráfico 11 - Métodos de Validação



Fonte: Elaborado pela autora.

O gráfico 11 demonstra que o método de validação cruzada e suas variações foram os mais utilizados pelos autores, presente em aproximadamente 53% dos estudos, seguido pelo método *holdout*, que considerando apenas sua utilização isolada, foi encontrado em 33% das pesquisas.

Cabe ressaltar que o estudo elaborado pelos autores Pallathadka et al. (2021) não informa os métodos utilizados para validação do modelo construído.

Após descrever as várias formas como os estudos dividem os dados em conjunto de treino e teste, ou seja, validam o modelo, a etapa seguinte é identificar as métricas utilizadas para avaliar o seu desempenho. Um quadro com todas as métricas utilizadas para avaliação das técnicas está disponível no Apêndice B.

Com exceção dos estudos realizados por Sarker, Tiropanis e Davis (2014); Iam-On e Boongoen, (2017); Cardona e Cudney (2019); Contini e Salza (2020) e Peralta et al. (2021), encontramos a utilização de mais de uma técnica para modelagem dos dados, o que gera a necessidade de comparar a eficiência entre elas. Mas mesmo nesses casos o uso de métricas são necessárias para saber se o modelo consegue realizar a tarefa proposta e prever a evasão dos alunos, ainda mais considerando o uso de experimentos ou de algoritmos diferentes.

Alguns estudos utilizaram métricas específicas para avaliação dos seus estudos. É o caso de Maldonado et al. (2021) que além da Curva ROC faz a comparação entre os classificadores para saber qual deles apresenta o melhor Lucro Máximo (MP – sigla do inglês para *Maximum Profit*). O MP avalia através da matriz de confusão os custos e os benefícios associados a campanha de retenção e do Valor do Tempo de Vida do Cliente/Aluno (CLV – sigla do inglês para *Costumer Lifetime Value),* que considera as taxas de matrículas dos alunos retidos pela campanha e o valor do financiamento do governo para aqueles alunos que se enquadram como tal.

Nandeshwar, Menzies e Nelson (2011) utilizam como métrica a probabilidade de detecção e a probabilidade de alarme falso (considerando os verdadeiros positivos e os falsos negativos), observando a variação entre os dois na validação cruzada e utilizando o teste Mann-Whitney (onde o que vencer o maior número de vezes vence).

As métricas utilizadas por Bello et al. (2020) são a MDG (Média da Diminuição da Gini) e a MDA (Média da diminuição da Acurácia). Olaya e colaboradores (2020) também utilizaram a Gini como métrica.

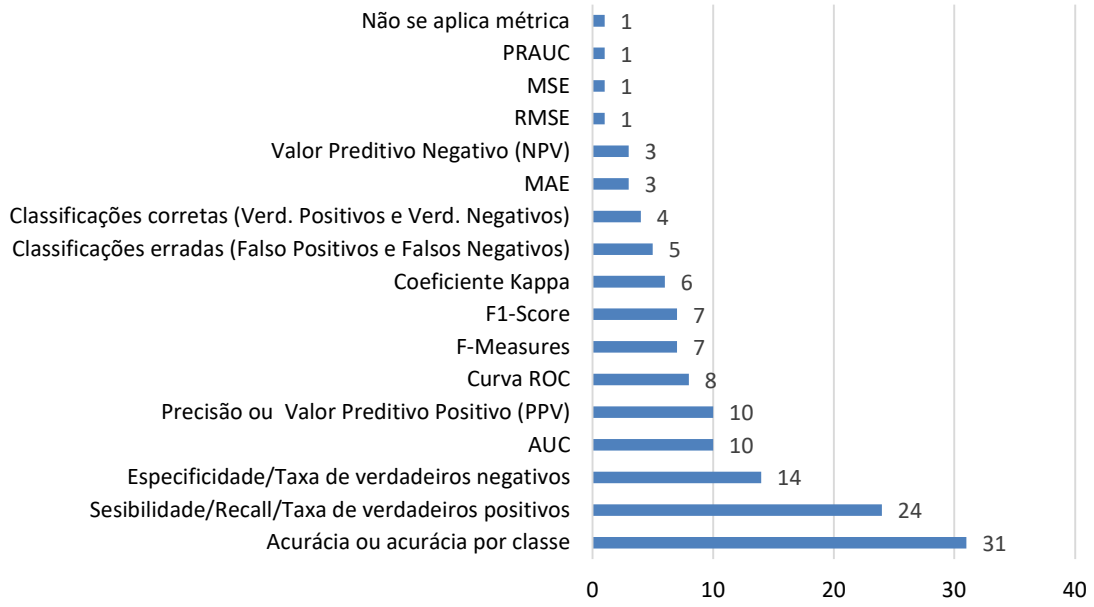
Beaulac e Rosenthal (2019) utilizaram diagrama de caixa[6] (em inglês *boxplot*) e Silva et al. (2019) além deste fizeram uso do Ganho Relativo (RG) e do teste Kolmogorov-Smirnov e Wilcoxon.

Os autores Chen e colaboradores (2018) utilizaram os modelos de análise de sobrevivência - Modelo de COX) e Modelo de Regressão Aditiva de Aalen (em inglês - *Aalen's*

[6] Ferramenta utilizada para visualizar anomalia nos dados (outlier). (BEAULAC e ROSENTHAL, 2019)

Additive model). Os estudos realizados por Palacios et al. (2021) utilizaram k-estatística e o teste de Friedman, enquanto Rovira, Puertas e Igual (2017) utilizaram o Kendall.

Gráfico 12 - Métricas



Fonte: Elaborado pela autora.

De um modo geral as métricas utilizadas são tradicionais quando se trata avaliar técnicas de aprendizagem de máquina, como o Erro Médio Quadrático da Raiz (RMSE[7]) e o Erro Médio Absoluto (MAE) para avaliar as tarefas de regressão e as métricas geradas a partir da Matriz de Confusão para validar as tarefas de classificação como a Acurácia, Precisão, F1-Score e F-Measure. O gráfico 12 ilustra o uso dessas métricas.

O último item examinado dentro do Protocolo Experimental diz respeito sobre os resultados. Se os artefatos propostos alcançam os resultados esperados e podem ser aplicados no mundo real.

Como se trata de trabalhos distintos cujo tema comum é o fenômeno da evasão, agrupar os estudos com relação à eficácia dos seus desempenhos não é uma tarefa fácil, devido ao uso de técnicas e métricas distintas, bem como o tamanho do conjunto de dados e das variáveis utilizadas.

[7] Erro Médio Quadrático (MSE) – nomenclatura encontrada em alguns estudos.

Ao escrutinar os textos constatou-se que todos alcançaram os resultados desejados no que diz respeito às tarefas preditivas. Os modelos propostos são capazes de prever a evasão, a retenção, o desempenho e as notas dos alunos de acordo com a proposta de cada estudo.

Para Peralta et al. (2021) prever a evasão é muito mais difícil do que prever a graduação. Os autores recomendam a seleção das variáveis que serão utilizadas, pois segundo eles, nem sempre um número maior representa um melhor desempenho. Quando se trata de Rede *Bayesian* seu desempenho depende do número de variáveis e sua construção da estrutura da rede.

Observou-se nos estudos, que algumas técnicas são mais utilizadas, se sobressaindo sobre outras. A Árvore de decisão, por exemplo é uma das técnicas individuais mais utilizadas, mesmo não apresentando o melhor desempenho dentre os modelos. Segundo Delen (2010); Berka e Marek (2020); Kemper, Vorhoff e Wigger (2020) ela é mais fácil de ser construída a partir do conjunto de dados e de ser utilizada, além disso ela favorece a interpretação dos seus resultados.

Berka e Marek (2020) sugerem que regras de associação são melhores para realizar a análise de dependência entre as variáveis quando comparada com a Regressão Logística, em inglês - *Logistc Regression* (LR), enquanto Wan Yaacob et al. (2020) atentam para o fato de que essa técnica possui um desempenho melhor, ainda que considere que os demais modelos são mais compreensíveis.

O algoritmo KNN foi relatado por Zeineddine, Braendle e Farah (2021) como um bom preditor capaz de desempenhar bem a partir da entrada de novos dados. Pérez et al (2018) constataram que o uso de mais variáveis nos modelos que utilizam Regressão Logística e Árvore de Decisão não trouxe muita diferença para o resultado dos seus experimentos. Ao passo que Al-Sudani e Palaniappan (2019), assim como Santos et al. (2019) acreditam que um conjunto maior de dados ou de mais parâmetros podem melhorar o desempenho do algoritmo KNN.

Quanto aos estudos que utilizaram variáveis de primeiro ano, os diversos autores que trabalharam com este grupo de variáveis consideraram ser possível sua utilização para prever alunos em risco de abandono. Segundo Alvarez, Callejas e Griol (2020) quando considerado somente variáveis de pré-matrícula a acurácia diminui, de 96,71% para 68,86%, já no estudo realizado por Zeineddine, Braendle e Farah (2021) utilizando Autoaprendizagem de Máquina o resultado obtido excedeu a acurácia de 70% apenas utilizando dados iniciais.

Gil et al. (2021), Cardona e Cudney (2019) consideraram que Máquina de Suporte de Vetores é a melhor técnica para trabalhar com variáveis de entrada e melhor para treinar com um número pequeno de dados. Os autores também constataram que Floresta Aleatória (em inglês *Random Forest*) desempenha melhor quando se trabalha com variáveis categóricas e numéricas juntas, além de auxiliar a encontrar padrões nos dados (VILA et al. 2019).

Palacios et al (2021) sugerem o uso de um modelo preditivo que utiliza níveis para prever a evasão baseado em seus experimentos que alcançaram uma acurácia de 80%. Tampakas et al (2019) propõem o uso de um classificador construído em duas etapas, a primeira para prever os alunos em risco de abandonar ou falhar em seus cursos e a segunda para prever a conclusão e 4 ou 6 anos, obtendo sucesso com os modelos adotados.

De acordo com Delen (2010) os métodos *ensembles* possuem um desempenho preditivo melhor do que os modelos individuais, no entanto os autores Acero, Achury e Morales Piñero (2021) apontam o bom desempenho dos dois tipos. Alguns artigos relataram o uso de técnicas de balanceamento de dados para buscar um equilíbrio entre as classes. Os autores Zeineddine, Braendle e Farah (2021) e Delen (2010) relataram que os conjuntos de dados balanceados produziram melhores resultados.

No que diz respeito à origem dos dados, Delen (2010) demonstrou que o uso de dados institucionais são boas fontes para a construção dos modelos, assim como Killian, Loose e Kelava (2020) que alcançaram bons resultados utilizando poucas informações oriundas do Sistema Acadêmico das instituições pesquisadas. No entanto, Sarker, Tiropanis e Davis (2014) com base em seus experimentos, recomendam o uso combinado, com fontes de dados abertos externos com conjunto de dados institucional, pois este modelo obteve um desempenho melhor do que os demais.

Zhang e Rangwala (2018) propuseram um modelo que consegue um bom desempenho preditivo utilizando poucas informações dos alunos do semestre corrente juntamente com informações dos alunos de períodos anteriores. Segundo eles, o método possui poucas vantagens com relação aos outros métodos, mas requer pouca informação. Patterson et al (2020) atingiram uma acurácia de 82% baseado em apenas quatro variáveis de entrada: GPA acumulado, lista do reitor, aluno adulto (não tradicional), e gênero.

Huo et al (2020) destacam em seu trabalho a importância de preditores para estudar a retenção de alunos não tradicionais[8]. Seus modelos alcançaram um bom desempenho nestes

[8] Alunos não tradicionais - alunos que possuem ao menos uma das sete características: atraso para se matricular no Ensino Superior, matriculado em meio período, trabalho em tempo integral, financeiramente independente

grupos. Jayaraman, Gerber e Garcia (2019) apontam para a necessidade de modelar grupos de alunos separadamente para obter sucesso na previsão. No caso do estudo realizado pelos autores, os dois grupos estudados foram de alunos nativos e de alunos transferidos de outras instituições.

Para Ahmada Tarmizis seus modelos podem ser utilizados no mundo real, pois produziram acurácia acima de 97 %. Os autores Gray e Perkins (2019) produziram um modelo capaz de identificar alvos para intervenção antecipada, mas salientam que para um coorte diferente de alunos será necessário utilizar diferentes padrões para conseguir identificá-los.

Iam-On e Boongoen (2015) utilizaram *Clustering ensemble* para construir uma matriz de dados resumida, ou seja, para redimensionar o conjunto de dados. Eles concluíram que este método é melhor do que as técnicas de redimensionamento de dados tradicionais (PCA e KPCA), pois o experimento que eles realizaram alcançaram resultados melhores utilizando a nova técnica para transformar o conjunto de dados original em um novo conjunto.

De acordo com os estudos feitos por Silva et al (2019) modelos de regressão *ensembles* baseados em *bagging* poderiam ser usados para prever com exatidão o desempenho dos alunos, pois eles são capazes de reduzir o erro na mesma medida em que antecipam a evasão.

No que diz respeito às variáveis que explicam a evasão ou a retenção os estudos podemos destacar o apontamento de alguns pesquisadores:

- NEM – Nota do Ensino Médio; (PEREZ et al 2018; AGUILAR-GONZALEZ e PALAFOX, 2019)
- O recebimento de benefícios como bolsa de estudos e créditos e o peso médio do Teste de Admissão Universitário - UTA Score (em inglês, *Weighted Average University Admission Test*) (VILORIA et al, 2019)
- Resultados acadêmicos anteriores podem ser bons preditores do desempenho acadêmico (Pallthadka et al, 2021)
- Informações acadêmicas, fatores demográficos, fatores psicológicos são fatores que contribuem para a evasão de alunos; (HEDGE e PRAGEETH 2018)
- Alunos com baixa nota de entrada, mais velhos, com grande lacuna entre o Ensino Médio e o Ensino Superior, alunos com baixo número de créditos, média abaixo de 7 e com baixo desempenho escolar anterior (GIL et al, 2020)
- Média do Ensino Médio e IELTS Band. (BILQUISE, ABDALLAH, KOBBAEY, 2020)

para receber proposta de ajuda financeira, possui outros dependentes além do cônjuge, foi pai/mãe solteira ou não tem diploma do Ensino Médio

- O comportamento do aluno ao realizar os exames; (KUZELI, ZDRAHAL, FUGLIK 2020)
- Desempenho acadêmico e ajuda financeiras; (BELLO et al, 2020)
- GPA (PATTERSON et al, 2020; PEREZ, CASTELLANOS, CORREAL, 2018)
- Notas (BELUAC e ROSENTHAL, 2019; ROVIRA, PUERTAS E IGUAL, 2017)
- SAT Matemática, SAT Inglês (alunos transferidos) e GPA faculdade e GPA Ensino Médio (alunos nativos); (JARAYAMAN, GERBER e GARCIA, 2019)
- Desempenho acadêmico, número de ausências, números de assuntos vistos, classificação em sala. (TSAI et al, 2020)
- Background familiar, status socioeconômico são critérios de persistência para alunos do 3º ano (NANDESHWAR, MENZIES E NESLON, 2011)
- GPA no primeiro ano, horas trabalhadas por semana, quantidade de empréstimos tomados, percepção do aluno sobre o tempo de conclusão, seletividade da instituição, controle institucional, seleção de transferência, matrícula em período integral ou meio período, durante o ano todo ou somente em uma parte. (HUO et al, 2020)

Quando se trata de tarefas descritivas, o trabalho de Contini e Salza (2020) constatou que a análise descritiva para a implementação de uma estimativa de risco para analisar carreiras acadêmicas com dados administrativos pode ser adaptada para estudar outros aspectos relacionados com as carreiras dos alunos, incluindo rematrícula após a evasão. Ullah et al (2019) conseguiram através da Regressão Logística encontrar relação entre a insatisfação do aluno e a retenção.

Santos et al (2019) constataram que alunos dos cursos de Engenharia da Computação e de Sistema de Informação abandonam no 4º semestre e que alunos do Ciência da Computação abandonam no 6º semestre, período nos quais os algoritmos apresentaram melhor desempenho. Em uma outra análise, um bom desempenho nas disciplinas de Matemática e Física está relacionada com um bom desempenho em Engenharia de Sistemas. (PEREZ, CASTELLANOS, CORREAL, 2018)

Os trabalhos prescritivos ou prognósticos produziram os seguintes resultados: Kilian, Losse e Kelava (2019) concluíram que o programa de estudos em si não contribui como indicador de grupo de risco, candidatos a professores começam com relativo pré-requisito com relação a matemática, mas não possuem vantagens com relação aos demais alunos. O sucesso

na disciplina *Analysis* I não depende somente de apenas um semestre estudado, mas sim da quantidade de semestres matemáticos com efeito positivo.

Olaya e colaboradores (2020) demonstraram que focar esforços de retenção (e.g. oferecendo tutoria) em alunos com probabilidade de ser retido pela intervenção amplia os efeitos do programa. A seleção através da *uplift modeling* é melhor do que a seleção aleatória. Os atributos de pré-matrícula demonstram que os esforços de retenção podem ser proativos.

Maldonado et al. (2021) alertam para a complexidade de utilizar a Análise de Negócios (em inglês, *Business Analytics*) para prever o *churn*/saída de alunos do sistema, pois na Educação os incentivos em sua maioria não são monetários (*e.g.* tutorias, assistência psicológica, mentoria) e o aluno que é selecionado para participar pode não aceitar o convite. No entanto, sua metodologia se mostrou promissora demonstrando a economia em se conseguir reter alunos com probabilidade de permanecer se forem direcionados para a intervenção.

De todos os artigos analisados, somente o estudo realizado por Perez et al. (2018) propõe um artefato e compara sua eficiência a um produto existente e já utilizado no mundo real e de acordo com os autores o modelo proposto possui um desempenho melhor do que o sistema implementado na instituição pesquisada.

4.3 Classificações das pesquisas segundo o critério da *Design Science Research*

A classificação das pesquisas segundo os critérios da DSR é uma tarefa difícil, pois em geral os estudos argumentam que suas propostas podem ser aplicadas em outros contextos, e por esse motivo além de contribuir com a teoria estariam contribuindo com a prática.

Este trabalho limita-se a fazer o enquadramento quanto aos tipos de pesquisas de acordo com os componentes observados nos estudos.

Sendo assim, ao analisar os estudos e observando os critérios de rigor e relevância foram encontradas três pesquisas que podem ser classificadas como pesquisas **indesejadas,** Pallathadka et al. (2021), Silva et al. (2019), Santos et. al. (2019), pois não apresentaram, descrito nos artigos, algum dos componentes metodológicos necessários para análise da *Data Analytics*, consequentemente não possuem sustentação teórica ou metodológica.

Nenhuma pesquisa foi considerada leviana.

As pesquisas que apresentaram todos os componentes: enquadramento do problema (incluindo a definição do termo evasão), teoria, modelagem e protocolo experimental foram consideradas **pesquisas necessárias**, dentre elas podemos citar os estudos realizados por Delen

(2010) e Rovira, Puertas e Igual (2017) e em especial o estudo realizado por Palácios e colaboradores (2021), onde os autores replicam um estudo realizado em uma outra instituição, demonstrando a aplicação prática de um artefato computacional, demonstrando que é possível diminuir a distância entre a teoria (academia) e a prática.

As demais pesquisas foram consideradas **pesquisas autocentradas**, apenas pelo fato de não apresentarem a definição do problema.

4.4 Mapa Mental

Após o estudo sistemático e a análise baseada na estrutura da *Design Science Research* foi possível elaborar um Mapa Mental (figura 14) para demonstrar de forma dinâmica como os diversos pontos se relacionam tendo como ponto central a evasão no Ensino Superior.

A partir do tema central "Evasão no Ensino Superior" encontramos os atores que estão envolvidos, os fatores que podem influenciar na decisão de abandono/permanência e as possíveis consequências deste fenômeno (pessoais, institucionais e governamentais). Estas questões geram inúmeros dados que podem ser analisados quantitativamente através de abordagens de *Data Analytics* (descritiva, diagnóstica, preditiva e prescritiva). Cada modelo implicará em determinado resultado de acordo com o seu objetivo: as causas, características dos alunos que evadem, quais variáveis são as variáveis preditivas ou aquelas que melhor explicam o fenômeno. Cada modelo deve cumprir critérios de rigor e relevância para a sua construção, conforme a *Design Science Research*, sendo assim as formas de validação e as métricas (representada pela acurácia) são essenciais para atender a esses critérios. Os trabalhos, em especial os prescritivos, permitem elaborar estratégias de prevenção com vistas a mitigar os efeitos do fenômeno da evasão.

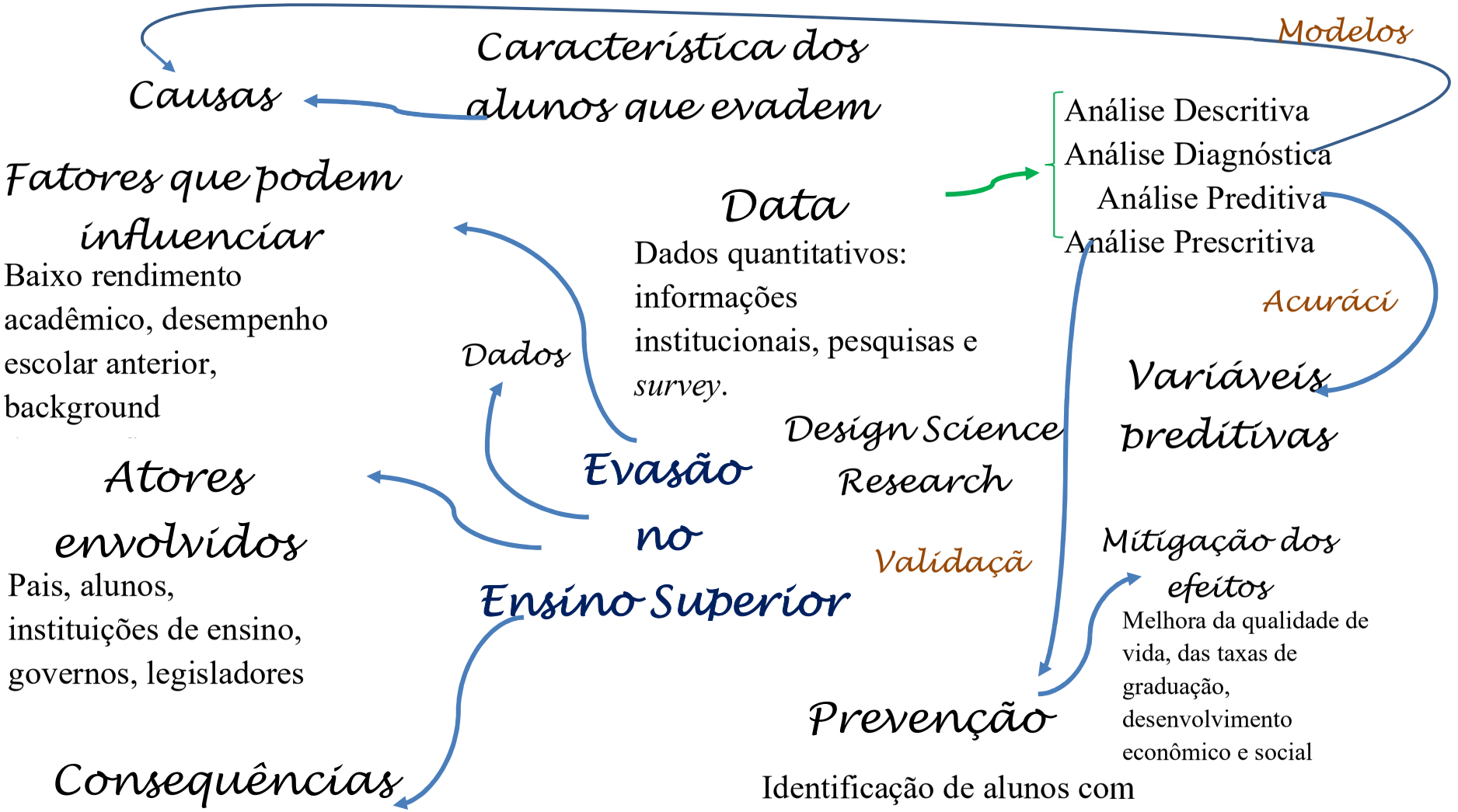


Figura 14 – Mapa Mental

5 CONCLUSÃO

O presente trabalho teve como objetivo investigar como a abordagem da *Data Analytics* é empregada nos estudos sobre o fenômeno da evasão no Ensino Superior, nos cursos de Bacharelado e Licenciatura, realizados na modalidade presencial. Esta investigação desenvolveu uma estrutura da *Design Science Research* adaptada para *Data Analytics* para analisar os textos que foram selecionados através de uma Revisão Sistemática.

Os resultados obtidos foram satisfatórios, 226 artigos retornaram a partir dos critérios de busca adotados e mesmo com as restrições muitos textos acabaram por ser descartados por fugirem do escopo desta pesquisa. Ainda assim, 42 publicações foram selecionadas para aplicação da análise metodológica.

Mesmo considerando uma proporção de aproximadamente 5% de aproveitamento com relação aos resultados iniciais, a qualidade dos artigos selecionados foi relevante para a próxima etapa da investigação. Cabe ressaltar que um viés desta revisão foi a execução feita por apenas uma revisora, o que dificultou a ampliação das buscas, revisões e discussões quanto aos resultados obtidos.

Quanto à análise da abordagem de *Data Analytics* utilizada nos estudos baseada na estrutura da DSR, pode-se observar que todos os estudos utilizam uma metodologia de análise quantitativa de dados, mas somente em poucos casos essa metodologia é definida. Os estudos em sua maioria utilizam dados institucionais, coletados no ato da matrícula e alimentados com informações acadêmicas e de desempenho dos alunos. Outros utilizam dados de pesquisas nacionais que são coletadas e armazenadas em grandes conjuntos de dados e disponibilizados para os pesquisadores e alguns fazem uso de *surveys* na tentativa de agregar informações e melhorar o desempenho dos modelos que estão sendo construídos.

A exceção foram as pesquisas feitas por Santos et al (2019), que não informa as variáveis utilizadas para a criação dos artefatos, Silva et al (2019) cujas variáveis não puderam ser enquadradas na estrutura de análise e de Pallathadka et al (2021) que não informaram as métricas de validação dos modelos.

Ao analisarmos o componente Enquadramento do Problema, observamos que 68% das tarefas de *Data Analytics* são de análises preditivas, o modelo mais utilizado para esta tarefa é o de classificação, sendo utilizada por 59% dos estudos. Observou-se que o uso da tarefa de agrupamento e de redimensionamento dos dados estão associados às tarefas de classificação.

Apesar do uso de modelos individuais serem empregados na maioria dos estudos, os autores que fizeram uso de modelos *ensembles* alegam que possuem um desempenho melhor. E sobre a definição do problema, somente 15 trabalhos definem o termo evasão.

A análise da Teoria revelou que as informações acadêmicas, informações sobre a educação anterior e o desempenho acadêmico são as mais utilizadas para a construção dos modelos. Levando-as em consideração, podemos fazer uma relação com os estudos de Vicent Tinto que aponta a Integração Acadêmica como um fator importante a ser investigado no que diz respeito às decisões do aluno de evadir ou permanecer em seus cursos ou instituições de ensino.

A Modelagem ou "Engenharia de Modelo" foi investigada em duas etapas: a primeira no que diz respeito às quais variáveis foram utilizadas para a construção do modelo, ou seja, como as informações relacionadas na teoria foram traduzidas na fase de construção do artefato. Foram identificadas cinco categorias: background demográfico, financeira, desempenho/informação escolar anterior, admissão, informações/desempenho acadêmico. O uso combinado de background demográfico, desempenho/informação escolar anterior e Informações/desempenho acadêmico foram as mais adotadas nos estudos. Gênero e idade são as variáveis demográficas mais utilizadas, no entanto, nenhum estudo relatou o gênero como preditor de evasão. Quanto a definição do alvo, ou seja, e o que é esperado prever, 11 estudos trabalharam com classes binárias, cuja variável dependente foi a situação final do aluno (evadiu/não evadiu).

O último componente analisado foi o Protocolo Experimental de Avaliação. Para efeito de organização foram criados 8 grupos para tentar agrupar os estudos, esses dizem respeito a como o modelo é construído e de que forma as variáveis são inseridas, quantos conjuntos de dados são utilizados. Foram encontrados 17 artigos pertencentes ao Grupo 1, aquele que trabalha com apenas um conjunto de dados nele foram observados 8 métodos de validação, 53% dos trabalhos utilizam a validação cruzada e 33% o método *holdout* (divisão em treino e teste). No que diz respeito às métricas de avaliação dos modelos, a mais utilizada foi a Acurácia, presente em 61% dos trabalhos, seguida pela Análise de Sensibilidade e de Especificidade. Os resultados obtidos no geral relatam o desempenho acadêmico anterior (notas do Ensino Médio), resultados de testes de admissão e desempenho acadêmico como fatores mais preditivos. Isto corrobora com a hipótese desta pesquisa de que dados institucionais podem ser utilizados para encontrar indicadores educacionais com forte valor preditivo.

A partir deste estudo foi possível elaborar um Mapa Mental que permite verificar os resultados deste estudo de uma forma dinâmica.

5.1 Resultados Alcançados

Ao iniciar este trabalho alguns resultados eram esperados e durante o desenvolvimento da pesquisa foi possível alcançá-los conforme demonstrado a seguir:

Os estudos acerca do fenômeno da evasão no ensino superior através da abordagem da *Data Analytics*, foram identificados e organizados. A literatura está repleta de pesquisas que utilizam análise de dados quantitativos para tratar o fenômeno da evasão, no entanto, nem todos estão classificados como estudos de *Data Analytics*.

Como dito anteriormente neste trabalho, a método da *Design Science Research* pode ser aplicado em diversas áreas. E ao elaborar uma metodologia para a análise dos artigos selecionados este estudo contribuiu com a ampliação e difusão da DSR aplica à área Educacional especialmente em pesquisas acadêmicas a respeito do fenômeno da evasão de alunos no Ensino Superior.

Ao reunir e organizar as pesquisas, foi possível demonstrar quais modelos, métricas, variáveis são utilizadas e quais resultados são alcançados. Este conjunto de informações é essencial para subsidiar a elaboração de Políticas Públicas que possam mitigar os problemas associados ao fenômeno da evasão de estudantes no ensino superior. Desta forma quando um estudo aponta que muitos alunos evadem por não terem condições financeiras para continuar em seus estudos, cabe aos legisladores e governantes criar políticas de permanência destinando verbas direta ou indiretamente através de financiamento estudantil. Da mesma forma, quando os estudos apontam que o um desempenho acadêmico é um fator relevante para a permanência do aluno, as instituições de ensino superior podem elaborar ações como programas de tutoria para auxiliar os alunos com baixo rendimento, aumentando suas chances de permanência e de diplomação.

Por último, este trabalho agrega conhecimento ao campo das Humanidades Digitais a partir da organização das pesquisas, onde pode-se verificar a aplicação de métodos computacionais para a solução do fenômeno da evasão, ao fornecer subsídios para a elaboração de políticas públicas e, ao concatenar os principais termos relacionados a este fenômeno, através da criação de um mapa mental que servirá como metadado para armazenamento e recuperação

deste conhecimento, para que outros pesquisadores o utilizem como ponto de partida para suas pesquisas.

5.2 Considerações Finais

De acordo com os resultados obtidos é possível concluir que os objetivos desta pesquisa foram alcançados. A metodologia da DSR foi útil para a análise dos textos, possibilitando a identificação das abordagens de *Data Analytics* presentes nos trabalhos investigados. Sendo assim fica a contribuição para que futuras pesquisas se utilizem desta metodologia para análise de estudos científicos, especialmente no que diz respeito à criação de artefatos computacionais.

O estudo demonstra que é possível implementar modelos capazes de prever a evasão de alunos mitigando os efeitos deste fenômeno no Ensino Superior. No entanto, não é possível escolher um modelo a ser implementado, pois é preciso levar em consideração as características institucionais e regionais de cada instituição, embora sejam passíveis de algum nível de generalização. Ainda nesta perspectiva, é preciso que mais pesquisas esclareçam em números os efeitos positivos de se prever a evasão de alunos antecipadamente. Os trabalhos apontam os melhores modelos, sugerem soluções para o problema (e.g. programas de tutoria, ajuda financeira) mas não informam quantos alunos foram beneficiados por tais programas. Olaya et al. (2020) analisam um programa de apoio acadêmico de uma Universidade Chilena que existe desde 2012, no entanto, não fornece os números de alunos que permaneceram/concluíram com o auxílio do programa.

Essa conclusão está em consonância com a *Design Science Research,* que segundo Dresch, Lacerda e Antunes (2015), não está em busca de uma solução ótima, mas sim para uma solução que seja a melhor para cada situação.

Esse estudo aponta para futuras pesquisas, como ampliação dos critérios de busca para tentar encontrar outras técnicas aplicadas à predição da evasão, como o uso de Regras de Associação, relatada discretamente no estudo feito por Berka e Marek (2020), a utilização do algoritmo K-NN como agrupamento, e não somente como técnica de redimensionamento do conjunto de dados, e o uso de Algoritmo Genético, como no estudo feito por Kalles e Pierrakeas (2006), mas que não foi utilizado nesta pesquisa por ser aplicado na educação à distância.

Outra questão que pode ser investigada é a utilização de fontes de dados externas como uso de redes sociais, como citado por Hirave et al. (2018), que argumenta que as plataformas de redes sociais (e.g. Twitter, Facebook, LinkedIn) fornecem informações importantes uma vez

que as comunidades de aprendizagem as utilizam para se expressar; e explorada também no trabalho realizado por Assis et al. (2021) que propõem investigar os fatores que levam os alunos a não se graduar utilizando as redes sociais como parâmetros de medição.

6 REFERÊNCIAS

ALBACO, S. et al. **Influence of Mathmatics in The Desertion of Higher Education**. Journal of Advances in Mathematics, vol 16. 2019. DOI https://doi.org/10.24297/jam.v16i0,8249

ALBAN, M.; MAURICIO, D. **Predicting University Dropout through Data Mining: A Systematic Literature**. Indian Journal of Science and Technology, Vol 12(4), DOI: 10.17485/ijst/2019/v12i4/139729, January 2019.

ALJOHANI, O. **A Comprehensive Review of the Major Studies and Theoretical Models of Student Retention in Higher Education**. Higher Education Studies, Vol. 6, Nº 2. 2016.

ALKHASAWNEH, R.; HARGRAVES, R. H. **Developing a Hybrid Model to Predict Student First Year Retention in STEM Disciplines Using Machine Learning Techniques**. Journal of STEM Education: Innovations and Research, v15 n3 p35-42 Oct-Dec 2014

ANDIFES. **Diplomação, Retenção e Evasão nos cursos de Graduação em Instituições de Ensino Superior Pública**s. 1996. Disponível em < http://www.andifes.org.br/diplomacao-retencao-e-evasao-nos-cursos-de-graduacao-em-instituicoes-de-ensino-superior-publicas/> Acesso em 18/09/19.

ASSIS et al. **Frequent pattern mining augmented by social network parameters for measuring graduation and dropout time factors:** A case study on a production engineering course. Socio-Economic Planning Sciences, Volume 81, 2022, ISSN 0038-0121, https://doi.org/10.1016/j.seps.2021.101200.

ASTIN, A. W. **Student Involvement: A Developmental Theory for Higher Education**. Journal of College Student Development. 1984.

BEAN, J. P. **Dropouts and Turnover: The Synthesis and Test of a Causal Model of Student Attrition. Research in Higher Education**. Vol. 12, Nº 2. 1980.
_____. **Interaction Effects Based on Class Level in a Explanatory Model of College Student Dropout Syndrome**. American Education Research Journal, Spring 1985, vol. 22, n.º 1, p. 35-64.

BEAULAC, C.; ROSENTHAL, J.S. **Predicting University Students' Academic Success and Major Using Random Forests**. *Res High Educ* **60,** 1048–1064 (2019). https://doi.org/10.1007/s11162-019-09546-y

BIOLCHINI, J.; MIAN, P. G. N.; TRAVASSOS, G. H. **Systematic Review in Software Engineering. Technical Report**. COPPE/UFRRJ. 2005.

BOTELHO, L. L. R.; CUNHA, C. C. de A.; · MACEDO, M. **O método da revisão integrativa nos estudos organizacionais**. Gestão e Sociedade. Belo Horizonte, v.5, n. 11, p. 121-136 · maio-ago. 2011 · ISSN 1980-5756. Disponível em: Acesso em: 04/03/2021.

CABRERA, A. F.; NORA, A; CASTAÑEDA, M. B. **The role of finances in the persistence process:** a structural model. Research in HigherEducation, 33(5), pp. 571-593. http://dx.doi.org/10.1007/ BF00973759, 1992.

CASANOVA, J. R. et al. **Factors that determine the persistence dropout of university students**. Psicothema. Vol. 30. no 4, p. 408-414, 2018.

CASTRO, A. K. S. S.; TEIXEIRA, M. A. P. **Evasão universitária:** modelos teóricos internacionais e panorama das pesquisas no Brasil. Psicologia Argumento. Curitiba, v. 32, n. 79, p. 9-17, Supl. 1-2014.

CHARITOPOULOS, A.; RANGOUSSI, M.; KOULOURIOTIS, D. **On the Use of Soft Computing Methods in Educational Data Mining and Learning Analytics Research:** a Review of Years 2010–2018. International Journal of Artificial Intelligence in Education. 2020. https://doi.org/10.1007/s40593-020-00200-8

COSTA, O. S.; GOUVEIA, L. B. **Modelos de Retenção de Estudantes:** abordagens e perspectivas. REAd – Porto Alegre – Vol. 24 – No3 – Setembro/Dezembro 2018 – p. 155 – 182.

DELEN, D. **A comparative analysis of machine learning techniques for student retention management.** Decision Support Systems. Volume 49, Issue 4, 2010, Pages 498-506, ISSN 0167-9236, https://doi.org/10.1016/j.dss.2010.06.003.

DRESCH, A.; LACERDA, D. P.; ANTUNES J. J. A. V. **Design Science Research. Método de Pesquisa para Avanço da Ciência e Tecnologia**. Bookman, 2015

ENGLE, J.; TINTO, V. **Moving Beyond Access College Sucess for Low-Income, First-Generation Students**. The Pell Institute. 2008.

ENRAGAL, A.; HADDARA, M. **Design Science Resarch:** Evaluation in the Lens of Big Data Analytics. Systems, 2019.

FREITAS JUNIOR et al. **Big Data e Gestão do Conhecimento**: definições e direcionamentos de pesquisa. Revista Alcance. – Eletrônica – vol. 23 – n. 4 – out./dez. 2016

HAN, J., KAMBER, M. AND PEI, J. **Data Mining: Concepts and Techniques (3rd ed.)**. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2011.

HEVNER, A. et al. **Design Science in Information Systems Research**. MIS Quarterly. Vol. 28, n. 1/March, 2004.

HIRAVE, T. et al. **Data Analytics Research Agenda: E-Learning & Its Integration With Other Platforms.** Fourth International Conference on Computing Communication Control and Automation. 2018.

IAM-ON, N.; BOONGOEN, T. **Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings**. International Journal of Machine Learning and Cybernetics. 8. 10.1007/s13042-015-0341-x, 2015.

KALLES, D; PIERRAKEAS, C**. Using Genetic Algorithms and Decision Trees for *a posteriori* Analysis and Evaluation of Tutoring Practices based on Student Failure Models**. In: Maglogiannis, I., Karpouzis, K., Bramer, M. (eds) Artificial Intelligence Applications and Innovations. AIAI 2006. IFIP International Federation for Information Processing, vol 204. Springer, Boston, MA . https://doi.org/10.1007/0-387-34224-9_2

KITCHENHAM, B.; CHARLES, S. **Guidelines for performing Systematic Reviews in Software Engineering**. Version 2.3. Technical Report, Keele University and University of Durham, 2007.

LACAVE, C.; MOLINA-DÍAZ, A; CRUZ-LEMUS, J. L**earning Analytics to identify dropout factors of Computer Science studies through Bayesian networks**. Behaviour & Information Technology. 1-15. 10.1080/0144929X.2018.1485053. 2018.

LEON, F. L. L.; MENEZES-FILHO, N. A. **REPROVAÇÃO, AVANÇO E EVASÃO ESCOLAR NO BRASIL.** Pesquisa e Planejamento Econômico - PPE, v.32, n.3, dez 2002

LIMA, F. S.; ZAGO, N. **EVASÃO NO ENSINO SUPERIOR: TENDÊNCIAS E RESULTADOS DE PESQUISA**. XI ANPED SUL. Curitiba, 2016. Disponível em < Introdução: Justificativa, delimitação e relevância do problema (pessoal acadêmica e social): Justificativa acerca da relevância pessoal (trajetória e história de vida na formação e profissional de forma metodologicamente argumentada e problematizada (mi (ufpr.br)> Acesso em 21/05/2021.

LIZ-DOMÍNGUEZ, M. et al. **Systematic Literature Review of Predictive Analysis Tools in Higher Education.** Applied Science. 2019.

LOZANO, J. M.; VIEITES, A. R.; CALABUIG, P. B. **Aplicación de árboles de clasificación a la detección precoz de abandono en los estudios universitarios de administración y dirección de empresas**. Revista Electrónica de Comunicaciones y Trabajos de ASEPUMA. 18. 177-201. 10.24309/recta. 2017.

MANHÃES, L. M. B; CRUZ, S. M. S. **Predição do Desempenho Acadêmico de Alunos da Graduação Utilizando Mineração de Dados**. XIX Simpósio de Pesquisa Operacional e Logística da Marinha. Rio de Janeiro, RJ, Brasil – 06 a 08 de novembro de 2019.

MINISTÉRIO DA EDUCAÇÃO. **Censo da Educação Superior 2019**: Divulgação de Resultados. Instituto Nacional de Estudos e Pesquisas Anísio Teixeira (INEP). Brasília - DF, 2020.

NEIVA, F. W.; SILVA, R. L. S. **Revisão da Literatura em Ciência da Computação** – Um Guia Prático. UFJF. MG, 2016.

POULSEN, C. J. B.; BANDEIRA, D. L. **Um Estudo Exploratório dos Regimes Acadêmicos Adotados por Instituições Privadas de Ensino Superior no Brasil**. XXXVIII Encontro ANPAD, Rio de Janeiro, 2014.

PROVOST, F.; FAWCETT, T. **Data Science for Business:** What You Need to Know about Data Mining and Data-Analytic Thinking. (1st ed.). O'Reilly Media, Inc, 2013.

RASTROLLO-GUERRERO, J.; GÓMEZ-PULIDO, J. A.; DURÁN-DOMÍNGUEZ, A. **Analyzing and Predicting Students' Performance by Means of Machine Learning:** A Review. Applied Science. 2020.

ROVIRA, S.; PUERTAS, E.; Igual, L. **Data-driven System to Predict Academic Grades and Dropout**. Plos - PloS one, 2017.

RUSSELL, S., NORVIG, P. **Artificial Intelligence** - A Modern Approach. 2009. 3ªed. Disponível em https://cs.calvin.edu/courses/cs/344/kvlinden/resources/AIMA-3rd-edition.pdf. Acesso em 28/05/2021.

SPADY, W. G. **Dropouts from Higher Education:** An Interdisciplinary Review and Synthesis. Interchange, 1970. 1, 64-85.

SPADY, W. G. **Dropouts from Higher Education:** Toward an Empirical Model. Interchange, 2, 68-62. 1971.

SALGANIK, M. J. **Bit by Bit**. Princeton University. 2018. ISBN 978-0-691-15864-8.

SARRA, A.; FONTANELLA, L.; DI ZIO, S. **Identifying Students at Risk of Academic Failure Within the Educational Data Mining Framework**. *Soc Indic Res* **146,** 41–60 (2019). https://doi.org/10.1007/s11205-018-1901-8

SILVA, G. P. **Análise da evasão no ensino superior: uma proposta de diagnóstico de seus determinantes**, 2013.

SOUZA, N. A.; MONTEIRO, A. J. **Os docentes da Universidade Federal do Ceará e a utilização de alguns dos recursos do sistema integrado de gestão de atividades acadêmica (SIGAA)**. Ensaio: aval. pol. públ. Educ., Rio de Janeiro, v. 23, n. 88, p. 611-630, jul./set. 2015. Disponível em O SIGAA na formação discente: Estudo de caso nos cursos da UFRR (nucleodoconhecimento.com.br)

STALLIVIERI, L. **O SISTEMA DE ENSINO SUPERIOR DO BRASIL CARACTERÍSTICAS, TENDÊNCIAS E PERSPECTIVAS.** Universidade Caxias do Sul. In book: Educación superior em América Latina y el Caribe: Sus estudiantes hoy (pp.79-100), 2007. Disponível em ,(PDF) O SISTEMA DE ENSINO SUPERIOR DO BRASIL CARACTERÍSTICAS, TENDÊNCIAS E PERSPECTIVAS (researchgate.net)> Acesso em 20/05/2021.

TINTO, V. **Dropout from Higher Education:** A Theoretical Synthesis of Recent Research, Review of Education Research, vol. 45, nº 1. P. 89-125, 1975.

ULRICHI, P. et al. **Data Analytics Systems and SME type - a Design Science Approach.** Procedia Computer Science. Volume 126, 2018, Pages 1162-1170.

VILORIA, A. et al. **Integration of Data Technology for Analyzing University Dropout**. Procedia Computer Science, Volume 155, Pages 569-574, ISSN 1877-0509, 2019. https://doi.org/10.1016/j.procs.2019.08.079.

ZOLTOWSKI, A. P. C. et al. **Qualidade Metodológica das Revisões Sistemáticas em Periódicos de Psicologia Brasileiros**. Psicologia: Teoria e Pesquisa. Jan-Mar 2014, vol. 30, n. 1. pp. 97-104.

Trabalhos Revisados

ACERO, A.; ACHURY, J.; MORALES PIÑERO, J. **University Dropout: A Prediction Model for an Engineering Program in Bogotá**, Colombia, 2019.

AGUILAR-GONZALEZ, S.; PALAFOX L. **Prediction of Student Attrition Using Machine Learning.** In: Martínez-Villaseñor L., Batyrshin I., Marín-Hernández A. (eds) Advances in Soft Computing. MICAI 2019. Lecture Notes in Computer Science, vol 11835. Springer, 2019. https://doi.org/10.1007/978-3-030-33749-0_18

AL-SUDANI, S.; PALANIAPPAN, R. **Predicting students' final degree classification using an extended profile**. Educ Inf Technol 24**,** 2357–2369. 2019. https://doi.org/10.1007/s10639-019-09873-8

ALVAREZ, N. L.; CALLEJAS, Z.; GRIOL, D. **Predicting Computer Engineering Students' Dropout in Cuban Higher Education with Pre-Enrollment and Early Performance Data.** Journal of Technology and Science Education, v10, n2, p. 241-258. 2020

AHMAD TARMIZI, S.S. et al. **A Case Study on Student Attrition Prediction in Higher Education Using Data Mining Techniques**. In: Berry M., Yap B., Mohamed A., Köppen M. (eds) Soft Computing in Data Science. SCDS 2019. Communications in Computer and Information Science, vol 1100. Springer, Singapore. https://doi.org/10.1007/978-981-15-0399-3_15

NANDESHWAR, A.; MENZIES, T.; NELSON, A. **Learning patterns of university student retention**. Expert Systems with Applications. Volume 38, Issue 12, 2011, p. 14984-14996, ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2011.05.048.

BEAULAC, C.; ROSENTHAL, J. S. **Predicting University Students' Academic Success and Major Using Random Forests**. Res High Educ. 60**,** 1048–1064. 2019. https://doi.org/10.1007/s11162-019-09546-y

BELLO, F. A. et al. **Using machine learning methods to identify significant variables for the prediction of first-year Informatics Engineering students dropout**. 39th International Conference of the Chilean Computer Science Society (SCCC)*, 2020, pp. 1-5, doi: 10.1109/SCCC51225.2020.9281280.*

BERKA, P.; MAREK, L. **Who tend to stay and who tend to leave?** Studies in Educational Evaluation, Volume 70, 2021, 100999, ISSN 0191-491X, https://doi.org/10.1016/j.stueduc.2021.100999.

BILQUISE G.; ABDALLAH S.; KOBBAEY T. **Predicting Student Retention Among a Homogeneous Population Using Data Mining**. In: Hassanien A., Shaalan K., Tolba M. (eds) Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2019. AISI 2019. Advances in Intelligent Systems and Computing, vol 1058. Springer, 2020. https://doi.org/10.1007/978-3-030-31129-2_4

GRAY, C. C.; PERKINS, D. **Utilizing early engagement and machine learning to predict student outcomes**. Computers & Education, Volume 131, p. 22-32, ISSN 0360-1315. 2019. https://doi.org/10.1016/j.compedu.2018.12.006.

CARDONA, T. A.; CUDNEY E. A**., Predicting Student Retention Using Support Vector Machines.** Procedia Manufacturing, Volume 39, 2019, p. 1827-1833, ISSN 2351-9789, https://doi.org/10.1016/j.promfg.2020.01.256.

CHEN, Y.; JOHRI, A.; RANGWALA, H. **Running out of STEM: a comparative study across STEM majors of college students at-risk of dropping out early**. 2018. DOI - 10.1145/3170358.3170410.

CONTINI, D.; SALZA, G. **Too few university graduates.** Inclusiveness and effectiveness of the Italian higher education system. Socio-Economic Planning Sciences, Volume 71, 2020, 100803, ISSN 0038-0121, https://doi.org/10.1016/j.seps.2020.100803.

DELEN, D. **A comparative analysis of machine learning techniques for student retention management.** Decision Support Systems. Volume 49, Issue 4, 2010, Pages 498-506, ISSN 0167-9236, https://doi.org/10.1016/j.dss.2010.06.003

SARKER, F.; TIROPANIS, T.; DAVIS, H. C. **Linked data, data mining and external open data for better prediction of at-risk students**, *2014 International Conference on Control, Decision and Information Technologies (CoDIT)*, 2014, pp. 652-657, doi: 10.1109/CoDIT.2014.6996973.

GIL, P.D. et al. **A data-driven approach to predict first-year students' academic success in higher education institutions**. Educ Inf Technol 26**,** 2165–2190. 2021. https://doi.org/10.1007/s10639-020-10346-6

HEGDE, V.; PRAGEETH, P. P. **Higher education student dropout prediction and analysis through educational data mining**. *2nd International Conference on Inventive Systems and Control (ICISC)*, 2018, pp. 694-699, doi: 10.1109/ICISC.2018.8398887.

HUO, H. D. et al. **Predicting Dropout for Nontraditional Undergraduate Students: A Machine Learning Approach.** Journal of College Student Retention: Research, Theory & Practice. 2020. https://doi.org/10.1177/1521025120963821

IAM-ON, N.; BOONGOEN, T. **Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings**. Int. J. Mach. Learn. & Cyber. 8**,** 497–510. 2017. https://doi.org/10.1007/s13042-015-0341-x

KUZILEK, J.; ZDRAHAL, Z.; FUGLIK, V. **Student success prediction using student exam behaviour**, Future Generation Computer Systems, Volume 125, 2021, p. 661-671, ISSN 0167-739X. https://doi.org/10.1016/j.future.2021.07.009.

JAYARAMAN, J. D.; GERBER, S.; GARCIA, J. **Supporting Minority Student Success by using Machine Learning to Identify At-Risk Students.** In Michel C. Desmarais, Collin F. Lynch, Agathe Merceron, Roger Nkambou, editors, *Proceedings of the 12th International Conference on Educational Data Mining, EDM 2019, Montréal, Canada, July 2-5, 2019*. International Educational Data Mining Society (IEDMS), 2019. [doi]

KILIAN, P.; LOOSE, F.; KELAVA, A. **Predicting Math Student Success in the Initial Phase of College With Sparse Information Using Approaches From Statistical Learning**. Frontiers in Education, Volume 5, DOI - 10.3389/feduc.2020.502698.

KEMPER, L.; VORHOFF, G.; WIGGER B.U. **Predicting student dropout:** A machine learning approach. European Journal of Higher Education. 2020. https://doi.org/10.1080/21568235.2020.1718520

MALDONADO, S. et al. **Redefining profit metrics for boosting student retention in higher education.** Decision Support Systems, Volume 143, 2021, 113493, ISSN 0167-9236, https://doi.org/10.1016/j.dss.2021.113493.

OLAYA, D. et al. **Uplift Modeling for preventing student dropout in higher education.** Decision Support Systems, Volume 134, 2020, 113320, ISSN 0167-9236, https://doi.org/10.1016/j.dss.2020.113320.

PALACIOS, C. A. et al. **Discovery for Higher Education Student Retention Based on Data Mining: Machine Learning Algorithms and Case Study in Chile**. Entropy (Basel). 2021 Apr 20;23(4):485. doi: 10.3390/e23040485. PMID: 33923879; PMCID: PMC8072774.

PATTERSON, J. et al. **Integrating Lean Six Sigma and Data Analytics to Improve Student** Bachelor's degree student dropouts: **Retention.** Proceedings of the 2020 IISE Annual Conference, 2020.

PERALTA, B. et al. **A causal modelling for desertion and graduation prediction using Bayesian networks**: **A Chilean case.**

PÉREZ, B.; CASTELLANOS, C.; CORREAL, D. **Predicting Student Drop-Out Rates Using Data Mining Techniques:** A Case Study. In: Orjuela-Cañón A., Figueroa-García J., Arias-Londoño J. (eds) Applications of Computational Intelligence. ColCACI 2018. Communications in Computer and Information Science, vol 833. Springer, 2018. https://doi.org/10.1007/978-3-030-03023-0_10

PEREZ, A. **Comparative Analysis of Prediction Techniques to Determine Student Dropout:** Logistic Regression vs Decision Trees, 2018. DOI - 10.1109/SCCC.2018.8705262.

SANTOS, K. J. de O. et al. **Supervised Learning in the Context of Educational Data Mining to Avoid University Students Dropout,** IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*,* 2019, pp. 207-208, doi: 10.1109/ICALT.2019.00068.

SILVA, P. M. da, et al. **Ensemble Regression Models Applied to Dropout in Higher Education***,* 8th Brazilian Conference on Intelligent Systems (BRACIS)*,* 2019, pp. 120-125, doi: 10.1109/BRACIS.2019.00030.

TAMPAKAS, V. et al. **Prediction of Students' Graduation Time Using a Two-Level Classification Algorithm**. In: Tsitouridou M., A. Diniz J., Mikropoulos T. (eds) Technology and Innovation in Learning, Teaching and Education. TECH-EDU 2018. Communications in Computer and Information Science, vol 993. Springer, 2019.

TSAI, S, C. et al*.* **Precision education with statistical learning and deep learning:** a case study in Taiwan. Int J Educ Technol High Educ 17**,** 12. 2020. https://doi.org/10.1186/s41239-020-00186-2

ULLAH, M. A. et al. **Predicting Factors of Students Dissatisfaction for Retention**. In: Abraham A., Dutta P., Mandal J., Bhattacharya A., Dutta S. (eds) Emerging Technologies in Data Mining and Information Security. Advances in Intelligent Systems and Computing, vol 755. Springer, Singapore, 2019. https://doi.org/10.1007/978-981-13-1951-8_45

VILA, D. et al. **Detection of Desertion Patterns in University Students Using Data Mining Techniques:** A Case Study. In: Botto-Tobar M., Pizarro G., Zúñiga-Prieto M., D'Armas M., Zúñiga Sánchez M. (eds) Technology Trends. CITT 2018. Communications in Computer and Information Science, vol 895. Springer, 2019. https://doi.org/10.1007/978-3-030-05532-5_31

VILORIA, A. et al. **Integration of Data Technology for Analyzing University Dropout**. Procedia Computer Science, Volume 155, Pages 569-574, ISSN 1877-0509, 2019. https://doi.org/10.1016/j.procs.2019.08.079.

WAN YAACOB, W.F. et al. **Predicting Student Drop-Out in Higher Institution Using Data Mining Techniques.** Journal of Physics: Conference Series, Volume 1496, International Conference of Mathematics, Statistics and Computing Technology 2019 28 October 2019, Aula Timur, Institut Teknologi Bandung, Indonesia.

ZEINEDDINE, H.; BRAENDLE, U.; FARAH, A. **Enhancing prediction of student success**: Automated machine learning approach. Computers & Electrical Engineering, Volume 89, 2021, 106903, ISSN 0045-7906. https://doi.org/10.1016/j.compeleceng.2020.106903.

ZHANG L.; RANGWALA H. **Early Identification of At-Risk Students Using Iterative Logistic Regression**. In: Penstein Rosé C. et al. (eds) Artificial Intelligence in Education. AIED 2018. Lecture Notes in Computer Science, vol 10947. Springer, 2018. https://doi.org/10.1007/978-3-319-93843-1_45

APÊNDICE A
PROTOCOLO DE REVISÃO SISTEMÁTICA

1. Formulação da Pergunta

1.1 **Foco da Pergunta**: Identificar estudos que se utilizam da abordagem da Data Analytics para mitigar os efeitos da evasão de alunos no Ensino Superior.

1.2 **Amplitude e qualidade da pergunta**:

- ***Problema:*** a evasão no Ensino Superior é um fenômeno que preocupa instituições em todo o mundo. O objetivo é identificar os estudos que estão sendo realizados para evitar o abandono de alunos que buscam concluir o bacharelado.

- ***Pergunta:*** μ0: É possível desenvolver uma estrutura de Design Science Research para analisar estudos sobre evasão que utilizam abordagens de *Data Analytics*? μ1: Como a literatura faz o enquadramento analítico do processo de abandono (Churn, Cluster ou Detecção de Anomalia? μ2: Quais teorias são utilizadas para construir o modelo e quais variáveis são mais utilizadas para transformar essa teoria dentro do modelo? μ3: Como os estudos realizam o protocolo experimental de avaliação do modelo construido?

- ***Palavras-chaves e Sinônimos***:

  - Dropout
  - Retention
  - Higher Education, College, University
  - Machine Learning
  - Predicting

- ***Intervenção:*** Análise do estudo segundo a Design Science Research.

- ***Controle***: Nenhum.

- ***Efeitos***: Identificar trabalhos que utilizem abordagens quantitativas para tratar os dados sobre Evasão do ensino Superior através da Data Analytics.

- ***Medida do Resultado***: Número de estudos selecionados.

- ***População***: Ensino Superior, Curso de Bacharel, modalidade presencial.

- ***Aplicação***: Educação, Machine Learning, Humanidades Digitais.

- ***Projeto Experimental***: Nenhum método estatístico está sendo aplicado.

2. Seleção das fontes

2.1 **Definição de critérios e seleção das fontes**: consulta de artigos utilizando mecanismo de busca na web através do uso de palavras-chaves em plataformas recomendadas (NEIVA; SILVA, 2016).

- ***Método de pesquisa da fonte***: Pesquisa através de busca na web.

- ***String da pesquisa***: (Dropout) AND ("Machine Learning") AND ("Higher Education" OR College OR University) AND (Retention) AND (Preventing OR Predicting)

- ***Lista de fontes***: diversas, nas bases: Scopus, Springer Link, ACM Digital Library, Science Direct, Web of Science e Plos One, através do acesso na Plataforma Capes.

2.2 **Idioma**: Inglês, português e espanhol.

2.3 **Identificação das fontes**:

2.4 **Seleção das fontes após avaliação**: não será aplicada como critério de exclusão, as fontes foram consideradas satisfatórias.

3. Seleção dos estudos

3.1 **Definição dos Estudos**:

- **Definição dos critérios de inclusão e exclusão de estudos**: os trabalhos devem abordar a construção de modelos para mitigar os problemas relacionados a evasão de alunos no Ensino Superior tradicional, ou seja, estudos realizados na Educação a Distância não serão considerados.

- **Definição dos tipos de estudos**: somente artigos revisados pelos pares e pesquisas primárias, excluindo, portanto, surveys e demais estudos de revisão.

- **Procedimento para seleção de estudos**: a *string* de busca deve retornar resultados nas bases pesquisadas na web. Leitura de todos os títulos e abstracts dos estudos encontrados através do mecanismo de busca de acordo com os critérios de inclusão e exclusão. Em seguida todos as introduções e por último os textos serão lidos na íntegra.

3.2 **Execução da Seleção**:

- **Seleção inicial dos estudos**: Auxílio do Excel

- **Avaliação da qualidade dos estudos**: 42 estudos atenderam aos critérios de inclusão e exclusão.

- **Revisão da seleção**: Confirmação dos estudos aprovados.

4. Extração da Informação

4.1 **Definição dos critérios de inclusão e exclusão da informação**: a informação extraída dos estudos deve conter abordagens de Data Analytics, de aprendizagem de máquina ou mineração de dados aplicadas à Educação, especificamente a tentativa de prever os alunos em risco de abandono.

4.2 **Forma de extração dos dados**: ver a Seção 4 (análise do Enquadramento do Problema, Teoria, Modelagem e Protocolo Experimental de Avaliação)

4.3 **Extração dos dados**:

- ***Extração dos resultados objetivos***

  - Identificação do estudo
  - Enquadramento do problema (Abordagem de *Data Analytics*, tipo de problema e definição de evasão.
  - Teoria.
  - Modelagem ou "Engenharia de Modelo:"
  - Variáveis.
  - Protocolo Experimental de Avaliação.

- ***Extração dos resultados subjetivo***s

  - Informação sobre os autores: Não foi aplicado.
  - Abstrações e impressões gerais: Não foi aplicado.

4.4 **Solução de divergência entre os revisores**: "não se aplica"

5. Resumo dos Resultados

5.1 **Cálculo estatístico dos resultados**: não foi utilizado.

5.2 **Tabela para apresentação dos resultados**: os resultados serão apresentados ao longo do texto.

5.3 **Análise sensitiva**: não foi utilizada.

5.4 **Gráficos**: apresentados ao longo do texto.

5.5 **Comentários finais**:

- Número de estudos: Estudos encontrados 226; 42 estudos selecionados.
- Viés da pesquisa, seleção e extração: ver Conclusão
- Variação entre revisores: Não se aplica.
- Aplicação dos resultados: ver Conclusão
- Recomendações: ver Conclusão

APÊNDICE B

RELAÇÃO MÉTRICAS x TÉCNIAS UTILIZADAS

Tabela A – Relação entre as Técnicas e as Métricas

Continua

Técnicas	Métricas
Adaptative Boosting	Taxa de verdadeiro Negativo, Taxa de verdadeiro positivo, AUC, F-Measure
Análise de sobrevivência	AAF - Aalens´S Additive Filter e COX - roportional hazards regression model, discrete-time hazard function, Competing risk (CR)
Árvore de Decisão	Matrix de confusão, Acurácia Geral, Acurácia por Classe, AUC, Taxa de verdadeiro Negativo, Taxa de verdadeiro positivo (sensibilidade), F-Measure ou F1-Score, Kappa, Recall, Precision, desvio padrão, especificidade, valores predito positivo, valores predito negativo, verdadeiros positivos, verdadeiros negativos, falso positivo, falso negativo, taxa de erro, RMSE, k-statistic, Friedman Value, Curva ROC, classificações certas, classificações incorretas, taxa de verdadeiros positivos, taxas de verdadeiros negativos, taxa de falso positivo, taxa de falso negativo.
Bagging (random forest)	Matrix de confusão, Acurática Geral, Acurácia por Classe, Taxa de verdadeiro Negativo, Taxa de verdadeiro positivo, AUC, F-Measure
Bayesian Network	Verdadeiros positivos, verdadeiros negativos, falso positivos, falsos negativos, acurácia, sensibilidade, especificidade, curva ROC, f-measure, classificações certas, classificações incorretas, precision, recall, taxa de verdadeiros positivos, taxas de verdadeiros negativos, taxa de falso positivo, taxa de falso negativo.
Boosting (boosted trees)	Matrix de confusão, Acurácia Geral, Acurácia por Classe
Deep Learning	Acurácia, Taxa de verdadeiro Negativo, Taxa de verdadeiro positivo, AUC, F-Measure
Ensemble Clustering (Técnicas de transformação: PCA, KPCA, LPP, NPE, IsoP, BA, WCT-T (fixed-K), WCT-T (random-k), WTQ-T (fixed-k) WTQ-T (random-k) + classificadores Arvore de Decisão, Naive Bayes, KNN)	Taxa de erro de classificação
Ensemble model (Gradient Boosting)	Especificidade, recall, acurácia, AUC
Ensemble Voting	Acurácia, Taxa de verdadeiro Negativo, Taxa de verdadeiro positivo, AUC, F-Measure, Acurácia por Classe, Kappa, recall, especificidade
Gradient Boosted Tree	Taxa de verdadeiro Negativo, Taxa de verdadeiro positivo, AUC, F-Measure
Information fusion (weighted average)	Matrix de confusão, Acurácia Geral, Acurácia por Classe
Interative Logistic Regression	PRAUC, AUC, F1-socore, Verdadeiros Positivo, Falso Negativo, Falso Positivo, taxa de verdadeiro positivo, taxa de falso positivo
k-meas clustering	Acurácia geral, acurácia por classe, Kappa
KNN	AUC, Acurácia, desvio padrão, sensibilidade, especificidade, valor predito positivo, valor predito negativo, RMSE, k-statistic, Friedman Value, Curva ROC, taxa de verdadeiro positivo, taxa de falso positivo.

Tabela A – Relação entre as Técnicas e as Métricas

Continuação

método em conjunto: Bagging + linear regression; bagging + ridge regression; bagging + robust regression; linear regression+ lasso regression + bagging (usando decision tree) + boosting+random forest, vector regression support e K-nearest neighbors	MAE, MSE, desvio padrão, Kolmogorov-Smirnov e Wilcoxon, boxplot
Modelo linear generalizado	AUC
Naïve Bayes	Acurácia, Taxa de verdadeiro Negativo, Taxa de verdadeiro positivo, AUC, F-Measure, F1-Score ou Measure, Recall, Precision, RMSE, k-statistic, Friedman Value, taxa de erro, taxa de falso negativo, ROC área.
One R (One Rule)	probabilidade de detecção, probabilidade de falso alarme, variância entre PD e PF visto através da validação cruzada
PCA	Acurácia, Especificidade, Sensibilidade
Random Forest	Acurácia, Taxa de verdadeiro Negativo, Taxa de verdadeiro positivo (sensibilidade), AUC, F-Measure ou F1-Score, Kappa, Recall, Precision, verdadeiros positivos, verdadeiros negativos, falso positivo, falso negativo, RMSE, k-statistic, Friedman Value, taxa de erro, taxa de falso negativo, ROC área, boxplot, Matriz de confusão, MDA - Diminuição Média da Acurácia e MDG - Diminuição Média da Gini
Randon Forest e XGBoosting, X-learner, R-Learner, KL Divergence, ED - Euclidean Distance, Chi square, CTS – Tree Structures in Clok – Aplicados para estimar a uplift modeling	Valor Qini, Ganho de Informação
Redes Neurais Artificiais	Matrix de confusão, Acurácia Geral, Acurácia por Classe, AUC, (taxa de verdadeiro positivo) sensibilidade, especificidade, verdadeiros positivos, verdadeiros negativos, falso positivo, falso negativo, precision, recall, f-measure, Curva ROC, classificações certas, classificações incorretas, taxas de verdadeiros negativos, taxa de falso positivo, taxa de falso negativo.
Regra de Associação (LISp-Miner)	Desconhecida.
Regressão Linear	MAE, Kendall
Regressão Logística	Matrix de confusão, Acurática Geral, Acurácia por Classe, Taxa de verdadeiro Negativo, Taxa de verdadeiro positivo (sensibilidade), AUC, F-Measure ou F1-Score, Kappa, Recall, Precision, verdadeiros positivos, verdadeiros negativos, falso positivo, falso negativo, RMSE, k-statistic, Friedman Value
Sistema de recomendação	MAE, Kendall
Support Vector Regression	MAE, Kendall

Tabela A – Relação entre as Técnicas e as Métricas

Continuação

Suporte Vector Machine	Matrix de confusão, Acurácia Geral, Acurácia por Classe, AUC, Taxa de verdadeiro Negativo, Taxa de verdadeiro positivo (sensibilidade), F-Measure ou F1-Score, Kappa, Recall, Precision, desvio padrão, sensibilidade, especificidade, valor predito positivo, valores preditos negativo, verdadeiros positivos, verdadeiros negativos, falso positivo, falso negativo, RMSE, k-statistic, Friedman Value, Curva ROC
XGBoosting	Acurácia, acurácia balanceada, sensibilidade/recall, especificidade, Precision/valor predito positivo, valor predito negativo, F1-score, AUC.