Universidade de Brasília – UnB
Faculdade UnB Gama – FGA
Engenharia de Software

Identificação de fatores que afetam a Evasão no Ensino Superior

Autoras: Amanda Emilly Muniz de Menezes, Letícia Karla Soares Rodrigues de Araújo

Orientador: Profł. Me. Cristiane Soares Ramos

Coorientador: Prof. Dr. Sergio Antonio Andrade de Freitas

Brasília, DF

2022

Amanda Emilly Muniz de Menezes, Letícia Karla Soares Rodrigues de Araújo

Identificação de fatores que afetam a Evasão no Ensino Superior

Monografia submetida ao curso de graduação em Engenharia de Software da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia de Software.

Universidade de Brasília – UnB

Faculdade UnB Gama – FGA

Orientador: Profł. Me. Cristiane Soares Ramos

Coorientador: Prof. Dr. Sergio Antonio Andrade de Freitas

Brasília, DF

2022

Amanda Emilly Muniz de Menezes, Letícia Karla Soares Rodrigues de Araújo
Identificação de fatores que afetam a Evasão no Ensino Superior/ Amanda Emilly Muniz de Menezes, Letícia Karla Soares Rodrigues de Araújo.– Brasília, DF, 2022-
111 p. : il. (algumas color.) ; 30 cm.

Orientador: Profł. Me. Cristiane Soares Ramos Prof. Dr. Sergio Antonio Andrade de Freitas

Trabalho de Conclusão de Curso – Universidade de Brasília – UnB
Faculdade UnB Gama – FGA , 2022.

1. Evasão. 2. Evasão acadêmica. 3. Ensino superior. 4. Graduação. 5. Fatores. 6. Indicador. 7. Previsão. I. Profł. Me. Cristiane Soares Ramos. II Prof. Dr. Sergio Antonio Andrade de Freitas. III. Universidade de Brasília. IV. Faculdade UnB Gama. V. Identificação de fatores que afetam a Evasão no Ensino Superior

CDU 02:141:005.6

Amanda Emilly Muniz de Menezes, Letícia Karla Soares Rodrigues de Araújo

Identificação de fatores que afetam a Evasão no Ensino Superior

Monografia submetida ao curso de graduação em Engenharia de Software da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia de Software.

Trabalho aprovado. Brasília, DF, 5 de maio de 2022:

Profł. Me. Cristiane Soares Ramos
Orientador

Prof. Dr. Sergio Antonio Andrade de Freitas
Coorientador

Profł. Me. Andrea Felippe cabello
Convidado 1

Profž. Dr. John Lenon Cardoso Gardenghi
Convidado 2

Brasília, DF
2022

Amanda Emilly Muniz de Menezes.

Dedico este trabalho à minha mãe, por ter me ensinado a ser a mulher que sou, pelos sacrifícios que fez para criar três meninas sozinha e por seu colo que sempre trouxe conforto e ternura.

Letícia Karla Soares Rodrigues de Araújo.

Dedico a Deus este trabalho, por ter me dado todas as oportunidades para chegar aqui. À minha mãe Glaice, ao meu pai Carlos, ao meu irmão Carlos Júnior, ao meu namorado Gabriel, à minha tia Graça e à minha Tia Glícia que me auxiliaram de todas as formas possíveis para que eu conseguisse chegar a essa etapa da minha vida.

Agradecimentos

Amanda Emilly Muniz de Menezes.

Agradeço aos professores que tive ao longo dessa jornada. Em especial, aos meus orientadores Profł Cristiane e Profž Sérgio. Obrigada por me desafiarem e mostrarem novas aventuras no mundo acadêmico. Agradeço pela confiança e pela sabedoria que concederam nas orientações.

Agradecimentos

Letícia Karla Soares Rodrigues de Araújo.

Aos professores Cris e Sérgio, meus agradecimentos pela orientação e pelas horas gastas em nos ensinar e motivar.

.

Pesquisar é acordar para o mundo. - Marcelo Lamy

Resumo

Contextualização: A identificação de fatores que afetam a evasão acadêmica é importante para garantir que coordenações consigam tomar ações necessárias para evitar a desistência estudantil.

Objetivo: O objetivo desta monografia é entender como a área de pesquisa atual se encontra, de forma a compreender os principais fatores relacionados à evasão acadêmica e como aplicá-los em modelos de aprendizado de máquina.

Método: Trata-se de uma pesquisa quantitativa, exploratória e explicativa, de natureza aplicada, com o uso dos procedimentos técnicos de estudo de caso e de pesquisa bibliográfica.

Resultados: A revisão sistemática da literatura mostrou que a evasão acadêmica pode ser prevista por meio de fatores acadêmicos, demográficos e de aprendizado. Com base na identificação destes, foram criados modelos a partir de algoritmos de aprendizado de máquina. Os modelos foram utilizados para demonstrar como os fatores podem ser usados.

Conclusão: Os fatores acadêmicos, demográficos e de aprendizado são capazes de realizar a previsão da evasão. A definição dos fatores que serão utilizados e a forma de seu uso é importante para obter bons resultados de previsão. Além disso, quando considerando fatores acadêmicos para precisão é importante considerar o contexto aplicados aos dados.

Palavras-chave: evasão. evasão acadêmica. ensino superior. graduação. fatores. indicador. previsão.

Abstract

Contextualization: Identifying factors that affect academic dropout is essential to ensure that coordinators can take the necessary actions to avoid student dropout.

Goal: This monograph's objective is to understand the current research area, the main factors related to academic dropout, and how to apply them in machine learning models.

Method: It is quantitative, exploratory, and explanatory research of an applied nature, using the technical procedures of case study and bibliographic research.

Results: A systematic literature review showed that academic dropout could be predicted by academic, demographic, and learning factors. Based on the identification, models were created from machine learning algorithms. The models were used to demonstrate how factors can be used.

Conclusion: Academic, demographic, and learning factors are capable of predicting dropout. Defining the factors used and how to use them is essential to obtain good forecasting results. Furthermore, when considering academic factors for accuracy, it is vital to consider the context applied to the data.

Key-words: academic dropout prediction. higher education. undergraduate. factors. indicator. prediction.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
DEED	Diretoria de Estatísticas Educacionais
MEC	Ministério da Educação
AM	Aprendizado de máquina
IBM SPSS	Software científico
TCC 1	Trabalho de Conclusão de Curso 1
TCC 2	Trabalho de Conclusão de Curso 2
ETL	Ferramentas de software para extração e transformação de dados
MDE	Mineração de Dados Educacionais
SIGAA	Sistema Integrado de Gestão de Atividades Acadêmicas
UnB	Universidade de Brasília
BCE	Biblioteca Central da Universidade de Brasília

Sumário

1 Introdução

1.1 Contextualização

Garantir a permanência de estudantes em cursos de níveis de formações universitárias, desde o início até o fim é um dos objetivos das instituições de ensino superior. Nos resultados anuais de desempenho de uma universidade um dos fatores mais importantes a ser levado em consideração é a sua taxa de evasão, pois essa gera impactos financeiros e de visibilidade externa. O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), a Diretoria de Estatísticas Educacionais (DEED) e o Ministério da Educação (MEC) são responsáveis pelo Censo da Educação Superior, os quais apresentam os resultados das universidades federais e das faculdades privadas do país.

O Sampaio et al. (2021) considera como definição de evasão acadêmica a taxa de alunos que desistiram da graduação ou que realizaram transferência entre cursos. Entender os motivos responsáveis pela evasão estudantil é importante, pois garante às instituições a realização de ações para o decremento da taxa de desistência, seja via mudanças curriculares, distribuição de auxílios socioeconômicos ou reestruturação interna.

Uma forma de compreender as razões que induzem a evasão é compreender como e quais fatores externos e internos impactam o desempenho acadêmico dos discentes. Autores como Fernández-Martín et al. (2019) e Alvarez, Callejas e Griol (2020) focam suas pesquisas nesta identificação. Já Kilian, Loose e Kelava (2020) e Niessen, Meijer e Tendeiro (2016) buscam entender como estes fatores podem ser utilizados para prever a evasão acadêmica.

A evasão pode ser prevista com a utilização de modelos criados por meio de algoritmos de aprendizado de máquina. Segundo Kondo, Okubo e Hatanaka (2017), - o aprendizado de máquina (AM) é a abordagem que busca capacitar computadores para aprender de forma automática, por meio de algoritmos que identificam padrões em dados reais. Esses modelos foram utilizados nesse trabalho para prever a evasão acadêmica a partir da análise de informações coletadas sobre os estudantes.

O processo de identificação de fatores que impactam o rendimento acadêmico e, por consequência, podem prever a evasão, foi realizado por meio de uma revisão sistemática da literatura. A partir da determinação desses indicadores modelos foram criados com as técnicas de Floresta Aleatória, Árvore de Decisão e C5.0, para aplicação em dados de alunos da Faculdade do Gama.

1.2 Justificativa

A evasão é um fenômeno comum e inerente do ensino superior. Instituições são avaliadas em vários aspectos incluindo os índices de desistência. Sampaio et al. (2021) apresentam no Resumo Técnico do Ensino Superior de 2019, no indicador de desistência de alunos, que ingressaram em universidades entre 2010 e 2015, foi observado que a maior taxa de evasão ocorre no segundo ano da graduação.

Descobrir quais fatores impactam o desempenho dos estudantes e que podem aumentar as chances de evasão é importante para garantir que as universidades e as faculdades possam tomar ações para evitar essa possível desistência. A análise de dados relacionados ao rendimento de um aluno, como sua posição no fluxo e a quantidade de reprovações traz algumas previsões sobre a evasão do discente.

Levando em consideração as informações de alunos dos cursos de Engenharia Aeroespacial, Engenharia Automotiva, Engenharia Eletrônica, Engenharia de Energia e Engenharia de Software e as taxas de evasão do campus Faculdade do Gama da Universidade de Brasília, surgiu a necessidade de identificar quais fatores podem auxiliar na previsão da evasão desses estudantes. Estes fatores foram aplicados em modelos de aprendizado de máquina, com o objetivo de demonstrar como utilizar esses fatores relacionados a desistência estudantil.

1.3 Objetivos

1.3.1 Objetivo Geral

O objetivo desta pesquisa é identificar os fatores que contribuem para a evasão de estudantes de graduação nos cursos de engenharia da Faculdade do Gama da Universidade de Brasília.

1.3.2 Objetivos Específicos

- Identificar os fatores que contribuem para a evasão no ensino superior;
- Estabelecer como os fatores são utilizados na previsão;
- Utilizar os fatores a partir de uma proposta de modelos.

1.4 Organização dos Capítulos

Esta monografia está organizada e dividida em cinco capítulos, incluindo esta introdução. O Capítulo 2 abrange a revisão bibliográfica, na qual são apresentados a

forma de condução e os resultados da revisão sistemática, o estudo bibliométrico e as análises quantitativas e qualitativas nos artigos da revisão. No Capítulo 3 a metodologia de pesquisa é definida e o processo de fluxo de trabalho é descrito. O Capítulo 4 apresenta o uso dos indicadores, com informações sobre a fonte de dados, explicação dos modelos de previsão determinados e seus respectivos desenvolvimentos na linguagem R. O Capítulo 6 exibe a análise final das informações e o fechamento dos objetivos.

2 Revisão Bibliográfica

Este capítulo apresenta os resultados da revisão sistemática, do estudo bibliométrico dos artigos obtidos e das análises qualitativas e quantitativas, para que seja viável entender o atual estado das pesquisas na área. Além disso, é realizada a tratativa dos dados e das métricas coletadas com a condução do protocolo de pesquisa, possibilitando, dessa forma, a compreensão do contexto, dos conceitos e das características da evasão no ensino superior.

A evasão de estudantes no ensino superior tem uma vasta área de pesquisa, com diferentes autores e artigos publicados anualmente. Nos últimos três anos, houve um aumento na quantidade e na frequência de publicações realizadas. Essa movimentação trouxe uma variedade de pesquisas, com uma pluralidade de opiniões e contextos. Entender essa diversidade é importante para conhecer como as pesquisas atuais estão e quais são as lacunas existentes para novos estudos.

Nesse contexto, desenvolveu-se uma revisão sistemática de literatura para que, a partir das indagações levantadas nesta pesquisa, fosse plausível inferir os conceitos aplicados e como são utilizados. E assim, identificar como os fatores e os algoritmos estudados e manipulados contribuem na previsão da evasão de estudantes no ensino superior.

Na Seção 2.1, é indicado como foi executado o protocolo da revisão sistemática. Na Seção 2.2, foi realizado um estudo bibliométrico sobre a lista de artigos obtidos na realização da pesquisa. Os objetivos da revisão sistemática e das questões secundárias são descritos na Seção de Planejamento 2.3. A execução do protocolo de pesquisa foi detalhada na Seção de Condução 2.4. Os resultados obtidos foram apresentados na Seção de Resultados 2.5. E por fim, na Seção de Discussão 2.6, os autores expõem suas opiniões acerca da análise dos resultados. Detalhes sobre o protocolo conduzido podem ser encontrados no apêndice A, e a lista de artigos obtidos na condução pode ser acessada no apêndice B.

2.1 Método de Condução da Revisão Sistemática

O processo de condução da revisão sistemática consistiu em três grandes fases: planejamento, condução e publicação dos resultados. Sendo que cada uma dessas abrange um conjunto de atividades necessárias para que fosse possível realizar a aplicação da revisão sistemática da literatura.

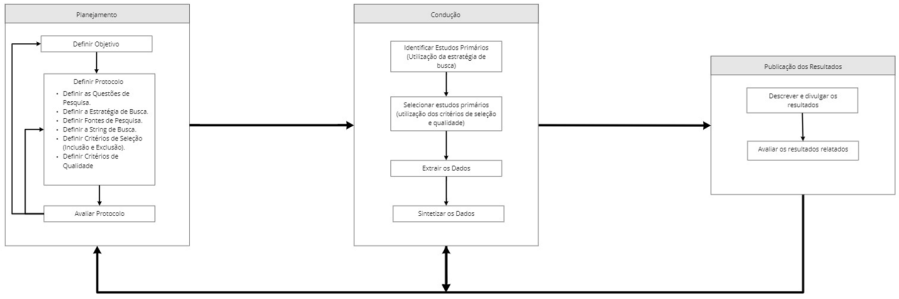


Figura 1 – Fases e atividades do processo de revisão sistemática.
Fonte: adaptado de Felizardo et al. (2020)

Os objetivos foram levantados durante a etapa de planejamento. Em seguida, foi criado o protocolo de pesquisa, no qual foram definidas as questões de estudo relacionadas à retenção e evasão e a *string* de busca. As fontes de pesquisa foram escolhidas como parte da estratégia de pesquisa, e, por fim os critérios de inclusão, exclusão e de qualidade foram selecionados. É importante ressaltar que o protocolo foi avaliado por um especialista da área e passou por vários refinamentos para melhor atendimento das necessidades desta pesquisa. Sua versão final pode ser encontrada para consulta no Apêndice A.

Na fase de condução, os estudos primários passaram a ser identificados com base em resultados obtidos na execução da *string* de busca, nas bases IEEE Xplore[1] e Scopus[2]. A seleção dos estudos foi executada a partir da leitura dos resumos das publicações e da aplicação dos critérios de inclusão e exclusão. Com os estudos selecionados, a extração e a síntese dos dados foram feitas com base na leitura integral dos artigos, de forma a coletar as informações pertinentes e definidas durante o processo de planejamento do protocolo. A condução da revisão sistemática pode ser encontrada na Seção 2.4.

Para a fase de publicação dos resultados, duas etapas de análise foram executadas. A primeira delas refere-se à aplicação do estudo bibliométrico, no qual buscou-se entender o estado atual da área de pesquisa, levando em consideração a quantidade de artigos e citações por autor e a frequência de publicação de revistas científicas. Assim, foi possível saber quais os principais autores da área e as fontes de publicação mais ativas. A segunda etapa contemplou as análises de resultados quantitativos e qualitativos coletados na leitura integral dos artigos, nos quais foi possível coletar métricas sobre os resultados de cada fase da condução do protocolo de pesquisa e, ainda, entender os conceitos aplicados pelos autores. Os resultados obtidos se encontram na Seção 2.5.

[1] Pode ser acessado em: https://www.ieee.org/
[2] Pode ser acessado em: https://elsevier.com/

2.2 Estudos bibliométricos

Segundo Filipe et al. (2016), para definir de forma correta das palavras-chave de um estudo bibliométrico é importante ter conhecimento prévio na área que será pesquisada. Diante disso, foi realizada uma síntese primária sobre evasão no ensino superior, utilizando termos base como: (1) análise de aprendizagem; (2) mineração de dados; (3) evasão; (4) retenção; (5) ensino superior; (6) graduação e (7) bacharelado. Com o resultado desse uso como *string* de busca, foram obtidos 252 artigos, dos quais os primeiros 27 foram lidos em sua totalidade e usados como base de conhecimento para a identificação das palavras-chave relevantes da área.

A definição dos termos para o estudo bibliométrico abrangeu as seguintes palavras: (1) estudante; (2) aluno de graduação; (3) predição; (4) evasão; (5) retenção; (6) atrito; (7) métrica; (8) medição; (9) indicador; (10) ensino superior e (11) bacharelado. E utilizadas na base de dados Scopus[2], para compreender os artigos que as incluíssem em títulos, resumos e palavras-chave. Por fim, os filtros foram aplicados nos tipos de documentos, limitando, assim, os resultados a artigos e papéis de conferência.

A lista final de estudos obtida continha 118 artigos com um total de 1.746 citações. Porém, 21 artigos não são citados e 7 são responsáveis por 51,8% das citações totais, isto pode ser observado na Figura 2. Ademais, dos 48,16% restantes, 35,5% possuem quantidades significativas de citações e contemplam 27 artigos totais.

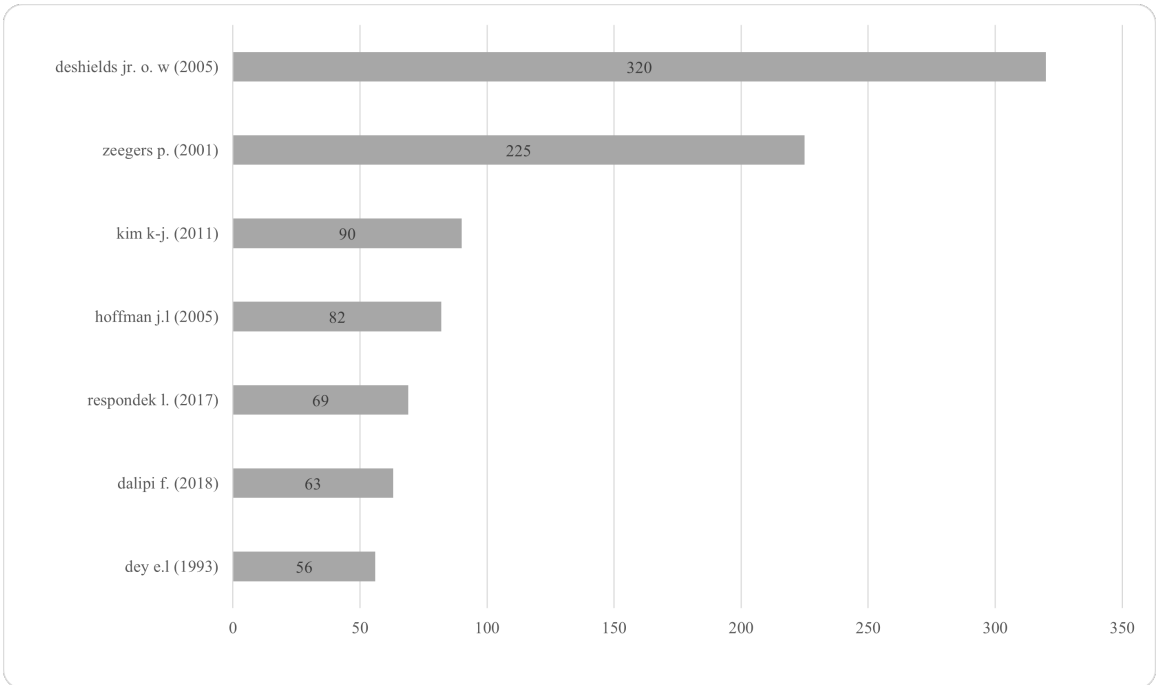


Figura 2 – Gráfico de Análise de Quantidade de Citações por Artigos.
Fonte: Autoras.

[2] Pode ser acessado em: https://elsevier.com/

O total de autores com publicações resultantes na pesquisa foi 354. Somente 9 publicaram dois artigos, que contemplam 3,7% das citações por autores. Na Figura 3, é possível observar a relação dos pesquisadores com maior número de publicações e suas citações. Porém, é importante ressaltar que quando considerado citações por autores, o número difere do total de citações por artigos. Isso ocorre pois os pesquisadores são considerados individualmente, enquanto a análise de artigos não considera os autores por artigos. Por isso, o número de citações por autores difere do de citações por documentos e tem um total maior englobando 2.295 citações. Do total 95,94% compreendem citações de 19 autores, isto pode ser observado na Figura 4.

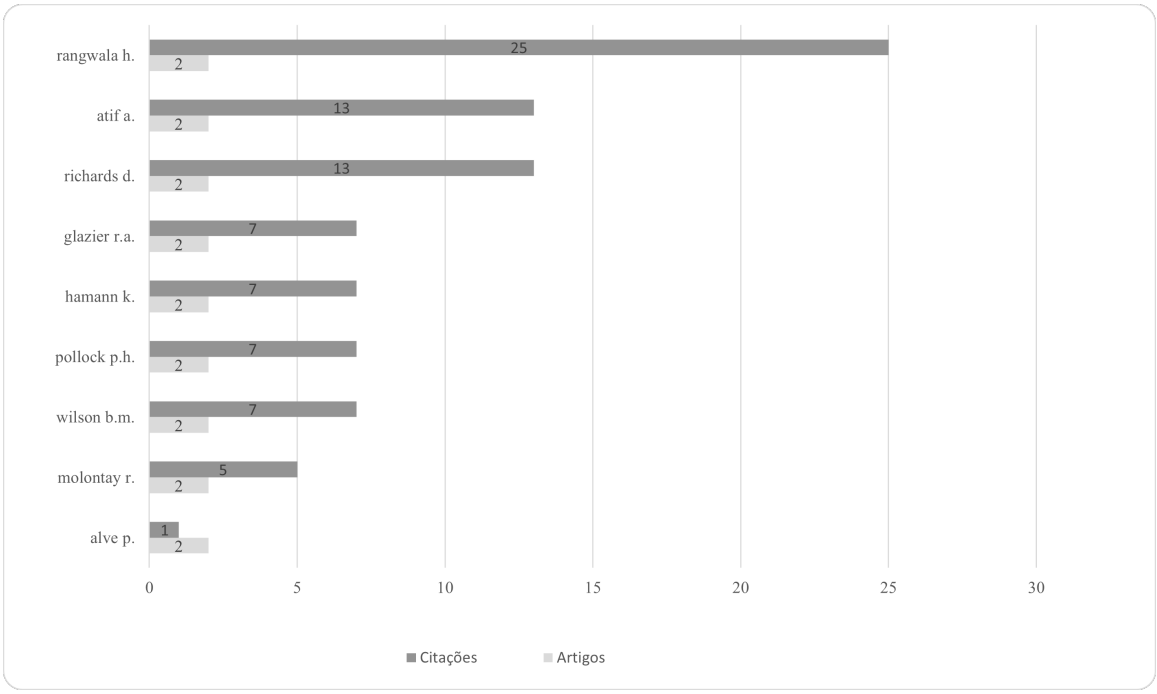


Figura 3 – Gráfico de Análise de Quantidade de Citações e Documentos por Autores. Fonte: Autoras.

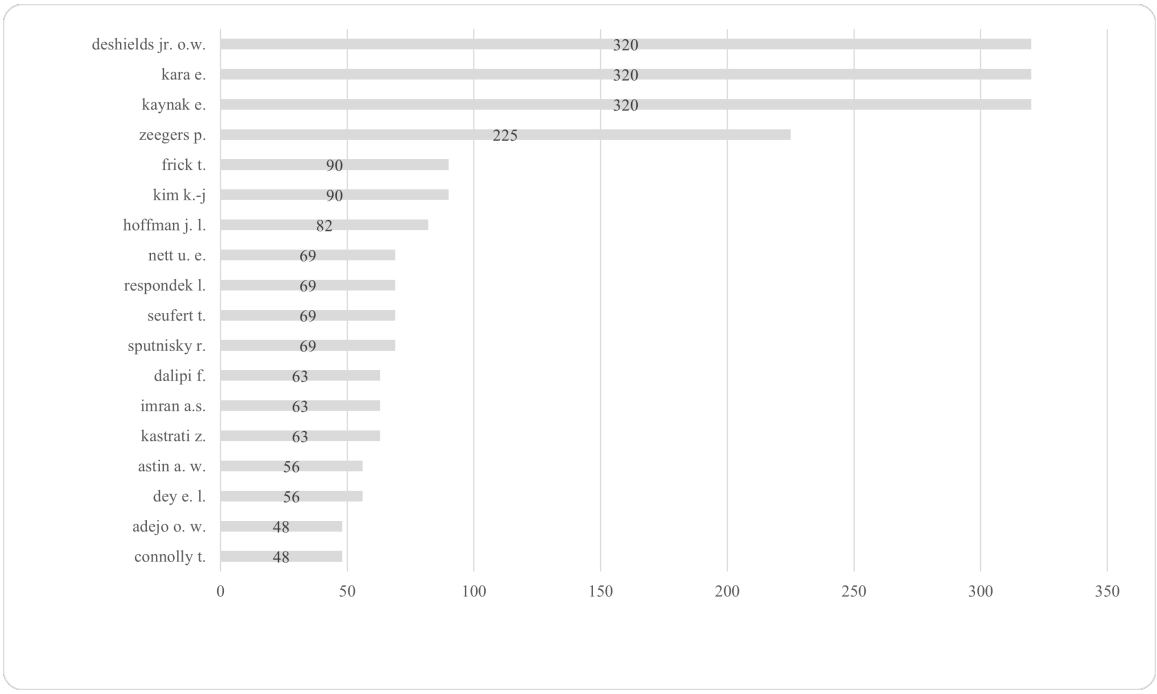


Figura 4 – Gráfico de Análise de Quantidade de Citações por Autoras.
Fonte: Autoras.

2.3 Fase: Planejamento

Durante o planejamento, foi definido como o objetivo de pesquisa, a identificação de fatores que contribuem para a retenção ou evasão de estudantes de graduação. Além disso, foi estabelecida a questão de pesquisa: Quais fatores podem ser identificáveis quando há a associação da previsão, de retenção e evasão no ensino superior?. Por fim, as perguntas secundárias foram escolhidas de forma a auxiliar a resolução do problema principal, sendo essas: (1) Quais são os fatores usados para prever a retenção no ensino superior?; (2) Quais são os fatores usados para prever a evasão no ensino superior?; e (3) Como os fatores são usados no processo da previsão?. Maiores detalhes sobre a fase de planejamento e sobre as questões de pesquisa podem ser encontrados no Apêndice A.

2.4 Fase: Condução da pesquisa

2.4.1 Identificar estudos primários - utilização da estratégia de busca

Neste estudo buscou-se na literatura publicações que atendessem temáticas relativas à evasão de alunos no ensino superior, e, fundamentado nessa pesquisa inicial, foi possível definir as questões de pesquisa, analisar a viabilidade do estudo e aumentar a proximidade com o tema.

As bases IEEE Xplore e Scopus foram selecionadas para as buscas de artigos.

Segundo Wazlawick (2009), a base IEEE Xplore é uma das maiores sociedades de computação do mundo e oferece à comunidade, especialmente aos seus associados, uma biblioteca digital de textos integrais na qual constam seus periódicos e eventos. Citada como uma das mais importantes do mundo por disponibilizar acesso a artigos de periódicos e anais de conferências das áreas de computação, engenharia elétrica e eletrônica. Sobre a base Scopus Wazlawick (2009), diz que ela contém resumos e citações bibliográficas de dezenas de milhares de periódicos.

A partir da alta taxa de evasão identificada e da leitura de artigos levantados para a revisão sistemática presente, foi levantada a seguinte questão de pesquisa: "Quais são os fatores usados para prever a evasão no ensino superior?". Com base nessa questão, foi desenvolvida a *string* de busca, a metodologia para a seleção de publicações, os critérios de inclusão e exclusão e, por fim, a avaliação de qualidade.

2.4.2 Selecionar estudos primários - Utilização dos critérios de seleção

Os 128 artigos resultantes da busca nas fontes de pesquisa IEEE Xplore[1] e Scopus[2], foram importados para a plataforma Parsifal[3], um filtro foi aplicado para a exclusão dos estudos duplicados. Dentre os artigos restantes, foi verificada a existência de pesquisas com enfoque em evasão acadêmica, retenção acadêmica e evasão e retenção acadêmica, como é possível observar no Gráfico da Figura 5.

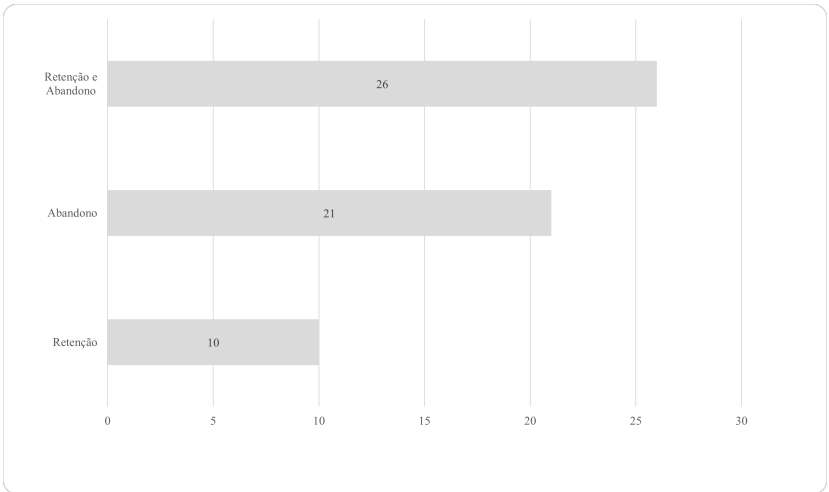


Figura 5 – Gráfico de Análise Temáticas das Publicações.
Fonte: Autoras.

A confirmação da existência de pesquisas com enfoque em evasão acadêmica, retenção acadêmica e evasão e retenção acadêmica foi obtida por meio da leitura dos resumos das publicações e da aplicação dos critérios de seleção. Essa triagem considerou apenas

[1] Pode ser acessado em: https://www.ieee.org/
[2] Pode ser acessado em: https://elsevier.com/
[3] Pode ser acessado em: https://parsif.al/

o conteúdo presente nos resumos, de forma a entender se os artigos estavam de alguma forma relacionados à retenção e à evasão acadêmica e se respondiam as perguntas secundárias deste estudo. Dessa seleção, 57 artigos foram selecionados para a etapa de leitura integral.

2.4.2.1 Critérios de inclusão

Com o intuito de identificar os fatores, indicadores e algoritmos relacionados à evasão e retenção estudantil, foram definidos os seguintes critérios de inclusão, com base nas questões secundárias de pesquisa:

- O artigo apresenta fator relacionado à evasão acadêmica.
- O artigo apresenta fator relacionado à evasão e retenção acadêmica.
- O artigo apresenta fator relacionado à retenção acadêmica.

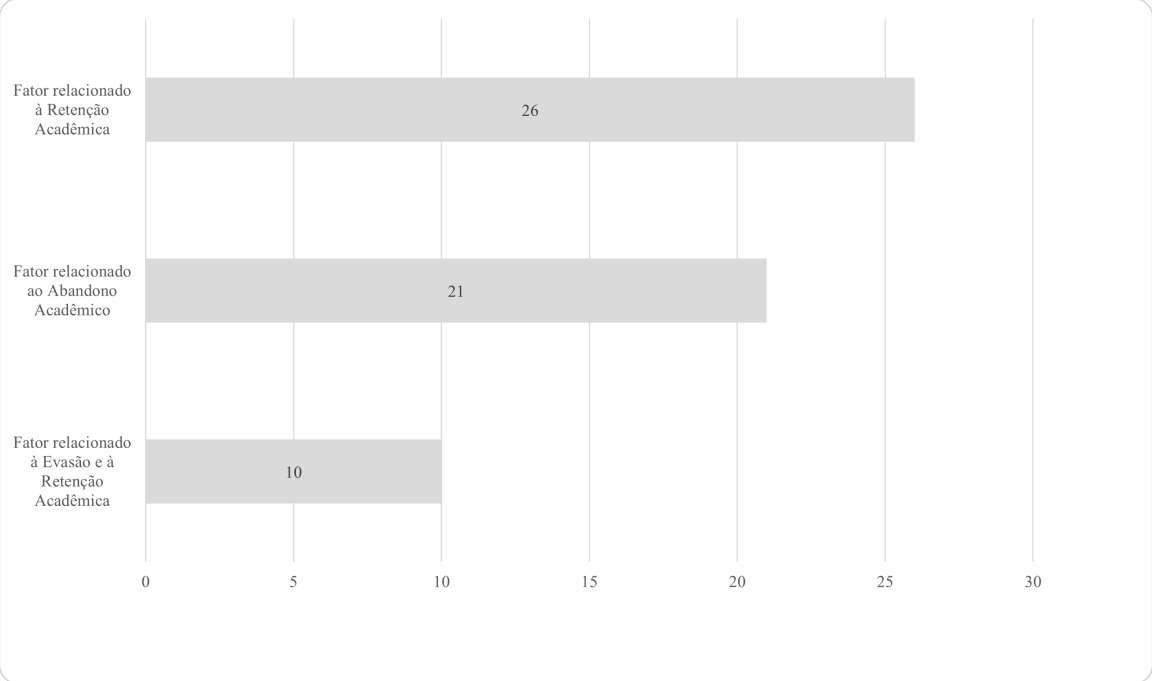


Figura 6 – Gráfico de Análise dos Critérios de Inclusão.
Fonte: Autoras.

No gráfico da Figura 6 é possível observar como os artigos foram selecionados de acordo com os critérios de inclusão. Dentre as publicações analisadas, 58 foram aceitas, sendo que aqueles relacionados à retenção compreenderam boa parte dos artigos selecionados e 21 artigos se relacionam à evasão acadêmica.

2.4.2.2 Critérios de exclusão

Os critérios de exclusão foram definidos levando em consideração a possibilidade de retirada de artigos que não respondessem uma das perguntas secundárias, que não estivessem disponíveis nos idiomas de preferência ou que não pudessem ser lidos de forma integral. Essa é uma etapa importante, pois garantiu que os artigos aceitos estivessem relacionados à área de estudo. Dessa forma, os critérios de exclusão levantados foram:

1. Artigo de outra área de estudo.
2. Não se enquadra nos critérios de aceitação.
3. Estudos duplicados.
4. Sem acesso ao texto completo do artigo.
5. Versões mais antigas do mesmo estudo.
6. Publicações redigidas em idiomas diferentes do inglês, português e espanhol.
7. Estudos secundários ou terciários.

Das 128 publicações analisadas, 61 não passaram nos critérios de aceitação. A maior parte dos artigos foi excluída pelo critério artigo de outra área de estudo (1), seguido por não se enquadra nos critérios de aceitação (2), que contempla artigos que não são enquadrados nos outros critérios de exclusão, mas também não se encaixam nos de inclusão. Observa-se também que o critério menos utilizado é o de estudos secundários ou terciários (3).

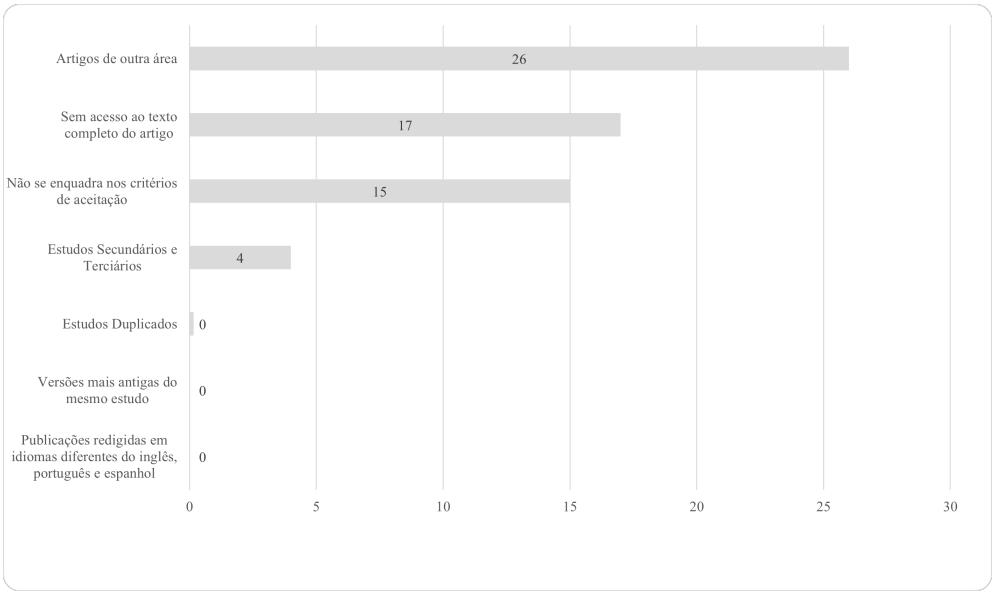


Figura 7 – Gráfico de Análise dos Critérios de Exclusão.
Fonte: Autoras.

2.4.3 Avaliação da Qualidade

A avaliação da Qualidade auxiliou na identificação da importância dos artigos para o estudo. Para isso, foram levantadas métricas, e os artigos poderiam receber notas mínimas de zero pontos e máximas de oito pontos. As métricas definidas são apresentadas abaixo.

- Os objetivos, questões de pesquisa e hipóteses (se aplicável) são claros e relevantes?
- Existe uma descrição adequada do contexto em que a pesquisa foi realizada?
- Os dados foram coletados de forma a abordar a questão de pesquisa?
- Existe uma declaração clara de resultados?
- Os autores descrevem as limitações do estudo?
- As conclusões, implicações para a prática e pesquisas futuras são adequadamente relatadas ao seu público?
- As descobertas foram claramente relatadas?
- As questões éticas são devidamente abordadas (intenções pessoais, integridade, confidencialidade, consentimento, aprovação do conselho de revisão)?

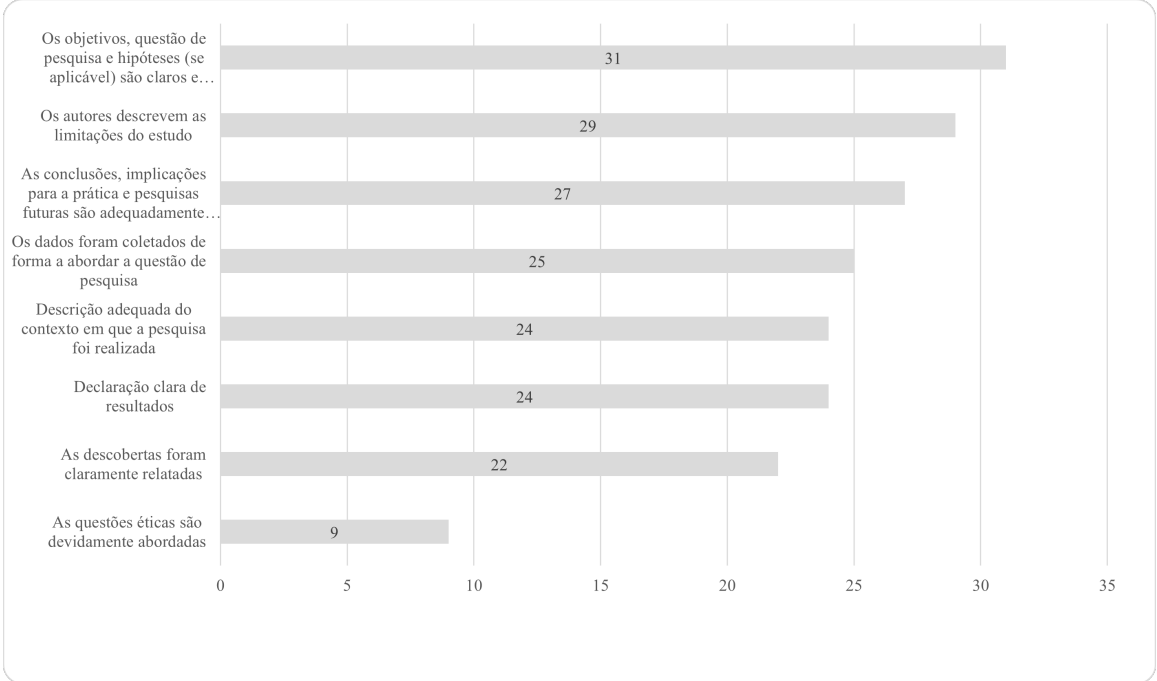


Figura 8 – Gráfico de Análise das Pontuações Altas dos Critérios de Qualidade.
Fonte: Autoras.

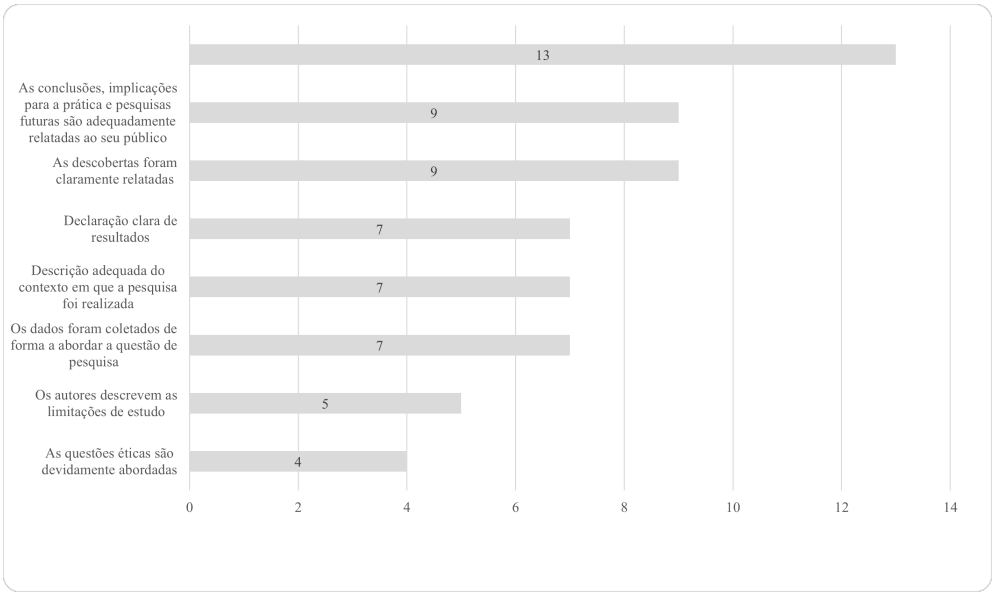


Figura 9 – Gráfico de Análise das Pontuações Parciais dos Critérios de Qualidade. Fonte: Autoras.

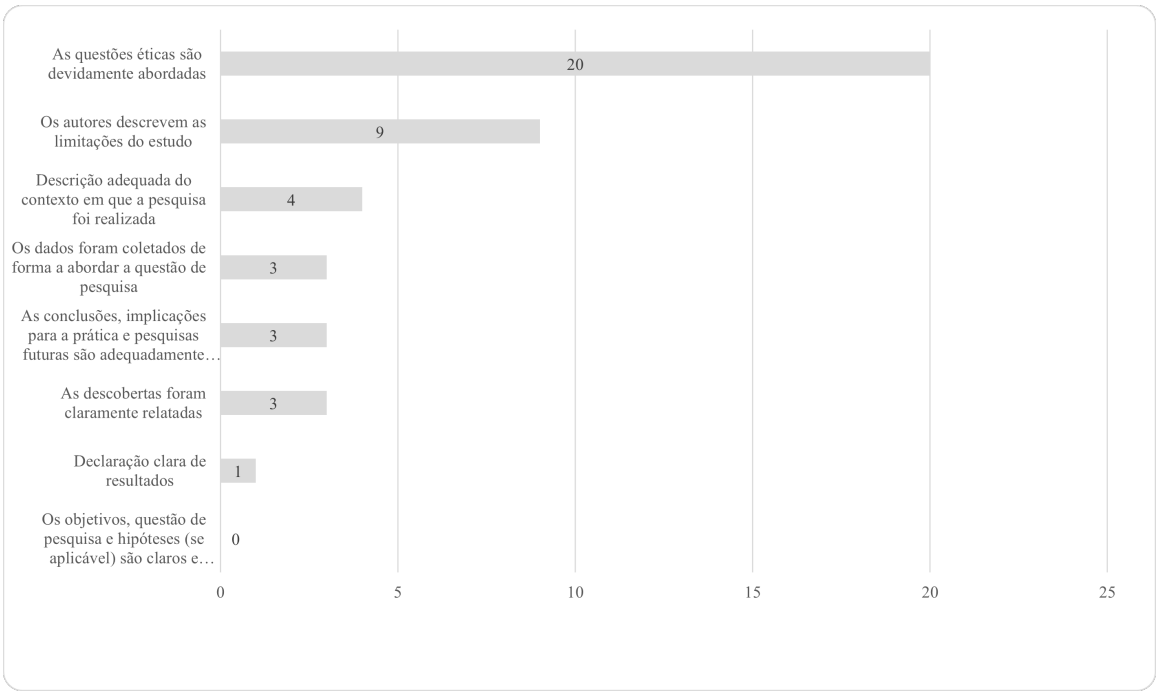


Figura 10 – Gráfico de Análise das Pontuações Baixas dos Critérios de Qualidade. Fonte: Autoras.

Das Figuras 8, 9 e 10 é possível aferir que os artigos fornecem informações claras sobre as questões de pesquisa, sendo que grande parte dos artigos traz informações sobre as limitações de artigos. Além disso, foram identificados 9 artigos que não trazem informações. Assim, foram levantadas quais limitações podem ser tratadas.

2.5 Fase: Resultados

2.5.1 Análise quantitativa

2.5.1.1 Análise por Região e Período

Não foram colocadas limitações na *string* de busca quanto às regiões em que foram produzidos os artigos. Por esse motivo, foi viável verificar de forma global o nível de preocupação com a Evasão Acadêmica. Na Figura 11 é possível ter uma visão mais detalhada de como os artigos estão separados por continentes.

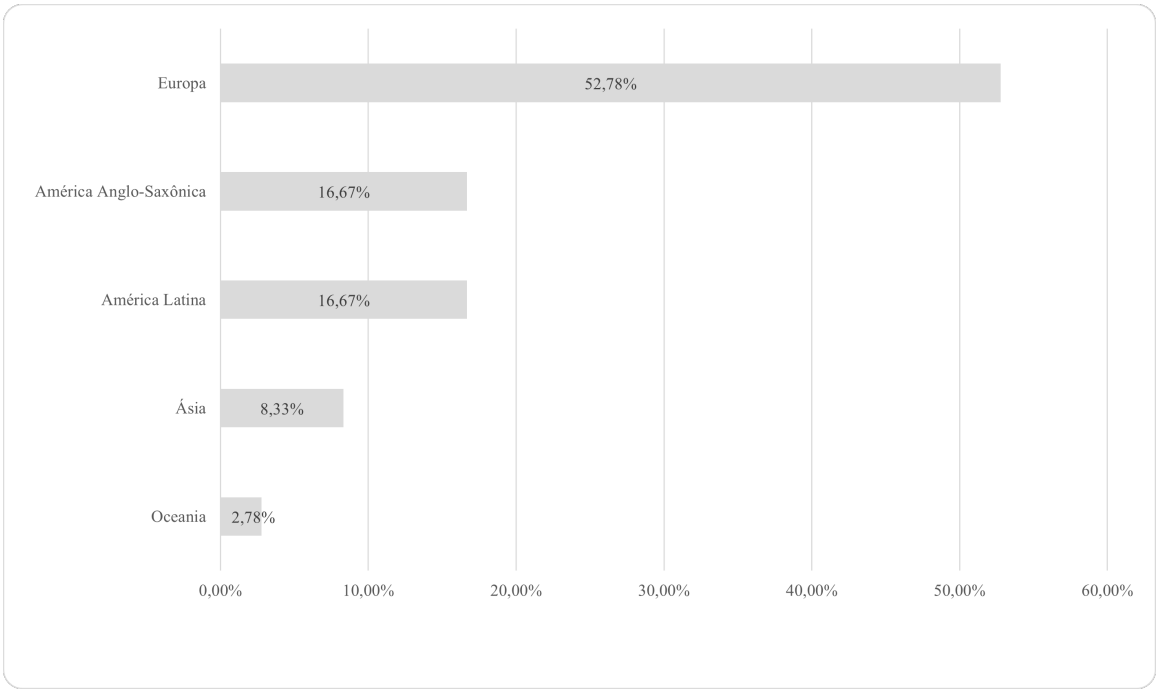


Figura 11 – Gráfico de Análise dos Continentes.
Fonte: Autoras.

Ao observar a Figura 11 percebe-se a falta de estudos sobre evasão acadêmica na América Latina comparado com a Europa.

2.5.1.2 Análise por Indicadores Acadêmicos, Demográficos e de Aprendizado

Foram levantadas, apoiado na leitura dos artigos, quais seriam as áreas dos indicadores, sendo estes acadêmicos, demográficos e de aprendizado, e além disso, como as áreas dos indicadores poderiam ser usadas como elementos de referência para a contribuição dos resultados.

Na Figura 12, é plausível realizar a verificação de como foi feito a catalogação dos indicadores por área. A área que teve menos indicadores foi a de aprendizado; a acadêmica e a demográfica tiveram a mesma quantidade de indicadores fichados.

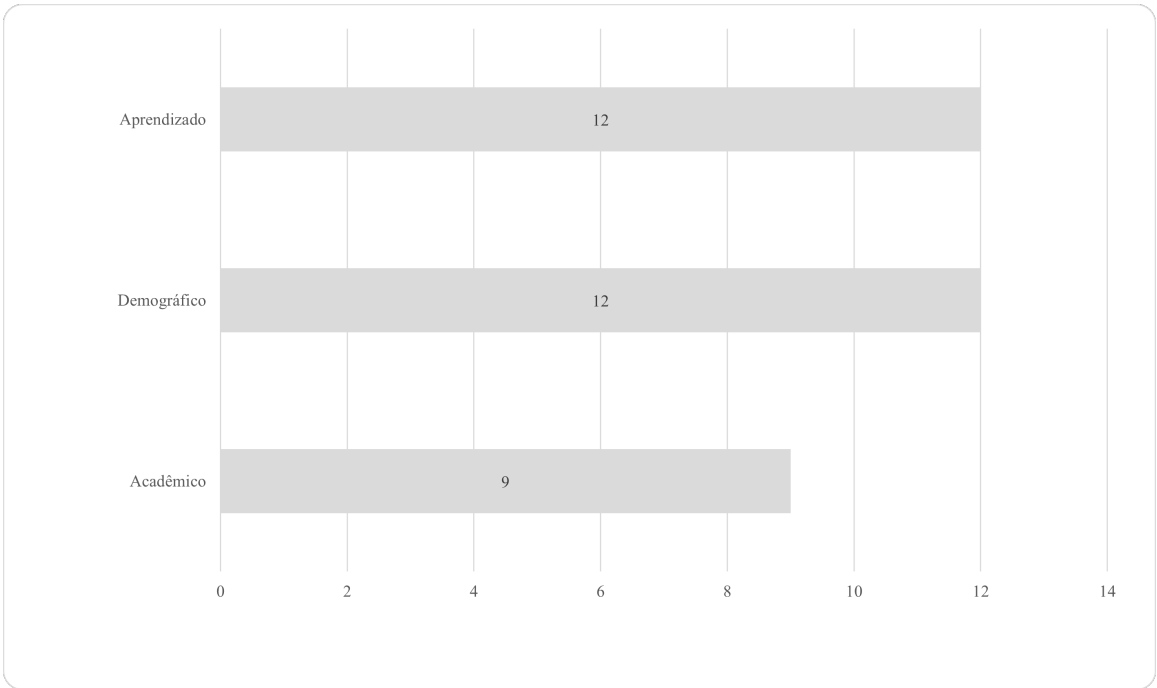


Figura 12 – Gráfico de Análise dos Indicadores.
Fonte: Autoras.

Na Figura 13, há a divisão dos artigos por indicadores acadêmicos. Notas do Curso e Pontuação Média foram os indicadores com mais artigos fichados, ambos com 14 artigos. Sendo estes fatores que podem levar o aluno à reprovação em uma determinada disciplina. É importante salientar, também, que 8 artigos estão sem indicadores acadêmicos. Como é possível analisar na Tabela 1, os artigos citados, via numeração, estão disponíveis no Apêndice B.

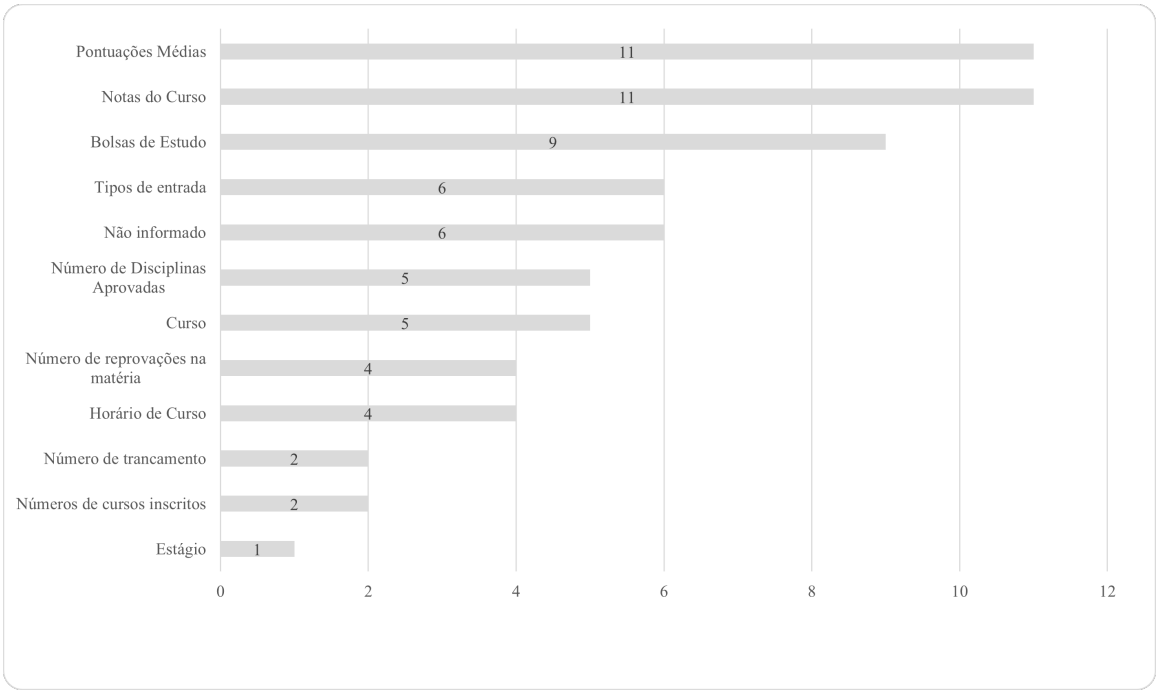


Figura 13 – Gráfico de Análise dos Indicadores Acadêmicos.
Fonte: Autoras.

Na Figura 14, os artigos estão divididos por indicadores demográficos. Os indicadores com maior número de artigos são os de Gênero e Idade, sendo o maior o de Gênero, com 17 artigos classificados, e Idade, com 14, como é possível analisar na Tabela 2.

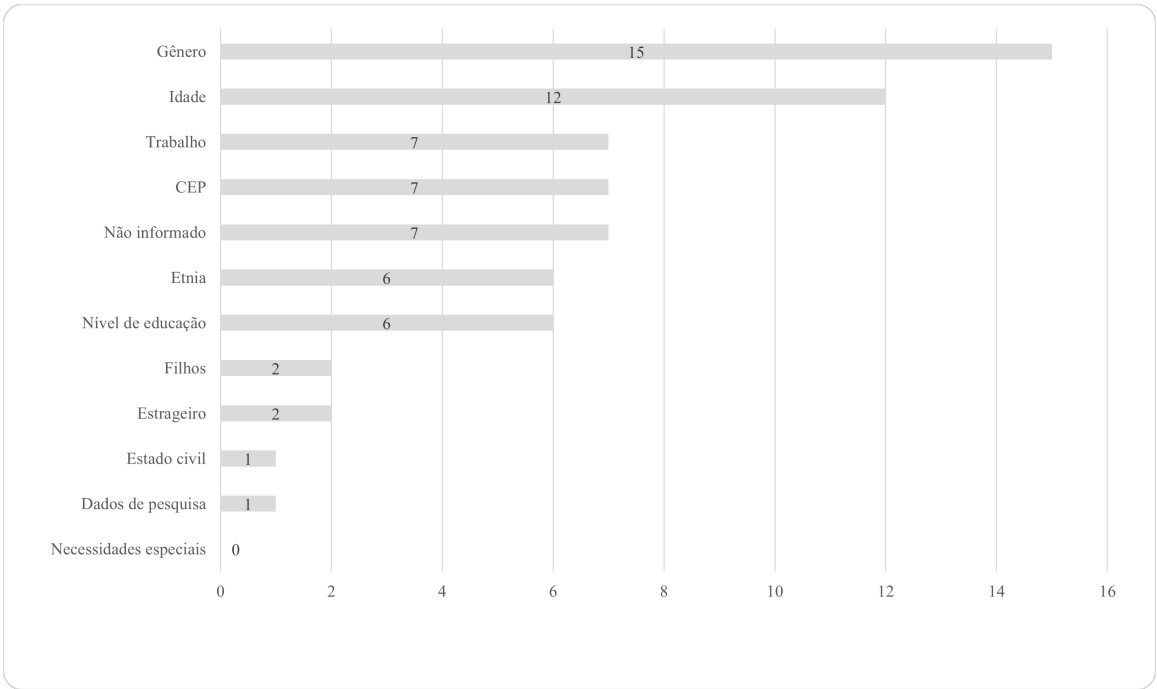


Figura 14 – Gráfico de Análise dos Indicadores Demográficos.
Fonte: Autores.

Na Figura 15, a divisão foi feita pelos fatores de Aprendizado. Os indicadores com

Fator	**Quantidade de Artigos**	**Artigos**
Pontuações Média	11 artigos	A-2, A-14, A-16, A-18, A-19, A-20, A-22, A-23, A-24, A-25 e A-27.
Curso	5 artigos	A-1, A-18, A-22, A-25 e A-27.
Notas do Curso	11 artigos	A-2, A-5, A-9, A-11, A-12, A-16, A-17, A-18, A-19, A-22 e A-23.
Horário do Curso	4 artigos	A-4, A-9, A-13 e A-17.
Tipos de Entrada	6 artigos	A-9, A-16, A-17, A-19, A-22 e A-23.
Estágio	1 artigos	A-13.
Número de Trancamento	2 artigos	A-13 e A-24.
Número de Disciplinas Aprovadas	5 artigos	A-4, A-13, A-18, A-22 e A-27.
Número de Cursos Inscritos	2 artigos	A-12 e A-13.
Número de Reprovações nas Matérias	4 artigos	A-9, A-13, A-17 e A-19.
Bolsa de Estudos	9 artigos	A-1, A-2, A-3, A-6, A-9, A-13, A-17, A-22 e A-26.
Não Informado	6 artigos	A-7, A-8, A-10, A-15, A-21 e A-28.

Tabela 1 – Tabela de Indicadores Acadêmicos

mais artigos classificados são os de Estudo regular e Acesso ao Fórum, sendo que, para Estudo Regular, existem 8 artigos, e Acesso ao Fórum, 6. Nesse indicador, houveram 15 artigos que não se encaixaram dentre os indicadores classificados, ou seja, houveram mais artigos não classificados do que classificados. Como é possível analisar na Tabela 4.

Fator	**Quantidade de Artigos**	**Artigos**
Filhos	2 artigos	A-8 e A-11.
Idade	12 artigos	A-2, A-8, A-9, A-11, A-14, A-15, A-17, A-19, A-20, A-22, A-23 e A-25.
Etnia	6 artigos	A-2, A-5, A-13, A-15, A-18 e A-26.
Estrangeiro	2 artigos	A-9, A-17 e A-22.
Gênero	15 artigos	A-2, A-3, A-5, A-6, A-8, A-11, A-14, A-15, A-16, A-18, A-20, A-22, A-23, A-25, A-26
Trabalho	7 artigos	A-2, A-3, A-6, A-8, A-11, A-14, A-26
Estado Civil	1 artigo	A-11
Nível de Educação	6 artigos	A-1, A-2, A-3, A-6, A-14, A-15
Dados de Pesquisa	1 artigo	A-14
Necessidades Especiais	0 artigos	N/A.
CEP	7 artigos	A-2, A-3, A-6, A-14, A-16, A-19, A-26
Não Informado	7 artigos	A-4, A-7, A-12, A-20, A-24, A-27, A-28

Tabela 2 – Tabela de Indicadores Demográficos

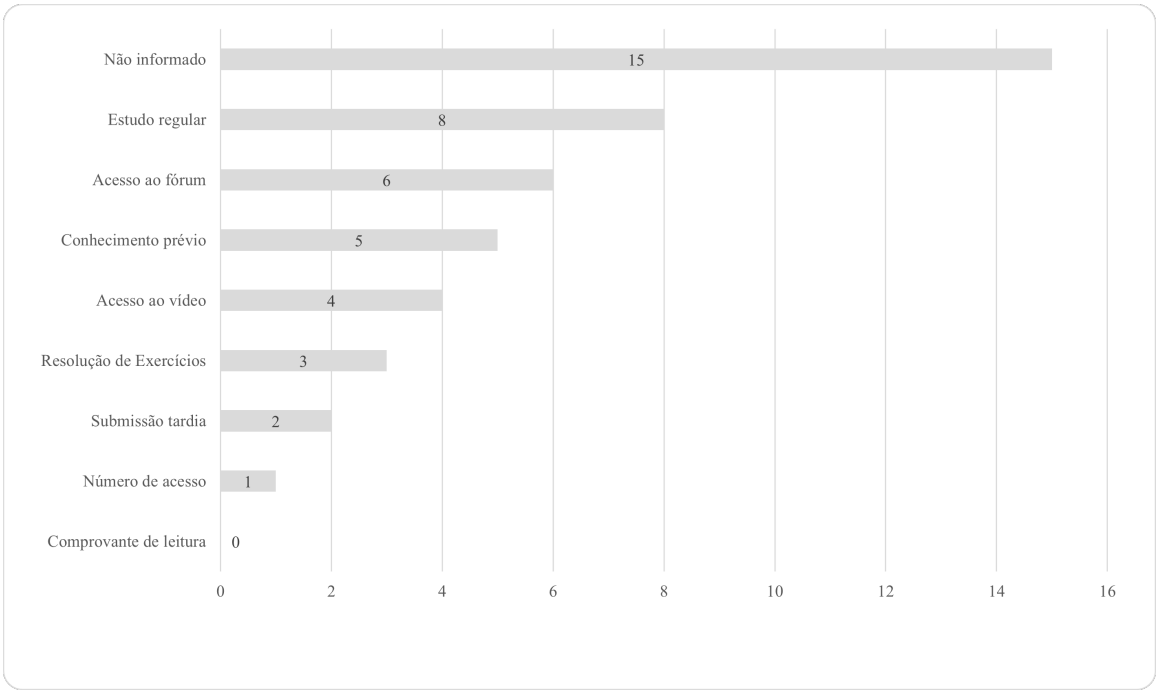


Figura 15 – Gráfico de Análise dos Indicadores de Aprendizado.
Fonte: Autores.

Fator	**Quantidade de Artigos**	**Artigos**
Conhecimento Prévio	3 artigos	A-11, A-19 e A-23.
Número de Acesso	1 artigo	A-12.
Submissão Tardia	1 artigo	A-10.
Resolução de Exercícios	2 artigos	A-10 e A-12.
Acesso ao Fórum	4 artigos	A-9, A-10, A-11 e A-12.
Comprovante de Leitura	0 artigos	N/A.
Estudo Regular	6 artigos	A-5, A-10, A-13, A-23, A-24 e A-25.
Acesso ao Vídeo	4 artigos	A-10 e A-20.
Não Informado	16 artigos	A-1, A-2, A-3, A-4, A-6, A-7, A-8, A-14, A-15, A-16, A-17, A-20, A-22, A-26, A-27, A-28

Tabela 3 – Tabela de Indicadores de Aprendizagem

Tabela 4 – Tabela de Decrição do PICOC

2.5.1.3 Algoritmos de Aprendizado de Máquina

O aprendizado de máquina é um subcampo ligado à Inteligência Artificial, o AM é um software que tem seu comportamento alterado de forma automatizada, a partir de modelos matemáticos e estatísticos, com o objetivo de atingir um resultado. Essa automação tem como intuito reduzir a interferência humana no programa e, para isso, são aplicados treinamentos que trazem ao software experiências. São estas experiências que dão conhecimento ao software.

Dentre os artigos levantados 15 não possuíam um algoritmo de Aprendizado de Máquina explícito. 8 artigos usaram o IBM SPSS, que é uma ferramenta para Mineração de Dados. Das 12 ferramentas levantadas, 4 não tinham citações, por isso foram zeradas e não adicionadas na Figura 2.5.1.3. Os outros algoritmos tiveram poucos levantamentos, sendo apenas 1 ou 2 artigos que traziam informações sobre eles.

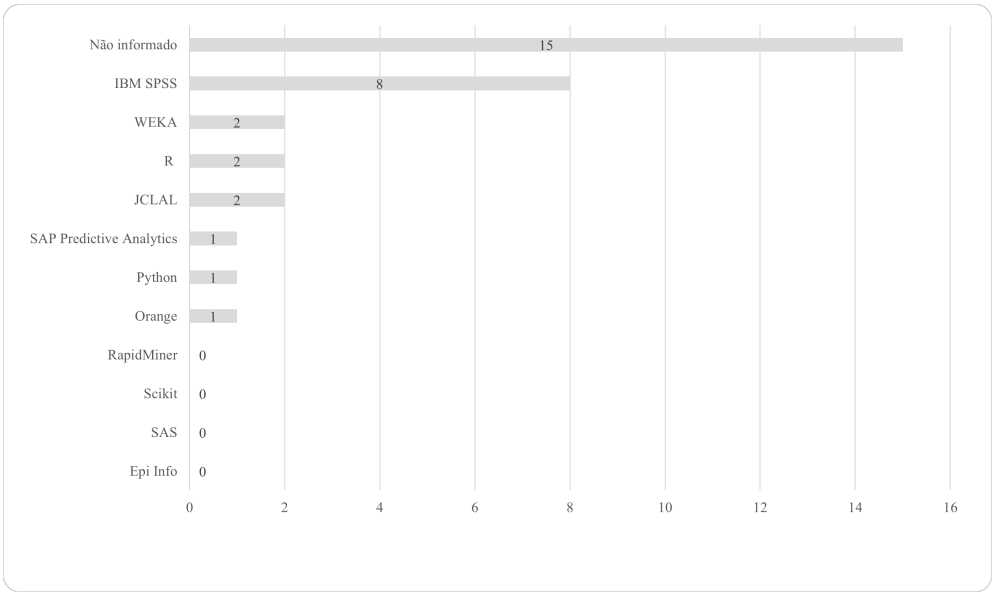


Gráfico de Análise das Ferramentas utilizadas pelos Algoritimos de Machine Learning. Fonte: Autores.

2.5.1.4 Modelos de Previsão

Foram levantados 18 possíveis modelos de previsão, dispostos na Figura 16. O mais utilizado foi o de Regressão, com 21 artigos. Esse modelo serve para prever comportamentos com base na associação entre duas variáveis que geralmente possuem uma boa correlação. O segundo mais utilizado foi o de Árvore de Decisão, com 11 artigos, os outros modelos ficaram com 7 ou menos artigos. 6 modelos não foram utilizados por falta de artigos que os cita.

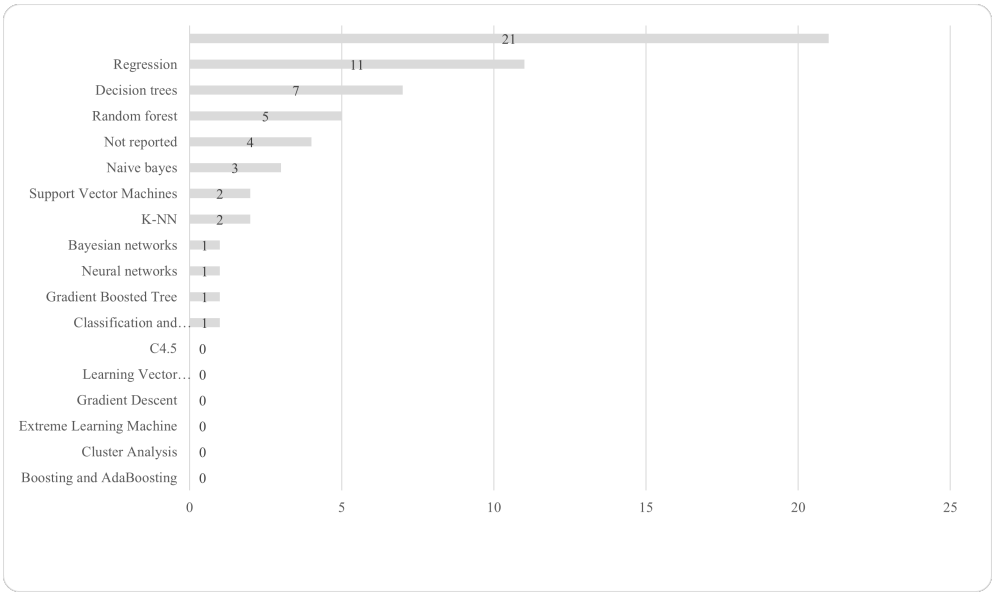


Figura 16 – Gráfico de Análise dos Modelos de Previsão mais utilizados. Fonte: Autores.

2.5.2 Análise qualitativa

2.5.2.1 Fatores usados para prever evasão no ensino superior

2.5.2.1.1 Fatores acadêmicos

Em relação a fatores acadêmicos é possível dividi-los em dois grupos distintos, o primeiro abrange aqueles considerados anteriores ao ingresso no ensino superior e o segundo contém fatores que são relacionados à faculdade. A forma de ingresso no ensino superior é encontrada e repetida ao longo dos artigos, por vezes só teve sua coleta de dados considerada e outras utilizada como fator importante. Svirina, Lopatin e Titko (2021) consideram unicamente dados sobre a forma de ingresso em suas análises e questionam a qualidade dos processos seletivos na qualificação dos alunos.

Mesmo com a diversidade de coletas de fatores acadêmicos, é possível inferir que considerar a forma de ingresso em uma instituição de ensino superior traz fatores que podem auxiliar na previsão de evasão. Porém, os entendimentos desses vão além, pois são capazes de contextualizar cada estudante por sua forma de entrada em uma faculdade. Alunos de instituições privadas muitas vezes precisam conciliar trabalho e estudos caso não tenham bolsa de estudos, enquanto discentes de universidades públicas variam conforme necessidades socioeconômicas. Essas diferenças podem impactar no desempenho acadêmico desses.

Silva et al. (2019) tentam entender esse impacto ao considerar, como algumas de suas variáveis, possíveis formas de remuneração do estudante, como bolsa para permanência na universidade, que é distribuída para alunos que necessitam de auxílio socioeconômico, e remuneração por alguma atividade desenvolvida academicamente como iniciativas em pesquisas científicas.

É importante entender também o desempenho dos estudantes em seus cursos. Para isso, autores como Niessen, Meijer e Tendeiro (2016) e Meens et al. (2018) consideram como fatores importantes aqueles que estão relacionados ao desempenho do estudante nas disciplinas. Notas e médias finais, matérias completadas com sucesso, número de trancamentos ao longo do semestre, número de reprovações, quantidade de créditos por semestre são alguns desses que, quando usados, são coletados por meio dos sistemas internos das faculdades.

Fatores acadêmicos se mostram essenciais para a previsão da evasão no ensino superior, uma vez que trazem consigo informações relacionadas diretamente ao desempenho acadêmico e fornecem pontos de atenção que, se considerados pelos docentes, podem impactar no possível sucesso acadêmico dos alunos.

2.5.2.1.2 Fatores demográficos

Os fatores demográficos têm seu peso de previsão avaliado por diversos autores na área. De acordo com Fernández-Martín et al. (2019), as taxas de evasão se relacionam com condições econômicas, familiares e pessoais. Atif, Richards e Bilgin (2013) afirmam que alunos que possuem desvantagens socioeconômicas apresentam dificuldades na criação de perspectivas profissionais.

Os autores Sani et al. (2020) contextualizam sua pesquisa em alunos de baixa renda e Frischenschlager, Haidinger e Mitterauer (2005) consideram informações sobre a educação e status sociais dos parentes do estudante, além das condições de vida deles. Respondek et al. (2017) relacionam os fatores demográficos com as faixas de evasão de alunos indígenas, portadores de deficiência e estudantes de áreas provinciais. Em contrapartida, Beck e Milligan (2014) indicam, em seus estudos, que a idade é o único fator que produz resultados significativos em suas previsões.

Além disso, Pérez et al. (2018) e Kostopoulos et al. (2018) consideram a quantidade de filhos de um estudante em suas análises. Kostopoulos et al. (2018) focam ainda no estado civil ao considerar o status familiar como um grande fator que impacta as condições de estudo de um discente. Oreshin et al. (2020) e Kiss et al. (2019) verificam a nacionalidade dos estudantes em suas análises, pois esses precisam passar por adaptações culturais e financeiras que podem afetar seus desempenhos.

Apesar da diversidade de uso e avaliação desses fatores, os resultados observados por Respondek et al. (2017) e Beck e Milligan (2014) são consideravelmente similares. Somente informações demográficas não são capazes de prever com eficácia a evasão estudantil, porém possuem influência nos resultados dos estudantes. Por essa razão, esses fatores são constantemente usados como variáveis nos modelos de previsão propostos.

Dos fatores mais utilizados, gênero, idade e raça são constantemente avaliados e utilizados para contextualizar a diversidade das faculdades. Informações sobre a possibilidade de o aluno trabalhar e a localização de sua moradia são abordadas como possíveis causas de evasão. Fatores como bolsa de estudos são consideradas para entender se os discentes possuem condições de dedicação integral aos estudos.

É importante ressaltar que autores como Baranyi et al. (2019) tiveram dificuldade em coletar fatores demográficos sobre os estudantes. Mesmo com acesso aos dados presentes nos sistemas das universidades, a presença desses fatores pode variar para cada aluno. Isso pode ser observado, pois Baranyi et al. (2019) e Kiss et al. (2019) apresentaram essas inconsistências de dados como dificuldades e limitações em suas pesquisas.

2.5.2.1.3 Fatores de aprendizado

Os fatores de aprendizagem estão relacionados com o comportamento dos alunos em plataformas de aprendizado online. A partir do acesso dos alunos a esses sistemas, é possível coletar dados como número de acessos, submissão de atividades, resolução de exercícios, acesso a fóruns e materiais de estudos.

Figueroa-Canas e Sancho-Vinuesa (2020) utilizam essas informações para entender os hábitos de estudos dos alunos e verificar sua possível associação com o desempenho deles. Já Kondo, Okubo e Hatanaka (2017) usam esses fatores para observar o comportamento de estudantes que possuem como pré-requisito organizar toda sua vida acadêmica em um sistema de gestão de aprendizagem.

Apesar da diferença de contextos entre as pesquisas dos autores Figueroa-Canas e Sancho-Vinuesa (2020) e Kondo, Okubo e Hatanaka (2017) é possível observar que alunos que usam ativamente essas plataformas estão empenhados em seus estudos, pois estão engajados nas atividades e nas discussões levantadas nos fóruns. Por consequência, eles estão comprometidos com seus estudos e possuem, assim, maiores chances de obter bons resultados.

Outro fator de aprendizagem importante é o conhecimento prévio dos alunos. Para Kostopoulos et al. (2018) entender esse fator é essencial para ter mais controle sobre os resultados de suas previsões. Em contrapartida, Kilian, Loose e Kelava (2020) restringem seus fatores iniciais para dados que possam ser coletados antes do início da disciplina e os avalia com os resultados obtidos. Kostopoulos et al. (2018) e Niessen, Meijer e Tendeiro (2016) aplicam testes para entender o conhecimento prévio dos estudantes sobre o conteúdo necessário nos cursos matriculados.

A oportunidade de coleta dos fatores acadêmicos abre um leque de possibilidades de previsão de evasão. Seus dados podem ser diretamente relacionados com fatores acadêmicos ao tentar prever as notas finais de uma disciplina, como, por exemplo, em Kondo, Okubo e Hatanaka (2017), e ainda, com os fatores demográficos, pois o acesso ao sistema avaliado pode variar de acordo com a condição do aluno. Além disso, quando considerado o conhecimento prévio dos estudantes, é possível separá-los em grupos distintos, levando em consideração a facilidade que possuem no conteúdo. Isso garante uma previsão mais detalhada em cima da dificuldade esperada por cada aluno. (KILIAN; LOOSE; KELAVA, 2020).

2.5.2.2 Como os fatores são usados para prever evasão a no ensino superior

Para prever a evasão, os fatores passam por um ciclo de coleta, preparo e utilização dos dados. A primeira fase contempla a coleta e geralmente é realizada por diferentes métodos, os principais são a aplicação de questionários e testes, como feito por Fényes,

Mohácsi e Pallay (2021), o acesso a dados privados das instituições e a utilização de informações presentes em sistemas de aprendizagem, como feito pelos autores Kondo, Okubo e Hatanaka (2017). Em seguida, esse material é refinado de forma a minimizar as inconsistências e os fatores são extraídos.

Os fatores, então, são aplicados em diferentes algoritmos de aprendizado de máquina utilizados para prever a evasão no ensino superior. Os modelos mais utilizados são regressão, árvores de decisão e floresta aleatória. Autores como Kiss et al. (2019) e Kostopoulos et al. (2018) realizam a aplicação de mais de um algoritmo e fazem comparação entre os resultados em suas pesquisas.

A aplicação dos algoritmos é feita com o auxílio de ferramentas, como IBM SPSS e Weka, e possuem o objetivo de verificar a probabilidade de previsão de um determinado indicador. É importante ressaltar que o uso de validação cruzada é frequentemente observado nos modelos de previsão apresentados por autores como Kostopoulos et al. (2018).

Os resultados geralmente são apresentados considerando cobertura, acurácia, precisão e outros tipos de pontuação. A partir destes, conclusões são realizadas de acordo com os fatores originais e suas potenciais eficácias na precisão de evasão do ensino superior.

2.6 Discussão

A partir da leitura das diferentes pesquisas, foi possível observar que a forma de uso dos fatores impacta diretamente seus resultados de previsão. Um fator utilizado de forma isolada, mesmo obtendo valores altos de acurácia, quando colocado no contexto da previsão pode não ser capaz de prever a evasão estudantil. Isso demonstra a importância da relação entre os fatores e o impacto que estes podem ter nos resultados gerais. Por isso, a definição dos fatores que serão utilizados e a forma de uso é importante para obter bons resultados de previsão. Saber definir quais dos fatores serão variáveis dependentes e independentes pode trazer melhor acurácia.

Além disso, utilizar mais de um algoritmo de aprendizado de máquina e compará-los utilizando técnicas como validação cruzada, pode trazer maior garantia na identificação dos fatores com maior impacto na evasão do ensino superior, já que fatores iguais podem obter resultados diferentes a partir do algoritmo utilizado no modelo de previsão.

Considerando os resultados quantitativos, foi possível observar que dentre as publicações analisadas, apenas uma era brasileira. Por isso, é de extrema importância o incentivo de pesquisas na área de evasão acadêmica, principalmente pelo elevado gasto do Governo Federal com alunos em Instituições de Ensino Superior públicas. Quando se estuda a evasão, é plausível promover ações de combate.

3 Metodologia de Pesquisa

Este capítulo abrange as metodologias de pesquisa, de forma a classificar a pesquisa quanto à abordagem, natureza, aos objetivos e procedimentos técnicos. Além disso, apresenta os fluxos de processo de trabalho e desenvolvimento.

3.1 Metodologias de Pesquisa

As metodologias de pesquisa podem ser classificadas por abordagem, natureza, objetivos e procedimentos técnicos. Entender cada uma dessas categorias é importante para definir o desenvolvimento científico que será aplicado para atingir os objetivos estabelecidos.

Prodanov e Freitas (2013) citam duas classificações quanto à abordagem, sendo essas: pesquisa quantitativa e pesquisa qualitativa. A primeira destas considera a quantificação das informações para análises, por meio da conversão de dados em números, enquanto a segunda busca compreender as possíveis relações entre contextos e entre sujeitos de forma e que não podem ser quantificadas (PRODANOV; FREITAS, 2013).

Quanto à natureza, a pesquisa pode ser classificada como básica, aquela que possui como resultado novos conhecimentos e teorias, ou aplicada, a que possui como objetivo a execução de conceitos de forma prática, com o intuito de solucionar problemas (PRODANOV; FREITAS, 2013).

Ao considerar os objetivos da pesquisa Gil (2002) define três possibilidades de classificação, sendo essas: exploratória, descritiva e explicativa. Pesquisas exploratórias buscam construir hipóteses e aprimorar ideias; pesquisas descritivas procuraram entender características de uma amostra e, por fim, as pesquisas explicativas tentam identificar fatores que contribuem ou são responsáveis por um fenômeno (GIL, 2002).

Os procedimentos técnicos definem a maneira pela qual os dados necessários para a elaboração da pesquisa são obtidos. Assim, são definidos dois grupos de procedimentos técnicos: (i) fontes de papel (pesquisa bibliográfica e pesquisa documental) e (ii) fornecidos por pessoas (pesquisa experimental, pesquisa ex-post facto, levantamento bibliográfico, estudo de caso, pesquisa-ação e pesquisa participante) (PRODANOV; FREITAS, 2013).

3.2 Classificação metodológica

Para a classificação da metodologia de pesquisa adotada, foram consideradas as categorias relacionadas a abordagem, natureza, objetivos e procedimentos técnicos. Esta

classificação da metodologia está representada na Figura 17.

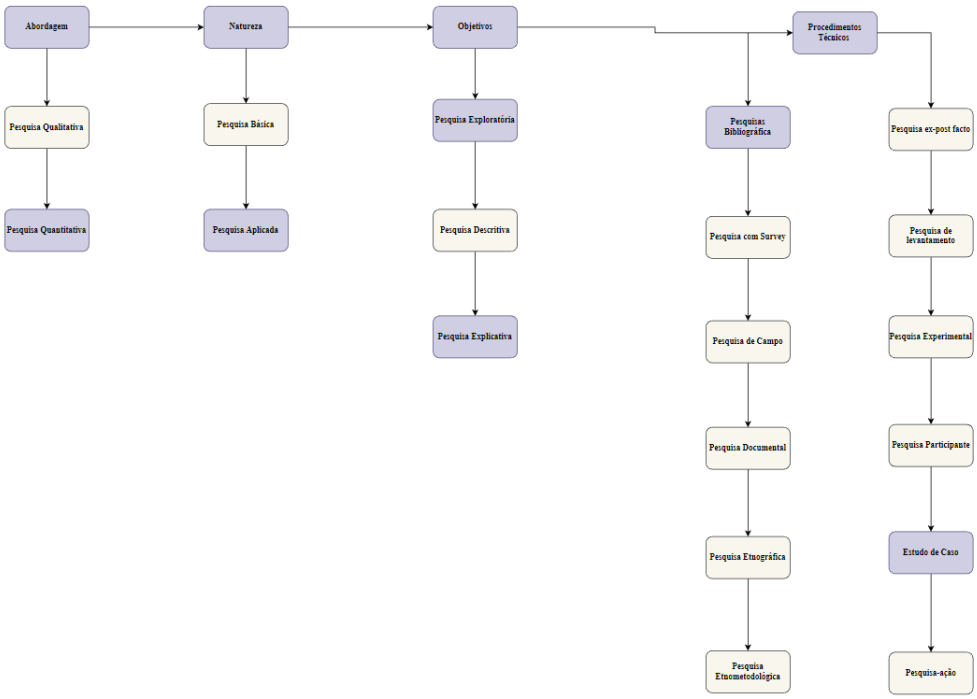


Figura 17 – Classificação da pesquisa quanto a abordagem, natureza, objetivos e procedimentos técnicos.

Fonte: Autores.

A abordagem desta pesquisa é classificada como quantitativa, pois os dados que serão coletados serão quantificados e, a partir da análise desses serão obtidos resultados numéricos que poderão ser representados como gráficos e tabelas. Quanto à natureza, a pesquisa é aplicada com sua execução focada na identificação do grau de previsão dos indicadores. Isso será realizado com a aplicação de algoritmos de aprendizado de máquina. Quanto aos objetivos, a pesquisa é classificada como pesquisa explicativa e exploratória; é explicativa, pois tem como objetivo a identificação de fatores relacionados à previsão no ensino superior; e é exploratória, por a forma de utilização destes indicadores ser realizada por meio de diferentes análises dos resultados obtidos de uma amostra de dados.

Os procedimentos técnicos utilizados para o desenvolvimento da pesquisa são:

Pesquisa bibliográfica realizada por meio de uma revisão sistemática da literatura com aplicação de um protocolo de pesquisa e um estudo bibliométrico. As bases de dados utilizadas durante a pesquisa bibliográfica foram IEEE Xplore[1] e Scopus[2].

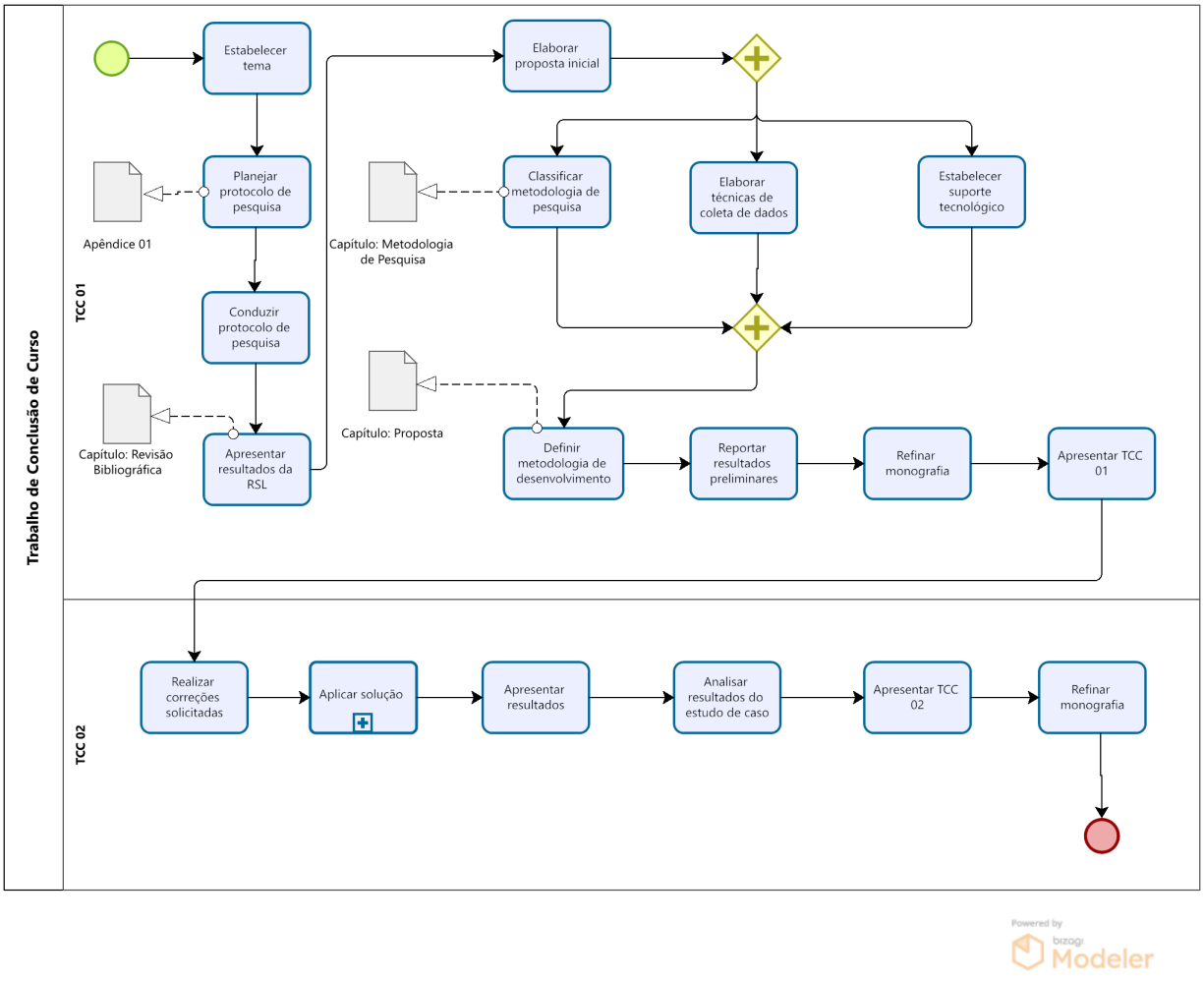
Estudo de Caso que, segundo Gil (2002), é uma análise minuciosa de poucos objetos buscando a preservação de suas características e a delimitação do contexto aos

[1] Pode ser acessado em: https://www.ieee.org/
[2] Pode ser acessado em: https://elsevier.com/

quais está inserido. Dessa forma, será utilizado no contexto da previsão do ensino superior dos alunos do campus Faculdade do Gama da Universidade de Brasília.

3.3 Fluxo de Trabalho

Com base na classificação da pesquisa e nas atividades necessárias para a realização do projeto, propõe-se um fluxo de trabalho. O processo é dividido em duas etapas, sendo estas: TCC 1 e TCC 2. Na Figura 3.3, é possível visualizar o processo de trabalho completo. Nesta seção, será apresentada uma breve descrição de cada atividade.



Estabelecer Tema: a primeira etapa compreende a definição do tema e a delimitação do escopo. Essas atividades são feitas com a cooperação dos orientadores e orientandos e com base no referencial teórico obtido sobre a área. O tema definido foca na

previsão da evasão de alunos do ensino superior, no contexto de alunos de engenharia do campus FGA. A partir da aplicação de diferentes algoritmos de aprendizado de máquina e dos resultados obtidos, será realizada a identificação dos fatores que impactam a evasão.

Planejar Protocolo: esta etapa busca a realização do planejamento do protocolo de pesquisa, como a definição das perguntas da revisão sistemática, da escolha das palavras-chave, do levantamento de dados importantes que precisam ser coletados, da definição dos critérios de inclusão e exclusão e dos filtros de qualidade.

Conduzir Protocolo de Pesquisa: esta etapa tem como objetivo a aplicação do protocolo de pesquisa planejado, a partir da execução das strings de busca nas bases selecionadas, filtragem dos artigos obtidos na pesquisa e leitura completa destes. Além disso, compreende a síntese dos artigos, na qual as informações relevantes para a revisão sistemática são extraídas.

Apresentar Resultados da Revisão Sistemática: esta etapa busca realizar análise nas informações obtidas por meio da síntese dos artigos. É nesta fase que gráficos são criados, resultados são quantificados e o estudo bibliométrico é aplicado.

Elaborar Proposta Inicial: esta etapa tem como objetivo definir uma proposta que viabilize a realização do comparativo entre os sistemas propostos na etapa de delimitar o tema.

Classificar Metodologia de Pesquisa: esta etapa tem como objetivo classificar a metodologia utilizada no projeto quanto à abordagem, natureza, aos objetivos e procedimentos técnicos.

Definir Metodologia de Desenvolvimento: após a classificação da metodologia, é realizada a definição da metodologia de desenvolvimento, onde é considerado todas as etapas necessárias para a realização da fase prática desta pesquisa.

Estabelecer Suporte Tecnológico: esta etapa tem como objetivo descrever as tecnologias utilizadas neste estudo. Dentre essas, estão incluídas ferramentas de pesquisa, ferramentas de avaliação e validação, e ferramentas de desenvolvimento.

Elaborar Técnicas de Coleta de Dados: esta etapa tem como objetivo desenvolver a forma como os dados devem ser coletados. É importante ressaltar que a coleta de dados está limitada aos dados disponíveis na plataforma saga.

Reportar Resultados Preliminares: esta etapa tem como objetivo acordar sobre os principais resultados obtidos com as atividades realizadas ao longo do TCC 1.

Refinar Monografia: esta etapa consiste em um novo ciclo de revisões com o objetivo de refinar ainda mais a documentação do projeto. Cabe ressaltar tambpem que o processo de revisão é algo realizado de forma constante ao longo do TCC 1 e não apenas nessa atividade.

Apresentar TCC 1: esta etapa consiste na apresentação do Trabalho de Conclusão de Curso 01 para os membros da banca avaliadora.

Realizar Correções Solicitadas: esta etapa tem como objetivo o refinamento do trabalho com as sugestões e correções propostas pelos membros da banca avaliadora.

Aplicar Solução: esta etapa tem como objetivo a execução dos algoritmos no indicadores selecionados previamente. A partir desta aplicação, será feita uma análise cruzada dos resultados obtidos. O fluxo da aplicação da solução é descrito na seção Metodologia de Desenvolvimento.

Apresentar Resultados: esta etapa tem como objetivo apresentar os resultados obtidos a partir da aplicação dos algoritmos e da análise dos resultados. Esta fase leva em consideração os indicadores de medição de precisão, recall, média ponderada, fscore e acurácia.

Analisar Resultados do Estudo de Caso: esta etapa tem como objetivo analisar os resultados obtidos e utilizá-los como base para a identificação dos fatores com maior impacto na evasão do ensino superior.

Apresentar TCC 2: esta etapa tem como objetivo apresentar o resultado final do trabalho de conclusão de curso para os membros da banca avaliadora.

Refinar Monografia: Esta etapa consiste na realização de correções e os refinamentos apontados pela banca após a apresentação do TCC 2.

4 Aplicação de Modelos de Aprendizado de Máquina

4.1 Visão Geral da Aplicação

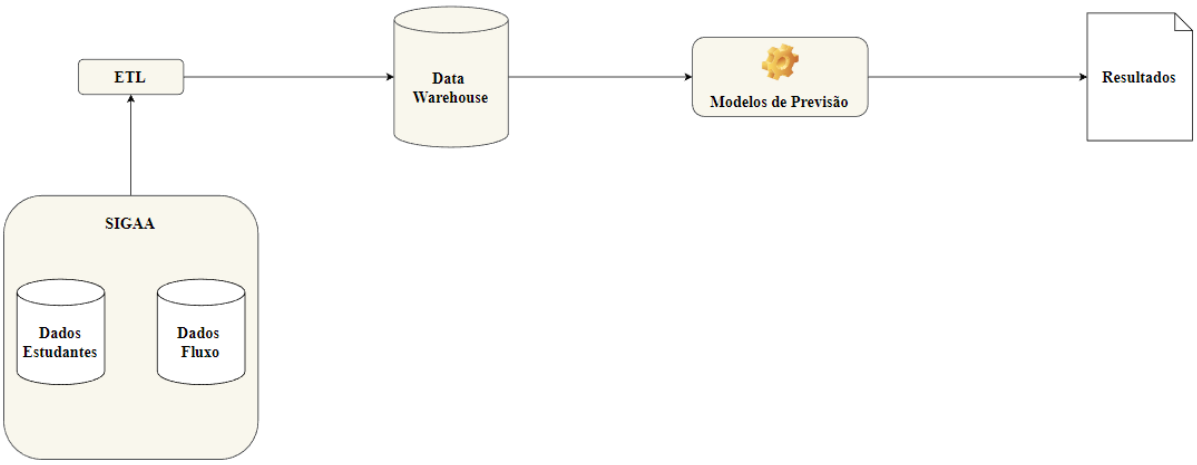


Figura 18 – Diagrama de blocos do uso dos indicadores. Fonte: Autoras

A proposta de uso dos indicadores identificados nesta monografia considera dados sobre alunos dos cursos Engenharia Aeroespacial, Engenharia Automotiva, Engenharia Eletrônica, Engenharia de Energia e Engenharia de Software do campus Faculdade do Gama da Universidade de Brasília. Esses dados foram coletados da plataforma SIGAA e disponibilizados pelos orientadores. São dados confidenciais e não estão abertos para fácil obtenção. Esses dados abrangem informações sobre os fluxos dos cinco cursos, as disciplinas obrigatórias e optativas, o número de vezes que um aluno fez determinada matéria e seu status atual em sua graduação.

Esses dados já estão tratados e presentes em uma *Dataware House* para seu uso. Apesar de não ser necessária a etapa de ETL nesta proposta, é importante ressaltar que, para as futuras extrações de dados será necessário passar pela etapa de preparação de dados antes de sua utilização. Por esse motivo, se fez relevante a inclusão da etapa de ETL no diagrama de blocos presente na Figura 18.

Os indicadores presentes no banco de dados foram utilizados em algoritmos de aprendizado de máquina. Os resultados obtidos dos modelos são descritos em relatórios para a coordenação dos cursos e apresentados em *dashboards* , para melhor visualização das informações obtidas. Cada um dos passos presentes na Figura 18 serão detalhados nas seções subsequentes.

4.2 Fonte de Dados

Segundo Niessen, Meijer e Tendeiro (2016), a Mineração de Dados Educacionais (MDE) utiliza técnicas de Mineração de dados para extrair informações importantes de dados educacionais. A MDE é uma área de pesquisa que perfeiçoa métodos de investigação para dados obtidos a partir de cenários educacionais.

O Sistema Integrado de Gestão de Atividades Acadêmicas - UnB (SIGAA) foi a fonte dos dados observada para esta pesquisa. Esses dados foram coletados da plataforma SIGAA e disponibilizados pelos orientadores. Com eles foi possível obter informações sobre cada uma das engenharias do campus Gama da Universidade de Brasília e sobre os estudantes que tiveram algum envolvimento com algum destes cursos entre os semestres 2015/02 e 2019/1.

4.2.1 ETL

O ETL é uma técnica de processamento de dados, que os extrai de uma base dados, os transforma de acordo com a necessidade ou regra de negócio e os armazena em outra base de dados. Os dados para esse trabalho extraídos do Sistema Integrado de Gestão de Atividades da UnB, foi realizada a sua transformação e a sua limpeza de acordo com os fatores identificados na RSL e, por fim, armazenados em relatórios para análise. Com o ETL, foi possível definir a qualidade das informações e regrar a usabilidade e a manipulação, desses dados, de forma estratégica e padronizada.

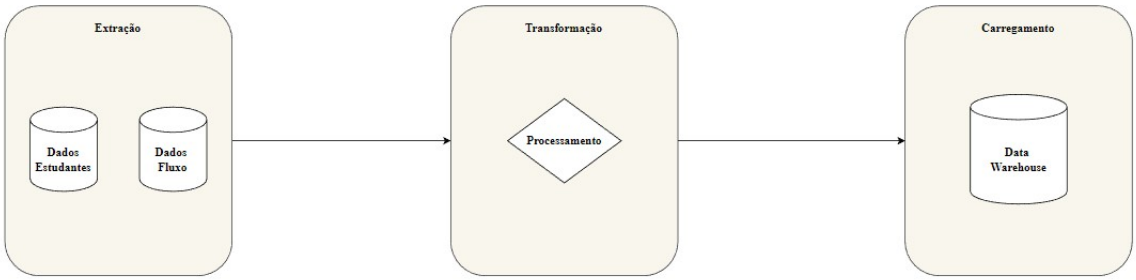


Figura 19 – Exemplificação de ETL. Fonte: Autoras

4.2.1.1 Extração

Nesse trabalho a etapa de extração foi relevante para padronizar os dados, o que auxiliou na manipulação desses em outras etapas da ETL. Os dados analisados vieram em duas planilhas, que foram inseridas em um Banco de Dados MySQL, para ampliação da segurança e usabilidade deles. A primeira tabela denominada “dados”, apresenta dados relativos ao estudante, que estão distribuídos em seis colunas.

- IDFluxo: identificação do fluxo, visto que um curso pode ter mais de um fluxo.

- Status: identificação do status do aluno, se ele está cursando a graduação, formado ou evadido.
- IDDisciplina: identificação do código da disciplina dentro do SIGAA.
- IDEstudante: identificação do estudante dentro do SIGAA.
- Matrículas: identificação de quantas vezes o estudante fez determinada disciplina.
- Ingresso: identificação do semestre em que o aluno ingressou na Universidade.

A segunda tabela nomeada de "fluxos", apresenta os dados relativos aos fluxos dos cursos, que estão distribuídos em cinco colunas.

- Abrev: identificação do Curso.
- IDFluxo: identificação do fluxo, visto que um curso pode ter mais de um fluxo.
- IDDisciplina: identificação do código da disciplina dentro do SIGAA.
- Período: identificação de qual período do curso a disciplina deve ser feita.
- Natureza: identificação se a matéria é obrigatória ou optativa.

4.2.1.2 Transformação

Foram feitas padronizações dos dados para que fosse possível realizar a categorização e a identificação de fatores associados a esses dados. Além disso, foram aplicados filtros para melhores análises dos modelos de Aprendizado de Máquina, na fase pós ETL, conforme a Figura 20.

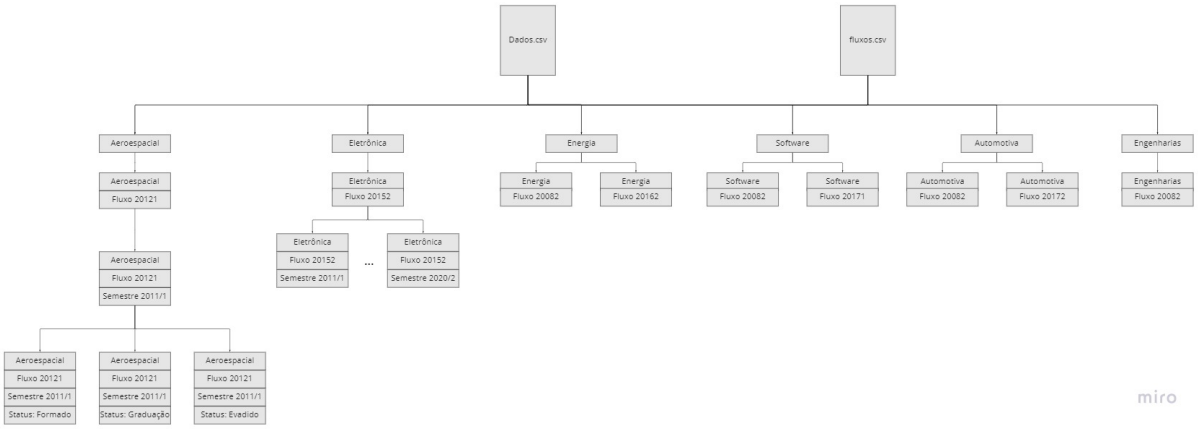


Figura 20 – Esquemático dos Filtros. Fonte: Autoras.

Nesse trabalho os filtros abaixo foram aplicados nas *queries* de busca dentro do Banco de Dados, para que fossem aplicados nos modelos, apenas, os mais interessantes para o projeto.

1. Curso.
2. Fluxo.
3. Semestre.
4. Status.

4.2.1.3 Carregamento

Nesta fase é preciso carregar os dados transformados, a tabelas do tipo csv tiveram seus dados extraídos, transformados, limpos e adicionados em uma Base de Dados. Esse carregamento é feito de forma automatizada após cada execução do ETL na plataforma RStudio.

4.2.2 Data Warehouse

Um sistema *Data Warehouse* está associado a um processo de ETL com grande volume de dados provenientes de fontes heterógeneas, sua estrutura foi desenvolvida para facilitar a análise desses dados e a sua aplicação em modelos de AM. Cada tabela desenvolvida está integrada às outras e esse sistema pode ser incrementado de acordo com a necessidade, conforme a Figura 21.

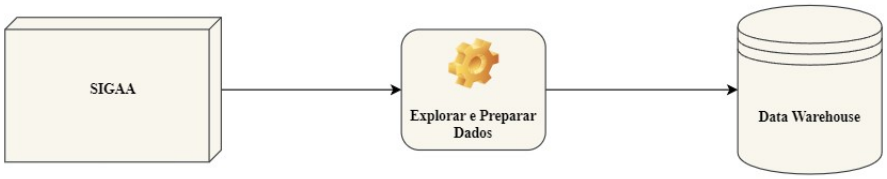


Figura 21 – Esquemático Data Warehouse. Fonte: AWS

Foi decidido por usar esse processo, pelos benefícios citados abaixo:

- Decisão a seguir com dados.
- Consolidação de Dados.
- Análise histórica de dados.
- Qualidade, consistência e precisão das informações.
- Melhor desempenho.

4.3 Ferramentas Utilizadas

4.3.1 Linguagem

R é uma linguagem de programação estatística e gráfica que pode ser utilizada para a manipulação, análise e visualização de dados. Possui código aberto, o que permite sua utilização para visualização, modificação e distribuição de forma gratuita. Essa característica colabora para seu progresso, visto que existe uma grande comunidade ativa de desenvolvedores que auxiliam para a aprimoramento do sistema.

Essa linguagem foi escolhida para esse trabalho por ser mais didático, simples e de fácil aprendizagem. O R tem funções mais intuitivas e a maior parte das funções necessárias para análise de dados já pertencer ao R-base, não necessitando instalação de bibliotecas, o que difere de outras linguagens, como por exemplo o Python.

4.3.2 Banco de Dados

4.3.3 MySQL 8.0

Segundo, Oracle (2022), o MySQL é um Sistema de Gerenciamento de Banco de Dados, que utiliza a linguagem SQL ou Linguagem de Consulta Estruturada. E atualmente é um dos sistemas de gerenciamento de bancos de dados mais utilizado.

O MySQL 8.0 foi escolhido para esse trabalho pelos motivos abaixo:

- Portabilidade, suporta quase todas as plataformas existentes atualmente, tais como, como Mac OS X, Solaris, Windows etc.
- Ótima estabilidade e desempenho.
- Fácil manipulação e aprendizado, por ser muito intuitivo e didático.

4.3.4 DBeaver 22.1.1

O DBeaver é um programa multiplataforma, que tem por objetivo conectar, administrar e manipular vários tipos de banco de dados. Ele pode ser utilizado em Windows, Linux e macOS, além de permitir a conexão com bancos MySQL, PostgreSQL, Firebird, SQL Server etc.

Para bancos de dados relacionais , ele usa a interface de programação de aplicativos (API) JDBC para interagir com bancos de dados por meio de um driver JDBC. Ele fornece um editor que suporta preenchimento de código e realce de sintaxe . Ele fornece uma arquitetura de *plug-in* que permite aos usuários modificar o comportamento do aplicativo

para fornecer funcionalidades específicas do banco de dados ou recursos independentes do banco de dados. Community (2022)

4.4 Modelos Utilizados

Foram utilizados modelos de Aprendizado de Máquina (AM) para a uso de fatores que podem prever a evasão de estudantes.

4.4.1 Aprendizado de Máquina (AM)

O AM é uma das categorias da Inteligência Artificial que tem como foco produzir softwares aptos para obter conhecimento de forma automática. Um sistema de AM toma decisões baseado em experiências de soluções bem-sucedidas. Existem diversos tipos de modelos de AM em que cada um possui características particulares que possibilita diversas implementações e adaptações a diversos cenários.

A inferência lógica denominada como indução é um processo em que um conceito específico, com um número alto de bons exemplos, é generalizado. Com a indução é possível gerar resultado genéricos sobre um grupo de análise, quando aplicada em Aprendizado de Máquina. O Aprendizado de Máquina Indutivo Supervisionado foi utilizado nesse trabalho, onde um conjunto de dados para treinamento foi disponibilizado para os modelos de AM.

O Objetivo da utilização dessa técnica foi construir um classificador, onde haviam duas possibilidades o estudante evadir ou não, a partir de fatores previamente levantados nesse trabalho.

4.4.2 Árvore de Decisão

A Árvore de Decisão é um dos modelos mais práticos e mais usados em inferência indutiva supervisionada. Os dados são divididos de forma sucessiva em conjuntos cada vez menores e intrínsecos em termos de características até alcançar uma dimensão simplificada e fichada.

Nesse modelo são estabelecidos nós que se co-relacionam por hierarquia. A raiz são os dados que vêm da base de informações, as folhas são os resultados finais, conforme a Figura 22. Na ligação entre os nós existem questões que abordam um preceito sobre a evasão estudantil. A árvore de decisão segue a técnica de recursividade, esse padrão é repetido em níveis cada vez mais profundos.

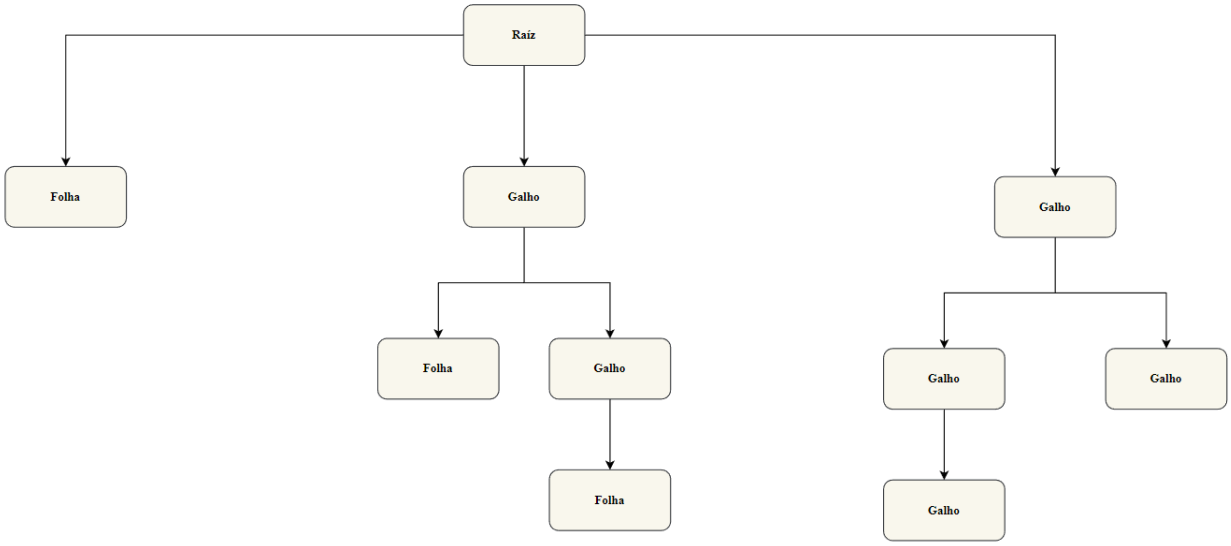


Figura 22 – Esquemático de Árvore de Decisão. Fonte: Autores.

4.4.3 Floresta Aleatória

Floresta aleatória é um modelo de aprendizado de máquina supervisionado. Ele foi utilizado nesse trabalho devido a sua precisão e flexibilidade, além do fato de poder ser usado para tarefas de classificação. É denominado como floresta pois são criadas combinações de árvores de decisão, a caracterização principal dessa técnica é a utilização do método *Bagging* ou *Bootstrap Aggregating* que consiste em construir um algoritmo de classificadores agregados, como representado na Figura 23. Dessa forma, a predição tem uma maior acurácia.

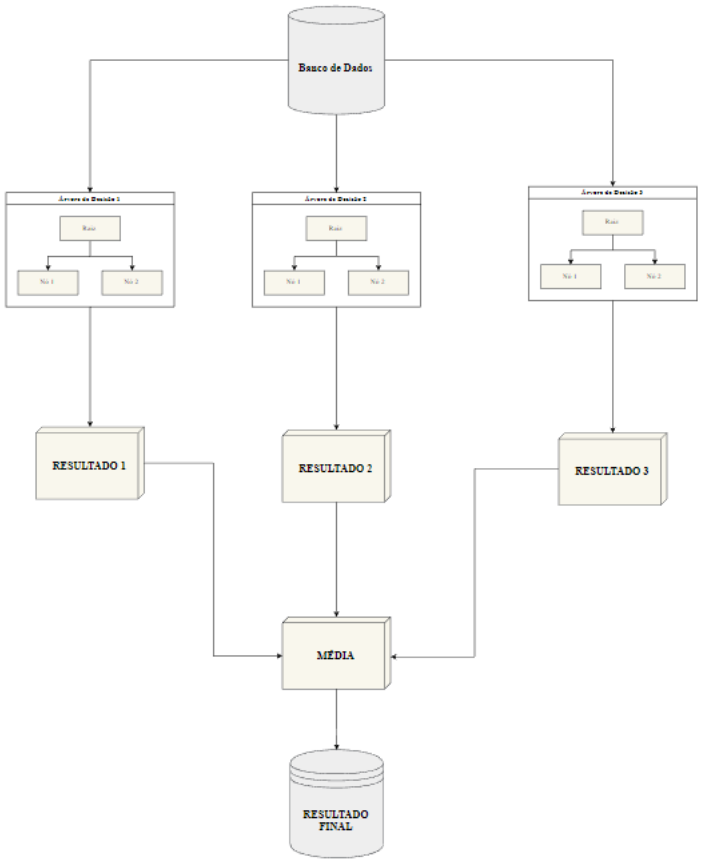


Figura 23 – Esquemático de Floresta Aleatória. Fonte: Autores.

Nesse modelo existe uma aleatoriedade a mais na busca de características, ou seja, a melhor característica é pesquisada em um subconjunto aleatório de características em vez de ser procurada na partição de nós. Com isso, o aumento da diversidade cria modelos superiores.

4.4.4 C5.0

O algoritmo C5.0 é uma técnica que utiliza de árvores de decisão para solucionar a maioria dos tipos de problemas. Em comparação com outros modelos de aprendizado de máquina, como por exemplo redes neurais, ele apresenta um desempenho superior aos demais e possui uma facilidade em sua implementação. Por conta dos fatores supracitados esse modelo foi um dos escolhidos para predição dos estudantes da FGA.

4.4.5 Aplicação dos Indicadores com Aprendizado de Máquina para prever a Evasão Acadêmica

A partir de três modelos distintos, foi criado um algoritmo para realizar a aplicação dos fatores relacionados a evasão dos estudantes de Engenharia. Assim, foi utilizado o pacote DBI, que consiste em uma biblioteca de definição de interface entre a linguagem

R e Sistemas de Gerenciamento de Bancos de Dados. Nesse trabalho o DBI foi utilizado para realizar a conexão com banco de dados através de chamadas de funções específicas.

```
library("DBI")
library(RMySQL)
```

A conexão com o banco de dados, que contém as informações necessárias para a previsão, foi realizada com o uso de duas funções específicas. A primeira delas é a dbDriver(), em que foi necessário informar qual tipo de banco estava sendo utilizado. A segunda é a dbConnect(), que recebe como argumentos as informações de usuário e senha do banco de dados.

```
drv <- dbDriver("MySQL")
mydb = dbConnect(
    drv,
    user='root',
    password='password',
    dbname='student_data',
    host='localhost')
```

Os modelos foram aplicados em oito semestres distintos de cinco cursos. Com isso, surgiu a necessidade de declarar variáveis globais para a execução do código. As variáveis cursos e semestres contém todas as diferentes combinações de ingresso possíveis, para as cinco graduações. Já a variável resultados inicia-se como lista vazia que é utilizada para guardar os resultados de acurácia dos modelos.

```
cursos <- c(
    "AUTOMOTIVA",
    "AEROESPACIAL",
    "SOFTWARE",
    "ENERGIA",
    "ELETRONICA")
semestres <- c(
    "2015/2",
    "2016/1",
    "2016/2",
    "2017/1",
    "2017/2",
    "2018/1",
    "2018/2",
```

```
    "2019/1")
resultados <- c()
```

As variáveis cursos e semestres foram utilizadas em dois laços *for*, dessa forma foi possível garantir que todos os semestres de um curso tivessem modelos únicos e aplicados em outros semestres.

```
for (curso in cursos) {
  for (semestre in semestres) {
    // código de criação de um modelo

    for (ingresso in semestres) {
      // código de previsão dos modelos
    }
  }
}
```

A primeira etapa para a criação dos modelos é o preparo dos dados. Com esse intuito, foi necessário realizar a busca dos dados que seriam utilizados. A busca da lista de disciplinas foi feita a partir de uma *query* na tabela fluxo e seu resultado foi guardado na variável listaDisciplina.

```
dataset <- dbGetQuery(
    mydb,
    paste0(
        "select Disciplina from student\_data.fluxos
        where Abrev ='", curso, "'")
    )
listaDisciplinas <- unique(dataset[, "Disciplina"])
```

Este mesmo processo, também, foi feito para pegar a relação de alunos de cada semestre e seu resultado está presente na variável listaEstudante.

```
dataset <- dbGetQuery(
    mydb,
    paste0(
        "SELECT IDFluxo, Status, Matriculas,
        Ingresso, IDDisciplina, IDEstudante
        FROM student_data.dados as dados
        INNER JOIN student_data.fluxos as fluxos
```

```
        ON dados.IDFluxo = fluxos.Fluxo
        where Abrev = '", curso,"'
        and Ingresso = '", semestre,"' ;")
    )
listaEstudantes <- unique(dataset[, "IDEstudante"])
```

As duas listas foram utilizadas para a criação de uma matriz, onde suas linhas recebem como nome a variável listaEstudante e as colunas a listaDisciplina.

```
dados <- data.frame(
    matrix(
        0,
        nrow = length(listaEstudantes),
        ncol = length(listaDisciplinas)
    )
)
rownames(dados) <- listaEstudantes
colnames(dados) <- listaDisciplinas
```

A matriz criada precisa ter a quantidade de vezes que um aluno realizou uma matéria. Isso foi feito com o uso de um laço *for*, que preenche a interseção de uma matéria com um estudante com o número de matrículas existentes, e caso não tenha, preenche com 0.

```
for (l in 1:nrow(dataset)) {
    i <- as.character(dataset[l, "IDEstudante"])
    j <- listaDisciplinas
    Tentativas <- as.integer(dataset[l, "Matriculas"])
    dados[i, j] <- Tentativas
}
```

A próxima fase de preparação de dados implica na utilização do status, pois é a variável que será prevista. Por essa razão, é necessário saber qual o atual status de um aluno. Então, uma lista vazia Status é inicializada, a variável listaStatus recebe todos os status dos alunos do semestre e um curso específico. Em um laço *for* a lista Status recebe os status atuais da listaStatus. E por fim, a lista de Status é adicionada à matriz dados criado anteriormente.

```
Status <- c()
listaStatus <- unique(dataset[ , c("IDEstudante", "Status")])
```

```
for (l in 1:nrow(listaStatus)) {
  Status <- c(Status, listaStatus[l, "Status"])
}

dados <- cbind(Status, dados)
```

Na execução de modelos de aprendizado de máquina é importante lidar com dados nulos. Consequentemente, foi utilizada a função is.na(), que verifica a existência de informações vazias na matriz dados, oportunizando, assim, sua substituição por 0.

```
dados[is.na(dados)] <- 0
```

O último passo de preparação de dados foi focado em transformar os possíveis status em níveis, dessa forma foi possível utilizá-los nos algoritmos de previsão.

```
dados$Status <- factor(dados$Status)
levels(dados$Status) <- c("Formado", "Matriculado", "Evadido")
```

Com os dados transformados e com as bibliotecas rpart, C5.0 e RandomForest, foi viável criar os três modelos. Todos os algoritmos utilizaram Status como variável a ser prevista e as disciplinas realizadas pelo aluno como variáveis exploratórias. É importante ressaltar que para o modelo de Floresta Aleatória foi necessário informar o número máximo de tentativas, o qual corresponde a raiz quadrada da quantidade de variáveis exploratórias.

```
library(rpart)
library(C50)
library(randomForest)

modeloArvoreDecisao <- rpart(
    Status ~ .,
    dados
)

modeloC50 <- C5.0(
    Status ~ .,
    data = dados,
    trials = 10
)
```

```
quantidadeVariaveis <- length(listaDisciplinas)
numero <- sqrt(quantidadeVariaveis)
max <- ceiling(numero)

dados$Status <- factor(dados$Status)
modeloRandomForest <- randomForest(
    Status ~ .,
    data = dados,
    ntree = 500,
    mtry = max,
    importance = TRUE
)
```

A previsão dos semestres foi realizada embasada nos modelos criados. Foi necessário buscar a informação dos alunos do semestre previsto e realizar a criação da matriz de dados de previsão de forma similar ao feito na criação. Em seguida a função de previsão foi chamada, a matriz de confusão foi desenhada e a acurácia foi calculada para cada um dos modelos.

```
previsaoAD <- predict(
    modeloArvoreDecisao,
    type = "class",
    newdata = dadosPrevisao
)

matrizConfArvore <- table(dadosPrevisao$Status, previsaoAD)
acuraciaAD <- sum(diag(matrizConfArvore))/sum(matrizConfArvore)

previsaoC50 <- predict(
    modeloC50,
    type = "class",
    newdata = dadosPrevisao
)

matrizConfC50 <- table(dadosPrevisao$Status, previsaoC50)
acuraciaC50 <- sum(diag(matrizConfC50))/sum(matrizConfC50)

previsaoRF <- predict(
    modeloRandomForest,
    type = "class",
```

```
    newdata = dadosPrevisao
)

matrizConfRF <- table(dadosPrevisao$Status, previsaoRF)
acuraciaRF <- sum(diag(matrizConfRF))/sum(matrizConfRF)
```

Os resultados obtidos foram guardados com o auxílio da função rbind(), foram passados como argumentos a lista resultados, o modelo aplicado, o semestre previsto e a acurácia dos três modelos.

```
resultados <- rbind(
    resultados,
    c(
        modelo,
        ingresso,
        acuraciaAD,
        acuraciaC50,
        acuraciaRF
    )
)
```

Depois da execução de todos os laços *for*, a lista de resultados foi transformada em um arquivo .csv e o banco de dados foi desconectado.

5 resultados

5.1 Engenharia Automotiva

Os modelos desenvolvidos para o curso de engenharia automotiva foram treinados com dados do semestre 2015/2 e 2016/1 e, tiveram acurácias iguais. O melhor resultado obtido com estes modelos foi no semestre de 2016/1, com uma acurácia total de 0,615 para todos os algoritmos utilizados. Nos demais semestres a acurácia foi menor que 0.267, como pode ser observado na Tabela 5.

As acurácias atingidas pelos modelos de 2016/2, 2017/1 e 2017/2 foram baixas para os semestres 2015/1 e 2015/2, porém altas para os demais. Cinco previsões conseguiram acurácias acima de 80%. Os resultados detalhados estão na Tabela 6.

O modelo 2018/1 conseguiu os mesmos valores de acurácia, nos algoritmos árvore de decisão e C5.0, que os semestres de 2016/2, 2017/1 e 2017/2. Porém, é considerado que na floresta aleatória houve variação nas acurácias. Na Tabela 7 é possível observar que a maior acurácia dos modelos, árvore de decisão e C5.0, foi de 0.863 na previsão do semestre 2016/2. Além disso, o desempenho da floresta aleatória foi, de forma geral, inferior aos demais modelos.

Esse comportamento de variação de acurácias entre árvore de decisão e c5.0 com floresta aleatória se repetiu nos últimos dois modelos de previsão. Neles houve o aumento de acurácia nos semestres 2016/2, 2017/1, 2017/2, e 2019/1, em relação ao último modelo aplicado, como observado na Tabela 8.

Algo importante de ser ressaltado e que pode ser notado quando os resultados dos modelos são comparados, é que a partir do semestre 2016/2 as acurácias obtidas pela árvore de decisão e C5.0 foram as mesmas. As possíveis causas para esse comportamento

Semestre	**Acurácia Árvore de Decisão**	**Acurácia C5.0**	**Acurácia Floresta Aleatória**
2015/2	0,590	0,590	0,590
2016/1	0,615	0,615	0,615
2016/2	0,136	0,136	0,136
2017/1	0,176	0,176	0,176
2017/2	0,167	0,167	0,167
2018/1	0,181	0,181	0,181
2018/2	0,173	0,173	0,173
2019/1	0,267	0,267	0,267

Tabela 5 – Tabela de Resultados do modelo Automotiva 2015/2 e Automotiva 2016/1

Semestre	**Acurácia Árvore de Decisão**	**Acurácia C5.0**	**Acurácia Floresta Aleatória**
2015/2	0,090	0,090	0,090
2016/1	0,153	0,153	0,153
2016/2	0,863	0,863	0,863
2017/1	0,823	0,823	0,823
2017/2	0,833	0,833	0,833
2018/1	0,818	0,818	0,818
2018/2	0,826	0,826	0,826
2019/1	0,733	0,733	0,733

Tabela 6 – Tabela de Resultados do modelo Automotiva 2016/2, 2017/1 e 2017/2

Semestre	**Acurácia Árvore de Decisão**	**Acurácia C5.0**	**Acurácia Floresta Aleatória**
2015/2	0,090	0,090	0,181
2016/1	0,153	0,153	0,307
2016/2	0,863	0,863	0,636
2017/1	0,823	0,823	0,588
2017/2	0,833	0,833	0,625
2018/1	0,818	0,818	0,909
2018/2	0,826	0,826	0,608
2019/1	0,733	0,733	0,333

Tabela 7 – Tabela de Resultados do modelo Automotiva 2018/1

Semestre	**Acurácia Árvore de Decisão**	**Acurácia C5.0**	**Acurácia Floresta Aleatória**
2015/2	0,090	0,090	0,136
2016/1	0,153	0,153	0,230
2016/2	0,863	0,863	0,772
2017/1	0,823	0,823	0,823
2017/2	0,833	0,833	0,791
2018/1	0,818	0,818	0,818
2018/2	0,826	0,826	0,860
2019/1	0,733	0,733	0,8

Tabela 8 – Tabela de Resultados do modelo Automotiva 2018/2 e 2019/1

Semestre	**Acurácia Árvore de Decisão**	**Acurácia C5.0**	**Acurácia Floresta Aleatória**
2015/2	0,391	0,0391	0,478
2016/1	0,045	0,045	0,045
2016/2	0,214	0,214	0,214
2017/1	0,267	0,267	0,267
2017/2	0	0	0,192
2018/1	0	0	0,104
2018/2	0	0	0,342
2019/1	0	0	0,38

Tabela 9 – Tabela de Resultados do modelo Aeroespacial 2015/2

Semestre	**Acurácia Árvore de Decisão**	**Acurácia C5.0**	**Acurácia Floresta Aleatória**
2015/2	0,347	0,347	0,347
2016/1	0,818	0,818	0,818
2016/2	0,714	0,714	0,714
2017/1	0,711	0,711	0,711
2017/2	0,153	0,153	0,153
2018/1	0,145	0,145	0,145
2018/2	0,078	0,078	0,078
2019/1	0,08	0,08	0,08

Tabela 10 – Tabela de Resultados do modelo Aeroespacial 2016/1, 2016/2 e 2017/1

podem estar relacionados a falta de mudança de fluxo e pelo uso de dados de alunos que estão relacionados com o curso desde 2016/2, em específico para aqueles que estão matriculados.

5.2 Engenharia Aeroespacial

O primeiro modelo utilizado nas previsões do aeroespacial não conseguiu bons resultados. Nos semestres a partir de 2017/2 os algoritmos árvore de decisão e C5.0 tiveram acurácia zerado, enquanto a floresta aleatória teve valores baixos, como pode ser observado na Tabela 9.

É notório na Tabela 10 que os modelos 2016/1, 2016/2 e 2017/1 obtiveram acurácias iguais entre si e entre os algoritmos implementados. É importante salientar que suas previsões para os demais semestres foram baixas, sendo a maior acurácia entre elas 0.345 quando aplicada no semestre 2015/2.

Os modelos dos semestres 2017/2, 2018/2 e 2019/1 seguem o mesmo comportamento e compartilham entre si os mesmos resultados de acurácia entre os algoritmos. Contudo, sua distinção em relação aos semestres anteriores está na acurácia em suas previsões, onde os modelos anteriores tinham sido previstos com boas acurácias e os modelos

Semestre	**Acurácia Árvore de Decisão**	**Acurácia C5.0**	**Acurácia Floresta Aleatória**
2015/2	0,260	0,260	0,260
2016/1	0,136	0,136	0,136
2016/2	0,071	0,071	0,071
2017/1	0,022	0,022	0,022
2017/2	0,846	0,846	0,846
2018/1	0,854	0,854	0,854
2018/2	0,921	0,921	0,921
2019/1	0,92	0,92	0,92

Tabela 11 – Tabela de Resultados do modelo Aeroespacial 2017/2, 2018/2 e 2019/1

Semestre	**Acurácia Árvore de Decisão**	**Acurácia C5.0**	**Acurácia Floresta Aleatória**
2015/2	0,260	0,260	0,260
2016/1	0,136	0,136	0,136
2016/2	0,071	0,071	0,071
2017/1	0,022	0,022	0,022
2017/2	0,846	0,846	0,846
2018/1	0,854	0,854	0,875
2018/2	0,921	0,921	0,921
2019/1	0,92	0,92	0,92

Tabela 12 – Tabela de Resultados do modelo Aeroespacial 2018/1

destes semestres previram com acurácias baixas. Isso pode ser observado na Tabela 11.

A Tabela 12 mostra os resultados do modelo 2018/1 que não segue o padrão dos dois últimos grupos. Ele possui os mesmos resultados que os modelos 2017/1, 2018/2 e 2019/1 para todas as previsões, com exceção da acurácia alcançada pela floresta aleatória de 2018/1.

O comportamento dos modelos aplicados no curso de Engenharia Aeroespacial é interessante. Ao observar a comparação de acurácias entre os grupos de modelos entre 2016/1 e 2019/1, com exceção de 2018/1. Os três primeiros semestres, 2016/1, 2016/2 e 2017/1 não conseguem ter acurácias altas para os modelos do segundo grupo, 2017/2, 2018/2 e 2019/1. E o inverso também ocorre.

5.3 Engenharia de Software

O modelo de 2015/2 teve acurácias medianas para os três primeiros semestres, 0,423, 0,526 e 0,737 respectivamente. Para os demais semestres as acurácias ficaram abaixo de 0,2. Evidenciando que as pontuações obtidas pela floresta aleatória se diferem dos resultados dos outros dois algoritmos. Esses detalhes estão presentes na Tabela 13

Semestre	**Acurácia Árvore de Decisão**	**Acurácia C5.0**	**Acurácia Floresta Aleatória**
2015/2	0,423	0,423	0,461
2016/1	0,526	0,526	0,385
2016/2	0,737	0,737	0,540
2017/1	0,149	0,149	0,119
2017/2	0,171	0,171	0,203
2018/1	0,033	0,033	0,033
2018/2	0,110	0,110	0,100
2019/1	0,042	0,042	0,008

Tabela 13 – Tabela de Resultados do modelo Software 2015/2

Semestre	**Acurácia Árvore de Decisão**	**Acurácia C5.0**	**Acurácia Floresta Aleatória**
2015/2	0,423	0,423	0,423
2016/1	0,526	0,526	0,526
2016/2	0,737	0,737	0,737
2017/1	0,149	0,149	0,149
2017/2	0,171	0,171	0,171
2018/1	0,033	0,033	0,033
2018/2	0,110	0,110	0,110
2019/1	0,042	0,042	0,042

Tabela 14 – Tabela de Resultados do modelo de Software 2016/1 e 2016/2

Semestre	**Acurácia Árvore de Decisão**	**Acurácia C5.0**	**Acurácia Floresta Aleatória**
2015/2	0,365	0,365	0,365
2016/1	0,298	0,298	0,298
2016/2	0,049	0,049	0,049
2017/1	0,850	0,850	0,850
2017/2	0,828	0,828	0,828
2018/1	0,967	0,967	0,967
2018/2	0,889	0,889	0,889
2019/1	0,957	0,957	0,957

Tabela 15 – Tabela de Resultados do modelo de Software 2017/1, 2017/2, 2018/1, 2018/2 e 2019/1

Ambos os semestres de 2016 conseguiram os mesmos resultados, sendo a acurácia mais alta obtida 0,737 para o semestre 2016/2. Os modelos para os semestres restantes também tiveram acurácias iguais. Para esse grupo, a maior acurácia atingida foi de 0,957 para o semestre 2019/1. Para ambos os grupos de modelos, os três algoritmos de aprendizado de máquina atingiram a mesma acurácia em suas previsões. A Tabela 14 apresenta os resultados do primeiro grupo e a Tabela 15 detalha as acurácias dos modelos de 2017/1 até 2019/1.

Semestre	**Acurácia Árvore de Decisão**	**Acurácia C5.0**	**Acurácia Floresta Aleatória**
2015/2	0,588	0,705	0,705
2016/1	0,631	0,421	0,421
2016/2	0,075	0,417	0,417
2017/1	0,8	0,6	0,6
2017/2	0,172	0,172	0,172
2018/1	0,8	0,6	0,6
2018/2	0,133	0,3	0,3
2019/1	0,161	0,290	0,290

Tabela 16 – Tabela de Resultados do modelo de Energia 2015/2

Semestre	**Acurácia Árvore de Decisão**	**Acurácia C5.0**	**Acurácia Floresta Aleatória**
2015/2	0,588	0,588	0,588
2016/1	0,631	0,631	0,631
2016/2	0,075	0,075	0,075
2017/1	0,8	0,8	0,8
2017/2	0,172	0,172	0,172
2018/1	0,8	0,8	0,8
2018/2	0,133	0,133	0,133
2019/1	0,161	0,161	0,161

Tabela 17 – Tabela de Resultados do modelo de Energia 2016/1, 2016/2, 2017/1, 2018/1

No geral, observando todos os resultados adquiridos nas previsões de curso de Engenharia de Software, é possível notar que modelos de cinco semestres obtiveram as mesmas acurácias. Quando considerado o contexto destes semestres, foi preciso levar em consideração que houve troca de fluxo em 2017/1 para este curso, essa informação tem peso nas previsões realizadas.

5.4 Engenharia de Energia

O modelo de energia 2015/2 conseguiu resultados similares aos demais modelos, considerando as acurácias da árvore de decisão. Entretanto, as acurácias de floresta aleatória e C5.0 se distinguem das obtidas pelos demais semestres. A maior acurácia obtida foi de 0,8 para os semestres 2017/1 e 2018/1. O detalhamento dos resultados atingidos estão na Tabela 16.

Os modelos dos semestres 2016/1, 2016/2, 2017/1 e 2018/1 dividem entre si os mesmos resultados. As maiores acurácias atingidas em suas previsões foram para os semestres 2017/1 e 2018/1. A principal diferença entre estes modelos e o de 2015/2 são os resultados da floresta aleatória e do C5.0, todos os outros são compartilhados entres os quatro modelos. Isso pode ser visto na Tabela 17.

Semestre	**Acurácia Árvore de Decisão**	**Acurácia C5.0**	**Acurácia Floresta Aleatória**
2015/2	0,352	0,352	0,352
2016/1	0,157	0,263	0,263
2016/2	0,083	0,25	0,25
2017/1	0,05	0,05	0,05
2017/2	0,827	0,896	0,896
2018/1	0,067	0,133	0,133
2018/2	0,867	0,833	0,833
2019/1	0,838	0,774	0,774

Tabela 18 – Tabela de Resultados do modelo de Energia 2017/2

Semestre	**Acurácia Árvore de Decisão**	**Acurácia C5.0**	**Acurácia Floresta Aleatória**
2015/2	0,352	0,352	0,352
2016/1	0,157	0,157	0,157
2016/2	0,083	0,083	0,083
2017/1	0,05	0,05	0,05
2017/2	0,827	0,827	0,827
2018/1	0,067	0,067	0,067
2018/2	0,867	0,867	0,867
2019/1	0,838	0,838	0,838

Tabela 19 – Tabela de Resultados dos modelos de Energia 2018/2 e 2019/1

O modelo de 2017/2 possuiu resultados diferentes dos modelos restantes, em todos os algoritmos utilizados. Como apresentado na Tabela 18 este modelo conseguiu as melhores acurácias para os semestres mais recentes, 2018/2 e 2019/1 com eficácia acima de 80%.

Já os modelos de 2018/2 e 2019/1 têm acurácias similares aos do modelo de 2017/2. A principal diferença entre eles é que os três algoritmos, árvore de decisão, C5.0 e floresta aleatória conseguiram as mesmas acurácias. Esses resultados são detalhados na Tabela 19.

Analisando todos os resultados obtidos pelas previsões do curso de energia é eminente como os modelos compartilham resultados próximos. Em cinco modelos as acurácias obtidas pela árvore de decisão e C5.0 são iguais, havendo diferença somente nas acurácias da floresta aleatória para um semestre em específico. Mesmo quando comparado os outros três semestres, é possível observar um comportamento semelhante.

5.5 Engenharia Eletrônica

O primeiro modelo aplicado foi o 2015/2 que obteve resultados acima de 50% para os semestres até 2017/1. Em contrapartida, para os demais a acurácia ficou abaixo de 20%. Como observado na Tabela 20. O modelo de 2016/1 possuiu como diferença os

Semestre	**Acurácia Árvore de Decisão**	**Acurácia C5.0**	**Acurácia Floresta Aleatória**
2015/2	0,629	0,629	0,629
2016/1	0,595	0,595	0,595
2016/2	0,75	0,75	0,75
2017/1	0,805	0,805	0,805
2017/2	0,032	0,032	0,032
2018/1	0,121	0,121	0,121
2018/2	0,137	0,137	0,137
2019/1	0,071	0,071	0,071

Tabela 20 – Tabela de Resultados dos modelos de Eletrônica 2015/2

Semestre	**Acurácia Árvore de Decisão**	**Acurácia C5.0**	**Acurácia Floresta Aleatória**
2015/2	0,629	0,629	0,629
2016/1	0,595	0,595	0,619
2016/2	0,75	0,75	0,678
2017/1	0,805	0,805	0,805
2017/2	0,032	0,032	0,032
2018/1	0,121	0,121	0,121
2018/2	0,137	0,137	0,137
2019/1	0,071	0,071	0,071

Tabela 21 – Tabela de Resultados dos modelos de Eletrônica 2016/1

Semestre	**Acurácia Árvore de Decisão**	**Acurácia C5.0**	**Acurácia Floresta Aleatória**
2015/2	0,629	0,481	0,481
2016/1	0,595	0,476	0,476
2016/2	0,75	0,857	0,857
2017/1	0,805	0,805	0,805
2017/2	0,032	0	0
2018/1	0,121	0,097	0,097
2018/2	0,137	0,103	0,103
2019/1	0,071	0,035	0,035

Tabela 22 – Tabela de Resultados dos modelos de Eletrônica 2016/2

resultados obtidos com o algoritmo floresta aleatória, suas acurácias são apresentadas na Tabela 21. Enquanto o modelo 2016/2 tem resultados diferentes para floresta aleatória e C5.0, presente na Tabela 22.

O modelo de 2017/1 conseguiu trazer resultados diferentes daqueles obtidos pelos três modelos anteriores. De forma geral, houve uma queda nas acurácias das previsões realizadas com árvore de decisão e C5.0 e um aumento com floresta aleatória. A menor delas sendo para o semestre 2017/2, onde árvore de decisão não conseguiu realizar as previsões. Isso pode ser observado na Tabela **??**.

Semestre	**Acurácia Árvore de Decisão**	**Acurácia C5.0**	**Acurácia Floresta Aleatória**
2015/2	0,370	0,629	0,518
2016/1	0,5	0,595	0,619
2016/2	0,678	0,75	0,571
2017/1	0,833	0,805	0,833
2017/2	0	0,032	0,032
2018/1	0,097	0,121	0,121
2018/2	0,103	0,137	0,137
2019/1	0,035	0,071	0,071

Tabela 23 – Tabela de Resultados dos modelos de Eletrônica 2017/1

Semestre	**Acurácia Árvore de Decisão**	**Acurácia C5.0**	**Acurácia Floresta Aleatória**
2015/2	0,148	0,148	0,148
2016/1	0,238	0,238	0,238
2016/2	0,035	0,035	0,035
2017/1	0,027	0,027	0,027
2017/2	0,967	0,967	0,967
2018/1	0,878	0,878	0,878
2018/2	0,862	0,862	0,862
2019/1	0,928	0,928	0,928

Tabela 24 – Tabela de Resultados dos modelos de Eletrônica 2017/2, 2018/1, 2018/2 e 2019/1

Os modelos 2017/2, 2018/1, 2018/2 e 2019/1 tiveram os mesmos resultados. Para os semestres inicias as acurácias foram menores e somente uma alcançou acurácia maior que 20%, já para os semestres finais as acurácias ficaram acima de 85%. O detalhamento destes resultados pode ser encontrado na Tabela 23.

5.6 Relatórios dos Cursos

Como artefatos de entrega desta monografia foram criados relatórios para os cursos de Engenharia da Universidade de Brasília, Campus Gama. Para a sua elaboração, foram utilizados oito modelos separados por graduação e suas respectivas acurácias. Viabilizando assim a demonstração do uso de indicadores na possível previsão dos alunos com probabilidade de evasão.

Os resultados obtidos a partir, da criação destes relatórios, mostraram que os modelos criados e aplicados não foram capazes de prever alunos que podem evadir, para os cursos de Engenharia Automotiva, Apêndice D, Engenharia de Software, disponível no F e Engenharia de Energia, Apêndice G.

Para os demais cursos os modelos retornaram alunos com o estado de provável

evasão, e portanto, foi possível identificar e detalhar os dados destes estudantes e de suas previsões. O relatório de Engenharia Aeroespacial pode ser encontrado no Apêndice E e o de Engenharia Eletrônica no Apêndice H.

Porém, é importante ressaltar que embora tenham sido gerados dois relatórios com alunos com status evadidos, estes não são resultados de uma previsão. Mas sim, resultados de uma demonstração de uso dos fatores identificados.

5.7 Resultados Gerais

Nas análises de acurácias por curso um comportamento foi repetido ao longo das previsões realizadas. Modelos criados com os semestres 2015/2, 2016/1, 2016/2 e 2017/1 muitas vezes não obtiveram acurácias altas, mesmo com previsões realizadas entre si. É possível observar que, geralmente, esses modelos atingiam acurácias acima de 70% no máximo em dois semestres.

Por outro lado, modelos criados com os semestres 2017/2, 2018/1, 2018/2 e 2019/1, de forma geral, tiveram melhores resultados quando as previsões eram realizadas entre si. Em alguns casos em específico, alguns desses modelos tiveram comportamento similar aos modelos de semestres iniciais.

Diante disso, é importante saber qual o contexto destes cursos ao longo dos semestres avaliados. Segundo informações disponíveis no SIGAA o curso de engenharia automotiva possui duas estruturas curriculares que aconteceram dentro do período sendo utilizado para as previsões. Essa mesma situação se repete para engenharia aeroespacial, engenharia de software e engenharia eletrônica. Já engenharia de energia possui três estruturas curriculares dentro do período sendo avaliado. Essas mudanças de fluxo precisam ser consideradas quando analisando os resultados obtidos.

Além disso, são altas as chances do mesmo grupo de alunos estarem sendo avaliados em cada previsão. Um exemplo claro disso, são estudantes que entraram em um dos cursos em 2015/2 e que até 2019/1 não tenham formado.

Com esses resultados, surge a necessidade de uso de novos fatores que talvez ajudem na acurácia destes modelos e numa possível previsão. Estes fatores podem ser aqueles relacionados ao hábito de uso do Aprender3 pelos estudantes e informações relacionadas pela quantidade trancamentos feitos pelos alunos. Ou seja, fatores relacionados a dados disponíveis no SIGAA.

6 Conclusão

A realização desta monografia proporcionou inicialmente um entendimento sobre como os fatores estão sendo utilizados nas pesquisas realizadas, e como eles estão associados com a previsão da retenção e a evasão acadêmica. Além disso, foi viável a análise e a compressão dos fatores que levam os estudantes à evasão acadêmica, dentro do ambiente do campus FGA da Universidade de Brasília, através da aplicação de modelos de aprendizado de máquina nos alunos das engenharias automotiva, aeroespacial, software, energia e eletrônica.

A realização dessas aplicações, possibilitou a observação de como o contexto de mudança de fluxo pode impactar o desempenho acadêmico. Essas análises proporcionaram ainda, a criação de relatórios com a relação de estudantes e suas respectivas possibilidades de evasão para os cursos de Engenharia Aeroespacial e Engenharia Eletrônica da FGA, obtidos através da demonstração de uso destes fatores.

Por meio da revisão sistemática da literatura realizada foi plausível identificar os fatores que contribuem para a evasão no ensino superior e entender como esses fatores são utilizados na previsão. A partir da criação de um código, com o uso dos modelos árvore de decisão, C5.0 e floresta aleatória, foi possível utilizar os fatores disponíveis nos dados dos alunos das cinco engenharias utilizadas como estudo de caso.

Assim, os três objetivos específicos desta monografia foram atingidos, possibilitando, desse modo, a conclusão da identificação dos fatores que contribuem para a evasão de estudantes de graduação nos cursos de engenharia da Universidade de Brasília, campus Gama.

Como os fatores utilizados nas aplicações realizadas nesta monografia estão ligadas ao fluxo do curso, ao status do aluno e a quantidade de vezes que ele fez uma matéria, acredita-se que exista a necessidade da criação de novos modelos com mais fatores disponíveis pelo SIGAA.

Por fim, é importante ressaltar que o estudo de caso aqui aplicado esteve limitado a dados de alunos pré-pandemia, onde o semestre mais recente considerado foi 2019/1. É possível que os modelos aqui criados tenham resultados diferentes para os semestres mais atuais.

Referências

ALVAREZ, N. L.; CALLEJAS, Z.; GRIOL, D. Factors that affect student desertion in careers in computer engineering profile. *Revista Fuentes*, v. 22, n. 1, p. 105–126, 2020. Cited By :2. Disponível em: <www.scopus.com>. Citado 2 vezes nas páginas 18 e 85.

ATIEH, E.; YORK, D.; MUNÌIZ, M. Beneath the surface: An investigation of general chemistry students' study skills to predict course outcomes. *Journal of Chemical Education*, v. 98, n. 2, p. 281–292, 2021. Disponível em: <www.scopus.com>. Citado 2 vezes nas páginas 77 e 84.

ATIF, A.; RICHARDS, D.; BILGIN, A. A student retention model: Empirical, theoretical and pragmatic considerations. In: *Proceedings of the 24th Australasian Conference on Information Systems*. [s.n.], 2013. Cited By :4. Disponível em: <www.scopus.com>. Citado 2 vezes nas páginas 39 e 84.

BARANYI, M. et al. Modeling students' academic performance using bayesian networks. In: *ICETA 2019 - 17th IEEE International Conference on Emerging eLearning Technologies and Applications, Proceedings*. [s.n.], 2019. p. 42–49. Cited By :1. Disponível em: <www.scopus.com>. Citado 2 vezes nas páginas 39 e 85.

BARGMANN, C.; THIELE, L.; KAUFFELD, S. *Motivation matters: predicting students career decidedness and intention to drop out after the first year in higher education*. 2022. 3845-861 p. Cited By :1. Disponível em: <www.scopus.com>. Citado na página 85.

BECK, H. P.; MILLIGAN, M. Factors influencing the institutional commitment of online students. *Internet and Higher Education*, v. 20, p. 51–56, 2014. Cited By :21. Disponível em: <www.scopus.com>. Citado 2 vezes nas páginas 39 e 85.

BRAHM T., J.-T. W. D. The crucial first year: a longitudinal study of students motivational development at a swiss business school. *Higher Education*, v. 73, p. 459–4780, 2017. Cited By :23. Disponível em: <www.scopus.com>. Citado na página 86.

CAMPBELL C., M.-J. Student perceptions matter: Early signs of undergraduate student retention/attrition. *Journal of College Student Retention: Research, Theory and Practice*, v. 14, n. 4, 2012. Cited By :16. Disponível em: <www.scopus.com>. Citado na página 86.

COMMUNITY, D. *About*. 2022. Disponível em: <https://dbeaver.io/about/>. Citado na página 52.

FELIZARDO, K. et al. A systematic mapping on the use of visual data mining to support the conduct of systematic literature reviews. 12 2020. Disponível em: <https://www.researchgate.net/publication/347534873_A_Systematic_Mapping_on_the_use_of_Visual_Data_Mining_to_Support_the_Conduct_of_Systematic_Literature_Reviewswww.scopus.com>. Citado na página 22.

FERNáNDEZ-MARTíN, T. et al. A multinomial and predictive analysis of factors associated with university dropout. *Revista Electronica Educare*, v. 23, n. 1, 2019. Cited By :3. Disponível em: <www.scopus.com>. Citado 3 vezes nas páginas 18, 39 e 84.

FIGUEROA-CANAS, J.; SANCHO-VINUESA, T. Early prediction of dropout and final exam performance in an online statistics course. *Revista Iberoamericana de Tecnologias del Aprendizaje*, v. 15, n. 2, p. 86–94, 2020. Cited By :6. Disponível em: <www.scopus.com>. Citado 2 vezes nas páginas 40 e 85.

FILIPE, Q.-S. et al. Estudo bibliométrico: Orientações sobre sua aplicação. *Revista Brasileira de Marketing*, v. 15, p. 246–262, 06 2016. Citado na página 23.

FRISCHENSCHLAGER, O.; HAIDINGER, G.; MITTERAUER, L. Factors associated with academic success at vienna medical school: Prospective survey. *Croatian medical journal*, v. 46, n. 1, p. 58–65, 2005. Cited By :39. Disponível em: <www.scopus.com>. Citado 2 vezes nas páginas 39 e 85.

FéNYES, H.; MOHáCSI, M.; PALLAY, K. Career consciousness and commitment to graduation among higher education students in central and eastern europe. *Economics and Sociology*, v. 14, n. 1, p. 61–75, 2021. Cited By :3. Disponível em: <www.scopus.com>. Citado 2 vezes nas páginas 41 e 84.

GIL, A. C. *Como elaborar projetos de pesquisa*. [S.l.]: Atlas São Paulo, 2002. v. 4. Citado 2 vezes nas páginas 42 e 43.

GUSTIAN, D.; HUNDAYANI, R. D. Combination of ahp method with c4.5 in the level classification level out students. In: IEEE (Ed.). *Conference: 2017 International Conference on Computing, Engineering, and Design (ICCED)*. [s.n.], 2017. v. 2017-November. Cited By :9. Disponível em: <www.ieeexplore.ieee.org/>. Citado na página 84.

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*, v. 5, 2020. Cited By :1. Disponível em: <www.scopus.com>. Citado 4 vezes nas páginas 18, 40, 77 e 86.

KISS, B. et al. Predicting dropout using high school and first-semester academic achievement measures. In: *ICETA 2019 - 17th IEEE International Conference on Emerging eLearning Technologies and Applications, Proceedings*. [s.n.], 2019. p. 383–389. Cited By :5. Disponível em: <www.scopus.com>. Citado 3 vezes nas páginas 39, 41 e 86.

KONDO, N.; OKUBO, M.; HATANAKA, T. Early detection of at-risk students using machine learning based on lms log data. In: *Proceedings - 2017 6th IIAI International Congress on Advanced Applied Informatics, IIAI-AAI 2017*. [s.n.], 2017. p. 198–201. Cited By :23. Disponível em: <www.scopus.com>. Citado 4 vezes nas páginas 18, 40, 41 e 85.

KOSTOPOULOS, G. et al. Early dropout prediction in distance higher education using active learning. In: *2017 8th International Conference on Information, Intelligence, Systems and Applications, IISA 2017*. [s.n.], 2018. v. 2018-January, p. 1–6. Cited By :6. Disponível em: <www.scopus.com>. Citado 4 vezes nas páginas 39, 40, 41 e 85.

LóPEZ, C. S.; J., D. S. Analyzing the influence of online behaviors and learning approaches on academic performance in first year engineering. In: PROCEEDINGS, C. W. (Ed.). *2nd Latin American Conference on Learning Analytics*. [s.n.], 2019. v. 2425-2019. Cited By :0. Disponível em: <www.scopus.com/>. Citado na página 84.

MEENS, E. E. M. et al. The association of identity and motivation with students' academic achievement in higher education. *Learning and Individual Differences*, v. 64, p. 54–70, 2018. Cited By :9. Disponível em: <www.scopus.com>. Citado 2 vezes nas páginas 38 e 86.

NIESSEN, A. S. M.; MEIJER, R. R.; TENDEIRO, J. N. Predicting performance in higher education using proximal predictors. *PLoS ONE*, v. 11, n. 4, 2016. Cited By :22. Disponível em: <www.scopus.com>. Citado 5 vezes nas páginas 18, 38, 40, 48 e 86.

ORACLE. *MySQL 8.0 Reference Manual*. 2022. Disponível em: <https://dev.mysql.com/doc/refman/8.0/en/>. Citado na página 51.

ORESHIN, S. et al. Implementing a machine learning approach to predicting students academic outcomes. In: *ACM International Conference Proceeding Series*. [s.n.], 2020. p. 78–83. Cited By :2. Disponível em: <www.scopus.com>. Citado 2 vezes nas páginas 39 e 85.

PRODANOV, C. C.; FREITAS, E. C. de. *Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico-2ª Edição*. [S.l.]: Editora Feevale, 2013. Citado na página 42.

PéREZ, A. et al. Comparative analysis of prediction techniques to determine student dropout: Logistic regression vs decision trees. In: *Proceedings - International Conference of the Chilean Computer Science Society, SCCC*. [s.n.], 2018. v. 2018-November. Cited By :1. Disponível em: <www.scopus.com>. Citado 2 vezes nas páginas 39 e 84.

R., C.; A.M., K. Mathematics as a factor in community college stem performance, persistence, and degree attainment. *Journal of Research in Science Teaching*, v. 57, p. 279–307, 2020. Cited By :6. Disponível em: <www.scopus.com>. Citado na página 85.

REPARAZ C., A.-S. M. M. G. Self-regulation of learning and mooc retention. *Computers in Human Behavior*, v. 111, n. 106423, 2020. Cited By :23. Disponível em: <www.scopus.com>. Citado na página 86.

RESPONDEK, L. et al. Perceived academic control and academic emotions predict undergraduate university student success: Examining effects on dropout intention and achievement. *Frontiers in Psychology*, v. 8, n. MAR, 2017. Cited By :68. Disponível em: <www.scopus.com>. Citado 2 vezes nas páginas 39 e 86.

SAMPAIO, C. E. M. et al. *Resumo Técnico do Censo da Educação Superior 2019*. [S.l.], 2021. Citado 2 vezes nas páginas 18 e 19.

SANI, N. S. et al. Drop-out prediction in higher education among b40 students. *International Journal of Advanced Computer Science and Applications*, v. 11, n. 11, p. 550–559, 2020. Cited By :6. Disponível em: <www.scopus.com>. Citado 2 vezes nas páginas 39 e 84.

SILVA, P. M. D. et al. Ensemble regression models applied to dropout in higher education. In: *Proceedings - 2019 Brazilian Conference on Intelligent Systems, BRACIS 2019*. [s.n.], 2019. p. 120–125. Cited By :2. Disponível em: <www.scopus.com>. Citado 2 vezes nas páginas 38 e 85.

SVIRINA, A.; LOPATIN, A.; TITKO, J. Analysis of students performance in relation to the results of state unified exam: The case of russian university. *Business, Management and Economics Engineering*, v. 19, n. 1, p. 170–179, 2021. Disponível em: <www.scopus.com>. Citado 2 vezes nas páginas 38 e 84.

WAZLAWICK, R. *Metodologia de Pesquisa*. [S.l.]: LTC, 2009. v. 3. Citado na página 26.

Apêndices

APÊNDICE A – Protocolo da Revisão Sistemática

A.1 Introdução

O desempenho estudantil está ligado a diferentes fatores com diferentes impactos na vida discente. Questões financeiras e sociais afetam os índices acadêmicos de um estudante, bem como as demandas universitárias. Prever os efeitos desses no rendimento de um aluno pode garantir que aações sejam criadas para evitar a retenção e evasão estudantil em universidades.

Essa necessidade abriu um leque de oportunidades de estudo na área. Pesquisadores como Atieh, York e MunÌiz (2021) tentam descobrir como os hábitos de estudos do discente pode impactar em seu desempenho em disciplinas. Já autores como Kilian, Loose e Kelava (2020) avaliam se o resultado obtido em matérias introdutórias pode ajudar na previsão de sucesso e evasão estudantil ao longo do curso.

Porém, essas não são as únicas pluralidades encontradas nas pesquisas na área. Existem variações no contexto, mesmo em estudos aplicados no mesmo país ou mesmo curso de graduação. Isso ocorre, pois características como a forma que os fatores são identificados, quais são selecionados para a realização da previsão e como os algoritmos de previsão são aplicados variam bastante entre as pesquisas; e entender as condições em comum entre as diferentes abordagens é importante para reconhecer o comportamento das pesquisas atuais.

A partir do conhecimento sobre esses procedimentos e hábitos será possível realizar uma revisão sistemática que entenda quais são os fatores que contribuem para a retenção e evasão de estudantes de graduação.

A.2 Processo de condução do RSL

A.2.1 Planejamento

A.2.1.1 Objetivo

Esta revisão sistemática da literatura busca identificar fatores contribuem para a retenção ou evasão de estudantes de graduação.

A.2.1.2 Protocolo

A.2.1.2.1 Definir as questões de pesquisa

A questão principal desse protocolo é "Quais fatores podem ser identificáveis quando há associação da previsão de retenção e evasão no ensino superior?". Além dessa, existem três perguntas secundárias definidas de forma a auxiliar o entendimento de como esses fatores se comportam nas pesquisas.

- Questão secundária 01: Quais são os fatores usados para prever a retenção ensino superior?
- Questão secundária 02: Quais são os fatores usados para prever a evasão no ensino superior?
- Questão secundária 03: Como os fatores são usados no processo de previsão?

A.2.1.2.2 Definir as fontes de pesquisa

As fontes de pesquisas que serão utilizadas são as bases de dados IEEE Xplore[1] e Scopus[2].

A.2.1.2.3 String de Busca

Petticrew e Roberts (2008) propõem a abordagem P.I.C.O.C para apoiar o desenvolvimento da expressão de busca. Essa corresponde à definição dos cinco elementos:

- (P) - População: refere-se ao que é afetado pela intervenção (um papel, uma categoria da engenharia de software, uma área de aplicação ou um determinado grupo etc);
- (I) - Intervenção: são os tipos de tratamento relacionados à população (metodologias, ferramentas, tecnologias, procedimentos etc);
- (C) - Comparação: refere-se às comparações feitas entre os tipos de intervenção em que há aspectos distintos, porém relacionados e são avaliados comparativamente. Não há intenção de se fazer um estudo comparativo nesta RSL, mas sim identificar o máximo possível de indicadores associados à retenção e evasão;
- (O) - Outcomes (resultados): são os aspectos de interesse que estão associados à população e à intervenção;

[1] Pode ser acessado em: https://www.ieee.org/
[2] Pode ser acessado em: https://elsevier.com/

- (C) - Contexto: contexto da pesquisa.

Termo	Palavra-chave	Termo correlato / sinônimo
População	Undergraduate	
Intervenção	Academic Dropout prediction	
	Retention prediction	
Comparação	—	—
Resultados	measurement	metric
	indicator	
Contexto	Higher Education	
	Bachelor Degree	

Tabela 25 – Tabela de Decrição do PICOC

A partir dos resultados obtidos na tabela acima, tem-se: (P) and (I) and (O) and (C), em que a string de busca original foi criada antes da aplicação da abordagem P.I.C.O.C. e foi utilizada para a leitura inicial sobre o tema. Foi a partir dessa que as palavras-chave importantes para este protocolo foram selecionadas.

TITLE-ABS-KEY (("learning analytics"OR "data mining") AND (dropout OR retention) AND ("higher education"OR undergraduate OR "bachelor degree"))

Com os resultados obtidos por meio da leitura dos artigos resultados da string original, foi possível realizar o primeiro refinamento na string de busca, em que a abordagem P.I.C.O.C foi aplicada.

TITLE-ABS-KEY ((undergraduate) AND ("academic dropout"OR retention) AND (measurement OR metric OR indicator) AND ("higher education"OR "bachelor degree")) <=> 34 resultados

O segundo refinamento da string de busca foi feito para adicionar as palavras-chave student, predict* e risk.

TITLE-ABS-KEY ((student OR undergraduate) AND (predict* AND (dropout OR retention)) AND (measurement OR indicator OR risk) AND ("higher education"OR "bachelor degree")) <=> 154 resultados

A partir dos resultados obtidos, foi identificado a necessidade de adicionar a palavra-chave metric* e mudar indicator para indicat* e risk para risk*.

TITLE-ABS-KEY ((student OR undergraduate) AND (predict* AND (dropout OR retention)) AND (metric* OR measurement OR indicat* OR risk*) AND ("higher education"OR "bachelor degree")) <=> 193 resultados

No quarto refinamento foi observado a necessidade de adicionar a palavra-chave attrition.

TITLE-ABS-KEY ((student OR undergraduate) AND (predict* AND (dropout OR retention OR attrition)) AND (metric* OR measurement OR indicat* OR risk*) AND ("higher education"OR "bachelor degree")) <=> 211 resultados

No quinto refinamento foi adicionado a restrição para tipos de documento. Para este protocolo de pesquisa é importante artigos e papéis de conferência.

TITLE-ABS-KEY ((student OR undergraduate) AND (predict* AND (dropout OR retention OR attrition)) AND (metric* OR measurement OR indicat* OR risk*) AND ("higher education"OR "bachelor degree")) AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cp")) <=> 203 resultados

No sexto e último refinamento foi retirada a palavra-chave risk e limitações de idiomas para inglês, espanhol e português.

TITLE-ABS-KEY ((student OR undergraduate) AND (predict* AND (dropout OR retention OR attrition)) AND (metric* OR measurement OR indicat*) AND ("higher education"OR "bachelor degree")) AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cp")) <=>112 resultados

A.2.1.2.4 Filtros de pesquisa

- Leitura dos resumos;
- Leitura do texto completo.

A.2.1.2.5 Critérios de inclusão e exclusão

Dos critérios de aceitação tem-se:

- O artigo apresenta fatores relacionados à evasão;
- O artigo apresenta fatores relacionados à retenção;
- O artigo apresenta fatores relacionados à evasão e retenção.

Dos critérios de exclusão tem-se:

- Estudos duplicados;
- Versões mais antigas de um mesmo estudo;
- Sem acesso ao texto completo do artigo;
- Publicações escritas em outros idiomas que não inglês, português e espanhol;
- Estudos secundários ou terciários.

A.2.1.2.6 Definir critérios de qualidade

Os critérios de qualidade foram definidos em forma de perguntas, de forma a classificar os artigos a partir das respostas obtidas. Para cada um dos itens é necessário indicar se o artigo avaliado responde à questão totalmente, parcialmente ou não responde. A lista de tópicos usados para essa análise são:

- Os objetivos, as perguntas de pesquisa e hipóteses (se aplicável) são claros e relevantes?
- Existe uma descrição do contexto na qual a pesquisa foi conduzida?
- Os dados foram coletados de forma a atender às perguntas de pesquisa?
- Existe uma declaração clara dos resultados?
- Os autores descrevem as limitações dos estudos?
- As conclusões, implicações para a prática e pesquisas futuras são adequadamente relatadas ao seu público?
- As descobertas foram claramente relatadas?
- As questões éticas são devidamente abordadas (intenções pessoais, integridade, confidencialidade, consentimento, aprovação do conselho de revisão)?

A.2.2 Condução

A.2.2.1 Extração dos Dados

A.2.2.1.1 Procedimento de Extração

O protocolo foi conduzido na ferramenta Parsifal[1] durante todas as etapas de condução do protocolo. Desde a inclusão dos artigos até a extração dos dados.

A.2.2.1.2 Filtros de Coleta

- Informação de citação: Autor(es), Autor(es) ID, Título, Ano, EID, Título da Fonte, Volume, Questões, Páginas, Contagem de citações, Fonte, Tipo de documento, Estágio de Publicação, DOI e Acesso livre.
- Informação bibliográfica: Afiliações, Identificadores de série (por exemplo, ISSN), PubMed ID, Publicador, Editor(es), Idioma do documento original, Endereço para correspondência e Título da Fonte abreviado.

[1] Pode ser acessado em: https://parsif.al/

- Resumo e Palavras-chave: Resumo, Palavras-chave do Autor e Palavras-chave do Índice.
- Detalhes de financiamento: Número, Acrônimo, Patrocinador e Texto de financiamento.
- Outras informações: Nomes comerciais e fabricantes, Números de adesão e produtos químicos, Informação da conferência e Inclusão de referências.

A.2.2.2 Síntese dos Dados

- Critérios de inclusão
- Pesquisador responsável pela seleção do artigo
- Resumo da publicação nas palavras do pesquisador responsável pela inclusão
- Classificação por facetas de pesquisa (Pesquisa de validação; Pesquisa de avaliação; Proposta de solução; Artigo filosófico; Relato de experiência; Artigo de opinião)
- Classificação por métodos de pesquisa (survey; estudo de caso; experimento controlado; pesquisa-ação; etnografia; simulação; prototipagem; análise matemática; e prova de propriedade)
- País em que ocorreu o estudo
- Curso(s) de graduação
- Disciplina - campo aberto
- Nome do indicador (fator) - dropdown (concatenar categoria e nome do indicador. Exemplo - acadêmico: reprovação)
- Outras informações relevantes para o pesquisador - campo aberto
- Ferramentas de previsão (fazer uma lista - dropdown) Exemplos: phyton; SPSS; R
- Algoritmos usados na previsão (fazer uma lista - dropdown) Exemplo: C5
- Descrever como o algoritmo foi usado (subsets, distribuição dos dados, indicadores utilizados etc)
- Indicadores de efetividade do algoritmo - campo aberto em que se espera respostas para: acurácia; cobertura; precision; recall; F1 score; e/ou outros relatados nos artigos
- Outras informações relevantes para o pesquisador

A.2.3 Publicação dos Resultados

A.2.3.1 Estratégia de publicação

- Artigo científico
- Dissertação de mestrado
- Trabalho de conclusão de curso
- Relatório para BCE

APÊNDICE B – Lista de artigos retornados pela string de busca

Nome do Artigo	Código	Artigo
A multinomial and predictive analysis of factors associated with university Dropout [Un análisis multinomial y predictivo de los factores asociados a la deserción universitaria]	A-1	(FERNáNDEZ-MARTíN et al., 2019)
A student retention model: Empirical, theoretical and pragmatic considerations	A-2	(ATIF; RICHARDS; BILGIN, 2013)
Analysis of students performance in relation to the results of state unified exam: The case of russian university	A-3	(SVIRINA; LOPATIN; TITKO, 2021)
Analyzing the influence of online behaviors and learning approaches on academic performance in first year engineering	A-4	(LóPEZ; J., 2019)
Beneath the Surface: An Investigation of General Chemistry Students' Study Skills to Predict Course Outcomes	A-5	(ATIEH; YORK; MUNÌIZ, 2021)
Career consciousness and commitment to graduation among higher education students in Central and Eastern Europe	A-6	(FéNYES; MOHáCSI; PALLAY, 2021)
Combination of AHP Method with C4.5 in the level classification level out students	A-7	(GUSTIAN; HUNDAYANI, 2017)
Comparative Analysis of Prediction Techniques to Determine Student Dropout: Logistic Regression vs Decision Trees	A-8	(PéREZ et al., 2018)
Drop-Out Prediction in Higher Education Among B40 Students	A-9	(SANI et al., 2020)

Early Detection of At-Risk Students Using Machine Learning Based on LMS Log Data	A-10	(KONDO; OKUBO; HATANAKA, 2017)
Early dropout prediction in distance higher education using active learning	A-11	(KOSTOPOULOS et al., 2018)
Early Prediction of Dropout and Final Exam Performance in an Online Statistics Course	A-12	(FIGUEROA-CANAS; SANCHO-VINUESA, 2020)
Ensemble Regression Models Applied to Dropout in Higher Education	A-13	(SILVA et al., 2019)
Factors associated with academic success at Vienna Medical School: Prospective survey	A-14	(FRISCHENSCHLAGER; HAIDINGER; MITTERAUER, 2005)
Factors influencing the institutional commitment of online students	A-15	(BECK; MILLIGAN, 2014)
Factors that affect student desertion in careers in Computer Engineering profile [Factores que inciden en la deserción estudiantil en carreras de perfil Ingeniería informática]	A-16	(ALVAREZ; CALLEJAS; GRIOL, 2020)
Implementing a Machine Learning Approach to Predicting Students Academic Outcomes	A-17	(ORESHIN et al., 2020)
Mathematics as a factor in community college STEM performance, persistence, and degree attainment	A-18	(R.; A.M., 2020)
Modeling Students Academic Performance Using Bayesian Networks	A-19	(BARANYI et al., 2019)
Motivation matters: predicting students career decidedness and intention to drop out after the first year in higher education	A-20	(BARGMANN; THIELE; KAUFFELD, 2022)

Perceived academic control and academic emotions predict undergraduate university student success: Examining effects on dropout intention and achievement	A-21	(RESPONDEK et al., 2017)
Predicting Dropout Using High School and First-semester Academic Achievement Measures	A-22	(KISS et al., 2019)
Predicting Math Student Success in the Initial Phase of College With Sparse Information Using Approaches From Statistical Learning	A-23	(KILIAN; LOOSE; KELAVA, 2020)
Predicting performance in higher education using proximal predictors	A-24	(NIESSEN; MEIJER; TENDEIRO, 2016)
Self-regulation of learning and MOOC retention	A-25	(REPARAZ C., 2020)
Student perceptions matter: Early signs of undergraduate student retention/attrition	A-26	(CAMPBELL C., 2012)
The association of identity and motivation with students' academic achievement in higher education	A-27	(MEENS et al., 2018)
The crucial first year: a longitudinal study of students motivational development at a Swiss Business School	A-28	(BRAHM T., 2017)
	N/A.	N/A.
Considering the role of the distance student experience in student satisfaction and retention	A-32	N/A.
Assessing the Validity of College Success Indicators for the At-Risk Student: Toward Developing a Best-Practice Model	A-29	N/A.
Leaks in Latina/o Studentsâ College-Going Pipeline: Consequences of Educational Expectation Attrition	N/A.	N/A.

Applying stress theory to higher education: lessons from a study of first-year students	N/A.	N/A.
MOOC dropout prediction using machine learning techniques: Review and research challenges	N/A.	N/A.
Planning and performance measurement in higher education: three case studies of operational research application [PlaneaciÃşn y mediciÃşn del desempeÃśo en educaciÃşn superior: tres casos de aplicaciÃşn de investigaciÃşn de operaciones]	N/A.	N/A.
Occupational self-efficacy and psychological capital amongst nursing students: A cross sectional study understanding the malleable attributes for success	N/A.	N/A.
Predicting students' college drop out and departure decisions by analyzing their campus-based social network text messages	N/A.	N/A.
The Relationship between Sense of Belonging and Student Outcomes in CS1 and beyond	N/A.	N/A.
Academic and diversity consequences of affirmative action in Brazil	N/A.	N/A.
Predicting college success with high school grades and test scores: Limitations for minority students	A-47	N/A.
Student dropout analysis with application of data mining methods [Analiza odustajanja od studija primjenom metoda rudarenja podataka]	N/A.	N/A.
Meta-analysis of the relationship between the Big Five and academic success at university	N/A.	N/A.

Are students really connected? Predicting college adjustment from social network usage	N/A.	N/A.
Technological barriers and incentives to learning analytics adoption in higher education: insights from users	A-56	N/A.
Predicting University Studentsâ Academic Success and Major Using Random Forests	N/A.	N/A.
Motivation in transition: Development and roles of expectancy, task values, and costs in early college engineering	N/A.	N/A.
Interpretable Multiview Early Warning System Adapted to Underrepresented Student Populations	A-41	N/A.
Effects of emotional intelligence and supportive text messages on academic outcomes in first-year undergraduates	A-36	N/A.
Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature	N/A.	N/A.
Transitioning from VET to HE in hospitality and tourism studies: VET grades as an indicator of performance in HE	N/A.	N/A.
Using academic analytics to predict dropout risk in engineering courses	N/A.	N/A.
The predictive role of gender and race on student retention	N/A.	N/A.
Embedding machine learning algorithm models in decision support system in predicting student academic performance using enrollment and admission data	N/A.	N/A.
Predicting student academic performance using multi-model heterogeneous ensemble approach	A-49	N/A.

Factors affecting the programme completion of pre-registration nursing students through a three year course: A retrospective cohort study	N/A.	N/A.
Customer service, university student segmentation and institutional commitment	N/A.	N/A.
5th European MOOCs Stakeholders Summit, EMOOCs 2017	N/A.	N/A.
What drives student success? Assessing the combined effect of transfer students and online courses	N/A.	N/A.
Exploring student characteristics of retention that lead to graduation in higher education using data mining models	A-38	N/A.
Validating the effectiveness of the moodle engagement analytics plugin to predict student academic performance	A-58	N/A.
Predicting educational success and attrition in problem-based learning: do first impressions count?	N/A.	N/A.
Finding predictors in higher education	A-39	N/A.
Changes in student motivation during online learning	N/A.	N/A.
Predicting persistence of urban commuter campus students utilizing student background characteristics from enrollment data	A-48	N/A.
Leaks in Latina/o Students College-Going Pipeline: Consequences of Educational Expectation Attrition	N/A.	N/A.
Predicting student enrolments and attrition patterns in higher educational	N/A.	N/A.
Predicting University Students Academic Success and Major Using Random Forests	N/A.	N/A.

Holistic factors related to student persistence at a large, public university	N/A.	N/A.
Redefining profit metrics for boosting student retention in higher education	N/A.	N/A.
Psychological resources, dropout risk and academic performance in university students pattern-oriented analysis and prospective study of Hungarian freshmen	N/A.	N/A.
Mathematics: A powerful pre- and post-admission variable to predict success in Engineering programmes at a University of Technology	N/A.	N/A.
Student selection for tertiary nursing courses: efficacy of the Anderson score as a performance predictor.	N/A.	N/A.
University freshman retention in North Carolina	A-57	N/A.
Factors affecting student progression and achievement: Prediction and intervention. A two-year study	N/A.	N/A.
Approaches to learning in science: A longitudinal study	N/A.	N/A.
Longitudinal study of drop-out and continuing students who attended the pre-university summer school at the university of Glasgow	N/A.	N/A.
The assessment of non-acade mic and academicservice quality in higher education [Yüksek Ögrenimde akademik ve akademik olmayan hizmet kalitesininin degerlendirilmesi]	N/A.	N/A.
Modeling students' academic performance using Bayesian networks	N/A.	N/A.

Some considerations related to the impact of undernutrition on brain development, intelligence and scholastic achievement [Algunas consideraciones sobre el impacto de la desnutricion en el desarrollo cerebral, inteligencia y rendimiento escolar]	N/A.	N/A.
Determinants of business student satisfaction and retention in higher education: Applying Herzberg's two-factor theory	N/A.	N/A.
How to retain students in higher engineering education? Findings of the ATTRACT project	A-40	N/A.
Modeling student success of international undergraduate engineers	A-45	N/A.
European nursing students' academic success or failure: A post-Bologna Declaration systematic review	N/A.	N/A.
Reducing the need for postsecondary remediation using self-efficacy to identify underprepared african-american and hispanic adolescents	N/A.	N/A.
Engagement vs performance: Using electronic portfolios to predict first semester engineering student retention	A-37	N/A.
Great Expectations: Examining the Discrepancy between Expectations and Experiences on College Student Retention	N/A.	N/A.
Early prediction of students' grade point averages at graduation: A data mining approach [Ö?rencinin mezuniyet notunun erken tahmini: Bir veri madencili?i yakla?idotlessmidotless]	A-34	N/A.
Can online student performance be forecasted by learning analytics?	N/A.	N/A.

Importance of physical health and health-behaviors in adolescence for risk of dropout from secondary education in young adulthood: An 8-year prospective study	N/A.	N/A.
Toward a new predictive model of student retention in higher education: An application of classical sociological theory	N/A.	N/A.
Tertiary education and its association with mental health indicators and educational factors among Arctic young adults: The NAAHS cohort study	N/A.	N/A.
Assessing the link between stress and retention and the existence of barriers to support service use within HE	N/A.	N/A.
Learning analytics in higher education: Assessing learning outcomes	A-42	N/A.
Development of an early alert system to predict students at risk of failing based on their early course activities	A-33	N/A.
Early predictors of study success in a Dutch advanced nurse practitioner education program: A retrospective cohort study	A-35	N/A.
Prediction model of first-year student desertion at Universidad Bernardo O'Higgins (UBO)	A-52	N/A.
Early identification of at-risk students using iterative logistic regression	N/A.	N/A.
Detection of desertion patterns in university students using data mining techniques: A case study	N/A.	N/A.
Modeling of student academic achievement in engineering education using cognitive and non-cognitive factors	A-44	N/A.

Bachelor completion and dropout rates of selected, rejected and lottery-admitted medical students in the Netherlands	A-30.	N/A.
Predictive modelling of student dropout using ensemble classifier method in higher education	A-53	N/A.
Do I think I'm an engineer? Understanding the impact of engineering identity on retention	N/A.	N/A.
Measuring, Manipulating, and Predicting Student Success: A 10-Year Assessment of Carnegie R1 Doctoral Universities Between 2004 and 2013	A-43	N/A.
Commitment in College Student Persistence	N/A.	N/A.
Persistence and engagement among first-year Hispanic students	A-46	N/A.
Looking for a dropout predictor based on the instructional design of online courses	N/A.	N/A.
Game-based student response system: The effectiveness of Kahoot! On junior and senior information science students' learning	N/A.	N/A.
Predictors of performance in business administration degrees: The effect of the high-school specialty [Predictores del rendimiento académico en las titulaciones de administración y dirección de empresas: El efecto de la especialidad en bachillerato]	A-54	N/A.
Using Big Data to Determine Potential Dropouts in Higher Education	N/A.	N/A.
Examining critical success factors augmenting quality of higher education institutes in India. A SEM$_PLSapproach$	N/A.	N/A.

Predicting economics student retention in higher education: The effects of students' economic competencies at the end of upper secondary school on their intention to leave their studies in economics	N/A.	N/A.
Predictors of dropout intentions in teacher education programmes compared with other study programmes	N/A.	N/A.
Review of Undergraduate Student Retention and Graduation Since 2010: Patterns, Predictions, and Recommendations for 2020	N/A.	N/A.
Predicting student degree completion using random forest	A-50	N/A.
Predicting success in an undergraduate exercise science program using science-based admission courses	A-51	N/A.
Gender gaps in the performance of Norwegian biology students: the roles of test anxiety and science confidence	N/A.	N/A.
PThe fit between dignity self-construal and independent university norms: Effects on university belonging, well-being, and academic success	N/A.	N/A.
Comparison of Predictive Models with Balanced Classes for the Forecast of Student Dropout in Higher Education	N/A.	N/A.
Identifying students at risk to academic dropout in higher education	N/A.	N/A.
Online teaching, student success, and retention in political science courses	N/A.	N/A.
Beyond grade point average and standardized testing: Incorporating a socio-economic factor in admissions to support minority success	A-31.	N/A.

Statistical alternatives for studying college student retention: A comparative analysis of logit, probit, and linear regression	A-55	N/A.

Tabela 27 – Tabela de Artigos

APÊNDICE C – Código-fonte

```
library("DBI")
library(RMySQL)
library(rpart)
library(C50)
library(randomForest)

drv <- dbDriver("MySQL")
mydb = dbConnect(
    drv,
    user='root',
    password='password',
    dbname='student_data',
    host='localhost'
)

cursos <- c(
    "AUTOMOTIVA",
    "AEROESPACIAL",
    "SOFTWARE",
    "ENERGIA",
    "ELETRONICA"
)

semestres <- c(
    "2015/2",
    "2016/1",
    "2016/2",
    "2017/1",
    "2017/2",
    "2018/1",
    "2018/2",
    "2019/1"
)

resultados <- c()
```

```
for (curso in cursos) {
  for (semestre in semestres) {

    dataset <- dbGetQuery(
        mydb,
        paste0(
            "select Disciplina
            from student_data.fluxos
            where Abrev ='", curso, "'"
        )
    )
    listaDisciplinas <- unique(dataset[, "Disciplina"])

    dataset <- dbGetQuery(
        mydb,
        paste0(
            "SELECT IDFluxo,
                    Status,
                    Matriculas,
                    Ingresso,
                    IDDisciplina,
                    IDEstudante
            FROM student_data.dados as dados
            INNER JOIN student_data.fluxos as fluxos
            ON dados.IDFluxo = fluxos.Fluxo
            where Abrev = '", curso,"'
            and Ingresso = '", semestre,"' ;"
        )
    )
    listaEstudantes <- unique(dataset[, "IDEstudante"])

    dados <- data.frame(
        matrix(
            0,
            nrow = length(listaEstudantes),
            ncol = length(listaDisciplinas)
        )
    )
```

```
rownames(dados) <- listaEstudantes
colnames(dados) <- listaDisciplinas

for (l in 1:nrow(dataset)) {
  i <- as.character(dataset[l, "IDEstudante"])
  j <- listaDisciplinas

  Tentativas <- as.integer(dataset[l, "Matriculas"])
  dados[i, j] <- Tentativas
}

Status <- c()
listaStatus <- unique(dataset[ , c("IDEstudante", "Status")])

for (l in 1:nrow(listaStatus)) {
  Status <- c(Status, listaStatus[l, "Status"])
}

dados <- cbind(Status, dados)
dados[is.na(dados)] <- 0
dados$Status <- factor(dados$Status)
levels(dados$Status) <- c("Formado", "Matriculado", "Evadido")

modeloArvoreDecisao <- rpart(Status ~ ., dados)

modeloC50 <- C5.0(Status ~ ., data = dados, trials = 10)

quantidadeVariaveis <- length(listaDisciplinas)
numero <- sqrt(quantidadeVariaveis)
max <- ceiling(numero)

dados$Status <- factor(dados$Status)
modeloRandomForest <- randomForest(
    Status ~ .,
    data = dados,
    ntree = 500,
    mtry = max,
    importance = TRUE
)
```

```
modelo <- paste0(curso, " ", semestre)
print(modelo)
for (ingresso in semestres) {

  dataset <- dbGetQuery(
    mydb,
    paste0(
        "SELECT IDFluxo,
                Status,
                Matriculas,
                Ingresso,
                IDDisciplina,
                IDEstudante
        FROM student_data.dados as dados
        INNER JOIN student_data.fluxos as fluxos
        ON dados.IDFluxo = fluxos.Fluxo
        where Abrev = '", curso,"'
        and Ingresso = '", ingresso,"' ;"
    )
  )
  listaEstudantes <- unique(dataset[, "IDEstudante"])

  dadosPrevisao <- data.frame(
    matrix(
        0,
        nrow = length(listaEstudantes),
        ncol = length(listaDisciplinas)
    )
)
  rownames(dadosPrevisao) <- listaEstudantes
  colnames(dadosPrevisao) <- listaDisciplinas

  for (l in 1:nrow(dataset)) {
    i <- as.character(dataset[l, "IDEstudante"])
    j <- listaDisciplinas

    Tentativas <- as.integer(dataset[l, "Matriculas"])
    dadosPrevisao[i, j] <- Tentativas
```

```
}

Status <- c()
statusPrevisao <- unique(dataset[ , c("IDEstudante", "Status")])

for (l in 1:nrow(statusPrevisao)) {
  Status <- c(Status, statusPrevisao[l, "Status"])
}

dadosPrevisao <- cbind(Status, dadosPrevisao)
dadosPrevisao[is.na(dadosPrevisao)] <- 0
dadosPrevisao$Status <- factor(dadosPrevisao$Status)
levels(dadosPrevisao$Status) <- c("Formado", "Matriculado", "Evadido")

previsaoAD <- predict(
  modeloArvoreDecisao,
  type = "class",
  newdata = dadosPrevisao
)
matrizConfArvore <- table(dadosPrevisao$Status, previsaoAD)
accuracyAD <- sum(diag(matrizConfArvore))/sum(matrizConfArvore)

previsaoC50 <- predict(
  modeloC50,
  type = "class",
  newdata = dadosPrevisao
)
matrizConfC50 <- table(dadosPrevisao$Status, previsaoC50)
accuracyC50 <- sum(diag(matrizConfC50))/sum(matrizConfC50)

previsaoRF <- predict(
  modeloRandomForest,
  type = "class",
  newdata = dadosPrevisao
)
matrizConfRF <- table(dadosPrevisao$Status, previsaoRF)
accuracyRF <- sum(diag(matrizConfRF))/sum(matrizConfRF)

resultados <- rbind(
```

```
        resultados,
        c(
            modelo,
            ingresso,
            accuracyAD,
            accuracyC50,
            accuracyRF
        )
      )
    }
  }
}

colnames(resultados) <- c(
    "Modelo",
    "Semestre",
    "Acurácia Árvore de Decisão",
    "Acurácia C5.0",
    "Acurácia Random Forest"
)

write.csv2(resultados, "resultados.csv")

#Desconectando do banco

all_cons <- dbListConnections(MySQL())

for(con in all_cons)
  +  dbDisconnect(con)
```

APÊNDICE D – Relatório de Evasão - Engenharia Automotiva

Para o curso de Engenharia Automotiva, os modelos não conseguiram prever evasão acadêmica. Impossibilitando assim a criação de uma relação de alunos com probabilidade de evasão.

Esse resultado pode indicar que melhoras nos modelos de previsões do algoritmo C5.0 podem ser aplicadas para uma melhor previsão. Esta melhoria pode indicar, ainda, possibilidades de futuros trabalhos.

APÊNDICE E – Relatório de Evasão - Engenharia Aeroespacial

Os dados, apresentados no relatório da relação de alunos com possibilidade de evasão do curso de Engenharia Aeroespacial, são resultados de previsões realizadas com o algoritmo C5.0. Estas previsões foram aplicadas nos semestres 2015/2, 2016/1, 2016/2, 2017/1. 2017/2, 2018/1, 2018/2 e 2019/1.

A probabilidade de evasão foi calculada a partir da quantidade de vezes que um estudante foi qualificado como evasor, por cada um dos oito modelos empregados. Assim, o percentual foi calculado da seguinte forma, o estudante que evadisse nos oito modelos teria percentual de 100% e os outros foram calculados com base nessa métrica.

Estudante	**Modelos Aplicados**	**Quantidade: Possível Evasor**	**Probabilidade de Evasão**
24597	8	1	12,5%
24600	8	1	12,5%
24637	8	1	12,5%
24674	8	1	12,5%
24679	8	1	12,5%
24680	8	1	12,5%
24149	8	1	12,5%
24580	8	1	12,5%
24599	8	1	12,5%
24645	8	1	12,5%
24675	8	1	12,5%
24677	8	1	12,5%
24681	8	1	12,5%
24581	8	1	12,5%
24596	8	1	12,5%
24598	8	1	12,5%
24670	8	1	12,5%
24671	8	1	12,5%
24672	8	1	12,5%
24676	8	1	12,5%
24678	8	1	12,5%
24682	8	1	12,5%

Tabela 28 – Relação de alunos de Engenharia Aeroespacial com chances de evasão em 2015/2

Estudante	**Modelos Aplicados**	**Quantidade: Possível Evasor**	**Probabilidade de Evasão**
18999	3	1	33,33%
19401	3	1	33,33%
18989	3	1	33,33%
18990	3	1	33,33%
18991	3	1	33,33%
18992	3	1	33,33%
18995	3	1	33,33%
18996	3	1	33,33%
18997	3	1	33,33%
18998	3	1	33,33%
19022	3	1	33,33%
19023	3	1	33,33%
19024	3	1	33,33%
19025	3	1	33,33%
19028	3	1	33,33%
19029	3	1	33,33%
19039	3	1	33,33%
19040	3	1	33,33%
19042	3	1	33,33%
19043	3	1	33,33%
19044	3	1	33,33%
19045	3	1	33,33%
18988	3	1	33,33%
18993	3	1	33,33%
18994	3	1	33,33%
19021	3	1	33,33%
19026	3	1	33,33%
19027	3	1	33,33%

Tabela 29 – Relação de alunos de Engenharia Aeroespacial com chances de evasão em 2016/2

Estudante	**Modelos Aplicados**	**Quantidade: Possível Evasor**	**Probabilidade de Evasão**
14162	3	1	33,33%
14118	3	1	33,33%
14119	3	1	33,33%
14120	3	1	33,33%
14121	3	1	33,33%
14123	3	1	33,33%
14124	3	1	33,33%
14126	3	1	33,33%
14127	3	1	33,33%
14128	3	1	33,33%
14129	3	1	33,33%
14131	3	1	33,33%
14132	3	1	33,33%
14133	3	1	33,33%
14134	3	1	33,33%
14137	3	1	33,33%
14138	3	1	33,33%
14140	3	1	33,33%
14141	3	1	33,33%
14142	3	1	33,33%
14044	3	1	33,33%
14045	3	1	33,33%
14046	3	1	33,33%
14149	3	1	33,33%
14153	3	1	33,33%
14158	3	1	33,33%
14161	3	1	33,33%
14163	3	1	33,33%

Tabela 30 – Relação de alunos de Engenharia Aeroespacial com chances de evasão em 2017/1

APÊNDICE F – Relatório de Evasão - Engenharia de Software

Para o curso de Engenharia de Software, os modelos não conseguiram prever a evasão acadêmica. Impossibilitando, assim, a criação de uma relação de alunos com probabilidade de evasão.

Esse resultado pode indicar que melhorias, nos modelos de previsões do algoritmo C5.0, que podem ser aplicadas para uma previsão superior. Esses aperfeiçoamento podem indicar, ainda, possibilidades de trabalhos futuros.

APÊNDICE G – Relatório de Evasão - Engenharia de Energia

Para o curso de Engenharia de Energia, os modelos não conseguiram prever a evasão acadêmica. Impossibilitando, assim, a criação de uma relação de alunos com probabilidade de evasão.

Esse resultado pode indicar que melhorias, nos modelos de previsões do algoritmo C5.0, que podem ser aplicadas para uma previsão superior. Esses aperfeiçoamento podem indicar, ainda, possibilidades de trabalhos futuros.

APÊNDICE H – Relatório de Evasão - Engenharia Eletrônica

Os dados, apresentados no relatório da relação de alunos com possibilidade de evasão do curso de Engenharia Eletrônica, são resultados de previsões realizadas com o algoritmo C5.0. Estas previsões foram aplicadas nos semestres 2015/2, 2016/1, 2016/2, 2017/1. 2017/2, 2018/1, 2018/2 e 2019/1.

A probabilidade de evasão foi calculada a partir da quantidade de vezes que um estudante foi qualificado como evasor, por cada um dos oito modelos empregados. Assim, o percentual foi calculado da seguinte forma, o estudante que evadisse nos oito modelos teria percentual de 100% e os outros foram calculados com base nessa métrica.

Estudante	**Modelos Aplicados**	**Quantidade: Possível Evasor**	**Probabilidade de Evasão**
25882	8	1	12,5%
25930	8	1	12,5%
25936	8	1	12,5%
25942	8	1	12,5%

Tabela 31 – Relação de alunos de Engenharia Eletrônica com chances de evasão em 2015/2

Estudante	**Modelos Aplicados**	**Quantidade: Possível Evasor**	**Probabilidade de Evasão**
18179	8	1	12,5%
18185	8	1	12,5%
18194	8	1	12,5%
18198	8	1	12,5%
18202	8	1	12,5%

Tabela 32 – Relação de alunos de Engenharia Eletrônica com chances de evasão em 2016/1

Estudante	**Modelos Aplicados**	**Quantidade: Possível Evasor**	**Probabilidade de Evasão**
18182	8	1	12,5%
18224	8	1	12,5%
18245	8	1	12,5%
18246	8	1	12,5%

Tabela 33 – Relação de alunos de Engenharia Eletrônica com chances de evasão em 2016/2

Estudante	**Modelos Aplicados**	**Quantidade: Possível Evasor**	**Probabilidade de Evasão**
14270	8	1	12,5%
14284	8	1	12,5%
14282	8	1	12,5%
14281	8	1	12,5%

Tabela 34 – Relação de alunos de Engenharia Eletrônica com chances de evasão em 2017/1

Estudante	**Modelos Aplicados**	**Quantidade: Possível Evasor**	**Probabilidade de Evasão**
14299	8	1	12,5%
14300	8	1	12,5%
14307	8	1	12,5%
14315	8	1	12,5%
14322	8	1	12,5%
14323	8	1	12,5%
14324	8	1	12,5%
14297	8	1	12,5%

Tabela 35 – Relação de alunos de Engenharia Eletrônica com chances de evasão em 2017/2

Estudante	**Modelos Aplicados**	**Quantidade: Possível Evasor**	**Probabilidade de Evasão**
5657	8	1	12,5%
5669	8	1	12,5%
5670	8	1	12,5%
5673	8	1	12,5%
5680	8	1	12,5%
5684	8	1	12,5%
5688	8	1	12,5%
5690	8	1	12,5%
5687	8	1	12,5%

Tabela 36 – Relação de alunos de Engenharia Eletrônica com chances de evasão em 2018/1

Estudante	**Modelos Aplicados**	**Quantidade: Possível Evasor**	**Probabilidade de Evasão**
5699	8	1	12,5%
5702	8	1	12,5%
5707	8	1	12,5%
5708	8	1	12,5%
5711	8	1	12,5%
5714	8	1	12,5%
5713	8	1	12,5%

Tabela 37 – Relação de alunos de Engenharia Eletrônica com chances de evasão em 2018/2

Estudante	**Modelos Aplicados**	**Quantidade: Possível Evasor**	**Probabilidade de Evasão**
4259	8	1	12,5%
4267	8	1	12,5%
4277	8	1	12,5%
4279	8	1	12,5%

Tabela 38 – Relação de alunos de Engenharia Eletrônica com chances de evasão em 2019/1